Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models

Gen Li* Yuting Wei* Yuxin Chen*† Yuejie Chi‡

July 11, 2023

Abstract

Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain far from mature. In this work, we develop a suite of non-asymptotic theory towards understanding the data generation process of diffusion models in discrete time, assuming access to reliable estimates of the (Stein) score functions. For a popular deterministic sampler (based on the probability flow ODE), we establish a convergence rate proportional to $1/T$ (with T the total number of steps), improving upon past results; for another mainstream stochastic sampler (i.e., a type of the denoising diffusion probabilistic model (DDPM)), we derive a convergence rate proportional to $1/\sqrt{T}$, matching the state-of-the-art theory. Our theory imposes only minimal assumptions on the target data distribution (e.g., no smoothness assumption is imposed), and is developed based on an elementary yet versatile non-asymptotic approach without resorting to toolboxes for SDEs and ODEs. Further, we design two accelerated variants, improving the convergence to $1/T^2$ for the ODE-based sampler and $1/T$ for the DDPM-type sampler, which might be of independent theoretical and empirical interest.

Keywords: diffusion models, score-based generative modeling, non-asymptotic theory, reverse SDE, probability flow ODE, denoising diffusion probabilistic model

Contents

*Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

†Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.

‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

1 Introduction

Diffusion models have emerged as a cornerstone in contemporary generative modeling, a task that learns to generate new data instances (e.g., images, text, audio) that look similar in distribution to the training data (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Dhariwal and Nichol, 2021; Jolicoeur-Martineau et al., 2021; Chen et al., 2021; Kong et al., 2021; Austin et al., 2021). Originally proposed by Sohl-Dickstein et al. (2015) and later popularized by Song and Ermon (2019); Ho et al. (2020), the mainstream diffusion generative models — e.g., denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) and denoising diffusion implicit models (DDIMs) (Song et al., 2020a) — have underpinned major successes in content generators like DALL·E 2 (Ramesh et al., 2022), Stable Diffusion (Rombach et al., 2022) and Imagen (Saharia et al., 2022), claiming state-of-the-art performance in the now broad field of generative artificial intelligence (AI). See Yang et al. (2022); Croitoru et al. (2023) for overviews of recent development.

In a nutshell, a diffusion generative model is based upon two stochastic processes in $\mathbb{R}^d$:

1) a forward process
$$X_0 \to X_1 \to \cdots \to X_T \tag{1}$$
that starts from a sample drawn from the target data distribution (e.g., of natural images) and gradually diffuses it into a noise-like distribution (e.g., standard Gaussians);

2) a reverse process
$$Y_T \to Y_{T-1} \to \cdots \to Y_0 \tag{2}$$
that starts from pure noise (e.g., standard Gaussians) and successively converts it into new samples sharing similar distributions as the target data distribution.

Transforming data into noise in the forward process is straightforward, often hand-crafted by increasingly injecting more noise into the data at hand. What is challenging is the construction of the reverse process: how to generate the desired information out of pure noise? To do so, a diffusion model learns to build a reverse process (2) that imitates the dynamics of the forward process (1) in a time-reverse fashion; more precisely, the design goal is to ascertain distributional proximity[1]

$$Y_t \overset{\mathrm{d}}{\approx} X_t, \qquad t = T, \cdots, 1 \tag{3}$$

through proper learning based on how the training data propagate in the forward process. Encouragingly, there often exist feasible strategies to achieve this goal as long as faithful estimates about the (Stein) score functions — the gradients of the log marginal density of the forward process — are available, an intriguing fact that can be illuminated by the existence and construction of reverse-time stochastic differential equations (SDEs) (Anderson, 1982) (see Section 2.2 for more precise discussions). Viewed in this light, a diverse array of diffusion models are frequently referred to as *score-based generative modeling (SGM)*. The popularity of SGM was initially motivated by, and has since further inspired, numerous recent studies on the problem of learning score functions, a subroutine that also goes by the name of score matching (e.g., Hyvärinen (2005, 2007); Vincent (2011); Song et al. (2020b)).

Nonetheless, despite the mind-blowing empirical advances, a mathematical theory for diffusion generative models is still in its infancy. Given the complexity of developing a full-fledged end-to-end theory, a divide-and-conquer approach has been advertised, decoupling the score learning phase (i.e., how to estimate score functions reliably from training data) and the generative sampling phase (i.e., how to generate new data instances given the score estimates). In particular, the past two years have witnessed growing interest and remarkable progress from the theoretical community towards understanding the generative sampling phase (Block et al., 2020; De Bortoli et al., 2021; Liu et al., 2022; De Bortoli, 2022; Lee et al., 2023; Pidstrigach, 2022; Chen et al., 2022b,a, 2023c). For instance, polynomial-time convergence guarantees have been established for stochastic samplers (e.g., Chen et al. (2022b,a)) and deterministic samplers (e.g., Chen et al. (2023c)), both of which accommodated a fairly general family of data distributions.

This paper. The present paper contributes to this growing list of theoretical endeavors by developing a new suite of non-asymptotic theory for several score-based generative modeling algorithms. We concentrate on two types of samplers (Song et al., 2021b) in discrete time: (i) a deterministic sampler based on a sort of ordinary differential equations (ODEs) called probability flow ODEs (which is closely related to the DDIM); and (ii) a DDPM-type stochastic sampler motivated by reverse-time SDEs. Only minimal assumptions are imposed on the target data distribution (e.g., no smoothness condition is needed). In comparisons to past works, our main contributions are three-fold.

- *Non-asymptotic convergence guarantees.* For a popular deterministic sampler, we demonstrate that the number of steps needed to yield ε-accuracy — meaning that the total variance (TV) distance between the distribution of X_1 and that of Y_1 is no larger than ε — is proportional to $1/\varepsilon$ (in addition to other polynomial dimension dependency), which improves upon the convergence guarantees in prior results Chen et al. (2023c) and does not rely on any sort of smoothness assumption. For another DDPM-type stochastic sampler, we establish an iteration complexity proportional to $1/\varepsilon^2$, matching existing theory Chen et al. (2022b,a) in terms of the ε-dependency.

- *Accelerating data generation processes.* In order to further speed up the sampling processes, we develop an accelerated variant for each of the above two samplers, taking advantage of estimates of a small number of additional quantities. As it turns out, these variants achieve more rapid convergence, with the deterministic (resp. stochastic) variant exhibiting a $1/\sqrt{\varepsilon}$ (resp. $1/\varepsilon$) scaling in the accuracy level ε (again measured in terms of the TV distance).

- *An elementary non-asymptotic analysis framework.* From the technical point of view, the analysis framework laid out in this paper is fully non-asymptotic in nature. In contrast to prior theoretical analyses that take a detour to study the continuum limits and then control the discretization error, our approach tackles the discrete-time processes directly using elementary analysis strategies. No knowledge of SDEs or ODEs is required for establishing our theory, thereby resulting in a more versatile framework and sometimes lowering the technical barrier towards understanding diffusion models.

[1]Two random vectors X and Y are said to obey $X \overset{\mathrm{d}}{=} Y$ (resp. $X \overset{\mathrm{d}}{\approx} Y$) if they are equivalent (resp. close) in distribution.

Notation. Before proceeding, we introduce a couple of notation to be used throughout. For any two functions $f(d,T)$ and $g(d,T)$, we adopt the notation $f(d,T) \lesssim g(d,T)$ or $f(d,T) = O(g(d,T))$ (resp. $f(d,T) \gtrsim g(d,T)$) to mean that there exists some universal constant $C_1 > 0$ such that $f(d,T) \leq C_1 g(d,T)$ (resp. $f(d,T) \geq C_1 g(d,T)$) for all d and T; moreover, the notation $f(d,T) \asymp g(d,T)$ indicates that $f(d,T) \lesssim g(d,T)$ and $f(d,T) \gtrsim g(d,T)$ hold at once. The notation $\widetilde{O}(\cdot)$ is defined similar to $O(\cdot)$ except that it hides the logarithmic dependency. Additionally, the notation $f(d,T) = o\big(g(d,T)\big)$ means that $f(d,T)/g(d,T) \to 0$ as d, T tend to infinity. We shall often use capital letters to denote random variables/vectors/processes, and lowercase letters for deterministic variables. For any two probability measures P and Q, the total variation (TV) distance between them is defined to be $\mathsf{TV}(P,Q) \coloneqq \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$. Throughout the paper, $p_X(\cdot)$ (resp. $p_{X\,|\,Y}(\cdot\,|\,\cdot)$) denotes the probability density function of X (resp. X given Y). For any matrix A, we denote by $\|A\|$ (resp. $\|A\|_{\mathrm{F}}$) the spectral norm (resp. Frobenius norm) of A.

2 Preliminaries

In this section, we introduce the basics of diffusion generative models. The ultimate goal of a generative model can be concisely stated: given data samples drawn from an unknown distribution of interest p_{data} in $\mathbb{R}^d$, we wish to generate new samples whose distributions closely resemble p_{data}.

2.1 Diffusion generative models

Towards achieving the above goal, a diffusion generative model typically encompasses two Markov processes: a forward process and a reverse process, as described below.

