Journal of Machine Learning Research 26 (2025) 1-55 Submitted 2/25; Revised 9/25; Published 9/25

$O(d/T)$ Convergence Theory for Diffusion Probabilistic Models under Minimal Assumptions*

Gen Li GENLI@CUHK.EDU.HK
Department of Statistics and Data Science
The Chinese University of Hong Kong

Yuling Yan YULING.YAN@WISC.EDU
Department of Statistics
University of Wisconsin-Madison

Editor: Tong Zhang

Abstract

Score-based diffusion models, which generate new data by learning to reverse a diffusion process that perturbs data from the target distribution into noise, have achieved remarkable success across various generative tasks. Despite their superior empirical performance, existing theoretical guarantees are often constrained by stringent assumptions or suboptimal convergence rates. In this paper, we establish a fast convergence theory for the denoising diffusion probabilistic model (DDPM), a widely used SDE-based sampler, under minimal assumptions. Our analysis shows that, provided ℓ_2-accurate estimates of the score functions, the total variation distance between the target and generated distributions is upper bounded by $O(d/T)$ (ignoring logarithmic factors), where d is the data dimensionality and T is the number of steps. This result holds for any target distribution with finite first-order moment. Moreover, we show that with careful coefficient design, the convergence rate improves to $O(k/T)$, where k is the intrinsic dimension of the target data distribution. This highlights the ability of DDPM to automatically adapt to unknown low-dimensional structures, a common feature of natural image distributions. These results are achieved through a novel set of analytical tools that provides a fine-grained characterization of how the error propagates at each step of the reverse process.

Keywords: score-based generative model, diffusion model, denoising diffusion probabilistic model, sampling

1. Introduction

Score-based generative models (SGMs) have emerged as a powerful class of generative frameworks, capable of learning and sampling from complex data distributions (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b; Song and Ermon, 2019; Dhariwal and Nichol, 2021). These models, including Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) and Denoising Diffusion Implicit Models (DDIM) (Song et al., 2021a), operate by gradually transforming a simple noise-like distribution (e.g., standard Gaussian) into a target data distribution through a series of diffusion steps. This transformation is achieved by learning a sequence of denoising processes governed by score functions, which

*. This work was presented in part at ICLR 2025 (Li and Yan, 2025).

estimate the gradient of the log-density of the data at each step. SGMs have demonstrated remarkable success in various generative tasks, including image generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), audio generation (Kong et al., 2021), video generation (Villegas et al., 2022), and molecular design (Hoogeboom et al., 2022). See e.g., Yang et al. (2022); Croitoru et al. (2023) for overviews of recent development.

At the core of SGMs are two stochastic processes: a forward process, which progressively adds noise to the data,

$$X_0 \to X_1 \to \cdots \to X_T,$$

where X_0 is drawn from the target data distribution p_{data} and is gradually transformed into X_T that resembles standard Gaussian noise; and a reverse process,

$$Y_T \to Y_{T-1} \to \cdots \to Y_0,$$

which starts from pure Gaussian noise Y_T and sequentially converts it into Y_0 that closely mimics the target data distribution p_{data}. At each step, the distributions of Y_t and X_t are kept close. The key challenge lies in constructing this reverse process effectively to ensure accurate sampling from the target distribution.

Motivated by classical results on the time-reversal of stochastic differential equations (SDEs) (Anderson, 1982; Haussmann and Pardoux, 1986), SGMs construct the reverse process using the gradients of the log marginal density of the forward process, known as score functions. At each step, Y_{t-1} is generated from Y_t with the help of the score function $\nabla \log p_{X_t}(\cdot)$, where p_{X_t} denotes the density of X_t. Both the DDPM sampler (Ho et al., 2020) and the DDIM sampler (Song et al., 2021a) follow this denoising framework, with the key distinction being whether additional random noise is injected when generating each Y_{t-1}. Although the score functions are not known explicitly, they are pre-trained using large neural networks through score-matching techniques (Hyvärinen, 2005, 2007; Vincent, 2011; Song and Ermon, 2019).

Despite their impressive empirical success, the theoretical foundations of diffusion models remain relatively underdeveloped. Early studies on the convergence of SGMs (De Bortoli et al., 2021; De Bortoli, 2022; Liu et al., 2022; Pidstrigach, 2022; Block et al., 2020) did not provide polynomial convergence guarantees. In recent years, a line of works have explored the convergence of the generated distribution to the target distribution, treating the score-matching step as a black box and focusing on the effects of the number of steps T and the score estimation error on the convergence of the sampling phase (Chen et al., 2023d,a,c; Benton et al., 2023a; Lee et al., 2022, 2023; Li et al., 2023, 2024b; Li and Yan, 2024; Gao and Zhu, 2025; Huang et al., 2025; Tang and Zhao, 2024; Liang et al., 2024; Chen et al., 2023e). Recent studies have investigated the performance guarantees of SGMs in the presence of low-dimensional structures (e.g., Li and Yan (2024); Azangulov et al. (2024); Huang et al. (2024); Potaptchik et al. (2024); Wang et al. (2024); Chen et al. (2023b); Tang and Yang (2024); Tang et al. (2025); Chen et al. (2024)) and the acceleration of SGMs (e.g., Li et al. (2024a); Liang et al. (2024); Gupta et al. (2024); Li and Cai (2024); Li and Jiao (2025); Li et al. (2025)). Following this general avenue, the goal of this paper is to establish a sharp convergence theory for diffusion models with minimal assumptions.

Prior convergence guarantees. In recent years, a flurry of work has emerged on the convergence guarantees for the DDPM and DDIM type samplers. However, these prior studies fall short of providing a fully satisfactory convergence theory due to at least one of the following three obstacles:

- *Stringent data assumptions.* Earlier works, such as Lee et al. (2022), required the target data distribution to satisfy the log-Sobolev inequality. Similarly, Chen et al. (2023d); Lee et al. (2023); Chen et al. (2023c,e) assumed that the score functions along the forward process must satisfy a Lipschitz smoothness condition. More recent work Gao and Zhu (2025) relied on the strong log-concavity assumption of the target distribution to establish convergence guarantees in Wasserstein distance. These assumptions are often impractical to verify and may not hold for complex distributions commonly seen in image data. Some newer studies on the DDPM sampler (e.g., Chen et al. (2023a); Benton et al. (2023a)) and the DDIM sampler (e.g., Li et al. (2024b)) have relaxed these stringent assumptions, and their results applied to more general target distributions with bounded second-order moments or sufficiently large support.

- *Slow convergence rate.* We follow most existing works and focus on the total variation (TV) distance between the target and the generated distributions.[1] Let T be the number of steps and d be the dimensionality of the data. For the DDPM sampler, Chen et al. (2023d) established a convergence rate of $O(L\sqrt{(d+M_2)/T})$, where L is the Lipschitz constant of the score functions along the forward process, and M_2 is the second-order moment of the target distribution. Later, Chen et al. (2023a) lifted the Lipschitz condition, but this came at the cost of a rate $O(d/\sqrt{T})$ with worse dimension dependence. The state-of-the-art result for the DDPM samplers is due to Benton et al. (2023a), achieving a convergence rate of $O\sqrt{d/T})$. However, this is still slower than the convergence rate for the DDIM sampler, achieved in Li et al. (2024b), which attains $O(d/T)$ in the regime $T \gg d^2$.

- *Additional score estimation requirements.* Convergence theory for diffusion models must also account for the impact of imperfect score estimation on performance. While recent results for the DDPM sampler (Chen et al., 2023d,a; Benton et al., 2023a) require only ℓ_2-accurate score function estimates, another line of work on the DDIM sampler (Li et al., 2023, 2024b; Huang et al., 2025) achieves faster convergence rates, albeit under stricter requirements for score estimates. Specifically, Li et al. (2023, 2024b) require not only that the score estimates be close to the true score functions, but also that the Jacobian of the score estimates be close to the Jacobian of the true score functions, which is a significantly stronger condition. Additionally, Huang et al. (2025) assumes higher-order smoothness in the score estimates.

From this discussion, it is evident that while the state-of-the-art convergence rates for the DDIM sampler surpass those for the DDPM sampler, they also rely on more restrictive assumptions. This motivates us to think whether it is possible to achieve the best of both worlds, namely,

> *Can we establish a convergence theory for diffusion models that achieves a fast convergence rate under minimal data and score estimation assumptions?*

1. Convergence rates in Kullback-Leibler (KL) divergence in Chen et al. (2023a); Benton et al. (2023a) are transferred to TV distance using Pinsker's inequality, because the KL divergence is not a distance.

Sampler	Convergence rate (in TV distance)	Data assumption ($X_0 \sim p_{\mathsf{data}}$, $s_t^\star = \nabla \log p_{X_t}$)	Requirements on score estimates s_t
DDPM (Chen et al., 2023d)	$L\sqrt{d/T}$	L-Lipschitz $s_t^\star$; $\mathbb{E}[\|X_0\|_2^2] < \infty$	$s_t \approx s_t^\star$ in $L^2(p_{X_t})$
DDPM (Chen et al., 2023a)	$\sqrt{d^2/T}$	$\mathbb{E}[\|X_0\|_2^2] < \infty$	$s_t \approx s_t^\star$ in $L^2(p_{X_t})$
DDPM (Benton et al., 2023a)	$\sqrt{d/T}$	$\mathbb{E}[\|X_0\|_2^2] < \infty$	$s_t \approx s_t^\star$ in $L^2(p_{X_t})$
DDIM (Chen et al., 2023c)	$L^2\sqrt{d}/T$	L-Lipschitz $s_t^\star$; $\mathbb{E}[\|X_0\|_2^2] < \infty$	L-Lipschitz s_t; $s_t \approx s_t^\star$ in $L^2(p_{X_t})$
DDIM (Li et al., 2023)	$d^2/T + d^6/T^2$	bounded support	$s_t \approx s_t^\star$ in $L^2(p_{X_t})$; $J_{s_t} \approx J_{s_t^\star}$ in $L^2(p_{X_t})$
DDIM (Li et al., 2024b)	d/T when $T \gtrsim d^2$	bounded support	$s_t \approx s_t^\star$ in $L^2(p_{X_t})$; $J_{s_t} \approx J_{s_t^\star}$ in $L^2(p_{X_t})$
DDPM **(this paper)**	d/T	$\mathbb{E}[\|X_0\|_2] < \infty$	$s_t \approx s_t^\star$ in $L^2(p_{X_t})$

Table 1: Comparison with prior convergence guarantees for diffusion models (ignoring log factors). Convergence rates in KL divergence are transferred to TV distance using Pinsker's inequality. Here $J_f : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ denotes the Jacobian matrix of a function $f : \mathbb{R}^d \to \mathbb{R}^d$.

As noted in Li et al. (2024b), a counterexample demonstrates that ℓ_2-accurate score estimation alone is insufficient for convergence of the DDIM sampler under TV distance. The current paper answers this question affirmatively by focusing on the DDPM sampler.

Our contributions. This paper develops a fast convergence theory for the DDPM sampler under minimal assumptions. We show that the TV distance between the generated and target distributions is bounded by:

$$\frac{d}{T} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\big\|s_t(X_t) - s_t^\star(X_t)\big\|_2^2\big]},$$

up to logarithmic factors. The first term reflects the discretization error, while the second term accounts for score estimation error. Compared to the two most relevant works (Benton et al., 2023a; Li et al., 2024b) , which provide state-of-the-art results for the DDPM and DDIM samplers, our main contributions are as follows:

- *$O(d/T)$ convergence rate.* Under perfect score function estimation, we establish an $O(d/T)$ convergence rate for the DDPM sampler in TV distance, improving on the previous best rate of $O(\sqrt{d/T})$ from Benton et al. (2023a). Our result also matches the convergence rate of the DDIM sampler achieved in Li et al. (2024b), but is more general, as their result only holds when $T \gg d^2$, while ours applies for arbitrary T and d.

- *Minimal assumptions.* Our theory requires only that the target distribution has finite first-order moment, which, to the best of our knowledge, is the weakest data assumption in the current literature. Additionally, we require only ℓ_2-accurate score estimates, which is a significantly weaker condition than the Jacobian accuracy required by Li et al. (2023, 2024b).

- *Adaptivity to unknown low-dimensional structures.* Extending our theory, we demonstrate that with carefully designed coefficients (Li and Yan, 2024), the DDPM sampler achieves a convergence rate of $O(k/T)$ in TV distance, where k is the intrinsic dimension of the target data distribution. This improves upon the previous best bound of $O(\sqrt{k/T})$ established in Potaptchik et al. (2024); Huang et al. (2024).

In summary, our results achieve the fastest convergence rate in the literature for both DDPM and DDIM samplers while requiring minimal assumptions. A comparative summary with prior work is presented in Table 1.

2. Problem set-up

In this section, we provide an overview of the diffusion model and the DDPM sampler.

Forward process. We consider a Markov process in $\mathbb{R}^d$ starting from $X_0 \sim p_{\mathsf{data}}$, evolving according to the recursion:

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}W_t \quad (t = 1, \dots, T), \tag{1}$$

where $W_1, \dots, W_T$ are independent draws from $\mathcal{N}(0, I_d)$, and $\beta_1, \dots, \beta_t \in (0,1)$ are the learning rates. For each $1 \le t \le T$, define $\alpha_t \coloneqq 1 - \beta_t$ and $\overline{\alpha}_t \coloneqq \prod_{i=1}^{t} \alpha_i$. This allows us to express X_t in closed form as:

$$X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t \quad \text{where} \quad \overline{W}_t \sim \mathcal{N}(0, I_d). \tag{2}$$

We select the learning rates such that (i) β_t is small for every $1 \le t \le T$; and (ii) $\overline{\alpha}_T$ is vanishingly small, ensuring that the distribution of X_T is exceedingly close to $\mathcal{N}(0, I_d)$. In this paper, we adopt the following learning rate schedule

$$\beta_1 = \frac{1}{T^{c_0}}, \qquad \beta_{t+1} = \frac{c_1 \log T}{T} \min\left\{\beta_1\Big(1 + \frac{c_1 \log T}{T}\Big)^t, 1\right\} \quad (t = 1, \dots, T-1), \tag{3}$$

for sufficiently large constants $c_0, c_1 > 0$. This schedule is commonly used in the diffusion model literature (see, e.g., Li et al. (2023, 2024b); Liang et al. (2025)), although the results in this paper hold for any learning rate schedule satisfying the properties in Lemma 18.

Reverse process. The crucial elements in constructing the reverse process are the score functions associated with the marginal distributions of the forward diffusion process (1). For each $t = 1, \dots, T$, we define the score function as:

$$s_t^\star(x) \coloneqq \nabla \log p_{X_t}(x) \quad (t = 1, \dots, T),$$

where $p_{X_t}(\cdot)$ represents the smooth probability density of X_t. Since the true score functions are typically unknown, we assume access to estimates $s_t(\cdot)$ for each $s_t^\star(\cdot)$. To quantify the error in these estimates, we define the averaged ℓ_2 score estimation error as:

$$\varepsilon_{\mathsf{score}}^2 \coloneqq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\|s_t(X_t) - s_t^\star(X_t)\|_2^2\big].$$

This error term quantifies the effect of imperfect score approximation in our theoretical analysis. Using these score estimates, we can construct the reverse process, which starts from $Y_T \sim \mathcal{N}(0, I_d)$ and proceeds as

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t s_t(Y_t) + \sigma_t Z_t\big) \quad (t = T, \dots, 1), \tag{4}$$

where $Z_1, \dots, Z_T$ are independent draws from $\mathcal{N}(0, I_d)$. This is the popular DDPM sampler (Ho et al., 2020).

Notation. The total variation (TV) distance between two probability measures P and Q on a probability space $(\Omega, \mathcal{F})$ is define as

$$\mathsf{TV}(P, Q) \coloneqq \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \frac{1}{2}\int_{\Omega} |p(x) - q(x)| \mathrm{d}x,$$

where the last relation holds if P and Q have probability density functions $p(x)$ and $q(x)$. Let $\mathsf{KL}(P \,\|\, Q)$ denote the Kullback-Leibler (KL) divergence of P from Q, then Pinsker's inequality states that

$$\mathsf{TV}(P, Q) \le \sqrt{\frac{1}{2}\mathsf{KL}(P \,\|\, Q)}.$$

For any matrix A, we use $\|A\|$ and $\|A\|_{\mathrm{F}}$ to denote its spectral norm and Frobenius norm. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the support set of p_{data}, i.e., the smallest closed set $C \subseteq \mathbb{R}^d$ such that $p_{\mathsf{data}}(C) = 1$.

3. Main results

3.1 General theory: an $O(d/T)$ convergence bound

In this section, we will establish a fast convergence theory for the DDPM sampler under minimal assumptions. Before proceeding, we introduce the only data assumption that our theory requires.

Assumption 1 *The target distribution p_{data} has finite first-order moment. Furthermore, we assume that there exists some constant $c_M > 0$ such that*

$$M_1 \coloneqq \mathbb{E}[\|X_0\|_2] \le T^{c_M}.$$

Here we require the first-order moment M_1 to be at most polynomially large in the number of steps T, which allows cleaner and more concise result that avoids unnecessary technical complicacy. Since $c_M > 0$ can be arbitrarily large, we allow the target data distribution to have exceedingly large first-order moment, which is a mild assumption. In comparison, Assumption 1 is weaker than the finite second-order moment condition in e.g., Chen et al. (2023d,a); Benton et al. (2023a) and bounded support condition in e.g., Li et al. (2023, 2024b).

Now we are positioned to present our convergence theory for the DDPM sampler.

Theorem 1 *Suppose that Assumption 1 holds, and take the coefficients of the DDPM sampler (4) to be* $\eta_t = \sigma_t^2 = 1 - \alpha_t$. *Then there exists some universal constant* $c > 0$ *such that*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \le c\frac{d \log^3 T}{T} + c\varepsilon_{\mathsf{score}}\sqrt{\log T}. \tag{5}$$

The two terms in the error bound (5) correspond to discretization error and score matching error, respectively. A few remarks are in order.

- *Sharp convergence guarantees.* Consider the setting with perfect score estimation (i.e., $\varepsilon_{\mathsf{score}} = 0$) and ignore any log factor. Theorem 1 reveals that the DDPM sampler converges at the order of $O(d/T)$ in total variation distance. This significantly improves the state-of-the-art convergence rate $O(\sqrt{d/T})$ for the DDPM sampler (Benton et al., 2023a). Turning to the DDIM sampler

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \frac{1-\alpha_t}{2} s_t(Y_t)\big) \quad (t = T, \dots, 1), \qquad Y_T \sim \mathcal{N}(0, I_d), \tag{6}$$

  Li et al. (2024b) achieved the same convergence rate $O(d/T)$ in the regime $T \gg d^2$. Our result holds for general T and d, including the regime $T \asymp d$, hence is more general.

- *Stability vis-à-vis imperfect score estimation.* The score estimation error in (5) is linear in $\varepsilon_{\mathsf{score}}$, which suggests that the performance of the DDPM sampler degrades gracefully when the score estimates become less accurate. In other words, our theory holds with ℓ_2-accurate score estimates, which aligns with recent line of work on the DDPM sampler (Chen et al., 2023d,a; Benton et al., 2023a). In comparison, the convergence bound in Li et al. (2024b) for the DDIM sampler reads

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d}{T} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}} \quad \text{where} \quad \varepsilon_{\mathsf{Jacobi}} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big[\Big\| \frac{\partial s_t^\star}{\partial x}(X_t) - \frac{\partial s_t}{\partial x}(X_t)\Big\|\Big],$$

  which exhibits worse stability against imperfect score estimation. First, the dependency of their bound on $\varepsilon_{\mathsf{score}}$ is $\sqrt{d}$ times worse than ours. Second, their error bound involves an additional term that is proportional to $\varepsilon_{\mathsf{Jacobi}}$, which means that their theory requires the Jacobian of s_t need to be accurate in estimating the Jacobian of $s_t^\star$, which is a stringent requirement.

- *Minimal data assumption.* The only assumption required in Theorem 1 is that M_1 is at most polynomially large in T. In fact, by slightly modifying the proof, we can further relax Assumption 1 to accommodate target data distributions with polynomially large δ-th order moment

$$M_\delta := \big(\mathbb{E}[\|X_0\|_2^\delta]\big)^{1/\delta} \le T^{c_M},$$

  for any constant $\delta > 0$. The same bound (5) holds provided that $T \gg \max\{1, \delta^{-1}\} d \log^2 T$.

- *Error metric.* Theorem 1 provides convergence guarantees to p_{X_1} instead of the target data distribution (i.e., the distribution of X_0), which is similar to the results in e.g., Chen et al. (2023a); Benton et al. (2023a); Li et al. (2023, 2024b). On one hand, since

$X_1 = \sqrt{1-\beta_1}X_0 + \sqrt{\beta_1}$ and $\beta_1 = T^{-c_0}$ is vanishingly small, the distributions of X_1 and X_0 are exceedingly close. Hence $\mathsf{TV}(p_{X_1}, p_{Y_1})$ is a valid error metric. On the other hand, the smoothness of p_{X_1} allows us to circumvent imposing any Lipschitz assumption on the score functions, which provides technical benefit for the analysis.

It is worth noting that most previous studies on the convergence of the DDPM sampler (e.g., Chen et al. (2023d,a); Benton et al. (2023a); Li et al. (2023)) typically begin by upper bounding the squared TV error using the KL divergence of the forward process from the reverse process. This is done through the following argument:

$$\mathsf{TV}^2(p_{X_1}, p_{Y_1}) \le \frac{1}{2}\mathsf{KL}\left(p_{X_1} \| p_{Y_1}\right) \le \frac{1}{2}\mathsf{KL}\left(p_{X_1,\dots,X_T} \| p_{Y_1,\dots,Y_T}\right), \tag{7}$$

where the first inequality follows from Pinsker's inequality and the second from the data-processing inequality. The KL divergence on the right-hand side of this bound is more tractable and can be further bounded, for example, using Girsanov's theorem. In fact, (Chen et al., 2023d, Theorem 7) provides theoretical evidence that the KL divergence between the forward and reverse processes is lower bound by $\Omega(d/T)$, even when the target distribution is as simple as a standard Gaussian and perfect score estimates are available. This suggests that such an approach cannot yield error bounds better than $O(\sqrt{d/T})$ in general.

