

STABLE GENERATIVE MODELING USING SCHRÖDINGER BRIDGES

GEORG A. GOTTWALD, FENGYI LI, YOUSSEF MARZOUK, AND SEBASTIAN REICH

ABSTRACT. We consider the problem of sampling from an unknown distribution for which only a sufficiently large number of training samples are available. Such settings have recently drawn considerable interest in the context of generative modelling and Bayesian inference. In this paper, we propose a generative model combining Schrödinger bridges and Langevin dynamics. Schrödinger bridges over an appropriate reversible reference process are used to approximate the conditional transition probability from the available training samples, which is then implemented in a discrete-time reversible Langevin sampler to generate new samples. By setting the kernel bandwidth in the reference process to match the time step size used in the unadjusted Langevin algorithm, our method effectively circumvents any stability issues typically associated with the time-stepping of stiff stochastic differential equations. Moreover, we introduce a novel split-step scheme, ensuring that the generated samples remain within the convex hull of the training samples. Our framework can be naturally extended to generate conditional samples and to Bayesian inference problems. We demonstrate the performance of our proposed scheme through experiments on synthetic datasets with increasing dimensions and on a stochastic subgrid-scale parametrization conditional sampling problem as well as generating sample trajectories of a dynamical system using conditional sampling.

1. INTRODUCTION

Generative modeling is the process of learning a mechanism for synthesizing new samples that resemble those of the original data-generating distribution, given only a finite set of samples. It has seen wide adoption and enormous success across diverse application domains, from image [5, 24] and text generation [59, 32], to drug discovery [3, 2] and anomaly detection [12, 49], to name but a few.

In this paper, we introduce a new nonparametric approach to generative modeling that combines ideas from Schrödinger bridges [43, 7] and reversible Langevin dynamics [41]. Suppose that we are given M training samples $x^{(i)} \sim \pi$, $i = 1, \dots, M$, from an unknown distribution π on \mathbb{R}^d . Perhaps the simplest nonparametric approach to generative modeling is to build a kernel density estimate (KDE) and then sample from it; the KDE is essentially a mixture model with M components. Alternatively, one could estimate the score function $s(x) = \nabla \log \pi(x)$, without directly estimating $\pi(x)$, and use this estimate in the Langevin dynamics

$$(1) \quad \dot{X}_\tau = s(X_\tau) + \sqrt{2}\dot{W}_\tau,$$

where W_τ denotes standard d -dimensional Brownian motion [41]. There is a plethora of ways of estimating the score function [26, 53], and given an estimate for it, one needs to discretize (1), for example using Euler–Maruyama, to obtain an implementable scheme. However, the step size needs to be carefully chosen: a small step size leads to slow convergence, while too large a step yields instability of the numerical scheme,

especially for data that are supported on or are strongly concentrated about a compact sub-manifold of \mathbb{R}^d , e.g.,

$$(2) \quad \mathcal{M} = \{x \in \mathbb{R}^d : g(x) = 0\},$$

for some unknown function $g(x)$; a situation commonly encountered in high-dimensional data and referred to as the manifold hypothesis [14, 57] and which is known to be challenging for generative models based on (1) [52]. Any estimated score function, denoted here by $\hat{s}_M(x)$, will take large values for x with $\|g(x)\|^2 \gg 0$, rendering the Langevin dynamics (1) stiff. This implies that an explicit time integration method such as Euler–Maruyama will require very small step sizes $\Delta\tau > 0$.

Example 1.1. *In order to illustrate this point consider the singular Gaussian distribution on \mathbb{R}^2 with $x = (x_1, x_2) \in \mathbb{R}^2$ satisfying $x_1 \sim \mathcal{N}(0, 1)$ and $x_2 = 0$. Let us assume that the data samples $x^{(i)}$, $i = 1, \dots, M$, have been polluted by noise from a Gaussian in \mathbb{R}^2 with mean zero and covariance νI for $0 < \nu \ll 1$. A standard score estimator $\hat{s}_M(x)$ will lead to*

$$(3) \quad \hat{s}_\infty(x) = \begin{pmatrix} -\frac{1}{1+\nu}x_1 \\ -\frac{1}{\nu}x_2 \end{pmatrix}$$

in the limit $M \rightarrow \infty$ and an Euler–Maruyama discretization of the corresponding Langevin dynamics (1) will require step-sizes $\Delta\tau < 2\nu$, which become arbitrarily small as $\nu \rightarrow 0$. We will return to this simple problem in Example 3.1 where it will be used to demonstrate certain advantages of the methodology proposed in this paper and, in particular, that we can circumvent the just described computational bottleneck as $\nu \rightarrow 0$.

Motivated by this illustrative example, we follow an alternative approach in this paper. Namely, instead of first estimating the score function $s(x)$ and then discretising the corresponding approximation to (1) in time, we employ Schrödinger bridges [43, 7] to directly estimate the conditional expectation value

$$(4) \quad \mu(x; \epsilon) := \mathbb{E}[X_\epsilon | X_0 = x]$$

from the given samples $\{x^{(i)}\}_{i=1}^M$ for given parameter $\epsilon > 0$. We denote the data-driven estimator by $m(x; \epsilon) : \mathbb{R}^d \rightarrow \mathbb{R}^d$.

More specifically, we first note that $\mu(x; \epsilon) = \exp(\epsilon\mathcal{L}) \text{id}(x)$, where $\text{id} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes the identity map and \mathcal{L} the generator of the Langevin dynamics (1) [41]. We then approximate the semi-group $\exp(\epsilon\mathcal{L})$ from the given samples $\{x^{(i)}\}_{i=1}^M$ using a Schrödinger bridge approximation [35, 58], which optimally couples the empirical measure of the training samples with itself over an appropriate random walk reference process. By solving the Schrödinger bridge problem, we construct a transition matrix whose state space encompasses all the training points, which in turn defines $m(x; \epsilon)$ for all $x = x^{(i)}$, $i = 1, \dots, M$. We then extend this approximation beyond the training data to all $x \in \mathbb{R}^d$, which leads to the desired approximation $m(x; \epsilon)$.

The second key ingredient of our method is to interpret ϵ as a step size and to read off a Gaussian transition kernel from the Schrödinger bridge approximation in the form of

$$(5) \quad X_{n+1} = m(X_n; \epsilon) + \sqrt{\Sigma(X_n)} \Xi_n,$$

with appropriately defined diffusion matrix $\Sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ and $\Xi_n \sim \mathcal{N}(0, \epsilon I)$. Broadly, $m(X_n; \epsilon)$ controls the drift, while $\sqrt{\Sigma(X_n)} \Xi_n$ introduces noise. An obvious choice for $\Sigma(x)$ is $\Sigma(x) = 2I$, which corresponds to (1) and its Euler–Maruyama

discretisation with step-size $\Delta\tau = \epsilon$. In addition, we explore data-informed choices of $\Sigma(x)$ in (5).

Comparing to directly discretizing (1) using Euler–Maruyama with step-size $\Delta\tau = \epsilon$, we will demonstrate that the scheme (5) is stable and ergodic for all step-sizes $\epsilon > 0$ and, hence, ϵ can be chosen solely on accuracy considerations. We also introduce a novel split-step time-stepping scheme, which ensures that the generated samples lie in the convex hull of the training samples.

In addition, we replace the constant diffusion matrix $\Sigma(x) = 2I$ with a scaled matrix

$$(6) \quad \Sigma(x) = 2\rho(x)I$$

for given bandwidth $\rho(x) > 0$, which requires appropriate modifications to the Schrödinger bridges considered in [58]. As we demonstrate in our numerical experiments, the resulting sampling scheme (5) provides a better representation of the underlying target distribution. More precisely, we assess the quality of the generated samples using a variable bandwidth kernel and using a fixed bandwidth kernel on synthetic data sets drawn from non-uniform distributions supported on irregular domains and on low-dimensional manifolds.

We then extend our method to cover Bayesian inference problems with $\pi(x)$ as prior and to create a conditional generative model. These extensions allow us to perform Bayesian inference in the “simulation-based” setting, i.e., without explicit evaluation of a prior density and, in the case of conditional sampling, even without evaluations of the data likelihood. We demonstrate the performance of our conditional generative model for a stochastic subgrid-scale parametrization problem and for the generation of synthetic time series of dynamical systems.

1.1. Related work. Langevin dynamics (1) characterizes the motion of particles as they experience a blend of deterministic and stochastic forces. Unlike in this paper, it is typically assumed that the deterministic forcing term $\nabla \log \pi(x)$ is given. Langevin dynamics has become a popular tool for sampling data from the target distribution $\pi(x)$. One variation of this is to introduce a symmetric preconditioning operator to the Langevin dynamics and to consider reversible processes of the general form

$$(7) \quad \dot{X}_\tau = K(X_\tau)\nabla \log \pi(X_\tau) + \nabla \cdot K(X_\tau) + \sqrt{2K(X_\tau)}\dot{W}_\tau,$$

which samples from the distribution $\pi(x)$ for any symmetric positive definite matrix $K(x)$. We adopt here the Itô interpretation of the multiplicative noise term [41]. While (7) has originally been discussed in molecular physics [15, 13, 25], popular choices of $K(x)$ arising from computational statistics include the empirical covariance [16] and the Riemannian metric [17, 31], making this method converge faster and more geometry-aware, while leaving the stationary distribution unchanged. The scaled diffusion matrix (6) corresponds to the choice $K(x) = \rho(x)I$. Optimal choice of $K(x)$ in terms of convergence to equilibrium have recently been discussed in [30].

