NON-DENOISING FORWARD-TIME DIFFUSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The scope of this paper is generative modeling through diffusion processes. An approach falling within this paradigm is the work of Song et al. (2021), which relies on a time-reversal argument to construct a diffusion process targeting the desired data distribution. We show that the time-reversal argument, common to all denoising diffusion probabilistic modeling proposals, is not necessary. We obtain diffusion processes targeting the desired data distribution by taking appropriate mixtures of diffusion bridges. The resulting transport is exact by construction, allows for greater flexibility in choosing the dynamics of the underlying diffusion, and can be approximated by means of a neural network via novel training objectives. We develop an unifying view of the drift adjustments corresponding to our and to time-reversal approaches and make use of this representation to inspect the inner workings of diffusion-based generative models. Finally, we leverage on scalable simulation and inference techniques common in spatial statistics to move beyond fully factorial distributions in the underlying diffusion dynamics. The methodological advances contained in this work contribute toward establishing a general framework for generative modeling based on diffusion processes.

1 INTRODUCTION

Denoising diffusion probabilistic modeling (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) is a recent generative modeling paradigm exhibiting strong empirical performance. Consider a dataset of N samples $\mathcal{D} = \{x^{(n)}\}_{n=1}^{N}$ with empirical distribution $\mathcal{P}_{\mathcal{D}}$. The unifying key steps underlying DDPM approaches are: (i) the definition of a stochastic process X on $[0, \tau]$, with initial distribution $\mathcal{P}_{\mathcal{D}}$, progressively transforming $\mathcal{P}_{\mathcal{D}}$ toward a simple data-independent distribution $\mathcal{P}_Z$; (ii) the derivation of the reverse-time dynamics of X progressively transforming $\mathcal{P}_Z$ toward $\mathcal{P}_{\mathcal{D}}$; (iii) the approximation of the reverse-time transitions by means of a neural network. Following the training step (iii), a sample Y, whose distribution approximates $\mathcal{P}_{\mathcal{D}}$, is drawn by (iv) simulating from the approximated reverse-time transitions starting with a sample from $\mathcal{P}_Z$. Both discrete-time (Ho et al., 2020) and continuous-time (Song et al., 2021) formulations of DDPM have been pursued and this work focuses on the latter case. As in (Song et al., 2021), we specify the continuous-time dynamics through diffusion processes, i.e. solutions to stochastic differential equations (SDE) with associated drift $f(\cdot)$ and diffusion $g(\cdot)$ coefficients.

A number of approximations are involved in the aforementioned steps. Firstly, as the process X evolves on a finite time interval $[0, \tau]$, its (forward-time) terminal distribution Q_τ retains a dependency on the initial distribution $\mathcal{P}_{\mathcal{D}}$. The exact reverse-time simulation of X requires to start with a sample from Q_τ, and $\mathcal{P}_Z$ is only an approximation for Q_τ. Secondly, while the reverse-time dynamics of (ii) are directly available for diffusions, they are approximated by means of a neural network in (iii). Thirdly, sampling from Y in (iv) is achieved through a discretization of the reverse-time transitions of X on a time-grid, which introduces a discretization error. De Bortoli et al. (2021, Theorem 1) links these three approximations to the total variation distance between the distribution of Y and $\mathcal{P}_{\mathcal{D}}$.

In the first methodological contribution of this paper we introduce a methodology for constructing diffusion processes targeting $\mathcal{P}_{\mathcal{D}}$ without relying on time-reversal arguments. The desired transport (coupling) between an initial distribution $\mathcal{P}_Z$ and a terminal distribution $\mathcal{P}_{\mathcal{D}}$ is achieved by: (1) specifying a diffusion process X on $[0, \tau]$ starting from x_0; (2) conditioning X on hitting x_τ at time τ, thus obtaining a diffusion bridge; (3) taking a bivariate mixture of diffusion bridges over (x_0, x_τ) resulting in a diffusion with initial distribution $\mathcal{P}_Z$ and terminal distribution $\mathcal{P}_{\mathcal{D}}$. The realized diffusion bridges mixture transport (DBMT) between $\mathcal{P}_Z$ and $\mathcal{P}_{\mathcal{D}}$ is exact by construction. We thus

sidestep the approximation common to all DDPM approaches due to the discrepancy between Q_τ and its approximating $\mathcal{P}_Z$. Moreover, the DBMT can be realized for almost arbitrary, and independently chosen, $\mathcal{P}_Z$, $f(\cdot)$ and $g(\cdot)$. This represents a departure from the diffusion time-reversal transport (DTRT) of Song et al. (2021), where Q_τ depends on $\mathcal{P}_\mathcal{D}$ thought the integration time τ, which can be considered fixed (Section 4.1), and through $f(\cdot)$ and $g(\cdot)$. The need to minimize the dependency of Q_τ from $\mathcal{P}_\mathcal{D}$, i.e. to achieve $Q_\tau \approx \mathcal{P}_Z$ for some simple $\mathcal{P}_Z$, places heavy constraints on the specifications of $f(\cdot)$ and $g(\cdot)$.

Similarly to the DTRT, achieving the DBMT requires the computation of a drift adjustment term which depends on $\mathcal{D}$. For a SDE class of interest, we develop an unified and interpretable representation of DTRT and DBMT drift adjustments as simple transformations of conditional expectations over $\mathcal{D}$. This novel result provides insights on the target mapping that we aim to approximate and on the quality of approximation achieved by the trained score models of Song et al. (2021). Having defined for the DBMT a Fisher divergence objective similarly to Song et al. (2021), we leverage on this unified representation to define two additional training objectives featuring appealing properties.

In our last methodological contribution we extend the class of SDEs that can be realistically employed in computer vision applications. Specifically, computational considerations have so far restricted the transitions of the stochastic processes employed in DDPM to be fully factorials. We view images at a given resolution as taking values over a 2D lattice which discretizes the continuous coordinate system $[0, 1]^2$ representing heights and widths. Diffusion processes are viewed as spatio-temporal processes with spatial support $[0, 1]^2$. Doing so, it is possible to leverage on scalable simulation and inference techniques from spatial statistics and consider more realistic diffusion transitions.

This paper is structured as follows. In Section 2 we review the DTRT of Song et al. (2021) and in Section 3 we introduce the DBMT. In order to implement the DTRT and the DBMT it is necessary to specify the underlying SDE, i.e the coefficients $f(\cdot)$ and $g(\cdot)$. We study a class of interest in Section 4. The unified view of drift adjustments is introduced in Section 5, while Section 6 develops training objectives for the DTRT and for the DBMT. In Section 7 we establish the connection with spatio-temporal processes. We conclude in Section 8. Appendices A to D contain the theoretical framework, assumptions, proofs, analytical formulas for the SDE class of Section 4 and additional material.

Notation and conventions: we use uppercase letters for probability distributions (measures, laws) and lowercase letters for densities; random elements are denoted by uppercase letters (an exception is made for times, always lowercase for typographical reasons); if P is the distribution of a stochastic process, we use subscript notation to refer to its finite dimensional distributions (densities with p), conditional or not, for some collection of times; for example $p_{t'|t}$ denotes a transition density, which is understood to be a function of four arguments $p_{t'|t}(y|x) = f(t, t', x, y)$; we refer directly to a given SDE instead of referring to the diffusion process satisfying such SDE when no ambiguity arises; we use $[a]_i$ and $[A]_{i,j}$ for vector and matrix indexing, $A^\top$ for matrix transposition.

2 Diffusion Time-Reversal Transport

The starting point of Song et al. (2021) is a diffusion process X satisfying a generic D-dimensional time-inhomogenous SDE with initial distribution $X_0 \sim \mathcal{P}_\mathcal{D}$

$$dX_u = f(X_u, u)du + g(X_u, u)dW_u, \tag{1}$$

over noising time $u \in [0, \tau]$. Thorough this paper we denote with Q the law of the diffusion solving (1) and with q the corresponding densities. Thus, let $q_{u'|u}(y|x)$, $0 \leq u < u' \leq \tau$, be the transition density of (1), and let $q_u(x)$, $0 < u \leq \tau$, be the marginal density of (1). As $X_0 \sim \mathcal{P}_\mathcal{D}$, we have

$$q_u(x) = \frac{1}{N} \sum_{n=1}^{N} q_{u|0}(x|x^{(n)}). \tag{2}$$

The dynamics of (1) over the reversed, i.e. sampling, time $t = \tau - u$, $t \in [0, \tau]$, are given by (Anderson, 1982; Haussmann & Pardoux, 1986; Millet et al., 1989)

$$dX_t = \left[-f(X_t, u) + \nabla \cdot G(X_t, u) + G(X_t, u)\, \nabla_{X_t} \ln q_u(X_t)\right] dt + g(X_t, u)dW_t, \tag{3}$$

where $u = u(t) = \tau - t$, $G(x, u) = g(x, u)g(x, u)^\top$ and the D-dimensional vector $\nabla \cdot G(x, u)$ is defined by $[\nabla \cdot G(x, u)]_i = \sum_{j=1}^{D} \nabla_{x_j} [G(x, u)]_{i,j}$. Approximating the terminal distribution Q_τ of

(1), i.e. the initial distribution of (3), with $\mathcal{P}_Z$, X_0 is sampled from $\mathcal{P}_Z$ and (3) is discretized and integrated over t to produce a sample X_τ approximately distributed as $\mathcal{P}_\mathcal{D}$.