To achieve a sharper convergence rate, we take a different approach by directly analyzing the total variation error without resorting to intermediate KL divergence bounds. Specifically, we establish a fine-grained recursive relation that tracks how the error $\mathsf{TV}(p_{X_t}, p_{Y_t})$ propagates through the reverse process as t decreases from T to 1. Roughly speaking, we demonstrate that the discretization error incurred in the t-th step is roughly of order

$$\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 d\log T + (1-\alpha_t)^2\Big\|\frac{\partial s_t^\star}{\partial x}(x_t)\Big\|_\mathsf{F}^2,$$

where $s_t^\star(x) = \nabla \log p_{X_t}(x)$ is the score function at step t. This recursion—established in Lemma 6 via a careful characterization of the algorithmic dynamics—together with auxiliary tools such as Lemma 4, yields a sharp bound on how discretization error accumulates along the reverse process. The proof can be found in Section 4.

3.2 Adapting to unknown low-dimensional structure

In this section, we will provide convergence guarantees for the DDPM sampler when the target data distribution p_{data} exhibits low-dimensional structures. This scenario is particularly important, as natural image data are often concentrated on or near low-dimensional manifolds (Pope et al., 2021; Simoncelli and Olshausen, 2001). To formalize this, we define the intrinsic dimension of $\mathcal{X} = \mathsf{supp}(p_{\mathsf{data}})$ as follows.

Definition 2 (Intrinsic dimension) *Fix $\varepsilon = T^{-c_\varepsilon}$, where $c_\varepsilon > 0$ is some sufficiently large constant. We define the intrinsic dimension of $\mathcal{X}$ to be some quantity $k > 0$ such that*

$$\log N_\varepsilon(\mathcal{X}) \le C_{\mathsf{cover}} k \log T$$

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

This definition relates the intrinsic dimension k to the metric entropy of $\mathcal{X}$ (see e.g., Wainwright (2019)) and is standard in existing literature (e.g., Li and Yan (2024); Huang et al. (2024)). Importantly, it accommodates approximate low-dimensional structures by requiring only that $\mathcal{X}$ is concentrated near low-dimensional manifolds, which is more general than assuming exact low-dimensionality.

Recent work by Li and Yan (2024) demonstrated that the following coefficient design is essential for achieving nearly dimension-free convergence bounds for the DDPM sampler:

$$\eta_t^\star = 1 - \alpha_t \qquad \text{and} \qquad \sigma_t^{\star 2} = \frac{(1-\alpha_t)(\alpha_t - \overline{\alpha}_t)}{1-\overline{\alpha}_t}. \tag{8}$$

Specifically, Theorem 1 in Li and Yan (2024) established that under this coefficient design, the DDPM sampler converges at a rate of $O(k^2/\sqrt{T})$ in TV distance. Furthermore, Theorem 2 provided evidence that (8) is the unique coefficient design enabling nearly (ambient) dimension-free convergence.

Building on the techniques developed in the proof of Theorem 1, we strengthen this result by proving a faster $O(k/T)$ convergence bound under the same coefficient design.

Theorem 3 *Suppose that Assumption 1 holds. Take the coefficients of the DDPM sampler (4) to be $\eta_t = \eta_t^\star$ and $\sigma_t^2 = \sigma_t^{\star 2}$ (cf. (8)). Then there exists some universal constant $c > 0$ such that*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \le c\frac{k \log^3 T}{T} + c\varepsilon_{\mathsf{score}}\sqrt{\log T}, \tag{9}$$

where k is the intrinsic dimension of $\mathcal{X}$ (see Definition 2).

Consider the setup with perfect score estimation (i.e., $\varepsilon_{\mathsf{score}} = 0$) and disregard log factors. Theorem 3 demonstrates that, under the coefficient design in (8), the convergence rate of the DDPM sampler is $O(k/T)$, extending Theorem 1 to target data distributions with low-dimensional structure. While the importance of this coefficient design for achieving ambient dimension-free convergence is not new (see Li and Yan (2024)), our result significantly improves upon prior rates, which are of order $O(\sqrt{\mathsf{poly}(k)/T})$ (Li and Yan, 2024; Azangulov et al., 2024; Huang et al., 2024; Potaptchik et al., 2024). For a detailed comparison, refer to Table 3.2.

A concurrent and independent work (Liang et al., 2025) achieved the same adaptive convergence rate $O(k/T)$ for the DDPM sampler. It is worth mentioning that Liang et al. (2025) also covers the low-dimensional adaption of the DDIM sampler, which is beyond the scope of the current paper (see also Tang and Yan (2025)).

4. Proof of Theorem 1

4.1 Preliminaries

For each $1 \le t \le T$ and any $x \in \mathbb{R}^d$, it is known that the score function $s_t^\star(x)$ associated with p_{X_t} admits the following expression

$$s_t^\star(x) = -\frac{1}{1-\overline{\alpha}_t}\int p_{X_0 \mid X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0 =: -\frac{1}{1-\overline{\alpha}_t}g_t(x).$$

Sampler	Convergence rate (in TV distance)	Data assumption ($X_0 \sim p_{\mathsf{data}}$)	Intrinsic dimension k of $\mathcal{X} = \mathsf{supp}(p_{\mathsf{data}})$
DDPM (Li and Yan, 2024)	$\sqrt{k^4/T}$	bounded support	metric entropy (Definition 2)
DDPM (Azangulov et al., 2024)	$\sqrt{k^3/T}$	bounded support; smooth $p_{\mathsf{data}} \asymp 1$	manifold dimension
DDPM (Potaptchik et al., 2024)	$\sqrt{k/T}$	bounded support; smooth $p_{\mathsf{data}} \asymp 1$	manifold dimension
DDPM (Huang et al., 2024)	$\sqrt{k/T}$	bounded support	metric entropy (Definition 2)
DDPM **(this paper)**	k/T	$\mathbb{E}[\\|X_0\\|_2] < \infty$	metric entropy (Definition 2)

Table 2: Comparison with prior convergence rates (ignoring log factors) for the DDPM sampler that adapts to intrinsic low-dimensional structures. Convergence rates in KL divergence are transferred to TV distance using Pinsker's inequality.

Let $J_t(x) = \partial g_t(x)/\partial x$ be the Jacobian matrix of $g_t(x)$, which can be expressed as

$$J_t(x) = \frac{1}{1-\overline{\alpha}_t}\bigg\{\Big(\int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)\Big(\int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)^{\top} - \int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^{\top}\mathrm{d}x_0\bigg\} + I. \tag{10}$$

It is straightforward to check that $I - J_t(x_t) \succeq 0$. The following lemma will be useful in the analysis.

Lemma 4 *Suppose that $x \in \mathbb{R}^d$ satisfies $-\log p_{X_t}(x) \le \theta d \log T$ for any given $\theta \ge 1$. Then we have*

$$\|s_t^{\star}(x)\|_2 \le 5\sqrt{\frac{(\theta + c_0)d\log T}{1-\overline{\alpha}_t}} \qquad \text{and} \qquad \mathsf{Tr}(I - J_t(x)) \le 12(\theta + c_0)d\log T,$$

where the constant $c_0 > 0$ is defined in (3). In addition, there exists universal constant $C_0 > 0$ such that

$$\sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\int_{x_t}\|J_t(x_t)\|_{\mathrm{F}}^2\, p_{X_t}(x_t)\mathrm{d}x_t \le C_0 d\log T.$$

Proof See Appendix A.1. ■

For some sufficiently large constants $C_1, C_2 > 0$, we define for each $2 \le t \le T$ the set

$$\mathcal{E}_{t,1} \coloneqq \big\{x_t : -\log p_{X_t}(x_t) \le C_1 d\log T, \|x_t\|_2 \le \sqrt{\overline{\alpha}_t}T^{2c_R} + C_2\sqrt{d(1-\overline{\alpha}_t)\log T}\big\}. \tag{11}$$

We view each $\mathcal{E}_{t,1}$ as the high-probability (or typical) set for both X_t and Y_t. Our analysis tracks the density evolution inside $\mathcal{E}_{t,1}$, while showing that the contribution from its complement is negligible. We also define the extended d-dimensional Euclidean space $\mathbb{R}^d \cup \{\infty\}$ by adding a point ∞ to $\mathbb{R}^d$. From now on, the random vectors can take value in $\mathbb{R}^d \cup \{\infty\}$, namely, they can be constructed in the following way:

$$X = \begin{cases} X', & \text{with probability } \theta, \\ \infty, & \text{with probability } 1-\theta, \end{cases}$$

where $\theta \in [0,1]$ and X' is a random vector in $\mathbb{R}^d$ in the usual sense. If X' has a density, denoted by $p_{X'}(\cdot)$, then the generalized density of X is

$$p_X(x) = \theta p_{X'}(x)\,\mathbb{1}\{x \in \mathbb{R}^d\} + (1-\theta)\delta_\infty.$$

To simplify presentation, we will abbreviate generalized density to density.

4.2 Step 1: introducing auxiliary sequences

We first define an auxiliary reverse process that uses the true score function:

$$Y_T^\star \sim \mathcal{N}(0, I_d), \qquad Y_{t-1}^\star = \frac{1}{\sqrt{\alpha_t}}\Big(Y_t^\star + (1-\alpha_t)s_t^\star(Y_t^\star) + \sqrt{1-\alpha_t}Z_t\Big) \quad \text{for } t = T, \dots, 1. \tag{12}$$

To control discretization error, we introduce an auxiliary sequence $\{\overline{Y}_t : t = T, \dots, 1\}$ along with intermediate variables $\{\overline{Y}_t^- : t = T, \dots, 1\}$ as follows.

1. (Initialization) Define $\overline{Y}_T^- = Y_T$ if $Y_T \in \mathcal{E}_{T,1}$ and $\overline{Y}_T^- = \infty$ otherwise. The density of $\overline{Y}_T^-$ is
$$p_{\overline{Y}_T^-}(y_T^-) = p_{Y_T}(y_T^-)\,\mathbb{1}\left\{y_T^- \in \mathcal{E}_{T,1}\right\} + \int_{y \in \mathcal{E}_{T,1}^{\mathrm{c}}} p_{Y_T}(y)\mathrm{d}y\,\delta_\infty. \tag{13a}$$

2. (Transition from $\overline{Y}_t^-$ to $\overline{Y}_t$) For $t = T, \dots, 1$, the conditional density of $\overline{Y}_t$ given $\overline{Y}_t^- = y_t^-$ is
$$p_{\overline{Y}_t \mid \overline{Y}_t^-}(y_t \,|\, y_t^-) = \min\left\{\frac{p_{X_t}(y_t^-)}{p_{\overline{Y}_t^-}(y_t^-)}, 1\right\}\delta_{y_t^-} + \left(1 - \min\left\{\frac{p_{X_t}(y_t^-)}{p_{\overline{Y}_t^-}(y_t^-)}, 1\right\}\right)\delta_\infty. \tag{13b}$$

3. (Transition from $\overline{Y}_t$ to $\overline{Y}_{t-1}^-$) For $t = T, \dots, 2$, the conditional density of $\overline{Y}_{t-1}^-$ given $\overline{Y}_t = y_t$ is defined as follows: if $y_t \in \mathcal{E}_{t,1}$, then
$$p_{\overline{Y}_{t-1}^- \mid \overline{Y}_t}(y_{t-1}^- \,|\, y_t) = p_{Y_{t-1}^\star \mid Y_t^\star}(y_{t-1}^- \,|\, y_t); \tag{13c}$$
otherwise, we let $p_{\overline{Y}_{t-1}^- \mid \overline{Y}_t}(y_{t-1}^- \,|\, y_t) = \delta_\infty$.

This defines a Markov chain

$$Y_T \to \overline{Y}_T^- \to \overline{Y}_T \to \overline{Y}_{T-1}^- \to \overline{Y}_{T-1} \to \cdots \to \overline{Y}_1^- \to \overline{Y}_1. \tag{14}$$

An important consequence of the construction (13b) is that, for any $y_t \neq \infty$,

$$p_{\overline{Y}_t}(y_t) = \int_{\mathbb{R}^d} p_{\overline{Y}_t|\overline{Y}_t^-}(y_t \,|\, y_t^-) p_{\overline{Y}_t^-}(y_t^-) \mathrm{d}y_t^- = \min\big\{p_{X_t}(y_t), p_{\overline{Y}_t^-}(y_t)\big\}. \tag{15}$$

The following remark provides intuition for introducing this auxiliary sequence.

Remark 5 *Our analysis controls the discretization error and estimation error separately, which leads us to bound the TV distance between Y_t and its idealized counterpart $Y_t^\star$. To control this term, we need an ℓ_2 score estimation error weighted by the density of $Y_t^\star$, which is different from $\varepsilon_{\mathsf{score}}$ weighted by the density of X_t. Fortunately, the property of TV distance allows us to focus only on the difference between $\min\{p_{Y_t^\star}, p_{X_t}\}$ and p_{X_t}. This motivates us to construct the auxiliary sequence $\{\overline{Y}_t : t = T, \dots, 1\}$, whose density $p_{\overline{Y}_t}$ serves as a lower bound of $\min\{p_{Y_t^\star}, p_{X_t}\}$ according to (15). Finally, we move any mass outside the typical set $\mathcal{E}_{t,1}$ to infinity, which further simplifies the analysis.*

To control estimation error, we introduce another auxiliary sequence $\{\widehat{Y}_t : t = T, \dots, 1\}$ along with intermediate variables $\{\widehat{Y}_t^- : t = T, \dots, 1\}$ as follows.

1. (Initialization) Let $\widehat{Y}_T^- = \overline{Y}_T^-$.

2. (Transition from $\widehat{Y}_t^-$ to $\widehat{Y}_t$) For $t = T, \dots, 1$, the conditional density of $\widehat{Y}_t$ given $\widehat{Y}_t^- = y_t^-$ is

$$p_{\widehat{Y}_t|\widehat{Y}_t^-}(y_t \,|\, y_t^-) = p_{\overline{Y}_t|\overline{Y}_t^-}(y_t \,|\, y_t^-). \tag{16a}$$

3. (Transition from $\widehat{Y}_t$ to $\widehat{Y}_{t-1}^-$) For $t = T, \dots, 2$, the conditional density of $\widehat{Y}_{t-1}^-$ given $\widehat{Y}_t = y_t$ is defined as follows: if $y_t \in \mathcal{E}_{t,1}$, then

$$p_{\widehat{Y}_{t-1}^-|\widehat{Y}_t}(y_{t-1}^- \,|\, y_t) = p_{Y_{t-1}|Y_t}(y_{t-1}^- \,|\, y_t); \tag{16b}$$

   otherwise, we let $p_{\widehat{Y}_{t-1}^-|\widehat{Y}_t}(y_{t-1}^- \,|\, y_t) = \delta_\infty$.

This defines another Markov chain

$$Y_T \to \widehat{Y}_T^- \to \widehat{Y}_T \to \widehat{Y}_{T-1}^- \to \widehat{Y}_{T-1} \to \cdots \to \widehat{Y}_1^- \to \widehat{Y}_1, \tag{17}$$

which is similar to (14) except that now the transitions from $\widehat{Y}_t$ to $\widehat{Y}_{t-1}^-$ are constructed using the estimated score functions. We can use induction to show that

$$p_{Y_t}(y_t) \geq p_{\widehat{Y}_t}(y_t), \qquad \forall\, y_t \neq \infty \tag{18}$$

holds for all $t = T, \dots, 1$. First, it is straightforward to check that (18) holds for $t = T$. Suppose that (18) holds for $t+1$. Then for any $y_t \neq \infty$, we have

$$\begin{aligned}
p_{\widehat{Y}_t}(y_t) &= \int_{\mathbb{R}^d} p_{\widehat{Y}_t|\widehat{Y}_t^-}(y_t \,|\, y_t^-) p_{\widehat{Y}_t^-}(y_t^-) \mathrm{d}y_t^- \overset{\text{(i)}}{=} \min\big\{p_{X_t}(y_t)/p_{\overline{Y}_t^-}(y_t), 1\big\} p_{\widehat{Y}_t^-}(y_t) \leq p_{\widehat{Y}_t^-}(y_t) \\
&= \int_{\mathbb{R}^d} p_{\widehat{Y}_t^-|\widehat{Y}_{t+1}}(y_t \,|\, y_{t+1}) p_{\widehat{Y}_{t+1}}(y_{t+1}) \mathrm{d}y_{t+1} \\
&\overset{\text{(ii)}}{\leq} \int p_{Y_t|Y_{t+1}}(y_t \,|\, y_{t+1}) p_{Y_{t+1}}(y_{t+1}) \mathrm{d}y_{t+1} = p_{Y_t}(y_t).
\end{aligned}$$

Here step (i) follows from (16a) and (13b), while step (ii) follows from the induction hypothesis and (16b).

4.3 Step 2: controlling discretization error

In this section, we will bound the total variation distance between p_{X_1} and $p_{\overline{Y}_1}$. For each $t = T, \dots, 1$, let

$$\Delta_t(x) \coloneqq p_{X_t}(x) - p_{\overline{Y}_t}(x), \qquad \forall\, x \in \mathbb{R}^d. \tag{19}$$

We emphasize that $\Delta_t(\cdot)$ is not defined at ∞. In view of (15), we know that $\Delta_t(x_t) \geq 0$ for any $x_t \neq \infty$. The following lemma characterizes the propagation of the error $\int \Delta_t(x)\mathrm{d}x$ through the reverse process.

Lemma 6 *There exists some universal constant $C_4 > 0$ such that, for $t = T, \dots, 2$,*

$$\int \Delta_{t-1}(x)\mathrm{d}x \leq \int \Delta_t(x)\mathrm{d}x + C_4\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t\in\mathcal{E}_{t,1}} \big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big) p_{X_t}(x_t)\mathrm{d}x_t + T^{-3}.$$

In addition, we have $\int \Delta_T(x)\mathrm{d}x \leq T^{-4}$.

Proof See Appendix A.2. ■

We can apply Lemma 6 recursively to achieve

$$\begin{aligned}
\int \Delta_1(x)\mathrm{d}x &\leq \int \Delta_T(x)\mathrm{d}x + T^{-2} + \sum_{t=2}^{T} C_4\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t\in\mathcal{E}_{t,1}} (d\log T + \|J_t(x_t)\|_{\mathrm{F}}^2) p_{X_t}(x_t)\mathrm{d}x_t \\
&\overset{\text{(a)}}{\leq} 8c_1C_4\frac{\log T}{T}\sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\int_{x_t\in\mathcal{E}_{t,1}} \|J_t(x_t)\|_{\mathrm{F}}^2 p_{X_t}(x_t)\mathrm{d}x_t + 64c_1^2C_4\frac{d\log^3 T}{T} + T^{-2} \\
&\overset{\text{(b)}}{\leq} 8c_1C_4C_0\frac{d\log^2 T}{T} + 64c_1^2C_4\frac{d\log^3 T}{T} + T^{-3} \leq C_5\frac{d\log^3 T}{T}.
\end{aligned}$$

Here step (a) utilizes Lemma 18; step (b) follows from Lemma 4; while step (c) holds provided that $C_5 \gg c_1^2C_4C_0$. This further implies that

$$\mathsf{TV}(p_{X_1}, p_{\overline{Y}_1}) = \int_{p_{X_1}(x) > p_{\overline{Y}_1}(x)} \big(p_{X_1}(x) - p_{\overline{Y}_1}(x)\big)\mathrm{d}x = \int \Delta_1(x)\mathrm{d}x \leq C_5\frac{d\log^3 T}{T}. \tag{20}$$

4.4 Step 3: controlling estimation error

In this section, we will bound the total variation distance between p_{Y_1} and $p_{\overline{Y}_1}$. Note that

$$\begin{aligned}
\mathsf{TV}\big(p_{Y_1}, p_{\overline{Y}_1}\big) &= \int_{\mathbb{R}^d} \big(p_{\overline{Y}_1}(x) - p_{Y_1}(x)\big)\,\mathbb{1}\big\{p_{\overline{Y}_1}(x) > p_{Y_1}(x)\big\}\mathrm{d}x + \mathbb{P}\big(\overline{Y}_1 = \infty\big) \\
&\overset{\text{(i)}}{\leq} \int_{\mathbb{R}^d} \big(p_{\overline{Y}_1}(x) - p_{\widehat{Y}_1}(x)\big)\,\mathbb{1}\big\{p_{\overline{Y}_1}(x) > p_{\widehat{Y}_1}(x)\big\}\mathrm{d}x + \mathbb{P}\big(\overline{Y}_1 = \infty\big) \\
&\overset{\text{(ii)}}{\leq} \mathsf{TV}\big(p_{\overline{Y}_1}, p_{\widehat{Y}_1}\big) + \mathsf{TV}\big(p_{X_1}, p_{\overline{Y}_1}\big) \overset{\text{(iii)}}{\leq} \sqrt{\mathsf{KL}\big(p_{\overline{Y}_1}\|p_{\widehat{Y}_1}\big)} + C_5\frac{d\log^3 T}{T}.
\end{aligned} \tag{21}$$

Here step (i) follows from (18); step (ii) follows from $\mathbb{P}(\overline{Y}_1 = \infty) \leq \mathsf{TV}(p_{X_1}, p_{\overline{Y}_1})$, which holds since X_1 does not take value at ∞; step (iii) utilizes Pinsker's inequality and (20). Hence it suffices to bound $\mathsf{KL}(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1})$, which can be decomposed into

$$\begin{aligned}
\mathsf{KL}\big(p_{\overline{Y}_1} \| p_{\widehat{Y}_1}\big) &\overset{\text{(a)}}{\leq} \mathsf{KL}\big(p_{\overline{Y}_1, \overline{Y}_1^-, \dots, \overline{Y}_T, \overline{Y}_T^-} \| p_{\widehat{Y}_1, \widehat{Y}_1^-, \dots, \widehat{Y}_T, \widehat{Y}_T^-}\big) \\
&\overset{\text{(b)}}{=} \mathsf{KL}\big(p_{\overline{Y}_T^-} \| p_{\widehat{Y}_T^-}\big) + \sum_{t=2}^{T} \mathop{\mathbb{E}}_{x_t \sim p_{\overline{Y}_t}} \big[\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \| p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big)\big] \\
&\qquad + \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t \sim p_{\overline{Y}_t^-}} \big[\mathsf{KL}\big(p_{\overline{Y}_t | \overline{Y}_t^- = x_t} \| p_{\widehat{Y}_t | \widehat{Y}_t^- = x_t}\big)\big] \\
&\overset{\text{(c)}}{=} \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t}} \big[\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \| p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big)\big].
\end{aligned} \tag{22}$$

Here step (a) follows from the data-processing inequality; step (b) uses the chain rule of KL divergence, where we use the fact that (14) and (17) are both Markov chains; step (c) follows from the facts that, by construction, $\overline{Y}_T^- = \widehat{Y}_T^-$, and for any $x \neq \infty$, the conditional distributions of $\widehat{Y}_t$ given $\widehat{Y}_t^- = x$ and $\overline{Y}_t$ given $\overline{Y}_t^- = x$ are identical. For any $x_t \in \mathcal{E}_{t,1}$, we have

$$\begin{aligned}
\mathsf{KL}\Big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\Big) &\overset{\text{(i)}}{=} \frac{1-\alpha_t}{2} \|s_t(x_t) - s_t^\star(x_t)\|_2^2 \\
&\overset{\text{(ii)}}{\leq} \frac{c_1 \log T}{2T} \|s_t(x_t) - s_t^\star(x_t)\|_2^2.
\end{aligned} \tag{23}$$

Here step (i) follows from the transition probability (13c) and (16b), which give

$$\begin{aligned}
\overline{Y}_{t-1}^- | \overline{Y}_t = x_t &\sim \mathcal{N}\left(\frac{x_t + (1-\alpha_t) s_t^\star(x_t)}{\sqrt{\alpha_t}}, \frac{1-\alpha_t}{\alpha_t} I_d\right) \quad \text{and} \\
\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t &\sim \mathcal{N}\left(\frac{x_t + (1-\alpha_t) s_t(x_t)}{\sqrt{\alpha_t}}, \frac{1-\alpha_t}{\alpha_t} I_d\right),
\end{aligned}$$

and the KL divergence between two Gaussian measures can be computed in closed-form; step (ii) utilizes Lemma 18. On the other hand, for any $x_t \in \mathcal{E}_{t,1}^{\mathrm{c}}$, we have

$$\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big) = 0. \tag{24}$$

Therefore we have

$$\mathsf{KL}\big(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1}\big) \overset{\text{(i)}}{\leq} \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{X_t}} \Big[\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big)\Big] \overset{\text{(ii)}}{\leq} \frac{c_1}{2} \varepsilon_{\mathsf{score}}^2 \log T. \tag{25}$$

Here step (i) follows from (22) and the relation $p_{\overline{Y}_t}(x) \leq p_{X_t}(x)$ for any $x \neq \infty$ (see (15)); while step (ii) follows from (23) and (24). Substitution of the bound (25) into (21) yields

$$\mathsf{TV}\big(p_{Y_1}, p_{\overline{Y}_1}\big) \leq \sqrt{\frac{c_1}{2} \log T} \varepsilon_{\mathsf{score}} + C_5 \frac{d \log^3 T}{T}. \tag{26}$$

Taking the two bounds (20) and (26) collectively, we achieve the desired result

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \mathsf{TV}(p_{X_1}, p_{\overline{Y}_1}) + \mathsf{TV}\big(p_{Y_1}, p_{\overline{Y}_1}\big) \le C\frac{d\log^3 T}{T} + C\varepsilon_{\mathsf{score}}\sqrt{\log T}$$

for some constant $C \gg \sqrt{c_1} + C_5$.

