

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Diffusion models for distribution learning: a statistical perspective

Jin Su, Shuyi Zhang, Yong Zhou & Jian Huang

To cite this article: Jin Su, Shuyi Zhang, Yong Zhou & Jian Huang (24 Nov 2026): Diffusion models for distribution learning: a statistical perspective, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2762679

To link to this article: https://doi.org/10.1080/24754269.2026.2762679

Published online: 24 Nov 2026.

Submit your article to this journal

Article views: 20

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2762679

Taylor & Francis
by informa

OPEN ACCESS Check for updates

Diffusion models for distribution learning: a statistical perspective

Jin Su [a], Shuyi Zhang [b], Yong Zhou [b] and Jian Huang [a,c]

[a]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong SAR, People's Republic of China; [b]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [c]Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR, People's Republic of China

ABSTRACT

Diffusion models have emerged as a cornerstone of generative learning, which seeks to recover an unknown target distribution from finite empirical samples and synthesize novel observations. These models achieve this by estimating the Stein score functions of a family of progressively perturbed marginal distributions, subsequently generating samples via reverse-time stochastic or deterministic dynamics. This review introduces the mathematical foundations of diffusion models and systematically synthesizes recent theoretical guarantees concerning the statistical discrepancy between the generated and target distributions. In particular, we examine how these guarantees are shaped by the interplay of finite-sample score estimation errors, the discretization and truncation of practical samplers, and the structural assumptions imposed on the underlying data distribution. Finally, we discuss extensions of this framework to conditional distribution estimation.

ARTICLE HISTORY
Received 31 July 2026
Revised 28 September 2026
Accepted 7 November 2026

KEYWORDS
Diffusion models; distribution learning; generative learning; score estimation; reverse-time sampling; conditional generation

1. Introduction

Recent advances in text generation, code generation, image synthesis, and scientific data generation have brought considerable attention to generative learning. Large language models can generate text (Brown et al., 2020), and code generation models can synthesize programs from natural-language descriptions (M. Chen et al., 2021). Text-to-image models, including latent diffusion models (Rombach et al., 2022) and Stable Diffusion 3 (Esser et al., 2024), can synthesize images from text prompts. Applications in science include the generation of molecular structures, protein structures, and precipitation fields (Hoogeboom et al., 2022; Ravuri et al., 2021; Watson et al., 2023). Although these tasks concern different types of data, they share a common statistical objective: to learn the data-generating distribution from a finite training sample and to generate new samples from the learned distribution. To formalize this objective, suppose that

$$X^{(1)}, \ldots, X^{(n)} \overset{\text{i.i.d.}}{\sim} P_0,$$

CONTACT Jian Huang j.huang@polyu.edu.hk Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR, People's Republic of China

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

where P_0 is the data-generating distribution. Given this training sample, the aim is to construct a distribution $\widehat{P}$ that approximates P_0, and then use $\widehat{P}$ to generate new samples. From this perspective, generative learning is a problem of distribution learning.

Within this distribution-learning framework, generative modelling approaches differ in how they represent distributions, estimate them from finite training data, and draw samples from the fitted models (Bond-Taylor et al., 2022; Ruthotto & Haber, 2021). Generative adversarial networks (GANs) represent $\widehat{P}$ implicitly as the distribution induced by a generator that maps latent variables to the data space. They fit the generator by adversarial training and generate samples by applying it to new draws from the latent distribution (Goodfellow et al., 2014; Mohamed & Lakshminarayanan, 2016). Normalizing flows define an explicit density by transforming a simple base distribution through a sequence of bijections with tractable Jacobian determinants (Rezende & Mohamed, 2015). Their parameters are commonly estimated by maximum likelihood, and samples are generated by transforming draws from the base distribution (Dinh et al., 2017; Papamakarios et al., 2021). Variational autoencoders (VAEs) specify latent-variable models and estimate the model and variational parameters by maximizing an evidence lower bound on the marginal log-likelihood. To generate data, they draw latent variables from the prior and then sample from the decoder distribution (Kingma & Welling, 2014). Energy-based models represent a distribution through an energy function, which determines its density up to a normalizing constant. They can be fitted using likelihood-based, score-matching, or related criteria, while sampling typically requires Markov chain Monte Carlo (Du & Mordatch, 2019; Grathwohl et al., 2020).

Diffusion-based frameworks, encompassing continuous-time score-based generative models (SGMs) and discrete-time denoising diffusion probabilistic models (DDPMs), have emerged as a dominant paradigm in generative modelling, successfully leveraging stochastic dynamics to bypass the optimization instabilities of GANs and the representation bottlenecks of VAEs. Instead of directly parameterizing a highly complex target density or solving a challenging minimax game, diffusion models represent the target distribution P_0 indirectly. They achieve this by estimating the Stein score function $\nabla_x \log p_t(x)$, the spatial gradient of the log marginal density, across a continuous family of noisy marginal distributions perturbed by a forward-time stochastic differential equation (SDE) (Hyvärinen, 2005; Song et al., 2021). Alternatively, this estimation procedure can be framed as learning optimal denoisers that are fundamentally linked to the score function via Tweedie's formula (Efron, 2011).

Generative sampling is subsequently formulated as an integration problem, wherein new samples are synthesized by simulating the reverse-time stochastic or deterministic trajectories starting from a tractable reference prior (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2021). Empirically, this paradigm has demonstrated unprecedented sample fidelity and mode coverage in high-dimensional continuous domains, outperforming GANs in image synthesis (Dhariwal & Nichol, 2021), while driving transformative breakthroughs in audio generation (Kong et al., 2021), text-to-image synthesis (Rombach et al., 2022), and structural biology (Watson et al., 2023).

Alongside these empirical successes, a growing body of theory provides guarantees for learning and sampling with diffusion models (M. Chen et al., 2023; Chen, Chewi, et al., 2023; Oko et al., 2023; Tang & Yang, 2024; Zhang et al., 2024). These guarantees are stated in different distributional metrics and under different assumptions on the target distribution, making their scope and relationships important to clarify.

Existing surveys and tutorials examine diffusion models from several complementary perspectives. Broad reviews cover model formulations, algorithmic developments, and

applications across domains and within computer vision (Ahsan et al., 2025; Cao et al., 2024; Croitoru et al., 2023; Yang et al., 2024). Other works emphasize component-wise design choices, training and sampling methods, and practical implementation (Chang et al., 2026; Ding & Jin, 2024; Yeğin & Amasyalı, 2024), or explain the mathematical foundations and connections among different generative formulations (Chan, 2024; Higham et al., 2025; Holderrieth & Erives, 2025; Lai et al., 2025; Luo, 2022; Nakkiran et al., 2025).

Several reviews and tutorials focus more specifically on the mathematical and statistical theory of diffusion models. Tang and Zhao (2025) discuss score matching and sampling guarantees within the SDE framework, while Lu (2026) develops a sampling-centred treatment of diffusion models and their error analysis. M. Chen et al. (2024) review score approximation, statistical estimation, and distribution recovery in both unconditional and conditional settings. Aamari and Stéphanovitch (2026) connect learned Markovian dynamics to finite-sample distributional guarantees, while Suh and Cheng (2025) discuss diffusion models within a broader review of deep learning theory. Dedicated surveys further examine conditioning mechanisms and controllable image generation (Cao et al., 2026; Zhan et al., 2025), as well as reward-guided generation at inference time (Uehara et al., 2025).

Building on these reviews, we adopt a distribution-learning perspective, viewing diffusion models as methods for learning an unknown target distribution P_0 from a finite sample. We organize the discussion around the discrepancy between the resulting generated distribution $\widehat{P}$ and P_0, examining how training objectives, score estimation, and reverse-time sampling jointly shape this discrepancy. In particular, we connect score approximation and statistical estimation errors with initialization, early-stopping, and discretization errors, and examine how distributional guarantees depend on the chosen metric and the structure of the target distribution. The value of this organization lies in helping readers relate theoretical results across the learning and sampling stages and understand how their assumptions and error bounds translate into guarantees for the generated distribution. We extend this perspective to conditional distribution learning, distinguishing recovery of the target conditional distribution from the distributional effects of guidance.

The remainder of this review is organized as follows. Section 2 reviews forward and reverse diffusion dynamics, probability-law evolution, and score estimation. Section 3 introduces diffusion probabilistic models and practical sampling procedures. Section 4 explains the population targets of variance-preserving (VP) diffusion training and the relations among them. Section 5 examines sources of distributional error and reviews statistical guarantees for the generated distribution. Section 6 considers conditional distribution learning and its statistical theory. Section 7 discusses open problems and future research directions.

Throughout, I_d denotes the $d \times d$ identity matrix, $\mathbb{E}_P$ denotes expectation under a probability law P, and, unless a subscript specifies otherwise, $\|\cdot\|$ denotes the Euclidean norm.

2. From diffusion processes to generative modelling

Diffusion models transform the data distribution into a tractable reference distribution through a forward noising process and generate samples by reversing this evolution. The reverse-time dynamics depend on the marginal score, which must be estimated from data. This section reviews the diffusion dynamics, probability-law evolution, score estimation, and sampling procedures underlying this construction.

2.1. Diffusion processes: forward and reverse dynamics

Early research in statistical physics investigated how microscopic random motion gives rise to macroscopic transport phenomena. Brownian motion provided the first physical system for this inquiry, characterizing how suspended particles trace highly irregular, non-differentiable trajectories in a fluid. By mathematically linking these microscopic fluctuations to the macroscopic diffusion equation, Einstein (1905), Einstein (1926) and Smoluchowski (1906) provided a probabilistic description of how the displacement distribution of particles evolves over time. This foundational insight decoupled the stochasticity of individual physical trajectories from the deterministic, analytical evolution of their global probability laws.

To describe these dynamics, Langevin (1908) introduced a Langevin equation for a Brownian particle, representing its motion as a balance between a deterministic frictional force and a rapidly fluctuating stochastic force representing the cumulative impact of molecular collisions. This formulation placed deterministic and random components within a unified dynamical equation. Subsequent mathematical formalization led to the rigorous construction of the stochastic integral by Itô (1944) and Feller's analysis of one-dimensional diffusion processes and boundary behaviour (Feller, 1954). In modern stochastic differential equation (SDE) notation, the deterministic component is specified by a drift vector field, while the stochastic component is driven by Brownian motion and scaled by a diffusion coefficient.

2.1.1. The forward noising process

The fundamental mechanism underlying modern diffusion models is a continuous-time forward noising process, which systematically corrupts a complex, structured data distribution into an analytically tractable prior. Let $X_0 \in \mathbb{R}^d$ be a random vector distributed according to the unknown data distribution P_0. We define the forward noising process as a continuous-time diffusion process $\{X_t\}_{t\in[0,T]}$ indexed over a finite time horizon $T > 0$. Under mild regularity assumptions on the drift and diffusion coefficients (e.g., Lipschitz continuity and bounded growth), this process is uniquely governed by the following Itô SDE (Øksendal, 2003):

$$\mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t, \quad t \in [0, T], \tag{1}$$

where $\{W_t\}_{t\geq 0}$ is a d-dimensional standard Brownian motion, $f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ represents the drift function, and $g(t) \in \mathbb{R}_+$ is a scalar-valued diffusion coefficient. We denote the marginal probability distribution of X_t by P_t, and its corresponding density (when it exists with respect to the Lebesgue measure) by $p_t(x)$.

The coefficients of the forward SDE (1) are carefully engineered to drive the initial distribution P_0 toward a pre-specified, computationally tractable reference prior distribution π_{ref} (typically a standard isotropic Gaussian $\mathcal{N}(0, I_d)$) at the terminal time T. We write $p_{t|0}(\cdot \mid x_0)$ for the density of $P_{t|0}(\cdot \mid x_0)$ whenever this density exists. Crucially, when the drift vector field $f(x, t)$ is chosen to be affine in x, these transition kernels are Gaussian, enabling exact closed-form evaluation and highly efficient training via denoising score matching (Song et al., 2021).

Important special case: variance-preserving diffusion. A mathematically and practically prominent instantiation of the forward SDE (1) is the *variance-preserving (VP) SDE* (Song et al., 2021). This continuous-time formulation corresponds to the infinitesimal limit of the popular Denoising Diffusion Probabilistic Models (DDPM) framework (Ho et al., 2020; Sohl-Dickstein et al., 2015). Let $\beta(t) \in \mathbb{R}_+$ denote a continuous, positive, non-decreasing

noise schedule over $[0, T]$. The VP SDE is defined by setting $f(X_t, t) = -\frac{1}{2}\beta(t)X_t$ and $g(t) = \sqrt{\beta(t)}$:

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

The linear drift term $f(X_t, t) = -\frac{1}{2}\beta(t)X_t$ acts as a mean-reverting force pulling the process toward the origin, which exactly counterbalances the expansionary effect of the stochastic diffusion term $g(t)\mathrm{d}W_t$. This unique balance ensures that if the initial data has unit variance, the variance of the marginal distribution P_t remains bounded and stable across the entire trajectory (hence, 'variance-preserving').

Applying Itô's formula to (2), we obtain the conditional transition density $p_{t|0}(x_t \mid x_0)$ in closed form:

$$p_{t|0}(x_t \mid x_0) = \mathcal{N}\left(x_t; \mathrm{e}^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}x_0, \left(1 - \mathrm{e}^{-\int_0^t \beta(s)\mathrm{d}s}\right) I_d\right).$$

Defining the scaling factor $\bar{\alpha}_t = \mathrm{e}^{-\int_0^t \beta(s)\mathrm{d}s} \in (0, 1]$, we can express this transition kernel more compactly as

$$X_t \mid X_0 \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}X_0, (1 - \bar{\alpha}_t)I_d). \tag{3}$$

As $t \to \infty$ (or for a sufficiently large time horizon T such that $\bar{\alpha}_T \approx 0$), the transition density $p_{T|0}(x_T \mid x_0)$ converges to $\mathcal{N}(0, I_d)$ independently of the starting point x_0. Consequently, for a sufficiently large time horizon T, the terminal marginal distribution P_T is close to the standard Gaussian distribution, providing a tractable approximate starting distribution for the generative reverse-time sampler.

