Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality

Zhihan Huang*   Yuting Wei*   Yuxin Chen*

October 24, 2024

Abstract

The denoising diffusion probabilistic model (DDPM) has emerged as a mainstream generative model in generative AI. While sharp convergence guarantees have been established for the DDPM, the iteration complexity is, in general, proportional to the ambient data dimension, resulting in overly conservative theory that fails to explain its practical efficiency. This has motivated the recent work Li and Yan (2024a) to investigate how the DDPM can achieve sampling speed-ups through automatic exploitation of intrinsic low dimensionality of data. We strengthen this prior work by demonstrating, in some sense, optimal adaptivity to unknown low dimensionality. For a broad class of data distributions with intrinsic dimension k, we prove that the iteration complexity of the DDPM scales nearly linearly with k, which is optimal when using KL divergence to measure distributional discrepancy. Our theory is established based on a key observation: the DDPM update rule is equivalent to running a suitably parameterized SDE upon discretization, where the nonlinear component of the drift term is intrinsically low-dimensional.

Contents

*Department of Statistics and Data Science, the Wharton School, University of Pennsylvania; email: {zhihanh,ytwei,yuxinc}@wharton.upenn.edu.

1 Introduction

In generative modeling, one is asked to produce novel samples that resemble, in distribution, the data on which the model has been trained. In order to fully unleash its power, a generative modeling algorithm must exploit — either implicitly or explicitly — the distinctive features of the data distribution of interest, rather than merely optimizing for worst-case scenarios. A prominent example of such distinctive features is the low-dimensional structure underlying the data of interest, which permeates modern statistical modeling. For instance, image data often resides on some low-dimensional manifold (Pope et al., 2021; Simoncelli and Olshausen, 2001), and genomics data might also be effectively modeled through manifolds (Moon et al., 2018; Zhu et al., 2018). While harnessing these low-dimensional data structures is generally believed to enable more efficient data generation, the theoretical underpinnings for this capability in contemporary generative modeling algorithms remain significantly under-explored.

In this paper, we study how low-dimensional data structure influences the performance of diffusion models, an emerging paradigm that facilitates remarkable progress across various subfields of modern generative modeling (Chen et al., 2024a; Dhariwal and Nichol, 2021; Ho et al., 2022; Ramesh et al., 2022; Sohl-Dickstein et al., 2015; Watson et al., 2023). While empirical evidence suggests that diffusion models might not suffer from the curse of dimensionality, a rigorous understanding of their efficacy in the presence of low-dimensional structure remains elusive. Consider, for concreteness, a mainstream diffusion model known as the denoising diffusion probabilistic model (DDPM) (Ho et al., 2020), which has received widespread adoption in practice and sparked considerable interest from the theoretical community to demystify its efficacy (e.g., Chen et al. (2022a,b); Lee et al. (2022, 2023); Li et al. (2024b); Li and Yan (2024b); Liang et al. (2024); Mbacke and Rivasplata (2023); Tang (2023); Tang and Zhao (2024)). In order for the DDPM sampler to generate a d-dimensional new sample that approximates well the target distribution in terms of KL divergence, the number of steps[1] needs to scale linearly with d, a scaling that has been shown to be un-improvable in general (Benton et al., 2024). However, in application like high-resolution image generation, the number of pixels in each image might be exceedingly large, far exceeding the number of steps typically needed to run the DDPM in practice. This indicates that even linear scaling with d does not adequately explain the observed efficiency of the DDPM in practice.

To bridge the theory-practice gap for the data generation stage, the seminal work Li and Yan (2024a) demonstrated that DDPM can automatically adapt to the intrinsic dimension of the target distribution p_{data}, which rigorously justifies sampling speed-ups if the support of p_{data} is intrinsically low-dimensional. As a particularly intriguing message of this prior result, the DDPM sampler, in its original form, achieves favorable adaptivity without explicitly modeling the low-dimensional structure.

Despite this remarkable theoretical progress, however, it remains unclear to what extent the low-dimensional structure can enhance the sampling efficiency of the DDPM. More precisely, Li and Yan (2024a) proved that when the support of p_{data} has intrinsic dimension k (to be made precise in Section 4.1.1), the iteration complexity of the DDPM scales at most with k^4 (up to some logarithmic factor and other dependency on the target accuracy level), provided that KL divergence is used to measure the distributional discrepancy. This result was recently improved to k^3 by Azangulov et al. (2024) under some low-dimensional smooth manifold assumption. This raises a natural question: is the k^3 scaling essential, and if not, how to sharpen it? Addressing this question would provide a quantitative understanding about DDPM's degree of adaptivity when tackling low-dimensional structure.

Main contributions. In this work, we develop a sharp convergence theory for the DDPM in the face of unknown low dimensionality. For a broad family of target data distributions p_{data} with intrinsic dimension k, we show that it takes DDPM at most $O(k/\varepsilon^2)$ (up to some logarithmic factor) steps to yield an ε^2-accurate distribution — measured by the KL divergence — assuming access to perfect score functions. This improves upon the state-of-the-art iteration complexity (Azangulov et al., 2024) by a factor of k^2, and achieves nearly optimal scaling in k without any burn-in requirement. Further, our convergence theory can be easily extended to accommodate imperfect score estimation.

Our convergence theory is established based on a key observation: the original DDPM update rule by Ho et al. (2020) is precisely equivalent to running a carefully parameterized stochastic differential equation

---

[1]Given that each step of the DDPM requires evaluating a score function (typically through computing the output of a large neural network or transformer), the sampling speed of the DDPM is largely dictated by the number of steps.

(SDE) upon discretization using the exponential integrator. Crucially, the drift term of this discretized SDE is semi-linear, with the non-linear component proportional to the posterior mean of the data given its Gaussian corruption. Intuitively, when the support of p_{data} is low-dimensional, this nonlinear drift term acts as a "projection" onto the low-dimensional manifold of interest, harnessing this intrinsic structure to enhance the smoothness of the solution path and thereby speed up the sampling process. Somewhat surprisingly, while DDPM (Ho et al., 2020) was initially proposed to maximize a variational lower bound on the log-likelihood, the resulting update rule inherently adopts the desirable time discretization schedule, in a way that is fully adaptive to unknown low-dimensional distributions. A key ingredient underlying our analysis of the discretized SDE lies in carefully calibrating the evolution of the posterior mean function along the forward process.

While finalizing this manuscript, we became aware of an interesting concurrent work Potaptchik et al. (2024) that develops a similar convergence theory as ours. We shall discuss the difference between our work and Potaptchik et al. (2024) in Section 4.2.

Notation. We introduce several notation to be used throughout. For any two functions f and g, we adopt the notation $f \lesssim g$ or $f = O(g)$ (resp. $f \gtrsim g$ or $f = \Omega(g)$) to mean that there exists some universal constant $C > 0$ such that $f \leq Cg$ (resp. $f \geq Cg$). The notation $f \asymp g$ or $f = \Theta(g)$ then indicates that $f \lesssim g$ and $f \gtrsim g$ hold at once. Additionally, $f = \widetilde{O}(g)$ (resp. $f = \widetilde{\Theta}(g)$) is defined analogously except that the logarithmic dependency is hidden. We denote by $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm of a matrix, and use $\|\cdot\|_2$ to denote the Euclidean norm of a vector. The notation $X \overset{\mathrm{d}}{=} Y$ indicates that the random objects X and Y are identical in distribution.

2 Preliminaries

In this section, we review some basics about the original DDPM sampler, introduce a useful perspective from SDEs, and describe some fundamental properties about the (Stein) score function.

The DDPM sampler. The procedure of the DDPM sampler is described in the sequel.

- *Forward process.* Consider a forward process $(X_t)_{t \in [0,T]}$ generated such that

$$X_t \overset{\mathrm{d}}{=} \sqrt{1 - \sigma_t^2} X_0 + \sigma_t W \qquad \text{with } \sigma_t := \sqrt{1 - e^{-2t}}, \tag{1}$$

where $X_0 \sim p_{\mathsf{data}}$ and $W \sim \mathcal{N}(0, I_d)$ are independently generated. In words, this forward process starts with a realization of X_0 drawn from the target data distribution p_{data}, and injects increasingly more Gaussian noise as t grows. Evidently, the distribution of X_T becomes exceedingly close to $\mathcal{N}(0, I_d)$ for a large enough time horizon T.

- *Data generation process.* In order to operate in discrete time, let us choose a set of N time discretization points $0 = t_0 < t_1 < \cdots < t_N < t_{N+1} = T$ for some large enough T, and adopt the update rule below:[2]

$$Y_{t_0} \sim \mathcal{N}(0, I_d),$$
$$Y_{t_{n+1}} = \frac{1}{\sqrt{\alpha_n}} \Big(Y_{t_n} + (1 - \alpha_n) \widehat{s}_{T-t_n}(Y_{t_n}) \Big) + \sqrt{\frac{(1 - \alpha_n)(1 - \overline{\alpha}_{n+1})}{1 - \overline{\alpha}_n}} Z_n, \qquad n = 0, \ldots, N-1, \tag{2}$$

where $\widehat{s}_t(\cdot)$ represents an estimate of the true score function $s_t(\cdot) = \nabla \log p_{X_t}(\cdot)$ w.r.t. the forward process (1), and the Z_n's are random vectors independently drawn from $Z_n \sim \mathcal{N}(0, I_d)$. Here and throughout, the coefficients $\{\alpha_n\}$ and $\{\overline{\alpha}_n\}$ are taken to be

$$\alpha_n := e^{-2(t_{n+1} - t_n)} \qquad \text{and} \qquad \overline{\alpha}_n := \prod_{i=n}^{N} \alpha_i = e^{-2(T - t_n)}, \tag{3}$$

---

[2]While the DDPM update rule was introduced in Ho et al. (2020) through the choices of $\{\alpha_n\}$ instead of the time discretization points $\{t_n\}$, it is equivalent to the version adopted herein.

which are fully determined by the locations of the time discretization points. In a nutshell, each iteration in (2) computes the weighted superposition of the current iterate, its corresponding score estimate, and an independent Gaussian vector, where the weights were originally proposed by Ho et al. (2020) to optimize a variational lower bound on the log-likelihood.

An SDE perspective. The feasibility to reverse the forward process with the aid of score functions can be elucidated in the continuous-time limits, an approach popularized by Song et al. (2021). To be more specific, construct the forward process as the standard Ornstein-Uhlenbeck (OU) process (Oksendal and Oksendal, 2003)

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0 \sim p_{\mathsf{data}}, \quad t \in [0, T] \tag{4}$$

with (W_t) the standard Brownian motion in $\mathbb{R}^d$. It is well-known this process (4) satisfies property (1). Crucially, process (4) can be reversed through the following SDE with the aid of score functions:

$$\mathrm{d}Y_t = \big(Y_t + 2s_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad t \in [0, T], \tag{5}$$

where (B_t) represents another standard Brownian motion in $\mathbb{R}^d$ independent from (X_t) and (W_t). If $Y_0 \sim X_T$, then classical results in the SDE literature (Anderson, 1982; Haussmann and Pardoux, 1986) tell us that

$$Y_{T-t} \overset{\mathrm{d}}{=} X_t, \qquad 0 \leqslant t \leqslant T \tag{6}$$

holds under fairly mild regularity conditions on p_{data}, thereby confirming the plausibility of score-based generative models. Note, however, that practical implementation of (5) has to operate in discrete time, and the sampling fidelity relies heavily upon the time discretization error. In fact, a large body of prior convergence theory for DDPM-type stochastic samplers (e.g., Benton et al. (2024); Chen et al. (2023a, 2022b)) was established by studying a certain time-discretized version of (5), with $s_{T-t}(\cdot)$ replaced by the score estimate $\widehat{s}_{T-t}(\cdot)$ to account for the effect of score estimation errors.

