Infinite-Dimensional Diffusion Models

Jakiw Pidstrigach   JAKIW.PIDSTRIGACH@STATS.OX.AC.UK
Institut für Mathematik
Universität Potsdam
Karl-Liebknecht-Str. 24/25
14476 Potsdam

Youssef Marzouk   YMARZ@MIT.EDU
Statistics and Data Science Center
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge, MA 02139 USA

Sebastian Reich   SEREICH@UNI-POTSDAM.DE
Institut für Mathematik
Universität Potsdam
Karl-Liebknecht-Str. 24/25
14476 Potsdam

Sven Wang   SVENWANG@MIT.EDU
Statistics and Data Science Center
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge, MA 02139 USA

Abstract

Diffusion models have had a profound impact on many application areas, including those where data are intrinsically infinite-dimensional, such as images or time series. The standard approach is first to discretize and then to apply diffusion models to the discretized data. While such approaches are practically appealing, the performance of the resulting algorithms typically deteriorates as discretization parameters are refined. In this paper, we instead directly formulate diffusion-based generative models in infinite dimensions and apply them to the generative modelling of *functions*. We prove that our formulations are well posed in the infinite-dimensional setting and provide *dimension-independent* distance bounds from the sample to the target measure. Using our theory, we also develop guidelines for the design of infinite-dimensional diffusion models. For image distributions, these guidelines are in line with current canonical choices. For other distributions, however, we can improve upon these canonical choices. We demonstrate these results both theoretically and empirically, by applying the algorithms to data distributions on manifolds and to distributions arising in Bayesian inverse problems or simulation-based inference.

Keywords: diffusion models, score-based generative models, infinite-dimensional analysis, hilbert spaces, bayesian inverse problems, function space

1. Introduction

Diffusion models (also score-based generative models or SGMs) (Sohl-Dickstein et al., 2015; Song et al., 2021) have recently shown great empirical success across a variety of domains.

In many applications, ranging from image generation (Nichol and Dhariwal, 2021; Dhariwal and Nichol, 2021), audio (Kong et al., 2021), and time series (Tashiro et al., 2021) to inverse problems (Kadkhodaie and Simoncelli, 2021; Batzolis et al., 2021), the signal to be modeled is actually a discretization of an *infinite-dimensional object* (i.e., a function of space and/or time). In such a setting, it is natural to apply the algorithm in high dimensions, corresponding to a fine discretization and a better approximation of the true quantity. Yet theoretical studies of current diffusion models suggest that performance guarantees deteriorate with increasing dimension (Chen et al., 2022; Bortoli, 2022).

When studying a discretization of an infinite-dimensional object, many application areas have found great success in directly studying the infinite-dimensional limit and only discretizing the problem in the last step, when implementing an algorithm on a computer. By accurately understanding the infinite-dimensional problem, one can gain valuable insights on how it should be discretized. Sometimes, this leads to algorithms that are *dimension-independent* in that their performance does not degrade when one chooses a finer discretization.

Important areas where it is now standard to study the infinite-dimensional object directly are, for example, Bayesian inverse problems (Stuart, 2010) and nonparametric statistics (Tsybakov, 2009; Giné and Nickl, 2015). Accordingly, many Markov chain Monte Carlo algorithms used for sampling, such as the Metropolis-adjusted Langevin (Cotter et al., 2013) or Hamiltonian Monte Carlo (Beskos et al., 2011) algorithms, have successfully been generalized to the infinite-dimensional setting; in addition to being an empirical success, these efforts have also led to dimension-independent convergence guarantees (see Hairer et al. (2014); Bou-Rabee and Eberle (2021); Pidstrigach (2022a)).

In the common implementation of the diffusion model algorithm, one first discretizes the data (for example images to pixels or wavelet coefficients, or functions to their evaluations on a grid) and then applies the algorithm in $\mathbb{R}^D$, as described in Song et al. (2021). When doing so, one does not consider the implications of the discretization dimension D. In particular, if there is no well-defined limiting algorithm as $D \to \infty$, one cannot expect the algorithm's performance to be stable as D becomes large. This instability can potentially be mitigated by *defining the diffusion model algorithm directly in infinite dimensions*, and studying its properties there. Once the algorithm is modified so that it exists in infinite dimensions, the discretized formulations that are implementable on a computer will possess *dimension-independent* properties.

1.1 Challenges in Extending Diffusion Models to Infinite Dimensions

Let us briefly recall the well-known finite-dimensional diffusion model setting. A forward SDE, typically an Ornstein–Uhlenbeck process, is used to diffuse the data μ_{data}:

$$\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}W_t, \qquad X_0 \sim \mu_{\text{data}}. \tag{1}$$

The densities of its marginal distributions are denoted by p_t. The following so-called "reverse SDE" will traverse the marginals of X_t backward:

$$\mathrm{d}Y_t = \frac{1}{2}Y_t\mathrm{d}t + \nabla \log p_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}W_t, \qquad Y_0 \sim p_T, \tag{2}$$

where W_t is a different Wiener process/Brownian motion than in (1). In particular, $Y_T \sim p_{T-T} = \mu_{\text{data}}$, where by a slight abuse of notation we denote both the density and the measure itself by μ_{data}. The goal of the diffusion model algorithm is to approximate paths of Y_t and use the realizations at time T as approximate samples from μ_{data}. Since the marginals of the forward SDE X_t converge to $\mathcal{N}(0, I)$ at an exponential speed, one can approximate the unknown term p_T by $\mathcal{N}(0, I)$. Furthermore, $\nabla \log p_t$ can be approximated using score-matching techniques (Vincent, 2011). There are three main challenges in generalizing this construction to infinite dimensions, which we highlight next.

1.1.1 CHOICE OF THE NOISING PROCESS

In finite dimensions, the process $(W_t : t \geq 0)$ in (1) is standard Brownian motion. Therefore, the noise increments for different coordinates i and j, e.g., $W_t^i - W_s^i$ and $W_t^j - W_s^j$, are independent and identically distributed. In infinite dimensions, one can associate a white-noise process W_t^U to each Hilbert space U, with the property that the coordinates of W_t^U in an orthonormal basis of U are independent and identically distributed. Therefore, one has to determine *which* white-noise process (i.e., which Hilbert space) to choose in the infinite-dimensional limit.

The common case of discretizing the infinite-dimensional target object to $\mathbb{R}^D$, e.g., discretizing a function onto a grid or real-life scenery into image pixels, and then choosing W_t as a standard Brownian motion on $\mathbb{R}^D$, means that at each grid point we will add independent noise values. In particular, as the grid grows finer, even for arbitrarily close values $a \approx b$, the evaluations $X_t(a)$ and $X_t(b)$ will be perturbed with independent noise. The limiting Wiener process will be $W_t^{L^2}$, i.e., the process associated to $U = L^2$, also called *space-time white noise*. We have depicted space-time white noise in Figure 1a.

In infinite dimensions, however, the choice of the noise process is a subtle issue, as it has a crucial impact on the space on which the diffusion process is supported. For instance, the above 'canonical' choice of space-time white noise will lead to X_t and W_t having such irregular samples that they are *not* supported in L^2 anymore. While we will also study such processes due to their widespread use in practice, we will see that other choices can be beneficial from a theoretical as well as a practical standpoint.

1.1.2 SCORE FUNCTION

The score function $\nabla \log p_t$ in (2) is typically defined via the Lebesgue density p_t of the law of X_t. Yet in infinite-dimensional vector spaces, the Lebesgue measure no longer exists; hence one can no longer specify the score functions in the same manner. Therefore, a key question is: How does one define and make sense of $\nabla \log p_t$ without relying on the notion of Lebesgue density, and still define an algorithm which provably samples from the correct measure?

1.1.3 DENOISING SCORE MATCHING OBJECTIVE

The score $\nabla \log p_t$ is typically approximated by a neural network $\tilde{s}(t, x)$ in some chosen neural network class, and identified by minimizing the denoising score-matching objective,

$$\text{Loss}(\tilde{s}) = \int_0^T \mathbb{E}[\|\nabla \log p_t(X_t) - \tilde{s}(t, X_t)\|_K^2]\mathrm{d}t,$$

over this class. Similarly to the choice of the noising process, it is not clear which Hilbert space K and norm $\|\cdot\|_K$ should be used for the analogous objective in infinite dimensions.

1.2 Contributions

Our paper, for the first time, formulates the diffusion model algorithm directly on infinite-dimensional spaces, and proves that this formulation is well-posed and satisfies crucial theoretical guarantees.

To formulate the reverse SDE in infinite dimensions, we must find a way to handle the $\nabla \log p_t$ term, as discussed in the last section. We do this *by replacing the score with a conditional expectation*, in Definition 2. This definition then carries over to the infinite-dimensional case. Furthermore, we are able to show under which circumstances one can generalize the denoising score matching objective to identify the neural network $\tilde{s}(t, x)$ in Lemma 7.

To justify approximating the reverse SDE to obtain samples from μ_{data}, we proceed in multiple steps. First, in Theorem 9, we show that the time-reversal of the forward SDE also satisfies an appropriate reverse SDE. The terminal condition of the reverse SDE will have distribution μ_{data}. To simulate this reverse SDE in practice, however, both its initial conditions and drift must be approximated. In Lemma 7 we establish under which conditions we can use the common denoising score matching objective to approximate the drift of the reverse SDE in infinite dimensions.

Second, we prove that the solution to such an SDE exists for general initial conditions— and in particular, for our approximate initial conditions. Moreover, we prove that the solution is unique; otherwise we could be approximating a different reverse SDE solution that does not sample μ_{data} at the terminal time T. We provide rigorous uniqueness results under two distinct scenarios: first, in Theorem 12, for μ_{data} which satisfies a manifold hypothesis; and second, in Theorem 13, under the assumption that μ_{data} has density with respect to a Gaussian measure. The first case is relevant for the typical use cases of diffusion models, as image data are usually supported on lower-dimensional manifolds or other substructures. The second case is relevant when, for example, applying diffusion models to Bayesian inverse problems or related problems of simulation-based inference.

Finally, building upon the preceding results, we establish *dimension-independent* convergence rates in Theorem 14. Our bound is quantitative and shows how the distance relies on different choices made in the diffusion model algorithm.

The theory described above guides choices for the noise process W_t^U and the loss norm $\|\cdot\|_K$. Both will depend on the properties of μ_{data}. In Section 6 we discuss the implications of the theory for implementing diffusion models in infinite dimensions. In Section 6.1, we work out guidelines for choosing W_t^U and K for a given μ_{data}. In Section 6.2, we study the case of image distributions and see that our theorems indeed apply for the typical properties of μ_{data} one expects in that setting; hence, we have proven that the standard diffusion model algorithm is well-defined for image distributions as $D \to \infty$. Moreover, we see that the choices $W_t^U = W_t^{L^2}$ and $K = L^2$ actually follow the guidelines developed in the preceding subsection. Therefore, the canonical choices made for diffusion models seem to be good default choices for image distributions.

For μ_{data} with different smoothness properties, however, the insights from our theory dictate other choices for U and K. In Section 7 we apply our guidelines to two specific data distributions μ_{data}. Our principled algorithms are compared to the common ad hoc implementation of diffusion models. These numerical findings confirm our theoretical insights: our modifications outperform the canonical choices, and the ways in which they do can be explained by the discussion in Section 6.

1.3 Related Work

The two efforts most related to ours are the concurrent works Hagemann et al. (2023) and Lim et al. (2023).

In Hagemann et al. (2023), methods are developed to train diffusion models simultaneously on multiple discretization levels of (infinite-dimensional) functions. They build upon our Wasserstein distance bounds to show that their multilevel approach is consistent.

Lim et al. (2023) are also able to generalize the trained model over multiple discretization levels. They propose to run the annealed Langevin algorithm in infinite dimensions, and use existing results for infinite-dimensional Langevin algorithms to justify their algorithms theoretically. The forward-reverse SDE framework is not treated.

Both of these efforts encounter difficulties when defining the infinite-dimensional score. Hagemann et al. (2023) circumvent this issue by only treating time-reversals of the *discretized* forward SDE. Lim et al. (2023), on the other hand, only analyze the case in which the measure is supported on the Cameron–Martin space of W_t^U. One can then simplify the problem by working with densities of X_t with respect to Gaussian measure. From a practical point of view, both of these works employ Fourier neural operators (Li et al., 2020) as their neural network architecture, while we work directly in the space domain and use the popular U-Net architecture for our neural networks.

In Kerrigan et al. (2022) an infinite-dimensional time-discrete version of the diffusion model algorithm is proposed. It is not studied whether the proposed algorithm is well defined in infinite dimensions.

Other works also transform data into a representation that is well suited to functions, e.g., by applying a wavelet (Guth et al., 2022; Phung et al., 2022) or spectral (Phillips et al., 2022) transform. After the transformation, however, these works employ the finite-dimensional formulation of the diffusion model algorithm; infinite-dimensional limits are not treated. We discuss how different spatial discretization schemes can be related to our results in Section 4.

Lastly, the subject of convergence of diffusion models to the target distribution has been a very active field of research recently; see Chen et al. (2022, 2023); Bortoli (2022); Lee et al. (2022); Yang and Wibisono (2022). In all these works, however, bounds on the distance to the target measure depend at least linearly on the discretization dimension D, rendering them vacuous in infinite dimensions.

1.4 A Primer on Probability in Hilbert Spaces

In this section, we will give a short summary of key concepts relating to probability theory on infinite-dimensional (Hilbert) spaces which are required to study the infinite-dimensional

formulation of SGMs rigorously. For an extensive introduction to this topic, see Hairer (2009, Chapter 3).

1.4.1 GAUSSIAN MEASURES ON HILBERT SPACES

Let $(H, \langle \cdot, \cdot \rangle_H)$ be a separable Hilbert space. We then say that a random variable X taking values in H is Gaussian if, for every $v \in H$, the real-valued random variable $\langle v, X \rangle_H$ is also Gaussian. If the $\langle v, X \rangle_H$ have mean zero, X is *centered*. The covariance operator of X is the symmetric, positive-definite operator $C : H \to H$ defined through

$$\langle g, Ch \rangle_H = \operatorname{Cov}(\langle X, g \rangle_H, \langle X, h \rangle_H) = \mathbb{E}_X[\langle X, g \rangle_H \langle X, h \rangle_H]. \tag{3}$$

We denote the law of X in this case by $\mathcal{N}(0, C)$. Since X takes values in H, C is guaranteed to be compact (Hairer, 2009). Therefore, there exists an orthonormal basis $(e_i : i \geq 1)$ of eigenvectors of C satisfying $Ce_i = c_i e_i$. Fixing this basis, the second moment of X is given by

$$\mathbb{E}[\|X\|_H^2] = \mathbb{E}\left[\sum_{i=1}^{\infty} \langle X, e_i \rangle_H^2\right] = \sum_{i=1}^{\infty} \mathbb{E}[\langle X, e_i \rangle_H^2] = \sum_{i=1}^{\infty} \langle e_i, Ce_i \rangle_H = \sum_{i=1}^{\infty} c_i.$$

Since a Gaussian measure is supported on H if and only if its second moment on H is finite (Hairer, 2009), and X takes values in H, the trace of C, $\operatorname{tr}(C) = \sum_{i=1}^{\infty} c_i$, will be finite. We then also say that C is of *trace class*. Note that this is not the case if one would choose $C = \operatorname{Id}$, since its trace is infinite. However, one could always just consider a larger space $H' \supset H$, such that H' supports $\mu := \mathcal{N}(0, C)$ and on which C would then have finite trace.

1.4.2 THE CAMERON–MARTIN SPACE

The covariance operator C plays a special role in that it characterizes the 'shape' of the Gaussian measure $\mathcal{N}(0, C)$. Indeed, one may define another canonical inner product space U associated to C, which is a (compactly embedded) subspace $U \subseteq H$ called the *Cameron–Martin space* of $\mathcal{N}(0, C)$. Intuitively speaking, with respect to the geometry of U, a random variable $X \sim \mathcal{N}(0, C)$ will have 'identity' covariance. Assuming that C is non-degenerate, the Cameron–Martin space is defined via the inner product

$$\langle g, h \rangle_U = \langle g, C^{-1}h \rangle_H = \langle C^{-1/2}g, C^{-1/2}h \rangle_H.$$

Since C^{-1} is unbounded, U is indeed a smaller space than H; more specifically, one can show that $U = C^{1/2}H$. In order to generate a realization of $X \sim \mathcal{N}(0, C)$, one may simply draw i.i.d. coefficients $(\xi_i \sim \mathcal{N}(0, 1) : i \geq 1)$ and set $X = \sum_{i=1}^{N} c_i^{1/2} \xi_i e_i$ where $(c_i, e_i)_{i=1}^{\infty}$ are the eigenpairs of C.[1]

It is important to note that X almost surely does *not* take values in U. As an example, let $H = L^2([0, 1])$, and consider a one-dimensional Brownian motion process $(B_t : t \in [0, 1])$. Of course, $B \in H$ almost surely. The Cameron–Martin space of B, however, is given as the space $U = H^1([0, 1])$ of weakly differentiable functions on $[0, 1]$. Since the sample paths of B are almost surely nowhere differentiable (Karatzas et al., 1991), we conclude that almost

---

1. This is also called the Karhunen–Loève expansion of X, and in finite dimensions relates to the simple fact that $C^{-1/2}X \sim \mathcal{N}(0, \operatorname{Id})$.

surely $B \notin U$.[2] Nevertheless, U does indicate the regularity of the Gaussian process at hand: the more regular U, the more regular the draws from the corresponding Gaussian measure.

1.4.3 C-Wiener processes in Hilbert spaces

The standard Brownian motion in $\mathbb{R}^D$ has increments $W_{t+\Delta t} - W_t \sim \mathcal{N}(0, \Delta t \mathrm{I}_D)$. However, for the case of a general infinite-dimensional Hilbert space H, the meaning of an identity covariance matrix depends on the choice of the scalar product with respect to which the Gaussian measure has identity covariance. Therefore, we will from now on fix two Hilbert spaces: the Cameron–Martin space U, with respect to which the increments of the Wiener process would have covariance $\Delta t I$, and a larger space H on which W_t^U takes values and has covariance operator C, i.e.

$$W_{t+\Delta t}^U - W_t^U \sim \mathcal{N}(0, \Delta t C).$$

In general, we will pick H large enough so that all of our objects take values in it (the target measure μ_{data} as well as the C-Wiener process W_t^U). The choice of U can then also be seen as being equivalent to choosing a covariance operator C of W_t^U on H.

1.4.4 Interpretation in finite dimensions

Given a Gaussian distribution $\mathcal{N}(0, C)$ on $\mathbb{R}^D$, its Cameron–Martin space will be again $\mathbb{R}^D$, but equipped with the scalar product

$$\langle x, y \rangle_U = \langle C^{-1/2} x, C^{-1/2} y \rangle_{\mathbb{R}^D} = x^T C^{-1} y.$$

Plugging U into definition (3), one sees that X has an identity covariance matrix with respect to U. If $X \sim \mathcal{N}(0, C)$, then it can also be represented as $\sqrt{C} Z$, for $Z \sim \mathcal{N}(0, \mathrm{I}_D)$. Similarly, a C-Wiener process with increments $\mathcal{N}(0, C)$ in finite dimensions can be constructed by using a standard Brownian motion W_t on $\mathbb{R}^D$ and multiplying it with $\sqrt{C}$.

Therefore, in finite dimensions, most of the discussions above can be simplified to choosing covariance matrices and representing objects of interest in terms of standard Gaussians (Z) or Brownian motions (W_t). The main technical difficulties in infinite dimensions arise because one has to choose a Hilbert space H on which Z would have the standard normal distribution, and because Z will not take values in H.

However, in infinite dimensions, one can still understand most concepts that relate to the choice of Gaussian measures by simply thinking about some large Hilbert space H' in which all quantities of interest take values and then identifying Gaussian random variables with their covariance operators on this space.

2. The Infinite-Dimensional Forward and Reverse SDEs

We will now formulate the forward and reverse SDEs of our generative model in infinite dimensions, and show that the reverse SDE is, in fact, well-posed with the correct terminal distribution.

---

2. Here, we have used that functions in H^1 are absolutely continuous, and therefore almost everywhere differentiable on $[0, 1]$.

To this end, let μ_{data} be our target measure, supported on a separable Hilbert space $(H, \langle \cdot, \cdot \rangle)$. Our goal is to generate samples from μ_{data}, which is done by first adding noise to given samples from μ_{data} using a forward SDE and then generating new samples using a learned reverse SDE (Song et al., 2021).

