LOCALIZED SCHRÖDINGER BRIDGE SAMPLER

GEORG A. GOTTWALD 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. In this paper, we build on previous work combining Schrödinger bridges and plug & play Langevin samplers. A key bottleneck of these approaches 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. In this context, a connection to multi-head self attention transformer architectures is established. As for the original Schrödinger bridge sampling 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 high-dimensional Gaussian problem, on a temporal stochastic process, and on a stochastic subgrid-scale parametrization conditional sampling problem. We also extend the idea of localization to plug & play Langevin samplers using kernel-based denoising in combination with Tweedie's formula.

Keywords: generative modeling, Langevin dynamics, denoising, 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 [28, 12, 27, 29, 36]. 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 $s(x;\theta) \approx \nabla \log \pi(x)$ by minimizing an appropriate loss function such as

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

in the parameters $\theta \in \mathbb{R}^{d_\theta}$ [13]. This estimate can then be used in combination with overdamped Langevin dynamics to yield

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

where $W(\tau)$ denotes standard d-dimensional Brownian motion [21]. The stochastic differential equation is typically discretized by the Euler–Maruyama (EM) method to yield an iterative

Date: November 19, 2024.

update of the form

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

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 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 $\pi(x)$.

Instead of first estimating the score function from samples and then discretizing (2) in time, it has been proposed in [10] 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

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

with appropriately defined diffusion matrix $S(x;\epsilon) \in \mathbb{R}^{d\times d}$. Broadly speaking, $m(x;\epsilon)$ controls the drift while $S(x;\epsilon)$ moderates the noise. An obvious choice for $S(x;\epsilon)$ is $S(x;\epsilon) = 2\epsilon I$, which corresponds to the EM discretization (3). A data-aware $S(x;\epsilon)$ has been introduced in [10], which is defined as the Schrödinger bridge approximation to the covariance matrix

$$\Sigma(x;\epsilon) := \mathbb{E}[X(\epsilon)X(\epsilon)^\top|X(0) = x] - \mu(x;\epsilon)\mu(x;\epsilon)^\top. \tag{6}$$

Provided the measure $\nu(\mathrm{d}x)$ possesses a smooth density $\pi(x)$, it holds asymptotically that

$$\Sigma(x;\epsilon) = 2\epsilon I + \mathcal{O}(\epsilon^2). \tag{7}$$

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 values (4) and (6) can be defined appropriately. The Schrödinger bridge approximation allows for such an extension [10]. Indeed, the, so called, *manifold hypothesis* states that many applications of generative modeling lead to measures $\nu(\mathrm{d}x)$ which concentrate on a low-dimensional manifold $\mathcal{M}$ in a high-dimensional ambient space $\mathbb{R}^d$ [8, 27, 33].

We note that (5) is closely related to the plug & play unadjusted Langevin sampler (PnP-ULA) of [14], where a denoiser $D(x;\epsilon)$ takes the role of $m(x;\epsilon)$ in (5), which should satisfy

$$\nabla \log \pi(x) \approx \frac{D(x;\epsilon) - x}{\epsilon} \tag{8}$$

and $S(x;\epsilon) = 2\epsilon I$. An additional stabilizing term of the form

$$\frac{\epsilon}{\lambda}(P_{\mathcal{C}}(X(n)) - X(n)) \tag{9}$$

is required for the associated PnP-ULA scheme to satisfy an appropriate growth condition. Here $\lambda > 0$ is a suitable parameter and $P_{\mathcal{C}}(x)$ projects x onto a compact set $\mathcal{C} \subset \mathbb{R}^d$ which should contain most of the probability mass of $\nu(\mathrm{d}x)$. In particular, if $\nu(\mathrm{d}x)$ is supported on a manifold $\mathcal{M} \subset \mathbb{R}^d$, then $\mathcal{C} \subseteq \mathcal{M}$. See [14] for more details and [5] for a related approach using a Moreau–Yoshida regularised score function. We note that the Schrödinger bridge sampler (5) has been shown to be stable and geometric ergodic [10] without any additional stabilization term.

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

$$X(n+1) = m(X(n);\epsilon) + \epsilon\nabla \log \pi(y|X(n)) + \sqrt{S(X(n);\epsilon)}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, I). \tag{10}$$

Furthermore, a particular choice of $\pi(y|x)$ can be used for conditional sampling [10]. We note that (10) fits into the general plug & play approach to data-aware Bayesian inference [31, 3, 14].

While it has been demonstrated in [10] that (5) and (10) work well for low-dimensional problems, the required number of training samples, M, increases exponentially in the dimension, d, of the samples [35]. 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 $m(x;\epsilon)$ and similarly for the diffusion matrix $S(x;\epsilon)$. The proposed localization strategy resembles localization strategies used in the ensemble Kalman filter (EnKF) [7, 25, 4]; but is fundamentally different in at least two ways: (i) For Gaussian measures with covariance matrix C, the EnKF would localize the empirical estimator of C while our approach relies on the localization of the precision matrix C^{-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. Furthermore, we extend the proposed localization strategy to conditional mean estimators based on kernel denoising [20].

The paper is organized as follows. The Schrödinger bridge formulation for $m(x;\epsilon)$ and $S(x;\epsilon)$ in (5) is summarized in the subsequent Section 2. There we also discuss connections to minimum mean square error (MMSE) denoising and kernel-based denoising, in particular to [20]. The localized variant is subsequently developed in Section 3 first for a Gaussian distribution for which C^{-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. Localization is extended to kernel-based denoising in Section 3.4. We discuss a connection between the Schrödinger bridge sampler and self attention transformers [30] in Remark 2 and its localized variant in the context of multi-head transformers in Remark 3. As applications, we consider sampling temporal stochastic processes in Section 4 and conditional sampling for a closure problem arising from the multi-scale Lorenz-96 model [16] in Section 5. The paper closes with some conclusions and suggestions for further work.

2. Plug & play Langevin sampler

In this section, we summarize two particular variants of plug & play Langevin samplers [14]. The first sampler has been proposed in [10] and is based on a Schrödinger bridge approximation of the Langevin semi-group with invariant measure $\nu(\mathrm{d}x)$ [35]. The second sampler builds upon kernel denoising [20] and Tweedie's formula [6].

2.1. Schrödinger bridge sampler.

In this subsection, we briefly recall how to approximate the conditional estimates (4) and (6) using Schrödinger bridges. One first introduces the symmetric matrix $T \in \mathbb{R}^{M\times M}$ of (unnormalized) transition probabilities

$$(T)_{jk} = \exp\left(-\frac{1}{4\epsilon}\|x^{(k)} - x^{(j)}\|^2\right) \tag{11}$$

for $j, k = 1, \ldots, M$. See [10] for a more general definition involving a state-dependent scaling matrix $K(x)$ and variable bandwidth implementation $K(x) = \rho(x)I$ with $\rho(x) > 0$ a suitable scaling function.

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

$$P = D(v)\, T\, D(v) \tag{12}$$

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

$$P\, w^* = w^*. \tag{13}$$

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 (12) by $\tilde{P} = M^{-1}P$ [22].

The next step is to extend the discrete Markov chain (12) to all $x \in \mathbb{R}^d$. For that purpose one introduces the vector-valued function $t(x) \in \mathbb{R}^M$ with entries

$$t^{(j)}(x) = \exp\left(-\frac{1}{4\epsilon}\|x - x^{(j)}\|^2\right) \tag{14}$$

for $j = 1, \ldots, M$. One then defines the probability vector $w(x) \in \mathbb{R}^M$ using the Sinkhorn weights, v, obtained in (12), i.e.,

$$w(x) = \frac{D(v)\, t(x)}{v^\top t(x)} \in \mathbb{R}^M \tag{15}$$

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

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

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

$$m(x;\epsilon) := \mathcal{X}\, w(x), \tag{17}$$

which provides a finite-dimensional approximation of the conditional expectation value $\mu(x;\epsilon)$ of the true underlying diffusion process (2). Note that the conditional mean $m(x;\epsilon)$ lies in the convex hull of the data since $0 \le w(x) \le 1$ is a probability vector for all x

We also recall a data-aware choice of the covariance matrix $S(x;\epsilon)$ [10]. Using (15) and (16), one can define the conditional covariance matrix

$$S(x;\epsilon) = \mathcal{X}\, D(w(x))\, \mathcal{X}^\top - m(x;\epsilon)\, m(x;\epsilon)^\top \in \mathbb{R}^{d\times d}, \tag{18}$$

which is the empirical covariance matrix associated with the probability vector $w(x)$; compare (6).