Score functions. Before proceeding, let us take a moment to delve into the score function defined as

$$s_t(x) := \nabla \log q_t(x), \tag{7}$$

where $q_t(\cdot)$ reprsents the marginal density function of X_t w.r.t. the Lebesgue measure. By defining the posterior mean and covariance of X_0 given X_t as follows:

$$\mu_t(x) := \mathbb{E}[X_0 \mid X_t = x] \quad \text{and} \quad \Sigma_t(x) := \mathrm{Cov}[X_0 \mid X_t = x],$$

one can readily derive from Tweedie's formula (Efron, 2011) that

$$\mu_t(X_t) = \frac{1}{\sqrt{1 - \sigma_t^2}}\left(X_t + \sigma_t^2 \nabla \log q_t(X_t)\right) = \frac{1}{\sqrt{1 - \sigma_t^2}}\left(X_t + \sigma_t^2 s_t(X_t)\right), \tag{8}$$

where we recall that $\sigma_t = \sqrt{1 - e^{-2t}}$. In other words, the score function is intimately connected with the posterior mean function. Throughout the paper, we shall denote by $\widehat{\mu}_t(\cdot)$ the estimate of the posterior mean $\mu_t(\cdot)$, which allows us to express the score estimate $\widehat{s}_t(\cdot)$ as

$$\widehat{\mu}_t(X_t) = \frac{1}{\sqrt{1 - \sigma_t^2}}\left(X_t + \sigma_t^2 \widehat{s}_t(X_t)\right). \tag{9}$$

3 Key observation: DDPM as an adaptively discretized SDE

While previous studies (e.g., Benton et al. (2024); Chen et al. (2023a, 2022b)) have explored the convergence properties of time discretization of the ideal reverse process (5) through the SDE framework, these results fall short of directly addressing the precise update rule (2) of the DDPM sampler. As pointed out by Li and Yan (2024a), the coefficients (or weights) adopted in the DDPM-type sampler are critical for effective adaptation

to unknown low dimensionality; in fact, even mild perturbation of the coefficients can lead to considerable performance degradation, a curious phenomenon that does not occur in the general full-dimensional setting.

In order to rigorously pin down the effectiveness of the DDPM update rule (2), we first uncover its equivalence to a carefully constructed SDE. To be precise, defining

$$\eta_t := \frac{\sqrt{1 - \sigma_t^2}}{\sigma_t^2} = \frac{e^{-t}}{1 - e^{-2t}} \tag{10}$$

for the sake of notational convenience, we introduce the following SDE:

- **An adaptively discretized reverse SDE.** Initialized at $Y_0 \sim \mathcal{N}(0, I_d)$, this SDE proceeds as

$$dY_t = \left(Y_t + 2 \left(\frac{\eta_{T-t}}{\eta_{T-t_n}} \cdot \frac{1}{\sigma_{T-t_n}^2} Y_{t_n} - \frac{1}{\sigma_{T-t}^2} Y_t \right) + \frac{2\eta_{T-t}}{\eta_{T-t_n}} \widehat{s}_{T-t_n}(Y_{t_n}) \right) dt + \sqrt{2} dB_t \tag{11a}$$

$$= \left(\left(1 - \frac{2}{\sigma_{T-t}^2} \right) Y_t + \frac{2\sqrt{1 - \sigma_{T-t}^2}}{\sigma_{T-t}^2} \widehat{\mu}_{T-t_n}(Y_{t_n}) \right) dt + \sqrt{2} dB_t \tag{11b}$$

in the n-th ($0 \leqslant n < N$) time interval $t \in [t_n, t_{n+1})$, where (B_t) stands for a standard Brownian motion in $\mathbb{R}^d$. Here, the equivalence of (11a) and (11b) follows immediately from (9).

Importantly, solving the above SDE (11) is equivalent to running the original DDPM update rule (2), as asserted by the following theorem.

Theorem 1. *Let $(Y_t)_{t \in [0,T]}$ be the solution to the SDE (11). Then for any $n = 0, \ldots, N - 1$, one can write*

$$Y_{t_{n+1}} = \frac{1}{\sqrt{\alpha_n}} \left(Y_{t_n} + (1 - \alpha_n) \widehat{s}_{T-t_n}(Y_{t_n}) \right) + \sqrt{\frac{(1 - \alpha_n)(1 - \overline{\alpha}_{n+1})}{1 - \overline{\alpha}_n}} \widetilde{Z}_n \tag{12}$$

for some random vectors $\{\widetilde{Z}_n\}$ in $\mathbb{R}^d$, where the $\widetilde{Z}_n$'s are statistically independent obeying $\widetilde{Z}_n \sim \mathcal{N}(0, I_d)$.

With this result in place, the equivalence of the solution (12) and the original DDPM update rule (2) is self-evident. The proof of Theorem 1 is postponed to Section 5.1.

Recall that in the general full-dimensional settings, previous work (e.g., Chen et al. (2023a)) suggested that discretizing different parametrization of (5) typically results in the same order of discretization errors, indicating that distinguishing between them may be unnecessary. This message, however, does not hold in the presence of low-dimensional structure, which merits further discussions as follows.

- *A reparameterized semi-linear SDE.* As can be readily seen, if one has access to the true score function (i.e., $\widehat{s}_t(\cdot) = s_t(\cdot)$), then SDE (11b) can be viewed as the discretization of the true reverse SDE under the exponential integrator scheme:

$$dY_t = \left(Y_t + 2s_{T-t}(Y_t) \right) dt + \sqrt{2} dB_t$$

$$= \left(\left(1 - \frac{2}{\sigma_{T-t}^2} \right) Y_t + \frac{2\sqrt{1 - \sigma_{T-t}^2}}{\sigma_{T-t}^2} \mu_{T-t}(Y_t) \right) dt + \sqrt{2} dB_t, \tag{13}$$

where the last line invokes the Tweedie formula (8). In other words, the key is to rearrange the drift term of the reverse process (5) into a different semi-linear form, with the nonlinear component reflecting the posterior mean rather than the score.

- *Adaptation to data geometry.* To elucidate why the semi-linear structure in (13) favors low-dimensional data structure, note that the discretization error depends heavily upon the smoothness of the non-linear drift term of the reparameterzied semi-linear SDE along the solution path. In (13), the non-linear drift term is proportional to the posterior mean of X_0 given X_{T-t}; informally, the posterior mean term acts as some sort of projection operator onto the support of the target data, which effectively exploits the (unknown) low-dimensional data geometry to enhance the smoothness of the solution path.

5

- *Parameterization matters.* It is worth noting that how we re-parameterize the drift term of (5) — namely, how we rearrange the semi-linear drift term therein — can exert significant influences upon the sampling efficiency. This can be elucidated via the inspiring example singled out in Li and Yan (2024a, Theorem 2). More concretely, when the target data distribution is isotropic Gaussian on a k-dimension linear subspace ($k \ll d$), the iteration complexity of the DDPM sampler is proportional to $O(\text{poly}(k))$ (Li and Yan, 2024a), but most of other DDPM-type update rules (i.e., the ones with different coefficients than the original DDPM sampler) can take $\Omega(\text{poly}(d))$ steps to yield the target sampling quality. Informally speaking, parameterizing the semi-linear drift term differently might result in a solution within $O(d/\text{poly}(N))$ KL divergence from the DDPM solution[3]; this discrepancy is negligible in the full-dimensional setting, but could blow up when k is exceedingly small compared to d and when N is proportional to k. All this underscores the critical importance of identifying a parameterization that adapts naturally to the intrinsic data geometry.

4 Main results

Armed with the equivalence between the original DDPM sampler and the SDE (11) (cf. Theorem 1), we can proceed to establish a sharp convergence theory for the DDPM sampler, leveraging the toolbox from the SDE literature. Before proceeding to our main theory, let us begin by imposing a couple of key assumptions.

4.1 Assumptions and key quantities

4.1.1 Low dimensionality of the target distribution

To introduce the intrinsic dimensionality of the target data distribution p_{data}, we take a look at how complex the support of p_{data} is. Here and throughout, we denote by $\mathcal{X}_{\mathsf{data}} \in \mathbb{R}^d$ the support of p_{data}, i.e., the closure of the intersection of all the sets $\mathcal{X}' \in \mathbb{R}^d$ such that $\mathbb{P}_{X_0 \sim p_{\mathsf{data}}}(X_0 \in \mathcal{X}') = 1$.

Given that we focus on data residing in the Euclidean space, we can measure the complexity through the Euclidean metric entropy below (Wainwright, 2019, Chapter 5):

Definition 1 (Covering number and metric entropy). *For any set $\mathcal{M} \subseteq \mathbb{R}^d$, the Euclidean covering number at scale $\varepsilon_0 > 0$, denoted by $N^{\mathsf{cover}}(\mathcal{M}, \|\cdot\|_2, \varepsilon_0)$, is the smallest s such that there exist points $x_1, \dots, x_s$ obeying*

$$\mathcal{M} \subseteq \bigcup_{i=1}^{s} B(x_i, \varepsilon_0),$$

where $B(x_i, \varepsilon_0) := \{x \in \mathbb{R}^d \mid \|x - x_i\|_2 \leqslant \varepsilon_0\}$. The metric entropy is defined as $\log N^{\mathsf{cover}}(\mathcal{M}, \|\cdot\|_2, \varepsilon_0)$.

Following Li and Yan (2024a), we make the following assumption to characterize the complexity of the data distribution. We shall often refer to k as the intrinsic dimension of $\mathcal{X}_{\mathsf{data}}$.

Assumption 1 (Intrinsic dimension). *Suppose that $C_0 > 0$ is some sufficiently large universal constant. For $\varepsilon_0 = k^{-C_0}$, the metric entropy of the support $\mathcal{X}_{\mathsf{data}}$ of the data distribution p_{data} satisfies*

$$\log N^{\mathsf{cover}}(\mathcal{X}_{\mathsf{data}}, \|\cdot\|_2, \varepsilon_0) \leqslant C_{\mathsf{cover}} k \log \frac{1}{\varepsilon_0}$$

for some universal constant $C_{\mathsf{cover}} > 0$.

Next, we adopt a fairly mild assumption regarding the boundedness of the support $\mathcal{X}_{\mathsf{data}}$, allowing the size of the data to scale polynomially in k.

Assumption 2. *Suppose that there exists some universal constant $C_R > 0$ such that*

$$\sup_{x \in \mathcal{X}_{\mathsf{data}}} \|x\|_2 \leqslant R, \qquad \text{where } R := k^{C_R}.$$

---

[3]Roughly speaking, in view of the Taylor expansion, we know that the discrepancy between different solutions is of order $O(1/\text{poly}(N))$; but since d-dimensional Gaussian noise is added in each step of the sampling procedure, this discrepancy can result in $O(d/\text{poly}(N))$ difference in KL divergence per step.

Remark 1. Note that in Assumption 2, c_R is allowed to be arbitrarily large, as long as it is a numerical constant. This means that R can be an arbitrarily large polynomial in k.

To help interpret the above assumptions, we single out several important examples below.

- *Linear subspace.* The metric entropy of a bounded subset of any k-dimensional linear subspace is at most $O(k \log(1/\varepsilon_0))$ (Vershynin, 2018, Section 4.2.1), and hence Assumption 1 covers linear subspaces studied by some prior work (e.g., Chen et al. (2023b)).

- *Low-dimensional manifold.* Assumption 1 covers various manifolds under mild regularity conditions, which have been commonly used in prior work to model low-dimensional structure. To make it precise, let us first introduce some geometric quantities of the manifold, whose geometric interpretations can be found in Lee (2018, Chapter 6). Suppose $\mathcal{M} \subseteq \mathbb{R}^d$ to be an isometrically embedded, compact, k-dimensional submanifold, then we can define:

  Definition 2. *The reach of $\mathcal{M}$, denoted by $\tau_{\mathcal{M}}$, is the distance between $\mathcal{M}$ and $\mathrm{Med}(\mathcal{M})$, where*

  $$\mathrm{Med}(\mathcal{M}) := \left\{ x \in \mathbb{R}^d \mid \text{there exists } p \neq q \in \mathcal{M} \text{ such that } \|p - x\|_2 = \|q - x\|_2 = \mathsf{dist}(x, \mathcal{M}) \right\},$$

  and $\mathsf{dist}(x, \mathcal{M})$ denotes the Euclidean distance of x to $\mathcal{M}$. Also, $\mathrm{vol}(\mathcal{M})$ is the volume, or equivalently, the k-dimensional Hausdorff measure, of the manifold $\mathcal{M}$.