5. Proof of Theorem 3

This section provides the proof of Theorem 3. While the high-level analysis idea is similar to the proof of Theorem 1, we need to carry out more careful analysis in order to precisely capture the low-dimensional structure. The constants $C_1, C_2, \ldots$ in this section are different from the ones in Section 4.

5.1 Preliminaries

For simplicity of presentation, we assume without loss of generality that $k \ge \log d$. In fact, if $k < \log d$, we can simply redefine $k \coloneqq \log d$, which does not change the desired bound (9). Let $\{x_i^\star\}_{1\le i\le N_\varepsilon}$ be an ε-net of $\mathcal{X} = \mathsf{supp}(p_{\mathsf{data}})$, where ε is sufficiently small

$$\varepsilon \ll \sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}}\min\left\{1, \sqrt{\frac{k\log T}{d}}\right\}, \tag{27}$$

and let $\{\mathcal{B}_i\}_{1\le i\le N_\varepsilon}$ be a disjoint ε-cover for $\mathcal{X}$ such that $x_i^\star \in \mathcal{B}_i$. Let

$$\begin{aligned}
\mathcal{I} &\coloneqq \{1 \le i \le N_\varepsilon : \mathbb{P}(X_0 \in \mathcal{B}_i) \ge \exp(-\theta k\log T)\}, \\
\mathcal{G} &\coloneqq \big\{\omega \in \mathbb{R}^d : \|\omega\|_2 \le 2\sqrt{d} + \sqrt{\theta k\log T}, \quad \text{and} \\
&\qquad\qquad |(x_i^\star - x_j^\star)^\top\omega| \le \sqrt{\theta k\log T}\|x_i^\star - x_j^\star\|_2 \quad \text{for all} \quad 1 \le i, j \le N_\varepsilon\big\},
\end{aligned}$$

where $\theta > 0$ is some sufficiently large constant. Then $\cup_{i\in\mathcal{I}}\mathcal{B}_i$ and $\mathcal{G}$ can be interpreted as high probability sets for the random variable X_0 and a standard Gaussian variable in $\mathbb{R}^d$. For each $t = 1, \ldots T$, we define a typical set for each X_t as follows

$$\mathcal{E}_{t,1} \coloneqq \big\{\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\omega : x_0 \in \cup_{i\in\mathcal{I}}\mathcal{B}_i, \omega \in \mathcal{G}\big\}. \tag{28}$$

This means that for any $x_t \in \mathcal{E}_{t,1}$, there exists an index $i(x_t) \in \mathcal{I}$ and two points $x_0(x_t) \in \mathcal{B}_{i(x_t)}$ and $\omega \in \mathcal{G}$ such that

$$x_t = \sqrt{\overline{\alpha}_t}x_0(x_t) + \sqrt{1-\overline{\alpha}_t}\omega. \tag{29}$$

It is worth mentioning that such $i(x_t)$, $x_0(x_t)$ and ω might not be unique, and we only need to arbitrarily choose and fix one of them. For any $x_t \in \mathcal{E}_{t,1}$ and any $r > 0$, define

$$\mathcal{I}\left(x_t; r\right) \coloneqq \left\{1 \le i \le N_\varepsilon : \overline{\alpha}_t\|x_i^\star - x_{i(x_t)}^\star\|_2^2 \le r\cdot k(1-\overline{\alpha}_t)\log T\right\}. \tag{30}$$

The following technical lemma will be crucial in the analysis.

Lemma 7 *There exists some universal constant $C_1 \gg \theta$ such that*

$$\mathbb{P}\left(X_0 \in \mathcal{B}_i \,|\, X_t = x_t\right) \le \exp\left(-\frac{\overline{\alpha}_t}{16\left(1-\overline{\alpha}_t\right)}\|x^\star_{i(x_t)} - x^\star_i\|_2^2\right)\mathbb{P}\left(X_0 \in \mathcal{B}_i\right)$$

for any $x_t \in \mathcal{E}_{t,1}$ and $i \notin \mathcal{I}(x_t; C_1\theta)$.

Proof The proof can be found in (Li and Yan, 2024, Appendix A.2) and is omitted here for brevity. ■

5.2 Main proof

We first define an auxiliary reverse process that uses the true score function:

$$Y_T^\star \sim \mathcal{N}(0, I_d), \qquad Y_{t-1}^\star = \frac{1}{\sqrt{\alpha_t}}\left(Y_t^\star + \eta_t^\star s_t^\star(Y_t^\star) + \sigma_t^\star Z_t\right) \quad \text{for } t = T, \dots, 1. \tag{31}$$

We introduce auxiliary sequences $\{\overline{Y}_t : t = T, \dots, 1\}$ and $\{\overline{Y}_t^- : t = T, \dots, 1\}$ as in (13), as well as $\{\widehat{Y}_t : t = T, \dots, 1\}$ and $\{\widehat{Y}_t^- : t = T, \dots, 1\}$ as in (16). It is worth mentioning that here we use $\mathcal{E}_{t,1}$ in (28) as well as the sequence $\{Y_t^\star : t = T, \dots, 1\}$ in (31) when defining these auxiliary sequences. In addition, we define $\Delta_t(x) = p_{X_t}(x) - p_{\overline{Y}_t}(x)$ as in (19).

The following lemma establishes bounds similar to Lemma 4. In order to avoid incurring polynomial dependency in d, it is important to focus on $I - J_t(x_t)$ instead of $J_t(x)$ itself.

Lemma 8 *There exists some universal constant $C_2 \gg C_1$ such that for any $x_t \in \mathcal{E}_{t,1}$,*

$$\|I - J_t(x_t)\| \le \|I - J_t(x_t)\|_{\mathrm{F}} \le |\mathsf{Tr}(I - J_t(x_t))| \le C_2\theta k \log T, \tag{32}$$

where $J_t(\cdot)$ is defined in (10). In addition, there exists universal constant $C_0 > 0$ such that

$$\sum_{t=2}^{T} \frac{1-\alpha_t}{1-\overline{\alpha}_t}\int_{x_t} \|I - J_t(x_t)\|_{\mathrm{F}}^2\, p_{X_t}(x_t)\mathrm{d}x_t \le C_0 k \log T. \tag{33}$$

Proof See Appendix B.1. ■

It is worth mentioning that unlike $I - J_t(x_t)$, the order of $s_t^\star(x_t)$ scales linearly with $\sqrt{d}$ even for $x_t \in \mathcal{E}_{t,1}$ as in Lemma 4. Therefore the key difficulty of this proof is to avoid introducing any error term that scales with $\|s_t^\star(x_t)\|_2$. Next, we establish the following lemma in analogy to Lemma 6.

Lemma 9 *There exists some universal constant $C_3 > 0$ such that, for $t = T, \dots, 2$,*

$$\int \Delta_{t-1}(x)\mathrm{d}x \le \int \Delta_t(x)\mathrm{d}x + C_4\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\int_{x_t \in \mathcal{E}_{t,1}}(|\mathsf{Tr}(I - J_t(x_t))| + \|I - J_t(x_t)\|_{\mathsf{F}}^2)p_{X_t}(x_t)\mathrm{d}x_t$$
$$+ T^{-3}.$$

In addition, we have $\int \Delta_T(x)\mathrm{d}x \le T^{-4}$.

Proof See Appendix B.2. ■

We can apply Lemma 6 recursively to achieve

$$\begin{aligned}
\int \Delta_1(x)\mathrm{d}x &\le \int \Delta_T(x)\mathrm{d}x + T^{-2} \\
&\quad + \sum_{t=2}^{T} C_4\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t\in\mathcal{E}_{t,1}} \big(|\mathsf{Tr}(I - J_t(x_t))| + \|I - J_t(x_t)\|_{\mathsf{F}}^2\big) p_{X_t}(x_t)\mathrm{d}x_t \\
&\overset{\text{(a)}}{\le} 8c_1C_4\frac{\log T}{T}\sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\int_{x_t\in\mathcal{E}_{t,1}} \|I - J_t(x_t)\|_{\mathsf{F}}^2 p_{X_t}(x_t)\mathrm{d}x_t + 64c_1^2C_4\frac{\theta k\log^3 T}{T} + T^{-2} \\
&\overset{\text{(b)}}{\le} 8c_1C_4C_0\frac{k\log^2 T}{T} + 64c_1^2C_4\frac{\theta k\log^3 T}{T} + T^{-3} \le C_5\frac{k\log^3 T}{T}.
\end{aligned}$$

Here step (a) utilizes Lemma 18; step (b) follows from Lemma 4; while step (c) holds provided that $C_5 \gg c_1^2C_4C_0\theta$. This further implies that

$$\mathsf{TV}(p_{X_1}, p_{\overline{Y}_1}) = \int_{p_{X_1}(x)>p_{\overline{Y}_1}(x)} \big(p_{X_1}(x) - p_{\overline{Y}_1}(x)\big)\mathrm{d}x = \int \Delta_1(x)\mathrm{d}x \le C_5\frac{k\log^3 T}{T}. \tag{34}$$

Equipped with (34), we can follow the same steps in Section 4.4 to control the estimation error, which gives

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \le C\frac{k\log^3 T}{T} + C\varepsilon_{\mathsf{score}}\sqrt{\log T}$$

as claimed, provided that $C \gg \sqrt{c_1} + C_5$.

6. Simulation study

We conducted a synthetic experiment to compare our adaptive schedule (3) against the constant step size used in prior work (e.g., Chen et al. (2023d)). We fix $\overline{\alpha}_1 = 0.99$ and $\overline{\alpha}_T = 0.005$ so that for any choice of T the learning rates α_t and β_t are uniquely determined for both schedules. We evaluate four different values of $T \in \{20, 50, 80, 100\}$.

Let the data distribution be a one-dimensional symmetric Gaussian mixture distribution

$$p_{\mathsf{data}} = \frac{1}{2}\mathcal{N}(1, \sigma^2) + \frac{1}{2}\mathcal{N}(-1, \sigma^2), \tag{35}$$

where $\sigma^2 \in [0, 1]$ is the parameter that controls the Lipschitz smoothness of the score function. In particular, $\sigma = 0$ yields a discrete two-point mass with a non-Lipschitz score, while larger σ^2 produces increasingly smooth scores.

Figure 1 plots the TV error as a funtion of σ^2 for each choice of T. As we can see, when the score is Lipschitz (i.e., σ^2 not too small), both schedules perform comparably. However, as $\sigma^2 \to 0$ (violating the Lipschitz score assumption), our adaptive schedule achieves substantially lower error. This behavior matches theory: prior works using the constant step size schedules (e.g., Chen et al. (2023d)) require a Lipschitz score assumption, whereas the theory in other works that use adaptive step size schedules (e.g., this work

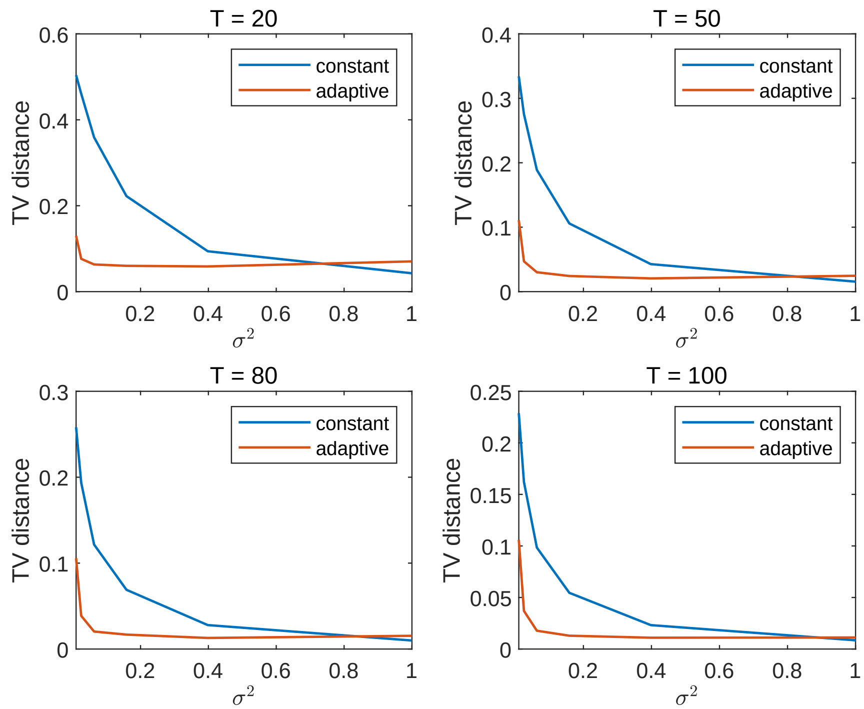

Figure 1: The TV distance between p_{X_1} and p_{Y_1} vs. the smoothness parameter σ^2 for DDPM with constant step size and the adaptive schedule (3), for $T \in \{20, 50, 80, 100\}$. The data distribution is $0.5\mathcal{N}(1, \sigma^2) + 0.5\mathcal{N}(-1, \sigma^2)$.

and Benton et al. (2023a)) do not. This suggests that adaptive learning rates might offer greater robustness to non-ideal score smoothness, accommodating general data distributions in practice.

We emphasize that this experiment is conducted on a simple one-dimensional Gaussian mixture. While it serves to illustrate the theoretical distinctions between constant and adaptive schedules, the conclusions may not fully reflect the practical behavior of the DDPM sampler on more complex, high-dimensional target distributions.

7. Discussion

In this paper, we establish an $O(d/T)$ convergence theory for the DDPM sampler, assuming access to ℓ_2-accurate score estimates. This significantly improves upon the state-of-the-art convergence rate of $O(\sqrt{d/T})$ in Benton et al. (2023a). Compared to the recent work of Li et al. (2024b), which also achieves an $O(d/T)$ rate for another DDIM sampler, our approach relaxes stringent score estimation requirements, such as the need for the Jacobian of the score estimates to closely match that of the true score functions. Furthermore, to account for low-dimensional structures in the target data distribution, we extend our theory to achieve an $O(k/T)$ convergence bound under careful coefficient design, where k is the intrinsic dimension. This improves upon the prior convergence rate of $O(\sqrt{k/T})$ established in Potaptchik et al. (2024); Huang et al. (2024).

This work opens several promising directions for future research. For example, it remains unclear whether the $O(d/T)$ convergence rate is tight for the DDPM sampler; it would be of interest to develop lower bounds on certain hard instances. Another intriguing direction is to explore whether the analysis in this paper can extend to developing convergence theory under KL divergence (e.g., Benton et al. (2023a)) or Wasserstein distance (e.g., Gao and Zhu (2025); Benton et al. (2023b)). A recent work (Jain and Zhang, 2025) achieved a sharp convergence theory for the DDPM sampler under KL divergence. Finally, while this paper focuses on analyzing the discretization error of the DDPM sampler and treats the score matching stage as a black box, it would be worthwhile to design score matching algorithms that adapt to unknown low-dimensional structures in the target data distribution.

Acknowledgments

Gen Li is supported in part by the Chinese University of Hong Kong Direct Grant for Research and the Hong Kong Research Grants Council ECS 2191363.

Appendix A. Proof of auxiliary lemmas in Section 4

A.1 Proof of Lemma 4

For any pairs $(x, x_0) \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying

$$\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2 \geq (6\theta + 3c_0) d (1 - \overline{\alpha}_t) \log T =: R^2 \tag{36}$$

where c_0 is defined in (3), we have

$$\begin{aligned} p_{X_0 | X_t}(x_0 \,|\, x) &= \frac{p_{X_0}(x_0)}{p_{X_t}(x)} p_{X_t | X_0}(x \,|\, x_0) \\ &\overset{\text{(i)}}{=} p_{X_0}(x_0) \cdot \big(2\pi(1 - \overline{\alpha}_t)\big)^{-d/2} \exp\Big(- \frac{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{2(1 - \overline{\alpha}_t)} - \log p_{X_t}(x) \Big) \\ &\overset{\text{(ii)}}{\leq} p_{X_0}(x_0) \exp\Big(- \frac{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{3(1 - \overline{\alpha}_t)} \Big). \end{aligned} \tag{37}$$

Here step (i) uses the fact that $X_t \,|\, X_0 = x_0 \sim \mathcal{N}(\sqrt{\overline{\alpha}_t} x_0, (1 - \overline{\alpha}_t) I_d)$, while step (ii) holds since

$$\begin{aligned} -\frac{d}{2} \log 2\pi(1 - \overline{\alpha}_t) - \frac{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{2(1 - \overline{\alpha}_t)} - \log p_{X_t}(x) &\overset{\text{(iii)}}{\leq} \frac{c_0}{2} d \log T - \frac{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{2(1 - \overline{\alpha}_t)} + \theta d \log T \\ &\overset{\text{(iv)}}{\leq} - \frac{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{3(1 - \overline{\alpha}_t)}, \end{aligned}$$

where step (iii) follows from the fact that $1 - \overline{\alpha}_t \geq 1 - \alpha_1 = \beta_1$ for any $1 \leq t \leq T$, and $-\log p_{X_t}(x) \leq \theta d \log T$; step (iv) follows from (36). Recall that

$$s_t^{\star}(x) = -\frac{1}{1 - \overline{\alpha}_t} \int_{x_0} p_{X_0 | X_t}(x_0 \,|\, x) \big(x - \sqrt{\overline{\alpha}_t} x_0\big) \mathrm{d} x_0 \tag{38}$$

and

$$\begin{aligned} \mathsf{Tr}\,(I - J_t(x)) = \frac{1}{1 - \overline{\alpha}_t} \Big(&\int_{x_0} p_{X_0 | X_t}(x_0 \,|\, x) \|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2 \mathrm{d} x_0 \\ &- \Big\| \int_{x_0} p_{X_0 | X_t}(x_0 \,|\, x) \big(x - \sqrt{\overline{\alpha}_t} x_0\big) \mathrm{d} x_0 \Big\|_2^2 \Big). \end{aligned} \tag{39}$$

Then we have

$$\begin{aligned} \|s_t^{\star}(x)\|_2 &\overset{\text{(a)}}{\leq} \frac{1}{1 - \overline{\alpha}_t} \int_{x_0} p_{X_0 | X_t}(x_0 \,|\, x) \|x - \sqrt{\overline{\alpha}_t} x_0\|_2 \mathrm{d} x_0 \\ &= \frac{1}{1 - \overline{\alpha}_t} \int p_{X_0 | X_t}(x_0 \,|\, x) \|x - \sqrt{\overline{\alpha}_t} x_0\|_2 \big(\mathbb{1}_{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2 \leq R} + \mathbb{1}_{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2 > R} \big) \mathrm{d} x_0 \\ &\overset{\text{(b)}}{\leq} \frac{R}{1 - \overline{\alpha}_t} + \frac{1}{1 - \overline{\alpha}_t} \int p_{X_0}(x_0) \exp\Big(- \frac{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{3(1 - \overline{\alpha}_t)} \Big) \|x - \sqrt{\overline{\alpha}_t} x_0\|_2 \, \mathbb{1}_{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2 > R} \, \mathrm{d} x_0 \\ &\overset{\text{(c)}}{\leq} \frac{R}{1 - \overline{\alpha}_t} + \sqrt{\frac{3}{1 - \overline{\alpha}_t}} \int p_{X_0}(x_0) \exp\Big(- \frac{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{6(1 - \overline{\alpha}_t)} \Big) \mathbb{1}_{\|x - \sqrt{\overline{\alpha}_t} x_0\|_2 > R} \, \mathrm{d} x_0 \\ &\leq \frac{R}{1 - \overline{\alpha}_t} + \sqrt{\frac{3}{1 - \overline{\alpha}_t}} \exp\Big(- \frac{R^2}{6(1 - \overline{\alpha}_t)} \Big) \overset{\text{(d)}}{\leq} \frac{2R}{1 - \overline{\alpha}_t}. \end{aligned} \tag{40}$$

Here step (a) utilizes Jensen's inequality; step (b) follows from (37); step (c) follows from the fact that $z\exp(-z^2) \le \exp\big(-z^2/2\big)$ holds for any $z \ge 0$; whereas step (d) holds provided that c_0 is sufficiently large. In addition, we have

$$\begin{aligned}\mathsf{Tr}(I - J_t(x)) &\le \frac{1}{1-\overline{\alpha}_t}\mathbb{E}\left[\|X_t - \sqrt{\overline{\alpha}_t}X_0\|_2^2 \,|\, X_t = x\right] \\ &= \frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0|X_t}(x_0\,|\,x)\|x - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \mathrm{d}x_0.\end{aligned}$$

Then we can use the analysis similar to (40) to show that

$$\begin{aligned}\mathsf{Tr}(I - J_t(x)) &\overset{\text{(i)}}{\le} \frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0|X_t}(x_0\,|\,x)\|x - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \mathrm{d}x_0 \\ &\le \frac{1}{1-\overline{\alpha}_t}\int p_{X_0|X_t}(x_0\,|\,x)\|x - \sqrt{\overline{\alpha}_t}x_0\|_2^2\, \mathbb{1}\left\{\|x - \sqrt{\overline{\alpha}_t}x_0\|_2 \le R\right\}\mathrm{d}x_0 \\ &\quad + \frac{1}{1-\overline{\alpha}_t}\int p_{X_0|X_t}(x_0\,|\,x)\|x - \sqrt{\overline{\alpha}_t}x_0\|_2^2\, \mathbb{1}\left\{\|x - \sqrt{\overline{\alpha}_t}x_0\|_2 > R\right\}\mathrm{d}x_0 \\ &\overset{\text{(ii)}}{\le} \frac{R^2}{1-\overline{\alpha}_t} + \frac{1}{1-\overline{\alpha}_t}\int p_{X_0}(x_0)\exp\Big(-\frac{\|x - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{3(1-\overline{\alpha}_t)}\Big)\|x - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \\ &\qquad\qquad \cdot \mathbb{1}\left\{\|x - \sqrt{\overline{\alpha}_t}x_0\|_2 > R\right\}\mathrm{d}x_0 \\ &\overset{\text{(iii)}}{\le} \frac{R^2}{1-\overline{\alpha}_t} + 3\int p_{X_0}(x_0)\exp\Big(-\frac{\|x - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{6(1-\overline{\alpha}_t)}\Big)\mathbb{1}\left\{\|x - \sqrt{\overline{\alpha}_t}x_0\|_2 > R\right\}\mathrm{d}x_0 \\ &\le \frac{R^2}{1-\overline{\alpha}_t} + 3\exp\Big(-\frac{R^2}{6(1-\overline{\alpha}_t)}\Big) \overset{\text{(iv)}}{\le} \frac{2R^2}{1-\overline{\alpha}_t}. \end{aligned} \tag{41}$$

Here step (i) follows from ((39)); step (ii) follows from (37); step (iii) follows from the fact that $x\exp(-x) \le \exp(-x/2)$ holds for any $z \ge 0$; while step (iv) holds provided that c_0 is sufficiently large.