The forward process. In the forward chain, one progressively injects noise into the data samples to diffuse and obscure the data. The distributions of the injected noise are often hand-picked, with the standard Gaussian distribution receiving widespread adoption. More specifically, the forward Markov process produces a sequence of d-dimensional random vectors $X_1 \to X_2 \to \cdots \to X_T$ as follows:

$$X_0 \sim p_{\mathsf{data}}, \tag{4a}$$

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\,W_t, \qquad 1 \leq t \leq T, \tag{4b}$$

where $\{W_t\}_{1\leq t\leq T}$ indicates a sequence of independent noise vectors drawn from $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. The hyper-parameters $\{\beta_t \in (0,1)\}$ represent prescribed learning rate schedules that control the variance of the noise injected in each step. If we define

$$\alpha_t \coloneqq 1 - \beta_t, \qquad \overline{\alpha}_t \coloneqq \prod_{k=1}^{t} \alpha_k, \qquad 1 \leq t \leq T, \tag{5}$$

then it can be straightforwardly verified that for every $1 \leq t \leq T$,

$$X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t \qquad \text{for some } \overline{W}_t \sim \mathcal{N}(0, I_d). \tag{6}$$

Clearly, if the covariance of X_0 is also equal to I_d, then the covariance of X_t is preserved throughout the forward process; for this reason, this forward process (4) is sometimes referred to as *variance-preserving* (Song et al., 2021b). Throughout this paper, we employ the notation

$$q_t \coloneqq \mathsf{law}\big(X_t\big) \tag{7}$$

to denote the distribution of X_t. As long as $\overline{\alpha}_T$ is vanishingly small, one has the following property for a general family of data distributions:

$$q_T \approx \mathcal{N}(0, I_d). \tag{8}$$

The reverse process. The reverse chain $Y_T \to Y_{T-1} \to \ldots \to Y_1$ is designed to (approximately) revert the forward process, allowing one to transform pure noise into new samples with matching distributions as the original data. To be more precise, by initializing it as

$$Y_T \sim \mathcal{N}(0, I_d), \tag{9a}$$

we seek to design a reverse-time Markov process with nearly identical marginals as the forward process, namely,

$$\text{(goal)} \qquad Y_t \overset{\mathrm{d}}{\approx} X_t, \qquad t = T, T-1, \cdots, 1. \tag{9b}$$

Throughout the paper, we shall often employ the following notation to indicate the distribution of Y_t:

$$p_t \coloneqq \mathsf{law}\big(Y_t\big). \tag{10}$$

2.2 Deterministic vs. stochastic samplers: a continuous-time interpretation

Evidently, the most crucial step of the diffusion model lies in effective design of the reverse process. Two mainstream approaches stand out:

- *Deterministic samplers.* Starting from $Y_T \sim \mathcal{N}(0, I_d)$, this approach selects a set of functions $\{\Phi_t(\cdot)\}_{1\leq t\leq T}$ and computes:

$$Y_{t-1} = \Phi_t\big(Y_t\big), \qquad t = T, \cdots, 1. \tag{11}$$

  Clearly, the sampling process is fully deterministic except for the initialization Y_T.

- *Stochastic samplers.* Initialized again at $Y_T \sim \mathcal{N}(0, I_d)$, this approach computes another collection of functions $\{\Psi_t(\cdot,\cdot)\}_{1\leq t\leq T}$ and performs the updates:

$$Y_{t-1} = \Psi_t\big(Y_t, Z_t\big), \qquad t = T, \cdots, 1, \tag{12}$$

  where the Z_t's are independent noise vectors obeying $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$.

In order to elucidate the feasibility of the above two approaches, we find it helpful to look at the continuum limit through the lens of SDEs and ODEs. It is worth emphasizing, however, that the development of our main theory does *not* rely on any knowledge of SDEs and ODEs.

- *The forward process.* A continuous-time analog of the forward diffusion process can be modeled as

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t \quad (0 \leq t \leq T), \qquad X_0 \sim p_{\mathsf{data}} \tag{13}$$

  for some functions $f(\cdot,\cdot)$ and $g(\cdot)$ (denoting respectively the drift and diffusion coefficient), where W_t denotes a d-dimensional standard Brownian motion. As a special example, the continuum limit of (4) takes the following form (Song et al., 2021b)

$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}W_t \quad (0 \leq t \leq T), \qquad X_0 \sim p_{\mathsf{data}} \tag{14}$$

  for some function $\beta(t)$. As before, we denote by q_t the distribution of X_t in (13).

- *The reverse process.* As it turns out, the following two reverse processes are both capable of reconstructing the distribution of the forward process, motivating the design of two distinctive samplers. Here and throughout, we use $\nabla \log q_t(X)$ to abbreviate $\nabla_X \log q_t(X)$ for notational simplicity.

  - One feasible approach is to resort to the so-called *probability flow ODE* (Song et al., 2021b)

$$\mathrm{d}Y_t^{\mathsf{ode}} = \Big(f\big(Y_t^{\mathsf{ode}}, T-t\big) - \frac{1}{2}g(T-t)^2\nabla \log q_{T-t}\big(Y_t^{\mathsf{ode}}\big)\Big)\mathrm{d}t \quad (0 \leq t \leq T), \qquad Y_0^{\mathsf{ode}} \sim q_T, \tag{15}$$

    which exhibits matching distributions as follows:

$$Y_{T-t}^{\mathsf{ode}} \overset{\mathrm{d}}{=} X_t, \qquad 0 \leq t \leq T.$$

    The deterministic nature of this approach often enables faster sampling. It has been shown that this family of deterministic samplers is closely related to the DDIM sampler (Karras et al., 2022; Song et al., 2021b).

– In view of the seminal result by Anderson (1982), one can also construct a "reverse-time" SDE

$$\mathrm{d}Y_t^{\mathsf{sde}} = \Big(f\big(Y_t^{\mathsf{sde}}, T-t\big) - g(T-t)^2 \nabla \log q_{T-t}\big(Y_t^{\mathsf{sde}}\big)\Big)\mathrm{d}t + g(T-t)\mathrm{d}Z_t^{\mathsf{sde}} \quad (0 \le t \le T) \tag{16}$$

with $Y_0^{\mathsf{sde}} \sim q_T$ and Z_t^{sde} being a standard Brownian motion. Strikingly, this process also satisfies

$$Y_{T-t}^{\mathsf{sde}} \overset{\mathrm{d}}{=} X_t, \qquad 0 \le t \le T.$$

The popular DDPM sampler (Ho et al., 2020; Nichol and Dhariwal, 2021) falls under this category.

Interestingly, in addition to the functions f and g that define the forward process, construction of both (15) and (16) relies only upon the knowledge of the gradient of the log density $\nabla \log q_t(\cdot)$ of the intermediate steps of the forward diffusion process — often referred to as the (Stein) score function. Consequently, a key enabler of the above paradigms lies in reliable learning of the score function, and hence the name *score-based generative modeling*.

3 Algorithms and main results

In this section, we analyze a couple of diffusion generative models, including both deterministic and stochastic samplers. While the proofs for our main theory are all postponed to the appendix, it is worth emphasizing upfront that our analysis framework directly tackles the discrete-time processes without resorting to any toolbox of SDEs and ODEs tailored to the continuous-time limits. This elementary approach might potentially be versatile for analyzing a broad class of variations of these samplers.

3.1 Assumptions and learning rates

Before proceeding, we impose some assumptions on the score estimates and the target data distributions, and specify the hypter-parameters $\{\alpha_t\}$, which shall be adopted throughout all cases.

Score estimates. Given that the score functions are an essential component in score-based generative modeling, we assume access to faithful estimates of the score functions $\nabla \log q_t(\cdot)$ across all intermediate steps t, thus disentangling the score learning phase and the data generation phase. This is made precise in the following assumption.

Assumption 1. *Suppose that we have access to the score estimates $s_t : \mathbb{R}^d \to \mathbb{R}^d$ $(1 \le t \le T)$ as follows:*

$$s_t \coloneqq \arg \min_{s:\mathbb{R}^d \to \mathbb{R}^d} \mathop{\mathbb{E}}_{X \sim q_t} \Big[\big\| s(X) - \nabla \log q_t(X) \big\|_2^2\Big], \qquad 1 \le t \le T. \tag{17}$$

As has been pointed out by previous works concerning score matching (e.g., Hyvärinen (2005); Vincent (2011); Chen et al. (2022b)), this score estimate admits an alternative form as follows (owing to properties of Gaussian distributions):

$$s_t \coloneqq \arg \min_{s:\mathbb{R}^d \to \mathbb{R}^d} \mathop{\mathbb{E}}_{W \sim \mathcal{N}(0, I_d), X_0 \sim p_{\mathsf{data}}} \left[\left\| s\big(\sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} W\big) + \frac{1}{\sqrt{1-\overline{\alpha}_t}} W \right\|_2^2\right], \tag{18}$$

which takes the form of the minimum mean square error estimator for $-\frac{1}{\sqrt{1-\overline{\alpha}_t}} W$ given $\sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} W$ and is often more amenable to training.

Target data distributions. Our goal is to uncover the effectiveness of diffusion models in generating a broad family of data distributions. Throughout this paper, the only assumptions we need to impose on the target data distribution p_{data} are the following:

- X_0 is a continuous random vector, and

$$\mathbb{P}\big(\|X_0\|_2 \leq R = T^{c_R} \mid X_0 \sim p_{\mathsf{data}}\big) = 1 \tag{19}$$

for some arbitrarily large constant $c_R > 0$.

This assumption allows the radius of the support of p_{data} to be exceedingly large (given that the exponent c_3 can be arbitrarily large).

Learning rate schedule. Let us also take a moment to specify the learning rates to be used for our theory and analyses. For some large enough numerical constants $c_0, c_1 > 0$, we set

$$\beta_1 = 1 - \alpha_1 = \frac{1}{T^{c_0}}; \tag{20a}$$

$$\beta_t = 1 - \alpha_t = \frac{c_1 \log T}{T} \min\left\{\beta_1\Big(1 + \frac{c_1 \log T}{T}\Big)^t, 1\right\}. \tag{20b}$$

3.2 Deterministic samplers

We begin by analyzing a deterministic sampler: a discrete-time version of the probability flow ODE.

3.2.1 An ODE-based deterministic sampler

Armed with the score estimates in Assumption 1, a discrete-time version of the probability flow ODE approach (cf. (15)) adopts the following update rule:

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \Phi_t\big(Y_t\big) \quad \text{ for } t = T, \cdots, 1, \tag{21a}$$

where $\Phi_t(\cdot)$ is taken to be

$$\Phi_t(x) := \frac{1}{\sqrt{\alpha_t}}\bigg(x + \frac{1-\alpha_t}{2} s_t(x)\bigg). \tag{21b}$$

This approach, based on the probability flow ODE (15), often achieves faster sampling compared to the stochastic counterpart (Song et al., 2021b). Despite the empirical advances, however, the theoretical understanding of this type of deterministic samplers remained far from mature.

We first derive non-asymptotic convergence guarantees — measured by the total variance distance between the forward and the reverse processes — for the above deterministic sampler (21). The proof of this result is postponed to Section 5.2.