  Roughly speaking, the reach of a manifold measures the scale of which the manifold can be regarded as a linear subspace. The larger the reach $\tau_{\mathcal{M}}$ is, the more regular the manifold appears. The volume of the manifold $\mathcal{M}$ is a measure of the size of the manifold, which generalizes the concept of volume in 3-dimensional Euclidean space. Instead of requiring data to be strictly supported on the manifold, Assumption 1 also allows the data to lie in the ε_0-enlargement of the manifold $\mathcal{M}$ defined as $\mathcal{M}^{\varepsilon_0} := \{x \in \mathbb{R}^d \mid \mathsf{dist}(x, \mathcal{M}) \leq \varepsilon_0\}$. This is stated formally in the following lemma, with the proof deferred to Appendix B.1.

  Lemma 1. *If the support of the target data distribution p_{data} is contained within the manifold $\mathcal{M}^{\varepsilon_0}$, i.e., $\mathcal{X}_{\mathsf{data}} \subseteq \mathcal{M}^{\varepsilon_0}$, and if $\tau_{\mathcal{M}} = \Omega(k^{-C_1})$ and $\mathrm{vol}(\mathcal{M}) = O(k^{C_2})$ for some universal constants $C_1, C_2 > 0$, then Assumption 1 holds.*

- *A general set with doubling dimension k.* Another notion for intrinsic dimension used in statistics for a general set in $\mathbb{R}^d$ is the doubling dimension (Dasgupta and Freund, 2008), as defined below.

  Definition 3 (Doubling dimension)**.** *The doubling dimension of $\mathcal{X}_{\mathsf{data}}$ is the smallest number s such that for all $x \in \mathcal{X}_{\mathsf{data}}$ and $r > 0$, $B(x, r) \cap \mathcal{X}_{\mathsf{data}}$ can be covered by 2^s Euclidean balls with radius $r/2$.*

  This notion of dimension can also be incorporated into our framework, as demonstrated by the following lemma (Kpotufe and Dasgupta, 2012, Lemma 6):

  Lemma 2. *If $\mathcal{X}_{\mathsf{data}}$ is contained within a ε_0-enlargement of a set that has doubling dimension k and if Assumption 2 holds, then Assumption 1 holds.*

- *Union of the above sets.* Our framework can also cover a union of the structures mentioned above, that is, Assumption 1 automatically holds when the data are supported on a union of $k^{O(k)}$ subsets, each of which satisfies Assumption 1.

4.1.2 Discretization time points

Recall that the discretization time points are $0 = t_0 < t_1 < \cdots < t_N < t_{N+1} = T$, with T the time horizon of the forward process. In addition, we also define throughout the associated time intervals as

$$\Delta_n := t_{n+1} - t_n \quad (0 \leq n \leq N) \qquad \text{and} \qquad \delta := \Delta_N = T - t_N < 1. \tag{14}$$

In particular, δ shall often be referred to as the early stopping time, the meaning of which shall be discussed momentarily when presenting our convergence theory.

Following prior diffusion model theory (e.g., Benton et al. (2024)), we introduce a key quantity, κ, defined to be the smallest quantity obeying

$$\Delta_n \leqslant \kappa \min\{1, T - t_n\} \qquad \text{for all } 0 \leqslant n \leqslant N. \tag{15}$$

To get a sense of the size of κ, it is helpful to bear in mind the following schedule that has been adopted in recent theoretical analysis (Benton et al., 2024; Chen et al., 2023a; Li et al., 2024b):

$$t_n = \begin{cases} \frac{2(T-1)n}{N}, & \text{if } 0 \leqslant n \leqslant N/2, \\ T - \delta^{2(n-N/2)/N}, & \text{if } N/2 < n \leqslant N, \end{cases} \tag{16}$$

where we assume N to be even for simplicity. In words, this schedule of time points can be divided into two phases: the first phase consists of the first $N/2$ time instances that are linearly spaced within $[0, T-1]$, whereas the second phase consists of the remaining $N/2$ time points that whose distance to T decays exponentially from 1 to δ. As shown in Benton et al. (2024, Corollary 1), this choice (16) satisfies

$$\kappa \asymp \frac{T + \log(1/\delta)}{N}. \tag{17}$$

For this reason, we can often interpret κ as the average time interval, or simply $\widetilde{O}(1/N)$ as T is typically taken to be $\widetilde{O}(1)$.

4.1.3 Score estimation errors

Given that typically only imperfect score estimates are available, we introduce another assumption concerning the mean squared score estimation errors.

Assumption 3. *Assume that the weighted sum of the mean squared score estimation errors is finite, i.e.,*

$$\sum_{n=0}^{N-1} \Delta_n \mathbb{E}\left[\left\|s_{T-t_n}(X_{T-t_n}) - \widehat{s}_{T-t_n}(X_{T-t_n})\right\|_2^2\right] \leqslant \varepsilon_{\mathsf{score}}^2 < \infty,$$

where the expectation is w.r.t. the distribution of the forward process (cf. (1)).

Treating the score learning phase as a black box, Assumption 3 singles out a single quantity, $\varepsilon_{\mathsf{score}} \geqslant 0$, that captures the accuracy of the pre-trained score functions, which is commonly used in recent development of diffusion model theory (Benton et al., 2024; Chen et al., 2023a; Dou et al., 2024; Li et al., 2024a; Li and Jiao, 2024; Li et al., 2023, 2024c; Li and Yan, 2024a,b; Liang et al., 2024). Notably, assuming a weighted average of the mean squared score errors is clearly less stringent than requiring the mean squared score estimation errors to hold uniformly over all time discretization points (Chen et al., 2023c, 2022b; Lee et al., 2022; Wu et al., 2024).

4.2 Convergence guarantees for DDPM

Armed with the aforementioned key assumptions, we are positioned to present our non-asymptotic convergence theory for the DDPM sampler (2). Here and throughout, we adopt the following convenient notation:

- q_t: the distribution of X_t of the forward process (1);
- p_t: the distribution of Y_t of the SDE specified by (11b);
- p_{output}: the distribution of Y_{t_N} (or equivalently, the distribution of the DDPM output after N steps).

Our convergence theory is stated below, whose proof is deferred to Section 5.2.

Theorem 2. *Suppose that Assumptions 1, 2 and 3 hold. Assume that $T > 1$, $\kappa \leqslant 0.9$, $1/\mathrm{poly}(k) \lesssim \delta < 1$ and $d = O(\mathrm{poly}(k))$. Then the DDPM sampler (2) satisfies*

$$\mathsf{KL}(q_\delta \,\|\, p_{\mathsf{output}}) \lesssim \kappa \min\left\{\mathbb{E}\big[\|X_0\|_2^2\big], Tk \log k\right\} + \kappa^2 Nk \log k + \varepsilon_{\mathsf{score}}^2 + \big(d + \mathbb{E}\big[\|X_0\|_2^2\big]\big)e^{-2T}. \tag{18}$$

The KL divergence upper bound in Theorem 2 can be divided into three parts. The first two terms on the right-hand side of (18) capture the time discretization error of the DDPM dynamics (namely, the error incurred when discretizing (13) as (11b)), the third term on the right-hand side of (18) reflects the effect of the score estimation error, whereas the fourth term corresponds to the initialization error (namely, the error induced by drawing Y_0 from $\mathcal{N}(0, I_d)$ rather than the distribution of X_T). Note that we are comparing the DDPM output with q_δ (i.e., the distribution of a slightly noisy version of the data) rather than $p_{\mathsf{data}} = q_0$, as adding a little Gaussian noise helps avoid some highly irregular cases. For this reason, $\delta > 0$ is commonly referred to as the early stopping time in prior theory. Note that δ can be taken to be very small, ensuring that q_δ is exceedingly close to p_{data}.

Next, we instantiate Theorem 2 to the specific discretization schedule described in (16), in order to obtain a clearer picture about how DDPM adapts to unknown low-dimensional structure.

Corollary 1. *Suppose that Assumptions 1, 2 and 3 hold. Choose $T = c_T \log(d/\varepsilon)$ for some large enough universal constant $c_T > 0$. Then the DDPM sampler (2) with the discretization schedule (16) achieves*

$$\mathsf{KL}(q_\delta \,\|\, p_{\mathsf{output}}) \lesssim \varepsilon^2 + \varepsilon_{\mathsf{score}}^2,$$

provided that $1/\mathrm{poly}(k) \lesssim \delta < 1$, $d = O\big(\mathrm{poly}(k)\big)$, and

$$N \geqslant \frac{c_N k \log^3(k/\varepsilon)}{\varepsilon^2} \tag{19}$$

for some large enough universal constant $c_N > 0$.

To interpret this result, consider, for simplicity, the case with perfect score estimation (i.e., $\varepsilon_{\mathsf{score}} = 0$) and an exceedingly small δ. Then Corollary 1 implies that: for fairly broad scenarios, it only takes the DDPM sampler

$$N = \widetilde{\Theta}(k/\varepsilon^2) \text{ steps}$$

to approximate the distribution of X_δ (i.e., the target data distribution p_{data} convoluted with exceedingly small Gaussian noise) to within ε^2 KL divergence. Encouragingly, this iteration complexity is linear in the intrinsic dimension k (modulo some logarithmic factor), which reduces to the state-of-the-art theory when $k = d$ (i.e., the case without low-dimensional structure) (Benton et al., 2024). It is noteworthy that all this is achieved without any prior knowledge about the potential low-dimensional structure in p_{data}, thereby unraveling the remarkable capability of DDPM to adapt to unknown low dimensionality. It is also worth noting that this linear dependency in k holds without any additional burn-in requirement; that is, its validity is guaranteed without requiring N to first exceed, say, some polynomial function in k or d.

To understand the optimality of DDPM in adapting to low dimensionality, it is helpful to first revisit the lower bounds established by Chen et al. (2022b, Theorem 7) and Benton et al. (2024, Appendix H) for the full-dimensional case $k = d$; namely, there exists a simple d-dimensional distribution p_{data} for which the KL error is lower bounded by some function scaling linearly in d. These arguments can be straightforwardly extended to the low-dimensional case, revealing the existence of a distribution p_{data} with intrinsic dimension k, that requires an iteration complexity scaling at least linearly with k, regardless of the sampler in use.

Comparison with the concurrent work Potaptchik et al. (2024). Assuming the data lies on a low-dimensional manifold, this concurrent work establishes an iteration complexity that scales linearly in the manifold dimension. While our results and theirs clearly share similarity, we point out several differences here. Firstly, in Potaptchik et al. (2024), the data is assumed to be strictly supported on a constant-scale manifold, and the density function (w.r.t. to the standard volume measure) needs to be both upper and lower bounded uniformly. In comparison, we consider a more general setting where the data support is only required to stay close to a low-dimensional structure whose size can grow polynomially in k; also, no boundedness requirement is imposed on the density function. In addition, our paper makes apparent an intimate connection between the DDPM sampler and a properly discretized SDE, which helps explain the rationale behind the adaptivity phenomenon. Finally, our proofs invoke the covering argument to accommodate broader low-dimensional structure, which differs from the techniques in Potaptchik et al. (2024).

5 Main analysis

5.1 Proof of Theorem 1

In what follows, we justify the equivalence of SDE (11b) and the update rule of the DDPM.