It has been found advantageous in [10] to replace the time-stepping method (5) by the split-step scheme

$$X(n+1/2) = X(n) + \sqrt{S(X(n);\epsilon)}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, I), \tag{19a}$$

$$X(n+1) = m(X(n+1/2);\epsilon), \tag{19b}$$

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 (19b) amounts to a projection into the convex hull of the samples, independent of the outcome of the noising step (19a). 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 $\nu(\mathrm{d}x)$ concentrates 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 [27].

Remark 2. We point to a connection of (19b) to self-attention transformer architectures [30]. We recall that the attention function acts on a matrix $Q \in \mathbb{R}^{N\times d}$ of N queries, a matrix $K \in \mathbb{R}^{M\times d}$ of M keys, and a matrix $V \in \mathbb{R}^{M\times d}$ of M values in the form of

$$\text{Attention}\,(Q,K,V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V. \tag{20}$$

In the context of (19b) we find that $Q = X(n+1/2)^{\mathrm{T}}$ and $K = V = \mathcal{X}^{\mathrm{T}}$. Hence $N = 1$ and the output of the softmax function becomes a probability vector of dimension $1\times M$ which we denote by $\breve{w}$. Note that $\breve{w} \in \mathbb{R}^{1\times M}$ multiplies $V = \mathcal{X}^{\mathrm{T}} \in \mathbb{R}^{M\times d}$ from the left resulting in essentially the transpose of what has been used in (17) with, however, a differently defined probability vector. Indeed, the Schrödinger bridge sampler defines the probability vector (15) in a manner closely related to what has been proposed as Sinkformer in [26] and the scaling factor $\sqrt{d}$ in (20) is substituted by 2ϵ. More specifically, we note that (11) could be replaced by

$$(T)_{jk} = \exp\left(\frac{(x^{(k)})^{\top}x^{(j)}}{2\epsilon}\right) \tag{21}$$

without changing the resulting Schrödinger bridge approximation P (albeit with a different scaling vector v compared to (12)). One would then also have to replace (14) and (15) by

$$\hat{t}^{(j)}(x) = v^{(j)}\exp\left(\frac{x^{\top}x^{(j)}}{2\epsilon}\right) = \exp\left(\frac{x^{\top}x^{(j)}}{2\epsilon} + \log v^{(j)}\right) \tag{22}$$

and

$$\hat{w}^{(j)}(x) = \frac{\hat{t}^{(j)}(x)}{\sum_{j=1}^{M}\hat{t}^{(j)}(x)} \tag{23}$$

for $j = 1,\dots,M$, respectively, in line with self-attention transformer architectures which do not involve the shift by $\log v^{(j)}$ in (22).

The denoising step (19b) has a gradient structure since

$$m(x;\epsilon) = x + \epsilon\nabla\log p_\epsilon(x) \tag{24}$$

with (unnormalised) density

$$p_\epsilon(x) = \big(v^{\top}t(x)\big)^2 \tag{25}$$

and

$$\nabla\log p_\epsilon(x) = -\frac{1}{\epsilon}\frac{\sum_{j=1}^{M}(x-x^{(j)})v^{(j)}t^{(j)}(x)}{v^{\mathrm{T}}t(x)} = \frac{1}{\epsilon}(\mathcal{X}\,w(x) - x). \tag{26}$$

Hence the proposed sampler can be viewed as an EM approximation of the modified Langevin dynamics

$$\dot{X}(\tau) = \nabla\log p_\epsilon(X(\tau)) + \sqrt{2}\,\dot{W}(\tau). \tag{27}$$

A modified score has also been considered in the form of Moreau–Yosida regularised score functions in [5] and smoothed score functions in the form of plug & play priors in [14]. Contrary to those approaches, the modified score $\nabla\log p_\epsilon(x)$ arises from the Schrödinger bridge approximation of the semi-group $\exp(\epsilon\mathcal{L})$ with generator $\mathcal{L}$ [21] given by

$$\mathcal{L}f = \nabla\log\pi(x)\cdot\nabla f + \Delta f. \tag{28}$$

2.2. **Kernel denoising and Tweedie's formula.** We note that (19b) is related to MMSE denoising as widely used to reduce random fluctuations in a signal. The connection between score estimation, autoencoders, and denoising has been discussed in [32, 1]. See also the recent survey [20]. However, while MMSE denoising typically considers conditional mean estimators in pseudo-linear form [20] or in the form of auto-encoders [1], our approach relies on (nonlinear) conditional mean estimators of the form (17), which also arise from kernel denoising [20], which is closely related to Tweedie's formula [6] as we explain next.

Given the data distribution π and a scale parameter $\gamma > 0$, consider the extended (unnormalized) distribution Π_γ in $(x, x') \in \mathbb{R}^{2d}$ defined by

$$\Pi_\gamma(x, x') = \exp\left(-\frac{1}{2\gamma}\|x - x'\|^2\right) \pi(x') \tag{29}$$

and its (unnormalized) marginal distribution

$$\pi_\gamma(x) = \int \Pi_\gamma(x, x')\, \mathrm{d}x'. \tag{30}$$

Tweedie's formula [6] states that

$$\nabla \log \pi_\gamma(x) = -\frac{1}{\gamma}\left(x - \mathbb{E}[x'|x]\right) \tag{31}$$

with the conditional expectation value defined by

$$\mathbb{E}[x'|x] = \frac{\int x'\, \Pi_\gamma(x, x')\, \mathrm{d}x'}{\pi_\gamma(x)}. \tag{32}$$

We may now replace the data distribution π_γ by the empirical measure over the training samples $\{x^{(j)}\}_{j=1}^M$ to obtain the equally weighted (unnormalized) Gaussian kernel density estimator (KDE)

$$\tilde{\pi}_\gamma(x) = \sum_{j=1}^M \exp\left(-\frac{1}{2\gamma}\|x - x^{(j)}\|^2\right), \tag{33}$$

which, according to (31), leads to the score function

$$s(x;\gamma) = \nabla \log \tilde{\pi}_\gamma(x) = -\frac{1}{\gamma}\left(x - \frac{\sum_{j=1}^M x^{(j)} \exp\left(-\frac{1}{2\gamma}\|x - x^{(j)}\|^2\right)}{\sum_{j=1}^M \exp\left(-\frac{1}{2\gamma}\|x - x^{(j)}\|^2\right)}\right). \tag{34}$$

Using this score function in (3) with $\gamma = \epsilon$ results in a scheme of the form (5) with $S(x;\epsilon) = 2\epsilon I$ and the conditional mean estimator $m(x;\epsilon)$ being replaced by the denoiser

$$D(x;\epsilon) := \frac{\sum_{j=1}^M x^{(j)} \exp\left(-\frac{1}{2\epsilon}\|x - x^{(j)}\|^2\right)}{\sum_{j=1}^M \exp\left(-\frac{1}{2\epsilon}\|x - x^{(j)}\|^2\right)} = \mathcal{X}\, \tilde{w}(x), \tag{35}$$

where the weight vector $\tilde{w}(x) \in \mathbb{R}^M$ is now defined by

$$\tilde{w}^{(j)}(x) = \frac{\exp\left(-\frac{1}{2\epsilon}\|x - x^{(j)}\|^2\right)}{\sum_{j=1}^M \exp\left(-\frac{1}{2\epsilon}\|x - x^{(j)}\|^2\right)}, \qquad j = 1, \ldots, M. \tag{36}$$

We find that $m(x;\epsilon)$ and $D(x;\epsilon)$ differ through the additional Sinkhorn weight vector $v \in \mathbb{R}^M$ in (15) and the scale parameter $2\gamma = 2\epsilon$ in (33) compared to the scale parameter 4ϵ used in (14). The connections drawn in Remark 2 to transformer architectures apply equally to (35).