The computation of the multiplicative drift adjustment $\nabla_x \ln q_u(x)$ entering (3), i.e. the score of the marginal density (2), requires in principle $\mathcal{O}(N)$ operations. Let $s_\phi(x, u)$ be a neural network for which we would like $s_\phi(x, u) \approx \nabla_x \ln q_u(x)$. It remains to find a suitable training objective for which unbiased gradients with respect to ϕ can be obtained at $\mathcal{O}(1)$ cost with respect to the dataset size N. As $q_u(x)$ has a mixture representation, the identity of Vincent (2011) for Fisher divergences provides us with the desired objective for a fixed $u \in (0, \tau]$

$$\begin{aligned}\mathbb{L}_{\text{FD,DTRT}}(\phi, u) &= \mathop{\mathbb{E}}_{X_u \sim Q_u} \left[\left\| \nabla_{X_u} \ln q_u(X_u) - s_\phi(X_u, u) \right\|^2 \right] \\ &= \mathop{\mathbb{E}}_{(X_0, X_u) \sim Q_{0,u}} \left[\left\| \nabla_{X_u} \ln q_{u|0}(X_u|X_0) - s_\phi(X_u, u) \right\|^2 \right]. \end{aligned} \tag{4}$$

The key point is that an unbiased, $\mathcal{O}(1)$ with respect to N, mini-batch Monte Carlo (MC) estimator for the expectation (4) can be trivially obtained by sampling a batch $X_0 \sim \mathcal{P}_\mathcal{D}$, $X_u \sim Q_{u|0}(dX_u|X_0)$, and evaluating the average loss over the bach. In order to achieve a global approximation over the whole time interval $(0, \tau]$, Song et al. (2021) proposes uniform sampling of time u

$$\mathbb{L}_{\text{FD,DTRT}}(\phi) = \mathop{\mathbb{E}}_{u \sim \mathcal{U}(0,\tau], (X_0, X_u) \sim Q_{0,u}} \left[r_u \left\| \nabla_{X_u} \ln q_{u|0}(X_u|X_0) - s_\phi(X_u, u) \right\|^2 \right], \tag{5}$$

where $r_u = \mathbb{E}[\|\nabla_{X_u} \ln q_{u|0}(X_u|X_0)\|^2]^{-1}$ is a regularization term. A MC estimator for (5) is constructed by augmenting the MC estimator for (4) with the additional sampling step $u \sim \mathcal{U}(0, \tau]$.

3 Diffusion Bridges Mixture Transport

Our starting point is a generic D-dimensional time-inhomogenous SDE which, in contrast to Song et al. (2021), is directly defined on the sampling time $t \in [0, \tau]$

$$dX_t = f(X_t, t)dt + g(X_t, t)dW_t. \tag{6}$$

We reserve $P_{\cdot|0}(\cdot|x_0)$ to denote the law of the diffusion solving (6) for a given starting value x_0 and $p_{\cdot|0}(\cdot|x_0)$ to denote the corresponding densities.

3.1 Diffusion Bridges

Diffusion bridges are central to the proposed methodology, in this Section we cover their basic theory. A diffusion bridge is a diffusion process starting from a given value which is conditioned on hitting a terminal value. It is a deep result, and consequence of Doob h-transforms (Särkkä & Solin (2019, Chapter 7.9), Rogers & Williams (2000, Chapter IV.6.39)), that a diffusions processes pinned down on both ends is still a diffusion process. In particular the Markov property is preserved. More precisely, (6) with initial value x_0 conditioned on hitting a terminal value x_τ at time τ is characterized the following SDE on $[0, \tau]$ with initial value x_0

$$dX_t = \left[f(X_t, t) + G(X_t, t) \nabla_{X_t} \ln p_{\tau|t}(x_\tau|X_t) \right] dt + g(X_t, t)dW_t, \tag{7}$$

where $G(x, t) = g(x, t)g(x, t)^\top$. The multiplicative adjustment factor $\nabla_{x_t} \ln p_{\tau|t}(x_\tau|x_t)$ forces the process to hit x_τ at time τ and the diffusion process solving (7) is known as the diffusion bridge from $(x_0, 0)$ to (x_τ, τ). As previously noted, $p_{\tau|t}(x_\tau|x_t)$ in (7) refers to the transition density of (6).

3.2 Diffusion Mixtures

The proposed transport construction relies on a result establishing SDE representations for diffusion mixtures. We present here an informal version and report the precise statement, the required assumptions, and the proof in Appendix A.

Theorem 1 (Mixture of diffusions — informal)**.** *Let $\{X^\lambda\}$, $\lambda \in \Lambda$ be a collection of diffusions with associated SDEs $\{dX_t^\lambda\}$ and marginal densities $\{\pi_t^\lambda\}$. Let $\mathcal{L}$ be a mixing distribution on Λ, π be the $\mathcal{L}$-mixture of $\{\pi_t^\lambda\}$. Then there exists a diffusion process X with marginal π. Moreover X follows a SDE whose drift and diffusion coefficients are weighted averages of the corresponding coefficients in $\{dX_t^\lambda\}$, where the weights are proportional to $\{\pi_t^\lambda\}$ and to the mixing density.*

Theorem 1 is first established in Brigo (2008, Corollary 1.3) limitedly to finite mixtures and 1-dimensional diffusions. The proof of Theorem 1 in Appendix A is more direct and extends the result to the required multivariate setting. In Section 3.1 we introduced diffusion bridges mapping arbitrary initial values X_0 to arbitrary final values X_τ. We now proceed to apply Theorem 1 to mix diffusion bridges and obtain a diffusion process for which (X_0, X_τ) is distributed according to a generic bivariate distribution $\Pi_{0,\tau}$. That is, we apply Theorem 1 to the collection of diffusion bridges (7) indexed by their initial and terminal values, $\lambda = (x_0, x_\tau)$, $\Lambda = \mathbb{R}^D \times \mathbb{R}^D$, with mixing distribution $\mathcal{L}(d\lambda) = \Pi_{0,\tau}(dx_0, dx_\tau)$. Then Theorem 1 yields the following SDE on $[0, \tau]$ with initial-terminal distribution $\Pi_{0,\tau}$

$$\begin{aligned} dX_t &= \mu(X_t, t)dt + g(X_t, t)dW_t, \\ \mu(x_t, t) &= f(x_t, t) + G(x_t, t) \underbrace{\int \nabla_{x_t} \ln p_{\tau|t}(x_\tau|x_t) \frac{p_{t|0,\tau}(x_t|x_0, x_\tau)}{\pi_t(x_t)} \Pi_{0,\tau}(dx_0, dx_\tau)}_{A(x_t,t)}, \\ \pi_t(x_t) &= \int p_{t|0,\tau}(x_t|x_0, x_\tau)\Pi_{0,\tau}(dx_0, dx_\tau). \end{aligned} \tag{8}$$

In (8), $A(x_t, t)$ gives the multiplicative drift adjustment factor for (6). While our most general construction allows for dependent X_0 and X_τ, we can achieve an independent transport via a factorial distribution $\Pi_{0,\tau} = \Pi_0 \otimes \Pi_\tau$. Another case of particular interest occurs when Π_0 puts all the mass on a single value x_0. In the following we refer to $A(x_t, t, x_0)$ in stance of $A(x_t, t)$, and to $\pi_{t|0}(x_t|x_0)$ in stance of $\pi_t(x_t)$, when it is necessary to distinguish this specific case. Having defined in (8) a processes with initial-terminal distribution $\Pi_{0,\tau}$ we extend the scope of Π to indicate the law of this process. Indeed, we have already denoted with π_t in (8) its marginal density. The transport of interest from $\mathcal{P}_Z$ to $\mathcal{P}_\mathcal{D}$ is achieved by $\Pi_\tau = \mathcal{P}_\mathcal{D}$ and $\mathcal{P}_Z = \Pi_0$ ($\mathcal{P}_Z$ can be arbitrarily defined). As $\mathcal{P}_\mathcal{D}$ is an empirical distribution, the integral in (8) with respect to x_τ reduces to averages over $\mathcal{D}$.