Step 1: Solving SDE (11b). We proceed by directly solving SDE (11b). Let $(Y_t)_{t\in[0,T-\delta]}$ be the solution to SDE (11b). Multiply Y_t by some differentiable function $f(\cdot)$ (to be specified shortly) and apply Itô's formula (Oksendal and Oksendal, 2003) to yield

$$\mathrm{d}\big(f(t)Y_t\big) = \big(f'(t) + f(t)(1 - 2\sigma_{T-t}^{-2})\big)Y_t\mathrm{d}t + 2f(t)\sqrt{1 - \sigma_{T-t}^2}\sigma_{T-t}^{-2}\widehat{\mu}_{T-t_n}(Y_{t_n})\mathrm{d}t + \sqrt{2}f(t)\mathrm{d}B_t, \qquad (20)$$

To get rid of the linear term in (20), a natural strategy is to choose $f(\cdot)$ to obey

$$f'(t) + f(t)\big(1 - 2\sigma_{T-t}^{-2}\big) = 0, \qquad \text{or equivalently,} \qquad f'(t) - f(t)\frac{1 + e^{-2(T-t)}}{1 - e^{-2(T-t)}} = 0, \qquad (21)$$

which forms a first-order ODE and admits a non-zero solution as follows

$$\frac{\mathrm{d}\big(\log f(t)\big)}{\mathrm{d}t} = \frac{1 + e^{-2(T-t)}}{1 - e^{-2(T-t)}} \qquad \implies \qquad f(t) = \frac{e^{-(T-t)}}{1 - e^{-2(T-t)}}. \qquad (22)$$

With this choice of $f(\cdot)$ in place, one can eliminate the linear term in (20) and obtain the simplified form

$$\mathrm{d}\big(f(t)Y_t\big) = 2\big(f(t)\big)^2\widehat{\mu}_{T-t_n}(Y_{t_n})\mathrm{d}t + \sqrt{2}f(t)\mathrm{d}B_t,$$

where we use the basic fact that $f(t) = \sqrt{1 - \sigma_{T-t}^2}\sigma_{T-t}^{-2}$. Integrating both sides from t_n to t_{n+1} results in

$$f(t_{n+1})Y_{t_{n+1}} = f(t_n)Y_{t_n} + 2\int_{t_n}^{t_{n+1}} \big(f(t)\big)^2\widehat{\mu}_{T-t_n}(Y_{t_n})\mathrm{d}t + \sqrt{2}\int_{t_n}^{t_{n+1}} f(t)\mathrm{d}B_t. \qquad (23)$$

In summary, this solves SDE (11b) with the aid of the choice (22) of $f(\cdot)$.

Step 2: Establishing the equivalence between (23) **and DDPM.** Armed with the solution (23), we are positioned to justify its equivalence to the DDPM update rule. From the quadratic variation of the Brownian motion, we can write, for each $0 \leqslant n < N$,

$$\sqrt{2}\int_{t_n}^{t_{n+1}} f(t)\mathrm{d}B_t = \left(\int_{t_n}^{t_{n+1}} 2\big(f(t)\big)^2\mathrm{d}t\right)^{1/2} \widetilde{Z}_n$$

for some vector $\widetilde{Z}_n \sim \mathcal{N}(0, I_d)$, where $\{\widetilde{Z}_n\}_{n=0,\dots,N-1}$ are statistically independent. This allows us to write

$$f(t_{n+1})Y_{t_{n+1}} = f(t_n)Y_{t_n} + \left(\int_{t_n}^{t_{n+1}} 2\big(f(t)\big)^2\mathrm{d}t\right)\widehat{\mu}_{T-t_n}(Y_{t_n}) + \left(\int_{t_n}^{t_{n+1}} 2\big(f(t)\big)^2\mathrm{d}t\right)^{1/2} \widetilde{Z}_n, \qquad (24)$$

By defining

$$\gamma_n = e^{-(T-t_n)},$$

we can explicitly calculate the integral as follows

$$\int_{t_n}^{t_{n+1}} 2\big(f(t)\big)^2\mathrm{d}t = \int_{t_n}^{t_{n+1}} \frac{2e^{-2(T-t)}}{(1 - e^{-2(T-t)})^2}\mathrm{d}t = \frac{\gamma_{n+1}^2 - \gamma_n^2}{(1 - \gamma_n^2)(1 - \gamma_{n+1}^2)}.$$

Substitution into equation (24) then gives

$$Y_{t_{n+1}} = \frac{1}{f(t_{n+1})}\left\{f(t_n)Y_{t_n} + \left(\int_{t_n}^{t_{n+1}} 2\big(f(t)\big)^2\mathrm{d}t\right)\widehat{\mu}_{T-t_n}(Y_{t_n}) + \left(\int_{t_n}^{t_{n+1}} 2\big(f(t)\big)^2\mathrm{d}t\right)^{1/2} \widetilde{Z}_n\right\}$$

$$= \frac{1 - \gamma_{n+1}^2}{\gamma_{n+1}} \left\{ \frac{\gamma_n}{1 - \gamma_n^2} Y_{t_n} + \frac{\gamma_{n+1}^2 - \gamma_n^2}{(1 - \gamma_n^2)(1 - \gamma_{n+1}^2)} \widehat{\mu}_{T-t_n}(Y_{t_n}) + \left(\frac{\gamma_{n+1}^2 - \gamma_n^2}{(1 - \gamma_n^2)(1 - \gamma_{n+1}^2)} \right)^{1/2} \widetilde{Z}_n \right\}$$

$$\overset{(a)}{=} \frac{1 - \gamma_{n+1}^2}{\gamma_{n+1}} \cdot \frac{\gamma_n}{1 - \gamma_n^2} Y_{t_n} + \frac{\gamma_{n+1}^2 - \gamma_n^2}{(1 - \gamma_n^2)\gamma_n\gamma_{n+1}} \left(Y_{t_n} + (1 - \gamma_n^2)\widehat{s}_{T-t_n}(Y_{t_n}) \right) + \sqrt{\frac{(\gamma_{n+1}^2 - \gamma_n^2)(1 - \gamma_{n+1}^2)}{(1 - \gamma_n^2)\gamma_{n+1}^2}} \widetilde{Z}_n$$

$$= \frac{\gamma_{n+1}}{\gamma_n} Y_{t_n} + \frac{\gamma_{n+1}^2 - \gamma_n^2}{\gamma_n\gamma_{n+1}} \widehat{s}_{T-t_n}(Y_{t_n}) + \sqrt{\frac{(\gamma_{n+1}^2 - \gamma_n^2)(1 - \gamma_{n+1}^2)}{(1 - \gamma_n^2)\gamma_{n+1}^2}} \widetilde{Z}_n,$$

where (a) follows from the relationship between $\widehat{\mu}_t$ and $\widehat{s}_t$ (cf. Eq. (9)). To finish up, recalling that $\alpha_n = \exp(2t_n - 2t_{n+1}) = \gamma_n^2/\gamma_{n+1}^2$ and $\overline{\alpha}_n = e^{-2(T-t_n)} = \gamma_n^2$ (cf. (3)), we arrive at

$$Y_{t_{n+1}} = \frac{1}{\sqrt{\alpha_n}} \left(Y_{t_n} + (1 - \alpha_n)\widehat{s}_{T-t_n}(Y_{t_n}) + \sqrt{\frac{(1 - \alpha_n)(\alpha_n - \overline{\alpha}_n)}{1 - \overline{\alpha}_n}} \widetilde{Z}_n \right),$$

which coincides with the DDPM update rule (2) introduced in Ho et al. (2020).

5.2 Proof of Theorem 2

We now present the proof of main convergence theory in Theorem 2. To control the KL divergence between the DDPM-type dynamics and our target distribution, a common approach is to connect the KL error with the ℓ_2 discretization error by means of Girsanov's theorem (e.g., Benton et al. (2024); Chen et al. (2023b, 2022b); Wu et al. (2024)). We shall follow this approach in this proof, while in the meantime taking into account the intrinsic low-dimensional structure of the target distribution. In particular, we shall invoke the covering argument to treat the low-dimensional structure approximately as the union of linear subspaces. In particular, As it turns out, the evolution of the posterior mean function, i.e., $\mu_t(\cdot)$, can be well calibrated along the forward process with high probability under this framework, which plays a key role in enabling our non-asymptotic convergence analysis.

Our main proof is composed of three steps, supported by a sequence of auxiliary lemmas whose proofs are deferred to Appendix A-B. Throughout the proof, we adopt the following convenient notation:

- Q: the measure of the reverse process (5);

- π_d: the standard Gaussian distribution in $\mathbb{R}^d$;

- P^T: the measure of the DDPM dynamics (11) when initialized with $Y_0 \sim X_T$;

- P^∞: the measure of DDPM dynamics (11) when initialized with $Y_0 \sim \pi_d \overset{\mathrm{d}}{=} X_\infty$, where X_∞ is the limit of the forward process (4) as $t \to \infty$.

Step 1: Bounding the KL divergence between two path measures. In this step, we list several lemmas that help connect the KL error of interest with the mean squared difference between the true and estimated posterior means; these results are now standard for analyzing DDPM-type samplers.

To begin with, we upper bound the KL error of interest, $\mathsf{KL}(q_\delta \,\|\, p_{\mathsf{output}})$, by means of the KL divergence between two path measures. This is accomplished through the following standard lemma; its proof is fairly standard and we refer the interested reader to Benton et al. (2024, Section 3.3).

Lemma 3. If $\mathsf{KL}(Q \,\|\, P^T) < \infty$, then we have

$$\mathsf{KL}(q_\delta \,\|\, p_{\mathsf{output}}) = \mathsf{KL}(q_\delta \,\|\, p_{t_N}) \leqslant \mathsf{KL}(Q \,\|\, P^\infty) = \mathsf{KL}(Q \,\|\, P^T) + \mathsf{KL}(q_T \,\|\, \pi_d).$$

In view of Lemma 3, to establish Theorem 2, it boils down to controlling $\mathsf{KL}(Q \,\|\, P^T)$ and $\mathsf{KL}(q_T \,\|\, \pi_d)$ separately. Let us start with the last term $\mathsf{KL}(q_T \,\|\, \pi_d)$, which is equivalent to analyzing the convergence rate of the forward process. The following lemma offers a simple upper bound on this term. Note that the exponential convergence of the OU process is well-known; we refer the reader to Chen et al. (2023a, Lemma 9) or Benton et al. (2024, Appendix G) for the proof this lemma.

11

Lemma 4. *Suppose that $\mathbb{E}[\|X_0\|^2] < \infty$. Then the forward OU process* (4) *converges exponentially fast to its stationary distribution in KL divergence, i.e., for any $T > 1$ one has*

$$\mathsf{KL}(q_T \,\|\, \pi_d) \lesssim \left(d + \mathbb{E}[\|X_0\|_2^2]\right) e^{-2T}.$$

As a result of Lemmas 3 and 4, it remains to control the term $\mathsf{KL}(Q \,\|\, P^T)$. The following lemma links $\mathsf{KL}(Q \,\|\, P^T)$ with the mean squared difference between the drift terms of SDEs (11b) and (13), a standard result that can be readily accomplished via the Girsanov theorem (Le Gall, 2016).