Our approximation of the semi-group $\exp(\epsilon\mathcal{L})$ relies on recent work on diffusion maps and an accelerated Sinkhorn algorithm [35, 58]. The Sinkhorn algorithm solves for the Markov transition kernel associated with a discrete Schrödinger bridge problem, where the coupling is between the empirical measure of the training samples with itself. This approach results in a symmetric bi-stochastic matrix that, notably, approximates the semi-group $\exp(\epsilon\mathcal{L})$ to higher accuracy in ϵ than standard diffusion map approximations [9, 8]. Separately, the idea of using variable bandwidth kernels can be found very early in the statistics community, for example, in the context of kernel density estimation [47, 54]. Recently, [1] replaces the original fixed bandwidth kernel with the variable bandwidth kernel in the construction of diffusion maps, making the

approximation of the generator accurate on unbounded domains. Inspired by this concept, we replace the fixed bandwidth kernel $\Sigma = 2I$ with a variable bandwidth kernel (6) for given $\rho(x) > 0$. The resulting Schrödinger bridge approximates the semi-group of the reversible Langevin diffusion process (7) with $K(x) = \rho(x)I$.

Several recent studies have combined a range of score function estimation techniques with Langevin dynamics. For example, [52] introduces a noise conditional score network to learn the score function and then uses annealed Langevin dynamics to generate samples. [4] studies the convergence rate of a Langevin based generative model, where the score is estimated using denoising auto-encoders. Such techniques are also studied within the Bayesian imaging community, commonly referred to as “plug and play” [29]. These approaches typically use neural networks for the estimation of the score function. The Schrödinger bridge sampler proposed in this paper is instead built upon a direct discrete-time approximation to (1). In addition, [29] uses an explicit projection to ensure that samples stay on the compact manifold given by (2) which is assumed to be explicitly known. In contrast, we do not assume any knowledge of \mathcal{M} .

We finally mention diffusion models or score generative models (SGM), which have been successfully used for generative modelling, in particular in image generation [23, 52]. These methods solve both a forward and a reverse stochastic differential equation (SDE). The forward SDE introduces noise to the sample, evolving the prior into the standard normal distribution, while the reverse SDE evolves samples from the standard normal distribution back to the original data distribution, yielding a different sample than the one initially fed into the forward SDE. During the training process, the score function is learned for the nonstationary distribution at each time. We mention that Schrödinger bridges have also been implemented in the context of SGMs in order to exactly couple the target distribution with the standard normal distribution [55, 11, 51, 6, 42]. Here we consider a completely different application of Schrödinger bridges; namely the estimation of conditional expectation values from training data.

1.2. Outline. In Section 2 we construct a Markov chain using a Schrödinger bridge approximation that samples from the given discrete data distribution. In Section 3, we extend this Markov chain to the continuous state space setting by constructing a Gaussian transition kernel which extracts its conditional mean and covariance matrix from the underlying diffusion map approximation. We introduce two discrete-time Langevin samplers; one with a data-unaware diffusion and one with a data-aware diffusion matrix in Section 3.1. Theoretical properties such as stability and ergodicity are discussed in Section 3.2. We further discuss the application of variable bandwidth kernels when constructing the Schrödinger bridge in Section 3.3. Here the goal is to approximate the tails of the reference distribution $\pi(x)$ better from the available training samples. In terms of practical applications, we explore the extension of our proposed scheme to a conditional sampling setting and Bayesian inference in Section 4. We demonstrate our proposed methods in Section 5 in a suite of examples. We start with a couple of synthetic examples demonstrating among others the benefit of variable bandwidth implementations. In terms of applications, we employ conditional sampling to provide a data-driven stochastic subgrid-scale parametrization for multi-scale systems. Furthermore, we show how our approach can be used to generate realistic synthetic trajectories of a dynamical system, given only a single time series for training. We conclude in Section 6 with a summary and an outlook.

2. DISCRETE SCHRÖDINGER BRIDGES

In this section, we collect some preliminary building blocks by considering the simpler task of building a discrete Markov chain over the samples $\{x^{(i)}\}_{i=1}^M$, which leaves

the associated empirical probability measure

$$(8) \quad \mu_{\text{em}}(\mathrm{d}x) = \frac{1}{M} \sum_{i=1}^M \delta_{x^{(i)}}(\mathrm{d}x)$$

in \mathbb{R}^d invariant. Here $\delta_x(\mathrm{d}x)$ denotes the Dirac delta distribution centred at x . In the subsequent section, we will generalise the finding from this section to approximately sample from $\pi(x)$, allowing for the generation of new samples which are distinct from the given training samples.

We consider the Schrödinger bridge problem of coupling $\mu_{\text{em}}(\mathrm{d}x)$ with itself along a reversible reference process with (unnormalized) transition probabilities

$$(9) \quad t_{ij} = \exp\left(-\frac{1}{2\epsilon}(x^{(i)} - x^{(j)})^\top \left(K(x^{(i)}) + K(x^{(j)})\right)^{-1} (x^{(i)} - x^{(j)})\right),$$

which we collect into a symmetric matrix $T \in \mathbb{R}^{M \times M}$. Here $\epsilon > 0$ is a tuneable parameter and $K(x)$ is a symmetric positive definite matrix for all $x \in \mathbb{R}^d$. Popular choices include $K = I$, $K = \Sigma_M$, where Σ_M is the empirical covariance matrix of the training samples $\{x^{(i)}\}_{i=1}^M$, and $K = \rho(x)I$, where $\rho(x) > 0$ is a scaling function representing variable bandwidth.

Instead of working with the empirical measure $\mu_{\text{em}}(\mathrm{d}x)$, we introduce the probability vector $p^* = (1/M, \dots, 1/M)^\top \in \mathbb{R}^M$ over $\{x^{(i)}\}_{i=1}^M$. Then 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

$$(10) \quad P = D(v)TD(v)$$

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

$$(11) \quad Pp^* = p^*.$$

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 bi-stochastic matrix \tilde{P} , which is related to (10) by $\tilde{P} = M^{-1}P$.

Given P , one can now construct a Monte Carlo scheme that samples from μ_{em} . Assume the Markov chain is currently in state $x^{(j)}$, then the transition probabilities to the next state $x \in \{x^{(i)}\}_{i=1}^M$ are given by

$$(12) \quad p_j = Pe_j \in \mathbb{R}^M,$$

where $e_j \in \mathbb{R}^M$ denotes the j -th unit vector in \mathbb{R}^M . Since all entries in P are bounded from below provided all samples satisfy $x^{(i)} \in \mathcal{C}$, where \mathcal{C} is a compact subset in \mathbb{R}^d , the constructed Monte Carlo scheme possesses a unique invariant measure given by p^* and is geometrically ergodic. The rate of convergence can be determined by the diffusion distance

$$(13) \quad d(x^{(i)}, x^{(j)}) = \|p_i - p_j\|^2.$$

If the diffusion distance is small, then $x^{(i)}$ and $x^{(j)}$ are well connected. Furthermore, if $d(x^{(i)}, x^{(j)})$ is small for all points, then the Markov chain will mix quickly. In particular, larger values of ϵ will lead to faster mixing.

However, the goal is to approximately sample from the underlying distribution $\pi(x)$ and not just the empirical distribution $\mu_{\text{em}}(\mathrm{d}x)$. The required extension of our baseline algorithm is discussed in the following section.

3. APPROXIMATING THE CONDITIONAL MEAN

In order to implement (5), we need to define $m(x; \epsilon)$ and $\Sigma(x)$ for any $x \in \mathbb{R}^d$. In this section, we discuss how one can obtain these functions from the training samples $\{x^{(i)}\}_{i=1}^M$ and the Markov chain approximation (10).

We introduce the vector-valued function $t(x) \in \mathbb{R}^M$ with entries

$$(14) \quad t_i(x) = \exp\left(-\frac{1}{2\epsilon}(x^{(i)} - x)^\top \left(K(x) + K(x^{(i)})\right)^{-1} (x^{(i)} - x)\right)$$

for $i = 1, \dots, M$. We then define the probability vector using the Sinkhorn weights, v , obtained in (10), i.e.,

$$(15) \quad p(x) = \frac{D(v)t(x)}{v^\top t(x)} \in \mathbb{R}^M$$

for all $x \in \mathbb{R}^d$. This vector gives the transition probabilities from x to training samples $\{x^{(i)}\}_{i=1}^M$ and provides a finite-dimensional approximation to the conditional probability distribution $\pi_\epsilon(\cdot|x)$ of the true underlying diffusion process; i.e., the semi-group $\exp(\epsilon\mathcal{L})$ with generator \mathcal{L} corresponding to the generalized reversible diffusion process (7).

Finally, using the probability vector $p(x)$, which implicitly depends on ϵ , and introducing the data matrix of training samples

$$(16) \quad \mathcal{X} = (x^{(1)}, \dots, x^{(M)}) \in \mathbb{R}^{d \times M},$$

our sample-based approximation of the conditional mean is provided by

$$(17) \quad m(x; \epsilon) := \mathcal{X}p(x).$$

Remark 3.1. *The construction of the conditional mean $m(x; \epsilon)$ is known as the barycentric projection of the entropy-optimally coupling [50, 44]. In optimal transport, v plays the role of the optimizer of the dual problem. We also mention the connection to de-noising schemes which utilize a structure similar to (17). See the recent review [40].*

3.1. Sampling algorithms. We now present our main Langevin sampling strategies based on the previously introduced probability vector $p(x)$ in (15). Sampling schemes of the form (5) have the same drift term (17), but differ in the way the diffusion matrix $\Sigma(x)$ is defined. We consider a data-unaware diffusion as well as a data-aware diffusion which turns out to be advantageous in generating new samples from the data distribution π (see the numerical experiments in Section 5).

3.1.1. Langevin sampler with data-unaware diffusion. Using $m_\epsilon(x)$, we propose the recursive sampler

$$(18) \quad X_{n+1} = X_n + \Delta\tau \left(\frac{m(X_n; \epsilon) - X_n}{\epsilon} \right) + \sqrt{2K(X_n)}\Xi_n$$

as an approximation to (7), where $\Delta\tau$ is the time step and $\Xi_n \sim \mathcal{N}(0, \Delta\tau I)$. If $K = I$, we obtain the score function approximation

$$(19) \quad s(x) = \frac{m(x; \epsilon) - x}{\epsilon}$$

in (1). Furthermore, by taking $\Delta\tau = \epsilon$, we have

$$(20) \quad X_{n+1} = m(X_n; \epsilon) + \sqrt{2K(X_n)}\Xi_n.$$

Note that (20) fits into the general formulation (5) with $\Sigma(x) = 2K(x)$.