2.1.2. The reverse-time generative process

Generative sampling requires the construction of a reverse-time process whose marginal distributions trace the forward noising family in reverse order, originating at the tractable prior $P_T \approx \pi_{\mathrm{ref}}$ and terminating at the target data distribution P_0. Because the forward process continuously injects Brownian noise, the terminal state X_T does not uniquely determine earlier states. Therefore, the required reverse dynamics cannot be obtained by deterministically inverting individual forward sample paths.

Under suitable regularity and boundary conditions, Anderson (1982) demonstrated that the time reversal of an Itô diffusion process is itself an Itô diffusion process. Crucially, its reverse-time drift depends on the Stein score function $s(x, t) = \nabla_x \log p_t(x)$ of the forward marginal density p_t.

Formally, we define the time-reversed process $Y_r = X_{T-r}$ for $r \in [0, T]$. If the initial state is distributed as $Y_0 \sim P_T$, the corresponding reverse-time process satisfies the following SDE (Anderson, 1982):

$$\mathrm{d}Y_r = [-f(Y_r, T - r) + g(T - r)^2 s(Y_r, T - r)]\mathrm{d}r + g(T - r)\,\mathrm{d}\overline{W}_r, \tag{4}$$

where $\{\overline{W}_r\}_{r\in[0,T]}$ is a standard Brownian motion operating in reverse time. At each reverse time $r \in [0, T]$, the marginal distribution of Y_r matches P_{T-r} exactly. Haussmann and Pardoux (1986) subsequently extended these time-reversal results to diffusion processes under more general regularity conditions.

2.2. Evolution of probability laws and dual representation

While the forward SDE models the individual, stochastic sample paths of the process $\{X_t\}_{t\geq 0}$, its global distributional dynamics are governed by partial differential equations (PDEs).

2.2.1. The forward Fokker–Planck equation

Under the assumption that the transition kernels are absolutely continuous with respect to the Lebesgue measure, the marginal probability density $p_t(x)$ evolves according to the Fokker–Planck (or Kolmogorov forward) equation (Fokker, 1914; Kolmogoroff, 1931; Risken, 1996):

$$\partial_t p_t(x) = -\nabla \cdot \{f(x,t)p_t(x)\} + \frac{1}{2}g(t)^2 \Delta p_t(x), \tag{5}$$

where $\nabla = (\partial/\partial x_1, \ldots, \partial/\partial x_d)^\top$ denotes the spatial gradient operator, $\nabla\cdot$ is the divergence operator, and $\Delta = \sum_{j=1}^{d} \partial^2/\partial x_j^2$ is the Laplacian. In this Eulerian formulation, the first-order advection term transports probability mass along the deterministic drift vector field $f(x,t)$, whereas the second-order diffusion term models isotropic smoothing driven by the scalar-valued diffusion coefficient $g(t)$. The SDE (Lagrangian perspective) and the Fokker–Planck PDE (Eulerian perspective) provide dual representations of the same underlying diffusion process (Øksendal, 2003).

2.2.2. The reverse Fokker–Planck equation

To characterize the marginal family in reverse time, we define the time-reversed density $q_r(x) = p_{T-r}(x)$ for $r \in [0,T]$. Standard results in stochastic regularization and time-reversal (Anderson, 1982) show that the reverse-time SDE corresponds to the following Fokker–Planck equation:

$$\partial_r q_r(x) = -\nabla \cdot ([-f(x,T-r) + g(T-r)^2 s(x,T-r)]q_r(x)) + \frac{1}{2}g(T-r)^2 \Delta q_r(x), \tag{6}$$

where $s(x,t) = \nabla_x \log p_t(x)$ is the Stein score function. By leveraging the identity $s(x,T-r)q_r(x) = \nabla_x q_r(x)$ and expanding the spatial divergence, we observe a crucial simplification:

$$\begin{aligned}\nabla \cdot (g(T-r)^2 s(x,T-r)q_r(x)) &= g(T-r)^2 \nabla \cdot (\nabla_x q_r(x)) \\ &= g(T-r)^2 \Delta q_r(x).\end{aligned}$$

Substituting the above identity for the divergence of the score term into Equation (6) yields

$$\partial_r q_r(x) = \nabla \cdot \{f(x,T-r)q_r(x)\} - \frac{1}{2}g(T-r)^2 \Delta q_r(x) = -\left.\partial_t p_t(x)\right|_{t=T-r}.$$

The resulting equality with $-\partial_t p_t(x)|_{t=T-r}$ demonstrates that the reverse-time SDE precisely traces the forward marginal densities in reverse temporal order, establishing the mathematical validity of score-based generative modelling.

2.3. Probability-flow ordinary differential equation

The Fokker–Planck representation allows us to construct deterministic dynamics with the same marginal densities $p_t(x)$ as the forward SDE. Using the score identity $\Delta p_t = \nabla \cdot$

$\{p_t \nabla_x \log p_t\}$, we can rewrite the second-order Fokker–Planck Equation (5) as a first-order continuity equation (Villani, 2009):

$$\partial_t p_t(x) = -\nabla \cdot \left(\left[f(x,t) - \frac{1}{2} g(t)^2 s(x,t) \right] p_t(x) \right).$$

This first-order PDE governs the transport of probability mass associated with the deterministic probability-flow ordinary differential equation (PF-ODE) (Song et al., 2021):

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = f(Z_t, t) - \frac{1}{2} g(t)^2 s(Z_t, t). \tag{7}$$

Under mild regularity conditions on the vector field and the score, if $Z_0 \sim P_0$, the solution of (7) satisfies $Z_t \sim P_t$ for all $t \in [0, T]$.

The corresponding time-reversed deterministic dynamics can be formulated by defining the reversed trajectory $\widetilde{Z}_r = Z_{T-r}$. Applying the chain rule directly yields the reverse-time probability-flow ODE:

$$\frac{\mathrm{d}\widetilde{Z}_r}{\mathrm{d}r} = -f(\widetilde{Z}_r, T-r) + \frac{1}{2} g(T-r)^2 s(\widetilde{Z}_r, T-r).$$

Because the marginal mapping is preserved, initializing $\widetilde{Z}_0 = Z_T \sim P_T$ ensures that $\widetilde{Z}_r \sim P_{T-r}$ for all $r \in [0, T]$. Consequently, both the reverse-time SDE and the reverse-time PF-ODE trace the same family of marginal distributions. The stochastic framework admits predictor–corrector samplers that combine numerical integration of the reverse-time SDE with Langevin MCMC corrections, whereas the deterministic PF-ODE supports fast adaptive sampling, exact likelihood evaluation via the instantaneous change-of-variables formula, and manipulation of latent representations (R. T. Q. Chen et al., 2018; Grathwohl et al., 2019; Song et al., 2021).

2.4. Explicit score matching and denoising score matching

A key challenge in implementing the reverse-time dynamics (or the probability-flow ODE) is that the marginal Stein score function $s(x,t) = \nabla_x \log p_t(x)$ depends on the unknown data distribution P_0 through the integral representation $p_t(x) = \int p_{t|0}(x \mid x_0) \mathrm{d}P_0(x_0)$. In practice, we only have access to a finite training sample $\mathcal{D}_n = \{X^{(i)}\}_{i=1}^n$, with $X^{(i)} \overset{\text{i.i.d.}}{\sim} P_0$. Consequently, we must utilize score matching techniques to learn a parameterized vector field $s_\theta(x,t)$ (typically represented by a deep neural network) that approximates $s(x,t)$.

2.4.1. Explicit score matching

Hyvärinen (2005) introduced score matching as an elegant framework for fitting non-normalized probability density models without calculating their partition functions. For a fixed time $t > 0$, let $u(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ be a candidate vector field in a function space $\mathcal{H}$. The population explicit score-matching (ESM) objective is defined as half the squared $L^2(P_t)$ distance to the ground-truth score:

$$J_{\mathrm{ESM}}(u; t) = \frac{1}{2} \mathbb{E}_{X_t \sim P_t} [\| u(X_t, t) - s(X_t, t) \|^2]. \tag{8}$$

Since $s(x,t)$ is unavailable, the objective (8) cannot be evaluated directly. However, assuming that $p_t(x)$ is continuously differentiable, and that $u(x,t)$ satisfies the mild boundary condition

$\lim_{\|x\|\to\infty} p_t(x)u(x,t) = 0$, multivariate integration by parts yields an equivalent, tractable objective:

$$J_{\mathrm{ESM}}(u;t) = \mathbb{E}_{X_t\sim P_t}\left[\frac{1}{2}\|u(X_t,t)\|^2 + \nabla_x \cdot u(X_t,t)\right] + \mathrm{const}(P_t), \tag{9}$$

where $\mathrm{const}(P_t) = \frac{1}{2}\mathbb{E}_{P_t}[\|s(X_t,t)\|^2]$ is a constant independent of the parameterized vector field u.

By replacing the expectation with an empirical average over samples $X_t^{(i)}$ generated by running the forward noising process on training data, we can optimize (9) without knowing $p_t(x)$. Despite its elegance, calculating the divergence $\nabla_x \cdot u = \sum_{j=1}^{d} \partial u_j/\partial x_j$ is computationally prohibitive in high dimensions, as standard backpropagation requires d separate backward passes. While stochastic estimators such as Hutchinson's trick (Hutchinson, 1989) or sliced score matching (Song et al., 2020) reduce this cost, explicit score matching remains challenging to scale.

2.4.2. Denoising score matching

To circumvent the computational bottleneck of evaluating the spatial divergence of neural networks, Vincent (2011) introduced denoising score matching (DSM). Rather than directly matching the marginal score, DSM leverages the tractability of the transition kernel $p_{t|0}(x_t \mid x_0)$ of the forward SDE. For the linear SDEs typical of diffusion models, this conditional transition density is Gaussian. With mean and covariance $\mu_t(x_0)$ and Σ_t, respectively, this Gaussian transition density has the following conditional score:

$$\nabla_{x_t} \log p_{t|0}(x_t \mid x_0) = -\Sigma_t^{-1}(x_t - \mu_t(x_0)).$$

The population DSM objective is defined as

$$J_{\mathrm{DSM}}(u;t) = \frac{1}{2}\mathbb{E}_{(X_0,X_t)\sim P_{0,t}}[\|u(X_t,t) - \nabla_{x_t} \log p_{t|0}(X_t \mid X_0)\|^2], \tag{10}$$

where the expectation is taken over the joint distribution $P_{0,t}(\mathrm{d}x_0, \mathrm{d}x_t) = P_0(\mathrm{d}x_0)p_{t|0}(x_t \mid x_0)\mathrm{d}x_t$.

The theoretical foundation linking the conditional score optimization in (10) to the marginal score $s(x,t)$ is established by the following critical conditional expectation identity. Under standard regularity conditions permitting the exchange of integration and differentiation, the marginal score is the orthogonal projection of the conditional score onto the space of functions of X_t:

$$\mathbb{E}[\nabla_{x_t} \log p_{t|0}(X_t \mid X_0) \mid X_t = x] = s(x,t).$$

This identity can be shown as follows. Using the definition of conditional expectation and Leibniz's integral rule, with posterior kernel $P_{0|t}(\mathrm{d}x_0 \mid x) = \{p_{t|0}(x \mid x_0)/p_t(x)\}P_0(\mathrm{d}x_0)$, we have

$$\begin{aligned}\mathbb{E}[\nabla_{x_t} \log p_{t|0}(X_t \mid X_0) \mid X_t = x] &= \int (\nabla_x \log p_{t|0}(x \mid x_0))P_{0|t}(\mathrm{d}x_0 \mid x) \\ &= \int \left(\frac{\nabla_x p_{t|0}(x \mid x_0)}{p_{t|0}(x \mid x_0)}\right)\frac{p_{t|0}(x \mid x_0)}{p_t(x)}P_0(\mathrm{d}x_0)\end{aligned}$$

$$= \frac{1}{p_t(x)} \nabla_x \int p_{t|0}(x \mid x_0)\, P_0(\mathrm{d}x_0)$$
$$= \frac{\nabla_x p_t(x)}{p_t(x)} = s(x,t).$$

By the law of total expectation and the projection property of L^2 spaces, any minimizer $u^*(\cdot, t)$ of the population DSM objective $J_{\mathrm{DSM}}(u;t)$ over an unrestricted function class must satisfy $u^*(x,t) = s(x,t)$ almost everywhere.

By parameterizing the score function $s_\theta(x,t)$ using a deep neural network and integrating the DSM objective across multiple noise levels with a positive weighting function $w(t) > 0$, we obtain the continuous-time denoising score-matching objective (Song et al., 2021):

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \int_0^T w(t) J_{\mathrm{DSM}}(s_\theta; t)\, \mathrm{d}t.$$

This formulation underpins the highly successful training protocols of modern score-based generative and diffusion models.

3. Diffusion probabilistic models and practical samplers

Having described continuous-time diffusion dynamics and score estimation, we next turn to finite-step sampling. The denoising diffusion probabilistic model (DDPM) is a natural starting point because its Gaussian construction links noise prediction to a score estimate that determines each discrete reverse transition. We then consider alternative discrete-time samplers and numerical methods for continuous-time reverse dynamics.

Diffusion probabilistic models were introduced by Sohl-Dickstein et al. (2015) as finite-step generative models with a fixed forward noising chain and a learned reverse chain. Building on this construction, Ho et al. (2020) introduced the DDPM formulation and established its connection to denoising score matching. Let K denote the number of diffusion steps, and let $\alpha_1, \ldots, \alpha_K \in (0,1)$ denote the coefficients of the Gaussian forward transitions. A DDPM uses

$$q(x_k \mid x_{k-1}) = \mathcal{N}\left(\sqrt{\alpha_k} x_{k-1}, (1-\alpha_k) I_d\right).$$

With $\bar{\alpha}_k = \prod_{j=1}^k \alpha_j$, the resulting forward marginal admits the representation

$$X_k = \sqrt{\bar{\alpha}_k} X_0 + \sqrt{1-\bar{\alpha}_k}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_d).$$

A time-conditioned neural network $\varepsilon_\theta(x_k, k)$ is trained to predict the injected noise. A commonly used simplified objective is

$$\mathbb{E}_{k, X_0, \varepsilon}[\|\varepsilon - \varepsilon_\theta(X_k, k)\|^2],$$

where k is sampled from $\{1, \ldots, K\}$. For Gaussian forward transitions, noise prediction is an equivalent parameterization of score prediction at the population level. In particular, the

noise predictor induces the corresponding score model:

$$s_\theta(x_k, k) = -\frac{\varepsilon_\theta(x_k, k)}{\sqrt{1-\bar{\alpha}_k}}.$$

We write the learned Gaussian reverse transition as

$$p_\theta(x_{k-1} \mid x_k) = \mathcal{N}(\mu_\theta(x_k, k), \sigma_k^2 I_d).$$

Its usual noise-parameterized mean can equivalently be expressed as

$$\mu_\theta(x_k, k) = \frac{1}{\sqrt{\alpha_k}}[x_k + (1-\alpha_k)s_\theta(x_k, k)].$$

The standard DDPM ancestral sampler therefore iterates

$$X_{k-1} = \mu_\theta(X_k, k) + \sigma_k Z_k, \quad Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d).$$

Starting from $X_K \sim \mathcal{N}(0, I_d)$, this update is applied for $k = K, \ldots, 1$ to obtain X_0. Because the learned reverse mean μ_θ is parameterized through s_θ, a DDPM can be viewed as a discrete-time score-based model.