4 SDE Class

The starting point of the proposed transport is the unconstrained SDE (6). In this Section we define SDEs which are realized through a time-change of simpler SDEs and which are general enough to subsume the SDEs introduced in Song et al. (2021). Consider the D-dimensional SDEs

$$dZ_t = \Gamma^{1/2} dW_t, \tag{9}$$

$$dZ_t = \alpha_t Z_t dt + \Gamma^{1/2} dW_t, \tag{10}$$

where $\alpha_t \neq 0$ is a scalar function and $G(X_t, t) = \Gamma$ introduces an arbitrary covariance structure. (9) is the SDE of a correlated and scaled Brownian motion and (10) is the SDE of an Ornstein-Uhlenbeck process driven by a correlated and scaled Brownian motion. The transition densities of (9) and (10) are Gaussian (Appendix B). We denote both with $\widetilde{p}_{t'|t}$, informally (9) is a special case of (10) with $\alpha_t = 0$. SDE (10) in the time-homogenous case $\alpha_t = -1/2$ has stationary distribution $\mathcal{N}_D(0, \Gamma)$. We now introduce the time-change. Let $\beta_t > 0$ be a continuous function on $[0, \tau]$. Then $b_t = \int_0^t \beta_u du$ defines a monotonically (strictly) increasing function $b_t : [0, \tau] \to [0, b_\tau]$. The following SDEs on $[0, \tau]$ represent the class of dynamics for (1) and (6) on which we focus in the rest of this paper

$$dX_t = \sqrt{\beta_t}\Gamma^{1/2} dW_t, \tag{11}$$

$$dX_t = \alpha_t \beta_t X_t dt + \sqrt{\beta_t}\Gamma^{1/2} dW_t, \tag{12}$$

and $G(x, t) = \beta_t \Gamma$. The standard time-change result for diffusions (Øksendal, 2003, Theorem 8.5.1) establishes that the processes X_t respectively from (11) and (12) are equivalent in law to their time-scaled counterparts Z_{b_t} from (9) and (10). That is, SDEs (11) and (12) correspond to the evolution of the simpler SDEs (9) and (10) under a non-linear time wrapping where time flows with instantaneous intensity β_t. For both (11) and (12) the time-change argument yields $p_{\tau|t}(y|x) = \widetilde{p}_{b_\tau|b_t}(y|x)$ for the transition density of (6), and equivalently for the transition density $q_{\tau|t}$ of (1). We thus obtain

$$p_{\tau|t}(x_\tau|x_t) = \mathcal{N}_D(x_\tau;\, x_t a(t, \tau),\, \Gamma v(t, \tau)) \tag{13}$$

for appropriate scalar functions $a(t, \tau), v(t, \tau)$ with $v(t, \tau) > 0$ (Appendix B). By direct computation

$$\nabla_{x_t} \ln p_{\tau|t}(x_\tau|x_t) = \Gamma^{-1} \left(\frac{x_\tau}{a(t, \tau)} - x_t \right) \frac{a^2(t, \tau)}{v(t, \tau)}, \tag{14}$$

$$\nabla_{x_\tau} \ln p_{\tau|t}(x_\tau|x_t) = \Gamma^{-1}\left(x_t a(t,\tau) - x_\tau\right)\frac{1}{v(t,\tau)}. \tag{15}$$

From Bayes theorem and the Markov property we have

$$p_{t|0,\tau}(x_t|x_0,x_\tau) = \mathcal{N}_D\left(x_t;\, x_0\underline{a}_{\mathrm{br}}(0,t,\tau) + x_\tau\overline{a}_{\mathrm{br}}(0,t,\tau),\, \Gamma v_{\mathrm{br}}(0,t,\tau)\right), \tag{16}$$

where once again $\underline{a}_{\mathrm{br}}(0,t,\tau)$, $\overline{a}_{\mathrm{br}}(0,t,\tau)$ and $v_{\mathrm{br}}(0,t,\tau) > 0$ are scalar functions given in Appendix B. Finally, once again by direct computation

$$\nabla_{x_t} p_{t|0,\tau}(x_t|x_0,x_\tau) = \Gamma^{-1}\frac{x_0\underline{a}_{\mathrm{br}}(0,t,\tau) + x_\tau\overline{a}_{\mathrm{br}}(0,t,\tau) - x_t}{v_{\mathrm{br}}(0,t,\tau)}. \tag{17}$$

These results provide all the analytical formulas required for the computation of the adjustment factors $A(x_t,t)$, $A(x_t,t,x_0)$ and of the training objectives used to approximate them (Section 6).

4.1 Interpretation of Denoising Time-Reversed SDEs

Song et al. (2021) introduces two specifications of (1), named VESDE and VPSDE, which are respectively given by

$$dX_u = \sqrt{\beta_{\mathrm{ve},u}}dW_u, \tag{18}$$

$$dX_u = -\frac{1}{2}\beta_{\mathrm{vp},u}X_u dt + \sqrt{\beta_{\mathrm{vp},u}}dW_u. \tag{19}$$

See Appendix B for the functional form of $\beta_{\mathrm{ve},u}$ and $\beta_{\mathrm{vp},u}$. We thus recover (18) and (19) from (11) and (12) with $\Gamma = I$ and $\alpha_t = -1/2$. That is, VESDE and VPSDE correspond to a time change of the much simpler SDEs for the standard Brownian motion and for the standard Langevin SDE

$$dZ_u = dW_u,$$

$$dZ_u = -\frac{1}{2}Z_u du + dW_u.$$

5 Unified View of Drift Adjustments

The linearity of SDEs (11) and (12), underlying our and Song et al. (2021) work, has the important consequence that (14) and (15) are linear in x_t. This in turn allow us to derive an alternative representation for the drift adjustment in (8). Indeed, substituting (14) in (8) gives (Appendix A)

$$G(x,t)A(x,t) = \beta_t\left(\frac{1}{a(t,\tau)}\mathop{\mathbb{E}}_{X_\tau\sim\Pi_{\tau|t}(dX_\tau|x)}[X_\tau] - x\right)\frac{a^2(t,\tau)}{v(t,\tau)}. \tag{20}$$

Similarly, for the time-reversal drift adjustment term in (3) we have (Appendix A)

$$G(x,u)\,\nabla_x \ln q_u(x) = \beta_u\left(a(0,u)\mathop{\mathbb{E}}_{X_0\sim Q_{0|u}(dX_0|x)}[X_0] - x\right)\frac{1}{v(0,u)}. \tag{21}$$

The relations (20) and (21) provide an unified view of the inner workings of the DTRT and of the DBMT targeting $\mathcal{P}_{\mathcal{D}}$. In the following we always refer to sampling time t. Let $\Delta t = \tau - t = u$ represent the remaining sampling time. For ease of exposition we assume $\beta_t = 1$, as shown in Section 4 the term β_t corresponds to a time warping of an underlying time scale. To start with, the expectation terms in (20) and (21), to which we commonly refer as $E[Y|x,t]$, are always convex linear combinations of the samples $x^{(n)}$ from $\mathcal{D}$. The weights $\omega(x,t)^{(n)}$, $n = 1,\ldots,N$, of the linear combination, i.e. the discrete probabilities defining $E[Y|x,t]$, are given by the probabilities, under the assumed stochastic process, of reaching each state $x^{(n)}$ at terminal time τ from x at time t. The term $a(t,\tau)$ gives the "integrated scaling" on Δt, which goes to 1 as $\Delta t \to 0$, and is not critically important. In the case of (11) it holds that $a(t,\tau) = 1$. The term $v(t,\tau)$ gives the "integrated variance" on Δt, which goes to 0 as $\Delta t \to 0$. In the case of (11), $v(t,\tau) = \tau - t$ and the same remains true asymptotically for (12) as $\Delta t \to 0$. In summary, both drift adjustments (20) and (21) are essentially of the form $(E[Y|x,t] - x)v_{\Delta t}^{-1}$ where the term $v_{\Delta t}^{-1}$ diverges as $\Delta t \to 0$.

By construction, the initial weights entering expectation (20) are all equal to $1/N$ when X starts from a fixed value x_0, and are so on average when X_0 is stochastic. The initial weights entering expectation

(21) are on average approximately equal to $1/N$, depending on the quality of the approximation $\mathcal{P}_Z \approx Q_\tau$. Thus $E[Y|X_0, 0]$ is an averaging of many samples $x^{(n)}$. As time progresses, changes in X_t correspond to changes in $E[Y|X_t, t]$ through changes in the weights $\omega(x,t)^{(n)}$. Eventually all mass concentrates on a single weight $\omega(x,t)^{(*)}$ corresponding to a sample $x^{(*)}$. Ultimately, the attractor dynamics implied by $(E[Y|x,t] - x)v_{\Delta t}^{-1}$ drive X_t to $x^{(*)}$. We provide an illustration in Figure 1 (top). From here onward $\mathcal{D}(\text{CIFAR})$ stands for the training portion of the CIFAR10 dataset (images are linearly scaled to $[0,1]$), and Euler(T) corresponds to the Euler scheme (Kloeden & Platen, 1992) applied with T discretization steps.

In the VESDE and VPSDE of Song et al. (2021) we have $G(x,u) = \beta_u I$ and reversing (21) gives

$$\mathop{\mathbb{E}}_{X_0 \sim Q_{0|u}(dX_0|x)}[X_0] = \frac{v(0,u)\,\nabla_x \ln q_u(x) + x}{a(0,u)}. \tag{22}$$

We can thus take a trained score model $s_\phi(x,u) \approx \nabla_x \ln q_u(x)$, plug it in (22), and verity the extend to which $E[Y|x,t]$ has been approximated. We provide a first investigation in Figure 1 (bottom).

