Stochastic Interpolants:
A Unifying Framework for Flows and Diffusions

Michael S. Albergo[*]   ALBERGO@NYU.EDU
Center for Cosmology and Particle Physics
New York University
New York, NY 10012, USA

Nicholas M. Boffi[*]   BOFFI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012, USA

Eric Vanden-Eijnden   EVE2@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012, USA

Editor: Maxim Raginsky

Abstract

A class of generative models that unifies flow-based and diffusion-based methods is introduced. These models extend the framework proposed in Albergo and Vanden-Eijnden (2023), enabling the use of a broad class of continuous-time stochastic processes called 'stochastic interpolants' to bridge any two probability density functions exactly in finite time. These interpolants are built by combining data from the two prescribed densities with an additional latent variable that shapes the bridge in a flexible way. The time-dependent probability density function of the stochastic interpolant is shown to satisfy a first-order transport equation as well as a family of forward and backward Fokker-Planck equations with tunable diffusion coefficient. Upon consideration of the time evolution of an individual sample, this viewpoint immediately leads to both deterministic and stochastic generative models based on probability flow equations or stochastic differential equations with an adjustable level of noise. The drift coefficients entering these models are time-dependent velocity fields characterized as the unique minimizers of simple quadratic objective functions, one of which is a new objective for the score of the interpolant density. We show that minimization of these quadratic objectives leads to control of the likelihood for generative models built upon stochastic dynamics, while likelihood control for deterministic dynamics is more stringent. We also construct estimators for the likelihood and the cross-entropy of interpolant-based generative models, and we discuss connections with other methods such as score-based diffusion models, stochastic localization processes, probabilistic denoising techniques, and rectifying flows. In addition, we demonstrate that stochastic interpolants recover the Schrödinger bridge between the two target densities when explicitly optimizing over the interpolant. Finally, algorithmic aspects are discussed and the approach is illustrated on numerical examples.

---

[*]. Author ordering alphabetical; authors contributed equally.

Contents

1. Introduction

1.1 Background and motivation

Dynamical approaches for deterministic and stochastic transport have become a central theme in contemporary generative modeling research. At the heart of progress is the idea to use ordinary or stochastic differential equations (ODEs/SDEs) to continuously transform samples from a base probability density function (PDF) ρ_0 into samples from a target PDF ρ_1 (or vice-versa), and the realization that inference over the velocity field in these equations can be formulated as an empirical risk minimization problem over a parametric class of functions (Grathwohl et al., 2019; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021b; Ben-Hamu et al., 2022; Albergo and Vanden-Eijnden, 2023; Liu et al., 2023b; Lipman et al., 2023).

A major milestone was the introduction of score-based diffusion methods (SBDM) (Song et al., 2021b), which map an arbitrary density into a standard Gaussian by passing samples through an Ornstein-Uhlenbeck (OU) process. The key insight of SBDM is that this process can be reversed by introducing a backwards SDE whose drift coefficient depends on the score of the time-dependent density of the process. By learning this score – which can be done by minimization of a quadratic objective function known as the denoising loss (Vincent, 2011) – the backwards SDE can be used as a generative model that maps Gaussian noise into data from the target. Though theoretically exact, the mapping takes infinite time in both directions, and hence must be truncated in practice.

While diffusion-based methods have become state-of-the-art for tasks such as image generation, there remains considerable interest in developing methods that bridge two *arbitrary* densities (rather than requiring one to be Gaussian), that accomplish the transport *exactly*, and that do so on a *finite* time interval. Moreover, while the highest quality results from score-based diffusion were originally obtained using SDEs (Song et al., 2021b), this has been challenged by recent works that find equivalent or better performance with ODE-based methods if the score is learned sufficiently well (Karras et al., 2022). If made to match the performance of their stochastic counterparts, ODE-based methods exhibit a number of desirable characteristics, such as an exact, computationally tractable formula for the likelihood and the easy application of well-developed adaptive integration schemes for sampling. It is an open question of significant practical importance to understand if

there exists a separation in sample quality between generative models based on deterministic dynamics and those based on stochastic dynamics.

In order to satisfy the desirable characteristics outlined in the previous paragraph, we develop a framework for generative modeling based on the method proposed in Albergo and Vanden-Eijnden (2023), which is built on the notion of a *stochastic interpolant* x_t used to bridge two arbitrary densities ρ_0 and ρ_1. We will consider more general designs below, but as one example the reader can keep in mind:

$$x_t = (1-t)x_0 + tx_1 + \sqrt{2t(1-t)}z, \quad t \in [0,1], \tag{1.1}$$

where x_0, x_1, and z are random variables drawn independently from ρ_0, ρ_1, and the standard Gaussian density $\mathsf{N}(0, Id)$, respectively. The stochastic interpolant x_t defined in (1.1) is a continuous-time stochastic process that, by construction, satisfies $x_{t=0} = x_0 \sim \rho_0$ and $x_{t=1} = x_1 \sim \rho_1$. Its paths therefore *exactly* bridge between samples from ρ_0 at $t=0$ and from ρ_1 at $t=1$. A key observation is that:

> *The law of the interpolant x_t at any time $t \in [0,1]$ can be realized by many different processes, including an ODE and forward and backward SDEs whose drifts can be learned from data.*

To see why this is the case, one must consider the probability distribution of the interpolant x_t. As shown below, for a large class of densities ρ_0 and ρ_1 supported on $\mathbb{R}^d$, this distribution is absolutely continuous with respect to the Lebesgue measure. Moreover, its time-dependent density $\rho(t)$ satisfies a first-order transport equation and a family of forward and backward Fokker-Planck equations in which the diffusion coefficient can be varied at will. Out of these equations, we can readily derive generative models that satisfy ODEs and SDEs, respectively, and whose densities at time t are given by $\rho(t)$.

Interestingly, the drift coefficients entering these ODEs/SDEs are the unique minimizers of quadratic objective functions that can be estimated empirically using data from ρ_0, ρ_1, and $\mathsf{N}(0, Id)$. The resulting least-squares regression problem allows us to estimate the drift coefficients of the ODE/SDEs, which can then be used to push samples from ρ_0 onto new samples from ρ_1 and vice-versa.

1.2 Main contributions and organization

The approach introduced here is a versatile way to build generative models that unifies and extends many existing algorithms. In Sec. 2, we develop the framework in full generality, where we emphasize the following key contributions:

- We prove that the stochastic interpolant defined in Section 2.1 has a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, and that its density $\rho(t)$ satisfies a first-order transport equation (TE) as well as a family of forward and backward Fokker-Planck equations (FPEs) with tunable diffusion coefficients.

- We show how the stochastic interpolant can be used to learn the drift coefficients that enter the TE and the FPEs. We characterize these coefficients as the minimizers of simple quadratic objective functions given in Section 2.2. We introduce a new objective for the score $\nabla \log \rho(t)$ of the interpolant density, as well as an objective function for learning a denoiser η_z, which we relate to the score.
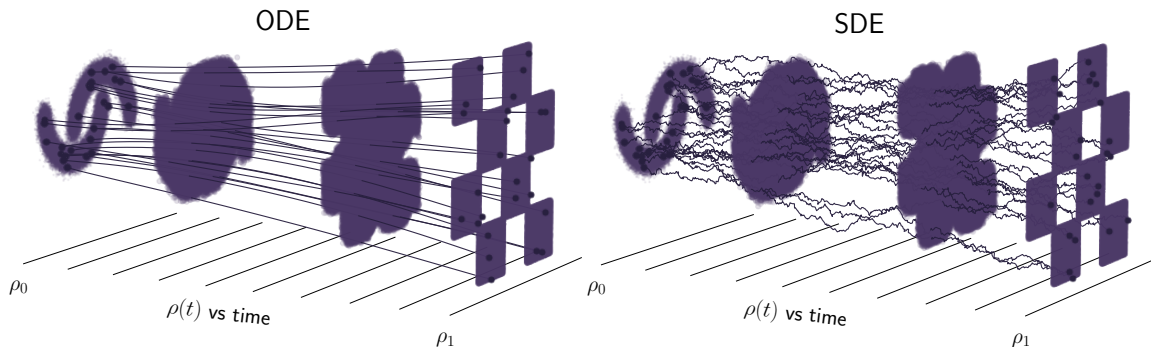
Figure 1: **The stochastic interpolant paradigm.** Example generative models based on the proposed framework, which connects two densities ρ_0 and ρ_1 using samples from both. The design of the time-dependent probability density $\rho(t)$ that bridges between ρ_0 and ρ_1 is separated from the choice of how to sample it, which can be accomplished with deterministic or stochastic generative models. *Left panel:* Sampling with a deterministic (ODE) generative model known as the probability flow equation. *Right panel:* Sampling with a stochastic generative model given by an SDE with a tunable diffusion coefficient. The probability flow equation and the SDE have different paths, but their time-dependent density $\rho(t)$ is the same. Moreover, the two equations rely on the same estimates for the velocity and the score.

- In Section 2.3, we derive ordinary and stochastic differential equations associated with the TE and FPEs that lead to deterministic and stochastic generative models. In Section 2.4, we show that regressing the drift for SDE-based models controls the likelihood, but that regressing the drift alone is not sufficient for ODE-based models, which must also minimize a Fisher divergence. We show how to optimally tune the diffusion coefficient to maximize the likelihood for SDEs.

- In Section 2.5, we develop a general formula to evaluate the likelihood of SDE-based generative models that serves as a natural counterpart to the continuous change-of-variables formula commonly used to compute the likelihood of ODE-based models. In addition, we give formulas to estimate the cross-entropy.

  In Section 3, we discuss instantiations of the stochastic interpolant method. In Section 3.4 we first show that interpolants are equivalent to a class of stochastic bridges, but that they avoid the need for Doob's h-transform, which is generically unknown; we show that this simplifies the construction of a broad class of generative models. In Section 3.2, we define the *one-sided interpolant*, which corresponds to the conventional setting in which the base ρ_0 is taken to be a Gaussian. With a Gaussian base, several aspects of the interpolant simplify, and we detail the corresponding objective functions. In Section 3.3, we introduce a *mirror interpolant* in which the base ρ_0 and the target ρ_1 are identical. Finally, in Section 3.4, we show how the interpolant framework leads to a natural formulation of the Schrödinger bridge problem between two densities.

  In Section 4, we discuss a special case in which the interpolant is spatially linear in x_0 and x_1. In this case, the velocity field can be factorized, which we show in Section 4.1 leads to a simpler learning problem. We detail specific choices of linear interpolants in Section 4.2,

and in Section 4.3 we illustrate how these choices influence the performance of the resulting generative model, with a particular focus on the role of the latent variable and the diffusion coefficient. For exposition, we focus on Gaussian mixture densities, for which the drift coefficients can be computed analytically. We provide the resulting formula in Appendix A. Finally, in Section 4.4, we discuss the case of spatially linear one-sided interpolants.

In Section 5, we formalize the connection between stochastic interpolants and related classes of generative models. In Section 5.1, we show that score-based diffusion models can be re-written as one-sided interpolants after a reparameterization of time; we highlight how this approach eliminates singularities that appear when naively compressing score-based diffusion onto a finite-time interval. In Section 5.2, we show how interpolants can be used to derive the Bayes-optimal estimator for a denoiser, and we show how this approach can be iterated to create a generative model. In Section 5.3, we consider the possibility of rectifying the flow map of a learned generative model. We show that the rectification procedure does not change the underlying generative model, though it may change the time-dependent density of the interpolant.

In Section 6, we provide the details of practical algorithms associated with the mathematical results presented above. In Section 6.1, we describe how to numerically estimate the objectives given empirical datasets from the base and the target. In Section 6.2, we complement this discussion on *learning* with algorithms for *sampling* with the ODE or an SDE.

We provide numerical demonstrations in line with these recommendations in Section 7, and we conclude with some remarks in Section 8.

1.3 Related work

Deterministic Transport and Normalizing Flows. Transport-based sampling and density estimation has its contemporary roots in Gaussianizing data via maximum entropy methods (Friedman, 1987; Chen and Gopinath, 2000; Tabak and Vanden-Eijnden, 2010; Tabak and Turner, 2013). The change of measure under such transformation is the backbone of normalizing flow models. The first neural network realizations of these methods arose through imposing clever structure on the transformation to make the change of measure tractable in discrete, sequential steps (Rezende and Mohamed, 2015; Dinh et al., 2017; Papamakarios et al., 2017; Huang et al., 2018; Durkan et al., 2019). A continuous time version of this procedure was made possible by viewing the map $T = X_t(x)$ as the solution of an ODE (Chen et al., 2018; Grathwohl et al., 2019), whose parametric drift defining the transport is learned via maximum likelihood estimation. Training this way is intractable at scale, as computing the gradient of the objective via the adjoint method requires simulating an ODE. Various methods have introduced regularization on the path taken between the two densities to make the ODE solves more efficient (Finlay et al., 2020; Onken et al., 2021; Tong et al., 2020), but the fundamental difficulty remains. We also work in continuous time; however, our approach allows us to learn the drift without simulation of the dynamics, and can be formulated at sample generation time through either deterministic or stochastic transport.

Stochastic Transport and Score-Based Diffusion Models (SBDMs). Complementary to approaches based on deterministic maps, recent works have realized that connecting

a data distribution to a Gaussian density can be viewed as the evolution of an Ornstein-Ulhenbeck (OU) process which gradually degrades samples from the distribution of interest to Gaussian noise (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song and Ermon, 2019; Song et al., 2021b). The OU process specifies a path in the space of probability densities; this path is simple to traverse in the forward direction by addition of noise, and can be reversed if access to the score of the time-dependent density $\nabla \log \rho(t)$ is available. This score can be approximated through solution of a least-squares regression problem (Hyvärinen, 2005; Vincent, 2011), and the target can be sampled by reversing the path once the score has been learned. Interestingly, the resulting forward and backward stochastic processes have an equivalent formulation (at the distribution level) in terms of a deterministic probability flow equation, first noted by Bakry and Émery (1985); Otto (2000); Kim and Milman (2010) and then applied in Maoutsa et al. (2020); Song et al. (2021a); Kingma et al. (2021); Boffi and Vanden-Eijnden (2023). The probability flow formulation is useful for density estimation and cross-entropy calculations, but it is worth noting that the probability flow and the reverse-time SDE will have densities that differ when using an approximate score. The SBDM framework, as it has been originally presented, has a number of features which are not *a priori* well motivated, including the dependence on mapping to a normal density, the complicated tuning of the time parameterization and noise scheduling (Xiao et al., 2022; Hoogeboom et al., 2023), and the choice of the underlying stochastic dynamics (Dockhorn et al., 2022; Karras et al., 2022).

Stochastic bridges. Starting with (Peluchetti, 2022) there has been some recent effort (Liu et al., 2022, 2023a; Somnath et al., 2023) to remove the dependence of SBDMs on the OU process via stochastic bridges, which can be used to connect two arbitrary densities in finite time. As another step in this direction, we observe here that the key idea behind SBDMs – the bridging of two densities via a time-dependent density whose evolution equation is available – can be generalized to a much wider class of processes in a straightforward and computationally accessible manner. This viewpoint highlights the key property that the construction of the bridge between the two densities is decoupled from the process used to sample it.

Stochastic Interpolants, Rectified Flows, and Flow matching. Variants of the stochastic interpolant method presented in Albergo and Vanden-Eijnden (2023) were also presented in Liu et al. (2023b); Lipman et al. (2023). In Liu et al. (2023b), a linear interpolant was proposed with a focus on straight paths. This was employed as a step toward rectifying the transport paths (Liu, 2022) through a procedure that improves sampling efficiency but introduces a bias. In Lipman et al. (2023), the interpolant picture was assembled from the perspective of conditional probability paths connecting to a Gaussian, where a noise convolution was used to improve the learning at the cost of biasing the method. Extensions of Lipman et al. (2023) were presented in Tong et al. (2023) that generalize the method beyond the Gaussian base density. In the method proposed here, we introduce an unbiased means to incorporate noise into the process, both via the introduction of a latent variable into the stochastic interpolant and the inclusion of a tunable diffusion coefficient in the associated stochastic generative models. We provide theoretical and practical motivation for the presence of these noise terms.

Optimal Transport and Schrödinger Bridges.   There is both theoretical and practical interest in minimizing the transport cost of connecting ρ_0 and ρ_1. In the case of deterministic maps, this is characterized by the optimal transport problem, and in the case of diffusive maps, by the Schrödinger Bridge problem (Villani, 2009; Chen et al., 2021). Formally, these two problems can be related by viewing the Schrödinger Bridge as an entropy-regularized optimal transport. Optimal transport has primarily been employed as a means to regularize flow-based methods by imposing either a path length penalty (Zhang et al., 2018; Onken et al., 2021; Finlay et al., 2020; Tong et al., 2020) or structure on the parameterization itself (Huang et al., 2021; Yang and Karniadakis, 2022). A variety of recent works have formulated the Schrödinger problem in the context of a learnable diffusion (Bortoli et al., 2021; Su et al., 2023; Chen et al., 2022b). For the case of Gaussians, recent work has also identified an analytical solution (Bunne et al., 2022). In the interpolant framework, (Albergo and Vanden-Eijnden, 2023; Liu et al., 2023b; Lipman et al., 2023; Tong et al., 2023) all propose optimal transport extensions to the learning procedure. The method proposed in Liu et al. (2023b); Liu (2022) allows one to sequentially lower the transport cost through rectification, at the cost of introducing a bias unless the velocity field is perfectly learned. The method proposed in Albergo and Vanden-Eijnden (2023) is an unbiased framework at the cost of solving an additional optimization problem over the interpolant function. The statement of optimal transport in Lipman et al. (2023) only applies to Gaussians, but is shown to be practically useful in experimental demonstrations.

In the method proposed below, we provide two approaches for optimizing the transport under a stochastic dynamics. Our primary approach, based on the scheme introduced in Albergo and Vanden-Eijnden (2023), is presented in Section 3.4. It offers an alternative route to solve the Schrödinger bridge problem under the Benamou-Brenier hydrodynamic formulation of transport by maximizing over the interpolant (Benamou and Brenier, 2000). However, we stress that this additional optimization step is not necessary in practice, as our approach leads to bias-free generative models for any fixed interpolant. In addition, Section 5.3 discusses an unbiased variant of the rectification scheme proposed in Liu et al. (2023b).

Convergence bounds.   Inspired by the successes of score-based diffusion, significant recent research effort has been expended to understand the control that can be obtained on suitable distances between the distribution of the generative model and the target data distribution, such as KL, W_2, or TV. Perhaps the first line of work in this direction is Song et al. (2021a), which showed that standard score-based diffusion training techniques bound the likelihood of the resulting SDE model. Importantly, as we show here, the likelihood of the corresponding probability flow is not bounded in general by this technique, as first highlighted in the context of SBDM by Lu et al. (2022). Control for SBDM-based techniques was later quantified more rigorously under the assumption of functional inequalities in a discretized setting by Lee et al. (2022), which were removed by Lee et al. (2023) and Chen et al. (2023) via Girsanov-based techniques. Most relevant to the PDE-based methods considered here is Chen et al. (2022a), which applies similar techniques to our own in the SBDM context to obtain sharp guarantees with minimal assumptions.

1.4 Notation

Throughout, we denote probability density functions as $\rho_0(x)$, $\rho_1(x)$, and $\rho(t,x)$, with $t \in [0,1]$ and $x \in \mathbb{R}^d$, omitting the function arguments when clear from the context. We proceed similarly for other functions of time and space, such as $b(t,x)$ or $I(t,x_0,x_1)$. We use the subscript t to denote the time-dependency of stochastic processes, such as the stochastic interpolant x_t or the Wiener process W_t. To specify that the random variable x_0 is drawn from the probability distribution with density ρ_0, say, with a slight abuse of notations we use $x_0 \sim \rho_0$. Similarly, we use $\mathsf{N}(0, Id)$ to denote both the density and the distribution of the Gaussian random variable with mean zero and covariance identity. We denote expectation by $\mathbb{E}$, and usually specify the random variables this expectation is taken over. With a slight abuse of terminology, we say that the law of the process x_t is $\rho(t)$ if $\rho(t)$ is the density of the probability distribution of x_t at time t.

We use standard notation for function spaces: for example, $C^1([0,1])$ is the space of continuously differentiable functions from $[0,1]$ to $\mathbb{R}$, $(C^2(\mathbb{R}^d))^d$ is the space of twice continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}^d$, and $C_0^p(\mathbb{R}^d)$ is the space of compactly supported functions from $\mathbb{R}^d$ to $\mathbb{R}$ that are continuously differentiable p times. Given a function $b : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ with value $b(t,x)$ at (t,x), we use $b \in C^1([0,1]; (C^2(\mathbb{R}^d))^d)$ to indicate that b is continuously differentiable in t for all $(t,x) \in [0,1] \times \mathbb{R}^d$ and that $b(t,\cdot)$ is an element of $(C^2(\mathbb{R}^d))^d$ for all $t \in [0,1]$.

2. Stochastic interpolant framework

2.1 Definitions and assumptions

We begin by defining the stochastic processes that are central to our approach:

Definition 1 (Stochastic interpolant) *Given two probability density functions $\rho_0, \rho_1 :$ $\mathbb{R}^d \to \mathbb{R}_{\geq 0}$, a stochastic interpolant between ρ_0 and ρ_1 is a stochastic process x_t defined as*

$$x_t = I(t, x_0, x_1) + \gamma(t)z, \qquad t \in [0,1], \tag{2.1}$$

where:

1. *$I \in C^2([0,1], C^2(\mathbb{R}^d \times \mathbb{R}^d)^d)$ satisfies the boundary conditions $I(0, x_0, x_1) = x_0$ and $I(1, x_0, x_1) = x_1$, as well as*

$$\exists C_1 < \infty \;:\; |\partial_t I(t, x_0, x_1)| \leq C_1 |x_0 - x_1| \qquad \forall (t, x_0, x_1) \in [0,1] \times \mathbb{R}^d \times \mathbb{R}^d. \tag{2.2}$$

2. *$\gamma : [0,1] \to \mathbb{R}$ satisfies $\gamma(0) = \gamma(1) = 0$, $\gamma(t) > 0$ for all $t \in (0,1)$, and $\gamma^2 \in C^2([0,1])$.*

3. *The pair (x_0, x_1) is drawn from a probability measure ν that marginalizes on ρ_0 and ρ_1, i.e.*

$$\nu(dx_0, \mathbb{R}^d) = \rho_0(x_0)dx_0, \qquad \nu(\mathbb{R}^d, dx_1) = \rho_1(x_1)dx_1. \tag{2.3}$$

4. *z is a Gaussian random variable independent of (x_0, x_1), i.e. $z \sim \mathsf{N}(0, Id)$ and $z \perp (x_0, x_1)$.*

Eq. (2.2) states that $I(t, x_0, x_1)$ does not move too fast along the way from x_0 at $t = 0$ to x_1 at $t = 1$, and as a result does not wander too far from either endpoint – this assumption is made for convenience but is not necessary for most arguments below. Later, we will find it useful to consider choices for I that are spatially nonlinear, which we show can recover the solution to the Schrödinger bridge problem. Nevertheless, a simple example that serves as a valid I in the sense of Definition 1 is given in (1.1). The measure ν allows for a coupling between the two densities ρ_0 and ρ_1, which affects the properties of the stochastic interpolant, but a simple choice is to take the product measure $\nu(dx_0, dx_1) = \rho_0(x_0)\rho_1(x_1)dx_0dx_1$, in which case x_0 and x_1 are independent. In Section 6 we discuss how to design the stochastic interpolant in (2.1) and state some properties of the corresponding process x_t. Examples of stochastic interpolants are also shown in Figure 2 for various choices of I and γ.

Remark 2 (Comparison with Albergo and Vanden-Eijnden (2023)) *The main difference between the stochastic interpolant defined in* (2.1) *and the one originally introduced in Albergo and Vanden-Eijnden (2023) is the inclusion of the latent variable $\gamma(t)z$. Many of the results below also hold when we set $\gamma(t)z = 0$, but the objective of the present paper is to elucidate the advantages that this additional term provides when neither of the endpoints are Gaussian. We note that we could generalize the construction by making $\gamma(t)$ a tensor; here we focus on the scalar case for simplicity. Another difference is the possibility to couple ρ_0 and ρ_1 via ν. While the latent variable can be drawn from any noise distribution, as we will see, it will be convenient to choose it to be a Gaussian.*

The stochastic interpolant x_t in (2.1) is a continuous-time stochastic process whose realizations are samples from ρ_0 at time $t = 0$ and from ρ_1 at time $t = 1$ by construction. As a result, it offers a way to bridge ρ_0 and ρ_1 – we are interested in characterizing the law of x_t over the full interval $[0, 1]$, as it will allow us to design generative models. Mathematically, we want to characterize the properties of the time-dependent probability distribution $\mu(t, dx)$ such that

$$\forall t \in [0, 1] \quad : \quad \int_{\mathbb{R}^d} \phi(x)\mu(t, dx) = \mathbb{E}[\phi(x_t)] \quad \text{for any test function} \quad \phi \in C_0^\infty(\mathbb{R}^d), \quad (2.4)$$

where x_t is defined in (2.1) and the expectation is taken independently over $(x_0, x_1) \sim \nu$, and $z \sim \mathsf{N}(0, Id)$. To this end, we will need to use conditional expectations over x_t[1], as described in the following definition.

Definition 3 *Given any $f : [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, its conditional expectation $\mathbb{E}\left[f(t, x_0, x_1, z)|x_t = x\right]$ is the function of x such that, for any test function $\phi \in C_0^\infty(\mathbb{R}^d)$, we have*

$$\forall t \in [0, 1] \quad : \quad \int_{\mathbb{R}^d} \phi(x)\mathbb{E}\left[f(t, x_0, x_1, z)|x_t = x\right]\mu(t, dx) = \mathbb{E}[\phi(x_t)f(t, x_0, x_1, z)], \quad (2.5)$$

---

1. Formally, in terms of the Dirac delta distribution, we can write

$$\mathbb{E}\left[f(t, x_0, x_1, z)|x_t = x\right] = \frac{\mathbb{E}\left[f(t, x_0, x_1, z)\delta(x - x_t)\right]}{\mathbb{E}[\delta(x - x_t)]}$$

and in this notation we also have $\mu(t, x) = \mathbb{E}[\delta(x - x_t)]$.
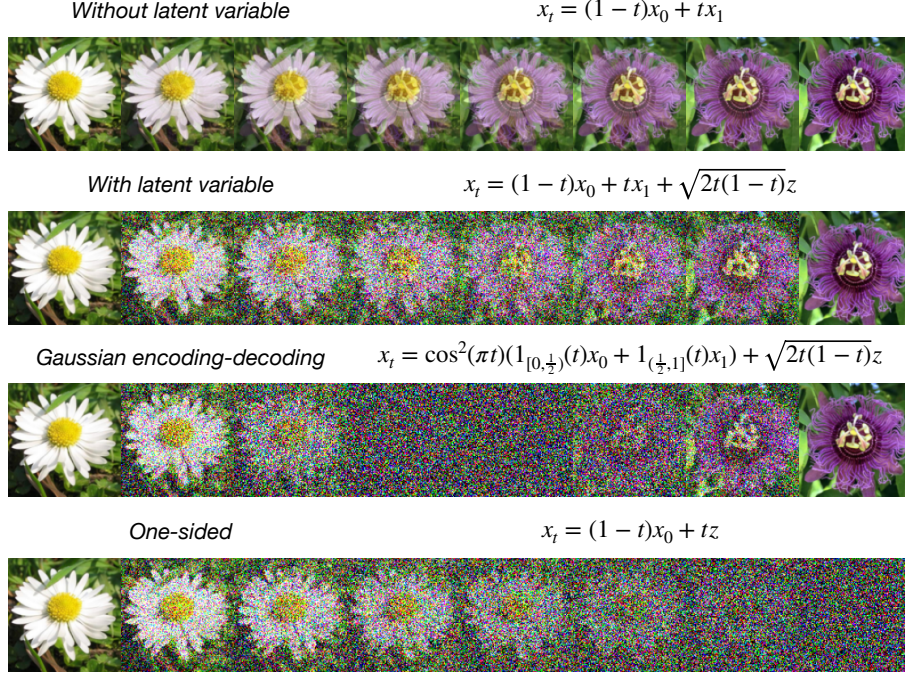
Figure 2: **Design flexibility.** An illustration of how stochastic interpolants can be tailored to specific aims. All examples show one realization of x_t with one $x_0 \sim \rho_0$, one $x_1 \sim \rho_1$ (the flowers at the left and right of the figures), and one $z \sim \mathsf{N}(0, Id)$. *Top, Upper middle, and lower middle*: various interpolants, ranging from direct interpolation with no latent variable (as in Albergo and Vanden-Eijnden (2023)) to Gaussian encoding-decoding in which the data transitions to pure noise at the mid-point. *Bottom*: one-sided interpolant, which connects with score-based diffusion methods.

where $\mu(t, dx)$ is the time-dependent distribution of x_t defined by (3), and the expectation on the right-hand side is taken independently over $(x_0, x_1) \sim \nu$ and $z \sim \mathsf{N}(0, Id)$ with x_t given by (2.1).

Vector-valued functions have conditional expectations that are defined analogously. Note that, with our definition, $\mathbb{E}[f(t, x_0, x_1, z)|x_t = x]$ is a deterministic function of $(t, x) \in [0, 1] \times \mathbb{R}^d$, not to be confused with the random variable $\mathbb{E}[f(t, x_0, x_1, z)|x_t]$ that can be defined analogously.

Remark 4 *Another seemingly more general way to define the stochastic interpolant is via*

$$x_t^{\mathsf{d}} = I(t, x_0, x_1) + N_t \tag{2.6}$$

where $N : [0, 1] \to \mathbb{R}^d$ is a zero-mean Gaussian stochastic process constrained to satisfy $N_{t=0} = N_{t=1} = 0$. As we will show below, our construction only depends on the single-time properties of N_t, which are completely specified by $\mathbb{E}[N_t N_t^{\mathsf{T}}]$. That is, if we take $\gamma(t)$ in (2.1) such that $\mathbb{E}[N_t N_t^{\mathsf{T}}] = \gamma^2(t) Id$, then the probability distribution of x_t will coincide with that of x_t' defined in (2.6), $x_t \stackrel{d}{=} x_t^{\mathsf{d}}$. For example, taking $\gamma(t) = \sqrt{t(1-t)}$ in (2.1) – a choice we
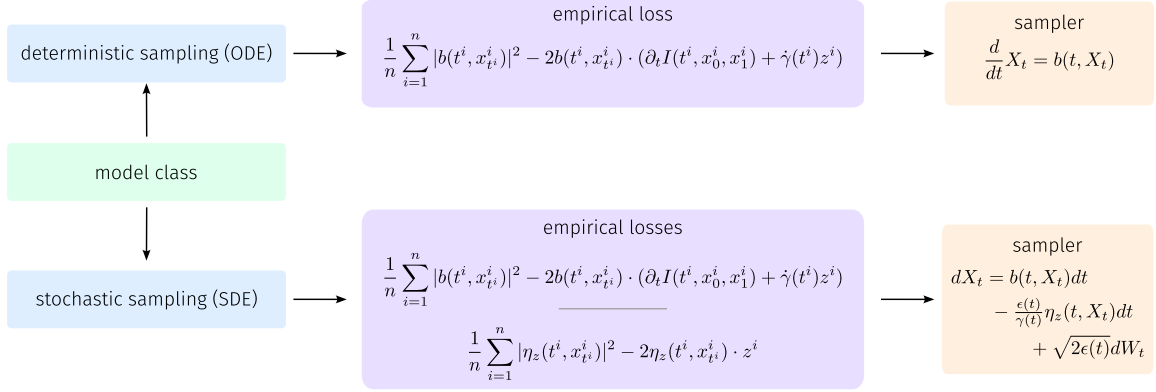
Figure 3: **Algorithmic implementation.** A simple overview of suggested implementation strategies. For deterministic sampling, a single velocity field b can be learned by minimizing the empirical loss in the top row. For stochastic sampling, the velocity field b, along with the denoiser η_z, can be learned by minimizing the two empirical losses specified in the bottom row. To sample deterministically, off-the-shelf ODE integrators can be used to integrate the probability flow equation. To sample stochastically, the listed SDE can be integrated using standard techniques such as the Euler-Maruyama method or the Heun sampler introduced in Karras et al. (2022). The time-dependent diffusion coefficient $\epsilon(t)$ can be specified *after learning* to maximize sample quality.

will consider below in Sec. 3.4 – is equivalent to choosing N_t to be a Brownian bridge in (2.6), i.e. the stochastic process realizable in terms of the Wiener process W_t as $N_t = W_t - tW_1$. This observation will also help us draw an analogy between our approach and the construction used in score-based diffusion models as well as methods based on stochastic bridges. As we will show in Sec. 3.4, it is simpler for both analysis and practical implementation to work with the definition (2.1) for x_t.

To proceed, we will make the following assumption on the densities ρ_0, ρ_1, and the interplay between the measure ν to the function I:

Assumption 5 *The densities ρ_0 and ρ_1 are strictly positive elements of $C^2(\mathbb{R}^d)$ and are such that*

$$\int_{\mathbb{R}^d} |\nabla \log \rho_0(x)|^2 \rho_0(x) dx < \infty \quad and \quad \int_{\mathbb{R}^d} |\nabla \log \rho_1(x)|^2 \rho_1(x) dx < \infty. \qquad (2.7)$$

The measure ν and the function I are such that

$$\exists M_1, M_2 < \infty \quad : \quad \mathbb{E}\big[|\partial_t I(t, x_0, x_1)|^4\big] \leq M_1; \quad \mathbb{E}\big[|\partial_t^2 I(t, x_0, x_1)|^2\big] \leq M_2, \quad \forall t \in [0, 1], \tag{2.8}$$

where the expectation is taken over $(x_0, x_1) \sim \nu$.

Note that for the interpolant (1.1), Assumption 5 holds if ρ_0 and ρ_1 both have finite fourth moments.