The DDPM ancestral sampler is stochastic because it injects Gaussian noise during reverse sampling. The same training objective and noise predictor can also be used to construct denoising diffusion implicit models (DDIMs), which are derived from non-Markovian forward processes (Song et al., 2021). Define the predicted clean sample by

$$\widehat{x}_{0,\theta}(x_k, k) = \frac{x_k - \sqrt{1-\bar{\alpha}_k}\,\varepsilon_\theta(x_k, k)}{\sqrt{\bar{\alpha}_k}}.$$

The deterministic DDIM update between adjacent noise levels is

$$X_{k-1} = \sqrt{\bar{\alpha}_{k-1}}\,\widehat{x}_{0,\theta}(X_k, k) + \sqrt{1-\bar{\alpha}_{k-1}}\,\varepsilon_\theta(X_k, k).$$

Because this update introduces no additional Gaussian noise, the sampling trajectory is deterministic conditional on X_K. The deterministic DDIM update can also be applied on a subsequence of noise levels to reduce the number of sampling steps.

DDPM and DDIM specify finite-step sampling recursions directly. We next consider practical sampling from the continuous-time reverse SDE in Equation (4), which requires numerical discretization. Let $\widehat{s}(x, t)$ denote a time-dependent score estimator obtained from the training data. In practice, for a small $\tau > 0$, the reverse-time SDE is run over $0 \le r \le T - \tau$, with initial distribution π_{ref}, typically $\mathcal{N}(0, I_d)$, and estimated score $\widehat{s}$:

$$\mathrm{d}\widehat{Y}_r = [-f(\widehat{Y}_r, T-r) + g(T-r)^2\widehat{s}(\widehat{Y}_r, T-r)]\mathrm{d}r + g(T-r)\,\mathrm{d}\overline{W}_r, \quad \widehat{Y}_0 \sim \pi_{\text{ref}}. \tag{11}$$

We discretize Equation (11) using Euler–Maruyama. For a reverse-time grid $0 = r_0 < r_1 < \cdots < r_L = T - \tau$, let $t_\ell = T - r_\ell$, $\Delta r_\ell = r_{\ell+1} - r_\ell$, and $h = \max_{0 \le \ell < L} \Delta r_\ell$. The resulting update is

$$\widehat{Y}_{\ell+1} = \widehat{Y}_\ell + \left[-f(\widehat{Y}_\ell, t_\ell) + g(t_\ell)^2\widehat{s}(\widehat{Y}_\ell, t_\ell)\right]\Delta r_\ell + g(t_\ell)\sqrt{\Delta r_\ell}\,Z_\ell, \tag{12}$$

where $Z_\ell \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. Let $\widehat{P}$ denote the distribution of the sampler output $\widehat{Y}_L$.

Beyond Euler–Maruyama, reverse-SDE sampling can use predictor–corrector schemes (Song et al., 2021) and higher-order stochastic solvers such as SA-solver and SEEDS (Gonzalez et al., 2023; Xue et al., 2023). Stochastic score-based sampling also encompasses annealed Langevin dynamics, which performs discretized Langevin updates at successively lower noise levels rather than integrating the reverse-time SDE (Song & Ermon, 2019). For deterministic continuous-time sampling, the probability-flow ODE in Section 2.2 can instead be integrated using standard methods such as Euler, Heun, and Runge–Kutta schemes, or diffusion-specific solvers such as PNDM, DEIS, DPM-Solver, and UniPC (Liu et al., 2022; Lu et al., 2022; Zhang & Chen, 2023; Zhao et al., 2023).

4. Population targets of VP-diffusion training

Sections 2 and 3 show that the forward noising process induces a family of marginal distributions P_t and that the reverse dynamics depend on their scores $s(x,t) = \nabla_x \log p_t(x)$. A key statistical problem is to use the sample $\mathcal{D}_n = \{X^{(i)}\}_{i=1}^n$ to estimate the score required by the reverse process.

In this section, we focus on the VP diffusion model (2). Section 5 also reviews results under other forward-noising processes, as specified below. Common training objectives use different prediction targets. Score-based models estimate the score directly, DDPMs predict the noise ε introduced by the forward process, and some diffusion models instead use X_0 as the regression target (Karras et al., 2022; Li & He, 2026; Tevet et al., 2023). Although these targets differ, their population minimizers under Gaussian noising are equivalent up to known transformations.

Given a clean sample $X_0 \sim P_0$, for $t \in (0,T]$, let $a_t \in \mathbb{R}$ and $\sigma_t > 0$ be deterministic scalar noising coefficients. For the VP diffusion (3), $a_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{1-\bar{\alpha}t}$. The corresponding noisy observation is

$$X_t = a_t X_0 + \sigma_t \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_d),$$

where ε is independent of X_0. Conditional on $X_0 = x_0$, X_t is Gaussian with mean $a_t x_0$ and covariance matrix $\sigma_t^2 I_d$. Integrating over P_0 gives the marginal density of X_t:

$$p_t(x) = \int \mathcal{N}(x; a_t x_0, \sigma_t^2 I_d)\, \mathrm{d}P_0(x_0).$$

Thus, P_t is the convolution of the distribution of $a_t X_0$ with $\mathcal{N}(0, \sigma_t^2 I_d)$. Gaussian convolution gives P_t a smooth density p_t, even when P_0 is not absolutely continuous with respect to Lebesgue measure.

Section 2.4 shows that denoising score matching constructs its objective from the gradient of the logarithm of the forward transition density and that the conditional mean of the transition score given the noisy observation equals the noisy marginal score (Vincent, 2011). Under the Gaussian noising model above,

$$\nabla_x \log p_{t|0}(x \mid X_0) = -\frac{x - a_t X_0}{\sigma_t^2}.$$

It follows that

$$s(x,t) = \mathbb{E}\left[-\frac{x - a_t X_0}{\sigma_t^2} \mid X_t = x \right].$$

Given $X_t = x$, the quantity $a_t X_0 - x$ is the correction from the current observation x to the scaled clean signal $a_t X_0$. The score $s(x, t)$ is the posterior mean of this correction, scaled by σ_t^{-2}. As the gradient of $\log p_t(x)$, the score also points in the direction of the steepest local increase of the noisy marginal density. Under Gaussian noising, the noisy marginal score therefore describes both the local direction of density increase and the average denoising direction (Alain & Bengio, 2014; Vincent, 2011).

The same relation connects noise prediction and score prediction. Since $X_t - a_t X_0 = \sigma_t \varepsilon$, we have

$$s(x, t) = -\frac{1}{\sigma_t}\mathbb{E}[\varepsilon \mid X_t = x],$$

or, equivalently,

$$\mathbb{E}[\varepsilon \mid X_t = x] = -\sigma_t s(x, t).$$

Under squared-error loss, noise prediction targets the conditional noise mean at the population level (Ho et al., 2020).

When $a_t \neq 0$, rearranging the score representation gives the posterior mean of the clean sample:

$$\mathbb{E}[X_0 \mid X_t = x] = \frac{1}{a_t}\{x + \sigma_t^2 s(x, t)\}.$$

In the additive-noise special case $a_t = 1$, this becomes Tweedie's formula for additive Gaussian noise:

$$\mathbb{E}[X_0 \mid X_t = x] = x + \sigma_t^2 \nabla_x \log p_t(x).$$

The correction term $\sigma_t^2 \nabla_x \log p_t(x)$ shifts the noisy observation toward higher-density regions of the Gaussian-smoothed data distribution (Efron, 2011; Robbins, 1956; Saremi & Hyvärinen, 2019).

More generally, suppose that a network is trained under squared-error loss to predict a target U_t. The unrestricted population minimizer of the squared-error risk is

$$m^\star(x, t) = \mathbb{E}[U_t \mid X_t = x].$$

For $U_t = \varepsilon$, the population target is the conditional noise mean. For $U_t = X_0$, the population target is the posterior mean of the clean sample. For score prediction, the population target is $s(x, t)$. At the population level, the three targets are connected by the identities derived above. Table 1 summarizes these targets and their relations to the score.

The population identities do not imply that the three parameterizations yield the same empirical estimators with finite samples, finite-capacity model classes, and specific optimization procedures. In addition, practical implementations differ in other design choices. For example, loss weighting determines how strongly each noise level contributes to the objective, while the noise schedule determines the signal-to-noise ratios encountered across time. Output parameterization and network preconditioning rescale the regression target and can alter the numerical behaviour of optimization. Kingma et al. (2021) analyse these choices through variational likelihood, signal-to-noise ratio, and loss weighting. Karras et al. (2022) organize output parameterization, preconditioning, loss weighting, and sampler design within a common design space. Li and He (2026) further examine how predicting clean data rather than noisy quantities can produce different learning behaviour in finite-capacity networks.

Table 1. Population targets of three common prediction parameterizations under Gaussian noising at $t > 0$ with $\sigma_t > 0$.

Prediction parameterization	Population target under squared-error loss, relation to the score, and statistical interpretation
Score prediction	*Population target:* $s(x,t)$. *Relation to the score:* Direct estimate of $s(x,t)$. *Interpretation:* Gradient of the noisy marginal log-density.
ε-prediction	*Population target:* $\mathbb{E}[\varepsilon \mid X_t = x]$. *Relation to the score:* $s(x,t) = -\mathbb{E}[\varepsilon \mid X_t = x]/\sigma_t$. *Interpretation:* Conditional mean of the injected noise.
X_0-prediction	*Population target:* $\mathbb{E}[X_0 \mid X_t = x]$. *Relation to the score:* $s(x,t) = \{a_t\mathbb{E}[X_0 \mid X_t = x] - x\}/\sigma_t^2$. *Interpretation:* Posterior mean of the clean sample; denoising interpretation.

5. Distributional error and statistical guarantees

This section studies the discrepancy $\mathsf{d}(\widehat{P}, P_0)$ between the generated distribution $\widehat{P}$ and the target distribution P_0. We review commonly used discrepancies, examine the main sources of distributional error, and compare convergence guarantees and statistical rates under different structural assumptions.

5.1. Measures of distributional error

Different choices of $\mathsf{d}(\cdot,\cdot)$ measure different aspects of the discrepancy between $\widehat{P}$ and P_0. We consider total variation distance, Kullback–Leibler divergence, Wasserstein distance, and integral probability metrics (IPMs) defined through classes of test functions.

Total variation distance is defined by

$$d_{\mathrm{TV}}(P,Q) = \sup_{A\in\mathcal{A}} |P(A) - Q(A)|,$$

where $\mathcal{A}$ is the sigma-algebra on the sample space (Gibbs & Su, 2002). If P and Q have densities p and q, respectively, then

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int |p(x) - q(x)|\,\mathrm{d}x.$$

Total variation uniformly controls differences in the probabilities assigned by P and Q to measurable events. It is commonly used to assess the output distribution of diffusion samplers and to derive finite-sample minimax rates for estimating P_0 with diffusion models (Oko et al., 2023; Zhang et al., 2024).

When P is absolutely continuous with respect to Q, the Kullback–Leibler divergence is defined as follows (Gibbs & Su, 2002):

$$\mathrm{KL}(P\|Q) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P.$$

KL divergence is directional, and $\mathrm{KL}(P\|Q)$ generally differs from $\mathrm{KL}(Q\|P)$. KL bounds often arise from the Girsanov formula, path-space comparisons, relative-entropy dissipation, or stability analyses of reverse processes (H. Chen et al., 2023; Chen, Chewi, et al., 2023; Yang

& Wibisono, 2022). Pinsker's inequality (Gibbs & Su, 2002; Tsybakov, 2009) gives

$$d_{\mathrm{TV}}(P, Q) \leq \left\{ \frac{1}{2} \mathrm{KL}(P \| Q) \right\}^{1/2}.$$

Thus, an upper bound on $\mathrm{KL}(P\|Q)$ also gives an upper bound on $d_{\mathrm{TV}}(P, Q)$.

Wasserstein distance measures the geometric transport cost between two probability distributions. For $p \geq 1$, it is defined by

$$W_p(P, Q) = \left\{ \inf_{\Gamma \in \Pi(P,Q)} \int \|x - y\|^p \, \mathrm{d}\Gamma(x, y) \right\}^{1/p},$$

where $\Pi(P, Q)$ is the set of couplings with marginals P and Q (Villani, 2009). If P and Q have finite p-th moments, then $W_p(P, Q)$ is finite. In particular, W_1 has the Kantorovich–Rubinstein dual representation

$$W_1(P, Q) = \sup_{\|\varphi\|_{\mathrm{Lip}} \leq 1} |\mathbb{E}_P \varphi(X) - \mathbb{E}_Q \varphi(X)|.$$

A small W_1 distance therefore implies that the expectations of all one-Lipschitz test functions are similar under P and Q.

Wasserstein distance remains informative when distributions have different supports. Suppose that P_0 is concentrated on a set M of zero ambient Lebesgue measure, whereas $\widehat{P}$ is absolutely continuous. Then $P_0(M) = 1$ and $\widehat{P}(M) = 0$, so $d_{\mathrm{TV}}(\widehat{P}, P_0) = 1$. Thus, total variation remains maximal even as $\widehat{P}$ concentrates near M, whereas Wasserstein distance can decrease. Accordingly, W_1 is often used in analyses under the manifold hypothesis and in low-dimensional distribution recovery (De Bortoli, 2022). The distance W_2 commonly appears in coupling arguments and stability analyses for SDEs (Gao et al., 2025) and probability-flow ODEs (Gao & Zhu, 2025).