The results of [35] suggest that (17) provides a more accurate approximation to the conditional expectation value (4) than (35), which is based on the equally weighted Gaussian mixture approximation (33). In particular, it holds that

$$m(x;\epsilon) := \exp(\epsilon\mathcal{L})\,\mathrm{id}(x) = x + \epsilon\nabla \log \pi(x) + \mathcal{O}(\epsilon^2) \tag{37}$$

in the limit $M \to \infty$, while Tweedie's formula formally leads to

$$D(x;\epsilon) := x + \epsilon\nabla \log \tilde{\pi}_\epsilon(x) = x + \epsilon\nabla \log \pi(x) + \mathcal{O}(\epsilon^2). \tag{38}$$

Here $\mathrm{id}(x) = x$ denotes the identity map. Hence, to leading order in ϵ, both approaches agree. However, while Tweedie's formula leads to an approximation error that arises from replacing π by a regularized density $\tilde{\pi}_\epsilon$, the Schrödinger bridge sampler leads to higher-order corrections which are consistent with the actual underlying Langevin dynamics. This becomes particularly appealing when implemented together with the data-aware covariance matrix (18) instead of a constant $S(x;\epsilon) = 2\epsilon I$ in (5) or when a variable bandwidth is implemented in (11) as was done in [10]. A precise statement will be the subject of future research. We also stress that the Schrödinger bridge sampler can easily be extended to Langevin dynamics with multiplicative noise [10] while such an extension is unclear when based on a KDE.

While (17) works well for low-dimensional problems and sufficiently large sample sizes M, applications to medium- or high-dimensional problems have remained an open challenge since accurate approximations of the Schrödinger bridge problem require an exponentially increasing number of samples as the dimension, d, of the sample space $\mathbb{R}^d$ increases (cf. [35]). The curse of dimensionality applies equally to the KDE-based approximation (35) and a failure to generalize has been discussed recently in the context of score-generative models [15].

The key observation of this paper is that the approximation of conditional expectations (4) via Schrödinger bridges does not necessarily require the full Markov chain (12) and that localization can be applied provided conditional independence can be established. This idea will be developed in the following section. Localization will subsequently be extended to kernel-based denoising in Subsection 3.4.

3. Localized Schrödinger bridge sampler

To introduce the main idea of localizing the Schrödinger bridge sampler developed in [10] we first consider an illustrative example of sampling from a multivariate Gaussian distribution. We will see that localization allows for a significant reduction of 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 is determined by the conditional independence which typically leads to a sequence of much lower dimensional estimation problems.

3.1. Motivational example: Gaussian setting.

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

$$C = (I - \Delta_h)^{-1}. \tag{39}$$

Instead of the distribution $\pi(x)$, we are given M samples $x^{(j)} \sim \mathrm{N}(0, C)$, $j = 1, \dots, M$, and denote their α-th entry by $x_\alpha^{(j)}$ for $\alpha = 1, \dots, d$. The goal is to produce more samples from $\mathrm{N}(0, C)$ using the time-stepping scheme (19) without making explicit reference to the unknown covariance matrix C. This particular setting of a generative model can become arbitrarily challenging by either increasing L for fixed mesh-size h or by decreasing the mesh-size $h = L/d$ for fixed L.

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

$$X(n+1) = X(n) - \epsilon(I - \Delta_h)X(n) + \sqrt{2\epsilon}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, I). \tag{40}$$

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

$$X_\alpha(n+1) = w_{-1}\, X_{\alpha-1}(n) + w_0\, X_\alpha(n) + w_1\, X_{\alpha+1}(n) + \sqrt{2\epsilon}\,\Xi_\alpha(n), \qquad \alpha = 1, \ldots, d, \tag{41}$$

with weights

$$w_{\pm 1} = \frac{\epsilon}{2h^2}, \qquad w_0 = 1 - \epsilon\left(1 + \frac{1}{h^2}\right) \tag{42}$$

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 (41) reveals that 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.,

$$\mathbb{E}[X_\alpha(n+1)\,|\,X(n)] = \mathbb{E}[X_\alpha(n+1)\,|\,(X_{\alpha-1}(n), X_\alpha(n), X_{\alpha+1}(n))] \tag{43a}$$

$$= w_{-1}\, X_{\alpha-1}(n) + w_0\, X_\alpha(n) + w_1\, X_{\alpha+1}(n). \tag{43b}$$

It is convenient to introduce the short-hand

$$X_{[\alpha]} := (X_{\alpha-1}, X_\alpha, X_{\alpha+1})^{\mathrm{T}} \in \mathbb{R}^{d_\alpha}, \tag{44}$$

with $d_\alpha = 3$, 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. x_α for $\alpha = 1, \ldots, d$. Subscripts $[\alpha]$ are reserved to denote localization around a component α; i.e., $x_{[\alpha]} \in \mathbb{R}^{d_\alpha}$.

The dependency of the conditional expectation value (43) on the neighboring points is to be exploited in the update step (19b), which we recall here in its component-wise formulation as

$$X_\alpha(n+1) = \sum_{j=1}^{M} x_\alpha^{(j)} w^{(j)}(X(n+1/2)), \tag{45}$$

for $\alpha = 1, \ldots, d$. We recall 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 whether we can find appropriately localized probability vectors $w(x)$ for the Schrödinger bridge sampler (19). The following formal argument can be made. We restrict $\mathrm{N}(0, C)$ to $x_{[\alpha]} \in \mathbb{R}^{d_\alpha}$ and truncate the samples $x^{(j)}$, $j = 1, \ldots, M$, accordingly to yield $x_{[\alpha]}^{(j)}$. The covariance matrix $C_{\mathrm{r}} \in \mathbb{R}^{d_\alpha \times d_\alpha}$ of the reduced random variables $X_{[\alpha]} \in \mathbb{R}^{d_\alpha}$ is simply the restriction of C to the corresponding sub-space, which in this particular example is independent of α. Furthermore, using the Schur complement, one finds

$$C_{\mathrm{r}}^{-1} = \begin{pmatrix} * & -\frac{1}{2h^2} & * \\ -\frac{1}{2h^2} & 1 + \frac{1}{h^2} & -\frac{1}{2h^2} \\ * & -\frac{1}{2h^2} & * \end{pmatrix}, \tag{46}$$

where $*$ denotes entries which differ from the matrix which would be obtained by restricting C^{-1} to the corresponding sub-space. The important point is that the central elements remain identical (cf. (42)) and that only those entries enter the approximation of the conditional expectation value (43).

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 (11) by localized matrices $T_\alpha \in \mathbb{R}^{M\times M}$ with entries

$$(T_\alpha)_{jk} = \exp\left(-\frac{1}{4\epsilon}\|x_{[\alpha]}^{(j)} - x_{[\alpha]}^{(k)}\|^2\right), \qquad j,k = 1,\ldots,M, \tag{47}$$

for fixed $\alpha \in \{1,\ldots,d\}$. For each of these localized matrices T_α we employ the local Sinkhorn algorithm to obtain the Sinkhorn weights $v_\alpha \in \mathbb{R}^M$ for $\alpha = 1,\ldots,d$, which render

$$P_\alpha = D(v_\alpha)T_\alpha D(v_\alpha) \tag{48}$$

bistochastic (cf.(13)). The key point is that the Euclidean norm in $\mathbb{R}^d$, $d \gg 1$, is replaced by the Euclidean norm in $\mathbb{R}^{d_\alpha}$ with $d_\alpha = 3$. Furthermore, in this particular example, the corresponding Schrödinger bridge approximately couples the restricted Gaussian distribution $\mathrm{N}(0, C_r)$ with itself. Next, the single M-dimensional probability vector (15) is replaced by d M-dimensional probability vectors

$$w_\alpha(x_{[\alpha]}) := \frac{D(v_\alpha)\, t_\alpha(x_{[\alpha]})}{v_\alpha^\top t_\alpha(x_{[\alpha]})}, \qquad \alpha = 1,\ldots,d, \tag{49}$$

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

$$t_\alpha^{(j)}(x_{[\alpha]}) = \exp\left(-\frac{1}{4\epsilon}\|x_{[\alpha]}^{(j)} - x_{[\alpha]}\|^2\right), \qquad j = 1,\ldots,M. \tag{50}$$

Note that (49) 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

$$w_\alpha^{(j)}(x_{[\alpha]}^{(k)}) = (P_\alpha)_{jk}, \qquad j,k = 1,\ldots,M, \tag{51}$$

holds.