Let us briefly discuss the qualitative behavior of the time-stepping method (20) as a function of $\epsilon > 0$. For large ϵ , the expected value $m(x; \epsilon)$ will become essentially independent of the current state X_n and the diffusion process will sample from a centred Gaussian. For $\epsilon \rightarrow 0$, on the other hand, the probability vector $p(x)$ can potentially degenerate into a vector with a single entry approaching one with all other entries essentially becoming zero. Hence a key algorithmic challenge is to find a good value for ϵ and a suitable $K(x)$, which guarantee both good mixing and accuracy, i.e., $X_n \sim \pi$ as $n \rightarrow \infty$.

In terms of initialisation, it is often best to initialise from one of the training data $x^{(j^*)}$ with $j^* \in \{1, \dots, M\}$ chosen at random.

3.1.2. Langevin sampler with data-aware diffusion. From (15) and (16), one can also define the scaled conditional covariance matrix,

$$(21) \quad C(x) = \epsilon^{-1}(\mathcal{X} - m(x; \epsilon)1_M^\top)D(p(x))(\mathcal{X} - m(x; \epsilon)1_M^\top)^\top \in \mathbb{R}^{d \times d},$$

which is the (scaled) covariance matrix associated with the probability vector $p(x)$. Here $1_M \in \mathbb{R}^M$ denotes the M -dimensional vector of ones. Therefore, one can more directly implement a Gaussian approximation associated with the transition probabilities $p(x)$ and introduce the update

$$(22) \quad X_{n+1} = X_n + \Delta\tau \left(\frac{m(X_n; \epsilon) - X_n}{\epsilon} \right) + \sqrt{C(X_n)}\Xi_n.$$

Similar to the previous case, setting $\Delta\tau = \epsilon$ implies

$$(23) \quad X_{n+1} = m(X_n; \epsilon) + \sqrt{C(X_n)}\Xi_n,$$

which we found to work rather well in our numerical experiments since it directly captures the uncertainty contained in the data-driven coupling P . The scheme (23) corresponds to setting $\Sigma(x) = C(x)$ in (5). Also note that the scheme (23) still depends on $K(x)$ through the probability vector $p(x)$.

3.2. Algorithmic properties. We briefly discuss several important properties on the stability and the ergodicity of the proposed Langevin samplers. The following Lemma establishes that, since each p is a probability vector, $m(x; \epsilon) = \mathcal{X}p(x)$ is a convex combination of the training sample $\{x^{(i)}\}_{i=1}^M$.

Lemma 3.1. *Let us denote the convex hull generated by the data points $\{x^{(i)}\}_{i=1}^M$ by \mathcal{C}_M . It holds that*

$$(24) \quad m(x; \epsilon) \in \mathcal{C}_M$$

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

Proof. The Lemma follows from the definition (17), which we write as $m(x; \epsilon) = \sum_{j=1}^M x^{(j)}p_j(x)$, and the fact that $p(x)$ is a probability vector with $0 \leq p_j(x) \leq 1$ for all $\epsilon > 0$ and all $x \in \mathbb{R}^d$. \square

This establishes stability of the Langevin samplers (20) and (23) for all step-sizes $\epsilon > 0$. The next lemma shows that the Langevin sampler (20) is geometrically ergodic.

Lemma 3.2. *Let us assume that the data generating density $\pi(x)$ has compact support. Then the time-stepping method (20) possesses a unique invariant measure and is geometrically ergodic provided the norm of the symmetric positive matrix $K(x)$ is bounded from above and below for all $x \in \mathbb{R}^d$.*

Proof. We consider $K(x) = I$ for simplicity. We introduce the Lyapunov function $V(x) = \|x\|^2$ and balls

$$(25) \quad \mathcal{B}_R = \{x \in \mathbb{R}^d : \|x\| \leq R\}$$

of radius $R > 0$ in \mathbb{R}^d . Since $m_\epsilon(X_n) \in \mathcal{C}_M$ and $\pi(x)$ has compact support, one can find radii $R_* > 0$ and $R > \sqrt{R_*^2 + 2\epsilon}$, which are independent of the training data $\{x^{(i)}\}_{i=1}^M$, such that $\mathcal{C}_M \subset \mathcal{B}_{R_*}$ and

$$(26) \quad \mathbb{E}[V(X_{n+1})|X_n] \leq \lambda V(X_n)$$

for all $X_n \notin \mathcal{B}_R$ and suitable $0 \leq \lambda < 1$. This follows from the fact that $\mathbb{E}[V(X_{n+1})|X_n] < R_*^2 + 2\epsilon$, while $V(x) > R^2$ for all $x \notin \mathcal{B}_R$. One then chooses $\lambda = (R_*^2 + 2\epsilon)/R^2 < 1$. Furthermore, there is a constant $\delta > 0$ such that

$$n(x'; m(x; \epsilon), 2\epsilon I) \geq \delta$$

for all $x, x' \in \mathcal{B}_R$. Here $n(x; m, \Sigma)$ denotes the Gaussian probability density function with mean m and covariance matrix Σ . In other words, \mathcal{B}_R is a small set in the sense of [39]. Geometric ergodicity follows from Theorem 15.0.1 in [39]. See also the self-contained presentation in [37]. \square

We note that extending Lemma 3.2 to the time-stepping scheme (23) with a data-aware diffusion is non-trivial since the covariance matrix (21) may become singular.

Lemma 3.3. *If $K(x) = I$, the conditional mean estimator $m(x; \epsilon)$ is equivalent to*

$$(27) \quad m(x; \epsilon) = x + \epsilon \nabla_x \log \Pi(x; \epsilon), \quad \Pi(x; \epsilon) := (v^T t(x))^2.$$

Hence, the update (20) is equivalent to an Euler–Maruyama discretization of Langevin dynamics (1) with modified probability density

$$(28) \quad \tilde{\pi}(x; \epsilon) := \frac{\Pi(x; \epsilon)}{\int \Pi(x; \epsilon) dx}.$$

In other words, for ϵ sufficiently small and $K(x) = I$, (20) samples approximately from $\tilde{\pi}(x; \epsilon)$.

Proof. Formula (27) follows from (15) and (14). \square

Lemma 3.1 suggests to replace the sampling step (20) by the associated split-step scheme

$$(29a) \quad X_{n+1/2} = X_n + \sqrt{2K(X_n)}\Xi_n,$$

$$(29b) \quad X_{n+1} = m(X_{n+1/2}; \epsilon).$$

This scheme now satisfies $X_n \in \mathcal{C}_M$ for all $n \geq 1$ and any choice of ϵ . Similarly, one can replace (23) by the split-step scheme

$$(30a) \quad X_{n+1/2} = X_n + \sqrt{C(X_n)}\Xi_n,$$

$$(30b) \quad X_{n+1} = m(X_{n+1/2}; \epsilon).$$

These split-step schemes have been used in our numerical experiments. We note that (29b) can be viewed as a de-noising step applied to a noisy $X_{n+1/2}$. See [40] for a recent survey of de-noising techniques used in image processing and their connection to SGM.

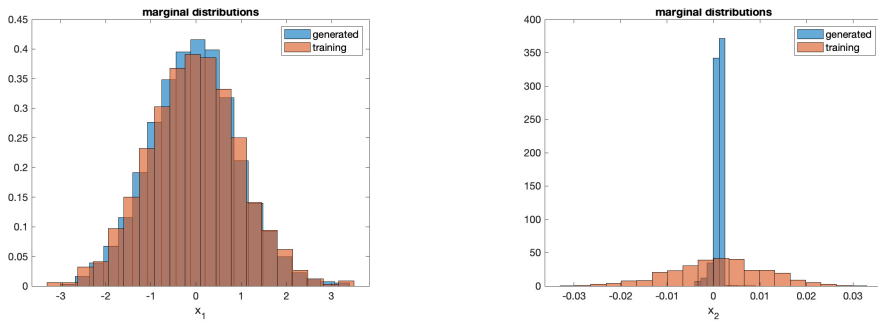


FIGURE 1. Histograms of the x_1 - and x_2 -components of the training as well as generated data for Example 3.1. One finds that the split-step scheme effectively denoises the x_2 -component while faithfully reproducing the standard normal distribution in the x_1 -component.

Example 3.1. We return to Example 1.1 and demonstrate the performance of the proposed split-step scheme (29) for this simple linear problem. We sample $M = 10^3$ training samples from the Gaussian distribution with mean zero and covariance matrix

$$(31) \quad C = \begin{pmatrix} 1 + \nu & 0 \\ 0 & \nu \end{pmatrix}$$

with $\nu = 10^{-4}$. The split-step scheme (29) is implemented with $\epsilon = 0.1$ and a total of 10^5 samples are generated. See Figure 1 for numerical results. It is found that the proposed scheme samples correctly from the normal distribution in x_1 and, at the same time, denoises the x_2 -component. We recall that a standard Euler–Maruyama discretisation would require step-sizes $\Delta\tau < 0.0002$ and would sample from the noisy distribution $N(0, C)$.

We finally discuss the computational complexity of the proposed algorithms in terms of the size, M , of the training data and the dimension, d , of the samples space \mathbb{R}^d . In the off-line phase of the algorithm, the accelerated Sinkhorn algorithm of [58] converges essentially in order $\mathcal{O}(M^0)$ iterations and each iteration requires matrix-vector multiplications; hence being of order $\mathcal{O}(M^2)$. Computing T involves calculating $\mathcal{O}(M^2)$ distances between vectors in \mathbb{R}^d . The online phase requires the multiplication of the $d \times M$ -dimensional data matrix \mathcal{X} with the M -dimensional weight vector $p(x)$. The computation of $p(x)$ in turn involves the computation of M distances $\|x - x^{(j)}\|^2$ and inner products of M -dimensional vectors. Overall the computational complexity of the online phase is of order $\mathcal{O}(dM)$.