An integral probability metric (IPM) defines a distributional discrepancy through a class of test functions. For a test-function class $\mathcal{G}$, Müller (1997) defines the corresponding IPM by

$$d_{\mathcal{G}}(P, Q) = \sup_{\varphi \in \mathcal{G}} |\mathbb{E}_P \varphi(X) - \mathbb{E}_Q \varphi(X)|.$$

Different choices of $\mathcal{G}$ give different distributional discrepancies. Taking $\mathcal{G}$ to be the class of all measurable functions with values in $[0, 1]$ gives total variation distance. Taking $\mathcal{G}$ to be the class of one-Lipschitz functions gives W_1, while taking $\mathcal{G}$ to be the unit ball of a reproducing kernel Hilbert space $\mathcal{H}_\kappa$ associated with a positive-definite kernel κ gives maximum mean discrepancy (MMD) (Gretton et al., 2012).

The choice of discrepancy is part of a convergence guarantee. Pinsker's inequality converts a KL bound in the stated direction into total variation control, whereas Wasserstein and IPM bounds do not generally imply total variation control. We therefore report each convergence result together with its distributional discrepancy, the direction of KL divergence when relevant, and the assumptions under which the bound holds. Table 2 summarizes the definitions, statistical interpretations, and main roles of these discrepancies in the theory of diffusion models.

Table 2. Distributional error measures in diffusion-model theory: definitions, statistical interpretations, and roles.

Measure	Definition	Statistical interpretation	Role in diffusion theory
Total variation	$d_{\mathrm{TV}}(P,Q) = \sup_A \lvert P(A) - Q(A)\rvert$	Errors in the probabilities of all measurable events	Distribution recovery and minimax distribution estimation
KL divergence	$\mathrm{KL}(P\Vert Q) = \int \log(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}P$	Directional relative-entropy discrepancy that is finite only if P is absolutely continuous with respect to Q	Girsanov arguments, path-space comparisons, entropy dissipation, and reverse-process stability
Wasserstein $W_p, p \geq 1$	$W_p(P,Q) = \{\inf_{\Gamma\in\Pi(P,Q)} \int \lVert x-y\rVert^p\,\mathrm{d}\Gamma\}^{1/p}$	Geometric transport discrepancy defined through couplings for distributions with finite pth moments	Manifold and low-dimensional targets, coupling arguments, and SDE or ODE stability
IPM	$d_{\mathcal{G}}(P,Q) = \sup_{\varphi\in\mathcal{G}} \lvert\mathbb{E}_P\varphi - \mathbb{E}_Q\varphi\rvert$	Differences in expectations over a specified test-function class	Test-function bounds, including total variation, W_1, and MMD as specific cases

Table 3. Reverse processes, endpoint distributions, and error components.

Process	Dynamics; score; initialization	Distribution	Comparison and associated error
Forward X_t	Equation (1); $X_0 \sim P_0$	P_t	Defines the noisy marginals
Ideal reverse Y_r	Equation (4); s; $Y_0 \sim P_T$	P_τ	Comparison with P_0: early-stopping error $E_{\mathrm{early}}(\tau)$
Score-perturbed $Y_r^{\widehat{s}}$	Equation (13); $\widehat{s}$; $Y_0^{\widehat{s}} \sim P_T$	$P_\tau^{\widehat{s}}$	Comparison with P_τ: score error $E_{\mathrm{score}}(\widehat{s}, s)$
Continuous $\widehat{Y}_r$	Equation (11); $\widehat{s}$; $\widehat{Y}_0 \sim \pi_{\mathrm{ref}}$	$\widehat{P}_\tau^{\mathrm{cont}}$	Comparison with $P_\tau^{\widehat{s}}$: initialization error $E_{\mathrm{init}}(T)$
Discrete $\widehat{Y}_\ell$	Equation (12); $\widehat{s}$; $\widehat{Y}_0 \sim \pi_{\mathrm{ref}}$	$\widehat{P}_\tau^{\mathrm{disc}} = \widehat{P}$	Comparison with $\widehat{P}_\tau^{\mathrm{cont}}$: discretization error $E_{\mathrm{disc}}(h)$

Note: Reverse-process laws are evaluated at $r = T - \tau$, with $r_L = T - \tau$ for the discrete sampler.

5.2. Sources of distributional error

Equations (4), (11), and (12) define, respectively, the ideal reverse process, the practical continuous sampler, and its Euler–Maruyama discretization. To separate score error from initialization error, consider an auxiliary continuous reverse process that retains the ideal initial distribution P_T but uses the estimated score $\widehat{s}$:

$$\begin{aligned} \mathrm{d}Y_r^{\widehat{s}} &= [-f(Y_r^{\widehat{s}}, T-r) + g(T-r)^2\widehat{s}(Y_r^{\widehat{s}}, T-r)]\mathrm{d}r + g(T-r)\,\mathrm{d}\overline{W}_r, \\ &0 \leq r \leq T-\tau, \quad Y_0^{\widehat{s}} \sim P_T. \end{aligned} \tag{13}$$

Let $P_\tau^{\widehat{s}}$ denote its endpoint distribution, so that $Y_{T-\tau}^{\widehat{s}} \sim P_\tau^{\widehat{s}}$.

Recall the continuous reverse sampler in Equation (11) and its Euler–Maruyama discretization in Equation (12). We denote their endpoint laws at reverse time $T-\tau$ by $\widehat{P}_\tau^{\mathrm{cont}}$ and $\widehat{P}_\tau^{\mathrm{disc}}$, respectively. Here, $\widehat{P}_\tau^{\mathrm{disc}} = \widehat{P}$ is the distribution of the discrete sampler output $\widehat{Y}_L$, as defined after Equation (12).

For a distributional metric $\mathrm{d}(\cdot,\cdot)$, the triangle inequality applied along the sequence $\widehat{P}_\tau^{\mathrm{disc}}$, $\widehat{P}_\tau^{\mathrm{cont}}$, $P_\tau^{\widehat{s}}$, P_τ, and P_0 gives the following error decomposition:

$$\mathrm{d}(\widehat{P}, P_0) \leq \underbrace{\mathrm{d}(\widehat{P}_\tau^{\mathrm{disc}}, \widehat{P}_\tau^{\mathrm{cont}})}_{E_{\mathrm{disc}}(h)} + \underbrace{\mathrm{d}(\widehat{P}_\tau^{\mathrm{cont}}, P_\tau^{\widehat{s}})}_{E_{\mathrm{init}}(T)} + \underbrace{\mathrm{d}(P_\tau^{\widehat{s}}, P_\tau)}_{E_{\mathrm{score}}(\widehat{s},s)} + \underbrace{\mathrm{d}(P_\tau, P_0)}_{E_{\mathrm{early}}(\tau)}. \tag{14}$$

Table 3 summarizes the processes and endpoint distributions appearing in Equation (14), together with their roles in the decomposition. The following subsections discuss these four sources of error.

5.3. Score error: propagation and sources

The analysis of the score-error term addresses two questions. The first is how the discrepancy between $\widehat{s}(x,t)$ and $s(x,t)$ propagates through the reverse dynamics to the output distribution under a given score-accuracy condition. The second is whether an estimator satisfying this condition can be learned from a finite sample. The latter depends on finite-sample estimation error and function-class approximation error.

5.3.1. From score error to distribution error

Recall from Table 3 that the ideal reverse process Y_r and the score-perturbed reverse process $Y_r^{\widehat{s}}$ both start from P_T and use the true score s and the estimated score $\widehat{s}$, respectively. For reverse SDEs with the same diffusion coefficient, a common propagation argument compares their path-space distributions using the Girsanov formula.

Let $\mathbf{P}$ and $\mathbf{P}^{\widehat{s}}$ denote the path-space distributions of Y_r and $Y_r^{\widehat{s}}$, whose endpoint marginals are P_τ and $P_\tau^{\widehat{s}}$, respectively. The difference between their reverse drifts, evaluated along the ideal process, is

$$g(T-r)^2\{\widehat{s}(Y_r, T-r) - s(Y_r, T-r)\}.$$

Under the integrability conditions required for the Girsanov formula (Øksendal, 2003),

$$\mathrm{KL}(\mathbf{P}\|\mathbf{P}^{\widehat{s}}) = \frac{1}{2}\int_\tau^T g(t)^2 \mathbb{E}_{P_t}[\|\widehat{s}(X_t,t) - s(X_t,t)\|^2]\mathrm{d}t.$$

Because the terminal marginal is the projection of the path-space distribution onto the terminal coordinate, the data processing inequality gives

$$\mathrm{KL}(P_\tau\|P_\tau^{\widehat{s}}) \le \frac{1}{2}\int_\tau^T g(t)^2 \mathbb{E}_{P_t}[\|\widehat{s}(X_t,t) - s(X_t,t)\|^2]\mathrm{d}t.$$

Pinsker's inequality further gives

$$d_{\mathrm{TV}}(P_\tau, P_\tau^{\widehat{s}}) \le \left\{\frac{1}{2}\mathrm{KL}(P_\tau\|P_\tau^{\widehat{s}})\right\}^{1/2}.$$

Thus, the time-integrated $L^2(P_t)$ score error, weighted by $g(t)^2$, controls the KL divergence between the terminal marginals and, through Pinsker's inequality, their total variation distance.

The time-integrated $L^2(P_t)$ condition above is one of several score-accuracy conditions used in the literature. More generally, obtaining a distributional error bound from score error requires specifying the error norm and how the error is aggregated over time. Let $\pi_\infty = \mathcal{N}(0, I_d)$ denote the stationary Gaussian distribution of the Ornstein–Uhlenbeck (OU) process, and let $I(P_t\|\pi_\infty)$ denote the relative Fisher information of P_t with respect to π_∞. Common score-accuracy conditions are summarized in Table 4, where MGF denotes moment-generating function.

Suppose that the uniform-in-time absolute-error conditions and the time-averaged error use the same time domain and corresponding noisy marginals, and that $\lambda_t > 0$. Assume that the relevant averaging weights are nonnegative and normalized: $\int_\tau^T w_{\mathrm{err}}(t)\,\mathrm{d}t = 1$ in continuous time and $\sum_k w_{\mathrm{err},k} = 1$ on a discrete grid. Jensen's inequality and the fact that a weighted

Table 4. Representative score-accuracy conditions in analyses of diffusion samplers.

Score-accuracy condition	Typical form	Interpretation	Selected references
Time-averaged error $L^2_{\mathrm{avg}}(P_t)$	$\int_\tau^T w_{\mathrm{err}}(t)\mathbb{E}_{P_t}\|\widehat{s}(X_t,t) - s(X_t,t)\|^2 \mathrm{d}t \le \eta^2$ or $\sum_k w_{\mathrm{err},k}\mathbb{E}_{P_{t_k}}\|\widehat{s}(X_{t_k},t_k) - s(X_{t_k},t_k)\|^2 \le \eta^2$	Averages the squared error across noise levels in continuous time or over a discrete grid	H. Chen et al. (2023); Conforti et al. (2025); Pedrotti et al. (2024); Huang et al. (2025)
Uniform-in-time error $L^2_{\mathrm{unif}}(P_t)$	$\sup_{t\in\mathcal{T}}\mathbb{E}_{P_t}\|\widehat{s}(X_t,t) - s(X_t,t)\|^2 \le \eta^2$	Controls the $L^2(P_t)$ error at every noise level in the specified interval or grid	Lee et al. (2022); Chen, Chewi, et al. (2023)
L^∞	$\sup_{t\in\mathcal{T},x\in\mathbb{R}^d}\|\widehat{s}(x,t) - s(x,t)\| \le \eta_\infty$	Gives uniform control over time and state space	De Bortoli et al. (2021)
Uniform-in-time marginal MGF error	$\sup_{t\in\mathcal{T}}\lambda_t^{-1}\log\mathbb{E}_{P_t}\exp\{\lambda_t\|\widehat{s}(X_t,t) - s(X_t,t)\|^2\} \le \eta^2_{\mathrm{mgf}}$	Controls marginal exponential moments of the score error uniformly over the specified time set; the cited result uses a grid-interpolated version with the estimated score frozen at a grid point over each interpolation interval	Yang and Wibisono (2022)
Terminal-noise-level MGF error	$\lambda_T^{-1}\log\mathbb{E}_{P_T}\exp\{\lambda_T\|\widehat{s}(X_T,T) - s(X_T,T)\|^2\} \le \eta^2_{\mathrm{mgf},T}$	Controls an exponential moment at the terminal noise level	Pedrotti et al. (2024)
Relative L^2	$\mathbb{E}_{P_{t_k}}\|\widehat{s}(X_{t_k},t_k) - s(X_{t_k},t_k)\|^2 \le \eta^2_{\mathrm{rel}} I(P_{t_k}\|\pi_\infty)$ for every grid time t_k	Scales the error by relative Fisher information in the Ornstein–Uhlenbeck setting	Conforti et al. (2025)

Note: Here, $\mathcal{T}$ denotes the relevant set of evaluation times, which may be a continuous interval or a discrete sampling grid. The quantity $\lambda_t > 0$ is the time-dependent exponential-moment parameter, with λ_T denoting its terminal value. For the time-averaged conditions, the continuous or discrete weights are nonnegative and normalized.

average is no larger than the corresponding supremum then give

$$L^\infty \Longrightarrow \mathrm{MGF}_{\mathrm{unif}} \Longrightarrow L^2_{\mathrm{unif}}(P_t) \Longrightarrow L^2_{\mathrm{avg}}(P_t).$$

In general, there is no implication in either direction between the relative L^2 condition and the absolute-error conditions above. The relative condition normalizes the error by the time-dependent quantity $I(P_t\|\pi_\infty)$. Across these formulations, L^2 conditions are most common because denoising score matching uses a squared loss, while the Girsanov formula and path-space KL comparisons naturally produce squared drift errors.

