DDPM-BASED GENERATION OF TILTED SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Given n independent samples from a d-dimensional probability distribution, our aim is to generate diffusion-based samples from a distribution obtained by tilting the original, where the degree of tilt is parameterized by $\theta \in \mathbb{R}^d$. We define a plug-in estimator and show that it is minimax-optimal. We develop an algorithm that, given original samples, generates exponentially tilted samples using a kerneled score estimator. The sample complexity of this algorithm is provably polynomial for bounded random variables, and avoids the curse of dimensionality. The proof uses tight Wasserstein bounds between the plug-in estimator distribution and the true distribution as a function of θ and n, and a novel score estimation bound. Furthermore, we also demonstrate sample complexity bounds for neural network-learned scores. Our theoretical results are supported by extensive simulations. Applications of our work include finance, weather and climate modeling, and many other domains, where the aim may be to generate samples from a tilted distribution that satisfies practically motivated moment constraints.

1 INTRODUCTION

Generative AI is a growing subfield of research responsible for the development of state-of-the-art video, text, and content generation. It typically uses samples from a high-dimensional distribution to generate more independent samples from the same distribution, with the samples often corresponding to images or text (Dhariwal & Nichol, 2021). Recently, there has also been research on the underlying samples corresponding to high-dimensional data in weather or climate modeling (Li et al., 2024; Ling et al., 2024) or in finance (see (Cont et al., 2025)).

In many practically important settings, it may happen that samples are available under one distribution, but our interest may be in generating samples from a related *tilted* distribution. To be more precise, let X be a random vector in $\mathbb{R}^d$. Given a vector $\theta \in \mathbb{R}^d$, the random vector X_θ is called an *exponential tilt* of X by the vector θ if

$$\mathbb{P}(X_\theta \in A) = \frac{\mathbb{E}[e^{\theta^T g(X)} \mathbf{1}_{X \in A}]}{\mathbb{E}[e^{\theta^T g(X)}]} \tag{1}$$

for all $A \subset \mathbb{R}^d$, and some suitable function g which can be considered a site-specific tilt.

There are several important applications of tilting. In portfolio optimization (Meucci, 2010) and option pricing, (Buchen & Kelly, 1996; Stutzer, 1996; Avellaneda, 1998), exponentially tilted versions of pricing distributions are sought to better model asset prices. Tilting of random variables or vectors is also a common technique in the field of rare event sampling and Monte Carlo simulation, having applications to phase transitions in physical phenomena (Lee et al., 2025), rare event sampling from discrete Markov processes (Aguilar & Gatto, 2024), and finding robust best response strategies in security games (Kong et al., 2025). It also appears as a solution to some well-motivated optimization problems (see (Csiszár, 1967; 2008)). We refer the reader to (Alvo, 2022, Chapter 1) for more applications of exponential tilting.

Previous works have largely approached tilted or biased sampling through importance sampling or MCMC biasing schemes, but these typically require some knowledge of the target density. More recently, diffusion-based methods have been proposed to enable sampling from tilted distributions via score guidance methods (see (Wang et al., 2024) and references therein), diffusion posterior

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sampling (DPS) (Chung et al., 2024) and loss-guided diffusion (Song et al., 2023). Diffusion is primarily used since it has a higher sample diversity than previous generative mechanisms.

However, each of these methods assume a differentiable tilting function and use alternate algorithms to produce tilted samples by diffusion. Furthermore, they typically rely on heuristic approximations of the score update. Consequently, theoretical support for these frameworks remains limited.

In this article, we consider samples from an unknown bounded high-dimensional probability distribution X, and construct an algorithm that uses diffusion to generate nearly tilted samples with high probability. In a large number of examples, we prove that our algorithm has a polynomial sample complexity in the support size, the parameter θ, and the desired accuracy. The proposed algorithm proceeds in two steps. We begin with i.i.d. samples $X_i, i = 1, 2, \ldots, n$ from X. We shall achieve the following objectives in this article.

- We create an estimator $X_{n,\theta}$ of X_θ and show that it is minimax in a particular sense.
- We provide Wasserstein bounds between X_θ and $X_{n,\theta}$ when X is bounded, and justify the choice of Wasserstein distance and the boundedness assumption.
- Using the DDPM framework in (Li & Yan, 2025) we construct a reverse process which, when instantiated from a normal and run according to a score estimator either created from $X_{n,\theta}$ or through a neural network), generates samples close to X_θ under suitable assumptions.
- We demonstrate examples for which the algorithm has polynomial sample complexity in the tilt.

The random vector $X_{n,\theta}$ referenced earlier has the form

$$\mathbb{P}(X_{n,\theta} \in A) = \frac{\frac{1}{n}\sum_{i=1}^n e^{\theta^T g(X_i)} \mathbf{1}_{X_i \in A}}{\frac{1}{n}\sum_{i=1}^n e^{\theta^T g(X_i)}}. \tag{2}$$

We can rewrite $X_{n,\theta}$ as

$$X_{n,\theta} = \sum_{i=1}^n \gamma_{\theta,i}\delta_{X_i}, \quad \gamma_{\theta,i} = \frac{e^{\theta^T g(X_i)}}{\sum_{i=1}^n e^{\theta^T g(X_i)}}. \tag{3}$$

Along with these objectives, we obtain multiple statements of independent interest which we elaborate on in the respective sections.

We briefly remark on related literature. The estimator $X_{n,\theta}$ is known in the literature as an SNIS (self-normalized importance sampler) (Owen, 2000), or a weighted importance sampler (Rubinstein, 1981). To our knowledge, the accuracy of the estimator in the Wasserstein distance has not been studied, although (Chatterjee & Diaconis, 2018) consider it in the L^1 distance. Its minimaxity appears to not have been considered before.

Finally, a recent article (Moitra et al., 2026) rigorously proves that generating from quadratically tilted distributions with polynomial sample complexity is impossible (if NP $\neq$ co-NP). However, this does not contradict our article, since our algorithm does not assert polynomial sample complexity uniformly. It only does so in a large, but limited number of cases.