Finally, we invoke Lemma 21 to achieve

$$\sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathsf{Tr}\big(\mathbb{E}\big[\big(\Sigma_{\overline{\alpha}_t}(X_t)\big)^2\big]\big) \le C_J d\log T, \tag{42}$$

where the matrix function $\Sigma_{\overline{\alpha}_t}(\cdot)$ is defined in Lemma 21 as

$$\Sigma_{\overline{\alpha}_t}(x) \coloneqq \mathsf{Cov}\big(Z\,|\,\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z = x\big)$$

for an independent $Z \sim \mathcal{N}(0, I_d)$. It is straightforward to check that $J_t(x) = I_d - \Sigma_{\overline{\alpha}_t}(x)$, therefore we have

$$\begin{aligned}\sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathsf{Tr}\big(\mathbb{E}\big[\big(\Sigma_{\overline{\alpha}_t}(X_t)\big)^2\big]\big) &= \sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathbb{E}\big[\mathsf{Tr}\big((I_d - J_t(X_t))^2\big)\big] \\ &= \sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathbb{E}\big[\|I_d - J_t(X_t)\|_{\mathrm{F}}^2\big]. \end{aligned} \tag{43}$$

Here the last relation holds since $\mathsf{Tr}(A^2) = \|A\|_{\mathrm{F}}^2$ for any symmetric matrix A. We conclude that

$$\begin{aligned}\sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\int_{x_t}\|J_t(x_t)\|_{\mathrm{F}}^2 p_{X_t}(x_t)\mathrm{d}x_t &= \sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathbb{E}\big[\|J_t(X_t)\|_{\mathrm{F}}^2\big] \\ &\overset{\text{(a)}}{\leq} \sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathbb{E}\big[2\|I_d - J_t(X_t)\|_{\mathrm{F}}^2 + 2\|I_d\|_{\mathrm{F}}^2\big] \\ &\overset{\text{(b)}}{\leq} 2C_J d\log T + 16c_1 d\log T \overset{\text{(c)}}{\leq} C_0 d\log T.\end{aligned}$$

Here step (a) utilizes the triangle inequality and the AM-GM inequality; step (b) follows from (42), (43) and Lemma 18; while step (c) holds provided that $C_0 \gg C_J + c_1$.

A.2 Proof of Lemma 6

We first observe that

$$\begin{aligned}p_{\overline{Y}_{t-1}^-}(x_{t-1}) &\geq \int_{\mathbb{R}^d} p_{\overline{Y}_{t-1}^-|\overline{Y}_t}(x_{t-1}\,|\,x_t)p_{\overline{Y}_t}(x_t)\mathrm{d}x_t \overset{\text{(i)}}{\geq} \int_{x_t\in\mathcal{E}_{t,1}} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1}\,|\,x_t)p_{\overline{Y}_t}(x_t)\mathrm{d}x_t \\ &\overset{\text{(ii)}}{=} \int_{x_t\in\mathcal{E}_{t,1}} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1}\,|\,x_t)p_{X_t}(x_t)\mathrm{d}x_t - \Delta_{t\to t-1}(x_{t-1}) \qquad (44)\end{aligned}$$

where we define

$$\Delta_{t\to t-1}(x_{t-1}) \coloneqq \int_{x_t\in\mathcal{E}_{t,1}} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1}\,|\,x_t)\Delta_t(x_t)\mathrm{d}x_t \geq 0.$$

Here step (i) follows from (13c), while step (ii) makes use of the definition (19). It is straightforward to check that

$$\int \Delta_{t\to t-1}(x)\mathrm{d}x = \int_{x_{t-1}}\int_{x_t\in\mathcal{E}_{t,1}} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1}\,|\,x_t)\Delta_t(x_t)\mathrm{d}x_t\mathrm{d}x_{t-1} \leq \int \Delta_t(x)\mathrm{d}x. \qquad (45)$$

For any x_{t-1} such that $\Delta_{t-1}(x_{t-1}) > 0$, we have

$$\begin{aligned}&p_{X_{t-1}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) \\ &\overset{\text{(a)}}{=} p_{\overline{Y}_{t-1}^-}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) \overset{\text{(b)}}{\geq} \int_{x_t\in\mathcal{E}_{t,1}} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1}\,|\,x_t)p_{X_t}(x_t)\mathrm{d}x_t \qquad (46) \\ &\overset{\text{(c)}}{=} \int_{x_t\in\mathcal{E}_{t,1}} p_{X_t}(x_t)\Big(\frac{\alpha_t}{2\pi(1-\alpha_t)}\Big)^{d/2}\exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1} - \big(x_t + (1-\alpha_t)s_t^\star(x_t)\big)\big\|^2}{2(1-\alpha_t)}\Big)\mathrm{d}x_t \\ &\overset{\text{(d)}}{=} \int_{x_t\in\mathcal{E}_{t,1}} \mathsf{det}\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1} p_{X_t}(x_t)\Big(\frac{\alpha_t}{2\pi(1-\alpha_t)}\Big)^{d/2}\exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1} - u_t\big\|^2}{2(1-\alpha_t)}\Big)\mathrm{d}u_t.\end{aligned}$$

Here step (a) utilizes the definition (19) and $p_{\overline{Y}_{t-1}}(x_{t-1}) = p_{\overline{Y}_{t-1}^-}(x_{t-1})$, which is a consequence of (15) and $\Delta_{t-1}(x_{t-1}) > 0$; step (b) follows from (44); step (c) follows from the definition (12); whereas step (d) applies the change of variable $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$. Moving forward, we need the following lemma.

Lemma 10 *For any $x_t \in \mathcal{E}_{t,1}$, we have*

$$\begin{aligned}
&\mathsf{det}\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1} p_{X_t}(x_t) \\
&\quad = \big(2\pi(2\alpha_t - 1 - \overline{\alpha}_t)\big)^{-d/2} \int_{x_0} p_{X_0}(x_0) \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big) \mathrm{d}x_0 \\
&\qquad \cdot \exp\Big(\xi_t(x_t) + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \big(d\log T + \|J_t(x_t)\|_\mathsf{F}^2\big)\Big)\Big),
\end{aligned} \tag{47}$$

where $\xi_t(x_t) \leq 0$ satisfies

$$\begin{aligned}
\int_{x_t \in \mathcal{E}_{t,1}} |\xi_t(x_t)| p_{X_t}(x_t) \mathrm{d}x_t \leq C_3 \Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t \in \mathcal{E}_{t,1}} \big(d\log T + \|J_t(x_t)\|_\mathsf{F}^2\big) p_{X_t}(x_t) \mathrm{d}x_t \\
+ T^{-4}
\end{aligned} \tag{48}$$

for some universal constant $C_3 > 0$.

Proof See Appendix A.3. ■

Taking the decomposition (47) and (46) collectively, we have

$$\begin{aligned}
&p_{X_{t-1}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) + \delta_{t-1}(x_{t-1}) \\
&\quad \geq \int_{x_0}\int_{x_t} \exp\Big(\Big[\xi_t(x_t) + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \big(d\log T + \|J_t(x_t)\|_\mathsf{F}^2\big)\Big)\Big] \mathbb{1}\left\{x_t \in \mathcal{E}_{t,1}\right\}\Big) \\
&\qquad \cdot p_{X_0}(x_0) \Big(\frac{\alpha_t}{4\pi^2(1-\alpha_t)(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)^{d/2} \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big) \\
&\qquad \cdot \exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1} - u_t\big\|^2}{2(1-\alpha_t)}\Big) \mathrm{d}u_t \mathrm{d}x_0,
\end{aligned} \tag{49}$$

where we define

$$\begin{aligned}
\delta_{t-1}(x_{t-1}) &:= \int_{x_0}\int_{x_t \notin \mathcal{E}_{t,1}} p_{X_0}(x_0) \Big(\frac{\alpha_t}{4\pi^2(1-\alpha_t)(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)^{d/2} \\
&\qquad \cdot \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big) \exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1} - u_t\big\|^2}{2(1-\alpha_t)}\Big) \mathrm{d}u_t \mathrm{d}x_0.
\end{aligned} \tag{50}$$

Moreover, it is straightforward to check that

$$\begin{aligned}
&\int_{x_0}\int_{x_t} p_{X_0}(x_0) \Big(\frac{\alpha_t}{4\pi^2(1-\alpha_t)(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)^{d/2} \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big) \\
&\qquad \cdot \exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1} - u_t\big\|^2}{2(1-\alpha_t)}\Big) \mathrm{d}u_t \mathrm{d}x_0 = p_{X_{t-1}}(x_{t-1}).
\end{aligned} \tag{51}$$

Then we can continue the derivation in (49):

$$
\begin{aligned}
& p_{X_{t-1}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) + \delta_{t-1}(x_{t-1}) \\
& \overset{\text{(i)}}{\geq} \int_{x_0}\int_{x_t} \Big(1 + \Big[\xi_t(x_t) + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big]\,\mathbb{1}\left\{x_t \in \mathcal{E}_{t,1}\right\}\Big)p_{X_0}(x_0) \\
& \quad \cdot \Big(\frac{\alpha_t}{4\pi^2(1-\alpha_t)(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)^{d/2} \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} - \frac{\big\|\sqrt{\alpha_t}x_{t-1} - u_t\big\|^2}{2(1-\alpha_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0 \\
& \overset{\text{(ii)}}{=} p_{X_{t-1}}(x_{t-1}) + \int_{x_0}\int_{x_t\in\mathcal{E}_{t,1}} \Big[\xi_t(x_t) + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big]p_{X_0}(x_0) \\
& \quad \cdot \Big(\frac{\alpha_t}{4\pi^2(1-\alpha_t)(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)^{d/2} \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} - \frac{\big\|\sqrt{\alpha_t}x_{t-1} - u_t\big\|^2}{2(1-\alpha_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0.
\end{aligned}
$$

Here step (i) follows from the fact that $e^x \geq 1 + x$ for all $x \in \mathbb{R}$, while step (ii) follows from (51). By rearranging terms and integrate over the variable x_{t-1}, we arrive at

$$
\begin{aligned}
\int_{x_{t-1}} \Delta_{t-1}(x_{t-1})\mathrm{d}x_{t-1} \leq & \int_{x_{t-1}} \big(\Delta_t(x_{t-1}) + \delta_{t-1}(x_{t-1})\big)\mathrm{d}x_{t-1} \\
& + \int_{x_0}\int_{x_t\in\mathcal{E}_{t,1}} \Big(|\xi_t(x_t)| + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big)p_{X_0}(x_0) \\
& \qquad \cdot \big(2\pi(2\alpha_t - 1 - \overline{\alpha}_t)\big)^{-d/2} \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0,
\end{aligned} \tag{52}
$$

where we used (45) and for any fixed u_t, the following is a density function of x_{t-1}:

$$
\Big(2\pi\frac{1-\alpha_t}{\alpha_t}\Big)^{-d/2} \exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1} - u_t\big\|_2^2}{2(1-\alpha_t)}\Big)
$$

To establish the desired result, we need the following two lemmas.

Lemma 11 *For $x_t \in \mathcal{E}_{t,1}$, we have*

$$
\begin{aligned}
\int_{x_0} p_{X_0}(x_0)\big(2\pi(2\alpha_t - 1 - \overline{\alpha}_t)\big)^{-d/2} \exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)\mathrm{d}x_0 \\
\leq 20\det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1}p_{X_t}(x_t).
\end{aligned}
$$

Proof See Appendix A.4. ■

Lemma 12 *For the function $\delta_{t-1}(\cdot)$ defined in (50), we have $\int_{x_{t-1}} \delta_{t-1}(x_{t-1})\mathrm{d}x_{t-1} \leq T^{-4}$.*

Proof See Appendix A.5. ■

Equipped with these two lemmas, we can continue the derivation in (52) as follows:

$$\begin{aligned}
\int_{x_{t-1}} \Delta_{t-1}(x_{t-1})\mathrm{d}x_{t-1} &\overset{\text{(a)}}{\leq} 20\int_{x_t\in\mathcal{E}_{t,1}} \Big(|\xi_t(x_t)| + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big) \\
&\qquad\qquad \cdot \det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1} p_{X_t}(x_t)\mathrm{d}u_t + \int_{x_t}\Delta_t(x_t)\mathrm{d}x_t + T^{-4} \\
&\overset{\text{(b)}}{=} \int_{x_t}\Delta_t(x_t)\mathrm{d}x_t + T^{-4} \\
&\qquad + 20\int_{x_t\in\mathcal{E}_{t,1}} \Big(|\xi_t(x_t)| + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big) p_{X_t}(x_t)\mathrm{d}x_t \\
&\overset{\text{(c)}}{\leq} \int_{x_t}\Delta_t(x_t)\mathrm{d}x_t + T^{-3} + C_4\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\int_{x_t\in\mathcal{E}_{t,1}}\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)p_{X_t}(x_t)\mathrm{d}x_t,
\end{aligned}$$

which establishes the desired recursive relation. Here step (a) follows from Lemmas 11 and 12; step (b) follows from $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$, hence

$$\mathrm{d}u_t = \mathsf{det}\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\mathrm{d}x_t;$$

whereas step (c) uses (48) in Lemma 10, and holds provided that $C_4 \gg C_3$.

Finally, we control the error $\int \Delta_T(x)\mathrm{d}x$ in the initial step of the reverse process. Notice that

$$\begin{aligned}
\int \Delta_T(x)\mathrm{d}x = \int_{x_T\neq\infty}\big(p_{X_T}(x_T) - p_{\overline{Y}_T^-}(x_T)\big)\mathrm{d}x_T &\overset{\text{(i)}}{=} \mathsf{TV}\big(p_{X_T}, p_{\overline{Y}_T^-}\big) \\
&\overset{\text{(ii)}}{\leq} \mathsf{TV}\big(p_{X_T}, p_{Y_T}\big) + \mathsf{TV}\big(p_{Y_T}, p_{\overline{Y}_T^-}\big), \qquad (53)
\end{aligned}$$

where step (i) follows from (15) and step (ii) utilizes the triangle inequality. The first term can be bounded by Lemma 20, so it boils down to bounding the second. By definition of $\overline{Y}_T^-$ in (13a), we have

$$\begin{aligned}
\mathsf{TV}\big(p_{Y_T}, p_{\overline{Y}_T^-}\big) &= \int_{y\in\mathcal{E}_{T,1}^{\mathrm{c}}} p_{Y_T}(y)\mathrm{d}y \qquad (54)\\
&\overset{\text{(a)}}{=} \int p_{Y_T}(y)\,\mathbb{1}\Big\{-\log p_{X_T}(y) > C_1 d\log T, \|y\|_2 \leq \sqrt{\overline{\alpha}_T}T^{2c_R} + C_2\sqrt{d(1-\overline{\alpha}_T)\log T}\Big\}\mathrm{d}y \\
&\qquad + \int p_{Y_T}(y)\,\mathbb{1}\Big\{\|y\|_2 > \sqrt{\overline{\alpha}_T}T^{2c_R} + C_2\sqrt{d(1-\overline{\alpha}_T)\log T}\Big\}\mathrm{d}y \\
&\overset{\text{(b)}}{\leq} \int p_{X_T}(y)\,\mathbb{1}\Big\{-\log p_{X_T}(y) > C_1 d\log T, \|y\|_2 \leq \sqrt{\overline{\alpha}_T}T^{2c_R} + C_2\sqrt{d(1-\overline{\alpha}_T)\log T}\Big\}\mathrm{d}y \\
&\qquad + \mathsf{TV}\big(p_{X_T}, p_{Y_T}\big) + \mathbb{P}\Big(\|Y_T\|_2 > \sqrt{\overline{\alpha}_T}T^{2c_R} + C_2\sqrt{d(1-\overline{\alpha}_T)\log T}\Big) \\
&\overset{\text{(c)}}{\leq} \Big[2\sqrt{\overline{\alpha}_T}T^{2c_R} + 2C_2\sqrt{d(1-\overline{\alpha}_T)\log T}\Big]^d\exp(-C_1 d\log T) + \mathbb{P}\Big(\|Y_T\|_2 > \frac{C_2}{2}\sqrt{d\log T}\Big) \\
&\qquad + \mathsf{TV}\big(p_{X_T}, p_{Y_T}\big) \overset{\text{(d)}}{\leq} \exp\Big(-\frac{C_1}{2}d\log T\Big) + \mathbb{P}\Big(\|Y_T\|_2 > \frac{C_2}{2}\sqrt{d\log T}\Big) + \mathsf{TV}\big(p_{X_T}, p_{Y_T}\big).
\end{aligned}$$

Here step (a) follows from the definition of $\mathcal{E}_{T,1}$ in (11); step (b) follows from the definition of total variation distance, i.e., $\mathsf{TV}(p,q) = \sup_B |p(B) - q(B)|$, where the supremum is taken over all Borel set B in $\mathbb{R}^d$; step (c) holds since $\overline{\alpha}_T \leq T^{-c_1/2}$ (see Lemma 18), provided that C_2 is sufficiently large; whereas step (d) holds provided that $C_1 \gg c_R$ and $T \gg d\log T$. By putting (53) and (54) together, we have

$$\int \Delta_T(x)\mathrm{d}x \leq 2\mathsf{TV}\big(p_{X_T}, p_{Y_T}\big) + \exp\Big(-\frac{C_1}{2}d\log T\Big) + \mathbb{P}\Big(\|Y_T\|_2 > \frac{C_2}{2}\sqrt{d\log T}\Big) \leq T^{-4},$$

where the last relation follows from Lemmas 20 and 19, provided that $C_1, C_2 > 0$ are both sufficiently large.