Lemma 5. *For any time discretization schedule with $t_0 = 0$ and $t_N = T - \delta$, it holds that*

$$\mathsf{KL}(Q \,\|\, P^T) \leqslant \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4} \mathbb{E}_Q\left[\left\|\mu_{T-t}(Y_t) - \widehat{\mu}_{T-t_n}(Y_{t_n})\right\|_2^2\right] \mathrm{d}t < \infty. \tag{25}$$

Proof. The proof follows from an application of Girsanov's Theorem to SDEs (11b) and (13), with the aid of a coupling argument. Given that this standard trick and its derivation has appeared in many recent papers (e.g., Chen et al. (2022b, Section 5.2) or Benton et al. (2024, Appendix F)), we omit it for brevity. □

In what follows, we shall often suppress the subscript Q in the expectation operator as long as it is clear from the context. To proceed, note that the middle expression in inequality (25) can be further bounded by the Cauchy-Schwarz inequality as

$$\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4} \mathbb{E}\left[\left\|\mu_{T-t}(Y_t) - \widehat{\mu}_{T-t_n}(Y_{t_n})\right\|_2^2\right] \mathrm{d}t$$

$$\lesssim \underbrace{\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \left(\frac{\eta_{T-t}}{\eta_{T-t_n}}\right)^2 \mathbb{E}\left[\left\|s_{T-t_n}(Y_{t_n}) - \widehat{s}_{T-t_n}(Y_{t_n})\right\|_2^2\right] \mathrm{d}t}_{=:\, T_1}$$

$$+ \underbrace{\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4} \mathbb{E}\left[\left\|\mu_{T-t}(Y_t) - \mu_{T-t_n}(Y_{t_n})\right\|_2^2\right] \mathrm{d}t}_{=:\, T_2}. \tag{26}$$

where we have used the definition (10) of η_t and the following decomposition (according to (9))

$$\mu_{T-t}(Y_t) - \widehat{\mu}_{T-t_n}(Y_{t_n}) = \left(\mu_{T-t_n}(Y_{t_n}) - \widehat{\mu}_{T-t_n}(Y_{t_n})\right) + \left(\mu_{T-t}(Y_t) - \mu_{T-t_n}(Y_{t_n})\right)$$

$$= \frac{\sigma_{t_n}^2}{\sqrt{1 - \sigma_{t_n}^2}}\left(s_{T-t_n}(Y_{t_n}) - \widehat{s}_{T-t_n}(Y_{t_n})\right) + \left(\mu_{T-t}(Y_t) - \mu_{T-t_n}(Y_{t_n})\right).$$

In the above decomposition (26), T_1 reflects the influence of the score estimation error, whereas T_2 captures the effect of the time discretization error. In the sequel, we shall upper bound T_1 and T_2 separately, and combine them to reach our final bound on $\mathsf{KL}(Q \,\|\, P^T)$.

Step 2: Bounding the effect of score estimation errors. To cope with the term T_1, we find it helpful to first establish the following result on the coefficients $\{\eta_t\}$, whose proof is provided in Appendix A.1.

Lemma 6. *There exists some universal constant $C_3 > 0$, independent of n, such that*

$$\eta_{T-t}/\eta_{T-t_n} \leqslant C_3 \qquad \text{for any } t \in [t_n, t_{n+1}).$$

With Lemma 6 in mind, we know the weight $(\eta_{T-t}/\eta_{T-t_n})^2$ can be uniformly bounded by $O(1)$. Thus,

$$T_1 = \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \left(\frac{\eta_{T-t}}{\eta_{T-t_n}}\right)^2 \mathbb{E}\left[\left\|s_{T-t_n}(Y_{t_n}) - \widehat{s}_{T-t_n}(Y_{t_n})\right\|_2^2\right] \mathrm{d}t$$

$$\lesssim \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \mathbb{E}\left[\left\|s_{T-t_n}(Y_{t_n}) - \widehat{s}_{T-t_n}(Y_{t_n})\right\|_2^2\right] \mathrm{d}t$$

$$= \sum_{n=0}^{N-1} (t_{n+1} - t_n) \mathbb{E}\left[\left\|s_{T-t_n}(Y_{t_n}) - \widehat{s}_{T-t_n}(Y_{t_n})\right\|_2^2\right] \mathrm{d}t \leqslant \varepsilon_{\mathsf{score}}^2, \tag{27}$$

where the last equality arises from Assumption 3.

Step 3: Bounding the effect of discretization errors. We now turn to the term T_2 in (26). Here, $(Y_t)_{t\in[0,T]}$ stands for the solution to the backward SDE (5) when initialized at $Y_0 \sim X_T$, which satisfies $X_t \stackrel{\mathrm{d}}{=} Y_{T-t}$ (cf. (6)). Towards this end, we intend to first analyze each term in the integral. Define

$$D_{s,t} := \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4} \mathbb{E}\left[\left\|\mu_{T-t}(Y_t) - \mu_{T-s}(Y_s)\right\|_2^2\right], \qquad t \geqslant s. \tag{28}$$

In the following, we shall demonstrate that $D_{s,t}$ is, in some sense, adaptive to the (unknown) low-dimension structure of the target distribution p_{data}. More specifically, we would like to prove that for a given s, $D_{s,t}$ changes slowly in t when p_{data} has a low-dimensional structure, i.e., $\partial_t D_{s,t}$ scales with k instead of d.

Towards this end, we begin with the following lemma, which leverages Itô's isometry to derive an alternative expression for $D_{s,t}$ that lends it well for subsequent analysis. The proof is provided in Appendix A.2.

Lemma 7. *Recall that $\Sigma_t(x) = \mathrm{Cov}[X_0|X_t = x]$. The quantity $D_{s,t}$ defined in (28) satisfies*

$$D_{s,t} = \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4}\left(\mathbb{E}\left[\mathrm{Tr}(\Sigma_{T-s}(Y_s))\right] - \mathbb{E}\left[\mathrm{Tr}(\Sigma_{T-t}(Y_t))\right]\right). \tag{29}$$

Given that $D_{s,t} \geqslant 0$ for $t \geqslant s$ (which follows directly from its definition (28)), Lemma 7 implies that

$$\mathbb{E}\left[\mathrm{Tr}\left(\Sigma_{T-t}(Y_t)\right)\right] \text{ is non-increasing in } t. \tag{30}$$

We are now ready to present a key lemma that can assist in upper bounding $D_{s,t}$; the proof can be found in Appendix A.3.

Lemma 8. *For all $t \in [0, T)$, there exists some universal constant $C_4 > 0$ such that*

$$\mathbb{E}\left[\mathrm{Tr}\left(\Sigma_{T-t}(Y_t)\right)\right] \leqslant C_4 \min\left\{\mathbb{E}[\|X_0\|_2^2], \frac{\sigma_{T-t}^2}{1 - \sigma_{T-t}^2} k \log k\right\}.$$

With the above derivations in mind, we are in a position to control T_2. First, let us rewrite T_2 using Lemma 7 as

$$T_2 = \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} D_{t_n,t}\mathrm{d}t = \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4}\left(\mathbb{E}\left[\mathrm{Tr}\left(\Sigma_{T-t_n}(Y_{t_n})\right)\right] - \mathbb{E}\left[\mathrm{Tr}\left(\Sigma_{T-t}(Y_t)\right)\right]\right)\mathrm{d}t. \tag{31}$$

In the following analysis, we decompose the time horizon $[0, T - \delta]$ into $[0, T - 1]$ and $[T - 1, T - \delta]$, and control the right-hand side of equation (31) on these two time intervals separately. Without loss of generality, suppose that the M-th discretization time point obeys $t_M = T - 1$.

(i) The case when $t \in [0, T - 1]$. By the monotonicity of σ_{T-t}, we have

$$\frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4} \leqslant \frac{1 - \sigma_{T-t}^2}{\sigma_1^4} = C_5(1 - \sigma_{T-t}^2) < C_5,$$

where $C_5 := 1/\sigma_1^4$ is a numerical constant. Direct calculation gives

$$\sum_{n=0}^{M-1} \int_{t_n}^{t_{n+1}} D_{t_n,t}\mathrm{d}t \leqslant \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} C_5\left(\mathbb{E}\left[\mathrm{Tr}\left(\Sigma_{T-t_n}(Y_{t_n})\right)\right] - \mathbb{E}\left[\mathrm{Tr}\left(\Sigma_{T-t}(Y_t)\right)\right]\right)\mathrm{d}t \tag{32}$$

$$\leqslant C_5\kappa\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_T(Y_0)\big)\big] \lesssim \kappa\mathbb{E}[\|X_0\|_2^2], \tag{33}$$

where the last line applies Lemma 8 to obtain an upper bound on $\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_T(Y_0)\big)\big]$. Another way to bound the left-hand side of inequality (32) is as follows:

$$
\begin{aligned}
\sum_{n=0}^{M-1}\int_{t_n}^{t_{n+1}} D_{t_n,t}\mathrm{d}t &\overset{(a)}{\leqslant} C_5\sum_{n=0}^{M-1}\int_{t_n}^{t_{n+1}}(1-\sigma_{T-t_{n+1}}^2)\Big(\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_n}(Y_{t_n})\big)\big] - \mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\big]\Big)\mathrm{d}t \\
&\overset{(b)}{\lesssim} (1-\sigma_{T-\kappa}^2)\kappa\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_T(Y_0)\big)\big] + \sum_{n=1}^{M-1}(\sigma_{T-t_n}^2 - \sigma_{T-t_{n+1}}^2)\kappa\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_n}(Y_{t_n})\big)\big] \\
&\overset{(c)}{\lesssim} \frac{(1-\sigma_{T-\kappa}^2)\sigma_T^2}{1-\sigma_T^2}\kappa k\log k + \sum_{n=1}^{M-1}(1-\sigma_{T-t_{n+1}}^2)\big(1-e^{-2(t_{n+1}-t_n)}\big)\cdot\frac{\sigma_{T-t_n}^2}{1-\sigma_{T-t_n}^2}\kappa k\log k \\
&\overset{(d)}{\lesssim} e^{2\kappa}\kappa k\log k + \sum_{n=1}^{M-1}(t_{n+1}-t_n)\kappa k\log k \\
&\overset{(e)}{\lesssim} \kappa k\log k + T\kappa k\log k, \tag{34}
\end{aligned}
$$

where (a) follows from Lemma 7 and the fact that $\sigma_{T-t}\leqslant 1$, (b) makes use of the fact that $t_{n+1}-t_n\leqslant\kappa$, (c) invokes Lemma 8, (d) relies on the facts that $\frac{1-\sigma_{T-t_{n+1}}^2}{1-\sigma_{T-t_n}^2} = e^{2(t_{n+1}-t_n)}\lesssim 1$, $\sigma_{T-t_n}\leqslant 1$ and $1-e^{-2(t_{n+1}-t_n)}\lesssim t_{n+1}-t_n$, and (e) holds since $\kappa\leqslant 0.9$ and $\sum_{n=1}^{M-1}(t_{n+1}-t_n)\leqslant T$. Combining inequalities (33) and (34) gives

$$\sum_{n=0}^{M-1}\int_{t_n}^{t_{n+1}} D_{t_n,t}\mathrm{d}t \lesssim \kappa\min\big\{\mathbb{E}[\|X_0\|_2^2], Tk\log k\big\}.$$

(ii) The case when $t\in[T-1, T-\delta]$. In this case, it is easily seen that

$$\frac{1-\sigma_{T-t}^2}{\sigma_{T-t}^4} = \frac{e^{-2(T-t)}}{(1-e^{-2(T-t)})^2} \lesssim \frac{1}{(T-t)^2}.$$

Taking this together with expression (31) yields

$$
\begin{aligned}
\sum_{n=M}^{N-1}\int_{t_n}^{t_{n+1}} D_{t_n,t}\mathrm{d}t &\lesssim \sum_{n=M}^{N-1}\int_{t_n}^{t_{n+1}}\frac{1}{(T-t)^2}\Big(\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_n}(Y_{t_n})\big)\big] - \mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\big]\Big)\mathrm{d}t \\
&\overset{(a)}{\leqslant} \sum_{n=M}^{N-1}\int_{t_n}^{t_{n+1}}\frac{1}{(T-t_{n+1})^2}\Big(\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_n}(Y_{t_n})\big)\big] - \mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_{n+1}}(Y_{t_{n+1}})\big)\big]\Big)\mathrm{d}t \\
&\overset{(b)}{\leqslant} \sum_{n=M}^{N-1}\frac{\kappa}{1-\kappa}\cdot\frac{1}{(T-t_{n+1})}\Big(\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_n}(Y_{t_n})\big)\big] - \mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_{n+1}}(Y_{t_{n+1}})\big)\big]\Big) \\
&\overset{(c)}{\leqslant} \frac{\kappa}{(1-\kappa)^2}\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_1(Y_{T-1})\big)\big] + \frac{\kappa^2}{(1-\kappa)^2}\sum_{n=M+1}^{N-1}\frac{1}{T-t_n}\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_{T-t_n}(Y_{t_n})\big)\big] \\
&\overset{(d)}{\lesssim} \frac{\kappa}{(1-\kappa)^2}\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_1(Y_{T-1})\big)\big] + \frac{\kappa^2}{(1-\kappa)^2}\sum_{n=M+1}^{N-1}\frac{1}{T-t_n}\cdot k\log k\frac{\sigma_{T-t}^2}{1-\sigma_{T-t}^2} \\
&\overset{(e)}{\lesssim} \kappa\min\big\{\mathbb{E}[\|X_0\|_2^2], k\log k\big\} + \kappa^2\sum_{n=M+1}^{N-1}k\log k \\
&\leqslant \kappa\min\big\{\mathbb{E}[\|X_0\|_2^2], k\log k\big\} + \kappa^2 Nk\log k. \tag{35}
\end{aligned}
$$

Here, (a) is due to the non-increasing property in (30); (b) and (c) invoke the following fact (see (15)):

$$t_{n+1}-t_n\leqslant\kappa(T-t_n) = \kappa(T-t_{n+1}) + \kappa(t_{n+1}-t_n) \implies t_{n+1}-t_n\leqslant\frac{\kappa}{1-\kappa}(T-t_{n+1});$$

(c) arises from Lemma 8; and (d) holds since $\kappa\leqslant 0.9$.