2.1 Forward SDE

We now define the infinite-dimensional forward SDE used to 'diffuse' the initial measure μ_{data}. As noted in Section 1.4, there is no natural Brownian motion process in infinite dimensions; instead there is one white noise process W_t^U for each Hilbert space U. From now on, we fix some Cameron–Martin space U, together with its Gaussian measure $\mathcal{N}(0, C)$. Furthermore, let H be large enough to not only support μ_{data}, but also $\mathcal{N}(0, C)$. In practice, an example would be to choose a Gaussian process (GP) with a Matérn covariance $\mathcal{N}(0, C)$ (which implicitly defines U). As the embedding space H, one could for example choose L^2. Then W_t^U would have increments that are samples from a Matérn GP.

We then define the forward SDE as

$$\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}W_t^U = -\frac{1}{2}X_t\mathrm{d}t + \sqrt{C}\mathrm{d}W_t^H, \qquad X_0 \sim \mu_{\text{data}}. \tag{4}$$

The marginal distributions of X_t will converge to the stationary distribution $\mathcal{N}(0, C)$ as $t \to \infty$ (Da Prato and Zabczyk, 2014, Theorem 11.11). We will denote the marginal distributions of X_t by $\mathbb{P}_t$.

The choice of U, or equivalently C, can be guided by the theory that we will develop and strongly impacts empirical performance. We discuss these choices in Section 6.

2.2 Definition of the Score Function

Analogously to score-based generative models in finite dimensions, we now wish to define the reverse SDE corresponding to (4); this SDE on H should approximately transform $\mathcal{N}(0, C)$ to μ_{data}. This can be achieved by time-reversing the SDE (4). In the finite-dimensional case, the drift of the time reversal SDE involves the score function $\nabla \log p_t$ (see (2)), where p_t is the density of $\mathbb{P}_t$ with respect to Lebesgue measure. More precisely, in the finite-dimensional case $H = \mathbb{R}^D$, the reverse SDE to the Ornstein–Uhlenbeck process

$$\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \sqrt{C}\mathrm{d}W_t$$

is given by

$$\mathrm{d}Y_t = \frac{1}{2}Y_t\mathrm{d}t + C\nabla \log p_{T-t}(Y_t)\mathrm{d}t + \sqrt{C}\mathrm{d}W_t\,;$$

see (Haussmann and Pardoux, 1986). In the infinite-dimensional case, the density p_t is no longer well-defined, since there is no Lebesgue measure. Hence, we need another way to make sense of the score function. Interestingly, in finite dimensions, there is an alternative way to express $C\nabla_H \log p_t$ via conditional expectations which is amenable to generalization to infinite dimensions.

8

Lemma 1 *Assume the finite-dimensional setting $H = \mathbb{R}^D$. Denote by p_t the Lebesgue density of X_t, where $X_{[0,T]}$ is a solution to (4). Then, we can express the function $C\nabla \log p_t$ as*

$$
\begin{aligned}
C\nabla \log p_t(x) &= -\frac{1}{1-e^{-t}}\left(\mathbb{E}\left[X_t - e^{-\frac{t}{2}}X_0 \mid X_t = x\right]\right) \\
&= -\frac{1}{1-e^{-t}}\left(x - e^{-\frac{t}{2}}\mathbb{E}[X_0 \mid X_t = x]\right)
\end{aligned}
$$

for $t > 0$, where $\mathbb{E}[f(X_\tau) \mid X_t = x]$ is the conditional expectation of the function $f(X_\tau)$ given $X_t = x$ and $\tau \in [0, T]$.

Conditional expectations are also *well-defined in infinite dimensions*. Therefore, we will give the conditional expectation from Lemma 1 a name and make use of it as the drift of the reverse SDE on Hilbert space H:

Definition 2 *Let H be a possibly infinite-dimensional Hilbert space and $X_{[0,T]}$ a solution to (4). We define the reverse drift as a map $s : [0, T] \times H \to H$,*

$$
s(t, x) := -\frac{1}{1-e^{-t}}\left(x - e^{-\frac{t}{2}}\mathbb{E}[X_0|X_t = x]\right).
$$

Remark 3 *For a definition of conditional expectations and measures for Hilbert-space valued random variables, see Bogachev (1997, Section 1.3).*

Remark 4 *Note that the above function is only defined up to $\mathbb{P}_t$-equivalence classes, where $\mathbb{P}_t$ is the distribution of the time-t marginal of the forward SDE. However, the loss function for diffusion models is a L^2 loss, integrated over $\mathbb{P}_t$. Therefore, without restricting the function class that one optimizes over, the minimizer is also only defined up to $\mathbb{P}_t$-equivalence. Neural networks are normally contained in the class of continuous functions in t and x. We will see that we can pick versions of $s(t, x)$ satisfying continuity properties, for example being locally Lipschitz continuous in x (see Section 3.2).*

We will also frequently use the fact that the drift of the reverse SDE is actually a rescaled martingale in reverse time. We will later show that this also holds in infinite dimensions, in Theorem 9.

Lemma 5 *Assume the finite-dimensional setting $H = \mathbb{R}^D$. Then, the quantity $M_t = e^{-t/2}\nabla \log p_t(X_t)$ is a time-continuous reverse time martingale, i.e.,*

$$
\nabla \log p_t(X_t) = e^{\frac{(t-\tau)}{2}}\mathbb{E}[\nabla \log p_\tau(X_\tau) \mid X_t] \quad \text{for all } 0 < \tau < t.
$$

The proofs of both of these lemmas can be found in Appendix D.1.

2.3 Reverse SDE

We are now able to write down the infinite-dimensional forward SDE:

$$
X_0 \sim \mu_{\text{data}}, \qquad \mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \sqrt{C}\mathrm{d}W_t^H, \tag{5}
$$

where $W_t^U = \sqrt{C}W_t^H$ are C-Wiener processes. Defining $Y_s := X_{T-t}$ as the time-reversal of a solution to (5), we want to show that it satisfies the following stochastic differential equation:

$$Y_0 \sim \mathbb{P}_T, dY_t = \frac{1}{2}Y_t dt + s(T-t, Y_t)dt + \sqrt{C}dW_t^H. \tag{6}$$

Here, the drift $s(t,x)$ of the reverse SDE is given by Definition 2.

In finite dimensions, one could also rewrite the reverse SDE as

$$\begin{aligned}
dY_t &= \frac{1}{2}Y_t dt + C\nabla \log p_{T-t}(X_t)dt + \sqrt{C}dW_t^H \\
&= \frac{1}{2}Y_t dt + C\nabla \log \frac{dp_{T-t}}{d\mathcal{N}(0,C)}(X_t)dt + C\nabla \log \mathcal{N}(0,C)(X_t) + \sqrt{C}dW_t^H \\
&= -\frac{1}{2}Y_t dt + C\nabla \log \frac{dp_{T-t}}{d\mathcal{N}(0,C)}(X_t)dt + \sqrt{C}dW_t^H,
\end{aligned} \tag{7}$$

where we denote by $\mathcal{N}(0,C)(x)$ the density of $\mathcal{N}(0,C)$ of evaluated at x. In finite as well as in infinite dimensions, if $X_0 \sim \mu_{\text{data}}$ has a density with respect to a Gaussian $\mathcal{N}(0,C)$, then so will the distribution of X_t (see the proof of Theorem 13). Therefore, under that assumption, the SDE in the last line of (7) can also be made sense of in infinite dimensions. Rewriting the SDE in this form is helpful in the proof of Theorem 13.

Remark 6 *Another forward SDE with invariant measure $\nu = \mathcal{N}(0,C)$ is*

$$dX_t = -\frac{1}{2}C^{-1}X_t dt + dW_t^H, \tag{8}$$

with corresponding reverse SDE

$$dY_t = -\frac{1}{2}C^{-1}Y_t dt + \nabla_H \log \frac{d\mathbb{P}_{T-t}}{d\mathcal{N}(0,C)}(Y_t)dt + dW_t^H. \tag{9}$$

The operator C^{-1} can often be identified with a differential operator, turning (9) into a stochastic partial differential equation (SPDE). One can then use numerical tools for SPDEs to approximate the above. This constitutes an interesting direction for future work.

Note, however, that if C^{-1} is an unbounded operator, for any fixed positive time $t > 0$, the high frequencies of Y_t will already have been smoothed out by the process. Since the reverse SDE has to be discretized, care must be taken on how to train and evaluate the diffusion model; otherwise one might lose all high-frequency information. By 'high frequencies,' here we mean the eigenvectors corresponding to large eigenvalues of C^{-1}.

2.4 Training Loss

To simulate the reverse SDE, we need a way to approximately learn the drift function $s(t,x)$. For another function $\tilde{s}(t,x)$ (a candidate approximation to s), we measure the goodness of the fit of $\tilde{s}$ using a score-matching objective, i.e.,

$$\text{SM}_t(\tilde{s}) = \mathbb{E}[\|s(t,X_t) - \tilde{s}(t,X_t)\|_K^2]. \tag{10}$$

On $\mathbb{R}^D$, the norm $\|\cdot\|_K$ to measure the misfit is typically the Euclidean norm. For training, this loss can be rewritten into the denoising score matching objective,

$$\text{DSM}_t(\tilde{s}) = \mathbb{E}[\|\tilde{s}(t, X_t) - (1 - e^{-t})^{-1/2}(X_t - e^{-t/2}X_0)\|_K^2] = \text{SM}_t(\tilde{s}) + V_t. \qquad (11)$$

One can then show (see (Vincent, 2011)), that SM and DSM only differ by a constant V_t and therefore one can use DSM as an optimization objective to optimize SM. The DSM loss is normally optimized on a sequence of times $\{t_m\}_{m=1}^M$ on which the reverse SDE is discretized (since the score will only be evaluated at these t_i values), i.e.,

$$\begin{aligned}
\text{Loss}(\tilde{s}) &= \sum_{m=1}^M \text{SM}_{t_m}(\tilde{s}) = \sum_{m=1}^M \text{DSM}_{t_m}(\tilde{s}) - V_{t_m} \\
&= \mathbb{E}_{t_m, X}[\|\tilde{s}(t_m, X_{t_m}) - \sigma_{t_m}^{-1}(X_{t_m} - e^{-t_m/2}X_0)\|_K^2] - V,
\end{aligned} \qquad (12)$$

where the last expectation is taken over $t_m \in \text{Unif}(\{t_1, \ldots, t_M\})$ and $V = \sum_{m=1}^M V_{t_m}$.

We will see that the equivalence of SM_t and DSM_t does not hold in general in infinite dimensions. Furthermore, we will study the choice of the norm $\|\cdot\|_K$. Two natural choices that come to mind are the norm of the embedding Hilbert space H and of the Cameron–Martin space U of C. In the following lemma, we study conditions under which we can rewrite the loss into the denoising score matching objective.

Lemma 7 *Let $(K, \langle \cdot, \cdot \rangle_K)$ be a separable Hilbert space. Furthermore, denote by $\tilde{s}$ an approximation to s, such that the score matching objective* (10) *is finite. Then,*

$$\text{SM}_t(\tilde{s}) = \text{DSM}_t(\tilde{s}) - V_t,$$

where DSM_t is defined in (11) *and V_t is given by the conditional variance of X_0,*

$$V_t = \frac{e^{-t}}{1 - e^{-t}}\mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t]\|_K^2].$$

Furthermore, DSM_t is infinite if V_t is.

Lemma 7 shows that in infinite dimensions there is the possibility that the true objective SM, which we are trying to optimize, might be finite, while DSM is not. One might argue that this is not relevant since in practice one always has to discretize and then both will be finite. However, as we will argue in the following paragraph, the discretization level will impact the variance of the gradients. In practice, we do not evaluate the full expectation values in SM or DSM, but take Monte Carlo estimates in the form of mini-batches. Assuming that we have already reached the optimum, i.e., $s = \tilde{s}$, then the SM objective would be zero and also the gradient of any mini-batch taken to approximate it would be zero. However, derivatives of Monte Carlo estimates of the DSM objective will have the form

$$\partial_{\theta_i} \text{DSM}(\tilde{s}_\theta) = \frac{1}{M} \sum_{i=1}^M \langle \partial_{\theta_i} \tilde{s}(t, x_t^M), \tilde{s}(t, X_t) - \sigma_t^{-1}(X_t - e^{-t/2}X_0) \rangle,$$

where me made the parameters θ (typically, the weights of a neural network) of $\tilde{s}_\theta$ explicit. The random variable $\tilde{s}(t, X_t) - \sigma_t^{-1}(X_t - e^{-t/2}X_0)$ has infinite variance, and therefore we

can expect the above gradient estimates to have infinite variance too. Hence, if V_t is not finite and therefore DSM_t is not finite in infinite dimensions, one can expect variance of the the gradient of the DSM to get arbitrarily large as the discretization gets finer, despite the fact that the true gradient should be zero. In the following lemma, we study some cases in which we can expect V to be finite.

Lemma 8 *The denoising score matching objective* (11) *is finite in infinite dimensions if one of the following two conditions holds:*

1. *We use the Cameron–Martin norm* $\| \cdot \|_K = \| \cdot \|_U$ *in the objective, and* μ_{data} *is supported on the Cameron–Martin space* U *of* $\mathcal{N}(0, C)$ *and has finite second moment, i.e.,*

$$\mathbb{E}[\|X_0 - \mathbb{E}[X_0]\|_U^2] < \infty.$$

2. *Both* μ_{data} *and* $\mathcal{N}(0, C)$ *are supported on* K.

A consequence of point 2 of Lemma 8 is that the norm of the embedding Hilbert space $K = H$ is always a valid choice. The proof of both lemmas above can be found in Appendix D.2.

3. Well-Posedness of the Reverse SDE

We need to show that the reverse SDE possesses solutions and that they are unique in order to prove that the reverse SDE samples from the target distribution in infinite dimensions. In Section 3.1, we show that the time-reversal of the forward SDE satisfies the reverse SDE in infinite dimensions and therefore samples the right final distribution μ_{data} at its final time. In Section 3.2, we will show strong uniqueness and existence of the reverse SDE for general initial conditions.

3.1 The Time Reversal Satisfies the Reverse SDE

Thus far, we have *formally* formulated the reverse SDE (6) without showing that it actually constitutes a time reversal of the stochastic dynamics from the forward equation. In the following theorem, we show that Y_t indeed constitutes a time reversal of X_t and that it recovers the correct target distribution at terminal time T.

Theorem 9 *Assume* X_t *is a solution to* (4). *Then, the time reversal* $Y_t := X_{T-t}$ *solves the SDE* (6). *Furthermore, if* H *is a Hilbert space such that* μ_{data} *and* $\mathcal{N}(0, C)$ *are both supported on* H, *we can choose* s *such that* $M_t = s(t, X_t)$ *is almost surely continuous in* t *with respect to the* H-norm.

Proof (Sketch) We approximate the infinite-dimensional forward-SDE in finite dimensions using a spectral approximation in the eigenbasis of the covariance operator C. The finite-dimensional approximations are denoted by X_t^D.

Next we show that the finite-dimensional time-reversals $Y_t^D := X_{T-t}^D$ satisfy an equation analogous to (6):

$$Y_t^D - Y_0^D - \frac{1}{2} \int_0^t Y_r^D \mathrm{d}r - \int_0^t s_{T-r}^D \mathrm{d}r = \sqrt{C^D} B_t^D.$$

We then show that all of those terms converge to their counterparts in (6), and therefore $Y_t := X_{T-t}$ satisfies the same equation. The convergence of Y_t^D to Y_t is trivial: the Y_t^D are spectral approximations. The convergence of the other terms is a bit more involved. Unfortunately, for $L > D$, the conditional expectations s_t^D are not the projections of s_t^L to a lower-dimensional space, and the same holds for the Brownian motions B_t^D.

We can, however, show that s_t^D is a martingale in D. Combining this with the fact that $e^{-t/2} s_t^D$ is also a martingale in time (see Lemma 5), we obtain uniform-in-time convergence of s_t^D to s_t. The convergence of $C^D B_t^D$ also follows, and we can identify the limit as a C-Wiener process.

The full proof can be found in Appendix F.1. ∎

Remark 10 *The work Föllmer and Wakolbinger (1986) studies time-reversal of more general forward SDEs. Due to the more general setting, the resulting SDE is only expressed coordinate-wise, and the SDE as well as the assumptions are more technical. Using our approach and the reverse drift $s(t, x)$, we prove that we can still use the common denoising score matching loss to approximate $s(t, x)$; see Lemma 7. Another related concept is vector logarithmic derivatives, as discussed in Bogachev (1997).*

Due to Theorem 9, we know that there is a solution to (6) that will sample μ_{data} at the final time. To motivate approximating (6) for sampling from μ_{data}, we also need to show that these solutions are unique; otherwise there could be other solutions that have different terminal conditions. We will achieve this in the following section.

3.2 Uniqueness and Existence of Solutions

We now study strong uniqueness and existence of the solutions to the reverse SDE. We say an SDE satisfies *strong existence* if we can construct a solution to the SDE for any driving Brownian motion and that solution will be adapted to the filtration of the Brownian motion. We say that an SDE satisfies *strong uniqueness* if, for any two solutions Y_t and $\tilde{Y}_t$ of that SDE, with the same driving Brownian motion, it holds that $\mathbb{P}[Y_t = \tilde{Y}_t \text{ for all } t] = 1$.

Remark 11 *Here we will prove strong uniqueness (also called pathwise uniqueness) of solutions to the reverse SDE. For sampling purposes, uniqueness in law of the reverse SDE would suffice and is generally easier to prove. However, for the Wasserstein distance bounds which we will prove later (see Theorem 14) we will employ coupling arguments. These arguments implicitly rely on strong existence of solutions to the reverse SDE and therefore we will prove strong existence. Strong existence together with uniqueness in law already imply strong uniqueness; see Karatzas et al. (1991, Section 5.3) (the result also holds here since H is separable). Therefore, in our case we can obtain strong uniqueness no matter which uniqueness we prove.*

We will treat two different settings. The first setting is tailored to distributions supported on substructures of the full space. The main motivation for this setting are measures which are supported on a manifold-like structure $\mathcal{M}$. Since many distributions that diffusion models are applied to satisfy the manifold hypothesis, understanding how diffusion

models interact with manifolds has been an active area of research (Pidstrigach, 2022b; De Bortoli, 2022; Batzolis et al., 2022).

Theorem 12 *Fix a covariance operator C in the forward SDE* (5) *together with its Cameron–Martin space U. Assume that the support of μ_{data} is contained in a ball B_R in U of radius $R \geq 0$:*

$$B_R = \{x \ : \ \|x\|_U \leq R\}.$$

Then there is a version of s which is Lipschitz continuous with respect to the Cameron–Martin norm, i.e.,

$$\|s(t,x) - s(t,y)\|_U \leq L_t \|x - y\|_U, \tag{13}$$

where $L_t \in \mathbb{R}^+$ is a time-dependent Lipschitz constant. Moreover, the reverse SDE with the Lipschitz continuous version of $s(t,x)$ has a unique strong solution.

Proof (Sketch) The transition kernel of the forward SDE is given by

$$p_t(x_0, \cdot) \sim \mathcal{N}(e^{-t}x_0, v_t C),$$

where we used the shorthand notation $v_t = 1 - e^{-t}$. If x_0 is an element of the Cameron–Martin space U of C, then the transition kernel is absolutely continuous with respect to $\mathcal{N}(0, (1 - e^{-t})C)$. The explicit formula for the density is

$$n_t(x_0, x) = \frac{\mathrm{d}\mathcal{N}(e^{-t}x_0, v_t C)}{\mathrm{d}\mathcal{N}(0, v_t C)}(x)$$

by the Cameron–Martin theorem. Since μ_{data} almost surely takes values in U, one can use the above to derive an explicit expression for the conditional expectation $\mathbb{E}[X_0|X_t = x]$ in terms of these densities:

$$\mathbb{E}[X_0|X_t = x] = \frac{\int x_0 n_t(x_0, x)\mathrm{d}\mu_{\text{data}}(x_0)}{\int n_t(x_0, x)\mathrm{d}\mu_{\text{data}}(x_0)}.$$

This formula can be used to derive local Lipschitzness of

$$s(t,x) = -\frac{1}{1 - e^{-t}}x + \frac{e^{-\frac{t}{2}}}{1 - e^{-t}}\mathbb{E}[X_0|X_t = x].$$

Interestingly, the local Lipschitzness is in terms of the norm of U. Even if x and y themselves are not in U, if their difference in is U, the U-norm of the difference $s(t,x) - s(t,y)$ will be bounded by (13). Taking some care, one can still use a fixed point argument to obtain existence, but not uniqueness. One can then apply Grönwall's lemma to obtain uniqueness.

Note that obtaining *weak* uniqueness would be easier, since under our assumptions the drift $s(t,x)$ will always map to the Cameron–Martin space of the C-Wiener process and one could apply a Girsanov-type argument.