A score-accuracy condition alone does not determine a distributional guarantee; it must be paired with the reverse dynamics, a propagation argument, and a target distributional discrepancy. For reverse SDEs, the Girsanov argument links time-integrated $L^2(P_t)$ error to KL divergence. Wasserstein analyses typically use coupling and require stability of the reverse drift, whereas KL stability results may rely on exponential-moment or relative-L^2 conditions. Analyses of deterministic probability-flow ODEs and Fokker–Planck formulations use arguments based on PDEs, such as continuity and Kolmogorov equations, and require corresponding regularity conditions. Table 5 summarizes the main combinations of reverse dynamics, score-accuracy conditions, propagation arguments, and distributional discrepancies.

5.3.2. Statistical and approximation errors in score estimation

Recall that $\mathcal{D}_n = \{X^{(1)},\ldots,X^{(n)}\}$ denotes the i.i.d. training sample from P_0. Under the Gaussian noising model $X_t = a_t X_0 + \sigma_t\varepsilon$, the conditional-score formula in Section 4 gives

Table 5. Score-error propagation arguments for different reverse dynamics and distributional discrepancies.

Reverse dynamics or comparison	Score-accuracy condition	Propagation argument	Distributional discrepancy	Selected references
Reverse SDE	Time-averaged $L^2(P_t)$ error or L^2 error at each grid time	Girsanov arguments based on path-space KL and related KL-interpolation arguments	KL and, through Pinsker's inequality, total variation	Chen, Chewi, et al. (2023); H. Chen et al. (2023)
Reverse SDE or discrete reverse process	Gridwise $L^2(P_{t_k})$ error, with fixed or time-dependent tolerances	Bad-set reduction, interval-wise analysis, and change of measure	Total variation or Wasserstein	Lee et al. (2022); Lee et al. (2023)
Coupled reverse SDEs	L^2 error along generated iterates; alternatively, $L^2(P_t)$ error with Lipschitz score estimates	Synchronous coupling, Itô's formula, and strong log-concavity or contraction estimates	W_2	Gao et al. (2025)
Langevin or OU-based dynamics	MGF-type or relative L^2 error	Entropy dissipation, change of measure, and functional inequalities	KL	Yang and Wibisono (2022); Conforti et al. (2025)
Probability-flow ODE	Integrated $L^2_{\mathrm{avg}}(P_t)$ error with vector-field regularity	Continuity equations, the method of characteristic lines, and interpolation inequalities	Total variation	Huang et al. (2025)
Fokker–Planck analysis of reverse diffusion	Uniform-in-time or time-integrated L^2 drift or score error, with drift regularity	Backward Kolmogorov equations, IPM duality, and HJB/Bernstein gradient estimates	W_1, total variation, or other IPMs	Mimikos-Stamatopoulos et al. (2024)
Two-stage prediction–correction scheme	Integrated $L^2(P_t)$ error with terminal MGF control	Corrector contraction and reverse-dynamics stability	W_2 or total variation	Pedrotti et al. (2024)

$\nabla_{x_t} \log p_{t|0}(X_t \mid X_0) = -\varepsilon/\sigma_t$. The empirical DSM objective is therefore

$$\widehat{\mathcal{L}}_{\mathrm{DSM},n}(u) = \frac{1}{n}\sum_{i=1}^{n}\int_{\tau}^{T} w(t)\mathbb{E}_{\varepsilon}\left[\left\| u(a_t X^{(i)} + \sigma_t \varepsilon, t) + \frac{\varepsilon}{\sigma_t}\right\|^2\right] \mathrm{d}t.$$

Given a function class $\mathcal{F}$, the corresponding score estimator satisfies

$$\widehat{s} \in \arg\min_{u\in\mathcal{F}} \widehat{\mathcal{L}}_{\mathrm{DSM},n}(u).$$

By the DSM identity in Section 2.4, the population counterpart of the empirical DSM objective defined above differs from the integrated score risk

$$\mathcal{R}(u) = \int_{\tau}^{T} w(t)\mathbb{E}_{P_t}[\|u(X_t,t) - s(X_t,t)\|^2]\mathrm{d}t,$$

only by a constant that does not depend on u. We therefore examine how finite-sample estimation and restriction to the function class $\mathcal{F}$ affect $\mathcal{R}(\widehat{s})$.

Ignoring optimization error, the expected score risk decomposes as

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\widehat{s})] = \underbrace{\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\widehat{s})] - \inf_{u\in\mathcal{F}} \mathcal{R}(u)}_{E_{\mathrm{stat}}} + \underbrace{\inf_{u\in\mathcal{F}} \mathcal{R}(u)}_{E_{\mathrm{approx}}}.$$

Here, E_{stat} is the statistical error due to the finite training sample, whereas E_{approx} is the approximation error induced by restricting the score estimator to $\mathcal{F}$.

Statistical error. The function class $\mathcal{F}$ is typically a neural-network class. In this setting, E_{stat} depends on the sample size n, the complexity of $\mathcal{F}$, the weighting $w(t)$ across noise levels, and the moment and tail properties of the DSM loss. Finite-sample analyses use empirical-process bounds, covering numbers, Rademacher complexity, and concentration inequalities to control the difference between the empirical DSM objective and its population counterpart (Block et al., 2020; Han et al., 2024; Oko et al., 2023). Analyses involving multiple noise levels may also use martingale decompositions and concentration arguments adapted to the time-dependent objective (Kumar et al., 2025). If optimization error is included, analysing an estimator produced by a specific training algorithm also requires algorithm-specific optimization and generalization arguments (Han et al., 2024).

Beyond neural-network DSM estimators, kernel and plug-in score estimators can exploit Gaussian smoothing, density estimation, or empirical Bayes identities. Score-estimation guarantees have been established under assumptions involving sub-Gaussian tails, Sobolev smoothness, or compact support (Dou et al., 2024; Wibisono et al., 2024; Zhang et al., 2024).

Note that in practice, the training procedure also samples the noise level t and the Gaussian noise ε to approximate the time integral and the expectation over the noise in the empirical DSM objective. These auxiliary draws are not additional observations from P_0; the finite-sample statistical error E_{stat} concerns the n independent observations in $\mathcal{D}_n$.

Approximation error. By definition,

$$E_{\text{approx}} = \inf_{u \in \mathcal{F}} \int_{\tau}^{T} w(t)\mathbb{E}_{P_t}[\|u(X_t, t) - s(X_t, t)\|^2]\mathrm{d}t.$$

This term measures the best-in-class approximation error for the time-dependent score $s(x, t)$ over $\mathcal{F}$. For each $t > 0$, Gaussian smoothing gives P_t a smooth density p_{t}. As t approaches zero and the smoothing weakens, however, approximation bounds for neural score models may deteriorate (Fu et al., 2024; Oko et al., 2023). Such bounds may depend on the regularity and tail behaviour of P_0, structural properties of the distribution, and the approximation properties of the chosen network architecture.

If the function class varies with the sample size, denote it by $\mathcal{F}_n$. Increasing its richness can reduce E_{approx} but typically makes E_{stat} more difficult to control. The score-estimation rate therefore reflects the trade-off between these two errors.

5.4. Discretization and endpoint errors

This section considers time discretization error and endpoint errors. Time discretization error arises from approximating continuous reverse dynamics on a finite time grid. Endpoint errors occur at both ends of the time interval. At the high-noise end, the sampler is initialized from the reference distribution π_{ref} rather than the exact forward marginal P_T, producing initialization error. At the low-noise end, the ideal reverse process is stopped when the forward-time coordinate reaches $t = \tau > 0$, corresponding to reverse time $r = T - \tau$, so its final marginal is P_τ rather than the target distribution P_0, producing early-stopping error.

5.4.1. Discretization of reverse-time dynamics

Recall from Section 3 that the Euler–Maruyama sampler uses a reverse-time grid $0 = r_0 < \cdots < r_L = T - \tau$, with $t_\ell = T - r_\ell$, $\Delta r_\ell = r_{\ell+1} - r_\ell$, and $h = \max_{0 \le \ell < L} \Delta r_\ell$. Table 3 identifies $\widehat{P}_\tau^{\text{cont}}$ and $\widehat{P}_\tau^{\text{disc}} = \widehat{P}$ as the endpoint distributions of the practical continuous sampler and its Euler–Maruyama discretization, respectively. Accordingly, the discretization error term in (14) is

$$E_{\text{disc}}(h) = \mathrm{d}(\widehat{P}_\tau^{\text{disc}}, \widehat{P}_\tau^{\text{cont}}) = \mathrm{d}(\widehat{P}, \widehat{P}_\tau^{\text{cont}}).$$

Reverse-SDE discretization. The Euler–Maruyama scheme in Equation (12) fixes the drift, diffusion coefficient, and estimated score at the left endpoint of each time interval. The discrepancy over one time step arises from variation in these quantities within the time interval.

Setting $f(x,t) = -x/2$ and $g(t) = 1$ in Equation (1) gives the OU forward process

$$\mathrm{d}X_t = -\frac{1}{2} X_t \,\mathrm{d}t + \mathrm{d}W_t.$$

The corresponding estimated reverse drift is $\frac{1}{2}y + \widehat{s}(y, T-r)$. The Euler–Maruyama update is then

$$\widehat{Y}_{\ell+1}^{\text{EM}} = \widehat{Y}_\ell^{\text{EM}} + \left[\frac{1}{2}\widehat{Y}_\ell^{\text{EM}} + \widehat{s}(\widehat{Y}_\ell^{\text{EM}}, t_\ell)\right] \Delta r_\ell + \sqrt{\Delta r_\ell}\, Z_\ell.$$

To analyse the discretization error, write the Euler–Maruyama chain as a continuous interpolation driven by the same Brownian motion as the continuous reverse process. Set $\widehat{Y}_0^{\text{EM}} = \widehat{Y}_0 \sim \pi_{\text{ref}}$, $\sqrt{\Delta r_\ell}\, Z_\ell = \overline{W}_{r_{\ell+1}} - \overline{W}_{r_\ell}$, and, for $r \in [r_\ell, r_{\ell+1}]$, define

$$\mathrm{d}\widetilde{Y}_r^{\text{EM}} = \left[\frac{1}{2}\widehat{Y}_\ell^{\text{EM}} + \widehat{s}(\widehat{Y}_\ell^{\text{EM}}, t_\ell)\right] \mathrm{d}r + \mathrm{d}\overline{W}_r, \quad \widetilde{Y}_{r_\ell}^{\text{EM}} = \widehat{Y}_\ell^{\text{EM}}.$$

At $r_{\ell+1}$, this interpolation reproduces the Euler–Maruyama update.

Its drift residual relative to the continuous reverse SDE is

$$\frac{1}{2}(\widetilde{Y}_r^{\text{EM}} - \widehat{Y}_\ell^{\text{EM}}) + \widehat{s}(\widetilde{Y}_r^{\text{EM}}, T-r) - \widehat{s}(\widehat{Y}_\ell^{\text{EM}}, t_\ell).$$

Some theoretical analyses use an exponential integrator (EI) as an alternative discretization strategy. Its update is

$$\widehat{Y}_{\ell+1}^{\text{EI}} = \mathrm{e}^{\Delta r_\ell/2}\widehat{Y}_\ell^{\text{EI}} + 2(\mathrm{e}^{\Delta r_\ell/2} - 1)\widehat{s}(\widehat{Y}_\ell^{\text{EI}}, t_\ell) + \sqrt{\mathrm{e}^{\Delta r_\ell} - 1}\, Z_\ell^{\text{EI}}.$$

Similarly, couple the EI chain and the continuous reverse process using the same initial value and Brownian motion. Set $\widehat{Y}_0^{\text{EI}} = \widehat{Y}_0 \sim \pi_{\text{ref}}$, $\sqrt{\mathrm{e}^{\Delta r_\ell} - 1}\, Z_\ell^{\text{EI}} = \int_{r_\ell}^{r_{\ell+1}} \mathrm{e}^{(r_{\ell+1}-v)/2}\,\mathrm{d}\overline{W}_v$. The stochastic integral has covariance $(\mathrm{e}^{\Delta r_\ell} - 1)I_d$, so $Z_\ell^{\text{EI}} \sim \mathcal{N}(0, I_d)$. For $r \in [r_\ell, r_{\ell+1}]$, define the EI interpolation by

$$\mathrm{d}\widetilde{Y}_r^{\text{EI}} = \left[\frac{1}{2}\widetilde{Y}_r^{\text{EI}} + \widehat{s}(\widehat{Y}_\ell^{\text{EI}}, t_\ell)\right] \mathrm{d}r + \mathrm{d}\overline{W}_r, \quad \widetilde{Y}_{r_\ell}^{\text{EI}} = \widehat{Y}_\ell^{\text{EI}}.$$

Its drift residual relative to the continuous reverse SDE is

$$\widehat{s}(\widetilde{Y}_r^{\text{EI}}, T-r) - \widehat{s}(\widehat{Y}_\ell^{\text{EI}}, t_\ell).$$

The Euler–Maruyama residual contains both linear-drift variation and score variation, whereas the EI has no separate linear-drift residual.

Continuous interpolation represents a discrete sampler as a continuous process with the same diffusion coefficient as the continuous reverse SDE. Girsanov or related KL arguments can then convert the accumulated squared drift residual into a path-space KL bound. This path-space comparison follows the same type of argument as in Section 5.3.1, but the drift difference here arises from time and state discretization rather than from replacing the true score with the estimated score. Data processing gives a KL bound for the final marginals, and Pinsker's inequality further gives a total variation bound (H. Chen et al., 2023; Chen, Chewi, et al., 2023).

For Wasserstein distance, the continuous reverse SDE and its numerical interpolation can be coupled using the same Brownian motion. The difference between their endpoints then gives

$$W_2^2(\widehat{P}_\tau^{\mathrm{disc}}, \widehat{P}_\tau^{\mathrm{cont}}) \le \mathbb{E}[\|\widehat{Y}_L^{\mathrm{EM}} - \widehat{Y}_{T-\tau}\|^2].$$

Under suitable smoothness and contractivity conditions, the right-hand side can be controlled using Itô's formula and contraction estimates (Gao et al., 2025).