We obtain the localized approximation

$$m_\alpha(x_{[\alpha]};\epsilon) = \mathcal{X}_\alpha\, w_\alpha(x_{[\alpha]}), \qquad \alpha = 1,\ldots,d, \tag{52}$$

of the conditional expectation values, where

$$\mathcal{X}_\alpha = (x_\alpha^{(1)},\ldots,x_\alpha^{(M)}) \in \mathbb{R}^{1\times M}. \tag{53}$$

We also introduce the localized data matrix

$$\mathcal{X}_{[\alpha]} = (x_{[\alpha]}^{(1)},\ldots,x_{[\alpha]}^{(M)}) \in \mathbb{R}^{d_\alpha\times M}, \tag{54}$$

which enters into the computation of T_α.

For constant diffusion $S(x;\epsilon) = 2\epsilon I$ the localized variant of the iteration scheme (5) becomes

$$X_\alpha(n+1) = m_\alpha(X_{[\alpha]}(n);\epsilon) + \sqrt{2\epsilon}\,\Xi_\alpha(n), \qquad \alpha = 1,\ldots,d, \tag{55}$$

for $\Xi(n) \sim \mathrm{N}(0, I)$. Similarly, the split-step scheme (19) becomes

$$X_{[\alpha]}(n+1/2) = X_{[\alpha]}(n) + \sqrt{2\epsilon}\,\Xi_{[\alpha]}(n), \tag{56a}$$

$$X_\alpha(n+1) = m_\alpha(X_{[\alpha]}(n+1/2);\epsilon). \tag{56b}$$

In other words, we have replaced a single Schrödinger bridge update in $\mathbb{R}^d$ by d Schrödinger bridge updates in $\mathbb{R}^{d_\alpha}$.

Remark 3. In line with the discussion on transformer architectures from Remark 2, we wish to point to a connection to multi-head attention [30]. More specifically, our localization procedure has introduced d heads each relying on $\mathcal{X}_{[\alpha]}$ as matrix of key vectors, $\mathcal{X}_\alpha$ as matrix of value vectors, and $X_{[\alpha]}(n+1/2)$ as query vector in order to produce an update in the scalar-valued entries $X_\alpha(n+1/2)$ for $\alpha = 1,\ldots,d$.

We finally discuss a localized version of the data-aware covariance matrix (18). Given localized weights $w_\alpha(x_{[\alpha]})$ and localized data matrices $\mathcal{X}_{[\alpha]}$, we define the $d_\alpha \times d_\alpha$-dimensional covariance matrices

$$S_\alpha(x_{[\alpha]};\epsilon) = \mathcal{X}_{[\alpha]} D(w_\alpha(x_{[\alpha]})\mathcal{X}_{[\alpha]}^\top - \mathcal{X}_{[\alpha]} w_\alpha(x_{[\alpha]})\, w_\alpha(x_{[\alpha]})^\top \mathcal{X}_{[\alpha]}^\top \tag{57}$$

for $\alpha = 1, \ldots, d$. Given a sample $\Xi(n) \sim \mathrm{N}(0, I)$, one first computes the d_α-dimensional vector $\sqrt{S_\alpha(x_{[\alpha]};\epsilon)}\,\Xi_{[\alpha]}(n)$ of which one picks its scalar entry corresponding to x_α, which we denote by $\xi_\alpha(n)$. The localized variant of (5) then becomes

$$X_\alpha(n+1) = m_\alpha(X_{[\alpha]}(n);\epsilon) + \xi_\alpha(n), \qquad \alpha = 1, \ldots, d. \tag{58}$$

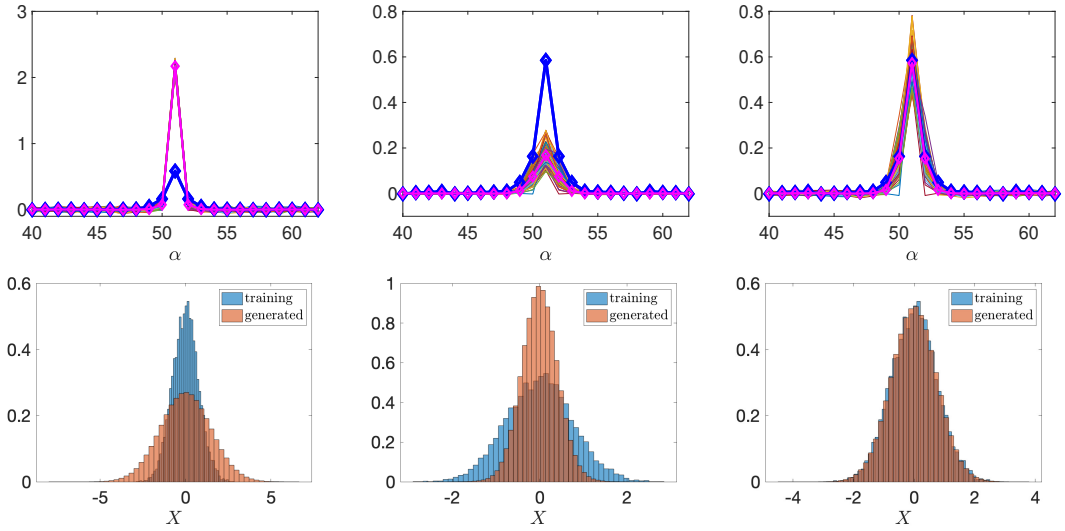

FIGURE 1. Comparison of the samples obtained from three different variants of localized Schrödinger bridge samplers. We show the centered rows of the empirical covariance matrix $\hat{C}$ (top row) and empirical histograms (bottom row) obtained from using all $d = 101$ components. The blue markers denote the empirical covariance for the given samples; the magenta markers show the average over all d rows. Left column: Localized EM-type sampler (55); middle column: Localized split-step sampler (56); right column: Localized sampler (58) with data-aware diffusion matrix (57). Given the large value of $\epsilon = 1$, only (58) is able to faithfully reproduce the target measure $\mathrm{N}(0, C)$.

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

$$x^{(j)} \sim \mathrm{N}(0, C) \tag{59}$$

for $j = 1, \ldots, M$ with a $d \times d$ covariance matrix of the form (6) with a tridiagonal precision matrix with $C^{-1}_{i,i} = 2$, $C^{-1}_{i,i\pm1} = -0.5$, and periodic conditions $C^{-1}_{1,d} = C^{-1}_{d,1} = -0.5$. The corresponding entries of the covariance matrix are $C_{i,i} \approx 0.58$, $C_{i,i\pm1} \approx 0.15$, $C_{i,i\pm2} \approx 0.04$, and $C_{i,i\pm3} \approx 0.01$.

We employ three different implementations of the localized Schrödinger bridge sampler with a localization set comprised of two neighboring grid points, i.e. $d_\alpha = 3$. These implementations are (i) the split-step scheme (56), (ii) the localized EM-type scheme (55), and (iii) the scheme (58) with data-aware noise update.