The more problematic part is the accuracy of the Schrödinger bridge approximation to the semi-group $\exp(\epsilon\mathcal{L})$. Optimal scaling of ϵ as a function of M leads to an approximation error of order $\mathcal{O}(M^{-2/(8+d)})$, which requires an exponential increase of the sample size M as a function of the dimension, d , of sample space in order to reach a desired accuracy. While improvements in accuracy can be obtained by the variable bandwidth techniques discussed next for low dimensional problems, a localized formulation of the Schrödinger bridge sampler, as proposed in [22], can lead to an effectively dimension-independent accuracy.

3.3. Variable bandwidth diffusion. It is well-known from the literature on diffusion maps that a variable bandwidth can improve the approximation quality for fixed sample size M [1]. Here we utilize the same idea. However, we no longer insist on approximating the standard generator with $K = I$, since we only wish to sample from

the distribution π rapidly. Hence, we consider reversible diffusion processes (7) with

$$(32) \quad K(x) = \rho(x)I.$$

It is an active area of research to select a ρ that increases the spectral gap of the associated generator \mathcal{L} , given by

$$(33) \quad \mathcal{L}f = \pi^{-1}\nabla \cdot (\pi\rho\nabla f) = \nabla \cdot (\rho\nabla f) + \rho\nabla \log \pi \cdot \nabla f,$$

while not increasing computational complexity. A larger spectral gap implies a faster convergence rate [46], indicating that the generated samples are closer to the reference at a finite time, exhibiting a high accuracy. We demonstrate numerically in Section 5 that ρ can indeed be used to increase the sampling accuracy. More specifically, the bandwidth $\rho(x)$ is chosen as

$$(34) \quad \rho(x) = \pi(x)^\beta,$$

where $\beta \leq 0$ is a parameter and the unknown sampling distribution π is approximated by an inexpensive low accuracy density estimator [47, 54]. One finds that the variable bandwidth parameter β in (34) and the scaling parameter ϵ both influence the effective step-size in the Markov chain approximation (10) for (32). In order to disentangle the two scaling effects we modify the construction of the entries (9) in T_ϵ as follows. We first compute $\pi_i \approx \pi(x^{(i)})$ over all data points and then rescale these typically unnormalized densities:

$$(35) \quad \tilde{\pi}_i = Z^{-1}\pi_i, \quad Z := \frac{1}{M} \sum_j \pi_j.$$

The variable scaling length is then set to

$$(36) \quad \rho_i = \tilde{\pi}_i^\beta = Z^{-\beta}\pi_i^\beta$$

for $i = 1, \dots, M$, i.e., $K(x^{(i)}) = \rho_i I$ in (9) and, more generally,

$$(37) \quad K(x) = Z^{-\beta}\pi(x)^\beta.$$

The proposed scaling implies that a constant target density $\pi(x)$ leads to $K(x^{(i)}) = I$ in (9) regardless of the chosen β value. See Section 5 below for our numerical findings.

In order to derive an appropriate time-stepping scheme, we note that the drift term in the Itô formulation (7) with $K(x) = \rho(x)I$ can be expressed as

$$(38) \quad \mathcal{L} \text{id}(x) = \rho\nabla \log \pi + \nabla \rho$$

and, hence, it holds that

$$(39) \quad m(x; \epsilon) = \mathcal{X}p(x) \approx \exp(\epsilon\mathcal{L}) \text{id}(x) = \mu(x; \epsilon),$$

and

$$(40) \quad \exp(\epsilon\mathcal{L}) \text{id}(x) \approx x + \epsilon\mathcal{L} \text{id}(x) = x + \epsilon s(x)$$

as desired (cf. (19)). We emphasize that the drift now satisfies $s(x) = \rho(x)\nabla \log \pi(x) + \nabla \rho(x)$ and is no longer equivalent to the score $\nabla \log \pi(x)$. For the numerical variable bandwidth experiments conducted in Section 5, we therefore used (29) with $K(x) = \rho(x)I$. We remark that our aim is not to reproduce the actual dynamics of an underlying stochastic process, i.e. the process associated with the generator (33), but rather to sample from the distribution $\pi(x)$.

4. BAYESIAN INFERENCE AND CONDITIONAL SAMPLING

We note that the previously discussed sampling methods can easily be extended to reversible diffusion processes of the form

$$(41) \quad \dot{X}_\tau = -\nabla V(X_\tau) + \nabla \log \pi(X_\tau) + \sqrt{2}\dot{W}_\tau,$$

where $V(x)$ denotes a known potential such as the negative log-likelihood function in case of Bayesian inference with $\pi(x)$ taking the role of the prior distribution. We obtain, for example, the adjusted time-stepping scheme

$$(42) \quad X_{n+1/2} = X_n - \epsilon \nabla V(X_n) + \sqrt{2} \Xi_n,$$

$$(43) \quad X_{n+1} = m(X_{n+1/2}; \epsilon)$$

for given samples $\{x^{(i)}\}_{i=1}^M$ from an unknown (prior) distribution $\pi(x)$. This numerical scheme approximately samples from the invariant distribution

$$\tilde{\pi}(x) \propto e^{-V(x)} \pi(x).$$

Remark 4.1. *We briefly divert to optimization of a regularised minimization problem of the form*

$$(44) \quad x^* = \arg \min_x \{V(x) - \log \pi(x)\}$$

for given potential $V(x)$ and regulariser $-\log \pi(x)$. Again assuming that only samples $\{x^{(i)}\}_{i=1}^M$ of $\pi(x)$ are available, an algorithm for approximating x^* can be defined as follows:

$$(45a) \quad x_{n+1/2} = x_n - \epsilon \nabla V(x_n),$$

$$(45b) \quad x_{n+1} = m(x_{n+1/2}; \epsilon).$$

Here $m(x; \epsilon) = \mathcal{X}p(x)$ approximates the optimization update associated with the regulariser $-\log \pi(x)$. The iteration is stable and any limiting point x_∞ is contained in the convex hull of the data points $\{x^{(i)}\}_{i=1}^M$.

The sampling schemes (29) and (30) can furthermore be extended to conditional generative modeling. More specifically, consider a random variable $x = (y, z)$, which we condition on the first component for given $y = y^*$. We assume here that the decomposition of x into the parts z and y is given, and we wish to sample from $\pi(z|y^*)$ given samples $x^{(i)} = (y^{(i)}, z^{(i)})$, $i = 1, \dots, M$, from the joint distribution $\pi(x) = \pi(y, z)$.

In order to perform the required conditional sampling, we propose a method which combines approximate Bayesian computation (ABC) with our diffusion map based sampling algorithm. As before, we construct vectors of conditional expectation value $m(x; \epsilon)$ based on the samples $\{x^{(i)}\}_{i=1}^M$. We assume that the bandwidth parameter ϵ used in the diffusion map approximation is also applied in the ABC misfit function, i.e.,

$$(46) \quad L(y, y^*) = \frac{1}{2\epsilon} \|y - y^*\|^2.$$

This suggests the following split-step approximation scheme. Given the last sample $X_n = (Y_n, Z_n)$, we first update the y -component using

$$(47) \quad \hat{Y}_n = Y_n - \epsilon \nabla_y L(Y_n, y^*) = y^*.$$

In other words, we replace the current $X_n = (Y_n, Z_n)$ by $\hat{X}_n = (y^*, Z_n)$. Next we apply the split-step scheme (29) to \hat{X}_n , i.e.,

$$(48) \quad X_{n+1/2} = \hat{X}_n + \sqrt{2K(\hat{X}_n)}\Xi_n,$$

$$(49) \quad X_{n+1} = m(X_{n+1/2}; \epsilon),$$

where $m(x; \epsilon) = \mathcal{X}p(x)$ with $\mathcal{X} = (x^{(1)}, \dots, x^{(M)})$, and the definition of the probability vectors $p(x)$ follows from (15). The split-step scheme (30) generalises along the same lines.

We show in Section 5.3 how conditional sampling with the Schrödinger bridge enables drawing inaccessible random variables compatible with a known macroscopic state, a problem known as stochastic subgrid-scale parametrization. We further employ conditional sampling in Section 5.4 to draw random trajectories of a dynamical system where a future state is conditioned on the current state.

5. NUMERICAL EXPERIMENTS

In this section, we illustrate our method through three numerical examples encompassing different ranges and focal points. In the first two examples, we generate samples using synthetic datasets with increasing dimensions. Our emphasis is on exploring the impact of different diffusions and of employing a variable bandwidth kernel. In the third example, we showcase the proposed conditional generative modeling in Section 4, applied to a stochastic subgrid-scale parametrization problem. In the final example in Section 5.4, we employ conditional sampling to generate trajectories of the Lorenz-63 system [33] from previously computed training samples.

5.1. One-dimensional manifold. To illustrate how well the proposed methods generate statistically reliable samples we consider first the case of M samples $x \in \mathbb{R}^2$. In particular, we consider samples with a polar representation with radius $r = 1 + \sigma_r \xi_r$ and angle $\theta = \pi/4 + \sigma_\theta \xi_\theta$ with $\sigma_r = 0.06$ and $\sigma_\theta = 0.6$ and $\xi_{r,\theta} \sim \mathcal{N}(0, 1)$. We used $M = 2,000$ samples to learn the transition kernel and then generated 10,000 new samples with an initial condition from the data-sparse tail of the distribution, chosen to be the data point corresponding to the smallest angle.

We begin by investigating the effect of the two noise models proposed, namely a constant diffusion as in (20) with constant bandwidth $K(x) = I$ and the case when the diffusion reproduces the sample covariance C as in (23). In both cases we use a constant bandwidth $K = I$ in (9) and (14) when forming the Schrödinger bridge. We employ a Langevin sampler with the splitting scheme (29) with $\epsilon = 0.009$ and (30), respectively. Figure 2 shows that choosing the sample covariance as the noise model is clearly advantageous. Whereas both noise models generate samples that reproduce the angular distribution the noise model using a constant diffusion is overdiffusive in the radial direction. In contrast the noise model using the sample covariance nicely reproduces the radial distribution.

