

Diffusion Bridge Mixture Transports, Schrödinger Bridge Problems and Generative Modeling

Stefano Peluchetti

Cogent Labs

106-0032, Tokyo, Japan.

SPELUCHETTI@COGENT.CO.JP

Editor: Aapo Hyvärinen

Abstract

The dynamic Schrödinger bridge problem seeks a stochastic process that defines a transport between two target probability measures, while optimally satisfying the criteria of being closest, in terms of Kullback-Leibler divergence, to a reference process. We propose a novel sampling-based iterative algorithm, the iterated diffusion bridge mixture (IDBM) procedure, aimed at solving the dynamic Schrödinger bridge problem. The IDBM procedure exhibits the attractive property of realizing a valid transport between the target probability measures at each iteration. We perform an initial theoretical investigation of the IDBM procedure, establishing its convergence properties. The theoretical findings are complemented by numerical experiments illustrating the competitive performance of the IDBM procedure. Recent advancements in generative modeling employ the time-reversal of a diffusion process to define a generative process that approximately transports a simple distribution to the data distribution. As an alternative, we propose utilizing the first iteration of the IDBM procedure as an approximation-free method for realizing this transport. This approach offers greater flexibility in selecting the generative process dynamics and exhibits accelerated training and superior sample quality over larger discretization intervals. In terms of implementation, the necessary modifications are minimally intrusive, being limited to the training loss definition.

Keywords: measure transport; coupling; Schrödinger bridge; iterative proportional fitting; diffusion process; stochastic differential equation; score-matching; generative modeling.

1. Introduction

Generating samples from a given distribution is a central topic in computational statistics and machine learning. In recent years, transportation of probability measures has gained considerable attention as a computational tool for sampling. The underlying concept entails generating samples from a complex distribution by first obtaining samples from a simpler distribution and subsequently computing a suitably constructed map of these samples (El Moselhy and Marzouk, 2012; Marzouk et al., 2016). Let us consider the following definition: given two probability measures Γ and Υ , a transport from Γ to Υ is defined as a map $h(X, \varepsilon)$, where ε is an independent auxiliary random element, such that if $X \sim \Gamma$ (i.e., X is distributed according to Γ), then $Y = h(X, \varepsilon)$ satisfies $Y \sim \Upsilon$. This broad definition encompasses both deterministic and random transports and includes a wide range of sampling methods, such as Markov Chain Monte Carlo (Brooks et al., 2011), normalizing flows (Papamakarios et al., 2021), and score-based generative modeling (Ho et al., 2020; Song et al., 2021).

To narrow down the extensive class of possible transports, we can identify the optimal transport according to a suitably chosen criterion. This paper addresses a more generalized version of the optimization problem initially posed by Schrödinger in 1931 (Léonard, 2014b). In this context, both Γ and Υ are defined on the same d -dimensional Euclidean space. The optimization space comprises d -dimensional diffusion processes on the time interval $[0, \tau]$, constrained to have initial distribution Γ and terminal distribution Υ . Optimality is achieved by minimizing the Kullback-Leibler divergence to a reference diffusion process. The optimal map $Y = h(X, \varepsilon)$ is thus obtained by defining ε as the d -dimensional Brownian motion on $[0, \tau]$ which drives the optimal diffusion process started at X , with terminal value Y . This seemingly intricate procedure of defining a transport is of practical relevance. When the reference diffusion is given by a scaled Brownian motion σW_t , for some $\sigma > 0$, $(X, h(X, \varepsilon))$ solves the Euclidean entropy-regularized optimal transport (EOT) problem (Peyré and Cuturi, 2020, Chapter 4) for the regularization level $2\sigma^2$. EOT provides a more tractable alternative to solving the optimal transport (OT) problem. However, in high-dimensional settings, solving the EOT problem remains challenging. The dynamic formulation, achieved by inflating the original space where Γ and Υ are supported so that they correspond to the initial-terminal distributions of a stochastic process, is essential to recent computational advancements in these demanding settings. Specifically, Bortoli et al. (2021); Vargas et al. (2021) rely on a time-reversal result for diffusion processes to implement sampling-based iterative algorithms aimed at solving the dynamic Schrödinger bridge problem.

1.1 Our Contributions

In this work, we begin by reviewing the theory of Schrödinger bridge problems in their various forms, as well as the approaches developed by Bortoli et al. (2021); Vargas et al. (2021). We establish that sampling-based time-reversal approaches suffer from a simulation-inference mismatch. Specifically, at each iteration, samples from the previous iteration are utilized to infer a new diffusion process, but relevant regions of the state space can be left unexplored. This issue is particularly prominent in scenarios characterized by a low level of randomness in the reference process. This is problematic when the goal is to solve the OT problem: for the EOT solution to closely approximate the OT solution, the scale σ of the reference diffusion σW_t must be small.

The primary contribution of this work is the development of a novel sampling-based iterative algorithm, the IDBM procedure, for addressing the dynamic Schrödinger bridge problem. We begin by noting that the problem's solution, i.e. the optimal diffusion process, can be represented as a mixture of diffusion bridges, with the mixing occurring over the bridges' endpoints. In general, processes constructed in this manner are not diffusion processes (Jamison, 1974, 1975). The IDBM procedure involves the following iterative steps: (i) constructing a stochastic process as a mixture of diffusion bridges such that its initial-terminal distribution forms a coupling of the target probability measures Γ and Υ ; (ii) matching the marginal distributions of the stochastic process generated in (i) with a diffusion process; (iii) updating the coupling from step (i) with the initial-terminal distributions of the diffusion obtained in step (ii). The diffusion process of step (ii) is inferred based on samples from the stochastic process of step (i). Crucially, as both processes share the same marginal distributions, no simulation-inference mismatch occurs. In this study, we conduct an initial

theoretical examination of the IDBM procedure, establishing its convergence properties toward the solution to the dynamic Schrödinger bridge problem. Additionally, we carry out empirical evaluations of the IDBM procedure in comparison to the IPF procedure, highlighting its robustness.

The recent advancements in score-based generative approaches (Song et al., 2021; Rombach et al., 2022) have demonstrated remarkable generative quality in visual applications by leveraging the time-reversal of a reference diffusion process to define the sampling process. To elaborate further, the reference diffusion’s dynamics are selected based on a decoupling criterion, which requires the process’s terminal value to be approximately independent of the process’s initial value. The reference process’s initial distribution is assumed to be the empirical distribution of a dataset of interest, or a slightly perturbed version of it. The time reversal of the reference diffusion, complemented by a dataset-independent initial distribution, defines the sampling process. In this work, we propose to utilize the first iteration of the IDBM procedure as an alternative method of defining a sampling process targeting a dataset of interest. Under this proposal, the reference diffusion’s dynamics are no longer constrained by decoupling considerations. We conduct an empirical evaluation of the resulting transport, demonstrating its competitiveness for visual applications in comparison to the approach of Song et al. (2021). Remarkably, the proposed approach exhibits accelerated training and superior sample quality at larger discretization intervals without fine-tuning the involved hyperparameters. From a practitioner standpoint, implementing the proposed method requires minimal changes, which are confined to the training loss definition.

1.2 Structure of the Paper

The paper is structured as follows. In Section 2 we introduce the Schrödinger bridge problem in its various forms, the IPF procedure, its time-reversal sampling-based implementations for the case of a reference diffusion process, and score-based generative modeling. The IDBM procedure, its convergence properties, and suitable training objectives are presented in Section 3. In Section 4 we define the class of reference stochastic differential equations (SDE) considered in the applications, which are carried out in Section 6. Section 5 discusses relevant connections and Section 7 provides concluding remarks for the paper. Proofs, auxiliary formulae and visualizations, code listings and a table summarizing the notation used throughout this work are deferred to the Appendices.

1.3 Notation and Preliminaries

In light of the methodological focus of the present work, we refrain from discussing some of the more technical aspects. The excellent treatises by Léonard (2014b,a) provide a formal account of the preliminaries discussed below and of the developments of Section 2.1. To aid the reader, a summary of the main notation is provided in Table 3 (Appendix E).

All the stochastic processes we are concerned with are defined over the continuous time interval $[0, \tau]$, have state space \mathbb{R}^d , and are continuous, for some constants $\tau > 0$ and $d \geq 1$. A realization of a stochastic process, or path, is an element $x \in \mathcal{C}([0, \tau], \mathbb{R}^d)$, the space of continuous functions from $[0, \tau]$ to \mathbb{R}^d .

We denote each probability measure (PM, or distribution, law) with an uppercase letter. For a PM P , $X \sim P$ denotes that X is a random element with distribution P . Let $n \geq 1$ and

$t_1, \dots, t_n \in [0, \tau]$. We denote the collection of PMs over $\mathcal{C}([0, \tau], \mathbb{R}^d)$ with $\mathcal{P}_{\mathcal{C}}$ and the collection of PMs over $\mathbb{R}^{d \times n}$ with \mathcal{P}_n . When $P \in \mathcal{P}_n$ admits a density with respect to the Lebesgue measure on $\mathbb{R}^{d \times n}$ we denote this density with the same lowercase letter (here p). For $P \in \mathcal{P}_{\mathcal{C}}$, we denote with $P_{t_1, \dots, t_n} \in \mathcal{P}_n$ the marginalization of P at t_1, \dots, t_n and with $P_{\bullet|t_1, \dots, t_n} \in \mathcal{P}_{\mathcal{C}}$ the conditioning of P given the values at times t_1, \dots, t_n . More explicitly, if $X \sim P \in \mathcal{P}_{\mathcal{C}}$, we have $P_{t_1, \dots, t_n}(A_1, \dots, A_n) = \Pr[X_{t_1} \in A_1, \dots, X_{t_n} \in A_n]$ and $P_{\bullet|t_1, \dots, t_n}(A|x_1, \dots, x_n) = \Pr[X \in A | X_{t_1} = x_1, \dots, X_{t_n} = x_n]$. A path PM can be defined via a mixture: if $\Psi \in \mathcal{P}_n$ and $P \in \mathcal{P}_{\mathcal{C}}$, let $Q = \Psi P_{\bullet|t_1, \dots, t_n}$ stand for $Q(A) = \int P_{\bullet|t_1, \dots, t_n}(A|x_1, \dots, x_n) \Psi(dx_1, \dots, dx_n)$. That is, Q is obtained by pinning down $X \sim P$ at times t_1, \dots, t_n , and taking a mixture with respect to Ψ over the pinned down values. With this notation, the marginal-conditional decomposition of $P \in \mathcal{P}_{\mathcal{C}}$ is $P = P_{t_1, \dots, t_n} P_{\bullet|t_1, \dots, t_n}$. If $P, Q \in \mathcal{P}_{\mathcal{C}}$, and P is absolutely continuous with respect to Q ($P \ll Q$) with density $(dP/dQ)(x)$, the marginal-conditional decomposition of $(dP/dQ)(x)$ is $(dP/dQ)(x) = (dP_{t_1, \dots, t_n}/dQ_{t_1, \dots, t_n})(x_{t_1}, \dots, x_{t_n}) \times (dP_{\bullet|t_1, \dots, t_n}/dQ_{\bullet|t_1, \dots, t_n})(x)$.

Given two PMs $\Gamma, \Upsilon \in \mathcal{P}_1$, we define $\mathcal{P}_{\mathcal{C}}(\Gamma, \Upsilon) \subseteq \mathcal{P}_{\mathcal{C}}$ as the collection of PMs having Γ as initial distribution and Υ as terminal distribution. Hence, $P \in \mathcal{P}_{\mathcal{C}}(\Gamma, \Upsilon)$ transports Γ to Υ ($P_0 = \Gamma, P_\tau = \Upsilon$), while its time reversal transports Υ to Γ . We write $\mathcal{P}_{\mathcal{C}}(\Gamma, \cdot)$ when only the initial distribution is fixed, and $\mathcal{P}_{\mathcal{C}}(\cdot, \Upsilon)$ when only the terminal distribution is fixed. We define $\mathcal{P}_2(\Gamma, \Upsilon) \subseteq \mathcal{P}_2$ as the collection of PMs with prescribed marginal distributions Γ and Υ , i.e. the collection of couplings between Γ and Υ . We write $\mathcal{P}_2(\Gamma, \cdot)$ ($\mathcal{P}_2(\cdot, \Upsilon)$) when only the first (second) marginal distribution is fixed. For $P \in \mathcal{P}_2$, we write P_0, P_τ to denote respectively the first and second marginal, and denote the conditional distributions with $P_{\bullet|0}$ and $P_{\bullet|\tau}$. With this notation, for $P \in \mathcal{P}_{\mathcal{C}} \cup \mathcal{P}_2$ we have $P = P_0 P_{\bullet|0} = P_\tau P_{\bullet|\tau}$.

We write $\mathcal{U}(0, \tau)$ for the uniform distribution on $(0, \tau)$, $\mathcal{N}_d(\mu, \Sigma)$ for the d -variate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$. We write $X_i \stackrel{\text{iid}}{\sim} P$ to denote that each random elements X_i is independent and identically distributed according to a PM P . For S, R two PMs on the same space, the Kullback-Leibler (KL) divergence from S to R is defined as $D_{KL}(S \| R) := \mathbb{E}_S[\log(\frac{dS}{dR})]$ whenever $S \ll R$, $+\infty$ otherwise.

We denote identity matrices with I , and transposition of a square matrix A with A^\top . We use the notation $[A]_i$ and $[A]_{i,j}$ to denote indexing respectively of a vector and of a matrix. The Euclidean norm of a vector V is denoted by $\|V\|$.

If $X \sim P \in \mathcal{P}_{\mathcal{C}}$, we denote the time reversal of X with \bar{X} , that is $\bar{X}_t := X_{\tau-t}$ ($t \in [0, \tau]$). The time reversal of P itself is defined as $\bar{P}[A] := P[\bar{A}], \bar{A} := \{x : x_{\tau-t} \in A\}$, so that $\bar{X} \sim \bar{P}$. Note that $E_P[f(X)] = E_{\bar{P}}[f(\bar{X})]$ for any integrable $f : \mathcal{C}([0, \tau], \mathbb{R}^d) \rightarrow \mathbb{R}$. Thorough the paper we set $\tau := \tau - t$.

Without further mention, all SDE drift coefficients are assumed to be \mathbb{R}^d -valued and all SDE diffusion coefficients are assumed to be $\mathbb{R}^{d \times d}$ -valued. Most SDEs, and associated diffusion solutions, are denoted with the same letter X . To avoid ambiguity, we clearly denote the corresponding probability laws. Similarly, we denote most \mathbb{R}^d -valued standard Brownian motions with W . It is understood that all Brownian motions driving different SDEs are independent nonetheless.

2. Schrödinger Bridge Problems

In Section 2.1 we review the theory of Schrödinger bridge problems in both dynamic and static formulations. We initially consider the generic setting of continuous stochastic processes

before specializing to the case of a diffusion reference measures. In Sections 2.2 and 2.3, we discuss the classical iterative algorithm employed in the numerical solution to Schrödinger bridge problems, focusing on recent contributions that rely on a time reversal argument for diffusion processes. Throughout our discussion, we adopt a path space, or continuous-time, perspective. This perspective not only provides clarity but also highlights the connection with a sequential estimation problem for diffusion processes. For the sake of completeness, we review the basics of simulation and inference for SDEs in Section 2.4. Finally, in Section 2.5, review score-based generative modeling, emphasizing its connections to the dynamic Schrödinger bridge problem.

2.1 Dynamic, Static and Half Versions

For $\Gamma, \Upsilon \in \mathcal{P}_1$ and $R \in \mathcal{P}_{\mathcal{C}}$, the solution to the dynamic Schrödinger bridge problem (S_{dyn}) is given by

$$S^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}}) := \arg \min_{S \in \mathcal{P}_{\mathcal{C}}(\Gamma, \Upsilon)} D_{KL}(S \| R).$$

(S_{dyn}) seeks a stochastic process transporting an initial distribution Γ to a terminal distribution Υ , while achieving minimum Kullback-Leibler (KL) divergence to the reference law R . For a given $S \in \mathcal{P}_{\mathcal{C}}$, $S \ll R$, the marginal-conditional decompositions for PMs and densities give

$$\begin{aligned} D_{KL}(S \| R) &= \mathbb{E}_{S_{0,\tau}} \left[\mathbb{E}_{S_{\bullet|0,\tau}} \left[\log \left(\frac{dS_{0,\tau}}{dR_{0,\tau}}(x_0, x_\tau) \frac{dS_{\bullet|0,\tau}}{dR_{\bullet|0,\tau}}(x) \right) \right] \right] \\ &= D_{KL}(S_{0,\tau} \| R_{0,\tau}) + \mathbb{E}_{S_{0,\tau}} [D_{KL}(S_{\bullet|0,\tau} \| R_{\bullet|0,\tau})]. \end{aligned} \quad (1)$$

The second term of (1) is minimized by $S_{\bullet|0,\tau}^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}}) = R_{\bullet|0,\tau}$, independently of Γ, Υ , yielding the representation of $S^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}})$ as a mixture of the reference process pinned down at its initial and terminal values. The mixing distribution $S_{0,\tau}^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}})$ over the initial and terminal values minimizes the first term of (1), i.e. it solves a specific instance of the following problem.

For $\Gamma, \Upsilon \in \mathcal{P}_1$ and $B \in \mathcal{P}_2$, the solution to the static Schrödinger bridge problem (S_{sta}) is given by

$$S^*(\Gamma, \Upsilon, B, \mathcal{P}_2) := \arg \min_{C \in \mathcal{P}_2(\Gamma, \Upsilon)} D_{KL}(C \| B).$$

(S_{sta}) differs from (S_{dyn}) in the optimization space: \mathcal{P}_2 instead of $\mathcal{P}_{\mathcal{C}}$. With this notation, $S_{0,\tau}^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}}) = S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2)$, the solution to (S_{sta}) for $B = R_{0,\tau}$. The solution to (S_{dyn}) is then $S^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}}) = S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2)R_{\bullet|0,\tau}$. Vice versa, if $S^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}})$ solves (S_{dyn}), then $S_{0,\tau}^*(\Gamma, \Upsilon, R, \mathcal{P}_{\mathcal{C}})$ solves (S_{sta}) for $B = R_{0,\tau}$. From this point onward, (S_{sta}) is always understood to be associated to the corresponding (S_{dyn}), i.e. we will only consider (S_{sta}) for $B = R_{0,\tau}$. It established in Rüschedorf and Thomsen (1993) that, if there is a $C \in \mathcal{P}_2(\Gamma, \Upsilon)$ such that $D_{KL}(C \| R_{0,\tau}) < \infty$, then (S_{sta}) admits a solution, which is unique. This hypothesis is assumed to hold thorough the paper.

We introduce an additional axis among which to classify Schrödinger bridge problems, covering both dynamic (H_{dyn}) and static (H_{sta}) variants jointly for conciseness. Thus, let \mathcal{P} be either $\mathcal{P}_{\mathcal{C}}$ (case of (H_{dyn}))) or \mathcal{P}_2 (case of (H_{sta}))). For $\Gamma, \Upsilon \in \mathcal{P}_1$, $Q \in \mathcal{P}$, the solutions to

the forward and to the backward Schrödinger half bridge problems are respectively given by

$$\begin{aligned} S^*(\Gamma, \cdot, Q, \mathcal{P}) &:= \arg \min_{H \in \mathcal{P}(\Gamma, \cdot)} D_{KL}(H \| Q), \\ S^*(\cdot, \Upsilon, Q, \mathcal{P}) &:= \arg \min_{H \in \mathcal{P}(\cdot, \Upsilon)} D_{KL}(H \| Q). \end{aligned}$$

In contrast to (S_{dyn}) and (S_{sta}) , only one of the marginal conditions is enforced. The forward and backward half bridge problems admit simpler solutions which form the basis of the developments of Section 2.2. Proceeding as in (1), it can be established that for $H \in \mathcal{P}$ such that $H \ll Q$

$$\begin{aligned} D_{KL}(H \| Q) &= D_{KL}(H_0 \| Q_0) + \mathbb{E}_{H_0}[D_{KL}(H_{\bullet|0} \| Q_{\bullet|0})] \\ &= D_{KL}(H_\tau \| Q_\tau) + \mathbb{E}_{H_\tau}[D_{KL}(H_{\bullet|\tau} \| Q_{\bullet|\tau})], \end{aligned} \tag{2}$$

from which we obtain the half bridge solutions $S^*(\Gamma, \cdot, Q, \mathcal{P}) = \Gamma Q_{\bullet|0}$ and $S^*(\cdot, \Upsilon, Q, \mathcal{P}) = \Upsilon Q_{\bullet|\tau}$. That is, one of the endpoint (or marginal, for (H_{sta})) distributions of $Q = Q_0 Q_{\bullet|0} = Q_\tau Q_{\bullet|\tau}$ is replaced by one of Γ, Υ while keeping the associated conditional distribution of Q constant. In the case of (H_{dyn}) , it thus suffices to propagate the initial (terminal) distribution Γ (Υ) through the forward (backward) dynamics of Q : $Q_{\bullet|0}$ ($Q_{\bullet|\tau}$).

R_0 plays a minor role in (S_{dyn}) . From (2) applied with $Q = R$, $H \in \mathcal{P}_C$, either R_0 is such that (S_{dyn}) do not admit any solution (when $D_{KL}(S \| R) = +\infty$ for every $S \in \mathcal{P}_C(\Gamma, \Upsilon)$), or the solution to (S_{dyn}) exists, is unique, and is independent of R_0 . This work is concerned with the case where R is given by the solution to a d -dimensional SDE, i.e. a diffusion process (Øksendal, 2003; Friedman, 1975). We consider the following reference SDE, with initial distribution Γ (without loss of generality),

$$\begin{aligned} dX_t &= \mu_R(X_t, t)dt + \sigma_R(X_t, t)dW_t, \quad t \in [0, \tau], \\ X_0 &\sim \Gamma. \end{aligned} \tag{3}$$

Throughout the paper, it is assumed that (3) admits a unique (weak, non-explosive) solution, whose law defines the reference law $R = \Gamma R_{\bullet|0}$. It is also assumed that $R \ll R^\circ$ where R° is the law of the unique solution to (3) with $\mu_R(x, t) = 0$, and that $\Sigma_R(x, t) := \sigma_R(x, t)\sigma_R(x, t)^\top$ is invertible. These are mild assumptions. The law $R_{\bullet|0}$ only depends on the drift coefficient $f(x, t)$ and on the diffusion coefficient $\sigma_R(x, t)$.