Theorem 1. *Suppose that* (19) *holds true. Equipped with the score estimates in Assumption 1 and the learning rate schedule* (20)*, the sampling process* (21) *satisfies*

$$\mathsf{TV}\big(q_1, p_1\big) \leq C_1 \frac{d^2 \log^4 T}{T} + C_1 \frac{d^6 \log^6 T}{T^2} \tag{22}$$

for some universal constants $C_1 > 0$, where we recall that p_1 (resp. q_1) represents the distribution of Y_1 (resp. X_1).

In other words, in order to achieve $\mathsf{TV}(q_1, p_1) \leq \varepsilon$, the number of steps T only needs to exceed

$$\widetilde{O}\bigg(\frac{d^2}{\varepsilon} + \frac{d^3}{\sqrt{\varepsilon}}\bigg). \tag{23}$$

To the best of our knowledge, the only non-asymptotic analysis for the probability flow ODE approach in prior literature was derived by a very recent work Chen et al. (2023c), which established the first non-asymptotic convergence guarantees that exhibit polynomial dependency in both d and $1/\varepsilon$ (see, e.g., Chen et al. (2023c, Theorem 4.1)). However, it fell short of providing concrete polynomial dependency in d and $1/\varepsilon$, and suffered from exponential dependency in the Lipschitz constant of the score function. In contrast, our result in Theorem 1 uncovers a concrete d^2/ε scaling (ignoring lower-order and logarithmic terms) without imposing any smoothness assumption, which was previously unavailable. Finally, while we were wrapping up the current paper, we became aware of the independent work Chen et al. (2023b) establishing improved polynomial dependency for two variants of the probability flow ODE. By inserting an additional stochastic corrector step — based on overdamped (resp. underdamped) Langevin diffusion — in each iteration of the probability flow ODE, Chen et al. (2023b) showed that $\widetilde{O}(L^3 d/\varepsilon^2)$ (resp. $\widetilde{O}(L^2\sqrt{d}/\varepsilon)$) steps are sufficient, where L denotes the Lipschitz constant of the score function. In comparison, our result demonstrates for the first time that the plain probability flow ODE already achieves the $1/\varepsilon$ scaling without requiring either a corrector step or the smoothness assumption; one limitation of our result, however, is the sub-optimal d-dependency compared to Chen et al. (2023b).

3.2.2 An accelerated deterministic sampler

Thus far, we have demonstrated that the iteration complexity of the deterministic sampler (21) is proportional to $1/\varepsilon$ (for small enough ε). A natural question is whether this convergence rate can be further improved.

As it turns out, if we have access to reliable estimates of two additional quantities, then a modified version of the sampler (21) is able to achieve much improved convergence guarantees. These estimates are made precise in the following assumption.

Assumption 2. *Suppose that we have access to the estimates $w_t : \mathbb{R}^d \to \mathbb{R}^d$ $(1 \le t \le T)$ defined as follows:*

$$w_t := \arg\min_{w:\mathbb{R}^d\to\mathbb{R}^d} \mathbb{E}\left[\Big\|\|W\|_2^2\Big(\frac{1}{\sqrt{1-\overline{\alpha}_t}}W + s_t(X_t)\Big) + \frac{1}{1-\overline{\alpha}_t}WW^\top s_t(X_t) - w(X_t)\Big\|_2^2\right], \tag{24}$$

where $s_t(\cdot)$ is defined in Assumption 1, and $X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W$. Here, the expectation is with respect to $W \sim \mathcal{N}(0, I_d)$ and $X_0 \sim p_{\mathsf{data}}$.

Armed with the score estimate in Assumption 1 and the additional estimates in Assumption 2, we are ready to introduce an accelerated variant of (21) as follows:

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \Phi_t(Y_t) \quad \text{for } t = T, \cdots, 1, \tag{25a}$$

where the mapping $\Phi_t(\cdot)$ is chosen to be

$$\Phi_t(x) = \frac{1}{\sqrt{\alpha_t}}\bigg(x + \Big(\frac{1-\alpha_t}{2} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)} - \frac{(1-\alpha_t)^2}{8}\big\|s_t(x)\big\|_2^2\Big)s_t(x) + \frac{(1-\alpha_t)^2}{8}w_t(x)\bigg). \tag{25b}$$

Notably, this new variant (25) is closely related to the original sampler (21); in fact, they both move along the direction specified by the score estimate s_t, except that the accelerated variant includes a proper correction term chosen based on higher-order expansion.

Encouragingly, our non-asymptotic analysis framework can be extended to derive enhanced convergence guarantees for the sampler (25), as stated below. The proof of this result is postponed to Section D.1.

Theorem 2. *Suppose that (19) holds true. Equipped with the estimates in Assumptions 1-2 and the learning rate schedule (20), the sampling process (25) obeys*

$$\mathsf{TV}(q_1, p_1) \le C_1 \frac{d^6 \log^6 T}{T^2} \tag{26}$$

for some universal constants $C_1 > 0$, where p_1 (resp. q_1) is the distribution of Y_1 (resp. X_1).

Theorem 2 reveals that: in order to achieve $\mathsf{TV}(q_1, p_1) \le \varepsilon$, the accelerated deterministic sampler (25) only requires the number of steps T to be on the order of

$$\widetilde{O}\bigg(\frac{d^3}{\sqrt{\varepsilon}}\bigg), \tag{27}$$

thus improving the dependency on ε from $\widetilde{O}(1/\varepsilon)$ (cf. (23)) to $\widetilde{O}(1/\sqrt{\varepsilon})$ for small enough ε. Consequently, the improved convergence result underscores the crucial role of bias correction when selecting the search direction.

3.3 Stochastic samplers

3.3.1 A DDPM-type stochastic sampler

Armed with the score estimates $\{s_t\}$ in Assumption 1, we can readily introduce the following stochastic sampler that operates in discrete time, motivated by the reverse-time SDE (16):

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \Psi_t(Y_t, Z_t) \quad \text{ for } t = T, \cdots, 1 \tag{28a}$$

where $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and

$$\Psi_t(y, z) = \frac{1}{\sqrt{\alpha_t}}\Big(y + (1-\alpha_t)s_t(y)\Big) + \sigma_t z \qquad \text{with } \sigma_t^2 = \frac{1}{\alpha_t} - 1. \tag{28b}$$

The key difference between this sampler and the deterministic sampler (21) is that: (i) there exists an additional pre-factor of $1/2$ on s_t in the deterministic sampler; and (ii) the stochastic sampler injects additional noise Z_t in each step.

In contrast to deterministic samplers, the stochastic samplers have received more theoretical attention, with the state-of-the-art results established by Chen et al. (2022b,a). The elementary approach developed in the current paper is also applicable towards understanding this type of samplers, leading to the following non-asymptotic theory.

Theorem 3. *Suppose that (19) holds true. Equipped with the estimates in Assumption 1 and the learning rate schedule (20), the stochastic sampler (28) achieves*

$$\mathsf{TV}\big(q_1, p_1\big) \le \sqrt{\frac{1}{2}\mathsf{KL}\big(q_1 \parallel p_1\big)} \le C_1 \frac{d^2 \log^3 T}{\sqrt{T}} \tag{29}$$

for some universal constants $C_1 > 0$, provided that $T \ge C_2 d^4 \log^6 T$ for some large enough constant $C_2 > 0$.

Theorem 3 establishes non-asymptotic convergence guarantees for the stochastic sampler (28). As asserted by the theorem, the number of steps needed to attain ε-accuracy (measured by the TV distance between p_1 and q_1) is proportional to $1/\varepsilon^2$, matching the state-of-the-art ε-dependency derived in Chen et al. (2022a), albeit exhibiting a worse dimensional dependency. Our analysis follows a completely different path compared with the SDE-based approach in Chen et al. (2022a), thus offering complementary interpretations for this important sampler. In order to further illustrate the versatility of our analysis approach, we shall demonstrate how it can be applied to study an accelerated version in the next subsection.

3.3.2 An accelerated stochastic sampler

In this subsection, we come up with a potential strategy to speed up the stochastic sampler (28), assuming access to reliable estimates of additional objects as described below.

Assumption 3. *Suppose that we have access to the estimates $v_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ $(1 \le t \le T)$ as follows:*

$$v_t := \arg\min_{v:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d} \mathbb{E}\left[\Big\|WW^\top Z - v\big(\sqrt{\overline{\alpha}_t}X + \sqrt{1-\overline{\alpha}_t}W, Z\big)\Big\|_2^2\right], \qquad 1 \le t \le T, \tag{30}$$

where X, W, Z are independently generated obeying $X \sim p_{\mathsf{data}}$, $W \sim \mathcal{N}(0, I_d)$, and $Z \sim \mathcal{N}(0, I_d)$.

With the estimates in Assumption 1 and Assumption 3 in place, we are positioned to introduce the proposed accelerated sampler as follows:

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \Psi_t(Y_t, Z_t) \quad \text{ for } t = T, \cdots, 1, \tag{31a}$$

where we choose the mapping $\Psi_t(\cdot,\cdot)$ as follows

$$\Psi_t(y, z) = \frac{1}{\sqrt{\alpha_t}}\Big(y + (1-\alpha_t)s_t(y)\Big) + \sigma_t\left\{z - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\left[z + (1-\overline{\alpha}_t)s_t(y)s_t(y)^\top z - v_t(y,z)\right]\right\} \tag{31b}$$

with

$$\sigma_t^2 = \frac{1}{\alpha_t} - 1. \tag{31c}$$

Clearly, the modified update mapping (31b) is still mainly a linear combination of the score estimate s_{t-1} and the additive noise Z_t, except that a correction term v_t (learned by solving (30)) needs to be included for acceleration purposes.

We now apply our analysis strategy to establish performance guarantees for the above stochastic sampler.

Theorem 4. *Suppose that* (19) *holds true. Equipped with the estimates in Assumption 1, 3 and the learning rate schedule* (20)*, the sampling process* (31) *satisfies*

$$\mathsf{TV}(q_1, p_1) \le \sqrt{\frac{1}{2}\mathsf{KL}(q_1 \parallel p_1)} \le C_1 \frac{d^3 \log^{4.5} T}{T} \tag{32}$$

for some universal constants $C_1 > 0$, provided that $T \ge C_2 d^3 \log^{4.5} T$ for some large enough constant $C_2 > 0$.