In Figure 1 we compare the generated new samples with the given samples for all three sampling strategies. We show the resulting empirical histograms as well as the rows of the empirical covariance matrix $\hat{C}$. The rows are centered at the middle point using periodicity. We use $M = 100$ training samples and a bandwidth of $\epsilon = 1$ for $d = 101$ to generate $10,000$ new samples.

While the split-step scheme (56) underestimates the variance of the distribution, the EM-type scheme (55) overestimates the variance. Only the localized scheme with data-aware diffusion (58) captures the marginal distribution and the covariance structure with desirable accuracy.

In Figure 2 we investigate the behavior of the localized split-step scheme (56) for varying parameter $\epsilon \in \{0.01, 0.1, 1\}$. We find that $\epsilon = 0.1$ improves the results while a further decrease to $\epsilon = 0.01$ leads to results comparable to $\epsilon = 1.0$. Similar results are obtained for the EM-type sampling scheme with constant diffusion (55). An optimal choice of ϵ depends, of course, on the number of available training samples, which has been set to $M = 100$ for these experiments.

We remark that unlocalized Schrödinger bridge samplers generate samples with a rather noisy correlation structure which is to be expected for $M = 100$ training samples.

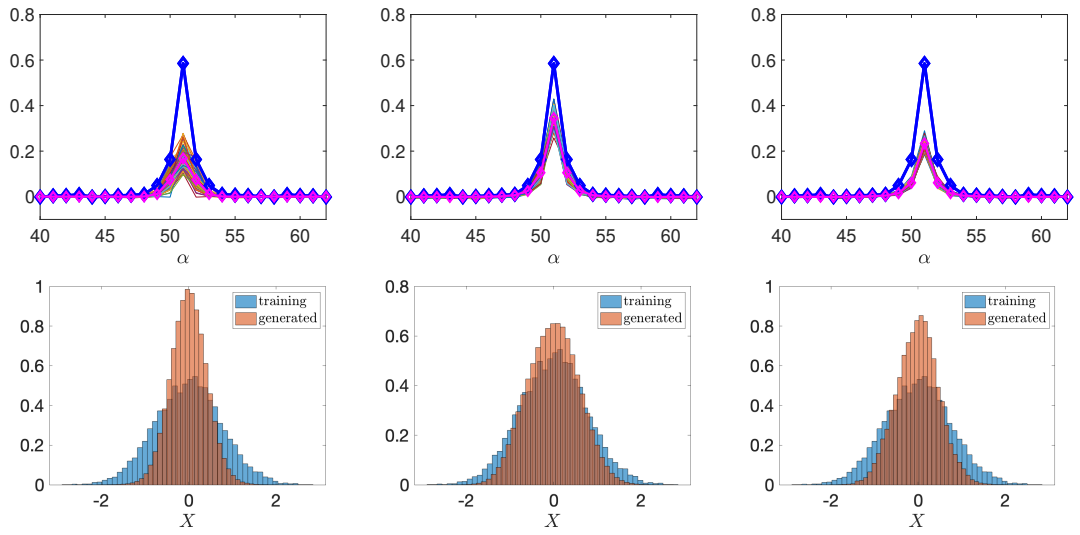

FIGURE 2. Comparison of the samples obtained from the localized split-step sampler (56) for varying parameter ϵ. Left column: $\epsilon = 1.0$; middle column: $\epsilon = 0.1$; right column: $\epsilon = 0.01$. While $\epsilon = 0.1$ leads to improved results, it is found that larger and smaller values of ϵ degrade the performance of the split-step sampler.

3.2. **Localized Schrödinger bridge sampler for general measures.** The strategy of constructing a dimension-reduced localized Schrödinger bridge sampler as presented in the previous example of a multivariate Gaussian readily extends to general target measures $\nu(\mathrm{d}x)$.

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 $\Lambda(\alpha) \subset \{1, \dots, d\}$ 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. We consider variables x'_α to be conditionally independent of $x_{\backslash[\alpha]}$ if the conditional distribution in x'_α, $p_\alpha(x_\alpha|x;\epsilon)$, satisfies

$$p_\alpha(x'_\alpha|x;\epsilon) = p_\alpha(x'_\alpha|x_{[\alpha]};\epsilon), \tag{60}$$

which implies $\mathbb{E}[x'_\alpha|x] = \mathbb{E}[x'_\alpha|x_{[\alpha]}]$. The conditional expectation value (52) 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

$$\mathbb{E}[X_\alpha(\epsilon)|X(0)=x] = \int x'_\alpha\, p(x'|x;\epsilon)\,\mathrm{d}x' = \int x'_\alpha\, p_\alpha(x'_\alpha|x;\epsilon)\,\mathrm{d}x'_\alpha \tag{61a}$$

$$= \int x'_\alpha\, p_\alpha(x'_\alpha|x_{[\alpha]};\epsilon)\,\mathrm{d}x'_\alpha = \mathbb{E}[X_\alpha(\epsilon)|X_{[\alpha]}(0)=x_{[\alpha]}], \tag{61b}$$

where the second line follows from the conditional independence assumption. 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

$$p_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)\,\pi_\alpha(x_{[\alpha]}) = p_\alpha(x_{[\alpha]}|x'_{[\alpha]};\epsilon)\,\pi_\alpha(x'_{[\alpha]}). \tag{62}$$

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

$$\mathbb{E}[X_\alpha(\epsilon)|X_{[\alpha]}(0)=x_{[\alpha]}] = \int x'_\alpha\, p_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)\,\mathrm{d}x'_{[\alpha]} \tag{63a}$$

$$= \int x'_\alpha\, \frac{p_\alpha(x_{[\alpha]}|x'_{[\alpha]};\epsilon)}{\pi_\alpha(x_{[\alpha]})}\,\pi_\alpha(x'_{[\alpha]})\,\mathrm{d}x'_{[\alpha]} \tag{63b}$$

$$= \int x'_\alpha\, \rho_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon)\,\pi(x'_{[\alpha]})\,\mathrm{d}x'_{[\alpha]}, \tag{63c}$$

with

$$\rho_\alpha(x'_{[\alpha]}|x_{[\alpha]};\epsilon) := \frac{p_\alpha(x_{[\alpha]}|x'_{[\alpha]};\epsilon)}{\pi_\alpha(x_{[\alpha]})}. \tag{64}$$

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 (63c) via Monte Carlo approximation using the restricted data samples $x^{(j)}_{[\alpha]} \sim \pi_\alpha(x_{[\alpha]})$, $j = 1, \ldots, M$; i.e.,

$$\mathbb{E}[X_\alpha(\epsilon)|X_{[\alpha]}(0)=x_{[\alpha]}] \approx \sum_{j=1}^{M} x^{(j)}_\alpha\, w^{(j)}_\alpha(x_{[\alpha]}), \qquad x^{(j)}_\alpha \sim \pi_\alpha, \tag{65}$$

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 (52). Furthermore, 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 (49). Algorithm 1 summarizes the localized Schrödinger bridge split-step sampler (56). The algorithm naturally extends to the sampler (58) with localized data-aware covariance matrix (57).

Algorithm 1: Localized Schrödinger bridge sampler

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

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

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

Step 1: Construct transition Sinkhorn weights $v_{[\alpha]}$
for $\alpha \leftarrow 1$ **to** d **do**
  construct localized data $\mathcal{X}_{[\alpha]}$;
  construct kernel matrix $T_\alpha \in \mathbf{R}^{M\times M}$ from localized data;
  construct Sinkhorn weights v_α from T_α;
end
Step 2: Generate N new samples $x_{\mathrm{s}}^{(j)}$ 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**
    $\Xi(n) \sim \mathrm{N}(0, I)$;
    for $\alpha \leftarrow 1$ **to** d **do**
      $X_{[\alpha]}(n) \leftarrow X(n)$;
      $\Xi_{[\alpha]}(n) \leftarrow \Xi(n)$;
      $X_{[\alpha]}(n+1/2) = X_{[\alpha]}(n) + \sqrt{2\epsilon}\,\Xi_{[\alpha]}(n)$; `/* 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
  $x_{\mathrm{s}}^{(j)} \leftarrow X(n_c)$;
end

Remark 4. We have assumed here a strong form of conditional independence by requesting that (61) holds for all $\epsilon > 0$. In general, such a condition will be satisfied approximately for sufficiently small ϵ only. Compare the EM sampler (40), 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 set $\Lambda(\alpha)$ which defines $x_{[\alpha]} \in \mathbb{R}^{d_\alpha}$.

