Operator-informed score matching for Markov diffusion models

Anonymous authors
Paper under double-blind review

Abstract

Diffusion models strive to invert a noising process via score matching, a learning objective agnostic to the underlying spatial and temporal dependencies given by an incremental noising schedule. This paper studies the spectral properties of the infinitesimal generators that govern *Markov* noising processes for better-informed score matching. Notably, we illustrate that it is possible to (i) reformulate the implicit score matching loss of Hyvärinen from an *operator-centric* viewpoint and (ii) obtain training-free parametric score estimates for all noise levels, using only sample averages with respect to the data distribution. The resulting *operator-informed score matching* provides both a standalone approach to sample generation for low-dimensional distributions, as well as a recipe for aiding the training of neural score estimators in practical high-dimensional settings.

1 Introduction

Diffusion models (DMs) (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) have emerged as powerful generative models, achieving strong performance on complex, high-dimensional data.Their conceptual appeal stems from the simple yet illuminating observation that "creating noise from data is easy" (Song et al., 2020b), as a wide range of tractable diffusion processes (often known as *forward* or *noising processes*) can be used to devolve a hard-to-sample data distribution p_{data} into an easy-to-sample noise distribution π. The forward process induces a corresponding *backward process*, whose law coincides with the time reversal of the forward process, which can be simulated provided that the score functions of its marginal distributions are known. As simulating the backward process produces samples from p_{data}, much effort has been devoted to *score matching*, a data-driven approach that infers the score function from independent samples (Hyvärinen, 2005; Vincent, 2011; Song et al., 2020a).

While Markov DMs remain the default choice for forward processes, they play a supporting role in degrading structured data into unstructured noise. In principle, a wide variety of forward processes can be reversed for generative modeling; for example, it is not required that the forward process is Markov, or even a diffusion process (despite the DM terminology) (Bansal et al., 2023; Holderrieth et al., 2025). Although the generality of the DM framework exceeds the breadth of *Markov* forward processes, we illustrate that a careful analysis of Markov diffusion forward processes confers the following new insight and methodology:

- **New insight into score matching:** The temporal evolution of the data distribution under the forward process is best understood from the perspective of *Markov diffusion operators* – an analytic approach to the study of conditional expectations. In particular, the most popular Markov forward processes have well-defined diffusion operators, which we leverage to derive an operator-based variant of the score matching loss.
- **Improved score matching:** The *spectrum* of diffusion operators characterizes the temporal evolution of conditional expectations, and for the most popular Markov forward processes the spectra can be completely characterized. Drawing parallels from *orthogonal function* methods, we propose *operator-informed score matching* (OISM); a novel, widely applicable approach to score matching that exploits an explicitly computable approximation to the score function based on the eigenfunctions of the infinitesimal generator for reversible (symmetric) diffusion Markov processes.Our approach is notable for its simplicity, as we require only estimates of the expected value of selected eigenfunctions with respect to the data distribution

to produce training-free score approximations across all noise perturbation levels. We showcase the validity of our approach with illustrative and practical examples, which demonstrate that OISM is expressive enough on its own to generate low-dimensional data, and low-order OISM accelerates the training of neural network score estimates for generating high-dimensional data.

The interplay between score matching methodology and the choice of the forward process is a relatively unexplored research direction in DMs. The recent works of Scarvelis et al. (2023); Lai et al. (2023); Benton et al. (2024); Ren et al. (2025) address this aspect, but to a lesser extent compared to the present paper, as discussed in detail in Section 6. Our work adopts the framework of *Markov diffusion operators* as set out in standard references such as Bakry et al. (2013). This formalism provides an aerial view of the corresponding forward process as an evolution of probability distributions guided by Markov semi-group operators and we contend that this perspective enables useful theoretical insights, as well as the development of improved score matching methodology, in the context of DMs.

The set-up and notation used in this paper are introduced in Section 2. Our main findings on DMs, obtained through the lens of Markov diffusion operators, are detailed in Section 3. The proposed OISM methodology is presented in Section 4 and empirically assessed in Section 5. An extended discussion is contained in Section 6.

2 Set-up and Notation

The central task of *generative modeling* is learning to produce novel samples from a data distribution p_{data}, given only a finite set $\{\mathbf{x}_m\}_{m=1}^M$ of independent samples from p_{data}. For simplicity, we presume in this paper that p_{data} is supported on $\mathbb{R}^d$, and for convenience, we overload the notation so that (e.g.,) p_{data} refers both to the distribution and its associated density on $\mathbb{R}^d$. Similarly, we let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, and $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the corresponding density function in shorthand. For the purpose of this paper, we assume all relevant densities exist and are positive and differentiable as required, and that all integrals we consider exist.

Here we briefly summarize the key elements of generative models based on DMs noised through a Markov diffusion process. Our aim is not to provide a rigorous mathematical treatment, as comprehensive introductions to the stochastic analysis tools underlying DMs can be found in Øksendal (2003); Stroock (2013); Särkkä & Solin (2019); Bakry et al. (2013). We consider the *forward process* of a DM to be a time-indexed stochastic process $(\mathbf{X}_t)_{t\geq 0}$ whose initial value $\mathbf{X}_0 \in \mathbb{R}^d$ is sampled from $\rho_0 := p_{\text{data}}$, the distribution which gives rise to the dataset, and for which the marginal distributions $(\rho_t)_{t\geq 0}$ of subsequent states $\mathbf{X}_t$ converge to an explicitly known noise distribution π as $t \to \infty$. In particular, we focus on autonomous Markov diffusion processes on $\mathbb{R}^d$, which can be represented as the solution to the Itô stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{X}_t = \mathbf{b}(\mathbf{X}_t)\,\mathrm{d}t + \mathbf{L}(\mathbf{X}_t)\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0 \sim \rho_0, \tag{1}$$

where $\mathbf{b}(\mathbf{x})$ is the drift vector, $\mathbf{L}(\mathbf{x})$ is the dispersion matrix and $\mathbf{W}_t$ is the standard d-dimensional Brownian motion (BM). In this paper, we restrict our focus on drift and dispersion coefficients that lead to *reversible* Markov diffusion processes, an assumption that is key to the development of our methodology. Prominent examples of Markov DMs include the *variance-exploding* DM described by the SDE:

$$\mathrm{d}\mathbf{X}_t = \sqrt{2}\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0 \sim \rho_0, \tag{2}$$

and the *variance-preserving* DM:

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0 \sim \rho_0, \tag{3}$$

also known as an Ornstein–Uhlenbeck (OU) process (Song et al., 2020b). In practice, the forward process is simulated until a finite time horizon T, large enough that ρ_T becomes indistinguishable from π. The *probability flow* ordinary differential equation (ODE) (Song et al., 2020b):

$$\frac{\mathrm{d}\mathbf{X}_t}{\mathrm{d}t} = -\nabla \log \frac{\rho_t}{\pi}(\mathbf{X}_t) = -\mathbf{X}_t - \nabla \log \rho_t(\mathbf{X}_t), \qquad \mathbf{X}_0 \sim \rho_0, \tag{4}$$

describes the deterministic evolution of the marginal distributions ρ_t; the final equality in (4) presents the evolution of ρ_t under OU dynamics, to illustrate the concept. The main observation in DM is that simulating (4) backwards in time (the *backward process*) with initialization $\mathbf{X}_T \sim \pi$ approximately yields $\mathbf{X}_0 \sim \rho_0$. To instantiate the backward process, the score of ρ_t, that is, $\nabla \log \rho_t(\mathbf{x})$, is required. We build upon the line of work that models and learns the marginal distributions of the diffusion process using energy-based formulations (Song & Kingma, 2021), therefore enabling direct evaluation of the score function.

Markov forward processes remain the default choice for DMs, with the BM process in (2) and the OU process in (3) most widely used. However, there are many other forward processes yielding a time-indexed sequence of distributions of decreasing complexity that can in principle be inverted (e.g. Bansal et al., 2023). This paper presents new theoretical insights that highlight particular advantages of *Markov* forward processes, relative to these alternatives, in the context of DMs. To this end, we next revisit DMs through the lens of Markov diffusion operators.

3 Diffusion Models through the Lens of Markov Diffusion Operators

This section outlines the basis for our theoretical insight into Markov DMs. Our main tools are Markov diffusion operators (Bakry et al., 2013), a set of operators that can be used to characterize the marginal distributions $(\rho_t)_{t\geq 0}$ of a Markov DM (Section 3.1). These operators will be exploited to shed new light on score matching and related objectives (Section 3.2).

3.1 Markov Diffusion Operators and their Spectral Decompositions

One of the fundamental ideas that we leverage in this paper is that the convergence of a Markov diffusion process to its stationary distribution π can be characterized using its Markov semigroup and infinitesimal generator, as explained below. The *Markov semigroup* $(P_t)_{t\geq 0}$ is a time-indexed sequence of operators, such that $P_t f(\mathbf{x})$ evaluates the conditional expected value of $f(\mathbf{X}_t)$ given that $\mathbf{X}_0$ is initialized at $\mathbf{x}$, that is,

$$P_t f(\mathbf{x}) := \mathbb{E}_{\mathbf{X}_t \sim \rho_t(\mathbf{X}_t|\mathbf{X}_0=\mathbf{x})} \left[f(\mathbf{X}_t) \right]. \tag{5}$$

From this definition, and using the tower rule, it is easy to derive the following identity for expectations under the time-marginal of a Markov process, which we will use throughout:

$$\mathbb{E}_{\mathbf{X}_t \sim \rho_t}[f(\mathbf{X}_t)] = \mathbb{E}_{\mathbf{X}_0 \sim \rho_0}[P_t f(\mathbf{X}_0)]. \tag{6}$$

Moreover, we can represent $(P_t)_{t\geq 0}$ more parsimoniously using its *infinitesimal generator* $\mathscr{L}$, which is defined as the derivative of P_t at $t = 0$:

$$\mathscr{L} f(\mathbf{x}) := \lim_{t \downarrow 0} \frac{P_t f(\mathbf{x}) - f(\mathbf{x})}{t}, \tag{7}$$

noticing that $P_0 f(\mathbf{x}) = f(\mathbf{x})$. As a result, we can formally express $P_t = e^{t\mathscr{L}}$, where the exponential of an operator is defined by the Taylor expansion of the exponential function, in analogy to the matrix exponential. The infinitesimal generator of the Markov diffusion processes (1) is:

$$\mathscr{L} f(\mathbf{x}) = \mathbf{b}(\mathbf{x})^\top \nabla f(\mathbf{x}) + \frac{1}{2} \operatorname{Tr}(\mathbf{G}(\mathbf{x}) \nabla^2 f(\mathbf{x})), \tag{8}$$

where $\mathbf{G}(\mathbf{x}) = \mathbf{L}(\mathbf{x})\mathbf{L}(\mathbf{x})^\top$ is the diffusion matrix and $\nabla^2 f$ is the Hessian of f, see Bakry et al. (2013, Section 1.11.1). In the sequel, our experiments focus on the OU and the BM forward processes (or rather, a variance-preserving BM variant, see Appendix B), ubiquitous in the practical application of DMs:

Example 1 (BM process, Markov operators)**.** *The Markov semigroup associated with the BM process in* (2) *is* $P_t f(\mathbf{x}) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{N}(\alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I})} \left[f(\mathbf{X}_t) \right]$*, where* $\alpha_t = 1$ *and* $\sigma_t^2 = 2t$ *and the infinitesimal generator is the Laplace operator* $\mathscr{L} f(\mathbf{x}) = \Delta f(\mathbf{x})$*; see Example 7.3.4 of Øksendal (2003). The invariant measure of BM, while no longer a probability measure, is nominally the Lebesgue measure* $\mathrm{d}\mathbf{x}$*, with infinite variance as* $t \to \infty$*.*

Example 2 (OU process, Markov operators)**.** *The Markov semigroup associated with the OU process in* (3) *is* $P_t f(\mathbf{x}) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{N}(\alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I})}[f(\mathbf{X}_t)]$*, where* $\alpha_t = e^{-t}$*,* $\sigma_t^2 = 1 - \alpha_t^2$*, and the infinitesimal generator is* $\mathscr{L} f(\mathbf{x}) = \langle -\mathbf{x}, \nabla f(\mathbf{x}) \rangle + \Delta f(\mathbf{x})$*; see Theorem 7.3.3 of Øksendal (2003). The invariant distribution* π *is* $\mathcal{N}(\mathbf{0}, \mathbf{I})$*.*

General expressions for the diffusion operators of Markov forward processes corresponding to the linear time-invariant (LTI) SDEs (Särkkä & Solin, 2019) are given in Appendix A.1, including the critically-dampened OU process, which can be interpreted as a DM on an extended space (Dockhorn et al., 2021). Further examples of autonomous SDEs include the Dirichlet (or Wright–Fisher) diffusion process for generating data on the simplex (Avdeyev et al., 2023; Chandra et al., 2026), also known as the Jacobi diffusion process in one dimension; we provide further details in Appendix A.2.

In summary, Markov operators characterize how conditional and marginal expectations evolve over time in a Markov process. While it is generally nontrivial to explore the spectral properties of an arbitrary diffusion process (Chewi et al., 2020), the spectra of Markov diffusion processes are well-understood (Bakry et al., 2013). The *spectrum* of a Markov diffusion processes consists of pairs of eigenvalues $\lambda_n \in \mathbb{C}$ and eigenfunctions $\phi_n : \mathbb{R}^d \to \mathbb{C}$, such that:

$$\mathscr{L} \phi_n(\mathbf{x}) = \lambda_n \phi_n(\mathbf{x}), \tag{9}$$

and will be denoted $(\lambda_n, \phi_n)_{n \in \mathcal{I}}$ for an appropriate index set $\mathcal{I}$. Because formally $P_t = e^{t\mathscr{L}}$, we have:

$$P_t \phi_n = e^{\lambda_n t} \phi_n, \tag{10}$$

and the spectrum of the semi-group operator P_t is $(e^{\lambda_n t}, \phi_n)_{n \in \mathcal{I}}$, $t \geq 0$. Therefore:

$$\mathbb{E}_{\mathbf{X}_t \sim \rho_t}[\phi_n(\mathbf{X}_t)] = \mathbb{E}_{\mathbf{X}_0 \sim \rho_0}[P_t \phi_n(\mathbf{X}_0)] = e^{\lambda_n t} \mathbb{E}_{\mathbf{X}_0 \sim \rho_0}[\phi_n(\mathbf{X}_0)], \tag{11}$$

where the first identity holds because of (6). This implies that that the expectations of the eigenfunctions ϕ_n *evolve predictably* in time under the dynamics of the forward process as governed by the exponent λ_n in (11). This observation is key to the OISM methodology proposed in Section 4: that is, the numerical approximation of quantities of the form of $\mathbb{E}_{\mathbf{X}_t \sim \rho_t}[f(\mathbf{X}_t)]$, a common occurrence in the training of DMs, can be bypassed for all f constructed as linear combinations of eigenfunctions $\{\phi_n\}_{n \in \mathcal{I}}$. This begs the questions: *how expressive are these eigenfunctions*, and *can an eigenfunction basis be harnessed to accelerate training of a Markov DM?*