The question of whether diffusion can generate accurate samples has long been a field of study (Chen et al., 2023b; Benton et al., 2024; Anonymous, 2026; Gao et al., 2025). The most recent advancement in this direction is that of (Li & Yan, 2025), whose framework we will use in this article. Of the listed citations, the latter two study diffusion accuracy in the Wasserstein distance. The former work, however, employs a log-concavity assumption, while the latter does not consider an initial perturbation of the data. Consequently, our work does not intersect significantly with these.

Notation: For a vector $x = (x_1, \ldots, x_d)$, the L^p norm for $p \geq 1$ is defined by $\|x\|_p^p = \sum_{i=1}^d |x_i|^p$. For a metric space $(\mathcal{X}, d)$ and $S \subset \mathcal{X}$, let $\text{diam}(S) = \sup_{x,y \in S} d(x, y)$ denote the diameter of S. The Dirac probability measure at a point $x \in \mathbb{R}^d$ will be denoted by δ_x. Let $B(x, r) = \{y \in \mathbb{R}^d : \|y - x\| < r\}$ and $B_r = B(0, r)$.

2 On the minimaxity of $X_{n,\theta}$

In this section, we justify our choice of estimator $X_{n,\theta}$ by demonstrating its asymptotic minimaxity. Informally, this means that for any other estimator, one can find a random vector X for which $X_{n,\theta}$ approximates X_θ better than the estimator does.

We proceed to define the notion of minimaxity: note that we place importance in the CDFs of random variables, rather than the random variables themselves. Let $C_{a,b}$ be the class of all 1-dimensional continuous CDFs supported on the set $[a, b]$. Let $w_\theta(x) = e^{\theta g(x)}$ for some $\theta \in \mathbb{R}$ and site-specific tilt function $g : \mathbb{R} \to \mathbb{R}$. Given iid samples $X_1, X_2, \cdots, X_n$ of a random variable X with CDF F, let F_θ, $F_{n,\theta}$ be the CDF of X_θ, $X_{n,\theta}$ respectively.

Let G be any subspace of real-valued functions on $\mathbb{R}$, with $\mathcal{C}_{a,b} \subset G$. Let B be the smallest Borel field on G such that every element of $\mathcal{C}_{a,b}$ is an element of B and such that, for every positive integer k and all sets of real numbers $\{t_1, \cdots, t_k\}$ and $\{a_1, \cdots, a_k\}$ with $t_1 < \cdots < t_k$, the set $\{g \,|g \in G; g(t_1) < a_1, \cdots, g(t_k) < a_k\}$ is in B. Let D_n be the class of all real-valued functions ϕ_n on $B \times R^n$ with the properties: for each $x^{(n)} \in \mathbb{R}^n$, $\phi_n(\cdot; x^{(n)})$ is a probability measure on G; and for each $\Delta \in B$, $\phi_n(\Delta; \cdot)$ is Borel-measurable on R^n.

For a random estimator $g \sim \phi \in D_n$ which attempts to estimate F_θ given F, define the error

$$r_n^*(F, g) = \mathbb{P}_F \left\{ \sup_x |g(x) - F_\theta(x)| > \frac{r}{\sqrt{n}} \right\}.$$

Let $r_n(F) = \mathbb{P}_F(\sqrt{n}\|F_{n,\theta} - F_\theta\| > r)$ be the CDF estimator for $g = F_{n,\theta}$.

Theorem 1 (Asymptotic Minimaxity)**.** *For every value $r > 0$,*

$$\lim_{n\to\infty} \frac{\sup_{F\in\mathcal{C}_{a,b}} r_n(F)}{\inf_{\phi\in D_n} \sup_{F\in\mathcal{C}_{a,b}} r_n^*(F, g)} = \lim_{n\to\infty} \frac{\sup\limits_{F\in\mathcal{C}_{a,b}} P_F \left\{ \sup_x |F_{n,\theta}(x) - F_\theta(x)| > \frac{r}{\sqrt{n}} \right\}}{\inf\limits_{\phi\in D_n} \sup\limits_{F\in\mathcal{C}_{a,b}} P_F \left\{ \sup_x |g(x) - F_\theta(x)| > \frac{r}{\sqrt{n}} \right\}} = 1.$$

where g is a random CDF estimator sampled from ϕ.

We discuss the relevance of this result and prove it in Appendix A.5. Having justified the importance of the estimator (2), we proceed to study its Wasserstein accuracy.

3 On Wasserstein accuracy of $X_{n,\theta}$

For a vector $\theta \in \mathbb{R}^d$ and a d-dimensional distribution X such that $\mathbb{E}[e^{\theta^T X}] < \infty$, recall X_θ from (1). Now, given i.i.d. samples $X_1, X_2, \ldots, X_n$ from X, we recall $X_{n,\theta}$ from (2). Note that $X_{n,\theta}$ is a random measure since the samples X_i are random. Throughout this article, we denote the expectation over $X_{n,\theta}$ by $\mathbb{E}_n$, and the expectation over other sources of randomness by $\mathbb{E}$.

3.1 Main result

In order to present our main result of this section, we define the p-Wasserstein distance W_p for $p \geq 1$ in generality, and refer the reader to (Villani, 2009; Peyré & Cuturi, 2019) for more details on these distances.

Definition 1. *Let $(\mathcal{X}, d)$ be a metric space and let $\mathcal{B}(\mathcal{X})$ be the Borel sigma-algebra. Given random elements Y_1, Y_2 taking values in X, a coupling Z of Y_1 and Y_2 is a random element on $\mathcal{X} \times \mathcal{X}$ such that*

$$\mathbb{P}(Z \in A \times \mathcal{X}) = \mathbb{P}(Y_1 \in A), \text{ and } \mathbb{P}(Z \in \mathcal{X} \times B) = \mathbb{P}(Y_2 \in B) \tag{4}$$

for all $A, B \in \mathcal{B}(\mathcal{X})$. We write $Z = (Z_1, Z_2)$ to represent the first and second components. Let Π be the set of all couplings of Y_1 and Y_2. Define for $p \geq 1$,

$$(W_p^{\mathcal{X}})(Y_1, Y_2) = \inf_{(Z_1,Z_2)\in\Pi} \mathbb{E}[d(Z_1, Z_2)^p]^{1/p}. \tag{5}$$

For the special case $\mathcal{X} = \mathbb{R}^d$ and $d(x, y) = \|x - y\|_2$, we write W_p in place of $W_p^{\mathcal{X}}$ to denote the p-Wasserstein distance between two d-dimensional random vectors under the standard Euclidean metric.