We now investigate the effect of a variable bandwidth $K(x)$. We employ the noise model (30) with the sample covariance but use $K(x)$ to determine the diffusion map (cf. (14)). The Langevin sampler (30) is again initialized with the coordinates of the data point corresponding to the smallest angle in the data-sparse tail. Figure 3 (left) shows the samples projected onto the convex hull of the data, i.e. outputs of step (30b), when a uniform bandwidth $K(x) = I$ with $\epsilon = 0.009$ is employed. Although the mean behaviour is well reproduced, it is seen that the generative model fails near the data-sparse tails for large and small angles. Here the value of ϵ is too small to generate significant diffusion and the samples are aligned on the (linear) convex

hull of the widely separated data samples. To mitigate against this behaviour we employ a variable bandwidth $\rho(x) = \pi^\beta$ with $\beta = -1/5$ and a kernel density estimate $\pi(x)$. Figure 3 (right) shows how the variable bandwidth kernel better reproduces the distribution in the data-sparse tail regions. Figure 4 shows the empirical histograms for the radius and the angle variables of the noisy samples corresponding to Figure 3 (i.e. outputs of step (30a)). Whereas both, the uniform and the variable bandwidth kernels, reproduce the radial distribution very well, the uniform bandwidth fails to reproduce the angular distribution in the tails where the diffusion is not sufficiently strong to allow for efficient mixing and the diffusion process gets stuck in the data sparse region.

We have seen that a constant uniform bandwidth generates samples which are concentrated in the bulk of the data and which are overly diffusive in the radial direction (cf. Figure 2 (right)). One may wonder if employing a smaller virtual time step $\Delta\tau < \epsilon$ in the Langevin sampler (18) will allow a Langevin sampler with constant bandwidth to generate more faithful samples. Figure 5 shows that choosing a smaller time step $\Delta\tau$ in (22), here with $\Delta\tau = \epsilon/4$ is indeed able to reproduce the radial distribution. However, if the Langevin sampler is initialised with a data point in the center of the data samples, it is not able to diffuse into the tail of the distribution distribution, leading to an under-diffusive empirical histogram for the angles.

The numerical experiments above suggest that we employ a noise model using the sample covariance C combined with a variable bandwidth $K(x)$ to control eventual data sparse regions. When employing a variable bandwidth our method contains two hyper-parameters which require tuning: the bandwidth factor ϵ and the exponent β in the arbitrary choice of the variable bandwidth $\rho(x)$. Their role will be explored in the following subsection.

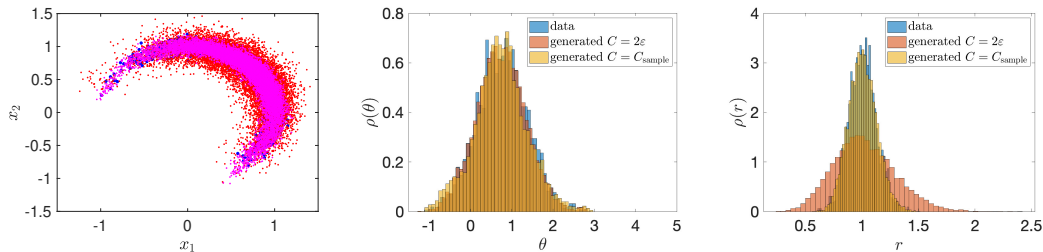


FIGURE 2. Comparison of the different noise models employed by the generative model. We employed a constant bandwidth with $\epsilon = 0.009$. Left: Original (blue) and generated data using a constant covariance (red) and the sample covariance $C(x)$ (magenta). Middle: Empirical histograms of the angular variable θ . Right: Empirical histograms of the radial variable r .

5.2. Multi-dimensional manifolds. In this numerical example, we show our proposed method on hyper semi-spheres of dimension $d = \{3, 4, 9\}$, using both a fixed bandwidth kernel and a variable bandwidth kernel. Data are generated by firstly sampling $z^{(i)} = (z_1^{(i)}, \dots, z_d^{(i)})$ from a d -dimensional standard normal distribution, and then setting $y^{(i)} = (z_1^{(i)}, \dots, \alpha z_d^{(i)})$, with $\alpha = 5$ to promote non-uniformity. Finally, the samples $x^{(i)}$ are obtained by normalizing $y^{(i)}$ to achieve the unit length, i.e., $y^{(i)} / \|y^{(i)}\|$, and perturbing $y^{(i)} / \|y^{(i)}\|$ in the radial direction with $U(0, 0.01)$ noise. An instance of the target samples of three dimensions can be seen in Figure 6. Given a bandwidth

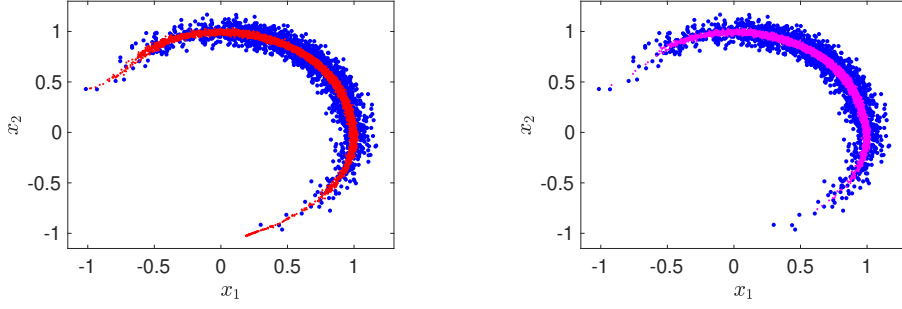


FIGURE 3. Effect of a variable bandwidth $K(x) = \rho(x)I$ in data-sparse regions. For the generative model the Langevin sampler (30) is used and we set $\epsilon = 0.009$. Results are shown for the output of step (30b). Left: Original (blue) and generated data for a constant bandwidth $K(x) = I$ (red). Right: Original (blue) and generated data for a variable bandwidth $K(x) = \rho(x)I$ with $\rho(x) = \pi(x)^\beta$ with $\beta = -1/5$ (magenta).

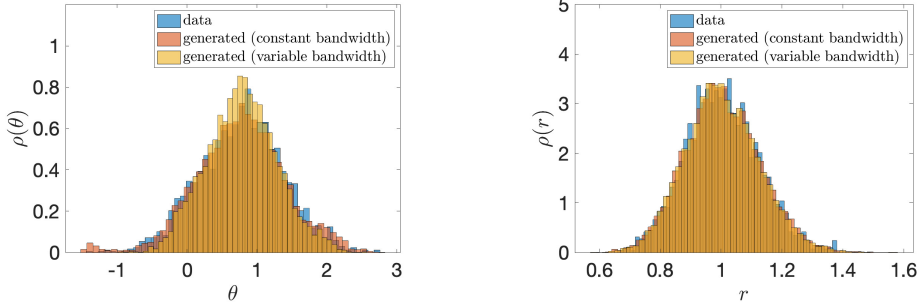


FIGURE 4. Effect of a variable bandwidth $K(x) = \rho(x)I$ on the angular and radial distributions (left and right, respectively). Shown are the original data (blue), generated data for a constant bandwidth $K(x) = I$ (red) and for a variable bandwidth $K(x) = \rho(x)I$ with $\rho(x) = \pi(x)^\beta$ with $\beta = -1/5$. The data were generated using a constant covariance noise model in (30) and $\epsilon = 0.009$.

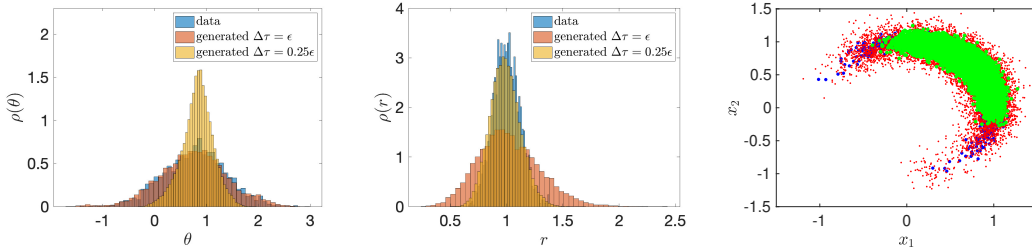


FIGURE 5. Effect of a variable time step $\Delta\tau$ in the Langevin sampler (29) with constant diffusion $K = 1$. Results are shown for the original data, and for $\Delta\tau = \epsilon$ and $\Delta\tau = \epsilon/4$. Throughout a constant bandwidth is used. Left: Empirical histogram of the angular variable θ . Middle: Empirical histograms of the radial variable r . Right: Original (blue) and generated data in the (x_1, x_2) -plane with $\Delta\tau = 1$ (red) and with $\Delta\tau = \epsilon/4$ (green).

ϵ and a bandwidth function $\rho(x)$, we implement the proposed scheme (30). For the fixed bandwidth kernel we set $\rho(x) = 1$. For the variable bandwidth kernel (37), we set $\rho(x) = \pi(x)^\beta$, with $\beta < 0$, and $\pi(x)$ is approximated using a kernel density estimator. We use $M = 1,000$ training samples to learn the transition kernel and run a Langevin sampler to generate 50,000 samples, with the initial data point being $(1, 0, \dots, 0) \in \mathbb{R}^d$. To obtain a better mixing of the Langevin sampler, we take one every 20 samples of the last 20,000 generated samples of the chain, resulting in a total of 1,000 samples. To evaluate the quality of the generated samples, we compute the regularized optimal transport (OT) distance between the generated samples $\{x_{\text{gen}}^{(i)}\}_{i=1}^{M_g}$ and the original reference samples $\{x_{\text{ref}}^{(j)}\}_{j=1}^{M_r}$. The regularized OT distance with entropy penalty $1/\lambda$ is defined as

$$d_\lambda(x_{\text{gen}}, x_{\text{ref}}) = \min_P \sum_{i,j} P_{ij} C_{ij} - \frac{1}{\lambda} h(P),$$

subject to the constraints that

$$\sum_{j=1}^{M_r} P_{ij} = M_g^{-1}, \quad \sum_{i=1}^{M_g} P_{ij} = M_r^{-1},$$

where

$$h(P) = - \sum_{i,j} P_{ij} \log P_{ij}$$

is the information entropy. The entries of $C \in \mathbb{R}^{M_g \times M_r}$ are set to be the pairwise Euclidean distances between $\{x_{\text{gen}}^{(i)}\}_{i=1}^{M_g}$ and $\{x_{\text{ref}}^{(j)}\}_{j=1}^{M_r}$, and each sample is assigned equal weight marginally. We compute the OT distance using the Sinkhorn–Knopp algorithm [10, 28]. The number of reference samples is chosen to be $M_r = 50,000$, and the entropic regularization penalty is set to be $1/\lambda = 100$. We consider the OT distance as a diagnostic to quantify the statistical accuracy of the sampling scheme.