In concluding this section, it is noteworthy that solving (S_{sta}) is equivalent to solving a corresponding EOT problem. We seek a solution $S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2)$, and for the sake of simplicity we examine the case where $\Gamma, \Upsilon, R_{0,\tau}$ and $C \in \mathcal{P}_2(\Gamma, \Upsilon)$ all admits densities with respect to Lebesgue measures. Therefore,

$$S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2) = \arg \min_{C \in \mathcal{P}_2(\Gamma, \Upsilon)} \mathbb{E}_C[-\log r_{\tau|0}(X_\tau | X_0)] + H(C).$$

Here $H(C) := \mathbb{E}_C[\log c(X_0, X_\tau)]$ is the entropy of C . Given the same assumption, we consider a cost function $\kappa(x, y) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ and two target marginal distributions Γ, Υ of interest. The solution to the EOT problem with regularization $\varepsilon > 0$ is then defined as

$$EOT^*(\Gamma, \Upsilon, \kappa, \varepsilon) := \arg \min_{C \in \mathcal{P}_2(\Gamma, \Upsilon)} \mathbb{E}_C[\kappa(X_0, X_\tau)] + \varepsilon H(C). \tag{4}$$

By judiciously selecting the reference measure R , it becomes feasible to solve a corresponding EOT problem of interest. For example, if R is associated to $dX_t = \sigma dW_t$ for a scalar $\sigma > 0$, then $S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2) = EOT^*(\Gamma, \Upsilon, \|x - y\|^2, 2\sigma^2)$. Consequently, $S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2)$ is the solution to the Euclidean EOT problem for regularization $\varepsilon = 2\sigma^2$. The arguments presented here carry over to the more generic measure-theoretic setting (Peyré and Cuturi, 2020, Remark 4.2).

2.2 Iterative Proportional Fitting (IPF)

The main tool employed in the numerical solution to (S_{dyn}) and (S_{sta}) is an iterative procedure, known in the literature under a variety of names. In (S_{sta}) , when Γ and Υ admit Lebesgue densities $\gamma(\cdot)$ and $\nu(\cdot)$, the term *Fortet iterations* (Fortet, 1940) is used. Instead, when Γ and Υ concentrate all mass on a finite set of values, the term *iterative proportional fitting procedure* (Deming and Stephan, 1940) is used. In the same setting, due to the equivalence between (S_{sta}) and EOT, the same procedure is also known as *Sinkhorn algorithm*. Finally, the term IPF can be used in the context of the measure-theoretic formulation of Ruschendorf (1995), including the dynamic version (S_{dyn}) (Bernton et al., 2019). In this work, we utilize the IPF term.

Algorithm 1 IPF

Input: $\Gamma, \Upsilon, R_{\bullet|0}, n$
Output: $\{F^{(i)}\}_{i=1}^n$

- 1: $F^{(0)} \leftarrow \Gamma R_{\bullet|0}$
- 2: **for** $i = 1, \dots, n$ **do**
- 3: **if** i is even **then**
- 4: $F^{(i)} \leftarrow \Gamma F_{\bullet|0}^{(i-1)}$ \triangleright forward IPF
- 5: **else**
- 6: $F^{(i)} \leftarrow \Upsilon F_{\bullet|\tau}^{(i-1)}$ \triangleright backward IPF
- 7: **end if**
- 8: **end for**

For (S_{sta}) or (S_{dyn}) , the IPF procedure (Algorithm 1) is obtained by iteratively solving the forward and backward half bridge problems starting from the reference measure R . More precisely, let \mathcal{P} be either \mathcal{P}_C or \mathcal{P}_2 , $R \in \mathcal{P}$, $\Gamma, \Upsilon \in \mathcal{P}_1$, and assume that $R = \Gamma R_{\bullet|0}$. The IPF procedure is defined by the iterates $F^{(0)} = R = S^*(\Gamma, \cdot, R, \mathcal{P})$, $F^{(1)} = S^*(\cdot, \Upsilon, F^{(0)}, \mathcal{P})$, $F^{(2)} = S^*(\Gamma, \cdot, F^{(1)}, \mathcal{P})$, and so on. Under appropriate conditions (Ruschendorf, 1995), the PMs $F^{(i)}$ associated with the IPF iterates of Algorithm 1 converge in total variation (TV) metric and in KL divergence to $S^*(\Gamma, \Upsilon, R, \mathcal{P})$ as $i \rightarrow \infty$. Note that our choice of indexing of the IPF iterations differs from the prior literature, where indexing starts from 1. In this work, indexing reflects the number of learning problems that needs to be solved in order to produce samples from the corresponding IPF iteration. As simulation from (3) is in general feasible (Section 2.4), and can be carried out exactly for the specifications of interest (Section 4), no learning is required to produce samples from $F^{(0)}$.

For (S_{dyn}) , and specifically for the case of a reference diffusion law R , Bortoli et al. (2021); Vargas et al. (2021) introduce implementations of Algorithm 1 that rely on a time reversal result for diffusion processes (Anderson, 1982). Let X be the d -dimensional diffusion process with law $P \in \mathcal{P}_C$ arising as unique solution to the SDE

$$\begin{aligned} dX_t &= \mu(X_t, t)dt + \sigma_R(X_t, t)dW_t, \quad t \in [0, \tau], \\ X_0 &\sim P_0. \end{aligned} \tag{5}$$

Under suitable conditions, the time reversed process $\bar{X}_t := X_{\tau-t}$ is still a diffusion process, associated to the SDE

$$\begin{aligned} d\bar{X}_t &= v(\bar{X}_t, t)dt + \sigma_R(\bar{X}_t, \tau) dW_t, \quad t \in [0, \tau], \\ v(x, t) &:= -\mu(x, \tau) + \nabla \cdot \Sigma_R(x, \tau) + \Sigma_R(x, \tau) \nabla_x \log p_\tau(x), \\ \bar{X}_0 &\sim P_\tau. \end{aligned} \tag{6}$$

In (6) $\tau := \tau - t$, $p_t(x)$ is the Lebesgue density of the marginal distribution P_t , and the d -dimensional vector $\nabla \cdot \Sigma_R(x, t)$ is defined as $[\nabla \cdot \Sigma_R(x, t)]_i := \sum_{j=1}^d \nabla_{x_j} [\Sigma_R(x, t)]_{i,j}$, $1 \leq i \leq d$. We refer to Haussmann and Pardoux (1986); Millet et al. (1989) for the conditions required on (5) for the time reversal (6) to hold.

2.3 Diffusion IPF as Iterative Simulation-Inference

Bortoli et al. (2021); Vargas et al. (2021) propose sampling-based implementations of Algorithm 1. At iteration $i \geq 1$, samples are generated from $F^{(i-1)}$ to construct an approximation to $F^{(i)}$. This approximation is used in turn to generate samples from $F^{(i)}$, which form the input to iteration $i+1$.

Let $F^{(i-1)}$ be a diffusion associated to a SDE (which holds for $i = 1$, as $F^{(0)} = R$). We show that $F^{(i)}$ is also a diffusion associated to a different SDE. First, lines 4 and 6 of Algorithm 1 are modified respectively to $F^{(i)} \leftarrow \Gamma \bar{F}_{\bullet|0}^{(i-1)}$ and $F^{(i)} \leftarrow \Upsilon \bar{F}_{\bullet|0}^{(i-1)}$. With this choice, the path PMs associated to backward IPF iterations are defined over a reverse (relatively to the forward IPF iterations) timescale. Denote the drift and diffusion coefficients associated to $F^{(i-1)}$ respectively with $\mu_F^{(i-1)}(x, t)$ and $\sigma_R(x, t)$, so that $\mu_F^{(0)}(x, t) = \mu_R(x, t)$. It will become clear shortly why the diffusion coefficient is independent of i . If i is even, we have $F_0^{(i)} = \Gamma$, otherwise $F_0^{(i)} = \Upsilon$. Thus, for all $i \geq 1$, $F^{(i)} = F_0^{(i)} \bar{F}_{\bullet|0}^{(i-1)}$. Crucially, from the time reversal result of Anderson (1982), we know that $\bar{F}^{(i-1)}$ is the law of the diffusion associated to

$$\begin{aligned} d\bar{X}_t^{(i-1)} &= \mu_F^{(i)}(\bar{X}_t^{(i-1)}, t)dt + \sigma_R(\bar{X}_t^{(i-1)}, \tau) dW_t, \quad t \in [0, \tau], \\ \mu_F^{(i)}(x, t) &:= -\mu_F^{(i-1)}(x, \tau) + \nabla \cdot \Sigma_R(x, \tau) + \Sigma_R(x, \tau) \nabla_x \log f_\tau^{(i-1)}(x), \\ \bar{X}_0^{(i-1)} &\sim F_\tau^{(i-1)}. \end{aligned} \tag{7}$$

Moreover, $\bar{F}^{(i-1)} = F_\tau^{(i-1)} \bar{F}_{\bullet|0}^{(i-1)}$ and $F^{(i)} = F_0^{(i)} \bar{F}_{\bullet|0}^{(i-1)}$ differs only in the initial distribution: $F^{(i)}$ is associated to (7) with initial distribution $F_0^{(i)}$. An ideal implementation would thus iteratively compute the drift coefficient $v^{(i)}(x, t)$ from $F^{(i-1)}$ for all $i \geq 1$.

It is known that the convergence of the IPF iterations becomes problematic in the small noise regime (Dvurechensky et al., 2018), i.e. for a vanishing diffusion coefficient $\sigma_R(x, t)$. The aforementioned theoretical construction of the IPF iterations provides some insight. Consider $F^{(0)}$, solution to (3). Under suitable conditions (Stroock and Varadhan, 2006, Chapter 11), if $\sigma_R(x, t)$ vanishes, the solution to (3) converges in law to the solution to the random ODE

$$\begin{aligned} dX_t &= \mu_R(X_t, t)dt, \quad t \in [0, \tau], \\ X_0 &\sim \Gamma. \end{aligned} \tag{8}$$

The time reversal of (8) is

$$\begin{aligned} d\bar{X}_t &= -\mu_R(\bar{X}_t, \tau)dt, \quad t \in [0, \tau], \\ \bar{X}_0 &\sim R_\tau, \end{aligned} \tag{9}$$

and $F^{(1)}$ is the solution to (9) with $\bar{X}_0 \sim \Upsilon$. But $F^{(2)}$ is given once again by the solution to (8), and the IPF iterations fail to converge. The key issue is the disappearance of the x -score $\nabla_x \log f_{\mathfrak{r}}^{(i-1)}(x)$ factor from (7), through which the laws of the IPF iterations propagate.

In practice, approximators of $\mu_F^{(i)}(x, t)$, $i \geq 1$, are required. Relying directly on the functional form of the drift coefficients $\mu_F^{(i)}(x, t)$ forms the basis of the approach of Section 2.5, which is limited to $i = 1$. Indeed, for $i > 1$ scalability issues arise, as the nested application of (7) results in i approximators being employed to match $\mu_F^{(i)}(x, t)$ at iteration i . Thus, Bortoli et al. (2021); Vargas et al. (2021) propose to directly infer the drift coefficients $\mu_F^{(i)}(x, t)$, as $F^{(i)}$ is the law of a diffusion process with known diffusion for every $i \geq 0$. The approach relies on two observations. First, generating a sample $\bar{X}^{(i-1)} \sim \bar{F}^{(i-1)}$ is trivial if we have access to a sample $X^{(i-1)} \sim F^{(i-1)}$, as $\bar{X}_t^{(i-1)} = X_{\tau-t}^{(i-1)}$. Second, samples from $\bar{F}^{(i-1)}$ suffice to carry out inference for $F^{(i)}$. Indeed, $\bar{F}^{(i-1)}$ and $F^{(i)}$ differing only in the initial distribution, share the same drift and diffusion coefficients. These considerations suggest the following strategy: (i) use samples from $\bar{F}^{(i-1)}$ to infer an approximator $\alpha_F^{(i)}(x, t) \approx \mu_F^{(i)}(x, t)$; (ii) construct the SDE

$$\begin{aligned} dX_t^{(i)} &= \alpha_F^{(i)}(X_t^{(i)}, t)dt + \sigma_R(X_t^{(i)}, \tau)dW_t, \quad t \in [0, \tau], \\ X_0^{(i)} &\sim F_0^{(i)}, \end{aligned} \tag{10}$$

whose solution is approximately distributed as $F^{(i)}$. Samples from (10) are used in turn as input to iteration $i + 1$. After n iterations a sequence of inferred drift coefficients $\alpha_F^{(1)}(x, t), \dots, \alpha_F^{(n)}(x, t)$ is obtained.

A number of approximations are involved in the aforementioned approach. First, the simulation from each SDE (10) results in a discretization error. Second, the inferred drift coefficients $\alpha_F^{(i)}(x, t)$, differing from their ideal counterparts, give raise to approximation errors. These errors arise from the finite amount of data simulated from $\bar{F}^{(i-1)}$ (Monte Carlo error), from the finite model capacity of an approximator $\alpha_F^{(i)}(x, t) \approx \mu_F^{(i)}(x, t)$, and from local minima in the optimization required to carry out inference. Both discretization and approximation errors can be well controlled by increasing the required computation effort.

However, at a more fundamental level, inference for $\alpha_F^{(i)}(x, t)$ is based on samples from $\bar{F}^{(i-1)}$. As such we can expect a good approximation of $\alpha_F^{(i)}(x, t)$ over the regions the path

space $\mathbb{R}^d \times [0, \tau]$ with non-negligible probability under $\bar{F}^{(i-1)}$. As noted, $\bar{F}^{(i-1)}$ differs from $F^{(i)}$ in its initial distribution, because in general $F_0^{(i)} \neq F_\tau^{(i-1)}$. The equality holds at convergence ($i \rightarrow \infty$) or when $F_\tau^{(0)} = R_\tau = \Upsilon$, in which case the solution to (S_{dyn}) is trivial: $S^*(\Gamma, \Upsilon, R, \mathcal{P}) = R$. Due to the mismatch between $\bar{F}^{(i-1)}$ and $F^{(i)}$, it is possible for the simulated solution to (10) to explore regions of the path space with negligible probability under $\bar{F}^{(i-1)}$, where essentially no information has been available at inference time about the value of $\mu_F^{(i)}(X_t, t)$. In Section 6.2 we show that, far from being just a theoretical concern, the simulation-inference distribution mismatch can have a concrete detrimental effect. Our proposal, introduced in Section 3, does not suffer from the aforementioned issue.

In addition to the discussion in this Section, Vargas et al. (2021) investigates error accumulation within Diffusion IPF (DIPF) approaches, as well as failure instances resulting from insufficient exploration and from the difficulty of bridging distant distributions. These issues are explored quantitatively via empirical experimentation, and initial strides are made towards the establishment of a theoretical framework. Furthermore, Fernandes et al. (2022) explores the challenge of “prior forgetting” encountered in DIPF procedures. In this context, the prior refers to the reference process that only affects the first iteration of DIPF algorithms. In contrast, in our proposed approach, the reference process directly affects every step of the procedure.

2.4 Inference and Simulation for SDEs

For a given drift coefficient $\mu(x, t)$, consider SDE (5) with corresponding diffusion solution P^μ . Sample paths can be generated with arbitrary accuracy using a variety of discretization schemes, the simplest being the Euler scheme. Given a number of time steps $m \geq 1$, corresponding to a time interval $\Delta t = \tau/m$, a discretization Y starting at $Y_0 \sim P_0^\mu$ is generated sequentially on the time grid $\{0, \Delta t, \dots, \tau\}$ as

$$Y_{t+\Delta t} = Y_t + \mu(Y_t, t)\Delta t + \sigma_R(Y_t, t)(W_{t+\Delta t} - W_t), \quad (W_{t+\Delta t} - W_t) \stackrel{\text{iid}}{\sim} \mathcal{N}_d(0, \Delta t I). \quad (11)$$

Convergence of discretization schemes can be assessed according to different metrics. Strong, or path-wise, convergence requires $E[\|X_\tau - Y_\tau\|] \rightarrow 0$ as $\Delta t \rightarrow 0$, where $X \sim P^\mu$ and the same Browning motion W drives both X and its discretization Y . More appropriate to our setting, weak convergence requires $|E[f(X_\tau) - f(Y_\tau)]| \rightarrow 0$ as $\Delta t \rightarrow 0$ for $f(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ belonging to a class of test functions. Kloeden and Platen (1992) contains a thorough coverage of discretization schemes for SDEs and of their convergence properties.

Inference for diffusion processes is a rich research area with long historical developments. We review only two basic approaches and refer to Hurn et al. (2007); Kessler et al. (2012); Fuchs (2013) and references therein for a broader overview. The key difficulty in performing maximum likelihood inference is that transition densities are seldom analytically available outside of restrictive SDEs’ families. Discrete time series data can be observed over arbitrarily long time intervals; however, simple approximations (such as (11)) are accurate only over short time intervals. In this sense, our setting simplifies inference since an arbitrary amount of data can be simulated at arbitrarily high frequencies. Moreover, only the drift coefficient needs to be inferred.

Under suitable conditions, the Cameron-Martin-Girsanov formula provides the density between $P_{\bullet|0}^\mu$ and $P_{\bullet|0}^\gamma$ for another drift coefficient $\gamma(x, t)$ (Liptser and Shiriaev, 1977, Chapter

7),

$$\frac{dP_{\bullet|0}^\mu}{dP_{\bullet|0}^\gamma}(x) = \exp \left\{ \int_0^\tau [(\mu - \gamma)^\top \Sigma_R^{-1}] (x_t, t) dx_t - \frac{1}{2} \int_0^\tau [(\mu - \gamma)^\top \Sigma_R^{-1} (\mu + \gamma)] (x_t, t) dt \right\}. \quad (12)$$

In the following, let $\alpha(x, t)$ be an approximating function for the drift coefficient $\mu(x, t)$. Maximum likelihood inference for μ can be implemented with a driftless SDE acting as dominating measure R° :

$$\mu(x, t) = \arg \max_{\alpha(x, t)} \mathbb{O}_{\text{MLE}}(\alpha, P^\mu, \Sigma_R), \quad \mathbb{O}_{\text{MLE}}(\alpha, P^\mu, \Sigma_R) := \mathbb{E}_{P^\mu} \left[\frac{dP_{\bullet|0}^\alpha}{dR_{\bullet|0}^\circ}(X) \right]. \quad (13)$$

The integrals in (12) need to be discretized, but the discretization errors can be controlled by simulating data at increasing frequencies.

An alternative approach is to consider one of the defining properties of diffusions, i.e. (Friedman, 1975, Chapter 5.4)

$$\mu(x, t) = \lim_{\Delta t \rightarrow 0^+} \frac{\mathbb{E}_{P^\mu}[X_{t+\Delta t} - X_t | X_t = x]}{\Delta t} \approx \mathbb{E}_{P^\mu} \left[\frac{X_{t+\Delta t} - X_t}{\Delta t} \mid X_t = x \right]$$

for suitably small Δt , hence

$$\begin{aligned} \mu(x, t) &\approx \arg \min_{\alpha(x, t)} \mathbb{L}_{\text{DM}}(\alpha, P^\mu, \Delta t), \\ \mathbb{L}_{\text{DM}}(\alpha, P^\mu, \Delta t) &:= \mathbb{E}_{t \sim \mathcal{U}(0, \tau)} \left[\mathbb{E}_{P^\mu} \left[\left\| \alpha(X_t, t) - \frac{X_{t+\Delta t} - X_t}{\Delta t} \right\|^2 \right] \right]. \end{aligned} \quad (14)$$

In the context of the Euler scheme, the estimator that is based on (14) corresponds to the estimator that relies on (13), under the condition that (12) undergoes a piece-wise constant discretization from the left. A difference in their implementation manifests in how (13) incorporates all discretized values for a simulated path into the computation, contrary to (14), which sub-samples the time component. The former approach recovers the drift estimator derived in Vargas et al. (2021), as well as the drift matching estimator of Bortoli et al. (2021) presented in Appendix E, up to a vanishing term as $\Delta t \rightarrow 0$. The remaining estimators presented in Bortoli et al. (2021) target either $\mu(Y_t, t)\Delta t$ or $Y_t + \mu(Y_t, t)\Delta t$ of the Euler discretization (11), again taking into account all discretization steps of a given path.

2.5 Score-based Generative Modeling (SGM)

The approach of Song et al. (2021) is simpler, corresponding to the computation of $F^{(1)}$ in Algorithm 2, i.e. the time reversal of (3). SDE (3) is chosen to ensure approximate conditional independence of X_τ from X_0 , such that $X_\tau \sim R_\tau \approx \Upsilon$ for a simple distribution Υ . For a dataset of interest, the distribution of X_0 is a smoothed version of the training data distribution, i.e. $D_\eta := \frac{1}{n} \sum_{s=1}^n \mathcal{N}_d(x_s; 0, \eta^2 I_d) \in \mathcal{P}_1$ for a small scalar $\eta \geq 0$ where $\{x_s\}_{s=1}^n$ are the n d -dimensional samples. $\eta = 0$ corresponds to the empirical data distribution: $D_0 = \frac{1}{n} \sum_{s=1}^n \delta_{x_s}$. As observed in Section 2.3, due to the (approximate) decoupling of X_τ from X_0 , $F^{(0)}$ approximately solves (S_{dyn}) , $F^{(0)} = R \approx S^*(D_\eta, \Upsilon, R, \mathcal{P}_C)$, and $F^{(1)}$ amounts

to computing its time reversal. This also implies that, for specifications of (3) such that X_τ is almost independent of X_0 , there is no advantage in solving (S_{dyn}) .

In Song et al. (2021), an inferential procedure is developed and carried out to compute the time reversal \bar{X} of (3), and \bar{X}_τ is simulated to produce samples with a distribution close to D_η . A shortcoming of this approach is that achieving the decoupling of X_τ from X_0 requires a large effective integration time (Section 4), which makes the generation process either computationally intensive or inaccurate. More precisely, the specification of (3) introduced in Song et al. (2021) for the CIFAR-10 datasets amounts to simulating a simpler SDE on the time interval $[0, 50^2]$. Consequently, 1,000 time steps, using an ad-hoc predictor-corrector discretization scheme, are employed for the simulation of \bar{X}_τ to ensure the samples' visual quality, as seen in Section 6.3.