2.2 Transport equations, score, and quadratic objectives

We now state a result that specifies some important properties of the probability distribution of the stochastic interpolant x_t:

Theorem 6 (Stochastic interpolant properties) *The probability distribution of the stochastic interpolant x_t defined in (2.1) is absolutely continuous with respect to the Lebesgue measure at all times $t \in [0, 1]$ and its time-dependent density $\rho(t)$ satisfies $\rho(0) = \rho_0$, $\rho(1) = \rho_1$, $\rho \in C^1([0, 1]; C^p(\mathbb{R}^d))$ for any $p \in \mathbb{N}$, and $\rho(t, x) > 0$ for all $(t, x) \in [0, 1] \times \mathbb{R}^d$. In addition, ρ solves the transport equation*

$$\partial_t \rho + \nabla \cdot (b\rho) = 0, \tag{2.9}$$

where we defined the velocity

$$b(t, x) = \mathbb{E}[\dot{x}_t | x_t = x] = \mathbb{E}[\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z | x_t = x]. \tag{2.10}$$

This velocity is in $C^0([0, 1]; (C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$, and such that

$$\forall t \in [0, 1] \quad : \quad \int_{\mathbb{R}^d} |b(t, x)|^2 \rho(t, x) dx < \infty. \tag{2.11}$$

Note that this theorem means that we can write (2.4) as

$$\forall t \in [0, 1] \quad : \quad \int_{\mathbb{R}^d} \phi(x)\rho(t, x) dx = \mathbb{E}\phi(x_t) \quad \text{for any test function} \quad \phi \in C_0^\infty(\mathbb{R}^d). \tag{2.12}$$

The transport equation (2.9) can be solved either forward in time from the initial condition $\rho(0) = \rho_0$, in which case $\rho(1) = \rho_1$, or backward in time from the final condition $\rho(1) = \rho_1$, in which case $\rho(0) = \rho_0$.

The proof of Theorem 6 is given in Appendix B.1; it mostly relies on manipulations involving the characteristic function of the stochastic interpolant x_t. The transport equation (2.9) for ρ lead to methods for generative modeling and density estimation, as explained in Secs. 2.3 and 2.5, provided that we can estimate the velocity b. This velocity is explicitly available only in special cases, for example when ρ_0 and ρ_1 are both Gaussian mixture densities: this case is treated in Appendix A. In general b must be calculated numerically, which can be performed via empirical risk minimization of a quadratic objective function, as characterized by our next result:

Theorem 7 (Objective) *The velocity b defined in (2.10) is the unique minimizer in $C^0([0, 1]; (C^1(\mathbb{R}^d))^d)$ of the quadratic objective*

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{b}(t, x_t)|^2 - (\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z) \cdot \hat{b}(t, x_t)\right) dt \tag{2.13}$$

where x_t is defined in (2.1) and the expectation is taken independently over $(x_0, x_1) \sim \nu$ and $z \sim \mathsf{N}(0, Id)$.

The proof of Theorem 7 is given in Appendix B.1: it relies on the definitions of b in (2.10), as well as the definition of ρ in (2.12) and some elementary properties of the conditional expectation. We discuss how to estimate the objective function (2.13) in practice in Section 6. Interestingly, we also have access to the score of the probability density, as shown by our next result:

13

Theorem 8 (Score) *The score of the probability density ρ specified in Theorem 6 is in $C^1([0,1];(C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$ and given by*

$$s(t,x) = \nabla \log \rho(t,x) = -\gamma^{-1}(t)\mathbb{E}(z|x_t = x) \quad \forall (t,x) \in (0,1) \times \mathbb{R}^d \tag{2.14}$$

In addition it satisfies

$$\forall t \in [0,1] \quad : \quad \int_{\mathbb{R}^d} |s(t,x)|^2 \rho(t,x)dx < \infty, \tag{2.15}$$

and is the unique minimizer in $C^1([0,1];(C^1(\mathbb{R}^d))^d)$ of the quadratic objective

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{s}(t,x_t)|^2 + \gamma^{-1}(t)z \cdot \hat{s}(t,x_t)\right)dt \tag{2.16}$$

where x_t is defined in (2.1) and the expectation is taken independently over $(x_0, x_1) \sim \nu$ and $z \sim \mathsf{N}(0, Id)$

The proof of Theorem 8 is given in Appendix B.1. We stress that the objective function is well defined despite the fact that $\gamma(0) = \gamma(1) = 0$: see Section 6 for more details about how to evaluate this objective in practice.

Remark 9 (Denoiser) *The quantity*

$$\eta_z(t,x) = \mathbb{E}(z|x_t = x), \tag{2.17}$$

will be referred as the denoiser, for reasons that will be made clear in Section 5.2. By (2.14), this quantity gives access to the score on $t \in (0,1)$ (where $\gamma(t) > 0$) since, from (2.14),

$$s(t,x) = -\gamma^{-1}(t)\eta_z(t,x). \tag{2.18}$$

This denoiser is the minimizer of an equivalent expression to (2.16),

$$\mathcal{L}_{\eta_z}[\hat{\eta}_z] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{\eta}_z(t,x_t)|^2 - z \cdot \hat{\eta}_z(t,x_t)\right)dt. \tag{2.19}$$

The denoiser η_z is useful for numerical realizations. In particular, the objective in (2.19) is easier to use than the one in (2.16) because it does not contain the factor $\gamma^{-1}(t)$, which needs careful handling as t approaches 0 and 1.

Having access to the score immediately allows us to rewrite the TE (2.9) as forward and backward Fokker-Planck equations, which we state as:

Corollary 10 (Fokker-Planck equations) *For any $\epsilon \in C^0([0,1])$ with $\epsilon(t) \geq 0$ for all $t \in [0,1]$, the probability density ρ specified in Theorem 6 satisfies:*

1. The forward Fokker-Planck equation

$$\partial_t \rho + \nabla \cdot (b_\mathsf{F}\rho) = \epsilon(t)\Delta\rho, \qquad \rho(0) = \rho_0, \tag{2.20}$$

where we defined the forward drift

$$b_\mathsf{F}(t,x) = b(t,x) + \epsilon(t)s(t,x). \tag{2.21}$$

Equation (2.20) is well-posed when solved forward in time from $t = 0$ to $t = 1$, and its solution for the initial condition $\rho(t = 0) = \rho_0$ satisfies $\rho(t = 1) = \rho_1$.

2. *The backward Fokker-Planck equation*

$$\partial_t \rho + \nabla \cdot (b_{\mathsf{B}}\rho) = -\epsilon(t)\Delta\rho, \qquad \rho(1) = \rho_1, \tag{2.22}$$

where we defined the backward drift

$$b_{\mathsf{B}}(t, x) = b(t, x) - \epsilon(t)s(t, x). \tag{2.23}$$

Equation (2.22) is well-posed when solved backward in time from $t = 1$ to $t = 0$, and its solution for the final condition $\rho(1) = \rho_1$ satisfies $\rho(0) = \rho_0$.

In Section 2.3 we will use the results of this theorem to design generative models based on forward and backward stochastic differential equations. Note that we can replace the diffusion coefficient $\epsilon(t)$ by a positive semi-definite tensor; also note that if we define $\rho_{\mathsf{B}}(t_{\mathsf{B}}, x) = \rho(1 - t_{\mathsf{B}}, x)$, the reversed FPE (2.22) can be written as

$$\partial_{t_{\mathsf{B}}}\rho_{\mathsf{B}}(t_{\mathsf{B}}, x) - \nabla \cdot (b_{\mathsf{B}}(1 - t_{\mathsf{B}}, x)\rho_{\mathsf{B}}(t_{\mathsf{B}}, x)) = \epsilon(1 - t_{\mathsf{B}})\Delta\rho_{\mathsf{B}}(t_{\mathsf{B}}, x), \qquad \rho_{\mathsf{B}}(t_{\mathsf{B}} = 0) = \rho_1, \tag{2.24}$$

which is now well-posed forward in (reversed) time t_{B}. So as to have only one definition of time t, it is more convenient to work with (2.22).

Let us make a few remarks about the statements made so far:

Remark 11 *If we set $\gamma(t) = 0$ in x_t (i.e, if we remove the latent variable), the stochastic interpolant (2.1) reduces to the one originally considered in Albergo and Vanden-Eijnden (2023). In this setup, the results above formally stand except that we cannot guarantee the spatial regularity of $b(t, x)$ and $s(t, x)$, since it relies on the presence of the latent variable (as shown in the proof of Theorem 6). Hence, we expect the introduction of the latent variable $\gamma(t)z$ to help for generative modeling, where the solution to the corresponding ODEs/SDEs will be better behaved, and for statistical approximation, since the targets b and s will be more regular. We will see in Section 6 that it also gives us much greater flexibility in the way we can bridge ρ_0 and ρ_1, which will enable us to design generative models with appealing properties.*

Remark 12 *We will see in Section 2.4 that the forward and backward FPE in (2.20) and (2.22) are more robust than the TE in (2.9) against approximation errors in the velocity b and the score s, which has practical implications for generative models based on these equations.*

Remark 13 *We could also obtain $b(t, \cdot)$ at any $t \in [0, 1]$ by minimizing*

$$\mathbb{E}\left(\tfrac{1}{2}|\hat{b}(t, x_t)|^2 - (\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z) \cdot \hat{b}(t, x_t)\right) \qquad t \in [0, 1] \tag{2.25}$$

and $s(t, \cdot)$ at any $t \in (0, 1)$ by minimizing

$$\mathbb{E}\left(\tfrac{1}{2}|\hat{s}(t, x_t)|^2 + \gamma^{-1}(t)z \cdot \hat{s}(t, x_t)\right) \qquad t \in (0, 1) \tag{2.26}$$

Using the time-integrated versions of these objectives given in (2.13) and (2.16) is more convenient numerically as it allows one to parameterize $\hat{b}$ and $\hat{s}$ globally for $(t, x) \in [0, 1] \times \mathbb{R}^d$.

Remark 14 *From (2.10) we can write*

$$b(t, x) = v(t, x) - \dot{\gamma}(t)\gamma(t)s(t, x), \tag{2.27}$$

where s is the score given in (2.14) and we defined the velocity field

$$v(t, x) = \mathbb{E}(\partial_t I(t, x_0, x_1)|x_t = x). \tag{2.28}$$

The velocity field $v \in C^0([0, 1]; (C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$ and can be characterized as the unique minimizer of

$$\mathcal{L}_v[\hat{v}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{v}(t, x_t)|^2 - \partial_t I(t, x_0, x_1) \cdot \hat{v}(t, x_t)\right) dt \tag{2.29}$$

Learning this velocity and the score separately may be useful in practice.

Remark 15 *The objectives in (2.13) and (2.16) (as well as the ones in (2.19) and (2.29)) are amenable to empirical estimation if we have samples $(x_0, x_1) \sim \nu$, since in that case we can generate samples of $x_t = I(t, x_0, x_1) + \gamma(t)z$ at any time $t \in [0, 1]$. We will use this feature in the numerical experiments presented below.*

Remark 16 *Since s is the score of ρ, an alternative objective to estimate it is (Hyvärinen, 2005)*

$$\int_0^1 \mathbb{E}\left(|\hat{s}(t, x_t)|^2 + 2\nabla \cdot \hat{s}(t, x_t)\right) dt. \tag{2.30}$$

The derivation of (2.30) is standard: for the reader's convenience we recall it at the end of Appendix B.1. The advantage of using (2.16) over (2.30) is that it does not require us to take the divergence of $\hat{s}$.

Remark 17 (Energy-based models) *By definition, the score $s(t, x) = \nabla \log \rho(t, x)$ is a gradient field. As a result, if we model $\hat{s}(t, x) = -\nabla \hat{E}(t, x)$, we can turn (2.16) into an objective function for $\hat{E}(t, x)$*

$$\mathcal{L}_E[\hat{E}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\nabla \hat{E}(t, x_t)|^2 + \gamma^{-1}(t)z \cdot \nabla \hat{E}(t, x_t)\right) dt \tag{2.31}$$

This objective is invariant to constant shifts in $\hat{E}$ and should therefore be minimized under some constraint, such as $\min_x \hat{E}(t, x) = 0$ for all $t \in [0, 1]$. The minimizer of (2.31) provides us with an energy-based model (EBM) (LeCun et al., 2007; Song and Kingma, 2021) that can in principle be used to sample the PDF of the stochastic interpolant, $\rho(t, x)$, at any fixed $t \in [0, 1]$ using e.g. Langevin dynamics. We will not exploit this possibility here, and instead rely on generative models to sample $\rho(t, x)$, as discussed next in Sec. 2.3.

2.3 Generative models

Our next result is a direct consequence of Theorem 6, and it shows how to design generative models using the stochastic processes associated with the TE (2.9), the forward FPE (2.20), and the backward FPE (2.22):

Corollary 18 (Generative models) *At any time $t \in [0, 1]$, the law of the stochastic interpolant x_t coincides with the law of the three processes X_t, X_t^{F}, and X_t^{B}, respectively defined as:*

1. *The solutions of the probability flow associated with the transport equation* (2.9)

$$\frac{d}{dt}X_t = b(t, X_t), \tag{2.32}$$

   solved either forward in time from the initial data $X_{t=0} \sim \rho_0$ or backward in time from the final data $X_{t=1} = x_1 \sim \rho_1$.

2. *The solutions of the forward SDE associated with the FPE* (2.20)

$$dX_t^{\mathsf{F}} = b_{\mathsf{F}}(t, X_t^{\mathsf{F}})dt + \sqrt{2\epsilon(t)}\, dW_t, \tag{2.33}$$

   solved forward in time from the initial data $X_{t=0}^{\mathsf{F}} \sim \rho_0$ independent of W.

3. *The solutions of the backward SDE associated with the backward FPE* (2.22)

$$dX_t^{\mathsf{B}} = b_{\mathsf{B}}(t, X_t^{\mathsf{B}})dt + \sqrt{2\epsilon(t)}\, dW_t^{\mathsf{B}}, \quad W_t^{\mathsf{B}} = -W_{1-t}, \tag{2.34}$$

   solved backward in time from the final data $X_{t=1}^{\mathsf{B}} \sim \rho_1$ independent of W^{B}; the solution of (2.34) *is by definition $X_t^{\mathsf{B}} = Z_{1-t}^{\mathsf{F}}$ where Z_t^{F} satisfies*

$$dZ_t^{\mathsf{F}} = -b_{\mathsf{B}}(1 - t, Z_t^{\mathsf{F}})dt + \sqrt{2\epsilon(t)}\, dW_t, \tag{2.35}$$

   solved forward in time from the initial data $Z_{t=0}^{\mathsf{F}} \sim \rho_1$ independent of W.

To avoid repeated applications of the transformation $t \mapsto 1 - t$, it is convenient to work with (2.34) directly using the reversed Itô calculus rules stated in the following lemma, which follows from the results in Anderson (1982) and is proven in Appendix B.2:

Lemma 19 (Reverse Itô Calculus) *If X_t^{B} solves the backward SDE* (2.34):

1. *For any $f \in C^1([0, 1]; C_0^2(\mathbb{R}_d))$ and $t \in [0, 1]$, the backward Itô formula holds*

$$df(t, X_t^{\mathsf{B}}) = \partial_t f(t, X_t^{\mathsf{B}})dt + \nabla f(X_t^{\mathsf{B}}) \cdot dX_t^{\mathsf{B}} - \epsilon(t)\Delta f(t, X_t^{\mathsf{B}})dt. \tag{2.36}$$

2. *For any $g \in C^0([0, 1]; (C_0(\mathbb{R}_d))^d)$ and $t \in [0, 1]$, the backward Itô isometries hold:*

$$\mathbb{E}_{\mathsf{B}}^x \int_t^1 g(t, X_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}} = 0; \qquad \mathbb{E}_{\mathsf{B}}^x \left| \int_t^1 g(t, X_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}} \right|^2 = \int_t^1 \mathbb{E}_{\mathsf{B}}^x \left| g(t, X_t^{\mathsf{B}}) \right|^2 dt, \tag{2.37}$$

   where $\mathbb{E}_{\mathsf{B}}^x$ denotes expectation conditioned on the event $X_{t=1}^{\mathsf{B}} = x$.

The relevance of Corollary 18 for generative modeling is clear. Assuming, for example, that ρ_0 is a simple density that can be sampled easily (e.g. a Gaussian or a Gaussian mixture density), we can use the ODE (2.32) or the SDE (2.33) to push these samples forward in time and generate samples from a complex target density ρ_1. In Section 2.5, we will show how to use the ODE (2.32) or the reverse SDE (2.34) to estimate ρ_1 at any $x \in \mathbb{R}^d$ assuming that we can evaluate ρ_0 at any $x \in \mathbb{R}^d$. We will also show how similar ideas can be used to estimate the cross entropy between ρ_0 and ρ_1.

Remark 20 *We stress that the stochastic interpolant x_t, the solution X_t to the ODE (2.32), and the solutions X_t^{F} and X_t^{B} of the forward and backward SDEs (2.33) and (2.34) are different stochastic processes, but their laws all coincide with $\rho(t)$ at any time $t \in [0, 1]$. This is all that matters when applying these processes as generative models. However, the fact that these processes are different has implications for the accuracy of the numerical integration used to sample from them at any t as well as for the propagation of statistical errors (see also the next remark).*

Generative models based on solutions X_t to the ODE (2.32), solutions X_t^{F} to the forward SDE (2.33), and solutions X_t^{B} to the backward SDE (2.34) will typically involve drifts b, b_{F}, and b_{B} that are, in practice, imperfectly estimated via minimization of (2.13) and (2.16) over finite datasets. It is important to estimate how this statistical estimation error propagates to errors in sample quality, and how the propagation of error depends on the generative model used, which is the object of our next section.

2.4 Likelihood control

In this section, we demonstrate that jointly minimizing the objective functions (2.29) and (2.16) (or the losses (2.13) and (2.16)) controls the KL-divergence from the target density ρ_1 to the model density $\hat{\rho}_1$. We focus on bounds involving the score, but we note that analogous results hold for learning the denoiser $\eta_z(t, x)$ defined in (2.17) by the relation $\eta_z(t, x) = -s(t, x)/\gamma(t)$. The derivation is based on a simple and exact characterization of the KL-divergence between two transport equations or two Fokker-Planck equations with different drifts. Remarkably, we find that the presence of a diffusive term determines whether or not it is sufficient to learn the drift to control KL. This can be seen as a generalization of the result for score-based diffusion models described in Song et al. (2021a) to arbitrary generative models described by ODEs or SDEs. The proofs of the statements in this section are provided in Appendix B.3. These proofs rely on manipulation of the time derivative of the KL divergence, which is a practice that has proven useful elsewhere in the literature (Vempala and Wibisono, 2019).

We first characterize the KL divergence between two densities transported by two different continuity equations but initialized from the same initial condition:

Lemma 21 *Let $\rho_0 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote a fixed base probability density function. Given two velocity fields $b, \hat{b} \in C^0([0, 1], (C^1(\mathbb{R}^d))^d)$, let the time-dependent densities $\rho : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and $\hat{\rho} : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote the solutions to the transport equations*

$$
\begin{aligned}
\partial_t \rho + \nabla \cdot (b\rho) = 0, \qquad \rho(0) = \rho_0, \\
\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\hat{\rho}) = 0, \qquad \hat{\rho}(0) = \rho_0.
\end{aligned}
\tag{2.38}
$$

Then, the Kullback-Leibler divergence of $\rho(1)$ from $\hat{\rho}(1)$ is given by

$$
\mathsf{KL}(\rho(1) \,\|\, \hat{\rho}(1)) = \int_0^1 \int_{\mathbb{R}^d} (\nabla \log \hat{\rho}(t, x) - \nabla \log \rho(t, x)) \cdot (\hat{b}(t, x) - b(t, x))\rho(t, x)dxdt. \tag{2.39}
$$

Lemma 21 shows that it is insufficient in general to match $\hat{b}$ with b to obtain control on the KL divergence. The essence of the problem is that a small error in $\hat{b} - b$ does not ensure

control on the Fisher divergence $\mathsf{FI}(\rho(t) \parallel \hat{\rho}(t)) = \int_{\mathbb{R}^d} |\nabla \log \rho(t, x) - \nabla \log \hat{\rho}(t, x)|^2 \rho(t, x) dx$, which is necessary due to the presence of $(\nabla \log \hat{\rho} - \nabla \log \rho)$ in (2.39).

In the next lemma, we study the case for two Fokker-Planck equations, and highlight that the situation becomes quite different.

Lemma 22 *Let $\rho_0 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote a fixed base probability density function. Given two velocity fields $b_\mathsf{F}, \hat{b}_\mathsf{F} \in C^0([0, 1], (C^1(\mathbb{R}^d))^d)$, let the time-dependent densities $\rho : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and $\hat{\rho} : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote the solutions to the Fokker-Planck equations*

$$
\begin{aligned}
\partial_t \rho + \nabla \cdot (b_\mathsf{F} \rho) &= \epsilon \Delta \rho, & \rho(0) &= \rho_0, \\
\partial_t \hat{\rho} + \nabla \cdot (\hat{b}_\mathsf{F} \hat{\rho}) &= \epsilon \Delta \hat{\rho}, & \hat{\rho}(0) &= \rho_0.
\end{aligned}
\tag{2.40}
$$

where $\epsilon > 0$. Then, the Kullback-Leibler divergence from $\rho(1)$ to $\hat{\rho}(1)$ is given by

$$
\mathsf{KL}(\rho(1) \parallel \hat{\rho}(1)) = \int_0^1 \int_{\mathbb{R}^d} (\nabla \log \hat{\rho}(t, x) - \nabla \log \rho(t, x)) \cdot \left(\hat{b}_\mathsf{F}(t, x) - b_\mathsf{F}(t, x) \right) \rho(t, x) dx dt
$$
$$
- \epsilon \int_0^1 \int_{\mathbb{R}^d} |\nabla \log \rho(t, x) - \nabla \log \hat{\rho}(t, x)|^2 \rho(t, x) dx dt,
\tag{2.41}
$$

and as a result

$$
\mathsf{KL}(\rho(1) \parallel \hat{\rho}(1)) \leq \frac{1}{4\epsilon} \int_0^1 \int_{\mathbb{R}^d} \left| \hat{b}_\mathsf{F}(t, x) - b_\mathsf{F}(t, x) \right|^2 \rho(t, x) dx dt.
\tag{2.42}
$$

Lemma 22 shows that, unlike for transport equations, the KL-divergence between the solutions of two Fokker-Planck equations is controlled by the error in their drifts. The diffusive term in each Fokker-Planck equation provides an additional negative term in the KL-divergence, which eliminates the need for explicit control on the Fisher divergence.

Putting the above results together, we can state the following result, which demonstrates that the losses (2.13) and (2.16) control the likelihood for learned approximations to the FPE (2.20).

Theorem 23 *Let ρ denote the solution of the Fokker-Planck equation (2.20) with $\epsilon(t) = \epsilon > 0$. Given two velocity fields $\hat{b}, \hat{s} \in C^0([0, 1], (C^1(\mathbb{R}^d))^d)$, define*

$$
\hat{b}_\mathsf{F}(t, x) = \hat{b}(t, x) + \epsilon \hat{s}(t, x), \qquad \hat{v}(t, x) = \hat{b}(t, x) + \gamma(t) \dot{\gamma}(t) \hat{s}(t, x)
\tag{2.43}
$$

where the function γ satisfies the properties listed in Definition 1. Let $\hat{\rho}$ denote the solution to the Fokker-Planck equation

$$
\partial_t \hat{\rho} + \nabla \cdot (\hat{b}_\mathsf{F} \hat{\rho}) = \epsilon \Delta \hat{\rho}, \qquad \hat{\rho}(0) = \rho_0.
\tag{2.44}
$$

Then,

$$
\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon} \left(\mathcal{L}_b[\hat{b}] - \min_{\hat{b}} \mathcal{L}_b[\hat{b}] \right) + \frac{\epsilon}{2} \left(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}] \right),
\tag{2.45}
$$

where $\mathcal{L}_b[\hat{b}]$ and $\mathcal{L}_s[\hat{s}]$ are the objective functions defined in (2.13) and (2.16), and

$$
\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon} \left(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}] \right) + \frac{\sup_{t \in [0,1]} (\gamma(t)\dot{\gamma}(t) - \epsilon)^2}{2\epsilon} \left(\mathcal{L}_s[\hat{s}] - \min_{\hat{v}} \mathcal{L}_s[\hat{s}] \right).
\tag{2.46}
$$

where $\mathcal{L}_v[\hat{v}]$ is the objective function defined in (2.29).

Remark 24 (Generative modeling) *The above results have practical ramifications for generative modeling. In particular, they show that minimizing either the losses (2.13) and (2.16) or (2.29) and (2.16) maximize the likelihood of the stochastic generative model*

$$d\hat{X}_t^{\mathsf{F}} = \left(\hat{b}(t, \hat{X}_t^{\mathsf{F}}) + \epsilon \hat{s}(t, \hat{X}_t^{\mathsf{F}}) \right) dt + \sqrt{2\epsilon}dW_t, \tag{2.47}$$

but that minimizing the objective (2.13) is insufficient in general to maximize the likelihood of the deterministic generative model

$$\dot{\hat{X}}_t = \hat{b}(t, \hat{X}_t). \tag{2.48}$$

Moreover, they show that, when learning $\hat{b}$ and $\hat{s}$, the choice of ϵ that minimizes the upper bound is given by

$$\epsilon^* = \left(\frac{\mathcal{L}_b[\hat{b}] - \min_{\hat{b}} \mathcal{L}_b[\hat{b}]}{\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}]} \right)^{1/2}, \tag{2.49}$$

so that $\epsilon^* > 1$ if the score is learned to higher accuracy than $\hat{b}$ and $\epsilon^* < 1$ in the opposite situation. Note that (2.49) suggests to take $\epsilon = 0$ if $\hat{b}$ is learned perfectly but $\hat{s}$ is not, and send $\epsilon \to \infty$ in the opposite situation. While taking $\epsilon = 0$ is achievable in practice and leads to the ODE (2.32), taking $\epsilon \to \infty$ is not, as increasing ϵ increases the expense of the numerical integration in (2.33) and (2.34).

2.5 Density estimation and cross-entropy calculation

It is well-known that the solution of the TE (2.9) can be expressed in terms of the solution to the probability flow ODE (2.32); for completeness, we now recall this fact:

Lemma 25 *Given the velocity field $\hat{b} \in C^0([0,1], (C^1(\mathbb{R}^d))^d)$, let $\hat{\rho}$ satisfy the transport equation*

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\hat{\rho}) = 0, \tag{2.50}$$

and let $X_{s,t}(x)$ solve the ODE

$$\frac{d}{dt} X_{s,t}(x) = b(t, X_{s,t}(x)), \qquad X_{s,s}(x) = x, \qquad t, s \in [0,1] \tag{2.51}$$

Then, given the PDFs ρ_0 and ρ_1:

1. The solution to (2.50) for the initial condition $\hat{\rho}(0) = \rho_0$ is given at any time $t \in [0,1]$ by

$$\hat{\rho}(t,x) = \exp\left(-\int_0^t \nabla \cdot b(\tau, X_{t,\tau}(x))d\tau \right) \rho_0(X_{t,0}(x)) \tag{2.52}$$

2. The solution to (2.50) for the final condition $\hat{\rho}(1) = \rho_1$ is given at any time $t \in [0,1]$ by

$$\hat{\rho}(t,x) = \exp\left(\int_t^1 \nabla \cdot b(\tau, X_{t,\tau}(x))d\tau \right) \rho_1(X_{t,1}(x)) \tag{2.53}$$

The proof of Lemma 25 can be found in Appendix B.4. Interestingly, we can obtain a similar result for the solution of the forward and backward FPEs in (2.20) and (2.22). These results make use of auxiliary forward and backward SDEs in which the roles of the forward and backward drifts are switched:

Theorem 26 *Given $\epsilon > 0$ and two velocity fields $\hat{b}, \hat{s} \in C^0([0,1], (C^1(\mathbb{R}^d))^d)$, define*

$$\hat{b}_{\mathsf{F}}(t, x) = \hat{b}(t, x) + \epsilon\hat{s}(t, x), \qquad \hat{b}_{\mathsf{B}}(t, x) = \hat{b}(t, x) - \epsilon\hat{s}(t, x), \tag{2.54}$$

and let Y_t^{F} and Y_t^{B} denote solutions of the following forward and backward SDEs:

$$dY_t^{\mathsf{F}} = b_{\mathsf{B}}(t, Y_t^{\mathsf{F}})dt + \sqrt{2\epsilon}dW_t, \tag{2.55}$$

to be solved forward in time from the initial condition $Y_{t=0}^{\mathsf{F}} = x$ independent of W; and

$$dY_t^{\mathsf{B}} = b_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt + \sqrt{2\epsilon}dW_t^{\mathsf{B}}, \quad W_t^{\mathsf{B}} = -W_{1-t}, \tag{2.56}$$

to be solved backwards in time from the final condition $Y_{t=1}^{\mathsf{B}} = x$ independent of W^{B}. Then, given the densities ρ_0 and ρ_1:

1. *The solution to the forward FPE*

$$\partial_t \hat{\rho}_{\mathsf{F}} + \nabla \cdot (\hat{b}_{\mathsf{F}} \hat{\rho}_{\mathsf{F}}) = \epsilon \Delta \hat{\rho}_{\mathsf{F}}, \qquad \hat{\rho}_{\mathsf{F}}(0) = \rho_0, \tag{2.57}$$

   can be expressed at $t = 1$ as

$$\hat{\rho}_{\mathsf{F}}(1, x) = \mathbb{E}_{\mathsf{B}}^x \left(\exp\left(-\int_0^1 \nabla \cdot \hat{b}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt \right) \rho_0(Y_{t=0}^{\mathsf{B}}) \right), \tag{2.58}$$

   where $\mathbb{E}_{\mathsf{B}}^x$ denotes expectation on the path of Y_t^{B} conditional on the event $Y_{t=1}^{\mathsf{B}} = x$.

2. *The solution to the backward FPE*

$$\partial_t \hat{\rho}_{\mathsf{B}} + \nabla \cdot (\hat{b}_{\mathsf{B}} \hat{\rho}_{\mathsf{B}}) = -\epsilon \Delta \hat{\rho}_{\mathsf{B}}, \qquad \hat{\rho}_{\mathsf{B}}(1) = \rho_1, \tag{2.59}$$

   can be expressed at any $t = 0$ as

$$\hat{\rho}_{\mathsf{B}}(0, x) = \mathbb{E}_{\mathsf{F}}^x \left(\exp\left(\int_0^1 \nabla \cdot \hat{b}_{\mathsf{B}}(t, Y_t^{\mathsf{F}})dt \right) \rho_1(Y_{t=1}^{\mathsf{F}}) \right), \tag{2.60}$$

   where $\mathbb{E}_{\mathsf{F}}^x$ denotes expectation on the path of Y_t^{F} conditional on $Y_{t=0}^{\mathsf{F}} = x$.

The proof of Theorem 26 can be found in Appendix B.4. Note that to generate data from either $\hat{\rho}_{\mathsf{F}}(1)$ or $\hat{\rho}_{\mathsf{B}}(0)$ assuming that we can sample exactly the PDF at the other end, i.e. ρ_0 and ρ_1 respectively, we would still rely on the equivalent of the forward and backward SDE in (2.33) and (2.34), now used with the approximate drifts in (2.54), i.e.

$$d\hat{X}_t^{\mathsf{F}} = \hat{b}_{\mathsf{F}}(t, \hat{X}_t^{\mathsf{F}})dt + \sqrt{2\epsilon}dW_t, \tag{2.61}$$

and

$$d\hat{X}_t^{\mathsf{B}} = \hat{b}_{\mathsf{B}}(t, \hat{X}_t^{\mathsf{B}})dt + \sqrt{2\epsilon}dW_t^{\mathsf{B}}, \quad W_t^{\mathsf{B}} = -W_{1-t}, \tag{2.62}$$

If we solve (2.61) forward in time from initial data $\hat{X}^{\mathsf{F}}_{t=0} \sim \rho_0$, we then have $\hat{X}^{\mathsf{F}}_{t=1} \sim \hat{\rho}_{\mathsf{F}}(1)$ where $\hat{\rho}_{\mathsf{F}}$ is the solution to the forward FPE (2.57). Similarly If we solve (2.62) backward in time from final data $\hat{X}^{\mathsf{B}}_{t=1} \sim \rho_1$, we then have $\hat{X}^{\mathsf{B}}_{t=0} \sim \hat{\rho}_{\mathsf{B}}(0)$ where $\hat{\rho}_{\mathsf{B}}$ is the solution to the backward FPE (2.59).

The results of Lemma 25 and Theorem 26 can be used to test the quality of samples generated by either the ODE (2.32) or the forward and backward SDEs (2.33) and (2.34). In particular, the following two results are direct consequences of Lemma 25 and Theorem 26, respectively:

Corollary 27 *Under the same conditions as Lemma 25, if $\hat{\rho}(0) = \rho_0$, the cross-entropy of $\hat{\rho}(1)$ relative to ρ_1 is given by*

$$
\begin{aligned}
\mathsf{H}(\rho_1 \parallel \hat{\rho}(1)) &= -\int_{\mathbb{R}^d} \log \hat{\rho}(1, x) \rho_1(x) dx \\
&= \mathbb{E}_1 \int_0^1 \nabla \cdot b(\tau, X_{1,\tau}(x_1)) d\tau - \mathbb{E}_1 \log \rho_0(X_{1,0}(x_1))
\end{aligned}
\tag{2.63}
$$

where $\mathbb{E}_1$ denotes an expectation over $x_1 \sim \rho_1$. Similarly, if $\hat{\rho}(1) = \rho_1$, the cross-entropy of $\hat{\rho}(0)$ relative to ρ_0 is given by

$$
\begin{aligned}
\mathsf{H}(\rho_0 \parallel \hat{\rho}(0)) &= -\int_{\mathbb{R}^d} \log \hat{\rho}(0, x) \rho_0(x) dx \\
&= -\mathbb{E}_0 \int_0^1 \nabla \cdot b(\tau, X_{0,\tau}(x_0)) d\tau - \mathbb{E}_0 \log \rho_1(X_{0,1}(x_0))
\end{aligned}
\tag{2.64}
$$

where $\mathbb{E}_0$ denotes an expectation over $x_0 \sim \rho_0$.