Example 3 (BM process, spectrum)**.** *The eigenfunctions of the infinitesimal generator are the complex exponentials (that is, trigonometric functions)* $\Delta \exp(\imath \langle \boldsymbol{\xi}, \mathbf{x} \rangle) = -\|\boldsymbol{\xi}\|^2 \exp(\imath \langle \boldsymbol{\xi}, \mathbf{x} \rangle)$*, indexed by* $\boldsymbol{\xi} \in \mathbb{R}^d$*. Notice that in this case the spectrum is uncountable (*$\mathcal{I} = \mathbb{R}^d$*) and the eigenvalues are* $\lambda_\xi = -\|\boldsymbol{\xi}\|^2$*.*

Example 4 (OU process, spectrum)**.** *In the one-dimensional setting (*$d = 1$*), the infinitesimal generator of the OU process has a spectrum consisting of eigenpairs* $(\lambda_n, \phi_n)_{n \in \mathbb{N}}$ *with* $\lambda_n = -n$ *and* $\phi_n(x) = \mathrm{He}_n(x)/\sqrt{n!}$*, with* $\mathrm{He}_n(x)$ *the* n*th probabilist's Hermite polynomial (Abramowitz & Stegun, 1965). In the multivariate setting (*$d > 1$*) the eigenpairs become tensorized as sums* $\lambda_{n_1} + \cdots + \lambda_{n_d}$ *and products* $\phi_{n_1}(x_1) \cdots \phi_{n_d}(x_d)$ *of the one-dimensional eigenpairs, indexed by* $(n_1, \ldots, n_d) \in \mathbb{N}^d$*. In each case, the eigenfunctions are a* rich approximating set*; they form an orthogonal basis for* $L^2(\pi)$ *(Bakry et al., 2013, Section 2.7.1).*

3.2 Score Matching via Generalized Integration-by-Parts

The accurate approximation of the score function $\nabla \log \rho_t(\mathbf{x})$ is the cornerstone to the success of a DM, enabling samples to be generated by (e.g.) simulating the time reversal of the probability flow ODE (4). The standard approach is to train a flexible time-dependent vector field $\mathbf{s}_t(\mathbf{x})$ by minimizing the *score matching loss*, an L^2 distance to the true score function:

$$\mathcal{J}(\mathbf{s}_t) := \frac{1}{2} \int_{t=0}^{T} \mathbb{E}_{\mathbf{X}_t \sim \rho_t} \left[w(t) \left\| \mathbf{s}_t(\mathbf{X}_t) - \nabla \log \rho_t(\mathbf{X}_t) \right\|^2 \right] \, \mathrm{d}t, \tag{12}$$

for some non-negative weight function $w(t)$. While (12) includes intractable terms, closed-form equivalent objectives can be obtained via implicit score matching (ISM) (Hyvärinen, 2005) or denoising score matching

(DSM) (Vincent, 2011). Our first contribution in this paper is to note that the ISM loss function can be derived elegantly through the lens of Markov diffusion operators.

To streamline the following derivation, we additionally introduce a bilinear operator induced by the infinitesimal generator $\mathscr{L}$, the *carré-du-champ* operator $\Gamma(\cdot,\cdot)$, defined as:

$$\Gamma(f,g) := \frac{1}{2}\left(\mathscr{L}(fg) - f\mathscr{L}g - g\mathscr{L}f\right), \tag{13}$$

with additional notation $\Gamma(f) := \Gamma(f,f)$; see Bakry et al. (2013, Section 1.4.2), for background. In particular, the carré-du-champ operator of the Markov diffusion processes (1) is the weighted inner product:

$$\Gamma(f,g)(\mathbf{x}) = \frac{1}{2}\langle \mathbf{L}(\mathbf{x})^T\nabla f(\mathbf{x}), \mathbf{L}(\mathbf{x})^T\nabla g(\mathbf{x})\rangle, \tag{14}$$

see Bakry et al. (2013, Section 1.11.1).

Example 5 (BM and OU processes, carré-du-champ operator)**.** *For the BM and OU processes, the carré-du-champ operator is* $\Gamma(f,g)(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x})\rangle$.

In the following, we assume that Markov semigroup associated to reversible Markov diffusions, is *symmetric* on $L^2(\pi)$ with vanishing boundary term (see Appendix A.3), yielding:

$$\int \Gamma(f,g)(\mathbf{x})\,\mathrm{d}\pi(\mathbf{x}) = -\int f(\mathbf{x})\mathscr{L}g(\mathbf{x})\,\mathrm{d}\pi(\mathbf{x}) = -\int g(\mathbf{x})\mathscr{L}f(\mathbf{x})\,\mathrm{d}\pi(\mathbf{x}). \tag{15}$$

Equation (15) holds, for example, for the BM, OU and Jacobi processes, and it can be seen as a generalization of the standard integration-by-parts formula, which is recovered in the special case of the BM process (where $\mathscr{L}$ is the Laplace operator Δ, and the invariant measure is $\mathrm{d}\pi = \mathrm{d}\mathbf{x}$). To exploit symmetry of the diffusion with respect to the invariant distribution π, we assume that the density ratio between the marginal distributions $(\rho_t)_{t\geq 0}$ and π is approximated by an *energy-based model* (LeCun et al., 2006):

$$\frac{\rho_t}{\pi} \approx \frac{\exp(f_t(\mathbf{x}))}{Z_{f_t}}, \tag{16}$$

for some negative energy function $f_t : \mathbb{R}^d \to \mathbb{R}$ (relative to π) and partition function Z_{f_t}. This implies that the vector field that we learn to approximate the score is re-parametrized as:

$$\mathbf{s}_t(\mathbf{x}) = \nabla f_t(\mathbf{x}) + \nabla \log \pi(\mathbf{x}). \tag{17}$$

Then, recalling the expression of the carré-du-champ operator (14), there is an apparent semblance between the score matching loss (12) and the generalized integration-by-parts formula (15), as they both involve the inner product of gradients. We formalize this observation in the following theorem, which is our first contribution and provides an equivalent form of the implicit score matching loss that is amenable to the novel score matching methodology that we are going to present:

Theorem 1 (Semigroup form of the score matching loss)**.** *Let* $(\mathbf{X}_t)_{t\geq 0}$ *be a forward reversible Markov diffusion process* (1)*, with energy-based representation* (16)*. The score matching loss* (12) *is equivalent to:*

$$\mathcal{J}_{\mathrm{OISM}}(f_t) := \frac{1}{2}\int_{t=0}^{T} \mathbb{E}_{\mathbf{X}_0\sim\rho_0}\left[P_t\Gamma(f_t)(\mathbf{X}_0) + 2P_t\mathscr{L}f_t(\mathbf{X}_0)\right]\mathrm{d}t, \tag{18}$$

up to an additive constant.

Proof. Song et al. (2021) prove that, when the weight in (12) is the diffusion matrix of the forward process (1), optimizing the score matching objective coincides with the maximum likelihood estimate of the corresponding energy-based model. Upon the identification (17), and using the characterization of the carré-du-champ operator (14), we therefore formulate (12) as:

$$\mathcal{J}(f_t) := \frac{1}{2}\int_{t=0}^{T} \mathcal{L}(f_t)\,\mathrm{d}t,$$

where

$$\mathcal{L}(f_t) := \mathbb{E}_{\mathbf{X}_t\sim\rho_t}\left[\left\|\nabla f_t(\mathbf{X}_t) - \nabla \log \frac{\rho_t}{\pi}(\mathbf{X}_t)\right\|^2_{\mathbf{G}(\mathbf{x})}\right]$$
$$\stackrel{+C}{=} \mathbb{E}_{\mathbf{X}_t\sim\rho_t}\left[\Gamma(f_t, f_t)(\mathbf{X}_t) - 2\Gamma\left(f_t, \log\frac{\rho_t}{\pi}\right)(\mathbf{X}_t)\right], \tag{19}$$

and, $\stackrel{+C}{=}$ indicates equality up to an f_t-independent additive constant. The second term in (19) can be further rewritten leveraging the chain rule for the carré-du-champ operator:

$$\Gamma(f, \varphi(g)) = \varphi'(g)\Gamma(f, g),$$

for $\varphi \in \mathcal{C}^2$ see Bakry et al. (2013, Section 1.11.1). Choosing $\varphi(u) = \log u$ in the above identity, followed by the generalized integration by part formula (15), leads to:

$$\int \Gamma\left(f_t, \log\frac{\rho_t}{\pi}\right)(\mathbf{x})\,\mathrm{d}\rho_t(\mathbf{x}) = \int \Gamma\left(f_t, \frac{\rho_t}{\pi}\right)(\mathbf{x})\,\mathrm{d}\pi(\mathbf{x})$$
$$= -\int \mathscr{L} f_t(\mathbf{x})\frac{\rho_t}{\pi}(\mathbf{x})\,\mathrm{d}\pi(\mathbf{x}) = -\mathbb{E}_{\mathbf{X}_t\sim\rho_t}\left[\mathscr{L} f_t(\mathbf{X}_t)\right].$$

Substituting back into (19), and using the shorthand notation $\Gamma(f) = \Gamma(f, f)$, we arrive at a formula resembling the familiar ISM loss of Hyvärinen (2005):

$$\mathcal{L}(f_t) \stackrel{+C}{=} \mathbb{E}_{\mathbf{X}_t\sim\rho_t}\left[\Gamma(f_t)(\mathbf{X}_t) + 2\mathscr{L} f_t(\mathbf{X}_t)\right].$$

Finally, using property (6) of the semigroup P_t, we can rewrite the above identity as an integral with respect to the noise-free data distribution:

$$\mathcal{L}(f_t) \stackrel{+C}{=} \mathbb{E}_{\mathbf{X}_0\sim\rho_0}\left[P_t\Gamma(f_t)(\mathbf{X}_0) + 2P_t\mathscr{L} f_t(\mathbf{X}_0)\right],$$

and obtain the stated result. □

Theorem 1 allows for a *streamlined, Markov-aware* and *operator-informed* derivation of score matching, by providing two key observations for the methodology we develop in Section 4: (i) the optimization of noised-perturbed scores (∇f_t in our re-parametrization) links to expectations of quantities with respect to the data distribution ρ_0 via the operator P_t; (ii) based on the definition (5) of P_t , sampling from the forward process is a general strategy to evaluate terms of the form $P_t\phi$, for those processes for which P_t cannot be computed in closed form. However, when ϕ is a linear combination of the eigenfunctions $\{\phi_n\}_{n\in\mathcal{I}}$, exact computation of $P_t\phi$ is possible, in principle enabling sampling from the forward process to be avoided, as explained in the following.

4 Operator-Informed Score Matching

Building on the theoretical insights from the previous section, we present *operator-informed score matching* (OISM), a recipe for score matching that makes use of the spectral decomposition of the Markov diffusion operators. At a high level, the concept of OISM is to approximate in the linear space spanned by eigenfunctions of the forward process, and leverage the spectral property (11) to compute both terms in (18) in closed form, obtaining a quadratic form corresponding to the score matching loss.

4.1 Score Matching as a Quadratic Form

To formulate OISM, we make one further assumption on the spectral properties of the infinitesimal generator $\mathscr{L}$: the eigenfunctions $\{\phi_n\}_{n\in\mathcal{I}}$ form a countable Hilbertian basis in $L^2(\pi)$ (Bakry et al., 2013). In fact, the eigenvalues represent a measure of importance of the corresponding eigenvectors; when non-positive and real-valued, larger (closer to 0) eigenvalues indicate a slower decay of $\mathbb{E}_{\mathbf{X}_t\sim\rho_t}[\phi_n(\mathbf{X}_t)]$, as t increases, following (11). If the spectrum is uncountable, it is not possible to enumerate the "top n most important

eigenfunctions", which is an essential task for accurately defining the energy-based model (16). The spectrum of the OU forward process is countable, while the spectrum of the BM forward process is uncountable, but becomes countable if its domain is modified from $\mathbb{R}^d$ to a finite interval; see Appendix B for a detailed explanation. Then:

Theorem 2 (Explicit form of score matching estimator)**.** *Let $[n] = \{1, \dots, n\}$, for some $n \in \mathbb{N}$, and let $\{\phi_k\}_{k\in[n]}$ be n eigenfunctions of the infinitesimal generator of* (1)*. Consider an energy-based model for the density ratio* (16)*, with $f_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{k\in[n]} \alpha_k \phi_k(\mathbf{x})$, parametrized by the time-varying coefficients $\boldsymbol{\alpha} \in \mathbb{R}^n$. Suppose that we can cast $\phi_k\phi_\ell = \sum_{h\in[m]} \beta_h^{(k,\ell)} \phi_h$ as a linear combination with appropriate coefficients $\beta_h^{(k,\ell)}$ for some $m \in \mathbb{N}$. Then the integrand of the score matching loss* (18) *takes the quadratic form:*

$$\mathcal{L}\left(f_{\boldsymbol{\alpha}}\right) \overset{+C}{=} \boldsymbol{\alpha}^\top \mathbf{A}_t \boldsymbol{\alpha} + 2\mathbf{b}_t^\top \boldsymbol{\alpha}, \tag{20}$$

where

$$\mathbf{A}_t^{(k,\ell)} = \sum_{h\in[m]} \left(\tfrac{\lambda_h - \lambda_k - \lambda_\ell}{2}\right) e^{\lambda_h t} \beta_h^{(k,\ell)} \mathbb{E}_{\rho_0}\left[\phi_h\right], \qquad \mathbf{b}_t^{(k)} := \lambda_k e^{\lambda_k t} \mathbb{E}_{\rho_0}[\phi_k],$$

which is explicitly minimised at $\widehat{\boldsymbol{\alpha}} := \mathbf{A}_t^{-1}\mathbf{b}_t$.