Probability-flow ODE discretization. For a probability-flow ODE using an estimated score, the discretization error is assessed by comparing the continuous flow with its finite-step numerical approximation (Chen, Daras, et al., 2023). Wasserstein bounds can be obtained by synchronously coupling the continuous flow and its numerical approximation from the same initial value (Gao & Zhu, 2025). For total variation, some analyses represent the numerical solution by a continuous interpolation and compare the induced density evolutions using transport-equation arguments (Huang et al., 2025).

5.4.2. Initialization and early-stopping errors

Initialization error. Recall from Table 3 and Equation (14) that $E_{\mathrm{init}}(T) = \mathsf{d}(\widehat{P}_\tau^{\mathrm{cont}}, P_\tau^{\widehat{s}})$ compares two continuous reverse processes that use the same estimated score $\widehat{s}$ but start from π_{ref} and P_T, respectively. This endpoint error is induced by the discrepancy between the two initial distributions.

For the OU forward process with $f(x,t) = -x/2$ and $g(t) = 1$,

$$X_T = \mathrm{e}^{-T/2} X_0 + \sqrt{1 - \mathrm{e}^{-T}}\, Z, \quad Z \sim \mathcal{N}(0, I_d), \quad Z \perp X_0.$$

As $T \to \infty$, P_T converges to the Gaussian distribution $\pi_\infty = \mathcal{N}(0, I_d)$. A common choice for the finite-horizon reference distribution is therefore $\pi_{\mathrm{ref}} = \pi_\infty$, which approximates P_T. If P_0 has a finite second moment,

$$W_2(P_T, \pi_\infty) \le \mathrm{e}^{-T/2} W_2(P_0, \pi_\infty).$$

For $T \ge 1$,

$$\mathrm{KL}(P_T \| \pi_\infty) \lesssim (d + \mathbb{E}_{P_0}\|X_0\|^2)\mathrm{e}^{-T},$$

and Pinsker's inequality gives the corresponding total variation bound (H. Chen et al., 2023). Because the two reverse processes use the same estimated reverse transition kernel, data processing carries these KL and total variation bounds to the corresponding endpoint distributions, with the KL direction preserved. Consequently, the initialization error in total variation admits an upper bound that decays exponentially with T.

Early-stopping error. Recall from Table 3 and Equation (14) that $E_{\mathrm{early}}(\tau) = \mathsf{d}(P_\tau, P_0)$. At the low-noise endpoint, the ideal reverse process runs to $t = 0$ and returns P_0. When P_0 is

nonsmooth or supported on a lower-dimensional set, the amount of Gaussian smoothing decreases as t approaches zero, making it more difficult to control score-estimation error and score regularity uniformly. Some theoretical analyses therefore stop the reverse process when the forward-time coordinate reaches $t = \tau > 0$ (De Bortoli, 2022).

Under the OU forward process, P_τ is the law of the early-stopped endpoint

$$X_\tau = \mathrm{e}^{-\tau/2} X_0 + \sqrt{1 - \mathrm{e}^{-\tau}}\, Z, \quad Z \sim \mathcal{N}(0, I_d), \quad Z \perp X_0.$$

A coupling constructed using the same X_0 gives

$$W_2^2(P_\tau, P_0) \le (1 - \mathrm{e}^{-\tau/2})^2 \mathbb{E}_{P_0} \|X_0\|^2 + d(1 - \mathrm{e}^{-\tau}).$$

If P_0 has a finite second moment, the W_2 early-stopping error therefore converges to zero as $\tau \to 0$.

The choice of distributional metric is important in the analysis of early-stopping error. Suppose that P_0 is supported on a lower-dimensional set M with zero ambient Lebesgue measure. Then $P_0(M) = 1$. For every $\tau > 0$, Gaussian noising gives P_τ a density with respect to ambient Lebesgue measure, so $P_\tau(M) = 0$. Consequently, $d_{\mathrm{TV}}(P_\tau, P_0) = 1$. This example illustrates the discussion in Section 5.1. For target distributions supported on lower-dimensional sets, Wasserstein distance can describe the geometric convergence of P_τ to P_0. Early-stopping guarantees must therefore specify the metric and the required regularity conditions.

With the metric and regularity conditions specified, finite-sample convergence rates for recovering P_0 are obtained by choosing $\tau = \tau_n \to 0$ and jointly controlling the error terms. As τ_n decreases, the early-stopping error decreases, while the statistical and approximation error bounds for score estimation may deteriorate near time zero. The following two examples illustrate different ways of controlling the dependence on the early-stopping time.

For the results of Zhang et al. (2024) discussed here and in Section 5.5, the time parameters and sampler output distributions follow their Brownian parameterization. Zhang et al. (2024) establish an explicit trade-off between score estimation and early stopping. Under their assumptions, with Sobolev smoothness $0 < \nu \le 2$, the TV error contribution from their kernel-based score estimator is bounded, up to logarithmic factors, by $n^{-1/2} \tau_n^{-d/4}$, whereas the early-stopping error is of order $\tau_n^{\nu/2}$. Balancing these bounds gives $\tau_n = n^{-2/(2\nu+d)}$ and the TV convergence rate $n^{-\nu/(2\nu+d)}$.

A different approach is taken by Oko et al. (2023), who control the early-stopping error within the target convergence rate. Under their assumptions, with Besov smoothness ν, a suitable network class $\mathcal{F}_n$ yields an integrated score-risk bound of order $n^{-2\nu/(2\nu+d)}$, up to logarithmic factors, accounting for both statistical and approximation errors. They choose τ_n to decay sufficiently fast with n so that the early-stopping error is no larger than $n^{-\nu/(2\nu+d)}$.

In addition, decreasing τ_n to reduce the early-stopping error may require a smaller sampling step size to control the discretization error. For example, the TV guarantee of M. Chen et al. (2023) for the early-stopped latent distribution holds with a uniform step size $h \le d_\star^{-1} \tau_n^2 n^{-2/(d_\star+5)}$, where $d_\star$ is the intrinsic dimension.

Overall, controlling the early-stopping error requires a suitable choice of metric, together with a choice of τ_n that accounts for the statistical, approximation, and discretization errors.

5.5. Guarantees under structural assumptions

We first consider the classical nonparametric setting in which P_0 admits a density p_0 with respect to Lebesgue measure on $\mathbb{R}^d$. Zhang et al. (2024) assume that P_0 is sub-Gaussian and that p_0 belongs to a Sobolev class of smoothness $0 < \nu \leq 2$. Using a kernel-based score estimator, with $\tau_n = n^{-2/(2\nu+d)}$ and $T = n^{2\nu/(2\nu+d)}$, their continuous-time reverse sampler satisfies

$$\mathbb{E}_{\mathcal{D}_n}[d_{\mathrm{TV}}(\widehat{P}^{\mathrm{cont}}_{\tau_n}, P_0)] = \widetilde{\mathcal{O}}(n^{-\nu/(2\nu+d)}).$$

Here, $\widetilde{\mathcal{O}}$ suppresses logarithmic factors in n. This rate matches the classical minimax rate up to logarithmic factors.

Oko et al. (2023) consider densities supported on $[-1,1]^d$ in the Besov class $B^{\nu}_{p,q}$, with $0 < p, q \leq \infty$ and $\nu > d(1/p - 1/2)_+$, assuming positive lower and finite upper bounds on the density and additional C^∞ smoothness near the boundary within the support. Their diffusion-based distribution estimator attains an expected TV error of order $\widetilde{\mathcal{O}}(n^{-\nu/(2\nu+d)})$. For $d \geq 3$, using a modified score-network construction, they also obtain an expected W_1 error of order $n^{-(\nu+1-\delta)/(2\nu+d)}$ for any fixed, arbitrarily small $\delta > 0$.

Beyond the full-dimensional density setting, low-dimensional support or factorization structure can lead to convergence rates governed by an intrinsic or effective dimension rather than the ambient dimension.

Linear-subspace support. M. Chen et al. (2023) consider distributions supported on an unknown low-dimensional linear subspace. Their analysis assumes a strictly positive, twice continuously differentiable density within the subspace, with a Gaussian upper bound on its tails, together with uniformly Lipschitz scores for the forward-noised latent distributions. Under these conditions, they establish score estimation and subspace recovery guarantees. With finite KL divergence from the latent distribution to a standard Gaussian, they further obtain distribution recovery guarantees. The convergence rates depend primarily on the subspace dimension rather than the ambient dimension.

Manifold-supported distributions. De Bortoli (2022) establishes W_1 convergence guarantees for compactly supported targets, including manifold-supported distributions, under a pointwise score-error bound that permits linear growth in the state variable and deterioration as the noise level decreases. Tang and Yang (2024) study distributions supported on compact, boundaryless manifolds of smoothness $\beta \geq 2$ and with reach bounded away from zero. They assume that the density with respect to manifold volume is bounded away from zero and has smoothness $0 \leq \nu \leq \beta - 1$. Under these assumptions, their suitably tuned forward-backward diffusion estimator achieves the minimax W_1 rate up to logarithmic factors, with the convergence exponent depending on the intrinsic dimension.

Densities with factorization structure. Low-dimensional structure may also arise through factorization or low-order interactions within a full-dimensional density. Suppose that the density p_0 is supported on $[-1,1]^d$, bounded above and below by positive constants on this support, and has a log-density composed of ν-Hölder-smooth components with maximal interaction order d^*. Fan et al. (2025) show that a diffusion-based distribution estimator can attain

$$\mathbb{E}_{\mathcal{D}_n}[d_{\mathrm{TV}}(\widehat{P}^{\mathrm{cont}}_{\tau}, P_0)] = \widetilde{\mathcal{O}}(n^{-\nu/(2\nu+d^*)}).$$

The dimension dependence is governed by d^* rather than the ambient dimension d. Kwon et al. (2026) also establish this TV rate for Hölder-smooth factorizable densities, assuming positive density lower bounds and uniformly bounded derivatives of all orders near the boundary within the support.

6. Conditional distribution learning with diffusion models

Many generative and statistical tasks involve distributions conditional on auxiliary information. Examples include class labels or text prompts in image generation, observed histories in probabilistic forecasting, noisy or incomplete measurements in inverse problems, and observed components in missing-data inference. In each setting, the target is a conditional law $P_{0|c}$, where c denotes the available information.

6.1. Incorporating conditioning information

6.1.1. Conditional diffusion models

Let C denote the conditioning variable. A conditional diffusion model uses paired training observations

$$(X^{(1)}, C^{(1)}), \ldots, (X^{(n)}, C^{(n)}).$$

Let $P_{t|c}$ denote the conditional distribution of X_t given $C = c$, and let $p_t(x \mid c)$ denote its density. The corresponding conditional score is

$$s(x, t; c) = \nabla_x \log p_t(x \mid c).$$

The model estimates this score through

$$\widehat{s}(x, t; c) \approx s(x, t; c).$$

The score network takes c as an additional input and learns the dependence of X_t on c across noise levels. For a condition c, substituting $\widehat{s}(x, t; c)$ into the reverse-time dynamics produces samples from the estimated conditional law.

6.1.2. Classifier guidance

Classifier guidance (CG) combines an unconditional diffusion model with a separately trained classifier to construct a conditional reverse process (Dhariwal & Nichol, 2021). At noise level t, Bayes' rule gives

$$p_t(x \mid c) = \frac{p_t(c \mid x) p_t(x)}{p_t(c)}.$$

Since $p_t(c)$ does not depend on x, taking logarithms and differentiating both sides with respect to x yields

$$\nabla_x \log p_t(x \mid c) = \nabla_x \log p_t(x) + \nabla_x \log p_t(c \mid x).$$

The first term is the unconditional noisy-marginal score and can be estimated by the diffusion-model score estimator $\widehat{s}(x, t)$. The second term can be estimated using an additional classifier. Specifically, a classifier $\widehat{p}_\phi(c \mid x, t)$ is trained to predict the condition c from

a noisy sample x, with t specifying the noise level. This gives the plug-in conditional-score estimator

$$\widehat{s}^{\mathrm{CG}}_{\mathrm{cond}}(x,t;c) = \widehat{s}(x,t) + \nabla_x \log \widehat{p}_\phi(c \mid x,t).$$

Substituting $\widehat{s}^{\mathrm{CG}}_{\mathrm{cond}}(x,t;c)$ into the reverse-time dynamics allows the classifier to guide generation toward samples consistent with condition c.

CG introduces a guidance scale $\gamma > 0$:

$$\widehat{s}_{\mathrm{CG},\gamma}(x,t;c) = \widehat{s}(x,t) + \gamma\, \nabla_x \log \widehat{p}_\phi(c \mid x,t).$$

When $\gamma = 1$, this expression reduces to the conditional-score plug-in estimator above. Increasing γ strengthens the contribution of the classifier gradient, producing a diversity–fidelity trade-off in conditional generation (Dhariwal & Nichol, 2021).

6.1.3. Classifier-free guidance

CG uses a separately trained classifier to estimate $\nabla_x \log p_t(c \mid x)$. The same Bayes identity gives

$$\nabla_x \log p_t(c \mid x) = s(x,t;c) - s(x,t).$$

The guidance term can therefore be obtained from the difference between the conditional and unconditional scores. Classifier-free guidance (CFG) uses this identity to learn both scores within a single diffusion model, without training a separate classifier (Ho & Salimans, 2022).

During training, the conditioning information is randomly dropped. When c is provided, the model estimates the conditional score,

$$\widehat{s}(x,t;c) \approx s(x,t;c).$$

When the condition is dropped, the same model estimates the unconditional score,

$$\widehat{s}(x,t) \approx s(x,t).$$

The difference

$$\widehat{s}(x,t;c) - \widehat{s}(x,t)$$

then replaces the classifier gradient used in classifier guidance.

Let $\omega \geq 0$ denote the CFG guidance scale. Classifier-free guidance modifies the conditional score estimate by adding the scaled difference between the conditional and unconditional score estimates:

$$\widehat{s}_{\mathrm{CFG},\omega}(x,t;c) = \widehat{s}(x,t;c) + \omega\{\widehat{s}(x,t;c) - \widehat{s}(x,t)\}, \qquad \omega \geq 0.$$

Equivalently,

$$\widehat{s}_{\mathrm{CFG},\omega}(x,t;c) = (1+\omega)\widehat{s}(x,t;c) - \omega\widehat{s}(x,t).$$

Substituting $\widehat{s}_{\mathrm{CFG},\omega}(x,t;c)$ into the reverse-time dynamics gives the classifier-free guided sampler. When $\omega = 0$, $\widehat{s}_{\mathrm{CFG},0}(x,t;c) = \widehat{s}(x,t;c)$. When $\omega > 0$, the vector field also contains the scaled difference $\omega\{\widehat{s}(x,t;c) - \widehat{s}(x,t)\}$.