Corollary 28 *Under the same conditions as Theorem 26, the cross-entropy of $\hat{\rho}_{\mathsf{F}}(1)$ relative to ρ_1 is given by*

$$
\begin{aligned}
\mathsf{H}(\rho_1 \parallel \hat{\rho}_{\mathsf{F}}(1)) &= -\int_{\mathbb{R}^d} \log \hat{\rho}_{\mathsf{F}}(1, x) \rho_1(x) dx \\
&= -\mathbb{E}_1 \log \mathbb{E}^{x_1}_{\mathsf{B}} \left(\exp\left(-\int_0^1 \nabla \cdot b_{\mathsf{F}}(t, Y^{\mathsf{B}}_t) dt \right) \rho_0(Y^{\mathsf{B}}_{t=0}) \right),
\end{aligned}
\tag{2.65}
$$

where $\mathbb{E}^{x_1}_{\mathsf{B}}$ denotes an expectation over Y^{B}_t conditioned on the event $Y^{\mathsf{B}}_{t=1} = x_1$, and $\mathbb{E}_1$ denotes an expectation over $x_1 \sim \rho_1$. Similarly, the cross-entropy of $\hat{\rho}_{\mathsf{B}}(0)$ relative to ρ_0 is given by

$$
\begin{aligned}
\mathsf{H}(\rho_0 \parallel \hat{\rho}_{\mathsf{B}}(0)) &= -\int_{\mathbb{R}^d} \log \hat{\rho}_{\mathsf{B}}(0, x) \rho_0(x) dx \\
&= -\mathbb{E}_0 \log \mathbb{E}^{x_0}_{\mathsf{F}} \left(\exp\left(\int_0^1 \nabla \cdot b_{\mathsf{B}}(t, Y^{\mathsf{F}}_t) dt \right) \rho_1(Y^{\mathsf{F}}_{t=1}) \right),
\end{aligned}
\tag{2.66}
$$

where $\mathbb{E}^{x_0}_{\mathsf{B}}$ denotes an expectation over Y^{F}_t conditioned on the event $Y^{\mathsf{F}}_{t=0} = x_0$, and $\mathbb{E}_0$ denotes an expectation over $x_0 \sim \rho_0$.

22

If in (2.63), (2.64), (2.65), and (2.66) we approximate the expectations $\mathbb{E}_0$ and $\mathbb{E}_1$ over ρ_0 and ρ_1 by empirical expectations over the available data, these equations allow us to cross-validate different approximations of $\hat{b}$ and $\hat{s}$, as well as to compare the cross-entropies of densities evolved by the TE (2.50) with those of the forward and backward FPEs (2.57) and (2.59).

Remark 29 *When using* (2.65) *and* (2.66) *in practice, taking the* log *of the expectations* $\mathbb{E}_{\mathsf{B}}^{x_1}$ *and* $\mathbb{E}_{\mathsf{F}}^{x_0}$ *may create difficulties, such as when using Hutchinson's trace estimator to compute the divergence of* b_{F} *or* b_{B}*, which will introduce a bias. One way to remove this bias is to use Jensen's inequality, which leads to the upper bounds*

$$\mathsf{H}(\rho_1 \parallel \hat{\rho}_{\mathsf{F}}(1)) \leq \int_0^1 \mathbb{E}_1 \mathbb{E}_{\mathsf{B}}^{x_1} \nabla \cdot b_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) dt - \mathbb{E}_1 \mathbb{E}_{\mathsf{B}}^{x_1} \log \rho_0(Y_{t=0}^{\mathsf{B}}), \tag{2.67}$$

and

$$\mathsf{H}(\rho_0 \parallel \hat{\rho}_{\mathsf{B}}(0)) \leq -\mathbb{E}_0 \mathbb{E}_{\mathsf{F}}^{x_0} \int_0^1 \nabla \cdot b_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt - \mathbb{E}_0 \mathbb{E}_{\mathsf{F}}^{x_0} \log \rho_1(Y_{t=1}^{\mathsf{F}}). \tag{2.68}$$

However, these bounds are not sharp in general – in fact, using calculations similar to the one presented in the proof of Theorem 28, we can derive exact expressions that capture precisely what is lost when applying Jensen's inequality:

$$\mathsf{H}(\rho_1 \parallel \hat{\rho}_{\mathsf{F}}(1)) = \int_0^1 \mathbb{E}_1 \mathbb{E}_{\mathsf{B}}^{x_1} \left(\nabla \cdot b_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) - \epsilon |\nabla \log \hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})|^2 \right) dt - \mathbb{E}_1 \mathbb{E}_{\mathsf{B}}^{x_1} \log \rho_0(Y_{t=0}^{\mathsf{B}}),$$
$$\tag{2.69}$$

and

$$\mathsf{H}(\rho_0 \parallel \hat{\rho}_{\mathsf{B}}(0)) = -\mathbb{E}_0 \mathbb{E}_{\mathsf{F}}^{x_0} \int_0^1 \left(\nabla \cdot b_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) + \epsilon |\nabla \log \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}})|^2 \right) dt - \mathbb{E}_0 \mathbb{E}_{\mathsf{F}}^{x_0} \log \rho_1(Y_{t=1}^{\mathsf{F}}).$$
$$\tag{2.70}$$

Unfortunately, since $\nabla \log \hat{\rho}_{\mathsf{F}} \neq \hat{s}$ *and* $\nabla \log \hat{\rho}_{\mathsf{B}} \neq \hat{s}$ *in general due to approximation errors, we do not know how to estimate the extra terms on the right-hand side of* (2.69) *and* (2.70). *One possibility is to use* $\hat{s}$ *as a proxy for* $\nabla \log \hat{\rho}_{\mathsf{F}}$ *and* $\nabla \log \hat{\rho}_{\mathsf{B}}$*, which may be useful in practice, but this approximation is uncontrolled in general.*

3. Instantiations and extensions

In this section, we instantiate the stochastic interpolant framework discussed in Section 2.

3.1 Diffusive interpolants

Recently, there has been a surge of interest in the construction of generative models through diffusive bridge processes (Peluchetti, 2022; Liu et al., 2022; Somnath et al., 2023). In this section, we connect these approaches with our own, highlighting that stochastic interpolants allow us to manipulate certain bridge processes in a simpler and more direct manner. We also show that this perspective leads to a generative process that samples any target density ρ_1 by pushing a point mass at any $x_0 \in \mathbb{R}^d$ through an SDE. We begin by introducing a new kind of interpolant:

Definition 30 (Diffusive interpolant) *Given two probability density functions* $\rho_0, \rho_1 :$ $\mathbb{R}^d \to \mathbb{R}_{\geq 0}$, *a diffusive interpolant between* ρ_0 *and* ρ_1 *is a stochastic process* x_t^{d} *defined as*

$$x_t^{\mathsf{d}} = I(t, x_0, x_1) + \sqrt{2a(t)}B_t, \qquad t \in [0, 1], \tag{3.1}$$

where:

1. $I(t, x_0, x_1)$ *is as in Definition 1;*

2. $(x_0, x_1) \sim \nu$ *with* ν *satisfying (2.3) in Definition 1;*

3. $a(t) \in C^2([0, 1])$ *with* $a(0) > 0$ *and* $a(t) \geq 0$ *for all* $t \in (0, 1]$, *and;*

4. B_t *is a standard Brownian bridge process, independent of* x_0 *and* x_1.

Pathwise, (3.1) is different from the stochastic interpolant introduced in Definition 1: in particular, x_t^{d} is continuous but not differentiable in time. At the same time, since B_t is a Gaussian process with mean zero and variance $\mathbb{E}B_t^2 = t(1-t)$, (3.1) has the same single-time statistics and time-dependent density $\rho(t, x)$ as the stochastic interpolant (2.1) if we set $\gamma(t) = \sqrt{2a(t)t(1-t)}$, i.e.

$$x_t = I(t, x_0, x_1) + \sqrt{2a(t)t(1-t)}z \quad \text{with} \quad (x_0, x_1) \sim \nu, \; z \sim \mathsf{N}(0, Id), \; (x_0, x_1) \perp z. \tag{3.2}$$

As a result, (3.1) and (3.2) lead to the *same* generative models. Technically, it is easier to work with (3.2) than with (3.1), because it avoids the use of Itô calculus, and enables direct sampling of x_t using samples from ρ_0, ρ_1, and $\mathsf{N}(0, Id)$. However, (3.1) sheds light on some interesting properties of the generative models based on (3.2), i.e. stochastic interpolants with $\gamma(t) = \sqrt{2a(t)t(1-t)}$. To see why, we now re-derive the transport equation for the density $\rho(t, x)$ shared by (3.1) and (3.2) using the relation (3.1) using Fourier analysis. For simplicity, we focus on the case where $a(t)$ is constant in time, i.e. we set $a(t) = a > 0$ in (3.1).

To begin, recall that the Brownian Bridge B_t can be expressed in terms of the Wiener process W_t as $B_t = W_t - tW_{t=1}$. Moreover, it satisfies the SDE obtained by conditioning on $B_{t=1} = 0$ via Doob's h-transform (Doob, 1984):

$$dB_t = -\frac{B_t}{1-t}dt + dW_t, \qquad B_{t=0} = 0. \tag{3.3}$$

A direct application of Itô's formula implies that

$$de^{ik \cdot x_t^{\mathsf{d}}} = ik \cdot \left(\partial_t I(t, x_0, x_1) - \frac{\sqrt{2a}B_t}{1-t}\right)e^{ik \cdot x_t^{\mathsf{d}}}dt - a|k|^2 e^{ik \cdot x_t^{\mathsf{d}}}dt + \sqrt{2a}ik \cdot dW_t e^{ik \cdot x_t^{\mathsf{d}}}. \tag{3.4}$$

Taking the expectation of this expression and using the independence between (x_0, x_1) and B_t, we deduce that

$$\partial_t \mathbb{E}e^{ik \cdot x_t^{\mathsf{d}}} = ik \cdot \mathbb{E}\left(\left(\partial_t I(t, x_0, x_1) - \frac{\sqrt{2a}B_t}{1-t}\right)e^{ik \cdot x_t^{\mathsf{d}}}\right) - a|k|^2 \mathbb{E}e^{ik \cdot x_t^{\mathsf{d}}}. \tag{3.5}$$

Since for all fixed $t \in [0,1]$ we have $B_t \overset{d}{=} \sqrt{t(1-t)}z$ and $x_t^{\mathsf{d}} \overset{d}{=} x_t$ with x_t defined in (3.2), the time derivative (3.5) can also be written as

$$\partial_t \mathbb{E}e^{ik\cdot x_t} = ik \cdot \mathbb{E}\left(\left(\partial_t I(t, x_0, x_1) - \frac{\sqrt{2at}\,z}{\sqrt{1-t}}\right)e^{ik\cdot x_t}\right) - a|k|^2 \mathbb{E}e^{ik\cdot x_t}. \tag{3.6}$$

Moreover, since by definition of their probability density we have $\mathbb{E}e^{ik\cdot x_t^{\mathsf{d}}} = \mathbb{E}e^{ik\cdot x_t} = \int_{\mathbb{R}^d} e^{ik\cdot x}\rho(t,x)dx$, we can deduce from (3.6) that $\rho(t)$ satisfies

$$\partial_t \rho + \nabla \cdot (u\rho) = a\Delta\rho, \tag{3.7}$$

where we defined

$$u(t,x) = \mathbb{E}\left(\partial_t I(t, x_0, x_1) - \frac{\sqrt{2at}\,z}{\sqrt{1-t}}\bigg| x_t = x\right). \tag{3.8}$$

For the interpolant x_t in (3.2), we have from the definitions of b and s in (2.10) and (2.14) that

$$b(t,x) = \mathbb{E}\left(\partial_t I(t, x_0, x_1) + \frac{a(1-2t)z}{\sqrt{2t(1-t)}}\bigg| x_t = x\right),$$

$$s(t,x) = \nabla\log\rho(t,x) = -\frac{1}{\sqrt{2at(1-t)}}\mathbb{E}(z|x_t = x), \tag{3.9}$$

As a result, $u - s = b$ and (3.7) can also be written as the TE (2.9) using $\Delta\rho = \nabla \cdot (s\rho)$.

Conditional sampling. Remarkably, the drift u defined in (3.8) remains non-singular for all $t \in [0,1]$ (including $t = 0$) even if ρ_0 is replaced by a point mass at x_0; by contrast, both b and s are singular at $t = 0$ in this case. Hence, the SDE associated with the FPE (3.7) provides us with a generative model that samples ρ_1 from a base measure concentrated at a single x_0 (i.e. such that the density ρ_0 is replaced by a point mass measure at $x = x_0$). We formalize this result in the following theorem:

Theorem 31 *Assume that $I(t, x_0, x_1) = x_0$ for $t \in [0, \delta]$ with some $\delta \in (0,1]$. Given any $a > 0$, let*

$$u^{\mathsf{d}}(t, x, x_0) = \mathbb{E}_{x_1, z}\left(\partial_t I(t, x_0, x_1) - \frac{\sqrt{2at}\,z}{\sqrt{1-t}}\bigg| x_t = x\right), \tag{3.10}$$

where x_t is given by (3.2) and where $\mathbb{E}_{x_1, z}(\cdot|x_t = x)$ denotes an expectation over $x_1 \sim \rho_1 \perp z \sim \mathsf{N}(0, Id)$ conditioned on $x_t = x$ with $x_0 \in \mathbb{R}^d$ fixed. Then $u^{\mathsf{d}}(\cdot, \cdot, x_0) \in C^0([0,1]; (C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$ and $x_0 \in \mathbb{R}^d$. Moreover, the solutions to the forward SDE

$$dX_t^{\mathsf{d}} = u^{\mathsf{d}}(t, X_t^{\mathsf{d}}, x_0)dt + \sqrt{2a}\,dW_t, \qquad X_{t=0}^{\mathsf{d}} = x_0, \tag{3.11}$$

are such that $X_{t=1}^{\mathsf{d}} \sim \rho_1$.

Note that the additional assumption we make on $I(t, x_0, x_1)$ is consistent with the requirements in Definition 1 and Assumption 5: this additional assumption is made for simplicity and can probably be relaxed to $\partial_t I(t = 0, x_0, x_1) = 0$.

The proof of Theorem 31 is given in Appendix B.5. It relies on the calculations that led to (3.8), along with the observation that at $t = 0$ and $x = x_0$,

$$u^{\mathsf{d}}(t = 0, x_0, x_0) = \mathbb{E}_{x_1}\left(\partial_t I(t = 0, x_0, x_1)\right), \tag{3.12}$$

whereas at $t = 1$ and any $x \in \mathbb{R}^d$, we have

$$u^{\mathsf{d}}(t = 1, x, x_0) = \partial_t I(t = 1, x_0, x) + 2a\nabla \log \rho_1(x), \tag{3.13}$$

which are both well-defined. To put the result in Theorem (31) in perspective, observe that no probability flow ODE with $b \in C^0([0, 1]; (C^p(\mathbb{R}^d))^d)$ can achieve the same feat as the diffusion in (3.11). This is because the solutions of such an ODE are unique, and therefore can only map x_0 onto a single point at time $t = 1$. Moreover, $u^{\mathsf{d}}(t, x, x_0)$ is the unique minimizer of the objective function

$$\mathcal{L}_{u^{\mathsf{d}}}[\hat{u}^{\mathsf{d}}] = \int_0^1 \mathbb{E}_{x_1, z} \left(|\hat{u}^{\mathsf{d}}(t, x_t, x_0)|^2 - 2\left(\partial_t I(t, x_0, x_1) - \frac{\sqrt{2at}z}{\sqrt{(1-t)}}\right) \cdot \hat{u}^{\mathsf{d}}(t, x_t, x_0) \right) dt.$$

3.2 One-sided interpolants for Gaussian ρ_0

A common choice of base density for generative modeling in the absence of prior information is to choose $\rho_0 = \mathsf{N}(0, Id)$. In this setting, we can group the effect of the latent variable z with x_0. This leads to a simpler type of stochastic interpolant that, in particular, will enables us to instantiate score-based diffusion within our general framework (see Section 3.4).

Definition 32 (One-sided stochastic interpolant) *Given a probability density function $\rho_1 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, a one-sided stochastic interpolant between $\mathsf{N}(0, Id)$ and ρ_1 is a stochastic process x_t^{os}*

$$x_t^{\mathsf{os}} = \alpha(t)z + J(t, x_1), \qquad t \in [0, 1] \tag{3.14}$$

that fulfills the requirements:

1. $J \in C^2([0, 1], C^2(\mathbb{R}^d)^d)$ satisfies the boundary conditions $J(0, x_1) = 0$ and $J(1, x_1) = x_1$.

2. x_1 and z are independent random variables drawn from ρ_1 and $\mathsf{N}(0, Id)$, respectively.

3. $\alpha : [0, 1] \to \mathbb{R}$ satisfies $\alpha(0) = 1$, $\alpha(1) = 0$, $\alpha(t) > 0$ for all $t \in [0, 1)$, and $\alpha^2 \in C^2([0, 1])$.

By construction, $x_{t=0}^{\mathsf{os}} = z \sim \mathsf{N}(0, Id)$ and $x_{t=1}^{\mathsf{os}} = x_1 \sim \rho_1$, so that the distribution of the stochastic process x_t^{os} bridges $\mathsf{N}(0, Id)$ and ρ_1. It is easy to see that the one-sided stochastic interpolant defined in (3.14) will have the same density as the stochastic interpolant defined in (2.1) if we set $I(t, x_0, x_1) = J_t(x_1) + \delta(t)x_0$ and take $\delta^2(t) + \gamma^2(t) = \alpha^2(t)$. Restricting to this case, our earlier theoretical results apply where the velocity field b defined in (2.10) becomes

$$b(t, x) = \mathbb{E}(\dot{\alpha}(t)z + \partial_t J(t, x_1)|x_t^{\mathsf{os}} = x), \tag{3.15}$$

and the quadratic objective in (2.13) becomes

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{b}(t, x_t^{\mathsf{os}})|^2 - (\dot{\alpha}(t)z + \partial_t J(t, x_1)) \cdot \hat{b}(t, x_t^{\mathsf{os}})\right) dt. \tag{3.16}$$

In the expression above, x_t^{os} is given by (3.14) and the expectation $\mathbb{E}$ is taken independently over $x_1 \sim \rho_1$ and $z \sim \mathsf{N}(0, Id)$. Similarly, the score is given by

$$s(t, x) = -\alpha^{-1}(t)\eta_z(t, x), \qquad \eta_z(t, x) = \mathbb{E}(z|x_t^{\mathsf{os}} = x), \tag{3.17}$$

26

where $\eta_z(t, x)$ is the equivalent of the denoiser defined in (2.17). These functions are the unique minimizers of the objectives

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{s}(t, x_t^{\mathsf{os}})|^2 + \gamma^{-1}(t)z \cdot \hat{s}(t, x_t^{\mathsf{os}})\right) dt, \tag{3.18}$$

$$\mathcal{L}_{\eta_z}[\hat{\eta}_z] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{\eta}_z(t, x_t^{\mathsf{os}})|^2 - z \cdot \hat{\eta}_z(t, x_t^{\mathsf{os}})\right) dt. \tag{3.19}$$

Moreover, we can weaken Assumption 5 to the following requirement:

Assumption 33 *The density $\rho_1 \in C^2(\mathbb{R}^d)$, satisfies $\rho_1(x) > 0$ for all $x \in \mathbb{R}^d$, and:*

$$\int_{\mathbb{R}^d} |\nabla \log \rho_1(x)|^2 \rho_1(x) dx < \infty. \tag{3.20}$$

The function J satisfies

$$\exists C_1 < \infty : \quad |\partial_t J(t, x_1)| \le C_1|x_1| \qquad \text{for all} \quad (t, x_1) \in [0, 1] \times \mathbb{R}^d, \tag{3.21}$$

and

$$\exists M_1, M_2 < \infty : \quad \mathbb{E}\left[|\partial_t J(t, x_1)|^4\right] \le M_1; \quad \mathbb{E}\left[|\partial_t^2 J(t, x_1)|^2\right] \le M_2, \quad \text{for all} \quad t \in [0, 1], \tag{3.22}$$

where the expectation is taken over $x_1 \sim \rho_1$.

Remark 34 *The construction above can easily be generalized to the case where $\rho_0 = \mathsf{N}(0, C_0)$ with C_0 a positive-definite matrix. Without loss of generality, we can then assume that C_0 can be represented as $C_0 = \sigma_0 \sigma_0^{\mathsf{T}}$ where σ_0 is a lower-triangular matrix and replace (3.14)*

$$x_t^{\mathsf{os}} = \alpha(t)\sigma_0 z + J(t, x_1), \qquad t \in [0, 1], \tag{3.23}$$

with J and α satisfying the conditions listed in Definition 32 and where $z \sim \mathsf{N}(0, Id)$.

3.3 Mirror interpolants

Another practically relevant setting is when the base and the target are the same density ρ_1. In this setting we can define a stochastic interpolant as:

Definition 35 (Mirror stochastic interpolant) *Given a probability density function $\rho_1 : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$, a mirror stochastic interpolant between ρ_0 and itself is a stochastic process x_t^{mir}*

$$x_t^{\mathsf{mir}} = K(t, x_1) + \gamma(t)z, \qquad t \in [0, 1] \tag{3.24}$$

that fulfills the requirements:

1. $K \in C^2([0, 1], C^2(\mathbb{R}^d)^d)$ satisfies the boundary conditions $K(0, x_1) = x_1$ and $K(1, x_1) = x_1$.

2. x_1 and z are random variables drawn independently from ρ_1 and $\mathsf{N}(0, Id)$, respectively.

3. $\gamma : [0, 1] \to \mathbb{R}$ satisfies $\gamma(0) = \gamma(1) = 0$, $\gamma(t) > 0$ for all $t \in (0, 1)$, and $\gamma^2 \in C^1([0, 1])$.

By construction, $x_{t=0}^{\text{mir}} = x_{t=1}^{\text{mir}} = x_1 \sim \rho_1$, so that the distribution of the stochastic process x_t^{mir} bridges ρ_1 to itself. Note that a valid choice is $K(t, x_1) = \alpha(t)x_1$ with $\alpha(0) = \alpha(1) = 1$ (e.g. $\alpha(t) = 1$): in this case, mirror interpolants are related to denoisers, as will be discussed in Section 5.2.

It is easy to see that our earlier theoretical results apply where the velocity field b defined in (2.10) becomes

$$b(t, x) = \mathbb{E}(\partial_t K(t, x_1) + \dot{\gamma}(t)z | x_t^{\text{mir}} = x), \tag{3.25}$$

and the quadratic objective in (2.13) becomes

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{b}(t, x_t^{\text{mir}})|^2 - (\partial_t K(t, x_1) + \dot{\gamma}(t)z) \cdot \hat{b}(t, x_t^{\text{mir}})\right) dt. \tag{3.26}$$

In the expression above, x_t^{mir} is given by (3.24) and the expectation $\mathbb{E}$ is taken independently over $x_1 \sim \rho_1$ and $z \sim \mathsf{N}(0, Id)$. Similarly, the score is given by

$$s(t, x) = -\gamma^{-1}(t)\eta_z(t, x), \qquad \eta_z(t, x) = \mathbb{E}(z | x_t^{\text{mir}} = x), \tag{3.27}$$

which are the unique minimizers of the objective functions

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{s}(t, x_t^{\text{mir}})|^2 + \gamma^{-1}(t)z \cdot \hat{s}(t, x_t^{\text{mir}})\right) dt. \tag{3.28}$$

$$\mathcal{L}_{\eta_z}[\hat{\eta}_z] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{\eta}_z(t, x_t^{\text{mir}})|^2 - z \cdot \hat{\eta}_z(t, x_t^{\text{mir}})\right) dt. \tag{3.29}$$

Moreover, we can weaken Assumption 5 to the following requirement:

Assumption 36 *The density $\rho_1 \in C^2(\mathbb{R}^d)$ satisfies $\rho_1(x) > 0$ for all $x \in \mathbb{R}^d$ and*

$$\int_{\mathbb{R}^d} |\nabla \log \rho_1(x)|^2 \rho_1(x)dx < \infty. \tag{3.30}$$

The function K satisfies

$$\exists C_1 < \infty \ : \ |\partial_t K(t, x_1)| \leq C_1|x_1| \qquad \text{for all} \quad (t, x_1) \in [0, 1] \times \mathbb{R}^d, \tag{3.31}$$

and

$$\exists M_1, M_2 < \infty \ : \ \mathbb{E}\big[|\partial_t K(t, x_1)|^4\big] \leq M_1; \quad \mathbb{E}\big[|\partial_t^2 K(t, x_1)|^2\big] \leq M_2, \quad \text{for all} \quad t \in [0, 1], \tag{3.32}$$

where the expectation is taken over $x_1 \sim \rho_1$.

Remark 37 *Interestingly, if we take $K(t, x_1) = x_1$, then $\partial_t K(t, x_1) = 0$, and the velocity field defined in (3.25) is completely defined by the denoiser η_z*

$$b(t, x) = \dot{\gamma}(t)\eta_z(t, x) \tag{3.33}$$

Since the score s also depends on η_z, this denoiser is the only quantity that needs to be learned.

Remark 38 *If ρ_1 is only accessible via empirical samples, mirror interpolants do not enable calculation of the functional form of ρ_1. A notable exception is if we set $K(t, x_1) = 0$ for $t \in [t_1, t_2]$ with $0 < t_1 \le t_2 < 1$: in that case, $x_t^{\text{mir}} = \gamma(t)z \sim \gamma(t)\mathsf{N}(0, Id)$ for $t \in [t_1, t_2]$, which gives us a reference density for comparison. In this setup, mirror interpolants essentially reduce to two one-sided interpolants glued together (with the second one time-reversed), or in fact a regular stochastic interpolant when $\rho_0 = \rho_1$ and we set $I(t, x_0, x_1) = 0$ for $t \in [t_1, t_2]$.*

3.4 Stochastic interpolants and Schrödinger bridges

The stochastic interpolant framework can also be used to solve the Schrödinger bridge problem. For background material on this problem, we refer the reader to Léonard (2014) and the references therein. Consistent with the overall viewpoint of this paper, we consider the hydrodynamic formulation of the Schrödinger bridge problem, in which the goal is to obtain a pair (ρ, u), that solves the following optimization problem for a fixed $\epsilon > 0$

$$\min_{\hat{u}, \hat{\rho}} \int_0^1 \int_{\mathbb{R}^d} |\hat{u}(t, x)|^2 \hat{\rho}(t, x) dx dt \tag{3.34}$$

$$\text{subject to:} \quad \partial_t \hat{\rho} + \nabla \cdot (\hat{u}\hat{\rho}) = \epsilon \Delta \hat{\rho}, \quad \hat{\rho}(0) = \rho_0 \quad \hat{\rho}(1) = \rho_1$$

Under our assumptions on ρ_0 and ρ_1 listed in Assumption 5, it is known (see e.g. Proposition 4.1 in Léonard (2014)) that (3.34) has a unique minimizer $(\rho, u = \nabla\lambda)$, with (ρ, λ) classical solutions of the Euler-Lagrange equations:

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\nabla\lambda\rho) &= \epsilon\Delta\rho, \quad \rho(0) = \rho_0, \quad \rho(1) = \rho_1, \\ \partial_t \lambda + \tfrac{1}{2}|\nabla\lambda|^2 &= -\epsilon\Delta\lambda. \end{aligned} \tag{3.35}$$

To proceed we will make the additional assumption that the solution ρ to (3.35) can be reversibly mapped to a standard Gaussian:

Assumption 39 *There exists a reversible map $T : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ with $T, T^{-1} \in C^1([0, 1], (C^d(\mathbb{R}^d))^d)$ such that:*

$$\forall t \in [0, 1] \quad : \quad z \sim \mathsf{N}(0, Id) \implies T(t, z) \sim \rho(t); \quad x_t \sim \rho(t) \implies T^{-1}(t, x_t) \sim \mathsf{N}(0, Id), \tag{3.36}$$

where ρ is the solution to (3.35).

We stress that the actual form of the map T is not important for the arguments below. Assumption 39 can be used to show the existence of a stochastic interpolant whose density solves (3.35):

Lemma 40 *If Assumption (39) holds, then the solution $\rho(t)$ to (3.35) is the density of the stochastic interpolant*

$$x_t = T(t, \alpha(t)T^{-1}(0, x_0) + \beta(t)T^{-1}(1, x_1)) + \gamma(t)z, \tag{3.37}$$

as long as $\alpha^2(t) + \beta^2(t) + \gamma^2(t) = 1$.

The proof is given in Appendix B.6: (3.37) corresponds to choosing $I(t, x_0, x_1) = T(t, \alpha(t)T^{-1}(t, x_0) + \beta(t)T^{-1}(t, x_1))$ in (2.1). With the help of Lemma 40, we can establish the following result, which shows how to optimize over the function I to solve the problem (3.34)

Theorem 41 *Pick some* $\gamma : [0, 1] \to [0, 1)$ *such that* $\gamma(0) = \gamma(1) = 0$, $\gamma(t) > 0$ *for* $t \in (0, 1)$, $\gamma \in C^2((0, 1))$ *and* $\gamma^2 \in C^1([0, 1])$, *and let* $\hat{x}_t = \hat{I}(t, x_0, x_1) + \gamma(t)z$, *with* $x_0 \sim \rho_0$, $x_1 \sim \rho_1$, *and* $z \sim \mathsf{N}(0, Id)$ *all independent. Consider the max-min problem over* $\hat{I} \in C^1([0, 1], (C^1(\mathbb{R}^d \times \mathbb{R}^d))^d)$ *and* $\hat{u} \in C^0([0, 1], (C^1(\mathbb{R}^d))^d)$:

$$\max_{\hat{I}} \min_{\hat{u}} \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{u}(t, \hat{x}_t)|^2 - \left(\partial_t \hat{I}(t, x_0, x_1) + (\dot{\gamma}(t) - \epsilon\gamma^{-1}(t))z \right) \cdot \hat{u}(t, \hat{x}_t) \right) dt. \qquad (3.38)$$

If Assumption 39 holds, then all the optimizers (I, u) *of* (3.38) *are such that the density of the associated* $x_t = I(t, x_0, x_1) + \gamma(t)z$ *is the solution* ρ *to* (3.35). *Moreover,* $u = \nabla\lambda$, *with* λ *the solution to* (3.35).

The proof is also given in Appendix B.6. Note that if we fix $\hat{I}$, the velocity u minimizing this objective is the forward drift b_F defined in (2.21). Note also that if we set $\epsilon \to 0$, the minimizing velocity field is b as defined in (2.10), and the max-min problem formally reduces to solving the optimal transport problem. In this case, Assumption 39 becomes more stringent, as we need to assume that that system (3.35) with $\epsilon = 0$ (i.e. in the absence of the diffusive terms) has a classical solution. Theorem (41) gives a practical route towards solving the Schrödinger bridge problem with stochastic interpolants, and we leave the numerical investigation of this formulation to future work.

4. Spatially linear interpolants

In this section, we study the stochastic interpolants that are obtained when we specialize the function I used in (2.1) to be linear in both x_0 and x_1, i.e. we consider

$$x_t^{\mathsf{lin}} = \alpha(t)x_0 + \beta(t)x_1 + \gamma(t)z, \qquad (4.1)$$

where $(x_0, x_1) \sim \nu$ and $z \sim \mathsf{N}(0, Id)$ with $(x_0, x_1) \perp z$, and $\alpha, \beta, \gamma^2 \in C^2([0, 1])$ satisfy the conditions

$$\alpha(0) = \beta(1) = 1; \quad \alpha(1) = \beta(0) = \gamma(0) = \gamma(1) = 0; \quad \forall t \in (0, 1) \; : \; \gamma(t) > 0. \qquad (4.2)$$

Despite its simplicity, this setup offers significant design flexibility. The discussion highlights how the presence of the latent variable $\gamma(t)z$ can simplify the structure of the intermediate density $\rho(t)$. Since our ultimate aim is to investigate the properties of practical generative models built upon ODEs or SDEs, we will also study the effect of time-dependent diffusion coefficient $\epsilon(t)$, which controls the amplitude of the noise in a generative SDE. Throughout, to build intuition, we choose ρ_0 and ρ_1 to be Gaussian mixture densities, for which the drift coefficients can be computed analytically (see Appendix A). This enables us to visualize the effect of each choice on the resulting generative models.

4.1 Factorization of the velocity field

When the stochastic interpolant is of the form (4.1), the velocity b and the score s defined in (2.10) and (2.14) can both be expressed in terms of the following three conditional expectations (the third is the denoiser already defined in (2.17)):

$$\eta_0(t, x) = \mathbb{E}(x_0|x_t^{\mathsf{lin}} = x), \quad \eta_1(t, x) = \mathbb{E}(x_1|x_t^{\mathsf{lin}} = x), \quad \eta_z(t, x) = \mathbb{E}(z|x_t^{\mathsf{lin}} = x). \qquad (4.3)$$

Specifically, we have

$$b(t,x) = \dot{\alpha}(t)\eta_0(t,x) + \dot{\beta}(t)\eta_1(t,x) + \dot{\gamma}(t)\eta_z(t,x), \qquad s(t,x) = -\gamma^{-1}(t)\eta_z(t,x). \quad (4.4)$$

The second relation above holds for $t \in (0,1)$ (i.e. when $\gamma(t) \neq 0$). Moreover, because $\mathbb{E}(x_t|x_t = x) = x$ by definition, the functions η_0, η_1, and η_z satisfy

$$\forall (t,x) \in [0,1] \times \mathbb{R}^d \quad : \quad \alpha(t)\eta_0(t,x) + \beta(t)\eta_1(t,x) + \gamma(t)\eta_z(t,x) = x. \quad (4.5)$$

This enables us to reduce computational expense: given two of the η's, the third can always be calculated via (4.5). Finally, it is easy to see that the functions η_0, η_1, and η_z are the unique minimizers of the objectives

$$\mathcal{L}_{\eta_0}(\hat{\eta}_0) = \int_0^1 \mathbb{E}\left[\tfrac{1}{2}|\hat{\eta}_0(t,x_t^{\mathsf{lin}})|^2 - x_0 \cdot \hat{\eta}_0(t,x_t^{\mathsf{lin}})\right]dt,$$

$$\mathcal{L}_{\eta_1}(\hat{\eta}_1) = \int_0^1 \mathbb{E}\left[\tfrac{1}{2}|\hat{\eta}_1(t,x_t^{\mathsf{lin}})|^2 - x_1 \cdot \hat{\eta}_1(t,x_t^{\mathsf{lin}})\right]dt, \qquad (4.6)$$

$$\mathcal{L}_{\eta_z}(\hat{\eta}_z) = \int_0^1 \mathbb{E}\left[\tfrac{1}{2}|\hat{\eta}_z(t,x_t^{\mathsf{lin}})|^2 - z \cdot \hat{\eta}_z(t,x_t^{\mathsf{lin}})\right]dt,$$

where x_t^{lin} is defined in (4.1) and the expectation is taken independently over $(x_0, x_1) \sim \nu$ and $z \sim \mathsf{N}(0, Id)$.