Proof. Using the assumption that the negative energy $f_{\boldsymbol{\alpha}}$ lies in the eigenbasis span, and the eigendecompositions (9) and (10), the terms in the integrand of (18) can be written as:

$$\mathbb{E}_{\rho_0}\left[P_t \Gamma\left(f_{\boldsymbol{\alpha}}\right)\right] = \sum_{k,\ell\in[n]} \alpha_k \alpha_\ell \mathbb{E}_{\rho_0}\left[P_t \Gamma(\phi_k, \phi_\ell)\right] = \boldsymbol{\alpha}^\top \mathbf{A}_t \boldsymbol{\alpha}, \qquad \mathbf{A}_t^{(k,\ell)} := \mathbb{E}_{\rho_0}\left[P_t \Gamma(\phi_k, \phi_\ell)\right];$$

$$\mathbb{E}_{\rho_0}\left[P_t \mathscr{L} f_{\boldsymbol{\alpha}}\right] = \sum_{k\in[n]} \alpha_k \lambda_k e^{\lambda_k t} \mathbb{E}_{\rho_0}[\phi_k] = \mathbf{b}_t^\top \boldsymbol{\alpha}, \qquad \mathbf{b}_t^{(k)} := \lambda_k e^{\lambda_k t} \mathbb{E}_{\rho_0}[\phi_k].$$

Since we assume that we can cast $\phi_k\phi_\ell = \sum_{h\in[m]} \beta_h^{(k,\ell)} \phi_h$ as a linear combination with appropriate coefficients $\beta_h^{(k,\ell)}$, we can use the definition of the carré-du-champ operator (13) and the eigendecompositions (9) and (10) to obtain the explicit formula:

$$\begin{aligned}
\mathbf{A}_t^{(k,\ell)} &= \mathbb{E}_{\rho_0}\left[P_t\, \Gamma(\phi_k, \phi_\ell)\right] = \frac{1}{2}\mathbb{E}_{\rho_0}\left[P_t\left(\mathscr{L}(\phi_k\phi_\ell) - \phi_k \mathscr{L}\phi_\ell - \phi_\ell \mathscr{L}\phi_k\right)\right] \\
&= \frac{1}{2}\mathbb{E}_{\rho_0}\left[P_t\left(\mathscr{L}(\phi_k\phi_\ell) - (\lambda_k + \lambda_\ell)\phi_k\phi_\ell\right)\right] \\
&= \sum_{h\in[m]} \frac{1}{2}\mathbb{E}_{\rho_0}\left[\beta_h^{(k,\ell)}\, P_t \mathscr{L}\phi_h - (\lambda_k + \lambda_\ell)\beta_h^{(k,\ell)} P_t \phi_h\right] \\
&= \sum_{h\in[m]} \left(\tfrac{\lambda_h - \lambda_k - \lambda_\ell}{2}\right) e^{\lambda_h t} \beta_h^{(k,\ell)} \mathbb{E}_{\rho_0}\left[\phi_h\right].
\end{aligned}$$

Consequently, the integrand of (18) is the quadratic form:

$$\mathcal{L}\left(f_{\boldsymbol{\alpha}}\right) \overset{+C}{=} \boldsymbol{\alpha}^\top \mathbf{A}_t \boldsymbol{\alpha} + 2\mathbf{b}_t^\top \boldsymbol{\alpha},$$

which is explicitly minimised at $\widehat{\boldsymbol{\alpha}} := \mathbf{A}_t^{-1}\mathbf{b}_t$, as claimed. □

The assumption that we can cast products of eigenfunctions as a linear combination with appropriate coefficients is satisfied for the OU and Jacobi processes, whose eigenfunctions are polynomials, and the BM process with its associated trigonometric eigenfunctions; see Appendices B.2 and C.1 for details.

From Theorem 2, and re-parametrization (17) a score matching estimator

$$\hat{\mathbf{s}}_t(\mathbf{x}) = \sum_{k\in[n]} \hat{\alpha}_k \nabla\phi_k(\mathbf{x}) + \nabla \log \pi(\mathbf{x}) \tag{21}$$

can therefore be learned *without requiring the forward process to be simulated*, by approximating the expectations $\mathbb{E}_{\rho_0}[\phi_h]$ using averages taken over the dataset and optimizing (20) to obtain the time-varying coefficients $\hat{\boldsymbol{\alpha}}$.

Example 6 (OU process, second order OISM)**.** *To build intuition for how the score function is approximated in OISM, recall from Example 4 that polynomials are eigenfunctions of the OU infinitesimal generator $\mathscr{L}$. For instance, one can show that, with arbitrary vector $\mathbf{v}$ and matrix $\mathbf{U}$, $\mathbf{v}^\top\mathbf{x}$ has eigenvalue 1 and $\mathbf{x}^\top\mathbf{U}\mathbf{x} - tr(\mathbf{U})$ has eigenvalue 2. Construction of a score matching estimator* (21) *based on the polynomial eigenfunctions of first and second orders, therefore, only requires computing the empirical mean $\widehat{\boldsymbol{\mu}}$ and covariance $\widehat{\boldsymbol{\Sigma}}$ of the dataset. In fact, the corresponding energy-based model coincides with a Gaussian approximation of the dataset, given that $\nabla \log \pi$ is also quadratic (Wang & Vastola, 2024).*

Regularised Estimation of Eigenfunctions via Shrinkage Our discussion so far has shown how to convert expectations $\mathbb{E}_{\rho_0}[\phi_k]$ into an explicit approximation of the score function $\hat{\mathbf{s}}_t(\mathbf{x})$, yet these expectations with respect to the data-generating distribution must still be estimated. The famous result of Stein (1956) showed that the sample mean is an inadmissible estimator for estimating three or more expectations from a dataset. Additionally, the tensorized structure of the spectrum in high dimensions (c.f. Example 3 and Example 4) involves estimating a potentially large number of expectations in OISM, motivating the consideration of *shrinkage* estimators; see Appendices C.2 and C.3 for a detailed explanation, and the following examples.

4.2 Illustrative Experiments

To illustrate the efficacy of OISM with low-dimensional distributions, we consider the truncated BM with a countable set of trigonometric eigenfunctions (see Appendix B.2). Our first example considers the 1-dimensional distribution for ρ_0 showcased in the top left panel of Figure 1, called the "Bart Simpson" distribution (Wasserman, 2005, Example 6.1). This distribution is somewhat non-trivial yet it allows for scores, densities and eigenfunction expectations to be computed. As a dataset we draw 2,000 samples, and consider the 50 eigenfunctions $\sqrt{2}\cos(kx)$ and $\sqrt{2}\sin(kx)$ for $k = 1, 2, \ldots, 25$. Specifically, we opt for the *modulation estimator* (Beran & Dumbgen, 1998), a multivariate extension of the James-Stein estimator that shrinks each sample mean, see Appendix C.3.

Results are shown in Figure 1. Most notably, OISM exhibits clear capacity to approximate the score function $\nabla \log \rho_t(\mathbf{x})$ in high-density regions of $\rho_t(\mathbf{x})$. As the area of high density region grows with t, the OISM score estimate smoothens and becomes more accurate in a wider domain. Furthermore, we witness a substantial reduction in the score matching loss when shrinkage is applied, thus enforcing its practical value in obtaining OISM.

A second example illustrating OISM in 2D is provided in Figure 2, with additional details given in Appendix D. For this setup, we use a dataset of 10,000 samples and enumerate all trigonometric eigenfunctions with eigenvalues satisfying $|\lambda_n| \leq 125$. This yields a score estimate in (21) consisting of $n = 200$ total eigenfunctions. For this low-dimensional target, the 'pure' OISM methodology combined with shrinkage is sufficient to learn an accurate generative model. However, this is not the case in high dimensions, where shrinkage estimation becomes significantly more complicated; there, we instead advocate for OISM as a powerful complement to neural sore estimators, as illustrated below.

4.3 Practical OISM in High Dimensions

OISM can be a powerful alternative to standard DM in the low-dimensional setting, as showcased in Section 4.2. However, a truncated eigenbasis cannot fully replace flexible learning of the score function when the data are high-dimensional, since these features (i.e. the eigenfunctions) are not adapted to the dataset. Our proposal to extend OISM to high-dimensional settings is quite simple: we construct a first approximation $\hat{\mathbf{s}}_t(\mathbf{x})$ to the score function as just described, and learn the *residual score*

$$\mathbf{r}_t(\mathbf{x}) := \nabla \log \rho_t(\mathbf{x}) - \hat{\mathbf{s}}_t(\mathbf{x}), \tag{22}$$

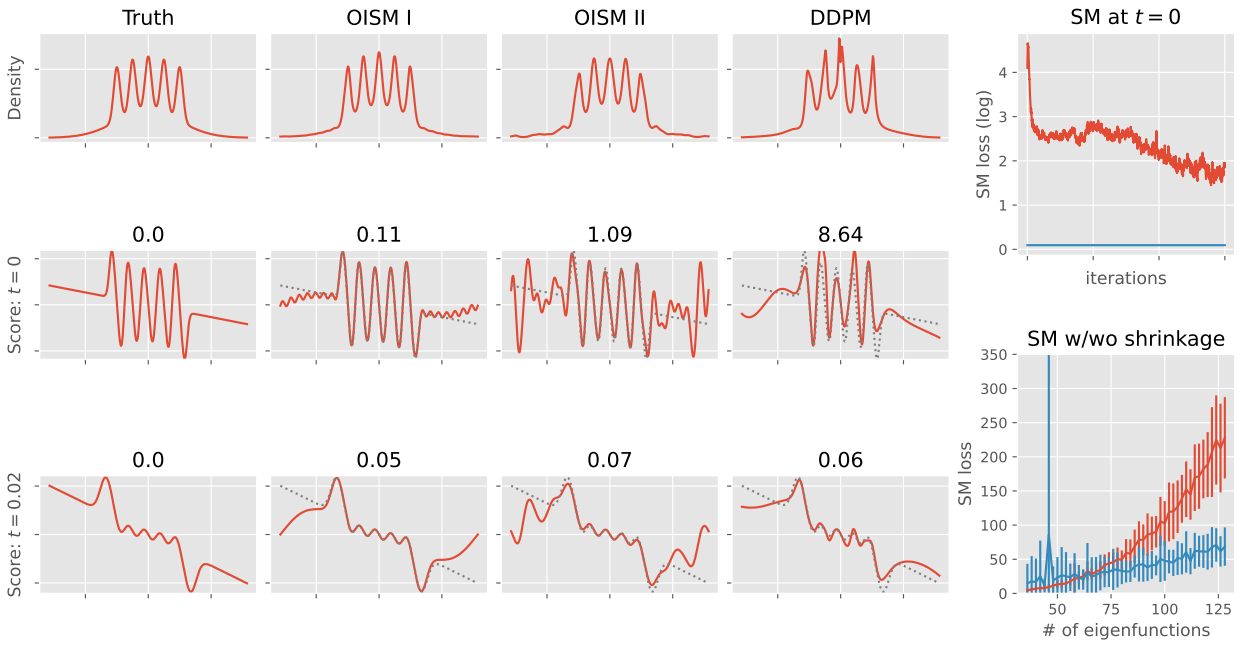

Figure 1: Illustration of OISM in 1D. The first column contains the ground truth density and score functions $\nabla \log \rho_t$ at $t = 0$ and $t = 0.02$. **OISM I** is based on exact $\mathbb{E}_{\rho_0}[\phi_k]$, highlighting the capacity for an eigenfuntion basis to provide high-quality score estimates across different levels of noise perturbation, which leads to accurate density estimates via the probability flow ODE. **OISM II** instead uses data-based shrinkage estimates for $\mathbb{E}_{\rho_0}[\phi_k]$; the density remains accurately estimated. As a baseline, the fourth column showcases the score and density estimates obtained via **DDPM**, a neural network trained via denoising score matching (SM), when training samples are produced using the BM as nosing process. [In the second and third rows the score matching loss in displayed in the subtitles.] The final column (top) shows the evolution of the SM loss for **DDPM** (red) and **OISM II** (blue) along training iterations; and (bottom) the SM loss for **OISM II** as a function of the number of eigenfunctions, corresponding to sample means (red) and shrinkage estimators (blue). [Error bars denote the standard error over 50 simulations.] The Stein effect is clearly visible, as we observe little difference between sample mean and shrinkage estimators when the dimensionality of eigenfunction is low, however, the performance gap becomes clearly visible as the number of eigenfunction increases.

using a neural network trained via denoising score matching (Ho et al., 2020). This set-up is appealing; it represents a minimal modification of existing code, which can be achieved by augmenting the output of the score function approximator in the DM; see also Appendix E.

An important issue is *which eigenfunctions to include* in the first approximation $\hat{\mathbf{s}}_t(\mathbf{x})$. This question is rather difficult, and we propose a simple heuristic in this work while leaving the idea of data-adaptive eigenfunction selection for future work. We note that if the data distribution ρ_0 factorizes, one only needs to estimate expectations of the *univariate* eigenfunctions of the form $\mathbb{E}_{\mathbf{x}\sim\rho_0}[\phi_k(x_i)]$, $i \in \{1, \dots, d\}$, $k \in [n]$, where x_i denotes the ith component of $\mathbf{x}$; c.f. Example 4. Our practical approach is to use the first n univariate eigenfunctions in each data dimension *without* shrinkage, denoting this **OISM** (n); by ignoring interaction terms, the need for shrinkage may be reduced, and empirical support for this heuristic is provided in Section 5. We emphasize that this independence assumption is merely a first-order approximation designed to guide the initial score estimate; we do not assume the true data distribution is independent across dimensions. While OISM handles the heavy-lifting of the structural score approximation, the neural network remains responsible for learning the fine, high-frequency details and cross-dimensional dependencies inherent in ρ_0.

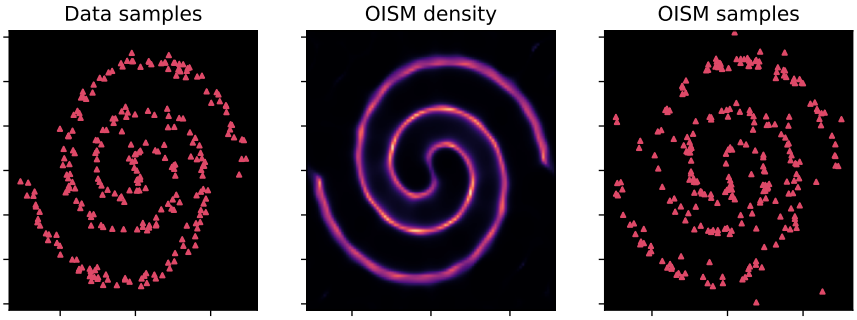

Figure 2: Illustration of OISM in 2D. The leftmost plot showcases a subset of training data. OISM was applied to estimate the score function, using 200 eigenfunctions in total. The density recovered by inverting the probability flow ODE (4) is quite accurate, which we attribute to the periodicity of the eigenfunctions of the truncated BM forward process.

5 Experimental Assessment

This section presents an empirical assessment of OISM on a prototypical high-dimensional generative modelling task. For this assessment we use the CIFAR-10 dataset (Krizhevsky et al., 2009, Chapter 3). Similar results are observed for ImageNet32 (Chrabaszcz et al., 2017), which we reserve for Appendix E.

Implementation detail Our implementation makes use of the OU (variance-preserving) process (3) as the noising process for high-dimensional simulation, following the weighted noising schedule as established by Song et al. (2021), and the neural network architecture DDPM++ proposed by Song et al. (2020b). The eigenfunctions of the OU process are Hermite polynomials, and throughout our simulations we considered a univariate polynomials of orders 3, 6, and 9. We opt for the simple solution of sample mean with no consideration of shrinkage, as we did not observe an improvement when shrinkage was applied in our testing (not shown). This finding indeed seems consistent with the lack of Stein's effect for small n, as shown in Figure 1. The rest of our training procedure is identical to the description in Song et al. (2020b). Code to reproduce these experiments is available in the electronic supplement accompanying the paper. Computation was performed on A100 GPUs and training of each model using a total of 500,000 iterations required approximately 30 hours in total. There was no increased cost at training time to using OISM as all linear systems involving $\mathbf{A}_t$ and $\mathbf{b}_t$ were pre-solved.

Results Full results are shown for **OISM** (3) and **OISM** (6) in Figure 3. In both cases, OISM significantly accelerates training, providing an informative initial guess of the score function, and, by extension, a rough estimate of the data distribution. Comparing the the score matching loss, log-likelihood and generated samples displayed in Figure 3, we see that the initial several thousand iterations can effectively be skipped over entirely with the help of OISM. On the other hand, higher order does not guarantee better performance, with **OISM** (6) performing slightly worse than **OISM** (3). Other standard performance metrics are reported in Appendix E, which confirm the substantial improvement of OISM over DDPM during the initial training period, while all methods perform comparably well after an extended training period.