We then optimize over the parameters ϵ for a fixed λ using grid search. To be more precise, for the Langevin sampler with a fixed bandwidth kernel, we vary ϵ , and compute the OT distance of the generated samples. The best performed ϵ is chosen to be the one that corresponds to the smallest OT distance, and we call it ϵ^* . For the variable bandwidth kernel, we fix $\epsilon = \epsilon^*$. In order to disentangle the effect of varying ϵ and varying β we use the normalized variable bandwidth as described in (37). We set $\beta = -0.01 \times 2^n$ with $n = \{0, \dots, 8\}$ for $d = \{3, 4, 9\}$. The optimal bandwidth ϵ^* is reported in Table 1 and the comparisons between the fixed bandwidth kernel and the variable bandwidth kernel are presented in Figure 7. We observe that by keeping ϵ fixed, the OT distance becomes smaller for a wide range of β .

We then examine the quality of generated samples at the optimal ϵ and β along the last coordinate (the nonuniform direction). Similar to the previous study, we compute the one dimensional OT distance of the marginal distribution (see Figure 7 (right)) and show the histograms and the cumulative density function (CDF) of the samples generated using the fixed bandwidth kernel and using the variable bandwidth kernel in Figure 8. The benefit of using the variable bandwidth kernel becomes more prominent when focusing on the marginal samples. In the case where we sample a 4-dimensional hyper-semisphere with non-uniformity along the last coordinate, we see in Figure 8 that the empirical CDF of the samples generated with the variable bandwidth kernel closely aligns with the reference (constructed using samples from the target distribution) for the most part. In contrast, the samples generated using the fixed variable bandwidth kernel noticeably diverge from the reference. This aligns

d	optimal bandwidth
3	$\epsilon^* = 0.008$
4	$\epsilon^* = 0.010$
9	$\epsilon^* = 0.050$

TABLE 1. Optimal bandwidth parameters leading to minimal OT distance for different dimensions d , obtained using grid search.

with what we observed in Figure 3 and Figure 4 — the data generated using the variable bandwidth kernel better resemble the original data. In the cases where the samples are drawn from a 9-dimensional hyper-semisphere, while both methods struggle to generate samples that mirror those from the target distribution, primarily due to the inherent limitations of kernel methods in high-dimensional scenarios, employing a variable bandwidth kernel produces samples that exhibit a closer resemblance to those from the target distribution.

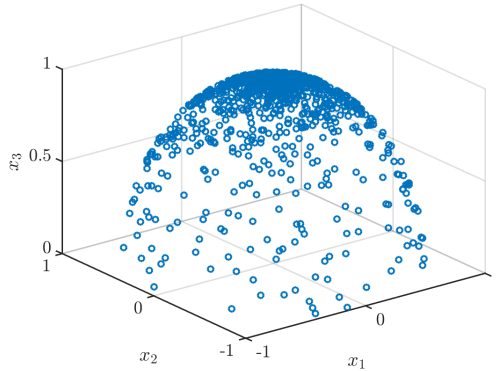


FIGURE 6. Nonuniform samples $x^{(i)}$ on a 2-dimensional semisphere. The non-uniformity is along the last coordinate x_3 .

5.3. Stochastic subgrid-scale parametrization. The conditional sampling algorithm described in Section 4 can be used to perform stochastic subgrid-scale parametrization, a central problem encountered in, for example, the climate sciences. The problem of subgrid-scale parametrization, or more generally of model closure, is the following: given a potentially stiff dynamical system

$$(50a) \quad \dot{z} = F_z(z) + g(z, y; \varepsilon)$$

$$(50b) \quad \dot{y} = F_y(z, y; \varepsilon),$$

where $\varepsilon < 1$ denotes the time scale separation between the slow resolved variables of interest $z \in \mathbb{R}^{d_s}$ and the unresolved fast degrees $y \in \mathbb{R}^{d_f}$. Note the notational difference between the time scale separation parameter ε and the bandwidth parameter ϵ used to define the diffusion map. For $\varepsilon \ll 1$ the system is stiff and to ensure numerical stability a small time step $\Delta t < \varepsilon$ is needed. This, together with the potential high-dimensionality of the fast subspace $d_f \gg d_s$, constitutes a computational barrier for simulating the dynamics (50a)–(50b) on the slow time scale of interest. Hence one is interested in obtaining an effective evolution equation for the slow resolved variables z

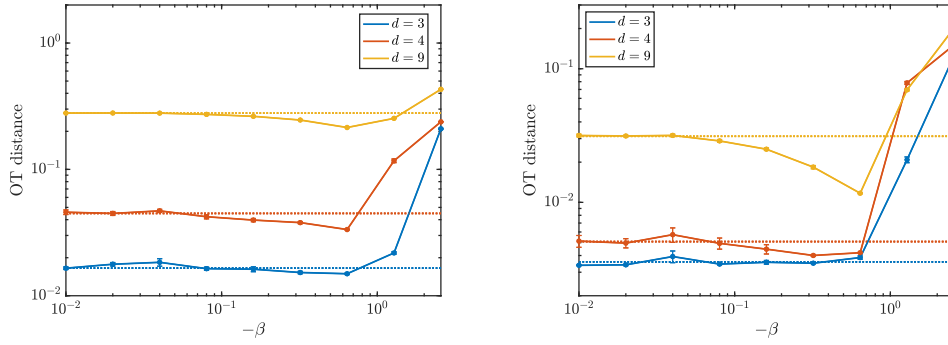


FIGURE 7. The straight dashed lines denote the OT distances of the samples using a fixed bandwidth kernel, and the solid lines denote the OT distances of the samples using a variable bandwidth kernel. Left: comparison between the OT distance of samples generated using a fixed bandwidth kernel and using a variable bandwidth kernel. Right: comparison between the *one-dimensional* marginal OT distance of samples generated using a fixed bandwidth kernel and using a variable bandwidth kernel, along the last coordinate (non-uniform direction). Here the Langevin sampler is initialized at $(1, 0, \dots, 0)$ for all cases.

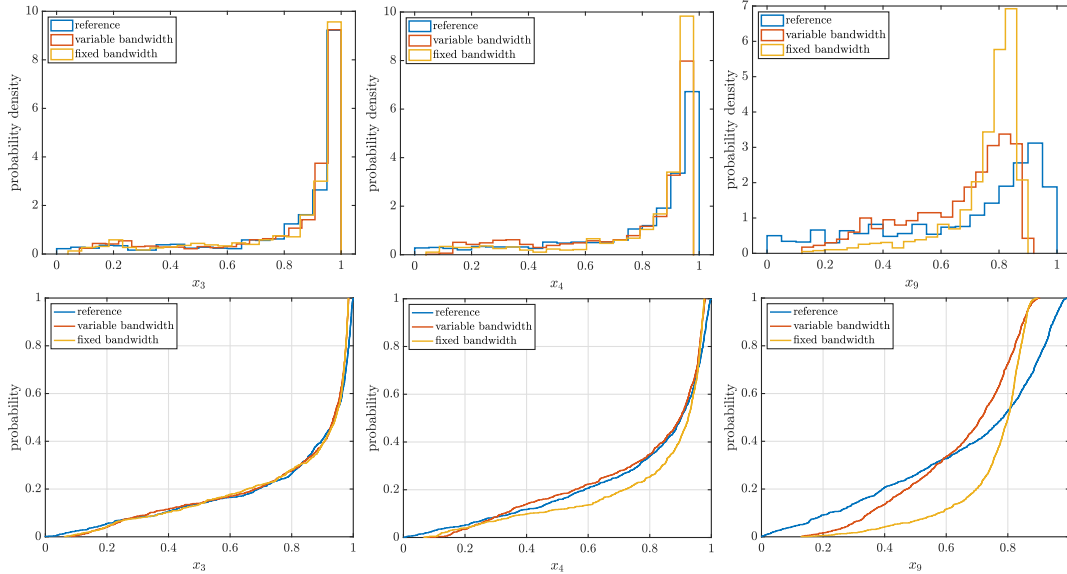


FIGURE 8. Comparisons of empirical histograms (top row) and CDFs (bottom row) of the marginal distribution of the generated samples along the last coordinate. From left to right: the data are sampled from a $\{3, 4, 9\}$ -dimensional (hyper-)semisphere.

only which captures the essential effect of the unresolved variables y . Hence, we seek to determine the effective reduced dynamics

$$\dot{z} = F_z(z) + \psi(z).$$

Here $F_z(z)$ denotes a deterministic drift which we assume to be known *a priori*, possibly based on physical reasoning. The term $\psi(z)$ denotes the unknown closure term to be learned which may be deterministic or stochastic, and which parametrizes the unknown

unresolved fast processes. Deterministic machine learning methods have previously been used to learn $\psi(z)$ as the average effect of the unresolved variables, i.e. the average of $g(z, y; \varepsilon)$ over the (conditional) invariant measure of the fast process [20, 21]. Deterministic maps, however, are not able to capture the resolved dynamics with sufficient statistical accuracy, and it is by now well established that the effective equation is often of a stochastic nature [38, 19, 27, 18]. In the case of infinite time scale separation there are explicit expressions for the effective drift and diffusion term of the effective slow dynamics. However, these terms include integrals over auto-correlation functions and are notoriously hard to estimate. Instead we propose to learn the closure term $\psi(z)$ and generate realisations $\psi(z)$ on the fly employing the conditional sampling algorithm described in Section 4. We consider the situation in which scientists have a good understanding of the resolved dynamics and know the slow vector field $F_z(z)$. Given data of the resolved variables $\{z^{(i)}\}_{i=0}^M$ sampled at equidistant times Δt , the associated closure term $\psi^{(i)}$ capturing the effect of the unresolved dynamics (50b) can then be estimated as

$$\psi^{(i)} := z^{(i)} - z^{(i-1)} - F_z(z^{(i-1)}) \Delta t,$$

for $i = 1, \dots, M$. This produces the training samples $\{x^{(i)} = (z^{(i-1)}, \psi^{(i)})\}_{i=1}^M$ which we collect in the $2d_s \times M$ data matrix \mathcal{X} .