4.2 Some specific design choices

It is useful to assume that both ρ_0 and ρ_1 have been scaled to have zero mean and identity covariance (which can be achieved in practice, for example, by an affine transformation of the data). In this case, the time-dependent mean and covariance of (4.1) are given by

$$\mathbb{E}[x_t^{\mathsf{lin}}] = 0, \qquad \mathbb{E}[x_t^{\mathsf{lin}}(x_t^{\mathsf{lin}})^{\mathsf{T}}] = (\alpha^2(t) + \beta^2(t) + \gamma^2(t))Id. \quad (4.7)$$

Preserving the identity covariance at all times therefore leads to the constraint

$$\forall t \in [0,1] \quad : \quad \alpha^2(t) + \beta^2(t) + \gamma^2(t) = 1. \quad (4.8)$$

This choice is also sensible if ρ_0 and ρ_1 have covariances that are not the identity but are on a similar scale. In this case we no longer need to enforce (4.8) exactly, and could, for example, take three functions whose sum of squares is of order one. For definiteness, in the sequel we discuss possible choices that satisfy (4.8) exactly, with the understanding that the corresponding functions α, β, and γ could all be slightly modified without significantly affecting the conclusions.

Linear and trigonometric α and β. One way to ensure that (4.8) holds while maintaining the influence of ρ_0 and ρ_1 everywhere on $[0,1]$ except at the endpoints is to choose

$$\alpha(t) = t, \qquad \beta(t) = 1 - t, \qquad \gamma(t) = \sqrt{2t(1-t)}. \quad (4.9)$$

This choice was advocated in Liu et al. (2023b), without the inclusion of the latent variable ($\gamma = 0$). Another possibility that gives more leeway is to pick any $\gamma : [0,1] \to [0,1]$ and set

$$\alpha(t) = \sqrt{1 - \gamma^2(t)}\cos(\tfrac{1}{2}\pi t), \qquad \beta(t) = \sqrt{1 - \gamma^2(t)}\sin(\tfrac{1}{2}\pi t). \quad (4.10)$$

Stochastic Interpolant		$\alpha(t)$	$\beta(t)$	$\gamma(t)$
Arbitrary ρ_0 (two-sided)	linear	$1-t$	t	$\sqrt{at(1-t)}$
	trig	$\cos\frac{\pi}{2}t$	$\sin\frac{\pi}{2}t$	$\sqrt{at(1-t)}$
	enc-dec	$\cos^2(\pi t)1_{[0,\frac{1}{2})}(t)$	$\cos^2(\pi t)1_{(\frac{1}{2},1]}(t)$	$\sin^2(\pi t)$
Gaussian ρ_0 (one-sided)	linear	$1-t$	t	0
	trig	$\cos\frac{\pi}{2}t$	$\sin\frac{\pi}{2}t$	0
	SBDM (VP)	$\sqrt{1-t^2}$	t	0
Mirror		0	1	$\sqrt{at(1-t)}$

Table 1: **Spatially linear interpolants.** A table suggesting various linear interpolants. In general, this paper describes methods for arbitrary ρ_0 and ρ_1. In Section 4.4, we detail linear interpolants for one-sided generation, where ρ_0 is a Gaussian and the latent variable z can be absorbed into x_0. Later, in Section 5.1, we discuss how to recast score-based diffusion models (SBDM) as linear one-sided interpolants, which leads to the expressions given in the table when using a variance-preserving parameterization. Linear mirror interpolants, where $\rho_0 = \rho_1$ are equal, are defined by (3.24) with the choice $K(t, x_1) = \alpha(t)x_1$.

With $\gamma = 0$, this was the choice preferred in Albergo and Vanden-Eijnden (2023). The PDF $\rho(t)$ obtained with the choices (4.9) and (4.10) when ρ_0 and ρ_1 are both Gaussian mixture densities are shown in Figure 4. As this example shows, when ρ_0 and ρ_1 have distinct complex features, these would be duplicated in $\rho(t)$ at intermediate times if not for the smoothing effect of the latent variable; this behavior is seen in Figure 4, where it is most prominent in the first row with $\gamma(t) = 0$. From a statistical learning perspective, eliminating the formation of spurious features will simplify the estimation of the velocity field b, which becomes smoother as the formation of such features is suppressed.

Gaussian encoding-decoding. A useful limiting case is to devolve the data from ρ_0 completely into noise by the halfway point $t = \frac{1}{2}$ and to reconstruct ρ_1 completely from noise starting from $t = \frac{1}{2}$. One choice that allows us to do so while satisfying (4.8) is

$$\alpha(t) = \cos^2(\pi t)1_{[0,\frac{1}{2})}(t), \qquad \beta(t) = \cos^2(\pi t)1_{(\frac{1}{2},1]}(t), \qquad \gamma(t) = \sin^2(\pi t), \qquad (4.11)$$

where $1_A(t)$ is the indicator function of A, i.e. $1_A(t) = 1$ if $t \in A$ and $1_A(t) = 0$ otherwise. With this choice, it is easy to see that $x_{t=\frac{1}{2}} = \gamma(\frac{1}{2})z \sim \mathsf{N}(0, \gamma^2(\frac{1}{2}))$, which seamlessly glues together two interpolants: one between ρ_0 and a standard Gaussian, and one between a standard Gaussian and ρ_1.

$$x_t = \cos(\tfrac{1}{2}\pi t)x_0 + \sin(\tfrac{1}{2}\pi t)x_1$$

$$x_t = \sqrt{1 - \gamma^2(t)}\cos(\tfrac{1}{2}\pi t)x_0 + \sqrt{1 - \gamma^2(t)}\sin(\tfrac{1}{2}\pi t)x_1 + \gamma(t)z, \qquad \gamma(t) = \sqrt{2t(1-t)}$$

$$x_t = \cos^2(\pi t)(1_{[0,\frac{1}{2})}(t)x_0 + 1_{(\frac{1}{2},1]}(t)x_1) + \sin^2(\pi t)z$$

ρ_0 ρ_1

$t = 0.00$ $t = 0.25$ $t = 0.50$ $t = 0.75$ $t = 1.00$
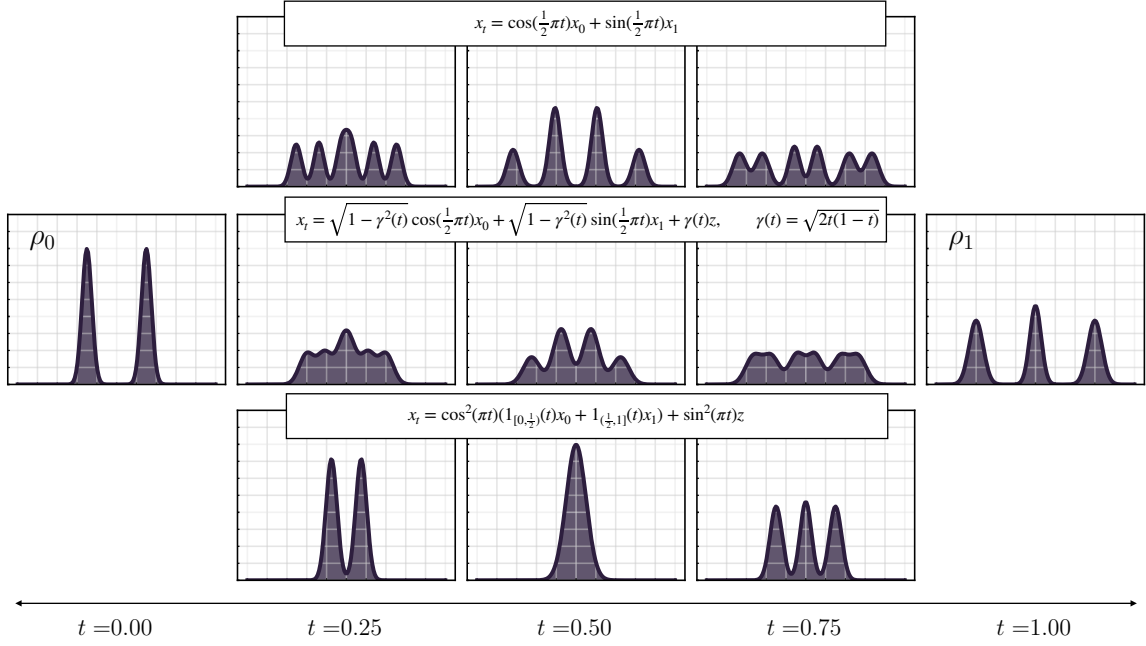
Figure 4: **The effect of $\gamma(t)z$ on $\rho(t)$.** A visualization of how the choice of $\gamma(t)$ changes the density $\rho(t)$ of $x_t^{\mathsf{lin}} = \alpha(t)x_0 + \beta(t)x_1 + \gamma(t)z$ when ρ_0 and ρ_1 are Gaussian mixture densities with two modes and three modes, respectively. The first row depicts $\gamma(t) = 0$, which reduces to the stochastic interpolant developed in Albergo and Vanden-Eijnden (2023). This case forms a valid transport between ρ_0 and ρ_1, but produces spurious intermediate modes in $\rho(t)$. The second row depicts the choice of $\gamma(t) = \sqrt{2t(1-t)}$. In this case, the spurious modes are partially damped by the addition of the latent variable, leading to a simpler $\rho(t)$. The final row shows the Gaussian encoding-decoding, which smoothly encodes ρ_0 into a standard normal distribution on the time interval $[0, 1/2]$, which is then decoded into ρ_1 on the interval $(1/2, 1]$. In this case, no intermediate modes form in $\rho(t)$: the two modes in ρ_0 collide to form $\mathsf{N}(0, 1)$ at $t = \frac{1}{2}$, which then spreads into the three modes of ρ_1. A visualization of individual sample trajectories from deterministic and stochastic generative models based on ODEs and SDEs whose solutions have density $\rho(t)$ can be seen in Figure 5.

Even though the choice (4.11) encodes ρ_0 into pure noise on the interval $[0, \frac{1}{2}]$, which is then decoded into ρ_1 on the interval $[\frac{1}{2}, 1]$ (and vice-versa when proceeding backwards in time), the resulting velocity b still defines a single continuity equation that maps ρ_0 to ρ_1 on $[0, 1]$. This is most clearly seen at the level of the probability flow (2.32), since its solution X_t is a bijection between the initial and final conditions $X_{t=0}$ and $X_{t=1}$, but a similar pairing can also be observed in the solutions to the forward and
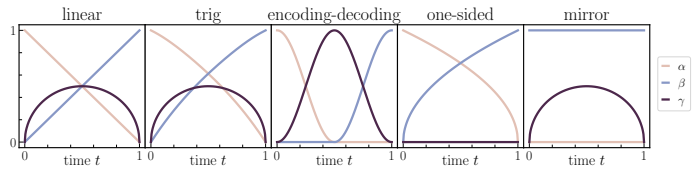

Figure 6: **The functions $\alpha(t)$, $\beta(t)$, and $\gamma(t)$** for the linear (4.9), trigonometric (4.10) with $\gamma(t) = \sqrt{2t(1-t)}$, Gaussian encoding-decoding (4.11), one-sided (3.14), and mirror (3.24) interpolants.
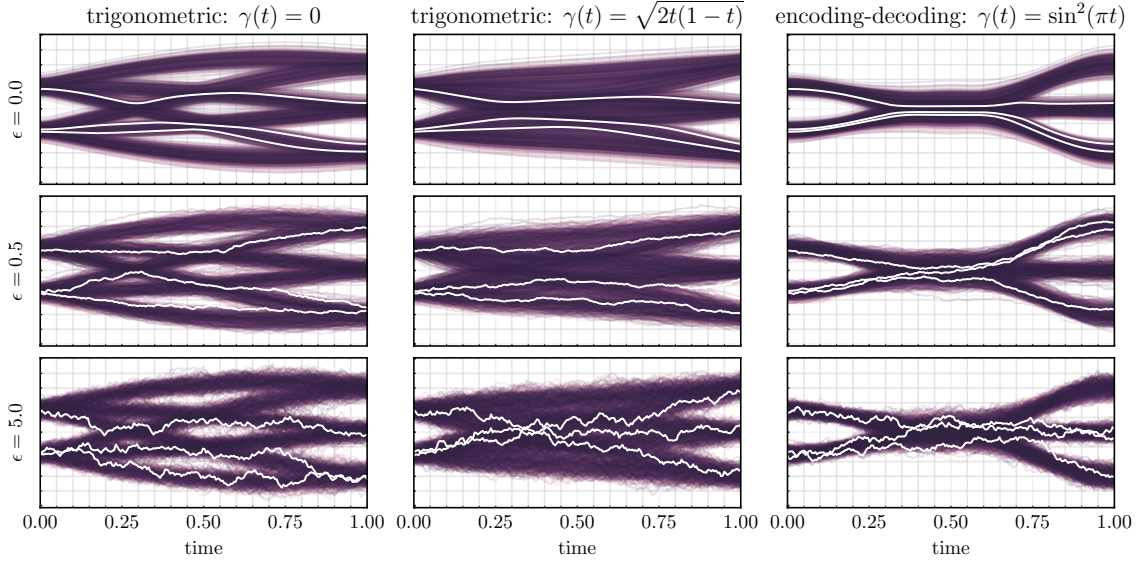
Figure 5: **The effect of ϵ on sample trajectories.** A visualization of how the choice of ϵ affects the sample trajectories obtained by solving the ODE (2.32) or the forward SDE (2.33). The set-up is the same as in Figure 4: ρ_0 and ρ_1 are taken to be the same Gaussian mixture densities, and the analytical expressions for b and s are used. In the three panels in each column the value of γ is the same, and each panel shows trajectories with different ϵ. Three specific trajectories from the same three initial conditions drawn from ρ_0 are also highlighted in white in every panel. As ϵ increases but γ stays the same, the density $\rho(t)$ is unchanged, but the individual trajectories become increasingly stochastic. While all choices are equivalent with exact b and s, Theorem 23 shows that nonzero values of ϵ provide control on the likelihood in terms of the error in b and s when they are approximate.

backward SDEs (2.33) and (2.34), whose solutions at time $t = 1$ or $t = 0$ remain correlated with the initial or final condition used. This allows for a more direct means of image-to-image translation with diffusions when compared to the recent approach described in Su et al. (2023). The choice (4.11) is depicted in the final row of Figure 4, where no spurious modes form at all; individual sample trajectories of the deterministic and stochastic generative models based on ODEs and SDEs whose solutions have this $\rho(t)$ as density can be seen in the panels forming the third column in Figure 5. We note that the elimination of spurious intermediate modes can also be implemented by use of a data-adapted coupling $\nu(dx_0, dx_1)$, as considered in Albergo et al. (2023).

Unsurprisingly, it is necessary to have $\gamma(t) > 0$ for the choice (4.11): for $\gamma(t) = 0$, the density $\rho(t)$ collapses to a Dirac measure at $t = \frac{1}{2}$. This consideration highlights that the inclusion of the latent variable $\gamma(t)z$ matters even for the deterministic dynamics (2.32), and its presence is distinct from the stochasticity inherent to the SDEs (2.33) and (2.34).

4.3 Impact of the latent variable $\gamma(t)z$ and the diffusion coefficient $\epsilon(t)$

The stochastic interpolant framework enables us to discern the independent roles of the latent variable $\gamma(t)z$ and the diffusion coefficient $\epsilon(t)$ we use in a generative model. As shown in Theorem 6, the presence of the latent variable $\gamma(t)z$ for $\gamma \neq 0$ smooths both the density $\rho(t)$ and the velocity b defined in (2.10) spatially. This provides a computational advantage at sample generation time because it simplifies the required numerical integration of (2.32), (2.33), and (2.34). Intuitively, this is because the density $\rho(t)$ of x_t can be represented exactly as the density that would be obtained with $\gamma(t) = 0$ convolved with $\mathsf{N}(0, \gamma^2(t)Id)$ at each $t \in (0,1)$. A comparison between the density $\rho(t)$ obtained with trigonometric interpolants with $\gamma(t) = 0$ and $\gamma(t) = \sqrt{2t(1-t)}$ can be seen in the first and second row of Figure 4.

By contrast, the diffusion coefficient $\epsilon(t)$ leaves the density $\rho(t)$ unchanged, and only affects the way we sample it. In particular, the probability flow ODE (2.32) results in a map that pushes every $X_{t=0} = x_0$ onto a single $X_{t=1} = x_1$ and vice-versa. The forward SDE (2.33) maps each $X_{t=0}^{\mathsf{F}} = x_0$ onto an ensemble $X_{t=1}^{\mathsf{F}}$ whose spread is controlled by the amplitude of $\epsilon(t)$ (and similarly for the reversed SDE (2.34) that maps each $X_{t=1}^{\mathsf{B}} = x_1$ onto an ensemble $X_{t=0}^{\mathsf{B}}$). This ensemble is not distributed according to ρ_1 for finite $\epsilon(t)$ – like with the ODE, we need to sample initial conditions from ρ_0 to get solutions at time $t = 1$ that sample ρ_1 – but its density converges towards ρ_1 as $\epsilon(t) \to \infty$. These features are illustrated in Figure 5.

Remark 42 *Another potential advantage of including the latent variable $\gamma(t)z$ is its impact on the velocity b at the end points. Since $x_{t=0} = x_0$ and $x_{t=1} = x_1$, it is easy to see that the velocity b of the linear interpolant x_t defined in (4.1) satisfies*

$$
\begin{aligned}
b(0, x) &= \dot{\alpha}(0)x + \dot{\beta}(0)\mathbb{E}[x_1|x_0 = x] - \lim_{t \to 0} \gamma(t)\dot{\gamma}(t)s_0(x), \\
b(1, x) &= \dot{\alpha}(1)\mathbb{E}[x_0|x_1 = x] + \dot{\beta}(1)x - \lim_{t \to 1} \gamma(t)\dot{\gamma}(t)s_1(x),
\end{aligned}
\tag{4.12}
$$

where $s_0 = \nabla \log \rho_0$ and $s_1 = \nabla \log \rho_1$. If $\gamma \in C^2([0,1])$, because $\gamma(0) = \gamma(1) = 0$, the terms involving the scores s_0 and s_1 in these expressions vanish. Choosing $\gamma^2 \in C^1([0,1])$ but γ not differentiable at $t = 0$ or $t = 1$ leaves open the possibility that the limits remain nonzero. For example, if we take one of the choices discussed in Section 3.1, i.e.

$$
\gamma(t) = \sqrt{at(1-t)}, \qquad a > 0,
\tag{4.13}
$$

we obtain

$$
\lim_{t \to 0} \gamma(t)\dot{\gamma}(t) = -\lim_{t \to 1} \gamma(t)\dot{\gamma}(t) = \frac{a}{2}.
\tag{4.14}
$$

As a result, the choice (4.13) ensures that the velocity b encodes information about the score of the densities ρ_0 and ρ_1 at the end points. We stress however that, while the choice of $\gamma(t)$ given in (4.13) is appealing because of its nontrivial influence on the velocity b at the endpoints, the user is free to explore a variety of alternatives. We present some examples in Table 2, specifying the differentiability of γ at $t = 0$ and $t = 1$. The function $\gamma(t)$ specified in (4.13) is the only featured case for which the contribution from the score is non-vanishing in the velocity b at the endpoints. In Section 7, we illustrate on numerical examples that

35

$\gamma(t):$	$\sqrt{at(1-t)}$	$t(1-t)$	$\hat{\sigma}(t)$	$\sin^2(\pi t)$
C^1 at $t = 0, 1$	✗	✓	✓	✓

Table 2: **Differentiability of $\gamma(t)z$.** A characterization of the possible choices of $\gamma(t)$ with respect to their differentiability. The column specified by $\hat{\sigma}(t)$ is sum of sigmoid functions, made compact by the notation
$\hat{\sigma}(t) = \sigma(f(t - \frac{1}{2}) + 1) - \sigma(f(t - \frac{1}{2}) - 1) - \sigma(-\frac{f}{2} + 1) + \sigma(-\frac{f}{2} - 1)$, where $\sigma(t) = e^t/(1 + e^t)$
and f is a scaling factor.

there are tradeoffs between different choices of γ, which might be directly related to this fact. When using the ODE as a generative model, the score is only felt through b, whereas it is explicit when using the SDE as a generative model.

4.4 Spatially linear one-sided interpolants

Much of the discussion above generalizes to one-sided interpolants if we take the function $J(t, x_1)$ in (3.14) to be linear in x_1 and define

$$x_t^{\text{os,lin}} = \alpha(t)z + \beta(t)x_1, \qquad t \in [0, 1] \tag{4.15}$$

where $\alpha^2, \beta \in C^2([0, 1])$ and $\alpha(0) = \beta(1) = 1$, $\alpha(1) = \beta(0) = 0$, and $\alpha(t) > 0$ for all $t \in [0, 1)$. The velocity b and the score s defined in (2.10) and (2.14) can now be expressed as

$$b(t, x) = \dot{\alpha}(t)\eta_z^{\text{os}}(t, x) + \dot{\beta}(t)\eta_1^{\text{os}}(t, x), \qquad s(t, x) = -\alpha^{-1}(t)\eta_z^{\text{os}}(t, x), \tag{4.16}$$

where the second expression holds for all $t \in [0, 1)$ and we defined:

$$\eta_z^{\text{os}}(t, x) = \mathbb{E}(z|x_t^{\text{os,lin}} = x), \qquad \eta_1^{\text{os}}(t, x) = \mathbb{E}(x_1|x_t^{\text{os,lin}} = x). \tag{4.17}$$

Note that, by definition of the conditional expectation, η_z^{os} and η_1^{os} satisfy

$$\forall (t, x) \in [0, 1] \times \mathbb{R}^d \quad : \quad \alpha(t)\eta_z^{\text{os}}(t, x) + \beta(t)\eta_1^{\text{os}}(t, x) = x. \tag{4.18}$$

As a result, only one of them needs to be estimated. For example, we can express η_1^{os} as a function of η_z^{os} for all t such that $\beta(t) \neq 0$, and use the result to express the velocity (4.16) as

$$b(t, x) = \dot{\beta}(t)\beta^{-1}(t)x + \left(\dot{\alpha}(t) - \alpha(t)\dot{\beta}(t)\beta^{-1}(t)\right)\eta_z^{\text{os}}(t, x) \qquad \forall t \; : \; \beta(t) \neq 0. \tag{4.19}$$

Assuming that $\beta(t) \neq 0$ for all $t \in (0, 1]$, this formula only needs to be supplemented at $t = 0$ with

$$b(0, x) = \dot{\alpha}(0)x + \dot{\beta}(0)\mathbb{E}[x_1] \tag{4.20}$$

which follows from (4.16) since $x_{t=0}^{\text{os,lin}} = z$. Later in Section 5.2 we will show that using the velocity b in (4.19) to solve the probability flow ODE (2.32) can be seen as using a denoiser to construct a generative model.

Finally note that η_z and/or η_1 can be estimated using the following two objective functions, respectively:

$$\mathcal{L}_{\eta_z}(\hat{\eta}_z^{\mathsf{os}}) = \int_0^1 \mathbb{E}\left[\tfrac{1}{2}|\hat{\eta}_z^{\mathsf{os}}(t, x_t^{\mathsf{os,lin}})|^2 - z \cdot \hat{\eta}_z^{\mathsf{os}}(t, x_t^{\mathsf{os,lin}})\right] dt,$$

$$\mathcal{L}_{\eta_1}(\hat{\eta}_1^{\mathsf{os}}) = \int_0^1 \mathbb{E}\left[\tfrac{1}{2}|\hat{\eta}_1^{\mathsf{os}}(t, x_t^{\mathsf{os,lin}})|^2 - x_1 \cdot \hat{\eta}_1^{\mathsf{os}}(t, x_t^{\mathsf{os,lin}})\right] dt.$$

(4.21)

5. Connections with other methods

In this section, we discuss connections between the stochastic interpolant framework and the score-based diffusion method (Song et al., 2021b), the stochastic localization framework (Eldan, 2013; El Alaoui and Montanari, 2022; Montanari, 2023)), denoising methods (Simoncelli and Adelson, 1996; Hyvärinen, 1999; Kadkhodaie and Simoncelli, 2021; Ho et al., 2020), and the rectified flow method (Liu et al., 2023b).

5.1 Score-based diffusion models and stochastic localization

Score-based diffusion models (SBDM) are based on variants of the Ornstein-Uhlenbeck process

$$dZ_\tau = -Z_\tau dt + \sqrt{2}dW_\tau, \qquad Z_{\tau=0} \sim \rho_1, \tag{5.1}$$

which has the property that the marginal density of its solution at time τ converges to a standard normal as τ tends towards infinity. By learning the score of the density of Z_τ, we can write the associated backward SDE for (5.1), which can then be used as a generative model – this backwards SDE is also the one that is used in the stochastic localization process, see Montanari (2023).

To see the connection with stochastic interpolants, notice that the solution of (5.1) from the initial condition $Z_{\tau=0} = x_1 \sim \rho_1$ can be written exactly as

$$Z_\tau = x_1 e^{-\tau} + \sqrt{2} \int_0^\tau e^{-\tau+s} dW_s. \tag{5.2}$$

As a result, the law of Z_τ conditioned on $Z_{\tau=0} = x_1$ is given by

$$Z_\tau \sim \mathsf{N}(x_1 e^{-\tau}, (1 - e^{-2\tau})Id), \tag{5.3}$$

for any time $\tau \in [0, \infty)$. This is also the law of the process

$$y_\tau = x_1 e^{-\tau} + \sqrt{1 - e^{-2\tau}}\, z, \qquad z \sim \mathsf{N}(0, Id), \qquad \tau \in [0, \infty). \tag{5.4}$$

If we let $x_1 \sim \rho_1$ with $x_1 \perp z$, the process y_τ is similar to a one-sided stochastic interpolant, except the density of y_τ only converges to $\mathsf{N}(0, Id)$ as $\tau \to \infty$; by contrast, the one-sided interpolants we introduced in Section 3.2 converge on the finite interval $[0, 1]$. In SBDM, this is handled by capping the evolution of Z_τ to a finite time interval $[0, T]$ with $T < \infty$, and then by using the backward SDE associated with (5.1) restricted to $[0, T]$. However, this introduces a bias that is not present with one-sided stochastic interpolants, because the final condition used for the backwards SDE in SBDM is drawn from $\mathsf{N}(0, Id)$ even though the density of the process (5.1) is not Gaussian at time T.

We can, however, turn (5.4) into a one-sided linear stochastic interpolant by defining $t = e^{-\tau}$ and by choosing $\alpha(t)$ and $\beta(t)$ in (4.15) to have a specific form. More precisely, evaluating (5.4) at $\tau = -\log t$,

$$y_{\tau=-\log t} = \sqrt{1 - t^2}z + tx_1 \equiv x_t^{\text{os,lin}} \quad \text{for} \quad \alpha(t) = \sqrt{1 - t^2}, \quad \beta(t) = t. \tag{5.5}$$

With this choice of $\alpha(t)$ and $\beta(t)$, from (4.16) we get the velocity field

$$b(t, x) = -\frac{t}{\sqrt{1 - t^2}}\eta_z^{\text{os}}(t, x) + \eta_1^{\text{os}}(t, x) \equiv ts(t, x) + \eta_1^{\text{os}}(t, x) \tag{5.6}$$

where η_z^{os} and η_1^{os} are defined in (4.17). This expression shows that the velocity b used in the probability flow ODE (2.32) is well-behaved at all times, including at $t = 1$ where $\dot{\alpha}(t)$ is singular. The same is true for the drift $b_\mathsf{F}(t, x) = b(t, x) + \epsilon(t)s(t, x)$ used in the forward SDE (2.33), regardless of the choice of $\epsilon \in C^0([0, 1])$ with $\epsilon(t) \geq 0$. This shows that casting SBDM into a one-sided linear stochastic interpolant (4.15) allows the construction of *unbiased* generative models that operate on $t \in [0, 1]$. This comes at no extra computational cost, since only one of the two functions defined in (4.17) needs to be estimated, which is akin to estimating the score in SBDM.

It is worth comparing the above procedure to an equivalent change of time at the level of the diffusion process (5.1), which we now show leads to singular terms that pose numerical and analytical difficulties. Indeed, if we define $Z_t^\mathsf{B} = Z_{\tau=-\log t}$, from (5.1) we obtain

$$dZ_t^\mathsf{B} = t^{-1}Z_t^\mathsf{B}dt + \sqrt{2t^{-1}}dW_t^\mathsf{B}, \quad Z_{t=1}^\mathsf{B} \sim \rho_1, \tag{5.7}$$

to be solved backwards in time. Because of the factor t^{-1}, this SDE cannot easily be solved until $t = 0$, which corresponds to $\tau = \infty$ in the original (5.1). For the same reason, the forward SDE associated with (5.7)

$$dZ_t^\mathsf{F} = t^{-1}Z_t^\mathsf{F}dt + 2t^{-1}s(t, Z_t^\mathsf{F})dt + \sqrt{2t^{-1}}dW_t, \tag{5.8}$$

cannot be solved from $t = 0$, where formally $Z_{t=0}^\mathsf{F} \overset{d}{=} Z_{t=0}^\mathsf{B} = Z_{\tau=\infty} \sim \mathsf{N}(0, Id)$. This means it cannot be used as a generative model unless we start from some $t > 0$, which introduces a bias. Importantly, this problem does not arise with the stochastic interpolant framework, because the construction of the density $\rho(t)$ connecting ρ_0 and ρ_1 is handled separately from the construction of the process that generates samples from $\rho(t)$. By contrast, SBDM combines these two operations into one, leading to the singularity at $t = 0$ in the coefficients in (5.7) and (5.8).

Remark 43 *To emphasize the last point made above, we stress that there is no contradiction between having a singular drift and diffusion coefficient in (5.8), and being able to write a nonsingular SDE with stochastic interpolants. To see why, notice that the stochastic interpolant tells us that we can change the diffusion coefficient in (5.8) to any nonsingular $\epsilon \in C^0([0, 1])$ with $\epsilon(t) \geq 0$ and replace this SDE with*

$$dX_t^\mathsf{F} = t^{-1}Z_t^\mathsf{F}dt + (t^{-1} + \epsilon(t))s(t, X_t^\mathsf{F})dt + \sqrt{2\epsilon(t)}dW_t, \tag{5.9}$$

38

This SDE has the property that $X^{\mathsf{F}}_{t=1} \sim \rho_1$ if $X^{\mathsf{F}}_{t=0} \sim \rho_0$, and its drift is also nonsingular at $t = 0$ and given precisely by (5.6). Indeed, using the constraint (4.18), which here reads $x = \sqrt{1-t^2}\eta^{\text{os}}_z(t,x) + t\eta^{\text{os}}_1(t,x) \equiv -(1-t^2)s(t,x) + t\eta^{\text{os}}_1(t,x)$, it is easy to see that

$$t^{-1}x + t^{-1}s(t,x) = ts(t,x) + \eta^{\text{os}}_1(t,x), \tag{5.10}$$

which is nonsingular at $t = 0$.

5.2 Denoising methods

Consider the spatially-linear one-sided stochastic interpolant defined in (4.15). By solving this equation for x_1, we obtain

$$x_1 = \beta^{-1}(t)\left(x^{\text{os,lin}}_t - \alpha(t)z\right) \qquad t \in (0, 1]. \tag{5.11}$$

Taking a conditional expectation at fixed $x^{\text{os,lin}}_t$ and using (4.17) implies that

$$\mathbb{E}\big(x_1|x^{\text{os,lin}}_t\big) = \eta^{\text{os}}_1(t, x^{\text{os,lin}}_t) = \beta^{-1}(t)\left(x^{\text{os,lin}}_t - \alpha(t)\eta^{\text{os}}_z(t, x^{\text{os,lin}}_t)\right) \qquad t \in (0, 1] \tag{5.12}$$

while trivially $\mathbb{E}(x_1|x^{\text{os,lin}}_{t=0}) = \mathbb{E}[x_1]$ since $x^{\text{os,lin}}_{t=0} = z$. This expression is commonly used in denoising methods (Simoncelli and Adelson, 1996; Kadkhodaie and Simoncelli, 2021), and it is Stein's unbiased risk estimator (SURE) for x_1 given the noisy information in $x^{\text{os,lin}}_t$ Stein (1981). Rather than considering the conditional expectation of x_1, we can consider an analogous quantity for $x^{\text{os,lin}}_s$ for any $s \in [0, 1]$; this leads to the following result.

Lemma 44 (SURE) *For $s \in [0, 1]$, we have*

$$\mathbb{E}(x^{\text{os,lin}}_s|x^{\text{os,lin}}_t) = \frac{\beta(s)}{\beta(t)}x^{\text{os,lin}}_t + \left(\alpha(s) - \frac{\alpha(t)\beta(s)}{\beta(t)}\right)\eta^{\text{os}}_z(t, x^{\text{os,lin}}_t) \qquad t \in (0, 1] \tag{5.13}$$

and $\mathbb{E}(x^{\text{os,lin}}_s|x^{\text{os,lin}}_{t=0}) = \alpha(s)x^{\text{os,lin}}_{t=0} + \beta(s)\mathbb{E}[x_1]$.