The full proof can be found in Appendix F.1. ∎

The other case of interest is applying diffusion models to Bayesian inverse problems or simulation-based inference. In this case, we assume that the true measure is given as a density with respect to a Gaussian reference measure. We treat it in the theorem below:

Theorem 13 *Fix a covariance operator C in the forward SDE (5). Assume μ_{data} is given as*

$$\mu_{\text{data}} \propto \exp(-\Phi(x)) \mathrm{d}\mathcal{N}(0, C_\mu).$$

Let $(H, \langle \cdot, \cdot \rangle_H)$ be a Hilbert space on which $\mathcal{N}(0, C_\mu)$ is supported and C is bounded. For the potential $\Phi \in C^1(H)$ we assume,

- $\Phi(x) \geqslant E_0$,

- $\Phi(x) \leqslant E_1 + E_2 \|x\|^2$, *and*

- $\|\nabla \Phi(x) - \nabla \Phi(y)\| \leqslant L\|x - y\|$,

where the gradient is the H-gradient. Then there is a version of $s(t, x)$ that is locally Lipschitz continuous with respect to the H-norm for each t, i.e., for $\|x\|, \|y\| \leq r$ there is a $L_{t,r} < \infty$ such that

$$\|s(t, x) - s(t, y)\| \leq L_{t,r} \|x - y\|,$$

and the reverse SDE with the locally Lipschitz continuous version of $s(t, x)$ has a unique strong solution.

Proof (Sketch) The proof holds for any C that is diagonalizable with respect to the same eigenbasis as C_μ (in particular, also for $C = \text{Id}$, i.e., H-white noise), but we will only treat the less technical case $C = C_\mu$ here.

In the case of $C = C_\mu$, the distribution $\mathbb{P}_t$ of X_t is absolutely continuous with respect to $\mathcal{N}(0, C)$. One can rewrite the reverse SDE as in (7). It will then hold that

$$\nabla_{x_t} \log \frac{\mathrm{d}p_t}{\mathrm{d}\mathcal{N}(0, C)}(x_t) = \mathbb{E}[C\nabla\Phi(X_0)|X_t = x_t],$$

and the proof will mainly translate the Lipschitzness properties of $\nabla\Phi(x)$ to $\mathbb{E}[\nabla\Phi(X_0)|X_t = x]$. The global Lipschitzness of $\nabla\Phi(x)$ only induces local Lipschitzness of $\mathbb{E}[\nabla\Phi(X_0)|X_t = x]$, but that is enough to apply a Grönwall argument and deduce strong uniqueness.

Furthermore, one can obtain weak existence to the reverse SDE. By Theorem 9, the time reversal will be a weak solution with initial condition $\mathbb{P}_T$. However, under the assumptions of the theorem, $\mathcal{N}(0, C)$ will be absolutely continuous with respect to p_T. Therefore, one can obtain a weak solution with initial conditions $\mathcal{N}(0, C)$ by reweighting the time reversal. However, weak existence together with strong uniqueness already imply strong uniqueness; see Karatzas et al. (1991, Section 5.3).

The full proof can be found in Appendix F.2. ∎

4. Algorithms and Discretizations

We state simplified versions of our proposed algorithms in Algorithms 1 and 2. There are many potential modifications one might make to the above algorithms, as for example discussed in Song et al. (2021); Song and Ermon (2020); Ho et al. (2020); we do not include these here since they are not the focus of the current work. To implement any algorithm on

---
Algorithm 1 Training

Require: Covariance operator C
Require: Training data $\{X^n\}_{n=1}^N$
Require: Loss Norm $\|\cdot\|$
Require: Batch size B
Require: Discretization grid $\{t_1, \ldots, t_M\}$
  1: **while** Metrics not good enough **do**
  2:     Sample $\{\xi^i\}_{i=1}^B \sim \mathcal{N}(0, C)$ i.i.d.
  3:     Subsample $\{x_0^i\}_{i=1}^B$ from $\{X^n\}_{n=1}^N$
  4:     Sample $t^i \in \text{Unif}(\{t_1, \ldots, t_M\})$
  5:     $x_t^i \leftarrow e^{-t^i} x_0^i + \sqrt{1 - e^{-t^i}} \xi^i$
  6:     $\text{Loss}(\theta) = \sum_{i=1}^B \|\tilde{s}_\theta(t^i, x_t^i) - \frac{1}{\sqrt{1 - e^{-t^i}}} \xi^i\|^2$
  7:     Perform gradient step on Loss.
  8: **end while**

Algorithm 2 Sampling

Require: Covariance operator C
Require: Discretization grid $\{t_1, \ldots, t_M\}$
Require: Number of samples to generate L
  1: $\{x_M^i\}_{i=1}^L \sim \mathcal{N}(0, C)$
  2: **for** $m \leftarrow M, \ldots, 1$ **do**
  3:     $\Delta t \leftarrow t_m - t_{m-1}$
  4:     Sample $\{\xi^i\}_{i=1}^L \sim \mathcal{N}(0, C)$ i.i.d.
  5:     $x_{m-1}^i \leftarrow x_m^i + \Delta t\, \tilde{s}_\theta(t_m, x_m^i) + \sqrt{\Delta t}\, \xi^i$
  6: **end for**
  7: **return** $\{x_M^i\}_{i=1}^M$

---

a computer, the functions have to be discretized in some way. Discretization also interacts with the covariance matrix C, as the same covariance matrix has different meanings in different discretizations. We discuss this briefly now.

If the functions are discretized on a grid, i.e., if the samples are of the form $\{f(x_d)\}_{d=1}^D$ for a fixed grid $\{x_d\}$, choosing an identity covariance matrix corresponds to adding independent noise at each grid point x_d. The limiting object of the noise as the grid gets finer is space-time white noise (recall Section 1.1). Furthermore, the Euclidean norm on $\mathbb{R}^D$ in the loss function (11) will correspond to using the L^2 loss in the limit—i.e., the Cameron–Martin norm of the noising process.

In $\mathbb{R}^D$ the choice of the white noise process is equivalent to choosing a covariance matrix C and adding $\sqrt{C}\mathrm{d}W_t$ with a standard $\mathbb{R}^D$-valued Brownian motion W_t. Any correlated Wiener noise process W_t^U can be represented in this way on $\mathbb{R}^D$. If one wants the limit of $\sqrt{C}\mathrm{d}W_t$ to be a Gaussian process, one needs to plug in for C the kernel matrix of that Gaussian process on the grid $\{x_d\}_{d=1}^D$. Alternatively, one can also use one of many available libraries to generate Gaussian process realizations for common kernels (such as Matérn or squared exponential).

Note that the meaning of C depends on the discretization. If f is discretized with respect to some basis e_i of a space U, then using the identity covariance matrix corresponds to using the white noise process with Cameron–Martin space U. Therefore, discretizing the functions in a wavelet or Fourier basis will also result in space-time white noise as these both form an orthonormal basis of L^2 (under the common scaling of the basis vectors). However, if one does not want to work in the spatial domain, one can also just discretize the functions in an orthonormal basis of the Cameron–Martin space U of the noise one is targeting. Therefore, we can translate the approaches in Guth et al. (2022); Phillips et al. (2022); Phung et al. (2022) into our setting.

5. Bounding the Distance to the Target Measure

We now study how far the samples generated by the diffusion model algorithm lie from the true target measure μ_{data}. We do this in the Wasserstein-2-distance,

$$
\mathcal{W}_2(\mu, \nu) = \left(\inf_{\kappa \in Q(\mu,\nu)} \int \|x - y\|_H^2 \mathrm{d}\kappa(x,y) \right)^{1/2},
$$

where κ runs over all measures on $H \times H$ which have marginals μ and ν. The Wasserstein-2 distance in some sense "lifts" the distance induced by $\|\cdot\|_H$ to the space of measures. In the following theorem, we give an upper bound for the Wasserstein distance between the sample measure and the true data-generating measure. The bound holds irrespective of $\|\cdot\|_H$, giving us the freedom to study how different choices of $\|\cdot\|_H$ affect the distance bound.

Theorem 14 *We denote the covariance of the forward noising process by C. Let $(H, \|\cdot\|_H)$ be any Hilbert space such that the support of μ_{data} and $\mathcal{N}(0,C)$ are contained in H. Assume that $\|\cdot\|_H$ is at least as strong as the norm $\|\cdot\|_K$ used in the training of the diffusion model (see (12)), i.e.,*

$$
\|x\|_H \leq a\|x\|_K
$$

for some constant a. Further, assume that $s(t,x)$ is Lipschitz on H with constant L, i.e.,

$$
\|s(t,x) - s(t,y)\|_H \leq L\|x - y\|_H
$$

and that the reverse SDE has a strong solution (see Theorem 12 or 13 for the requirements). Let the reverse SDE (6) be discretized using an exponential integrator (see Appendix C). Then,

$$
\mathcal{W}_2(\mu_{\text{data}}, \mu_{\text{sample}}) \leqslant \left(\exp(-T/2)\, \mathcal{W}_2(\mu_{\text{data}}, \mathcal{N}(0,C)) + \varepsilon_{\text{Num}}^{1/2} + a\varepsilon_{\text{Loss}}^{1/2} \right) \exp\left(\frac{1}{4}L^2 T \right),
$$
(14)

where $\varepsilon_{\text{Loss}}$ is the value of the loss objective (12) and ε_{Num} denotes the error due to the numerical integration procedure,

$$
\varepsilon_{\text{Num}} = O(\Delta t) \sup_{0 < t \leq T} \mathbb{E}_{X_t \sim p_t}[\|s(t, X_t)\|_H^2].
$$

Proof (Sketch) We define two strong SDE solutions: Y_t, which is a solution to (6) with the correct drift $s(t,x)$ and started in $\mathbb{P}_T$; and $\tilde{Y}_t$, which uses the approximate drift $\tilde{s}$ and is started in $\mathcal{N}(0,I)$.

Both solutions are run to time T. We couple them by using the same Brownian motion process for both and starting them in $\mathcal{W}_2$-optimally coupled initial conditions.

We then obtain a bound on $\mathbb{E}[\|Y_T - \tilde{Y}_T\|_H^2]$. Since we know that $Y_T \sim \mu_{\text{data}}$, $\tilde{Y}_T \sim \mu_{\text{sample}}$ by definition, this gives us a coupling between μ_{data} and μ_{sample} and therefore upper bounds the Wasserstein-2 distance between those two.

We make use of the fact that the score is a martingale to obtain an upper bound for the numerical integration error, depending only on the quantity $\sup_{0 < t \leq T} \mathbb{E}_{X_t \sim p_t}[\|s(t, X_t)\|_H^2]$.

The full proof can be found in Appendix G. ∎

Since the choice of the embedding space $(H, \|\cdot\|_H)$ is left open in Theorem 14, we briefly discuss the implications of that choice. Controlling the Wasserstein distance with respect to a stronger underlying norm always implies the same Wasserstein-bound w.r.t. any weaker underlying norm. Of course, there is the possibility to obtain a better bound by directly applying the theorem for a weaker norm.

Picking stronger norms for H will in general result in the Wasserstein distance also factoring in differences in sample smoothness as well as deviations in function values. For example, picking $H = L^2$ means that the bound only implies closeness of the function evaluations while, for example, samples being too rough is not factored in. Picking positive Sobolev spaces will punish deviations in function values and deviations in the derivatives of the samples from the true samples. Picking negative Sobolev spaces for H (as we will need to for the common implementation of diffusion models; see Section 6.2) means that the samples are only close in a distributional sense. See Section B in the appendix for further discussion.

6. Implementing Infinite-Dimensional Diffusion Models

Our theory yields several suggestions on how one should design infinite-dimensional diffusion model algorithms. Section 6.1 gives a concise summary of those design principles, while Section 6.2 discusses the extent to which those design principles align with common implementations of diffusion models. In Section 7, we then show how to implement the guidelines in some explicit examples.

The two main design choices we will discuss are

1. The choice of the forward noising process W_t^U in (4).

2. The choice of the norm $\|\cdot\|_K$ in the denoising score matching objective in (12).

The first choice (of U and hence W_t^U) is equivalent to the choice of an invariant Gaussian distribution $\mathcal{N}(0, C)$, and we will use these choices interchangeably. In general, picking a C such that $\mathcal{N}(0, C)$ produces smoother samples corresponds to picking a smaller space U, while rougher samples correspond to larger Cameron–Martin spaces U; see also the examples in Section 7.1 and Figure 1. Furthermore, after discretizing the problem to finite dimensions, the choice of U or C corresponds to nothing else than specifying a covariance matrix C, i.e., W_t is replaced by $\sqrt{C}W_t$ in the diffusion processes. See Section 4 for more details.

After discretization, the second design choice of $\|\cdot\|_K$ corresponds to specifying a loss norm of the form $\|K^{-1/2}\cdot\|$ in place of the typical Euclidean norm $\|\cdot\|$.

6.1 Practical Implementation Guidance

6.1.1 MATCH C TO μ_{data}

First, we begin by pointing out the implications of Theorem 14 on the choice of C, or equivalently, U. The term $\mathcal{W}_2(\mu_{\text{data}}, \mathcal{N}(0, C))$ appearing in the error bound (14) clearly indicates choosing C such that $\mathcal{N}(0, C)$ is as close as possible to μ_{data}.

6.1.2 Choosing C such that we can pick a strong H-norm

Second, as discussed at the end of Section 5, we would like to choose as strong an $\|\cdot\|_H$-norm as possible in Theorem 14. This suggests not picking C too rough (U too large) as the norm space H in Theorem 14 has to support $\mathcal{N}(0, C)$.

This last points seems to suggest that we would want to pick C as smooth as possible to allow for stronger H-norms. However, since H has to support μ_{data}, there is a restriction on how strong an H-norm can be chosen. Therefore, this suggests matching $\mathcal{N}(0, C)$ to μ_{data} so that they are supported on the same space H, similarly to our first observation.

6.1.3 Choosing the loss-norm $\|\cdot\|_K$

The H-norm in Theorem 14 also has to be stronger than the loss norm $\|\cdot\|_K$, again suggesting choosing K as small as possible. However, besides numerical issues, also here there are lower bounds on how strong we can choose K.

To that end, we take another look at Lemma 8 in which we study two separate cases. In the first case, if C is rough enough such that U contains the support of μ_{data}, we can choose the Cameron–Martin norm of $\mathcal{N}(0, C)$ as the loss norm, i.e., $K = U$. In the second case, K has to support both $\mathcal{N}(0, C)$ and μ_{data}, which is the same condition as for the space H chosen in Theorem 14.

Hence, there are predominantly two natural ways to design the algorithm:

1. Choose $\mathcal{N}(0, C)$ as smooth as possible / U as small as possible, but large enough such that the support of μ_{data} is contained in its Cameron–Martin space U. Then choose the loss norm $\|\cdot\|_K$ in (12) equal to the Cameron–Martin norm, $K = U$. This algorithm design is called *Infinite-Dimensional Diffusion Model 1 (IDDM1)*

2. Match C to the data, i.e., choose $\mathcal{N}(0, C)$ such that its samples are as similar to the samples from μ_{data} as possible. Then choose the loss norm $\|\cdot\|_K$ in (12) such that it supports both μ_{data} and $\mathcal{N}(0, C)$. Let us call this algorithm design *Infinite-Dimensional Diffusion Model 2 (IDDM2)*.

Note that by Theorem 12, if not much is known about the distribution, and if in particular it might be supported on manifold-like structures, we must pick U large enough to contain the support of μ_{data} anyway. We will therefore use IDDM1 in these cases; see Section 7.2. If one has more structural information, for example the knowledge that μ_{data} has density with respect to a Gaussian measure, we will use IDDM2; see Section 7.3.

6.2 Image Distributions and White Noise Diffusion Models

The common implementation of the diffusion model algorithm will converge as $D \to \infty$ to $U = L^2$, i.e., use space-time white noise in the forward noising process. Furthermore, the loss function will also approach the L^2 loss, which means we are in the setting where we use the Cameron–Martin norm in the loss. For more details, see Section 4. We will call this algorithm *White Noise Diffusion Model (WNDM)*.

If μ_{data} is an image distribution, we can expect it to lie on a manifold, or more generally some lower-dimensional substructure. Furthermore, since the function values of an image are bounded on $[0, 1]$, the image samples are all contained in L^2. Therefore, we are in the

setting of Theorem 12, where the data are contained in the Cameron–Martin space U of the noise. Furthermore, we can apply Lemma 8 (bullet point 1) to see that we can use the Cameron–Martin norm, i.e., the L^2 norm, and obtain a well-defined denoising score-matching objective. Therefore, under these assumptions, we have shown that applying WNDM to image distributions *has a well-defined infinite-dimensional limit*.

Coming back to the discussion in Section 6.1 (in particular the design guidance for IDDM1), however, our theory suggests that we should try to pick U as small as possible, while still containing the typical image distribution. However, this U cannot be too regular, since images are quite irregular—for example, they can be discontinuous. If we identify images with two-dimensional functions, then already the L^2-Sobolev spaces H^α of order $\alpha > 1$ only contain continuous functions. Therefore, on the Sobolev scale ($H^\alpha : -\infty < \alpha < \infty$), the 'optimal' Cameron–Martin space would possess regularity of at most $\alpha = 1$. In light of this, setting $U = H^0 = L^2$ indeed seems like a natural choice that is close to matching the maximal possible regularity. Strikingly, this is in line with the huge empirical success of the WNDM algorithm for image distributions. To further refine the optimal choice of the space U beyond L^2 is an interesting avenue, both for theoretical and empirical future study.

7. Numerical Illustrations

In this section, we illustrate our results through numerical experiments. We sample functions defined on $[0, 1]$, and we discretize this spatial domain into a uniform grid with $D = 256$ evenly spaced points. For other discretization schemes, see the discussion in Section 4. We employ a grid-based spatial discretization since it allows us to use the popular U-Net architecture. Other common discretization schemes 'whiten' the data, rendering the convolutional layers of the U-Net unnecessary. For implementation details, see Appendix A.

Section 7.1 introduces some common preliminaries needed for both of the subsequent numerical examples. Section 7.2 then compares various diffusion model constructions in the setting of distributions that are not defined via Gaussian reference distributions, but rather supported on submanifolds of the infinite-dimensional space. Section 7.3 demonstrates the use of infinite-dimensional diffusion models for solving Bayesian inverse problems via a simulation-based (i.e., conditional sampling) approach.

7.1 Families of Gaussian Measures

We first construct a family $(\pi^\alpha, H^\alpha : -\infty < \alpha < \infty)$ of Gaussian measures π^α and their Cameron–Martin spaces H^α. This construction allows us to interpolate between measures with different sample smoothness and compare between the algorithms described in Section 6.1 and the canonical implementation of diffusion models described in Section 6.2.

To that end, we fix an orthonormal basis e_k of $L^2([0, 1])$. Then, we construct a family $(\pi^\alpha : -\infty < \alpha < \infty)$ of Gaussian measures as the distributions of

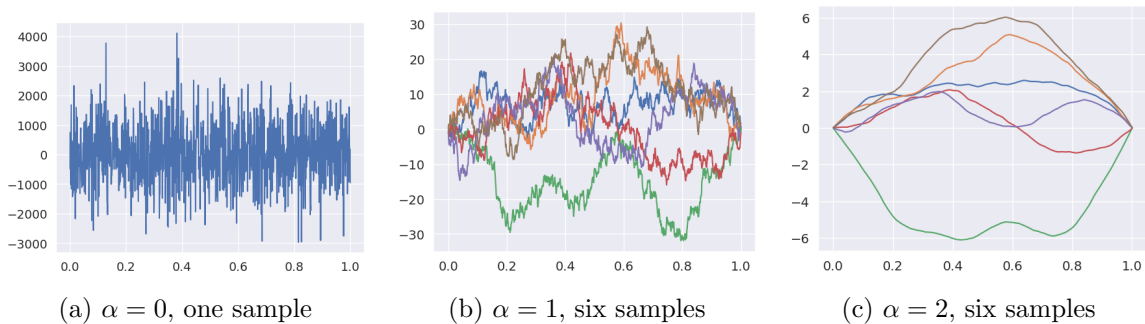$$\sum_{k=1}^{\infty} k^{-\alpha} Z_k e_k \sim \pi^\alpha, \tag{15}$$

(a) $\alpha = 0$, one sample   (b) $\alpha = 1$, six samples   (c) $\alpha = 2$, six samples

Figure 1: In each panel, we plot samples from π^α for different values of α, where π^α is defined in Section 7.1. We chose $e_k(\cdot) = \sqrt{2}\sin(2\pi k \cdot)$ as an orthonormal basis of L^2. For $\alpha = 0$ we see a sample of space-time white noise, where no function value is correlated to any of its neighboring function values. For $\alpha = 1$ and our specific choice of e_k, the sampled measure is the Brownian bridge measure.

where $Z_k \sim \mathcal{N}(0,1)$ i.i.d. The Cameron–Martin space of π^α is denoted by H^α and has norm

$$\|x\|_\alpha^2 = \sum_{k=1}^\infty k^{2\alpha} \langle x, e_k \rangle_{L^2([0,1])}^2. \tag{16}$$

Note that $H^0 = L^2([0,1])$ and therefore π^0 is space-time white noise. Furthermore, $H^\alpha \subset H^\beta$ for $\alpha > \beta$. As we have discussed before, samples of π^α will (almost surely) not be elements of the corresponding Cameron–Martin space H^α. Nevertheless, the distribution π^α is supported on $H^{\alpha - \kappa}$ as long as $\kappa > \frac{1}{2}$; see Beskos et al. (2011, Proposition 3.1).