Interpretation Even a low-order approximation, such as **OISM** (3), has the potential to substantially accelerate training the generative DM. However, and somewhat counter-intuitively, a benefit from using higher-order approximations was not observed. This counterintuitive result might stem from a lack of understanding of the interplay between our novel score estimates and the implicit regularisation provided by the neural network. One can speculate that **DDPM** outperforms OISM in low-density regions thanks to its inductive bias, as observed in Figure 1. Indeed, as the order of polynomial increases in OISM, the neural network might be forced to fit a residual that has greater complexity and curvature in the tail, for which the usual linear extrapolation of the neural network may be ill-suited.

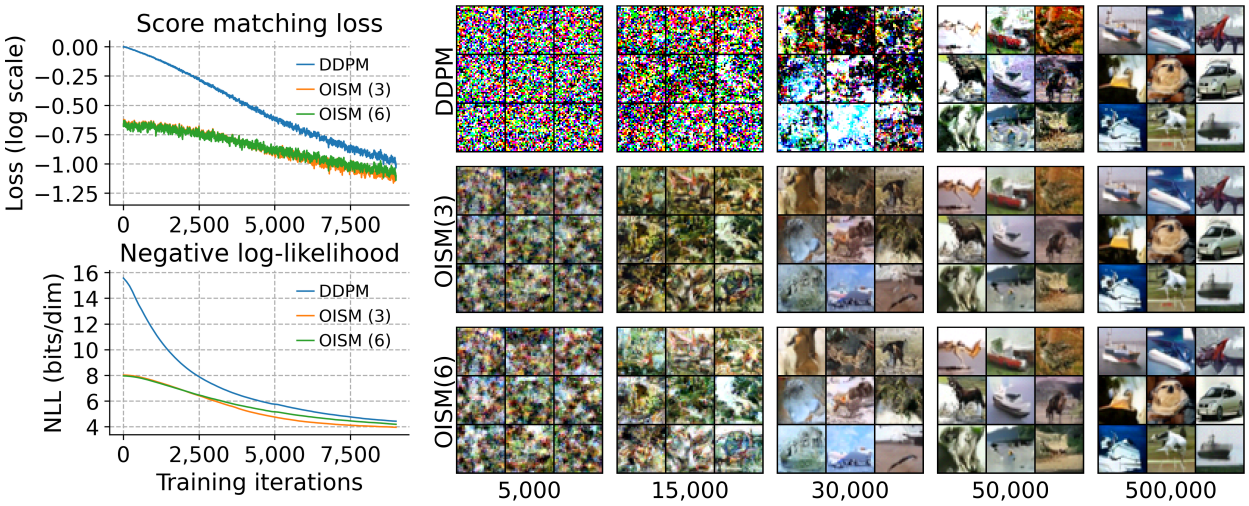

Figure 3: Performance evaluation on CIFAR-10. Top left: Smoothed score matching (SM) loss for **DDPM** (the maximum likelihood estimator of Song et al., 2021), **OISM** (3), and **OISM** (6). Bottom left: Negative log-likelihood evaluated on the test dataset. Right: Samples corresponding to different training iterations, obtained by reversing the probability flow ODE (4).

6 Discussion

This paper contributes to an improved theoretical understanding of DMs, demonstrating the potential of OISM as a practical tool to accelerate training of Markov DMs. In this section we discuss related work, and the limitations and potential of the methods we have proposed.

Related work The proliferation of diffusion models has prompted the use of a continuous-time noise injection for distributional approximation, and many efforts in this area can be viewed through the lens of Markov diffusion operators, e.g., *time score matching* (Choi et al., 2022), a generalization of the *score Fokker–Planck equation* (Lai et al., 2023), generalized score matching on non-Euclidean spaces proposed by Benton et al. (2024), and the use of target score identity for variance reduction (Duston & Bui-Thanh, 2026), another direct result from Markov semigroups. Since these observations are not critical to our methodology, they are reserved for the interested reader in Appendix F. Notably, Scarvelis et al. (2023) considered a simple forward process taking the form of a weighted mixture of the data distribution p_{data} and the noise distribution π, so that both the forward process and the associated scores can be exactly computed. This approach circumvents neural approximation of score functions, but memorizes the finite dataset when applied naively. Our regularized feature-based approach obviates the need for iterating the dataset and alleviates the issue of memorization. The loss (18) relates to the loss proposed by Benton et al. (2024), where Markov operators are used as a means to generalise Markov DMs to Riemannian manifolds, but the authors of that work did not consider (18) as a tool to derive explicit approximation to the score function, as instead proposed in this work. Outside the scope of generative modeling, our work closely relates to the use of orthogonal series in density estimation. Whittle (1958); Cencov (1962); Schwartz (1967) observed a density estimate could be obtained provided that $\mathbb{E}_{\rho_0}[\phi_k]$ were known, and its error can be improved with the introduction of shrinkage estimators; see Efromovich (2010).

Limitations and potential This work presented a novel viewpoint of encoding diffusion operators into score-based generative modeling. While our empirical assessment was intentionally restricted to 'vanilla' DMs to isolate and emphasize the foundational aspects of the framework, extending OISM to more sophisticated architectures on challenging tasks remains an important avenue for future research. One compelling direction is exploring whether further performance gains can be achieved via higher-order approximations. Although straightforward to implement, increasing the order introduces scaling complexities that require both the

choice of eigenfunctions better adapted to the data, and effective shrinkage strategies to penalize higher-order eigenfunction coefficients. However, pursuing such complex high-order extensions may not even be necessary if OISM is instead applied to latent representation diffusion models (Vahdat et al., 2021; Rombach et al., 2022). Because latent spaces inherently prioritize simplicity and independence among individual dimensions, they create an ideal environment for low-order approximations to thrive. Ultimately, our contribution supports the view that Markov diffusion operators yield valuable insight into DMs, holding the potential both to enhance existing methods and to inspire new, latent-focused methodological developments.

References

Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Courier Corporation, 1965.

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, pp. 1276–1301. PMLR, 2023.

J Bakosi and JR Ristorcelli. A stochastic diffusion process for the Dirichlet distribution. *International Journal of Stochastic Analysis*, 2013(1):842981, 2013.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, New York, 2013. ISBN 978-3-319-00226-2.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. From denoising diffusions to denoising Markov models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):286–301, 2024.

Rudolf Beran and Lutz Dumbgen. Modulation of Estimators and Confidence Sets. *The Annals of Statistics*, 26(5):1826–1856, 1998.

Diana Cai, Chirag Modi, Charles Margossian, Robert Gower, David Blei, and Lawrence Saul. Eigenvi: score-based variational inference with orthogonal function expansions. *Advances in Neural Information Processing Systems*, 37:132691–132721, 2024.

Nikolai N Cencov. Estimation of an unknown distribution density from observations. *Soviet Math.*, 3: 1559–1566, 1962.

Nuria Alina Chandra, Yucen Lily Li, Alan Nawzad Amin, Alex Ali, Joshua Rollins, Sebastian W. Ober, Aniruddh Raghu, and Andrew Gordon Wilson. A Unification of Discrete, Gaussian, and Simplicial Diffusion. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2098–2109, 2020.

Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density Ratio Estimation via Infinitesimal Classification. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 2552–2573. PMLR, 2022.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. *CoRR*, 2017.

Richard Courant and David Hilbert. Methods of mathematical physics, vol. i. *Physics Today*, 7(5):17, 1954.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Alois Duston and Tan Bui-Thanh. Variance-Reduced Diffusion Sampling via Target Score Identity. *arXiv:2601.01594*, 2026.

Sam Efromovich. Orthogonal series density estimation. *WIREs Computational Statistics*, 2(4):467–476, 2010. ISSN 1939-0068. doi: 10.1002/wics.97.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In *International Conference on Learning Representations*, September 2018.

RC Griffiths. A transition density expansion for a multi-allele diffusion model. *Advances in Applied Probability*, 11(2):310–325, 1979.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceesings of the 34th Conference on Neural Information Processing Systems*, pp. 6840–6851, 2020.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky TQ Chen, and Yaron Lipman. Generator Matching: Generative modeling with arbitrary Markov processes. In *International Conference on Learning Representations (ICLR)*, Singapore, 2025.

Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 361–379. University of California Press, 1961.

Samuel Karlin and Howard E Taylor. *A second course in stochastic processes*. Elsevier, 1981.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. FP-Diffusion: Improving score-based diffusion models by enforcing the underlying score Fokker–Planck equation. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin Heidelberg New York Dordrecht London, 6th edition, 2003. ISBN 978-3-540-04758-2.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, NY, 2014. ISBN 978-1-4939-1322-0 978-1-4939-1323-7. doi: 10.1007/978-1-4939-1323-7.

Yinuo Ren, Grant M Rotskoff, and Lexing Ying. A Unified Approach to Analysis and Design of Denoising Markov Models. *arXiv preprint arXiv:2504.01938*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019. ISBN 978-1-316-64946-6.

Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-Form Diffusion Models. *arXiv:2310.12395*, 2023.

Stuart C. Schwartz. Estimation of Probability Density by an Orthogonal Series. *The Annals of Mathematical Statistics*, 38(4):1261–1265, August 1967.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.

Yang Song and Diederik P. Kingma. How to Train Your Energy-Based Models, February 2021.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced Score Matching: A Scalable Approach to Density and Score Estimation. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pp. 574–584. PMLR, August 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *Proceedings of the 8th International Conference on Learning Representations*, 2020b.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428. Curran Associates, Inc., 2021.

Charles Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 197–207. University of California Press, 1956.

Daniel W Stroock. *An Introduction to Markov Processes*. Springer Science & Business Media, 2013.

JD Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International conference for learning representations*, volume 6, 2018.

Ch. Tsitouras. Runge–Kutta pairs of order 5 (4) satisfying only the first column simplifying assumption. *Computers & Mathematics with Applications*, 62(2):770–775, 2011.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021.

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

Binxu Wang and John Vastola. The unreasonable effectiveness of Gaussian score approximation for diffusion models and its applications. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, New York, 2005. ISBN 978-0-387-25145-5.

P. Whittle. On the Smoothing of Probability Density Functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):334–343, 1958.

Appendices

These appendices supplement the paper *Operator-informed score matching for Markov diffusion models.*

Appendix A details the properties of a broad class of forward diffusion processes, and it elaborates on the symmetry assumption required for OISM to be valid. Appendix B takes a closer look at the peculiarities involving BM, and proposes a truncated variant to address them. Appendix C gives additional details about OISM, by provides pragmatic implementation details of the corresponding score, revisiting the use of orthogonal functions in nonparametric density estimation, and laying out the details about the specific shrinkage estimator. Full details required to reproduce our two-dimensional experiment are contained in Appendix D, and full details to reproduce our image generation experiments are contained in Appendix E. Finally, Appendix F explores further consequences of Markov diffusion operators in the context of related work, including time score matching, the score Fokker–Planck equation and generalized score matching.

A Markov Diffusion Processes

Appendix A.1 details the general class of linear time-invariant (LTI) SDEs processes encompassed by equation (1); Appendix A.2 gives examples of autonomous SDEs corresponding to the Jacobi and Dirichlet diffusion processes; finally Appendix A.3 gives the definition of a symmetric Markov diffusion, and Appendix A.4 discusses operators of asymmetric processes.

A.1 Markov LTIs SDEs

This appendix presents Markov diffusion operators for a wider class of LTI SDEs, generalizing beyond the OU and BM processes featured in the main text; Särkkä & Solin (2019, Chapter 6) contains helpful explanations on this topic. LTI SDEs take the form of the following Itô diffusion:

$$\mathrm{d}\mathbf{X}_t = \mathbf{F}\mathbf{X}_t + \mathbf{L}\,\mathrm{d}\mathbf{W}_t, \tag{23}$$

for which existence and uniqueness of the solution is guaranteed for any choice of the drift matrix $\mathbf{F}$ and of the dispersion matrix $\mathbf{L}$, while its stability, and existence of and convergence to a stationary distribution π depend on these matrices.

Stationary measure and Markov transition kernel The existence of a finite stationary measure π for (23) is determined by the eigenvalues of $\mathbf{F}$. If the real parts of the eigenvalues of $\mathbf{F}$ are strictly negative, then (23) evolves towards π a zero-mean Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\infty)$, where the covariance matrix $\mathbf{\Sigma}_\infty$ satisfies the Lyapunov equation (Särkkä & Solin, 2019, equation 6.69):

$$\mathbf{F}\mathbf{\Sigma}_\infty + \mathbf{\Sigma}_\infty\mathbf{F}^\top + \mathbf{G} = \mathbf{0}.$$

where $\mathbf{G} = \mathbf{L}\mathbf{L}^\top$ is the diffusion matrix. Furthermore, the conditional distributions of (23) have the following closed form solution:

$$\rho_t\left(\tilde{\mathbf{x}}|\mathbf{x}\right) = \mathcal{N}\left(\tilde{\mathbf{x}}|\mathbf{S}_t\mathbf{x}, \mathbf{\Sigma}_t\right), \quad \mathbf{S}_t := \exp(t\mathbf{F}), \quad \mathbf{\Sigma}_t := \mathbf{\Sigma}_\infty - \mathbf{S}_t\mathbf{\Sigma}_\infty\mathbf{S}_t^\top,$$

that is, the Markov transition kernel of LTI SDEs, remains conditionally Gaussian.

Example 7 (Common LTI SDEs)**.** *Formula* (23) *for LTI SDEs covers most diffusion forward processes, for example BM is obtained with* $\mathbf{F} = \mathbf{0}, \mathbf{L} = \sqrt{2}\mathbf{I}$*, OU with* $\mathbf{F} = -\mathbf{I}, \mathbf{L} = \sqrt{2}\mathbf{I}$*, and the critically-dampened OU process with:*

$$\mathbf{F} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & -c\mathbf{I} \end{pmatrix}, \qquad \mathbf{L} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt{2c}\mathbf{I} \end{pmatrix},$$

corresponding to recent efforts of dampening the forward processes with momentum variables, see Dockhorn et al. (2021). When (23) *corresponds to a generalized version of the OU process, that is,* $\mathbf{F} = -\theta\mathbf{I}$*,* $\mathbf{L} = \sigma\mathbf{I}$*, for some* $\theta, \sigma > 0$*, the stationary distribution* π *becomes* $\mathcal{N}(\mathbf{0}, \frac{\sigma^2}{2\theta}\mathbf{I})$*.*

Infinitesimal generator and spectrum The infinitesimal generator takes the following form:

$$\mathscr{L} f(\mathbf{x}) = \langle \mathbf{F}\mathbf{x}, \nabla f(\mathbf{x}) \rangle + \frac{1}{2} \langle \nabla, \mathbf{G} \nabla f(\mathbf{x}) \rangle. \tag{24}$$

A thorough analysis of the spectral decomposition of general diffusion process, including LTI SDE, can be complicated (Bakry et al., 2013). In general, a diffusion process possesses a purely discrete spectrum if its invariant measure π decays fast enough at infinity or if the domain is strictly compact. If the domain is unbounded or the stationary distribution has heavy, non-integrable polynomial tails, the generator's spectrum shifts from discrete eigenvalues to a continuous spectral band.