3.3. **Algorithmic properties.** We briefly discuss a few important results on the stability and ergodicity of the proposed localized Langevin samplers, which they essentially inherit from the unlocalized Schrödinger bridge sampler [10].

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 (56b) 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 (52).

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

$$\mathcal{C}_M = \{x \in \mathbb{R}^d : |x_\alpha| \leq |\mathcal{X}_\alpha|_\infty\}. \tag{66}$$

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

$$m_{\rm loc}(x;\epsilon) \in \mathcal{C}_M \tag{67}$$

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 (52) 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 $m_{\rm loc}(x;\epsilon)$ and $S(x;\epsilon) = 2\epsilon I$, i.e.,

$$X(n+1) = m_{\rm loc}(X(n);\epsilon) + \sqrt{2\epsilon}\,\Xi(n), \qquad \Xi(n) \sim \mathrm{N}(0, I), \tag{68}$$

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. (19)b), but instead is confined to $\mathcal{C}_M$ in expectation. The exact gradient structure of the conditional expectation value (24) does no longer hold for the localized $m_{\rm loc}(x;\epsilon)$. The next lemma shows that the localized sampler (68) remains geometrically ergodic.

Lemma 2. Let us assume that the data generating density π has compact support. Then the localized time-stepping method (68) possesses a unique invariant measure and is geometrically ergodic.

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

$$\mathcal{B}_R = \{x \in \mathbb{R}^d : \|x\| \leq R\} \tag{69}$$

of radius $R > 0$. 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 $\mathcal{C}_M \subset \mathcal{B}_R$ and

$$\mathbb{E}[V(X(n+1))|X(n)] \leq \lambda V(X(n)) \tag{70}$$

for all $X(n) \notin \mathcal{B}_R$ with $0 \leq \lambda < 1$. Furthermore, because of the additive Gaussian noise in (68), there is a constant $\delta > 0$ such that

$$\mathrm{n}(x'; m_{\rm loc}(x;\epsilon), 2\epsilon I) \geq \delta \tag{71}$$

for all $x, x' \in \mathcal{B}_R$. Here $\mathrm{n}(x; m, C)$ denotes the Gaussian probability density function with mean m and covariance matrix C. In other words, $\mathcal{B}_R$ is a small set in the sense of [19]. Geometric ergodicity follows from Theorem 15.0.1 in [19]. See also the self-contained presentation in [18]. □

Remark 5. We emphasize that, contrary to the unlocalized Schrödinger bridge sampler, the localized $m_{\rm loc}(x;\epsilon)$ is not restricted to the linear subspace of $\mathbb{R}^d$ spanned by the training data $\mathcal{X} \in \mathbb{R}^{d\times M}$ in case $M < d$. The localized sampler shares this desirable property with the localized EnKF [25, 4, 7].

3.4. **Localized kernel-denoising.** Localizing the KDE-based denoiser (35) follows along the same lines. We first note that a component-wise formulation of Tweedie's formula (31) leads to

$$\partial_{x_\alpha} \log \pi_\epsilon(x) = -\frac{1}{\epsilon} \left(x_\alpha - \mathbb{E}[x'_\alpha | x] \right), \tag{72}$$

$\alpha = 1, \ldots, d$. Upon assuming the conditional independence relation

$$\mathbb{E}[x'_\alpha | x] = \mathbb{E}[x'_\alpha | x_{[\alpha]}] \tag{73}$$

one finds that

$$\partial_{x_\alpha} \log \pi_\epsilon(x) = \partial_{x_\alpha} \log \pi_\epsilon(x_{[\alpha]}). \tag{74}$$

We approximate the restricted density $\pi_\epsilon(x_{[\alpha]})$ by the localized KDE estimator

$$\tilde{\pi}_\epsilon(x_{[\alpha]}) \propto \sum_{j=1}^{M} \exp\left(-\frac{1}{2\epsilon} \| x_{[\alpha]} - x^{(j)}_{[\alpha]} \|^2 \right), \tag{75}$$

which results in

$$\partial_{x_\alpha} \log \pi_\epsilon(x) \approx -\frac{1}{\epsilon} \left(x_\alpha - \sum_{j=1}^{M} x^{(j)}_\alpha \, \tilde{w}^{(j)}_\alpha(x_{[\alpha]}) \right) \tag{76}$$

with localized weights

$$\tilde{w}^{(j)}_\alpha(x_{[\alpha]}) := \frac{\exp\left(-\frac{1}{2\epsilon} \| x_{[\alpha]} - x^{(j)}_{[\alpha]} \|^2 \right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2\epsilon} \| x_{[\alpha]} - x^{(j)}_{[\alpha]} \|^2 \right)} \tag{77}$$

for $j = 1, \ldots, M$. We collect these weights in the vector $\tilde{w}_\alpha(x; \epsilon) \in \mathbb{R}^M$. One finally obtains the following localized KDE update step for $X_\alpha(n + 1/2)$:

$$X_\alpha(n+1) = D_\alpha(X(n+1/2); \epsilon) := \mathcal{X}_\alpha \, \tilde{w}_\alpha(X_{[\alpha]}(n+1/2)), \qquad \alpha = 1, \ldots, d, \tag{78}$$

which can be employed whenever (73) holds to sufficient accuracy.

4. Localised Schrödinger bridge sampler for temporal stochastic processes

In this section, we consider temporal stochastic processes $Z(t_k) \in \mathbb{R}^s$ with $t_k = k\Delta t$ and $k = 0, \ldots, K$, and assume that M realizations

$$x^{(j)} = \{ Z^{(j)}(t_k) \}_{k=0}^{K} \in \mathbb{R}^d, \qquad d = (K+1)s, \tag{79}$$

$j = 1, \ldots, M$, of such a process have become available. We furthermore assume that the generating process is Markovian, i.e., $Z(t_{k+1})$ is conditionally independent of all $Z(t_l)$ with $l < k$ and $l > k+1$. Such a setting provides a perfect application of the localization strategy proposed in Section 3. More specifically, we obtain the following subsets for localization in terms of the entries x_α of the augmented state vector $x \in \mathbb{R}^d$: $\Lambda(\alpha) = \{1, \ldots, s\}$ for $\alpha \in \{1, \ldots, s\}$ and

$$\Lambda(\alpha) = \{ sl + 1, \ldots, s(l+2) \} \tag{80}$$

for $\alpha \in \{ s(l+1) + 1, \ldots, s(l+2) \}$ and $l = 0, \ldots, K-2$.

As a numerical illustration we consider the bimodal stochastic differential equation (SDE)

$$\frac{d}{dt} Z(t) = -Z(t)^3 + Z(t) + \sqrt{0.2} \frac{d}{dt} B(t), \qquad Z(0) \sim \mathrm{N}(0, 1), \tag{81}$$

where $B(t)$ denotes standard Brownian motion. Here $s = 1$ and we sample solutions in time-intervals of length $\Delta t = 5$ over $K = 100$ intervals; hence the dimension of the augmented state vector $x \in \mathbb{R}^d$ becomes $d = 101$. The training data consists of $M = 1,000$ independent realizations, which were obtained with a small step-size EM algorithm applied to (81).

Results from the localized Schrödinger bridge split-step sampler (56) with constant diffusion $S(x;\epsilon) = 2\epsilon I$ with $\epsilon = 0.0025$ and $N = 25,000$ generated samples can be found in Figures 3 and 4, respectively. The numerical results demonstrate that the localized Schrödinger bridge sampler can successfully generated samples for this rather high-dimensional ($d = 101$) and nonlinear problem given only $M = 1,000$ training samples.