14

(iii) Combining these two cases. With inequalities (32) and (35) in place, we reach

$$T_2 = \sum_{n=0}^{M-1} \int_{t_n}^{t_{n+1}} D_{t_n,t} \mathrm{d}t + \sum_{n=M}^{N-1} \int_{t_n}^{t_{n+1}} D_{t_n,t} \mathrm{d}t$$
$$\lesssim \kappa \min\left\{ \mathbb{E}[\|X_0\|_2^2], Tk\log k \right\} + \kappa^2 Nk\log k. \tag{36}$$

Step 4: Putting all pieces together. Combining inequalities (27) and (36) with Lemma 5, we obtain

$$\mathsf{KL}(Q\,\|\,P^T) \leqslant T_1 + T_2 \lesssim \kappa \min\left\{ \mathbb{E}[\|X_0\|^2], Tk\log k \right\} + \kappa^2 Nk\log k + \varepsilon_{\mathsf{score}}^2. \tag{37}$$

Finally, taking this result together with Lemma 4 and Lemma 3 completes the proof of Theorem 2.

6 Discussion

This paper has made progress towards demystifying the unreasonable effectiveness of then DDPM when the target distribution has a low-dimensional structure. We have sharpened the past result by Li and Yan (2024a), by proving that the iteration complexity scales linearly with the intrinsic dimension k. In some sense, our results reveal that the DDPM sampler is optimally adaptive to unknown low dimensionlity, provided that a suitable discretization schedule is adopted. Our result is derived by exploiting the connection between the DDPM sampler and an adaptively discretized SDE, where the non-linear component of the drift time of this SDE is inherently low-dimensional.

We conclude this paper by pointing out several directions worthy of future investigation. To begin with, it remains unclear how the DDPM sampler adapts to unknown low dimensionality when other metrics (e.g., the total-variation distance or the Wasserstein distance) are used to measure distributional discrepancy. Another interesting direction is to examine the adaptivity of other mainstream samplers (e.g., the probability flow ODE (Chen et al., 2023d; Gao and Zhu, 2024; Huang et al., 2024; Li et al., 2024c; Song et al., 2021)) in the face of unknown low-dimensional structure. Moreover, this paper treats the training procedure as a black box and directly assumes a mean squared score estimation error bound. For distributions with low intrinsic dimensions, a desirable score learning algorithm woulclud adapt to the low-dimensional structure too, which has been the focus of a recent line of work (Azangulov et al., 2024; Chen et al., 2023b, 2024b; Kadkhodaie et al., 2023; Tang and Yang, 2024; Wang et al., 2024). It would thus be interesting to develop an end-to-end theory to accommodate both the pre-training phase and the sampling process. Finally, our analysis, which highlights the crucial role of proper parameterization and discretization, might shed light on how to design the coefficients for other algorithms in order to accommodate inherently structured data distributions.

Acknowledgements

A Proofs of the main lemmas

A.1 Proof of Lemma 6

For any $t \in [t_n, t_{n+1})$, the definition of κ gives $t - t_n < \kappa$. Therefore,

$$\frac{\eta_{T-t}}{\eta_{T-t_n}} = \frac{e^{-(T-t)}}{1-e^{-2(T-t)}} \cdot \frac{1-e^{-2(T-t_n)}}{e^{-(T-t_n)}} = e^{t-t_n} \cdot \frac{1-e^{-2(T-t_n)}}{1-e^{-2(T-t)}}$$
$$\leqslant e^{\kappa} \cdot \frac{1-e^{-2(T-t_n)}}{1-e^{-2(T-t_{n+1})}}.$$

Observing that

$$T - t_{n+1} = T - t_n - (t_{n+1} - t_n) \geqslant (1 - \kappa)(T - t_n),$$

we can obtain

$$\frac{\eta_{T-t}}{\eta_{T-t_n}} \leqslant e^\kappa \cdot \frac{1 - e^{-2(T-t_n)}}{1 - e^{-2(1-\kappa)(T-t_n)}} \lesssim e^\kappa \cdot \frac{1}{1 - \kappa} \lesssim 1.$$

A.2   Proof of Lemma 7

For notational simplicity, set $R_t := s_{T-t}(Y_t)$. We first present the SDE representation of R_t as follows

$$\mathrm{d}R_t = -R_t\mathrm{d}t + \sqrt{2}\nabla s_{T-t}(Y_t)\mathrm{d}B_t \tag{38}$$

with (B_t) the standard Brownian motion in $\mathbb{R}^d$, which can be derived using Itô's formula and can be found in the proof of Benton et al. (2024, Lemma 3). Recognizing that $\sigma_{T-t}^2 = 1 - e^{-2(T-t)}$, one can easily obtain

$$\mathrm{d}(\sigma_{T-t}) = -\frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}}\mathrm{d}t. \tag{39}$$

With equations (38) and (39) in mind, we arrive at the following SDE representation of $\mu_{T-t}(Y_t)$:

$$\begin{aligned}
\mathrm{d}\big(\mu_{T-t}(Y_t)\big) &= \mathrm{d}\left(\frac{1}{\sqrt{1 - \sigma_{T-t}^2}}Y_t + \frac{\sigma_{T-t}^2}{\sqrt{1 - \sigma_{T-t}^2}}R_t\right) \\
&= e^{-t}\big(-Y_t\mathrm{d}t + \mathrm{d}Y_t\big) - \frac{2 - \sigma_{T-t}^2}{\sqrt{1 - \sigma_{T-t}^2}}R_t\mathrm{d}t + \frac{\sigma_{T-t}^2}{\sqrt{1 - \sigma_{T-t}^2}}\mathrm{d}R_t \\
&= \frac{\sqrt{2}\sigma_{T-t}^2}{\sqrt{1 - \sigma_{T-t}^2}}\left(\nabla s_{T-t}(Y_t) + \frac{1}{\sigma_{T-t}^2}I_D\right)\mathrm{d}B_t. \tag{40}
\end{aligned}$$

Now, let us recall some useful properties for a general diffusion process $\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}B_t$, for which Itô's formula applied to $f(\cdot) = \|\cdot\|^2$ gives

$$\frac{\mathrm{d}(\|X_t\|_2^2)}{\mathrm{d}t} = 2\langle\mu(X_t, t), X_t\rangle + \|\sigma(X_t, t)\|_\mathrm{F}^2.$$

Thus, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[\|X_t\|_2^2\big] = 2\mathbb{E}\big[\langle\mu(X_t, t), X_t\rangle\big] + \mathbb{E}\big[\|\sigma(X_t, t)\|_\mathrm{F}^2\big].$$

Applying this general formula to the special case described in equation (40), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\big\|\mu_{T-t}(Y_t) - \mu_{T-s}(Y_s)\big\|_2^2\right] = \frac{2\sigma_{T-t}^4}{1 - \sigma_{T-t}^2}\mathbb{E}\left[\left\|\nabla s_{T-t}(Y_t) + \frac{1}{\sigma_{T-t}^2}I_D\right\|_\mathrm{F}^2\right]. \tag{41}$$

To further simplify the expression above, we resort to the following lemma that provides a useful characterization of the score function; see Benton et al. (2024, Lemma 5) for the proof.

Lemma 9. *The score function satisfies*

$$\nabla s_{T-t}(Y_t) = -\frac{1}{\sigma_{T-t}^2}I_d + \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4}\Sigma_{T-t}(Y_t).$$

Taking Lemma 9 collectively with equation (41), we can deduce that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\left\|\mu_{T-t}(Y_t) - \mu_{T-s}(Y_s)\right\|_2^2\right] = 2\frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4}\mathbb{E}\left[\left\|\Sigma_{T-t}(Y_t)\right\|_{\mathrm{F}}^2\right]. \tag{42}$$

Now, it remains to cope with the Frobenius norm of the posterior covariance in (42). To do so, we adopt the following result introduced in Benton et al. (2024, Lemma 1), a crucial property borrowed from the stochastic localization literature (El Alaoui and Montanari, 2022; Eldan, 2020).

Lemma 10. *For all $t \in [0, T)$, one has*

$$\mathbb{E}\left[\left\|\Sigma_{T-t}(Y_t)\right\|_{\mathrm{F}}^2\right] = -\frac{\sigma_{T-t}^4}{2(1 - \sigma_{T-t}^2)}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\right].$$

Combining all the results above, we arrive at

$$\begin{aligned}
D_{s,t} &= \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4}\mathbb{E}\left[\left\|\mu_{T-t}(Y_t) - \mu_{T-s}(Y_s)\right\|_2^2\right] \\
&= \frac{2(1 - \sigma_{T-t}^2)}{\sigma_{T-t}^4}\int_s^t \frac{1 - \sigma_{T-u}^2}{\sigma_{T-u}^4}\mathbb{E}\left[\left\|\Sigma_{T-u}(Y_u)\right\|_{\mathrm{F}}^2\right]\mathrm{d}u \\
&\overset{\mathrm{(a)}}{=} -\frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4}\int_s^t \mathrm{d}\Big(\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-u}(Y_u)\big)\right]\Big) \\
&= \frac{1 - \sigma_{T-t}^2}{\sigma_{T-t}^4}\Big(\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-s}(Y_s)\big)\right] - \mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\right]\Big),
\end{aligned}$$

where equality (a) applies Lemma 10. We have thus established the advertised result in Lemma 7.

A.3   Proof of Lemma 8

Recalling the fact that $(Y_t)_{t\in[0,T]} \overset{\mathrm{d}}{=} (X_{T-t})_{t\in[0,T]}$ (see (6)), we have

$$\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\right] = \mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-t}(X_{T-t})\big)\right]. \tag{43}$$

In view of this relation, we are allowed to prove Lemma 8 regarding the forward process $(X_t)_{t\in[0,T]}$.

By the monotonicity of $\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\right]$ stated in (30), we know that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\right] &\leqslant \mathbb{E}\left[\mathrm{Tr}\big(\Sigma_T(Y_0)\big)\right] = \mathbb{E}\left[\mathrm{Tr}\big(\Sigma_T(X_T)\big)\right] \\
&\leqslant \mathbb{E}_{X_T}\left[\mathbb{E}[\|X_0\|^2|X_T]\right] \leqslant \mathbb{E}[\|X_0\|^2],
\end{aligned} \tag{44}$$

where the last line has used the definition of Σ_T. Thus, everything comes down to showing that

$$\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_{T-t}(Y_t)\big)\right] \lesssim \frac{\sigma_{T-t}^2}{1 - \sigma_{T-t}^2}k\log k, \qquad \forall t \in [0, T),$$

which, according to equation (43), is equivalent to proving that

$$\mathbb{E}\left[\mathrm{Tr}\big(\Sigma_t(X_t)\big)\right] \lesssim \frac{\sigma_t^2}{1 - \sigma_t^2}k\log k, \qquad \forall t \in [0, T). \tag{45}$$

In the sequel, we would like to prove inequality (45) using Assumption 1 as well as a covering argument.