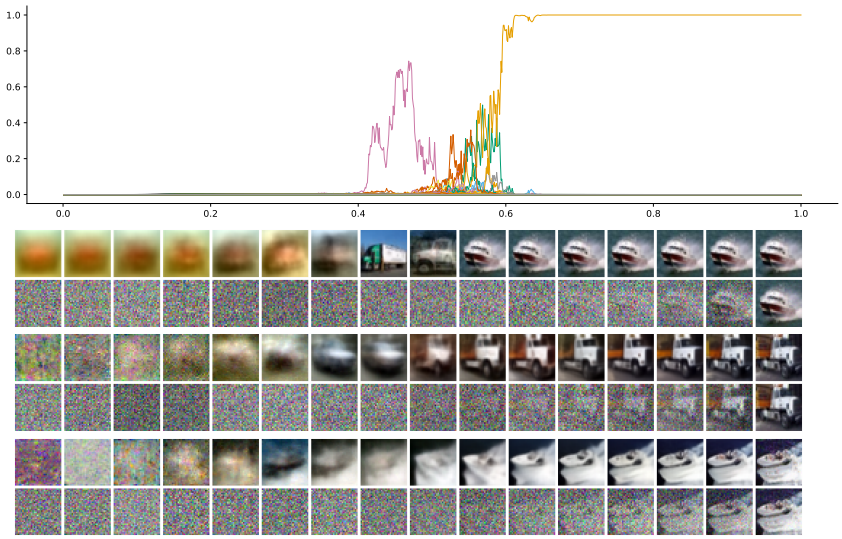

Figure 1: X follows (3) over sampling time t, VPSDE model, $X_0 \sim \mathcal{P}_Z^{\text{vp}}$; (top): evolution of $\omega(X_t,t)^{(n)}$ for all samples $x^{(n)}$ in $\mathcal{D}(\text{CIFAR})$, true score, Euler(1000); (bottom): $E[Y|X_t,t]$ and X_t (interleaved rows) for: true score, Euler(1000) (1st and 2nd rows, correspond to (top)); trained score model, Euler(1000) (3rd and 4th rows); trained score model, Euler(100) (5th and 6th rows).

We pause for a moment to review the findings of Figure 1 (see Appendix D for additional related plots). First of all, we can classify the dynamics of $E[Y|X_t,t]$ and of the associated weights in three stages. In the 1st stage the weights do not move much. During the 2nd stage, roughly $t \in [0.4, 0.6]$, the weights' mass gets distributed over a limited number of samples. Interestingly, the weights' dynamics are not monotonic. As time progresses the weights' mass shifts to different objects from different classes. From the beginning of the 3rd stage all mass is allocated to a single weight, the terminal image gets decided well in advance of the terminal time. These dynamics are suboptimal. We would like to shorten the 1st stage but it is associated with large values of β_t (i.e. quick time passing) which are required to decouple Q_τ from $\mathcal{P}_\mathcal{D}$. This is an intrinsic limitation of time-reversal approaches. It is also dubious that (partially) sampling multiple objects over t is beneficial for efficient generative modeling when we make use only of the terminal sample. This issue applies to trained models as well, as the 3rd row of Figure 1 shows. An interesting open question is how to obtain more suitable dynamics, where perhaps class transitions happen rarely. Secondly, $E[Y|X_t,t]$ provides a denoised representation of X_t across the whole 3rd stage. An alternative to the noise removal step applied to X_τ in Song et al. (2021) is to consider $E[Y|X_\tau,\tau]$ as the terminal sample instead. Thirdly, Figure 1 makes it clear that lowering the number of discretization steps affects generative sampling in multiple ways. On the one hand the terminal sample X_τ is more noisy. This is not very surprising: close to τ the drift adjustment is approximately $(x^{(*)} - X_t)(\tau - t)^{-1}$, for a sample $x^{(*)}$, which is the

drift of a Brownian bridge. Bridge sampling is notoriously problematic (Bladt et al., 2016). On the other hand larger discretization errors significantly affect the dynamics of $E[Y|X_t, t]$ resulting in less coherent samples. We remark that none of these insights could have been gained by observing X_t alone, i.e. the odd rows of Figure 1 (bottom). To conclude, (20) and (21) give an additional meaning to "denoising". Neural network approximators need to map from a noisy input X_t to an adjustments toward a smoother superimposition of samples. The desire to minimize the discrepancy between the smoothness properties of X_t and of $\mathbb{E}[Y|X_t, t]$ motivates the developments of Section 7.

6 Transports Approximation

As in Song et al. (2021), computing the multiplicative drift adjustment $A(x_t, t)$ requires $\mathcal{O}(N)$ operations. In this Section we introduce three training objectives for which unbiased and scalable, i.e. $\mathcal{O}(1)$ with respect to N, MC estimators can be immediately derived.

The first training objective applies only to $A(x_t, t, x_0)$. It relies on the identity (Appendix A)

$$A(x_t, t, x_0) = \nabla_{x_t} \ln \pi_{t|0}(x_t|x_0) - \nabla_{x_t} \ln p_{t|0}(x_t|x_0). \tag{23}$$

It is advantageous to consider the right hand side of (23) because from (8) we know that $\pi_{t|0}(x_t|x_0)$ has mixture representation. As in Song et al. (2021), we can rely on Vincent (2011) to obtain a scalable objective to train a neural network approximator $s_\phi(x_t, t) \approx \nabla_{x_t} \ln \pi_{t|0}(x_t|x_0)$, i.e.

$$\begin{aligned}\mathbb{L}_{\text{FD,DBMT}}(\phi) &= \mathop{\mathbb{E}}_{t\sim\mathcal{U}(0,\tau), X_t\sim\Pi_{t|0}} \left[r_t \left\| \nabla_{X_t} \ln \pi_{t|0}(X_t|x_0) - s_\phi(X_t, t) \right\|^2 \right] \\ &= \mathop{\mathbb{E}}_{t\sim\mathcal{U}(0,\tau), (X_t,X_\tau)\sim\Pi_{t,\tau|0}} \left[r_t \left\| \nabla_{X_t} \ln p_{t|0,\tau}(X_t|x_0, X_\tau) - s_\phi(X_t, t) \right\|^2 \right],\end{aligned} \tag{24}$$

where $r_t = \mathbb{E}[\|\nabla_{X_t} \ln p_{t|0,\tau}(X_t|x_0, X_\tau)\|^2]^{-1}$ is a regularization term.

The remaining training objectives rely on the identities (20) and (21). The goal is directly approximate the expectations of (20) and (21) which, as in Section 5, we denote with a generic $\mathbb{E}[Y|x, t]$. That is, we aim to train a neural network approximator $s_\phi(x, t) \approx \mathbb{E}[Y|x, t]$. As conditional expectations are mean squared error minimizers, suitable objectives for the expectation terms of (20) and (21) are

$$\mathbb{L}_{\text{CE,DBMT}}(\phi) = \mathop{\mathbb{E}}_{t\sim\mathcal{U}(0,\tau], (X_t,X_\tau)\sim\Pi_{t,\tau}} \left[\left\| X_\tau - s_\phi(X_t, t) \right\|^2 \right], \tag{25}$$

$$\mathbb{L}_{\text{CE,DTRT}}(\phi) = \mathop{\mathbb{E}}_{u\sim\mathcal{U}(0,\tau], (X_0,X_u)\sim Q_{0,u}} \left[\left\| X_0 - s_\phi(X_u, u) \right\|^2 \right]. \tag{26}$$

In Table 1 we summarize the operations needed to implement the plain MC estimators for the four objectives considered in this work. We reference where to find the required quantities for SDEs (11) and (12). The MC estimators for the Fisher divergence losses $\mathbb{L}_{\text{FD},*}$ involve multiplications by Γ^{-1} (by (15) and (17)). Moreover, computing the drift adjustment at generation time requires multiplications by Γ. In Section 7 we discuss how to manage the computational burden. An appealing property of $\mathbb{L}_{\text{CE},*}$ is that computing the drift adjustment only requires the application of simple scalar functions (see (20) and (21)), and that their MC estimators only requires sampling operations. A further advantage of $\mathbb{L}_{\text{CE},*}$ is that no regularization is required. In contrast, in the absence of regularization terms, $\mathbb{L}_{\text{FD},*}$ are divergent for $t \approx \tau$ due to the term $v_{\Delta t}^{-1}$ discussed in Section 5.

$\mathbb{L}$	Sampling ($t,u\sim\mathcal{U}(0,\tau]$)	Evaluation
$\mathbb{L}_{\text{FD,DTRT}}$	$X_0\sim\mathcal{P}_\mathcal{D}$, $X_u\sim Q_{u\|0}(dX_u\|X_0)$(13)	$\nabla_{X_u} \ln q_{u\|0}(X_u\|X_0)$(15)
$\mathbb{L}_{\text{FD,DBMT}}$	$X_\tau\sim\mathcal{P}_\mathcal{D}$, $(X_0{=}x_0)$, $X_t\sim P_{t\|0,\tau}(dX_t\|x_0,X_\tau)$(16)	$\nabla_{X_t} p_{t\|0,\tau}(X_t\|X_0,X_\tau)$(17)
$\mathbb{L}_{\text{CE,DTRT}}$	$X_0\sim\mathcal{P}_\mathcal{D}$, $X_u\sim Q_{u\|0}(dX_u\|X_0)$(13)	
$\mathbb{L}_{\text{CE,DBMT}}$	$X_\tau\sim\mathcal{P}_\mathcal{D}$, $X_0\sim\Pi_{0\|\tau}(dX_0\|X_\tau)$, $X_t\sim P_{t\|0,\tau}(X_T\|X_0,X_\tau)$(16)	

Table 1: Sampling and evaluation operations required to implement the proposed MC estimators.