Even though the inference techniques of Section 2.3 could also be applied to infer \bar{X} , SGM approaches leverage the specific form of the time reversal drift coefficient of (7), with the aim of learning the x -score $\nabla_x \log r_t(x)$. Inference is based on the minimization of a scalable version of the x -score matching objective (Hyvärinen, 2005), i.e.

$$\mathbb{L}_{\text{SSM}}(\alpha, R, t) := \mathbb{E}_R[\|\alpha(X_t, t) - \nabla_{X_t} \log r_{t|0}(X_t | X_0)\|^2] = \mathbb{E}_R[\|\alpha(X_t, t) - \nabla_{X_t} \log r_t(X_t)\|^2] + c,$$

where $\alpha(x, t) : \mathbb{R}^d \times [0, \tau] \rightarrow \mathbb{R}^d$ is function approximating $\nabla_x \log r_t(x)$ and c is independent of α . The equality, which established in Vincent (2011), holds thanks to the mixture representation of $R_{0,t}$. While computing $\nabla_x \log r_t(x)$ has computational cost $\mathcal{O}(n)$, which is impractical for large datasets, computing $\nabla_x \log r_{t|0}(x|y)$ is $\mathcal{O}(1)$ with respect to n . We provide a simpler derivation:

$$\begin{aligned} \nabla_{x_t} \log r_t(x_t) &= \frac{\nabla_{x_t} \int r_{t|0}(x_t | x_0) R_0(dx_0)}{r_t(x_t)} = \int \nabla_{x_t} \log r_{t|0}(x_t | x_0) \frac{r_{t|0}(x_t | x_0)}{r_t(x_t)} R_0(dx_0) \\ &= \mathbb{E}_R[\nabla_{X_t} \log r_{t|0}(X_t | X_0) | X_t = x_t], \end{aligned} \quad (15)$$

and by the projection property of conditional expectations,

$$\nabla_x \log r_t(x) = \arg \min_{\alpha(x, t)} \mathbb{E}_R[\|\alpha(X_t, t) - \nabla_{X_t} \log r_{t|0}(X_t | X_0)\|^2].$$

Both our derivation and the one of Vincent (2011) rely on an exchange of limits which trivially holds when $R_0 = D_\eta$. In order to infer $\nabla_x \log r_t(x)$ over all $t \in [0, \tau]$, the following objective is considered in Song et al. (2021)

$$\mathbb{L}_{\text{SSM}}(\alpha, R) = \mathbb{E}_{t \sim \mathcal{U}(0, \tau)} [\lambda_t \mathbb{L}_{\text{SSM}}(\alpha, R, t)],$$

where $\lambda_t : (0, \tau) \rightarrow \mathbb{R}_{>0}$ is a time-dependent regularizer. Indeed, the conditional x -scores $\nabla_x \log r_{t|0}(x|y)$ always diverge as $t \rightarrow 0$, contrary to the x -score $\nabla_x \log r_t(x)$, whose behavior for $t \rightarrow 0$ is governed by R_0 . Therefore, it is sensible to normalize the orders of magnitude of $\mathbb{L}_{\text{SSM}}(\alpha, R, t)$ over the range of t .

Due to the identity (15), the generative process can be represented as

$$\begin{aligned} d\bar{X}_t &= \mu_F^{(1)}(\bar{X}_t, \mathbf{r}) dt + \sigma_R(\bar{X}_t, \mathbf{r}) dW_t, \quad t \in [0, \tau], \\ \mu_F^{(1)}(x, t) &= -\mu_R(x, t) + \nabla \cdot \Sigma_R(x, t) + \Sigma_R(x, t) \mathbb{E}_R[\nabla_{X_t} \log r_{t|0}(X_t | X_0) | X_t = x], \\ \bar{X}_0 &\sim \Upsilon. \end{aligned} \quad (16)$$

In Song et al. (2021) a neural network $\alpha_\theta(x, t)$ parametrized by θ is employed as approximating function $\alpha(x, t)$. Having obtained an approximation $\alpha_\theta(x, t) \approx \nabla_x \log r_t(x)$, generation is achieved by numerically integrating (16), with $\alpha_\theta(x, t)$ in place of $\mathbb{E}_R[\nabla_{X_t} \log r_{t|0}(X_t|X_0)|X_t = x]$, to sample \overline{X}_τ .

3. Diffusion Bridge Mixture Transport

In this section, we develop the IDBM procedure. From Section 2.1, we know that the solution to (S_{dyn}) admits the representation $S^*(\Gamma, \Upsilon, R, \mathcal{P}_C) = S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2)R_{\bullet|0,\tau}$. We are concerned with the case where R corresponds to the diffusion process solution to (3). In Section 3.1 we characterize $R_{\bullet|0,\tau}$ as the law of a diffusion bridge. Considering the class of processes $CR_{\bullet|0,\tau}$, indexed by $C \in \mathcal{P}_2(\Gamma, \Upsilon)$, offers a natural means of approximating $S^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$. At the same time, it is advantageous to exploit the dynamic nature of (S_{dyn}) , rather than attempting to directly solve (S_{sta}) , by constructing a sequence of diffusion approximations to $S^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$, as in Section 2.3. Processes formulated as $CR_{\bullet|0,\tau}$ are investigated in Jamison (1974, 1975). In these studies (Section 3.2), it is established that $CR_{\bullet|0,\tau}$ constitutes a diffusion process if and only if $C = S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2)$, which occurs at the optimum of (S_{dyn}) . To address this central issue, we rely on a result that allows the construction of a diffusion process matching the marginal distributions of a mixture of diffusion processes (Section 3.3), with $CR_{\bullet|0,\tau}$ being a special case. The resulting transport is elaborated upon in Section 3.4, where we establish that its iterative application, the IDBM procedure, converges in law to $S^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$. Additionally, we present suitable inference objectives that enable sampling-based implementations of the IDBM procedure.

3.1 Diffusion Bridges

Consider (3) with initial value x_0 , i.e. $R_0 = \delta_{x_0}$. Probabilistically conditioning (3) on hitting a terminal value x_τ at time τ results in the following SDE¹ with initial value x_0 and terminal value x_τ (Särkkä and Solin, 2019, Theorem 7.11)

$$\begin{aligned} dX_t &= b_R(X_t, t)dt + \sigma_R(X_t, t)dW_t, \quad t \in [0, \tau], \\ b_R(x, t) &:= \mu_R(x, t) + \Sigma_R(x, t) \nabla_x \log r_{\tau|t}(x_\tau|x), \\ X_0 &= x_0. \end{aligned} \tag{17}$$

The drift adjustment $\Sigma_R(x, t) \nabla_x \log r_{\tau|t}(x_\tau|x)$ forces the process to hit x_τ at time τ and the diffusion process solving (17) is known as the diffusion bridge from $(x_0, 0)$ to (x_τ, τ) . Consistently with the adopted notation we write $R_{\bullet|0,\tau}$ for its law.

3.2 Reciprocal Processes and Solution to (S_{dyn})

Jamison (1974) studies the properties of reciprocal processes. A d -dimensional process X on $[0, \tau]$ is reciprocal if $\forall s, t : 0 \leq s < t \leq \tau$,

$$\Pr[A_{(s,t)^c} \cap B_{(s,t)}|X_s, X_t] = \Pr[A_{(s,t)^c}|X_s, X_t] \Pr[B_{(s,t)}|X_s, X_t],$$

1. It is a particular case of (19), obtained by Doob h-transform, for $\Gamma = \delta_{x_0}$ and $\Upsilon = \delta_{x_\tau}$.

whenever $A_{(s,t)^c}$ belongs to the σ -algebra generated by $\{X_r : 0 \leq r < s\}$ or $\{X_r : t < r \leq \tau\}$ and $B_{(s,t)}$ to the σ -algebra generated by $\{X_r : s < r < t\}$. A Markov process is reciprocal, but the converse does not hold without further conditions. In Jamison (1974), it is demonstrated that tying down a Markov process at its initial and terminal values and taking a mixture over such values results in a reciprocal process. For a process obtained through this construction, Jamison (1974, Theorem 3.1) characterizes the cases in which the resulting process is not only reciprocal but also Markov. Let $X \sim Q \in \mathcal{P}_C$ be a Markov process, let $q_{t|s}$ denote the associated family of transition densities, here assumed to exist, be strictly positive and continuous, and let $C \in \mathcal{P}_2(\Gamma, \Upsilon)$ for some $\Gamma, \Upsilon \in \mathcal{P}_1$. Then $CQ_{\bullet|0,\tau}$ is Markov if and only if $\exists V_0, V_\tau$ σ -finite positive measures over \mathbb{R}^d such that $C \ll V_0 \otimes V_\tau$, for the product measure $V_0 \otimes V_\tau$, with density

$$\frac{dC}{dV_0 \otimes V_\tau}(x_0, x_\tau) = q_{\tau|0}(x_\tau | x_0). \quad (18)$$

In particular, if at least one of the marginal distributions of C , i.e. Γ, Υ , concentrates all mass to a single point, (18) is satisfied and $CQ_{\bullet|0,\tau}$ is Markov. Moreover, given $\Gamma, \Upsilon \in \mathcal{P}_1$, there are unique $C \in \mathcal{P}_2(\Gamma, \Upsilon)$ and V_0, V_τ σ -finite positive measures such that (18) holds (Jamison, 1974, Theorem 3.2). Finally, C of (18) equivalently solves (S_{sta}) for $B = Q_{0,\tau}$, i.e. $C = S^*(\Gamma, \Upsilon, Q_{0,\tau}, \mathcal{P}_2)$. Within this setting, (18) is often presented in the alternative form $(dC/dQ_{0,\tau})(x_0, x_\tau) = \varphi_0(x_0)\varphi_\tau(x_\tau)$, where $\varphi_0, \varphi_\tau : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$ are the (Schrödinger) potentials (Rüschenhoff and Thomsen, 1993; Pavon et al., 2018; Bernton et al., 2019).

Jamison (1975) specializes these result to the case where Q is the law of a diffusion process. Thus, let $Q = R$, where R is given by the solution to (3). Assuming (18), it is shown that $CR_{\bullet|0,\tau}$, and hence the solution to (S_{dyn}) , is realized by the diffusion

$$\begin{aligned} dX_t &= [\mu_R(X_t, t) + \Sigma_R(X_t, t) \nabla_{X_t} \log h(X_t, t)]dt + \sigma_R(X_t, t)dW_t, \\ h(x, t) &\coloneqq \int r_{\tau|t}(x_\tau | x_t) V_\tau(dx_\tau), \\ X_0 &\sim \Gamma, \end{aligned} \quad (19)$$

i.e. by means of Doob h-transform (Rogers and Williams, 2000, Chapter IV.6.39). Typically, (19) is not directly applicable, V_τ being analytically unavailable.

Finally, Dai Pra (1991) characterizes $h(x, t)$ entering the SDE's drift in (19) as the solution to a stochastic optimal control problem. More precisely, consider R^u associated to

$$\begin{aligned} dX_t &= [\mu_R(X_t, t) + u(X_t, t)]dt + \sigma_R(X_t, t)dW_t, \\ X_0 &\sim \Gamma, \end{aligned}$$

where $u(x, t)$ represents the Markov control. Then, $u(x, t) = \Sigma_R(X_t, t) \nabla_{X_t} \log h(X_t, t)$ minimizes

$$\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R) \coloneqq \mathbb{E}_{R^u} \left[\int_0^\tau \|u(X_t, t)\|_{\Sigma_R^{-1}}^2 dt \right] \quad (20)$$

for the weighted squared norm $\|x\|_A^2 \coloneqq x^\top Ax$, under the constraint that $X_\tau \sim \Upsilon$. We refer to Dai Pra (1991) for the required conditions and precise statement of this result.

3.3 Diffusion Mixture Matching

The following result establishes that a mixture of diffusion processes can be matched in terms of marginal distributions by a single diffusion process. Theorem 1 is established in Brigo (2002, Corollary 1.3) limitedly to finite mixtures and 1-dimensional diffusions. The proof and required assumptions are deferred to Appendix A.

Theorem 1 (Diffusion mixture matching) *Consider the family of d -dimensional SDEs indexed by $\lambda \in \Lambda$*

$$\begin{aligned} dX_t^\lambda &= \mu^\lambda(X_t^\lambda, t)dt + \sigma^\lambda(X_t^\lambda, t)dW_t^\lambda, \quad t \in [0, \tau], \\ X_0^\lambda &\sim P_0^\lambda, \end{aligned} \tag{21}$$

corresponding to the family of path PMs $\{P^\lambda\}_{\lambda \in \Lambda}$. For a mixing PM Ψ on Λ , let $\Pi \in \mathcal{P}_C$ be obtained by taking the Ψ -mixture of (21) over $\lambda \in \Lambda$. In particular, define the mixture marginal densities π_t , $t \in (0, \tau)$, and the mixture initial PM Π_0 by

$$\pi_t(x) := \int_{\Lambda} p_t^\lambda(x)\Psi(d\lambda), \quad \Pi_0(dx) := \int_{\Lambda} P_0^\lambda(dx)\Psi(d\lambda). \tag{22}$$

Consider the d -dimensional SDE

$$\begin{aligned} dX_t &= \mu(X_t, t)dt + \sigma(X_t, t)dW_t, \quad t \in [0, \tau], \\ \mu(x, t) &:= \frac{\int_{\Lambda} \mu^\lambda(x, t)p_t^\lambda(x)\Psi(d\lambda)}{\pi_t(x)}, \\ \sigma(x, t) &:= \frac{\int_{\Lambda} \sigma^\lambda(x, t)p_t^\lambda(x)\Psi(d\lambda)}{\pi_t(x)}, \\ X_0 &\sim \Pi_0, \end{aligned} \tag{23}$$

with law P . Then, under mild conditions, for all $t \in [0, \tau]$ it holds that $P_t = \Pi_t$.

3.4 Diffusion Bridge Mixture Transports

In Section 3.1 we introduced diffusion bridges interpolating initial values x_0 to terminal values x_τ . For $C \in \mathcal{P}_2(\Gamma, \Upsilon)$, consider $\Pi(C, R_{\bullet|0,\tau}) \in \mathcal{P}_C(\Gamma, \Upsilon)$ given by $\Pi(C, R_{\bullet|0,\tau}) := CR_{\bullet|0,\tau}$, i.e. the mixture of diffusion bridges (17) over $(X_0, X_\tau) \sim C$. We apply Theorem 1 to $\Pi(C, R_{\bullet|0,\tau})$ with $\lambda = (x_0, x_\tau)$, $\Lambda = \mathbb{R}^d \times \mathbb{R}^d$, and mixing distribution $\Psi(d\lambda) = C(dx_0, dx_\tau)$. The resulting mixture matching SDE has diffusion coefficient $\sigma_R(x, t)$ and drift coefficient $\mu_M(x, t) := \mu_R(x, t) + \Sigma_R(x, t)e(x, t)$ where

$$\begin{aligned} e(x_t, t) &= \frac{\int \nabla_{x_t} \log r_{\tau|t}(x_\tau|x_t) \pi_{t|0,\tau}(x_t|x_0, x_\tau) \Pi_{0,\tau}(dx_0, dx_\tau)}{\int \pi_{t|0,\tau}(x_t|x_0, x_\tau) \Pi_{0,\tau}(dx_0, dx_\tau)} \\ &= \int \nabla_{x_t} \log r_{\tau|t}(x_\tau|x_t) \frac{\pi_{t|0,\tau}(x_t|x_0, x_\tau)}{\pi_t(x_t)} \Pi_{0,\tau}(dx_0, dx_\tau) \\ &= \int \nabla_{x_t} \log r_{\tau|t}(x_\tau|x_t) \Pi_{\tau|t}(dx_\tau|x_t) \\ &= \mathbb{E}_{\Pi}[\nabla_{X_t} \log r_{\tau|t}(X_\tau|X_t)|X_t = x]. \end{aligned}$$

In conclusion, let $M(\Pi(C, R_{\bullet|0,\tau})) \in \mathcal{P}_C(\Gamma, \Upsilon)$ be associated to

$$\begin{aligned} dX_t &= \mu_M(X_t, t)dt + \sigma_R(X_t, t)dW_t, \quad t \in [0, \tau], \\ \mu_M(x, t) &= \mu_R(x, t) + \Sigma_R(x, t) \mathbb{E}_{\Pi}[\nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t) | X_t = x], \\ X_0 &\sim \Gamma. \end{aligned} \tag{24}$$

Then $M(\Pi(C, R_{\bullet|0,\tau}))$ satisfies $M_t(\Pi(C, R_{\bullet|0,\tau})) = \Pi_t(C, R_{\bullet|0,\tau})$ for all $t \in [0, \tau]$. In particular $M(\Pi(C, R_{\bullet|0,\tau}))$ transports Γ to Υ . We refer to $M(\Pi(C, R_{\bullet|0,\tau}))$ as the diffusion bridge mixture (DBM) transport based on C .

Before turning our attention to the iterated application of the DBM transport, we consider two additional related transports. The first transport is simply a specialization of the DBM transport to the case of a degenerate initial distribution, $\Gamma = \delta_{x_0}$, resulting in

$$\begin{aligned} dX_t &= \mu_{M,x_0}(X_t, t)dt + \sigma_R(X_t, t)dW_t, \quad t \in [0, \tau], \\ \mu_{M,x_0}(x, t) &= \mu_R(x, t) + \Sigma_R(x, t) \mathbb{E}_{\Pi}[\nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t) | X_t = x, X_0 = x_0], \\ X_0 &= x_0. \end{aligned} \tag{25}$$

In view of the results presented in Section 3.2, $\Pi(\delta_{x_0} \otimes \Upsilon, R_{\bullet|0,\tau})$ is the law of a diffusion process (not only a reciprocal process, as in the general case), specifically $S^*(\delta_{x_0}, \Upsilon, R, \mathcal{P}_C)$. Moreover, it is easy to see that Theorem 1 applied to a single diffusion yields that same diffusion, not only a diffusion with matching marginal distributions, i.e. $M(\Pi(\delta_{x_0} \otimes \Upsilon, R_{\bullet|0,\tau})) = \Pi(\delta_{x_0} \otimes \Upsilon, R_{\bullet|0,\tau})$. Indeed, it can be verified by direct calculation that (25) and (19) share the same drift coefficient when $\Gamma = \delta_{x_0}$. In summary, when $\Gamma = \delta_{x_0}$, the DBM transport solves a simple version of (S_{dyn}) . The use of the resulting SDE, sometimes termed Schrödinger-Föllmer Sampler, has been extensively explored in the literature, in generative (Wang et al., 2021; Ye et al., 2022), sampling (Tzen and Raginsky, 2019; Vargas et al., 2022; Huang et al., 2021; Ruzayqat et al., 2023), and optimal control (Dai Pra, 1991; Zhang and Chen, 2022) contexts. These arguments carry over to a more general setting. Whenever the initial coupling $C \in \mathcal{P}_2(\Gamma, \Upsilon)$ is optimal, i.e. solves (S_{sta}) , the resulting mixture process $\Pi(C, R_{\bullet|0,\tau})$ is a diffusion process, and $M(\Pi(C, R_{\bullet|0,\tau})) = \Pi(C, R_{\bullet|0,\tau})$. That is, the DBM transport preserves the optimal coupling.

The second transport involves constructing the DBM in the reversed time direction. More precisely, the following two approaches equivalently yield the same additional transport: (i) considering as reference SDE the time reversal of (3) with initial distribution Υ and constructing the DBM transport based on $\bar{C} \in \mathcal{P}_2(\Upsilon, \Gamma)$ or; (ii) performing a time reversal of (24). For a given $C \in \mathcal{P}_2(\Gamma, \Upsilon)$, we refer to the resulting transport as backward DBM (BDBM) transport based on C . Its law, $\bar{M}(\Pi(C, R_{\bullet|0,\tau})) \in \mathcal{P}_C(\Upsilon, \Gamma)$, is associated to

$$\begin{aligned} d\bar{X}_t &= v_M(\bar{X}_t, \mathfrak{r})dt + \sigma_R(\bar{X}_t, \mathfrak{r})dW_t, \quad t \in [0, \tau], \\ v_M(x, t) &= -\mu_R(x, t) + \nabla \cdot \Sigma_R(x, t) + \Sigma_R(x, t) \mathbb{E}_{\Pi}[\nabla_{X_t} \log r_{t|0}(X_t | X_0) | X_t = x], \\ \bar{X}_0 &\sim \Upsilon. \end{aligned} \tag{26}$$

Then $\bar{M}(\Pi(C, R_{\bullet|0,\tau}))$ satisfies $\bar{M}_t(\Pi(C, R_{\bullet|0,\tau})) = \bar{\Pi}_t(C, R_{\bullet|0,\tau})$ for all $t \in [0, \tau]$. In particular $\bar{M}(\Pi(C, R_{\bullet|0,\tau}))$ transports Υ to Γ . Comparing (16) with (26) reveals that the only difference

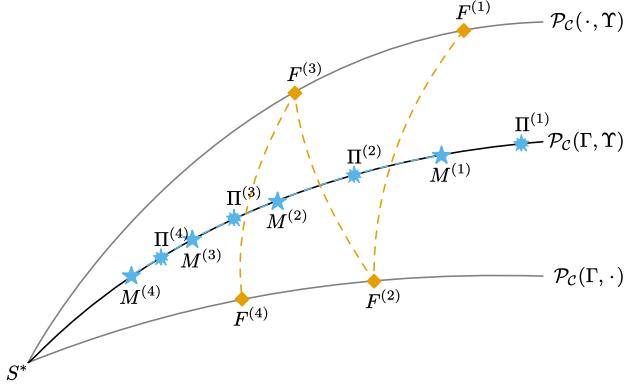


Figure 1: Conceptual representation of IPF and IDBM iterations (inspired by Figure 1 of Bernton et al. (2019)), the ordering of $\Pi^{(i)}$ and $M^{(i)}$ is justified by Theorem 2.

between the SGM model and the corresponding BDBM transport is the measure with respect to which the expectation in the drift adjustment factor is taken: Π instead of R .

Algorithm 2 IDBM

Input: $\Gamma, \Upsilon, R_{\bullet|0,\tau}, C^{(0)}, n$

Output: $\{M^{(i)}\}_{i=1}^n$

- 1: **for** $i = 1, \dots, n$ **do**
 - 2: $\Pi^{(i)} \leftarrow \Pi(C^{(i-1)}, R_{\bullet|0,\tau})$
 - 3: $M^{(i)} \leftarrow M(\Pi^{(i)})$
 - 4: $C^{(i)} \leftarrow M_{0,\tau}^{(i)}$
 - 5: **end for**
-

We consider the iterated application of the DBM transport, specified by Algorithm 2. Starting from an initial coupling $C^{(0)}$, the initial-terminal distribution of the DBM transport of each iteration i is employed to construct the diffusion bridge mixture of iteration $i + 1$. As the BDBM transport is the time reversal of the DBM transport, any iteration i of Algorithm 2 can be equivalently formulated on the reverse timescale, resulting in $\bar{\Pi}^{(i)}$, $\bar{M}^{(i)}$ and $\bar{C}^{(i)}$. In Theorem 2 we establish the convergence properties of Algorithm 2.

Theorem 2 (IDBM convergence) *For $\Gamma, \Upsilon \in \mathcal{P}_1$, $R \in \mathcal{P}_C$ associated to (3) with $\sigma_R(x, t) = I$, consider the iterates $\Pi^{(i)}, M^{(i)} \in \mathcal{P}_C(\Gamma, \Upsilon)$ of Algorithm 2. Assume that²: (i) $D_{KL}(C^{(0)} \| S_{0,\tau}^*) < \infty$; (ii) for each $i \geq 1$ the Cameron-Martin-Girsanov theorem hold for $M^{(i)}$ yielding $dM^{(i)} / dS^*$ (implying that (24) for $\Pi = \Pi^{(i)}$ has a unique diffusion solution with law $M^{(i)}$, and that $M^{(i)} \ll S^*$). Then: (i) $\Pi^{(i)} \xrightarrow{\mathcal{L}} S^*$ and $M^{(i)} \xrightarrow{\mathcal{L}} S^*$ as $i \rightarrow \infty$, where $\xrightarrow{\mathcal{L}}$ denotes converge in law; (ii) $D_{KL}(\Pi^{(i)} \| S^*) \geq D_{KL}(M^{(i)} \| S^*) \geq D_{KL}(\Pi^{(i+1)} \| S^*)$ for $i \geq 1$ (implying that $D_{KL}(M^{(i)} \| S^*)$ is non-increasing in i); (iii) $D_{KL}(\Pi(C) \| S^*) = D_{KL}(M(C) \| S^*)$ if and only if $\Pi(C) = M(C) = S^*$.*

². As in the proof, we denote $\Pi(C) := \Pi(C, R_{\bullet|0,\tau})$, $M(C) := M(\Pi(C, R_{\bullet|0,\tau}))$, $S^* := S^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$.