We also implemented the localized KDE-based denoiser (78). The results are virtually indistinguishable from the results obtained from the localized Schrödinger bridge sampler.

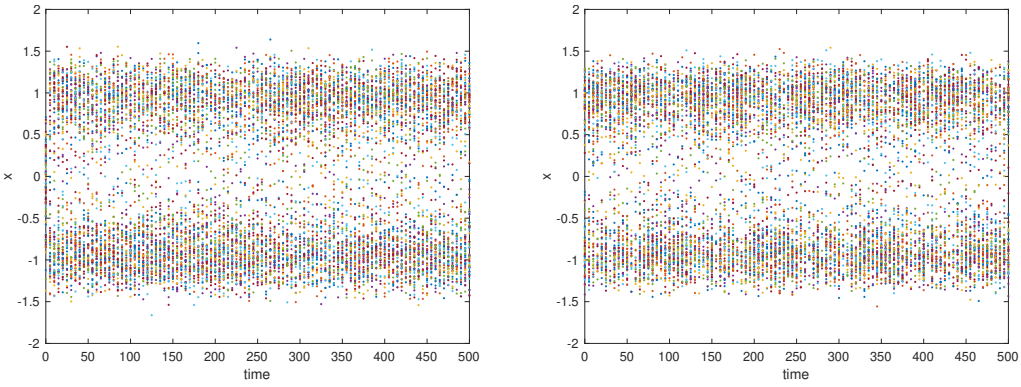

FIGURE 3. Generated trajectories for the bimodal SDE (81) using the localized Schrödinger bridge split-step sampler with constant diffusion (56). Left panel: 100 trajectories out of $M = 1,000$ training samples; right panel: 100 trajectories out of $N = 25,000$ generated samples. The computed transition rates (relative number of sign changes along trajectories) agree well with 9% for the training data and 11% for the generated data.

5. Conditional localized Schrödinger bridge sampler

As for the standard Schrödinger bridge sampler [10] the localized sampler lends itself to conditional sampling. Consider samples $x^{(j)} = (z^{(j)}, \psi^{(j)})$ for $j = 1, \ldots, M$. The localized Schrödinger bridge sampler described in Section 3 and Algorithm 1 allows us to learn the joint probability measure $\nu(\mathrm{d}z, \mathrm{d}\psi)$. To draw samples from the conditional probability measure $\nu(\mathrm{d}\psi|z)$ 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 $X(n)$ are set to the value z^* on which we wish to condition. To achieve this we add the conditioning assignment, $Z(n) \leftarrow z^*$, between lines 11 and 12 of Algorithm 1. We demonstrate the performance of the conditional sampler for the multi-scale Lorenz-96 model in the next subsection.

5.1. **Conditional sampling for a closure problem.**

We apply the conditional localized Schrödinger bridge sampler to the multi-scale Lorenz-96 model for K slow variables z_k which are

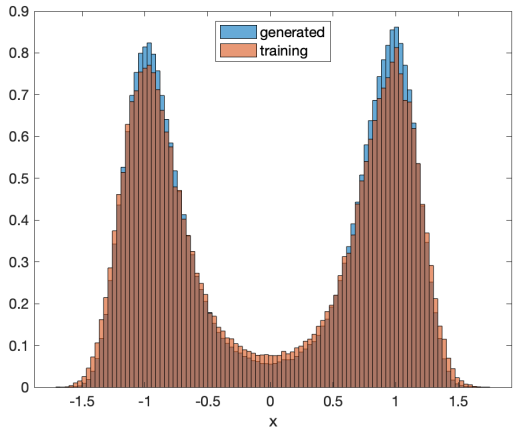

FIGURE 4. Normalized empirical histograms of training and generated data for the bimodal SDE (81) using the localized Schrödinger bridge split-step sampler with constant diffusion (56). We show results over all $1,000$ training and $25,000$ generated data points. The invariant distribution of the bimodal SDE is well reproduced by the generated data; the dispersion of the generated data in each of its two modes being slightly smaller than the one from the training data, which has also been observed for the split-step scheme in Figure 2.

each coupled to J fast variables $y_{j,k}$ and evolve according to

$$\frac{d}{dt}z_k = -z_{k-1}(z_{k-2} - z_{k+1}) - z_k + F - \frac{hc}{b}\sum_{j=1}^{J} y_{j,k}, \tag{82a}$$

$$\frac{d}{dt}y_{j,k} = -cby_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - cy_{j,k} + \frac{hc}{b}z_k \tag{82b}$$

with periodic boundary conditions $z_{k+K} = z_k$, $y_{j,k+K} = y_{j,k}$ 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 [16]. 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 [34, 2, 9]. 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 [11].

In the climate science and other disciplines one is typically only interested in the slow large-scale dynamics. A direct simulation of the multi-scale system (82), 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

$$\frac{d}{dt}z_k = G_k(z) + \psi_k(z), \tag{83}$$

for $z = (z_1, z_2, \ldots, z_K)$ with $G_k(z) = -z_{k-1}(z_{k-2} - z_{k+1}) - z_k + F$. We assume 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 (82). 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 $x^{(j)} = (z^{(j)}, \psi^{(j)})$, $j = 1, \ldots, M$, which consists of a time series with $x^{(j)} = x(t_j)$ with $t_j = j\Delta t$ and $\Delta t = 5 \times 10^{-3}$.

We obtain $M = 40,000$ samples $z^{(j)} \in \mathbb{R}^K$ by integrating (82) using a fourth-order Runge–Kutta method with a fixed time step $\delta t = 5 \times 10^{-4}$ and collecting the state in time intervals of $\Delta t = 10\,\delta t$. Samples of the closure term $\psi^{(j)} \in \mathbb{R}^K$ are then determined from the samples $z^{(j)}$ via

$$\psi^{(j)} := \frac{z^{(j+1)} - z^{(j)}}{\Delta t} - G(z^{(j)}), \tag{84}$$

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

To numerically integrate the closure model (83) in terms of the state vector $z \in \mathbb{R}^K$ we employ an Euler discretization

$$z(m+1) = z(m) + (G(z(m)) + \psi(m|z(m)))\,\Delta t \tag{85}$$

with a time step Δt. At each time step $m \geq 0$ we generate a sample $\psi(m|z(m))$ 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 parameter of $\epsilon = 0.1$ and consider a nearest neighbor localization 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 (47) and (50) for fixed parameter ϵ, we replace the standard Euclidean product with a scaled one where we divide the inner product in 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 $x^{(j)}$.

Figures 5 and 6 show a comparison of the outputs of the localized Schrödinger bridge sampler with data obtained from simulating the full multi-scale Lorenz-96 system (82). 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 (82), and of the discretization of the closure scheme (85). We show in Figure 5 a comparison of the empirical histograms of z obtained by integrating the closure model (85) with the original samples $\{z^{(j)}\}_{j=1}^M$ which were obtained from a simulation of the full multi-scale Lorenz-96 system (82). The non-stiff trained stochastic closure model (85) is able to reproduce the actual histogram well. We further show a scatter plot of the stochastic closure term ψ obtained from the full Lorenz-96 system (82) and obtained from the localized conditional Schrödinger bridge. The closure term is well represented by the localized conditional Schrödinger bridge. In Figure 6 we show the entries of the rows of the empirical covariance matrix, centered 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 that

the temporal autocorrelation structure of the Lorenz-96 system is well reproduced by the localized conditional Schrödinger bridge.