Proof (5.13) follows from inserting (5.11) in the expression for $x^{\text{os,lin}}_s$ and taking the conditional expectation using the definition of η^{os}_1 and η^{os}_z in (4.17). $\mathbb{E}(x^{\text{os,lin}}_s|x^{\text{os,lin}}_{t=0}) = \alpha(s)x^{\text{os,lin}}_{t=0} + \beta(s)\mathbb{E}[x_1]$ follows from $x^{\text{os,lin}}_{t=0} = z$ together with (5.11). $\blacksquare$

At this stage, equations (5.11) and (5.13) cannot be used as generative models: the random variable $\mathbb{E}(x_1|x^{\text{os,lin}}_t)$ is not a sample of ρ_1, and the random variable $\mathbb{E}(x^{\text{os,lin}}_s|x^{\text{os,lin}}_t)$ is not a sample from $\rho(s)$, the density of $x^{\text{os,lin}}_s$. However, the following result shows that if we iterate upon formula (5.13) by taking infinitesimal steps, we obtain a generative model consistent with the probability flow equation (2.32) associated with $x^{\text{os,lin}}_t$.

Theorem 45 *Let $t_j = j/N$ with $j \in \{1, \ldots, N\}$, set $X^{\text{den}}_1 = z$, and define for $j = 1, \ldots, N-1$,*

$$X^{\text{den}}_{j+1} = \frac{\beta(t_{j+1})}{\beta(t_j)}X^{\text{den}}_j + \left(\alpha(t_{j+1}) - \frac{\alpha(t_j)\beta(t_{j+1})}{\beta(t_j)}\right)\eta^{\text{os}}_z(t_j, X^{\text{den}}_j). \tag{5.14}$$

Then, (5.14) is a consistent integration scheme for the probability flow equation (2.32) associated with the velocity field (4.16) expressed as in (4.19). That is, if $N, j \to \infty$ with $j/N \to t \in [0, 1]$, then $X_j^{\mathsf{den}} \to X_t$ where

$$\dot{X}_t = b(t, X_t) = \frac{\dot{\beta}(t)}{\beta(t)} X_t + \left(\dot{\alpha}(t) - \frac{\alpha(t)\dot{\beta}(t)}{\beta(t)} \right) \eta_z^{\mathsf{os}}(t, X_t), \quad X_{t=0} = z. \tag{5.15}$$

In particular, if $z \sim \mathsf{N}(0, Id)$, then $X_N^{\mathsf{den}} \to x_1 \sim \rho_1$ in this limit.

The proof of this theorem is given in Appendix B.7, and proceeds by Taylor expansion of the right-hand side of (5.14).

5.3 Rectified flows

We now discuss how stochastic interpolants can be *rectified* according to the procedure proposed in Liu et al. (2023b). Suppose that we have perfectly learned the velocity field b in the probability flow equation (2.32) for a given stochastic interpolant. Denote by $X_t(x)$ the solution to this ODE with the initial condition $X_{t=0}(x) = x$, i.e.

$$\frac{d}{dt} X_t(x) = b(t, X_t(x)), \qquad X_{t=0}(x) = x. \tag{5.16}$$

We can use the map $X_{t=1} : \mathbb{R}^d \to \mathbb{R}^d$ to define a new stochastic interpolant

$$x_t^{\mathsf{rec}} = \alpha(t)x_0 + \beta(t)X_{t=1}(x_0), \tag{5.17}$$

where $\alpha^2, \beta \in C^2([0, 1])$ satisfy $\alpha(0) = \beta(1) = 1$, $\alpha(1) = \beta(0)w = 0$, and $\alpha(t) > 0$ for all $t \in [0, 1)$. Clearly, we then have $x_{t=0}^{\mathsf{rec}} = x_0 \sim \rho_0$ since $X_{t=0}(x_0) = x_0$ and $x_{t=1}^{\mathsf{rec}} = X_{t=1}(x_0) \sim \rho_1$ by definition of the probability flow equation. We can define a new probability flow equation associated with the velocity field

$$b^{\mathsf{rec}}(t, x) = \mathbb{E}[\dot{x}_t^{\mathsf{rec}} | x_t^{\mathsf{rec}} = x] = \dot{\alpha}(t)\mathbb{E}[x_0 | x_t^{\mathsf{rec}} = x] + \dot{\beta}(t)\mathbb{E}[X_{t=1}(x_0) | x_t^{\mathsf{rec}} = x]. \tag{5.18}$$

It is easy to see that this velocity field is amenable to estimation, since it is the unique minimizer of

$$\mathcal{L}_{b^{\mathsf{rec}}}[\hat{b}^{\mathsf{rec}}] = \int_0^1 \mathbb{E}\left[\frac{1}{2} |\hat{b}^{\mathsf{rec}}(t, x_t^{\mathsf{rec}})|^2 - (\dot{\alpha}(t)x_0 + \dot{\beta}(t)X_{t=1}(x_0)) \cdot \hat{b}^{\mathsf{rec}}(t, x_t^{\mathsf{rec}}) \right] dt, \tag{5.19}$$

where x_t^{rec} is given in (5.17) and the expectation is now only on $x_0 \sim \rho_0$. Our next result show that the probability flow equation associated with the velocity field (5.18) has straight line solutions, but ultimately it leads to a generative model that is identical to the one based on (5.16). To phrase this result, we first make an assumption on the invertibility of x_t^{rec}.

Assumption 46 *The map $x \to M(t, x)$ where $M(t, x) = \alpha(t)x + \beta(t)X_{t=1}(x)$ with $X_t(x)$ solution to (5.16) is invertible for all $(t, x) \in [0, 1] \times \mathbb{R}^d$, i.e. $\exists N(t, \cdot,) : \mathbb{R}^d \to \mathbb{R}^d$ such that*

$$\forall (t, x) \in [0, 1] \times \mathbb{R}^d \quad : \quad N(t, M(t, x)) = M(t, N(t, x)) = x. \tag{5.20}$$

This is equivalent to requiring that the determinant of the Jacobian of $M(t, x)$ is nonzero for all $(t, x) \in [0, 1] \times \mathbb{R}^d$; put differently, Assumption 46 requires that the Jacobian of $X_{t=1}(x)$ never has eigenvalues precisely equal to $-\alpha(t)/\beta(t)$, which is generic. Under this assumption, we state the following theorem.

Theorem 47 *Consider the probability flow equation associated with* (5.18),

$$\frac{d}{dt} X_t^{\mathsf{rec}}(x) = b^{\mathsf{rec}}(t, X_t^{\mathsf{rec}}(x)), \qquad X_{t=0}^{\mathsf{rec}}(x) = x. \tag{5.21}$$

Then, all solutions are such that $X_{t=1}^{\mathsf{rec}}(x_0) \sim \rho_1$ if $x_0 \sim \rho_0$. In addition, if Assumption (46) *holds, the velocity field defined in* (5.18) *reduces to*

$$b^{\mathsf{rec}}(t, x) = \dot{\alpha}(t) N(t, x) + \dot{\beta}(t) X_{t=1}(N(t, x)) \tag{5.22}$$

and the solution to the probability flow ODE (5.21) *is simply*

$$X_t^{\mathsf{rec}}(x) = \alpha(t)x + \beta(t) X_{t=1}(x). \tag{5.23}$$

The proof is given in Appendix B.8.

Theorem 47 implies that $X_t^{\mathsf{rec}}(x)$ is a simpler flow than $X_t(x)$, but we stress that they give the *same* map, $X_{t=1}^{\mathsf{rec}} = X_{t=1}$. In particular, $X_t^{\mathsf{rec}}(x)$ reduces to a straight line between x and $X_{t=1}(x)$ for $\alpha(t) = 1 - t$ and $\beta(t) = t$. We also note that the approach can be used to learn a single-step map, since (5.22) and $N(t = 0, x) = x$ give

$$b^{\mathsf{rec}}(t = 0, x) = \dot{\alpha}(0)x + \dot{\beta}(0) X_{t=1}(x), \tag{5.24}$$

which expresses $X_{t=1}(x)$ in terms of known quantities as long as $\dot{\beta}(0) \neq 0$. For example, if $\dot{\alpha}(0) = 0$ and $\dot{\beta}(0) = 1$, we obtain $b^{\mathsf{rec}}(t = 0, x) = X_{t=1}(x)$.

Remark 48 (Optimal transport) *The discussion above highlights the fact that a probability flow equation can have straight line solutions and lead to a map that exactly pushes ρ_0 onto ρ_1 but is not the optimal transport map. That is, straight line solutions is a necessary condition for optimal transport, but it is not sufficient.*

Remark 49 (Gradient fields) *The map is unaffected by the rectification procedure because we do not impose that the velocity $b^{\mathsf{rec}}(t, x)$ be a gradient field. If we do impose this structure by setting $b^{\mathsf{rec}}(t, x) = \nabla \phi(t, x)$ for some $\phi : \mathbb{R}^d \to \mathbb{R}$, then $X_{t=1}^{\mathsf{rec}} \neq X_{t=1}$. As shown in Liu (2022), iterating over this procedure eventually gives the optimal transport map. That is, implemented over gradient fields and iterated infinitely, rectification computes Brenier's polar decomposition of the map (Brenier, 1991).*

Remark 50 (Consistency models) *Recent work has introduced the notion of consistency models (Song et al., 2023), which distill a velocity field learned via score-based diffusion into a single-step map. Section 5.1 and the previous discussion provide an alternative perspective on consistency models, and show how they may be computed in the framework of stochastic interpolants via rectification.*

6. Algorithmic aspects

The methods described in the previous sections have efficient numerical realizations. Here, we detail algorithms and practical recommendations for an implementation. These suggestions can be split into two complementary tasks: learning the drift coefficients, and sampling with an ODE or an SDE.

6.1 Learning

As described in Section 2.2, there are a variety of algorithmic choices that can be made when learning the drift coefficients in (2.9), (2.20), and (2.22). While all choices lead to exact generative models in the absence of numerical and statistical errors, in practice, the presence of these errors ensures that different choices lead to different generative models, some of which may perform better for specific applications. Here, we describe the various realizations explicitly.

Deterministic generative modeling: Learning b versus learning v and s. Recall from Section 2.2 that the drift b of the transport equation (2.9) can be written as $b(t, x) = v(t, x) - \gamma(t)\dot{\gamma}(t)s(t, x)$. This raises the practical question of whether it would be better to learn an estimate $\hat{b}$ of b by minimizing the empirical risk

$$\hat{\mathcal{L}}_b(\hat{b}) = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{2} |\hat{b}(t_i, x_{t_i}^i)|^2 - \hat{b}(t_i, x_{t_i}^i) \cdot \left(\partial_t I(t_i, x_0^i, x_1^i) + \dot{\gamma}(t_i)z^i \right) \right), \qquad (6.1)$$

or to learn estimates of $\hat{v}$ and $\hat{s}$ by minimizing the empirical risks

$$\hat{\mathcal{L}}_v(\hat{v}) = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{2} |\hat{v}(t_i, x_{t_i}^i)|^2 - \hat{v}(t_i, x_{t_i}^i) \cdot \partial_t I(t_i, x_0^i, x_1^i) \right) \qquad (6.2)$$

and

$$\hat{\mathcal{L}}_s(\hat{s}) = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{2} |\hat{s}(t_i, x_{t_i}^i)|^2 + \gamma(t_i)^{-1}\hat{s}(t_i, x_{t_i}^i) \cdot z^i \right) \qquad (6.3)$$

and construct the estimate $\hat{b}(t, x) = \hat{v}(t, x) - \gamma(t)\dot{\gamma}(t)\hat{s}(t, x)$. Above, $x_{t_i}^i = I(t_i, x_0^i, x_1^i) + \gamma(t_i)z^i$, and N denotes the number of samples t^i, x_0^i, and x_1^i. If the practitioner is interested in a deterministic generative model (for example, to exploit adaptive integration or exact likelihood computation), learning the estimate $\hat{b}$ directly only requires learning a single model, and hence will typically lead to greater efficiency. This recommendation is captured in Algorithm 1. If both stochastic and deterministic generative models are of interest, it is necessary to learn two models for most choices of the interpolant; we discuss more suggestions for stochastic case below.

Antithetic sampling and capping. In practice, the losses for b (2.13) and s (2.16) can become high-variance near the endpoints $t = 0$ and $t = 1$ due to the presence of the singular term $1/\gamma(t)$ and the (potentially) singular term $\dot{\gamma}(t)$. This issue can be eliminated by using antithetic sampling, which we found necessary for stable training of objectives involving $\gamma^{-1}(t)$. To show why, we consider the loss (2.16) for s, but an analogous calculation can be

---

Algorithm 1: Learning b with arbitrary ρ_0 and ρ_1.

Input: Batch size N, interpolant function $I(t, x_0, x_1)$, coupling $\nu(dx_0, dx_1)$ to
  sample x_0, x_1, noise function $\gamma(t)$, initial parameters θ_b, gradient-based optimization
  algorithm to minimize $\mathcal{L}_b$ defined in (2.13), number of gradient steps N_g.
Returns: An estimate $\hat{b}$ of b.

for $j = 1, \ldots, N_g$ **do**

> Draw N samples $(t_i, x_0^i, x_1^i, z^i) \sim \mathsf{Unif}([0, 1]) \times \nu \times \mathsf{N}(0, 1)$, for $i = 1, \ldots, N$.
>
> Construct samples $x_t^i = I(t_i, x_0^i, x_1^i) + \gamma(t_i)z^i$, for $i = 1, \ldots, N$.
>
> Take gradient step with respect to $\mathcal{L}_b\left(\theta_b, \{x_t^i\}_{i=1}^N\right)$.

Return: $\hat{b}$.

---

performed for the loss (2.13) for b or (2.19) for η_z (even though it is not necessary for this
last quantity). We first observe that, by definition of x_t and by Taylor expansion, as $t \to 0$
or $t \to 1$

$$
\begin{aligned}
\frac{1}{\gamma(t)} z \cdot s(t, x_t) &= \frac{1}{\gamma(t)} z \cdot s(t, I(t, x_0, x_1) + \gamma(t)z), \\
&= \frac{1}{\gamma(t)} z \cdot \left(s(t, I(t, x_0, x_1)) + \gamma(t)\nabla s(t, I(t, x_0, x_1))z + o(\gamma(t))\right), \qquad (6.4) \\
&= \frac{1}{\gamma(t)} z \cdot s(t, I(t, x_0, x_1)) + z \cdot \nabla s(t, I(t, x_0, x_1))z + o(1).
\end{aligned}
$$

Even though the conditional mean of the first term at the right-hand side is finite in the
limit as $t \to 0$ or $t \to 1$, its variance diverges. By contrast, let $x_t^+ = I(t, x_0, x_1) + \gamma(t)z$ and
$x_t^- = I(t, x_0, x_1) - \gamma(t)z$ with x_0, x_1, and z fixed. Then,

$$
\begin{aligned}
&\frac{1}{2\gamma(t)}\left(z \cdot s(t, x_t^+) - z \cdot s(t, x_t^-)\right) \\
&= \frac{1}{2\gamma(t)}\left(z \cdot s(t, I(t, x_0, x_1) + \gamma(t)z)) - z \cdot s(t, I(t, x_0, x_1) - \gamma(t)z)\right), \\
&= \frac{1}{2\gamma(t)} z \cdot \left(s(t, I(t, x_0, x_1)) + \gamma(t)\nabla s(t, I(t, x_0, x_1))z + o(\gamma(t))\right) \qquad (6.5) \\
&\quad - \frac{1}{2\gamma(t)} z \cdot \left(s(t, I(t, x_0, x_1)) - \gamma(t)\nabla s(t, I(t, x_0, x_1))z + o(\gamma(t))\right), \\
&= z \cdot \nabla s(t, I(t, x_0, x_1))z + o(1),
\end{aligned}
$$

so that both the conditional mean and variance are finite in the limit as $t \to 0$ or $t \to 1$
despite the singularity of $1/\gamma(t)$. In practice, this can be implemented by using x_t^+ and x_t^-
for every draw of x_0, x_1, and z in the empirical discretization of the population loss.

Learning the score s versus learning a denoiser η_z. When learning s, an alternative
to antithetic sampling is to consider learning the denoiser η_z defined in (2.17), which is
related to the score by a factor of γ. Note that the objective function for the denoiser in
(2.19) is well behaved for all $t \in [0, 1]$, and can be thought of as a generalization of the

---

Algorithm 2: Learning η_z with arbitrary ρ_0 and ρ_1.

Input: Batch size N, interpolant function $I(t, x_0, x_1)$, a coupling $\nu(dx_0, dx_1)$ to sample x_0, x_1, noise function $\gamma(t)$, initial parameters θ_{η_z}, gradient-based optimization algorithm to minimize $\mathcal{L}_{\eta_z}$ defined in (2.19), number of gradient steps N_g.

Returns: An estimate $\hat{\eta}_z$ of η_z.

for $j = 1, \ldots, N_g$ **do**

    Draw N samples $(t_i, x_0^i, x_1^i, z^i) \sim \mathsf{Unif}([0,1]) \times \nu \times \mathsf{N}(0,1)$, for $i = 1, \ldots, N$.

    Construct samples $x_t^i = I(t_i, x_0^i, x_1^i) + \gamma(t_i)z^i$, for $i = 1, \ldots, N$.

    Take gradient step with respect to $\mathcal{L}_{\eta_z}\left(\theta_{\eta_z}, \{x_t^i\}_{i=1}^N\right)$.

Return: $\hat{\eta}_z$.

Algorithm 3: Learning η_z^{os} with Gaussian ρ_0.

Input: Batch size N, interpolant function $x_t^{\mathsf{os,lin}}$, a coupling $\nu(dx_0, dx_1)$ to sample z, x_1, initial parameters $\theta_{\eta_z^{\mathsf{os}}}$, gradient-based optimization algorithm to minimize $\mathcal{L}_{\eta_z^{\mathsf{os}}}$ defined in (2.19), number of gradient steps N_g.

Returns: An estimate $\hat{\eta}_z^{\mathsf{os}}$ of η_z^{os}.

for $j = 1, \ldots, N_g$ **do**

    Draw N samples $(t_i, z^i, x_1^i) \sim \mathsf{Unif}([0,1]) \times \nu$, for $i = 1, \ldots, N$.

    Construct samples $x_t^i = \alpha(t_i)z + \beta(t_i)x_1^i$, for $i = 1, \ldots, N$.

    Take gradient step w.r.t $\mathcal{L}_{\eta_z^{\mathsf{os}}}\left(\theta_{\eta_z^{\mathsf{os}}}, \{x_t^i\}_{i=1}^N\right)$.

Return: $\hat{\eta}_z^{\mathsf{os}}$.

DDPM loss introduced in Ho et al. (2020). The empirical risk associated with this loss reads

$$\hat{\mathcal{L}}_{\eta_z}(\hat{\eta}_z) = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{2}|\hat{\eta}_z(t_i, x_{t_i}^i)|^2 - \hat{\eta}_z(t_i, x_{t_i}^i) \cdot z^i \right) \tag{6.6}$$

A detailed procedure for learning the denoiser η_z, e.g. for its use in an SDE-based generative model is given in Algorithm 2. For the case of one-sided spatially-linear interpolants, the procedure becomes particularly simple, which is highlighted in Algorithm 3.

6.2 Sampling

We now discuss several practical aspects of sampling generative models based on stochastic interpolants. These are intimately related to the choice of objects that are learned, as well as to the specific interpolant used to build a path between ρ_0 and ρ_1. A general algorithm for sampling models built on either ordinary or stochastic differential equations is presented in Algorithm 4.

Using the denoiser η_z instead of the score s. We remarked in Section 6.1 that learning the denoiser η_z is more numerically stable than learning the score s directly. We note that

---

Algorithm 4: Sampling general stochastic interpolants.

Input: Number of samples n, timestep Δt, drift estimates $\hat{b}$ and $\hat{\eta}_z$, initial time t_0, final time t_f, noise function $\gamma(t)$, diffusion coefficient $\epsilon(t)$, SDE or ODE timestepper `TakeStep`.

Returns: $\{\hat{x}_1^{(i)}\}_{i=1}^n$, a batch of model samples.

Initialize

    Set time $t = t_0$.

    Draw initial conditions $\hat{x}_{t_0}^{(i)} \sim \rho_0$ for $i = 1, \ldots, n$.

    Construct $\hat{s}(t, x) = -\hat{\eta}_z(t, x)/\gamma(t)$.

    Construct $\hat{b}_{\mathsf{F}}(t, x) = \hat{b}(t, x) + \epsilon(t)\hat{s}(t, x)$. `// Reduces to` $\hat{b}$ `for` $\epsilon(t) = 0$ `(ODE).`

while $t < t_f$ **do**

    Propagate $\hat{x}_{t+\Delta t}^{(i)} = \texttt{TakeStep}(t, \hat{x}_t^{(i)}, b_{\mathsf{F}}, \epsilon, \Delta t)$ for $i = 1, \ldots, n$. `// ODE or SDE`
        `integrator.`

    Update $t = t + \Delta t$.

Return: $\{\hat{x}^{(i)}\}_{i=1}^n$.

Algorithm 5: Sampling spatially-linear one-sided interpolants with Gaussian ρ_0.

Input: Number of samples n, timestep Δt, denoiser estimate $\hat{\eta}_z$, initial time t_0, final time t_f, noise function $\gamma(t)$, diffusion coefficient $\epsilon(t)$, interpolant functions $\alpha(t)$ and $\beta(t)$, SDE or ODE timestepper `TakeStep`.

Returns: $\{\hat{x}_1^{(i)}\}_{i=1}^n$, a batch of model samples.

Initialize

    Set time $t = t_0$.

    Draw initial conditions $\hat{x}_{t_0}^{(i)} \sim \rho_0$ for $i = 1, \ldots, n$.

    Construct $\hat{s}(t, x) = -\hat{\eta}_z(t, x)/\alpha(t)$.

    Construct $\hat{b}(t, x) = \dot{\alpha}(t)\hat{\eta}_z^{\mathsf{os}}(t, x) + \frac{\dot{\beta}(t)}{\beta(t)}(x - \alpha(t)\hat{\eta}_z^{\mathsf{os}}(t, x))$.

    Construct $\hat{b}_{\mathsf{F}}(t, x) = \hat{b}(t, x) + \epsilon(t)\hat{s}(t, x)$. `// Reduces to` $\hat{b}$ `for` $\epsilon(t) = 0$ `(ODE).`

while $t < t_f$ **do**

    Propagate $\hat{x}_{t+\Delta t}^{(i)} = \texttt{TakeStep}(t, \hat{x}_t^{(i)}, b_{\mathsf{F}}, \epsilon, \Delta t)$ for $i = 1, \ldots, n$. `// ODE or SDE`
        `integrator.`

    Update $t = t + \Delta t$.

Return: $\{\hat{x}^{(i)}\}_{i=1}^n$.

---

while the objective for η_z is well-behaved for all $t \in [0, 1]$, the resulting drifts can become singular at $t = 0$ and $t = 1$ when using $s(t, x) = -\eta_z(t, x)/\gamma(t)$. There are several ways to avoid this singularity in practice. One method is to choose a time-varying $\epsilon(t)$ that vanishes in a small interval around the endpoints $t = 0$ and $t = 1$, which avoids this numerical instability. An alternative option is to integrate the SDE up to a final time t_f with $t_f < 1$, and then to perform a step of denoising using (5.13). We use this approach in Section 7 below when sampling the SDE.

A denoiser is all you need for spatially-linear one-sided interpolants. As shown in (4.19), and as considered in Section 5.2, the denoiser η_z^{os} is sufficient to represent the velocity field b appearing in the probability flow equation (2.32).

Using this definition for b and the relationship $s(t, x) = -\eta_z(t, x)/\gamma(t)$, we state the following ordinary and stochastic differential equations for sampling

$$\text{ODE}: \quad \dot{X}_t = \dot{\alpha}(t)\eta_z^{\mathsf{os}}(t, X_t) + \frac{\dot{\beta}(t)}{\beta(t)}\big(X_t - \alpha(t)\eta_z^{\mathsf{os}}(t, X_t)\big)$$

$$\text{SDE}: \quad dX_t^{\mathsf{F}} = \big(\dot{\alpha}(t)\eta_z^{\mathsf{os}}(t, X_t^{\mathsf{F}}) + \frac{\dot{\beta}(t)}{\beta(t)}\big(X_t^{\mathsf{F}} - \alpha(t)\eta_z^{\mathsf{os}}(t, X_t^{\mathsf{F}})\big) - \frac{\epsilon(t)}{\alpha(t)}\eta_z^{\mathsf{os}}(t, X_t^{\mathsf{F}})\big)dt \qquad (6.7)$$

$$\qquad\qquad + \sqrt{2\epsilon(t)}dW_t.$$

Because $\beta(0) = 0$, the drift is numerically singular in both equations. However, $b(t = 0, x)$ has a finite limit

$$b(t = 0, x) = \dot{\alpha}(0)x + \dot{\beta}(0)\mathbb{E}[x_1], \qquad (6.8)$$

as originally given in (4.20). Equation (6.8) can be estimated using available data, which means that when learning a one-sided interpolant, ODE and SDE-based generative models can be defined exactly on the interval $t \in [0, 1]$ using only a score or a denoiser without singularity.

The factor of $\alpha(t)^{-1}$ in the final term of the SDE could pose numerical problems at $t = 1$, as $\alpha(1) = 0$. As discussed in the paragraph above, a choice of $\epsilon(t)$ which is such that $\epsilon(t)/\alpha(t) \to C$ for some constant C as $t \to 1$ avoids any issue.

An algorithm for sampling with only the denoiser η_z^{os} is given in Algorithm 5.

7. Numerical results

So far, we have been focused on the impact of α, β, and γ in (4.1) on the density $\rho(t)$, which we illustrated analytically. In this section, we study examples where the drift coefficients must be learned over parametric function classes. In particular, we explore numerically the tradeoffs between generative models based on ODEs and SDEs, as well as the various design choices introduced in Sections 3, 4, and 6. In Section 7.1, we consider simple two-dimensional distributions that can be visualized easily. In Section 7.2, we consider high-dimensional Gaussian mixtures, where we can compare our learned models to analytical solutions. Finally in Section 7.3 we perform some experiments in image generation.

7.1 Deterministic versus stochastic models: 2D

As shown in Section 2.2, the evolution of $\rho(t)$ can be captured exactly by either the transport equation (2.9) or by the forward and backward Fokker-Planck equations (2.20) and (2.22). These perspectives lead to generative models that are either based on the deterministic dynamics (2.32) or the forward and backward stochastic dynamics (2.33) and (2.34), where the level of stochasticity can be tuned by varying the diffusion coefficient $\epsilon(t)$. We showed in Section 2.4 that setting a constant $\epsilon(t) = \epsilon > 0$ can offer better control on the likelihood when using an imperfect velocity b and an imperfect score s. Moreover, the optimal choice of ϵ is determined by the relative accuracy of the estimates $\hat{b}$ and $\hat{s}$. Having laid out the
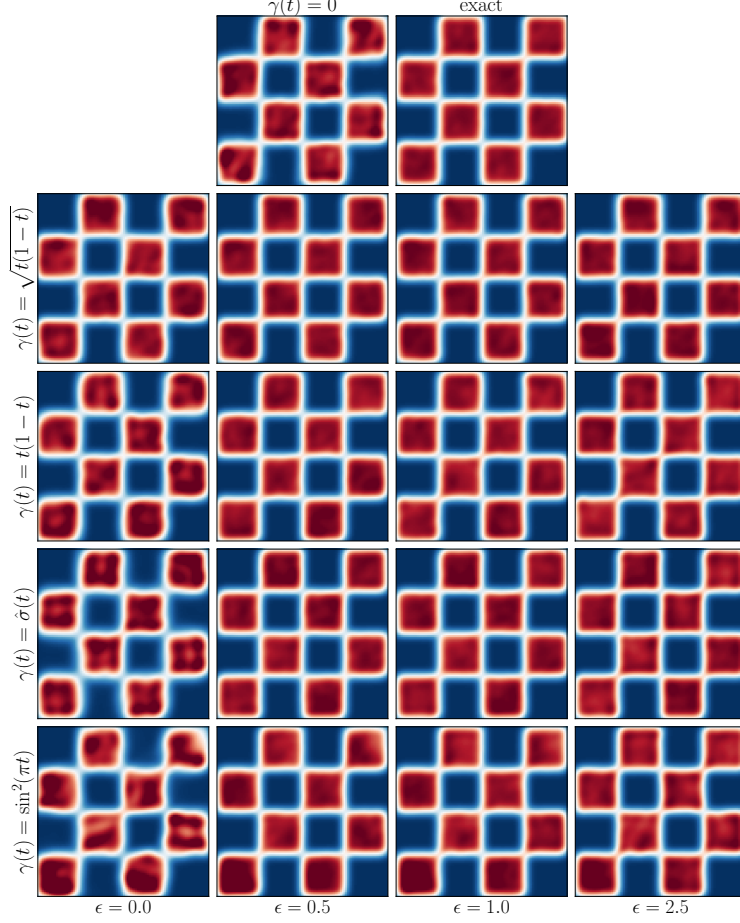
Figure 7: **The effects of $\gamma(t)$ and ϵ on sample quality: qualitative comparison.**
Kernel density estimates of $\hat{\rho}(1)$ for models with different choices of γ and ϵ. Sampling with $\epsilon = 0$ corresponds to using the probability flow with the learned drift $\hat{b} = \hat{v} - \gamma\dot{\gamma}\hat{s}$, whereas sampling with $\epsilon > 0$ corresponds to using the SDE with the learned drift $\hat{b}$ and score $\hat{s}$. We find empirically that SDE sampling is generically better than ODE sampling for this target density, though the gap is smallest for the probability flow specified with $\gamma(t) = \sqrt{t(1-t)}$, in agreement with the Remark 42 regarding the influence of γ on b at the endpoints. The SDE performs well at any noise level, though numerically integrating it for higher ϵ requires a smaller step size.

evolution of $\rho(t)$ for different choices of γ in the previous section, we now show how different values of ϵ can build these densities up from individual trajectories. The stochasticity intrinsic to the sampling process increases with ϵ, but by construction, the marginal density $\rho(t)$ for fixed α, β and γ is independent of ϵ.

The roles of $\gamma(t)$ and ϵ for 2D density estimation. To explore the roles of γ and ϵ, we consider a target density ρ_1 whose mass concentrates on a two-dimensional checkerboard and a base density $\rho_0 = \mathsf{N}(0, Id)$; here, the target was chosen to highlight the ability of the
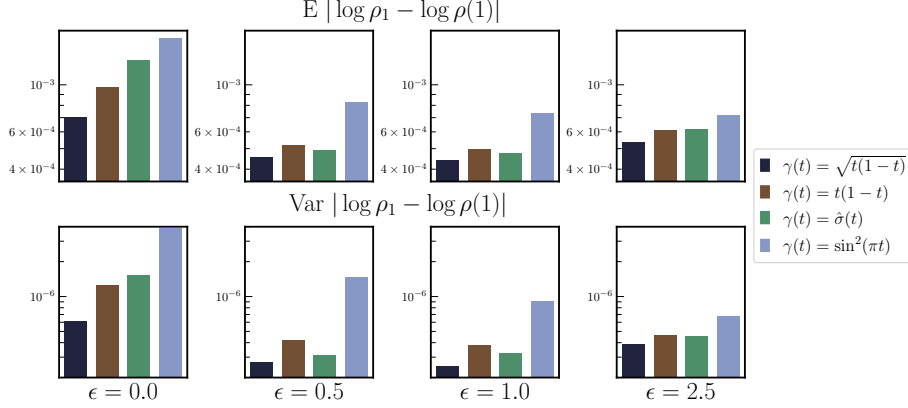
Figure 8: **The effects of $\gamma(t)$ and ϵ on sample quality: quantitative comparison.** For each γ and each ϵ specified in Figure 7, we compute the mean and variance of the absolute value of the difference of $\log \rho_1$ (exact) and $\log \hat{\rho}(1)$ (model). The model specified with $\gamma(t) = \sqrt{t(1-t)}$ is the best performing probability flow ($\epsilon = 0$). At large ϵ, SDE sampling with the same learned drift $\hat{b}$ and score $\hat{s}$ performs better, complementing the observations in the previous figure.

method to learn a challenging density with sharp boundaries. The same model architecture and training procedure was used to learn both v and s for several choices of γ given in Table 2. The feed-forward network was defined with 4 layers, each of size 512, and with the ReLU Nair and Hinton (2010) as an activation function.

After training, we draw 300,000 samples using either an ODE ($\epsilon = 0$) or an SDE with $\epsilon = 0.5$, $\epsilon = 1.0$, or $\epsilon = 2.5$. We compute kernel density estimates for each resulting density, which we compare to the exact density and to the original stochastic interpolant from Albergo and Vanden-Eijnden (2023) (obtained by setting $\gamma = 0$). Results are given in Figure 7 for each γ and each ϵ. Sampling with $\epsilon > 0$ empirically performs better, though the gap is smallest when using the γ specified in (4.13). Moreover, even when $\epsilon = 0$, using the probability flow with γ given by (4.13) performs better than the original interpolant from Albergo and Vanden-Eijnden (2023). Numerical comparisons of the mean and variance of the absolute value of the difference of $\log \rho_1$ (exact) from $\log \hat{\rho}(1)$ (model) for the various configurations are given in Figure 8, which corroborate the above observations.

7.2 Deterministic versus stochastic models: 128D Gaussian mixtures

We now study the performance of the stochastic interpolant method in the case where the target is a high-dimensional Gaussian mixture. Gaussian mixtures (GMs) are a convenient class of target distributions to study, because they can be made arbitrarily complex by increasing the number of modes, their separation, and the overall dimensionality. Moreover, by considering low-dimensional projections, we can compute quantitative error metrics such as the KL-divergence between the target and the model as a function of the (constant) diffusion coefficient ϵ. This enables us to quantify the tradeoffs of ODE and SDE-based samplers.