6.2. Statistical theory for conditional diffusion models

In conditional distribution learning, the target is a family of conditional laws indexed by c. Existing theory primarily studies whether learned conditional samplers approximate these laws and how guidance changes the distributions they produce.

6.2.1. Estimation of conditional laws

Given paired observations $\mathcal{D}_n^{\mathrm{cond}} = \{(X^{(i)}, C^{(i)})\}_{i=1}^n$, let $\widehat{P}_{0|c}$ denote the endpoint law of the learned conditional reverse-time sampler run under $C = c$. Define the average conditional risk by

$$\mathbb{E}_C[\mathsf{d}(\widehat{P}_{0|C}, P_{0|C})],$$

where $\mathsf{d}(\cdot,\cdot)$ may be one of the distributional discrepancies introduced in Section 5.1, such as total variation or Wasserstein distance.

Suppose that $X_0 \in \mathbb{R}^d$ and $C \in [0,1]^{d_c}$. Fu et al. (2024) consider conditional densities of the form $p_0(x \mid c) = \mathrm{e}^{-\kappa\|x\|^2/2} f(x,c)$, where $\kappa > 0$ and f is jointly Hölder smooth of order ν in (x,c), uniformly bounded, and bounded away from zero. With their prescribed score-network and time-horizon choices, the early-stopped continuous-time conditional sampler satisfies

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{cond}}}\mathbb{E}_C[d_{\mathrm{TV}}(\widehat{P}_{0|C}, P_{0|C})] = \widetilde{\mathcal{O}}(n^{-\nu/(2\nu+d+d_c)}).$$

Tang et al. (2025) consider $X_0 \in [-1,1]^d$ and $C \in [-1,1]^{d_c}$, with a conditional density $p_0(x \mid c)$ that is anisotropically Hölder smooth of orders ν_x and ν_c in the x- and c-directions, respectively, and uniformly bounded away from zero. The truncated output of their early-stopped continuous-time sampler satisfies, with probability at least $1 - n^{-1}$ over the training sample,

$$\mathbb{E}_C[d_{\mathrm{TV}}(\widehat{P}_{0|C}, P_{0|C})] = \widetilde{\mathcal{O}}(n^{-1/(2+d/\nu_x+d_c/\nu_c)}),$$

and

$$\mathbb{E}_C[W_1(\widehat{P}_{0|C}, P_{0|C})] = \widetilde{\mathcal{O}}(\max\{n^{-1/(2+d_c/\nu_c)}, n^{-(1+1/\nu_x)/(2+d/\nu_x+d_c/\nu_c)}\}).$$

When $\nu_x = \nu_c = \nu$, their TV bound has the same polynomial order in n as that of Fu et al. (2024), although the distributional assumptions and the forms of the guarantees differ. They also show that, under the low-dimensional structural conditions (a compact covariate support satisfying a local dimension bound, smoothly varying conditional manifolds, and smooth conditional densities bounded away from zero relative to manifold volume), the convergence rate for conditional distribution estimation in W_1 depends on intrinsic rather than ambient dimensions.

6.2.2. Guidance error and the induced distribution

Bayes' rule gives

$$s(x,t) + \nabla_x \log p_t(c \mid x) = s(x,t;c).$$

This identity shows that classifier guidance with $\gamma = 1$ and CFG with $\omega = 0$ both use the conditional score $s(x,t;c)$. In either case, the corresponding reverse process targets $P_{0|c}$ (exact recovery assumes initialization from $P_{T|c}$; using a reference distribution incurs initialization error, as in Section 5.4.2).

For classifier guidance, a classifier is trained to estimate the conditional class probability $p_t(c \mid x)$ through $\widehat{p}_\phi(c \mid x, t)$. The reverse sampler uses the gradient

$$\nabla_x \log \widehat{p}_\phi(c \mid x, t).$$

However, accurate estimation of $p_t(c \mid x)$ does not automatically imply accurate estimation of this gradient. Sahu et al. (2026) study this issue for bounded data support and a target label of positive probability. For twice continuously differentiable classifiers satisfying noise-dependent bounds on the gradient and Laplacian of their log probabilities, they bound the mean squared guidance-gradient error in terms of the classifier's excess cross-entropy risk at each noise level. They further incorporate the accumulated guidance error into a KL bound for the discretized sampler.

When $\gamma \neq 1$ or $\omega > 0$, the guided reverse process may produce a distribution different from $P_{0|c}$. Classifier guidance with $\gamma \neq 1$ uses

$$s(x, t) + \gamma \nabla_x \log p_t(c \mid x).$$

For CFG with $\omega > 0$, the score used by the reverse process is

$$s_{\mathrm{CFG},\omega}(x, t; c) = (1 + \omega)s(x, t; c) - \omega s(x, t) = \nabla_x \log \left\{ \frac{p_t(x \mid c)^{1+\omega}}{p_t(x)^\omega} \right\}.$$

For each t such that the normalizing constant is finite, define

$$\widetilde{p}_{t,\omega}(x \mid c) = \frac{p_t(x \mid c)^{1+\omega} p_t(x)^{-\omega}}{\displaystyle\int p_t(z \mid c)^{1+\omega} p_t(z)^{-\omega} \, \mathrm{d}z}.$$

Then

$$s_{\mathrm{CFG},\omega}(x, t; c) = \nabla_x \log \widetilde{p}_{t,\omega}(x \mid c).$$

In particular,

$$\widetilde{p}_{0,\omega}(x \mid c) \propto p_0(x \mid c)^{1+\omega} p_0(x)^{-\omega}.$$

Let $\widetilde{P}_{0,\omega|c}$ denote the distribution with density $\widetilde{p}_{0,\omega}(\cdot \mid c)$; this is a natural candidate for the distribution targeted by CFG. Applying the forward noising process to this candidate gives a density $p^{\mathrm{fwd}}_{t,\omega}(\cdot \mid c)$ at time t:

$$\widetilde{p}_{0,\omega}(\cdot \mid c) \xrightarrow{\text{forward noising}} p^{\mathrm{fwd}}_{t,\omega}(\cdot \mid c).$$

A reverse process targeting $\widetilde{P}_{0,\omega|c}$ would use the score $\nabla_x \log p^{\mathrm{fwd}}_{t,\omega}(x \mid c)$ at time t, whereas CFG uses $\nabla_x \log \widetilde{p}_{t,\omega}(x \mid c)$. The density $\widetilde{p}_{t,\omega}(\cdot \mid c)$ is obtained by first noising the conditional and unconditional distributions and then reweighting

$$(p_0(\cdot \mid c), p_0) \xrightarrow{\text{forward noising}} \big(p_t(\cdot \mid c), p_t\big) \xrightarrow{\text{reweighting}} \widetilde{p}_{t,\omega}(\cdot \mid c).$$

Because forward noising and reweighting generally do not commute,

$$p^{\mathrm{fwd}}_{t,\omega}(\cdot \mid c) \neq \widetilde{p}_{t,\omega}(\cdot \mid c).$$

We give an example to illustrate this point.

Example 6.1: (Forward noising and reweighting do not commute). Consider the Gaussian construction of Bradley and Nakkiran (2025). Let C, ε be independent standard Gaussian variables and $X_0 = C + \varepsilon$. Then $P_0 = \mathcal{N}(0, 2)$ and $P_{0|C=0} = \mathcal{N}(0, 1)$. Consider the OU forward process

$$X_t = \mathrm{e}^{-t/2} X_0 + \sqrt{1 - \mathrm{e}^{-t}} Z,$$

where $Z \sim \mathcal{N}(0, 1)$ is independent of (X_0, C).

Reweighting followed by forward noising. For $\omega > 0$,

$$\widetilde{p}_{0,\omega}(x \mid 0) \propto p_0(x \mid 0)^{1+\omega} p_0(x)^{-\omega} \propto \exp\left[-\frac{x^2}{2}\left(1 + \omega - \frac{\omega}{2}\right)\right].$$

Thus, the reweighted distribution is Gaussian with mean zero and variance $\{(1 + \omega) - \omega/2\}^{-1} = 2/(\omega + 2)$. Forward noising gives

$$p^{\mathrm{fwd}}_{t,\omega}(\cdot \mid 0) = \mathcal{N}\left(0, \mathrm{e}^{-t}\frac{2}{\omega + 2} + 1 - \mathrm{e}^{-t}\right) = \mathcal{N}\left(0, 1 - \frac{\omega}{\omega + 2}\mathrm{e}^{-t}\right).$$

Forward noising followed by reweighting. The noisy marginals are $P_t = \mathcal{N}(0, 1 + \mathrm{e}^{-t})$ and $P_{t|C=0} = \mathcal{N}(0, 1)$. Reweighting gives a zero-mean Gaussian with variance

$$\left\{(1 + \omega) - \frac{\omega}{1 + \mathrm{e}^{-t}}\right\}^{-1} = \frac{1 + \mathrm{e}^{-t}}{1 + (1 + \omega)\mathrm{e}^{-t}},$$

and hence

$$\widetilde{p}_{t,\omega}(\cdot \mid 0) = \mathcal{N}\left(0, \frac{1 + \mathrm{e}^{-t}}{1 + (1 + \omega)\mathrm{e}^{-t}}\right).$$

The difference between the two variances is

$$\frac{\omega(1 + \omega)\mathrm{e}^{-t}(1 - \mathrm{e}^{-t})}{(\omega + 2)\{1 + (1 + \omega)\mathrm{e}^{-t}\}} > 0 \quad (\omega > 0,\ 0 < t < \infty).$$

Thus, the two operations produce different distributions at every finite positive time. When $\omega = 0$, both constructions recover the same conditional marginal.

Example 6.1 shows that the CFG score need not coincide with the score of the forward-noised reweighted distribution. The standard time-reversal argument therefore does not guarantee recovery of $\widetilde{P}_{0,\omega|c}$. Going beyond this intermediate-time comparison, Bradley and Nakkiran (2025) explicitly compute the final CFG output laws in this Gaussian setting: DDPM and DDIM samplers driven by the same CFG score can yield different endpoint laws, neither of which equals $\widetilde{P}_{0,\omega|c}$. Thus, the CFG output is sampler-dependent and need not equal $\widetilde{P}_{0,\omega|c}$. They further interpret continuous-time CFG as the combination of a conditional predictor and a Langevin corrector.

Chidambaram et al. (2024) analyse the effect of guidance strength on probability-flow ODE sampling in two one-dimensional mixture models with compactly supported or Gaussian components. They show that stronger guidance concentrates the output distribution in increasingly extreme regions associated with the target component and can magnify the effect of score-estimation error on the output distribution.

7. Discussion

In recent years, diffusion-based generative models have emerged as a dominant paradigm in statistical machine learning, demonstrating remarkable versatility across a broad spectrum of complex downstream tasks. In tabular and structured domain settings, diffusion samplers have been applied to synthetic data generation (Kim et al., 2023; Kotelnikov et al., 2023), with empirical evaluations focussed on reproducing joint dependencies and marginal distributions. For sequential and temporal data, they facilitate probabilistic forecasting under highly non-linear and multi-modal dynamics (Rasul et al., 2021; Shen & Kwok, 2023). Furthermore, they provide conditional generative models for missing-data imputation (Liu et al., 2023; Tashiro et al., 2021) by modelling conditional distributions over corrupted observation spaces. Beyond purely generative applications, diffusion models have revolutionized the solution of ill-posed linear and non-linear inverse problems (Chung et al., 2023; Kawar et al., 2022), serving as powerful, data-driven regularizers.

Across these diverse domains, the fundamental statistical inquiry remains the same: does the distribution induced by the reverse-time sampler rigorously approximate the true, underlying target distribution? Quantifying this approximation error, typically measured in terms of the Wasserstein distance, Kullback-Leibler divergence, or total variation, requires a meticulous mathematical analysis of several interacting systems. This review systematically deconstructs these theoretical guarantees, focussing on (i) the propagation of score estimation errors (typically bounded in L^2) through the reverse dynamics; (ii) the discretization error introduced by numerical reverse-time SDE (Lee et al., 2023) and probability-flow ODE solvers (Gao & Zhu, 2025; Huang et al., 2025); and (iii) the structural and geometric assumptions, such as the manifold hypothesis or bounded support (Oko et al., 2023; Tang & Yang, 2024; Zhang et al., 2024). Functional inequalities, such as the log-Sobolev inequality, also yield convergence guarantees under different regimes (Yang & Wibisono, 2022).

This foundational inquiry into distributional convergence becomes increasingly critical, and mathematically challenging, in complex statistical regimes characterized by auxiliary conditioning, partial observability, heterogeneous data modalities, and domain shifts. Extending both the algorithmic frameworks and the corresponding non-parametric statistical theory of diffusion models to these settings represents a highly active research frontier. For instance, in conditional generation and guided sampling, the target distribution is a conditional slice $P_{0|c}$, requiring a precise analysis of how errors in guided score matching propagate under classifier or classifier-free guidance schemes (Chidambaram et al., 2024; Sahu et al., 2026). Similarly, solving inverse problems and recovering missing data require diffusion models to function as data-driven priors (Chung et al., 2023; Song et al., 2022), necessitating rigorous bounds on posterior contraction rates under partial, noisy measurements. Furthermore, translating distributions across domains or aligning heterogeneous data sources shifts the paradigm from standard diffusion processes to Schrödinger bridges, introducing complex boundary-value problems on path spaces that demand novel statistical analysis (Chen et al., 2022; De Bortoli et al., 2021).

A critically complementary direction concerns the downstream consequences of integrating synthetic samples into scientific, clinical, and statistical workflows. While generative models are frequently deployed for data augmentation, a fundamental challenge is ensuring that downstream estimators trained on synthetic surrogates preserve the statistical validity of the original data-generating process. This requires quantifying how synthetic data impacts classical inferential properties, such as the nominal coverage of confidence intervals, type-I

error control in hypothesis testing, and asymptotic efficiency. This concern is further amplified by the emerging threat of 'model collapse' or 'self-consuming loops', where subsequent generations of models trained recursively on their own synthetic outputs suffer from progressive losses of sample quality and/or diversity (Alemohammad et al., 2024; Shumailov et al., 2024).