In generative modeling applications, the simplest choice is to set $C^{(0)}$ to the independent coupling $\Gamma \otimes \Upsilon$, from which samples are trivially obtainable. Dependent couplings are the natural choice in other applications. In inverse problems, $C^{(0)}$ would be the distribution of pairs of clean and corrupted (or latent and partially observed) data. Applications of the DBM transport to inverse problems and to aligned data are reviewed in Section 5. We briefly consider dependent initial couplings in Section 6.1.

We still need to address the problem of inferring the drift coefficients of (24) or (26) for each iteration i of Algorithm 2. The projection property of conditional expectations once again yields suitable objectives for the drift adjustments:

$$\begin{aligned} \mathbb{E}_{\Pi}[\Sigma_R(X_t, t) \nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t) | X_t = x] &= \arg \min_{\alpha(x,t)} \mathbb{L}_{\text{DBM}}(\alpha, \Pi, R, \Sigma_R), \\ \mathbb{L}_{\text{DBM}}(\alpha, \Pi, R, \Sigma_R) &:= \mathbb{E}_{t \sim \mathcal{U}(0, \tau)} [\varrho_t \mathbb{E}_{\Pi}[\|\alpha(X_t, t) - \Sigma_R(X_t, t) \nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t)\|^2]], \\ \mathbb{E}_{\Pi}[\Sigma_R(X_t, t) \nabla_{X_t} \log r_{t|0}(X_t | X_0) | X_t = x] &= \arg \min_{\alpha(x,t)} \mathbb{L}_{\text{BDBM}}(\alpha, \Pi, R, \Sigma_R), \\ \mathbb{L}_{\text{BDBM}}(\alpha, \Pi, R, \Sigma_R) &:= \mathbb{E}_{t \sim \mathcal{U}(0, \tau)} [\lambda_t \mathbb{E}_{\Pi}[\|\alpha(X_t, t) - \Sigma_R(X_t, t) \nabla_{X_t} \log r_{t|0}(X_t | X_0)\|^2]]. \end{aligned} \quad (27)$$

In (27), $\varrho_t : (0, \tau) \rightarrow \mathbb{R}_{>0}$ and $\lambda_t : (0, \tau) \rightarrow \mathbb{R}_{>0}$ are time-dependent regularizers that compensates for the diverging $\nabla_x \log r_{\tau|t}(y|x)$ as $t \rightarrow \tau$ and $\nabla_y \log r_{t|0}(y|x)$ as $t \rightarrow 0$.

Although we rely exclusively on (27) in the numerical experiments of Section 6, we can leverage additional inferential objectives. Indeed, under mild conditions, performing inference following the description in Section 2.4 on data simulated from Π allows the recovery of the drift coefficients $\mu_M(x, t)$ and $v_M(x, t)$. For instance, plugging-in the representation (25) in (12) yields $\mu_M(x, t) = \arg \max_{\alpha(x,t)} \mathbb{O}_{\text{MLE}}(\alpha, \Pi, \Sigma_R)$, while that $\mu_M(x, t) \approx \arg \min_{\alpha(x,t)} \mathbb{L}_{\text{DM}}(\alpha, \Pi, \Delta t)$ follows directly from the definition of Π by interchanging limits. Proofs are omitted here.

We take a moment to provide several comments on the proposed IDBM procedure, contrasting it with the IPF procedure detailed in Section 2.3:

- (i) the IPF iterations generate a sequence of initial-terminal measures $F_{0,\tau}^{(i)} \in \mathcal{P}_2(\Gamma, \cdot) \cup \mathcal{P}_2(\cdot, \Upsilon)$ that matches one of the target marginal distributions Γ, Υ at a time, thus producing a valid coupling between Γ and Υ only in the limit where (S_{sta}) is solved; in contrast, each IDBM iteration produces a valid initial-terminal coupling $C^{(i)} \in \mathcal{P}_2(\Gamma, \Upsilon)$, with the sequence of coupling being progressively optimal toward solving (S_{sta}) ; this crucial difference is depicted in the sketch of Figure 1;
- (ii) IPF iterations necessarily alternates between forward and backward time directions; each IDBM iterations can be solved in either of the time direction (or both);
- (iii) approaches relying on samples from $F^{(i-1)}$ to infer the drift coefficient of $F^{(i)}$ suffer from the simulation-inference distribution mismatch $\bar{F}^{(i-1)} \neq F^{(i)}$ which can negatively impact inferential efficiency; by construction $\Pi^{(i)}$ and $M^{(i)}$ share the same marginal distributions, so samples from $\Pi^{(i)}$ are guaranteed to cover regions of high probability according to $M^{(i)}$, thus making inference of the corresponding drift coefficient reliable;

- (iv) the measures resulting from the IPF iterations do not admit a simple marginal-conditional decomposition; in contrast each $\Pi^{(i)}$ is given by a mixture of diffusion bridges.

The aforementioned points suggest that using the IDBM can be advantageous within the context of generative models. Point (i) establishes that truncating the IDBM iterations at a finite iteration number does not bias the transport. Even the first iteration of the IDBM procedure produces a valid, even if suboptimal, transport. In Section 6.3 we follow this program as an alternative to the SGM approach. In contrast, the IPF procedure requires employing many iterations, which can be computationally expensive. Nevertheless, when multiple IDBM iterations are desirable, it might prove beneficial to solve the iterations alternating between the forward and backward time directions. As the approximation and discretization errors cumulate over the iterates, repeatedly solving the IDBM procedure in the same time direction could lead to a terminal distribution progressively deviating from the corresponding target marginal distribution. Points (iii) and (iv) point to potential efficiency gains of the IDBM procedure. Indeed, at iteration $i > 1$, inferring the drift coefficients of both the IPF and the IDBM requires the numerical discretization and simulation from an SDE associated to iteration $i - 1$. This is computationally expensive: in Bortoli et al. (2021) paths are sampled, cached, and re-used for 100 steps of gradient descent. With this choice, path sampling still amounts to roughly 50% of the total computational time. In the IDBM procedure, simulated paths, which are similarly costly to produce, give raise to samples from $C^{(i-1)}$. However, this time, each sample from $C^{(i-1)}$ can be used to generate multiple samples at arbitrary time points from $\Pi^{(i)} = C^{(i-1)} R_{\bullet|0,\tau}$ which, for SDEs typically used in generative modeling (Section 4), can be done inexpensively and exactly (see (32)). We implement this approach in the empirical comparison of Section 6.4.

4. Reference SDE Class

In this section we introduce a class of reference SDEs. This is achieved in two steps: first we formulate a simple SDE, (28), then we show that SDEs commonly employed in generative models such as Song et al. (2021), i.e. SDE (33), are realized through a time change.

Consider the d -dimensional linear SDE

$$\begin{aligned} dY_t &= -\alpha Y_t dt + \Sigma^{1/2} dW_t, \quad t \geq 0, \\ Y_0 &\sim \Gamma, \end{aligned} \tag{28}$$

with associated path PM P , where $\alpha \geq 0$ is a scalar and Σ is a positive definite covariance matrix. For $\alpha = 0$, (28) yields a correlated and scaled Brownian motion, otherwise (28) corresponds to an Ornstein-Uhlenbeck process. For $\alpha = \frac{1}{2}$, (28) has stationary distribution $\mathcal{N}_d(0, \Sigma)$. The transition densities $p_{t|s}$ of (28) are Gaussian:

$$\begin{aligned} (Y_t | Y_s) &\sim \mathcal{N}_d(Y_s a(s, t), \Sigma v(s, t)), \\ a(s, t) &= e^{-\alpha(t-s)}, \\ v(s, t) &= \begin{cases} t - s, & \text{if } \alpha = 0, \\ \frac{1}{2\alpha}(1 - e^{-2\alpha(t-s)}), & \text{if } \alpha > 0, \end{cases} \end{aligned} \tag{29}$$

and

$$\nabla_{y_s} \log p_{t|s}(y_t|y_s) = \Sigma^{-1} \left(\frac{y_t}{a(s,t)} - y_s \right) \frac{a^2(s,t)}{v(s,t)}, \quad (30)$$

$$\nabla_{y_t} \log p_{t|s}(y_t|y_s) = \Sigma^{-1} \left(y_s a(s,t) - y_t \right) \frac{1}{v(s,t)}, \quad (31)$$

which shows that the diffusion bridge (17) corresponding to (28) remains a linear SDE. From Bayes theorem and the Markov property, for $0 \leq s < t < u$,

$$\begin{aligned} (Y_t|Y_s, Y_u) &\sim \mathcal{N}_d(Y_s \hat{a}(s,t,u) + Y_u \check{a}(s,t,u), \Sigma \tilde{v}(s,t,u)), \\ \hat{a}(s,t,u) &= \frac{v(t,u)a(s,t)}{v(s,t)a^2(t,u) + v(t,u)}, \\ \check{a}(s,t,u) &= \frac{v(s,t)a(t,u)}{v(s,t)a^2(t,u) + v(t,u)}, \\ \tilde{v}(s,t,u) &= \frac{v(s,t)v(t,u)}{v(s,t)a^2(t,u) + v(t,u)}. \end{aligned} \quad (32)$$

We consider the following time change of (28). Let $\beta_t : [0, \tau] \rightarrow \mathbb{R}_{>0}$ be a continuous strictly positive function and $b_t := \int_0^t \beta_u du$. Then $b_t : [0, \tau] \rightarrow [0, b_\tau]$ is differentiable strictly increasing function and $\beta_t = \frac{db_t}{dt}$. An application of (Øksendal, 2003, Theorem 8.5.1) establishes that the time-changed process $X_t := Y_{b_t}$ is equivalent in law to the solution to

$$\begin{aligned} dX_t &= -\alpha \beta_t X_t dt + \sqrt{\beta_t} \Sigma^{1/2} dW_t, \quad t \in [0, \tau], \\ X_0 &\sim \Gamma. \end{aligned} \quad (33)$$

That is, SDE (33) corresponds to the evolution of the simpler SDE (28) under a non-linear time wrapping where time flows with instantaneous intensity β_t .

We conclude this section by specializing (16), (24) and (26) to the case of a reference SDE given by $dX_t = \sigma dW_t$, $\sigma > 0$, referring to Appendix B for the general setting. The generative time reversal process (16) is given by

$$\begin{aligned} d\bar{X}_t &= \frac{\mathbb{E}_R[\bar{X}_\tau | \bar{X}_t] - \bar{X}_t}{\tau - t} dt + \sigma dW_t, \quad t \in [0, \tau], \\ \bar{X}_0 &\sim \Upsilon. \end{aligned}$$

The DBM transport (24) is given by

$$\begin{aligned} dX_t &= \frac{\mathbb{E}_\Pi[X_\tau | X_t] - X_t}{\tau - t} dt + \sigma dW_t, \quad t \in [0, \tau], \\ X_0 &\sim \Gamma, \end{aligned} \quad (34)$$

while the BDBM transport (26) is given by

$$\begin{aligned} d\bar{X}_t &= \frac{\mathbb{E}_\Pi[\bar{X}_\tau | \bar{X}_t] - \bar{X}_t}{\tau - t} dt + \sigma dW_t, \quad t \in [0, \tau], \\ \bar{X}_0 &\sim \Upsilon. \end{aligned} \quad (35)$$

In this case the DBM and BDBM transports are symmetric: the BDM transport based on $C \in \mathcal{P}_2(\Gamma, \Upsilon)$ is equivalent in law to the BDBM transport based on $\bar{C} \in \mathcal{P}_2(\Upsilon, \Gamma)$.

5. Additional Related Literature

The DBM transport is initially introduced in the unpublished manuscript (Peluchetti, 2021), which, however, lacks empirical validations. This shortcoming is subsequently addressed by Wu et al. (2022); Liu et al. (2023b), who successfully implements the DBM transport across a multitude of applications, among other significant contributions. Liu et al. (2023b) demonstrates that the DBM transport corresponds to an optimal Markovianization of the mixture of diffusion bridges it matches. The theoretical developments associated with this finding contribute to our proof of Theorem 2. The proposal put forth by Liu et al. (2023a) is equivalent to the DBM transport for a reference scaled Brownian motion, and its application is shown to yield competitive results in image restoration problems. Similarly, the proposal of Somnath et al. (2023) is equivalent to the DBM for the same reference dynamics. It also assumes that the initial coupling is optimal, and thus preserved. The resulting methodology is successfully applied to the case of aligned data. While both works are equivalent to a specific instance of the DBM transport, they differ in certain aspects of implementation, such as the choice of the discretization scheme for sampling SDE paths and the definition of the loss regularizer. These studies further substantiate the satisfactory performance we observe in the dataset transfer experiment of Section 6.4 for the BDBM transport.

Concurrently and independently of our research, Shi et al. (2023) introduces the *Diffusion Schrödinger Bridge Matching—Iterative Markovian Fitting (DSBM-IMF)* procedure, which is equivalent to the IDBM procedure developed in this work. Shi et al. (2023) conducts a preliminary theoretical assessment of the convergence properties of the *DSBM-IMF* procedure and empirically compares it with the proposal of Bortoli et al. (2021). One of the numerical experiments pertains to the Gaussian setting detailed in Section 6.1, with $\Sigma_0 = \Sigma_1 = I$ and $d = 50$. In contrast to our investigation, neural networks approximators are utilized and trained iteratively. The subsequent results are consistent with our findings of Section 6.1.

The Rectified Flow (RF) method, introduced in Liu et al. (2022) and further explored in Liu (2022), is particularly pertinent to the IDBM procedure discussed in this paper due to their significant similarities. The RF procedure commences with an initial coupling $C \in \mathcal{P}_2(\Gamma, \Upsilon)$, from which it constructs a mixing process, denoted as $\Phi := CL_{\bullet|0,1}$, via the deterministic linear interpolant process $L_{\bullet|0,1}$. The latter is defined by $X_t := (1-t)X_0 + tX_1$, where $t \in [0, 1]$. A rectification of this mixing process is then introduced, given by the solution to the RF ODE

$$\begin{aligned} dV_t &= \nu(V_t, t)dt, \quad t \in [0, 1], \\ V_0 &\sim \Gamma, \end{aligned}$$

where $\nu(x, t) := \mathbb{E}_{\Phi}[X_1 - X_0 | X_t = x]$.

It is shown in Liu et al. (2022) that this rectification procedure results in a valid coupling: $\text{Law}(V_0, V_1) \in \mathcal{P}_2(\Gamma, \Upsilon)$. Moreover, when the rectification process is iterated, it yields a sequence of couplings, $\text{Law}(V_0^{(i)}, V_1^{(i)})$, where $i \geq 1$, such that $\mathbb{E}[\kappa(V_1^{(i)} - V_0^{(i)})]$ is non-increasing across all convex cost functions $\kappa : \mathbb{R}^d \rightarrow \mathbb{R}$. For additional properties of the RF procedure, we direct the reader to Liu et al. (2022). Finally, Liu (2022) establishes that while the RF method successfully solves the one-dimensional Euclidean OT problem, it does not address the multidimensional variant. Consequently, a modification of the RF approach is introduced, which is proven to effectively solve the multidimensional Euclidean OT problem.

The RF method can be understood as the limiting case of the IDBM procedure for a reference scaled Brownian motion as its randomness vanishes. Indeed, consider the reference measure R associated to $dX_t = \sigma dW_t$ over $t \in [0, 1]$. $X_t \sim R_{t|0,1}$ is realized by $X_t = (1-t)X_0 + tX_1 + \sigma\sqrt{t(1-t)}Z_t$, where $Z_t \sim N(0, 1)$. The DBM drift is given by

$$\mu_M(x, t) = \mathbb{E}_{\Pi} \left[\frac{X_1 - X_t}{1-t} \mid X_t = x \right] = \mathbb{E}_{\Pi} \left[X_1 - X_0 - \sigma \sqrt{\frac{t}{1-t}} Z_t \mid X_t = x \right],$$

for $\Pi = CR_{\bullet|0,1}$. As $\sigma \rightarrow 0$ the Brownian bridge's marginal variances converge to zero, while its marginal means are independent of σ . Informally, $\Pi \rightarrow \Phi$ and we obtain the aforementioned limiting equivalence. Section 6.1 expands on this connection within a fully Gaussian setting. Moreover, the conclusions drawn from the numerical experiment of Section 6.4 indicate that the optimal value of σ depends on the specific application. In particular, our attempts to apply the RF procedure to this dataset transfer experiment have not yielded satisfactory results.

As shown in Shi et al. (2023), under certain conditions, both the Flow Matching (FM) proposed by Lipman et al. (2023) and the Conditional Flow Matching (CFM) suggested by Tong et al. (2023) correspond to the first iteration of the RF for generative modeling with a standard Gaussian as the initial distribution. The OT-CFM variant, as introduced by Tong et al. (2023), initially attempts to solve the EOT problem to derive an optimal coupling, followed by fitting a stochastic process to preserve this optimal coupling. This approach is analogous to computing the DBM transport starting from the optimal coupling.

Finally, Chen et al. (2022), while sharing similarities with the work of Bortoli et al. (2021); Vargas et al. (2021), introduces divergence-based objectives, while Thornton et al. (2022) extends the scope of the DIPF method of (Bortoli et al., 2021) by generalizing it to non-Euclidean settings.

6. Applications

In this section we consider four applications of the IDBM procedure.

6.1 Gaussian Transports

We investigate in depth the case where both the initial and terminal distributions are d -dimensional Gaussian distributions, $\Gamma = \mathcal{N}_d(\mu_0, \Sigma_0)$, $\Upsilon = \mathcal{N}_d(\mu_1, \Sigma_1)$, and the reference measure R is associated to $dX_t = \sigma dW_t$ over the time interval $[0, 1]$. The solution to the Euclidean EOT problem, or to (S_{sta}) , defined by (Γ, Υ, R) , is the Gaussian coupling

$$S_{0,1}^*(\Gamma, \Upsilon, R, \mathcal{P}_C) = \mathcal{N}_{2d} \left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_S(\sigma) \\ \Sigma_S(\sigma)^T & \Sigma_1 \end{bmatrix} \right), \quad (36)$$

where $\Sigma_S(\sigma) := (\Sigma_0 \Sigma_1 + \frac{\sigma^4}{4} I)^{1/2} - \frac{\sigma^2}{2} I$. The solution $OT^*(\Gamma, \Upsilon)$ to the OT problem is obtained setting $\sigma = 0$ in (36)³, with corresponding OT plan $\varphi_{OT}(x) := \mu_1 + \Sigma_0^{-1} \Sigma_S(0)(x - \mu_0)$. See for instance Peyré and Cuturi (2020, Section 2.6) and Janati et al. (2020, Section 2).

3. Gaussian distributions with positive semi-definite but not positive definite covariance matrices are well-defined through their characteristic function.

Consider the DBM transport based on the Gaussian coupling $C \in \mathcal{P}_2(\Gamma, \Upsilon)$,

$$C = \mathcal{N}_{2d} \left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_C \\ \Sigma_C^\top & \Sigma_1 \end{bmatrix} \right). \quad (37)$$

The mixture of diffusion bridges with law $\Pi(C, R_{\bullet|0,1})$ has a joint distribution $\Pi_{0,t,1}(C, R_{\bullet|0,1})$ which is again Gaussian,

$$\Pi_{0,t,1}(C, R_{\bullet|0,1}) = \mathcal{N}_{3d} \left(\begin{bmatrix} \mu_0 \\ (1-t)\mu_0 + t\mu_1 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_{\Pi;0,t} & \Sigma_C \\ \Sigma_{\Pi;0,t}^\top & \Sigma_{\Pi;t,t} & \Sigma_{\Pi;t,1} \\ \Sigma_C^\top & \Sigma_{\Pi;t,1}^\top & \Sigma_1 \end{bmatrix} \right),$$

where $\Sigma_{\Pi;0,t} := (1-t)\Sigma_0 + t\Sigma_C$, $\Sigma_{\Pi;t,t} := (1-t)^2\Sigma_0 + t^2\Sigma_1 + t(1-t)(\Sigma_C + \Sigma_C^\top + \sigma^2 I)$ and $\Sigma_{\Pi;t,1} := (1-t)\Sigma_C + t\Sigma_1$. It follows that

$$\mathbb{E}_{\Pi}[X_1|X_t] = \mu_1 + \Sigma_{\Pi;t,1}^\top \Sigma_{\Pi;t,t}^{-1} (X_t - \mu_t)$$

and thus the DBM transport with law $M(\Pi(C, R_{\bullet|0,1}))$ is given by the solution to

$$\begin{aligned} dX_t &= \frac{\mathbb{E}_{\Pi}[X_1|X_t] - X_t}{1-t} dt + \sigma dW_t, \quad t \in [0, 1], \\ X_0 &\sim \Gamma. \end{aligned} \quad (38)$$

SDE (38) is of the form $dX_t = (A_t X_t + b_t)dt + \sigma dW_t$ for $A_t : [0, 1] \rightarrow \mathbb{R}^{d \times d}$, $b_t : [0, 1] \rightarrow \mathbb{R}^d$, i.e. it is linear and time-inhomogenous with Gaussian transition probabilities. The following representation holds: $X_1|X_0$ is distributed as $P_1 X_0 + \varepsilon$, where P_t is given by the solution to the matrix-valued ODE

$$\begin{aligned} dP_t &= A_t P_t, \quad t \in [0, 1], \\ P_0 &= I, \end{aligned} \quad (39)$$

and ε is a d -dimensional Gaussian distribution whose parameters depend on the functions A_t and b_t , but not on X_0 . In conclusion,

$$M_{0,1}(\Pi(C, R_{\bullet|0,1})) = \mathcal{N}_{2d} \left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_{C'} \\ \Sigma_{C'}^\top & \Sigma_1 \end{bmatrix} \right), \quad \Sigma_{C'} := \Sigma_0 P_1^\top. \quad (40)$$

Starting from the Gaussian coupling $C^{(0)} = \Gamma \otimes \Upsilon$, i.e. $\Sigma_{C^{(0)}} = 0I$, the IDBM procedure iteratively computes the updates $\Sigma_C \rightarrow \Sigma_{C'}$ resulting in a sequence of Gaussian couplings $C^{(i)}$.