The effective dynamics is then provided by the discrete stochastic surrogate model

$$(51) \quad z_k = z_{k-1} + F_z(z_{k-1}) \Delta t + \psi_k$$

for $k \geq 1$ and given z_0 , where the subgrid-scale terms ψ_k are generated as follows: Given the conditional mean $m(x; \varepsilon)$ as described in Section 3, we perform the discrete Langevin sampler

$$(52a) \quad \hat{X}_n = (z_{k-1}, P_2 X_n)$$

$$(52b) \quad X_{n+1/2} = \hat{X}_n + \sqrt{2} \Xi_n$$

$$(52c) \quad X_{n+1} = m(X_{n+1/2}; \varepsilon),$$

for $n = 1, \dots, n_s$ with $n_s = 100$ and $\Xi_n \sim \mathcal{N}(0, \varepsilon I)$. Here $P_2 : \mathbb{R}^{2d_s} \rightarrow \mathbb{R}^{d_s}$ is a projector with $P_2 x = \psi$. We finally set $\psi_k = P_2 X_{n_s}$. The assignment of the first component of \hat{X}_n to z_{k-1} ensures the conditioning of ψ_k on z_{k-1} , and $n_s = 100$ ensures that the generated samples ψ_k are close to independent from ψ_{k-1} . We choose a fixed bandwidth with $\varepsilon = 0.001$.

We consider here the particular example with $d_s = 1$ and $d_f = 3$ given by

$$(53) \quad \dot{z} = z(1 - z^2) + \frac{4}{90\varepsilon} h(z) y_2,$$

where the fast dynamics is given by the Lorenz-63 system [33]

$$(54a) \quad \varepsilon^2 \dot{y}_1 = 10(y_2 - y_1)$$

$$(54b) \quad \varepsilon^2 \dot{y}_2 = 28y_1 - y_2 - y_1 y_3$$

$$(54c) \quad \varepsilon^2 \dot{y}_3 = -\frac{8}{3} y_3 + y_1 y_2.$$

We look at the case of effective additive noise with $h(z) = 1$ which does not require conditioning on the slow variable z , as well as at the case of multiplicative noise with $h(z) = z$. We used MATLAB's built-in ode45 routine [36] to generate the time series with a time-scale separation parameter of $\varepsilon = 0.01$. The time series is subsequently sub-sampled with $\Delta t = 0.1$. Figure 9 shows a comparison between the full multi-scale system (53)–(54) with additive noise $h(z) = 1$ and the stochastic surrogate

model (51). The slow z -dynamics exhibits stochastic bimodal dynamics. Figure 10 shows a comparison for the multiplicative case $h(z) = z$ which yields unimodal slow dynamics. It is clearly seen that the surrogate model (51) obtained by the generative conditional sampler generates statistically reliable dynamics.

We remark that the above proposed method to draw random samples from the closure term does not require the knowledge of the time-scale separation parameter ε . The closure term ψ only depends on the time series $\{z^{(i)}\}_{i=0}^M$ of the resolved variables and the knowledge of the slow vector field F_z (which may be zero).

In the next subsection we show that one can indeed draw full sample trajectories of a dynamical system even without any time-scale separation using the conditional Schrödinger bridge sampler.

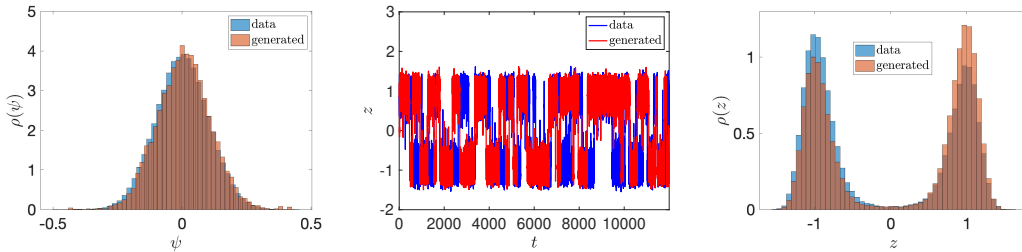


FIGURE 9. Results for the stochastic subgrid-scale parametrization for the multi-scale system (53)–(54) with $\varepsilon = 0.01$ and with additive noise $h(z) = 1$. Shown are results obtained by integrating the full multi-scale system and by the stochastic subgrid-scale parametrization scheme using our generative sampler (52) trained with $M = 120,000$. Left: Empirical histograms of the closure term $\psi = \psi(z)$. Middle: Time series of the slow variable $z(t)$. Right: Empirical histograms of the slow variable z .

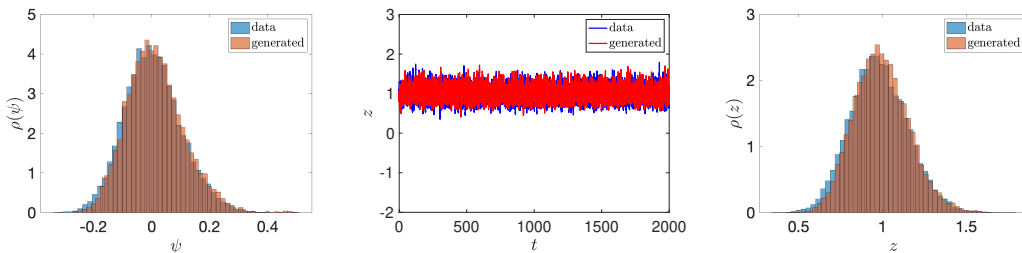


FIGURE 10. Results for the stochastic subgrid-scale parametrization for the multi-scale system (53)–(54) with $\varepsilon = 0.01$ and with multiplicative noise $h(z) = z$. Shown are results obtained by integrating the full multi-scale system and by the stochastic subgrid-scale parametrization scheme using our generative sampler (52), trained with $M = 20,000$. Left: Empirical histograms of the closure term $\psi = \psi(z)$. Middle: Time series of the slow variable $z(t)$. Right: Empirical histograms of the slow variable z .

5.4. Generative modelling of dynamical systems. We now employ the Schrödinger bridge sampler to perform sequential conditional sampling and generate typical trajectories from a dynamical system. We consider the Lorenz-63 system [33] for $y =$

(y_1, y_2, y_3) given by (54) with $\varepsilon = 1$. Given a long time trajectory $\{y^{(i)}\}_{i=0}^M$ with equidistantly sampled data points $y^{(i)} = y(t_i)$ with $t_{i+1} - t_i = \Delta t$ for all i , we construct the Schrödinger bridge as a coupling for the joint probability function $\pi(y(t), y(t + \Delta t))$ from the M training pairs $x^{(i)} = (y^{(i-1)}, y^{(i)}) \in \mathbb{R}^6$. Our aim is to generate new trajectories $\{y_k\}_{k \geq 0}$ in time intervals of Δt for given $y_0 = y(t_0)$ and $y_1 = y(t_1)$ sequentially, similarly to the generative modelling of texts: Given $x_{k-1} = (y_{k-1}, y_k)$, generate a new $x_k = (y_k, y_{k+1})$, which is not part of the initial training samples. Obviously the first component of x_k needs to be conditioned on the second component of x_{k-1} . As in Section 5.3, to decorrelate we perform $n_s = 20$ Langevin sampling steps

$$(55a) \quad \hat{X}_n = (y_k, P_2 X_n)$$

$$(55b) \quad X_{n+1/2} = \hat{X}_n + \sqrt{2} \Xi_n$$

$$(55c) \quad X_{n+1} = m(X_{n+1/2}; \epsilon),$$

for given y_k with $\Xi_n \sim N(0, \epsilon I)$. Here $P_2 : \mathbb{R}^6 \rightarrow \mathbb{R}^3$ denotes the project onto the second three components of $x \in \mathbb{R}^6$. We finally set $y_{k+1} = P_2 X_{n_s}$.

In Figure 11 we show that the conditional sampling produces trajectories which resemble those of the actual Lorenz-63 system, as well as having the same asymptotic statistical behaviour as seen by the reconstruction of the famous butterfly attractor. We choose a fixed bandwidth with $\epsilon = 0.05$, and used 10,000 training data $x(t_n)$ with $\Delta t = 0.1$.

The proposed approach can be compared to a direct approximation of the time- Δt -propagator

$$(56) \quad y_{n+1} = \Psi(y_n)$$

via, for example, a random feature map approximation as proposed in [20]. More specifically, using a smaller observation interval of $\Delta t = 0.02$ and noisy data, highly accurate predictions were achieved in [20] using $D_r = 300$ randomly chosen feature maps. The output weights $W \in \mathbb{R}^{3 \times D_r}$ were learned using the ensemble Kalman filter [45]. In comparison, the proposed Schrödinger bridge sampler requires a larger training set ($M = 10,000$ versus $M = 4,000$ in [20]), but involves only a single tuning parameter ϵ and repeated computation of (17).