A.3 Proof of Lemma 10

Consider any $x_t \in \mathcal{E}_{t,1}$. Recall the definition $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$, and we decompose

$$\begin{aligned}
\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} &= \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{(2\alpha_t - 1 - \overline{\alpha}_t)(1-\overline{\alpha}_t)} \\
&\quad + \frac{(1-\alpha_t)s_t^\star(x_t)^\top(x_t - \sqrt{\overline{\alpha}_t}x_0)}{2\alpha_t - 1 - \overline{\alpha}_t} + \frac{(1-\alpha_t)^2\|s_t^\star(x_t)\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} \\
&= \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)^2\|s_t^\star(x_t)\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} + \zeta_t(x_t, x_0) \\
&\quad + \frac{1-\alpha_t}{(2\alpha_t - 1 - \overline{\alpha}_t)(1-\overline{\alpha}_t)}\int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t)\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2\mathrm{d}x_0 \\
&\quad + \frac{1-\alpha_t}{2\alpha_t - 1 - \overline{\alpha}_t}s_t^\star(x_t)^\top\int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0,
\end{aligned}$$

where we let

$$\begin{aligned}
\zeta_t(x_t, x_0) &:= \frac{(1-\alpha_t)\big(\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 - \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t)\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2\mathrm{d}x_0\big)}{(2\alpha_t - 1 - \overline{\alpha}_t)(1-\overline{\alpha}_t)} \\
&\quad + \frac{(1-\alpha_t)s_t^\star(x_t)^\top\big[(x_t - \sqrt{\overline{\alpha}_t}x_0) - \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big]}{2\alpha_t - 1 - \overline{\alpha}_t}.
\end{aligned} \tag{55}$$

In view of (38) and (39), we can further derive

$$\begin{aligned}
\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} &= \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{1-\alpha_t}{2\alpha_t - 1 - \overline{\alpha}_t}\mathsf{Tr}\left(I - J_t(x_t)\right) + \frac{(1-\alpha_t)^2\|s_t^\star(x_t)\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} + \zeta_t(x_t, x_0) \\
&\overset{\text{(i)}}{=} \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \Big(1 + O\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)\Big)\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathsf{Tr}\left(I - J_t(x_t)\right) + \frac{(1-\alpha_t)^2\|s_t^\star(x_t)\|_2^2}{2(1-\overline{\alpha}_t)}\Big) + \zeta_t(x_t, x_0) \\
&\overset{\text{(ii)}}{=} \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathsf{Tr}\left(I - J_t(x_t)\right) + O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 d\log T\bigg) + \zeta_t(x_t, x_0) \\
&\overset{\text{(iii)}}{=} \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \log\det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big) - \frac{d}{2}\log\frac{2\alpha_t - 1 - \overline{\alpha}_t}{1-\overline{\alpha}_t} \\
&\quad + \zeta_t(x_t, x_0) + O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_\mathsf{F}^2\big)\bigg).
\end{aligned} \tag{56}$$

Here, step (i) utilizes an immediate consequence of Lemma 18

$$\frac{1-\overline{\alpha}_t}{2\alpha_t-1-\overline{\alpha}_t} = 1+\frac{2(1-\alpha_t)/(1-\overline{\alpha}_t)}{1-2(1-\alpha_t)/(1-\overline{\alpha}_t)} = 1+O\left(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\right) = 1+O\left(\frac{\log T}{T}\right), \tag{57}$$

which holds provided that $T \gg c_1 \log T$; step (ii) follows from $x_t \in \mathcal{E}_{t,1}$ and Lemma 4; whereas step (iii) follows from the following two facts:

$$\log \mathsf{det}\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big) = -\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathsf{Tr}\big(J_t(x_t)\big) + O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\|J_t(x_t)\|_{\mathsf{F}}^2\bigg),$$

and

$$\frac{d}{2}\log\frac{2\alpha_t-1-\overline{\alpha}_t}{1-\overline{\alpha}_t} = \frac{d(1-\alpha_t)}{1-\overline{\alpha}_t} + O\bigg(\frac{d(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^2}\bigg) = O\left(\frac{d\log T}{T}\right). \tag{58}$$

Then we can use (56) to achieve

$$\begin{aligned}
&\int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0 \\
&\quad= \int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} - \zeta_t(x_t,x_0)\Big)\mathrm{d}x_0 \cdot \exp\Big(-\log\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big) \\
&\qquad\qquad + \frac{d}{2}\log\frac{2\alpha_t-1-\overline{\alpha}_t}{1-\overline{\alpha}_t} + O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)\bigg)\Big).
\end{aligned}$$

Define a function $\xi_t(\cdot)$ as follows

$$\xi_t(x_t) := -\log\frac{\int_{x_0} p_{X_0}(x_0)\exp\big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} - \zeta_t(x_t,x_0)\big)\mathrm{d}x_0}{\int_{x_0} p_{X_0}(x_0)\exp\big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\big)\mathrm{d}x_0}. \tag{59}$$

Then we can write

$$\begin{aligned}
&\int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0 \qquad (60)\\
&\quad= \exp\left(-\xi_t(x_t) + O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)\bigg)\right)\int_{x_0} p_{X_0}(x_0) \\
&\qquad\quad \cdot\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} - \log\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big) + \frac{d}{2}\log\frac{2\alpha_t-1-\overline{\alpha}_t}{1-\overline{\alpha}_t}\Big)\mathrm{d}x_0,
\end{aligned}$$

and $\xi_t(x_t) \le 0$ for any $x_t \in \mathcal{E}_{t,1}$ since

$$\begin{aligned}
\exp(-\xi_t(x_t)) &= \int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\exp\big(-\zeta_t(x_t,x_0)\big)\mathrm{d}x_0 \\
&\ge 1 - \int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\zeta_t(x_t,x_0)\mathrm{d}x_0 = 1,
\end{aligned}$$

where we have used the fact that $e^x \ge 1+x$ for any $x\in\mathbb{R}$. Notice that

$$p_{X_t}(x_t) = \big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0, \tag{61}$$

we can rearrange terms in (60) to achieve

$$\begin{aligned}
&\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1}p_{X_t}(x_t)\\
&\quad=\big(2\pi(2\alpha_t-1-\overline{\alpha}_t)\big)^{-d/2}\int_{x_0}p_{X_0}(x_0)\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0\\
&\qquad\cdot\exp\Bigg(\xi_t(x_t)+O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T+\|J_t(x_t)\|_{\mathsf{F}}^2\big)\bigg)\Bigg),
\end{aligned}\tag{62}$$

which gives the desired decomposition (47). To establish (48), we need the following result.

Lemma 13 *We have*

$$\int_{x_0}\int_{x_t\notin\mathcal{E}_{t,1}}(2\pi(2\alpha_t-1-\overline{\alpha}_t))^{-d/2}p_{X_0}(x_0)\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0\mathrm{d}u_t\le T^{-4}\tag{63a}$$

and

$$\int_{x_t\in\mathcal{E}_{t,1}^{\mathrm{c}}}p_{X_t}(x_t)\mathrm{d}x_t\le T^{-4}.\tag{63b}$$

Proof See Appendix A.6. ■

Then we have

$$\begin{aligned}
1&\overset{\text{(i)}}{\ge}\int_{x_t\in\mathcal{E}_{t,1}}\int_{x_0}(2\pi(2\alpha_t-1-\overline{\alpha}_t))^{-d/2}p_{X_0}(x_0)\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0\mathrm{d}u_t\\
&\overset{\text{(ii)}}{=}\int_{x_t\in\mathcal{E}_{t,1}}\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1}p_{X_t}(x_t)\\
&\qquad\qquad\cdot\exp\Big(O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T+\|J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)-\xi_t(x_t)\Big)\mathrm{d}u_t\\
&\overset{\text{(iii)}}{=}\int_{x_t\in\mathcal{E}_{t,1}}p_{X_t}(x_t)\exp\Bigg(-\xi_t(x_t)+O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T+\|J_t(x_t)\|_{\mathsf{F}}^2\big)\bigg)\Bigg)\mathrm{d}x_t\\
&\overset{\text{(iv)}}{\ge}\int_{x_t\in\mathcal{E}_{t,1}}\Bigg(1-\xi_t(x_t)+O\bigg(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(d\log T+\|J_t(x_t)\|_{\mathsf{F}}^2\big)\bigg)\Bigg)p_{X_t}(x_t)\mathrm{d}x_t.
\end{aligned}$$

Here step (i) follows from (63a); step (ii) utilizes (62); step (iii) holds since $u_t=x_t+(1-\alpha_t)s_t^\star(x_t)$, namely

$$\mathrm{d}u_t=\mathsf{det}\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\mathrm{d}x_t;$$

while step (iv) follows from the fact that $e^x\ge1+x$ for any $x\in\mathbb{R}$. Recall that $\xi_t(x_t)\le0$ for any $x_t\in\mathcal{E}_{t,1}$. By rearranging terms, we have

$$\begin{aligned}
&\int_{x_t\in\mathcal{E}_{t,1}}|\xi_t(x_t)|p_{X_t}(x_t)\mathrm{d}x_t\\
&\qquad\le\int_{x_t\in\mathcal{E}_{t,1}^{\mathrm{c}}}p_{X_t}(x_t)\mathrm{d}x_t+C_3\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\int_{x_t\in\mathcal{E}_{t,1}}\big(d\log T+\|J_t(x_t)\|_{\mathsf{F}}^2\big)p_{X_t}(x_t)\mathrm{d}x_t\\
&\qquad\le C_3\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\int_{x_t\in\mathcal{E}_{t,1}}\big(d\log T+\|J_t(x_t)\|_{\mathsf{F}}^2\big)p_{X_t}(x_t)\mathrm{d}x_t+T^{-4}
\end{aligned}$$

for some universal constant $C_3 > 0$, where the last step follows from (63b).

A.4 Proof of Lemma 11

Recall the definition of $\zeta_t(x_t, x_0)$ from (55) in Appendix A.3. For any $x_t \in \mathcal{E}_{t,1}$, we have

$$\begin{aligned}
-\zeta_t(x_t, x_0) &\overset{\text{(i)}}{\leq} 2\frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t)\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \mathrm{d}x_0 + 2\frac{1-\alpha_t}{1-\overline{\alpha}_t}\big|s_t^\star(x_t)^\top(x_t - \sqrt{\overline{\alpha}_t}x_0)\big| \\
&\overset{\text{(ii)}}{\leq} 4\frac{1-\alpha_t}{1-\overline{\alpha}_t}(6C_1 + 3c_0)d\log T + (1-\alpha_t)\|s_t^\star(x_t)\|_2^2 + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \\
&\overset{\text{(iii)}}{\leq} 50\frac{1-\alpha_t}{1-\overline{\alpha}_t}(C_1 + c_0)d\log T + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \\
&\overset{\text{(iv)}}{\leq} 1 + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2.
\end{aligned} \tag{64}$$

Here step (i) utilizes (38), (55) and (57); step (ii) follows from the AM-GM inequality and an intermediate step in (41):

$$\frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t)\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \mathrm{d}x_0 \leq 2(6C_1 + 3c_0)d\log T,$$

where we also use the fact that $-\log p_{X_t}(x_t) \leq C_1 d\log T$ for $x_t \in \mathcal{E}_{t,1}$; step (iii) follows from Lemma 4; while step (iv) follows from Lemma 18 and holds provided that $T \gg c_1(C_1 + c_0)$. In addition, we also have

$$\begin{aligned}
\|J_t(x_t)\|_{\mathsf{F}}^2 \leq 2\|I_d - J_t(x_t)\|_{\mathsf{F}}^2 + 2\|I_d\|_{\mathsf{F}}^2 &\overset{\text{(a)}}{\leq} 2\big[\mathsf{Tr}\big(I_d - J_t(x_t)\big)\big]^2 + 2d \\
&\overset{\text{(b)}}{\leq} 288(C_1 + c_0)^2 d^2 \log^2 T + 2d,
\end{aligned} \tag{65}$$

for $x_t \in \mathcal{E}_{t,1}$, where step (a) holds since $I_d - J_t(x_t) \succeq 0$ and step (b) follows from Lemma 4. Substituting the bounds (64), (65) and (58) into (56) gives

$$\begin{aligned}
-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} \leq -\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} - \log\det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big) \\
+ \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 + 2,
\end{aligned} \tag{66}$$

provided that $T \gg c_1(C_1 + c_0)d\log^2 T$. Taking (66) and (58) collectively yields

$$\begin{aligned}
&\det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\int_{x_0} p_{X_0}(x_0)\big(2\pi(2\alpha_t - 1 - \overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)\mathrm{d}x_0 \\
&\quad \leq 10\int_{x_0} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2\Big)\mathrm{d}x_0.
\end{aligned} \tag{67}$$

provided that $T \gg d\log T$. To achieve the desired result, it suffices to connect the above expression with

$$p_{X_t}(x_t) = \int_{x_0} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0.$$

For any $x_t \in \mathcal{E}_{t,1}$, define a set

$$\mathcal{A}(x_t) \coloneqq \Big\{x_0 : \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 > (6C_1 + 3c_0)\frac{1-\alpha_t}{1-\overline{\alpha}_t}d\log T\Big\}.$$

We have

$$\begin{aligned}
&\int_{x_0\in\mathcal{A}(x_t)} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2\Big)\mathrm{d}x_0 \\
&\quad = p_{X_t}(x_t)\int_{x_0\in\mathcal{A}(x_t)} p_{X_0|X_t}(x_0\,|\,x_t)\exp\Big(\frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2\Big)\mathrm{d}x_0 \\
&\quad \overset{\text{(i)}}{\leq} p_{X_t}(x_t)\int_{x_0\in\mathcal{A}(x_t)} p_{X_0}(x_0)\exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{3(1-\overline{\alpha}_t)} + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2\Big)\mathrm{d}x_0 \\
&\quad \overset{\text{(ii)}}{\leq} p_{X_t}(x_t)\int_{x_0\in\mathcal{A}(x_t)} p_{X_0}(x_0)\exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{4(1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0 \\
&\quad \overset{\text{(iii)}}{\leq} p_{X_t}(x_t)\exp\Big(-\frac{(6C_1+3c_0)d\log T}{4}\Big)\int_{x_0\in\mathcal{A}(x_t)} p_{X_0}(x_0)\mathrm{d}x_0 \overset{\text{(iv)}}{\leq} \frac{1}{2}p_{X_t}(x_t). \qquad (68)
\end{aligned}$$

Here step (i) follows from (37); step (ii) utilizes Lemma 18 and holds provided that $T \gg c_1 \log T$; step (iii) follows from the definition of $\mathcal{A}(x_t)$; while step (iv) holds provided that C_1 is sufficiently large. On the other hand, we have

$$\begin{aligned}
&\int_{x_0\in\mathcal{A}(x_t)^{\mathrm{c}}} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2\Big)\mathrm{d}x_0 \\
&\quad \overset{\text{(a)}}{\leq} \exp\Big((6C_1+3c_0)\frac{1-\alpha_t}{1-\overline{\alpha}_t}d\log T\Big) \\
&\qquad \cdot \int_{x_0} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0 \\
&\quad \overset{\text{(b)}}{\leq} \exp\Big((6C_1+3c_0)\frac{8c_1 d\log^2 T}{T}\Big)p_{X_t}(x_t) \overset{\text{(c)}}{\leq} \frac{3}{2}p_{X_t}(x_t). \qquad (69)
\end{aligned}$$

Here step (a) follows from the definition of $\mathcal{A}(x_t)$; step (b) utilizes Lemma 18; whereas step (c) holds provided that $T \gg c_1(C_1 + c_0)d\log^2 T$. Taking (67), (68) and (69) collectively gives

$$\begin{aligned}
&\det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\int_{x_0} p_{X_0}(x_0)\big(2\pi(2\alpha_t - 1 - \overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0 \\
&\qquad \leq 20 p_{X_t}(x_t).
\end{aligned}$$

Rearrange terms to achieve the desired result.

A.5 Proof of Lemma 12

By definition of $\delta_{t-1}(x_{t-1})$ in (50), we have

$$\int_{x_{t-1}} \delta_{t-1}(x_{t-1})\mathrm{d}x_{t-1} = \int_{x_0}\int_{x_{t-1}}\int_{x_t\notin\mathcal{E}_{t,1}} p_{X_0}(x_0)\Big(\frac{\alpha_t}{4\pi^2(1-\alpha_t)(2\alpha_t-1-\overline{\alpha}_t)}\Big)^{d/2} \tag{70}$$
$$\cdot\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|_2^2}{2(1-\alpha_t)}\Big)\mathrm{d}x_{t-1}\mathrm{d}u_t\mathrm{d}x_0$$
$$\overset{\text{(i)}}{=}\int_{x_0}\int_{x_t\notin\mathcal{E}_{t,1}}(2\pi(2\alpha_t-1-\overline{\alpha}_t))^{-d/2}p_{X_0}(x_0)\exp\Big(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t-1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0\mathrm{d}u_t \overset{\text{(ii)}}{\le} T^{-4}.$$

Here step (i) holds since for fixed u_t, the following function

$$\Big(2\pi\frac{1-\alpha_t}{\alpha_t}\Big)^{-d/2}\exp\Big(-\frac{\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|_2^2}{2(1-\alpha_t)}\Big)$$

is a density function w.r.t. x_{t-1}, while step (ii) was established in (63a).

A.6 Proof of Lemma 13

Proof of (63). We first prove (63b). Recall that

$$\mathcal{E}_{t,1}=\big\{x_t: -\log p_{X_t}(x_t)\le C_1 d\log T, \|x_t\|_2\le\sqrt{\overline{\alpha}_t}T^{2c_R}+C_2\sqrt{d(1-\overline{\alpha}_t)\log T}\big\}.$$

Then we can decompose

$$\int_{x_t\in\mathcal{E}_{t,1}^{\mathrm{c}}} p_{X_t}(x_t)\mathrm{d}x_t = \int p_{X_t}(x_t)\,\mathbb{1}\big\{\|x_t\|_2>\sqrt{\overline{\alpha}_t}T^{2c_R}+C_2\sqrt{d(1-\overline{\alpha}_t)\log T}\big\}\mathrm{d}x_t$$
$$+\int p_{X_t}(x_t)\,\mathbb{1}\big\{-\log p_{X_t}(x_t)>C_1 d\log T, \|x_t\|_2\le\sqrt{\overline{\alpha}_t}T^{2c_R}+C_2\sqrt{d(1-\overline{\alpha}_t)\log T}\big\}\mathrm{d}x_t$$
$$\overset{\text{(i)}}{\le}\exp\Big(-\frac{C_1}{2}d\log T\Big)+\mathbb{P}\big(\|X_t\|_2>\sqrt{\overline{\alpha}_t}T^{2c_R}+C_2\sqrt{d(1-\overline{\alpha}_t)\log T}\big)$$
$$\overset{\text{(ii)}}{\le}\exp\Big(-\frac{C_1}{2}d\log T\Big)+\mathbb{P}\big(\|X_0\|_2>T^{2c_R}\big)+\mathbb{P}\big(\|\overline{W}_t\|_2>C_2\sqrt{d\log T}\big)$$
$$\overset{\text{(iii)}}{\le}\exp\Big(-\frac{C_1}{2}d\log T\Big)+\frac{\mathbb{E}[\|X_0\|_2]}{T^{2c_R}}+\mathbb{P}\big(\|\overline{W}_t\|_2>C_2\sqrt{d\log T}\big)\overset{\text{(iv)}}{\le}T^{-4}.$$

Here step (i) follows from a simple volume argument

$$\int p_{X_t}(x_t)\,\mathbb{1}\big\{-\log p_{X_t}(x_t)>C_1 d\log T, \|x_t\|_2\le\sqrt{\overline{\alpha}_t}T^{2c_R}+C_2\sqrt{d(1-\overline{\alpha}_t)\log T}\big\}\mathrm{d}x_t$$
$$\le\big(2\sqrt{\overline{\alpha}_t}T^{2c_R}+2C_2\sqrt{d(1-\overline{\alpha}_t)\log T}\big)^d\exp\left(-C_1 d\log T\right)\le\exp\Big(-\frac{C_1}{2}d\log T\Big),$$

provided that $C_1\gg c_R$ and $T\gg d\log T$; step (ii) follows from $X_t=\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}\,\overline{W}_t$; step (iii) utilizes Markov's inequality; while step (iv) holds provided that $C_1, C_2, c_R>0$ are large enough. This establishes (63b).

Then we prove (63a). Define

$$\mathcal{B}_t := \big\{x : \|x\|_2 \leq \sqrt{\overline{\alpha}_t}T^{2c_R} + C_2\sqrt{d(2\alpha_t - 1 - \overline{\alpha}_t)\log T}\big\},$$

and for each $k \geq 1$,

$$\mathcal{L}_{t,k} := \big\{x_t : 2^{k-1}C_1 d\log T < -\log p_{X_t}(x_t) \leq 2^k C_1 d \log T\big\}.$$

We first decompose

$$\begin{aligned}
I &:= \int_{x_0}\int_{x_t\notin\mathcal{E}_{t,1}} (2\pi(2\alpha_t - 1 - \overline{\alpha}_t))^{-d/2} p_{X_0}(x_0)\exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)\mathrm{d}x_0\mathrm{d}u_t \\
&\overset{\text{(a)}}{\leq} \underbrace{\int_{x_0}\int_{u_t\notin\mathcal{B}_t} p_{X_0}(x_0)\big(2\pi(2\alpha_t - 1 - \overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0}_{=:I_0} \\
&\quad + \sum_{k=1}^{\infty}\underbrace{\int_{x_0}\int_{x_t\in\mathcal{L}_{t,k},u_t\in\mathcal{B}_t} p_{X_0}(x_0)\big(2\pi(2\alpha_t - 1 - \overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)\mathrm{d}x_0\mathrm{d}u_t}_{=:I_k},
\end{aligned}$$

where step (a) holds since $\mathcal{E}_{t,1}^{\mathrm{c}} = \cup_{k=1}^{\infty}\mathcal{L}_{t,k}$. The first term I_0 can be upper bounded as follows:

$$\begin{aligned}
I_0 &\leq \Big(\int_{\|x_0\|_2\geq T^{2c_R}}\int_{u_t} + \int_{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2 \geq C_2\sqrt{d(2\alpha_t - 1 - \overline{\alpha}_t)\log T}}\int_{x_0}\Big)p_{X_0}(x_0) \\
&\qquad \cdot\Big(\frac{1}{2\pi(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)^{d/2}\exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0 \\
&\overset{\text{(i)}}{\leq} \mathbb{P}\big(\|X_0\|_2 \geq T^{2c_R}\big) + \mathbb{P}\big(\|Z\|_2 \geq C_2\sqrt{d\log T}\big) \overset{\text{(ii)}}{\leq} \frac{\mathbb{E}[\|X_0\|_2]}{T^{2c_R}} + \frac{1}{2}T^{-5} \overset{\text{(iii)}}{\leq} T^{-5}.
\end{aligned} \tag{71}$$

Here step (i) holds since

$$(2\pi(2\alpha_t - 1 - \overline{\alpha}_t))^{-d/2}p_{X_0}(x_0)\exp\Big(-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)}\Big)$$

is the joint density of $(X_0, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{2\alpha_t - 1 - \overline{\alpha}_t}Z)$ where $Z \sim \mathcal{N}(0, I_d)$ is independent of X_0; step (ii) utilizes Markov's inequality and holds when C_2 is sufficiently large; whereas step (iii) holds provided that c_R is sufficiently large. Regarding I_k, we first show that

$$\begin{aligned}
-\frac{\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} &\overset{\text{(a)}}{\leq} -\frac{(\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2 - (1-\alpha_t)\|s_t^{\star}(x_t)\|_2)^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} \\
&\leq -\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(2\alpha_t - 1 - \overline{\alpha}_t)} + \frac{1-\alpha_t}{2\alpha_t - 1 - \overline{\alpha}_t}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2\|s_t^{\star}(x_t)\|_2 \\
&\overset{\text{(b)}}{\leq} -\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{1-\alpha_t}{(1-\overline{\alpha}_t)(2\alpha_t - 1 - \overline{\alpha}_t)}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \\
&\qquad + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)(2\alpha_t - 1 - \overline{\alpha}_t)}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 + \frac{(1-\alpha_t)(1-\overline{\alpha}_t)}{4(2\alpha_t - 1 - \overline{\alpha}_t)}\|s_t^{\star}(x_t)\|_2^2 \\
&\overset{\text{(c)}}{\leq} -\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + (1-\alpha_t)\,\|s_t^{\star}(x_t)\|_2^2.
\end{aligned} \tag{72}$$