Define

$$s = N^{\mathsf{cover}}(\mathcal{X}_{\mathsf{data}}, \|\cdot\|_2, \varepsilon_0) \leqslant \exp(C'_{\mathsf{cover}}k\log k),$$

where C'_{cover} is a universal constant chosen as $C'_{\mathsf{cover}} = C_0 \cdot C_{\mathsf{cover}}$. Let $\{x_i^*\}_{1\leqslant i\leqslant s}$ be an ε_0-net of the data support $\mathcal{X}_{\mathsf{data}}$ (see the definition of an epsilon-net in Wainwright (2019)), and let $\{\mathcal{B}_i\}_{1\leqslant i\leqslant s}$ be the corresponding ε_0-covering of $\mathcal{X}_{\mathsf{data}}$ obeying $x_i^* \in \mathcal{B}_i$ for all $1 \leqslant i \leqslant s$. Without loss of generality, we choose

the ε-covering in a way such that $\mathcal{B}_i$ and $\mathcal{B}_j$ are disjoint for all $1 \leqslant i \neq j \leqslant s$, and $B(x_i^*, \varepsilon_0/2) \subset \mathcal{B}_i$ for all $1 \leqslant i \leqslant s$.

To make our analysis more clear, we find it helpful to define the following sets:

$$\mathcal{I} := \big\{ 1 \leqslant i \leqslant s \mid \mathbb{P}(X_0 \in \mathcal{B}_i) \geqslant \exp(-C_6 k \log k) \big\}, \tag{46a}$$

$$\mathcal{G} := \big\{ z \in \mathbb{R}^d \mid \|z\|_2 \leqslant 2\sqrt{d} + \sqrt{C_6 k \log k}, \quad \text{and}$$

$$\langle x_i^* - x_j^*, z \rangle \leqslant \sqrt{C_6 k \log k} \, \|x_i^* - x_j^*\|_2 \quad \text{for all } 1 \leqslant i, j \leqslant s \big\}, \tag{46b}$$

where $C_6 > 0$ is some sufficiently large constant. The index set $\mathcal{I}$ can be regarded as the collection of sets in ε_0-covering with large enough probability, and we can see later that $\mathcal{G}$ is a high-probability region of a standard Gaussian random vector. Further, define a high-probability region of the random vector X_t as:

$$\mathcal{T}_t := \left\{ \sqrt{1 - \sigma_t^2} x_0 + \sigma_t z \, \middle| \, x_0 \in \bigcup_{i \in \mathcal{I}} \mathcal{B}_i, z \in \mathcal{G} \right\},$$

whose high-probability property is confirmed by the following lemma. The proof of this lemma is provided in Appendix B.2.

Lemma 11. *For all $0 < t \leqslant T$, it holds that*

$$\mathbb{P}(X_t \notin \mathcal{T}_t) \leqslant \exp\left(-\frac{C_6}{4} k \log k \right).$$

To obtain tight control of $\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_t(X_t)\big) \big]$, we look at the events $\{X_t \in \mathcal{T}_t\}$ and $\{X_t \notin \mathcal{T}_t\}$ separately.

(i) The case when $X_t \in \mathcal{T}_t$. The following lemma characterizes the typical performance of X_0, whose proof can be found in Appendix B.3.

Lemma 12. *Consider any $x \in \mathcal{T}_t$, and suppose $x = \sqrt{1 - \sigma_t^2} x_0' + \sigma_t z$ with $x_0' \in \mathcal{B}_{i(x)}$ and $z \in \mathcal{G}$. We have*

$$\mathbb{P}\left(\|X_0 - x_{i(x)}^*\|_2 \geqslant \frac{\sigma_t}{\sqrt{1 - \sigma_t^2}} \cdot \sqrt{Ck \log k} \, \middle| \, X_t = x \right) \leqslant \exp\left(-\frac{C}{20} k \log k \right)$$

holds for all $C \geqslant C_7$ provided $C_7 > 0$ is some large enough constant.

Armed with the sub-Gaussian tail probability bound in Lemma 12, one can directly check that, for all $x \in \mathcal{T}_t$,

$$\mathrm{Tr}\big(\Sigma_t(x)\big) \leqslant \mathbb{E}\big[\|X_0 - x_{i(x)}^*\|_2^2 \mid X_t = x \big] \lesssim \frac{\sigma_t^2}{1 - \sigma_t^2} k \log k, \tag{47}$$

with the proviso that $\sigma_t^2/(1 - \sigma_t^2) \geqslant \sigma_\delta^2 = \Omega(1/\mathrm{poly}(k))$.

(ii) The case when $X_t \notin \mathcal{T}_t$. In this case, we can simply recall Assumption 2 to derive

$$\mathrm{Tr}\left(\Sigma_t(x)\right) \leqslant \mathbb{E}\big[\|X_0\|_2^2 \mid X_t = x \big] \leqslant \sup_{x \in \mathcal{X}_{\mathrm{data}}} \|x\|_2^2 \leqslant k^{2C_R}. \tag{48}$$

(iii) Putting all this together. Combining inequalities (47) and (48) with Lemma 11, we arrive at

$$\mathbb{E}\big[\mathrm{Tr}\big(\Sigma_t(X_t)\big) \big] \lesssim \mathbb{P}(X_t \in \mathcal{T}_t) \cdot \frac{\sigma_t^2}{1 - \sigma_t^2} k \log k + \mathbb{P}(X_t \notin \mathcal{T}_t) \cdot k^{2C_R}$$

$$\lesssim \frac{\sigma_t^2}{1 - \sigma_t^2} k \log k + \exp\left(-\frac{C_6}{4} k \log k \right) k^{2C_R}$$

$$\leqslant \frac{\sigma_t^2}{1 - \sigma_t^2} k \log k + \exp\left(-\frac{C_6}{5} k \log k \right)$$

$$\asymp \frac{\sigma_t^2}{1 - \sigma_t^2} k \log k, \tag{49}$$

provided that C_6 is sufficiently large and $\sigma_t^2/(1 - \sigma_t^2) = \Omega(1/\mathrm{poly}(k))$. Inequality (49) taken together with inequality (44) concludes the proof.

18

B Proofs of the auxiliary lemmas

B.1 Proof of Lemma 1

In this proof, we are in need of the following lemma concerning the covering number of a compact manifold.

Lemma 13. *For any $0 < \varepsilon_0 \leqslant \tau_{\mathcal{M}}$, the covering number of a compact manifold satisfies*

$$N^{\text{cover}}(\mathcal{M}, \|\cdot\|_2, \varepsilon_0) \leqslant \frac{\text{vol}(\mathcal{M})}{\omega_k} \left(\frac{\pi}{2}\right)^k \varepsilon_0^{-k},$$

where ω_k is the volume of the unit ball in $\mathbb{R}^k$.

The proof of Lemma 13 can be found in Block et al. (2022, Appendix A). It is worth noting that in the proof, the ε_0-covering is constructed as

$$\mathcal{M} \subseteq \bigcup_{i=1}^{s} B(x_i, \varepsilon_0),$$

where $s = N^{\text{cover}}(\mathcal{M}, \|\cdot\|_2, \varepsilon_0)$ and $x_i \in \mathcal{M}$ for all $i = 1, \ldots, s$. By direct calculation and Lemma 13, we have

$$N^{\text{cover}}(\mathcal{M}^{\varepsilon_0}, \|\cdot\|_2, \varepsilon_0) \leqslant N^{\text{cover}}(\mathcal{M}, \|\cdot\|_2, \varepsilon_0/2) \leqslant \frac{\text{vol}(\mathcal{M})}{\omega_k} \pi^k \varepsilon_0^{-k}.$$

Thus, we obtain an upper bound on the metric entropy:

$$\log N^{\text{cover}}(\mathcal{M}^{\varepsilon_0}, \|\cdot\|_2, \varepsilon_0) \leqslant k \log \varepsilon_0^{-1} + \log \text{vol}(\mathcal{M}) - \log \omega_k + k \log \pi$$
$$\leqslant k \left(\log \varepsilon_0^{-1} + (\text{vol}(\mathcal{M}))^{1/k} + C_4\right),$$

where C_4 is some numerical constant. Therefore, if $\mathcal{M}$ satisfies $\tau_{\mathcal{M}} = \Omega(k^{-C_1})$ and $\text{vol}(\mathcal{M}) = O(k^{C_2})$ for some large enough constants $C_1, C_2 > 0$, we arrive at

$$\log N^{\text{cover}}(\mathcal{M}^{\varepsilon_0}, \|\cdot\|_2, \varepsilon_0) \leqslant (C_0 + C_1 + C_2)k \log k \asymp k \log k,$$

which verifies Assumption 1.

B.2 Proof of Lemma 11

Since $\mathcal{T}_t$ is characterized by two components $\bigcup_{i \in \mathcal{I}} \mathcal{B}_i$ and $\mathcal{G}$, we analyze them separately. Let us begin with the set $\bigcup_{i \in \mathcal{I}} \mathcal{B}_i$, which satisfies

$$\mathbb{P}\left(X_0 \notin \bigcup_{i \in \mathcal{I}} \mathcal{B}_i\right) \overset{(a)}{\leqslant} \sum_{i \in \mathcal{I}} \exp(-C_6 k \log k)$$
$$\leqslant \exp\left(-(C_6 - C'_{\text{cover}})k \log k\right)$$
$$\overset{(b)}{\leqslant} \frac{1}{2} \exp\left(-\frac{C_6}{4} k \log k\right). \tag{50}$$

Here, we apply the union bound in inequality (a), and invoke the fact that C_6 is large enough (e.g., $C_6 > 8C'_{\text{cover}}$) in equality (b).

Regarding the set $\mathcal{G}$, we analyze it in a similar manner. First, we recall some standard concentration bounds below; the proof can be found in Wainwright (2019, Chapter 2).

Lemma 14. *For $Z \sim \mathcal{N}(0, I_d)$ and any $t > 1$, we have the following upper bound on the tail probability of Gaussian and chi-square random variables:*

$$\mathbb{P}(\|Z_1\|_2 \geqslant t) \leqslant e^{-t^2/2} \qquad \text{and} \qquad \mathbb{P}(\|Z\|_2^2 \geqslant \sqrt{d} + t) \leqslant e^{-t^2/2},$$

where Z_1 denotes the first coordinate of Z.

By virtue of Lemma 14, we can demonstrate that

$$
\begin{aligned}
\mathbb{P}(Z \notin \mathcal{G}) &\overset{(a)}{\leqslant} \mathbb{P}\left(\|Z\|_2 > \sqrt{d} + \sqrt{C_6 k \log k}\right) + \sum_{i=1}^{s}\sum_{j=1}^{s} \mathbb{P}\left(\langle x_i^* - x_j^*, Z\rangle > \sqrt{C_6 k \log k}\|x_i^* - x_j^*\|_2\right) \\
&\leqslant (s^2 + 1)\exp\left(-\frac{C_6}{2}k\log k\right) \\
&\leqslant \left(\exp(2C'_{\mathsf{cover}} k \log k) + 1\right)\exp\left(-\frac{C_6}{2}k\log k\right) \\
&\overset{(b)}{\leqslant} \frac{1}{2}\exp\left(-\frac{C_6}{4}k\log k\right),
\end{aligned}
\tag{51}
$$

where (a) applies the union bound, and (b) holds for large enough C_6 (e.g., $C_6 > 8C'_{\mathsf{cover}}$).