Note that if Z_1, Z_2 are any random elements, and $Z = (Z_1, Z_2)$ is any coupling of Z_1 and Z_2, then for $p < q$ we have $(\mathbb{E}\left[\|Z_1 - Z_2\|^p\right])^{\frac{1}{p}} < (\mathbb{E}\left[\|Z_1 - Z_2\|^q\right])^{\frac{1}{q}}$ by Lyapunov's inequality. Thus, $W_p(Z_1, Z_2)$ is an increasing function of p.

In order to state our central assumption in this article, we define precisely what we mean by low-dimensionality.

Definition 2. *For a d-dimensional random vector X, let $S = supp(X)$ denote the support of X. Let K be the set of all $k > 0$ satisfying the following condition: there exists a constant $C_k > 0$ such that for every $\epsilon > 0$, there exist $x_1, x_2, \ldots, x_{C_k\epsilon^{-k}} \in \mathbb{R}^d$ such that $S \subseteq \bigcup_{i=1}^{C_k\epsilon^{-k}} B(x_i, \epsilon)$. Then, $d_{int} = \inf K$ is called the intrinsic dimension of X.*

With the intent of creating an algorithm with polynomial sample complexity, we assume that the data distribution X is a *bounded* random vector with a specified intrinsic dimension.

Assumption 1. *The intrinsic dimension of X is $d_{int} \in [0, \infty)$, and we have $\|X\|_2 \leq \sqrt{d_{int}D}$ a.s. for some constant $D > 0$.*

Before stating our main result of this section, we define the key quantities

$$M(\theta) = \mathbb{E}[e^{\theta^T g(X)}]; \quad \rho_\theta = \frac{M(2\theta)}{M(\theta)^2}. \tag{6}$$

Observe that $M(\theta), \rho_\theta < \infty$ since X is bounded.

Theorem 2. *For each $p \geq 1, d_{int} \geq 0$, there exists a constant $C_{d_{int},p}$ independent of θ, X and n such that*

$$\frac{\mathbb{E}[W_p^p(X_\theta, X_{n,\theta})]}{\max\{1, (d_{int}D)^{p/2}\}} \leq \begin{cases} C_{d_{int},p} \max\left\{\sqrt{\frac{\rho_\theta}{n}}, \frac{\rho_\theta}{n}\right\} & d_{int} < 2p, \\ C_{d_{int},p} \max\left\{\sqrt{\frac{\rho_\theta}{n}} \ln \frac{n}{\rho_\theta}, \frac{\rho_\theta}{n}\right\} & d_{int} = 2p, \\ C_{d_{int},p} \max\left\{\left(\frac{\rho_\theta}{n}\right)^{p/d_{int}}, \frac{\rho_\theta}{n}\right\} & d_{int} > 2p \end{cases}$$

for every X satisfying Assumption 1 with intrinsic dimension d_{int} and D as in the assumption.

A few remarks are in order. First, when $\theta = 0$, $X_0 = X$ has intrinsic dimension d, and $X_{n,0}$ is the empirical distribution of X, the rates obtained above match completely with those in (Fournier & Guillin, 2013, Theorem 1), for which the case $d \neq 2p$ is known to be tight, and the case $d = 2p$ nearly tight. In particular, $\mu_{n,\theta}$ approximates μ_θ, up to constants, as well as the empirical distribution of X_θ over n samples would.

Secondly, $C_{d_{\text{int}},p}$ depends exponentially on d_{int} and p, even if it were chosen only to satisfy each of the given parts

Finally, for a large class of bounded random vectors, the quantity (6) actually grows polynomially in $\|\theta\|$. As a result, obtaining highly Wasserstein-accurate samples requires only polynomially many samples in the tilt and accuracy in a large number of cases. We elaborate on this in Section A.4.

In contrast, we have the following asymptotic lower bound for the distance W_1 on the real line. We are not aware if this extends to higher dimensions, but expect it to do so. This is merely proposed as a counterweight to Theorem 2, and is not claimed to be optimal.

Proposition 1. *Let X be a possibly unbounded random variable, and $\theta \in \mathbb{R}$ be such that $\mathbb{E}[e^{2\theta g(X)}(1 + X^2)] < \infty$. Then,*

$$\liminf_{n\to\infty} \sqrt{n}\mathbb{E}[W_1(X_\theta, X_{n,\theta})] \geq \sqrt{\frac{2}{\pi}} \left[\frac{\mathbb{E}[e^{2\theta g(X)}(X - \mathbb{E}[X_\theta])^2]}{\mathbb{E}[e^{\theta g(X)}]^2}\right]^{\frac{1}{2}}.$$

In particular, if for some constant $C_X > 0$ independent of θ (but can depend on X) we have $\mathbb{P}((x - \mathbb{E}[X_\theta])^2 > C_X) > \frac{1}{2}$, then

$$\liminf_{n\to\infty} \sqrt{n}\mathbb{E}[W_1(X_\theta, X_{n,\theta})] \geq \frac{\sqrt{C_X}}{2} \sqrt{\frac{2}{\pi}} \sqrt{\frac{\rho_\theta}{n}}.$$

The above estimate holds for any $W_p, p > 1$ since $W_p(X, Y)$ is an increasing function of p. This strongly motivates Assumption 1 as necessary for accurate tilting with low sample complexity: indeed, for a large number of unbounded random variables, $\sqrt{\frac{p_\theta}{n}}$ grows exponentially in θ. This will be expounded on in Section A.4.