The proof of this result is provided in Section E.1. In comparison to the stochastic sampler (28), Theorem 4 asserts that the iteration complexity of the sampler (31) is at most

$$\widetilde{O}\left(\frac{d^3}{\varepsilon}\right), \tag{33}$$

thus significantly reducing the scaling $\widetilde{O}(1/\varepsilon^2)$ for the original sampler (28) to $\widetilde{O}(1/\varepsilon)$ regarding the ε-dependency. All in all, our theory reveals that having information about a small number of additional objects might substantially speed up the data generation process.

4 Other related works

Theory for SGMs. Early theoretical efforts in understanding the convergence of score-based stochastic samplers suffered from being either not quantitative (De Bortoli et al., 2021; Liu et al., 2022; Pidstrigach, 2022), or the curse of dimensionality (e.g., exponential dependencies in the convergence guarantees) (Block et al., 2020; De Bortoli, 2022). The recent work Lee et al. (2022) provided the first polynomial convergence guarantee in the presence of L_2-accurate score estimates, for any smooth distribution satisfying the log-Sobelev inequality, effectively only allowing unimodal distributions though. Chen et al. (2022b); Lee et al. (2023); Chen et al. (2022a) subsequently lifted such a stringent data distribution assumption. More concretely, Chen et al. (2022b) accommodated a broad family of data distributions under the premise that the score functions over the entire trajectory of the forward process are Lipschitz; Lee et al. (2023) only required certain smoothness assumptions but came with worse dependence on the problem parameters; and more recent results in Chen et al. (2022a) applied to literally any data distribution with bounded second-order moment. In addition, Wibisono and Yang (2022) also established a convergence theory for score-based generative models, assuming that the error of the score estimator has a bounded moment generating function and that the data distribution satisfies the log-Sobelev inequality. Turning attention to samplers based on the probability flow ODE, Chen et al. (2023c) derived the first non-asymptotic bounds for this type of samplers. Improved convergence guarantees have recently been provided by a concurrent work Chen et al. (2023b), with the assistance of additional corrector steps inerspersed in each iteration of the probability flow ODE. It is worth noting that the corrector steps proposed therein are based on Langevin-type diffusion and inject additive noise, and hence the resulting sampling processes are not deterministic. Additionally, theoretical justifications for DDPM in the context of image in-painting have been developed by Rout et al. (2023).

Score matching. Hyvärinen (2005) showed that the score function can be estimated via integration by parts, a result that was further extended in Hyvärinen (2007). Song et al. (2020b) proposed sliced score matching to tame the computational complexity in high dimension. The consistency of the score matching estimator was studied in Hyvärinen (2005), with asymptotic normality established in Forbes and Lauritzen

(2015). Optimizing the score matching loss has been shown to be intimately connected to minimizing upper bounds on the Kullback-Leibler divergence (Song et al., 2021a) and Wasserstein distance (Kwon et al., 2022) between the generated distribution and the target data distribution. From a non-asymptotic perspective, Koehler et al. (2022) studied the statistical efficiency of score matching by connecting it with the isoperimetric properties of the distribution.

Other theory for diffusion models. Oko et al. (2023) studied the approximation and generalization capabilities of diffusion modeling for distribution estimation. Assuming that the data are supported on a low-dimensional linear subspace, Chen et al. (2023a) developed a sample complexity bound for diffusion models. Moreover, Ghimire et al. (2023) adopted a geometric perspective and showed that the forward and backward processes of diffusion models are essentially Wasserstein gradient flows operating in the space of probability measures. Recently, the idea of stochastic localization, which is closely related to diffusion models, is adopted to sample from posterior distributions (Montanari and Wu, 2023; El Alaoui et al., 2022), which has been implemented using the approximate message passing algorithm (Donoho et al. (2009); Li and Wei (2022)).

5 Analysis

In this section, we describe our non-asymptotic proof strategies for two simpler samplers (i.e., (21) and (28)). The analyses for the two accelerated variants follow similar arguments as their non-accelerated counterparts, and are hence postponed to the appendices.

5.1 Preliminary facts

Before proceeding, we gather a couple of facts that will be useful for the proof, with most proofs postponed to Appendix A.

Properties related to the score function. First of all, in view of the alternative expression (18) for the score estimate s_t and the property of the minimum mean square error (MMSE) estimator (e.g., Hajek (2015, Section 3.3.1)), we know that s_t is given by the conditional expectation

$$\begin{aligned} s_t(x) &= \mathbb{E}\left[-\frac{1}{\sqrt{1-\overline{\alpha}_t}}W \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right] = \frac{1}{1-\overline{\alpha}_t}\mathbb{E}\left[\sqrt{\overline{\alpha}_t}X_0 - x \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right] \\ &= -\frac{1}{1-\overline{\alpha}_t}\underbrace{\int_{x_0}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)p_{X_0 \mid X_t}(x_0 \mid x)\mathrm{d}x_0}_{=:\, g_t(x)}. \end{aligned} \tag{34}$$

Let us also introduce the Jacobian matrix associated with $g_t(\cdot)$ as follows:

$$J_t(x) \coloneqq \frac{\partial g_t(x)}{\partial x}, \tag{35}$$

which can be equivalently rewritten as

$$\begin{aligned} J_t(x) = I_d + \frac{1}{1-\overline{\alpha}_t}\Big\{&\mathbb{E}\big[X_t - \sqrt{\overline{\alpha}_t}X_0 \mid X_t = x\big]\Big(\mathbb{E}\big[X_t - \sqrt{\overline{\alpha}_t}X_0 \mid X_t = x\big]\Big)^{\top} \\ &- \mathbb{E}\Big[\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)^{\top} \mid X_t = x\Big]\Big\}. \end{aligned} \tag{36}$$

Properties about the learning rates. Next, we isolate a few useful properties about the learning rates as specified by $\{\alpha_t\}$ in (20):

$$\alpha_t \geq 1 - \frac{c_1 \log T}{T} \geq \frac{1}{2}, \qquad\qquad 1 \leq t \leq T \tag{37a}$$

$$\frac{1}{2}\frac{1-\alpha_t}{1-\overline{\alpha}_t} \le \frac{1}{2}\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t} \le \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \le \frac{4c_1\log T}{T}, \qquad 2\le t\le T \tag{37b}$$

$$1 \le \frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} \le 1+\frac{4c_1\log T}{T}, \qquad 2\le t\le T \tag{37c}$$

$$\overline{\alpha}_T \le \frac{1}{T^{c_2}}, \tag{37d}$$

where c_1 defined in (20), and $c_2>0$ is some large enough numerical constant. In addition, if $\frac{d(1-\alpha_t)}{\alpha_t-\overline{\alpha}_t}\lesssim 1$, then one has

$$\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} = 1+\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{d(d-2)(1-\alpha_t)^2}{8(\alpha_t-\overline{\alpha}_t)^2}+O\bigg(d^3\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\bigg). \tag{37e}$$

The proof of these properties is postponed to Appendix A.2.

Properties of the forward process. Additionally, recall that the forward process satisfies $X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W$ with $W\sim\mathcal{N}(0,I_d)$. We have the following tail bound concerning the random vector X_0 conditional on X_t, whose proof can be found in Appendix A.3.

Lemma 1. *Suppose that there exists some numerical constant $c_R>0$ obeying*

$$\mathbb{P}\big(\|X_0\|_2 \le R\big) = 1 \qquad \textit{and} \qquad R = T^{c_R}. \tag{38}$$

Consider any $y\in\mathbb{R}$ obeying

$$-\log p_{X_t}(y) \le c_6 d\log T \tag{39}$$

for some large enough constant $c_6>0$. Then for any quantity $c_5 \ge \sqrt{2+c_0/2+c_R} =: \overline{c}_5$ (with c_0 defined in (20)), conditioned on $X_t=y$ one has

$$\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2 \le 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T} \tag{40}$$

with probability at least $1-\exp\big(-c_5^2 d\log T\big)$. In addition, it holds that

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2 \,\big|\, X_t=y\right] \le 6\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T}, \tag{41a}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^2 \,\big|\, X_t=y\right] \le 30\overline{c}_5^2 d(1-\overline{\alpha}_t)\log T, \tag{41b}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^3 \,\big|\, X_t=y\right] \le 130\overline{c}_5^3\big(d(1-\overline{\alpha}_t)\log T\big)^{3/2}, \tag{41c}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^4 \,\big|\, X_t=y\right] \le 630\overline{c}_5^4\big(d(1-\overline{\alpha}_t)\log T\big)^{2}. \tag{41d}$$

In words, Lemma 1 implies that conditional on a *typical* value of X_t, the vector $\sqrt{\overline{\alpha}_t}X_0 - X_t = \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t$ (see (6)) might still follow a sub-Gaussian tail, whose expected norm remains as the same order of that of an unconditional Gaussian vector $\mathcal{N}(0,(1-\overline{\alpha}_t)I_d)$.

The next lemma singles out another useful fact that controls the tail of p_{X_t} of the forward process; the proof is postponed to Appendix A.4.

Lemma 2. *Consider any two points $x_t, x_{t-1}\in\mathbb{R}^d$ obeying*

$$-\log p_{X_t}(x_t) \le \frac{1}{2}c_6 d\log T, \quad \textit{and} \quad \Big\|x_{t-1}-\frac{x_t}{\sqrt{\alpha_t}}\Big\|_2 \le c_3\sqrt{d(1-\alpha_t)\log T} \tag{42}$$

for some large constants $c_6, c_3>0$. If we define $x_t(\gamma) \coloneqq \gamma x_{t-1} + (1-\gamma)x_t/\sqrt{\alpha_t}$ for any $\gamma\in[0,1]$, then

$$-\log p_{X_{t-1}}\big(x_t(\gamma)\big) \le c_6 d\log T, \qquad \forall\gamma\in[0,1]. \tag{43}$$

In other words, if x_t falls within a typical range of X_t and if the point x_{t-1} is not too far away from $x_t/\sqrt{\alpha_t}$, then x_{t-1} is also a typical value of the previous point X_{t-1}. As an immediate consequence, combining Lemma 2 with Lemma 1 reveals that: if the assumption (42) holds, then conditional on $X_{t-1} = x_t(\gamma)$ for any $\gamma \in [0,1]$, one has

$$\big\|\sqrt{\overline{\alpha}_{t-1}}X_0 - x_t(\gamma)\big\|_2 \le 5c_5\sqrt{d(1-\overline{\alpha}_{t-1})\log T} \tag{44a}$$

$$\bigg\|\sqrt{\overline{\alpha}_{t-1}}X_0 - \frac{x_t}{\sqrt{\alpha_t}}\bigg\|_2 \le \big\|\sqrt{\overline{\alpha}_{t-1}}X_0 - x_t(\gamma)\big\|_2 + \bigg\|x_t(\gamma) - \frac{x_t}{\sqrt{\alpha_t}}\bigg\|_2 \le (5c_5+c_3)\sqrt{d(1-\overline{\alpha}_{t-1})\log T} \tag{44b}$$

with probability exceeding $1-\exp\big(-c_5^2 d\log T\big)$, where the last inequality invokes the property (37b).