The exact form of samples of π^α depends on the chosen basis e_k in (15). In our examples H^α will be L^2-Sobolev spaces with either zero or periodic boundary conditions. For the case of zero boundary conditions, we have visualized samples of π^α for different values of α in Figure 1.

As discussed in Section 6.1, we want to study two main modeling choices: First, we must select a Gaussian measure $\mathcal{N}(0, C)$, or equivalently its Cameron–Martin space U, for the noising process. We do that by fixing an α_{noise} and setting

$$U = H^{\alpha_{\text{noise}}}.$$

Second, a loss norm $\|\cdot\|_K$ must be chosen. We do so by fixing an α_{loss} and setting

$$K = H^{\alpha_{\text{loss}}}.$$

As recommended in Section 6.1, these choices should depend on the structure of μ_{data}. Therefore, we choose an α_{data} and define μ_{data} through a nonlinear transformation of $\pi^{\alpha_{\text{data}}}$. This way, μ_{data} will be *non-Gaussian*, but we still have perfect knowledge about where its samples are supported. In particular, we will have

$$\text{support}(\mu_{\text{data}}) \approx H^{\alpha_{\text{data}} - \frac{1}{2}}.$$

This gives us the possibility to match U and K to μ_{data} in different ways. In realistic examples, knowledge about μ_{data} could come from prior information or by studying the

training samples—the empirical covariance matrix is, for example, a natural candidate for specifying C in IDDM2.

Lastly, we will be able to make explicit statements about the norm of H for which the distance bounds in Theorem 14 hold, which we will also quantify by choosing an α_{dist}. The larger α_{dist}, the better, since the underlying norm for the distance measurement gets stronger (see also the discussion in Section 6.1).

Note that the limit of the common implementation of diffusion models, which we called WNDM (see Section 6.2) will use white noise for the noising process as well as the loss, i.e., $\alpha_{\text{noise}} = \alpha_{\text{loss}} = 0$. In that case, $\mathcal{N}(0, C)$ will only be supported on any H^α with $\alpha < -\frac{1}{2}$. Therefore, so that we can apply Theorem 14 with $H^{\alpha_{\text{dist}}}$ we have to choose $\alpha_{\text{dist}} < -\frac{1}{2}$, i.e., use the norm of a negative Sobolev space.

For the two numerical experiments in Sections 7.2 and 7.3, we will proceed as follows:

1. We have information about the sample smoothness and support of μ_{data}, in this case in the form of an α_{data}.

2. Based on Section 6.1, we then choose U and K, which boils down to the choice of α_{noise} and α_{loss}.

3. We then know for which norms $\| \cdot \|_H$ our Wasserstein bound in Theorem 14 holds. In our interpolation family, this boils down to an upper bound for α_{dist}. Therefore, we can make statements about which properties of μ_{data} the diffusion model should successfully approximate.

7.2 Manifold Distribution on a Cameron–Martin Sphere

In this section, we will study a distribution which lies on an infinite-dimensional submanifold of $L^2([0, 1])$, namely the unit sphere of some Cameron–Martin space. To that end, choose $e_k(\cdot) = \sqrt{2} \sin(k\pi \cdot)$ in the construction of Section 7.1. For this choice, the Cameron–Martin spaces H^α will be the Sobolev spaces $W_0^{\alpha, 2}$ of functions vanishing at the boundary, and π^1 is proportional to the distribution of a Brownian bridge. Here we see that we can not only capture smoothness but also structural information, such as boundary conditions, through the choice of an appropriate Gaussian measure. For a more in-depth study of this, see Mathieu et al. (2023).

We draw $N = 50\,000$ samples from $\pi^{\alpha_{\text{data}}}$, where the data-generating α_{data} was set to

$$\alpha_{\text{data}} = 2.$$

By our discussion in Section 7.1, these samples are supported on any H^α with $\alpha < \frac{3}{2}$, in particular H^1. Now define $\alpha_{\text{supp}} = 1 < \frac{3}{2}$. The target distribution μ_{data} is created by projecting $\pi^{\alpha_{\text{data}}}$ onto the 10-sphere in $H^{\alpha_{\text{supp}}}$, i.e., applying the map

$$H^{\alpha_{\text{supp}}} \to H^{\alpha_{\text{supp}}}, \quad x \mapsto 10 \frac{x}{\|x\|_{H^\alpha_{\text{supp}}}}$$

to all samples. We depict some of the training samples and a heatmap of their marginal densities in Figure 2.

As described in Section 6.1, we use the theory to guide the choices for α_{noise} and α_{loss}. The data distribution was not absolutely continuous with respect to a Gaussian, since we
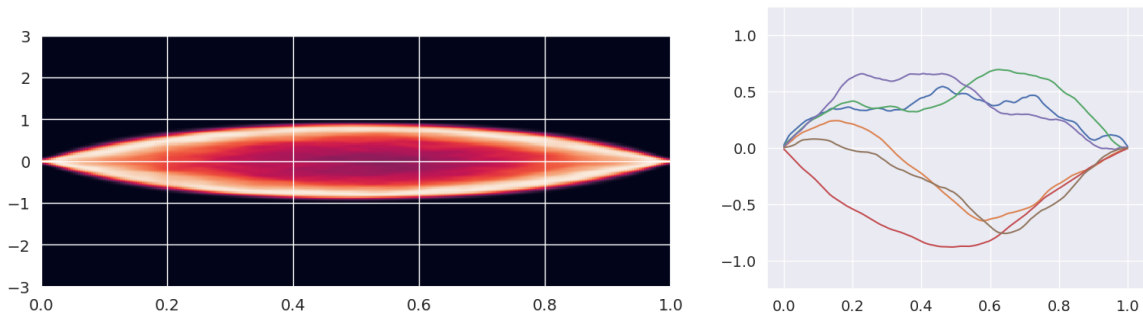
Figure 2: We generated 50 000 training examples from the distribution described in Section 7.2. On the left, we show a heatmap of the resulting marginal densities of function values at each point in the domain $[0, 1]$. On the right, we plot a few training samples.
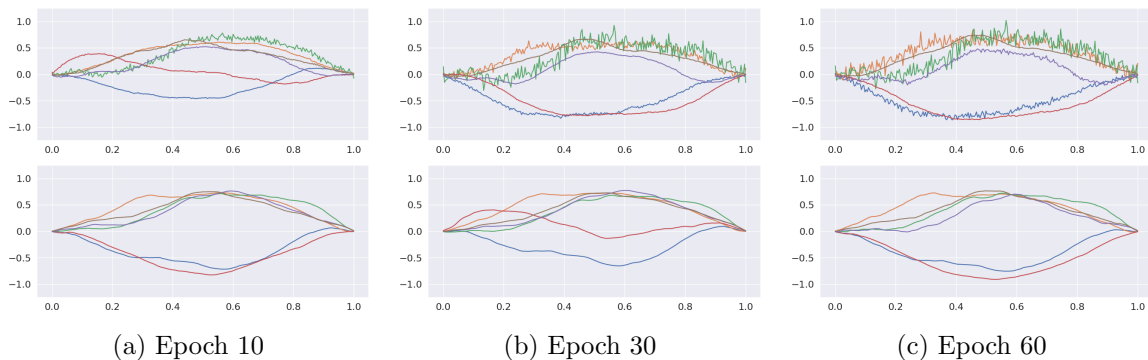

(a) Epoch 10 (b) Epoch 30 (c) Epoch 60

Figure 3: Example of Section 7.2: samples generated by WNDM (row 1) and IDDM1 (row 2) after increasing numbers of training epochs. Samples from the true measure can be compared in Figure 2.

projected it to a submanifold (the sphere). Therefore, we must apply Theorem 12 to obtain uniqueness. To satisfy the assumptions of Theorem 12, however, the Cameron–Martin space U has to contain μ_{data}. Hence, we will apply the IDDM1 from Section 6.1.

To apply IDDM1, we choose U so that it contains the support of μ_{data} and $\mathcal{N}(0, C)$. This is accomplished by setting $\alpha_{\text{noise}} = 1$. Then, following the design principles of IDDM1, we pick the loss norm to be $K = U$, i.e., $\alpha_{\text{loss}} = \alpha_{\text{noise}} = 1$, and learn the score by using the Cameron–Martin norm in the loss.

Note that the bound in Theorem 14 holds for any α_{dist} smaller than

$$\alpha_{\text{dist}} < \min\left\{\alpha_{\text{data}} - \frac{1}{2}, \alpha_{\text{noise}} - \frac{1}{2}, \alpha_{\text{loss}}\right\} = \min\left\{\frac{3}{2}, \frac{1}{2}, 1\right\} = \frac{1}{2}.$$

For WNDM, i.e., the canonical implementation of diffusion models described in Section 6.2 with $\alpha_{\text{loss}} = \alpha_{\text{noise}} = 0$, the upper bound is $-\frac{1}{2}$. Therefore, while we do not expect the samples of WNDM to match the smoothness class of μ_{data}, we expect the samples of IDDM1 to at least partially retain the smoothness.

Figure 3 shows samples generated by the two models. We see that our theoretical findings are confirmed: WNDM fails to learn the smoothness or correlation structure of the
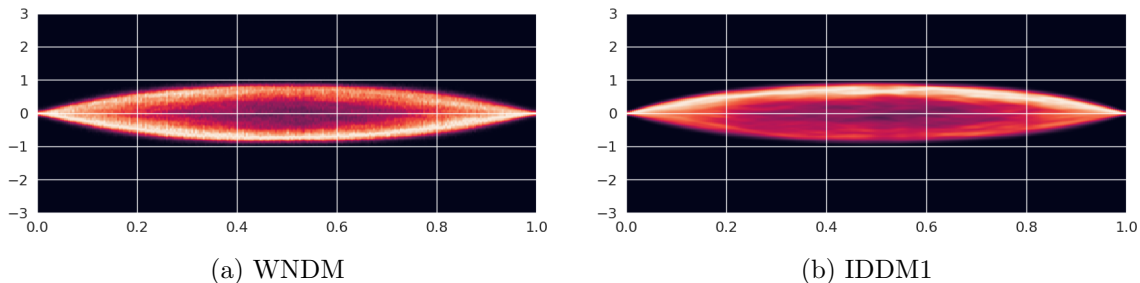
(a) WNDM

(b) IDDM1

Figure 4: Example of Section 7.2: each vertical slice shows a heatmap of the marginal density estimated from 2048 samples generated by each of the diffusion models, after 60 epochs of training. For comparison, the heatmap of the 50 000 training examples is plotted in Figure 2. The one-dimensional marginals are matched well by both algorithms.

samples. Solely at training epoch 10 the WNDM algorithm generated some samples that seemed to have the right smoothness, but even those actually contain jitter if one looks closely. Overall, the training process was very unstable regarding the data smoothness, and minimizing the loss did not seem to correlate with also matching the derivatives of the functions. On the other hand, IDDM1 produces samples from the correct smoothness class, from the start of training onwards.

Note that both algorithms matched the marginals quite well, as can be seen in the heatmap plots of Figure 4, which is also suggested by the theory: even if Theorem 14 only holds for an underlying negative Sobolev norm, the overall distribution and in particular its marginals should still match the true marginals (see Section B).

7.3 Conditional Sampling and Infinite-Dimensional Bayesian Inverse Problems: Volatility Estimation

The following numerical experiment is inspired by Bayesian inverse problems (BIPs) (Stuart, 2010), which here we approach via the paradigm of *simulation-based inference*. In this setting, we will use our infinite-dimensional diffusion models for conditional sampling.

We assume that we have some knowledge about an unknown random variable $X \in \mathcal{H}$ in the form of a measurement $y \in \mathbb{R}^l$ drawn from

$$Y \sim q(X, \cdot),$$

where q is an observation kernel. Furthermore, we have some prior information on X, formalized through a prior probability distribution:

$$X \sim \pi := \mathcal{N}(0, C_\mu).$$

By Bayes' theorem, the posterior distribution ν of X given some observations $Y = y$ is given by

$$d\pi(\,\cdot\,|Y = y) \propto q(\cdot, y)\, d\pi. \tag{17}$$

Gold-standard methods for asymptotically exact sampling of distributions like (17) involve Markov chain Monte Carlo (MCMC), e.g., Hamiltonian Monte Carlo (Duane et al., 1987) or,

24

in the infinite-dimensional case, Hilbert space Hamiltonian Monte Carlo (HSHMC) (Beskos et al., 2011), or other geometry-exploiting infinite-dimensional MCMC methods (Cotter et al., 2013; Cui et al., 2016; Kim et al., 2023).

These MCMC methods, however, rely on having an explicit formula for the density of ν (up to a normalizing constant). In many cases this is not possible—for example, if q or any of its components is given as a black-box model. To train a conditional diffusion model, on the other hand, we only need *samples* from the joint distribution of (X, Y). These can be generated by sampling X^i from the prior measure and sampling $Y \sim q(X^i, \cdot)$. We then train a conditional diffusion model to generate samples from $X|Y = y$ for any y. This is done by making the score model s not only depend on X_t, but also on Y, i.e., we have a model $s(t, X_t, Y)$, which predicts X_0 given $Y = y$. The only modification to Algorithm 1 is that one sub-samples paired states and observations (x^i, y^i) in line 3 from the training data, and then inputs y^i into the diffusion model on line 6. During generation, one can then input the observation value y that one wants to condition on during simulation of the reverse SDE (Batzolis et al., 2021). In Algorithm 2, this would correspond to inputting a fixed value of y for all times t in line 5. Hence the entire procedure is sample-driven, and an example of simulation-based inference (Cranmer et al., 2020).

We now proceed to a specific instance of a Bayesian inverse problem. The experiment is inspired by volatility estimation. We assume that we observe a path of a time series, for example a stock price, modeled as

$$\mathrm{d}S_\tau = \sigma_\tau S_\tau \mathrm{d}B_\tau,$$

with no drift and a time-dependent volatility σ_τ. The solution to the above equation is given by a geometric Brownian motion, i.e.,

$$S_\tau = S_0 \exp\left(\int_0^\tau \sigma_r \mathrm{d}B_r - \frac{1}{2}\int_0^\tau \sigma_r^2 \mathrm{d}r\right).$$

We simulate paths of the above and observe S_τ at discrete times $\tau_1 = \frac{1}{4}, \tau_2 = \frac{2}{4}, \tau_3 = \frac{3}{4}, \tau_4 = 1$. Then, we apply a log-transformation and define r_i as the log-returns:

$$r_i := \log S_{\tau_i} - \log S_{\tau_{i-1}} = \int_{\tau_{i-1}}^{\tau_i} \sigma_r \mathrm{d}B_r - \frac{1}{2}\int_{\tau_{i-1}}^{\tau_i} \sigma_r^2 \mathrm{d}r \sim \mathcal{N}\left(-\frac{1}{2}v_i, v_i\right), \text{ with } v_i := \int_{\tau_{i-1}}^{\tau_i} \sigma_r^2 \mathrm{d}r.$$

Here, we set $\tau_0 = 0$ for notational convenience. Since σ should be positive, we model it as

$$\sigma_\tau = \exp(a_\tau),$$

and seek to infer the log-volatility $a : [0, 1] \to \mathbb{R}$.

Again, we define a family of Gaussian measures as in Section 7.1. This time we use a different orthonormal basis of $L^2([0, 1])$, given by

$$e_k(\tau) = \begin{cases} \sqrt{2}\cos(k\pi\tau), & \text{if } k \text{ even} \\ \sqrt{2}\sin((k+1)\pi\tau), & \text{otherwise} \end{cases}.$$

This leads to Gaussian measures whose samples have periodic boundary conditions. Since e_k and e_{k+1} (for k uneven) should have the same 'magnitude,' we slightly modify (15) and

(16): for k uneven, we replace $(k+1)^{-\alpha}$ by $k^{-\alpha}$. All the discussed properties of the family π^α are not affected by this change, since the decay of the eigenvalues is asymptotically the same. We put a prior on a. It's covariance is given by $\frac{1}{2}C_{\text{prior}}$, where C_{prior} is the covariance of $\pi^{\alpha_{\text{data}}}$, with $\alpha_{\text{data}} = 4$:

$$a \sim \mathcal{N}(0, C_{\text{prior}}).$$

The goal of a conditional diffusion model is to generate samples from the posterior

$$d\pi^{\alpha_{\text{data}}}(a_\tau | r_1, r_2, r_3, r_4) \propto \prod_{i=1}^{4} \mathcal{N}\left(r_i; -\frac{1}{2}v_i, v_i\right) d\pi^{\alpha_{\text{data}}}, \tag{18}$$

for a fixed observation $r = (r_1, r_2, r_3, r_4)$. Via the model defined above, each v_i is a functional of σ_τ and thus a_τ.

For training, we generate $N = 50\,000$ samples from the prior $\{a^n\}_{n=1}^N$ together with simulated observations $\{r^n\}_{n=1}^N$. The trained diffusion models should, for any input $r \in \mathbb{R}^4$, generate samples from (18).

To assess the performance of the trained models, we drew a random $\tilde{a}_\tau$ and corresponding observations $\tilde{r} = (\tilde{r}_1, \tilde{r}_2, \tilde{r}_3, \tilde{r}_4)$. We used the HSHMC algorithm to generate $50\,000$ "reference" posterior samples from (18) for this fixed observation value $\tilde{r}$. We plot these posterior samples and their heatmap, as well as the data-generating value $\tilde{a}$ of the log-volatility, in Figure 5. After training, we input $\tilde{r}$ (which the diffusion models have not seen before) to the conditional diffusion models and compare the generated samples to those from HSHMC.

As in Section 7.2, we again compare the canonical diffusion model implementation WNDM against an implementation motivated by the infinite-dimensional theory. In this case, since we are sampling from a Bayesian inverse problem with a Gaussian prior, we are in the setting of Theorem 13. Therefore, we will implement the IDDM2 algorithm from Section 6.1. We match the noise structure to the data by setting $\alpha_{\text{noise}} = \alpha_{\text{data}} = 4$, which is justified by the form of (18) of μ_{data}. Then we set $\alpha_{\text{loss}} = 2$, such that K supports μ_{data} and $\pi^{\alpha_{\text{noise}}}$. Note that any $\alpha_{\text{loss}} < \frac{7}{2}$ would also have been a valid choice. The choice $\alpha_{\text{loss}} = 3$ worked comparably well in our numerical experiments.

We compare samples generated by the two diffusion models in Figure 6. Again, the WNDM algorithm did not match the smoothness class of μ_{data} in a stable way. While during training, there were times at which the network generated smooth samples, it later unlearned to do so. The IDDM2 algorithm outputs samples of the correct class at every point during training. Both algorithms are able to match the marginal distributions, although IDDM2 does slightly better, as seen in Figure 7. Therefore, as in Section 7.2, these numerical experiments confirm the theoretical predictions made in Section 6.

Remark 15 *Note that for our choice of α_{noise}, the reverse SDE now starts with initial condition $\pi^{\alpha_{data}}(a_\tau) = \mathcal{N}(0, C_{prior})$ and ends in the posterior $\pi^{\alpha_{data}}(a_\tau | r)$. Therefore, it has learned to transport the prior to the posterior. Furthermore, in this case, we can interpret the reverse SDE as a smoothed version of the forward process, i.e., an Ornstein–Uhlenbeck process conditioned on its terminal values. This also opens up the way to interpret the training of the reverse SDE as an infinite-dimensional control problem.*
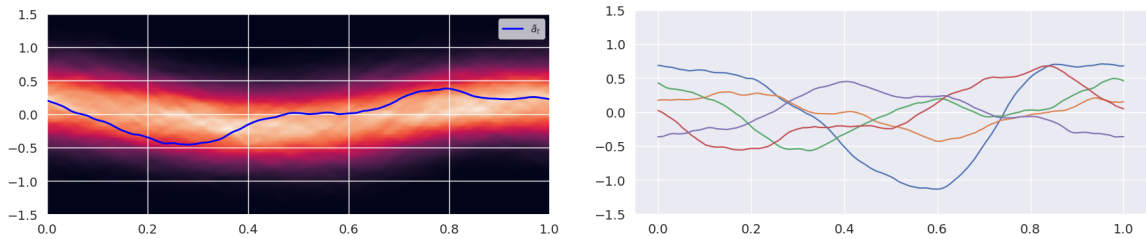
Figure 5: Example of Section 7.3. As a reference/comparison, we generate $50\,000$ high-quality posterior samples from $d\pi^{\alpha_{\mathrm{data}}}(a_\tau|\tilde{r})$ using the Hilbert space Hamiltonian Monte Carlo algorithm. On the left is a heatmap of posterior marginal densities of a_τ, at each point in the domain $\tau \in [0, 1]$. On the right, we plot a few example posterior samples.