Note that the upper and lower bounds provided in Theorem 2 and Proposition 1 respectively are nearly matching as $n \to \infty$ for $p = 1$, but do not match completely due to the $(x - \mathbb{E}[x_\theta]^2)$ term in the proposition.

4 Main algorithm and sample complexity

In this section, we will describe our main algorithm, Algorithm 1 and state its sample complexity bound Theorem 4, the main result of this article. First, we shall give some background on diffusion models, including the theoretical result used to prove our result.

4.1 DDPM and its sample complexity

Score-generative models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Dhariwal & Nichol, 2021) are state-of-the-art generative mechanisms used to sample from complex distributions. Two of the popular models in this class are DDIM (Song et al., 2021a) and DDPM (Ho et al., 2020). We present the version of the latter algorithm found in (Li & Yan, 2025), which is also commonly used by many other articles.

Suppose we wish to sample from a d-dimensional data distribution p_{data}. Let $T \geq 1$, and consider a Markov process in $\mathbb{R}^d$ starting from the data distribution $X_0 \sim p_{\text{data}}$ such that

$$X_t = \sqrt{1 - \beta_t} X_{t-1} + \sqrt{\beta_t} W_t \quad t = 1, 2, \ldots, T \tag{7}$$

where W_i are i.i.d. $N(0, I)$ and $\beta_i \in (0, 1), i = 1, 2, \ldots, T$ are learning rates. While explicit formulas for β_i are specified by (Li & Yan, 2025, (3)), we will only assume $\beta_1 = T^{-c_0}$ in our work, where c_0 is a large independent constant. Define the constants

$$\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i. \tag{8}$$

We can express X_t in closed form as

$$X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \bar{W}_t, \quad \bar{W}_t \sim N(0, I). \tag{9}$$

Next, we define the reverse process. Let p_{X_t} be the density of X_t for $t = 1, 2, \ldots, T$ and $s_t^*(x) = \nabla \log p_{X_t}(x)$ be the score of X_t for $x \in \mathbb{R}^d$ and $t = 1, 2, \ldots, T$. The functions s_t^* are inaccessible since X is unknown. Therefore, s_t^* is typically estimated by statistical methods or neural networks (Hyvärinen & Dayan, 2005; Vincent, 2011; Block et al., 2022; Chewi et al., 2026). Denote by $s_t(x)$ the score estimator for $s_t^*(x)$, and by

$$\epsilon_{score}^2 := \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}[\|s_t(X_t) - s_t^*(X_t)\|_2^2] \tag{10}$$

the averaged L^2 error of the score estimator across the forward process. We define the reverse process Y_t for $t = T, T-1, \ldots, 0$ by $Y_T \sim N(0, I)$, and

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}} (Y_t + \eta_t s_t(Y_t) + \sigma_t Z_t), \quad t = T, T-1, \ldots, 1 \tag{11}$$

where $Z_i, i = 1, \ldots, T$ are i.i.d. $N(0, I)$ draws.

The main result of (Li & Yan, 2025) follows: note that the bound is in the TV distance $TV(Z_1, Z_2) = \sup_A |\mathbb{P}(Z_1 \in A) - \mathbb{P}(Z_2 \in A)|$. Recall the intrinsic dimension from Definition 2.

Theorem 3. *Assume that $\mathbb{E}\|p\|_2 < \infty$ and that for some constant c_M independent of the distribution X_0, we have $T^{c_M} \geq \mathbb{E}\|p\|_2$. Suppose that d_{int} is the intrinsic dimension of p. Then, for the forward process given by (11) run with $\eta_t = 1 - \alpha_t$ and $\sigma_t^2 = \frac{(1-\alpha_t)(\alpha_t - \bar{\alpha}_t)}{1 - \bar{\alpha}_t}$,*

$$TV(X_1, Y_1) \leq c \frac{d_{int} \log^3 T}{T} + c \epsilon_{score} \sqrt{\log T}$$

where ϵ^2_{score} is as in (10) and X_t, Y_t are as in (9) and (11) respectively.

A short remark on the result follows. Theorem 3 is the tightest known DDPM accuracy bound with the most minimal assumptions on the data distribution, superceding older bounds (Chen et al., 2023b; Benton et al., 2024). Furthermore, it incorporates early stopping and intrinsic dimensionality in its bounds, making it highly suitable for sampling from low-dimensional bounded distributions . Early stopping is a common feature of DDPM sampling from random vectors with bounded support (Song et al., 2021b), and can be reconciled either through projection or choosing a weaker metric, as our main result does.

4.2 Main algorithm

We are ready to describe our complete algorithm for generating tilted samples. Given samples $X_1, \ldots, X_n$ from a random vector X and $\theta \in \mathbb{R}^d$ such that $\mathbb{E}[e^{\theta^T g(X)}] < \infty$, consider the random vector $X_{n,\theta}$ as in (2). This provides a strong justification for the choice of the estimator $X_{n,\theta}$: its score can be explicitly calculated.

To demonstrate the second point, let $X_{n,\theta,t}, t = 0, 1, \ldots, T$ be the forward process (7) started from $X_{n,\theta,0} \sim X_{n,\theta}$. Let $p_{n,\theta,t}(x)$ be the density of $X_{n,\theta,t}$ and $s_{n,\theta,t}(x) = \nabla \log p_{X_{n,\theta,t}}(x)$ be the score of X_t. We have the following explicit formula for $s_{n,\theta,t}$, thereby demonstrating its easy evaluation at any point $t = 1, \ldots, T$.