Experimental details. We consider the problem of mapping $\rho_0 = \mathsf{N}(0, Id)$ to a Gaussian mixture with five modes in dimension $d = 128$. The mean $m_i \in \mathbb{R}^d$ of each mode is drawn i.i.d. $m_i \sim \mathsf{N}(0, \sigma^2 Id)$ with $\sigma = 7.5$. To maximize performance at high ϵ, the timestep should be adapted to ϵ; here, we chose to use a fixed computational budget that performs well for moderate levels of ϵ to avoid computational effort that may become unreasonable in practice. Each covariance $C_i \in \mathbb{R}^{d \times d}$ is also drawn randomly with $C_i = \frac{1}{d} W_i^\mathsf{T} W_i + Id$ and $(W_i)_{kl} \sim \mathsf{N}(0,1)$ for $k, l = 1, \ldots, d$; this choice ensures that each covariance is a positive definite perturbation of the identity with diagonal entries that are $O(1)$ with respect to dimension d. For a *fixed* random draw of the means $\{m_i\}_{i=1}^5$ and covariances $\{C_i\}_{i=1}^5$, we study the four combinations of learning b or v and s or η_z to form a stochastic interpolant
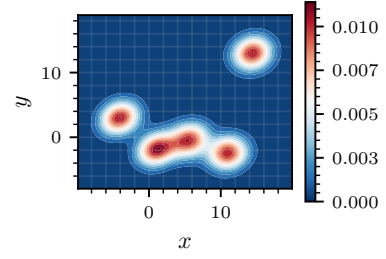

Figure 9: **Gaussian mixtures: target projection.** Low-dimensional marginal of the target density ρ_1 for the Gaussian mixture experiment, visualized via KDE.

from ρ_0 to ρ_1. In each case, we consider a linear interpolant with $\alpha(t) = 1 - t$, $\beta(t) = t$, and $\gamma(t) = \sqrt{t(1-t)}$. For visual reference, a projection of the target density ρ_1 onto the first two coordinates is depicted in Figure 9 – it contains significant multimodality, several modes that are difficult to distinguish, and one mode that is well-separated from the others, which requires nontrivial transport to resolve. In the following experiments, all samples were generated with the fourth-order Dormand-Prince adaptive ODE solver (`dopri5`) for $\epsilon = 0$ and by using one thousand timesteps of the Heun SDE integrator introduced in Karras et al. (2022) for $\epsilon \neq 0$. When learning η_z, to avoid singularity at $t = 0$ and $t = 1$ when dividing by $\gamma(t)$ in the formula $s(t, x) = -\eta(t, x)/\gamma(t)$, we set $t_0 = 10^{-4}$ and $t_f = 1 - t_0$ in Algorithm 4. For all other cases, we set $t_0 = 0$ and $t_f = 1$.

Quantitative metric To quantify performance, we make use of an error metric given by a KL-divergence between kernel density estimates (KDE) of low-dimensional marginals of ρ_1 and the model density $\hat{\rho}_1$; this error metric was chosen for computational tractability and interpretability. To compute it, we draw $50,000$ samples from ρ_1 and each $\hat{\rho}_1$. We obtain samples from the marginal density over the first two coordinates by projection, and then compute a Gaussian KDE with bandwidth parameter chosen by Scott's rule. We then draw a fresh set of $N_e = 50,000$ samples $\{x_i\}_{i=1}^{N_e}$ with each $x_i \sim \rho_1$ for evaluation. To compute the KL-divergence, we form a Monte-Carlo estimate with control variate

$$\mathsf{KL}(\rho_1 \parallel \hat{\rho}_1) \approx \frac{1}{N_e} \sum_{i=1}^{N_e} \left(\log \rho_1(x_i) - \log \hat{\rho}_1(x_i) - \left(\frac{\hat{\rho}_1(x_i)}{\rho_1(x_i)} - 1 \right) \right). \tag{7.1}$$

We found use of the control variate $\hat{\rho}_1/\rho_1 - 1$ helpful to reduce variance in the Monte-Carlo estimate; moreover, by concavity of the logarithm, use of the control variate ensures that the Monte-Carlo estimate cannot become negative.

Results. Figures 10 and 11 display two-dimensional projections (computed via KDE) of the model density error $\hat{\rho}_1 - \rho_1$ and the model density $\hat{\rho}_1$ itself, respectively, for different instantiations of Algorithms 1 and 2 and different choices of ϵ in Algorithm 4. Taken
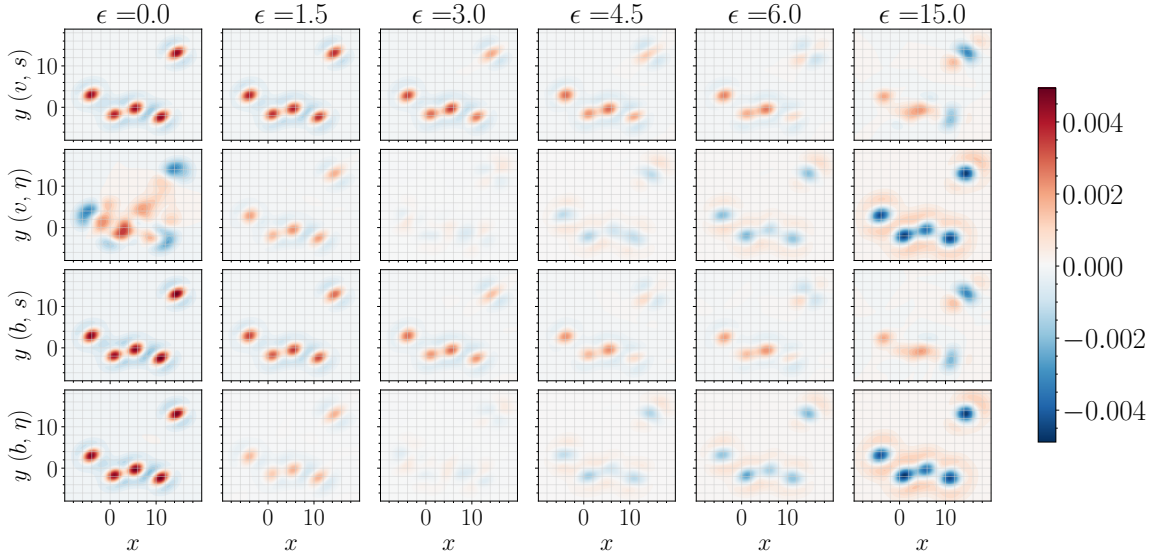
Figure 10: **Gaussian mixtures: density errors.** Errors $\hat{\rho}_1(x,y) - \rho_1(x,y)$ in the marginals over the first two coordinates for all four variations of learning b or v and s or η, computed via kernel density estimation. For small ϵ, the model densities tend to be overly-concentrated, and overestimate the density within the modes and underestimate the densities in the tails. As ϵ increases, the model becomes less concentrated and more accurately represents the target density. For ϵ too large, the model becomes overly spread out and under-estimates the density within the modes. *Note: visualizations show two-dimensional slices of a 128-dimensional density.*

together with Figure 9, these results demonstrate qualitatively that small values of ϵ tend to over-estimate the density within the modes and under-estimate the density in the tails. Conversely, when ϵ is taken too large, the model tends to under-estimate the modes and over-estimate the tails. Somewhere in between (and for differing levels of ϵ), every model obtains its optimal performance. Figure 12 makes these observations quantitative, and displays the KL-divergence from the target marginal to the model marginal $\mathsf{KL}(\rho_1 \parallel \hat{\rho}_1^\epsilon)$ as a function of ϵ, with each data point on the curve matching the models depicted in Figures 10 and 11. We find that for each case, there is an optimal value of $\epsilon \neq 0$, in line with the qualitative picture put forth by Figures 10 and 11. Moreover, we find that learning b generically performs better than learning v, and that learning η generically performs better than learning s (except when ϵ is taken large enough that performance starts to degrade). With proper treatment of the singularity in the sampling algorithm when using the denoiser in the construction of $s(t,x) = -\eta(t,x)/\gamma(t)$ – either by capping $t_0 \neq 0$ and $t_f \neq 1$ or by properly tuning $\epsilon(t)$ as discussed in Section 6.2 – our results suggest that learning the denoiser is best practice.

7.3 Image generation

In the following, we demonstrate that the proposed method scales straightforwardly to high-dimensional problems like image generation. To this end, we illustrate the use of our
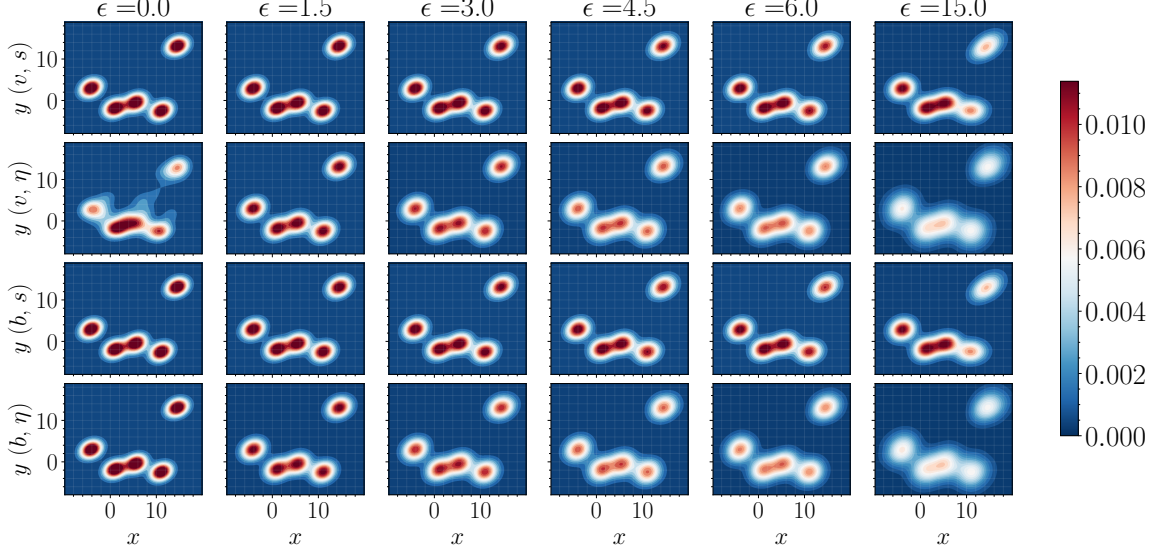
Figure 11: **Gaussian mixtures: densities.** A visualization of the marginal densities in the first two variables of the model density $\hat{\rho}_1$ computed for all four variations of learning b or v and s or η, computed via kernel density estimation. *Note: visualizations show two-dimensional slices of a 128-dimensional density.*
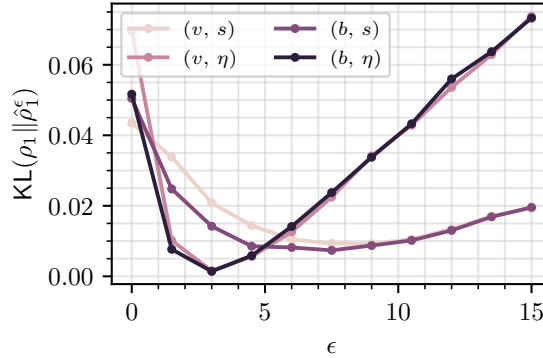

Figure 12: **Gaussian mixtures: quantitative comparison.** $\mathsf{KL}(\rho_1 \| \hat{\rho}_1^\epsilon)$ as a function of ϵ when learning each of the four possible sets of drift coefficients (b, s), (b, η), (v, s), (v, η). The best performance is achieved when learning b and η, along with a proper choice of $\epsilon > 0$.

approach on the 128×128 Oxford flowers dataset Nilsback and Zisserman (2006) by testing two different variations of the interpolant for image generation: the one-sided interpolant, using $\rho_0 = \mathsf{N}(0, Id)$, as well as the mirror interpolant, where $\rho_0 = \rho_1$ both represent the data distribution. The purpose of this section is to demonstrate that our theory is well-motivated, and that it provides a framework that is both scalable and flexible. In this regard, image generation is a convenient exercise, but is not the main focus of this work, and we will leave a more thorough study on other datasets such as ImageNet with standard benchmarks such as the Frechet Inception Distance (FID) for a future study.
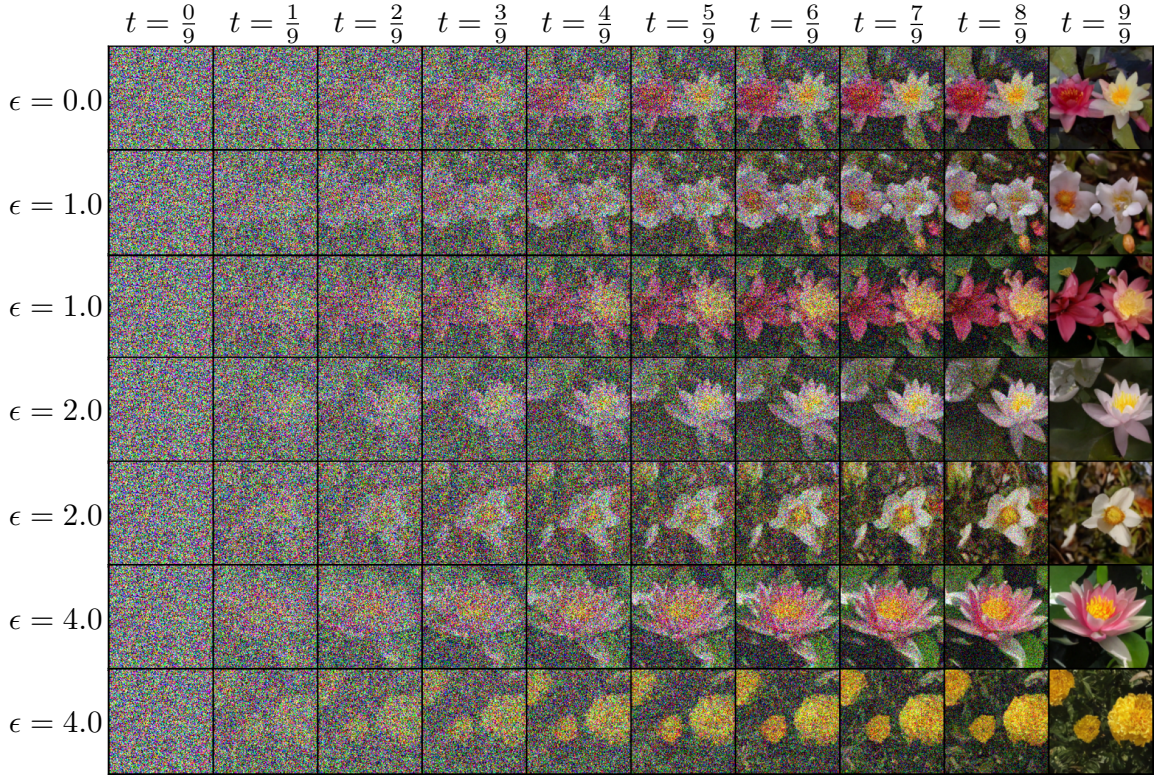
Figure 13: **Image generation: Oxford flowers.** Example generated flowers from the same initial condition x_0 using either the ODE with $\epsilon = 0$ and learned b or the SDE for various increasing values of ϵ with learned b and s. For $\epsilon = 0$, sampling is done using the `dopri5` solver and therefore the number of steps is adaptive. Otherwise, 2000, 2500, and 4000 steps were taken using the Heun solver for $\epsilon = 1.0$, 2.0, and 4.0 respectively.

Generation from Gaussian ρ_0. We train spatially-linear one-sided interpolants $x_t = (1-t)z + tx_1$ and $x_t = \cos(\frac{\pi}{2}t)z + \sin(\frac{\pi}{2}t)x_1$ on the 128×128 Oxford flowers dataset, where we take $z \sim \mathsf{N}(0, Id)$ and x_1 from the data distribution. Based on our results for Gaussian mixtures, we learn the drift $b(t, x)$, the score $s(t, x)$, and the denoiser $\eta_z(t, x)$ to benchmark our generative models based on ODEs or SDEs. In all cases, we parameterize the networks representing $\hat{\eta}$, $\hat{s}$ and $\hat{b}$ using the U-Net architecture used in Ho et al. (2020). Minimization of the objective functions given in Section 3.2 is performed using the Adam optimizer. Details of the architecture in both cases and all training hyperparameters are provided in Appendix C.1.

Like in the case of learning Gaussian mixtures, we use the fourth-order `dopri5` solver when sampling with the ODE and the Heun method for the SDE, as detailed in Algorithm 4. When learning a denoiser η_z, we found it beneficial to complete the image generation with a final denoising step, in which we set $\epsilon = 0$ and switch the integrator to the one given in (5.14).

Exemplary images generated from the model using the ODE and the SDE with various value of the diffusion coefficient ϵ are shown in Figure 13, starting from the same sample

Generated Image

5 Nearest neighbors in training set

Figure 14: **No memorization.** Top: An example generated image from a trained linear one-sided interpolant. Bottom: Five nearest neighbors (in ℓ_1 norm) to the generated image in the dataset. The nearest neighbors are visually distinct, highlighting that the interpolant does not overfit on the dataset.

Figure 15: **Mirror interpolant.** Example trajectory from a learned denoiser model $\hat{b} = \dot{\gamma}(t)\hat{\eta}_z$ for the mirror interpolant $x_t = x_1 + \gamma(t)z$ on the 128×128 Oxford flowers dataset with $\gamma(t) = \sqrt{t(1-t)}$. The parametric form for $\hat{\eta}_z$ is the U-Net from Ho et al. (2020), with hyperparameter details given in Appendix C. The choice of ϵ in the generative SDE given in (2.33) influences the extent to which the generated image differs from the original. Here, $\epsilon(t) = 10.0$.

from ρ_0. The result illustrates that different images can be generated from the same sample when using the SDE, and their diversity increases as we increase the diffusion coefficient ϵ. To highlight that the model does not memorize the training set, in Figure 14 we compare an example generated image to its five nearest neighbors in the training set (measured in ℓ_1 norm), which appear visually quite different.

Mirror interpolant. We consider the mirror interpolant $x_t = x_1 + \gamma(t)z$, for which (3.33) shows that the drift b is given in terms the denoiser η_z by $b(t, x) = \dot{\gamma}(t)\eta_z(t)$; this means that it is sufficient to only learn an estimate $\hat{\eta}_z$ to construct a generative model. Similar to the previous section, we demonstrate this on the Oxford flowers dataset, again making use of a U-Net parameterization for $\hat{\eta}_z(t, x)$. Further experimental details can be found in Appendix C.1. In this setup the output image at time $t = 1$ is the same as the input image if we use the ODE (2.32); with the SDE, however, we can generate new images from the same input. This is illustrated in Figure 15, where we show how a sample image from the dataset ρ_1 is pushed forward through the SDE (2.33) with $\epsilon(t) = \epsilon = 10$. As can be seen the original image is resampled to a proximal flower not seen in the dataset.

53

8. Conclusion

The above exposition provides a full treatment of the stochastic interpolant method, as well as a careful consideration of its relation to existing literature. Our goal is to provide a general framework that can be used to devise generative models built upon dynamical transport of measure. To this end, we have detailed mathematical theory and efficient algorithms for constructing both deterministic and stochastic generative models that map between two densities exactly in finite time. Along the way, we have illustrated the various design parameters that can be used to shape this process, with connections, for example, to optimal transport and Schrödinger bridges. While we detail specific instantiations, such as the mirror and one-sided interpolants, we highlight that there is a much broader space of possible designs that may be relevant for future applications. Several candidate application domains include the solution of inverse problems such as image inpainting and super-resolution, spatiotemporal forecasting of dynamical systems, and scientific problems such as sampling of molecular configurations and machine learning-assisted Markov chain Monte-Carlo.

Acknowledgements

Appendix A. Bridging two Gaussian mixture densities

In this appendix, we consider the case where ρ_0 and ρ_1 are both Gaussian mixture densities. We denote by

$$
\begin{aligned}
\mathsf{N}(x|m,C) &= (2\pi)^{-d/2}[\det C]^{-1/2}\exp\left(-\tfrac{1}{2}(x-m)^\mathsf{T}C^{-1}(x-m)\right), \\
&= (2\pi)^{-d}\int_{\mathbb{R}^d} e^{ik\cdot(x-m)-\frac{1}{2}k^\mathsf{T}Ck}dk,
\end{aligned}
\tag{A.1}
$$

the Gaussian probability density with mean vector $m \in \mathbb{R}^d$ and positive-definite symmetric covariance matrix $C = C^\mathsf{T} \in \mathbb{R}^{d\times d}$. We assume that

$$
\rho_0(x) = \sum_{i=1}^{N_0} p_i^0 \mathsf{N}(x|m_i^0, C_i^0), \qquad \rho_1(x) = \sum_{i=1}^{N_1} p_i^1 \mathsf{N}(x|m_i^1, C_i^1)
\tag{A.2}
$$

where $N_0, N_1 \in \mathbb{N}$, $p_i^0 > 0$ with $\sum_{i=1}^{N_0} p_i^0 = 1$, $m_i^0 \in \mathbb{R}^d$, $C_i^0 = (C_i^0)^\mathsf{T} \in \mathbb{R}^{d\times d}$ positive-definite, and similarly for p_i^1, m_i^1, and C_i^1. We have:

Proposition 51 *Consider the process x_t defined in (2.1) using the probability densities in (A.2) and the interpolant in (4.1), i.e.*

$$x_t^{\mathsf{lin}} = \alpha(t)x_0 + \beta(t)x_1 + \gamma(t)z, \tag{4.1}$$

where $x_0, \sim \rho_0$, $x_1 \sim \rho_1$, and $z \sim \mathsf{N}(0, Id)$ with $x_0 \perp x_1 \perp z$, and $\alpha, \beta, \gamma^2 \in C^2([0, 1])$ satisfy the conditions in (4.2). Denote

$$m_{ij}(t) = \alpha(t)m_i^0 + \beta(t)m_j^1, \quad C_{ij}(t) = \alpha^2(t)C_i^0 + \beta^2(t)C_j^1 + \gamma^2(t)Id, \tag{A.3}$$

where $i = 1, \ldots, N_0$, $j = 1, \ldots, N_1$. Then the probability density ρ of x_t is the Gaussian mixture density

$$\rho(t, x) = \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 \mathsf{N}(x|m_{ij}(t), C_{ij}(t)) \tag{A.4}$$

and the velocity b and the score s defined in (2.10) and (2.14) are

$$b(t, x) = \frac{\sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 \left(\dot{m}_{ij}(t) + \frac{1}{2}\dot{C}_{ij}(t)C_{ij}^{-1}(t)(x - m^{ij}(t)) \right) \mathsf{N}(x|m_{ij}(t), C_{ij}(t))}{\sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 \mathsf{N}(x|m_{ij}(t), C_{ij}(t))}, \tag{A.5}$$

and

$$s(t, x) = -\frac{\sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 C_{ij}^{-1}(t)(x - m_{ij}(t)) \mathsf{N}(x|m_{ij}(t), C_{ij}(t))}{\sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 \mathsf{N}(x|m_{ij}(t), C_{ij}(t))}. \tag{A.6}$$

This proposition implies that b and s grow at most linearly in x, and are approximately linear in regions where the modes of $\rho(t, x)$ remain well-separated. In particular, if ρ_0 and ρ_1 are both Gaussian densities, $\rho_0 = \mathsf{N}(m_0, C_0)$ and $\rho_1 = \mathsf{N}(m_1, C_1)$, we have

$$b(t, x) = \dot{m}(t) + \tfrac{1}{2}\dot{C}(t)C^{-1}(t)(x - m(t)), \tag{A.7}$$

and

$$s(t, x) = -C^{-1}(t)(x - m(t)), \tag{A.8}$$

where

$$m(t) = \alpha(t)m_0 + \beta(t)m_1, \quad C(t) = \alpha^2(t)C_0 + \beta^2(t)C_1 + \gamma^2(t)Id. \tag{A.9}$$

Note that the probability flow ODE (2.32) associated with the velocity (A.7) is the linear ODE

$$\frac{d}{dt}X_t = \dot{m}(t) + \tfrac{1}{2}\dot{C}(t)C^{-1}(t)(X_t - m(t)). \tag{A.10}$$

This equation can only be solved analytically if $\dot{C}(t)$ and $C(t)$ commute (which is the case e.g. if $C_0 = Id$), but it is easy to see that it always guarantees that

$$\mathbb{E}_0 X_t(x_0) = m(t), \qquad \mathbb{E}_0 \left[(X_t(x_0) - m(t))(X_t(x_0) - m(t))^{\mathsf{T}} \right] = C(t), \tag{A.11}$$

where $X_t(x_0)$ denotes the solution to (A.10) for the initial condition $X_{t=0}(x_0) = x_0$ and $\mathbb{E}_0$ denotes expectation over $x_0 \sim \rho_0$. A similar statement is true if we solve (A.10) with final

conditions at $t = 1$ drawn from ρ_1. Similarly, the forward SDE (2.33) associated with the velocity (A.7) and the score (A.8) is the linear SDE

$$dX_t^{\mathsf{F}} = \dot{m}(t)dt + \left(\tfrac{1}{2}\dot{C}(t) - \epsilon\right)C^{-1}(t)(X_t^{\mathsf{F}} - m(t))dt + \sqrt{2\epsilon}dW_t. \tag{A.12}$$

and its solutions are such that

$$\mathbb{E}_0\mathbb{E}_{\mathsf{F}}^{x_0} X_t^{\mathsf{F}} = m(t), \qquad \mathbb{E}_0\mathbb{E}_{\mathsf{F}}^{x_0}\left[(X_t^{\mathsf{F}} - m(t))(X_t^{\mathsf{F}} - m(t))^{\mathsf{T}}\right] = C(t), \tag{A.13}$$

where $\mathbb{E}_{\mathsf{F}}^{x_0}$ denotes expectation over the solution of (A.12) conditional on the event $X_{t=0}^{\mathsf{F}} = x_0$ and $\mathbb{E}_0$ denotes expectation over $x_0 \sim \rho_0$. A similar statement also holds for the backward SDE (2.34).

Proof The characteristic function of $\rho(t, x)$ is given by

$$g(t, k) = \mathbb{E}e^{ik \cdot x_t} = \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 e^{ik \cdot m_{ij}(t) - \frac{1}{2}k^{\mathsf{T}}C_{ij}(t)k}. \tag{A.14}$$

whose inverse Fourier transform is (A.4). This automatically implies (A.6) since we know from 2.14 that $s = \nabla \log \rho$. To derive (A.7) use the function m defined below in (B.12):

$$m(t, k) = \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 (\dot{m}_{ij}(t) + \tfrac{1}{2}i\dot{C}_{ij}(t)k)e^{ik \cdot m_{ij}(t) - \frac{1}{2}k^{\mathsf{T}}C_{ij}^{\gamma}(t)k}. \tag{A.15}$$

From (B.13), we know that the inverse Fourier transform of this function is $b\rho$, so that we obtain

$$b(t, x)\rho(t, x) = \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_i^0 p_j^1 \left(\dot{m}_{ij}(t) + \tfrac{1}{2}\dot{C}_{ij}(t)C_{ij}^{-1}(t)(x - m_{ij}(t))\right) \mathsf{N}(x|m_{ij}(t), C_{ij}^{\epsilon}(t)).$$
$$\tag{A.16}$$

This gives (A.5). ∎

Appendix B. Proofs

In this appendix, we provide the details for proofs omitted from the main text. For ease of reading, a copy of the original theorem statement is provided with the proof.

B.1 Proof of Theorems 6, 7, and 8, and Corollary 10.

Theorem 6 (Stochastic interpolant properties) *The probability distribution of the stochastic interpolant x_t defined in (2.1) is absolutely continuous with respect to the Lebesgue measure at all times $t \in [0, 1]$ and its time-dependent density $\rho(t)$ satisfies $\rho(0) = \rho_0$, $\rho(1) = \rho_1$, $\rho \in C^1([0, 1]; C^p(\mathbb{R}^d))$ for any $p \in \mathbb{N}$, and $\rho(t, x) > 0$ for all $(t, x) \in [0, 1] \times \mathbb{R}^d$. In addition, ρ solves the transport equation*

$$\partial_t \rho + \nabla \cdot (b\rho) = 0, \tag{2.9}$$

where we defined the velocity

$$b(t,x) = \mathbb{E}[\dot{x}_t | x_t = x] = \mathbb{E}[\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z | x_t = x]. \tag{2.10}$$

This velocity is in $C^0([0,1]; (C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$, and such that

$$\forall t \in [0,1] \quad : \quad \int_{\mathbb{R}^d} |b(t,x)|^2 \rho(t,x) dx < \infty. \tag{2.11}$$

Proof Let $g(t,k) = \mathbb{E}e^{ik \cdot x_t}$, $k \in \mathbb{R}^d$, be the characteristic function of $\rho(t,x)$. From the definition of x_t in (2.1),

$$g(t,k) = \mathbb{E}e^{ik \cdot (I(t,x_0,x_1) + \gamma(t)z)}. \tag{B.1}$$

Using the independence between (x_0, x_1) and z, we have

$$g(t,k) = \mathbb{E}\left(e^{ik \cdot I(t,x_0,x_1)}\right) \mathbb{E}\left(e^{i\gamma(t)k \cdot z}\right) \equiv g_0(t,k)e^{-\frac{1}{2}\gamma^2(t)|k|^2} \tag{B.2}$$

where we defined

$$g_0(t,k) = \mathbb{E}\left(e^{ik \cdot I(t,x_0,x_1)}\right) \tag{B.3}$$

The function $g_0(t,k)$ is the characteristic function of $I(t, x_0, x_1)$ with $(x_0, x_1) \sim \nu$. From (B.2), we have

$$|g(t,k)| = |g_0(t,k)|e^{-\frac{1}{2}\gamma^2(t)|k|^2} \leq e^{-\frac{1}{2}\gamma^2(t)|k|^2} \tag{B.4}$$

Since $\gamma(t) > 0$ for all $t \in (0,1)$ by assumption, this shows that

$$\forall p \in \mathbb{N} \text{ and } t \in (0,1) \quad : \quad \int_{\mathbb{R}^d} |k|^p |g(t,k)| dk < \infty, \tag{B.5}$$

implying that $\rho(t, \cdot)$ is in $C^p(\mathbb{R}^d)$ for any $p \in \mathbb{N}$ and all $t \in (0,1)$. From (B.2), we also have

$$\begin{aligned}
|\partial_t g(t,k)|^2 &= \left|\mathbb{E}[(ik \cdot \partial_t I_t(x_0, x_1) - \gamma(t)\dot{\gamma}(t)|k|^2)e^{ik \cdot I_t(x_0,x_1)}]\right|^2 e^{-\gamma^2(t)|k|^2} \\
&\leq 2\left(|k|^2 \mathbb{E}[|\partial_t I_t(x_0, x_1)|^2] + |\gamma(t)\dot{\gamma}(t)|^2|k|^4\right)e^{-\gamma^2(t)|k|^2} \\
&\leq 2\left(|k|^2 M_1 + 4|\gamma(t)\dot{\gamma}(t)|^2|k|^4\right)e^{-\gamma^2(t)|k|^2}
\end{aligned} \tag{B.6}$$

and

$$\begin{aligned}
|\partial_t^2 g(t,k)|^2 &\leq 4\left(|k|^2 \mathbb{E}[|\partial_t^2 I_t(x_0, x_1)|^2] + (|\dot{\gamma}(t)|^2 + \gamma(t)\ddot{\gamma}(t))^2|k|^4\right)e^{-\gamma^2(t)|k|^2} \\
&\quad + 8\left(|k|^2 \mathbb{E}[|\partial_t I_t(x_0, x_1)|^4] + (\gamma(t)\dot{\gamma}(t))^4|k|^8\right)e^{-\gamma^2(t)|k|^2} \\
&\leq 4\left(|k|^2 M_2 + (|\dot{\gamma}(t)|^2 + \gamma(t)\ddot{\gamma}(t))^2|k|^4\right)e^{-\gamma^2(t)|k|^2} \\
&\quad + 8\left(|k|^2 M_1 + (\gamma(t)\dot{\gamma}(t))^4|k|^8\right)e^{-\gamma^2(t)|k|^2}
\end{aligned} \tag{B.7}$$

where in both cases we used (2.8) in Assumption 5 to get the last inequalities. These imply that

$$\forall p \in \mathbb{N} \text{ and } t \in (0,1) \quad : \quad \int_{\mathbb{R}^d} |k|^p |\partial_t g(t,k)| dk < \infty; \quad \int_{\mathbb{R}^d} |k|^p |\partial_t^2 g(t,k)| dk < \infty \tag{B.8}$$

57

indicating that $\partial_t \rho(t, \cdot)$ and $\partial_t^2 \rho(t, \cdot)$ are in $C^p(\mathbb{R}^d)$ for any $p \in \mathbb{N}$, i.e. $\rho \in C^1((0, 1); C^p(\mathbb{R}^d))$ as claimed. To show that ρ is also positive, denote by $\mu_0(t, dx)$ the unique (by the Fourier inversion theorem) probability measure associated with $g_0(t, k)$, i.e. the measure such that

$$g_0(t, k) = \int_{\mathbb{R}^d} e^{ik \cdot x} \mu_0(t, dx). \tag{B.9}$$

From (B.2) and the convolution theorem it follows that we can express ρ as

$$\rho(t, x) = \int_{\mathbb{R}^d} \frac{e^{-|x-y|^2/(2\gamma^2(t))}}{(2\pi\gamma^2(t))^{d/2}} \mu_0(t, dy), \tag{B.10}$$

This shows that $\rho > 0$ for all $(t, x) \in (0, 1) \times \mathbb{R}^d$. Since $x_{t=0} = x_0$ and $x_{t=1} = x_1$ by definition of the interpolant, we also have $\rho(0) = \rho_0$ and $\rho(1) = \rho_1$, which shows that ρ is also positive and in $C^p(\mathbb{R}^d)$ at $t = 0, 1$ by Assumption 5. Note that since $\rho \in C^1((0, 1); C^p(\mathbb{R}^d))$ and is positive, we also immediately deduce that $s = \nabla \log \rho = \nabla \rho / \rho \in C^1((0, 1); (C^p(\mathbb{R}^d))^d)$.