A.2 Additional Examples of Autonomous SDEs

Here we provide the example of the Dirichlet and Jacobi diffusion processes, Markov diffusion processes governed by the autonomous SDEs (1), where the drift and diffusion parameters are be state-dependent. The OISM methodology applies also to these cases. The Dirichlet diffusion process is the stochastic process supported on the simplex $\mathcal{S}^{d-1} = \{\mathbf{x} \in [0,1]^d : \sum_{i=1}^d x_i = 1\}$, with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) > 0$, defined by:

$$\mathrm{d}\mathbf{X}_t{}^{(i)} = \left(\alpha_i - \alpha_0 \mathbf{X}_t^{(i)}\right) \mathrm{d}t + \sum_{j=1}^d \sqrt{2\mathbf{X}_t^{(i)}(\delta_{ij} - \mathbf{X}_t^{(j)})}\, \mathrm{d}\mathbf{W}_t{}^{(j)}, \qquad i = 1, \ldots, d,$$

where $\mathbf{X}_t^{(i)}$ denotes the i-th entry of the vector $\mathbf{X}_t$, the starting point $\mathbf{X}_0$ is sampled in the simplex, $\alpha_0 = \sum_{i=1}^d \alpha_i$ and δ_{ij} is the Kronecker delta. The associated Markov semigroup cannot be represented in closed form, but the infinitesimal generator has form (8), with drift vector $\mathbf{b}(\mathbf{x}) = \boldsymbol{\alpha} - \alpha_0 \mathbf{x}$ and diffusion matrix $\mathbf{G}(\mathbf{x}) = 2(\mathrm{diag}(\mathbf{x}) - \mathbf{x}\mathbf{x}^T)$; see Karlin & Taylor (1981); Bakosi & Ristorcelli (2013) for early derivations and Avdeyev et al. (2023); Chandra et al. (2026) for modern applications. When $d = 2$, the system collapses to a single degree of freedom on the unit interval $[0, 1]$, and the Dirichlet diffusion process reduces to the so-called Jacobi diffusion process.

The eigenfunctions of the Jacobi process are the (shifted) Jacobi polynomials $\phi_n(x) = P_n^{(\alpha_2-1,\alpha_1-1)}(2x - 1)$, $n \in \mathbb{N}$ orthogonal in $L^2(\pi)$, where π is the Beta(α_1, α_2) distribution. The corresponding eigenvalues are $\lambda_n = -n(n + \alpha_0 - 1)$, where $\alpha_0 = \alpha_1 + \alpha_2$; see Karlin & Taylor (1981, Chapter 15). When $d > 2$, the eigenfunctions of the Dirichlet diffusion process generalize to the multivariate Jacobi (Griffiths) polynomials, an orthogonal basis in $L^2(\pi)$, where π is the Dirichlet distribution with parameter $\boldsymbol{\alpha}$ (Griffiths, 1979).

A.3 Symmetric Markov Diffusions

While the generator (8) applies to all autonomous SDEs (1), our OISM methodology is most applicable when the associated Markov semigroup is *symmetric*. Following Bakry et al. (2013, Definition 1.6.1) a Markov semigroup $(P_t)_{t\geq 0}$ is *symmetric* with respect to the invariant measure π, if for all $f, g \in L^2(\pi)$ and $t \geq 0$:

$$\int f(\mathbf{x}) P_t g(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) = \int g(\mathbf{x}) P_t f(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}).$$

Example 8 (Symmetry of LTI SDEs)**.** *The Markov semigroup of a LTI SDE* (23) *is symmetric when* $-\mathbf{F}$ *and* $\mathbf{L}$ *are positive definite matrices, which excludes forward processes with a more complex structure, such as underdamped systems (Dockhorn et al., 2021).*

We could infer symmetry by checking if the transition kernel $k_t(\mathbf{x}, \mathbf{y}) := \frac{\rho_t(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{x})}$ is symmetric. In fact, when expressing P_t as a kernel integral operator, the symmetry becomes evident:

$$\int f(\mathbf{x}) P_t g(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) = \iint f(\mathbf{x}) k_t(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) \mathrm{d}\pi(\mathbf{y}) \mathrm{d}\pi(\mathbf{x}) = \int g(\mathbf{y}) P_t f(\mathbf{y}) \mathrm{d}\pi(\mathbf{y}).$$

Example 9 (OU process, transition kernel)**.** *For instance, the OU forward process has a Markov transition kernel* $\rho_t(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|e^{-t}\mathbf{x}, (1-e^{-2t})\mathbf{I})$*, which gives the transition kernel w.r.t.* π *as the* Mehler kernel:

$$k_t(\mathbf{x},\mathbf{y}) = \frac{1}{(1-e^{-2t})^{d/2}} \exp\left(-\frac{\|e^{-t}\mathbf{x}\|^2 + \|e^{-t}\mathbf{y}\|^2 - 2e^{-t}\langle \mathbf{x},\mathbf{y}\rangle}{2(1-e^{-2t})}\right).$$

Once we have established the symmetry of the Markov semigroup, from the definition of the infinitesimal generator (7), it follows that:

$$\int f(\mathbf{x})\mathscr{L}g(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) = \int g(\mathbf{x})\mathscr{L}f(\mathbf{x})\mathrm{d}\pi(\mathbf{x}), \qquad \forall f \in L^2(\pi),\, g \in \mathcal{D}(\mathscr{L}), \tag{25}$$

where $\mathcal{D}(\mathscr{L})$ is the *domain* of the infinitesimal generator, the set of functions $g \in L^2(\pi)$ and satisfying the following:

$$\int \mathscr{L}g(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) = 0. \tag{26}$$

Substituting (25) into the definition of the carré-du-champ operator (13), under the assumption that $fg \in \mathcal{D}(\mathscr{L})$, leads to the generalized integration-by-parts formula (15) underpinning OISM.

A.4 Operators of Asymmetric Markov Diffusions

When a Markov semigroup violates the symmetry constraint, the introduction of *adjoint* and *dual* operators becomes necessary (Bakry et al., 2013). Given the the infinitesimal generator $\mathscr{L}$ (8), the *Hermitian adjoint* of $\mathscr{L}$ in the inner product space $L^2(\pi)$ is defined as the linear operator $\mathscr{L}^*$ such that:

$$\int f(\mathbf{x})\mathscr{L}g(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) = \int g(\mathbf{x})\mathscr{L}^*f(\mathbf{x})\mathrm{d}\pi(\mathbf{x}), \qquad \forall f \in L^2(\pi),\, g \in \mathcal{D}(\mathscr{L}).$$

Observe that when the diffusion process is symmetric $\mathscr{L}^* = \mathscr{L}$

Example 10 (Adjoint generator for LTI SDEs)**.** *We can determine that the adjoint generator of* (24) *for LTI SDEs has the following form:*

$$\mathscr{L}^*f(\mathbf{x}) := \langle \mathbf{\Sigma}_\infty^{-1}\mathbf{x}, \mathbf{F}\mathbf{\Sigma}_\infty \nabla f(\mathbf{x})\rangle + \frac{1}{2}\langle \nabla, \mathbf{G}\nabla f(\mathbf{x})\rangle.$$

The adjoint generator is useful in defining the *dual semigroup* $\left(P_t^\dagger\right)_{t\geq 0}$, which acts on the set of probability measures. The semigroup $P_t^\dagger$ is implicitly defined by the duality below:

$$\int P_t f(\mathbf{x})\mathrm{d}\rho_0(\mathbf{x}) = \int f(\mathbf{x})\mathrm{d}\rho_t(\mathbf{x}) = \int f(\mathbf{x})\mathrm{d}\left(P_t^\dagger \rho_0\right)(\mathbf{x}).$$

Therefore, provided that ρ_0 is the law of $\mathbf{X}_0$, $P_t^\dagger\rho_0$ is the law of $\mathbf{X}_t$. When the Markov semigroup is symmetric and the density ratio $\frac{\rho_0}{\pi}$ is within the domain of P_t, $P_t^\dagger\rho_0$ is formally $\pi \cdot P_t\frac{\rho_0}{\pi}$, as evidenced by the following relation:

$$\int f(\mathbf{x})\mathrm{d}\left(P_t^\dagger\rho_0\right)(\mathbf{x}) = \int P_t f(\mathbf{x})\frac{\rho_0}{\pi}(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) = \int f(\mathbf{x})\mathrm{d}\left(\pi \cdot P_t\frac{\rho_0}{\pi}\right)(\mathbf{x}).$$

When the Markov semigroup is no longer symmetric, we can use $\left(e^{t\mathscr{L}}\right)^* = e^{t\mathscr{L}^*}$ for the derivation below:

$$\begin{aligned}
\int f(\mathbf{x})\mathrm{d}\left(P_t^\dagger\rho_0\right)(\mathbf{x}) &= \int P_t f(\mathbf{x})\frac{\rho_0}{\pi}(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) \\
&= \int e^{t\mathscr{L}}f(\mathbf{x})\frac{\rho_0}{\pi}(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) = \int f(\mathbf{x})\left(e^{t\mathscr{L}}\right)^*\frac{\rho_0}{\pi}(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) \\
&= \int f(\mathbf{x})e^{t\mathscr{L}^*}\frac{\rho_0}{\pi}(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) = \int f(\mathbf{x})\mathrm{d}\left(\pi \cdot e^{t\mathscr{L}^*}\frac{\rho_0}{\pi}\right)(\mathbf{x}).
\end{aligned}$$

Therefore, the dual semigroup can be similarly expressed with the adjoint operator: $P_t^\dagger \rho_0(\mathbf{x}) = \pi(\mathbf{x}) \cdot e^{t\mathscr{L}^*} \frac{\rho_0}{\pi}(\mathbf{x})$, or more concisely,

$$\frac{\rho_t}{\pi} = e^{t\mathscr{L}^*} \frac{\rho_0}{\pi}.$$

Differentiating both sides with respect to t, we obtain a partial differential equation (PDE) that describes the time evolution of the density ratio $\frac{\rho_t}{\pi}$, which naturally leads to the *Fokker-Planck equation*:

$$\frac{\partial}{\partial t} \frac{\rho_t}{\pi} = \mathscr{L}^* \frac{\rho_t}{\pi}, \qquad \frac{\partial}{\partial t} \rho_t = \pi \mathscr{L}^* \frac{\rho_t}{\pi}. \tag{27}$$

When the Markov semigroup lacks symmetry, we cannot use the same derivation adopted in this paper that exploits the carré-du-champ operator, and the generator $\mathscr{L}$ usually has a more complex, though not fully intractable, spectral structure. For example, the eigenfunctions can still take a polynomial form, and whether or not the eigenvalues of $\mathscr{L}$ are complex-valued depends on whether the system is underdamped or overdamped; see Pavliotis (2014, Section 6.3) for a full explanation.

B The Special Case of the Brownian Motion Process

The Brownian motion does not have finite invariant measure, given that π is the Lebesgue measure. As a consequence, even if its infinitesimal generator has a simple form, its spectral decomposition is less straightforward, compared to stable Markov diffusion processes, as detailed in Appendix B.1. In Appendix B.2 we propose an alternative diffusion process with the same operators, but finite support and countable spectrum.

B.1 Spectrum of the BM Infinitesimal Generator

As $\mathscr{L}$ is conceptually the time derivative of P_t, and f should satisfy the invariant condition $\mathbb{E}_{\mathbf{X}\sim\pi}[f(\mathbf{X})] = \mathbb{E}_{\mathbf{X}\sim\pi}[P_t f(\mathbf{X})]$, we can see that even if a function might satisfy $\mathscr{L} f(x) = \lambda f(x)$ pointwise for some $\lambda \in \mathbb{C}$, it might violate square integrability with respect to the Lebesgue measure or the invariant condition (26), and thus be excluded from the domain of $\mathscr{L}$. In fact, for the one-dimensional BM, functions $\phi_\xi(x) = \exp(\imath \xi x)$, $\xi \in \mathbb{R}$, formally have eigenvalue $\lambda_\xi = -\xi^2$, but they do not belong to $\mathcal{D}(\mathscr{L})$ as they are not square integrable and $\int \phi_\xi(x) \mathrm{d}\pi(x)$ do not exist.

Despite the above caveat, we are free to assume that $\int \phi_\xi \mathrm{d}\rho_0(x)$ does exist, and $P_t \phi_\xi(x)$ can be calculated in closed form:

$$\begin{aligned} P_t \phi_\xi(x) &= \int \exp(\imath \xi y) \mathcal{N}(y|x, 2t) \mathrm{d}y \\ &= \exp(-\xi^2 t + \imath \xi x) \\ &= \exp(-\xi^2 t) \phi_\xi(x), \end{aligned} \tag{28}$$

which is a direct result following the characteristic function of a Gaussian distribution. Therefore, the predictable time evolution $\mathbb{E}_{\rho_t}[\phi_\xi] = e^{-\xi^2 t} \mathbb{E}_{\rho_0}[\phi_\xi]$ still holds despite the domain issue.

Apart from the lack of square integrability, the uncountable nature of the BM eigenfunctions poses another issue for the enumeration "top n most important eigenfunctions" to be used to represent the energy-based model (16), as explained in Section 4.1.

B.2 Truncated BM Process

We propose to address both issues encountered by using the BM, by considering a different but related forward process, with modified domain and invariant measure. Our reasoning broadly follows that of Bakry et al. (2013, Chapter 3), which outlines that (symmetric) Markov diffusion operators could be effectively abstracted into the *Markov triple* (E, π, Γ), that is, the domain E of $\mathbf{x}$, a σ-finite invariant measure $\pi(\mathrm{d}\mathbf{x})$

and the carré-du-champ operator $\Gamma(\cdot,\cdot)$. Conceptually, given the Markov triple and a suitable function set within the domain of Γ, it is possible extrapolate the infinitesimal generator $\mathscr{L}$ and the Markov semigroup P_t.

The Markov triple of the one-dimensional Brownian motion is formally $(\mathbb{R}, \mathrm{d}x, f'(x)g'(x))$, but the Lebesgue measure $\mathrm{d}x$ is not σ-finite. However, if we constrain the domain of x to be a finite interval, chosen as $E = [-\pi, \pi]$, and specify the invariant measure π as the uniform distribution on E, the resulting Markov triple satisfies all conditions, namely, the σ-finiteness of the invariant measure.

Strictly speaking, this altered Markov triple induces a different generator and semigroup, and by extension, a completely new forward process, which we call *truncated Brownian motion*. We demonstrate that this forward process largely coincides with Brownian motion with one single revision. Our description focuses on the 1-dimensional setting, but our argument extends trivially to high-dimensions.