Lemma 1. *Let $\gamma_{\theta,i} = \frac{e^{\theta^T g(X_i)}}{\sum_{i=1}^n e^{\theta^T g(X_i)}}, i = 1, \ldots, n$ and $\alpha_t, t = 1, 2, \ldots, T$ be as in (8). For a number $K > 0$, let*

$$\varphi_K(z) = \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{\|z\|_2^2}{2K^2}\right)$$

the density of $N(0, K^2 I)$ evaluated at $z \in \mathbb{R}^d$. The following formula for $s_{n,\theta,t}(x)$ holds:

$$s_{n,\theta,t}(x) = -\frac{x}{1-\bar{\alpha}_t} + \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t} \frac{\sum_{i=1}^n \gamma_{\theta,i} X_i \varphi_{(1-\bar{\alpha}_t)}(x - \sqrt{\bar{\alpha}_t} X_i)}{\sum_{j=1}^n \gamma_{\theta,j} \varphi_{(1-\bar{\alpha}_t)}(x - \sqrt{\bar{\alpha}_t} X_j)}.$$

The proof is standard and postponed to the appendix. Now, the algorithm proceeds by simply executing the reverse process (11) from a fresh sample $Y_T \sim N(0, I)$, but using $s_{n,\theta,t}$ in place of s_t, till we obtain Y_1, which we output.

Algorithm 1 DDPM for tilted distributions

Input: Samples $X_i, i = 1, \ldots, n$, tilt vector $\theta \in \mathbb{R}^d$
Parameters: $T \in \mathbb{N}$
Output: Random vector Y_1

1: $Y_T \sim N(0, I)$
2: **for** $t = T, T-1, \ldots, 2$ **do**
3: $\quad Z_t \sim N(0, I)$
4: $\quad Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}(Y_t + (1-\alpha_t) s_{n,\theta,t}(Y_t) + \sqrt{\frac{(1-\alpha_t)(\alpha_t - \bar{\alpha}_t)}{1-\bar{\alpha}_t}} Z_t)$
5: **end for**
6: **return** Y_1

The idea of using $s_{n,\theta,t}$ as a score estimator is natural; this is precisely kernelized score estimation (Song & Ermon, 2020; Liu et al., 2024; Chewi et al., 2026; Yu & Yu, 2026) adapted to tilted distributions.

The sample complexity of Algorithm 1 follows if we can contrast it with a perfect algorithm, which would ideally have access to the score s_t^* of the forward process (7) instantiated from $X_0 \sim X_\theta$. That this replacement does not affect the accuracy very much is a general artifact of a much more broad result on score estimation which we establish for bounded random vectors.

Proposition 2. *Let P, Q be any two probability measures supported on $B_{\sqrt{d_{int}} D}$, and let $X_P \sim P, X_Q \sim Q$ be the associated random vectors. Let $X_t^P, X_t^Q, t = 0, 1, \ldots, T$ be the forward processes (11) started from $X_0^P = X_P$ and $X_0^Q = X_Q$ respectively. Let s_t^P, s_t^Q denote the scores of*

X_t^P, X_t^Q *respectively. We have*

$$\frac{1}{T}\sum_{t=1}^{T} \|s_t^P(X_t^P) - s_t^Q(X_t^P)\|_2^2 \leq A_T(\sqrt{d_{int}D})W_2^2(X_P, X_Q),$$

where $A_T(\sqrt{d_{int}D})$ *is a constant independent of* P *and* Q*.*

Observe that our result assumes no prior relation between P and Q, and is thus of interest outside of this article. Indeed, all we need is for the measure Q to be close to P in Wasserstein distance, and for s_t^Q to have an explicit formula. Our main result is a polynomial sample complexity bound for Algorithm 1.

Theorem 4. *Fix* $d_{int} > 0, d \in \mathbb{N}$, $p \geq 1$*. Let* c_0 *be some appropriately large constant as in the formula for* β_1*. There exists a constant* $C_{d_{int},D,p}$ *depending only on* d_{int}*,* D *and* p *such that the following result holds. Given* $\epsilon > 0$ *small enough, random vector* X *satisfying Assumption 1 and* $\theta \in \mathbb{R}^d$*, if* ρ_θ *is as in (6) and* Y *is the output of Algorithm 1 given* n *i.i.d. samples from* X*, then*

1. *If* $n > C_{d_{int},D,p}\epsilon^{-(8.8c_0+2)/0.51}\rho_\theta$*, we have* $\mathbb{E}[d_{BL}(X_\theta, Y)] < \epsilon$ *for* ϵ *smaller than some polynomial in* d^{-1}*.*
2. *If* $n > C_{d_{int},D,p}\epsilon^{-p(8.8c_0+2)/0.51}\rho_\theta$*, we have* $\mathbb{E}[W_p(X_\theta, pr(Y))] < \epsilon$ *for* ϵ *smaller than some polynomial in* d^{-1} *and* p^{-1}*, where pr is the projection onto* $B_{\sqrt{d_{int}D}}$*.*

In both cases, $C_{d_{int},D,p}$ *has polynomial dependence on* D *but exponential dependence on the others.*

Thus, in many situations this algorithm needs only polynomially many samples in θ (See Section A.4) and ϵ to produce accurately tilted samples. Dependence on the ambient dimension d is completely absent for small enough ϵ, although the threshold beyond which the independence holds is itself some polynomial in d^{-1}. We will give examples of this phenomena in the appendix.

The estimator $s_{n,\theta,t}$ is both computationally intensive to calculate (since it requires every sample in its construction), and can produce a measure which lacks sample diversity, as most kernelized estimators tend to do. Therefore, we also prove that one can approximate the score of $X_{\theta,t}$ using a ReLu neural network with the number of parameters depending on θ, and subsequently obtain accuracy guarantees for network-learned score versions of Algorithm 1.