To show that ρ satisfies the TE (2.9), we take the time derivative of (B.1) to deduce that

$$\partial_t g(t, k) = ik \cdot m(t, k) \tag{B.11}$$

where $m : [0, 1] \times \mathbb{R}^d \to \mathbb{C}^d$ is the vector-valued function defined as

$$m(t, k) = \mathbb{E}\left((\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z)e^{ik \cdot x_t}\right). \tag{B.12}$$

By definition of the conditional expectation, $m(t, k)$ can be expressed as

$$\begin{aligned} m(t, k) &= \int_{\mathbb{R}^d} \mathbb{E}\left((\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z)e^{ik \cdot x_t} | x_t = x\right) \rho(t, x) dx \\ &= \int_{\mathbb{R}^d} e^{ik \cdot x} \mathbb{E}\left((\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z) | x_t = x\right) \rho(t, x) dx \\ &= \int_{\mathbb{R}^d} e^{ik \cdot x} b(t, x) \rho(t, x) dx \end{aligned} \tag{B.13}$$

where the last equality follows from the definition of b in (2.10). Inserting (B.13) in (B.11), we deduce that this equation can be written in real space as the TE (2.9).

Let us now investigate the regularity of b. To that end, we go back to m and use the independence between x_0, x_1, and z, as well as Gaussian integration by parts to deduce that

$$m(t, k) = \mathbb{E}\left((\partial_t I(t, x_0, x_1) - i\gamma(t)\dot{\gamma}(t)k)e^{ik \cdot I(t, x_1, x_0)}\right) e^{-\frac{1}{2}\gamma^2(t)|k|^2}, \tag{B.14}$$

As a result

$$\begin{aligned} |m(t, k)|^2 &= \left|\mathbb{E}\left((\partial_t I(t, x_0, x_1) - i\gamma(t)\dot{\gamma}(t)k)e^{ik \cdot I(t, x_1, x_0)}\right)\right|^2 e^{-\gamma^2(t)|k|^2} \\ &\leq 2\left(\mathbb{E}[|\partial_t I(t, x_0, x_1)|^2] + |\gamma(t)\dot{\gamma}(t)|^2|k|^2\right) e^{-\gamma^2(t)|k|^2} \\ &\leq 2M_1 e^{-\gamma^2(t)|k|^2}, \end{aligned} \tag{B.15}$$

and

$$
\begin{aligned}
|\partial_t m(t,k)|^2 &\leq 4 \left(\mathbb{E}\big[|\partial_t^2 I(t,x_0,x_1)|^2 + (\gamma(t)\ddot{\gamma}(t) + \dot{\gamma}^2(t))^2\big) e^{-\gamma^2(t)|k|^2} \\
&\quad + 8|k|^2 \left(\mathbb{E}|\partial_t I(t,x_0,x_1)|^4\big] + (\gamma(t)\dot{\gamma}(t))^4|k|^4 \right) e^{-\gamma^2(t)|k|^2} \\
&\leq 4 \left(M_1 + (\gamma(t)\ddot{\gamma}(t) + \dot{\gamma}^2(t))^2 \right) e^{-\gamma^2(t)|k|^2} \\
&\quad + 8|k|^2 \left(M_2 + (\gamma(t)\dot{\gamma}(t))^4|k|^4 \right) e^{-\gamma^2(t)|k|^2},
\end{aligned}
\tag{B.16}
$$

where in both cases the last inequalities follow from (2.8). Therefore

$$
\forall p \in \mathbb{N} \ \text{ and } \ t \in (0,1) \quad : \quad \int_{\mathbb{R}^d} |k|^p |m(t,k)| dk < \infty, \quad \int_{\mathbb{R}^d} |k|^p |\partial_t m(t,k)| dk < \infty, \tag{B.17}
$$

which implies that the inverse Fourier transform of m is a function $j : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ that satisfies $j(t,\cdot) \in (C^p(\mathbb{R}^d))^d$ for any $p \in \mathbb{N}$ and any $t \in (0,1)$ and can be expressed as

$$
j(t,x) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-ik\cdot x} m(t,k) dk = \mathbb{E}\left(\partial_t I(t,x_0,x_1) + \dot{\gamma}(t)z | x_t = x \right) \rho(t,x) \equiv b(t,x)\rho_t(x)
\tag{B.18}
$$

where the last equality follows from the definition of b in (2.10). We deduce that $b \in C^0([0,1]; (C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$ since $j \in C^0([0,1]; (C^p(\mathbb{R}^d))^d)$ and $\rho \in C^1([0,1]; C^p(\mathbb{R}^d))$, and $\rho > 0$.

Finally, let us establish (2.11). By (2.8) we have

$$
\begin{aligned}
\int_{\mathbb{R}^d} |b(t,x)|^2 \rho(t,x) dx &= \int_{\mathbb{R}^d} |\mathbb{E}\left(\partial_t I(t,x_0,x_1) + \dot{\gamma}(t)z | x_t = x \right)|^2 \rho(t,x) dx \\
&\leq 2 \int_{\mathbb{R}^d} \mathbb{E}\left(|\partial_t I(t,x_0,x_1)|^2 + |\dot{\gamma}(t)|^2|z|^2 | x_t = x \right) \rho(t,x) dx \\
&\leq 2\mathbb{E}\big[|\partial_t I(t,x_0,x_1)|^2 + |\dot{\gamma}(t)|^2|z|^2\big] \\
&< 2M_1^{1/2} + 2d|\dot{\gamma}(t)|^2,
\end{aligned}
\tag{B.19}
$$

so that this integral is bounded for all $t \in (0,1)$. To analyze its behavior at the end points, notice that the decomposition (2.27) implies that

$$
\begin{aligned}
b_0(x) &\equiv \lim_{t \to 0} b(t,x) = \mathbb{E}_1[\partial_t I(0,x,x_1)] - \lim_{t \to 0} \dot{\gamma}(t)\gamma(t)s_0(x), \\
b_1(x) &\equiv \lim_{t \to 1} b(t,x) = \mathbb{E}_0[\partial_t I(0,x_0,x)] - \lim_{t \to 1} \dot{\gamma}(t)\gamma(t)s_1(x),
\end{aligned}
\tag{B.20}
$$

where $s_0 = \nabla \log \rho_0$, $s_1 = \nabla \log \rho_1$, $\mathbb{E}_0$ and $\mathbb{E}_1$ denote expectations over $x_0 \sim \rho_0$ and $x_1 \sim \rho_1$, respectively, and we used the property that $x_{t=0} = x_0$ and $x_{t=1} = x_1$. Since $\lim_{t \to 0,1} \dot{\gamma}(t)\gamma(t)$ exists by our assumption that $\gamma^2 \in C^1([0,1])$, b_0 and b_1 are well defined, and

$$
\int_{\mathbb{R}^d} |b_0(x)|^2 \rho_0(x) dx < \infty, \qquad \int_{\mathbb{R}^d} |b_1(x)|^2 \rho_1(x) dx < \infty. \tag{B.21}
$$

by Assumption 5. As a result, the integral in (B.19) is continuous at $t = 0$ and $t = 1$, so it must be integrable on $[0,1]$, and (2.11) holds. ∎

Theorem 7 (Objective) *The velocity b defined in (2.10) is the unique minimizer in $C^0([0,1]; (C^1(\mathbb{R}^d))^d)$ of the quadratic objective*

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{b}(t, x_t)|^2 - (\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z) \cdot \hat{b}(t, x_t)\right) dt \qquad (2.13)$$

where x_t is defined in (2.1) and the expectation is taken independently over $(x_0, x_1) \sim \nu$ and $z \sim \mathsf{N}(0, Id)$.

Proof By definition of ρ, the objective $\mathcal{L}_b$ defined in (2.13) can also be written as

$$
\begin{aligned}
\mathcal{L}_b[\hat{b}] &= \int_0^1 \int_{\mathbb{R}^d} \left(\tfrac{1}{2}|\hat{b}(t, x)|^2 - \mathbb{E}\left((\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)z | x_t = x) \cdot \hat{b}(t, x)\right)\right) \rho(t, x) dx dt \\
&= \int_0^1 \int_{\mathbb{R}^d} \left(\tfrac{1}{2}|\hat{b}(t, x)|^2 - b(t, x) \cdot \hat{b}(t, x)\right) \rho(t, x) dx dt
\end{aligned}
\qquad (\text{B}.22)
$$

where we used the definition of b in (2.10). This quadratic objective is bounded from below since

$$
\begin{aligned}
\mathcal{L}_b[\hat{b}] &= \tfrac{1}{2} \int_0^1 \int_{\mathbb{R}^d} \left|\hat{b}(t, x) - b(t, x)\right|^2 \rho(t, x) dx dt - \tfrac{1}{2} \int_0^1 \int_{\mathbb{R}^d} |b(t, x)|^2 \rho(t, x) dx dt \\
&\geq -\tfrac{1}{2} \int_0^1 \int_{\mathbb{R}^d} |b(t, x)|^2 \rho(t, x) dx dt > -\infty
\end{aligned}
\qquad (\text{B}.23)
$$

where the last inequality follows from (2.11). Since ρ_t is positive the minimizer of (B.22) is unique and given by $\hat{b} = b$. $\blacksquare$

Theorem 8 (Score) *The score of the probability density ρ specified in Theorem 6 is in $C^1([0,1]; (C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$ and given by*

$$s(t, x) = \nabla \log \rho(t, x) = -\gamma^{-1}(t)\mathbb{E}(z | x_t = x) \quad \forall (t, x) \in (0, 1) \times \mathbb{R}^d \qquad (2.14)$$

In addition it satisfies

$$\forall t \in [0, 1] \quad : \quad \int_{\mathbb{R}^d} |s(t, x)|^2 \rho(t, x) dx < \infty, \qquad (2.15)$$

and is the unique minimizer in $C^1([0,1]; (C^1(\mathbb{R}^d))^d)$ of the quadratic objective

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{s}(t, x_t)|^2 + \gamma^{-1}(t)z \cdot \hat{s}(t, x_t)\right) dt \qquad (2.16)$$

where x_t is defined in (2.1) and the expectation is taken independently over $(x_0, x_1) \sim \nu$ and $z \sim \mathsf{N}(0, Id)$

Proof Since $\rho \in C^1((0,1); C^p(\mathbb{R}^d))$ and is positive by Theorem 6, we already know that $s = \nabla \log \rho = \nabla \rho / \rho \in C^1((0,1); (C^p(\mathbb{R}^d))^d)$. To establish (2.14), note that, for $t \in (0,1)$ where $\gamma(t) > 0$, we have

$$\mathbb{E}\left(ze^{i\gamma(t)k\cdot z}\right) = -\gamma^{-1}(t)(i\partial_k)\mathbb{E}e^{i\gamma(t)k\cdot z} = -\gamma^{-1}(t)(i\partial_k)e^{-\frac{1}{2}\gamma^2(t)|k|^2} = i\gamma(t)ke^{-\frac{1}{2}\gamma^2(t)|k|^2}. \tag{B.24}$$

As a result, using the independence between x_0, x_1, and z, we have

$$\mathbb{E}\left(ze^{ik\cdot x_t}\right) = i\gamma(t)kg(t,k) \tag{B.25}$$

where g is the characteristic function of x_t defined in (B.1). Using the properties of the conditional expectation, the left-hand side of this equation can be written as

$$\mathbb{E}\left(ze^{ik\cdot x_t}\right) = \int_{\mathbb{R}^d} \mathbb{E}\left(ze^{ik\cdot x_t}|x_t = x\right)\rho(t,x)dx = \int_{\mathbb{R}^d} \mathbb{E}\left(z|x_t = x\right)e^{ikx_t}\rho(t,x)dx \tag{B.26}$$

Since the left hand side of (B.25) is the Fourier transform of $-\gamma(t)\nabla\rho(t,x)$, we deduce that

$$\mathbb{E}\left(z|x_t = x\right)\rho(t,x) = -\gamma(t)\nabla\rho(t,x) = -\gamma(t)s(t,x)\rho(t,x). \tag{B.27}$$

Since $\rho(t,x) > 0$, this implies (2.14) for $t \in (0,1)$ where $\gamma(t) > 0$.

To establish (2.15), notice that

$$\begin{aligned}
\int_{\mathbb{R}^d} |s(t,x)|^2 \rho(t,x)dx &= \int_{\mathbb{R}^d} |\mathbb{E}\left((\gamma^{-1}(t)z|x_t = x\right)|^2\rho(t,x)dx \\
&\leq \int_{\mathbb{R}^d} \gamma^{-2}(t)\mathbb{E}\left(|z|^2|x_t = x\right)\rho(t,x)dx \\
&\leq d\gamma^{-2}(t).
\end{aligned} \tag{B.28}$$

This means that this integral is bounded for all $t \in (0,1)$. Since the integral is also continuous at $t = 0$ and $t = 1$, with values given by (2.7), it must be integrable on $[0,1]$ and (2.15) holds.

The objective $\mathcal{L}_s$ defined in (2.16) can also be written as

$$\begin{aligned}
\mathcal{L}_s[\hat{s}] &= \int_0^1 \int_{\mathbb{R}^d} \left(\tfrac{1}{2}|\hat{s}(t,x)|^2 + \gamma^{-1}(t)\mathbb{E}\left(z|x_t = x\right)\cdot\hat{s}(t,x)\right)\rho(t,x)dxdt \\
&= \int_0^1 \int_{\mathbb{R}^d} \left(\tfrac{1}{2}|\hat{s}(t,x)|^2 - s(t,x)\cdot\hat{s}(t,x)\right)\rho(t,x)dxdt,
\end{aligned} \tag{B.29}$$

where we used the definition of s in (2.14). This quadratic objective is bounded from below since

$$\begin{aligned}
\mathcal{L}_s[\hat{s}] &= \tfrac{1}{2}\int_0^1 \int_{\mathbb{R}^d} |\hat{s}(t,x) - s(t,x)|^2 \rho(t,x)dxdt - \tfrac{1}{2}\int_0^1 \int_{\mathbb{R}^d} |s(t,x)|^2 \rho(t,x)dxdt \\
&\geq -\tfrac{1}{2}\int_0^1 \int_{\mathbb{R}^d} |s(t,x)|^2 \rho(t,x)dxdt > -\infty
\end{aligned} \tag{B.30}$$

where the last inequality follows from (2.15). Since ρ is positive the minimizer of (B.29) is unique and given by $\hat{s} = s$. ∎

Corollary 52 (Fokker-Planck equations) *For any $\epsilon \in C^0([0,1])$ with $\epsilon(t) \geq 0$ for all $t \in [0,1]$, the probability density ρ specified in Theorem 6 satisfies:*

1. *The forward Fokker-Planck equation*

$$\partial_t \rho + \nabla \cdot (b_{\mathsf{F}} \rho) = \epsilon(t)\Delta\rho, \qquad \rho(0) = \rho_0, \tag{2.20}$$

   where we defined the forward drift

$$b_{\mathsf{F}}(t,x) = b(t,x) + \epsilon(t)s(t,x). \tag{2.21}$$

   Equation (2.20) is well-posed when solved forward in time from $t=0$ to $t=1$, and its solution for the initial condition $\rho(t=0) = \rho_0$ satisfies $\rho(t=1) = \rho_1$.

2. *The backward Fokker-Planck equation*

$$\partial_t \rho + \nabla \cdot (b_{\mathsf{B}} \rho) = -\epsilon(t)\Delta\rho, \qquad \rho(1) = \rho_1, \tag{2.22}$$

   where we defined the backward drift

$$b_{\mathsf{B}}(t,x) = b(t,x) - \epsilon(t)s(t,x). \tag{2.23}$$

   Equation (2.22) is well-posed when solved backward in time from $t=1$ to $t=0$, and its solution for the final condition $\rho(1) = \rho_1$ satisfies $\rho(0) = \rho_0$.

Proof The forward FPE (2.20) and the backward FPE (2.22) are direct consequences of the TE (2.9) and (2.14), since the equality

$$\epsilon(t)\Delta\rho = \epsilon(t)\nabla \cdot (\rho \nabla \log \rho) = \epsilon(t)\nabla \cdot (s\rho) \tag{B.31}$$

can be used to convert between these equations.

$\blacksquare$

B.2 Proof of Lemma 19

Lemma 53 (Reverse Itô Calculus) *If X_t^{B} solves the backward SDE (2.34):*

1. *For any $f \in C^1([0,1]; C_0^2(\mathbb{R}_d))$ and $t \in [0,1]$, the backward Itô formula holds*

$$df(t, X_t^{\mathsf{B}}) = \partial_t f(t, X_t^{\mathsf{B}})dt + \nabla f(X_t^{\mathsf{B}}) \cdot dX_t^{\mathsf{B}} - \epsilon(t)\Delta f(X_t^{\mathsf{B}})dt. \tag{2.36}$$

2. *For any $g \in C^0([0,1]; (C_0(\mathbb{R}_d))^d)$ and $t \in [0,1]$, the backward Itô isometries hold:*

$$\mathbb{E}_{\mathsf{B}}^x \int_t^1 g(t, X_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}} = 0; \qquad \mathbb{E}_{\mathsf{B}}^x \left| \int_t^1 g(t, X_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}} \right|^2 = \int_t^1 \mathbb{E}_{\mathsf{B}}^x \left| g(t, X_t^{\mathsf{B}}) \right|^2 dt, \tag{2.37}$$

   where $\mathbb{E}_{\mathsf{B}}^x$ denotes expectation conditioned on the event $X_{t=1}^{\mathsf{B}} = x$.

Proof The SDE (2.33) and the ODE (2.32) are the evolution equations for the processes whose densities solve (2.20) and (2.9), respectively. The equation that requires some explanation is the backwards SDE (2.34), which can be solved backwards in time from $t = 1$ to $t = 0$. As discussed in the main text, by definition, its solution is $X_t^\mathsf{B} = Z_{1-t}^\mathsf{F}$ where Z_t^F solves the forward SDE

$$dZ_t^\mathsf{F} = -b_\mathsf{B}(1 - t, Z_t^\mathsf{F})dt + \sqrt{2\epsilon}dW_t. \tag{B.32}$$

To see how to write the backward Itô formula (2.36) note that given any $f \in C^1([0, 1]; C_0^2(\mathbb{R}^d))$, we have

$$
\begin{aligned}
df(1 - t, Z_t^\mathsf{F}) &= -\partial_t f(1 - t, Z_t^\mathsf{F})dt + \nabla f(1 - t, Z_t^\mathsf{F}) \cdot dZ_t^\mathsf{F} + \epsilon \Delta f(1 - t, Z_t^\mathsf{F})dt \\
&= -\partial_t f(1 - t, Z_t^\mathsf{F})dt + \left(-b_\mathsf{B}(1 - t, Z_t^\mathsf{F}) \cdot \nabla f(1 - t, Z_t^\mathsf{F}) + \epsilon \Delta f(1 - t, Z_t^\mathsf{F})\right)dt \\
&\quad + \sqrt{2\epsilon}\nabla f(1 - t, Z_t^\mathsf{F}) \cdot dW_t
\end{aligned}
\tag{B.33}
$$

In integral form, this equation can be written as

$$
\begin{aligned}
f(1, Z_1^\mathsf{F}) &= f(1 - t, Z_{1-t}^\mathsf{F}) \\
&\quad - \int_{1-t}^1 \left(\partial_t f(1 - s, Z_s^\mathsf{F}) + b_\mathsf{B}(1 - s, Z_s^\mathsf{F}) \cdot \nabla f(1 - s, Z_s^\mathsf{F}) - \epsilon \Delta f(1 - s, Z_s^\mathsf{F})\right) ds \\
&\quad - \sqrt{2\epsilon} \int_{1-t}^1 \nabla f(1 - s, Z_s^\mathsf{F}) \cdot dW_s.
\end{aligned}
\tag{B.34}
$$

Using $X_t^\mathsf{B} = Z_{1-t}^\mathsf{F}$ and $W_t^\mathsf{B} = -W_{1-t}$ and changing integration variable from s to $1 - s$, this is

$$
\begin{aligned}
f(1, X_0^\mathsf{B}) &= f(1 - t, X_t^\mathsf{B}) - \int_t^1 \left(\partial_t f(s, X_s^\mathsf{B}) + b_\mathsf{B}(s, X_s^\mathsf{B}) \cdot \nabla f(s, X_s^\mathsf{B}) - \epsilon \Delta f(s, X_s^\mathsf{B})\right) ds \\
&\quad - \sqrt{2\epsilon} \int_t^1 \nabla f(s, X_s^\mathsf{B}) \cdot dW_s^\mathsf{B},
\end{aligned}
\tag{B.35}
$$

In differential form, this is equivalent to saying that

$$
\begin{aligned}
df(t, X_t^\mathsf{B}) &= \partial_t f(t, X_t^\mathsf{B})dt + \nabla f(t, X_t^\mathsf{B}) \cdot dX_t^\mathsf{B} - \epsilon \Delta f(X_t^\mathsf{B})dt \\
&= \left(\partial_t f(t, X_t^\mathsf{B}) + b_\mathsf{B}(t, X_t^\mathsf{B}) \cdot \nabla f(t, X_t^\mathsf{B}) - \epsilon \Delta f(t, X_t^\mathsf{B})\right) dt + \sqrt{2\epsilon}\nabla f(t, X_t^\mathsf{B}) \cdot dW_t
\end{aligned}
\tag{B.36}
$$

which is the backward Itô formula (2.36). Similarly, by the Itô isometries we have: for any $g \in C^0([0, 1]; (C_0^0(\mathbb{R}_d))^d)$ and $t \in [0, 1]$

$$\mathbb{E}^x \int_{1-t}^1 g(s, Z_s^\mathsf{F}) \cdot dW_s = 0, \qquad \mathbb{E}^x \left| \int_{1-t}^1 g(s, Z_s^\mathsf{F}) \cdot dW_s \right|^2 = \int_{1-t}^1 \mathbb{E}^x |g(s, Z_s^\mathsf{F})|^2 ds. \tag{B.37}$$

Written in terms of X_t^B, these are (2.37). ∎

Derivation of (2.30). For any $t \in [0, 1]$ the score is the minimizer of

$$
\int_{\mathbb{R}^d} |\hat{s}(t, x) - \nabla \log \rho(t, x)|^2 \rho(t, x) dx
$$

$$
= \int_{\mathbb{R}^d} \left(|\hat{s}(t, x)|^2 - 2\hat{s}(t, x) \cdot \nabla \log \rho(t, x) + |\nabla \log \rho(t, x)|^2 \right) \rho(t, x) dx \tag{B.38}
$$

$$
= \int_{\mathbb{R}^d} \left(|\hat{s}(t, x)|^2 + 2\nabla \cdot \hat{s}(t, x) + |\nabla \log \rho(t, x)|^2 \right) \rho(t, x) dx,
$$

where we used the identity $\hat{s} \cdot \nabla \log \rho \, \rho = \hat{s} \cdot \nabla \rho$ and integration by parts to obtain the second equality. The last term involving $|\nabla \log \rho|^2$ is a constant in $\hat{s}$ that can be neglected for optimization. Expressing the remaining terms as an expectation over x_t and integrating the result in time gives (2.30).

B.3 Proofs of Lemmas 21 and 22, and Theorem 23.

Lemma 54 *Let $\rho_0 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote a fixed base probability density function. Given two velocity fields $b, \hat{b} \in C^0([0, 1], (C^1(\mathbb{R}^d))^d)$, let the time-dependent densities $\rho : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and $\hat{\rho} : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote the solutions to the transport equations*

$$
\begin{aligned}
\partial_t \rho + \nabla \cdot (b\rho) &= 0, & \rho(0) &= \rho_0, \\
\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\hat{\rho}) &= 0, & \hat{\rho}(0) &= \rho_0.
\end{aligned} \tag{2.38}
$$

Then, the Kullback-Leibler divergence of $\rho(1)$ from $\hat{\rho}(1)$ is given by

$$
\mathsf{KL}(\rho(1) \,\|\, \hat{\rho}(1)) = \int_0^1 \int_{\mathbb{R}^d} (\nabla \log \hat{\rho}(t, x) - \nabla \log \rho(t, x)) \cdot (\hat{b}(t, x) - b(t, x)) \rho(t, x) dx dt. \tag{2.39}
$$

Proof Using (2.38), we compute analytically

$$
\begin{aligned}
\frac{d}{dt} \mathsf{KL}(\rho(t) \,\|\, \hat{\rho}(t)) &= \frac{d}{dt} \int_{\mathbb{R}^d} \log \left(\frac{\rho}{\hat{\rho}} \right) \rho dx \\
&= \int_{\mathbb{R}^d} \hat{\rho} \left(\frac{\partial_t \rho}{\hat{\rho}} - \frac{\rho}{(\hat{\rho})^2} \partial_t \hat{\rho} \right) dx + \int \log \left(\frac{\rho}{\hat{\rho}} \right) \partial_t \rho dx \\
&= -\int_{\mathbb{R}^d} \left(\frac{\rho}{\hat{\rho}} \right) \partial_t \hat{\rho} dx + \int_{\mathbb{R}^d} \log \left(\frac{\rho}{\hat{\rho}} \right) \partial_t \rho dx \\
&= \int_{\mathbb{R}^d} \left(\frac{\rho}{\hat{\rho}} \right) \nabla \cdot \left(\hat{b}\hat{\rho} \right) dx - \int \log \left(\frac{\rho}{\hat{\rho}} \right) \nabla \cdot (b\rho) dx \\
&= -\int_{\mathbb{R}^d} \nabla \left(\frac{\rho}{\hat{\rho}} \right) \cdot \hat{b}\hat{\rho} dx + \int (\nabla \log \rho - \nabla \log \hat{\rho}) \cdot b\rho dx \\
&= -\int_{\mathbb{R}^d} \left(\frac{\nabla \rho}{\hat{\rho}} - \frac{\rho \nabla \hat{\rho}}{\hat{\rho}^2} \right) \cdot \hat{b}\hat{\rho} dx + \int_{\mathbb{R}^d} (\nabla \log \rho - \nabla \log \hat{\rho}) \cdot b\rho dx \\
&= \int_{\mathbb{R}^d} (\nabla \log \hat{\rho} - \nabla \log \rho) \cdot \hat{b}\rho dx + \int_{\mathbb{R}^d} (\nabla \log \rho - \nabla \log \hat{\rho}) \cdot b\rho dx \\
&= \int_{\mathbb{R}^d} (\nabla \log \hat{\rho} - \nabla \log \rho) \cdot (\hat{b} - b)\rho dx.
\end{aligned}
$$

where we omitted the argument (t, x) of all functions for simplicity of notation. Integrating both sides from 0 to 1 completes the proof. ∎

Lemma 55 *Let $\rho_0 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote a fixed base probability density function. Given two velocity fields $b_\mathsf{F}, \hat{b}_\mathsf{F} \in C^0([0, 1], (C^1(\mathbb{R}^d))^d)$, let the time-dependent densities $\rho : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and $\hat{\rho} : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote the solutions to the Fokker-Planck equations*

$$
\begin{aligned}
\partial_t \rho + \nabla \cdot (b_\mathsf{F} \rho) &= \epsilon \Delta \rho, & \rho(0) &= \rho_0, \\
\partial_t \hat{\rho} + \nabla \cdot (\hat{b}_\mathsf{F} \hat{\rho}) &= \epsilon \Delta \hat{\rho}, & \hat{\rho}(0) &= \rho_0.
\end{aligned}
\tag{2.40}
$$

where $\epsilon > 0$. Then, the Kullback-Leibler divergence from $\rho(1)$ to $\hat{\rho}(1)$ is given by

$$
\begin{aligned}
\mathsf{KL}(\rho(1) \,\|\, \hat{\rho}(1)) = {} & \int_0^1 \int_{\mathbb{R}^d} (\nabla \log \hat{\rho}(t, x) - \nabla \log \rho(t, x)) \cdot \left(\hat{b}_\mathsf{F}(t, x) - b_\mathsf{F}(t, x) \right) \rho(t, x) dx dt \\
& - \epsilon \int_0^1 \int_{\mathbb{R}^d} |\nabla \log \rho(t, x) - \nabla \log \hat{\rho}(t, x)|^2 \rho(t, x) dx dt,
\end{aligned}
\tag{2.41}
$$

and as a result

$$
\mathsf{KL}(\rho(1) \,\|\, \hat{\rho}(1)) \leq \frac{1}{4\epsilon} \int_0^1 \int_{\mathbb{R}^d} \left| \hat{b}_\mathsf{F}(t, x) - b_\mathsf{F}(t, x) \right|^2 \rho(t, x) dx dt.
\tag{2.42}
$$

Proof Similar to the proof of Lemma 21, we can use the FPE in (2.40) to compute $\frac{d}{dt}\mathsf{KL}(\rho(t) \,\|\, \hat{\rho}(t))$, which leads to the main result. Instead, we take a simpler approach, leveraging the result in Lemma 21. We re-write the Fokker-Planck equations in (2.40) as the (score-dependent) transport equations

$$
\begin{aligned}
\partial_t \rho &= -\nabla \cdot ((b_\mathsf{F} - \epsilon \nabla \log \rho)\rho), \\
\partial_t \hat{\rho} &= -\nabla \cdot ((\hat{b}_\mathsf{F} - \epsilon \nabla \log \hat{\rho})\hat{\rho}).
\end{aligned}
\tag{B.39}
$$

Applying Lemma 21 directly, we find that

$$
\mathsf{KL}(\rho(1) \,\|\, \hat{\rho}(1)) = \int_0^1 \int_{\mathbb{R}^d} \left[(\nabla \log \hat{\rho} - \nabla \log \rho) \cdot \left(\left[\hat{b}_\mathsf{F} - \epsilon \nabla \log \hat{\rho} \right] - [b_\mathsf{F} - \epsilon \nabla \log \rho] \right) \right] \rho \, dx dt.
\tag{B.40}
$$

Expanding the above,

$$
\begin{aligned}
\mathsf{KL}(\rho(1) \,\|\, \hat{\rho}(1)) = {} & \int_0^1 \int_{\mathbb{R}^d} \left[(\nabla \log \hat{\rho} - \nabla \log \rho) \cdot (\hat{b}_\mathsf{F} - b_\mathsf{F}) \right] \rho \, dx dt \\
& - \epsilon \int_0^1 \int_{\mathbb{R}^d} \left[|\nabla \log \rho - \nabla \log \hat{\rho}|^2 \right] \rho \, dx dt,
\end{aligned}
\tag{B.41}
$$

which proves the first part of the lemma. Now, by Young's inequality, it holds for any fixed $\eta > 0$ that

$$
\begin{aligned}
\mathsf{KL}(\rho_1 \,\|\, \hat{\rho}(1)) \leq {} & \frac{1}{2\eta} \int_0^1 \int_{\mathbb{R}^d} \left| \hat{b}_\mathsf{F} - b_\mathsf{F} \right|^2 \rho_t \, dx dt \\
& + \left(\tfrac{1}{2}\eta - \epsilon \right) \int_0^1 \int_{\mathbb{R}^d} |\nabla \log \rho_t - \nabla \log \hat{\rho}_t|^2 \rho_t \, dx dt.
\end{aligned}
\tag{B.42}
$$

65

Hence, for $\eta = 2\epsilon$,

$$\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{4\epsilon} \int_0^1 \int_{\mathbb{R}^d} \left| \hat{b}_{\mathsf{F}} - b_{\mathsf{F}} \right|^2 \rho dx dt, \tag{B.43}$$

which proves the second part of the lemma. ∎

Theorem 23 *Let ρ denote the solution of the Fokker-Planck equation (2.20) with $\epsilon(t) = \epsilon > 0$. Given two velocity fields $\hat{b}, \hat{s} \in C^0([0,1], (C^1(\mathbb{R}^d))^d)$, define*

$$\hat{b}_{\mathsf{F}}(t,x) = \hat{b}(t,x) + \epsilon \hat{s}(t,x), \qquad \hat{v}(t,x) = \hat{b}(t,x) + \gamma(t)\dot{\gamma}(t)\hat{s}(t,x) \tag{2.43}$$

where the function γ satisfies the properties listed in Definition 1. Let $\hat{\rho}$ denote the solution to the Fokker-Planck equation

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}_{\mathsf{F}} \hat{\rho}) = \epsilon \Delta \hat{\rho}, \qquad \hat{\rho}(0) = \rho_0. \tag{2.44}$$

Then,

$$\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon} \left(\mathcal{L}_b[\hat{b}] - \min_{\hat{b}} \mathcal{L}_b[\hat{b}] \right) + \frac{\epsilon}{2} \left(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}] \right), \tag{2.45}$$

where $\mathcal{L}_b[\hat{b}]$ and $\mathcal{L}_s[\hat{s}]$ are the objective functions defined in (2.13) and (2.16), and

$$\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon} \left(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}] \right) + \frac{\sup_{t \in [0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon)^2}{2\epsilon} \left(\mathcal{L}_s[\hat{s}] - \min_{\hat{v}} \mathcal{L}_s[\hat{s}] \right). \tag{2.46}$$

where $\mathcal{L}_v[\hat{v}]$ is the objective function defined in (2.29).