5.2 Analysis for the sampler based on probability flow ODE (Theorem 1)

We now present the proof for our main result (i.e., Theorem 1) tailored to the deterministic sampler (21) based on probability flow ODE. Given that the total variation distance is always bounded above by 1, it suffices to assume

$$T \ge C_1 d^2\log^4 T + \sqrt{C_1}d^3\log^3 T \tag{45}$$

throughout the proof (otherwise the claimed result (22) becomes trivial). Before proceeding, we find it convenient to introduce a function

$$\phi_t(x) = x + \frac{1-\alpha_t}{2}s_t(x) = x - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\int_{x_0}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0, \tag{46}$$

where the second identity follows from (34). This allows us to express the update rule (21) as follows:

$$Y_{t-1} = \Phi_t(Y_t) = \frac{1}{\sqrt{\alpha_t}}\phi_t(Y_t). \tag{47}$$

Our proof consists of three steps below.

Step 1: bounding the density ratios of interest. To begin with, we note that for any vectors y_{t-1} and y_t, elementary properties about transformation of probability distributions give

$$\begin{aligned}\frac{p_{Y_{t-1}}(y_{t-1})}{p_{X_{t-1}}(y_{t-1})} &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})}\\ &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{Y_t}(y_t)}\cdot\left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{X_t}(y_t)}\right)^{-1}\cdot\frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)},\end{aligned} \tag{48}$$

thus converting the density ratio of interest into the product of three other density ratios. Noteworthily, this observation (48) connects the target density ratio $\frac{p_{Y_{t-1}}}{p_{X_{t-1}}}$ at the $(t-1)$-th step with its counterpart $\frac{p_{Y_t}}{p_{X_t}}$ at the t-th step, motivating us to look at the density changes within adjacent steps in both the forward and the reverse processes (i.e., $p_{X_{t-1}}$ vs. p_{X_t} and $p_{Y_{t-1}}$ vs. p_{Y_t}). Motivated by this expression, we develop a key lemma related to some of these density ratios, which plays a central role in establishing Theorem 1. The proof of this result is postponed to Appendix B.1.

Lemma 3. *Suppose* $\frac{d^2(1-\alpha_t)\log T}{\alpha_t-\overline{\alpha}_t}\lesssim 1$. *For every* $x\in\mathbb{R}$ *obeying* $-\log p_{X_t}(x)\le c_6 d\log T$ *for some large enough constant* $c_6>0$*, it holds that*

$$\begin{aligned}\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} &= 1+\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg)\\ &\quad+\frac{(1-\alpha_t)\Big(\big\|\int\big(x-\sqrt{\overline{\alpha}_t}x_0\big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\big\|_2^2-\int\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}.\end{aligned} \tag{49a}$$

Moreover, for any random vector Y, one has

$$\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)} = 1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2 \log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3 \log^3 T\bigg)$$
$$+ \frac{(1-\alpha_t)\Big(\big\|\int\big(x-\sqrt{\overline{\alpha}_t}x_0\big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\big\|_2^2 - \int\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}. \tag{49b}$$

Step 2: bounding the TV distance with the aid of a high-probability event. In order to bound the TV distance of interest, we find it helpful to single out the following set:

$$\mathcal{E}_t := \bigg\{y : \bigg|\frac{q_t(y)}{p_t(y)} - 1\bigg| \le c_5\Big(\frac{d^2\log^4 T}{T^2} + \frac{d^6\log^6 T}{T^3}\Big)(T-t+1)\bigg\} \tag{50}$$

for some large enough constant $c_5 > 0$. Informally speaking, this set $\mathcal{E}_t$ contains all points such that $q_t(y) \approx p_t(y)$. We now claim that the event $Y_t \in \mathcal{E}_t$ occurs with high probability, in the sense that

$$\mathbb{P}(Y_t \in \mathcal{E}_t) \ge 1 - (T-t+1)\exp\big(-c_3 d\log T\big), \qquad t \ge 1 \tag{51}$$

for some constant $c_3 > 0$. The key to proving this claim is to invoke Lemma 3 and the transformation relation (48) to justify — with high probability — the proximity of $q_t(Y_t)$ and $p_t(Y_t)$ throughout the reverse process; the proof is deferred to Step 3 to streamline presentation.

Suppose for the moment that the above claim (51) is valid. Then taking $t = 1$ leads to

$$\mathbb{P}(Y_1 \in \mathcal{E}_1) = \mathbb{P}\bigg(\bigg|\frac{q_1(Y_1)}{p_1(Y_1)} - 1\bigg| \le c_5\Big(\frac{d^2\log^4 T}{T} + \frac{d^6\log^6 T}{T^2}\Big)\bigg) \ge 1 - T\exp\big(-c_3 d\log T\big),$$

which also reveals that $\int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y \le T\exp\big(-c_3 d\log T\big)$. Additionally, we make the observation that

$$\begin{aligned}
\int_{y\notin\mathcal{E}_1}\big|p_1(y)-q_1(y)\big|\mathrm{d}y &\le \int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y + \int_{y\notin\mathcal{E}_1} q_1(y)\mathrm{d}y = \int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y + 1 - \int_{y\in\mathcal{E}_1} q_1(y)\mathrm{d}y \\
&\le \int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y + 1 - \int_{y\in\mathcal{E}_1} p_1(y)\mathrm{d}y + \int_{y\in\mathcal{E}_1}\big|p_1(y)-q_1(y)\big|\mathrm{d}y \\
&= 2\int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y + \int_{y\in\mathcal{E}_1}\big|p_1(y)-q_1(y)\big|\mathrm{d}y.
\end{aligned}$$

The above results together with the definition of the total variation distance gives

$$\begin{aligned}
\mathsf{TV}\big(q_1, p_1\big) &= \frac{1}{2}\int_{y\in\mathcal{E}}\big|q_1(y)-p_1(y)\big|\mathrm{d}y + \frac{1}{2}\int_{y\notin\mathcal{E}}\big|q_1(y)-p_1(y)\big|\mathrm{d}y \\
&\le \int_{y\in\mathcal{E}}\big|q_1(y)-p_1(y)\big|\mathrm{d}y + \int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y \\
&= \mathbb{E}_{Y_1\sim p_1}\bigg[\bigg|\frac{q_1(Y_1)}{p_1(Y_1)} - 1\bigg|\cdot \mathbb{1}\left\{Y_1\in\mathcal{E}_1\right\}\bigg] + \int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y \\
&\le c_5\left(\frac{d^2\log^4 T}{T} + \frac{d^6\log^6 T}{T^2}\right) + \exp\big(-c_3 d\log T\big) \\
&\asymp \frac{d^2\log^4 T}{T} + \frac{d^6\log^6 T}{T^2}.
\end{aligned} \tag{52}$$

This establishes the advertised result in Theorem 1, provided that Claim (51) can be verified.

Step 3: justifying the claim (51). We would like to prove this claim by induction, for which we start with the base case with $t = T$. Recall that $X_T \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_T}X_0 + \sqrt{1-\overline{\alpha}_T}B$ and $Y_T \overset{\mathrm{d}}{=} B$ with $B \sim \mathcal{N}(0, I_d)$ independent of X_0, and that $\|X_0\|_2 \le R$ with $R = T^{c_R}$ for some constant $c_R > 0$. For large enough T, it immediately follows from (37d) that

$$\mathbb{P}(Y_T \in \mathcal{E}_T) \ge 1 - \exp(-c_3 d \log T) \tag{53}$$

for some constant $c_3 > 0$ large enough.

Suppose now that the claim (51) holds for some $t \ge 2$, and we wish to prove the claim for $t-1$. We would first like to claim that with probability at least $1-(T-t)\exp(-c_3 d\log T)$ for some constant $c_3 > 0$, one has

$$q_t(Y_t) \ge \exp\big(-c_6 d \log T\big). \tag{54}$$

With (54) in place, one sees that Lemma 3 is applicable to $x = Y_t$ with high probability.

Next, consider any y obeying $-\log p_{X_t}(y) \le c_6 d \log T$. It is seen from Lemma 1 and (37) that

$$\frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} \lesssim \frac{d\log T}{T} = o(1)$$

and

$$\left|\frac{(1-\alpha_t)\Big(\big\|\int \mathbb{E}\big[X_t - \sqrt{\overline{\alpha}_t}X_0 \,|\, X_t = y\big]\big\|_2^2 - \int \mathbb{E}\big[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \,|\, X_t = y\big]\Big)}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}\right|$$
$$\le \left|\frac{(1-\alpha_t)\int \mathbb{E}\big[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \,|\, X_t = y\big]}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}\right| \lesssim \frac{(1-\alpha_t)d\log T}{\alpha_t - \overline{\alpha}_t} \lesssim \frac{d\log T}{T} = o(1).$$

With these two bounds in mind, applying relations (49a) and (49b) in Lemma 3 leads to

$$\begin{aligned}\frac{p_{\sqrt{\alpha_t}Y_{t-1}}\big(\phi_t(y)\big)}{p_{Y_t}(y)}\left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(y)\big)}{p_{X_t}(y)}\right)^{-1} &= \frac{p_{\phi_t(Y_t)}\big(\phi_t(y)\big)}{p_{Y_t}(y)}\left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(y)\big)}{p_{X_t}(y)}\right)^{-1} \\ &= 1 + O\left(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3 \log^3 T\right).\end{aligned}$$

Replacing y with Y_t in the above display, using the fact that $Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\phi_t(Y_t)$, and invoking the relation (48), we immediately arrive at

$$\frac{p_{t-1}(Y_{t-1})}{q_{t-1}(Y_{t-1})} = \left\{1 + O\left(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3 \log^3 T\right)\right\} \cdot \frac{p_t(Y_t)}{q_t(Y_t)}$$

with probability exceeding $1-(T-t)\exp(-c_3 d\log T)$. This concludes the proof of Claim (51) via standard induction arguments.