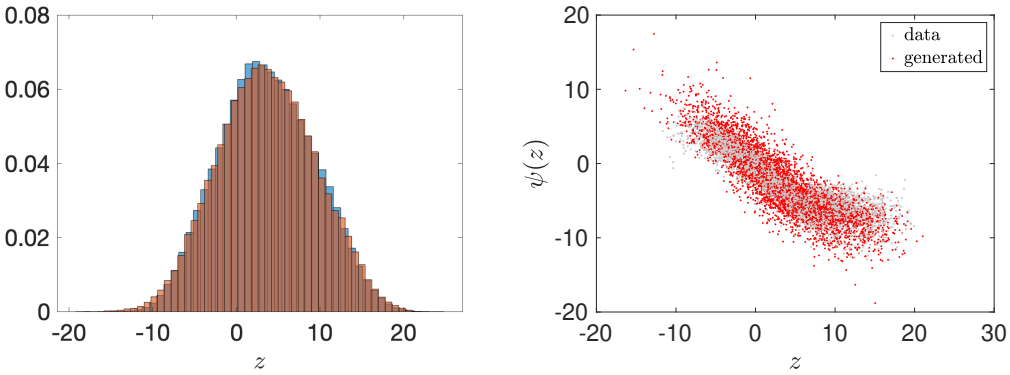

FIGURE 5. Comparison of the samples obtained from the localized Schrödinger bridge sampler and given samples drawn from the multi-scale Lorenz-96 system (82) using nearest neighbor localization with $\Lambda(\alpha) = \{\alpha - 1, \alpha, \alpha + 1\, K + \alpha - 1, K + \alpha, K + \alpha + 1\}$ with the obvious periodic extensions for $\alpha = 1$ and $\alpha = d$. We consider $40,000$ new and given samples. Left: Empirical histograms. Right: Scatter plot of the closure term ψ as a function of z.

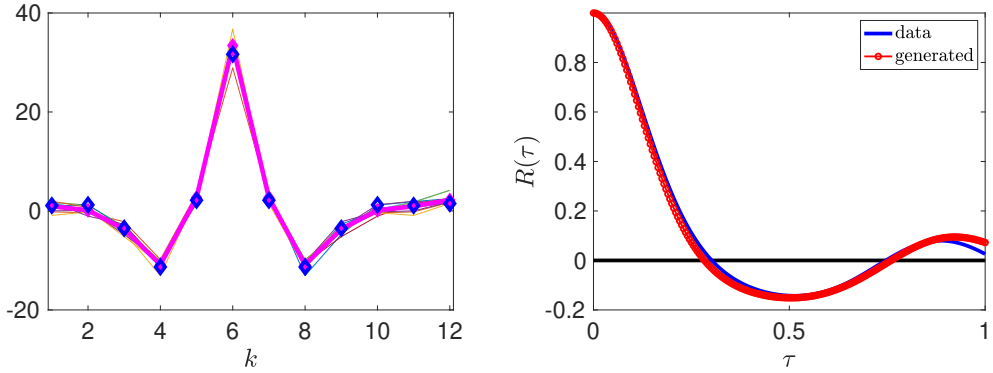

FIGURE 6. Comparison of the samples obtained from the localized Schrödinger bridge sampler and given samples drawn from the multi-scale Lorenz-96 system (82) using nearest neighbor localization with $\Lambda(\alpha) = \{\alpha - 1, \alpha, \alpha + 1\, K + \alpha - 1, K + \alpha, K + \alpha + 1\}$ with the obvious periodic extensions for $\alpha = 1$ and $\alpha = d$. We consider $40,000$ new and given samples. Left: Centered 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: Autocovariance function $R(\tau)$.

6. Conclusions

The construction of the previously proposed Schrödinger bridge sampler [10] is fraught with an unfavorable dependency in the dimension d. The required number of samples scales for a desired accuracy exponentially on the underlying intrinsic dimensionality of the data [35]. We have shown here that for data which satisfy conditional independence one can successfully employ localization to express the single Schrödinger bridge problem for d-dimensional data to d localized Schrödinger bridge problems of smaller size $d_\alpha \ll d$. The localized Schrödinger bridge sampler can be used to generate samples from an unknown prior and readily lends itself to conditional sampling and Bayesian inference.

We have numerically demonstrated the advantage of localization for several examples. We considered a Gaussian distribution for which the inverse covariance matrix has tri-diagonal structure, a bimodal SDE and a conditional sampling problem of determining a closure term in a nonlinear multi-scale system.

We have established theoretically that the proposed sampler is stable and geometric ergodic under relatively mild conditions. The stability of our sampler allows for applications to data drawn from a singular measure which arise when data are concentrated on a lower-dimensional manifold. This sets it apart from score-generative models which rely on Tweedie's formula and the differentiability of a regularized measure.

We have established several connections with other sampling strategies. The Schrödinger bridge sampler was shown to be closely related to kernel-based denoising. The Schrödinger bridge sampler, however, has the advantage that it can employ a data-aware noising step, which was demonstrated to be advantageous in Section 3.1.1, and can be constructed using a variable bandwidth [10], which is desirable with training data that involve data-sparse regions in the state space. Further, 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 [12, 27, 29, 36] and transformers [30, 26]. We have shown that the conditional mean of a Schrödinger bridge sampler is formally akin to self-attention in transformer architectures and that localization naturally leads to multi-head self attention. It will be interesting to further explore these connections.

The framework of localized Schrödinger bridges lends itself to numerous applications. In particular, we mention here sequential data assimilation [25, 4, 7], feedback particle filter and homotopy methods [37, 24, 23], which are implemented utilizing Schrödinger bridges, and interacting particle sampling methods, which rely on grad-log density estimators such as (1) [17]. Finally, the proposed localized conditional estimator $m_{\rm loc}(x)$ as well as its KDE-based variant (78) could be of independent interest for MMSE denoising [20].

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] G. Alain and Y. Bengio. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning*, 15:3743–3773, 2014.

[2] 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.

[3] S. Arridge, P. Maas, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

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

[5] A. Durmus, E. Moulines, and M. Pereayra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. *SIAM J. Imaging Sciences*, 11: 473–506, 2018.

[6] B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106:1602–1614, 12 2011. doi: 10.1198/jasa.2011.tm11181.

[7] 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.

[8] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *J. Amer. Math. Soc.*, 29:983–1049, 2016. doi: 10.1090/jams/852.

[9] 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. doi: https://doi.org/10.1016/j.physd.2021.132911.

[10] G. A. Gottwald, F. Li, S. Reich, and Y. Marzouk. Stable generative modeling using Schrödinger bridges. Technical report, arXiv:2401.04372, 2024.

[11] S. Herrera, J. Fernández, M. Rodríguez, and J. Gutiérrez. Spatio-temporal error growth in the multi-scale Lorenz'96 model. *Nonlinear Processes in Geophysics*, 17:329, 2010.

[12] 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.

[13] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 2005.

[14] 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. doi: 10.1137/21M1406349.

[15] S. Li, S. Chen, and Q. Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.

[16] 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.

[17] D. Maoutsa, S. Reich, and M. Opper. Interacting particle solutions of Fokker–Planck equations through gradient-log-density estimation. *Entropy*, 22(8), 2020. doi: 10.3390/e22080802.

[18] 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.

[19] S. Meyn and R. T. Tweedy. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition, 2009. doi: 10.1017/CBO9780511626630.

[20] P. Milanfar and M. Delbracio. Denoising: A powerful building-block for imaging, inverse problems, and machine learning. *arXiv prepint arXiv:2409.06219*, 2024.

[21] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer Verlag, New York, 2016.

[22] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 5–6:355–607, 2019.

[23] J. Pidstrigach and S. Reich. Affine-invariant ensemble transform methods for logistic regression. *Found. Comput. Math*, 2022. published online 21 January 2022.

[24] S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, 2011.

[25] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

[26] M. E. Sander, P. Ablin, M. Blondel, and G. Peyre. Sinkformers: Transformers with doubly stochastic attention. *PMLR*, 151:1–16, 2022.

[27] 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.

[28] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] 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.

[30] 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.

[31] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

[32] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23:1661–1674, 2011. doi: 10.1162/NECO_a_00142.

[33] N. Whiteley, A. Gray, and P. Rubin-Delanchy. Statistical exploration of the manifold hypothesis, 2024.

[34] 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.

[35] 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.

[36] 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.

[37] 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