Since $X_t \overset{\mathrm{d}}{=} \sqrt{1 - \sigma_t^2}X_0 + \sigma_t Z$, combining inequalities (50) and (51) yields

$$
\mathbb{P}(X_t \notin \mathcal{T}_t) \leqslant \mathbb{P}\left(X_0 \notin \bigcup_{i\in\mathcal{I}}\mathcal{B}_i\right) + \mathbb{P}(Z \notin \mathcal{G}) \leqslant \exp\left(-\frac{C_6}{4}k\log k\right).
$$

B.3 Proof of Lemma 12

For ease of our presentation, we fine it convenient to define the following set

$$
\mathcal{E}_t(x) := \left\{ x_0 \,\Big|\, \|x_0 - x^*_{i(x)}\|_2 \geqslant \frac{\sigma_t}{\sqrt{1 - \sigma_t^2}} \cdot \sqrt{Ck\log k} \right\},
$$

where $C > C_7$ for some sufficiently large constant $C_7 > 0$ (to be specified momentarily). It suffices to bound $\mathbb{P}(\mathcal{E}_t(x) \mid X_t = x)$. Denote by $q_{X_t|X_0}(\cdot|x_0)$ the conditional density function of X_t given $X_0 = x$, and q_0 the measure induced by X_0. Making use of the Bayes formula, we can obtain

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_t(x) \mid X_t = x) &= \frac{\int_{\mathcal{E}_t(x)} q_{X_t|X_0}(x \mid x_0)\mathrm{d}q_0(x_0)}{\int_{\mathcal{X}_{\mathsf{data}}} q_{X_t|X_0}(x \mid \widetilde{x}_0)\mathrm{d}q_0(\widetilde{x}_0)} \\
&\leqslant \frac{\int_{\mathcal{E}_t(x)} q_{X_t|X_0}(x|x_0)\mathrm{d}q_0(x_0)}{\int_{\mathcal{B}_{i(x)}} q_{X_t|X_0}(x \mid \widetilde{x}_0)\mathrm{d}q_0(\widetilde{x}_0)} \\
&\leqslant \frac{\int_{\mathcal{E}_t(x)} q_{X_t|X_0}(x \mid x_0)\mathrm{d}q_0(x_0)}{\mathbb{P}_{X_0}(\mathcal{B}_{i(x)}) \inf_{\widetilde{x}_0\in\mathcal{B}_{i(x)}} q_{X_t|X_0}(x \mid \widetilde{x}_0)} \\
&\leqslant \frac{1}{\mathbb{P}_{X_0}(\mathcal{B}_{i(x)})} \cdot \frac{\sup_{x_0\in\mathcal{E}_t(x)} q_{X_t|X_0}(x \mid x_0)}{\inf_{\widetilde{x}_0\in\mathcal{B}_{i(x)}} q_{X_t|X_0}(x \mid \widetilde{x}_0)} \\
&\leqslant \exp\left(C_6 k\log k\right) \sup_{x_0\in\mathcal{E}_t(x),\widetilde{x}_0\in\mathcal{B}_{i(x)}} \exp\left\{\frac{1}{2\sigma_t^2}\left(\left\|x - \sqrt{1-\sigma_t^2}\widetilde{x}_0\right\|_2^2 - \left\|x - \sqrt{1-\sigma_t^2}x_0\right\|_2^2\right)\right\},
\end{aligned}
\tag{52}
$$

where the last inequality holds by exploiting $i(x) \in \mathcal{I}$ and the definition of $\mathcal{I}$ in (46a). The proof then boils down to upper bounding the exponential term above.

Given that $x \in \mathcal{T}_t$, we can decompose it as $x = \sqrt{1-\sigma_t^2}x'_0 + \sigma_t z$ for some $x'_0 \in \mathcal{B}_{i(x)}$ and $z \in \mathcal{G}$. For any $x_0 \in \mathcal{E}_t(x)$ and $\widetilde{x}_0 \in \mathcal{B}_{i(x)}$, suppose that $x_0 \in \mathcal{B}_j$, then direct expansion gives

$$
\begin{aligned}
&\left\|x - \sqrt{1-\sigma_t^2}\widetilde{x}_0\right\|_2^2 - \left\|x - \sqrt{1-\sigma_t^2}x_0\right\|_2^2 \\
&= -(1-\sigma_t^2)\|x'_0 - x_0\|_2^2 + 2\sqrt{\sigma_t^2(1-\sigma_t^2)}\langle x_0 - \widetilde{x}_0, z\rangle + (1-\sigma_t^2)\|x'_0 - \widetilde{x}_0\|_2^2 \\
&\overset{(a)}{\leqslant} -(1-\sigma_t^2)\left(\|x^*_{i(x)} - x^*_j\|_2 - 2\varepsilon_0\right)^2 + 2\sqrt{\sigma_t^2(1-\sigma_t^2)}\langle x_0 - \widetilde{x}_0, z\rangle + 4(1-\sigma_t^2)\varepsilon_0^2
\end{aligned}
$$

20

$$\overset{(b)}{\leqslant} -(1-\sigma_t^2)\big\|x_{i(x)}^* - x_j^*\big\|_2^2 + 4(1-\sigma_t^2)\big\|x_{i(x)}^* - x_j^*\big\|_2\varepsilon_0$$
$$+ 2\sqrt{\sigma_t^2(1-\sigma_t^2)}\left(\big\langle x_j^* - x_{i(x)}^*, z\big\rangle + 2\left(\sqrt{d} + \sqrt{C_6 k\log k}\right)\varepsilon_0\right), \qquad (53)$$

where inequality (a) holds for the property of the ε_0-net, and inequality (b) comes from the definition of $z \in \mathcal{G}$. Again, recall the definition of $\mathcal{G}$ to obtain

$$\big\langle x_j^* - x_{i(x)}^*, z\big\rangle \leqslant \sqrt{C_6 k\log k}\,\big\|x_j^* - x_{i(x)}^*\big\|_2. \qquad (54)$$

For a sufficiently large C_0, it can be easily checked that

$$\varepsilon_0 = k^{-C_0} \leqslant \sigma_\delta \min\left\{1, \sqrt{\frac{k\log k}{d}}\right\} \leqslant \frac{\sigma_t}{\sqrt{1-\sigma_t^2}}\min\left\{1, \sqrt{\frac{k\log k}{d}}\right\}, \qquad (55)$$

and we have

$$4\sqrt{\sigma_t^2(1-\sigma_t^2)}\left(\sqrt{d} + \sqrt{C_6 k\log k}\right)\varepsilon_0 \leqslant 5\sigma_t^2 k\log k. \qquad (56)$$

Substituting inequalities (54), (55) and (56) into (53), and rearranging terms, we arrive at

$$\left\|x - \sqrt{1-\sigma_t^2}\,\widetilde{x}_0\right\|_2^2 - \left\|x - \sqrt{1-\sigma_t^2}\,x_0\right\|_2^2 \leqslant -\frac{1-\sigma_t^2}{2}\big\|x_{i(x)}^* - x_j^*\big\|_2^2 + 5\sigma_t^2 k\log k$$
$$\overset{(a)}{\leqslant} -\frac{1-\sigma_t^2}{8}\big\|x_{i(x)}^* - x_0\big\|_2^2 + 5\sigma_t^2 k\log k$$
$$\leqslant -\frac{C}{8}\sigma_t^2 k\log k + 5\sigma_t^2 k\log k$$
$$\overset{(b)}{\leqslant} -\frac{C}{9}\sigma_t^2 k\log k,$$

where (a) holds by $\big\|x_{i(x)}^* - x_j^*\big\|_2 \geqslant 1/2\cdot\big\|x_{i(x)}^* - x_0\big\|_2$, a simple property of the covering, and (b) holds provided that $C \geqslant C_7$ is large enough. Taking the above bound collectively with inequality (52) yields

$$\mathbb{P}(\mathcal{E}_t(x) \mid X_t = x) \leqslant \exp\left(C_6 k\log k\right) \sup_{x_0\in\mathcal{E}_t(x),\widetilde{x}_0\in\mathcal{B}_{i(x)}} \exp\left\{\frac{1}{2\sigma_t^2}\left(\left\|x - \sqrt{1-\sigma_t^2}\,\widetilde{x}_0\right\|_2^2 - \left\|x - \sqrt{1-\sigma_t^2}\,x_0\right\|_2^2\right)\right\}$$
$$\leqslant \exp\left(C_6 k\log k\right)\cdot\exp\left(-\frac{C}{18}k\log k\right) \leqslant \exp\left(-\frac{C}{20}k\log k\right),$$

thus concluding the proof of this lemma.

References

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

Azangulov, I., Deligiannidis, G., and Rousseau, J. (2024). Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*.

Benton, J., Bortoli, V. D., Doucet, A., and Deligiannidis, G. (2024). Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*.

Block, A., Jia, Z., Polyanskiy, Y., and Rakhlin, A. (2022). Intrinsic dimension estimation using wasserstein distance. *Journal of Machine Learning Research*, 23(313):1–37.

Chen, H., Lee, H., and Lu, J. (2022a). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*.

Chen, H., Lee, H., and Lu, J. (2023a). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR.

Chen, M., Huang, K., Zhao, T., and Wang, M. (2023b). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR.

Chen, M., Mei, S., Fan, J., and Wang, M. (2024a). Opportunities and challenges of diffusion models for generative ai. *National Science Review*, page nwae348.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2023c). The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2022b). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.

Chen, S., Daras, G., and Dimakis, A. G. (2023d). Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. *arXiv preprint arXiv:2303.03384*.

Chen, S., Zhang, H., Guo, M., Lu, Y., Wang, P., and Qu, Q. (2024b). Exploring low-dimensional subspaces in diffusion models for controllable image editing. *arXiv preprint arXiv:2409.02374*.

Dasgupta, S. and Freund, Y. (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 537–546.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794.

Dou, Z., Kotekal, S., Xu, Z., and Zhou, H. H. (2024). From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

El Alaoui, A. and Montanari, A. (2022). An information-theoretic view of stochastic localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426.

Eldan, R. (2020). Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755.

Gao, X. and Zhu, L. (2024). Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*.

Haussmann, U. G. and Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. (2022). Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646.

Huang, D. Z., Huang, J., and Lin, Z. (2024). Convergence analysis of probability flow ode for score-based generative models. *arXiv preprint arXiv:2404.09730*.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. (2023). Generalization in diffusion models arises from geometry-adaptive harmonic representation. *arXiv preprint arXiv:2310.02557*.

Kpotufe, S. and Dasgupta, S. (2012). A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515.

Le Gall, J.-F. (2016). *Brownian motion, martingales, and stochastic calculus*. Springer.

Lee, H., Lu, J., and Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882.

Lee, H., Lu, J., and Tan, Y. (2023). Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR.

Lee, J. M. (2018). *Introduction to Riemannian manifolds*, volume 2. Springer.

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. (2024a). Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*.

Li, G., Huang, Z., and Wei, Y. (2024b). Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*.

Li, G. and Jiao, Y. (2024). Improved convergence rate for diffusion probabilistic models. *arXiv preprint arXiv:2410.13738*.

Li, G., Wei, Y., Chen, Y., and Chi, Y. (2023). Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2024c). A sharp convergence theory for the probability flow ODEs of diffusion models. *arXiv preprint arXiv:2408.02320*.

Li, G. and Yan, Y. (2024a). Adapting to unknown low-dimensional structures in score-based diffusion models. *arXiv preprint arXiv:2405.14861*.

Li, G. and Yan, Y. (2024b). $O(d/T)$ convergence theory for diffusion probabilistic models under minimal assumptions. *arXiv preprint arXiv:2409.18959*.

Liang, Y., Ju, P., Liang, Y., and Shroff, N. (2024). Broadening target distributions for accelerated diffusion models via a novel analysis approach. *arXiv preprint arXiv:2402.13901*.

Mbacke, S. D. and Rivasplata, O. (2023). A note on the convergence of denoising diffusion probabilistic models. *arXiv preprint arXiv:2312.05989*.

Moon, K. R., Stanley III, J. S., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46.

Oksendal, B. and Oksendal, B. (2003). *Stochastic differential equations*. Springer.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. (2021). The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*.

Potaptchik, P., Azangulov, I., and Deligiannidis, G. (2024). Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265.

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

Tang, R. and Yang, Y. (2024). Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR.

Tang, W. (2023). Diffusion probabilistic models. *preprint*.

Tang, W. and Zhao, H. (2024). Score-based diffusion models via stochastic differential equations–a technical tutorial. *arXiv preprint arXiv:2402.07487*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. (2024). Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*.

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., et al. (2023). De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100.

Wu, Y., Chen, Y., and Wei, Y. (2024). Stochastic runge-kutta methods: Provable acceleration of diffusion models. *arXiv preprint arXiv:2410.04760*.

Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., Peng, J., Kaplan, T., and Zeng, J. (2018). Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic acids research*, 46(8):e50–e50.