Proof of property (54). First, for any y obeying $y \in \mathcal{E}_t$ and $p_t(y) \ge 2\exp(-c_6 d\log T)$, one has

$$q_t(y) \ge \left(1 - \left|\frac{q_t(y)}{p_t(y)} - 1\right|\right) p_t(y) \ge \frac{1}{2}p_t(y) \ge \exp(-c_6 d \log T)$$

and, as a result,

$$\Big\{y \,|\, p_t(y) \ge 2\exp\big(-c_6 d\log T\big)\Big\} \cap \mathcal{E}_t \subseteq \Big\{y \,|\, q_t(y) \ge \exp\big(-c_6 d\log T\big)\Big\}. \tag{55}$$

We can then deduce that

$$\mathbb{P}\Big(q_t(Y_t) \ge \exp\big(-c_6 d\log T\big)\Big) \ge \mathbb{P}\Big(p_t(Y_t) \ge 2\exp\big(-c_6 d\log T\big) \text{ and } Y_t \in \mathcal{E}_t\Big)$$

$$
\begin{aligned}
&\geq \mathbb{P}\Big(Y_t \in \big\{y : p_t(y) \geq 2\exp\big(-c_6 d\log T\big)\big\} \cap \mathcal{E}_t \cap \big\{y : \|y\|_2 \leq T^{c_y}\big\}\Big) \\
&\geq 1 - \mathbb{P}\left\{Y_t \notin \mathcal{E}_t\right\} - \mathbb{P}\Big(Y_t \in \big\{y : p_t(y) < 2\exp\big(-c_6 d\log T\big)\big\} \cap \big\{y : \|y\|_2 \leq T^{c_y}\big\}\Big) - \mathbb{P}\big\{\|Y_t\|_2 > T^{c_y}\big\} \\
&\geq 1 - (T-t-1)\exp\big(-c_3 d\log T\big) - \int_{y:\|y\|_2 \leq T^{c_y}} 2\exp\big(-c_6 d\log T\big)\mathrm{d}y - \mathbb{P}\big\{\|Y_T\|_2 > T^{c_y/2}\big\} \\
&\geq 1 - (T-t)\exp\big(-c_3 d\log T\big).
\end{aligned}
$$

Here, the last line makes use of the fact $y_T \sim \mathcal{N}(0, I_d)$ and holds as long as c_6 is large enough; regarding the penultimate line, it suffices to recognize that (cf. (19), (46) and (37))

$$
\big\|\phi_t(x)\big\|_2 \leq \left(1 - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\right)\|x\|_2 + \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}T^{c_R},
$$

and hence it is straightforward to show that $\|y_t\|_2 \leq T^{c_y}$ $(1 \leq t < T)$ holds for lage enough constant $c_y > 0$ unless $\|y_T\|_2 > T^{c_y/2}$ (a condition that happens with exponentially small probability given that $Y_T \sim \mathcal{N}(0, I_d)$). □

5.3 Analysis for the DDPM-type sampler (Theorem 3)

Turning attention to the DDPM-type stochastic sampler (28), we now present the main steps for the proof of Theorem 3. Towards this, let us first introduce the following mapping

$$
\begin{aligned}
\mu_t(x_t) &:= \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0 = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}g_t(x_t) \\
&= \frac{1}{\sqrt{\alpha_t}}\Big(x_t + (1-\alpha_t)s_t(x_t)\Big),
\end{aligned} \tag{56}
$$

where the last identity relies on the expression (34) for $s_t(\cdot)$. Recalling the update rule (28), we can write

$$
p_{Y_{t-1} \,|\, Y_t}(x_{t-1} \,|\, x_t) = \frac{1}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{d/2}}\exp\left(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1} - \mu_t(x_t)\big\|_2^2\right) \tag{57}
$$

for any two points $x_t, x_{t-1} \in \mathbb{R}^d$. For notational simplicity, we shall also use the following notation throughout:

$$
\widehat{x}_t := \frac{1}{\sqrt{\alpha_t}}x_t. \tag{58}
$$

The proof of Theorem 3 can be divided into several steps below.

Step 1: decomposition of the KL divergence. The celebrated Pinsker inequality (see, e.g., Tsybakov (2009, Lemma 2.5)) tells us that

$$
\mathsf{TV}(p_{X_1}, p_{Y_1}) \leq \sqrt{\frac{1}{2}\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1})}, \tag{59}
$$

and hence it suffices to work with the KL divergence. Recall that $X_1 \to \cdots \to X_T$ and $Y_T \to \cdots \to Y_1$ are both Markov chains (so are their reverse processes). In order to compute the KL divergence between p_{X_1} and p_{Y_1}, we make note of the following elementary relations:

$$
\begin{aligned}
\mathsf{KL}(p_{X_1,\dots,X_T} \,\|\, p_{Y_1,\dots,Y_T}) &= \mathsf{KL}(p_{X_1} \,\|\, p_{Y_1}) + \sum_{t=2}^{T}\mathbb{E}_{x\sim X_{t-1}}\Big[\mathsf{KL}\Big(p_{X_t \,|\, X_{t-1}}(\cdot \,|\, x) \,\Big\|\, p_{Y_t \,|\, Y_{t-1}}(\cdot \,|\, x)\Big)\Big] \\
&= \mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) + \sum_{t=2}^{T}\mathbb{E}_{x\sim X_t}\Big[\mathsf{KL}\Big(p_{X_{t-1} \,|\, X_t}(\cdot \,|\, x) \,\Big\|\, p_{Y_{t-1} \,|\, Y_t}(\cdot \,|\, x)\Big)\Big],
\end{aligned}
$$

which combined with the non-negativity of the KL divergence indicates that

$$\mathsf{KL}(p_{X_1} \parallel p_{Y_1}) \le \mathsf{KL}(p_{X_T} \parallel p_{Y_T}) + \sum_{t=2}^{T} \mathbb{E}_{x\sim X_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x)\,\Big\|\,p_{Y_{t-1}\,|\,Y_t}(\cdot\,|\,x)\Big)\Big]. \tag{60}$$

This allows us to focus attention on the transition probabilities at each time instance t.

On the right-hand side of (60), the term that is the easiest to bound is $\mathsf{KL}(p_{X_T} \parallel p_{Y_T})$. Recognizing that $Y_T \sim \mathcal{N}(0, I_d)$ and that $X_T \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_T} X_0 + \sqrt{1-\overline{\alpha}_T}\,\overline{W}_t$ with $\overline{W}_t \sim \mathcal{N}(0, I_d)$ (independent from X_0), one has

$$\begin{aligned}
\mathsf{KL}(p_{X_T} \parallel p_{Y_T}) &= \int p_{X_T}(x)\log\frac{p_{X_T}(x)}{p_{Y_T}(x)}\mathrm{d}x \\
&\overset{\text{(i)}}{=} \int p_{X_T}(x)\log\frac{\int_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}} p_{\sqrt{\overline{\alpha}_T}X_0}(y)\,p_{\sqrt{1-\overline{\alpha}_T}\,\overline{W}_t}(x-y)\mathrm{d}y}{p_{Y_T}(x)}\mathrm{d}x \\
&\le \int p_{X_T}(x)\log\frac{\sup_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}} p_{\sqrt{1-\overline{\alpha}_T}\,\overline{W}_t}(x-y)}{p_{Y_T}(x)}\mathrm{d}x \\
&= \int p_{X_T}(x)\bigg(-d/2\log(1-\overline{\alpha}_T) + \sup_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}}\Big(-\frac{\|x-y\|_2^2}{2(1-\overline{\alpha}_T)} + \frac{\|x\|_2^2}{2}\Big)\bigg)\mathrm{d}x \\
&\overset{\text{(ii)}}{\le} \int p_{X_T}(x)\bigg(-d/2\log(1-\overline{\alpha}_T) + \|x\|_2 \sup_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}}\frac{\|y\|_2}{1-\overline{\alpha}_T}\bigg)\mathrm{d}x \\
&\le -d/2\log(1-\overline{\alpha}_T) + \frac{\sqrt{\overline{\alpha}_T}T^{c_R}}{2(1-\overline{\alpha}_T)}\mathbb{E}\left[\|X_T\|_2\right] \\
&\overset{\text{(iii)}}{\lesssim} \overline{\alpha}_T d + \frac{\sqrt{\overline{\alpha}_T}T^{c_R}}{2(1-\overline{\alpha}_T)}\Big(\sqrt{\overline{\alpha}_T}T^{c_R} + \sqrt{d}\Big) \overset{\text{(iv)}}{\lesssim} \frac{1}{T^{200}},
\end{aligned} \tag{61}$$

where (i) arises from the assumption that $\|X_0\|_2 \le T^{c_R}$, (ii) applies the Cauchy-Schwarz inequality, (iii) holds true since

$$\mathbb{E}\left[\|X_T\|_2\right] \le \sqrt{\overline{\alpha}_T}\|X_0\|_2 + \mathbb{E}\left[\|\overline{W}_t\|_2\right] \le \sqrt{\overline{\alpha}_T}T^{c_R} + \sqrt{\mathbb{E}\left[\|\overline{W}_t\|_2^2\right]} \le \sqrt{\overline{\alpha}_T}T^{c_R} + \sqrt{d},$$

and (iv) makes use of (37d) when c_2 is sufficiently large. Thus, in view of (60), it suffices to focus attention on bounding $\mathsf{KL}\big(p_{X_{t-1}\,|\,X_t}(\cdot\mid x) \parallel p_{Y_{t-1}\,|\,Y_t}(\cdot\mid x)\big)$ for each $1 < t \le T$, which forms the main content of the subsequent proof.