7 NON-DENOISING DIFFUSIONS

In computer vision applications, color images of resolution $H \times W$ have $S = HW$ pixels and correspond to $D = 3S$. Even in the case of datasets featuring images with the very modest resolution of 32×32 pixels, such as CIFAR10, the resulting dimensionality is already $D = 3072$. A direct modeling approach allowing an arbitrary covariance matrix Γ in (11) and (12) would require to store a $D \times D$ matrix Γ, and perform its Cholesky (or equivalent) decomposition with cost $\mathcal{O}(D^3)$. As the resolution increases the computational burden gets intractable very quickly. Indeed, to the best of the authors knowledge, all prior DDPM literature only considers independent transitions, that is $\Gamma = I$.

We propose to view SDEs (11) and (12) as corresponding to the space discretization on a $H \times W$ grid of a spatio-temporal process defined over the spatial domain $[0, 1]^2$. This different perspective can be formalized (Prato & Zabczyk, 2014) giving raise to a stochastic partial differential equation (SPDE) featuring derivatives with respect to time and space (despite the misleading nomenclature, both SDEs and SPDEs solutions are actually nowhere differentiable with respect to time, and often with respect to space, and are interpreted as solutions to stochastic integral equations). Doing this rigorously is not trivial, but for our needs an heuristical application of the principle suffices. Consider the Euler discretization of (12): $X_{t+\Delta t} = X_t + \alpha_t \beta_t X_t \Delta t + \sqrt{\Delta t} \mathcal{E}_t$, where $\mathcal{E}_t \sim \mathcal{N}_D(0, \Gamma)$. The idea is adopt a functional perspective: $X(t + \Delta t, s) = X(t, s) + \alpha_t \beta_t X(t, s) \Delta t + \sqrt{\Delta t} \mathcal{E}(t, s)$ where $s \in [0, 1]^2$ defines space coordinates. That is, both $X(t)$ and $\mathcal{E}(t)$ at each time t are random processes over $[0, 1]^2$. We assume the innovations $\mathcal{E}(t)$ to be distributed as a Gaussian process (GP) for each t.

To start with, it is necessary to model the GP. For simplicity, we assume a factorial distribution over the channels and an isotopic stationary covariance function. We rely on the semi-variogram approach (Cressie, 1993) to compare how different covariance functions fit $\mathcal{D}(\text{CIFAR})$. A semi-variogram is a measure of dependency across space. In the case of isotropic stationary covariances it simplifies to a scalar function of the Euclidean distance between points: $\gamma(\|\Delta s\|) = \mathbb{E}[(\Delta x_s)^2]/2$ with $\Delta x_s = x_{s+\Delta s} - x_s$. The rate of decrease of $\gamma(\|\Delta s\|)$ toward 0 as $\|\Delta s\| \to 0$ gives a measure of the strength of the spatial dependency, i.e. the smoothness of the spatial process. Semi-variograms corresponding to different covariance functions are here fitted to their empirical counterparts via a weighted minimum-least-squares procedure (Cressie, 1993). Figure 2 illustrates the exponential and RBF semi-variogram fits for two images of $\mathcal{D}(\text{CIFAR})$. The exponential covariance, which corresponds to rougher paths, provides a much better fit to the shown samples. This results is consistent across $\mathcal{D}(\text{CIFAR})$. We remark that $\Gamma = I$ corresponds to a pure white-noise process with a perfectly flat semi-variogram which would clearly result in a very poor fit to the empirical semi-variograms shown in Figure 2. Based on these findings, we model the innovations of each image channel as a GP with exponential covariance function with length-scale $\theta = 0.205$, the median estimated value (Figure 2 (right)). We match the marginal variance to that of $\mathcal{D}(\text{CIFAR})$, $\sigma^2 = 0.063$.

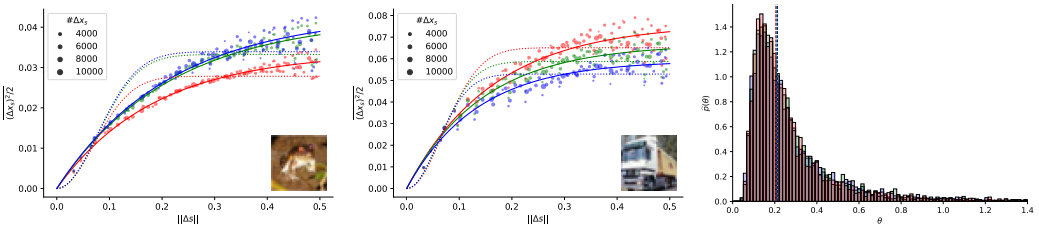

Figure 2: Empirical semi-variograms (dots) and fitted exponential (solid lines) and RBF (doted lines) semi-variograms for the 1st (left) and 2nd (center) image of $\mathcal{D}(\text{CIFAR})$; histograms (bins) and medians (lines) of the distributions of the length-scale parameters in the exponential variogram model over $\mathcal{D}(\text{CIFAR})$ (right); colors represent the RGB channels.

As the GPs $\mathcal{E}(t)$ are defined on a 2D domain we can leverage on scalable inference techniques fom spatial statistics. As an example, we consider the circulant embedding method (CEM) (Wood & Chan, 1994; Dietrich & Newsam, 1997) which exploits a connection with the fast Fourier transform (FFT). See Appendix C for a cursory review of the CEM. Consider a $H \times W$ uniform grid $\mathcal{S}$ of size $S = HW$ discretizing $[0, 1]^2$, i.e. the support of images. For a stationary covariance function, such as the exponential covariance, the CEM samples $\mathcal{E}(t)$ on $\mathcal{S}$ with cost $\mathcal{O}(S \ln(S))$. This is close

to the $\mathcal{O}(S)$ cost of sampling from a pure white-noise process, and compares very favourably to the $\mathcal{O}(S^3)$ cost that a Cholesky decomposition entails. One limitation of CEM is that generated samples, while always Gaussian, might not have the correct covariances. Whether this happens, and in that case the quality of the approximation, depends on the covariance function. For the estimated exponential covariance function with length-scale $\theta = 0.205$ sampling is exact. Figure 3 shows samples from a pure white-noise GP, i.e. $\Gamma = I$, (1st row) and from the estimated GP using CEM (2nd row). As noted in Section 6, sampling is enough to implement the MC estimators for $\mathbb{L}_{\text{CE},*}$, but the MC estimators and drift adjustments for $\mathbb{L}_{\text{CE},*}$ involve additional matrix multiplications by Γ and Γ^{-1}. The CEM allows to compute these at the same $\mathcal{O}(S \ln(S))$ cost if we define the GP $\mathcal{E}(t)$ on a 2D torus (Rue & Held, 2005, Chapter 2.1). This corresponds to introducing dependencies between "opposing" boundaries of $[0, 1]^2$. Figure 3 (3rd row) shows some samples, in the highlighted patch the opposing-boundaries dependency is evident. Either way, all samples from the 2nd and 3rd rows of Figure 3 share smoothness properties matching the empirical findings of Figure 2.

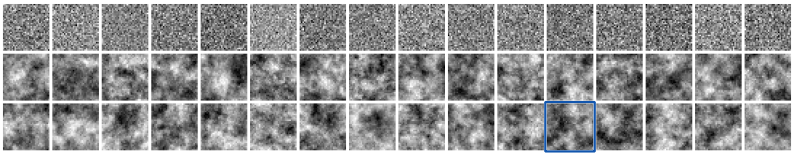

Figure 3: Spatial GP samples, see the main text for the description.

8 Final Considerations

The DBMT construction of Section 3 is general, and applies to essentially arbitrary initial-terminal distributions $\Pi_{0,\tau}$ and SDE coefficients $f(\cdot)$, $\sigma(\cdot)$. The SDE class considered in Section 4 is tractable as it results in linear in the narrow-sense (Kloeden & Platen, 1992) SDE bridges. Within this SDE class, the time-space factorization of the diffusion coefficient $g(x, t) = \sqrt{\beta_t}\Gamma^{1/2}$ separates modeling concerns. β_t corresponds to a non-linear time-wrapping which impacts the transport's dynamics. An analysis for the DTRT of Song et al. (2021) is carried out in Section 5. As shown in Section 7, Γ can be efficiently modeled by fitting the micro-scale properties of $\mathcal{P}_{\mathcal{D}}$. The spatial modeling could definitely be improved by introducing across-channel dependencies. The widespread availability of GPU-accelerated FFT implementations stands behind our focus on the CEM. Within this context, we stress the importance of the connection with spatio-temporal modeling. Alternative scalable approaches abound, from Gaussian Markov Random Fields (Rue & Tjelmeland, 2002; Rue, 2001) to Karhunen–Loève expansions (Betz et al., 2014).