Theorem 5. *There exists a universal constant* $C > 0$ *such that the following holds. Given* $\epsilon > 0$ *small enough, there exists a ReLu neural network with at most* $C_{d,\epsilon,d_{int}}$ *parameters such that, for every* X *satisfying Assumption 1 and* $\theta \in \mathbb{R}^d$*, if we run a version of Algorithm 1 with* $s_{n,\theta,t}$ *learned by a ReLu neural network* $\tilde{S}_t$*, then* $n > C_{d_{int},D}\epsilon^{-C}$ *implies that* $\mathbb{E}[d_{BL}(X_\theta, Y)] < \epsilon$*, and for* $p \geq 1$*,* $n > C_{d_{int},D,p}\epsilon^{-Cp}$ *implies that* $\mathbb{E}[W_p(X_\theta, pr(Y))] < \epsilon$*. The constant* $C_{d,\epsilon,d_{int}}$ *has only linear dependence on* d *and polynomial dependence on other parameters, while* $C_{d_{int},D,p}$ *has polynomial dependence on* D *and exponential dependence on others.*

We prove this result completely in Appendix A.3. Note that our result does not invoke tilting anywhere, and we briefly explain why. Primarily, our construction, which strongly uses the low-dimensional nature of X, differs from ordinary neural network score approximation guarantees such as Yakovlev & Puchkin (2025), in that most work on score matching approximate the score arising from a continuous time process, and in the process place significant assumptions on the underlying random vector. We wished to avoid these assumptions.

Secondly, our result places no restriction on the *sizes* of the weights in the score learning neural network. Thus, our claim is that parametrically, since tilting preserves the support, the low-dimensionality of the support plays a more crucial role than the degree of the tilt. Furthermore, it is also possible that tilting a random vector X can produce a more "smooth" version of X (for instance, $X = Y_\theta$ implies $Y = X_{-\theta}$, so one of the two is "smoother" but they are both tilts of each other). Thus, our architecture does not involve the tilt in any manner.

5 Experiments

We break our experiments into two categories a) score learning through the kernelized score estimator, and b) score learning through a neural network.

We consider four random vectors $X_1 = U[-1,1]$, $X_2 = (V_1, V_2)$ where V_i are independent Beta$(2,2)$ random variables, $X_3 = U([0,1]^3)$, and X_4 is uniform on a tilted rectangle $\{y + z = 0, 0 \leq y \leq 1, 0 \leq x \leq 1\}$ in three dimensions (and hence has intrinsic dimension 2).

We consider five values of $\|\theta\|$, namely $\Theta = \{0.1, 0.5, 1, 3, 6\}$. In each dimension, we design tilt vectors with the given $\|\theta\|$ as follows: in one dimension, we will tilt by each $c \in \Theta$. In two dimensions, we will tilt by $c(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ where $c \in \Theta$. In three dimensions, we tilt by $c(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$, for $c \in \Theta$.

5.1 On the accuracy of Algorithm 1

In this experiment, we observe the behaviour of the error of Algorithm 1 with kernelized score estimation, with decreasing values of $\frac{\rho_\theta}{n}$. We pin the value of $\frac{\rho_\theta}{n}$ to five values on the x-axis, namely $0.5, 0.1, 0.05, 0.01, 0.005$. For example, for each random vector X_i and tilt amount (say 0.1) we compute the number of samples n so that ρ_θ / n equals the corresponding value (say, 0.5). The errors are then plotted against the value of $\frac{\rho_\theta}{n}$ in Figure 1. The figur confirms a decrease in error with a drop in $\frac{\rho_\theta}{n}$, but also confirms that the intrinsic dimension plays a role in the error.

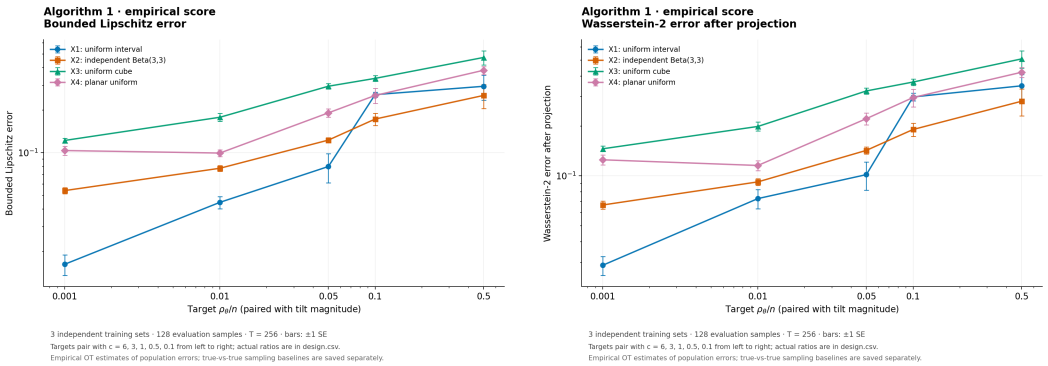

Figure 1: As the value of ρ_θ / n goes down, the error goes down. Furthermore, the intrinsic dimension plays a role in how much the error goes down, with higher intrinsic dimension being harder to tilt

5.2 Neural-network approximated score learning

This experiment breaks into two parts. First, we fix the architecture throughout the various tilts that we do, in particular the number of parameters. In the first part of our experiment, we increase the number of samples proportionally to the tilt while not changing the architecture, while in the second part of the experiment we do not change the number of samples with the tilt.

We see that tilting can be performed accurately by our architecture, assuming that the number of parameters is constant. One sees in Figure 2 that if the number of samples increases, then the error drops despite the architecture not changing with the tilt, demonstrating that the architecture is able to capture higher tilts without a change in architecture, if the number of samples is increased. On the other hand, if the number of samples is held constant, then Figure 3 demonstrates that the error does not change significantly with the tilt. This demonstrates that tilted random variables can be learned as well in our architecture as untilted ones, within the sample sample and neural network complexity.

6 Conclusion

In this article, we have provided a diffusion-based algorithm for generating tilted samples that, under a boundedness assumption, has a provably polynomial sample complexity in a large number of cases. We achieve this by creating an empirical estimator of the score function in the reverse process, and then asserting the accuracy of the score using an intermediate Wasserstein-distance based argument.