The IPF updates for (S_{sta}) can be computed analytically thanks to the following property of Gaussian distributions. Let X and Y be d -dimensional random variables with $(X, Y) \sim P_{X,Y}$, where

$$P_{X,Y} := \mathcal{N}_{2d} \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_{yy} \end{bmatrix} \right),$$

and denote with $P_{Y|X}$ the conditional distribution of Y given X under $P_{X,Y}$. For $Q_X := \mathcal{N}_d(\lambda_x, \Gamma_{xx})$, we have

$$\begin{aligned} Q_X P_{Y|X} &= \mathcal{N}_{2d} \left(\begin{bmatrix} \lambda_x \\ \lambda_y \end{bmatrix}, \begin{bmatrix} \Gamma_{xx} & \Gamma_{xy} \\ \Gamma_{xy}^\top & \Gamma_{yy} \end{bmatrix} \right), \\ \lambda_y &:= \mu_y + \Sigma_{xy}^\top \Sigma_{xx}^{-1} (\lambda_x - \mu_x), \\ \Gamma_{xy} &:= \Gamma_{xx} \Sigma_{xx}^{-1} \Sigma_{xy}, \\ \Gamma_{yy} &:= \Sigma_{yy} + \Sigma_{xy}^\top (\Sigma_{xx}^{-1} \Gamma_{xx} \Sigma_{xx}^{-1} - \Sigma_{xx}^{-1}) \Sigma_{xy}. \end{aligned} \tag{41}$$

Therefore, starting from

$$F_{0,1}^{(0)} = \mathcal{N}_{2d} \left(\begin{bmatrix} \mu_0 \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_0 \\ \Sigma_0 & \Sigma_0 + \sigma^2 I \end{bmatrix} \right),$$

the IPF iterations update one of the marginal distributions of $F_{0,1}^{(i)}$ at a time via (41).

When $d = 1$, ODE (39) can be solved analytically. The coupling distribution (37) has a single degree of freedom, its correlation coefficient, which we denote with ρ_C . Then, $M_{0,1}(\Pi(C, R_{\bullet|0,1})) \in \mathcal{P}_2(\Gamma, \Upsilon)$ is Gaussian with correlation coefficient

$$\begin{aligned} \rho_M(\rho_C, \Sigma_0, \Sigma_1, \sigma) &:= \exp \left\{ -\sigma^2 \frac{\tanh^{-1}(\frac{c_1}{c_3}) + \tanh^{-1}(\frac{c_2}{c_3})}{c_3} \right\} \\ c_1 &= \sigma^2 + 2\Sigma_1(\rho_C \Sigma_0 - \Sigma_1), c_3 = \sqrt{(\sigma^2 + 2(\rho_C + 1)\Sigma_0 \Sigma_1)(\sigma^2 + 2(\rho_C - 1)\Sigma_0 \Sigma_1)}, \\ c_2 &= \sigma^2 + 2\Sigma_0(\rho_C \Sigma_1 - \Sigma_0), \end{aligned} \tag{42}$$

which is independent of μ_0, μ_1 . The following results hold: (i) for fixed $\sigma, \Sigma_0, \Sigma_1$, the map $\rho_M(\rho, \Sigma_0, \Sigma_1, \sigma) : \rho \mapsto \rho'$ is a contraction over $\rho \in [-1, 1]$ with limiting value $(\sqrt{\Sigma_0 \Sigma_1 + \frac{\sigma^4}{4}} - \frac{\sigma^2}{2}) / \sqrt{\Sigma_0 \Sigma_1}$, i.e. the correlation coefficient of $S_{0,1}^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$; (ii) for fixed ρ, Σ_0, Σ_1 , $\lim_{\sigma \rightarrow 0} \rho_M(\rho, \Sigma_0, \Sigma_1, \sigma) = \rho_M(\rho, \Sigma_0, \Sigma_1, 0) = 1$, i.e. the correlation coefficient of $OT^*(\Gamma, \Upsilon)$. Result (i) establishes that the couplings produced by the IDBM iterations converge to $S_{0,1}^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$, as expected from Theorem 2. Result (ii) shows that the DBM transport remains well-posed for vanishing regularization σ , and that $M_{0,1}^{(1)} \xrightarrow{\mathcal{L}} OT^*(\Gamma, \Upsilon)$ as $\sigma \rightarrow 0$, i.e. convergence (in law) is attained by the first iteration of the IDBM procedure. For further insight, reconsider the DBM SDE $dX_t = (A_t X_t + b_t) dt + \sigma dW_t$. In the noiseless limit $\sigma \rightarrow 0$, the dynamics are specified by the DBM ODE $dx_t = (A_t^0 x_t + b_t^0) dt$, which corresponds to the RF ODE (Section 5), where $A_t^0 := \lim_{\sigma \rightarrow 0} A_t$ and $b_t^0 := \lim_{\sigma \rightarrow 0} b_t$. Moreover, the solution M to the DBM SDE converges in law to the solution to the DBM ODE (Stroock and Varadhan, 2006, Chapter 11). Solving the DBM ODE and computing the solution at terminal time yields $x_1 = \mu_1 + \sqrt{\Sigma_1 / \Sigma_0}(x_0 - \mu_0) = \varphi_{OT}(x_0)$, the OT plan.

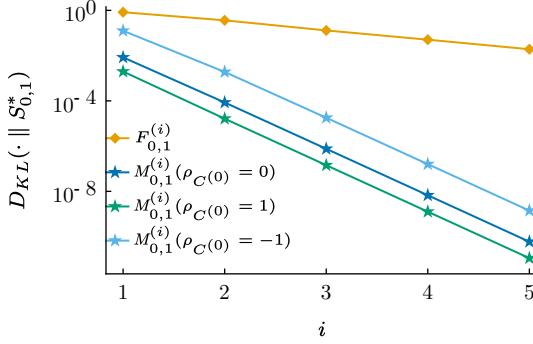


Figure 2: Convergence of IPF and IDBM initial-terminal PMs to $S_{0,1}^*$ in KL divergence; $d = 1$; log-scale.

lim	$D_{KL}(M_{0,1}^{(1)} \parallel S_{0,1}^*)^4$	$D_{KL}(F_{0,1}^{(1)} \parallel S_{0,1}^*)$
$\mu_{\{0,1\}} \rightarrow \pm\infty$	k	∞
$\Sigma_{\{0,1\}} \rightarrow \infty$	K_{Π}	∞
$\Sigma_{\{0,1\}} \rightarrow 0$	0	k
$\sigma \rightarrow \infty$	0	0
$\sigma \rightarrow 0$	k	∞

Table 1: Summary of limiting behavior of the first iteration of the IDBM and IPF procedures; $d = 1$.

We consider the scenario $\mu_0 = -1, \mu_1 = 1, \Sigma_0 = \Sigma_1 = \sigma^2 = 1$. We compute the KL divergence from the IPF PMs $F_{0,1}^{(i)}$ and from the IDBM couplings $M_{0,1}^{(i)}$ to $S_{0,1}^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$ as function of the iterations. The KL divergence between multivariate Gaussian distributions can be calculated analytically, so no approximation is required. The results are shown in Figure 2 where, in addition to the standard initial independent coupling $C^{(0)} = \Gamma \otimes \Upsilon$, we consider two additional initial couplings with correlations $\rho_{C^{(0)}} = \pm 1$. It is observed that the IDBM iterations exhibit a higher rate of convergence with respect to the KL divergence. We note that with our choice of indexing, we slightly favor the IPF iterations, as $F_{0,1}^{(0)}$ already incorporates the reference measure R , while $C^{(0)}$ does not depend on R .

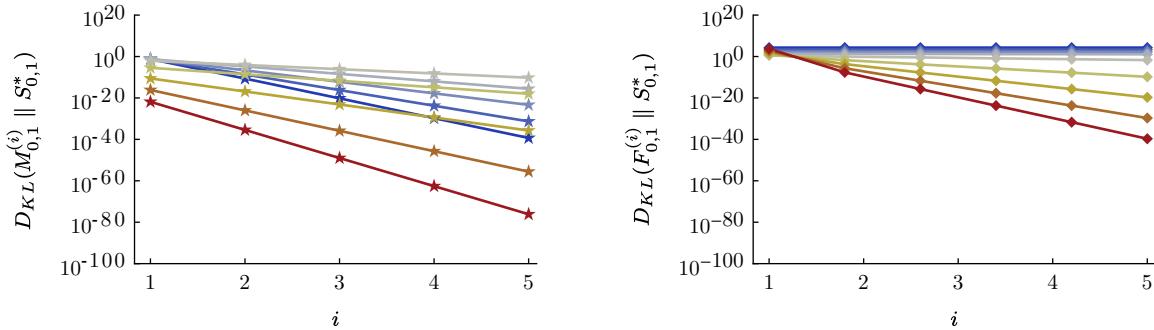


Figure 3: Same as Figure 2 for varying levels of regularization σ ; (left): IDBM; (right): IPF; $d = 1$; shared log-scale.

In Figure 3 we modify the considered scenario by varying the level of regularization: $\sigma = 10^s$ for $s = -2, \dots, 2$, with $\sigma = 10^{-2}$ corresponding to the blue color and $\sigma = 10^2$ to the red color. While the IDBM iterations converge quickly for both low and high levels of regularization (with $\sigma = 1$, as in Figure 2, exhibiting the slowest convergence), the IPF iterations display increasing KL divergence for vanishing σ .

4. $D_{KL}(M_{0,1}^{(1)} \parallel S_{0,1}^*)$ is independent of $\mu_{\{0,1\}}$.

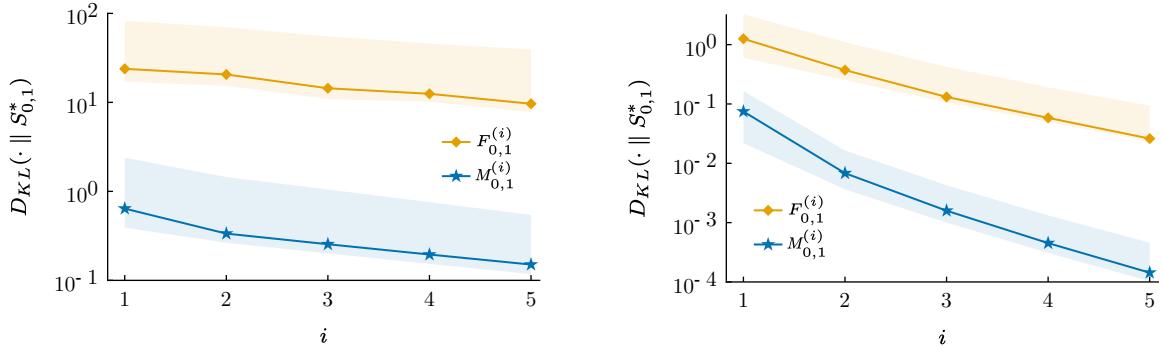


Figure 4: Convergence of IPF and IDBM initial-terminal PMs to $S_{0,1}^*$ in KL divergence; (left): $d = 5$; (right): $d = 10$; log-scale.

We complete the analysis of the one-dimensional setting by studying the behavior of $D_{KL}(M_{0,1}^{(1)} \parallel S_{0,1}^*)$ and $D_{KL}(F_{0,1}^{(1)} \parallel S_{0,1}^*)$ at the boundaries of the parameters' space spanned by $\theta = (\mu_0, \mu_1, \Sigma_0, \Sigma_1, \sigma)$. The results are reported in Table 1, where k denotes positive finite values, $K_{\Pi} := \frac{1}{4}(\pi + \log[4] - 2(1 + \log[\pi]))$ is the supremum of $D_{KL}(M_{0,1}^{(1)} \parallel S_{0,1}^*)$ over θ , $\mu_{\{0,1\}}$ stands for either of μ_0, μ_1 , and $\Sigma_{\{0,1\}}$ stands for either of Σ_0, Σ_1 . When taking each limit all other parameters are kept constant. Note that all of $M_{0,1}^{(1)}, F_{0,1}^{(1)}$ and $S_{0,1}^*$ depend on θ , even though the notation does not make this explicit. The results of Table 1 support the robustness of the IDBM procedure in the one-dimensional setting. In particular, the IDBM procedure does not suffer from the difficulties in bridging distant distributions which are inherent to DIPF approaches (Vargas et al., 2021).

For $d > 1$, we numerically integrate the matrix-valued ODE (39). We generate 20 scenarios for each of $d = 5, 10$. In each scenario, we sample μ_0, μ_1 uniformly on $[-1, 1]^d$ and Σ_0, Σ_1 from a Wishart distribution with d degrees of freedom and covariance matrix $0.2I$ (as in Janati et al. (2020)). We set $\sigma = 0.2$ when $d = 5$ and $\sigma = 1$ when $d = 10$. In the multivariate instance, $M_{0,1}^{(1)}$ does not converge to $OT^*(\Gamma, \Upsilon)$ as $\sigma \rightarrow 0$, and the convergence rate of the IDBM procedure worsens as σ approaches 0. Indeed, in the Gaussian setting defined for this study, the $\sigma \rightarrow 0$ limit of the DBM transport is equivalent to the RF of Liu et al. (2022) (Section 5), and Liu (2022) establishes that the RF solves the OT problem only in the one-dimensional case. Nonetheless, the IDBM procedure consistently outperforms the IPF procedure in all scenarios. In Figure 4 we plot the average KL divergences from $M_{0,1}^{(i)}$ and from $F_{0,1}^{(i)}$ to $S_{0,1}^*$ as function of the iterations with solid lines, where each average is over the 20 sampled scenarios. The banded regions correspond to the ranges between the highest and the lowest KL divergence values across the scenarios.

6.2 Mixture Transports

We consider $\Gamma = \frac{1}{3}(\mathcal{N}_1(-3, 0.2^2) + \mathcal{N}_1(0.5, 0.2^2) + \mathcal{N}_1(3, 0.2^2))$, $\Upsilon = \mathcal{N}_1(0, 2^2)$, and R associated to $dX_t = \sigma dW_t$ over the time interval $[0, 1]$ for $\sigma = 0.2$. This configuration is selected to illustrate the limitation of the sampling-based DIPF procedures discussed in Section 2.3.

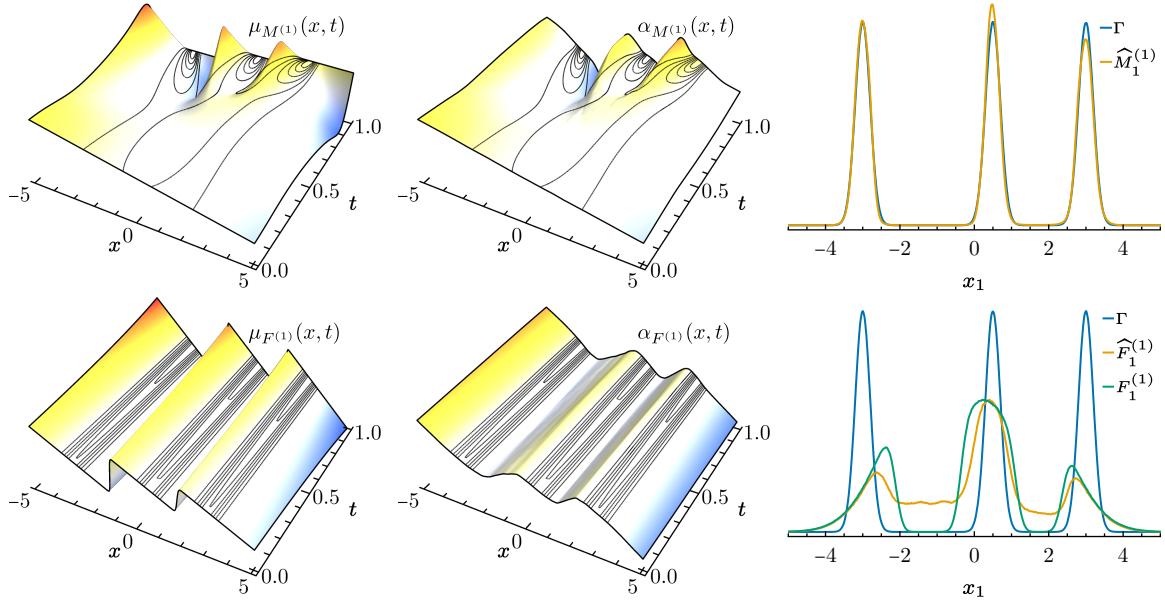


Figure 5: Drift (left), estimated drift (center) and terminal density (right) resulting from the first iteration of the IPF (bottom) and IDBM (top) procedures; generative time.

The goal of this application is to construct a transport from Υ to Γ . This toy experiment can be seen as a simplified instance of the generative setting introduced in Section 2.5, where the training data consist of the values $-3, 0.5, 3$ and $\Gamma = D_\eta$ for $\eta = 0.2$. Compared with typical applications, σ has been set to a small value to introduce a strong dependency between X_0 and X_1 under R , making SGM approaches inapplicable.

In this application, we always assume time t on the generative timescale. The first iteration of the IPF procedure relies on samples from R to infer $\mu_{F^{(1)}}(x, t)$. We display $\mu_{F^{(1)}}(x, t)$ in Figure 5 (bottom-left), superimposed with gray lines representing the level sets of the marginal densities r_t , $t \in [0, 1]$. Samples from R have negligible probability of falling outside narrowly defined regions. We employ as drift approximator a fully-connected neural network $\alpha(\theta, x, t)$ with 3 hidden layers, the ReLU ($x \rightarrow \max(0, x)$) activation function, and a width of 512 units. Utilizing objective (14), $\alpha(\theta, x, t)$ is optimized via stochastic gradient descent (SGD). We display the resulting inferred drift coefficient at convergence, $\alpha_{F^{(1)}}(x, t)$, in Figure 5 (bottom-center). Again we superimpose with the level sets of r_t . The approximation is accurate in regions of high probability under R , but significantly deteriorates outside these regions. We contrast the true density of $F_1^{(1)}$ with a kernel density estimate based on 10 million samples from $\widehat{F}_1^{(1)}$, which are obtained by applying the Euler scheme, $\Delta t = 10^{-4}$, to

$$\begin{aligned} dX_t &= \alpha_{F^{(1)}}(X_t, t)dt + \sigma dW_t, \quad t \in [0, 1], \\ X_0 &\sim \Upsilon, \end{aligned}$$

and collecting the terminal states. As illustrated in Figure 5 (bottom-right), the approximation errors have a significant impact.

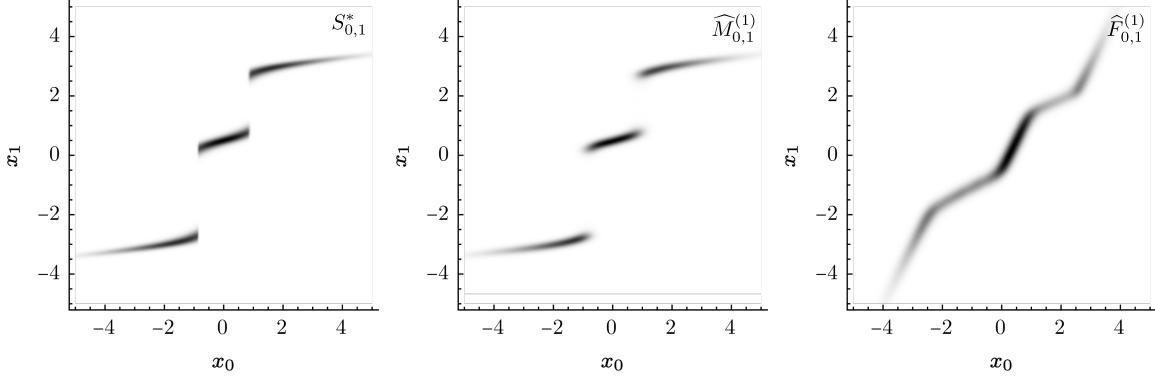


Figure 6: Optimal coupling (left), coupling from the first iteration of the IDBM procedure (center), PM from the first iteration of the IPF procedure (right); generative time.

Implementing sampling-based DIPF procedures requires carefully tuning the level of regularization. On the one hand, an excessively high level of σ renders IPF iterations after the first superfluous, with $F^{(1)}$ already approximately solving (S_{dyn}) . On the other hand, an excessively low level of σ risks incurring the difficulties just exposed. We remark that, in particular instances, the simulation-inference mismatch can become irrelevant. This is the case for the Gaussian transports of Section 6.1, where $F_t^{(i)}$ is a multivariate Gaussian distribution for each i and t (Mallasto et al., 2022). Therefore, $\mu_{F^{(i)}}(x, t)$ is an affine function in x for each i and t , in which case local information about $\mu_{F^{(i)}}(x, t)$ is global as well.

We construct the IDBM procedure started from $C^{(0)} = \Upsilon \otimes \Gamma$. The drift coefficient $\mu_{M^{(1)}}(x, t)$ corresponding to the first IDBM iteration is illustrated in Figure 5 (top-left), superimposed with gray lines representing the level sets of the marginal densities $\pi_t^{(1)}$, $t \in [0, 1]$. Drift inference is based on samples from $\Pi^{(1)}$. The same neural network architecture is employed, and utilizing objective (27) $\alpha(\theta, x, t)$ is optimized via SGD. We display the resulting inferred drift coefficient at convergence $\alpha_{M^{(1)}}(x, t)$ in Figure 5 (top-center), superimposing the level sets of $\pi_t^{(1)} = m_t^{(1)}$. As expected, $\alpha_{M^{(1)}}(x, t)$ closely approximates $\mu_{M^{(1)}}(x, t)$ in regions of high probability under $\Pi^{(1)}$. Crucially, this suffices for learning $\mu_{M^{(1)}}(x, t)$ on the regions of the path space relevant for the simulation from $M^{(1)}$. An interesting observation is that the drift coefficient $\mu_{M^{(1)}}(x, t)$ is smoother⁵ than $\mu_{F^{(1)}}(x, t)$. This aspect makes it easier to both approximate $\mu_{M^{(1)}}(x, t)$ with a neural network and accurately simulate from the resulting SDE. We contrast the true density of Γ with a kernel density estimate based on 10 million samples from $\widehat{M}_1^{(1)}$, which are obtained by applying the Euler scheme, $\Delta t = 10^{-4}$, to

$$\begin{aligned} dX_t &= \alpha_{M^{(1)}}(X_t, t)dt + \sigma dW_t, \quad t \in [0, 1], \\ X_0 &\sim \Upsilon, \end{aligned}$$

and collecting the terminal states. The close agreement between Γ and $\widehat{M}_1^{(1)}$ is illustrated in Figure 5 (top-right).

We discretize Υ and Γ into 5,000 equally spaced bins on $[-5, 5]$ and solve the corresponding EOT problem using the Sinkhorn algorithm from the POT library (Flamary et al., 2021).

5. Both drift coefficients are analytical functions.

For $\sigma = 0.2$, convergence is attained in around 2,000 iterations with default tolerances. We display the resulting 2D histogram in Figure 6 (left). We additionally plot the 2D kernel density estimates of $\widehat{F}_{0,1}^{(1)}$ (center) and $\widehat{M}_{0,1}^{(1)}$ (right). It can be seen that the first iteration of the IDBM procedure suffices to produce a coupling capturing the overall shape of $S_{0,1}^*$.