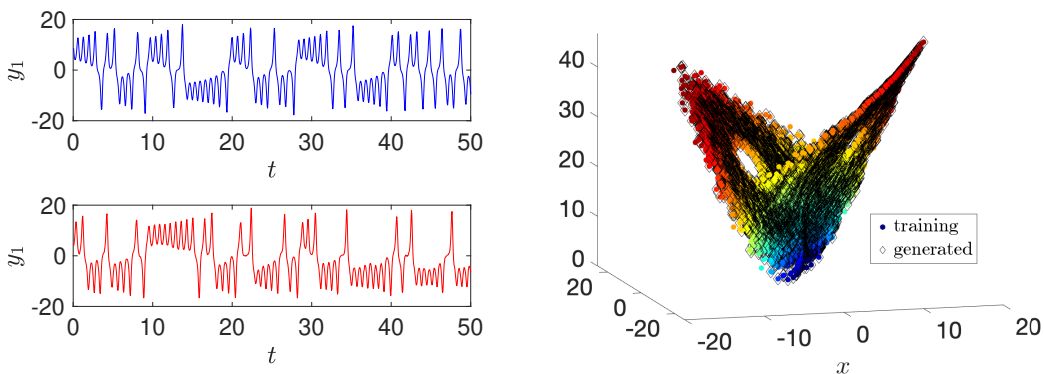


FIGURE 11. Comparison of generated data from the Schrödinger bridge sampler with the original dynamics of the Lorenz-63 system (54). Left: Typical trajectories of y_1 obtained from a simulation of (54) (top) and the generated data using conditional sampling (55) (bottom). Right: The corresponding attractor for the simulated and generated data.

6. SUMMARY AND OUTLOOK

We have introduced a Schrödinger bridge based Langevin scheme for generative modeling. Our method combines sample-based Markov chain approximations with discrete-time Langevin-type sampling, yielding a non-parametric generative model that is unconditionally stable and geometrically ergodic. Even though the approach is entirely data-driven, the generative model possesses a single tuning parameter, which is the choice of the step-size $\epsilon > 0$. Practical choices of ϵ depend both on the number of training samples, M , and the properties of the target distribution $\pi(x)$ on \mathbb{R}^d . We showed numerically that employing a variable bandwidth kernel, in contrast to a fixed bandwidth kernel, results in generated samples with enhanced accuracy. However, employing variable a bandthwith kernel comes with an added layer of complexity and requires delicate tuning.

In terms of practical applications, the performance of the conditional generative model was showcased through its applications to a stochastic subgrid-scale parametrization problem and the generation of trajectories of the chaotic Lorenz-63 system. In both cases, the conditional sampler relied on the availability of a single simulated trajectory, which served as the training samples. A constant bandwith kernel has been employed in all these examples.

We considered in this paper only equally weighted training samples $x^{(i)} \sim \pi(x)$. If there is a change of measure due to, for example, observed data with a given likelihood, then the measure and the couplings can be appropriately modified to take into account the non-uniform measure, i.e, weighted training samples.

Furthermore, all computational examples considered in this paper were low-dimensional, which is intrinsic to Schrödinger bridge approximations of the semi-group $\exp(\epsilon\mathcal{L})$ [58]. This could be seen as a major disadvantage in comparison to the sophisticated deep neural network architectures typically used in high-dimensional SGMs [23, 52]. A first step to close this gap has been taken in [22], where a localized Schrödinger bridge sampler has been proposed and implemented for high-dimensional generative sampling problems. Interestingly, and worth further investigation, the localized Schrödinger bridge sampler can be viewed as a variant of multi-head self attention [56, 48].

Additional future research will delve into the theoretical foundations of the proposed scheme, including its convergence rate and scalability extending results from [58]. Finally, the proposed methodology can be extended to score generative modeling by replacing all required score functions by Schrödinger bridge-based approximations and to interacting particle approximations of the Fokker–Planck equation, where the score function represents now diffusion [34].

Acknowledgements. This work has been funded by Deutsche Forschungsgemeinschaft (DFG) - Project-ID 318763901 - SFB1294. GAG acknowledges funding from the Australian Research Council, grant DP220100931. FL and YMM acknowledge support from the US Department of Energy, Office of Advanced Scientific Computing Research, award DE-SC0023188. We thank Ricardo Baptista for pointing out the relationship to the barycentric projection.

REFERENCES

- [1] T. Berry and J. Harlim. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 40:68–96, 2016.
- [2] Y. Bian and X.-Q. Xie. Generative chemistry: drug discovery with deep learning generative models. *Journal of Molecular Modeling*, 27:1–18, 2021.

- [3] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12:e1608, 2022.
- [4] A. Block, Y. Mroueh, and A. Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. Technical report, arXiv:2002.00107, 2020.
- [5] R. Cai, G. Yang, H. Averbuch-Elor, Z. Hao, S. J. Belongie, N. Snavely, and B. Hariharan. Learning gradient fields for shape generation. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 364–381. Springer, 2020.
- [6] T. Chen, G.-H. Liu, and E. Theodorou. Likelihood training of Schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nioAdKCEdXB>.
- [7] Y. Chen, T. T. Georgiou, and M. Pavon. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Review*, 63(2): 249–313, 2021.
- [8] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.
- [9] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005.
- [10] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [11] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709. Curran Associates, Inc., 2021.
- [12] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft. Image anomaly detection with generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 3–17. Springer, 2019.
- [13] D. L. Ermak and J. McCammon. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.*, 69:1352–1360, 1978.
- [14] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *J. Amer. Math. Soc.*, 29(4):983–1049, 2016. doi: 10.1090/jams/852. URL <https://doi.org/10.1090/jams/852>.
- [15] M. Fixman. Simulation of polymer dynamics. I. General theory. *J. Chem. Phys.*, 69:1527–1537, 1978.
- [16] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19:412–441, 2020.
- [17] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:123–214, 2011.

- [18] G. Gottwald, D. Crommelin, and C. Franzke. Stochastic climate theory. In C. L. E. Franzke and T. J. O’Kane, editors, *Nonlinear and Stochastic Climate Dynamics*, pages 209–240. Cambridge University Press, Cambridge, 2017.
- [19] G. A. Gottwald and I. Melbourne. Homogenization for deterministic maps and multiplicative noise. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 469(2156), 2013.
- [20] 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.
- [21] G. A. Gottwald and S. Reich. Combining machine learning and data assimilation to forecast dynamical systems from noisy partial observations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31:101103, 2021.
- [22] G. A. Gottwald and S. Reich. Localized schrödinger bridge sampler. Technical report, arXiv:2409.07968, 2024.
- [23] 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.
- [24] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23: 47:1–47:33, 2021.
- [25] M. Hütter and H. C. Öttinger. Fluctuation-dissipation theorem, kinetic stochastic integrals and efficient simulations. *J. Chem. Soc. Faraday Trans.*, 94:1403–1405, 1998.
- [26] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 2005.
- [27] D. Kelly and I. Melbourne. Deterministic homogenization for fast–slow systems with chaotic noise. *Journal of Functional Analysis*, 272:4063–4102, 2017.
- [28] P. A. Knight. The Sinkhorn-Knopp algorithm: Convergence and applications. *SIAM J. Matrix Anal. Appl.*, 30:261–275, 2008.
- [29] R. Laumont, V. D. Bortoli, A. Almansa, J. Delon, A. Durmus, and M. Pereyra. Bayesian imaging using plug & play priors: When Langevin meets Tweedie. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022.
- [30] T. Lelièvre, G. A. Pavliotis, G. Robin, R. Santet, and G. Stoltz. Optimizing the diffusion coefficient of overdamped Langevin dynamics. Technical report, arXiv:2404.12087, 2024.
- [31] C. Li, C. Chen, D. E. Carlson, and L. Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI Conference on Artificial Intelligence*, 2015.
- [32] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto. Diffusion-LM improves controllable text generation. Technical report, arXiv:2205.14217, 2022.
- [33] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963.
- [34] D. Maoutsa, S. Reich, and M. Opper. Interacting particle solutions of Fokker–Planck equations through gradient-log-density estimation. *Entropy*, 22(8), 2020.
- [35] N. F. Marshall and R. R. Coifman. Manifold learning with bi-stochastic kernels. *IMA J. Appl. Maths.*, 84:455–482, 2019.
- [36] MATLAB. *version 9.13.0.2049777 (R2022b)*. The MathWorks Inc., Natick, Massachusetts, 2022.

- [37] 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.
- [38] I. Melbourne and A. Stuart. A note on diffusion limits of chaotic skew-product flows. *Nonlinearity*, 24:1361–1367, 2011.
- [39] S. Meyn and R. T. Tweedy. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition, 2009.
- [40] P. Milanfar and M. Delbracio. Denoising: A powerful building-block for imaging, inverse problems, and machine learning. *arXiv preprint arXiv:2409.06219*, 2024.
- [41] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer Verlag, New York, 2016.
- [42] S. Peluchetti. Diffusion bridge mixture transports, Schrödinger bridge problems and generative modeling. *J. Mach. Learn. Res.*, 24(1), 2024.
- [43] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 5–6:355–607, 2019.
- [44] A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. Technical report, arXiv:2109.12004, 2022.
- [45] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, Cambridge, 2015.
- [46] L. Rey-Bellet and K. V. Spiliopoulos. Improving the convergence of reversible samplers. *Journal of Statistical Physics*, 164:472–494, 2016.
- [47] I. H. Salgado-Ugarte and M. A. Pérez-Hernández. Exploring the use of variable bandwidth kernel density estimators. *The Stata Journal*, 3:133–147, 2003.
- [48] M. E. Sander, P. Ablin, M. Blondel, and G. Peyre. Sinkformers: Transformers with doubly stochastic attention. *PMLR*, 151:1–16, 2022.
- [49] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.
- [50] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. In *Proceedings of the International Conference in Learning Representations*, 2018.
- [51] Y. Shi, V. D. Bortoli, A. Campbell, and A. Doucet. Diffusion Schrödinger bridge matching. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=qy070HsJT5>.
- [52] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Álché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [53] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020.
- [54] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.
- [55] F. Vargas, P. Thodoroff, A. Lamacraft, and N. Lawrence. Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9), 2021. doi: 10.3390/e23091134.
- [56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates,

- Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [57] N. Whiteley, A. Gray, and P. Rubin-Delanchy. Statistical exploration of the manifold hypothesis, 2024. URL <https://arxiv.org/abs/2208.11665>.
- [58] 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.
- [59] D. Yogatama, C. Dyer, W. Ling, and P. Blunsom. Generative and discriminative text classification with recurrent neural networks. Technical report, arXiv:1703.01898, 2017.