Step 2: controlling the conditional distributions $p_{X_{t-1}\,|\,X_t}$ and $p_{Y_{t-1}\,|\,Y_t}$. In order to compute the KL divergence of interest in (60), one needs to calculate the two conditional distributions $p_{X_{t-1}\,|\,X_t}$ and $p_{Y_{t-1}\,|\,Y_t}$, which we study in this step. To do so, we find it helpful to first introduce the following set

$$\mathcal{E} := \bigg\{(x_t, x_{t-1})\,\big|\, -\log p_{X_t}(x_t) \le \frac{1}{2}c_6 d\log T,\ \|x_{t-1} - \widehat{x}_t\|_2 \le c_3\sqrt{d(1-\alpha_t)\log T}\bigg\}, \tag{62}$$

where the two numerical constants $c_3, c_6 > 0$ are introduced in Lemma 2. Informally, $\mathcal{E}$ encompasses a typical range of the values of (X_t, X_{t-1}), and our analysis shall often proceed by studying the points in $\mathcal{E}$ and those outside $\mathcal{E}$ separately.

The first result below quantifies the conditional density $p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)$ for those points residing within $\mathcal{E}$, which plays a central role in comparing $p_{X_{t-1}\,|\,X_t}$ against $p_{Y_{t-1}\,|\,Y_t}$ (see (57)). The proof can be found in Appendix C.1.

Lemma 4. *There exists some large enough numerical constant $c_\zeta > 0$ such that: for every $(x_t, x_{t-1}) \in \mathcal{E}$,*

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = \frac{1}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{d/2}}\exp\bigg(-\frac{\alpha_t\big\|x_{t-1}-\mu_t(x_t)\big\|_2^2}{2(1-\alpha_t)} + \zeta_t(x_{t-1}, x_t)\bigg) \tag{63}$$

holds for some residual term $\zeta_t(x_{t-1}, x_t)$ obeying

$$\big|\zeta_t(x_{t-1}, x_t)\big| \le c_\zeta d^2 \bigg(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\bigg) \log^2 T. \tag{64}$$

Here, we recall the definition of $\mu_t(x_t)$ in (56).

By comparing Lemma 4 with expression (57), we see that when restricted to the set $\mathcal{E}$, the two conditional distributions $p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)$ and $p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)$ (i.e., informally, the time-reversed transition kernels) are fairly close to each other, a crucial observation that suggests the validity of the diffusion generative model.

Furthermore, we are also in need of bounding the ratio of the two conditional distributions when going beyond the set $\mathcal{E}$. As it turns out, it suffices to develop a crude bound on the logarithm of such ratios (which are used in defining the KL divergence), as stated in the following lemma.

Lemma 5. *For all $(x_t, x_{t-1}) \in \mathbb{R}^d \times \mathbb{R}^d$, it holds that*

$$\log \frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} \le 2T\left(\|x_{t-1} - \widehat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}\right). \tag{65}$$

The proof of Lemma 5 is provided in Appendix C.2.

Step 3: bounding the KL divergence between $p_{X_{t-1}\,|\,X_t}$ and $p_{Y_{t-1}\,|\,Y_t}$. Armed with the above two lemmas, we are ready to control the KL divergence between $p_{X_{t-1}\,|\,X_t}$ and $p_{Y_{t-1}\,|\,Y_t}$. It is first seen from Lemma 4 and (57) that: for any $(x_t, x_{t-1}) \in \mathcal{E}$,

$$\begin{aligned}
\frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} &= \exp\Big(O\Big(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\log^2 T\Big)\Big) = 1 + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\log^2 T\bigg) \\
&= 1 + O\bigg(d^2 \frac{\log^3 T}{T}\bigg) \in \Big[\frac{1}{2}, 2\Big],
\end{aligned} \tag{66}$$

where the last line results from (37b) and the assumption that $T \ge c_{10} d^2 \log^3 T$ for some large enough constant $c_{10} > 0$. We can then calculate

$$\begin{aligned}
&\mathbb{E}_{x_t \sim X_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t) \;\|\; p_{Y_{t-1}\,|\,Y_t}(\cdot\,|\,x_t)\Big)\Big] \\
&= \Big(\int_{\mathcal{E}} + \int_{\mathcal{E}^c}\Big) p_{X_t}(x_t) p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \log \frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} \mathrm{d}x_{t-1}\mathrm{d}x_t, \\
&\overset{\text{(i)}}{=} \int_{\mathcal{E}} p_{X_t}(x_t) \bigg\{ p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) - p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t) \\
&\qquad\qquad + p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \cdot O\bigg(\bigg(\frac{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)}{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)} - 1\bigg)^2\bigg)\bigg\} \mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\quad + \int_{\mathcal{E}^c} p_{X_t}(x_t) p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \log \frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} \mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\overset{\text{(ii)}}{=} \int_{\mathcal{E}} p_{X_t}(x_t) \bigg\{ p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) - p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t) + p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) O\bigg(d^4\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2 \log^4 T\bigg)\bigg\} \mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\quad + \int_{\mathcal{E}^c} p_{X_t}(x_t) p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \Big\{2T\big(\|x_t\|_2^2 + \|x_{t-1} - \widehat{x}_t\|_2^2 + T^{2c_R}\big)\Big\} \mathrm{d}x_{t-1}\mathrm{d}x_t.
\end{aligned} \tag{67}$$

Here, (i) results from the elementary fact that: if $\big|\frac{p_Y(x)}{p_X(x)} - 1\big| < \frac{1}{2}$, then the Taylor expansion gives

$$p_X(x) \log \frac{p_X(x)}{p_Y(x)} = -p_X(x) \log\Big(1 + \frac{p_Y(x) - p_X(x)}{p_X(x)}\Big)$$

$$= p_X(x) - p_Y(x) + p_X(x) O\Big(\Big(\frac{p_Y(x)}{p_X(x)} - 1\Big)^2\Big);$$

regarding (ii), we invoke (66) and Lemma 5.

To continue, let us bound each term on the right-hand side of (67) separately. From the definition of the set $\mathcal{E}$ (cf. (62)), direct calculations yield

$$\begin{aligned}\mathbb{P}\big((X_t, X_{t-1}) \notin \mathcal{E}\big) &= \int_{(x_t, x_{t-1}) \notin \mathcal{E}} p_{X_{t-1}}(x_{t-1}) p_{X_t \,|\, X_{t-1}}(x_t \,|\, x_{t-1}) \mathrm{d}x_{t-1} \mathrm{d}x_t \\ &= \int_{(x_t, x_{t-1}) \notin \mathcal{E}} p_{X_{t-1}}(x_{t-1}) \frac{1}{\big(2\pi(1-\alpha_t)\big)^{d/2}} \exp\Big(-\frac{\|x_t - \sqrt{\alpha_t} x_{t-1}\|_2^2}{2(1-\alpha_t)}\Big) \mathrm{d}x_{t-1} \mathrm{d}x_t \\ &\le \exp\big(-c_3 d \log T\big), \end{aligned} \tag{68}$$

and similarly,

$$\int_{(x_{t-1}, x_t) \notin \mathcal{E}} p_{X_t}(x_t) p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t) \Big(2T\big(\|x_t\|_2^2 + \|x_{t-1} - \widehat{x}_t\|_2^2\big) + T^{2c_R}\Big) \mathrm{d}x_{t-1} \mathrm{d}x_t \le \exp\big(-c_3 d \log T\big). \tag{69}$$

In addition, for every (x_t, x_{t-1}) obeying $\|x_{t-1} - x_t/\sqrt{\alpha_t}\|_2 > c_3\sqrt{d(1-\alpha_t)\log T}$ and $-\log p_{X_t}(x_t) \le \frac{1}{2} c_6 d \log T$, one can use the definition (56) of $\mu_t(\cdot)$ to obtain

$$\begin{aligned}\|x_{t-1} - \mu_t(x_t)\|_2 &= \Big\|x_{t-1} - \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)} \mathbb{E}\Big[x_t - \sqrt{\overline{\alpha}_t} X_0 \mid X_t = x_t\Big]\Big\|_2 \\ &\ge \Big\|x_{t-1} - \frac{1}{\sqrt{\alpha_t}} x_t\Big\|_2 - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)} \mathbb{E}\Big[\big\|x_t - \sqrt{\overline{\alpha}_t} X_0\big\|_2 \mid X_t = x_t\Big] \\ &\ge c_3\sqrt{d(1-\alpha_t)\log T} - 6\overline{c}_5 \frac{1-\alpha_t}{\sqrt{\alpha_t(1-\overline{\alpha}_t)}} \sqrt{d \log T} \\ &= \left(c_3 - 6\overline{c}_5 \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t(1-\overline{\alpha}_t)}}\right) \sqrt{d(1-\alpha_t)\log T} \ge \frac{c_3}{2}\sqrt{d(1-\alpha_t)\log T}, \end{aligned}$$

where the third line results from (41a) in Lemma 1, and the last line applies (37) and holds true as long as c_3 is large enough. In turn, this combined with (57) indicates that: for any x_t obeying $-\log p_{X_t}(x_t) \le \frac{1}{2} c_6 d \log T$,

$$\int_{x_{t-1}: \|x_{t-1} - x_t/\sqrt{\alpha_t}\|_2 > c_3\sqrt{d(1-\alpha_t)\log T}} p_{Y_{t-1} \,|\, Y_t}(x_{t-1} \,|\, x_t) \mathrm{d}x_{t-1} \le \exp\Big(-\frac{c_3}{2} d \log T\Big). \tag{70}$$

As a result, (68) and (70) taken collectively demonstrate that

$$\begin{aligned}&\Big|\int_{\mathcal{E}} p_{X_t}(x_t) \Big\{p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t) - p_{Y_{t-1} \,|\, Y_t}(x_{t-1} \,|\, x_t)\Big\} \mathrm{d}x_{t-1} \mathrm{d}x_t\Big| \\ &\quad = \Big|1 - \mathbb{P}\big((X_t, X_{t-1}) \notin \mathcal{E}\big) - \int_{(x_t, x_{t-1}) \notin \mathcal{E}} p_{X_t}(x_t) \Big\{1 - p_{Y_{t-1} \,|\, Y_t}(x_{t-1} \,|\, x_t)\Big\} \mathrm{d}x_{t-1} \mathrm{d}x_t\Big| \\ &\quad \le 2\exp\Big(-\frac{c_3}{2} d \log T\Big). \end{aligned} \tag{71}$$

Substituting (69) and (71) into (67) yields: for each $t \ge 2$,

$$\mathbb{E}_{x_t \sim X_t}\Big[\mathsf{KL}\Big(p_{X_{t-1} \,|\, X_t}(\cdot \,|\, x_t) \parallel p_{Y_{t-1} \,|\, Y_t}(\cdot \,|\, x_t)\Big)\Big] \lesssim d^4 \Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \log^4 T + 3\exp\Big(-\frac{c_3}{2} d \log T\Big) \lesssim \frac{d^4 \log^6 T}{T^2}, \tag{72}$$

where the last inequality utilizes the properties (37) of the learning rates.