6.3 Generative Modeling

In this application, we explore the use of the BDBM transport as an alternative to the score-based generative paradigm (Section 2.5) for image generation purposes. The reference SDE is given by (33) with $\alpha = 0$ and $\Sigma = I$, i.e. by a time-scaled d -dimensional Brownian motion. We consider the CIFAR-10 dataset, hence $d = 28 \times 28 \times 3$. As baseline modeling choice we consider the “VE SDE” parametrization from Song et al. (2021) for β_t , and the corresponding smoothed training data distribution D_η ,

$$\begin{aligned} dX_t &= \sqrt{\beta_t} dW_t, \quad t \in [0, 1], \\ X_0 &\sim D_\eta, \\ \beta_t &:= \sigma_{\min}^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^{2t} 2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right), \\ b_t &= \sigma_{\min}^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^{2t} - \sigma_{\min}^2, \end{aligned} \tag{43}$$

where $\eta = \sigma_{\min} = 10^{-2}$ and $\sigma_{\max} = 50$. We recall that $D_\eta := \frac{1}{n} \sum_{s=1}^n \mathcal{N}_d(x_s; 0, \eta^2 I_d)$. Here x_1, \dots, x_n are the 50,000 samples of CIFAR-10’s training dataset $D_{\text{CIFAR-10}}^{\text{train}}$. We select the NCSN++ (smaller) architecture, and training is performed using the official PyTorch implementation⁶ of Song et al. (2021). We review the SGM training procedure in Algorithm 3.

Algorithm 3 SGM training

Input: $\Gamma, R_{t|0}, \nabla_y \log r_{t|0}(y, x), \alpha(\theta, x, t)$

Output: $\alpha_{\text{SGM}}(x, t)$

- 1: **repeat**
 - 2: $t \sim \mathcal{U}(0, \tau)$
 - 3: $X_0 \sim \Gamma$
 - 4:
 - 5: $X_t \sim R_{t|0}(\cdot | X_0)$
 - 6: $Y_t \leftarrow \nabla_{X_t} \log r_{t|0}(X_t | X_0)$
 - 7: $\mathcal{L} \leftarrow \|Y_t - \alpha(\theta, X_t, t)\|^2 \lambda_t$
 - 8: $\theta \leftarrow \text{sgdstep}(\theta, \mathcal{L})$
 - 9: **until** convergence
-

Algorithm 4 BDBM training

Input: $\Gamma, \Upsilon, R_{t|0,\tau}, \nabla_y \log r_{t|0}(y, x), \alpha(\theta, x, t)$

Output: $\alpha_{\text{BDBM}}(x, t)$

- 1: **repeat**
 - 2: $t \sim \mathcal{U}(0, \tau)$
 - 3: $X_0 \sim \Gamma$
 - 4: $X_\tau \sim \Upsilon$
 - 5: $X_t \sim R_{t|0,\tau}(\cdot | X_0, X_\tau)$
 - 6: $Y_t \leftarrow \nabla_{X_t} \log r_{t|0}(X_t | X_0)$
 - 7: $\mathcal{L} \leftarrow \|Y_t - \alpha(\theta, X_t, t)\|^2 \lambda_t$
 - 8: $\theta \leftarrow \text{sgdstep}(\theta, \mathcal{L})$
 - 9: **until** convergence
-

Implementing the BDBM approach requires minimal changes, as delineated in Algorithm 4. PyTorch implementations of SGM and BDBM losses are provided for a reference SDE described by (33) with $\Sigma = I$ and for the regularizer $\lambda_t = v(0, t)$ in Listings 1 and 2 (Appendix C).

6. https://github.com/yang-song/score_sde_pytorch.

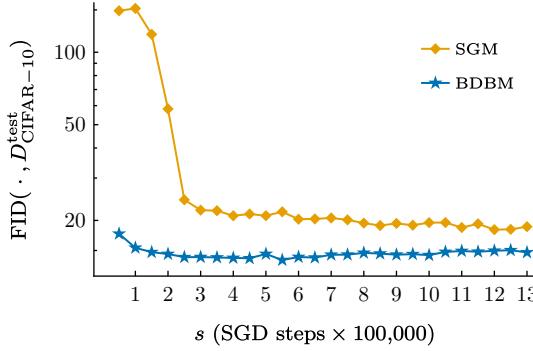


Figure 7: FID between samples generated from the SGM and BDBM models and CIFAR-10 test empirical distribution at $\Delta t = 10^{-3}$; log-scale.

By design, the reference SDE (43) strongly decouples X_1 from X_0 , in which case the BDBM and SGM approaches yield similar generative models. We set $\sigma_{\max} = 1$ and keep $\Upsilon = \mathcal{N}_d(0, \sigma_{\max}^2 I)$ for the generative process’s initial distribution, without making efforts to optimize these hyperparameters. Aside from the modifications to the loss function, no additional changes are introduced to the training code, and training proceeds as for the baseline SGM.

The trained neural network approximators obtained from Algorithm 3 and Algorithm 4 corresponds to the drift adjustment factors appearing in (16) and (26) as conditional expectations. As the remaining terms in (16) and (26) are identical, no code changes are necessary when transitioning from SGM to BDBM at generation time.

To assess the visual quality of generated samples, we employ the Fréchet Inception Distance (FID, Heusel et al. (2017)). FID, which aligns with human-perceived visual quality, is calculated as the 2-Wasserstein Euclidean distance in features space, where image features are derived form the Inception model architecture. CIFAR-10 training occurs over 1,300,000 SGD steps. After every 50,000 SGD steps, we compute the FID between the test portion of the CIFAR-10 dataset, $D_{\text{CIFAR-10}}^{\text{test}}$, and the corresponding 5,000 samples generated from the SGM and BDBM models. Sampling is performed using the Euler scheme (11), $\Delta t = 10^{-3}$. The results are presented in Figure 7, illustrating the competitive performance of the BDBM model for this discretization interval. Furthermore, the BDBM approach exhibits significantly accelerated training.

For each of the considered models and their 26 parameter states (checkpoints, at intervals of 50,000 SGD steps), generating 5,000 samples from either the SGM or the BDBM model using the Euler scheme with a discretization interval $\Delta t = 10^{-4}$ necessitates approximately one day of computation on the NVIDIA RTX 3090 GPU utilized in this experiment. With the same discretization interval, the predictor-corrector scheme of Song et al. (2021) demands twice the computation time. As such, we have chosen to circumvent the intensive computations required to identify an “optimal” checkpoint through grid-search across parameter states using the predictor-corrector scheme and $\Delta t = 10^{-4}$, as is done in Song et al. (2021). Our analysis is instead confined to the Euler scheme and a single “optimal” checkpoint is selected for both the SGM and BDBM models relying on the findings of Figure 7. For the BDBM model, we opt for the model trained for 250,000 steps, while for the SGM model we select



Figure 8: Samples from the SGM and BDBM models with different discretization intervals, compared to samples from the CIFAR-10 training dataset.

the model trained for 1,200,000 steps, which aligns with the checkpoint chosen in Song et al. (2021). Subsequently, we calculate the FID values corresponding to $\Delta t = 1/25, 1/100, 1/1000$, obtaining 93.7, 14.0, 12.1 for the BDBM model and 420.3, 18.3, 12.0 for the SGM model, respectively. Corresponding randomly selected samples are shown in Figure 8 and contrasted with the first 16 CIFAR-10 training samples. Although the visual quality of the SGM and BDBM models is comparable at finer discretization steps, the BDBM model retains superior visual quality as the discretization interval increases.

6.4 Dataset Transfer

In this section, we investigate an application where both the initial and the terminal distributions are complex, rendering SGM approaches inapplicable. Following Bortoli et al. (2021), we explore the scenario where the initial distribution Γ is represented by the MNIST dataset (D_{MNIST}), whereas the terminal distribution Υ is derived from the first five lowercase and the first five uppercase characters, specifically, a, \dots, e, A, \dots, E , of the EMNIST dataset (D_{EMNIST}). Consequently, $d = 28 \times 28$. The reference process law R is associated to $dX_t = \sigma dW_t$ over the time interval $[0, 1]$. Our objective is to approximately solve (S_{dyn}) .

We evaluate the performance of the DIPF and IDBM procedures, alternating between solving the IDBM iterations in backward and forward time directions—a requirement for the DIPF procedure. As in Bortoli et al. (2021), we utilize a lightweight configuration of the U-Net neural network architecture as proposed by Dhariwal and Nichol (2021). We employ two neural network models $\alpha(\theta, x, t)$, $\alpha(\phi, x, t)$, which act as approximators to the drifts corresponding to the forward and backward iterations respectively.

Each of these models, $\alpha(\theta, x, t)$ and $\alpha(\phi, x, t)$, is associated with an independent instance of the Adam optimizer due to its adaptive nature, and a model copy $(\hat{\alpha}(\theta, x, t))$ to $\alpha(\theta, x, t)$,

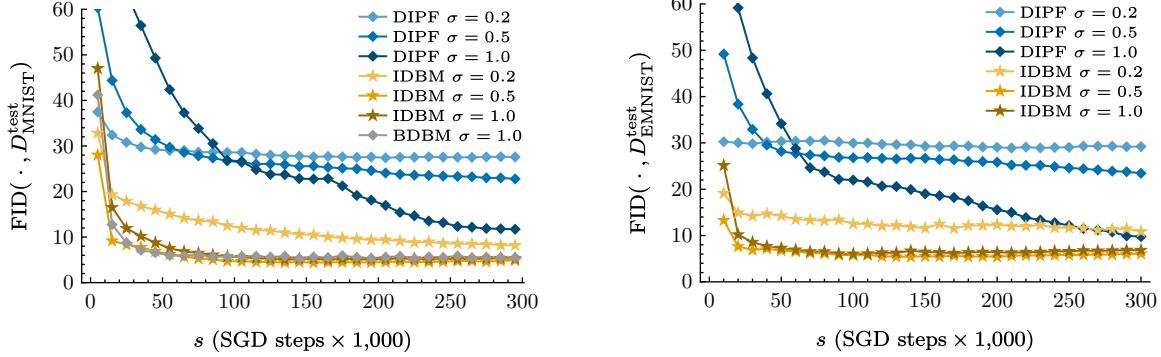


Figure 9: IDBM and DIPF test FID values as function of SGD steps for (left): $D_{\text{EMNIST}}^{\text{test}} \rightarrow D_{\text{MNIST}}^{\text{test}}$, i.e. backward time; (right): $D_{\text{MNIST}}^{\text{test}} \rightarrow D_{\text{EMNIST}}^{\text{test}}$, i.e. forward time; linear-scale, truncated at 60.

$\alpha(\hat{\phi}, x, t)$ to $\alpha(\phi, x, t)$). The parameters of these copies, $\hat{\theta}$ and $\hat{\phi}$, are updated according to the Exponential Moving Averaging (EMA) scheme. Models $\alpha(\hat{\theta}, x, t)$ and $\alpha(\hat{\phi}, x, t)$, whose parameters evolve more stably, are used to simulate the required SDE paths using the Euler scheme and a discretization interval of $\Delta t = 1/30$. To increase efficiency, we implement caching of sampled paths. For the DIPF algorithm, entire discretized paths are cached, while for the IDBM algorithm, only the initial and terminal values are cached. Sampling from the reference diffusion bridge corresponding to R at arbitrary time points can be performed quickly and exactly (Section 4).

We utilize the following training methods. For the DIPF procedure, we rely on the drift matching estimator (14). For the IDBM procedure, we employ the regression estimators (27). For the reference process $dX_t = \sigma dW_t$, both the BDM (34) and BDBM (35) have the same target $(X_1 - X_t)/(1 - t)$ entering the forward and backward losses in (27). Instead of using time-dependent regularizers as in Section 6.3, we limit the simulation of t to $t \sim \mathcal{U}(0, 1 - \Delta t/2)$, which allows us to recover the Rectified Flow loss with $\sigma = 0$.

Test FID values are calculated by initializing X_0 from $D_{\text{EMNIST}}^{\text{test}}$ ($D_{\text{MNIST}}^{\text{test}}$) and subsequently sampling the discretized path X_t to obtain model samples that are approximately distributed as $D_{\text{MNIST}}^{\text{test}}$ ($D_{\text{EMNIST}}^{\text{test}}$). The specific procedures corresponding to the IDBM and DIPF methods are as follows. For the DIPF procedure, we remove the stochastic component from the terminal Euler discretization step as customary. For the IDBM procedure, we employ the estimator $E_t := \mathbb{E}[X_1 | X_t]$, which is directly obtained from $\alpha(\hat{\theta}, x, t)$ and $\alpha(\hat{\phi}, x, t)$. Evaluating $E_{1-\Delta t}$ is equivalent to removing the stochastic component from the terminal Euler discretization step, but we find that $E_{1-2\Delta t}$ empirically performs better.

We consider three different levels of regularization $\sigma = 1.0, 0.5, 0.2$, and compare the IDBM and DIPF procedures over 60 iterations, with each iteration comprising 5,000 SGD steps. Additionally, we assess the simpler application of the BDBM procedure, which is trained for an equivalent amount of 300,000 SGD steps. Detailed settings of all hyperparameters defining this experiment, along with implementation details, are provided in the accompanying code implementation⁷.

7. <https://github.com/stepelu/idbm-pytorch>

Procedure	IDBM			DIPF			BDBM
	1.0	0.5	0.2	1.0	0.5	0.2	
Regularization σ	1.0	0.5	0.2	1.0	0.5	0.2	1.0
FID forward	10.9	6.1	6.8	29.2	23.4	9.8	
FID backward	8.2	4.9	5.2	27.6	22.8	11.7	5.3
$\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R)/d$ forward	4.0	5.2	12.5	65.9	31.2	22.5	
$\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R)/d$ backward	3.9	4.8	12.5	62.1	33.6	22.5	4.2

Table 2: Test FID and $\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R)$ values for fully trained IDBM, DIPF and BDBM procedures.

The findings from this experiment are presented in Table 2 and Figure 9. The IDBM procedure consistently displays superior convergence properties and lower test FID values compared to the DIPF procedure. For both $\sigma = 0.5$ and $\sigma = 0.2$, the DIPF procedure fails to make significant progress, with model samples almost identical to the input samples. Indeed, a FID value around 30 corresponds to $\text{FID}(D_{\text{MNIST}}^{\text{test}}, D_{\text{EMNIST}}^{\text{test}})$. Visualizations of trajectories of sampled paths X_t , and of estimators E_t for the IDBM and BDBM approaches, are included in Appendix D.

The results from the IDBM procedure also illustrates a dependency on the regularization level σ , with performance deteriorating for the lowest regularization level $\sigma = 0.2$. We were unable to achieve satisfactory results for the case $\sigma = 0$, which corresponds to the Rectified Flow method. In this instance a valid transport is not achieved as well, with model samples strongly resembling the input samples. We hypothesize that while the RF method offers the advantageous feature of progressively straightening the inferred paths, the ensuing inference problem may present considerable difficulty. Indeed, the neural network model has to predict the terminal sample exactly at $t = 0$ based on the initial sample, which commences to be the case at low levels of σ (see E_t in Appendix D).

The optimal SDE solving (S_{dyn}) accomplishes a valid coupling while minimizing the drift norm functional $\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R)$, as outlined in Section 3.2. Lower FID values serve as indicators of the accuracy of the inferred coupling. As demonstrated by the results presented in Table 2, the IDBM procedure not only infers an accurate coupling but also effectively minimizes $\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R)$. Estimations of the test values for $\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R)$ are computed through Monte Carlo sampling of the drift norm functional (20). In this process, the relevant neural network approximator is substituted for $\mu(x, t)$. Finally, the simpler BDBM procedure yields FID and $\mathbb{L}_{\text{OC}}(u, R^u, \Sigma_R)$ values that are comparable to those produced by the more complex IDBM procedure. This suggests that the BDBM crude approximation to (S_{dyn}) might be acceptable within this context.

7. Discussion

This paper introduces a novel iterative algorithm, the IDBM procedure, aimed at solving the dynamic Schrödinger bridge problem (S_{dyn}), and performs a preliminary theoretical analysis of its convergence properties. The theoretical findings are complemented by various numerical simulations and analytical results, demonstrating the competitive performance of the IDBM procedure in comparison to the IPF procedure.

As in Pavon et al. (2018); Bortoli et al. (2021); Vargas et al. (2021), we assume that samples from the target initial and terminal distributions are either readily available, belonging to some dataset of interest, or can be generated without further approximations. The IDBM procedure is particularly suitable in scenarios where the reference diffusion admits simple analytical transition densities. These considerations suggest that the IDBM procedure is ideally suited for current generative applications. Since the IDBM procedure produces a valid coupling at each iteration, we utilize the first IDBM iteration, i.e. the (backward) DBM transport, as an alternative to the approach of Song et al. (2021). This alternative achieves accelerated training and superior sample quality at larger discretization intervals. An additional advantage of this proposal is the simplicity of its implementation, which differs from the approach of Song et al. (2021) only in the training loss definition. The time-reversal sampling-based implementations of Bortoli et al. (2021); Vargas et al. (2021) suffer from the simulation-inference mismatch demonstrated in the present work, and might be difficult to scale to demanding generative applications due to the ensuing lower efficiency and higher computational cost.

The outcomes of our study give rise to additional avenues for future research. From a theoretical perspective, our initial analysis falls short of establishing the convergence in KL divergence of the IDBM procedure’s iterates to the solution to (S_{dyn}) . The numerical and analytical results of Section 6.1 support that, under suitable conditions, this stronger form of convergence is achieved. Non-asymptotic results would be particularly valuable, as they could elucidate the conditions under which either the IPF or IDBM methods are more favorable. Moreover, it would be desirable to introduce more easily verifiable conditions guaranteeing convergence of the IDBM procedure. Finally, when $d > 1$, neither the IPF nor the IDBM procedures are robust to vanishing randomness in the reference diffusion R . Consequently, the design of an iterative procedure solving (S_{dyn}) , while being robust to a vanishing regularization level, remains an open problem.

On the empirical side, computational constraints limited the scope of our simulation study comparing the DBM transport to the time-reversal approach of Song et al. (2021). Specifically, no hyperparameter optimization was performed, neither for DBM-specific parameters, nor for the employed neural network architecture, while Figure 7 suggests that the tested configuration might suffer from overfitting later in training. This is conceivable, as the DBM transport approximates the solution to (S_{dyn}) , which is associated to an SDE that has a “simpler” optimal drift (Dai Pra, 1991). Since the exact solution to the DBM transport, and of the competing approaches, perfectly recovers the training data distribution, a less powerful architecture might be required for superior generalization properties. Overfitting to the CIFAR-10 dataset is not a novel concern (Nichol and Dhariwal, 2021), and it is typically addressed through hyperparameter search.

At one extreme, when the reference process’s dynamics imply a terminal value almost independent of the initial value, the DBM results in a generative model akin to that of Song et al. (2021), while rectifying the terminal distribution mismatch inherent in all time-reversal approaches (Bortoli et al., 2021). More generally, the DBM transport enlarges the space of feasible (exact) transports. Consequently, it is plausible that improvements in current state-of-the-art solutions can be attained by empirically exploring this enlarged space, optimizing for a given metric of interest, such as FID. In particular, high-resolution images pose challenges for time-reversal approaches and necessitate an increase in the reference

process randomness, thus increasing the implicit integration time (Hoogeboom et al., 2023). The DBM transport, which precisely matches the initial and terminal marginal distributions, does not face this issue. Lastly, computational constraints precluded an evaluation of the application of further iterations of the IDBM procedure in generative modeling applications, which would trade off increased training time with a more efficient generation process.

The approaches of Bortoli et al. (2021); Vargas et al. (2021) are directly applicable to a broad spectrum of reference SDEs. Conversely, the IDBM procedure detailed in this work necessitates an analytically solvable reference SDE. However, we remark that the IDBM methodology can potentially be extended to accommodate complex reference dynamics. A promising approach involves harnessing the findings of De Bortoli et al. (2021) to construct neural network approximators for $\nabla_x \log(r_{\tau|t}(y|x))$ and $\nabla_y \log(r_{t|0}(y|x))$. These approximators allow both the sampling of the reference diffusion bridge via a discretization of (17), and the calculation of the targets in the estimators (27). An alternative strategy employs the Exact Algorithm (EA) proposed by Beskos and Roberts (2005); Beskos et al. (2008) to sample exactly from the reference diffusion bridge (17). The unavailability of $\nabla_x \log(r_{\tau|t}(y|x))$ and $\nabla_y \log(r_{t|0}(y|x))$ can be circumvented by performing drift inference as outlined in Section 2.4, rather than relying on (27). We refer to the discussion following (27) for more details. However, we must underscore that the application of the EA is restricted to reference SDEs with a “reducible” diffusion coefficient (Ait-Sahalia, 2008), and to low-dimensional spaces due to scalability concerns (Peluchetti and Roberts, 2012). We refer to Beskos et al. (2008) for the additional requirements on the drift coefficient of the reference SDE. The empirical evaluation of these extensions is left to future research.

Appendices

A. Proofs

Proof of Theorem 1

The following assumptions are sufficient for the application of Theorem 1. We assume that (23) and each member of (21) admit a unique solution, with strictly positive marginal densities over \mathbb{R}^d . We assume that each marginal density is the unique solution to the corresponding Fokker-Plank PDE, and that the exchanges of limits denoted with $(\star\star)$ hold. When Theorem 1 is applied to match $\Pi = CR_{\bullet|0,\tau}$ in Section 3.4, these conditions are satisfied if R is given by (33) and the marginal distributions of C are given by mixtures of Gaussian distributions, which covers the setting of generative modelling applications.

Theorem 1 (Diffusion mixture matching) *Consider the family of d -dimensional SDEs indexed by $\lambda \in \Lambda$*

$$\begin{aligned} dX_t^\lambda &= \mu^\lambda(X_t^\lambda, t)dt + \sigma^\lambda(X_t^\lambda, t)dW_t^\lambda, \quad t \in [0, \tau], \\ X_0^\lambda &\sim P_0^\lambda, \end{aligned} \tag{21}$$

corresponding to the family of path PMs $\{P^\lambda\}_{\lambda \in \Lambda}$. For a mixing PM Ψ on Λ , let $\Pi \in \mathcal{P}_C$ be obtained by taking the Ψ -mixture of (21) over $\lambda \in \Lambda$. In particular, define the mixture marginal densities π_t , $t \in (0, \tau)$, and the mixture initial PM Π_0 by

$$\pi_t(x) := \int_{\Lambda} p_t^\lambda(x)\Psi(d\lambda), \quad \Pi_0(dx) := \int_{\Lambda} P_0^\lambda(dx)\Psi(d\lambda). \tag{22}$$

Consider the d -dimensional SDE

$$\begin{aligned} dX_t &= \mu(X_t, t)dt + \sigma(X_t, t)dW_t, \quad t \in [0, \tau], \\ \mu(x, t) &:= \frac{\int_{\Lambda} \mu^\lambda(x, t)p_t^\lambda(x)\Psi(d\lambda)}{\pi_t(x)}, \\ \sigma(x, t) &:= \frac{\int_{\Lambda} \sigma^\lambda(x, t)p_t^\lambda(x)\Psi(d\lambda)}{\pi_t(x)}, \\ X_0 &\sim \Pi_0, \end{aligned} \tag{23}$$

with law P . Then, under mild conditions, for all $t \in [0, \tau]$ it holds that $P_t = \Pi_t$.