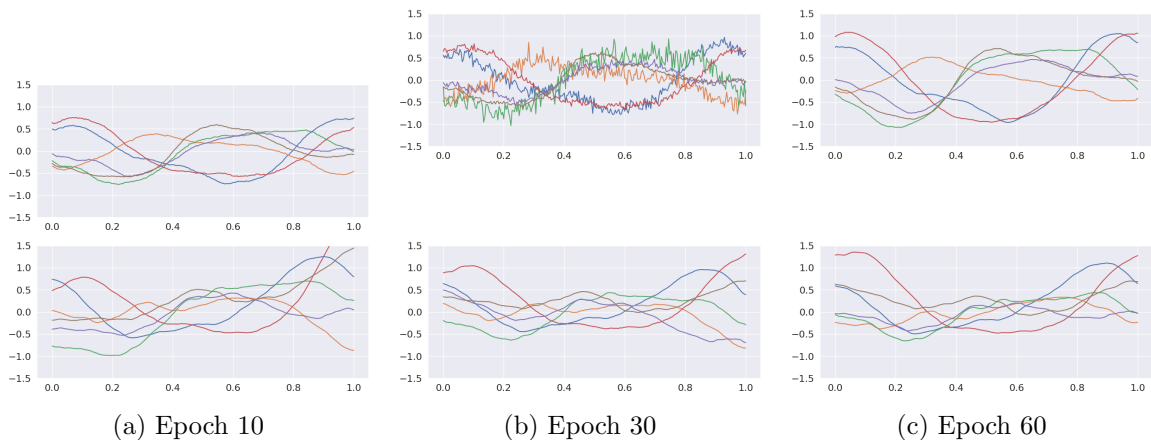(a) Epoch 10   (b) Epoch 30   (c) Epoch 60

Figure 6: Example of Section 7.3. Conditional samples from WNDM (upper row) and IDDM2 (lower row) after varyings number of training epochs. Compare to the high-quality posterior samples generated using Hamiltonian Monte Carlo in Figure 5.

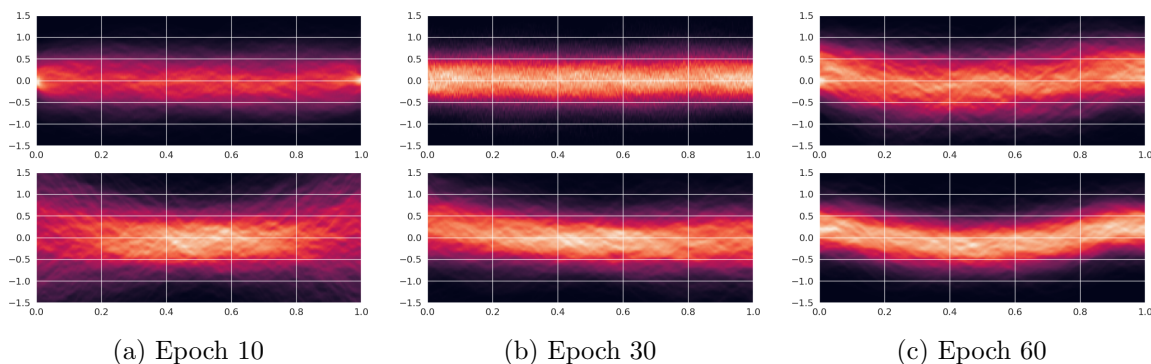(a) Epoch 10   (b) Epoch 30   (c) Epoch 60

Figure 7: Example of Section 7.3. Heatmaps of the 2048 conditional samples generated by WNDM (upper row) and IDDM2 (lower row), for increasing numbers of training epochs. Reference heatmaps generated via Hilbert space Hamiltonian Monte Carlo are in Figure 5 for comparison.

7.4 Limitations

It is important to point out that, generally, the white noise diffusion models were able to generate samples from the appropriate smoothness class after several rounds of retraining. However, the sample smoothness was not robust. While the models typically fitted the moments and marginals of the distributions quite well, the smoothness of the samples proved inconsistent—varying with the initial conditions and the duration of the training. Specifically, depending on the network's initial training parameters, the samples could either be smooth, become less smooth over time, or exhibit initial smoothness that diminished as the training progressed and the marginals were better fitted.

We also note that the numerics here are intended as an illustration of the preceding theory and the resulting guidelines. A more comprehensive numerical study could train multiple models with different random initial conditions on the same training data set, perform ablation studies over individual design choices, and compare numerical measures of the smoothness of paths. It would also be of interest to evaluate the impact of different neural network architectures. Such studies are outside the scope of this article, however.

8. Summary

We have formulated the diffusion-based generative modeling approach directly on infinite-dimensional Hilbert spaces. Our formulation involves specifying infinite-dimensional forward and reverse SDEs and an associated denoising score matching objective. We prove that our formulation is well-posed. To that end, we show that the reverse SDE we wish to approximate has a unique solution; furthermore, we show under which conditions the denoising score matching objective generalizes to an infinite-dimensional setting. Building on these results, we are able to prove dimension-independent convergence bounds for diffusion models, which hold in the infinite-dimensional case.

These theoretical developments reveal an intricate relationship between the properties of the target/data-generating measure μ_{data} and the choices of the Wiener process W_t^U and the loss norm in the denoising score matching objective $\|\cdot\|_K$. We utilize this knowledge to develop guidelines on how to make such choices for a given μ_{data}. For image distributions, these guidelines are in line with the canonical choices made in practice. For other target distributions μ_{data}, however, the algorithm design should be modified. We apply these modifications to two generative modeling tasks that are discretizations of infinite-dimensional problems, and the numerical results confirm our theoretical findings.

Acknowledgements

Appendix A. Numerical Details

We list some implementation details:

1. Instead of running the forward SDE with a uniform speed, we instead ran it using a speed function $\alpha(t)$,

$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)C}\mathrm{d}W_t.$$

The SDE was then run on the interval $[0, 1]$. This corresponds to a time-change and is common practice for diffusion models; see for example Song et al. (2021). We used the time-change function $\beta(t)$ as in Song et al. (2021), i.e.,

$$\beta(t) = 0.001 + t(20 - 0.001).$$

2. We discretized the unit interval $[0, 1]$ into $M = 1000$ evenly spaced points for training and generation.

3. We added a last denoising step, as is common practice. That means, that in the last step of the Euler-Maruyama integrator, we did not add any extra noise any more, but just evaluated the drift and took a step in that direction. This is even more important in our case for comparison than normally, since the added noise has correlation structure $\mathcal{N}(0, C)$ which is close to μ_{data}, while the added noise in WNDM has structure $\mathcal{N}(0, I)$. Therefore, adding this noise to all samples right before comparing them would have given an unfair disadvantage to WNDM.

4. Furthermore, we added $\varepsilon_{\mathrm{reg}}\mathrm{Id}$ onto the covariance matrices for numerical stability, where $\varepsilon_{\mathrm{reg}} = 0.0001$.

5. Our experiments were implemented in JAX, and we used the U-Net architecture from Song et al. (2021) for the neural network.

Appendix B. Negative Sobolev Wasserstein Distances

We briefly and heuristically explore what it means if μ and π are close in $\mathcal{W}_2$ when the underlying norm is a negative Sobolev norm. Denote by H^α the spaces defined in Section 7.1. Now let $f \in H^\alpha$. Assume that X, Y form a $\mathcal{W}_2^{-\alpha}$-optimal coupling, i.e., $X \sim \mu$ and $Y \sim \pi$ and that

$$\mathbb{E}[\|X - Y\|_{-\alpha}] \leq \mathcal{W}_2^{-\alpha}(\pi, \mu).$$

Now, $H^{-\alpha}$ can be viewed as the dual of H^α and therefore we can evaluate X or Y on f. Then,

$$\mathbb{E}[|X(f) - Y(f)|] \leq \|f\|_\alpha \mathbb{E}[\|X - Y\|_{-\alpha}] \leq \|f\|_\alpha \mathcal{W}_2^{-\alpha}(\mu, \pi)$$

Therefore, we can expect the evaluations of X and Y on test-functions from H^α to be close. The larger α gets, the fewer test functions are in H^α. Note that for any $\alpha \geq 0$ (and in particular for our typical case $H^0 = L^2$), H^α will not contain point evaluations, and therefore we cannot expect point evaluations of X and Y to be close (in case they are well-defined).

Appendix C. Exponential Integrator

The exponential integrator (Certaine, 1960) is derived by splitting the following SDE

$$dY_t = \frac{1}{2}Y_t + s(t, Y_t)dt + \sqrt{C}dW_t = \frac{1}{2}Y_t + s(t, Y_t)dt + \sqrt{C}dW_t \tag{19}$$

into the linear and nonlinear part. The exact solution is then given by

$$Y_{t+\Delta t} = e^{t/2}Y_t = e^{t/2}Y_t + (e^{\Delta t/2} - 1)s(t, Y_t) + \sqrt{e^{\Delta t} - 1}\xi.$$

The exponential integrator was first applied to diffusion models in Zhang and Chen (2023).

Appendix D. Proofs for Section 2

D.1 Proofs for Section 2.2

We first prove Lemma 1
Proof In finite dimensions we can explicitly write p_t as

$$p_t(x) = \int p_{t|0}(x|x_0)d\mu_{\text{data}}(x_0) \tag{20}$$

where $p_{t|0}$ is the time t-transition kernel of the forward SDE, given by

$$p_{t|0}(x|x_0) = \frac{1}{\sqrt{(2\pi v_t)^D \det(C)}} \exp\left(-\frac{1}{2(1-e^{-t})}\left\langle\left(x - e^{-\frac{t}{2}}x_0\right), C^{-1}\left(x - e^{-\frac{t}{2}}x_0\right)\right\rangle_H^2\right).$$

We can exchange the derivative with the integral by Leibniz rule since the derivative of the integrand is bounded. Therefore, we have that

$$
\begin{aligned}
\nabla \log p_t(x) &= \frac{1}{p_t(x)}\nabla_x \int p_{t|0}(x|x_0)d\mu_{\text{data}}(x_0) \\
&= -\frac{1}{(1-e^{-t})}\int C^{-1}\left(x - e^{-\frac{t}{2}}x_0\right)\frac{p_{t|0}(x|x_0)}{p_t(x)}d\mu_{\text{data}}(x_0) \\
&= -\frac{1}{(1-e^{-t})}C^{-1}\left(\mathbb{E}\left[X_t - e^{-\frac{t}{2}}X_0|X_t = x\right]\right).
\end{aligned}
$$

In the last equation we used the formula for the conditional density; see Durrett (2005, Section 4.1.c). ■

We now prove Lemma 5.
Proof We first treat continuity. Since p_t can be written as the convolution of μ_{data} with a Gaussian kernel, we know that it is smooth in space and time on $(0, \infty]$. Furthermore, it holds that $p_t > 0$ everywhere. Due to that, we can deduce that $\nabla \log p_t = \frac{\nabla p_t}{p_t}$ is continuous in t. Since X_t is also continuous in time, we get that $\nabla \log p_t(X_t)$ is time-continuous.

Now we prove the reverse-time martingale property. Since we can write

$$p_t(x_t) = \int p_s(x_s)p_{t|s}(x_t|x_s)dx_s$$

and by using since the transition kernel $p_{t|s}(x_t|x_s)$ is given by $\mathcal{N}(e^{-(t-s)/2}x_s, \frac{1}{1-e^{-(t-s)}}C)$,

$$
\begin{aligned}
&\nabla p_t(x_t) \\
=\ & \nabla_{x_t} \int \exp\left(-\frac{1}{2(1-e^{-(t-s)})}\left\langle x_t - e^{-\frac{(t-s)}{2}}x_s, x_t - e^{-\frac{(t-s)}{2}}x_s\right\rangle\right) p_s(x_s)\mathrm{d}x_s \\
=\ & \int e^{\frac{(t-s)}{2}} \nabla_{x_s} \exp\left(-\frac{1}{2(1-e^{-(t-s)})}\left\langle x_t - e^{-\frac{(t-s)}{2}}x_s, x_t - e^{-\frac{(t-s)}{2}}x_s\right\rangle\right) p_s(x_s)\mathrm{d}x_s \\
=\ & \int e^{\frac{(t-s)}{2}} \exp\left(-\frac{1}{2(1-e^{-(t-s)})}\left\langle x_t - e^{-\frac{(t-s)}{2}}x_s, x_t - e^{-\frac{(t-s)}{2}}x_s\right\rangle\right) \nabla_{x_s} p_s(x_s)\mathrm{d}x_s \\
=\ & \int e^{\frac{(t-s)}{2}} p_{t|s}(x_t|x_s)p_s(x_s)\nabla_{x_s} \log p_s(x_s)\mathrm{d}x_s \\
=\ & e^{\frac{(t-s)}{2}} \int p_{s,t}(x_s, x_t)\nabla_{x_s} \log p_s(x_s)\mathrm{d}x_s.
\end{aligned}
$$

Since $\nabla \log p_t(x_t) = \frac{\nabla p_t(x_t)}{p_t(x_t)}$ and $p_{s|t}(x_s|x_t) = \frac{p_{s,t}(x_s,x_t)}{p_t(x_t)}$, we get that

$$
\begin{aligned}
&\nabla \log p_t(x_t) \\
=\ & e^{\frac{(t-s)}{2}} \int \frac{p_{s,t}(x_s, x_t)}{p_t(x_t)}\nabla_{x_s} \log p_s(x_s)\mathrm{d}x_s = e^{\frac{(t-s)}{2}} \int p_{s|t}(x_s|x_t)\nabla_{x_s} \log p_s(x_s)\mathrm{d}x \\
=\ & e^{\frac{(t-s)}{2}} \mathbb{E}[\nabla \log p_s(X_s)|X_t = x_t].
\end{aligned}
$$

$\blacksquare$

The above calculations have already been done in Chen et al. (2022) to bound the difference $\mathbb{E}[\|\nabla \log p_t(x_t) - \nabla \log p_s(x_s)\|^2]$.

D.2 Proofs for Section 2.4

We start by proving Lemma 7.

Proof Let e_i be a basis of K and $K^D = \text{span}\langle e_1, \ldots, e_D\rangle$. We denote by P^D the projection onto K^D and by $X_t^D = P^D X_t$ the projection of X_t onto K^D. We will denote by

$$\|\cdot\| = \|\cdot\|_K$$

throughout this proof. Let

$$s^D = P^D \mathbb{E}[s(t, X_t)|X_t^D] = \mathbb{E}[\sigma_t^{-1}(X_t^D - e^{-t/2}X_t^D)|X_t^D],$$

where $\sigma_t^{-1} = \frac{1}{\sqrt{1-e^{-t}}}$. We have that

$$\mathbb{E}[\|s^D\|^2] \leqslant \mathbb{E}[\|\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2] = \sigma_t^{-2}\mathbb{E}[\|\mathcal{N}(0, P^D C P^D)\|^2] < \infty$$

since the right hand side is the expectation of the norm of a finite dimensional Gaussian, which is finite. Let $\tilde{s}^D(t, X_t) = P^D \mathbb{E}[\tilde{s}(t, X_t)|X_t^D]$.

Now,

$$
\begin{aligned}
\mathbb{E}[\|s^D - \tilde{s}^D\|_K^2] &= \mathbb{E}[\|P^D(\mathbb{E}[\tilde{s}(t, X_t)|X_t^D] - \mathbb{E}[s(t, X_t)|X_t^D])\|^2] \\
&\leqslant \mathbb{E}[\|\mathbb{E}[\tilde{s}(t, X_t) - s(t, X_t)|X_t^D]\|^2] \leqslant \mathbb{E}[\|s(t, X_t) - \tilde{s}(t, X_t)\|^2] < \infty,
\end{aligned}
$$

and

$$
\mathbb{E}[\|\tilde{s}^D\|^2] \leqslant 2(\mathbb{E}[\|s^D - \tilde{s}^D\|^2] + \mathbb{E}[\|s^D\|^2]) < \infty.
$$

Therefore, we get that

$$
\mathbb{E}[\|s^D - \tilde{s}^D\|^2] = \mathbb{E}[\|s^D\|^2 + \|\tilde{s}^D\| - \langle s^D, \tilde{s}^D \rangle] = \mathbb{E}[\|s^D\|^2] + \mathbb{E}[\|\tilde{s}^D\|] - 2\mathbb{E}[\langle s^D, \tilde{s}^D \rangle],
$$

where we used that all the terms are finite in the last equality, so that we are no adding and subtracting infinities. Now, for

$$
\mathbb{E}[\langle s^D, \tilde{s}^D \rangle] = \mathbb{E}[\langle \mathbb{E}[\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)|X_t^D], \tilde{s}^D \rangle] = \mathbb{E}[\langle \sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D), \tilde{s}^D \rangle],
$$

and therefore, since $\mathbb{E}[\|X_t^D - e^{-t/2}X_0^D\|^2]$ is finite we can do a zero-addition of $\mathbb{E}[\|X_t^D - e^{-t/2}X_0^D\|^2]$ and get that

$$
\mathbb{E}[\|s^D - \tilde{s}^D\|^2] = \mathbb{E}[\|\tilde{s}^D - \sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2] + \mathbb{E}[\|s^D\|^2] - \mathbb{E}[\|\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2].
$$

By Lemma 16 we see that the left hand side converges to $\mathbb{E}[\|s - \tilde{s}\|^2]$. Therefore we get that $\mathbb{E}[\|\tilde{s}^D - \sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2]$ converges to something finite, if and only if

$$
\mathbb{E}[\|s^D\|^2] - \mathbb{E}[\|\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2]
$$

does. For this term we get that since

$$
\begin{aligned}
\mathbb{E}[\langle s^D, \sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D) \rangle] &= \mathbb{E}[\langle \mathbb{E}[\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)|X_t^D], \sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D) \rangle] \\
&= \mathbb{E}[\|\mathbb{E}[\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)|X_t^D]\|^2] \\
&= \mathbb{E}[\|s^D\|^2],
\end{aligned}
$$

we can deduce

$$
\begin{aligned}
&\mathbb{E}[\|s^D\|^2] - \mathbb{E}[\|\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2] \\
&= -\mathbb{E}[\|s^D\|^2] + 2\mathbb{E}[\langle s^D, \sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D) \rangle] - \mathbb{E}[\|\sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2] \\
&= -\mathbb{E}[\|s^D - \sigma_t^{-1}(X_t^D - e^{-t/2}X_0^D)\|^2] \to_{L^2} -\mathbb{E}[\|s - \sigma_t^{-1}(X_t - e^{-t/2}X_0)\|^2]
\end{aligned}
$$

where the last convergence is implied by Proposition 16. The last result now follows by rewriting

$$
V_t = \mathbb{E}[\|\mathbb{E}[\sigma_t^{-1}(X_t - e^{-t/2}X_0)|X_t] - \sigma_t^{-1}(X_t - e^{-t/2}X_0)\|^2]
$$

and using that X_t can be pulled out of the conditional expectation. ∎

Now we prove Lemma 8:

Proof **Item 1:** We know from Lemma 7 that DSM is finite if and only if V is finite. Therefore, we will prove that V is finite:

$$V_t = \frac{e^{-t}}{1 - e^{-t}} \mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t]\|_U^2]$$

Now,

$$
\begin{aligned}
&\mathbb{E}[\|X_0 - \mathbb{E}[X_0]\|_U^2] \\
=\ & \mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t] + \mathbb{E}[X_0|X_t] - \mathbb{E}[X_0]\|_U^2] \\
=\ & \mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t]\|_U^2] + \mathbb{E}[\|\mathbb{E}[X_0|X_t] - \mathbb{E}[X_0]\|_U^2] \\
& + \mathbb{E}[\langle X_0 - \mathbb{E}[X_0|X_t], \mathbb{E}[X_0|X_t] - \mathbb{E}[X_0]\rangle_U] \\
=\ & \mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t]\|_U^2] + \mathbb{E}[\|\mathbb{E}[X_0|X_t] - \mathbb{E}[X_0]\|_U^2],
\end{aligned}
$$

where the last term drops by taking the conditional expectation with respect to X_t. Therefore,

$$\mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t]\|_U^2] \leqslant \mathbb{E}[\|X_0 - \mathbb{E}[X_0]\|_U^2] < \infty$$

and so V_t is finite.