Proof Observe that by Proposition 18, the target density $\rho_1 = \rho(1, \cdot)$ is the density of the process X_t that evolves according to SDE (2.33). By Lemma 22 and an additional application of Young's inequality, we then have that

$$\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon} \int_0^1 \int_{\mathbb{R}^d} \left(|\hat{b} - b|^2 + \epsilon^2 |s - \hat{s}|^2 \right) \rho dx dt. \tag{B.44}$$

Using the definition of $\mathcal{L}_b[\hat{b}]$ and $\mathcal{L}_s[\hat{s}]$ in (2.13), and (2.16), we conclude that

$$\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon} \left(\mathcal{L}_b[\hat{b}] - \min_{\hat{b}} \mathcal{L}_b[\hat{b}] \right) + \frac{\epsilon}{2} \left(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}] \right), \tag{B.45}$$

which is (2.45). Using that $b = v - \gamma\dot{\gamma}s$ and $\hat{b} = \hat{v} - \gamma\dot{\gamma}\hat{s}$, we can write (B.43) in this case as

$$\mathsf{KL}(\rho_1 \parallel \hat{\rho}(1)) \leq \frac{1}{4\epsilon} \int_0^1 \int_{\mathbb{R}^d} |\hat{v} - v - (\gamma(t)\dot{\gamma}(t) - \epsilon)(\hat{s} - s)|^2 \rho dx dt, \tag{B.46}$$

$$\leq \frac{1}{2\epsilon} \int_0^1 \int_{\mathbb{R}^d} \left(|\hat{v} - v|^2 + (\gamma(t)\dot{\gamma}(t) - \epsilon)^2 |\hat{s} - s|^2 \right) \rho dx dt, \tag{B.47}$$

$$\leq \frac{1}{2\epsilon} \left(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}] \right) + \frac{\max_{t \in [0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon)^2}{2\epsilon} \left(\mathcal{L}_s[\hat{s}] - \min_{\hat{v}} \mathcal{L}_s[\hat{s}] \right), \tag{B.48}$$

which is the final part of the theorem. ∎

B.4 Proofs of Lemma 25 and Theorem 26

Lemma 56 *Given the velocity field $\hat{b} \in C^0([0,1],(C^1(\mathbb{R}^d))^d)$, let $\hat{\rho}$ satisfy the transport equation*

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\hat{\rho}) = 0, \tag{2.50}$$

and let $X_{s,t}(x)$ solve the ODE

$$\frac{d}{dt}X_{s,t}(x) = b(t, X_{s,t}(x)), \qquad X_{s,s}(x) = x, \qquad t, s \in [0,1] \tag{2.51}$$

Then, given the PDFs ρ_0 and ρ_1:

1. The solution to (2.50) for the initial condition $\hat{\rho}(0) = \rho_0$ is given at any time $t \in [0,1]$ by

$$\hat{\rho}(t,x) = \exp\left(-\int_0^t \nabla \cdot b(\tau, X_{t,\tau}(x))d\tau\right)\rho_0(X_{t,0}(x)) \tag{2.52}$$

2. The solution to (2.50) for the final condition $\hat{\rho}(1) = \rho_1$ is given at any time $t \in [0,1]$ by

$$\hat{\rho}(t,x) = \exp\left(\int_t^1 \nabla \cdot b(\tau, X_{t,\tau}(x))d\tau\right)\rho_1(X_{t,1}(x)) \tag{2.53}$$

Proof If $\hat{\rho}$ solves the TE (2.50) and $X_{s,t}$ solves the ODE (2.51), we have

$$\frac{d}{dt}\hat{\rho}(t, X_{s,t}(x)) = \partial_t \hat{\rho}(t, X_{s,t}(x)) + b(t, X_{s,t}(x)) \cdot \nabla \hat{\rho}(t, X_{s,t}(x))$$
$$= -\nabla \cdot b(t, X_{s,t}(x))\hat{\rho}(t, X_{s,t}(x)) \tag{B.49}$$

This equation implies that

$$\frac{d}{dt}\left(\exp\left(\int_s^t \nabla \cdot b(\tau, X_{s,\tau}(x))d\tau\right)\hat{\rho}(t, X_{s,t}(x))\right) = 0. \tag{B.50}$$

Integrating (B.50) over $[0,t]$ and setting $s = t$ in the result gives

$$\hat{\rho}(t,x) = \exp\left(-\int_0^t \nabla \cdot b(\tau, X_{t,\tau}(x))d\tau\right)\hat{\rho}(0, X_{t,0}(x)) \tag{B.51}$$

If we use the initial condition $\hat{\rho}(0) = \rho_0$, this gives (2.52). Similarly, Integrating (B.50) on $[t,1]$ and setting $s = t$ in the result gives

$$\hat{\rho}(t,x) = \exp\left(\int_t^1 \nabla \cdot b(\tau, X_{t,\tau}(x))d\tau\right)\hat{\rho}(1, X_{t,1}(x)) \tag{B.52}$$

If we use the final condition $\hat{\rho}(1) = \rho_1$, this gives (2.53).

∎

Theorem 26 *Given $\epsilon > 0$ and two velocity fields $\hat{b}, \hat{s} \in C^0([0,1], (C^1(\mathbb{R}^d))^d)$, define*

$$\hat{b}_{\mathsf{F}}(t,x) = \hat{b}(t,x) + \epsilon\hat{s}(t,x), \qquad \hat{b}_{\mathsf{B}}(t,x) = \hat{b}(t,x) - \epsilon\hat{s}(t,x), \tag{2.54}$$

and let Y_t^{F} and Y_t^{B} denote solutions of the following forward and backward SDEs:

$$dY_t^{\mathsf{F}} = b_{\mathsf{B}}(t, Y_t^{\mathsf{F}})dt + \sqrt{2\epsilon}dW_t, \tag{2.55}$$

to be solved forward in time from the initial condition $Y_{t=0}^{\mathsf{F}} = x$ independent of W; and

$$dY_t^{\mathsf{B}} = b_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt + \sqrt{2\epsilon}dW_t^{\mathsf{B}}, \quad W_t^{\mathsf{B}} = -W_{1-t}, \tag{2.56}$$

to be solved backwards in time from the final condition $Y_{t=1}^{\mathsf{B}} = x$ independent of W^{B}. Then, given the densities ρ_0 and ρ_1:

1. *The solution to the forward FPE*

   $$\partial_t\hat{\rho}_{\mathsf{F}} + \nabla \cdot (\hat{b}_{\mathsf{F}}\hat{\rho}_{\mathsf{F}}) = \epsilon\Delta\hat{\rho}_{\mathsf{F}}, \qquad \hat{\rho}_{\mathsf{F}}(0) = \rho_0, \tag{2.57}$$

   can be expressed at $t = 1$ as

   $$\hat{\rho}_{\mathsf{F}}(1,x) = \mathbb{E}_{\mathsf{B}}^x \left(\exp\left(-\int_0^1 \nabla \cdot \hat{b}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt \right) \rho_0(Y_{t=0}^{\mathsf{B}}) \right), \tag{2.58}$$

   where $\mathbb{E}_{\mathsf{B}}^x$ denotes expectation on the path of Y_t^{B} conditional on the event $Y_{t=1}^{\mathsf{B}} = x$.

2. *The solution to the backward FPE*

   $$\partial_t\hat{\rho}_{\mathsf{B}} + \nabla \cdot (\hat{b}_{\mathsf{B}}\hat{\rho}_{\mathsf{B}}) = -\epsilon\Delta\hat{\rho}_{\mathsf{B}}, \qquad \hat{\rho}_{\mathsf{B}}(1) = \rho_1, \tag{2.59}$$

   can be expressed at any $t = 0$ as

   $$\hat{\rho}_{\mathsf{B}}(0,x) = \mathbb{E}_{\mathsf{F}}^x \left(\exp\left(\int_0^1 \nabla \cdot \hat{b}_{\mathsf{B}}(t, Y_t^{\mathsf{F}})dt \right) \rho_1(Y_{t=1}^{\mathsf{F}}) \right), \tag{2.60}$$

   where $\mathbb{E}_{\mathsf{F}}^x$ denotes expectation on the path of Y_t^{F} conditional on $Y_{t=0}^{\mathsf{F}} = x$.

Proof Evaluating $d\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})$ via the backward Itô formula (2.36) we obtain

$$\begin{aligned}
d\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) &= \partial_t\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt + \nabla\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) \cdot dY_t^{\mathsf{B}} - \epsilon\Delta\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt \\
&= \partial_t\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt + \nabla\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) \cdot \hat{b}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt + \sqrt{2\epsilon}\nabla\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}} - \epsilon\Delta\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt \\
&= -\nabla \cdot \hat{b}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}})dt + \sqrt{2\epsilon}\nabla\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}}.
\end{aligned} \tag{B.53}$$

where we used (2.56) in the second step and (2.57) in the last one. This equation can be written as a total differential in the form

$$\begin{aligned}
d&\left(\exp\left(-\int_t^1 \nabla \cdot \hat{b}_{\mathsf{F}}(\tau, Y_\tau^{\mathsf{B}})d\tau \right) \hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) \right) \\
&= \sqrt{2\epsilon}\exp\left(-\int_t^1 \nabla \cdot \hat{b}_{\mathsf{F}}(\tau, Y_\tau^{\mathsf{B}})d\tau \right) \nabla\hat{\rho}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}},
\end{aligned} \tag{B.54}$$

which after integration over $t \in [0, 1]$ becomes

$$\hat{\rho}_{\mathsf{F}}(1, Y_{t=1}^{\mathsf{B}}) - \exp\left(-\int_0^1 \nabla \cdot \hat{b}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) dt\right) \rho_0(Y_{t=0}^{\mathsf{B}})$$
$$= \sqrt{2\epsilon} \int_0^1 \exp\left(-\int_t^1 \nabla \cdot \hat{b}_{\mathsf{F}}(\tau, Y_\tau^{\mathsf{B}}) d\tau\right) \nabla \hat{\rho}(t, Y_t^{\mathsf{B}}) \cdot dW_t^{\mathsf{B}}. \tag{B.55}$$

where we used $\hat{\rho}(0) = \rho_0$. Taking an expectation conditioned on the event $Y_{t=1}^{\mathsf{B}} = x$ and using that the term on the right-hand side has mean zero, we find that

$$\hat{\rho}_{\mathsf{F}}(1, Y_{t=1}^{\mathsf{B}}) - \mathbb{E}_{\mathsf{B}}^x \exp\left(-\int_0^1 \nabla \cdot \hat{b}_{\mathsf{F}}(t, Y_t^{\mathsf{B}}) dt\right) \rho_0(Y_{t=0}^{\mathsf{B}}) = 0. \tag{B.56}$$

This gives (2.58).

Similarly, evaluating $d\hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}})$ via Itô's formula, we obtain

$$d\hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) = \partial_t \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt + \nabla \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) \cdot dY_t^{\mathsf{F}} + \epsilon \Delta \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt$$
$$= \partial_t \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt + \nabla \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) \cdot \hat{b}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt + \sqrt{2\epsilon} \nabla \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) \cdot dW_t + \epsilon \Delta \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt$$
$$= -\nabla \cdot \hat{b}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt + \sqrt{2\epsilon} \nabla \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) \cdot dW_t. \tag{B.57}$$

where we used (2.55) in the second step and (2.59) in the last one. This equation can be written as a total differential in the form

$$d\left(\exp\left(\int_0^t \nabla \cdot \hat{b}_{\mathsf{B}}(\tau, Y_\tau^{\mathsf{F}}) d\tau\right) \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}})\right)$$
$$= \sqrt{2\epsilon} \exp\left(\int_0^t \nabla \cdot \hat{b}_{\mathsf{B}}(\tau, Y_\tau^{\mathsf{F}}) d\tau\right) \nabla \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) \cdot dW_t. \tag{B.58}$$

Integrating the above on $t \in [0, 1]$, we find that

$$\exp\left(\int_0^1 \nabla \cdot \hat{b}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt\right) \rho_1(Y_1^{\mathsf{F}}) - \hat{\rho}_{\mathsf{B}}(0, Y_{t=0}^{\mathsf{F}})$$
$$= \sqrt{2\epsilon} \int_0^1 \exp\left(\int_0^t \nabla \cdot \hat{b}_{\mathsf{B}}(\tau, Y_\tau^{\mathsf{F}}) d\tau\right) \nabla \hat{\rho}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) \cdot dW_t. \tag{B.59}$$

where we used $\hat{\rho}(1) = \rho_1$. Taking an expectation conditioned on the event $Y_{t=0}^{\mathsf{F}} = x$ and applying the Itô isometry, we deduce that

$$\mathbb{E}_{\mathsf{F}}^x\left(\exp\left(\int_0^1 \nabla \cdot \hat{b}_{\mathsf{B}}(t, Y_t^{\mathsf{F}}) dt\right) \rho_1(Y_{t=1}^{\mathsf{F}})\right) - \hat{\rho}_{\mathsf{B}}(0, y) = 0. \tag{B.60}$$

This gives (2.60).

∎

B.5 Proof of Theorem 31

Theorem 31 *Assume that $I(t, x_0, x_1) = x_0$ for $t \in [0, \delta]$ with some $\delta \in (0, 1]$. Given any $a > 0$, let*

$$u^{\mathsf{d}}(t, x, x_0) = \mathbb{E}_{x_1, z}\Big(\partial_t I(t, x_0, x_1) - \frac{\sqrt{2at}\, z}{\sqrt{1-t}} \Big| x_t = x\Big), \tag{3.10}$$

where x_t is given by (3.2) and where $\mathbb{E}_{x_1, z}(\cdot | x_t = x)$ denotes an expectation over $x_1 \sim \rho_1 \perp z \sim \mathsf{N}(0, Id)$ conditioned on $x_t = x$ with $x_0 \in \mathbb{R}^d$ fixed. Then $u^{\mathsf{d}}(\cdot, \cdot, x_0) \in C^0([0, 1]; (C^p(\mathbb{R}^d))^d)$ for any $p \in \mathbb{N}$ and $x_0 \in \mathbb{R}^d$. Moreover, the solutions to the forward SDE

$$dX_t^{\mathsf{d}} = u^{\mathsf{d}}(t, X_t^{\mathsf{d}}, x_0)dt + \sqrt{2a}\, dW_t, \qquad X_{t=0}^{\mathsf{d}} = x_0, \tag{3.11}$$

are such that $X_{t=1}^{\mathsf{d}} \sim \rho_1$.

Proof Let us first consider what happens on the interval $t \in [0, \delta]$ where $I(t, x_0, x_1) = x_0$ by assumption and the stochastic interpolant (3.2) with x_0 fixed and $a(t) = a > 0$ reduces to

$$x_t = x_0 + \sqrt{2at(1-t)}z \quad \text{with} \quad x_0 \text{fixed}, \ z \sim \mathsf{N}(0, Id). \tag{B.61}$$

The law of this process is simply

$$x_t \sim \mathsf{N}(x_0, 2at(1-t)Id), \tag{B.62}$$

which means that its density satisfies for all $t > 0$ the FPE

$$\partial_t \rho + 2at\nabla \cdot (s(t, x)\rho) = a\Delta\rho, \tag{B.63}$$

where $s(t, x) = \nabla \log \rho(t, x)$ is the score, which is explicitly given by

$$s(t, x) = -\frac{x - x_0}{2at(1-t)} \qquad \text{for} \ \ t \in (0, \delta]. \tag{B.64}$$

This means that the drift term in the FPE (B.63) can be written as

$$2ats(t, x) = -\frac{x - x_0}{1-t} \qquad \text{for} \ \ t \in (0, \delta]. \tag{B.65}$$

and this equality also holds in the limit as $t \to 0$. It is also easy to check that

$$-\frac{x - x_0}{2at(1-t)} = -\frac{\sqrt{2at}}{\sqrt{(1-t)}}\mathbb{E}(z | x_t = x) \qquad \text{for} \ \ t \in [0, \delta] \tag{B.66}$$

which is consistent with the drift in (3.10) on $t \in [0, \delta]$ since $\partial_t I(t, x_0, x_1) = 0$ on this interval. Since the right-hand side of (B.65) is nonsingular at $t = 0$ it also means that the SDE (3.11) is well-defined for $t \in [0, \delta]$ and the law of its solutions coincide with that of the process x_t defined (B.61), $X_t^{\mathsf{d}} \sim \mathsf{N}(x_0, 2at(1-t)Id)$ for $t \in [0, \delta]$.

Considering next what happens on $t \in (\delta, 1]$, since $\gamma(t) = \sqrt{2at(1-t)}$ is only zero at the endpoint $t = 1$ where we assume that the density ρ_1 satisfies Assumption 5 we can mimick all the arguments in the proof of of Theorem 6 and Corollaries 10 and 18 to terminate the proof. ∎

Note that this proof shows that is enough to have $\partial_t I(t = 0, x_0, x_1) = 0$, since this implies that $I(t, x_0, x_1) = x_0 + O(t^2)$. It also shows that, while it is key to use an SDE on $t \in [0, \delta]$ so that the generative process can spread the mass away from x_0, the diffusive noise is no longer necessary and we could switch back to a probability flow ODE on $t \in (\delta, 1]$ (using a time-dependent $a(t)$ to that effect with $a(t) = 0$ for $t \in (\delta, 1]$).

B.6 Proof of Lemma 40 and Theorem 41.

Lemma 57 *If Assumption (39) holds, then the solution $\rho(t)$ to (3.35) is the density of the stochastic interpolant*

$$x_t = T(t, \alpha(t)T^{-1}(0, x_0) + \beta(t)T^{-1}(1, x_1)) + \gamma(t)z, \tag{3.37}$$

as long as $\alpha^2(t) + \beta^2(t) + \gamma^2(t) = 1$.

Proof By definition of the map T in (3.36), if $x_0 \sim \rho_0$ and $x_1 \sim \rho_1$, then $T^{-1}(0, x_0) \sim \mathsf{N}(0, Id)$ and $T^{-1}(1, x_1) \sim \mathsf{N}(0, Id)$. As a result, since x_0, x_1, and z are independent, and $z \sim \mathsf{N}(0, Id)$, we have

$$\alpha(t)T^{-1}(0, x_0) + \beta(t)T^{-1}(1, x_1) + \gamma(t)z \sim \mathsf{N}(0, \alpha^2(t)Id + \beta^2(t)Id + \gamma^2(t)Id) = \mathsf{N}(0, Id) \tag{B.67}$$

where the second equality follows from the condition $\alpha^2(t) + \beta^2(t) + \gamma^2(t) = 1$. Therfore, using again the definition of the map T

$$T(t, \alpha(t)T^{-1}(0, x_0) + \beta(t)T^{-1}(1, x_1) + \gamma(t)z) \sim \rho(t), \tag{B.68}$$

and we are done. ∎

Theorem 41 *Pick some $\gamma : [0, 1] \to [0, 1)$ such that $\gamma(0) = \gamma(1) = 0$, $\gamma(t) > 0$ for $t \in (0, 1)$, $\gamma \in C^2((0, 1))$ and $\gamma^2 \in C^1([0, 1])$, and let $\hat{x}_t = \hat{I}(t, x_0, x_1) + \gamma(t)z$, with $x_0 \sim \rho_0$, $x_1 \sim \rho_1$, and $z \sim \mathsf{N}(0, Id)$ all independent. Consider the max-min problem over $\hat{I} \in C^1([0, 1], (C^1(\mathbb{R}^d \times \mathbb{R}^d))^d)$ and $\hat{u} \in C^0([0, 1], (C^1(\mathbb{R}^d))^d)$:*

$$\max_{\hat{I}} \min_{\hat{u}} \int_0^1 \mathbb{E}\left(\tfrac{1}{2}|\hat{u}(t, \hat{x}_t)|^2 - \left(\partial_t \hat{I}(t, x_0, x_1) + (\dot{\gamma}(t) - \epsilon \gamma^{-1}(t))z\right) \cdot \hat{u}(t, \hat{x}_t)\right) dt. \tag{3.38}$$

If Assumption 39 holds, then all the optimizers (I, u) of (3.38) are such that the density of the associated $x_t = I(t, x_0, x_1) + \gamma(t)z$ is the solution ρ to (3.35). Moreover, $u = \nabla\lambda$, with λ the solution to (3.35).

Proof Denote by $\hat{\rho}(t, x)$ the density of $\hat{x}_t$ and define the current $\hat{j} : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ as

$$\hat{j}(t, x) = \mathbb{E}\left(\partial_t \hat{I}_t + \gamma(t)z | x = \hat{x}_t\right)\hat{\rho}(t, x). \tag{B.69}$$

The max-min problem (3.38) can then be formulated as the constrained optimization problem:

$$\max_{\hat{\rho},\hat{j}} \min_{\hat{u}} \int_0^1 \int_{\mathbb{R}^d} \left(\tfrac{1}{2}|\hat{u}(t, x)|^2 \hat{\rho}(t, x) - \hat{u}(t, x) \cdot \hat{j}(t, x)\right) dxdt \tag{B.70}$$

$$\text{subject to:} \quad \partial_t \hat{\rho} + \nabla \cdot \hat{j} = \epsilon \Delta \hat{\rho}, \quad \hat{\rho}(t = 0) = \rho_0, \quad \hat{\rho}(t = 1) = \rho_1$$

To solve this problem we can use the extended objective

$$\max_{\hat{\rho},\hat{j}} \min_{\hat{u}} \left(\int_0^1 \int_{\mathbb{R}^d} \left(\tfrac{1}{2}|\hat{u}(t, x)|^2 \hat{\rho}(t, x) - \hat{u}(t, x) \cdot \hat{j}(t, x)\right) dxdt \right.$$

$$- \int_0^1 \int_{\mathbb{R}^d} \lambda(t, x) \left(\partial_t \hat{\rho}(t, x) + \nabla \cdot \hat{j}(t, x) - \epsilon \Delta \hat{\rho}(t, x)\right) dxdt \tag{B.71}$$

$$\left. + \int_{\mathbb{R}^d} \eta_0(x) \left(\hat{\rho}(0, x) - \rho_0(x)\right) dx - \int_{\mathbb{R}^d} \eta_1(x) \left(\hat{\rho}(1, x) - \rho_1(x)\right) dx \right)$$

71

where $\lambda(t,x)$, $\eta_0(x)$, and $\eta_1(x)$ are Lagrange multipliers used to enforce the constraints. The unique minimizer (ρ, j, λ) of this optimization problem solves the Euler-Lagrange equations:

$$
\begin{aligned}
&\partial_t \rho + \nabla \cdot j = \epsilon \Delta \rho, \quad \rho(t=0) = \rho_0 \quad \rho(t=1) = \rho_1 \\
&\partial_t \lambda + \tfrac{1}{2}|u|^2 = -\epsilon \Delta \lambda, \\
&j = u\rho \\
&u = \nabla \lambda
\end{aligned}
\tag{B.72}
$$

We can use the last two equations to write the first two as (3.35), with $u = \nabla\lambda$. Since under Assumption 39 there is an interpolant that realizes the density $\rho(t)$ that solves (3.35), we conclude that an optimizer (I, u) of the the max-min problem (3.38) exists. For any optimizer, I will be such that $\rho(t)$ is the density of $x_t = I(t, x_0, x_1) + \gamma(t)z$, and u will satisfy $u = \nabla\lambda$. ∎

B.7 Proof of Theorem 45

Theorem 45 *Let $t_j = j/N$ with $j \in \{1, \ldots, N\}$, set $X_1^{\mathsf{den}} = z$, and define for $j = 1, \ldots, N-1$,*

$$
X_{j+1}^{\mathsf{den}} = \frac{\beta(t_{j+1})}{\beta(t_j)} X_j^{\mathsf{den}} + \left(\alpha(t_{j+1}) - \frac{\alpha(t_j)\beta(t_{j+1})}{\beta(t_j)} \right) \eta_z^{\mathsf{os}}(t_j, X_j^{\mathsf{den}}).
\tag{5.14}
$$

Then, (5.14) is a consistent integration scheme for the probability flow equation (2.32) associated with the velocity field (4.16) expressed as in (4.19). That is, if $N, j \to \infty$ with $j/N \to t \in [0,1]$, then $X_j^{\mathsf{den}} \to X_t$ where

$$
\dot{X}_t = b(t, X_t) = \frac{\dot{\beta}(t)}{\beta(t)} X_t + \left(\dot{\alpha}(t) - \frac{\alpha(t)\dot{\beta}(t)}{\beta(t)} \right) \eta_z^{\mathsf{os}}(t, X_t), \quad X_{t=0} = z.
\tag{5.15}
$$

In particular, if $z \sim \mathsf{N}(0, Id)$, then $X_N^{\mathsf{den}} \to x_1 \sim \rho_1$ in this limit.

Proof Use

$$
\beta(t_{j+1})\beta^{-1}(t_j) = 1 + \dot{\beta}(t_j)\beta^{-1}(t_i)(t_{j+1} - t_j) + O\big((t_{j+1} - t_j)^2\big)
\tag{B.73}
$$

and

$$
\alpha(t_{j+1}) - \alpha(t_j)\beta(t_{j+1})\beta^{-1}(t_j) = \big(\dot{\alpha}(t_j) - \alpha(t_j)\dot{\beta}(t_j)\beta^{-1}(t_i)\big)(t_{j+1} - t_j) + O\big((t_{j+1} - t_j)^2\big)
\tag{B.74}
$$

to deduce that (5.14) implies

$$
\begin{aligned}
X_{j+1}^{\mathsf{den}} = X_j^{\mathsf{den}} &+ \dot{\beta}(t_j)\beta^{-1}(t_j)X_j^{\mathsf{den}}(t_{j+1} - t_j) \\
&+ \big(\dot{\alpha}(t_j) - \alpha(t_j)\dot{\beta}(t_j)\beta^{-1}(t_i)\big)\eta_z^{\mathsf{os}}(t_j, X_j^{\mathsf{den}})(t_{j+1} - t_j) + O\big((t_{j+1} - t_j)^2\big).
\end{aligned}
\tag{B.75}
$$

or equivalently

$$
\frac{X_{j+1}^{\mathsf{den}} - X_j^{\mathsf{den}}}{t_{j+1} - t_j} = \dot{\beta}(t_j)\beta^{-1}(t_j)X_j^{\mathsf{den}} + \big(\dot{\alpha}(t_j) - \alpha(t_j)\dot{\beta}(t_j)\beta^{-1}(t_i)\big)\eta_z^{\mathsf{os}}(t_j, X_j^{\mathsf{den}}) + O(t_{j+1} - t_j).
\tag{B.76}
$$

Taking the limit as $N, j \to \infty$ with $j/N \to t \in [0, 1]$, we recover (5.15) and deduce that $X_j^{\text{den}} \to X_t$. ∎

Note that the proof shows that the result of Theorem 45 also holds if we use a nonuniform grid of times t_j, $j \in \{1, \dots, N\}$.

B.8 Proof of Theorem 47

Theorem 47 *Consider the probability flow equation associated with (5.18),*

$$\frac{d}{dt} X_t^{\text{rec}}(x) = b^{\text{rec}}(t, X_t^{\text{rec}}(x)), \qquad X_{t=0}^{\text{rec}}(x) = x. \tag{5.21}$$

Then, all solutions are such that $X_{t=1}^{\text{rec}}(x_0) \sim \rho_1$ if $x_0 \sim \rho_0$. In addition, if Assumption (46) holds, the velocity field defined in (5.18) reduces to

$$b^{\text{rec}}(t, x) = \dot{\alpha}(t) N(t, x) + \dot{\beta}(t) X_{t=1}(N(t, x)) \tag{5.22}$$

and the solution to the probability flow ODE (5.21) is simply

$$X_t^{\text{rec}}(x) = \alpha(t)x + \beta(t) X_{t=1}(x). \tag{5.23}$$

Proof The first part of the statement can be established by following the same steps as in the proof of Theorem 6 and Corollary 10. For the proof of the second part, use first (5.17) written as $x_t^{\text{rec}} = M(t, z)$ in (5.22) to deduce that

$$b^{\text{rec}}(t, x_t^{\text{rec}}) = \dot{\alpha}(t) N(t, M(t, z)) + \dot{\beta}(t) X_{t=1}(N(t, M(t, z))) = \dot{\alpha}(t)z + \dot{\beta}(t) X_{t=1}(z) = \dot{x}_t^{\text{rec}} \tag{B.77}$$

This shows that (5.22) is the unique minimizer of (5.19). Next, use (5.23) written as $X_t^{\text{rec}}(x) = M(t, x)$ in (5.22) to deduce that

$$b^{\text{rec}}(t, X_t^{\text{rec}}(x)) = \dot{\alpha}(t) N(t, M(t, x)) + \dot{\beta}(t) X_{t=1}(N(t, M(t, x))) = \dot{\alpha}(t)x + \dot{\beta}(t) X_{t=1}(x) \tag{B.78}$$

This implies that (5.23) solves (5.21), and since the solution of this ODE is unique, we are done. ∎

Appendix C. Experimental Specifications

Details for the experiments in Section 7.1 are provided here. Feed forward neural networks of depth 4 and width 512 are used for each model of the velocities b, v, and s. Training was done for 7000 iterations on batches comprised of 25 draws from the base, 400 draws from the target, and 100 time slices. At each iteration, we used a variance reduction technique based on antithetic sampling, in which two samples $\pm z$ are used for each evaluation of the loss. The objectives given in (2.13) and (2.16) were optimized using the Adam optimizer. The learning rate was set to .002 and was dropped by a factor of 2 every 1500 iterations of training. To integrate the ODE/SDE when drawing samples, we used the Heun-based integrator as suggested in Karras et al. (2022).

	ImageNet 32×32	Flowers	Flowers (Mirror)
Dimension	32×32	128×128	128×128
# Training point	1,281,167	315,123	315,123
Batch Size	512	64	64
Training Steps	8×10^5	3.5×10^5	8×10^5
Hidden dim	256	128	128
Attention Resolution	64	64	64
Learning Rate (LR)	0.0002	0.0002	0.0002
LR decay (1k epochs)	0.995	0.995	0.985
U-Net dim mult	[1,2,2,2]	[1,1,2,3,4]	[1,1,2,3,4]
Learned t sinusoidal embedding	Yes	Yes	Yes
t_0, t_f for $t \sim \text{Unif}[t_0, t_f]$, learning η	[0.0, 1.0]	n/a	n/a
t_0, t_f for $t \sim \text{Unif}[t_0, t_f]$, learning s	n/a	[.0002, .9998]	[.0002, .9998]
t_0, t_f when sampling with ODE	[0.0, 1.0]	[0.0001, 0.9999]	[.0001, .9999]
t_0, t_f when sampling with SDE, η	[0.0, 0.97] + denoising	n/a	n/a
t_0, t_f when sampling with SDE, s	n/a	[.0001, .9999]	[.0001, .9999]
$\gamma(t)$ in x_t	$\sqrt{(t(1-t)}$	$\sqrt{(t(1-t)}$	$\sqrt{(10t(1-t)}$
EMA decay rate	0.9999	0.9999	0.9999
EMA start iteration	10000	10000	10000
# GPUs	2	4	2

Table 3: Hyperparameters and architecture for image datasets.

C.1 Image Experiments

Network Architectures. For all image generation experiments, the U-Net architecture originally proposed in Ho et al. (2020) is used. The specification of architecture hyperparameters as well as training hyperparameters are given in Table 3. The same architecture is used regardless of whether learning $b, v, s,$ or η.

When using the SDE and learning η_z, we found that integrating to a time slightly before $t_f = 1.0$ and using the denoising formula (5.14) beyond this point provided the best results, as described in the main text.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Michael S. Albergo, Mark Goldstein, Nicholas M. Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. *arXiv*, October 2023.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In Jacques Azéma and Marc Yor, editors, *Séminaire de Probabilités XIX 1983/84*, pages 177–206, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39397-9.

Heli Ben-Hamu, Samuel Cohen, Joey Bose, Brandon Amos, Maximilian Nickel, Aditya Grover, Ricky T. Q. Chen, and Yaron Lipman. Matching normalizing flows and probability paths on manifolds. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1749–1763. PMLR, 2022.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Nicholas M Boffi and Eric Vanden-Eijnden. Probability flow solution of the fokker–planck equation. *Machine Learning: Science and Technology*, 4(3):035012, jul 2023.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Charlotte Bunne, Ya-Ping Hsieh, Marco Cuturi, and Andreas Krause. The Schr\"odinger Bridge between Gaussian Measures has a Closed Form. *arXiv:2202.05722*, February 2022.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *ICML*, 2022a.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Scott Chen and Ramesh Gopinath. Gaussianization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023.

Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022b.

Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrödinger bridge. *SIAM Review*, 63(2):249–313, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations (ICLR)*, 2022.

Joseph L. Doob. *Classical potential theory and its probabilistic counterpart*, volume 262 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1984. ISBN 0-387-90881-1.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ahmed El Alaoui and Andrea Montanari. An information-theoretic view of stochastic localization. *IEEE Trans. Inf. Theory*, 68(11):7423–7426, 2022.

Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

Chris Finlay, Joern-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3154–3164. PMLR, 13–18 Jul 2020.

Jerome H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397):249–266, 1987.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019.

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. Curran Associates, Inc., 2020.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, 2023.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2078–2087. PMLR, 10–15 Jul 2018.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2021.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Aapo Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11(7):1739–1768, 1999.

Zahra Kadkhodaie and Eero P. Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser. *arXiv:2007.13640*, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022.

Young-Heon Kim and Emanuel Milman. A generalization of caffarelli's contraction theorem via (reverse) heat flow. *Mathematische Annalen*, 354:827–862, 2010.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. On density estimation with diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Yann LeCun, Sumit Chopra, and Raia Hadsell. A tutorial on energy-based learning. In Gökhan BakIr, Thomas Hofmann, Alexander J Smola, Bernhard Schölkopf, and Ben Taskar, editors, *Predicting structured data*, chapter 10. MIT press, 2007.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *arXiv:2206.06227*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona, editors, *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 946–985, 2023.

Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

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.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^{2}sb: Image-to-image schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023a.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv:2209.14577*, 2022.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv:2208.14699*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion ODEs by high order denoising score matching. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of fokker–planck equations through gradient–log–density estimation. *Entropy*, 22(8), 2020.

Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv:2305.10690*, 2023.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1447–1454, 2006.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9223–9232, May 2021.

Felix Otto. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2335–2344, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Stefano Peluchetti. Non-denoising forward-time diffusions. *https://openreview.net/forum?id=oVfIKuhqfC*, 2022.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

Eero P. Simoncelli and Edward H. Adelson. Noise removal via bayesian wavelet coring. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 1, pages 379–382 vol.1, 1996.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David

Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned diffusion schrödinger bridges. In *The 39th Conference on Uncertainty in Artificial Intelligence*, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yang Song and Diederik P. Kingma. How to train your energy-based models. *arXiv:2101.03288*, 2021.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1415–1428. Curran Associates, Inc., 2021a.

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*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. *ICML*, 2023.

Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135 – 1151, 1981.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*, 2023.

Esteban G. Tabak and Cristina V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217 – 233, 2010.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9526–9536. PMLR, 13–18 Jul 2020.

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. *arXiv:2302.00482*, 2023.

Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Log-sobolev suffices. *arXiv:1903.08568*, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

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

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, 2022.

Liu Yang and George Em Karniadakis. Potential flow generator with l2 optimal transport regularity for generative models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):528–538, 2022.

Linfeng Zhang, Weinan E, and Lei Wang. Monge-Ampère flow for generative modeling. *arXiv.org:1809.10188*, 2018.