Example 11 (Truncated BM)**.** *To sample X_t from a truncated BM, given the initial state $X_0 = x$, we first sample $\tilde{X}_t \sim \mathcal{N}(x, 2t)$ according to the Brownian motion* (2)*, and then shift $\tilde{X}_t$ using multiples of 2π until it falls into the interval $[-\pi, \pi]$. Specifically,*

$$X_t = \begin{cases} \tilde{X}_t \bmod 2\pi, & \text{if } \left(\tilde{X}_t \bmod 2\pi\right) \leq \pi, \\ \tilde{X}_t \bmod 2\pi - 2\pi, & \text{otherwise.} \end{cases}$$

Therefore, the conditional distribution $\rho_t(y|x)$ follows an infinite sum:

$$\rho_t(y|x) = \sum_{z\in\mathbb{Z}} \mathcal{N}(y|x + 2\pi z, 2t). \tag{29}$$

We could see that, as $t \to \infty$, the conditional distribution converges to a uniform distribution, which is therefore the stationary distribution. Since a truncated BM is indistinguishable from an unconstrained BM for a small t, the generator $\mathscr{L}$ and the carré-du-champ operator coincide with those of the Brownian motion. Moreover, the functions $\left\{\sqrt{2}\exp(\imath nx)\right\}_{n\in\mathbb{Z}}$ form an orthonormal basis on $L^2([-\pi, \pi])$, and they have period 2π when extended onto the real line $\mathbb{R}$ and therefore satisfy (26). Furthermore,

$$\begin{aligned} P_t\phi_n(x) &= \sum_{z\in\mathbb{Z}} \int_{-\pi}^{\pi} \exp(\imath ny)\mathcal{N}(y|x + 2\pi z, 2t)\mathrm{d}y \\ &= \int_{\mathbb{R}} \exp(\imath ny)\mathcal{N}(y|x, 2t)\mathrm{d}y \\ &= \exp(-n^2 t + \imath nx) \\ &= \exp(-n^2 t)\phi_n(x). \end{aligned}$$

Most notably, in the truncated setting (29), the eigen-relation (28) does not apply to all functions $\phi_\xi = \exp(\imath\xi x)$, $\xi \in \mathbb{R}$, as re-writing the infinite sum into an integral on $\mathbb{R}$ in the second identity above applies to only functions with a period of 2π, adding the additional constraint that $\xi \in \mathbb{Z}$. Therefore, the eigen-pairs contained within $\mathcal{D}(\mathscr{L})$ only include $\left(-n^2, \sqrt{2}\exp(\imath nx)\right)_{n\in\mathbb{Z}}$, and the eigenfunctions satisfy both square integrability with respect to π and the invariance property.

As a consequence, by constraining the BM on a finite interval, we obtain a diffusion process with a finite stationary measure, as well as a better-defined countable set of eigenvectors, which can be sorted by importance.

Enumerating the eigenfunctions The truncated Brownian motion allows for the enumeration of the "top-n most important eigenfunctions". In order to circumvent working with complex numbers, we make the practical choice to use the sine and cosine functions instead of complex exponentials, that is, we consider the eigenpairs:

$$\left\{\left(-k^2, \sqrt{2}\cos(kx)\right)\right\}_{k\in[n]} \bigcup \left\{\left(-k^2, \sqrt{2}\sin(kx)\right)\right\}_{k\in[n]}$$

for the spectrum of the generator $\mathscr{L}$. Additionally, the representation of the product of two trigonometric as sums requested in Theorem 2 can be obtained via simple trigonometric identities. In the multi-dimensional

setting, we first set up a threshold of the minimum eigenvalue $-n_d$, and denote $\boldsymbol{\Xi}(n_d) := \{\boldsymbol{\xi} \in \mathbb{Z}^d : \|\boldsymbol{\xi}\|^2 \leq n_d\}$. Then the set of eigenpairs consists of:

$$\left\{\left(-\|\boldsymbol{\xi}\|^2, \sqrt{2}\cos(\boldsymbol{\xi}^\top \mathbf{x})\right)\right\}_{\boldsymbol{\xi}\in\boldsymbol{\Xi}(n_d)} \bigcup \left\{\left(-\|\boldsymbol{\xi}\|^2, \sqrt{2}\sin(\boldsymbol{\xi}^\top \mathbf{x})\right)\right\}_{\boldsymbol{\xi}\in\boldsymbol{\Xi}(n_d)}.$$

Boundary conditions The nature of the truncated Brownian motion dictates that the data distribution ρ_0 needs to satisfy the boundary condition: $\rho_0(-\pi) = \rho_0(\pi)$. A practical workaround when this condition is violated by the data distribution is to scale the samples so that these are compactly supported within $[-\pi, \pi]^d$, while simultaneously increasing the truncation order n_d of the eigenfunctions required to resolve fine-scale details of the data distribution.

C Additional Notes on OISM

This section outlines pragmatic implementation details for obtaining the OISM score in Appendix C.1, the usage of orthogonal functions in density estimation and its connection to Stein's normal means problem in Appendix C.2, the specific form of shrinkage estimators in Appendix C.3.

C.1 Computational Implementation

To apply Theorem 2, we need to enumerate eigenfunctions of the Markov diffusion process, express their products as linear combinations of an extended basis, and solve a linear system.

Hermite polynomial recurrences In Appendix B.2 we explained how trigonometric identities allow to easily handle the eigenfunctions of the truncated BM. OISM for the OU forward process requires the calculation of normalized Hermite polynomials $\phi_n(x) = \frac{\mathrm{He}_n(x)}{\sqrt{n!}}$. We mainly use the recurrence relation below, for $\ell \geq 1$,

$$\phi_{\ell+1}(x) = \frac{1}{\sqrt{\ell+1}} x\phi_\ell(x) - \sqrt{\frac{\ell}{\ell+1}}\phi_{\ell-1}(x), \qquad \phi_0(x) = 1, \quad \phi_1(x) = x; \tag{30}$$

see Abramowitz & Stegun (1965). The recurrence relation circumvents the explicit expression of $\phi_n(x)$ as a polynomial, which might include large numerical coefficients for some x^p.

As we require the polynomial decomposition of $\phi_k\phi_\ell$ in order to evaluate the carré-du-champ operator, we further make use of the recurrence relation to obtain a dynamic programming approach to evaluate $\phi_k\phi_\ell = \sum_{h=1}^m \beta_h^{(k,l)}\phi_h$. Firstly, we note that $\boldsymbol{\beta}^{(k,0)} = \mathbb{1}^{(k)}$ by definition, where $\mathbb{1}^{(k)}$ refers to the one-hot vector representation where the k-th entry is 1, with all other entries being 0. In addition, as $x = \phi_1(x)$, the recurrence relation (30) implies the following:

$$\boldsymbol{\beta}^{(k,1)} = \sqrt{k+1}\boldsymbol{\beta}^{(k+1,0)} + \sqrt{k}\boldsymbol{\beta}^{(k-1,0)}.$$

Multiplying both sides of equation (30) with ϕ_k, we derive the high-order recurrence to calculate $\boldsymbol{\beta}^{(k,\ell)}$ for $\ell \geq 2$:

$$\boldsymbol{\beta}^{(k,\ell+1)} = \sqrt{\frac{k+1}{\ell+1}}\boldsymbol{\beta}^{(k+1,\ell)} + \sqrt{\frac{k}{\ell+1}}\boldsymbol{\beta}^{(k-1,\ell)} - \sqrt{\frac{\ell}{\ell+1}}\boldsymbol{\beta}^{(k,\ell-1)}.$$

Preconditioned Linear Solver OISM requires solving the linear system $\mathbf{A}_t\widehat{\boldsymbol{\alpha}} = \mathbf{b}_t$. To enhance numerical stability, we adopt a simple preconditioner. Using similar steps as in the proof of Theorem 2

$$\begin{aligned}\int \Gamma(\phi_k(\mathbf{x}), \phi_\ell(\mathbf{x}))\mathrm{d}\pi(\mathbf{x}) &= \frac{1}{2}\int \mathscr{L}(\phi_k\phi_\ell)(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) - \frac{\lambda_k+\lambda_\ell}{2}\int \phi_k(\mathbf{x})\phi_\ell(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) \\ &= -\frac{\lambda_k+\lambda_\ell}{2}\delta_{k\ell},\end{aligned}$$

where the second equality is based on the fact that $\{\phi_n\}$ are orthonormal in $L^2(\pi)$ so that $\mathbb{E}_\pi[\mathscr{L}(\phi_k\phi_\ell)] = 0$. Therefore, as ρ_t converges to π as $t \to \infty$, the matrix $\mathbf{A}_t$ converges to a diagonal matrix $\boldsymbol{\Lambda} := \mathrm{diag}(-\lambda_1, -\lambda_2, \ldots, -\lambda_n)$, and we opt to solve the *preconditioned* system $\boldsymbol{\Lambda}^{-1}\mathbf{A}_t\widehat{\boldsymbol{\alpha}} = \boldsymbol{\Lambda}^{-1}\mathbf{b}_t$.

C.2 Orthogonal Series Density Estimation

Orthogonal functions in density estimation was first proposed by Whittle (1958); Cencov (1962); Schwartz (1967), and it remains a topic of interest, see the recent works in variational inference (Cai et al., 2024). Given a set of basis functions $(\phi_n)_{n\in\mathbb{N}}$ in the space of square-integrable functions $L^2(\pi)$ and assuming the density ratio $\frac{\rho}{\pi}$ is square-integrable, we can decompose the density ratio into the following:

$$\frac{\rho}{\pi}(\mathbf{x}) = \sum_{n\in\mathbb{N}} \mathbb{E}_\rho\left[\phi_n\right]\cdot\phi_n(\mathbf{x}) \approx \sum_{n\in\mathbb{N}}\left(\frac{1}{M}\sum_{m=1}^{M}\phi_n(\mathbf{x}_m)\right)\cdot\phi_n(\mathbf{x}) =: \frac{\widehat{\rho}}{\pi}(\mathbf{x}). \tag{31}$$

The right-hand side of (31) refers to the infinite series with sample means using M samples plugged in for each coefficient, which is typically numerically unstable, or even not exist at all. For example, in the case of Hermite polynomials, the infinite series holds in the sense of distributions; see Courant & Hilbert (1954, Section II.9):

$$\sum_{n\in\mathbb{N}}\phi_n(\mathbf{x})\phi_n(\mathbf{x}_0) = \delta(\mathbf{x}-\mathbf{x}_0).$$

Therefore, the right-hand side of the infinite series yields the density estimate $\frac{1}{M}\sum_{m=1}^{M}\delta(\mathbf{x}-\mathbf{x}_m)$, indicating the superfluous "density estimate" that simply involves Dirac delta's on the M samples themselves. This is an undesirable outcome, as it defeats the purpose of density estimation and generative modeling: a model that generates exact copies from within the dataset $\{\mathbf{x}_m\}_{m=1}^{M}$ lacks the ability to generate novel, unseen samples.

The infeasibility of replacing the orthogonal series coefficients with sample averages (31) could also be seen from the inadmissibility of the sample mean estimator in normal means estimation (Stein, 1956). Employing the central limit theorem, we could derive that when ρ and π are sufficiently similar, the sample means estimator $\widehat{\boldsymbol{\theta}}\in\mathbb{R}^n$ with components $\widehat{\boldsymbol{\theta}}^{(k)} := \frac{1}{M}\sum_{m=1}^{M}\phi_k(\mathbf{x}_m)$ converges in distribution to the Gaussian $\mathcal{N}\left(\boldsymbol{\theta}, \frac{1}{M}\boldsymbol{\Sigma}\right)$, where

$$\boldsymbol{\theta}^{(k)} := \mathbb{E}_\rho[\phi_k], \qquad \boldsymbol{\Sigma}^{(k,\ell)} := \mathbb{E}_\rho\left[\left(\phi_k-\boldsymbol{\theta}^{(k)}\right)\left(\phi_\ell-\boldsymbol{\theta}^{(\ell)}\right)\right] \approx \delta_{k\ell},$$

thus recovering the exact same scenario as Stein's paradox. In fact, when orthogonal functions are used for nonparametric regression and density estimation, the need for lower-risk shrinkage estimators has been widely acknowledged; see Wasserman (2005, Chapter 8) and Efromovich (2010) for detailed background review on suitable choices for linear shrinkage.

C.3 Modulation Estimator

We explain in this section how we implement the linear shrinkage estimator. Given the sample means estimator $\widehat{\boldsymbol{\theta}}$, the simplest linear shrinkage would be the James-Stein estimator (James & Stein, 1961), which uses the estimator $\gamma\widehat{\boldsymbol{\theta}}$, where $0\le\gamma\le 1$ is a scalar-valued shrinkage parameter. We opt for the slightly more flexible *modulation estimator* (Beran & Dumbgen, 1998), which uses $\boldsymbol{\gamma}\circ\widehat{\boldsymbol{\theta}}$, where $\boldsymbol{\gamma}\in[0,1]^n$ is a vector-valued shrinkage estimator, and $\circ$ denotes Hadamard (element-wise) product.

To construct the modulation estimator, we require an estimate of the variance of the sample mean, which is obtained as follows:

$$\hat{\sigma}_k^2 = \frac{1}{M^2}\sum_{m=1}^{M}\left(\phi_k(\mathbf{x}_m)-\widehat{\boldsymbol{\theta}}^{(k)}\right)^2.$$

Then the empirical risk is obtained as:

$$\widehat{R}(\boldsymbol{\gamma}) = \sum_{k=1}^{n}\gamma_k^2\hat{\sigma}_k^2 + (1-\gamma_k)^2\max\left\{\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)^2-\hat{\sigma}_k^2, 0\right\}, \tag{32}$$

which is minimized at $\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in[0,1]^n}\widehat{R}(\boldsymbol{\gamma})$.

D Experimental Settings: Low-Dimensional Illustrations

This appendix contains full details required to reproduce the one- and two-dimensional experiments that we reported in the main text.

Computational resources This experiment was implemented in JAX and performed on the CPU of an Apple MacBook M2 laptop. OISM takes virtually no training as we only need to compute the shrinkage estimator of the expected eigenfunctions, so the main computational cost comes from density estimation and data generation. For the one-dimensional toy problem, the score network takes $x \in \mathbb{R}$ and diffusion time t as inputs. The time t is mapped to a 128-dimensional sinusoidal embedding and encoded by an MLP ($128 \to 64 \to 256$), while x is mapped to periodic features $(\cos x, \sin x) \in \mathbb{R}^2$ and encoded by an MLP ($2 \to 64 \to 256$). The two embeddings are concatenated and passed through a decoder MLP ($512 \to 64 \to 64 \to 1$), with SiLU activations in all hidden layers and no final activation, producing the scalar score estimate $\hat{s}_t(x) \approx \nabla_x \log \rho_t(x)$.

Forward process and selection of eigenfunctions To generate samples for training the score in DDPM, we use the BM forward process with a changed noise schedule, such that $0 \le \tau \le 1$, $\sigma_{\min} = 0.01$, $\sigma_{\max} = 50$:

$$\mathrm{d}\mathbf{X}_\tau = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\tau}\sqrt{2\log\frac{\sigma_{\max}}{\sigma_{\min}}}\mathrm{d}\mathbf{W}_\tau, \qquad \mathbf{X}_0 \sim \rho_0,$$

For the OISM score (21), in the 1-dimensional example we choose a combined total of $n = 50$ eigenfunctions of the corresponding truncated BM, as explained in Appendix B, that is $\sqrt{2}\cos(kx)$ and $\sqrt{2}\sin(kx)$ for $k = 1, 2, \ldots, 25$. As we require calculation of $\mathbb{E}_{\rho_0}[\phi_\ell]$ on an extended set to account for the carré-du-champ operator, the extended set involves $k = 1, 2, \ldots, 50$.

For the 2-dimensional example, we enumerate all eigenfunctions (λ_k, ϕ_k) such that $\lambda_k \ge -125$, which includes 200 eigenfunctions. Calculating the carré-du-champ operator requires eigenfunctions corresponding to eigenvalues no less than $-125 \times 2^2 = -500$, which includes 790 eigenfunctions. After calculating the sample mean, we apply the dimension-wise linear shrinkage estimator (32), as explained in Appendix C.3.