Item 2: In this case we use that we can also write V_t as

$$V_t = \frac{1}{1 - e^{-t}} \mathbb{E}[\|X_t - e^{-t/2} X_0 - \mathbb{E}[X_t - e^{-t/2} X_0|X_t]\|_H^2]$$

Similarly as above,

$$\mathbb{E}[\|X_t - e^{-t/2} X_0 - \mathbb{E}[X_t - e^{-t/2} X_0|X_t]\|_H^2] \leqslant \mathbb{E}[\|X_t - e^{-t/2} X_0\|_H^2]$$

where we used that $\mathbb{E}[X_t - e^{-t/2} X_0] = 0$. Now, since $X_t - e^{-t/2} X_0 \sim \mathcal{N}(0, (1 - e^{-t})C)$, which is supported on H, the above expectation is finite. ∎

We used the following lemma in the proofs of the two above lemmas:

Lemma 16 *Let $(K, \langle \cdot, \cdot \rangle)$ be a separable Hilbert space, and $Z, \tilde{Z}$ random variables taking values in K. Let e_i be an orthonormal basis of K. Denote by $K^D = \mathrm{span}\langle e_1, \ldots, e_D \rangle$ and by P^D the projection onto K^D. Furthermore, let Z^D be given by $Z^D = P^D \mathbb{E}[Z|P^D \tilde{Z}]$. Then, if $\mathbb{E}[\|\mathbb{E}[Z|\tilde{Z}]\|_K^2] < \infty$, $Z^D \to \mathbb{E}[Z|\tilde{Z}]$ in L^2 and almost surely.*

Proof We have that

$$
\begin{aligned}
\mathbb{E}[\|Z^D - \mathbb{E}[Z|\tilde{Z}]\|_K^2] &= \mathbb{E}[\|P^D(\mathbb{E}[Z|\tilde{Z}^D] - \mathbb{E}[Z|\tilde{Z}])\|_K^2] + \mathbb{E}[\|(I - P^D)\mathbb{E}[Z|\tilde{Z}]\|_K^2] \\
&\leqslant \mathbb{E}[\|\mathbb{E}[Z|\tilde{Z}^D] - \mathbb{E}[Z|\tilde{Z}]\|_K^2] + \mathbb{E}[\|(I - P^D)\mathbb{E}[Z|\tilde{Z}]\|_K^2] \qquad (21)
\end{aligned}
$$

The cross term in the first equality is 0 since P^D is the orthogonal projection. The first term in the (21) converges to 0, since $\mathbb{E}[Z|\tilde{Z}^D] = \mathbb{E}[\mathbb{E}[Z|\tilde{Z}]|\tilde{Z}^D]$ is a family of conditional expectations of the L^2-random variable $\mathbb{E}[Z|\tilde{Z}]$. The result follows by the L^2-martingale convergence theorem. The second term converges to 0 since

$$\mathbb{E}[\|\mathbb{E}[Z|\tilde{Z}]\|_K^2] = \sum_{d=1}^{\infty} \mathbb{E}[\langle e_i, \mathbb{E}[Z|\tilde{Z}]\rangle_K^2] < \infty.$$

But $\mathbb{E}[\|(I - P^D)\mathbb{E}[Z|\tilde{Z}]\|_K^2]$ is equal to $\sum_{d=D}^{\infty} \mathbb{E}[\langle e_i, \mathbb{E}[Z|\tilde{Z}]\rangle_K^2]$, which converges to 0 since the full sum is finite.

Furthermore, we can write

$$\|Z^D - \mathbb{E}[Z|\tilde{Z}]\|_K \leqslant \|P^D(\mathbb{E}[Z|\tilde{Z}^D] - \mathbb{E}[Z|\tilde{Z}])\|_K^2 + \|(I - P^D)\mathbb{E}[Z|\tilde{Z}]\|_K^2$$
$$\leqslant \|\mathbb{E}[Z|\tilde{Z}^D] - \mathbb{E}[Z|\tilde{Z}]\|_K^2 + \|(I - P^D)\mathbb{E}[Z|\tilde{Z}]\|_K^2$$

The second term on the right-hand vanishes as $D \to \infty$ since $\mathbb{E}[Z|\tilde{Z}] \in K$. The first term almost surely converges to 0 due to the almost sure martingale convergence theorem. ∎

Appendix E. Existence Proof

E.1 Spectral Approximation of C

Let $\nu = \mathcal{N}(0, C)$ be a Gaussian measure with values in $(H, \langle\cdot, \cdot\rangle_H)$. C has an orthonormal basis e_i of eigenvectors and corresponding non-negative eigenvalues $c_i \geqslant 0$, i.e.,

$$Ce_i = c_i e_i.$$

We define the linear span of the first D eigenvectors as

$$H^D = \left\{ \sum_{i=1}^{D} f_i e_i | f_1, \ldots, f_D \in \mathbb{R} \right\} \subset H$$

Let $P^D : H \to H^D$ be the orthogonal projection onto H^D. If we write an element f of H as

$$f = \sum_{i=1}^{\infty} \langle f, \varphi_i \rangle_H \varphi_i,$$

P^D is equivalent to restricting f to its first D coefficients:

$$P^D : H \to H^D, \quad f \mapsto \sum_{i=1}^{D} \langle f, \varphi_i \rangle_H \varphi_i.$$

The push-forwards $(P^D)_* \nu$ of ν under P^D are denoted by

$$\nu^D := (P^D)_* \nu, \quad \text{where} \quad (P^D)_* \nu(A) = \nu((P^D)^{-1}(A)).$$

It is a Gaussian measure with covariance operator $P^D C P^D$.

By sending $v \in \mathbb{R}^D$ to $\hat{v} = v_1 e_1 + \cdots + v_D e_D$ we can identify H^D with $\mathbb{R}^D$. Under these identifications, ν^D would have distribution $\mathcal{N}(0, C^D)$ on $\mathbb{R}^D$, where C^D is a diagonal matrix with entries $c_1, \ldots, c_D$.

E.2 Spectral Approximation of the SDEs

We define the finite-dimensional approximations of μ_{data} by $\mu_{\text{data}}^D = (P^D)_* \mu_{\text{data}}$. We discretize the forward SDE (5) by

$$\mathrm{d}X_t^D = -\frac{1}{2}X_t^D \mathrm{d}t + \sqrt{P^D C P^D}\mathrm{d}W_t, \quad X_0^D \sim \mu_{\text{data}}^D \tag{22}$$

Since μ_{data} is supported on H^D, $P^D C P^D$ projects the noise down to H^D, and the operation $X_t \to -\frac{1}{2}X_t$ keeps H^D invariant, X_t^D will stay in H^D for all times. Therefore we can view X_t^D as process on $\mathbb{R}^D$ and define the Lebesgue densities p_t^D of X_t^D there.

E.3 Proof of Theorem 9

We can now prove Theorem 9:

Proof The forward SDE is just a standard Ornstein–Uhlenbeck process and existence and uniqueness of that is standard; see, for example, Da Prato and Zabczyk (2014, Theorem 7.4). We will now show that the time reversal

$$Y_t := X_{T-t}$$

is a solution to (6).

The solution to the forward SDE is given as the stochastic convolution,

$$X_t = e^{-t}X_0 + \int_0^t e^{-(t-s)}\sqrt{C}\mathrm{d}W_s.$$

The processes $X_t^D := P^D(X_t)$ (see Section E.1) are solutions to (22), since the SDE coefficients are decoupled. We now show that they converge to X_t almost surely in the supremum norm. We define

$$X_t^{D:\infty} = X_t - X_t^D.$$

Then

$$X_t^{D:\infty} = e^{-t}X_0^{D:\infty} + \int_0^t e^{-(t-s)}\sqrt{C}\mathrm{d}W_s^{D:\infty},$$

where $W_s^{D:\infty}$ is the projection of W_s onto $\text{span}\{e_D, e_{D+1}\ldots\}$. It holds that

$$\mathbb{E}[\sup_{t\leqslant T}\|X_t^{D:\infty}\|_H^2] \leqslant 4e^{-2t}\mathbb{E}[\|X_0^{D:\infty}\|_H^2] + 4(1-e^{-t})\sum_{i=D}^{\infty} c_i \to 0$$

for $D \to \infty$, where we used Doob's L^2 inequality to bound the stochastic integral. The first term will converge to 0 almost surely, since X_0 is H-valued and therefore the sum $\|X_0\|_H^2 = \sum_{i=1}^{\infty}\langle X_0, \varphi_i\rangle^2$ is almost surely finite, where the φ_i are defined in Section E.1. Therefore, $\|X_0^{D:\infty}\|_H^2 = \sum_{i=D}^{\infty}\langle X_0, \varphi_i\rangle^2$ will almost surely converge to zero. An analogous argumentation holds for the second term since $\sum_{i=1}^{\infty} c_i$ is finite because C is trace-class on H.

Denote by E the Banach space of continuous, H-valued paths with the supremum norm, $E = C([0,T], H)$. Then we can view X. as an E-valued random variable; see Da Prato and

Zabczyk (2014, Theorem 4.12). Then we have just proven that X^N converges to X in $L^2(\Omega, E)$.

We denote the time-reversals of X_t^N by $Y_t^N := X_{T-t}^N$. These converge to their infinite-dimensional counterpart $Y_t = X_{T-t}$ in the same way as X_t^N converge to X_t. The main difficulty is to show Y_t solves the SDE (6). We do this using an approximation argument.

We define p_t^D as in Section E.2. Furthermore, we define the finite-dimensional time reversals of X_t^D as $Y_t^D = X_{T-t}^D$. By Proposition 17, we know that they satisfy

$$Y_t^D - Y_0^D - \frac{1}{2}\int_0^t Y_r^D \mathrm{d}r - \int_0^t s_{T-r}^D \mathrm{d}r = \sqrt{P^D C P^D} B_t^D,$$

for a H^D-Brownian motion $B_{D,t}$. It is important to note, that B_D will not be equal to the projection of B_{D+L} onto H^D in general, and the same hold for the $\nabla \log p_t^D$. However, as we will see now, we can prove some martingale-like properties for them to obtain convergence to their infinite-dimensional counterpart.

By, Lemma 1, we know that we can replace the $C\nabla \log p_t$ term by a conditional expectation s^D,

$$s^D(t, x^D) = \frac{1}{1-e^{-t}}\mathbb{E}[X_t^D - e^{-\frac{t}{2}}X_0^D | X_t^D = x^D].$$

However, since the forward SDE decouples, we can also write the above conditional expectation in terms of the infinite-dimensional process X_t:

$$s^D(t, x^D) = \frac{1}{1-e^{-t}}P^D\mathbb{E}\left[X_t - e^{-\frac{t}{2}}X_0 | P^D X_t = x^D\right],$$

where we use that the projections $P^D X_t$ are solutions to (22). In particular, due to the tower property of conditional expectations,

$$s_t^D = s^D(t, X_t^D) = P^D\mathbb{E}[s(t, X_t)|X_t^D].$$

By Lemma 16, which makes use of the fact that $\mathbb{E}[s(t, X_t)|P^D X_t]$ is a martingale in D, the s_t^D converge to $s(t, X_t)$ in L^2. Furthermore, by Lemma 5, we know that the $e^{-t/2}s_t^D$ form a reverse-time martingale in t. However, due to Doob's L^2-inequality, we get that for any $\epsilon > 0$,

$$\mathbb{E}[\sup_{\varepsilon \leqslant t \leqslant T}\|s_t^D - s_t^L\|^2] \leqslant e^T\mathbb{E}[\|s_\varepsilon^D - s_\varepsilon^L\|^2].$$

The right hand side is Cauchy and therefore is the left-hand side is too. Therefore, the convergence of s_t^D to s_t in L^2 is uniform on $[\varepsilon, T]$, i.e., continuous martingales s_t^D form a Cauchy sequence in the norm

$$\|N\|_{I,\epsilon} = \mathbb{E}[\sup_{t\geq\epsilon}|N_t|^2].$$

The continuous martingales are closed with respect to that norm (see Karatzas et al. (1991, Section 1.3)), and s_t is also a continuous martingale on $[\varepsilon, T]$. Since ϵ was arbitrary, we have shown that $s(t, X_t)$ is a continuous local martingale (in reverse time) up to $t = 0$.

Furthermore, since all the terms on the left-hand side converge in L^2, uniformly in t, so does the right-hand side. The right-hand side is $P^D C P^D$ Brownian motion for each D. Using again the that the spaces of martingales is closed and furthermore the Levy

36

characterization of Brownian motion, we find that the B_t^D have to converge to a C-Brownian motion B_t. Therefore,

$$Y_t = Y_0 + \frac{1}{2} \int_0^t Y_r \mathrm{d}r + \int_0^t s_{T-r} \mathrm{d}r + B_t \tag{23}$$

is indeed a weak solution to (6). It is a *weak* solution since Y_s is not necessarily measurable with respect to the filtration generated by B_s. In general, it can even be the other way around, see Proposition 17. ∎

Proposition 17 *Let X_t be a solution to (4). Assume $H = \mathbb{R}^D$. Then, the time-reversal $Y_t = X_{T-t}$ of the SDE satisfies the SDE*

$$\mathrm{d}Y_t = \frac{1}{2} Y_t \mathrm{d}t + C \nabla \log p_{T-t}(Y_t) \mathrm{d}t + \sqrt{C} \mathrm{d}B_t. \tag{24}$$

Here, B_t is a different Brownian motion B_t to W_t. If C has full rank, B_t can be defined on the same probability space as X_t itself.

Proof The above is the usual time-reversal formula. All we need to show is that for the special case of the forward SDE (4) the conditions from Haussmann and Pardoux (1986) are always satisfied. Assumption $(A)(i)$ in Haussmann and Pardoux (1986) is satisfied since b and σ are linear. Assumption $(A)(ii)$ is that for each $t_0 > 0$, it holds that

1. $\int_{t_0}^T \int_{B_R} |p(t,x)|^2 \mathrm{d}x < \infty$, and

2. $\int_{t_0}^T \int_{B_R} |\partial_{x_i} p(t,x)|^2 \mathrm{d}x < \infty$ for all im

where B_R is the ball of Radius R on $\mathbb{R}^D$ and $p(t,x)$ is the Lebesgue-density of $\mathbb{P}_t$. We now prove that both of these conditions hold. We have the explicit formula

$$p(t,x) = \frac{1}{\sqrt{2\pi(v_t \det C)^D}} \int e^{-\frac{\|x-x_0\|_U^2}{2}} \mathrm{d}\mu_{\mathrm{data}}(x_0)$$

and therefore, in particular $|p(t,x)| \leqslant \frac{1}{\sqrt{2\pi(v_t \det C)^D}}$. Since $\frac{1}{v_t} = \frac{1}{1-e^{-t}}$ is integrable on $[t_0, T]$ for $t_0 > 0$, this implies 1. Furthermore, we get that

$$\nabla p(t,x) = \frac{1}{\sqrt{2\pi(v_t \det C)^D}} \int C^{-1} x e^{-\frac{\|x-x_0\|_U^2}{2}} \mathrm{d}\mu_{\mathrm{data}}(x_0),$$

where we used the Leibniz rule to exchange differentiation and integration, since the integrand is bounded in x_0. On B_R this can be upper bounded by $\|\nabla p(t,x)\| \leqslant \frac{\|C^{-1}\| R}{\sqrt{2\pi(v_t \det C)^D}}$ which is again integrable on $[t_0, T]$.

By Haussmann and Pardoux (1986) we know that the time reversal Y_t will have the same generator as the SDE (24). In particular $M_t = Y_t - Y_0 - \int_0^t \frac{1}{2} Y_r + C \nabla \log p_{T-r}(Y_r) \mathrm{d}r$ is a continuous martingale with quadratic variation C with respect to the canonical filtration of Y_s.

We want to apply the Martingale representation theorem to express this martingale in terms of a Brownian motion. In general, one might need to extend the probability space to do so. However, since C has full rank, we can express the Brownian motion as $B_t = C^{-1/2} M_t$, which is defined on the same probability space as Y_t. ∎

Appendix F. Uniqueness Proofs

F.1 Proof of Theorem 12

First we prove Theorem 12:
Proof

Step 1: Prove that s is locally Lipschitz with respect to the Cameron-Martin Norm Recall that

$$s(t, x) = -\frac{1}{1 - e^{-t}} x + \frac{e^{-\frac{t}{2}}}{1 - e^{-t}} \mathbb{E}[X_0 | X_t = x]. \tag{25}$$

The $-\frac{1}{1-e^{-t}} x$ term is Lipschitz for any $t \in [\epsilon, T]$. Therefore, our goal is to show that $\mathbb{E}[X_0 | X_t = x]$ is Lipschitz in x too. We will frequently use that for $u \in U$, we can write the Radon-Nikodym derivative of $\mathcal{N}(u, v_t C)$ with respect to $\mathcal{N}(0, v_t C)$ as

$$\frac{\mathrm{d}\mathcal{N}(u, v_t C)}{\mathrm{d}\mathcal{N}(0, v_t C)}(x_t) = \exp\left(\frac{\langle u, x_t \rangle_U - \|u\|_U^2}{v_t} \right),$$

by the Cameron-Martin theorem (see, for example, Hairer (2009)). Here v_t is a shorthand notation for

$$v_t = 1 - e^{-t}.$$

To simplify notation, we will define

$$n(x_0, x_t) := \frac{\mathrm{d}\mathcal{N}(e^{-t} x_0, v_t C)}{\mathrm{d}\mathcal{N}(0, v_t C)}(x_t).$$

Then, the joint distribution of X_0 and X_t is given by

$$\mathrm{d}n(x_0, x_t) = \mathrm{d}(\mathcal{N}(0, v_t C)(x_t) \otimes \mu_{\mathrm{data}}(x_0)).$$

This can be seen by the following calculation:

$$\int_A \int_B n(x_0, x_t) \, \mathrm{d}\mathcal{N}(0, v_t C)(x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)$$

$$= \int_A \int_B \frac{\mathrm{d}\mathcal{N}(e^{-t} x_0, v_t C)}{\mathrm{d}\mathcal{N}(0, v_t C)}(x_t) \mathrm{d}\mathcal{N}(0, v_t C)(x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)$$

$$= \int_A \int_B \mathrm{d}\mathcal{N}(e^{-t} x_0, v_t C)(x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)$$

$$= \mathbb{P}[X_0 \in A, X_t \in B],$$

where we used that $\mathcal{N}(e^{-t}x_0, v_t C)$ is the transition kernel of the forward SDE (5). We show that

$$f(x_t) = \frac{\int x_0 n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)}{\int n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)}$$

is a version of the conditional expectation $\mathbb{E}[X_0 | X_t = x]$. The function f is $\sigma(X_t)$ measurable by Fubini's theorem. Furthermore, for $A \in \sigma(X_t)$,

$$
\begin{aligned}
\mathbb{E}_{X_t}[1_A f(X_t)] &= \int_A \frac{\int_H x_0 n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)}{\int_H n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)} \, \mathrm{d}\mathbb{P}_t(x_t) \\
&= \int_H \int_A \frac{\int_H x_0 n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)}{\int_H n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)} n(\tilde{x}_0, x_t) \, \mathrm{d}\mathcal{N}(0, v_t C)(x_t) \, \mathrm{d}\mu_{\mathrm{data}}(\tilde{x}_0) \\
&= \int_A \int_H x_0 n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0) \frac{\int_H n(\tilde{x}_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(\tilde{x}_0)}{\int_H n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)} \, \mathrm{d}\mathcal{N}(0, v_t C)(x_t) \\
&= \int_A \int_H x_0 n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0) \, \mathrm{d}\mathcal{N}(0, v_t C)(x_t) = \mathbb{E}[1_A X_0].
\end{aligned}
$$

Since these two properties define the conditional expectation, we have shown that

$$\mathbb{E}[X_0 | X_t = x] = f(x)$$

almost surely. We will now proceed to show that f is Lipschitz with respect to the Cameron-Martin norm $\| \cdot \|_U$. For notational convenience, we will define

$$\pi_t(x_t) = \int n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0) = \int \exp\left(\frac{2\langle e^{-t/2} x_0, x_t \rangle_U - e^{-t} \|x_0\|_U^2}{2v_t} \right) \mathrm{d}\mu_{\mathrm{data}}(x_0). \quad (26)$$