Π_τ needs to match $\mathcal{P}_{\mathcal{D}}$, but the choice of $\Pi_{0|\tau}$ has so far received little attention. An interesting result from Song et al. (2021) is the derivation of a random ordinary differential equation (RODE) matching the marginal distribution of a given SDE. This result requires the SDE to start from an initial distribution with a density. From this perspective, a non degenerate Π_0 is preferable as the application of the RODE approach leads to faster sampling and to likelihood evaluation. A natural implementation is then $\Pi_0 = \mathcal{N}_D(0, \Gamma)$ for (12) with $\alpha_t = 1/2$ which preserves Π_0. We now consider a DBMT starting from a fixed value x_0, i.e. $\Pi_0 = \delta_{x_0}$. In this case, the goal of minimizing discretization errors suggests to choose the initial value that cancels the drift adjustment at $t = 0$. This is achieved by $x_0 = 1/N \sum_{n=1}^{N} x^{(n)} a(0, \tau)^{-1}$, see (20). Finally, a dependent $\Pi_{0|\tau}$ can reduce the effort required to transport samples from Π_0 to Π_τ. As an example, we could first sample a length-scale from the histogram of Figure 2 (right), and then condition both Π_0 and Π_τ to match this length-scale.

This paper has been entirely dedicated to the methodological study of generative modeling through diffusion processes. The next logical step is to apply the insights contained in this work to the actual training of neural networks. Three novel training objectives have been proposed in Section 6, two of which offering desireable properties compared to the Fisher divergence objective of Song et al. (2021), especially in the case of non-factorial transitions. We argue that the increased understanding of the target mappings of (20) and (21) can, and should, guide the development of neural network architectures more closely matching the target structure compared to the U-Net default choice.

`https://github.com/?` links to the code accompanying this paper which is made available under the `MIT` license (link to be shared with a comment directed to the reviewers and area chairs).

REFERENCES

Brian D.O. Anderson. Reverse-Time Diffusion Equation Models. *Stochastic Processes and their Applications*, 12(3):313–326, May 1982.

Wolfgang Betz, Iason Papaioannou, and Daniel Straub. Numerical Methods for the Discretization of Random Fields by Means of the Karhunen–Loève Expansion. *Computer Methods in Applied Mechanics and Engineering*, 271:109–129, April 2014.

Mogens Bladt, Samuel Finch, and Michael Sørensen. Simulation of Multivariate Diffusion Bridges. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(2):343–369, 2016.

Damiano Brigo. The General Mixture-Diffusion SDE and Its Relationship with an Uncertain-Volatility Option Model with Volatility-Asset Decorrelation, December 2008.

Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling, June 2021.

C. R. Dietrich and G. N. Newsam. Fast and Exact Simulation of Stationary Gaussian Processes through Circulant Embedding of the Covariance Matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, July 1997.

U. G. Haussmann and E. Pardoux. Time Reversal of Diffusions. *The Annals of Probability*, 14(4): 1188–1205, October 1986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Number 113 in Graduate Texts in Mathematics. Springer, New York, 2nd ed edition, 1996. ISBN 978-0-387-97655-6 978-3-540-97655-4.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1992. ISBN 978-3-642-08107-1 978-3-662-12616-5.

NV Krylov. *Introduction to the Theory of Diffusion Processes*, volume 142. Providence, 1995.

Annie Millet, David Nualart, and Marta Sanz. Integration by Parts and Time Reversal for Diffusion Processes. *The Annals of Probability*, pp. 208–238, 1989.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. Second edition edition, 2014.

L Chris G Rogers and David Williams. *Diffusions, Markov Processes and Martingales: Volume 2: Itô Calculus*, volume 2. Cambridge university press, 2000.

Havard Rue. Fast Sampling of Gaussian Markov Random Fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2):325–338, 2001.

Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. Number 104 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, 2005. ISBN 978-1-58488-432-3.

Hååvard Rue and Hååkon Tjelmeland. Fitting Gaussian Markov Random Fields to Gaussian Fields. *Scandinavian Journal of Statistics*, 29(1):31–49, 2002.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, first edition, April 2019. ISBN 978-1-108-18673-5 978-1-316-51008-7 978-1-316-64946-6.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011.

Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep Generative Learning via Schrödinger Bridge, June 2021.

Andrew T. A. Wood and Grace Chan. Simulation of Stationary Gaussian Processes in [0,1]d. *Journal of Computational and Graphical Statistics*, 3(4):409–432, 1994.

B. K. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, Berlin ; New York, 6th ed. edition, 2003. ISBN 978-3-540-04758-2.

A THEORETICAL FRAMEWORK

A.1 ASSUMPTIONS

Assumption 1 (SDE solution)**.** *A given D-dimensional SDE(f, g) with associated initial distribution $\mathcal{V}_0$ and integration interval $[0, \tau]$ admits an unique strong solution on $[0, \tau]$.*

Assumption 1 can be checked through the application of the standard existence and uniqueness theorems for SDE solutions. Of particular relevance to our setting is the formulation of Krylov (1995, Chapter 5, Theorem 1) that limits the monotonicity requirement to, informally speaking, drifts that pull the process toward infinities.

Assumption 2 (SDE density)**.** *A given D-dimensional SDE(f, g) with associated initial distribution $\mathcal{V}_0$ and integration interval $[0, \tau]$ admits a marginal / transition density on $(0, \tau)$ with respect to the D-dimensional Lebesgue measure that uniquely satisfies the Fokker-Plank / Kolmogorov-Forward partial differential equation (PDE).*

We refer to Särkkä & Solin (2019, Chapter 5) and to Karatzas & Shreve (1996, Chapter 5.7) for connections between SDEs and PDEs.

All theoretical results of this work rely on simple algebraic manipulations and re-arrangements of quantities of interest. The main complication stems from the need to justify differentiation and integration exchanges, i.e. exchange of limits.

Assumption 3 (exchange of limits)**.** *We assume that limits exchanges are justified in the steps marked with $(\star)$ and $(\star\star)$.*

Similarly, various steps in the derivations involve considering fractional quantities with densities appearing in the denominators.

Assumption 4 (positivity)**.** *For a given stochastic process, all finite-dimensional densities, conditional or not, are strictly positive.*

Assumption 4 is easy to verify. We resorted to the practical but somewhat unsatisfactory formulation of Assumption 3 because in full generality it is complicated to give easily checkable conditions. We just note that when $\Pi_T = \mathcal{P}_{\mathcal{D}}$, the limit exchange marked with $(\star)$ is always justified. So are the limits exchanges marked with $(\star\star)$ when in addition Π_0 puts all the mass to a fixed initial value, or (by direct verification) when Π_0 is Gaussian for the SDE class of Section 4. Thorough this paper, both in the main text and in the proofs that follow, it is supposed that Assumptions 1, 2 and 4 are satisfied by SDEs (1), (6) and (7). This is the case for the SDE class of Section 4, i.e. (11) and (12), for any Π_0 with finite variance.

Remark: For ease of exposition it is assumed thorough this paper that all diffusions take values in the state space $\mathbb{R}^D$. There is no impediment in extending the presented results to the case of diffusions taking values in a subset $\mathcal{X} \subset \mathbb{R}^D$. The obvious changes to Assumptions 1 to 4 apply, the proofs carry over without substantial modifications. This extension could be of practical interest as images are often represented as floating point values in $[0, 1]$.

A.2 STATEMENT AND PROOF OF MIXTURE OF DIFFUSIONS THEOREM

Theorem 2 (Mixture of diffusions)**.** *Consider the family of D-dimensional SDEs on $t \in [0, \tau]$ indexed by $\lambda \in \Lambda$*

$$\begin{aligned} dX_t^\lambda &= \mu^\lambda(X_t^\lambda, t)dt + \sigma^\lambda(X_t^\lambda, t)dW_t^\lambda, \\ X_0^\lambda &\sim \mathcal{V}_0^\lambda, \end{aligned} \tag{27}$$

where the initial distributions $\mathcal{V}_0^\lambda$ and the BMs W_t^λ are all independent. Let $\nu_t^\lambda, t \in (0, \tau)$ denote the marginal density of X_t^λ. For a generic mixing distribution $\mathcal{L}$ on Λ, define the mixture marginal density ν_t for $t \in (0, \tau)$ and the mixture initial distribution $\mathcal{V}_0$ by

$$\nu_t(x) = \int_\Lambda \nu_t^\lambda(x)\mathcal{L}(d\lambda), \quad \mathcal{V}_0(dx) = \int_\Lambda \mathcal{V}_0^\lambda(dx)\mathcal{L}(d\lambda). \tag{28}$$

Consider the D-dimensional SDE on $t \in [0, \tau]$ defined by

$$\begin{aligned}
\mu(x,t) &= \frac{\int_\Lambda \mu^\lambda(x,t)\nu_t^\lambda(x)\mathcal{L}(d\lambda)}{\nu_t(x)}, \\
\sigma(x,t) &= \frac{\int_\Lambda \sigma^\lambda(x,t)\nu_t^\lambda(x)\mathcal{L}(d\lambda)}{\nu_t(x)}, \\
dX_t &= \mu(X_t,t)dt + \sigma(X_t,t)dW_t, \\
Y_0 &\sim \mathcal{V}_0.
\end{aligned} \tag{29}$$

It is assumed that all diffusion processes X^λ and the diffusion process X solving (29) satisfy the regularity assumptions Assumptions 1, 2 and 4 and that Assumption 3 holds. Then the marginal distribution of the diffusion X is ν_t.