Here step (a) utilizes the triangle inequality and $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$; step (b) invokes the AM-GM inequality; whereas step (c) follows from (57). Therefore we have

$$
\begin{aligned}
I_k &\overset{\text{(i)}}{\leq} \int_{x_t\in\mathcal{L}_{t,k},u_t\in\mathcal{B}_t}\int_{x_0} p_{X_0}(x_0)\Big(\frac{1}{2\pi(1-\overline{\alpha}_t)}\Big)^{d/2} \\
&\qquad\qquad\qquad\qquad \cdot \exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + (1-\alpha_t)\left\|s_t^\star(x_t)\right\|_2^2\Big)\mathrm{d}x_0\mathrm{d}u_t \\
&= \int_{x_t\in\mathcal{L}_{t,k},u_t\in\mathcal{B}_t}\int_{x_0} p_{X_0,X_t}(x_0,x_t)\exp\Big((1-\alpha_t)\left\|s_t^\star(x_t)\right\|_2^2\Big)\mathrm{d}x_0\mathrm{d}u_t \\
&\overset{\text{(ii)}}{=} \exp\Big(200c_1(2^kC_1+c_0)\frac{d\log^2 T}{T}\Big)\int_{x_t\in\mathcal{L}_{t,k},u_t\in\mathcal{B}_t} p_{X_t}(x_t)\mathrm{d}u_t \\
&\overset{\text{(iii)}}{\leq} \exp\Big(200c_1(2^kC_1+c_0)\frac{d\log^2 T}{T}\Big)\int_{u_t\in\mathcal{B}_t}\exp\Big(-2^{k-1}C_1 d\log T\Big)\,\mathrm{d}u_t \\
&\overset{\text{(iv)}}{\leq} \exp\Big(200c_1(2^kC_1+c_0)\frac{d\log^2 T}{T} - 2^{k-1}C_1 d\log T + 4dc_R\log T + 4d\log(C_2 d)\Big) \\
&\overset{\text{(v)}}{\leq} \exp\Big(-\frac{C_1}{4}2^k d\log T\Big) = T^{-(C_1/4)2^k d}.
\end{aligned}
\tag{73}
$$

Here step (i) follows from (72); step (ii) uses a consequence of Lemma 4 and Lemma 18: for $x_t \in \mathcal{L}_{t,k}$,

$$
(1-\alpha_t)\left\|s_t^\star(x_t)\right\|_2^2 \leq 25\frac{1-\alpha_t}{1-\overline{\alpha}_t}(2^kC_1+c_0)d\log T \leq 200c_1(2^kC_1+c_0)\frac{d\log^2 T}{T};
$$

step (iii) follows from the definition of $\mathcal{L}_{t,k}$, which ensures tht $p_{X_t}(x_t) \leq \exp(-2^{k-1}C_1 d\log T)$ for any $x_t \in \mathcal{L}_{t,k}$; step (iv) follows from

$$
\begin{aligned}
\log \mathsf{vol}(\mathcal{B}_t) &\leq d\log\left(2\sqrt{\overline{\alpha}_t}T^{2c_R} + 2C_2\sqrt{d(2\alpha_t-1-\overline{\alpha}_t)\log T}\right) \\
&\leq 4c_R d\log T + 4d\log(C_2 d);
\end{aligned}
$$

and finally, step (v) holds provided that $C_1 \gg c_R + c_0$ and $T \gg d\log^2 T$. Taking (72) and (73) collectively yields

$$
I \leq I_0 + \sum_{k=1}^{\infty} I_k \leq T^{-5} + \sum_{k=1}^{\infty} T^{-(C_1/4)2^k d} \leq T^{-4},
$$

provided that C_1 is sufficiently large.

Appendix B. Proof of auxiliary lemmas in Section 5

B.1 Proof of Lemma 8

We start with the following decomposition

$$
\begin{aligned}
&\mathsf{Tr}\left(I - J_t(x_t)\right) \\
&\overset{\text{(i)}}{=} \frac{1}{1-\overline{\alpha}_t}\Big(\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2\mathrm{d}x_0 - \Big\|\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big\|_2^2\Big) \\
&\overset{\text{(ii)}}{=} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\Big[\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\|x_0(x_t) - x_0\|_2^2\mathrm{d}x_0 - \Big\|\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\big(x_0(x_t) - x_0\big)\mathrm{d}x_0\Big\|_2^2\Big] \\
&\quad + 2\sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}}\Big[\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\omega^\top\big(x_0(x_t) - x_0\big)\mathrm{d}x_0 - \omega^\top\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\big(x_0(x_t) - x_0\big)\mathrm{d}x_0\Big] \\
&\le \underbrace{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\|x_0(x_t) - x_0\|_2^2\mathrm{d}x_0}_{=:\xi} + \underbrace{4\sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}}\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\big|\omega^\top\big(x_0(x_t) - x_0\big)\big|\mathrm{d}x_0}_{=:\zeta}.
\end{aligned}
$$

Here step (i) follows from the definition of $J_t(\cdot)$ in (10), while step (ii) utilizes the decomposition (29). Then we bound ξ and ζ respectively.

- Regarding ξ, we have

$$
\begin{aligned}
\xi &\le \sum_{i=1}^{N_\varepsilon} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\sup_{x_0\in\mathcal{B}_i}\|x_0(x_t) - x_0\|_2^2\mathbb{P}\left(X_0\in\mathcal{B}_i\,|\,X_t = x_t\right) \\
&\overset{\text{(a)}}{\le} \sum_{i=1}^{N_\varepsilon} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\big(\|x^\star_{i(x_t)} - x^\star_i\|_2 + 2\varepsilon\big)^2\mathbb{P}\left(X_0\in\mathcal{B}_i\,|\,X_t = x_t\right) \\
&\le 2\underbrace{\sum_{i\in\mathcal{I}(x_t;C_1\theta)} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\|x^\star_{i(x_t)} - x^\star_i\|_2^2\mathbb{P}\left(X_0\in\mathcal{B}_i\,|\,X_t = x_t\right)}_{=:\xi_1} \\
&\qquad + 2\underbrace{\sum_{i\notin\mathcal{I}(x_t;C_1\theta)} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\|x^\star_{i(x_t)} - x^\star_i\|_2^2\mathbb{P}\left(X_0\in\mathcal{B}_i\,|\,X_t = x_t\right)}_{=:\xi_2} + 4\underbrace{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\varepsilon^2}_{=:\xi_3},
\end{aligned}
$$

  where the constant C_1 was specified in Lemma 7. Here step (a) follows from the fact that, for $x_0\in\mathcal{B}_i$, we have

$$
\|x_0(x_t) - x_0\|_2 \le \|x_0(x_t) - x^\star_{i(x_t)}\|_2 + \|x^\star_{i(x_t)} - x^\star_i\|_2 + \|x^\star_i - x_0\|_2 \le \|x^\star_{i(x_t)} - x^\star_i\|_2 + 2\varepsilon; \tag{74}
$$

  In view of the definition of $\mathcal{I}(x_t;C_1\theta)$, we have

$$
\xi_1 \le C_1\theta k\log T\sum_{i\in\mathcal{I}(x_t;C_1\theta)}\mathbb{P}\left(X_0\in\mathcal{B}_i\,|\,X_t = x_t\right) \le C_1\theta k\log T.
$$

To bound ξ_2, we have

$$
\begin{aligned}
\xi_2 &\overset{\text{(i)}}{\leq} \sum_{i \notin \mathcal{I}(x_t; C_1\theta)} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t} \|x^\star_{i(x_t)} - x^\star_i\|_2^2 \exp\left(-\frac{\overline{\alpha}_t}{16\,(1-\overline{\alpha}_t)} \|x^\star_{i(x_t)} - x^\star_i\|_2^2\right) \mathbb{P}\left(X_0 \in \mathcal{B}_i\right) \\
&\overset{\text{(ii)}}{\leq} \sum_{i \notin \mathcal{I}(x_t; C_1\theta)} \exp\left(-\frac{\overline{\alpha}_t}{32\,(1-\overline{\alpha}_t)} \|x^\star_{i(x_t)} - x^\star_i\|_2^2\right) \mathbb{P}\left(X_0 \in \mathcal{B}_i\right) \\
&\overset{\text{(iii)}}{\leq} \sum_{i \notin \mathcal{I}(x_t; C_1\theta)} \exp\left(-\frac{1}{32} C_1 \theta k \log T\right) \mathbb{P}\left(X_0 \in \mathcal{B}_i\right) \leq \exp\left(-\frac{1}{32} C_1 \theta k \log T\right).
\end{aligned}
$$

Here step (i) follows from Lemma 7, while step (ii) holds when $\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\|x^\star_{i(x_t)} - x^\star_i\|_2^2$ is large enough, which can be guaranteed by taking $C_1 > 0$ to be sufficiently large; step (iii) follows from the definition of $\mathcal{I}(x_t; C_1\theta)$. In addition, $\xi_3 \ll 1$ as long as ε is sufficiently small (see (27). Therefore we have

$$
\xi \leq 2\xi_1 + 2\xi_2 + \xi_3 \leq 3C_1 \theta k \log T \tag{75}
$$

provided that $C_1 > 0$ is sufficiently large.

- Regarding ζ, we have

$$
\begin{aligned}
\zeta &\leq \sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}} \sum_{i=1}^{N_\varepsilon} \sup_{x_0 \in \mathcal{B}_i} \left|\omega^\top\left(x_0(x_t) - x_0\right)\right| \mathbb{P}\left(X_0 \in \mathcal{B}_i \,|\, X_t = x_t\right) \\
&\overset{\text{(a)}}{\leq} \sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}} \sum_{i=1}^{N_\varepsilon} \left(\left|\omega^\top\left(x^\star_i - x^\star_{i(x_t)}\right)\right| + \varepsilon\|\omega\|_2\right) \mathbb{P}\left(X_0 \in \mathcal{B}_i \,|\, X_t = x_t\right) \\
&\overset{\text{(b)}}{\leq} \underbrace{\sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}} \sum_{i \in \mathcal{I}(x_t; C_1\theta)} \sqrt{\theta k \log T} \|x^\star_i - x^\star_{i(x_t)}\|_2 \mathbb{P}\left(X_0 \in \mathcal{B}_i \,|\, X_t = x_t\right)}_{=:\zeta_1} \\
&\quad + \underbrace{\sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}} \sum_{i \notin \mathcal{I}(x_t; C_1\theta)} \sqrt{\theta k \log T} \|x^\star_i - x^\star_{i(x_t)}\|_2 \mathbb{P}\left(X_0 \in \mathcal{B}_i \,|\, X_t = x_t\right)}_{=:\zeta_2} \\
&\quad + \underbrace{\sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}} \varepsilon(2\sqrt{d} + \sqrt{\theta k \log T})}_{=:\zeta_3}.
\end{aligned}
$$

Here step (a) uses Cauchy-Schwarz inequality, while step (b) follows from the definition of $\mathcal{G}$. By the definition of $\mathcal{I}(x_t; C_1\theta)$, we have

$$
\zeta_1 \leq \sum_{i \in \mathcal{I}(x_t; C_1\theta)} \sqrt{C_1} \theta k \log T \mathbb{P}\left(X_0 \in \mathcal{B}_i \,|\, X_t = x_t\right) \leq \sqrt{C_1} \theta k \log T.
$$

To bound ζ_2, we have

$$\begin{aligned}\zeta_2 &\overset{\text{(i)}}{\leq} \sum_{i \notin \mathcal{I}(x_t; C_1\theta)} \sqrt{\theta k \log T} \exp\left(-\frac{\overline{\alpha}_t}{32\left(1-\overline{\alpha}_t\right)}\|x^\star_{i(x_t)} - x^\star_i\|_2^2\right) \mathbb{P}\left(X_0 \in \mathcal{B}_i\right) \\ &\overset{\text{(ii)}}{\leq} \sqrt{\theta k \log T} \exp\left(-\frac{1}{32} C_1 \theta k \log T\right) \overset{\text{(iii)}}{\leq} \exp\left(-\frac{1}{64} C_1 \theta k \log T\right).\end{aligned}$$

Here step (i) holds when $\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\|x^\star_{i(x_t)} - x^\star_i\|_2^2$ is large enough, which can be guaranteed by taking $C_1 > 0$ to be sufficiently large; step (ii) follows from the definition of $\mathcal{I}(x_t; C_1\theta)$; and step (iii) holds when C_1 is large enough. In addition, we have $\xi_3 \ll 1$ as long as ε is sufficiently small (see (27). Hence we have

$$\zeta \leq 2\sqrt{C_1}\theta k \log T \tag{76}$$

provided that $C_1 > 0$ is sufficiently large.

Taking the bounds on ξ and ζ collectively leads to

$$\mathsf{Tr}\left(I - J_t(x_t)\right) \leq \xi + 4\zeta \leq 4C_1 \theta k \log T$$

provided that $C_1 > 0$ is large enough. In addition, since $I - J_t(x_t) \succeq 0$, we have

$$\|I - J_t(x_t)\|_{\mathrm{F}}^2 \leq \mathsf{Tr}\left(I - J_t(x_t)\right)^2,$$

hence we have

$$\|I - J_t(x_t)\| \leq \|I - J_t(x_t)\|_{\mathrm{F}} \leq \mathsf{Tr}\left(I - J_t(x_t)\right) \leq C_2 \theta k \log T$$

provided that $C_2 \geq 4C_1$. This finishes the proof of the first relation (32).

Finally, we invoke Lemma 22 to obtain

$$\sum_{t=2}^{T} \frac{1-\alpha_t}{1-\overline{\alpha}_t} \mathsf{Tr}\big(\mathbb{E}\big[\big(\Sigma_{\overline{\alpha}_t}(X_t)\big)^2\big]\big) \leq C_J k \log T, \tag{77}$$

where the matrix function

$$\Sigma_{\overline{\alpha}_t}(x) = \mathsf{Cov}\big(Z \mid \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} Z = x\big) = I_d - J_t(x).$$

Here $Z \sim \mathcal{N}(0, I_d)$ is independent of X_0. By noticing that

$$\mathsf{Tr}\big(\mathbb{E}\big[\big(\Sigma_{\overline{\alpha}_t}(X_t)\big)^2\big]\big) = \mathsf{Tr}\big(\mathbb{E}\big[\big(I_d - J_t(X_t)\big)^2\big]\big) = \mathbb{E}\big[\|I - J_t(X_t)\|_{\mathrm{F}}^2\big] = \int_{x_t} \|I - J_t(x_t)\|_{\mathrm{F}}^2\, p_{X_t}(x_t) \mathrm{d}x_t,$$

we finish the proof of the second relation (33).

B.2 Proof of Lemma 9

The proof of Lemma 9 is similar to the proof of Lemma 6 in Appendix A.2. We will only highlight the differences due to the different update rule (8). Equation (46) should be changed to

$$
\begin{aligned}
& p_{X_{t-1}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) && (78)\\
& \geq \int_{x_t \in \mathcal{E}_{t,1}} \mathsf{det}\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t} J_t(x_t)\Big)^{-1} p_{X_t}(x_t) \Big(\frac{\alpha_t(1-\overline{\alpha}_t)}{2\pi(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)^{d/2} && (79)\\
& \qquad \cdot \exp\Big(-\frac{(1-\overline{\alpha}_t)\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|^2}{2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}u_t.
\end{aligned}
$$

Lemma 10 need to be changed to the following version.

Lemma 14 *For any $x_t \in \mathcal{E}_{t,1}$, we have*

$$
\begin{aligned}
& \mathsf{det}\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t} J_t(x_t)\Big)^{-1} p_{X_t}(x_t) \\
& = \big(2\pi(\alpha_t-\overline{\alpha}_t)\big)^{-d/2} \int_{x_0} p_{X_0}(x_0) \exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0 \\
& \qquad \cdot \exp\Big(\xi_t(x_t) + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))| + \|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big), && (80)
\end{aligned}
$$

where $\xi_t(x_t) \leq 0$ satisfies

$$
\begin{aligned}
\int_{x_t\in\mathcal{E}_{t,1}} |\xi_t(x_t)| p_{X_t}(x_t)\mathrm{d}x_t \leq C_3 \Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t\in\mathcal{E}_{t,1}} (|\mathsf{Tr}(I-J_t(x_t))| + \|I-J_t(x_t)\|_{\mathrm{F}}^2) p_{X_t}(x_t)\mathrm{d}x_t \\
+ T^{-4} \qquad (81)
\end{aligned}
$$

for some universal constant $C_5 > 0$.

Proof See Appendix B.3. ■

Taking the decomposition (80) and (78) collectively, we have

$$
\begin{aligned}
& p_{X_{t-1}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) + \delta_{t-1}(x_{t-1}) && (82)\\
& \geq \int_{x_0}\int_{x_t} \exp\Big(\Big[\xi_t(x_t) + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))| + \|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big]\mathbb{1}_{x_t\in\mathcal{E}_{t,1}}\Big) \\
& \qquad \cdot p_{X_0}(x_0)\Big(\frac{\alpha_t(1-\overline{\alpha}_t)^2}{4\pi^2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)^3}\Big)^{d/2} \\
& \qquad \cdot \exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2} - \frac{(1-\overline{\alpha}_t)\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|^2}{2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0,
\end{aligned}
$$

where we define

$$
\begin{aligned}
\delta_{t-1}(x_{t-1}) := \int_{x_0}\int_{x_t\notin\mathcal{E}_{t,1}} p_{X_0}(x_0)\Big(\frac{\alpha_t(1-\overline{\alpha}_t)^2}{4\pi^2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)^3}\Big)^{d/2} && (83)\\
\cdot \exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2} - \frac{(1-\overline{\alpha}_t)\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|^2}{2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0.
\end{aligned}
$$

Moreover, it is straightforward to check that

$$\int_{x_0}\int_{x_t} p_{X_0}(x_0)\Big(\frac{\alpha_t(1-\overline{\alpha}_t)^2}{4\pi^2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)^3}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)$$
$$\cdot\exp\Big(-\frac{(1-\overline{\alpha}_t)\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|^2}{2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0 = p_{X_{t-1}}(x_{t-1}). \tag{84}$$

Then we can continue the derivation in (82):

$$\begin{aligned}
&p_{X_{t-1}}(x_{t-1})-\Delta_{t-1}(x_{t-1})+\Delta_{t\to t-1}(x_{t-1})+\delta_{t-1}(x_{t-1})\\
&\overset{\text{(i)}}{\geq}\int_{x_0}\int_{x_t}\Big(1+\Big[\xi_t(x_t)+O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big]\mathbb{1}_{x_t\in\mathcal{E}_{t,1}}\Big)\\
&\qquad\cdot p_{X_0}(x_0)\Big(\frac{\alpha_t(1-\overline{\alpha}_t)^2}{4\pi^2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)^3}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\\
&\qquad\cdot\exp\Big(-\frac{(1-\overline{\alpha}_t)\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|^2}{2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0\\
&\overset{\text{(ii)}}{=}p_{X_{t-1}}(x_{t-1})+\int_{x_0}\int_{x_t\in\mathcal{E}_{t,1}}\Big[\xi_t(x_t)+O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big]\\
&\qquad\cdot p_{X_0}(x_0)\Big(\frac{\alpha_t(1-\overline{\alpha}_t)^2}{4\pi^2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)^3}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\\
&\qquad\cdot\exp\Big(-\frac{(1-\overline{\alpha}_t)\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|^2}{2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}u_t\mathrm{d}x_0.
\end{aligned}$$

By rearranging terms and integrate over the variable x_{t-1}, we arrive at

$$\begin{aligned}
\int_{x_{t-1}}\Delta_{t-1}(x_{t-1})\mathrm{d}x_{t-1}&\leq\int_{x_{t-1}}\big(\Delta_t(x_{t-1})+\delta_{t-1}(x_{t-1})\big)\mathrm{d}x_{t-1}\\
&+\int_{x_0}\int_{x_t\in\mathcal{E}_{t,1}}\Big(|\xi_t(x_t)|+O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big)p_{X_0}(x_0)\\
&\cdot\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t-\overline{\alpha}_t)^2}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}u_t\mathrm{d}x_0,
\end{aligned}\tag{85}$$

where we used (45) and for any fixed u_t, the function

$$\Big(2\pi\frac{(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}{(1-\overline{\alpha}_t)\alpha_t}\Big)^{-d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\big\|\sqrt{\alpha_t}x_{t-1}-u_t\big\|^2}{2(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}\Big)$$

is a density function of x_{t-1}. To establish the desired result, we need the following two lemmas.

Lemma 15 *Suppose that $T\gg\theta k\log^2 T$. For any $x_t\in\mathcal{E}_{t,1}$, we have*

$$\int_{x_0}p_{X_0}(x_0)\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t-\overline{\alpha}_t)^2}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0$$
$$\leq 20\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1}p_{X_t}(x_t).$$

Proof See Appendix B.4. ■

Lemma 16 *For the function $\delta_{t-1}(\cdot)$ defined in (50), we have $\int_{x_{t-1}} \delta_{t-1}(x_{t-1})\mathrm{d}x_{t-1} \le T^{-4}$.*

Proof The proof is the same as that of Lemma 12, and is hence omitted. ■

Equipped with Lemmas 15 and 16, we can continue the derivation in (85) as follows:

$$
\begin{aligned}
&\int_{x_{t-1}} \Delta_{t-1}(x_{t-1})\mathrm{d}x_{t-1} \\
&\overset{\text{(a)}}{\le} \int_{x_t} \Delta_t(x_t)\mathrm{d}x_t + 20\int_{x_t\in\mathcal{E}_{t,1}} \Big(|\xi_t(x_t)| + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))| + \|I-J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big) \\
&\qquad\qquad\qquad\qquad \cdot \det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1} p_{X_t}(x_t)\mathrm{d}u_t + T^{-4} \\
&\overset{\text{(b)}}{=} \int_{x_t} \Delta_t(x_t)\mathrm{d}x_t + T^{-4} \\
&\qquad + 20\int_{x_t\in\mathcal{E}_{t,1}} \Big(|\xi_t(x_t)| + O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))| + \|I-J_t(x_t)\|_{\mathsf{F}}^2\big)\Big)\Big)p_{X_t}(x_t)\mathrm{d}x_t \\
&\overset{\text{(c)}}{\le} \int_{x_t} \Delta_t(x_t)\mathrm{d}x_t + T^{-3} + C_4\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t\in\mathcal{E}_{t,1}} \big(d\log T + \|J_t(x_t)\|_{\mathsf{F}}^2\big)p_{X_t}(x_t)\mathrm{d}x_t,
\end{aligned}
$$

which establishes the desired recursive relation. Here step (a) follows from Lemmas 15 and 16; step (b) follows from $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$, hence

$$
\mathrm{d}u_t = \mathsf{det}\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\mathrm{d}x_t;
$$

whereas step (c) uses (81) in Lemma 14, and holds provided that $C_4 \gg C_3$ is sufficiently large. In addition, the relation $\int \Delta_T(x)\mathrm{d}x \le T^{-4}$ can be established in the same way as the proof of Lemma 6, and is hence omitted here.