Synthetic datasets The Bart Simpson example, described in Wasserman (2005, Chapter 6), is a Gaussian mixture distribution $\rho_0(x) = \frac{1}{2}\mathcal{N}(x; 0, 1) + \frac{1}{10}\sum_{j=0}^{4}\mathcal{N}(x; j/2 - 1, 1/100)$, which yields a closed-form marginal distribution $\rho_\tau(x) = \frac{1}{2}\mathcal{N}(x; 0, 1 + \sigma_\tau^2) + \frac{1}{10}\sum_{j=0}^{4}\mathcal{N}(x; j/2 - 1, 1/100 + \sigma_\tau^2)$, where $\sigma_\tau := \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\tau}$. The expected values $\mathbb{E}_{\rho_0}[\phi_k]$ can be calculated in closed form given the characteristic function of Gaussian distributions. A total of $N = 2,000$ data points are simulated from the Bart Simpson example. The 2-dimensional synthetic data were generated using existing code that was downloaded from `https://github.com/rtqichen/ffjord` (Grathwohl et al., 2018), under the MIT license. For all experiments, we generated synthetic datasets of size $N = 20,000$.

Simulation of the probability flow ODE In order to generate samples via probability flow ODE (4) adapted for the truncated BM, we need to initialize $\mathbf{X}_1 \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\max}^2\mathbf{I}\right)$, and rescale $\mathbf{X}_1$ within the interval $[-\pi, \pi]^d$. It is also possible to directly sample $\mathbf{X}_1$ from the invariant measure Uniform$[-\pi, \pi]^d$. The exact form of the probability flow ODE is then:

$$\frac{\mathrm{d}\mathbf{X}_\tau}{\mathrm{d}\tau} = -\sigma_{\min}^2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2\tau}\log\frac{\sigma_{\max}}{\sigma_{\min}}\hat{\mathbf{s}}_{t(\tau)}(\mathbf{X}_\tau) := \mathbf{f}_\tau(\mathbf{X}_\tau), \tag{33}$$

where $\hat{\mathbf{s}}_{t(\tau)}(\mathbf{x})$ is the OISM solution (21), solved via a preconditioner as described in Appendix C.1, for the correct noise level, which corresponds to $t(\tau) = \frac{1}{2}\sigma_{\min}^2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2\tau}$. The generation of samples is achieved via first sampling from $\mathbf{X}_1 \sim \rho_1$ (the invariant measure) and then simulating (33) backwards from $\tau = 1$ to 0; if the end-point $\mathbf{X}_0$ is outside the range $[-\pi, \pi]^d$, we need to rescale $\mathbf{X}_0$ following the procedure in Example 11. Simulating (33) forward from $\tau = 0$ yields a density estimate via the infinitesimal change-of-variables formula

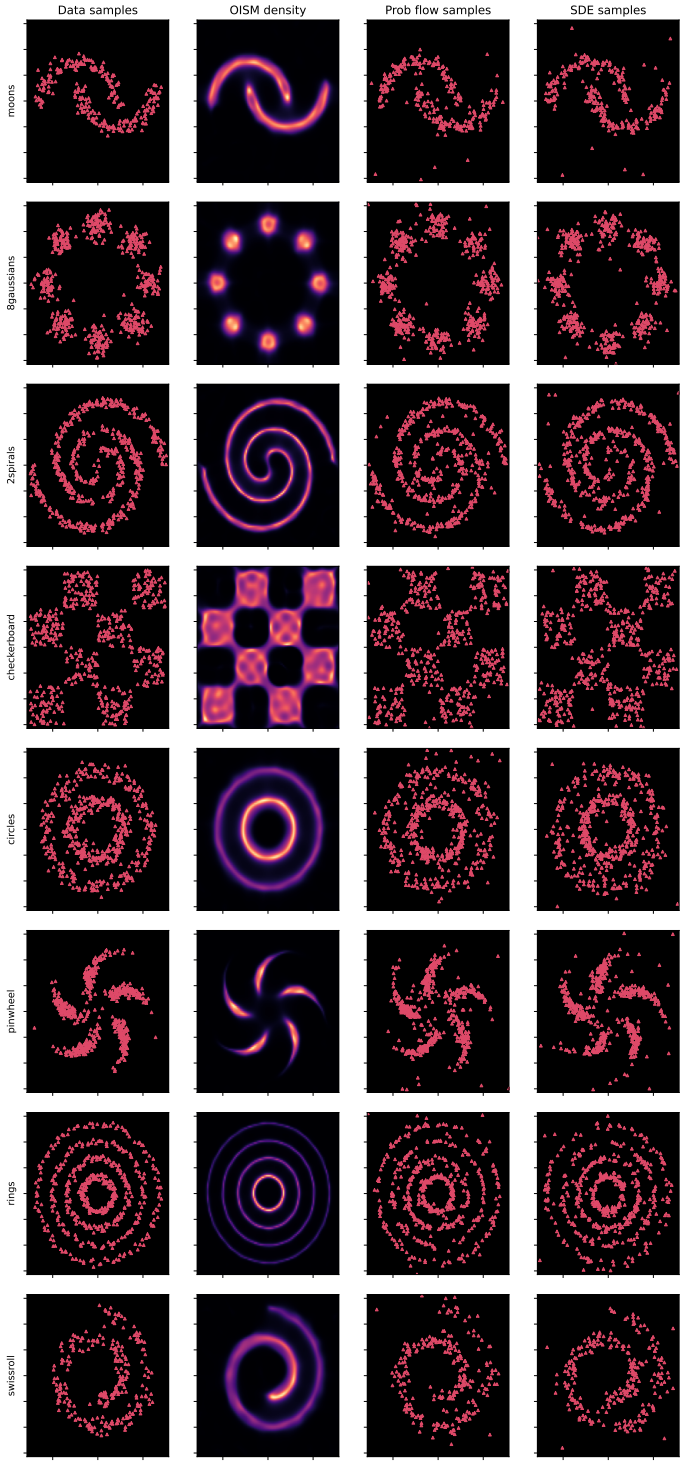

Figure 4: Additional illustrations on toy dataset. The first column shows (a subset of) ground truth samples from the data distribution; The second column visualizes the OISM density obtained via probability flow ODE; The third column plots OISM sample points simulated via probability flow ODE; The last column plots OISM samples generated by a reverse SDE sampler.

(Chen et al., 2018):

$$\log \rho_0^{(\mathrm{ODE})}(\mathbf{x}_0) := \log \pi(\mathbf{x}_1) + \int_0^1 \nabla \cdot \mathbf{f}_\tau(\mathbf{x}_\tau) \mathrm{d}\tau,$$

where $\log \pi(\mathbf{x}_1) = -d \log 2\pi$. We use the adaptive ODE solver `tsit5` (Tsitouras, 2011) for both tasks of data generation and density estimation.

Additional results Figure 4 demonstrates results obtained via OISM for a wide range of 2-dimensional datasets from `https://github.com/rtqichen/ffjord` (Grathwohl et al., 2018), as well as samples obtained via simulation of the reverse SDE, which analogously requires rescaling to the interval $[-\pi, \pi]^d$:

$$\mathrm{d}\mathbf{X}_\tau = -2\sigma_{\min}^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2\tau} \log \frac{\sigma_{\max}}{\sigma_{\min}} \, \hat{\mathbf{s}}_{t(\tau)}(\mathbf{X}_\tau) \, \mathrm{d}\tau + \sqrt{2\sigma_{\min}^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2\tau} \log \frac{\sigma_{\max}}{\sigma_{\min}}} \, \mathrm{d}\mathbf{W}_\tau.$$

E Experimental Settings: Image Generation

This appendix contains full details required to reproduce the image generation experiments that we reported in the main text.

Computational resources For our neural network architecture, we adapted the JAX implementation of the score-based diffusion model from Song et al. (2021). The original repository was obtained under the MIT license from `https://github.com/yang-song/score_flow` and subsequently modified using OISM. All experiments were executed on an NVIDIA A100 GPU.

Forward process and eigenfunctions The OU forward process takes the following form for $\tau \in [0, 1]$:

$$\mathrm{d}\mathbf{X}_\tau = -\frac{\beta_\tau}{2}\mathbf{X}_\tau \mathrm{d}\tau + \sqrt{\beta_\tau}\mathrm{d}\mathbf{W}_\tau, \qquad \beta_\tau = \beta_0 + \tau(\beta_1 - \beta_0), \qquad \mathbf{X}_0 \sim \rho_0, \tag{34}$$

where $\beta_0 = 0.1$ and $\beta_1 = 20$. The above forward process links to the original OU forward process (3) by the change of time variable formula $t(\tau) = \frac{\beta_0 \tau}{2} + \frac{(\beta_1 - \beta_0)\tau^2}{4}$. As OISM for image generation requires Hermite polynomials, we mainly use the Hermite recurrence relations mentioned in Appendix C.1 to calculate quantities related to OISM.

Image datasets The CIFAR-10 (Krizhevsky et al., 2009, Chapter 3) and ImageNet32 (Chrabaszcz et al., 2017) data is obtained via the TensorFlow Datasets API `https://github.com/tensorflow/datasets`; this fetches the CIFAR-10 data from its original source at `https://www.cs.toronto.edu/~kriz/cifar.html` under the terms of the MIT License, and the ImageNet32 data from its catalog under `imagenet_resized/32x32`. The CIFAR-10 dataset consists of 60,000 colour images of dimension 32×32, partitioned into 50,000 training and 10,000 test images. The ImageNet32 dataset consists of 1,331,167 colour images of dimension 32×32, partitioned into 1,281,167 training and 50,000 test images. Although the original images in both datasets have integer pixel values ranging from 0 to 255, we normalize them to between -1 and 1, so they fall in the high density of the invariant distribution of the OU forward process.

Implementation of standard DDPM The traditional approach of (Ho et al., 2020) trains a neural network to provide a flexible time-dependent vector field $\boldsymbol{\epsilon}_\tau(\mathbf{x})$ by minimizing the following objective:

$$\mathcal{J}(\boldsymbol{\epsilon}_\tau) = \frac{1}{2}\int_0^1 \mathbb{E}_{\substack{\mathbf{X}_0 \sim \rho_0 \\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}} \left[w(\tau) \left\| \boldsymbol{\epsilon}_\tau(\mathbf{X}_\tau) - \boldsymbol{\epsilon} \right\|^2 \right] \mathrm{d}\tau, \tag{35}$$

where $\boldsymbol{\epsilon}$ is the perturbation vector of the forward noise process, in our experiments re-parametrized as $\mathbf{X}_\tau = \alpha_\tau \mathbf{X}_0 + \sigma_\tau \boldsymbol{\epsilon}$, and the signal and noise scale factors α_τ and σ_τ are determined by solving the forward variance-preserving SDE (34):

$$\alpha_\tau = \exp\left(-\frac{1}{2}\int_0^\tau \beta_s \mathrm{d}s\right) = \exp\left(-\frac{1}{2}\left[\beta_0 \tau + \frac{\tau^2}{2}(\beta_1 - \beta_0)\right]\right), \qquad \sigma_\tau = \sqrt{1 - \alpha_\tau^2}. \tag{36}$$

It is possible to prove that (35) is equivalent to the score matching loss (12), upon (approximately) identifying the score that drives the probability flow ODE (used in our experiments) or reverse SDE for image generation as:

$$\boldsymbol{s}_\tau(\mathbf{x}) = -\boldsymbol{\epsilon}_\tau(\mathbf{x})/\sigma_\tau.$$

Additionally, Song et al. (2021) discover that minimizing (35) with weighting function $w(\tau) = \beta_\tau/\sigma_\tau^2$ is equivalent to maximizing $\mathbb{E}_{\mathbf{X}_0\sim\rho_0}\left[\log \rho_0^{(\mathrm{SDE})}(\mathbf{X}_0)\right]$, where $\rho_0^{(\mathrm{SDE})}$ represents the marginal distribution of the SDE-based sampler driven by the score corresponding to $\boldsymbol{\epsilon}_\tau$. We adopt this maximum likelihood weighting scheme, and use the importance sampling variance reduction techniques proposed in the same work by the authors, to stabilize optimization under the increased variance introduced by the likelihood weight.

Following common experimental practice for image generation tasks, we use a UNet architecture for score estimation. Specifically, our backbone consists of four ResNet blocks integrated with learned time-step embeddings to condition the network across the temporal trajectory.[1] The network parameters are optimized using the Adam optimizer for 500,000 iterations at a constant learning rate of 2×10^{-4} and a batch size of 128. The training can be completed in approximately 25 hours on a single NVIDIA A100 GPU.

Implementation of OISM with learned residuals OISM provides $\hat{\mathbf{s}}_\tau(\mathbf{x})$ (21) as an estimate of $\nabla \log \rho_\tau(\mathbf{x})$, for $\tau \in [0,1]$. The OISM score estimates are fixed throughout the training and evaluation processes, and the sequence of linear systems $\left(\mathbf{A}_\tau^{-1}\mathbf{b}_\tau\right)_{\tau\in[0,1]}$ is correlated temporally, as illustrated in Figure 5. For both training and evaluation, we obviate the need to call a linear solver for every score evaluation by a pre-processing step, that is, we pre-solve the linear systems on an evenly-spaced grid of 1,000 τs between 0 and 1, and extend the solution to continuous times by linear interpolation. The training time via pre-solving and interpolation was similar to that of a neural network only DDPM. The extra computation required for interpolating the OISM solutions constituted a small fraction of the overall computational cost; see Table 1. As in Appendix D, the linear systems for finding the optimal OISM coefficients are solved using a preconditioner as per Appendix C.1. However, unlike Appendix D, we do not apply shrinkage estimators but instead evaluate $\mathbb{E}_{\rho_0}[\phi_k]$ using plain sample means, because we do not observe practical benefit from shrinkage (not reported here).

As explained in Section 4.3, once $\hat{\mathbf{s}}_\tau(\mathbf{x})$ is available from OISM, in high-dimensions we train a neural network to learn the residual score (22). In principle, neural networks that optimize loss functions equivalent up to a constant (such as DSM and ISM), or whose optimal point can be recovered from these (as in OISM) should have similar performance. However, because optimization is performed with respect to different variables (the random perturbation vector in DSM, the score in ISM, or yet the gradient of the log-density ratio in OISM), one is required to designate what target function the neural network is attempting to estimate. As the neural network estimation target changes, the training process requires different hyperparameters; see Salimans & Ho (2022, Table 1). In our scenario, we opt for the maximum likelihood weighting function, and choose the neural network output as an approximation of the perturbation vector $\boldsymbol{\epsilon}$, with the same neural network configuration as for standard DDPM. Following the parameterization established in (35), the residual score that we wish to learn is:

$$\begin{aligned}\mathbf{r}_\tau(\mathbf{x}) &= \nabla \log \rho_\tau(\mathbf{x}) - \hat{\mathbf{s}}_\tau(\mathbf{x}) \\ &= -\frac{\boldsymbol{\epsilon}}{\sigma_\tau} - \hat{\mathbf{s}}_\tau(\mathbf{x}),\end{aligned}$$

where σ_τ is defined in (36). We can equivalently write $\boldsymbol{\epsilon} = -\sigma_\tau(\hat{\mathbf{s}}_\tau(\mathbf{x}) + \mathbf{r}_\tau(\mathbf{x}))$, leading to define the neural residual score estimator $\boldsymbol{e}_\tau(\mathbf{x})$, in the $\boldsymbol{\epsilon}$-parametrization, as the minimizer of the objective:

$$\mathcal{J}(\boldsymbol{e}_\tau) = \frac{1}{2}\int_0^1 \mathbb{E}_{\substack{\mathbf{X}_0\sim\rho_0 \\ \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}}\left[w(\tau)\left\|\boldsymbol{\epsilon} - \left(-\sigma_\tau\hat{\mathbf{s}}_\tau(\mathbf{X}_\tau) + \boldsymbol{e}_\tau(\mathbf{X}_\tau)\right)\right\|^2\right]\mathrm{d}\tau,$$

We train this neural network using the same architecture as for standard DDPM, while the learning rate is tuned to optimize the results over a grid of candidate values. Once $\boldsymbol{e}_\tau$ is obtained, samples can be generated solving the probability flow ODE or the reverse SDE governed by the total score $\boldsymbol{s}_\tau(\mathbf{x}) = \left(\sigma_\tau\hat{\mathbf{s}}_\tau(\mathbf{X}_\tau) - \boldsymbol{e}_\tau(\mathbf{X}_\tau)\right)/\sigma_\tau$. In all the experiments, samples are generated using the probability flow ODE.