Proof of Theorem 2. We start by establishing that the law of X is indeed given by the solution of (29). In this proof we make use of the following notation: for f scalar-valued $(f)_t = \frac{d}{dt}f$, for a vector-valued $(a)_x = \sum_{i=1}^D \frac{d}{dx_i}a$, for A matrix-valued $(A)_{xx} = \sum_{i,j=1}^D \frac{d^2}{dx_i dx_j}A$. This notation allows for a compact representation of PDEs reminiscent of the 1-dimensional setting. Then, for $0 < t < \tau$ we have that

$$\begin{aligned}
(\nu(x,t))_t &= \left(\int_\Lambda \nu^\lambda(x,t)\mathcal{L}(d\lambda)\right)_t \\
&= \int_\Lambda \left(\nu^\lambda(x,t)\right)_t \mathcal{L}(d\lambda) && (\star\star) \\
&= \int_\Lambda \left(\mu^\lambda(x,t)\nu^\lambda(x,t)\right)_x + \frac{1}{2}\left(\sigma^\lambda(x,t)\nu^\lambda(x,t)\right)_{xx}\mathcal{L}(d\lambda) \\
&= \int_\Lambda \left(\frac{\mu^\lambda(x,t)\nu^\lambda(x,t)}{\nu(x,t)}\nu(x,t)\right)_x + \frac{1}{2}\left(\frac{\sigma^\lambda(x,t)\nu^\lambda(x,t)}{\nu(x,t)}\nu(x,t)\right)_{xx}\mathcal{L}(d\lambda) \\
&= \left(\int_\Lambda \frac{\mu^\lambda(x,t)\nu^\lambda(x,t)}{\nu(x,t)}\mathcal{L}(d\lambda)\nu(x,t)\right)_x + \frac{1}{2}\left(\int_\Lambda \frac{\sigma^\lambda(x,t)\nu^\lambda(x,t)}{\nu(x,t)}\mathcal{L}(d\lambda)\nu(x,t)\right)_{xx}. && (\star\star)
\end{aligned}$$

The second line is an exchange of limits, the third line is the application of the Fokker-Plank PDEs for the collection of processes X^λ, the fourth line is a rewriting in terms of $\nu(y,t)$, the last line is another exchange of limits. The result follows by noticing that the last line gives the Fokker-Plank representation of (29). □

A.3 Drift Adjustments

Limitedly to this section, we lighten the notation by removing subscripts from probability measures and densities. The missing time points can be inferred without ambiguity from the variables.

A.3.1 Drift Adjustment Identities for Constant Initial Value

First identity:

$$\begin{aligned}
&\nabla_{x_t} \ln \int \frac{p(x_\tau|x_t)}{p(x_\tau|x_0)}\Pi(dx_\tau) \\
&= \int \frac{\nabla_{x_t} p(x_\tau|x_t)}{p(x_\tau|x_0)} p(x_t|x_0)\Pi(dx_\tau) \Big/ \int \frac{p(x_\tau|x_t)}{p(x_\tau|x_0)} p(x_t|x_0)\Pi(dx_\tau) && (\star) \\
&= \int \nabla_{x_t} \ln p(x_\tau|x_t) \frac{p(x_\tau, x_t|x_0)}{p(x_\tau|x_0)}\Pi(dx_\tau) \Big/ \int \frac{p(x_\tau, x_t|x_0)}{p(x_\tau|x_0)}\Pi(dx_\tau) \\
&= \int \nabla_{x_t} \ln p(x_\tau|x_t) p(x_t|x_0, x_\tau)\Pi(dx_\tau) \Big/ \pi(x_t|x_0) \\
&= A(x_t, t, x_0).
\end{aligned}$$

Second identity:

$$\begin{aligned}
&\nabla_{x_t} \ln \int \frac{p(x_\tau|x_t)}{p(x_\tau|x_0)} \Pi(dx_\tau) \\
&= \nabla_{x_t} \ln \int \frac{p(x_\tau|x_t)}{p(x_\tau|x_0)} p(x_t|x_0) \Pi(dx_\tau) - \nabla_{x_t} \ln p(x_t|x_0) \\
&= \nabla_{x_t} \ln \pi(x_t|x_0) - \nabla_{x_t} \ln p(x_t|x_0).
\end{aligned}$$

A.3.2 Drift Adjustments as Expectations

To establish (20) notice that from (14) we have

$$\begin{aligned}
&G(x_t,t)A(x_t,t) \\
&= \beta_t \Gamma \Gamma^{-1} \int \left(\frac{x_\tau}{a(t,\tau)} - x_t \right) \frac{a^2(t,\tau)}{v(t,\tau)} \frac{p(x_t|x_0,x_\tau)}{\pi(x_t)} \Pi_{0,\tau}(dx_0, dx_\tau) \\
&= \beta_t \left(\frac{1}{a(t,\tau)} \int x_\tau \frac{\pi(x_t|x_0,x_\tau)}{\pi(x_t)} \Pi_{0,\tau}(dx_0, dx_\tau) - x_t \right) \frac{a^2(t,\tau)}{v(t,\tau)} \\
&= \beta_t \left(\frac{1}{a(t,\tau)} \mathop{\mathbb{E}}_{X_\tau \sim \Pi(dX_\tau|x_t)} [X_\tau] - x_t \right) \frac{a^2(t,\tau)}{v(t,\tau)}.
\end{aligned}$$

To establish (21) note that from (15) we have

$$\begin{aligned}
G(x_u,u)\, \nabla_{x_u} \ln q(x_u) &= \Gamma \int \nabla_{x_u} \ln q(x_u|x_0) \frac{q(x_u|x_0)}{q(x_u)} \mathcal{P}_{\mathcal{D}}(dx_0) \\
&= \beta_u \Gamma \Gamma^{-1} \int \left(\frac{a(0,u)x_0 - x_u}{v(0,u)} \right) \frac{q(x_u|x_0)}{q(x_u)} \mathcal{P}_{\mathcal{D}}(dx_0) \\
&= \beta_u \left(a(0,u) \int x_0 \frac{q(x_u|x_0)}{q(x_u)} \mathcal{P}_{\mathcal{D}}(dx_0) - x_u \right) \frac{1}{v(0,u)} \\
&= \beta_u \left(a(0,u) \mathop{\mathbb{E}}_{X_0 \sim Q(dX_0|x_u)} [X_0] - x_u \right) \frac{1}{v(0,u)}.
\end{aligned}$$

B SDEs Class Formulas

The transition densities of (9) and (10) are given respectively by

$$\begin{aligned}
\widetilde{p}_{\mathrm{bm},\tau|t}(z_\tau|z_t) &= \mathcal{N}_D\left(z_\tau;\, z_t,\, \Gamma(\tau - t)\right), \\
\widetilde{p}_{\mathrm{ou},\tau|t}(z_\tau|z_t) &= \mathcal{N}_D\left(z_\tau;\, z_t e^{\overline{\alpha}_{t:\tau}(\tau - t)},\, \Gamma\left(\frac{1}{2\alpha_t} e^{2\overline{\alpha}_{t:\tau}(\tau-t)} - \frac{1}{2\alpha_\tau}\right)\right).
\end{aligned}$$

Here we used the notation $\overline{f}_{t:\tau} = \frac{1}{\tau - t}\int_t^\tau f_u du$, i.e. $\overline{f}_{t:\tau}$ is the average value of a function f_u on the interval $[t,\tau]$. The time-homogenous case of (10), where $\overline{\alpha}_{t:\tau} = \alpha$, is thus immediately recovered.