B.3 Proof of Lemma 14

The proof is similar to that of Lemma 10. Recall that $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$, we start with the following decomposition

$$
\begin{aligned}
\frac{(1-\overline{\alpha}_t)\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2} &= \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)} \\
&\quad + \frac{(1-\alpha_t)(1-\overline{\alpha}_t)s_t^\star(x_t)^\top(x_t-\sqrt{\overline{\alpha}_t}x_0)}{(\alpha_t-\overline{\alpha}_t)^2} + \frac{(1-\alpha_t)^2(1-\overline{\alpha}_t)\|s_t^\star(x_t)\|_2^2}{2(\alpha_t-\overline{\alpha}_t)^2} \\
&= \frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)}{2(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)}\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2\mathrm{d}x_0 \\
&\quad - \frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)}{2(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)}\Big\|\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\big(x_t-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big\|_2^2 + \zeta_t(x_t,x_0),
\end{aligned}
$$

where we let

$$\zeta_t(x_t, x_0) := \frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)\big(\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2 - \int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2 \mathrm{d}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)}$$

$$+\frac{(1-\alpha_t)\big[\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)\big(x_t-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big]^\top\sqrt{\overline{\alpha}_t}\big(x_0-\int_{x_0} p_{X_0|X_t}(x_0\,|\,x_t)x_0\mathrm{d}x_0\big)}{(\alpha_t-\overline{\alpha}_t)^2}. \tag{86}$$

We can further derive

$$\begin{aligned}
\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2} &\overset{\text{(i)}}{=} \frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)}{2(\alpha_t-\overline{\alpha}_t)^2}\mathsf{Tr}\left(I-J_t(x_t)\right)+\zeta_t(x_t,x_0) \\
&\overset{\text{(ii)}}{=} \frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\mathsf{Tr}\left(I-J_t(x_t)\right)+\zeta_t(x_t,x_0)+O\Big(\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2|\mathsf{Tr}(I-J_t(x_t))|\Big) \\
&\overset{\text{(iii)}}{=} \frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \log\det\Big(I+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}(I-J_t(x_t))\Big) \\
&\qquad +\zeta_t(x_t,x_0)+O\Big(\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big).
\end{aligned} \tag{87}$$

Here step (i) follows from (38) and (39); step (ii) holds since

$$\frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)}{2(\alpha_t-\overline{\alpha}_t)^2} = \frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\left(1+\frac{1-\alpha_t}{2(\alpha_t-\overline{\alpha}_t)}\right),$$

while step (iii) uses the fact that

$$\log\mathsf{det}\Big(I+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}(I-J_t(x_t))\Big) = \frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\mathsf{Tr}\big(I-J_t(x_t)\big)+O\Big(\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\|I-J_t(x_t)\|_{\mathrm{F}}^2\Big).$$

Then we have

$$\begin{aligned}
&\int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0 \\
&\quad= \int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}-\zeta_t(x_t,x_0)\Big)\mathrm{d}x_0 \\
&\qquad\quad\cdot\exp\bigg(-\log\det\Big(I+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}(I-J_t(x_t))\Big) \\
&\qquad\qquad\qquad +O\Big(\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\bigg).
\end{aligned}$$

Recall the definition of $\xi_t(x_t)$ in (59) and (61), which allows us to write

$$\begin{aligned}
&\int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0 \\
&\quad= \det\Big(I+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}(I-J_t(x_t))\Big)^{-1}p_{X_t}(x_t)\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2} \\
&\qquad\quad\cdot\exp\bigg(-\xi_t(x_t)+O\Big(\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\bigg).
\end{aligned}$$

Using the fact that

$$\det\Big(I+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}(I-J_t(x_t))\Big)=\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^d\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big), \tag{88}$$

we arrive at

$$\begin{aligned}
&\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1}p_{X_t}(x_t)\\
&\quad=\big(2\pi(\alpha_t-\overline{\alpha}_t)\big)^{-d/2}\int_{x_0}p_{X_0}(x_0)\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0\\
&\qquad\cdot\exp\Big(\xi_t(x_t)+O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big),
\end{aligned}\tag{89}$$

which gives the desired decomposition (80).

In order to establish (81), we need the following lemma.

Lemma 17 *Suppose that $\theta \gg C_{\mathsf{cover}}$ and $T \gg c_1C_1\log T$. Then we have*

$$\int_{x_0}\int_{x_t\notin\mathcal{E}_{t,1}}p_{X_0}(x_0)\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t-\overline{\alpha}_t)^2}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0\mathrm{d}u_t\le T^{-4} \tag{90a}$$

and

$$\int_{x_t\in\mathcal{E}_{t,1}^{\mathrm{c}}}p_{X_t}(x_t)\mathrm{d}x_t\le T^{-4}. \tag{90b}$$

Proof See Appendix B.5. ■

Then we have

$$\begin{aligned}
1&\overset{\text{(i)}}{\ge}\int_{x_t\in\mathcal{E}_{t,1}}\int_{x_0}\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t-\overline{\alpha}_t)^2}\Big)^{d/2}p_{X_0}(x_0)\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0\mathrm{d}u_t\\
&\overset{\text{(ii)}}{=}\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2}\int_{x_t\in\mathcal{E}_{t,1}}\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)^{-1}p_{X_t}(x_t)\\
&\qquad\qquad\cdot\exp\Big(-\xi_t(x_t)+O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big)\mathrm{d}u_t\\
&\overset{\text{(iii)}}{=}\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2}\int_{x_t\in\mathcal{E}_{t,1}}\exp\Big(-\xi_t(x_t)+O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big)\\
&\qquad\qquad\cdot p_{X_t}(x_t)\mathrm{d}x_t\\
&\overset{\text{(iv)}}{\ge}\int_{x_t\in\mathcal{E}_{t,1}}\Big(1-\xi_t(x_t)+O\Big(\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2\big(|\mathsf{Tr}(I-J_t(x_t))|+\|I-J_t(x_t)\|_{\mathrm{F}}^2\big)\Big)\Big)p_{X_t}(x_t)\mathrm{d}x_t.
\end{aligned}$$

Here step (i) follows from (90a); step (ii) utilizes (89); step (iii) holds since $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$, namely

$$\mathrm{d}u_t=\det\Big(I-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\mathrm{d}x_t;$$

while step (iv) follows from the facts that $1 > \alpha_t$ and $e^x \geq 1 + x$ for any $x \in \mathbb{R}$. Recall that $\xi_t(x_t) \leq 0$ for any $x_t \in \mathcal{E}_{t,1}$. By rearranging terms, we have

$$\begin{aligned}
&\int_{x_t \in \mathcal{E}_{t,1}} |\xi_t(x_t)| p_{X_t}(x_t) \mathrm{d}x_t \\
&\quad \leq \int_{x_t \in \mathcal{E}_{t,1}^{\mathrm{c}}} p_{X_t}(x_t) \mathrm{d}x_t + C_3 \Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t \in \mathcal{E}_{t,1}} \big(|\mathsf{Tr}(I - J_t(x_t))| + \|I - J_t(x_t)\|_{\mathrm{F}}^2\big) p_{X_t}(x_t) \mathrm{d}x_t \\
&\quad \leq C_3 \Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}\Big)^2 \int_{x_t \in \mathcal{E}_{t,1}} \big(|\mathsf{Tr}(I - J_t(x_t))| + \|I - J_t(x_t)\|_{\mathrm{F}}^2\big) p_{X_t}(x_t) \mathrm{d}x_t + T^{-4}
\end{aligned}$$

for some universal constant $C_3 > 0$, where the last step follows from (90b).

B.4 Proof of Lemma 15

To begin with, we record the following two results from Li and Yan (2024). For any $x_t \in \mathcal{E}_{t,1}$, we have

$$\int_{x_0} p_{X_0 | X_t}(x_0 \,|\, x_t) x_0 \mathrm{d}x_0 = \overline{x}_0 + \delta \quad \text{where} \quad \overline{x}_0 \in \bigcup_{i \in \mathcal{I}(x_t; C_1\theta)} \mathcal{B}_i \tag{91a}$$

and

$$\|\delta\|_2 \leq \sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}} \exp\left(-\frac{1}{32} C_1 \theta k \log T\right). \tag{91b}$$

In addition, for any $x, x' \in \mathcal{X}_t(x_t)$, we have

$$\overline{\alpha}_t \|x - x'\|_2^2 \leq 9 C_1 \theta k \,(1 - \overline{\alpha}_t) \log T \tag{92}$$

and

$$\left|\omega^\top \left(x - x'\right)\right| \leq \sqrt{\theta k \log T} \|x - x'\|_2 + \big(4\sqrt{d} + 4\sqrt{\theta k \log T}\big)\varepsilon \tag{93}$$

See (Li and Yan, 2024, Equations (A.4), (A.5) and (A.27)) for the proof.

Recall the definition of $\zeta_t(x_t, x_0)$ in (86), which can be written as

$$\zeta_t(x_t, x_0) = \frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)}{2(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)} \theta_1(x_t, x_0) + \frac{1-\alpha_t}{(\alpha_t-\overline{\alpha}_t)^2} \theta_2(x_t, x_0),$$

where

$$\begin{aligned}
\theta_1(x_t, x_0) &= \|x_t - \sqrt{\overline{\alpha}_t} x_0\|_2^2 - \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t) \|x_t - \sqrt{\overline{\alpha}_t} x_0\|_2^2 \mathrm{d}x_0, \\
\theta_2(x_t, x_0) &= \sqrt{\overline{\alpha}_t} \Big[\int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t} x_0\big) \mathrm{d}x_0\Big]^\top \Big(x_0 - \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t) x_0 \mathrm{d}x_0\Big).
\end{aligned}$$

For any $x_t \in \mathcal{E}_{t,1}$, recall the decomposition $x_t = \sqrt{\overline{\alpha}_t} x_0(x_t) + \sqrt{1-\overline{\alpha}_t}\omega$ in (29), we have

$$\begin{aligned}
\theta_1(x_t, x_0) &= \|x_t - \sqrt{\overline{\alpha}_t} x_0(x_t) + \sqrt{\overline{\alpha}_t} x_0(x_t) - \sqrt{\overline{\alpha}_t} x_0\|_2^2 \\
&\qquad - \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t) \|x_t - \sqrt{\overline{\alpha}_t} x_0(x_t) + \sqrt{\overline{\alpha}_t} x_0(x_t) - \sqrt{\overline{\alpha}_t} x_0\|_2^2 \mathrm{d}x_0 \\
&= \overline{\alpha}_t \Big(\|x_0 - x_0(x_t)\|_2^2 - \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t) \|x_0 - x_0(x_t)\|_2^2 \mathrm{d}x_0 \Big) \\
&\qquad - 2\sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)} \Big[\omega^\top \big(x_0 - x_0(x_t)\big) - \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t) \omega^\top \big(x_0 - x_0(x_t)\big) \mathrm{d}x_0 \Big].
\end{aligned}$$

In view of (93), we have

$$\big|\omega^\top \big(x_0 - x_0(x_t)\big)\big| \le \sqrt{\theta k \log T} \|x_0 - x_0(x_t)\|_2 + 4\varepsilon(\sqrt{d} + \sqrt{\theta k \log T}).$$

We also learn from (75) and (76) in the proof of Lemma 8 that

$$\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t} \int_{x_0} p_{X_0|X_t}(x_0 \,|\, x_t) \|x_0(x_t) - x_0\|_2^2 \mathrm{d}x_0 \le 3C_1 \theta k \log T$$

and

$$\sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}} \int p_{X_0|X_t}(x_0 \,|\, x_t) \big|\omega^\top \big(x_0(x_t) - x_0\big)\big| \mathrm{d}x_0 \le 2\sqrt{C_1} \theta k \log T.$$

Taking the above bounds collectively yields

$$-\theta_1(x_t, x_0) \le 7C_1(1-\overline{\alpha}_t)\theta k \log T + 2\sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)} \sqrt{\theta k \log T} \|x_0 - x_0(x_t)\|_2$$

provided that $\varepsilon > 0$ is sufficiently small (see (27)) and $C_1 > 0$ is sufficiently large. Regarding $\theta_2(x_t, x_0)$, we first use the decomposition (91a) to achieve

$$\begin{aligned}
\theta_2(x_t, x_0) &= \sqrt{\overline{\alpha}_t} \big(x_t - \sqrt{\overline{\alpha}_t}\overline{x}_0 - \sqrt{\overline{\alpha}_t}\delta\big)^\top \big(x_0 - \overline{x}_0 - \delta\big) \\
&= \sqrt{\overline{\alpha}_t} \big(\sqrt{\overline{\alpha}_t} x_0(x_t) + \sqrt{1-\overline{\alpha}_t}\omega - \sqrt{\overline{\alpha}_t}\overline{x}_0 - \sqrt{\overline{\alpha}_t}\delta\big)^\top \big(x_0 - x_0(x_t) + x_0(x_t) - \overline{x}_0 - \delta\big) \\
&= \overline{\alpha}_t \big(x_0(x_t) - \overline{x}_0 - \delta\big)^\top \big(x_0 - x_0(x_t)\big) + \sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)} \omega^\top \big(x_0 - \overline{x}_0 - \delta\big) \\
&\qquad + \overline{\alpha}_t \|x_0(x_t) - \overline{x}_0 - \delta\|_2^2.
\end{aligned}$$

Hence we have

$$\begin{aligned}
-\theta_2(x_t, x_0) &\overset{\text{(i)}}{\le} \overline{\alpha}_t \|x_0(x_t) - \overline{x}_0 - \delta\|_2 \|x_0 - x_0(x_t)\|_2 + \sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)} \big(\big|\omega^\top \big(x_0 - \overline{x}_0\big)\big| + \|\omega\|_2 \|\delta\|_2\big) \\
&\overset{\text{(ii)}}{\le} \overline{\alpha}_t (\|x_0(x_t) - \overline{x}_0\|_2 + \|\delta\|_2) \|x_0 - x_0(x_t)\|_2 \\
&\qquad + \sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)} \big(\sqrt{\theta k \log T} (\|x_0 - x_0(x_t)\|_2 + \|x_0(x_t) - \overline{x}_0\|_2) + \|\omega\|_2 \|\delta\|_2\big) \\
&\overset{\text{(iii)}}{\le} 4\sqrt{C_1 \overline{\alpha}_t \left(1-\overline{\alpha}_t\right) \theta k \log T} \|x_0 - x_0(x_t)\|_2 + 4\sqrt{C_1}(1-\overline{\alpha}_t)\theta k \log T.
\end{aligned}$$

Here step (i) utilizes the Cauchy-Schwarz inequality; step (ii) follows from (93); step (iii) uses (92) and (91b), and holds provided that $C_1 > 0$ is sufficiently large. Hence we have

$$
\begin{aligned}
-\zeta_t(x_t, x_0) &= -\frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)}{2(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)}\theta_1(x_t,x_0) - \frac{1-\alpha_t}{(\alpha_t-\overline{\alpha}_t)^2}\theta_2(x_t,x_0) \\
&\overset{\text{(a)}}{\leq} \frac{2(1-\alpha_t)}{(1-\overline{\alpha}_t)^2}\Big(8C_1(1-\overline{\alpha}_t)\theta k\log T + 5\sqrt{C_1\overline{\alpha}_t\,(1-\overline{\alpha}_t)\,\theta k\log T}\|x_0 - x_0(x_t)\|_2\Big) \\
&\overset{\text{(a)}}{\leq} 66C_1\frac{1-\alpha_t}{1-\overline{\alpha}_t}\theta k\log T + \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)^2}\overline{\alpha}_t\|x_0 - x_0(x_t)\|_2^2. \qquad (94)
\end{aligned}
$$

provided that $C_1 > 0$ is sufficiently large. Here step (a) follows from consequences of Lemma 18

$$
\frac{(1-\alpha_t)(1+\alpha_t-2\overline{\alpha}_t)}{2(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)} = \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\Big(1+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\Big(1+\frac{1-\alpha_t}{2(\alpha_t-\overline{\alpha}_t)}\Big) \leq \frac{2(1-\alpha_t)}{(1-\overline{\alpha}_t)^2}
$$

and

$$
\frac{1-\alpha_t}{(\alpha_t-\overline{\alpha}_t)^2} = \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\Big(1+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2 \leq \frac{2(1-\alpha_t)}{(1-\overline{\alpha}_t)^2}
$$

as long as T is sufficiently large. Finally, notice that

$$
\begin{aligned}
\overline{\alpha}_t\|x_0 - x_0(x_t)\|_2^2 - \|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 &= \overline{\alpha}_t\|x_0 - x_0(x_t)\|_2^2 - \|\sqrt{\overline{\alpha}_t}x_0(x_t) + \sqrt{1-\overline{\alpha}_t}\omega - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \\
&\leq -2\sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)}\omega^\top(x_0(x_t) - x_0) \\
&\overset{\text{(i)}}{\leq} 2\sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)\theta k\log T}\|x_0 - x_0(x_t)\|_2 + 1 - \overline{\alpha}_t \\
&\overset{\text{(ii)}}{\leq} \frac{1}{2}\overline{\alpha}_t\|x_0 - x_0(x_t)\|_2^2 + 3(1-\overline{\alpha}_t)\theta k\log T
\end{aligned}
$$

where the last step follows from (93) and (27). By rearranging terms we have

$$
\overline{\alpha}_t\|x_0 - x_0(x_t)\|_2^2 \leq 2\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 + 6(1-\overline{\alpha}_t)\theta k\log T. \qquad (95)
$$

Taking (94) and (95) collectively yields

$$
-\zeta_t(x_t, x_0) \leq 69C_1\frac{1-\alpha_t}{1-\overline{\alpha}_t}\theta k\log T + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \qquad (96)
$$

provided that $C_2 \geq 1$. Armed with this relation, we can follow the same analysis in the proof of Lemma 11 to establish the desired result under the condition $T \gg \theta k\log^2 T$.

B.5 Proof of Lemma 17

Proof of (90b)**.** We have

$$
\int_{x_t\in\mathcal{E}_{t,1}^{\mathrm{c}}} p_{X_t}(x_t)\mathrm{d}x_t = \mathbb{P}\left(X_t \notin \mathcal{E}_{t,1}\right) \leq \mathbb{P}\left(X_0 \notin \cup_{i\in\mathcal{I}}\mathcal{B}_i\right) + \mathbb{P}\left(\overline{W}_t \notin \mathcal{G}\right),
$$

where we use the decomposition $X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\overline{W}_t$ for $\overline{W}_t \sim \mathcal{N}(0, I_d)$. It is straightforward to check that

$$\mathbb{P}\left(X_0 \notin \cup_{i\in\mathcal{I}}\mathcal{B}_i\right) \le N_\varepsilon \exp\left(-\theta k \log T\right) \le \exp\left(C_{\mathsf{cover}} k \log T - \theta k \log T\right) \le \frac{1}{2}\exp\left(-\frac{\theta}{4}k\log T\right)$$

provided that $\theta \gg C_{\mathsf{cover}}$. In addition, since $\overline{W}_t \sim \mathcal{N}(0, I_d)$, by the definition of $\mathcal{G}$ we know that

$$\begin{aligned}
\mathbb{P}\left(\overline{W}_t \notin \mathcal{G}\right) &\le \mathbb{P}\left(\|\overline{W}_t\|_2 > \sqrt{d} + \sqrt{C_1 k \log T}\right) \\
&\quad + \sum_{i=1}^{N_\varepsilon}\sum_{j=1}^{N_\varepsilon} \mathbb{P}\left(|(x_i^\star - x_j^\star)^\top \overline{W}_t| > \sqrt{\theta k \log T}\|x_i^\star - x_j^\star\|_2\right) \\
&\overset{\text{(i)}}{\le} \left(N_\varepsilon^2 + 1\right)\exp\left(-\frac{\theta}{2}k\log T\right) \le \left(\exp\left(2C_{\mathsf{cover}}k\log T\right)+1\right)\exp\left(-\frac{\theta}{2}k\log T\right) \\
&\overset{\text{(ii)}}{\le} \frac{1}{2}\exp\left(-\frac{\theta}{4}k\log T\right).
\end{aligned}$$

Here step (i) follows from concentration bounds for Gaussian and chi-square variables (see Lemma 19); while step (ii) holds as long as $C_1 \gg C_{\mathsf{cover}}$. Taking the above bounds collectively yields

$$\int_{x_t \in \mathcal{E}_{t,1}^{\mathrm{c}}} p_{X_t}(x_t)\mathrm{d}x_t \le \exp\left(-\frac{\theta}{4}k\log T\right) \le T^{-4} \tag{97}$$

when $\theta > 0$ is sufficiently large.