Proof In this proof we make use of the following notation: for $f(x, t) : \mathbb{R}^d \times [0, \tau] \rightarrow \mathbb{R}$, $(f(x, t))_t := \frac{d}{dt}f(x, t)$, for $f(x, t) : \mathbb{R}^d \times [0, \tau] \rightarrow \mathbb{R}^d$, $(f(x, t))_x := \sum_{i=1}^d \frac{d}{dx_i}[f]_i(x, t)$, for $f(x, t) : \mathbb{R}^d \times [0, \tau] \rightarrow \mathbb{R}^d \times \mathbb{R}^d$, $(f(x, t))_{xx} := \sum_{i,j=1}^d \frac{d^2}{dx_i dx_j}[f]_{i,j}(x, t)$. Let $\Sigma^\lambda(x, t) := \sigma^\lambda(x, t)\sigma^\lambda(x, t)^\top$. Then, for $0 < t \leq \tau$,

$$\begin{aligned} (\pi_t(x))_t &= \left(\int_{\Lambda} p_t^\lambda(x)\Psi(d\lambda) \right)_t \\ &= \int_{\Lambda} \left(p_t^\lambda(x) \right)_t \Psi(d\lambda) \end{aligned} \tag{\star\star}$$

$$\begin{aligned}
 &= \int_{\Lambda} \left(\mu^\lambda(x, t) p_t^\lambda(x) \right)_x + \frac{1}{2} \left(\Sigma^\lambda(x, t) p_t^\lambda(x) \right)_{xx} \Psi(d\lambda) \\
 &= \int_{\Lambda} \left(\frac{\mu^\lambda(x, t) p_t^\lambda(x)}{\pi_t(x)} \pi_t(x) \right)_x + \frac{1}{2} \left(\frac{\Sigma^\lambda(x, t) p_t^\lambda(x)}{\pi_t(x)} \pi_t(x) \right)_{xx} \Psi(d\lambda) \\
 &= \left(\int_{\Lambda} \frac{\mu^\lambda(x, t) p_t^\lambda(x)}{\pi_t(x)} \Psi(d\lambda) \pi_t(x) \right)_x + \frac{1}{2} \left(\int_{\Lambda} \frac{\Sigma^\lambda(x, t) p_t^\lambda(x)}{\pi_t(x)} \Psi(d\lambda) \pi_t(x) \right)_{xx}. \quad (\star\star)
 \end{aligned}$$

The lines denoted with $(\star\star)$ consist of exchange of limits, the third line results from the application of the Fokker-Plank PDEs for $\{P^\lambda\}_{\lambda \in \Lambda}$. The Fokker-Plank PDE for (23) is

$$(p_t(x))_t = \left(\int_{\Lambda} \frac{\mu^\lambda(x, t) p_t^\lambda(x)}{\pi_t(x)} \Psi(d\lambda) p_t(x) \right)_x + \frac{1}{2} \left(\int_{\Lambda} \frac{\Sigma^\lambda(x, t) p_t^\lambda(x)}{\pi_t(x)} \Psi(d\lambda) p_t(x) \right)_{xx}.$$

As $P_0 = \Pi_0$ and p and π satisfy the same Fokker-Plank PDE, it follows that P and Π share the same marginal distribution. \blacksquare

Proof of Theorem 2

Theorem 2 (IDBM convergence) *For $\Gamma, \Upsilon \in \mathcal{P}_1$, $R \in \mathcal{P}_C$ associated to (3) with $\sigma_R(x, t) = I$, consider the iterates $\Pi^{(i)}, M^{(i)} \in \mathcal{P}_C(\Gamma, \Upsilon)$ of Algorithm 2. Assume that⁸: (i) $D_{KL}(C^{(0)} \| S_{0,\tau}^*) < \infty$; (ii) for each $i \geq 1$ the Cameron-Martin-Girsanov theorem hold for $M^{(i)}$ yielding $dM^{(i)} / dS^*$ (implying that (24) for $\Pi = \Pi^{(i)}$ has a unique diffusion solution with law $M^{(i)}$, and that $M^{(i)} \ll S^*$). Then: (i) $\Pi^{(i)} \xrightarrow{\mathcal{L}} S^*$ and $M^{(i)} \xrightarrow{\mathcal{L}} S^*$ as $i \rightarrow \infty$, where $\xrightarrow{\mathcal{L}}$ denotes converge in law; (ii) $D_{KL}(\Pi^{(i)} \| S^*) \geq D_{KL}(M^{(i)} \| S^*) \geq D_{KL}(\Pi^{(i+1)} \| S^*)$ for $i \geq 1$ (implying that $D_{KL}(M^{(i)} \| S^*)$ is non-increasing in i); (iii) $D_{KL}(\Pi(C) \| S^*) = D_{KL}(M(C) \| S^*)$ if and only if $\Pi(C) = M(C) = S^*$.*

Proof Let $P \in \mathcal{P}_C$ be the law of the diffusion solving

$$\begin{aligned}
 dX_t &= \mu(X_t, t) dt + \sigma_R(X_t, t) dW_t, \quad t \in [0, \tau], \\
 X_0 &\sim P_0.
 \end{aligned}$$

Let $C \in \mathcal{P}_2(\Gamma, \Upsilon)$. Within this proof, we use the following abbreviations: $\Pi(C) := \Pi(C, R_{\bullet|0,\tau})$, $M(C) := M(\Pi(C, R_{\bullet|0,\tau}))$, $S^* := S^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$. Assume that $D_{KL}(\Pi(C) \| P) < \infty$ and that the conditions required for the application of the Cameron-Martin-Girsanov theorem to obtain $dM(C) / dP$ are satisfied. As $M(C) \ll P$ as well, following Csiszar (1975), it holds that

$$\begin{aligned}
 D_{KL}(\Pi(C) \| P) - D_{KL}(\Pi(C) \| M(C)) &= \mathbb{E}_{\Pi(C)} \left[\log \frac{d\Pi(C)}{dP} \right] - \mathbb{E}_{\Pi(C)} \left[\log \frac{d\Pi(C)}{dM(C)} \right] \\
 &= \mathbb{E}_{\Pi(C)} \left[\log \frac{dM(C)}{dP} \right],
 \end{aligned}$$

and we want to show that

$$\mathbb{E}_{\Pi(C)} \left[\log \frac{dM(C)}{dP} \right] = \mathbb{E}_{M(C)} \left[\log \frac{dM(C)}{dP} \right].$$

8. As in the proof, we denote $\Pi(C) := \Pi(C, R_{\bullet|0,\tau})$, $M(C) := M(\Pi(C, R_{\bullet|0,\tau}))$, $S^* := S^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$.

As $\Pi_0(C) = M_0(C) = \Gamma$ it suffices to show that

$$\mathbb{E}_{\Pi(C)} \left[\log \frac{dM_{\bullet|0}(C)}{dP_{\bullet|0}} \right] = \mathbb{E}_{M(C)} \left[\log \frac{dM_{\bullet|0}(C)}{dP_{\bullet|0}} \right],$$

where

$$\log \frac{dM_{\bullet|0}(C)}{dP_{\bullet|0}}(x) = \int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}](x_t, t) dx_t - \frac{1}{2} \int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}(\mu_M + \mu)](x_t, t) dt$$

is given by (12). As $\Pi(\delta_{x_0} \otimes \Upsilon, R_{\bullet|0,\tau}) = \Pi_{\bullet|0}(C, R_{\bullet|0,\tau})$, from (25),

$$\begin{aligned} & \mathbb{E}_{\Pi(C)} \left[\int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}](X_t, t) dX_t \right] \\ &= \mathbb{E}_{\Pi(C)} \left[\int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}](X_t, t) (\mu_R(X_t, t) + \Sigma_R(X_t, t)) \mathbb{E}_{\Pi(C)} [\nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t) | X_t, X_0] dt \right] \\ &= \mathbb{E}_{\Pi(C)} \left[\int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}](X_t, t) (\mu_R(X_t, t) + \Sigma_R(X_t, t)) \mathbb{E}_{\Pi(C)} [\nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t) | X_t] dt \right], \end{aligned}$$

by the tower property of conditional expectations. On the other hand, from (24),

$$\begin{aligned} & \mathbb{E}_{M(C)} \left[\int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}](X_t, t) dX_t \right] \\ &= \mathbb{E}_{M(C)} \left[\int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}](X_t, t) (\mu_R(X_t, t) + \Sigma_R(X_t, t)) \mathbb{E}_{\Pi(C)} [\nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t) | X_t] dt \right] \\ &= \mathbb{E}_{\Pi(C)} \left[\int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}](X_t, t) (\mu_R(X_t, t) + \Sigma_R(X_t, t)) \mathbb{E}_{\Pi(C)} [\nabla_{X_t} \log r_{\tau|t}(X_\tau | X_t) | X_t] dt \right], \end{aligned}$$

as $M_t(C) = \Pi_t(C)$ for all $t \in [0, \tau]$. In the same way, equality in expectations of $-\frac{1}{2} \int_0^\tau [(\mu_M - \mu)^\top \Sigma_R^{-1}(\mu_M + \mu)](X_t, t) dt$ is established. We thus obtain a version of the Pythagorean law for (reverse) KL-projections (Csiszar, 1975; Nielsen, 2018),

$$D_{KL}(\Pi(C) \| P) = D_{KL}(\Pi(C) \| M(C)) + D_{KL}(M(C) \| P).$$

See also Liu et al. (2023b) for another derivation of this result. When $P = S^*$ (S^* is law of the diffusion solving (19)), under the same assumptions,

$$D_{KL}(\Pi(C) \| S^*) \geq D_{KL}(M(C) \| S^*), \quad (44)$$

with equality if and only if $\Pi(C) = M(C)$.

If $\Pi(C) = S^*$, then $M(C) = S^*$ as $\Pi(C) = CR_{\bullet|0,\tau}$ is equal in law to the diffusion process solving (19) and Theorem 1 applied to a (single) diffusion results in that same diffusion. On the other hand, if $\Pi(C) \neq S^*$, then $\Pi(C)$ is not a diffusion (it is an Ito process in the sense of Øksendal (2003)), see Section 3.2, and $M(C) \neq \Pi(C)$ (Øksendal, 2003, Theorem 8.4.3). That is $\Pi(C) = M(C)$ if and only if $\Pi(C) = M(C) = S^*$.

For $i \geq 0$,

$$\begin{aligned} D_{KL}(\Pi^{(i)} \| S^*) &= D_{KL}(\Pi_{0,\tau}^{(i)} \| S_{0,\tau}^*) + \mathbb{E}_{\Pi_{0,\tau}^{(i)}} [D_{KL}(\Pi_{\bullet|0,\tau}^{(i)} \| S_{\bullet|0,\tau}^*)] \\ &= D_{KL}(\Pi_{0,\tau}^{(i)} \| S_{0,\tau}^*), \end{aligned}$$

as $\Pi_{\bullet|0,\tau}^{(i)} = S_{\bullet|0,\tau}^* = R_{\bullet|0,\tau}$, and

$$\begin{aligned} D_{KL}(M^{(i)} \parallel S^*) &= D_{KL}(M_{0,\tau}^{(i)} \parallel S_{0,\tau}^*) + \mathbb{E}_{M_{0,\tau}^{(i)}} [D_{KL}(M_{\bullet|0,\tau}^{(i)} \parallel S_{\bullet|0,\tau}^*)] \\ &\geq D_{KL}(\Pi_{0,\tau}^{(i+1)} \parallel S_{0,\tau}^*), \end{aligned}$$

as $M_{0,\tau}^{(i)} = \Pi_{0,\tau}^{(i+1)}$. Thus,

$$D_{KL}(M^{(i)} \parallel S^*) \geq D_{KL}(\Pi^{(i+1)} \parallel S^*). \quad (45)$$

It is assumed that $D_{KL}(\Pi^{(0)} \parallel S^*) = D_{KL}(C^{(0)} \parallel S_{0,\tau}^*) < \infty$, and that the Cameron-Martin-Girsanov theorem applies to each $M^{(i)}$ yielding $dM^{(i)}/dS^*$. Therefore, we can iteratively apply (44) and (45) obtaining $D_{KL}(\Pi^{(i)} \parallel S^*) \geq D_{KL}(M^{(i)} \parallel S^*) \geq D_{KL}(\Pi^{(i+1)} \parallel S^*)$ for $i \geq 0$. Being non-increasing and bounded below, the sequence $D_{KL}(\Pi^{(i)} \parallel S^*), D_{KL}(M^{(i)} \parallel S^*), D_{KL}(\Pi^{(i+1)} \parallel S^*)$ converges, and $\lim_{i \rightarrow \infty} (D_{KL}(\Pi^{(i)} \parallel S^*) - D_{KL}(M^{(i)} \parallel S^*)) = \lim_{i \rightarrow \infty} D_{KL}(\Pi^{(i)} \parallel M^i) = 0$.

The sequences $\{\Pi^{(i)}\}_{i \geq 0}, \{M^{(i)}\}_{i \geq 0}$ are tight. We consider $\{\Pi^{(i)}\}_{i \geq 0}$, the case of $\{M^{(i)}\}_{i \geq 0}$ being identical. By the conditional Jensen inequality, for any measurable set K ,

$$\begin{aligned} D_{KL}(\Pi^{(i)} \parallel S^*) &= \mathbb{E}_{\Pi^{(i)}} \left[-\log \frac{dS^*}{d\Pi^{(i)}} \mid K^c \right] \Pi^{(i)}[K^c] + \mathbb{E}_{\Pi^{(i)}} \left[-\log \frac{dS^*}{d\Pi^{(i)}} \mid K \right] \Pi^{(i)}[K] \\ &\geq -\log(S^*[K^c]/\Pi^{(i)}[K^c]) \Pi^{(i)}[K^c] + -\log(S^*[K]/\Pi^{(i)}[K]) \Pi^{(i)}[K]. \end{aligned}$$

For $\varepsilon > 0$, choose K compact such $S^*[K^c] < \varepsilon, S^*[K] \geq 1 - \varepsilon$ by the tightness of S^* (as it is defined on a Polish space, see Léonard (2014a)). Assume that $\Pi^{(i)}$ is not tight. That is, there is $\gamma > 0$ such that for each compact K there is $i \geq 0$ with $\Pi^{(i)}[K^c] \geq \gamma, \Pi^{(i)}[K] < 1 - \gamma$. Hence, for each $\varepsilon > 0$, there is $i \geq 0$ such that

$$-\log(S^*[K^c]/\Pi^{(i)}[K^c]) \Pi^{(i)}[K^c] \geq -\log(\varepsilon/\gamma)\gamma,$$

which can be made arbitrarily large by a suitable small $\varepsilon > 0$. On the other hand, $-\log(S^*[K]/x)x$ is bounded below for $x \in [0, 1]$, so $-\log(S^*[K]/\Pi^{(i)}[K]) \Pi^{(i)}[K]$ is bounded below. But we know that for i large enough $D_{KL}(\Pi^{(i)} \parallel S^*)$ is bounded above, hence $\Pi^{(i)}$ must be tight.

Therefore, $\{\Pi^{(i)}\}_{i \geq 0}$ and $\{M^{(i)}\}_{i \geq 0}$ are relatively compact. Each subsequence of $\{\Pi^{(i)}\}_{i \geq 0}$ has a further subsequence $\{\Pi^{(l)}\}_{l \geq 1}$ which converges in law: $\Pi^{(l)} \xrightarrow{\mathcal{L}} \Pi^{(\infty)}$ as $l \rightarrow \infty$ for some $\Pi^{(\infty)} \in \mathcal{P}_C(\Gamma, \Upsilon)$. Each subsequence of $\{M^{(i)}\}_{i \geq 0}$ has a further subsequence $\{M^{(l)}\}_{l \geq 1}$ which converges in law: $M^{(l)} \xrightarrow{\mathcal{L}} M^{(\infty)}$ for some $M^{(\infty)} \in \mathcal{P}_C(\Gamma, \Upsilon)$. By the lower semi-continuity of the KL divergence with respect to the topology of weak convergence (van Erven and Harremos, 2014, Theorem 19), $0 = \liminf_{l \rightarrow \infty} D_{KL}(\Pi^{(l)} \parallel M^l) \geq D_{KL}(\Pi^{(\infty)} \parallel M^\infty)$ and therefore $\Pi^{(\infty)} = M^{(\infty)}$. Denote this common PM with $S^{(\infty)}$.

Due to the convergence in law of $\Pi^{(l)}$, and the form of its disintegration, $S^{(\infty)}$ is of the form $S_{0,\tau}^{(\infty)} R_{\bullet|0,\tau}$ for some $S_{0,\tau}^{(\infty)} \in \mathcal{P}_2(\Gamma, \Upsilon)$. If $S^{(\infty)}$ is also diffusion, it follows that $S_{0,\tau}^{(\infty)} = S_{0,\tau}^{(*)}$ (Section 3.2) and thus $S^{(\infty)} = S^*$. It remains to show that $S^{(\infty)}$ is a diffusion when $\sigma_R(x, t) = I$. Again by the lower semi-continuity of the KL divergence, we have that $D_{KL}(S^{(\infty)} \parallel S^*) < \infty$ and thus $S^{(\infty)} \ll S^*$. But $S^* \ll R$ and $R \ll R^\circ$ by assumption where

R° is the d -dimensional Brownian measure on $[0, \tau]$. Then, by Liptser and Shiriaev (1977, Theorem 7.11 and following Note 1), $S^{(\infty)}$ is a diffusion.

As convergence in law to S^* has been established for arbitrary convergent subsequences, $\{\Pi^{(i)}\} \xrightarrow{\mathcal{L}} S^{(\infty)}$ and $\{M^{(i)}\} \xrightarrow{\mathcal{L}} S^{(\infty)}$ by Billingsley (1999, Theorem 2.6). \blacksquare

B. Additional SDE Class Formulae

The linearity of (30) and (31) stands behind the form of the conditional expectations entering the drift coefficients that follow. The generative time reversal process (16) is given by

$$\begin{aligned} d\bar{X}_t &= \left[\alpha \beta_{\mathfrak{r}} \bar{X}_t + \beta_{\mathfrak{r}} \frac{\mathbb{E}_R[\bar{X}_\tau | \bar{X}_t] a(0, b_{\mathfrak{r}}) - \bar{X}_t}{v(0, b_{\mathfrak{r}})} \right] dt + \sqrt{\beta_{\mathfrak{r}}} \Gamma^{1/2} dW_t, \quad t \in [0, \tau], \\ \bar{X}_0 &\sim R_\tau \approx \Upsilon, \end{aligned}$$

the DBM transport (24) is given by

$$\begin{aligned} dX_t &= \left[-\alpha \beta_t X_t + \beta_t \frac{\mathbb{E}_{\Pi}[X_\tau | X_t] a(b_t, b_\tau) - X_t a(b_t, b_\tau)^2}{v(b_t, b_\tau)} \right] dt + \sqrt{\beta_t} \Gamma^{1/2} dW_t, \quad t \in [0, \tau], \\ X_0 &\sim \Gamma, \end{aligned}$$

while the BDBM transport (26) is given by

$$\begin{aligned} d\bar{X}_t &= \left[\alpha \beta_t \bar{X}_t + \beta_{\mathfrak{r}} \frac{\mathbb{E}_{\Pi}[\bar{X}_\tau | \bar{X}_t] a(0, b_{\mathfrak{r}}) - \bar{X}_t}{v(0, b_{\mathfrak{r}})} \right] dt + \sqrt{\beta_{\mathfrak{r}}} \Gamma^{1/2} dW_t, \quad t \in [0, \tau], \\ \bar{X}_0 &\sim \Upsilon. \end{aligned}$$

For $\alpha = 0$, the DBM and BDBM transports are symmetric in the following sense: the BDM transport based on $C \in \mathcal{P}_2(\Gamma, \Upsilon)$ and β_t is equivalent in law to the BDBM transport based on $\bar{C} \in \mathcal{P}_2(\Upsilon, \Gamma)$ and $\bar{\beta}_t := \beta_{\mathfrak{r}}$.

C. Generative Modeling Code

```

1 # requires:
2 #     b_t(t): b_t(t)
3 #     a_s_t(s, t): a(s,t) from (29)
4 #     v_s_t(s, t): v(s,t) from (29)
5 # inputs:
6 #     x_0: [B,C,H,W]
7 #     model: ([B,C,H,W], [B]) -> [B,C,H,W]
8 # outputs:
9 #     losses: [B]
10 def sgm_loss(x_0, model, T=1.0):
11     t = torch.rand(x_0.shape[0], device=x_0.device) * T           # [B]
12     scaled_t = b_t(t)                                              # [B]
13     z = torch.randn_like(x_0)                                         # [B,C,H,W]
14     a_0_t, v_0_t = a_s_t(0.0, scaled_t), v_s_t(0.0, scaled_t)      # [B]
15     s_0_t = torch.sqrt(v_0_t)                                         # [B]
16     x_t = a_0_t[:, None, None, None] * x_0 + s_0_t[:, None, None, None] * z    # [B,C,H,W]
17     score_t = (x_0 * a_0_t[:, None, None, None] - x_t) / v_0_t[:, None, None, None] # [B,C,H,W]
18     losses = (model(x_t, scaled_t) - score_t)**2 * v_0_t[:, None, None, None]      # [B,C,H,W]
19     losses = torch.sum(losses.view(losses.shape[0], -1), dim=1)                      # [B]
20     return losses

```

Listing 1: SGM loss sampling; x_0 is a batch of B images, each of which has C channels, height H and width W .

```

1 # requires:
2 #     b_t(t): b_t(t)
3 #     a_s_t(s, t): a(s,t) from (29)
4 #     v_s_t(s, t): v(s,t) from (29)
5 #     al_s_t_u(s, t, u):  $\hat{a}(s,t,u)$  from (32)
6 #     ar_s_t_u(s, t, u):  $\check{a}(s,t,u)$  from (32)
7 #     vb_s_t_u(s, t, u):  $\tilde{v}(s,t,u)$  from (32)
8 # inputs:
9 #     x_0: [B,C,H,W]
10 #     model: ([B,C,H,W], [B]) -> [B,C,H,W]
11 # outputs:
12 #     losses: [B]
13 def bdbm_loss(x_0, model, T=1.0, sigma_T=1.0):
14     t = torch.rand(x_0.shape[0], device=x_0.device) * T           # [B]
15     scaled_t, scaled_T = b_t(t), b_t(T)                           # [B]
16     z = torch.randn_like(x_0)                                         # [B,C,H,W]
17     a_0_t, v_0_t = a_s_t(0.0, scaled_t), v_s_t(0.0, scaled_t)      # [B]
18     x_T = torch.randn_like(x_0) * sigma_T                         # [B,C,H,W]
19     al_0_t_T, ar_0_t_T, = al_s_t_u(0.0, t, scaled_T), ar_s_t_u(0.0, t, scaled_T) # [B]
20     vb_0_t_T = vb_s_t_u(0.0, t, scaled_T)                         # [B]
21     sb_0_t_T = torch.sqrt(vb_0_t_T)                                # [B]
22     x_t = (
23         + al_0_t_T[:, None, None, None] * x_0
24         + ar_0_t_T[:, None, None, None] * x_T
25         + sb_0_t_T[:, None, None, None] * z
26     )                                                               # [B,C,H,W]
27     score_t = (x_0 * a_0_t[:, None, None, None] - x_t) / v_0_t[:, None, None, None] # [B,C,H,W]
28     losses = (model(x_t, scaled_t) - score_t)**2 * v_0_t[:, None, None, None]      # [B,C,H,W]
29     losses = torch.sum(losses.view(losses.shape[0], -1), dim=1)                      # [B]
30     return losses

```

Listing 2: BDBM loss sampling; x_0 is a batch of B images, each of which has C channels, height H and width W .