The scalar functions $a_{\mathrm{bm}}(t,\tau)$, $a_{\mathrm{ou}}(t,\tau)$, $v_{\mathrm{bm}}(t,\tau)$ and $v_{\mathrm{ou}}(t,\tau)$ are given by

$$\begin{aligned}
a_{\mathrm{bm}}(t,\tau) &= 1, & v_{\mathrm{bm}}(t,\tau) &= b_\tau - b_t, \\
a_{\mathrm{ou}}(t,\tau) &= e^{\overline{\alpha}_{b_t:b_\tau}(b_\tau - b_t)}, & v_{\mathrm{ou}}(t,\tau) &= \frac{1}{2\alpha_{b_t}} e^{2\overline{\alpha}_{b_t:b_\tau}(b_\tau - b_t)} - \frac{1}{2\alpha_{b_\tau}}.
\end{aligned}$$

The scalar functions $v_{\mathrm{br}}(0,t,\tau)$, $\underline{a}_{\mathrm{br}}(0,t,\tau)$ and $\overline{a}_{\mathrm{br}}(0,t,\tau)$ are given by

$$\begin{aligned}
v_{\mathrm{br}}(0,t,\tau) &= \frac{v(0,t)v(t,\tau)}{v(0,t)a^2(t,\tau) + v(t,\tau)}, \\
\underline{a}_{\mathrm{br}}(0,t,\tau) &= \frac{v(t,\tau)a(0,t)}{v(0,t)a^2(t,\tau) + v(t,\tau)}, \\
\overline{a}_{\mathrm{br}}(0,t,\tau) &= \frac{v(0,t)a(t,\tau)}{v(0,t)a^2(t,\tau) + v(t,\tau)}.
\end{aligned}$$

The scalar functions $\beta_{\mathrm{ve},u}$ and $\beta_{\mathrm{vp},u}$ are given by

$$\beta_{\mathrm{ve},u} = \sigma_{\min}^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2u} 2\log\frac{\sigma_{\max}}{\sigma_{\min}}, \quad \beta_{\mathrm{vp},u} = \left(\bar{\beta}_{\min} + u\left(\bar{\beta}_{\max} - \bar{\beta}_{\min}\right)\right).$$

The constants $\sigma_{\min}, \sigma_{\max}, \bar{\beta}_{\min}, \bar{\beta}_{\max}$ depend in part on the dataset considered, but are consistently chosen to have $\beta_{\mathrm{ve},u}, \beta_{\mathrm{vp},u}$ small for $u \approx 0$ and large for $u \approx \tau$.

The approximating distributions $\mathcal{P}_Z$ in Song et al. (2021) are $\mathcal{P}_Z^{\mathrm{ve}} = \mathcal{N}_D(0, I\sigma_{\max}^2)$ for VESDE, $\mathcal{P}_Z^{\mathrm{vp}} = \mathcal{N}_D(0, I)$ for VPSDE.

C Additional Material

C.1 Closely Related Work

A work closely related to the present paper is that of Wang et al. (2021) as it similarly avoids the time-reversal construction. Wang et al. (2021) construct a 2-stages diffusion process from a constant initial value x_0 to $\mathcal{P}_{\mathcal{D}}$ by relying on the theory of Schrödinger bridges. The most notable differences with respect to the DBMT transport are: (i) the dynamics considered in Wang et al. (2021) are less general, in our notation they correspond to $f(\cdot) = 0, g(\cdot) = \sigma I$ for a fixed scalar σ; (ii) the transport proposed in Wang et al. (2021) necessarily starts from x_0, the general result of (8) allows for (almost) arbitrary initial distributions and initial-terminal dependencies. For an initial x_0, and for the more limited dynamics considered in Wang et al. (2021), the achieved transport is the same. In this sense the DBMT generalizes the first stage diffusion of Wang et al. (2021). It is interesting to note that in the case of a constant x_0 the DBMT can also be obtained by an application of Doob h-transforms as we show in the following section.

C.2 Connection with Doob h-transforms

The previously established identity

$$A(x_t, t, x_0) = \nabla_{x_t} \ln \int \frac{p_{\tau|t}(x_\tau|x_t)}{p_{\tau|0}(x_\tau|x_0)} \Pi_\tau(dx_\tau),$$

shows that the drift adjustment can be equivalently expressed as

$$\mu(x_t, t) = f(x_t, t) + G(x_t, t)\, \nabla_{x_t} h(x_t, t), \quad h(x_t, t) = \ln \int \frac{p_{\tau|t}(x_\tau|x_t)}{p_{\tau|0}(x_\tau|x_0)} \Pi_\tau(dx_\tau),$$

as x_0 is a constant. It can be verified that the h function satisfies the required space-time regularity property (Särkkä & Solin, 2019, Eq. (7.73)). As such, it is a genuine Doob h-transform. That $p^h_{t'|t}(x_{t'}|x_t) = p_{t'|t}(x_{t'}|x_t) h(x_{t'}, t')/h(x_t, t)$ is the transition density of the DBMT transport from δ_{x_0} to Π_τ follows by direct computation.

C.3 Circulant Embedding Method

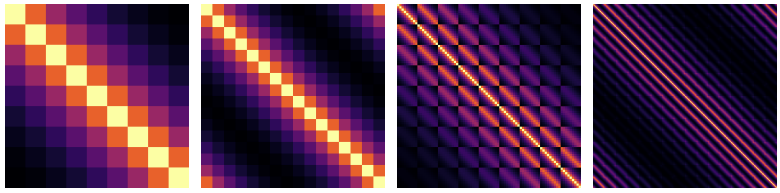

Figure 4: Circulant embedding covariance matrices, see the appendix's main text for the description.

We start by providing a cursory explanation leading to efficient sampling in the 1D case, before giving the intuition behind the extension to the 2D case. We refer to Wood & Chan (1994); Dietrich & Newsam (1997) for a complete explanation and to Rue & Held (2005) for the results underlying

efficient density (likelihood) computation. Let $[0, 1]$ be the spatial domain of interest. Let $\mathcal{S} = \{s_i\}_{i=1}^M$ be an uniform grid (regular lattice) discretizing $[0, 1]$, where the points s_i are assumed to be ordered. The first key observation is that for a stationary covariance function $\rho(\cdot, \cdot)$ the covariance matrix C with entries $C_{i,j} = \rho(s_j, s_j)$ is symmetric and Toeplitz, i.e. with constant-diagonals. See Figure 4 (leftmost) for an example where $M = 9$. A property of symmetric Toeplitz matrices is that they can always be embedded in larger symmetric circulant matrices. A circulant matrix of size $M' \times M'$ is defined by the property that all its rows (and columns) are obtained by cycling through the same M'-dimensional vector. Circulant matrices correspond to covariance matrices of GPs defined on a (here 1D) torus (Rue & Held, 2005, Chapter 2.1). The circulant embedding matrix just introduced corresponds to an artificial enlargement of the spatial domain $[0, 1]$ to a larger interval leading to a torus. See Figure 4 (2nd from left) for a circulant embedding of C. The second key observation is that a circulant matrix is diagonalized by the 1D FFT matrix. Having obtained the eigenvalues of C, efficient sampling on the enlarged domain is achieved by the 1D FFT applied to complex standard random numbers multiplied by the (square root of the) eigenvalues. The real and imaginary part of the generated samples are independent. The main issue wih the CEM is that the circulant matrix embedding might fail to be positive definite. The issue can be avoided by considering progressively larger embeddings, see the theoretical and empirical findings of Dietrich & Newsam (1997). Otherwise a level of approximation can be accepted by modifying the covariance function or by truncating the eigenvalues to be positive.

The development of the 2D CEM follows very similar steps. The domain of interest is now $[0, 1]^2$, the uniform grid is $\mathcal{S} = \{s_{i,j}\}_{i,j=1}^M$ and the points $s_{i,j}$ are assumed to be lexicographically ordered. The stationarity of the covariance functions results in a symmetric block-Toeplitz covariance matrix, as shown in Figure 4 (3rd from left). Again, symmetric block-Toeplitz matrices can be embedded in symmetric block-circulant matrices, as exemplified by Figure 4 (rightmost, zooming might be required to see the block structure). Block-circulant matrices can be shown the be diagonalized by the 2D FFT matrix, and efficient sampling follows from similar steps to the ones seen in the 1D case.

D Additional Figures

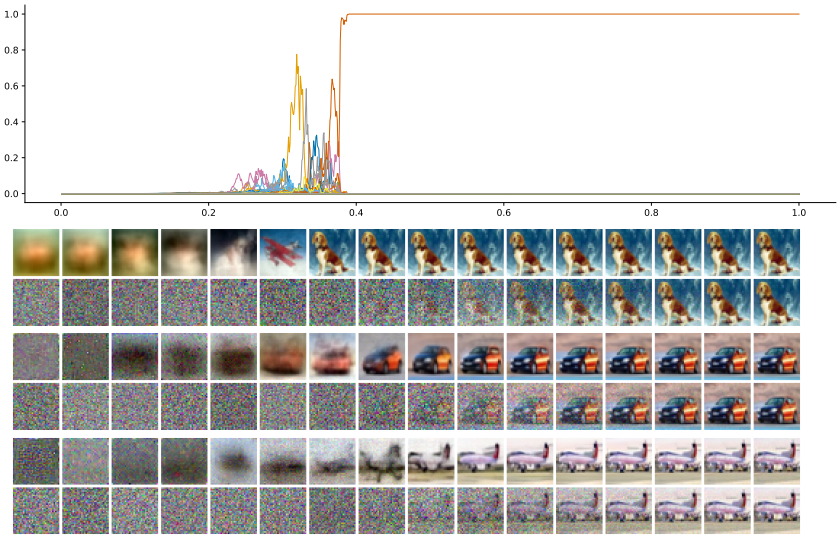

Figure 5: Same as Figure 1 for VESDE model.

Figure 6: Additional samples from the trained VPSDE model of Song et al. (2021), $E[Y|X_t, t]$ and X_t (interleaved rows) over sampling time t, Euler(1000) (top 16 rows) and Euler(100) (bottom 16 rows).

Figure 7: Additional samples from the trained VESDE model of Song et al. (2021), $E[Y|X_t, t]$ and X_t (interleaved rows) over sampling time t, Euler(1000) (top 16 rows) and Euler(100) (bottom 16 rows).