[1] Full architectural configurations and hyper-parameters are provided in the source code within the electronic supplement.

Table 1: Average runtime for each training step (in seconds) for CIFAR-10. Interpolating the solutions of the OISM linear system introduces only a small computational overhead, compared to DDPM.

Method	OISM (3)	OISM (6)	DDPM
Exact solution	0.428	0.799	0.308
Linear interpolation	0.315	0.312	

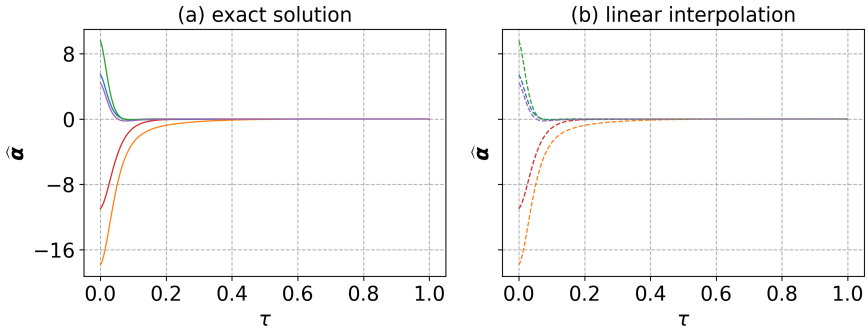

Figure 5: Coefficients $\widehat{\boldsymbol{\alpha}}$ from **OISM** (6) for CIFAR-10, as a function of the normalized time τ. Left: exact solution on the grid of points at which the linear systems were solved during training. Right: linearly interpolated solutions obtained on a grid of 1,000 evenly-spaced time points. Simple linear interpolation is sufficient for training and evaluation of the neural network, due to the temporal correlation of the coefficients.

Table 2: Performance at the end of training. NLL, FID, IS, and KID of DDPM, **OISM** (3), and **OISM** (9) evaluated at 950k training iterations.

Dataset	CIFAR10				ImageNet32			
Metric	NLL (↓)	FID (↓)	IS (↑)	KID (↓)	NLL(↓)	FID (↓)	IS (↑)	KID (↓)
DDPM	2.95	6.00	8.97	3.30e-03	3.59	14.52	9.98	1.37e-02
OISM(3)	2.95	7.60	8.57	5.48e-03	3.59	15.30	9.92	1.48e-02
OISM(9)	3.56	14.14	8.65	6.19e-03	3.99	18.22	9.46	1.37e-02

Performance evaluation To evaluate scalability, we extended the CIFAR-10 experiments shown in Figure 3 to the ImageNet32 dataset, displayed in Figure 6. For this task, we also evaluated OISM (9) as an alternative configurations to OISM (6); however, the empirical trends remained highly consistent with our CIFAR-10 observations. We further evaluated standard DDPM and OISM using negative log-likelihood (NLL) derived from the probability flow ODE, alongside three standard generative metrics: Inception Score (IS) Salimans et al. (2016), Fréchet Inception Distance (FID) Heusel et al. (2017), and Kernel Inception Distance (KID) Sutherland et al. (2018). Here, IS captures generation quality and diversity, whereas FID and KID evaluate how closely the generated data distribution matches the true distribution.

Figures 3 and 6 demonstrate that OISM offers a substantial performance advantage over DDPM during early training stages, while Table 2 shows that both approaches ultimately converge to comparable performance levels as training progresses.

F Further Consequences of Markov Diffusion Operators

The aim of this appendix is to revisit objectives related to score matching through the lens of Markov diffusion operators, specifically *time score matching* (Choi et al., 2022), the *score Fokker–Planck equation* (Lai et al., 2023), and *generalized* score matching (Benton et al., 2024). Our main tool of analysis is the *diffusion property*

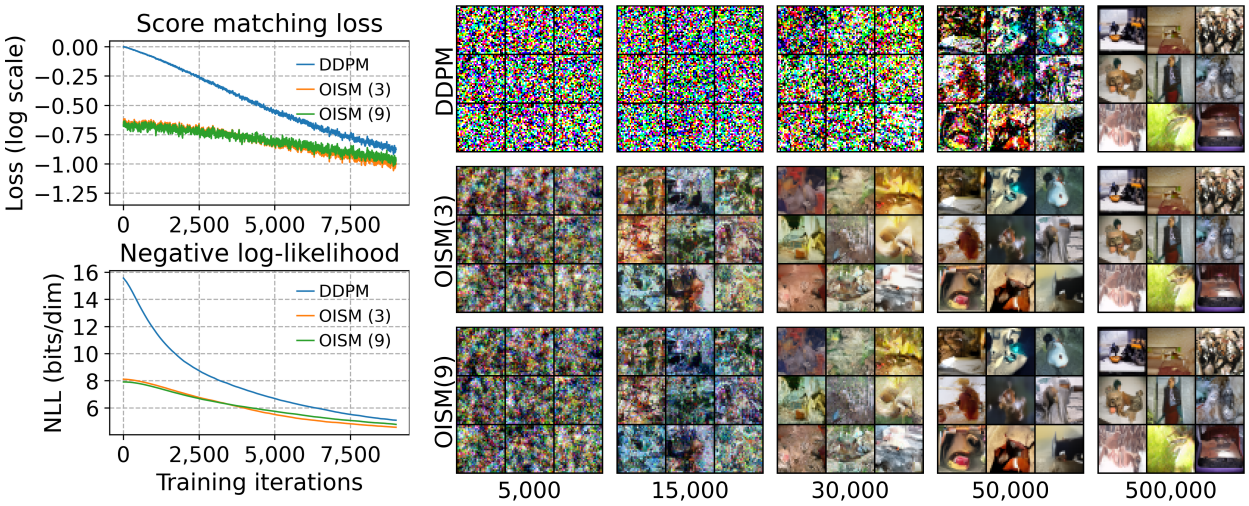

Figure 6: Performance evaluation on ImageNet32. Top left: Smoothed score matching (SM) loss for **DDPM** (the maximum likelihood estimator of Song et al., 2021), **OISM** (3), and **OISM** (9). Bottom left: Negative log-likelihood evaluated on the test dataset. Right: Samples corresponding to different training iterations, obtained by reversing the probability flow ODE (4).

of the generator $\mathscr{L}$ which states that, for sufficiently smooth f and $\psi : \mathbb{R} \to \mathbb{R}$, the following change of variable formula applies:

$$\mathscr{L}\psi(f) = \psi'(f)\mathscr{L}f + \psi''(f)\Gamma(f, f); \tag{37}$$

see Bakry et al. (2013, Definition 1.11.1).

Insight into time score matching An alternative to the score matching objective is to match the time derivative of the log density ratio, an approach termed *time score matching* in Choi et al. (2022), which, analogous to score matching, minimizes the expected squared loss with respect to the *temporal* derivative:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}_t\sim\rho_t}\left[\frac{\partial}{\partial t}f_t(\mathbf{X}_t) - \frac{\partial}{\partial t}\log\frac{\rho_t}{\pi}(\mathbf{X}_t)\right]^2 \overset{+C}{=} \mathbb{E}_{\mathbf{X}_0\sim\rho_0}&\left[P_t\left(\frac{\partial}{\partial t}f_t\right)^2(\mathbf{X}_0)\right] \\ &- 2\mathbb{E}_{\mathbf{X}_t\sim\rho_t}\left[\frac{\partial}{\partial t}f_t(\mathbf{X}_t)\right]\left[\frac{\partial}{\partial t}\log\frac{\rho_t}{\pi}(\mathbf{X}_t)\right].\end{aligned}$$

The intractable second term can similarly be resolved via integration by parts, and the time evolution of density ratio (27) assuming symmetry of the diffusion, so that $\mathscr{L}^* = \mathscr{L}$:

$$\begin{aligned}\int\left[\frac{\partial}{\partial t}f_t(\mathbf{x})\right]\left[\frac{\partial}{\partial t}\log\frac{\rho_t}{\pi}(\mathbf{x})\right]\mathrm{d}\rho_t(\mathbf{x}) &= \int\left[\frac{\partial}{\partial t}f_t(\mathbf{x})\right]\left[\frac{\partial}{\partial t}\frac{\rho_t}{\pi}(\mathbf{x})\right]\mathrm{d}\pi(\mathbf{x}) \\ &= \int\left[\frac{\partial}{\partial t}f_t(\mathbf{x})\right]\mathscr{L}\frac{\rho_t}{\pi}(\mathbf{x})\,\mathrm{d}\pi(\mathbf{x}) \\ &= \int\mathscr{L}\left(\frac{\partial}{\partial t}f_t\right)(\mathbf{x})\,\mathrm{d}\rho_t(\mathbf{x}) \\ &= \mathbb{E}_{\mathbf{X}_0\sim\rho_0}\left[P_t\mathscr{L}\left(\frac{\partial}{\partial t}f_t\right)(\mathbf{X}_0)\right].\end{aligned}$$

Therefore, matching the time derivative similarly offers a perspective analogous to the implicit score matching derived via OISM (20). Time score matching and score matching optimize the temporal and spatial partial derivatives of the forward process, respectively. We illustrate that the process of removing intractability from the original mean squared error form follows a similar pattern. We also obtain a notably different solution compared to Choi et al. (2022), which does not make use of the density ratio PDE (27).

Generalizing the score Fokker–Planck equation Sharing a similar starting point to this work, Lai et al. (2023) noted that traditional score matching is agnostic to the specific type of the forward process. The authors reasoned that the true score function ought to satisfy a PDE and derived an explicit PDE in the case of a Markov diffusion forward process. Here we present a streamlined and more general derivation of that result. The diffusion property (37) enables us to study the time evolution of the log-density ratio ρ_t/π, and we have that:

$$\begin{aligned}\frac{\partial}{\partial t}\log\frac{\rho_t}{\pi}(\mathbf{x}) &= \frac{\frac{\partial}{\partial t}P_t\frac{\rho_0}{\pi}(\mathbf{x})}{\frac{\rho_t}{\pi}(\mathbf{x})}\\ &= \log'\left(\frac{\rho_t}{\pi}(\mathbf{x})\right)\mathscr{L}\frac{\rho_t}{\pi}(\mathbf{x})\\ &= \mathscr{L}\log\frac{\rho_t}{\pi}(\mathbf{x}) + \left\|\nabla_{\mathbf{x}}\log\frac{\rho_t}{\pi}(\mathbf{x})\right\|^2,\end{aligned} \tag{38}$$

where the final equality follows from (37) with $\psi(f) = \log f$ and the explicit form of the carré-du-champ operator for the OU process in Example 5. The time derivative of the true score function of the density ratio $\mathbf{s}_t^\star(\mathbf{x}) := \nabla\log\frac{\rho_t}{\pi}(\mathbf{x})$ in (12) can be obtained by taking the $\mathbf{x}$ derivative of (38), leading in the case of the OU process to the PDE:

$$\frac{\partial}{\partial t}\mathbf{s}_t^\star(\mathbf{x}) = \nabla_{\mathbf{x}}\left[\langle -\mathbf{x}, \mathbf{s}_t^\star(\mathbf{x})\rangle + \operatorname{tr}(\mathbf{J}(\mathbf{s}_t^\star)(\mathbf{x}))\right] + 2\mathbf{J}(\mathbf{s}_t^\star)(\mathbf{x})\mathbf{s}_t^\star(\mathbf{x}), \tag{39}$$

where $\mathbf{J}(\mathbf{s}_t^\star)$ denotes the Jacobean of $\mathbf{s}_t^\star$, which recovers the *score Fokker–Planck equation* as coined by Lai et al. (2023). In that work the authors noted that standard score matching typically violates (39), especially for small values of t where the regularity of ρ_t/π is most pronounced (c.f. Fig. 1 in Lai et al., 2023). Based on this observation, the authors proposed to add a regularization term to the score matching objective that aims to explicitly enforce (39) to (approximately) hold. The derivation that we have presented allows to generalize beyond the OU process by substituting alternative expressions for the infinitesimal generator and the carré-du-champ operator.

Generalized score matching Benton et al. (2024) propose a *generalized* score matching objective that does not include a L^2 norm and can thus be used for non-euclidean data. The intuition behind it relates to the *diffusion* property (37). With $\psi(f) = \log f$, we can rewrite (37) as:

$$\Gamma(\log f) = \frac{\mathscr{L}f}{f} - \mathscr{L}\log f.$$

Therefore, as we match the score function of ρ_T and an energy-based model $\nu_t(\mathbf{x}) = \exp(-\varphi_t(\mathbf{x}))/Z_t$ of ρ_t, rewriting the carré-du-champ operator yields the generalized form of *explicit* score matching:

$$\mathcal{J}_{\mathrm{ESM}} := \mathbb{E}_{\rho_t}\left[\Gamma(\log\rho_t/\nu_t)\right] = \mathbb{E}_{\rho_t}\left[\frac{\mathscr{L}\frac{\rho_t}{\nu_t}}{\frac{\rho_t}{\nu_t}} - \mathscr{L}\log\frac{\rho_t}{\nu_t}\right].$$

Additionally, using again the diffusion property (37), and upon the identification $\nu_t(\mathbf{x}) \propto \pi(\mathbf{x})\exp(f_t(\mathbf{x}))$, where f_t is defined in (16), the implicit score matching $\mathcal{J}_{\mathrm{ISM}}$ proposed by Benton et al. (2024) is equivalent to our implicit score matching (18),

$$\begin{aligned}\mathcal{J}_{\mathrm{ISM}} &:= \mathbb{E}_{\rho_t}\left[\frac{\mathscr{L}\nu_t}{\nu_t} + \mathscr{L}\log\nu_t\right]\\ &= \mathbb{E}_{\rho_t}\left[\Gamma(\log\nu_t) + 2\mathscr{L}\log\nu_t\right]\\ &= \mathbb{E}_{\rho_0}\left[P_t\Gamma(\log\nu_t) + 2P_t\mathscr{L}\log\nu_t\right]\\ &= \text{integrand of (18).}\end{aligned}$$

Benton et al. (2024) focus on the appeal of generator-only formula for score matching, and apply it to diffusion models defined on manifolds.