D. Additional Experimental Results

		X_t		IDBM, $\sigma = 1.0$		E_t		IDBM, $\sigma = 0.5$		X_t		IDBM, $\sigma = 0.2$		E_t		BDBM, $\sigma = 1.0$		X_t	
C	d	a	d	b	C	d	a	d	b	C	d	a	d	b	C	d	a	d	b
e	c	b	e	E	e	c	b	e	E	e	c	b	e	E	e	c	b	e	E
C	e	e	e	C	C	e	e	e	C	C	e	e	e	C	C	e	e	C	C
a	s	d	A	d	a	s	d	A	d	a	s	d	A	d	a	s	d	A	d
D	e	e	a	e	D	e	e	a	e	D	e	e	a	e	D	e	e	a	e
7	7	9	8	5	7	7	9	8	5	7	7	9	8	5	7	7	9	8	5
9	7	5	9	5	4	7	5	9	5	4	7	5	9	5	4	7	5	9	5
1	9	4	9	7	1	9	4	9	7	1	9	4	9	7	1	9	4	9	7
2	8	3	4	7	2	6	3	4	7	2	6	3	4	7	2	6	3	4	7
0	9	1	9	0	0	9	1	9	0	0	9	1	9	0	0	9	1	9	0
C	d	a	d	b	C	d	a	d	b	C	d	a	d	b	C	d	a	d	b
e	c	b	e	E	e	c	b	e	E	e	c	b	e	E	e	c	b	e	E
C	e	e	e	C	C	e	e	e	C	C	e	e	e	C	C	e	e	C	C
a	s	d	A	d	a	s	d	A	d	a	s	d	A	d	a	s	d	A	d
D	e	e	a	e	D	e	e	a	e	D	e	e	a	e	D	e	e	a	e
7	4	9	8	5	7	4	9	8	5	7	4	9	8	5	7	4	9	8	5
8	7	3	8	5	8	7	3	8	5	8	7	3	8	5	8	7	3	8	5
4	9	9	9	8	1	9	9	9	8	1	9	9	9	8	1	9	9	9	8
7	8	3	4	3	7	8	3	4	3	7	8	3	4	3	7	8	3	4	3
0	9	1	9	0	0	9	1	9	0	0	9	1	9	0	0	9	1	9	0
C	d	a	d	b	C	d	a	d	b	C	d	a	d	b	C	d	a	d	b
e	c	b	e	E	e	c	b	e	E	e	c	b	e	E	e	c	b	e	E
C	e	e	e	C	C	e	e	e	C	C	e	e	e	C	C	e	e	C	C
a	s	d	A	d	a	s	d	A	d	a	s	d	A	d	a	s	d	A	d
D	e	e	a	e	D	e	e	a	e	D	e	e	a	e	D	e	e	a	e
3	4	9	7	8	3	4	9	7	3	3	4	9	7	3	3	4	9	7	3
8	7	3	8	5	8	7	3	8	5	8	7	3	8	5	8	7	3	8	5
7	9	9	9	8	7	9	9	9	8	7	9	9	9	8	7	9	9	9	8
2	8	4	4	4	2	8	4	4	4	2	8	4	4	4	2	8	4	4	4
0	9	3	9	0	0	9	3	9	0	0	9	3	9	0	0	9	3	9	0
C	d	a	d	b	C	d	a	d	b	C	d	a	d	b	C	d	a	d	b
e	c	b	e	E	e	c	b	e	E	e	c	b	e	E	e	c	b	e	E
C	e	e	e	C	C	e	e	e	C	C	e	e	e	C	C	e	e	C	C
a	s	d	A	d	a	s	d	A	d	a	s	d	A	d	a	s	d	A	d
D	e	e	a	e	D	e	e	a	e	D	e	e	a	e	D	e	e	a	e
3	4	9	7	8	3	4	9	7	3	3	4	9	7	3	3	4	9	7	3
8	7	3	8	5	8	7	3	8	5	8	7	3	8	5	8	7	3	8	5
7	9	9	9	8	7	9	9	9	8	7	9	9	9	8	7	9	9	9	8
2	8	4	4	4	2	8	4	4	4	2	8	4	4	4	2	8	4	4	4
0	9	3	9	0	0	9	3	9	0	0	9	3	9	0	0	9	3	9	0
C	d	a	d	b	C	d	a	d	b	C	d	a	d	b	C	d	a	d	b
e	c	b	e	E	e	c	b	e	E	e	c	b	e	E	e	c	b	e	E
C	e	e	e	C	C	e	e	e	C	C	e	e	e	C	C	e	e	C	C
a	s	d	A	d	a	s	d	A	d	a	s	d	A	d	a	s	d	A	d
D	e	e	a	e	D	e	e	a	e	D	e	e	a	e	D	e	e	a	e
3	4	9	7	8	3	4	9	7	3	3	4	9	7	3	3	4	9	7	3
8	7	3	8	5	8	7	3	8	5	8	7	3	8	5	8	7	3	8	5
7	9	9	9	8	7	9	9	9	8	7	9	9	9	8	7	9	9	9	8
2	8	4	4	4	2	8	4	4	4	2	8	4	4	4	2	8	4	4	4
0	9	3	9	0	0	9	3	9	0	0	9	3	9	0	0	9	3	9	0

Figure 10: Random paths X_t and corresponding terminal value estimators E_t for fully trained BDBM and IDBM procedures corresponding to different regularization levels; backward time direction; t uniformly spaced on $[0, 1]$ for X_t , $[0, 1 - \Delta t]$ for E_t ; $X_0 \sim D_{\text{EMNIST}}^{\text{test}}$ is fixed.

X_t	IDBM $\sigma = 1.0$	E_t	X_t	IDBM $\sigma = 0.5$	E_t	X_t	IDBM $\sigma = 0.2$	E_t
7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	C b d e B d C e e C e a c c C c e D B b b b C d b	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	C b d e B d C e e C e a c c C c e D B b b b C d b	C b d e B d C e e C e a c c C c e D B b b b C d b	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	C b d e d d C e e C e a c c C c e D B b b b C d b	C b d e d d C e e C e a c c C c e D B b b b C d b
7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	C b d e d d C e e C e a c c C c e D B b b b C d b	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	C b d e d d C e e C e a c c C c e D B b b b C d b	C b d e d d C e e C e a c c C c e D B b b b C d b	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	d b e e A 2 C e e C e a c c C d e c c b C b b a b	d b e e A 2 C e e C e a c c C d e c c b C b b a b
7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	d b e e A 2 C e e C e a c c C d e c c b C b b a b	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	d b e e A 2 C e e C e a c c C d e c c b C b b a b	d b e e A 2 C e e C e a c c C d e c c b C b b a b	7 3 3 9 7 2 9 8 0 5 8 4 5 3 9 9 3 0 3 6 5 8 1 2 6	d b e e A 2 C e e C e a c c C d e c c b C b b a b	d b e e A 2 C e e C e a c c C d e c c b C b b a b

Figure 11: Random paths X_t and corresponding terminal value estimators E_t for fully trained IDBM procedures corresponding to different regularization levels; forward time direction; t uniformly spaced on $[0, 1]$ for X_t , $[0, 1 - \Delta t]$ for E_t ; $X_0 \sim D_{\text{MNIST}}^{\text{test}}$ is fixed.

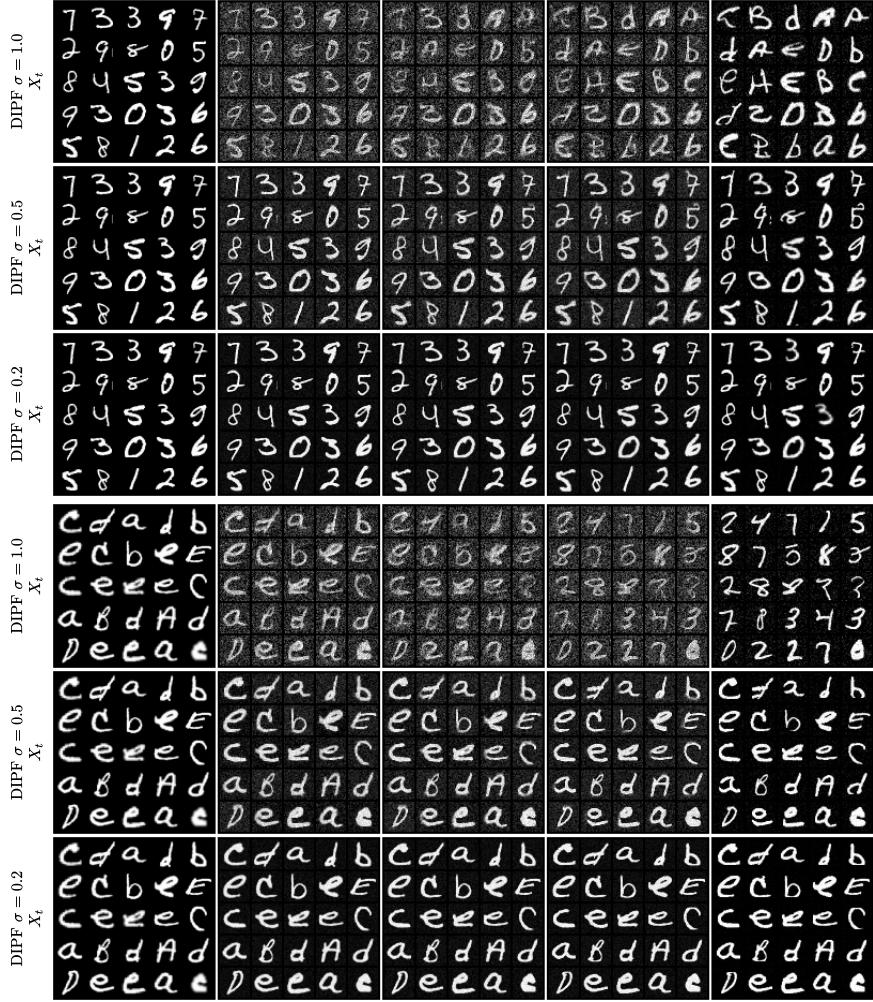


Figure 12: Random paths X_t for fully trained DIPF procedures corresponding to different regularization levels; forward (first three rows) and backward (last three rows) time directions; t uniformly spaced on $[0, 1]$; $X_0 \sim D_{\text{MNIST}}^{\text{test}}$ (forward time) and $X_0 \sim D_{\text{EMNIST}}^{\text{test}}$ (backward time) are fixed.

E. Notation

Notation	Description
$[0, \tau]$	Time interval
d	Dimension of state space
P, Q, \dots	Probability measure (PM)
\mathcal{P}_C	PMs over $\mathcal{C}([0, \tau], \mathbb{R}^d)$
\mathcal{P}_n	PMs over $\mathbb{R}^{d \times n}$
$\mathcal{P}_C(\Gamma, \Upsilon)$	PMs with initial-terminal distributions Γ, Υ ; $\mathcal{P}_C(\Gamma, \Upsilon) \subseteq \mathcal{P}_C$
$\mathcal{P}_2(\Gamma, \Upsilon)$	PMs with marginal distributions Γ, Υ ; $\mathcal{P}_2(\Gamma, \Upsilon) \subseteq \mathcal{P}_2$
$\mathcal{U}(0, \tau)$	Uniform distribution on $(0, \tau)$
$\mathcal{N}_d(\mu, \Sigma)$	d -variate normal distribution with mean μ and covariance Σ
dP/dQ	Density of PM P with respect to PM Q
p	Lebesgue density of PM P
P_{t_1, \dots, t_n}	Marginalization of $P \in \mathcal{P}_C$ at t_1, \dots, t_n
$P_{\bullet t_1, \dots, t_n}$	Conditioning of $P \in \mathcal{P}_C$ given values at t_1, \dots, t_n
$P = P_{t_1, \dots, t_n} P_{\bullet t_1, \dots, t_n}$	Marginal-conditional decomposition of $P \in \mathcal{P}_C$
$Q = \Psi P_{\bullet t_1, \dots, t_n}$	Mixing of $P \in \mathcal{P}_C$ via $\Psi \in \mathcal{P}_n$ at times t_1, \dots, t_n ; $Q \in \mathcal{P}_C$
$D_{KL}(S \ R)$	Kullback-Leibler divergence from PM S to PM R
I	Identity matrix
A^\top	Transposition of a square matrix A
$\ V\ $	Euclidean norm of a vector V
\bar{X}	Time reversal of stochastic process X
\bar{P}	Time reversal of PM P ; $P \in \mathcal{P}_C$
\mathfrak{r}	Reverse timescale; $\mathfrak{r} := \tau - t$
W	d -dimensional standard Brownian motion
Γ, Υ	Target PMs; $\Gamma, \Upsilon \in \mathcal{P}_1$
R	Reference PM; $R \in \mathcal{P}_C$
$\mu_R(x, t), \sigma_R(x, t)$	Reference SDE drift and diffusion coefficients
$\Sigma_R(x, t)$	Reference SDE covariance coefficient; $\Sigma_R := \sigma_R \sigma_R^\top$
$S^*(\Gamma, \Upsilon, R, \mathcal{P}_C)$	Solution (S_{dyn}) (dynamic)
$S^*(\Gamma, \Upsilon, R_{0,\tau}, \mathcal{P}_2)$	Solution to (S_{sta}) (static) (corresponding to (S_{dyn}))
$S^*(\Gamma, \cdot, Q, \mathcal{P})$	Solution to forward (H_{dyn}) ($\mathcal{P} = \mathcal{P}_C$), (H_{sta}) ($\mathcal{P} = \mathcal{P}_2$); $Q \in \mathcal{P}$
$S^*(\cdot, \Upsilon, Q, \mathcal{P})$	Solution to backward (H_{dyn}) ($\mathcal{P} = \mathcal{P}_C$), (H_{sta}) ($\mathcal{P} = \mathcal{P}_2$); $Q \in \mathcal{P}$
$F^{(i)}$	i -th IPF iteration; $F^{(i)} \in \mathcal{P}_C \cup \mathcal{P}_2$, $i \geq 0$
$\mu_{F^{(i)}}$	Drift corresponding to $F^{(i)} \in \mathcal{P}_C$
$\Pi(C, R_{\bullet 0,\tau})$	Diffusion mixture; $C \in \mathcal{P}_2(\Gamma, \Upsilon)$, $\Pi \in \mathcal{P}_C(\Gamma, \Upsilon)$
$M(\Pi(C, R_{\bullet 0,\tau}))$	Forward DBM based on C ; $C \in \mathcal{P}_2(\Gamma, \Upsilon)$, $M \in \mathcal{P}_C(\Gamma, \Upsilon)$
$\bar{M}(\Pi(C, R_{\bullet 0,\tau}))$	Backward DBM based on C ; $C \in \mathcal{P}_2(\Gamma, \Upsilon)$, $\bar{M} \in \mathcal{P}_C(\Upsilon, \Gamma)$
$M^{(i)}$	i -th (forward) IDBM iteration; $M^{(i)} \in \mathcal{P}_C$, $i \geq 0$
$C^{(i)}$	i -th IDBM coupling; $C^{(i)} \in \mathcal{P}_2(\Gamma, \Upsilon)$, $C^{(i)} := M_{0,\tau}^{(i)}$
$\Pi^{(i)}$	i -th IDBM diffusion mixture; $\Pi^{(i)} \in \mathcal{P}_C$, $\Pi^{(i)} := \Pi(C^{(i-1)}, R_{\bullet 0,\tau})$
$\mu_{M^{(i)}}$	Drift corresponding to $M^{(i)} \in \mathcal{P}_C$

Table 3: Main notation summary.

References

- Brian D.O. Anderson. Reverse-Time Diffusion Equation Models. *Stochastic Processes and their Applications*, 12(3):313–326, May 1982.
- Yacine Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906–937, 2008. ISSN 0090-5364, 2168-8966.
- Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E. Jacob. Schrödinger Bridge Samplers, 2019.
- Alexandros Beskos and Gareth O. Roberts. Exact simulation of diffusions. *The Annals of Applied Probability*, 15(4):2422–2444, 2005. ISSN 1050-5164, 2168-8737.
- Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O. Roberts. A Factorisation of Diffusion Measure and Finite Sample Path Constructions. *Methodology and Computing in Applied Probability*, 10(1):85–104, 2008. ISSN 1573-7713.
- Patrick Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Probability and Statistics Section. Wiley, 2nd ed edition, 1999.
- Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- Damiano Brigo. The General Mixture-Diffusion SDE and Its Relationship with an Uncertain-Volatility Option Model with Volatility-Asset Decorrelation, December 2002.
- Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.
- Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representations*, 2022.
- I. Csiszar. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146–158, 1975. ISSN 0091-1798.
- Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied Mathematics and Optimization*, 23(1):313–329, 1991. ISSN 1432-0606.
- Valentin De Bortoli, Arnaud Doucet, Jeremy Heng, and James Thornton. Simulating Diffusion Bridges with Score Matching, 2021.
- W. Edwards Deming and Frederick F. Stephan. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940. ISSN 0003-4851.
- Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn’s Algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1367–1376. PMLR, 2018.

Tarek A. El Moselhy and Youssef M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012. ISSN 0021-9991.

David Lopes Fernandes, Francisco Vargas, Carl Henrik Ek, and Neill D. F. Campbell. Shooting Schrödinger’s Cat. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. ISSN 1533-7928.

Robert Fortet. Résolution d’un système d’équations de M. Schrödinger. *J. Math. Pure Appl. IX*, 1:83–105, 1940.

Avner Friedman. *Stochastic Differential Equations and Applications - Volume 1*. Probability and Mathematical Statistics Series ; v. 28. Academic Press, 1975.

Christiane Fuchs. *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Science & Business Media, 2013.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images, 2023.

Jian Huang, Yuling Jiao, Lican Kang, Xu Liao, Jin Liu, and Yanyan Liu. Schrödinger-Föllmer Sampler: Sampling without Ergodicity, 2021.

A. S. Hurn, J. I. Jeisman, and K. A. Lindsay. Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations. *Journal of Financial Econometrics*, 5(3):390–455, 2007. ISSN 1479-8409.

- Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. ISSN 1533-7928.
- Benton Jamison. Reciprocal processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 30(1):65–86, 1974. ISSN 1432-2064.
- Benton Jamison. The Markov processes of Schrödinger. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(4):323–331, 1975. ISSN 1432-2064.
- Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10468–10479. Curran Associates, Inc., 2020.
- Mathieu Kessler, Alexander Lindner, and Michael Sørensen. *Statistical Methods for Stochastic Differential Equations*. Chapman and Hall/CRC, 2012.
- 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.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023.
- Robert Shevilevich Liptser and Al'bert Nikolaevich Shiriaev. *Statistics of Random Processes: General Theory*, volume 394. Springer, 1977.
- Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. I²SB: Image-to-Image Schrödinger Bridge, 2023a.
- Qiang Liu. Rectified Flow: A Marginal Preserving Approach to Optimal Transport, 2022.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, 2022.
- Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Learning Diffusion Bridges on Constrained Domains. In *The Eleventh International Conference on Learning Representations*, 2023b.
- Christian Léonard. Some Properties of Path Measures. In Catherine Donati-Martin, Antoine Lejay, and Alain Rouault, editors, *Séminaire de Probabilités XLVI*, Lecture Notes in Mathematics, pages 207–230. Springer International Publishing, 2014a.
- Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems*, 34(4):1533, 2014b.
- Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Information Geometry*, 5(1):289–323, 2022. ISSN 2511-249X.

Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. Sampling via Measure Transport: An Introduction. In Roger Ghanem, David Higdon, and Houman Owhadi, editors, *Handbook of Uncertainty Quantification*, pages 1–41. Springer International Publishing, 2016.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Frank Nielsen. What is an information projection. *Notices of the AMS*, 65(3):321–324, 2018.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. ISSN 1533-7928.

Michele Pavon, Esteban G. Tabak, and Giulio Trigila. The data-driven Schrödinger bridge, 2018.

Stefano Peluchetti. Non-Denoising Forward-Time Diffusions. 2021.

Stefano Peluchetti and Gareth O. Roberts. A Study of the Efficiency of Exact Methods for Diffusion Simulation. In Leszek Plaskota and Henryk Woźniakowski, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2010*, Springer Proceedings in Mathematics & Statistics, pages 161–187. Springer, 2012.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. 2020.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ludger Ruschendorf. Convergence of the Iterative Proportional Fitting Procedure. *The Annals of Statistics*, 23(4):1160–1174, 1995. ISSN 0090-5364.

Hamza Ruzayqat, Alexandros Beskos, Dan Crisan, Ajay Jasra, and Nikolas Kantas. Unbiased estimation using a class of diffusion processes. *Journal of Computational Physics*, 472: 111643, 2023. ISSN 0021-9991.

L. Rüschenhof and W. Thomsen. Note on the Schrödinger equation and I-projections. *Statistics & Probability Letters*, 17(5):369–375, 1993. ISSN 0167-7152.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion Schrödinger Bridge Matching, 2023.

Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned Diffusion Schrödinger Bridges. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 1985–1995. PMLR, 2023.

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.

Daniel W Stroock and S. R. S Varadhan. *Multidimensional Diffusion Processes*. 2006.

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.

James Thornton, Michael Hutchinson, Emile Mathieu, Valentin De Bortoli, Yee Whye Teh, and Arnaud Doucet. Riemannian Diffusion Schrödinger Bridge, 2022.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional Flow Matching: Simulation-Free Dynamic Optimal Transport, 2023.

Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 3084–3114. PMLR, 2019.

Tim van Erven and Peter Harremos. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. ISSN 1557-9654.

Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving Schrödinger Bridges via Maximum Likelihood. *Entropy*, 23(9):1134, September 2021.

Francisco Vargas, Andrius Ovsianas, David Fernandes, Mark Girolami, Neil D. Lawrence, and Nikolas Nüsken. Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):3, 2022. ISSN 1573-1375.

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. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10794–10804. PMLR, 2021.

Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-based Molecule Generation with Informative Prior Bridges. In *Advances in Neural Information Processing Systems*, 2022.

Mao Ye, Lemeng Wu, and Qiang Liu. First Hitting Diffusion Models for Generating Manifold, Graph and Categorical Data. In *Advances in Neural Information Processing Systems*, 2022.

Qinsheng Zhang and Yongxin Chen. Path Integral Sampler: A Stochastic Control Approach For Sampling. In *International Conference on Learning Representations*, 2022.

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.