Proof of (90a)**.** For any $j \ge 1$, define

$$\begin{aligned}
\mathcal{I}_j &:= \left\{1 \le i \le N_\varepsilon : \mathbb{P}(X_0 \in \mathcal{B}_i) \ge \exp(-2^{j-1}\theta k \log T)\right\}, \\
\mathcal{G}_j &:= \Big\{\omega \in \mathbb{R}^d : \|\omega\|_2 \le 2\sqrt{d} + \sqrt{2^{j-1}\theta k \log T}, \quad \text{and} \\
&\qquad\qquad |(x_i^\star - x_j^\star)^\top \omega| \le \sqrt{2^{j-1}\theta k\log T}\|x_i^\star - x_j^\star\|_2 \quad \text{for all} \quad 1 \le i, j \le N_\varepsilon\Big\},
\end{aligned}$$

and let

$$\mathcal{L}_{t,j} := \left\{\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\omega : x_0 \in \cup_{i\in\mathcal{I}_j}\mathcal{B}_i, \omega \in \mathcal{G}_j\right\}.$$

We know that $\mathcal{L}_{t,1} \subseteq \mathcal{L}_{t,2} \subseteq \cdots$ and $\cup_{j=1}^\infty \mathcal{L}_{t,j} = \mathbb{R}^d$. Notice that $\mathcal{E}_{t,1} = \mathcal{L}_{t,1}$. By defining $\mathcal{E}_{t,j} = \mathcal{L}_{t,j+1} \setminus \mathcal{L}_{t,j}$ for each $j \ge 2$, we know that

$$\bigcup_{j=1}^\infty \mathcal{E}_{t,j} = \mathbb{R}^d \qquad \text{where } \mathcal{E}_{t,1}, \mathcal{E}_{t,2}, \dots \text{ are disjoint.}$$

For any $x_t \in \mathcal{E}_{t,j}$, there exists an index $i(x_t) \in \mathcal{I}_j$, two points $x_0(x_t) \in \mathcal{B}_{i(x_t)}$ and $\omega \in \mathcal{G}_j$ such that $x_t = \sqrt{\overline{\alpha}_t}x_0(x_t) + \sqrt{1-\overline{\alpha}_t}\omega$. We learn from (94) that,

$$-\zeta_t(x_t, x_0) \le 66C_1\frac{1-\alpha_t}{1-\overline{\alpha}_t}2^j\theta k\log T + \frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\overline{\alpha}_t\|x_0 - x_0(x_t)\|_2^2. \tag{98}$$

This implies that for any $x_t \in \mathcal{E}_{t,j}$, we have

$$-\frac{(1-\overline{\alpha}_t)\|u_t - \sqrt{\overline{\alpha}_t}x_0\|^2}{2(\alpha_t - \overline{\alpha}_t)^2} \overset{\text{(i)}}{=} -\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\mathsf{Tr}\left(I - J_t(x_t)\right) - \zeta_t(x_t, x_0) \tag{99}$$
$$+ O\Big(\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 |\mathsf{Tr}(I - J_t(x_t))|\Big)$$
$$\overset{\text{(ii)}}{\leq} -\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0 - x_0(x_t)\|_2^2$$
$$- \frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\mathsf{Tr}\left(I - J_t(x_t)\right) + 66C_1\frac{1-\alpha_t}{1-\overline{\alpha}_t}2^j\theta k \log T$$
$$\overset{\text{(iii)}}{\leq} -\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0 - x_0(x_t)\|_2^2$$
$$- \log\det\Big(I + \frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}(I - J_t(x_t))\Big) + 530c_1C_1\frac{\log^2 T}{T}2^j\theta k.$$

Here step (i) follows from (87); step (ii) follows from (98) and Lemma 8, and holds provided that T is sufficiently large; step (iii) uses the relation $\log(1+x) \leq x$ for any $x \geq 0$ and $I - J_t(x_t) \succeq 0$. Therefore for any $j \geq 2$, we have

$$I_j := \int_{x_0}\int_{x_t \in \mathcal{E}_{t,j}} p_{X_0}(x_0)\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t - \overline{\alpha}_t)^2}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \overline{\alpha}_t)^2}\Big)\mathrm{d}x_0\mathrm{d}u_t$$
$$\overset{\text{(a)}}{=} \int_{x_0}\int_{x_t \in \mathcal{E}_{t,j}} p_{X_0}(x_0)\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t - \overline{\alpha}_t)^2}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \overline{\alpha}_t)^2}\Big)$$
$$\cdot \det\Big(I - \frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\mathrm{d}x_0\mathrm{d}x_t$$
$$\overset{\text{(b)}}{\leq} \int_{x_t \in \mathcal{E}_{t,j}}\int_{x_0} \exp\Big(-\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0 - x_0(x_t)\|_2^2\Big)$$
$$\cdot p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\mathrm{d}x_0\mathrm{d}x_t \cdot \exp\Big(530c_1C_1\frac{\log^2 T}{T}2^j\theta k\Big). \tag{100}$$

Here step (a) follows from the relation $u_t = x_t + (1-\alpha_t)s_t^\star(x_t)$; step (b) utilizes (99) and (88). Recall the defintion (30) and let

$$\mathcal{X}_j(x_t) = \bigcup_{i \in \mathcal{I}(x_t; C_1 2^j\theta)} \mathcal{B}_i \quad \text{and} \quad \mathcal{Y}_j(x_t) = \bigcup_{i \notin \mathcal{I}(x_t; C_1 2^j\theta)} \mathcal{B}_i.$$

Then we have

$$\int_{x_0 \in \mathcal{X}_j(x_t)} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0 - x_0(x_t)\|_2^2\Big)\mathrm{d}x_0$$
$$\overset{\text{(i)}}{=} p_{X_t}(x_t)\int_{x_0 \in \mathcal{X}_j(x_t)} p_{X_0|X_t}(x_0 \,|\, x_t)\exp\Big(\frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0 - x_0(x_t)\|_2^2\Big)\mathrm{d}x_0$$
$$\overset{\text{(ii)}}{\leq} p_{X_t}(x_t)\int_{x_0 \in \mathcal{X}_j(x_t)} p_{X_0|X_t}(x_0 \,|\, x_t)\exp\Big(4\frac{1-\alpha_t}{1-\overline{\alpha}_t}C_1 2^j\theta k \log T\Big)\mathrm{d}x_0$$
$$\overset{\text{(ii)}}{\leq} \exp\Big(32c_1C_1\frac{\log^2 T}{T}2^j\theta k\Big)p_{X_t}(x_t). \tag{101}$$

Here step (i) uses the following relation

$$p_{X_t}(x_t) = \int_{x_0} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0; \tag{102}$$

step (ii) follows from a direct consequence of $x_0 \in \mathcal{X}_j(x_t)$ and (74):

$$\|x_0 - x_0(x_t)\|_2 \le \sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}C_1 2^j\theta k\log T} + 2\varepsilon \le 2\sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}C_1 2^j\theta k\log T}.$$

In addition, we also have

$$\begin{aligned}
&\int_{x_0\in\mathcal{Y}_j(x_t)} p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0-x_0(x_t)\|_2^2\Big)\mathrm{d}x_0\\
&\quad\overset{\text{(a)}}{=} p_{X_t}(x_t)\int_{x_0\in\mathcal{Y}_j(x_t)} p_{X_0|X_t}(x_0\,|\,x_t)\exp\Big(\frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0-x_0(x_t)\|_2^2\Big)\mathrm{d}x_0\\
&\quad\le p_{X_t}(x_t)\sum_{i\notin\mathcal{I}(x_t;C_1 2^j\theta)}\mathbb{P}\left(X_0\in\mathcal{B}_i\,|\,X_t=x_t\right)\exp\Big(\sup_{x_0\in\mathcal{B}_i}\frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0-x_0(x_t)\|_2^2\Big)\\
&\quad\overset{\text{(b)}}{\le} p_{X_t}(x_t)\sum_{i\notin\mathcal{I}(x_t;C_1 2^j\theta)}\exp\Big(-\frac{\overline{\alpha}_t}{32\,(1-\overline{\alpha}_t)}\|x_{i(x_t)}^\star - x_i^\star\|_2^2\Big)\mathbb{P}\left(X_0\in\mathcal{B}_i\right)\\
&\quad\overset{\text{(c)}}{\le} \exp\Big(-\frac{1}{32}C_1 2^j\theta k\log T\Big)p_{X_t}(x_t).
\end{aligned}\tag{103}$$

Here step (a) uses the (102); step (b) follows from Lemma 7 and the following relation:

$$\begin{aligned}
\sup_{x_0\in\mathcal{B}_i}\frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0-x_0(x_t)\|_2^2 &\overset{\text{(i)}}{\le} \frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}(\|x_{i(x_t)}^\star - x_i^\star\|_2 + 2\varepsilon)^2\\
&\overset{\text{(ii)}}{\le} \frac{\overline{\alpha}_t}{32\,(1-\overline{\alpha}_t)}\|x_{i(x_t)}^\star - x_i^\star\|_2^2,
\end{aligned}$$

where step (i) uses the relation (74) and step (ii) follows from Lemma 18 and (27), and holds provided that T is sufficiently large; step (c) follows from the definition of $\mathcal{I}(x_t;C_1 2^j\theta)$. Taking (100), (101) and (103) collectively leads to

$$\begin{aligned}
I_j &= \exp\left(530c_1C_1\frac{\log^2 T}{T}2^j\theta k\right)\int_{x_t\in\mathcal{E}_{t,j}}\left(\int_{x_0\in\mathcal{X}_j(x_t)}+\int_{x_0\in\mathcal{Y}_j(x_t)}\right)p_{X_0}(x_0)\big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\\
&\qquad\cdot\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\|x_0-x_0(x_t)\|_2^2\Big)\mathrm{d}x_0\mathrm{d}x_t\\
&\le \exp\left(562c_1C_1\frac{\log^2 T}{T}2^j\theta k\right)\int_{x_t\in\mathcal{E}_{t,j}}p_{X_t}(x_t)\mathrm{d}x_t\\
&\overset{\text{(a)}}{\le}\exp\left(562c_1C_1\frac{\log^2 T}{T}2^j\theta k - \frac{1}{4}2^{j-1}\theta k\log T\right)\overset{\text{(b)}}{\le}\exp\left(-\frac{1}{8}2^{j-1}\theta k\log T\right).
\end{aligned}$$

Here step (a) follows from the relation (97) that we have already proved (by replacing θ with $2^{j-1}\theta$, since $\mathcal{E}_{t,j} \in \mathcal{L}_{t,j}^{\mathrm{c}}$); step (b) holds provided that $T \gg c_1 C_1 \log T$. Hence we have

$$\begin{aligned}
&\int_{x_0}\int_{x_t \notin \mathcal{E}_{t,1}} p_{X_0}(x_0)\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t-\overline{\alpha}_t)^2}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0\mathrm{d}u_t \\
&\quad \overset{\text{(i)}}{=} \sum_{j=2}^{\infty}\int_{x_0}\int_{x_t \in \mathcal{E}_{t,j}} p_{X_0}(x_0)\Big(\frac{1-\overline{\alpha}_t}{2\pi(\alpha_t-\overline{\alpha}_t)^2}\Big)^{d/2}\exp\Big(-\frac{(1-\overline{\alpha}_t)\|u_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_t-\overline{\alpha}_t)^2}\Big)\mathrm{d}x_0\mathrm{d}u_t \\
&\quad = \sum_{j=2}^{\infty} I_j \le \sum_{j=2}^{\infty}\exp\left(-\frac{1}{8}2^{j-1}\theta k\log T\right) \le T^{-4}
\end{aligned}$$

provided that $\theta > 0$ is sufficiently large.

Appendix C. Technical lemmas

In this section, we gather a couple of useful technical lemmas.

Lemma 18 *When T is sufficiently large, for $1 \le t \le T$, we have*

$$\alpha_t \ge 1 - \frac{c_1 \log T}{T} \ge \frac{1}{2}.$$

For $2 \le t \le T$, we have

$$\frac{1-\alpha_t}{1-\overline{\alpha}_t} \le \frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t} \le \frac{8c_1\log T}{T}.$$

In addition, we have

$$\overline{\alpha}_T \le T^{-c_1/2}.$$

Proof See Li et al. (2023, Appendix A.2). ■

Lemma 19 *For $Z \sim \mathcal{N}(0,1)$ and any $t \ge 1$, we know that*

$$\mathbb{P}\left(|Z| \ge t\right) \le e^{-t^2/2}, \qquad \forall\, t \ge 1.$$

In addition, for a chi-square random variable $Y \sim \chi^2(d)$, we have

$$\mathbb{P}(\sqrt{Y} \ge \sqrt{d} + t) \le e^{-t^2/2}, \qquad \forall\, t \ge 1.$$

Proof See Vershynin (2018, Proposition 2.1.2) and Laurent and Massart (2000, Section 4.1). ■

Lemma 20 *Suppose that Assumption 1 holds, and that T and c_2 are sufficiently large. Then we have*

$$\mathsf{TV}\big(p_{X_T}\|p_{Y_T}\big) \le T^{-99}.$$

Proof Define a random variable $X_0^- := X_0 \mathbb{1}\{\|X_0\|_2 \le T^{c_M+100}\}$ by truncating X_0. Let

$$X_T^- = \sqrt{\overline{\alpha}_T} X_0^- + \sqrt{1-\overline{\alpha}_T} Z,$$

where $Z \sim \mathcal{N}(0, I_d)$ is independent of X_0^-. Notice that X_0^- has bounded support, which allows us to invoke (Li et al., 2023, Lemma 3) to achieve

$$\mathsf{TV}(p_{\overline{X}_T}, p_{Y_T}) = O(T^{-100}), \tag{104}$$

provided that c_2 and T are sufficiently large. In addition, we have

$$\begin{aligned}
\mathsf{TV}(p_{\overline{X}_T}, p_{X_T}) &= \frac{1}{2}\int |p_{\overline{X}_T}(x) - p_{X_T}(x)| \mathrm{d}x \\
&= \frac{1}{2}\int_x \Big| \int_{x_0} \big(p_{\overline{X}_0}(x_0) - p_{X_0}(x_0)\big)\big(2\pi(1-\overline{\alpha}_T)\big)^{-d/2} \exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_T}x_0\|_2^2}{2(1-\overline{\alpha}_T)}\Big)\mathrm{d}x_0\Big|\mathrm{d}x \\
&\le \frac{1}{2}\int_x \int_{x_0} \big|p_{\overline{X}_0}(x_0) - p_{X_0}(x_0)\big|\big(2\pi(1-\overline{\alpha}_T)\big)^{-d/2} \exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_T}x_0\|_2^2}{2(1-\overline{\alpha}_T)}\Big)\mathrm{d}x_0\mathrm{d}x \\
&\overset{\text{(i)}}{=} \frac{1}{2}\int_{x_0} \big|p_{\overline{X}_0}(x_0) - p_{X_0}(x_0)\big|\mathrm{d}x_0 = \mathsf{TV}(p_{\overline{X}_0}, p_{X_0}) = \mathbb{P}\big(\|X_0\|_2 > T^{c_M+100}\big) \\
&\overset{\text{(ii)}}{\le} \frac{\mathbb{E}[\|X_0\|_2]}{T^{c_M+100}} = T^{-100}.
\end{aligned} \tag{105}$$

Here step (i) invokes Tonelli's theorem, while step (ii) follows from Markov's inequality. Taking (104) and (105) collectively yields the desired result, provided that T is sufficiently large. ■

Lemma 21 *Suppose that Assumption 1 holds, and that $T \gg d \log T$. Then we have*

$$\sum_{t=2}^{T} \frac{1-\alpha_t}{1-\overline{\alpha}_t} \mathsf{Tr}\Big(\mathbb{E}\big[\big(\Sigma_{\overline{\alpha}_t}(X_t)\big)^2\big]\Big) \le C_J d \log T \tag{106}$$

for some universal constant $C_J > 0$. Here the matrix function $\Sigma_{\overline{\alpha}_t}(\cdot)$ is defined as

$$\Sigma_{\overline{\alpha}_t}(x) := \mathsf{Cov}\big(Z \mid \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} Z = x\big), \tag{107}$$

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

Proof The result (106) was established in Li et al. (2024b, Lemma 2) under the stronger assumption that

$$\mathbb{P}(\|X_0\|_2 < T^{c_R}) = 1 \tag{108}$$

for some universal constant $c_R > 0$. The assumption (108) is used to prove part (a) of their Lemma 2, which states that for any $\overline{\alpha}', \overline{\alpha} \in [\overline{\alpha}_t, \overline{\alpha}_{t-1}]$ with $1 \le t \le T$, one has

$$\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}'}\big(\sqrt{\overline{\alpha}'}X_0 + \sqrt{1-\overline{\alpha}'}Z\big)\Big)^2\Big] \preceq c_1' \mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\Big)^2\Big] + c_1' \exp(-c_2' d \log T) I_d.$$

for some universal constants $c_1', c_2' > 0$. Through a similar truncation argument as in the proof of Lemma 20, we can show that

$$\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}'}\big(\sqrt{\overline{\alpha}'}X_0 + \sqrt{1-\overline{\alpha}'}Z\big)\Big)^2\Big] \preceq c_1'\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\Big)^2\Big] + c_1' T^{-100} I_d.$$

Armed with this result, we can use the same analysis for proving part (b) of Li et al. (2024b, Lemma 2) to establish (106) under our Assumption 1. The details are omitted here for simplicity. ∎

Lemma 22 *Let k be the intrinsic dimension (cf. Definition 2) of the support of p_{data}, and suppose that $T \gg k \log T$. Then we have*

$$\sum_{t=2}^{T} \frac{1-\alpha_t}{1-\overline{\alpha}_t} \mathsf{Tr}\Big(\mathbb{E}\big[\big(\Sigma_{\overline{\alpha}_t}(X_t)\big)^2\big]\Big) \le C_J k \log T \tag{109}$$

for some universal constant $C_J > 0$. Here the matrix function $\Sigma_{\overline{\alpha}_t}(\cdot)$ is defined in (107).

Proof This lemma can be proved by modifying the first part of the proof of Li et al. (2024b, Lemma 2), and we describe these modification as follows. For any $\overline{\alpha}, \overline{\alpha}' \in (0,1)$, define

$$X_{\overline{\alpha}} = \sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z \qquad \text{and} \qquad X_{\overline{\alpha}'} = \sqrt{\overline{\alpha}'}X_0 + \sqrt{1-\overline{\alpha}'}Z,$$

where $X_0 \sim p_{\mathsf{data}}$ and $Z \sim \mathcal{N}(0, I_d)$ are independently random variables. Li et al. (2024b, Lemma 2, Part (a)) demonstrated that as long as $T \gg d \log T$, for any

$$\frac{|\overline{\alpha}' - \overline{\alpha}|}{\overline{\alpha}(1-\overline{\alpha})} = O\Big(\frac{1}{d \log T}\Big)$$

and any pair (x, x') where x is in a certain typical set (see Equation (79) therein) and $x = \sqrt{\overline{\alpha}/\overline{\alpha}'}x'$, it holds that $p_{X_{\overline{\alpha}'}}(x') \asymp p_{X_{\overline{\alpha}}}(x)$; see Equation (81) therein. However here we only assume that $T \gg k \log T$, and we want such a result to hold for any

$$\frac{|\overline{\alpha}' - \overline{\alpha}|}{\overline{\alpha}(1-\overline{\alpha})} = O\Big(\frac{1}{k \log T}\Big) \tag{110}$$

in order to improve the dimension factor d in Li et al. (2024b, Lemma 2, Part (b)) to the intrinsic dimension k. To this end, we instead consider any pair (x, x') where

$$x = h(x') \coloneqq \sqrt{\overline{\alpha}/\overline{\alpha}'}x' + \Big(\sqrt{\overline{\alpha}/\overline{\alpha}'}(1-\overline{\alpha}') - \sqrt{(1-\overline{\alpha})(1-\overline{\alpha}')}\Big)s^{\star}_{\overline{\alpha}'}(x'), \tag{111}$$

Here $s^{\star}_{\overline{\alpha}'}(\cdot)$ is the score function of $X_{\overline{\alpha}'}$, namely

$$s^{\star}_{\overline{\alpha}'}(x') = -\frac{1}{1-\overline{\alpha}_t}\int p_{X_0 | X_{\overline{\alpha}'}}(x_0 \,|\, x')\big(x' - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0.$$

Let $\mathcal{E}_{\overline{\alpha},1}$ be the typical set of $X_{\overline{\alpha}}$ defined as replacing the $\overline{\alpha}_t$ in (28) with $\overline{\alpha}$. Following similar analysis as in Lemmas 14 and 15, we can show that

$$p_{X_{\overline{\alpha}'}}(x')\mathrm{d}x' \asymp p_{X_{\overline{\alpha}}}(x)\mathrm{d}x, \quad \text{i.e.,} \quad p_{X_{\overline{\alpha}'}}(x') \asymp p_{X_{\overline{\alpha}}}(x)|\det J_h(x')|,$$

holds for any $x \in \mathcal{E}_{\overline{\alpha}}$, where J_h is the Jacobian matrix of h (see (111)). Equipped with this relation, we can follow the steps in the proof of Li et al. (2024b, Lemma 2, Part (a)) to show that

$$p_{X_0|X_{\overline{\alpha}'}}(x_0 \,|\, x') \asymp p_{X_0|X_{\overline{\alpha}}}(x_0 \,|\, x),$$

which corresponds to Equation (82) therein, and this further leads to

$$\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}'}\big(\sqrt{\overline{\alpha}'}X_0 + \sqrt{1-\overline{\alpha}'}Z\big)\Big)^2\Big] \preceq C_3'^2\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\Big)^2\Big] + C_8' \exp(-C_9' k \log T) I_d$$

holds for all $\overline{\alpha}, \overline{\alpha}' \in (0,1)$ satisfying (110). Using the above result, we can follow the same proof as in Li et al. (2024b, Lemma 2, Part (b)) to establish the desired result. The detailed proof is omitted here for brevity. ■

References

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

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023a.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023b.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

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

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023b.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36:68552–68575, 2023c.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023d.

Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023e.

Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspaces in diffusion models for controllable image editing. *arXiv preprint arXiv:2409.02374*, 2024.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

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

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow ODEs of diffusion models in wasserstein distances. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Shivam Gupta, Linda Cai, and Sitan Chen. Faster diffusion-based sampling with randomized midpoints: Sequential and parallel. *arXiv preprint arXiv:2406.00924*, 2024.

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

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

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*, 2024.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

Nishant Jain and Tong Zhang. A sharp kl-convergence analysis for diffusion models under minimal assumptions. *arXiv preprint arXiv:2508.16306*, 2025.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985, 2023.

Gen Li and Changxiao Cai. Provable acceleration for diffusion models under minimal assumptions. *arXiv preprint arXiv:2410.23285*, 2024.

Gen Li and Yuchen Jiao. Improved convergence rate for diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Gen Li and Yuling Yan. O(d/t) convergence theory for diffusion probabilistic models under minimal assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. In *International Conference on Machine Learning*, pages 27942–27954. PMLR, 2024a.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024b.

Gen Li, Yuchen Zhou, Yuting Wei, and Yuxin Chen. Faster diffusion models via higher-order approximation. *arXiv preprint arXiv:2506.24042*, 2025.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*, 2025.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*, 2024.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

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

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

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

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021b.

Jiaqi Tang and Yuling Yan. Adaptivity and convergence of probability flow odes in diffusion generative models. *arXiv preprint arXiv:2501.18863*, 2025.

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

Rong Tang, Lizhen Lin, and Yun Yang. Conditional diffusion models are minimax-optimal and manifold-adaptive for conditional distribution estimation. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

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

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022.

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

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*, 2024.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.