Taken together, these directions broaden the statistical study of diffusion models beyond the approximation of a single unconditional target distribution. A comprehensive theory must also account for conditional and partially observed distributions, the validity of downstream procedures based on synthetic data, and the stability of recursive training on model-generated samples. Developing guarantees that connect score estimation, sampling dynamics, distributional accuracy, and downstream statistical validity will be essential for deploying diffusion models as reliable tools in scientific and high-stakes applications.

Author contributions

CRediT: **Jin Su:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing; **Shuyi Zhang:** Conceptualization, Writing – review & editing; **Yong Zhou:** Conceptualization, Writing – review & editing; **Jian Huang:** Conceptualization, Methodology, Writing – review & editing

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Jin Su http://orcid.org/0009-0005-7255-6010
Shuyi Zhang http://orcid.org/0000-0002-8039-6636
Yong Zhou http://orcid.org/0000-0001-9751-3141
Jian Huang http://orcid.org/0000-0002-5218-9269

References

Aamari, E., & Stéphanovitch, A. (2026). Statistical analysis of Markovian generative modeling. arXiv: 2604.22712.

Ahsan, M. M., Raman, S., Liu, Y., & Siddique, Z. (2025). A comprehensive survey on diffusion models and their applications. *Applied Soft Computing*, *181*, 113470.

Alain, G., & Bengio, Y. (2014). What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research*, *15*(110), 3743–3773.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. (2024). Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*.

Anderson, B. D. O. (1982). Reverse-time diffusion equation models. *Stochastic Processes and Their Applications*, *12*(3), 313–326.

Block, A., Mroueh, Y., & Rakhlin, A. (2020). Generative modeling with denoising auto-encoders and Langevin sampling. arXiv: 2002.00107.

Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2022). Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7327–7347.

Bradley, A., & Nakkiran, P. (2025). Classifier-free guidance is a predictor-corrector. *Transactions on Machine Learning Research*. https://openreview.net/forum?id = zrWNtzSZsf

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P. A., & Li, S. Z. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 2814–2830.
Cao, P., Zhou, F., Song, Q., & Yang, L. (2026). Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *48*(4), 4771–4791.
Chan, S. H. (2024). Tutorial on diffusion models for imaging and vision. *Foundations and Trends in Computer Graphics and Vision*, *16*(4), 322–471.
Chang, Z., Koulieris, G. A., Chang, H. J., & Shum, H. P. H. (2026). On the design fundamentals of diffusion models: A survey. *Pattern Recognition*, *169*, 111934.
Chen, H., Lee, H., & Lu, J. (2023). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 4735–4763). PMLR.
Chen, M., Huang, K., Zhao, T., & Wang, M. (2023). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 4672–4712). PMLR.
Chen, M., Mei, S., Fan, J., & Wang, M. (2024). Opportunities and challenges of diffusion models for generative AI. *National Science Review*, *11*(12), nwae348.
Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., & Ray, A. (2021). Evaluating large language models trained on code. arXiv: 2107.03374.
Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 6571–6583).
Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., & Zhang, A. R. (2023). Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*.
Chen, S., Daras, G., & Dimakis, A. (2023). Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 4462–4484). PMLR.
Chen, T., Liu, G. H., & Theodorou, E. A. (2022). Likelihood training of Schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*.
Chidambaram, M., Gatmiry, K., Chen, S., Lee, H., & Lu, J. (2024). What does guidance do? A fine-grained analysis in a simple setting. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 84968–85005).
Chung, H., Kim, J., McCann, M. T., Klasky, M. L., & Ye, J. C. (2023). Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*.
Conforti, G., Durmus, A., & Gentiloni Silveri, M. (2025). KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, *7*(1), 86–109.
Croitoru, F. A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850–10869.
De Bortoli, V. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*. https://openreview.net/forum?id = MhK5aXo3gB
De Bortoli, V., Thornton, J., Heng, J., & Doucet, A. (2021). Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 17695–17709).
Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 8780–8794).
Ding, Z., & Jin, C. (2024). A practical handbook. arXiv: 2412.17162.
Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using real NVP. In *International Conference on Learning Representations*.
Dou, Z., Kotekal, S., Xu, Z., & Zhou, H. H. (2024). From optimal score matching to optimal sampling. arXiv: 2409.07032.
Du, Y., & Mordatch, I. (2019). Implicit generation and modeling with energy-based models. In *Advances in Neural Information Processing Systems* (Vol. 32, pp. 3603–3613).
Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, *106*(496), 1602–1614.

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, *322*(8), 549–560.
Einstein, A. (1926). On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat (A.D. Cowper, Trans.). In R. Fürth (Ed.), *Investigations on the theory of the Brownian movement* (pp. 1–18). Methuen & Co. Ltd. Reprinted by Dover Publications, New York, in 1956.
Esser, P., Kulal, S., Blattmann, A., Entezari, R., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., & Podell, D. (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 12606–12633). PMLR.
Fan, J., Gu, Y., & Li, X. (2025). Optimal estimation of a factorizable density using diffusion models with ReLU neural networks. arXiv: 2510.03994.
Feller, W. (1954). Diffusion processes in one dimension. *Transactions of the American Mathematical Society*, *77*(1), 1–31.
Fokker, A. D. (1914). Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Annalen der Physik*, *348*(5), 810–820.
Fu, H., Yang, Z., Wang, M., & Chen, M. (2024). Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. arXiv: 2403.11968.
Gao, X., Nguyen, H. M., & Zhu, L. (2025). Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, *26*(43), 1–54.
Gao, X., & Zhu, L. (2025). Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 1009–1017). PMLR.
Gibbs, A. L., & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, *70*(3), 419–435.
Gonzalez, M., Fernandez Pinto, N., Tran, T. V. D., Gherbi, E., Hajri, H., & Masmoudi, N. (2023). SEEDS: exponential SDE solvers for fast high-quality sampling from diffusion models. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 68061–68120).
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Vol. 27).
Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., & Duvenaud, D. (2019). FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations (ICLR)*.
Grathwohl, W., Wang, K. C., Jacobsen, J. H., Duvenaud, D., Norouzi, M., & Swersky, K. (2020). Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*.
Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*, 723–773.
Han, Y., Razaviyayn, M., & Xu, R. (2024). Neural network-based score estimation in diffusion models: Optimization and generalization. In *International Conference on Learning Representations*.
Haussmann, U. G., & Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, *14*(4), 1188–1205.
Higham, C. F., Higham, D. J., & Grindrod, P. (2025). Diffusion models for generative artificial intelligence: An introduction for applied mathematicians. *SIAM Review*, *67*(3), 607–623.
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 6840–6851).
Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv: 2207.12598.
Holderrieth, P., & Erives, E. (2025). An introduction to flow matching and diffusion models. arXiv: 2506.02070.
Hoogeboom, E., Garcia Satorras, V., Vignac, C., & Welling, M. (2022). Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 8867–8887). PMLR.
Huang, D. Z., Huang, J., & Lin, Z. (2025). Convergence analysis of probability flow ODE for score-based generative models. *IEEE Transactions on Information Theory*, *71*(6), 4581–4601.

Hutchinson, M. F. (1989). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, *18*(3), 1059–1076.
Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, *6*, 695–709.
Itô, K. (1944). Stochastic integral. *Proceedings of the Imperial Academy*, *20*(8), 519–524.
Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 26565–26577).
Kawar, B., Elad, M., Ermon, S., & Song, J. (2022). Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 23593–23606).
Kim, J., Lee, C., & Park, N. (2023). STaSy: Score-based tabular data synthesis. In *International Conference on Learning Representations*.
Kingma, D. P., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 21696–21707).
Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.
Kolmogoroff, A. N. (1931). Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, *104*(1), 415–458.
Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2021). DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*.
Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 17564–17579). PMLR.
Kumar, S., Nagaraj, D., & Sarkar, P. (2025). Dimension-free score matching and time bootstrapping for diffusion models. In *Advances in Neural Information Processing Systems* (Vol. 38, pp. 32224–32277).
Kwon, H. K., Kim, D., Ohn, I., & Chae, M. (2026). Nonparametric estimation of a factorizable density using diffusion models. *Journal of Machine Learning Research*, *27*(22), 1–125.
Lai, C. H., Song, Y., Kim, D., Mitsufuji, Y., & Ermon, S. (2025). The principles of diffusion models. arXiv: 2510.21890.
Langevin, P. (1908). Sur la théorie du mouvement brownien. *Comptes Rendus Hebdomadaires des Séances de L'Académie Des Sciences*, *146*, 530–533.
Lee, H., Lu, J., & Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 22870–22882).
Lee, H., Lu, J., & Tan, Y. (2023). Convergence of score-based generative modeling for general data distributions. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory* (Vol. 201, pp. 946–985). PMLR.
Li, T., & He, K. (2026). Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 36115–36125).
Liu, L., Ren, Y., Lin, Z., & Zhao, Z. (2022). Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*.
Liu, M., Huang, H., Feng, H., Sun, L., Du, B., & Fu, Y. (2023). PriSTI: A conditional diffusion framework for spatiotemporal imputation. In *2023 IEEE 39th International Conference on Data Engineering* (pp. 1927–1939).
Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022). DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 5775–5787).
Lu, J. (2026). A mathematical introduction to diffusion models. arXiv: 2607.01693.
Luo, C. (2022). Understanding diffusion models: A unified perspective. arXiv: 2208.11970.
Mimikos-Stamatopoulos, N., Zhang, B. J., & Katsoulakis, M. A. (2024). Score-based generative models are provably robust: An uncertainty quantification perspective. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 63154–63183).
Mohamed, S., & Lakshminarayanan, B. (2016). Learning in implicit generative models. arXiv: 1610.03483.
Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, *29*(2), 429–443.

Nakkiran, P., Bradley, A., Zhou, H., & Advani, M. (2025). Step-by-step diffusion: An elementary tutorial. *Foundations and Trends in Computer Graphics and Vision*, *17*(1), 1–75.

Oko, K., Akiyama, S., & Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 26517–26582). PMLR.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications* (6th ed.). Springer.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, *22*(57), 1–64.

Pedrotti, F., Maas, J., & Mondelli, M. (2024). Improved convergence of score-based diffusion models via prediction-correction. *Transactions on Machine Learning Research*. https://openreview.net/forum?id = 0zKvH7YiAq

Rasul, K., Seward, C., Schuster, I., & Vollgraf, R. (2021). Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8857–8868). PMLR.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., & Prudden, R. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677.

Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 1530–1538). PMLR.

Risken, H. (1996). *The Fokker-Planck Equation: Methods of Solution and Applications* (2nd ed.). Springer.

Robbins, H. (1956). An empirical Bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1: Contributions to the Theory of Statistics, pp. 157–163). University of California Press.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. *GAMM-Mitteilungen*, *44*(2), e202100008.

Sahu, S., Banerjee, A., & Wu, Y. (2026). Provably reliable classifier guidance via cross-entropy control. arXiv: 2601.21200.

Saremi, S., & Hyvärinen, A. (2019). Neural empirical Bayes. *Journal of Machine Learning Research*, *20*(181), 1–23.

Shen, L., & Kwok, J. (2023). Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 31016–31029). PMLR.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759.

Smoluchowski, M. von. (1906). Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Annalen der Physik*, *326*(14), 756–780.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 2256–2265). PMLR.

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems* (Vol. 32).

Song, Y., Garg, S., Shi, J., & Ermon, S. (2020). Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence* (Vol. 115, pp. 574–584). PMLR.

Song, Y., Shen, L., Xing, L., & Ermon, S. (2022). Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Suh, N., & Cheng, G. (2025). A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. *Annual Review of Statistics and Its Application*, *12*, 177–207.

Tang, R., Lin, L., & Yang, Y. (2025). Conditional diffusion models are minimax-optimal and manifold-adaptive for conditional distribution estimation. In *International Conference on Learning Representations*.

Tang, R., & Yang, Y. (2024). Adaptivity of diffusion models to manifold structures. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics* (Vol. 238, pp. 1648–1656). PMLR.

Tang, W., & Zhao, H. (2025). Score-based diffusion models via stochastic differential equations. *Statistics Surveys*, *19*, 28–64.

Tashiro, Y., Song, J., Song, Y., & Ermon, S. (2021). CSDI: conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 24804–24816).

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., & Bermano, A. H. (2023). Human motion diffusion model. In *International Conference on Learning Representations*.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Uehara, M., Zhao, Y., Wang, C., Li, X., Regev, A., Levine, S., & Biancalani, T. (2025). Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. arXiv: 2501.09685.

Villani, C. (2009). *Optimal Transport: Old and New* (Vol. 338). Springer.

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

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., & Wicky, B. I. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, *620*(7976), 1089–1100.

Wibisono, A., Wu, Y., & Yang, K. Y. (2024). Optimal score estimation via empirical Bayes smoothing. In *Proceedings of Thirty Seventh Conference on Learning Theory* (Vol. 247, pp. 4958–4991). PMLR.

Xue, S., Yi, M., Luo, W., Zhang, S., Sun, J., Li, Z., & Ma, Z. M. (2023). SA-solver: Stochastic Adams solver for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 77632–77674).

Yang, K. Y., & Wibisono, A. (2022). Convergence of the inexact Langevin algorithm in KL divergence with application to score-based generative models. arXiv: 2211.01512.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang M. H. (2024). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, *56*(4), 105:1–105:39.

Yeğin, M. N., & Amasyalı, M. F. (2024). Generative diffusion models: A survey of current theoretical developments. *Neurocomputing*, *608*, 128373.

Zhan, Z., Chen, D., Mei, J. P., Zhao, Z., Chen, J., Chen, C., Lyu, S., & Wang, C. (2025). Conditional image synthesis with diffusion models: A survey. *Transactions on Machine Learning Research*. https://openreview.net/forum?id = ewwNKwh6SK

Zhang, K., Yin, H., Liang, F., & Liu, J. (2024). Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 60134–60178). PMLR.

Zhang, Q., & Chen, Y. (2023). Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*.

Zhao, W., Bai, L., Rao, Y., Zhou, J., & Lu, J. (2023). UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 49842–49869).