We see that

$$
\begin{aligned}
\pi_t(x_t + z) &= \int \exp\left(\frac{2\langle e^{-t/2} x_0, x_t + z \rangle_U - e^{-t} \|x_0\|_U^2}{2v_t} \right) \mathrm{d}\mu_{\mathrm{data}}(x_0) \\
&= \int \exp\left(\frac{\langle e^{-t/2} x_0, z \rangle_U}{v_t} \right) n(x_0, x_t) \, \mathrm{d}\mu_{\mathrm{data}}(x_0),
\end{aligned}
\quad (27)
$$

which differs from (26) only by $\exp\left(\frac{\langle e^{-t/2} x_0, z \rangle_U}{v_t} \right)$. By our assumption that the support of μ_{data} is contained in a Cameron-Martin ball of size R. Therefore,

$$\frac{\langle z, e^{-t/2} x_0 \rangle_U}{v_t} \le \frac{e^{-t/2}}{v_t} \|z\|_U R, \quad (28)$$

and

$$\exp\left(-R\|z\|_U \frac{e^{-t/2}}{v_t} \right) \le \frac{\pi_t(x_t + z)}{\pi_t(x_t)} \le \exp\left(R\|z\|_U \frac{e^{-t/2}}{v_t} \right). \quad (29)$$

With these estimates out of the way, let us show local Lipschitz continuity:

$$
\begin{aligned}
\|f(x_t + z) - f(x_t)\|_U &= \left\| \frac{\int x_0 \, n(x_0, x_t + z) \mathrm{d}\mu_{\mathrm{data}}(x_0)}{\pi_t(x_t + z)} - \frac{\int x_0 \, n(x_0, x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)}{\pi_t(x_t)} \right\|_U \\
&=: \left\| \frac{A'}{\pi_t(x_t + z)} - \frac{A}{\pi_t(x_t)} \right\|_U.
\end{aligned}
$$

We rewrite the above as

$$\left\| \frac{A'}{\pi_t(x_t + z)} - \frac{A}{\pi_t(x_t)} \right\|_U \leq \left| 1 - \frac{\pi_t(x_t + z)}{\pi_t(x_t)} \right| \left\| \frac{A'}{\pi_t(x_t + z)} \right\|_U + \frac{1}{\pi_t(x_t)} \|A' - A\|_U.$$

We see that

$$\left| 1 - \frac{\pi_t(x_t)}{\pi_t(x_t)} \right| \left\| \frac{A'}{\pi_t(x_t + z)} \right\|_U \leq \left(\exp\left(\frac{e^{-t/2}}{v_t} \|z\|_U R \right) - 1 \right) \|\mathbb{E}[X_0 | X_t = x_t + z]\|_U,$$

where we used (29) for the first term. We also get that

$$\frac{\|A - A'\|_U}{\pi_t(x_t)} \leq \left(\exp\left(\frac{e^{-t/2}}{v_t} \|z\|_U R \right) - 1 \right) \frac{1}{p_t(x_t)} \int \|x_0\|_U \frac{\mathrm{d}\mathcal{N}(e^{-t/2}x_0, v_t C)}{\mathrm{d}\mathcal{N}(0, v_t C)}(x_t) \mathrm{d}\mu_{\mathrm{data}}(x_0)$$

$$\leq \left(\exp\left(\frac{e^{-t/2}}{v_t} \|z\|_U R \right) - 1 \right) R$$

where we used (28) and (27) and our assumption that $\|x_0\|_U \leq R$. Putting it all together, we get that

$$\|f(x_t + z) - f(x_t)\| \leq 2 \left(\exp\left(\frac{e^{-t/2}}{v_t} \|z\|_U R \right) - 1 \right) R,$$

where we again used that $\|x_0\| \leq R$ to bound $f(x_t + z)$. However, we can strengthen this bound. For any N, it holds that

$$\|f(x_t + z) - f(x_t)\|_U \leq \sum_{i=1}^{N} \|f(x_t + z\frac{i}{N}) - f(x_t + z\frac{i-1}{N})\|_U$$

$$\leq N \, 2 \left(\exp\left(\frac{e^{-t/2}}{v_t} \frac{\|z\|_U}{N} R \right) - 1 \right) R.$$

In particular, we can take the limit $N \to \infty$ and get that

$$\|f(x_t + z) - f(x_t)\|_U \leq \frac{\mathrm{d}}{\mathrm{d}h}|_{h=0} 2 \left(\exp\left(\frac{e^{-t/2}}{v_t} h\|z\|_U R \right) - 1 \right) R = 2R^2 \frac{e^{-t/2}}{v_t} \|z\|_U.$$

From this we can conclude that f has the global Lipschitz constant $2R^2 \frac{e^{-t/2}}{v_t}$. From (25) we see that there is a version of $s(t, \cdot)$, such that for any $x_t, y_t \in H$

$$\|s(t, x_t) - s(t, y_t)\|_U \leq L_t \|x_t - y_t\|_U, \quad L_t = \frac{1}{(1 - e^{-t})^2} \max\{1, 2R^2 e^{-t}\}.$$

Step 2: Existence of solutions Fix a C-Wiener process W_t. Denote by $M : \mathcal{C}([0, T - u], H) \to \mathcal{C}([0, T - u])$ the map

$$(My)(t) = \int_0^t s(T - r, y_r)\mathrm{d}r + W_t.$$

Then we have if $u < T - \varepsilon$,

$$\sup_{t \leqslant u} \|My(t) - M\tilde{y}(t)\|_U \leqslant \int_0^u \|s(T - r, y_r) - s(T - r, \tilde{y}_r)\|_U \mathrm{d}r \leqslant u L_{T-\varepsilon} \sup_{t \leqslant u} \|y_t - \tilde{y}_t\|_U. \quad (30)$$

Here we used that we can apply Jensen's inequality because $\| \cdot \|_U : H \to [0, \infty]$ is lower-semicontinuous on H; see Proposition 18. We now choose u smaller than $\frac{1}{L_{T-\varepsilon}}$. Starting with any y^0, we can now define the sequence $y^{n+1} = My^n$. Applying (30) and noting that $\frac{u}{L_{T-\varepsilon}} < 1$, we see that y_n is Cauchy with respect to $\sup \| \cdot \|_U$ and therefore also with respect to $\sup \| \cdot \|_H$ and has a limit. We then have a strong solution w.r.t. to the fixed Wiener process W_t on $[0, u]$. We can extend the solution to $[0, T - \varepsilon]$ by repeating this process and gluing the solutions together. However, we cannot apply Banach's fix point theorem to get uniqueness, since $\|y - \tilde{y}\|_U$ might be infinite for two solutions y and $\tilde{y}$.

Step 3: Strong uniqueness of solutions We now assume we have two solutions to the reverse SDE, Y_t and $\tilde{Y}_t$ solving the reverse SDE (6) with respect to the same Wiener process B_t and $Y_0 = \tilde{Y}_0$. Since Y_t and $\tilde{Y}_t$ are not necessarily in U, we have to be a bit careful before directly applying Grönwall. Again, in Proposition 18 we have proven that $\| \cdot \|_U : H \to [0, \infty]$ is lower-semicontinuous on H. We define the seminorms

$$\|x\|_{U^D} = \|P^D x\|_U$$

where P^D is the projection operator defined in Section . Those are smooth functions on H and therefore the map $t \mapsto \|Y_t - \tilde{Y}_t\|_{U^D}$ is continuous and we can apply Grönwall:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Y_t - \tilde{Y}_t\|_{U^D} \leq \frac{e^{-\frac{t}{2}}}{1 - e^{-t}} \|\mathbb{E}[X_0|X_t = Y_t] - \mathbb{E}[X_0|X_t = \tilde{Y}_t]\|_{U^D} \leq \frac{e^{-\frac{T-t}{2}}}{1 - e^{-(T-t)}} 2R.$$

which in particular shows that

$$\|Y_t - \tilde{Y}_t - (Y_s - \tilde{Y}_s)\|_{U^D} \leq \frac{e^{-\frac{T-t}{2}}}{1 - e^{-T-t}} 2R(t - s)$$

for $t \geq s$. By taking the limit, we see that $t \mapsto \|Y_t - \tilde{Y}_t\|_U$ is continuous. Therefore, we can apply Grönwall to that quantity (continuity is a requirement for Grönwall).

Furthermore, by using the above calculation for $\tau = 0$, we find that $\|Y_t - \tilde{Y}_t\|_U$ will be finite for all t, if $\|Y_0 - \tilde{Y}_0\|_U$ is finite (even if Y_t and $\tilde{Y}_t$ are almost surely not in U). Now, since $Y_0 = \tilde{Y}_0$ and therefore $\|Y_0 - \tilde{Y}_0\|_U = 0$,

$$\|Y_t - \tilde{Y}_t\|_U \leqslant \int_0^t \|s(T - r, Y_r) - s(T - r, \tilde{Y}_r)\|_{U^d} \mathrm{d}r \leqslant \int_0^t L_{T-r}^2 \|Y_t - \tilde{Y}_t\|_U \mathrm{d}r.$$

If $t < T$, the Lipschitz constant L_{T-t} is finite. Therefore, we can apply Grönwall to see that $\|Y_t - \tilde{Y}_t\|_U = 0$ to prove uniqueness on $[0, t]$ for any $t < T$. Hence, we have shown that there is a unique strong solution on $[0, T - \varepsilon]$. Since ε was arbitrary, we have shown strong uniqueness on $[0, T]$. ∎

F.2 Proof of Theorem 13

Now we prove Theorem 13:

Proof

Step 0: A priori bounds   Let H be any Hilbert space on which $\mathcal{N}(0, C_\mu)$ is supported. Let e_i be an eigenbasis von C_μ and C, which exists by our assumptions. Let c_i and μ_i be the eigenvalues associated with C and C_μ respectively, i.e.,

$$Ce_i = c_i e_i, \quad C_\mu e_i = \mu_i e_i.$$

Furthermore, we define C_t by

$$C_t = (e^{-t} C_\mu + (1 - e^{-t})C), \tag{31}$$

which would be the covariance of X_t at time t in case $\Phi = 0$. The following operators are all bounded: $C_\mu C_t^{-1}$, $C C_t^{-1}$ and $C C_t C_\mu^{-1}$. We will show it for the first operator, and the others follow by similar arguments. The first operator will have eigenvalues

$$\lambda_i = \frac{\mu_i}{e^{-t} c_i + (1 - e^{-t})\mu_i}.$$

We know that $\mu_i \to 0$ since C_μ is trace class. For $c_i \to 0$, the eigenvalues converge to $\lambda_i \to \frac{1}{(1-e^{-t})}$, while as $c_i \to \infty$, we get that $\lambda_i \to 0$. Furthermore, the derivative with respect to c_i is negative, which shows that the λ_i are bounded by $\frac{1}{1-e^{-t}}$. A similar calculation can be done for the other operators listed.

Step 1: Rewrite the reverse SDE   The goal of this section is to show that we can write the drift in infinite dimensions as

$$s(t, x_t) = -e^{\frac{t}{2}} \mathbb{E}[C(C_\mu C_t^{-1})^{-1} \nabla \Phi(X_0) | X_t = x_t] + C C_t^{-1} x_t$$

without assuming any more conditions on Φ or $\nabla \Phi$. To that end, we will argue in finite dimensions and take the limit in the end. Henceforth, unless we say otherwise, everything will be in finite dimensions for this step. The projection of μ_{data} to H^D will be defined by

$$\mu_{\text{data}}^D = \exp(-\Phi^D) \mathcal{N}(0, C_\mu^D),$$

where C_μ^D is defined as in Section E.1 and

$$
\begin{aligned}
\exp(-\Phi^D(x^D)) &= \mathbb{E}_{\mathcal{N}(0,C)}[\exp(-\Phi(X)) | X^D = x^D] \\
&= \int \exp(-\Phi(x^D, x^{D+1:\infty})) \, d\mathcal{N}(0, C^{D+1:\infty})(x^{D+1:\infty})
\end{aligned}
$$

For the gradient it will then hold that

$$
\begin{aligned}
\nabla \Phi^D(x^D) &= \nabla \log \exp(\Phi^D(x^D)) \\
&= \frac{\int \nabla_{x^D} \exp(-\Phi(x^D, x^{D+1:\infty})) \, d\mathcal{N}(0, C^{D+1:\infty})(x^{D+1:\infty})}{\int \exp(-\Phi(x^D, x^{D+1:\infty})) \, d\mathcal{N}(0, C^{D+1:\infty})(x^{D+1:\infty})} \\
&= \frac{-\int \nabla_{x^D} \Phi(x^D, x^{D+1:\infty}) \exp(-\Phi(x^D, x^{D+1:\infty})) \, d\mathcal{N}(0, C^{D+1:\infty})(x^{D+1:\infty})}{\int \exp(-\Phi(x^D, x^{D+1:\infty})) \, d\mathcal{N}(0, C^{D+1:\infty})(x^{D+1:\infty})} \\
&= \mathbb{E}_{\mu_{\text{data}}}[\nabla \Phi(x) | X^D = x^D].
\end{aligned}
$$

We denote by $(\tilde{X}_t : 0 \leq t \leq T)$ the Gaussian solution, started in $\tilde{X}_0 \sim \mathcal{N}(0, C)$ and as always by $(X_t : 0 \leq t \leq T)$ the solution to the forward process started in $X_0 \sim \mu_{\text{data}}$. To be precise,

$$
\begin{pmatrix} \tilde{X}_0 \\ \tilde{X}_t \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} C_\mu & e^{-\frac{t}{2}}C_\mu \\ e^{-\frac{t}{2}}C_\mu & C_t \end{pmatrix} \right),
$$

where C_t is defined in (31). Therefore,

$$
\tilde{X}_0 | \tilde{X}_t \sim \mathcal{N}\left(e^{-\frac{t}{2}} C_\mu C_t^{-1} \tilde{X}_t, C_\mu - e^{-t} C_\mu C_t^{-1} C_\mu \right).
$$

Then we have that

$$
\frac{\mathrm{d}p_t}{\mathrm{d}\mathcal{N}(0, C_t)}(x) = \mathbb{E}[\exp(-\Phi(\tilde{X}_0)) | \tilde{X}_t = x] = \mathbb{E}_{\mathcal{N}\left(e^{-\frac{t}{2}} C_\mu C_t^{-1} x, C_\mu - e^{-t} C_\mu C_t^{-1} C_\mu \right)}[\exp(-\Phi(\tilde{X}_0))].
$$

$$(32)$$

We denote by

$$
A_t = e^{-\frac{t}{2}} C_\mu C_t^{-1}, \quad Q_t = C_\mu - e^{-t} C_\mu C_t^{-1} C_\mu.
$$

Note that for $C = C_\mu$ all of the definitions simplify to easier terms. Taking $\nabla \log$ of the above leads to

$$
\begin{aligned}
\nabla_{x_t} \frac{\mathrm{d}p_t}{\mathrm{d}\mathcal{N}(0, C_t)}(x_t) &= \frac{1}{Z} \int \nabla_{x_t} \exp(-\|A_t x_0 - x_t\|_{Q_t}^2) \exp(-\Phi(x_0)) \mathrm{d}x_0 \\
&= \frac{1}{Z} \int -A_t^{-1} \nabla_{x_0} \exp(-\|A_t x_0 - x_t\|_{Q_t}^2) \exp(-\Phi(x_0)) \mathrm{d}x_0 \\
&= \frac{1}{Z} \int A_t^{-1} \exp(-\|A_t x_0 - x_t\|_{Q_t}^2) \nabla_{x_0} \exp(-\Phi(x_0)) \mathrm{d}x_0 \\
&= -\frac{1}{Z} \int A_t^{-1} \exp(-\|A_t x_0 - x_t\|_{Q_t}^2) \nabla_{x_0} \Phi(x_0) \exp(-\Phi(x_0)) \mathrm{d}x_0,
\end{aligned}
$$

where B is the normalizing constant. Therefore,

$$
\begin{aligned}
C\nabla_{x_t} \log \frac{\mathrm{d}p_t}{\mathrm{d}\mathcal{N}(0, C_t)}(x_t) &= \frac{-\int CA_t^{-1} \exp(-\|A_t x_0 - x_t\|_{Q_t}^2) \nabla_{x_0} \Phi(x_0) \exp(-\Phi(x_0)) \mathrm{d}x_0}{\int \exp(-\|A_t x_0 - x_t\|_{Q_t}^2) \exp(-\Phi(x_0)) \mathrm{d}x_0} \\
&= \frac{\int CA_t^{-1} \nabla\Phi(x) \exp(-\Phi(x)) \mathrm{d}\mathcal{N}(A_t x_t, Q_t)}{\int \exp(\Phi(x)) \mathrm{d}\mathcal{N}(A_t x_t, Q_t)} \\
&= \mathbb{E}[-CA_t^{-1} \nabla\Phi(X_0) | X_t = x_t].
\end{aligned}
$$

We make the dimension dependence explicit and get that

$$
\begin{aligned}
C\nabla_{x_t^D} \log \frac{\mathrm{d}p_t^D}{\mathrm{d}\mathcal{N}(0, C_t^D)}(x_t^D) &= \mathbb{E}[-C^D (A_t^{-1})^D \nabla\Phi^D(X_0^D) | X_t^D = x_t^D] \\
&= \mathbb{E}[-C^D (A_t^{-1})^D \mathbb{E}[P^D \nabla\Phi(X_0) | X_0^D] | X_t^D = x_t^D] \\
&= \mathbb{E}[-C^D (A_t^{-1})^D P^D \nabla\Phi(X_0) | X_t^D = x_t^D] \\
&= P^D \mathbb{E}[-CA_t^{-1} \nabla\Phi(X_0) | X_t^D = x_t^D].
\end{aligned}
$$

Furthermore,

$$s^D(t, X_t^D) = C\nabla \log p_t^D(x_t^D) = C\nabla \log \frac{p_t^D(x_t^D)}{\mathcal{N}(0, C_t)} + C\nabla \log \mathcal{N}(0, C_t)(x_t^D)$$
$$= P^D \mathbb{E}[-CA_t^{-1}\nabla\Phi(X_0)|X_t^D = x_t^D] + C^D(C_t^{-1})^D x_t^D. \quad (33)$$

By Lemma 16 s^D converges almost surely to s. Furthermore, since CA_t^{-1} is bounded and $\nabla\Phi(X_0)$ is Lipschitz, we get that

$$\mathbb{E}[\|\mathbb{E}[CA_t^{-1}\nabla \log \Phi(X_0)|X_t]\|^2] \lesssim \mathbb{E}[\|\nabla \log \Phi(X_0)\|^2] \lesssim \mathbb{E}[\|X_0\|^2] < \infty.$$

Therefore, also the first term in (33) converges to its infinite dimensional counterpart by Lemma 16. The second term in (33) also converges. Therefore, we can take the limit on both sides and obtain

$$s(t, x_t) = \mathbb{E}[-CA_t^{-1}\nabla\Phi(X_0)|X_t = x_t] + CC_t^{-1}x_t.$$

The formula (32) for p_t holds in infinite dimensions too, and therefore we can write the conditional expectation as

$$f(t, x_t) = \mathbb{E}[-CA_t^{-1}\nabla\Phi(\tilde{X}_0)|\tilde{X}_t = x_t] = \frac{\int CA_t^{-1}\nabla\Phi(x)\exp(-\Phi(x))\mathrm{d}\mathcal{N}(A_t x_t, Q_t)}{\int \exp(-\Phi(x))\mathrm{d}\mathcal{N}(A_t x_t, Q_t)}.$$

Step 2: Local Lipschitzness in with respect to $\|\cdot\|$ Since CC_t^{-1} is bounded by step 0, it suffices to show that f is locally Lipschitz. We will now bound the difference

$$f(t, y_t) - f(t, x_t)$$
$$= \frac{\int CA_t^{-1}\nabla\Phi(x)\exp(-\Phi(x))\mathrm{d}\mathcal{N}(A_t x_t, Q_t)}{\int \exp(-\Phi(x))\mathrm{d}\mathcal{N}(A_t x_t, Q_t)} - \frac{\int CA_t^{-1}\nabla\Phi(x)\exp(-\Phi(x))\mathrm{d}\mathcal{N}(A_t y_t, Q_t)}{\int \exp(-\Phi(x))\mathrm{d}\mathcal{N}(A_t y_t, Q_t)}$$
$$=: \frac{B_1}{Z_1} - \frac{B_2}{Z_2} = \left(\frac{1}{Z_1} - \frac{1}{Z_2}\right)B_2 - \frac{1}{Z_1}(B_1 - B_2) = \left(\frac{Z_1 - Z_2}{Z_1 Z_2}\right)B_2 - \frac{1}{Z_1}(B_1 - B_2).$$

We will fix an $R \geq 0$ and assume that $\|x_t\|, \|y_t\| \leq R$. Then, since A_t is bounded, there exists an $\tilde{R}$ such that $\|A_t x_t\|, \|A_t y_t\| \leq \tilde{R}$.