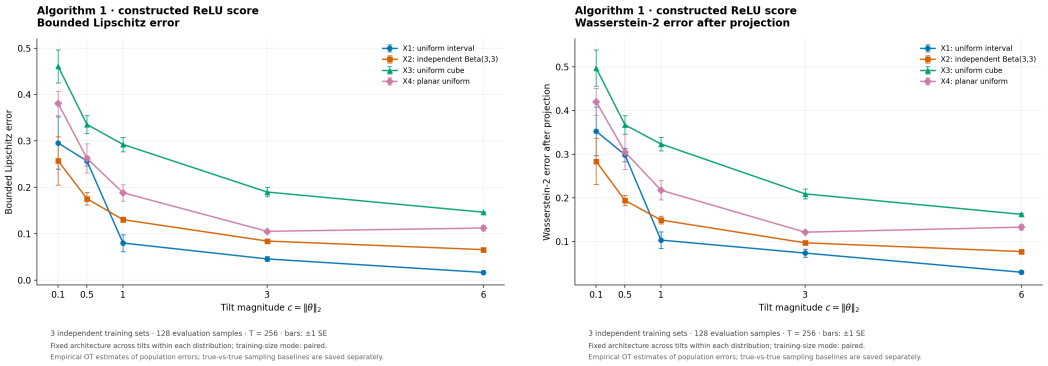

Figure 2: As the tilt goes up, if the number of samples goes up then the error goes down, even if the architecture is fixed throughout the experiment.

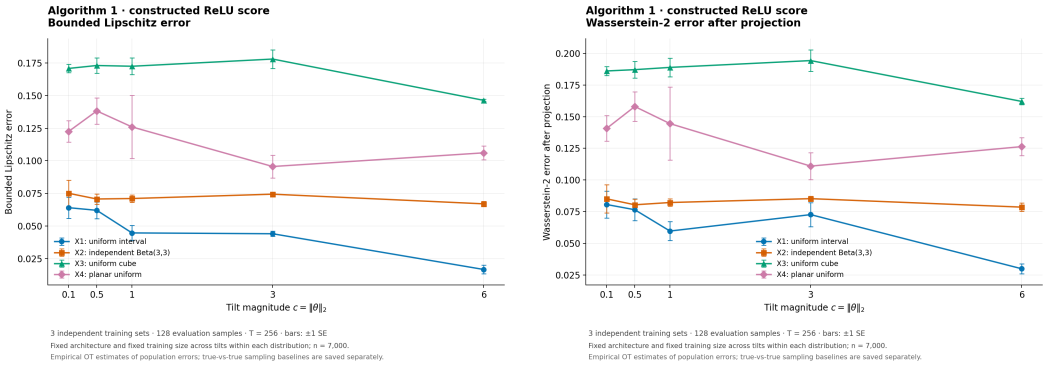

Figure 3: As the value of θ changes but the number of samples and architecture do not change, the error remains stable in the tilt regardless of the nature of the random vector

.

Along the way, we developed results of independent interest. Theorem 15 proves that the estimator $X_{n,\theta}$ in (2) is the best choice for score estimation. The Wasserstein bounds obtained in Theorem 2 extend results in the empirical literature to SNIS literature, which we complement with asymptotic lower bounds in Proposition 1. Finally, Proposition 2 is a simple yet fundamental result in score estimation that we have not seen before in the literature. We also include Theorem 5 which provides an architecture that can learn tilted scores as well as untilted ones.

The result strongly uses Assumption 1, without which all our results would break down. While there are fundamental limits to tilted sampling as shown in (Moitra et al., 2026) and instance-specific lower bounds such as Proposition 1, it remains an interesting question to ask which unbounded random vectors can be tilted easily.

Furthermore, our neural network architecture does not take extreme tilts and weights into account; it would be interesting to ask precisely how much the weights vary if the tilt θ is very high. The dimensions of all random vectors considered in the experiments did not cross 3: the primary reason for this was unstable computation of the Wasserstein and BL distances in high dimensions. We leave the analysis of theoretical behaviour of weights for high tilts and dimension to future work, along with any other form of architecture that may take advantage of the low-dimensional assumption.

AI USE STATEMENT

The use of generative AI and LLMs was primarily used for providing a literature review. The central idea behind the main results was ru using simple examples using AI. Finally, the entire codebase has been generated using AI, including the generation of the images.

Ethics and Reproducibility statement: Our work carries no ethical concerns. Our code is attached as a zip file to our submission, and Algorithm 1 is described completely, while the neural network version described in Theorem 5 is fairly simple to execute.

REFERENCES

Javier Aguilar and Riccardo Gatto. Unified perspective on exponential tilt and bridge algorithms for rare trajectories of discrete Markov processes. *Phys. Rev. E*, 109:034113, Mar 2024. doi: 10.1103/PhysRevE.109.034113. URL https://link.aps.org/doi/10.1103/PhysRevE.109.034113.

Mayer Alvo. *Exponential Tilting and Its Applications*, pp. 171–193. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06784-6. doi: 10.1007/978-3-031-06784-6_6. URL https://doi.org/10.1007/978-3-031-06784-6_6.

Anonymous. The accumulation of score estimation error in diffusion models. In *International Conference on Learning Representations (ICLR) 2026*, 2026. URL https://openreview.net/forum?id=end8EBwFOU. Under review as ICLR 2026 submission.

M Avellaneda. Minimum entropy calibration of asset pricing models, internat. *J. Theoret. Appl. Finance*, 1:447472, 1998.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising autoencoders and langevin sampling. *CoRR*, abs/2002.00107, 2022. URL https://arxiv.org.

Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'I.H.P. Probabilités et statistiques*, 50(2):539–563, 2014. doi: 10.1214/12-AIHP517. URL https://www.numdam.org/articles/10.1214/12-AIHP517/.

Peter W Buchen and Michael Kelly. The maximum entropy distribution of an asset inferred from option prices. *Journal of Financial and Quantitative Analysis*, 31(1):143–159, 1996.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018. doi: 10.1214/17-AAP1326. URL https://doi.org/10.1214/17-AAP1326.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/chen23o.html.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023b. URL https://arxiv.org/abs/2209.11215.

Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, and Omar Montasser. Ddpm score matching and distribution learning. In Francesco Locatello and John N. Tsitsiklis (eds.), *Proceedings of the Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1434–1435. PMLR, 09–12 Jul 2026. URL https://proceedings.mlr.press/v336/chewi26a.html.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024. URL https://arxiv.org/abs/2209.14687.

Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. Tail-gan: Learning to simulate tail risk scenarios. *Management Science*, 2025.

Imre Csiszár. On topology properties of f-divergences. *Studia Scientifica Mathematica Hungerica*, 2:329–339, 1967.

Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: Approximation by empirical measures. *Annales de l'I.H.P. Probabilités et statistiques*, 49(4):1183–1203, 2013. doi: 10.1214/12-AIHP489. URL https://www.numdam.org/item/10.1214/12-AIHP489.pdf.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34*, pp. 8780–8794, Virtual Conference, 2021. Curran Associates, Inc.

V. Dobrić and J. E. Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, 1995. doi: 10.1007/BF02213456. URL https://link.springer.com/article/10.1007/BF02213456.

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *Ann. Math. Statist.*, 40(1):40–50, 1968.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure, 2013. URL https://arxiv.org/abs/1312.2128.

Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43): 1–54, 2025. URL https://jmlr.org/papers/v26/24-0902.html. Published version with convergence bounds for SGMs in 2-Wasserstein distance.

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. doi: 10.1111/j.1751-5823.2002.tb00178.x.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=h8GeqOxtd4.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695–709, 2005.

Sarvesh Ravichandran Iyer, Himadri Mandal, Dhruman Gupta, Rushil Gupta, Agniv Bandhyopadhyay, Achal Bassamboo, Varun Gupta, and Sandeep Juneja. Fundamental limits for weighted empirical approximations of tilted distributions. *CoRR*, abs/2512.23979, 2025. doi: 10.48550/arXiv.2512.23979. URL https://arxiv.org/abs/2512.23979.

Lingkai Kong, Haichuan Wang, Yuqi Pan, Cheol Woo Kim, Mingxiao Song, Alayna Nguyen, Tonghan Wang, Haifeng Xu, and Milind Tambe. Robust optimization with diffusion models for green security. *arXiv preprint arXiv:2503.05730*, 2025.

Suemin Lee, Ruiyu Wang, Lukas Herron, and Pratyush Tiwary. Exponentially tilted thermodynamic maps (exptm): Predicting phase transitions across temperature, pressure, and chemical potential, 2025. URL https://arxiv.org/abs/2503.15080.

Gen Li and Yuling Yan. $\mathcal{O}(d/t)$ convergence theory for diffusion probabilistic models under minimal assumptions. *Journal of Machine Learning Research*, 26(292):1–55, 2025. URL https://www.jmlr.org/papers/v26/25-0272.html.

Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John R. Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13):eadk4489, 2024. doi: 10.1126/sciadv.adk4489. URL https://www.science.org/doi/10.1126/sciadv.adk4489.

Fenghua Ling, Zeyu Lu, Jing-Jia Luo, Lei Bai, Swadhin K. Behera, Dachao Jin, Baoxiang Pan, Huidong Jiang, Toshio Yamagata, et al. Diffusion model-based probabilistic downscaling for 180-year East Asian climate reconstruction. *npj Climate and Atmospheric Science*, 7:131, 2024. doi: 10.1038/s41612-024-00679-1. URL https://www.nature.com/articles/s41612-024-00679-1.

Haoyu Liu, Tingyu Zhu, Nanshan Jia, Jinghai He, and Zeyu Zheng. Learning to simulate from heavy-tailed distribution via diffusion model. *Operations Research*, 2024. doi: 10.2139/ssrn.4975931. Available at SSRN 4975931.

Attilio Meucci. Fully flexible views: Theory and practice, 2010. URL https://arxiv.org/abs/1012.2848.

Ankur Moitra, Andrej Risteski, and Dhruv Rohatgi. Steering diffusion models with quadratic rewards: a fine-grained analysis. In *Submitted to International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3tEt6VNJTU.

Xinhe Mu, Zaijiu Shang, Zhaoqi Zhou, Chuan Zhou, Qi Meng, Guiying Yan, and Zhiming Ma. Score approximation for diffusion models on arbitrary low-dimensional structures, 2026. URL https://arxiv.org/abs/2606.19894.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26633–26681. PMLR, 2023. URL https://proceedings.mlr.press/v202/oko23a.html.

Art B. Owen. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. doi: 10.1080/01621459.2000.10473909.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073.

Reuven Y. Rubinstein. *Simulation and the Monte Carlo Method*. John Wiley & Sons, 1981. ISBN 9780471089179. doi: 10.1002/9780470316511.

Johannes Schmidt-Hieber. Deep ReLU network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019. URL https://arxiv.org/abs/1908.00695.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. doi: 10.1214/19-AOS1875. URL https://doi.org.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *Proceedings of the International Conference on Learning Representations (ICLR) 2021*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32483–32498. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23k.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448. Curran Associates, Inc., 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021b. URL https://arxiv.org/abs/2011.13456.

Michael Stutzer. A simple nonparametric approach to derivative security valuation. *The Journal of Finance*, 51(5):1633–1652, 1996. doi: https://doi.org/10.1111/j.1540-6261.1996.tb05220.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb05220.x.

Matus Telgarsky. Neural networks and rational functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3387–3393. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/telgarsky17a.html.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011. doi: 10.1162/NECO_a_00142.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided SE(3) diffusion models, 2024. URL https://arxiv.org/abs/2403.14088.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2628–2648, 2019. doi: 10.3150/18-BEJ1065. URL https://doi.org.

E. Weinan and Qingcan Wang. Exponential convergence of the deep neural network approximation for analytic functions. *Science China Mathematics*, 61(10):1733–1740, 2018. doi: 10.1007/s11425-018-9387-x. URL https://doi.org/10.1007/s11425-018-9387-x.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5824–5891. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/yakovlev25a.html.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 0893-6080. doi: 10.1016/j.neunet.2017.07.002.

Yifeng Yu and Lu Yu. Diffusion models with heavy-tailed targets: Score estimation and sampling guarantees. *arXiv preprint arXiv:2601.06715*, 2026.