Then

$$Z_1 = \int \exp(-\Phi(x))\mathrm{d}\mathcal{N}(A_t x_t, Q_t)(x) = \int \exp(-\Phi(x + A_t x_t))\mathrm{d}\mathcal{N}(0, Q_t)(x)$$
$$\geq \int \exp(-(E_1 + E_2\|x + A_t x_t\|^2))\mathrm{d}\mathcal{N}(0, Q_t)(x)$$
$$= \exp(-E_1 + 2\|A_t x_t\|^2) \int \exp(-E_2\|x\|^2)\mathrm{d}\mathcal{N}(0, Q_t)(x)$$
$$\gtrsim E\exp(-2\|A_t x\|^2) \geq E\exp(-2\tilde{R}^2)$$

where E is a finite constant that only depends on C_μ, C, L and the E_i. A similar bound holds for Z_2. For

$$Z_1 - Z_2 = \int \exp(-\Phi(x + A_t x_t)) - \exp(-\Phi(x + A_t y_t))\mathrm{d}\mathcal{N}(0, Q_t)(x)$$
$$\leq \int \exp(-E_0)L\|A_t x_t - A_t y_t\|\mathrm{d}\mathcal{N}(0, Q_t)(x)$$
$$= \exp(-E_0)L\|A_t x_t - A_t y_t\| \leq EL\|x_t - y_t,\|$$

where we used that if Φ is C^1 and its derivative is bounded by L, then $\exp(-\Phi)$ has a derivative bounded by $\exp(-\inf \Phi)L$. Furthermore, we get that

$$\int CA_t^{-1} \nabla \Phi(x) \exp(-\Phi(x)) \mathrm{d}\mathcal{N}(A_t x_t, Q_t)$$

$$\leqslant \exp(-E_0) \int CA_t^{-1} \nabla \Phi(x + A_t x_t) \mathrm{d}\mathcal{N}(0, Q_t)$$

$$\leqslant \exp(-E_0) \int \|CA_t^{-1}\| (\|\nabla \Phi(0)\| + L\|x + A_t x_t\|) \mathrm{d}\mathcal{N}(0, Q_t)$$

$$\leqslant \exp(-E_0) \|CA_t^{-1}\| \left(\|\nabla \Phi(0)\| + L\|A_t x_t\| + L \int \|x\| \mathrm{d}\mathcal{N}(0, Q_t) \right)$$

$$\leqslant E(1 + \|x_t\|) \leqslant E(1 + R),$$

where we used that CA_t^{-1} and A_t are bounded. We also get that

$$\|B_1 - B_2\|$$

$$= \int \|CA_t^{-1} \nabla \Phi(x + A_t x_t)\| |\exp(-\Phi(x + A_t x_t)) - \exp(-\Phi(x + A_t y_t))| \mathrm{d}\mathcal{N}(0, Q_t)(x)$$

$$+ \int \|CA_t^{-1} \nabla \Phi(x + A_t y_t) - CA_t^{-1} \nabla \Phi(x + A_t x_t)\| \exp(-\Phi(x + A_t y_t)) \mathrm{d}\mathcal{N}(0, Q_t)(x)$$

$$\leqslant \exp(-E_0) L\|A_t x_t - A_t y_t\| \|CA_t^{-1}\| \left(\|\nabla \Phi(0)\| + L \left(\|A_t x_t\| + \int \|x\| \mathrm{d}\mathcal{N}(0, Q_t)(x) \right) \right)$$

$$+ L\|CA_t^{-1}\| \|A_t x_t - A_t y_t\| \exp(-E_0)$$

$$\leqslant E\|x_t - y_t\|(1 + R),$$

where we again used the boundedness of CA_t^{-1} and A_t. Putting it all together, we get that

$$\left\| C\nabla \log \frac{\mathrm{d}p_t}{\mathrm{d}\mathcal{N}(0, C_t)}(x_t) - C\nabla \log \frac{\mathrm{d}p_t}{\mathrm{d}\mathcal{N}(0, C_t)}(y_t) \right\|$$

$$\leqslant E\exp(4\tilde{R}^2)L\|x_t - y_t\| + E\exp(2\tilde{R})\|x_t - y_t\|(1 + R) \leqslant E\exp(4\tilde{R}^2)\|x_t - y_t\|.$$

Step 3: Strong uniqueness and existence Using the local Lipschitzness, we apply Grönwall to obtain strong uniqueness of solutions. This is a standard argument and similar to what we did in Step 3 of the proof of Theorem 13. Alternatively, see, for example, Karatzas et al. (1991, Theorem 2.5 in Section 5.2.B).

We can now prove weak existence of the reverse SDE. By Theorem 9, the time reversal will be a weak solution with initial condition $\mathbb{P}_T$. Denote the path measure of Y by $\mathbb{Q}$. Under the assumptions of the Theorem, $\mathcal{N}(0, C)$ will be absolutely continuous with respect to p_T. We define $\tilde{\mathbb{Q}}$ by

$$\frac{\mathrm{d}\tilde{\mathbb{Q}}}{\mathrm{d}\mathbb{Q}}(y_{[0,T]}) = \frac{\mathrm{d}\mathcal{N}(0, C)}{\mathrm{d}\mathbb{P}_T}(y_0).$$

$\tilde{\mathbb{Q}}$ is then the path measure of a solution $\tilde{Y}$ to 6 which has initial condition $\mathcal{N}(0, C)$, therefore we have constructed a weak solution. With that we conclude the proof since, weak existence together with strong uniqueness imply strong existence, see Karatzas et al. (1991, Section 5.3). ■

Appendix G. Wasserstein-Bound Proof

We now prove Theorem 14:

Proof We prove the theorem with the finite-dimensional notation, but one can replace $\nabla \log p_t$ by $s(t, \cdot)$ and nothing changes. We will partition $[0, T]$ into $\tau = \{0 = t_0, \ldots, t_N = T\}$. For the given partition we denote

$$\lfloor t \rfloor = \max_{t_i \in \tau}\{t_i \leqslant t\}, \quad \lceil t \rceil = \min_{t_i \in \tau}\{t_i \geqslant t\}, \quad \Delta = \max_{k=0,\ldots,N-1} t_{k+1} - t_k.$$

We couple two strong solutions of Y_t and $\tilde{Y}_t$ for the same Brownian motion B_t. These strong solutions exist because of the assumptions of the theorem and Theorem 12 or Theorem 13. The difference between Y_t and $\tilde{Y}_t$ can be bounded as follows:

$$\begin{aligned}
\mathrm{d}\|Y_t - \tilde{Y}_t\| &= \frac{1}{2}\|Y_t - \tilde{Y}_t\| + \frac{1}{\|Y_t - \tilde{Y}_t\|}\langle Y_t - \tilde{Y}_t, C\nabla \log p_{T-t}(Y_t) - \tilde{s}(T - \lfloor t \rfloor, \tilde{Y}_{\lfloor t \rfloor})\rangle \\
&\leqslant \frac{1}{2}\|Y_t - \tilde{Y}_t\| + \|C\nabla \log p_{T-t}(Y_t) - s_\theta(T - t, \tilde{Y}_{\lfloor t \rfloor})\|.
\end{aligned}$$

We bound

$$\begin{aligned}
&\|C\nabla \log p_{T-t}(Y_t) - \tilde{s}(T - t, \tilde{Y}_t)\| \\
&\leqslant \|C\nabla \log p_{T-t}(Y_t) - C\nabla \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\| \\
&\quad + \|C\nabla \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor}) - \tilde{s}(\lfloor t \rfloor, Y_{\lfloor t \rfloor})\| + \|\tilde{s}(\lfloor t \rfloor, Y_{\lfloor t \rfloor}) - \tilde{s}(\lfloor t \rfloor, \tilde{Y}_{\lfloor t \rfloor})\| \\
&\leqslant \|\nabla_U \log p_{T-t}(Y_t) - \nabla_U \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\| \\
&\quad + \|C\nabla \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor}) - \tilde{s}(\lfloor t \rfloor, Y_{\lfloor t \rfloor})\| + L_s\|Y_{\lfloor t \rfloor} - \tilde{Y}_{\lfloor t \rfloor}\|.
\end{aligned}$$

We take the supremum to get rid of the delay term $\|Y_{\lfloor t \rfloor} - \tilde{Y}_{\lfloor t \rfloor}\|$ and obtain

$$\begin{aligned}
\sup_{\tau \leqslant s}\|Y_r - \tilde{Y}_r\| &\leqslant L_s'\int_0^s \sup_{r \leqslant t}\|Y_r - \tilde{Y}_r\|\mathrm{d}t + \int_0^s \|\nabla_U \log p_{T-t}(Y_t) - \nabla_U \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\|\mathrm{d}t \\
&\quad + \int \left\|\nabla_U \log \frac{p_{T-\lfloor t \rfloor}}{\nu}(Y_{\lfloor t \rfloor}) - \tilde{s}(\lfloor t \rfloor, Y_{\lfloor t \rfloor})\right\|\mathrm{d}t,
\end{aligned}$$

where $L_s' = L + \frac{1}{2}$. Squaring the above expression and taking expectations, we arrive at

$$\begin{aligned}
&\mathbb{E}[\sup_{\tau \leqslant s}\|Y_r - \tilde{Y}_r\|^2] \\
&\lesssim L_s'^2\int_0^s \mathbb{E}[\sup_{r \leqslant t}\|Y_r - \tilde{Y}_r\|^2]\mathrm{d}t + \int_0^s \mathbb{E}[\|\nabla_U \log p_{T-t}(Y_t) - \nabla_U \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\|^2]\mathrm{d}t \\
&\quad + \mathbb{E}\left[\left\|\nabla_U \log \frac{p_{T-\lfloor t \rfloor}}{\nu}(Y_{\lfloor t \rfloor}) - \tilde{s}(\lfloor t \rfloor, Y_{\lfloor t \rfloor})\right\|^2\right]\mathrm{d}t \\
&= L_s'^2\int_0^s \mathbb{E}[\sup_{r \leqslant t}\|Y_r - \tilde{Y}_r\|^2]\mathrm{d}t + \int_0^s B_1 + B_2\mathrm{d}t
\end{aligned}$$

We start by bounding B_1:

$$
\begin{aligned}
B_1 &\leqslant \mathbb{E}\left[\left\|\nabla_U \log p_{T-t}(Y_t) - e^{\frac{(t-\lfloor t \rfloor)}{2}}\nabla_U \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\right\|^2\right] \\
&\quad + \left(1 - e^{\frac{(t-\lfloor t \rfloor)}{2}}\right)^2 \mathbb{E}[\|\nabla_U \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\|^2] \\
&= \mathbb{E}[\|\nabla_U \log p_{T-t}(Y_t)\|^2] - \mathbb{E}\left[\left\|e^{\frac{(t-\lfloor t \rfloor)}{2}}\nabla_U \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\right\|^2\right] \\
&\quad + \left(1 - e^{\frac{(t-\lfloor t \rfloor)}{2}}\right)^2 \mathbb{E}[\|\nabla_U \log p_{T-\lfloor t \rfloor}(Y_{\lfloor t \rfloor})\|^2]
\end{aligned}
$$

where we used that the L^2 norm of the a martingale M_t difference is the difference of the L^2 norms, i.e., $\mathbb{E}[\|M_t - M_s\|^2] = \mathbb{E}[\|M_t\|^2] - \mathbb{E}[\|M_s\|^2]$ for $t \geqslant s$. Then

$$
\begin{aligned}
\int_0^T B_1 \mathrm{d}t &\leqslant \sum_{i=1}^N (\mathbb{E}[\|\nabla_U \log p_{T-t_{k+1}}(Y_{t_{k+1}})\|^2] - \mathbb{E}[\|e^{(t_{k+1}-t_k)}\nabla_U \log p_{T-t_k}(Y_{t_k})\|^2])(t_{k+1} - t_k) \\
&\quad + \left(1 - e^{\frac{\Delta t}{2}}\right)^2 \mathbb{E}[\|\nabla_U \log p_T(Y_T)\|^2] \\
&\leqslant \Delta t \sum_{i=1}^N (\mathbb{E}[\|\nabla_U \log p_{T-t_{k+1}}(Y_{t_{k+1}})\|^2] - \mathbb{E}[\|e^{(t_{k+1}-t_k)}\nabla_U \log p_{T-t_k}(Y_{t_k})\|^2]) \\
&\quad + \left(1 - e^{\frac{\Delta t}{2}}\right)^2 \mathbb{E}[\|\nabla_U \log p_T(Y_T)\|^2] \\
&\leqslant \Delta t \mathbb{E}[\|\nabla_U \log p_0(Y_T)\|^2] + \left(1 - e^{\frac{\Delta t}{2}}\right)^2 \mathbb{E}[\|\nabla_U \log p_T(Y_T)\|^2] \\
&= O(\Delta t)\mathbb{E}[\|\nabla_U \log p_0(Y_T)\|^2]
\end{aligned}
$$

where we used that the L^2 norm of a martingale is increasing. The term B_2 is nothing more than the loss.

Putting it all together, we arrive at

$$
\begin{aligned}
\mathbb{E}[\sup_{r \leqslant s} \|Y_r - \tilde{Y}_r\|^2] &\leqslant L^2 \int_0^s \mathbb{E}[\sup_{r \leqslant t} \|Y_r - \tilde{Y}_r\|^2]\mathrm{d}t + O(\Delta t)\mathbb{E}[\|\nabla_U \log p_0(Y_T)\|^2] + \text{Loss} \\
&= L^2 \int_0^s \mathbb{E}[\sup_{r \leqslant t} \|Y_r - \tilde{Y}_r\|^2] + \text{Error}
\end{aligned}
$$

and can apply Grönwall to get that

$$
\mathbb{E}[\sup_{r \leqslant s} \|Y_r - \tilde{Y}_r\|^2] \leqslant (\mathbb{E}[\|Y_0 - \tilde{Y}_0\|^2] + \text{Error})\exp(L^2 s).
$$

Since $Y_T \sim \mu_{\text{data}}$ and $\tilde{Y}_T \sim \hat{\mu}_{\text{sample}}$ we found a coupling of μ_{data} and $\hat{\mu}_{\text{sample}}$ and bounded its L^2 distance. We have not picked the coupling of $Y_0 \sim p_T$ and $\tilde{Y}_0 \sim \mathcal{N}(0, C)$ yet. Therefore we just pick a ε-optimal coupling in the squared Wasserstein distance, i.e., $\mathbb{E}[\|Y_0 - \tilde{Y}_0\|^2] \leqslant \mathcal{W}_2^2(p_T, \mathcal{N}(0, C)) + \varepsilon$ and obtain

$$
\mathcal{W}_2^2(\hat{\mu}_{\text{sample}}, \mu_{\text{data}}) \leqslant \mathbb{E}[\sup_{r \leqslant T} \|Y_r - \tilde{Y}_r\|^2] \leqslant (\mathcal{W}_2(p_T, \mathcal{N}(0, C)) + \varepsilon + \text{Error})\exp(L^2 T).
$$

Since ε was arbitrary, the statement of the theorem follows, we actually get

$$\mathcal{W}_2^2(\hat{\mu}_{\text{sample}}, \mu_{\text{data}}) \leqslant \mathbb{E}[\sup_{r \leqslant T} \|Y_r - \tilde{Y}_r\|^2] \quad \leqslant \quad (\mathcal{W}_2(p_T, \mathcal{N}(0, C)) + \text{Error}) \exp(L^2 T).$$

Finally, $\mathcal{W}_2^2(p_T, \mathcal{N}(0, C))$ can be upper bounded by

$$\mathcal{W}_2^2(p_T, \mathcal{N}(0, C)) \leq \exp(-T) \mathcal{W}_2^2(\mu_{\text{data}}, \mathcal{N}(0, C))$$

since the Ornstein-Uhlenbeck forward process is contracting with rate $\exp(-t)$ in the squared $\mathcal{W}_2^2$-distance. From this, the statement of the theorem follows. $\blacksquare$

Proposition 18 *Let U be the Cameron-Martin space associated to a measure $\mathcal{N}(0, C)$ taking values in H. Then, $\|\cdot\|_U : H \to [0, \infty]$ is lower-semicontinuous and convex on H.*

Proof Let (e_i, c_i) be the eigenvectors and eigenvalues of C. Let $f_k \to f$ in H. We will prove lower semicontinuity for $\|\cdot\|_U^2$, the result for $\|\cdot\|_U$ then follows. Then,

$$\|f\|_U^2 = \sum_{d=1}^{\infty} \lim_{k \to \infty} \langle f_k, e_i \rangle c_i^{-1} \leqslant \liminf_{k \to \infty} \sum_{d=1}^{\infty} \langle f_k, e_i \rangle c_i^{-1} = \liminf_{k \to \infty} \|f_k\|_U^2,$$

which proves lower semi-continuity. Convexity follows since $\|\cdot\|_U$ is convex when restricted to U, and infinite otherwise. $\blacksquare$

References

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

Georgios Batzolis, Jan Stanczuk, and Carola-Bibiane Schönlieb. Your diffusion model secretly knows the dimension of the data manifold. *arXiv preprint arXiv:2212.12611*, 2022.

Alexandros Beskos, Frank J Pinski, Jesús Maria Sanz-Serna, and Andrew M Stuart. Hybrid Monte Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121(10): 2201–2230, 2011.

VI Bogachev. Differentiable measures and the Malliavin calculus. *Journal of Mathematical Sciences*, 87(4):3577–3731, 1997.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=MhK5aXo3gB.

Nawaf Bou-Rabee and Andreas Eberle. Two-scale coupling for preconditioned hamiltonian monte carlo in infinite dimensions. *Stochastics and Partial Differential Equations: Analysis and Computations*, 9:207–242, 2021.

John Certaine. The solution of ordinary differential equations with large time constants. *Mathematical methods for digital computers*, 1:128–132, 1960.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022.

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. URL `https://openreview.net/forum?id=zyLVMgsZOU_`.

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013. doi: 10.1214/13-STS421.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Tiangang Cui, Kody JH Law, and Youssef M Marzouk. Dimension-independent likelihood-informed mcmc. *Journal of Computational Physics*, 304:109–137, 2016.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*. Cambridge university press, 2014.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Richard Durrett. *Probability: Theory and Examples*. Probability: Theory & Examples. Duxbury Press, 3 edition, 2005. ISBN 0534424414; 9780534424411.

H Föllmer and A Wakolbinger. Time reversal of infinite-dimensional diffusions. *Stochastic processes and their applications*, 22(1):59–77, 1986.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015. doi: 10.1017/CBO9781107337862.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stephane Mallat. Wavelet score-based generative modeling. *arXiv preprint arXiv:2208.05003*, 2022.

Paul Hagemann, Lars Ruthotto, Gabriele Steidl, and Nicole Tianjiao Yang. Multilevel diffusion: Infinite dimensional score-based diffusion models for image generation. *arXiv preprint arXiv:2303.04772*, 2023.

Martin Hairer. An introduction to stochastic PDEs. *arXiv preprint arXiv:0907.4178*, 2009.

Martin Hairer, Andrew M Stuart, and Sebastian J Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Annals of Applied Probability*, 2014.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Zahra Kadkhodaie and Eero P Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=x5hh6N9bUUb.

Ioannis Karatzas, Ioannis Karatzas, Steven Shreve, and Steven E Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. *arXiv preprint arXiv:2212.00886*, 2022.

Ki-Tae Kim, Umberto Villa, Matthew Parno, Youssef Marzouk, Omar Ghattas, and Noemi Petra. hippylib-muq: A bayesian inference software framework for integration of data with complex predictive models under uncertainty. *ACM Transactions on Mathematical Software*, 2023.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=dUSI4vFyMK.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space. *arXiv preprint arXiv:2302.07400*, 2023.

Emile Mathieu, Vincent Dutordoir, Michael J Hutchinson, Valentin De Bortoli, Yee Whye Teh, and Richard E Turner. Geometric neural diffusion processes. *arXiv preprint arXiv:2307.05431*, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Angus Phillips, Thomas Seror, Michael Hutchinson, Valentin De Bortoli, Arnaud Doucet, and Emile Mathieu. Spectral diffusion processes. *arXiv preprint arXiv:2209.14125*, 2022.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. *arXiv preprint arXiv:2211.16152*, 2022.

Jakiw Pidstrigach. Convergence of preconditioned Hamiltonian Monte Carlo on Hilbert spaces. *IMA Journal of Numerical Analysis*, page drac052, 10 2022a. ISSN 0272-4979. doi: 10.1093/imanum/drac052. URL `https://doi.org/10.1093/imanum/drac052`.

Jakiw Pidstrigach. Score-based generative models detect manifolds. *arXiv preprint arXiv:2206.01018*, 2022b.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Andrew M Stuart. Inverse problems: a Bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=VzuIzbRDrum`.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, Dordrecht, 2009. doi: 10.1007/b13794. URL `https://cds.cern.ch/record/1315296`.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Kaylee Yingxi Yang and Andre Wibisono. Convergence in KL and Rényi divergence of the unadjusted Langevin algorithm using estimated score. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. URL `https://openreview.net/forum?id=RSNMAMiPFTM`.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Loek7hfb46P`.