Step 4: putting all this together. To finish up, substitute (72) into the decomposition (60) to obtain

$$\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1}) \lesssim \mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) + \sum_{2\leq t\leq T} \frac{d^4 \log^6 T}{T^2} \asymp \frac{d^4 \log^6 T}{T},$$

where the last relation applies the bound on $\mathsf{KL}(p_{X_T} \,\|\, p_{Y_T})$ as in (61). This establishes Theorem 3.

6 Discussion

In this paper, we have developed a new suite of non-asymptotic theory for establishing the convergence and faithfulness of diffusion generative modeling, assuming access to reliable estimates of the (Stein) score functions. Our analysis framework seeks to track the dynamics of the reverse process directly using elementary tools, which eliminates the need to look at the continuous-time limit and invoke the SDE and ODE toolboxes. Only the very minimal assumptions on the target data distribution are imposed. In addition to demonstrating the non-asymptotic iteration complexities of two mainstream discrete-time samplers — a deterministic sampler based on the probability flow ODE, and a DDPM-type stochastic sampler — we have discovered potential strategies to further accelerate the sampling process, taking advantage of estimates of a small number of additional objects. The analysis framework laid out in the current paper might shed light on how to analyze other variants of score-based generative models as well.

Moving forward, there are plenty of questions that require in-depth theoretical understanding. For instance, the dimension dependency in our convergence results remains sub-optimal; can we further refine our theory in order to reveal tight dependency in this regard? To what extent can we further accelerate the sampling process, without requiring much more information than the score functions? Ideally, one would hope to achieve accleration with the aid of the score functions only. It would also be of paramount interest to establish end-to-end performance guarantees that take into account both the score learning phase and the sampling phase.

Acknowledgements

Y. Wei is supported in part by the the NSF grants DMS-2147546/2015447, CAREER award DMS-2143215, CCF-2106778, and the Google Research Scholar Award. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009 and CCF-1907661. Y. Chi are supported in part by the grants ONR N00014-19-1-2404, NSF CCF-2106778, DMS-2134080 and CNS-2148212.

A Proof for several preliminary facts

A.1 Proof of properties (36)

Elementary calculations reveal that: the (i, j)-th entry of $J_t(x)$ is given by

$$\begin{aligned}\big[J_t(x)\big]_{i,j} = \mathbb{1}\{i=j\} + \frac{1}{1-\overline{\alpha}_t}\bigg\{ \Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_i - \sqrt{\overline{\alpha}_t}x_{0,i}\big)\mathrm{d}x_0\Big)\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_j - \sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\Big) \\ - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_i - \sqrt{\overline{\alpha}_t}x_{0,i}\big)\big(x_j - \sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\bigg\}. \qquad (73)\end{aligned}$$

This immediately establishes the matrix expression (36).

A.2 Proof of properties (37) regarding the learning rates

Proof of property (37a). From the choice of β_t in (20), we have

$$\alpha_t = 1 - \beta_t \geq 1 - \frac{c_1 \log T}{T} \geq \frac{1}{2}, \qquad t \geq 2.$$

The case with $t = 1$ holds trivially since $\beta_1 = 1/T^{c_0}$ for some large enough constant $c_0 > 0$.

Proof of properties (37b) and (37c). We start by proving (37b). Let τ be an integer obeying

$$\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{\tau}\leq 1<\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{\tau+1}, \tag{74}$$

and we divide into two cases based on τ.

- Consider any t satisfying $t\leq\tau$. In this case, it suffices to prove that

$$1-\overline{\alpha}_{t-1}\geq\frac{1}{3}\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{t}. \tag{75}$$

  Clearly, if (75) is valid, then any $t\leq\tau$ obeys

$$\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}=\frac{\beta_t}{1-\overline{\alpha}_{t-1}}\leq\frac{\frac{c_1\log T}{T}\beta_1\big(1+\frac{c_1\log T}{T}\big)^{t}}{\frac{1}{3}\beta_1\big(1+\frac{c_1\log T}{T}\big)^{t}}=\frac{3c_1\log T}{T}$$

  as claimed. Towards proving (75), first note that the base case with $t=2$ holds true trivially since $1-\overline{\alpha}_1=1-\alpha_1=\beta_1\geq\beta_1\big(1+\frac{c_1\log T}{T}\big)^2/3$. Next, let $t_0>2$ be *the first time* that Condition (75) fails to hold and suppose that $t_0\leq\tau$. It then follows that

$$1-\overline{\alpha}_{t_0-2}=1-\frac{\overline{\alpha}_{t_0-1}}{\alpha_{t_0-1}}\leq1-\overline{\alpha}_{t_0-1}<\frac{1}{3}\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{t_0}\leq\frac{1}{2}\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{t_0-1}<\frac{1}{2}, \tag{76}$$

  where the last inequality result from (74) and the assumption $t_0\leq\tau$. This taken together with the assumptions (75) and $t_0\leq\tau$ implies that

$$\frac{(1-\alpha_{t_0-1})\overline{\alpha}_{t_0-1}}{1-\overline{\alpha}_{t_0-2}}\geq\frac{\frac{c_1\log T}{T}\beta_1\min\Big\{\big(1+\frac{c_1\log T}{T}\big)^{t_0-1},1\Big\}\cdot\big(1-\frac{1}{2}\big)}{\frac{1}{2}\beta_1\big(1+\frac{c_1\log T}{T}\big)^{t_0-1}}=\frac{\frac{c_1\log T}{T}\beta_1\big(1+\frac{c_1\log T}{T}\big)^{t_0-1}}{\beta_1\big(1+\frac{c_1\log T}{T}\big)^{t_0-1}}=\frac{c_1\log T}{T}.$$

  As a result, we can further derive

$$\begin{aligned}1-\overline{\alpha}_{t_0-1}&=1-\alpha_{t_0-1}\overline{\alpha}_{t_0-2}=1-\overline{\alpha}_{t_0-2}+(1-\alpha_{t_0-1})\overline{\alpha}_{t_0-2}\\&=\Big(1+\frac{(1-\alpha_{t_0-1})\overline{\alpha}_{t_0-2}}{1-\overline{\alpha}_{t_0-2}}\Big)(1-\overline{\alpha}_{t_0-2})\\&\geq\Big(1+\frac{c_1\log T}{T}\Big)(1-\overline{\alpha}_{t_0-2})\geq\Big(1+\frac{c_1\log T}{T}\Big)\cdot\Big\{\frac{1}{3}\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{t_0-1}\Big\}\\&=\frac{1}{3}\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{t_0},\end{aligned}$$

  where the penultimate line holds since (75) is first violated at $t=t_0$; this, however, contradicts with the definition of t_0. Consequently, one must have $t_0>\tau$, meaning that (75) holds for all $t\leq\tau$.

- We then turn attention to those t obeying $t>\tau$. In this case, it suffices to make the observation that

$$1-\overline{\alpha}_{t-1}\geq1-\overline{\alpha}_{\tau-1}\geq\frac{1}{3}\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^{\tau}=\frac{\frac{1}{3}\beta_1\big(1+\frac{c_1\log T}{T}\big)^{\tau+1}}{1+\frac{c_1\log T}{T}}\geq\frac{1}{4}, \tag{77}$$

  where the second and the third inequalities come from (75). Therefore, one obtains

$$\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}\leq\frac{\frac{c_1\log T}{T}}{1/4}\leq\frac{4c_1\log T}{T}.$$

The above arguments taken together establish property (37b).

In addition, it comes immediately from (37b) that

$$1\leq\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}=1+\frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}=1+\frac{\overline{\alpha}_{t-1}(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\leq1+\frac{4c_1\log T}{T},$$

thereby justifying property (37c).

Proof of property (37d). Turning attention to the second claim (37d), we note that for any t obeying $t \geq \frac{T}{2} \gtrsim \frac{T}{\log T}$, one has

$$1 - \alpha_t = \frac{c_1 \log T}{T} \min \left\{ \beta_1 \Big(1 + \frac{c_1 \log T}{T}\Big)^t, 1 \right\} = \frac{c_1 \log T}{T}.$$

This in turn allows one to deduce that

$$\overline{\alpha}_T \leq \prod_{t: t \geq T/2} \alpha_t \leq \Big(1 - \frac{c_1 \log T}{T}\Big)^{T/2} \leq \frac{1}{T^{c_2}}$$

for an arbitrarily large constant $c_2 > 0$.

Proof of property (37e). Finally, it is easily seen from the Taylor expansion that the learning rates $\{\alpha_t\}$ satisfy

$$\begin{aligned}
\Big(\frac{1 - \overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\Big)^{d/2} &= \left(1 + \frac{1 - \alpha_t}{\alpha_t - \overline{\alpha}_t}\right)^{d/2} \\
&= 1 + \frac{d(1 - \alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + \frac{d(d-2)(1 - \alpha_t)^2}{8(\alpha_t - \overline{\alpha}_t)^2} + O\left(d^3 \Big(\frac{1 - \alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\right),
\end{aligned}$$

provided that $\frac{d(1-\alpha_t)}{\alpha_t - \overline{\alpha}_t} \lesssim 1$.

A.3 Proof of Lemma 1

To establish this lemma, we first make the following claim, whose proof is deferred to the end of this subsection.

Claim 1. *Consider any $c_5 \geq 5$. There exist some $x_0 \in \mathbb{R}$ such that*

$$\|\sqrt{\overline{\alpha}_t} x_0 - y\|_2 \leq c_5 \sqrt{d(1 - \overline{\alpha}_t) \log T} \qquad \text{and} \tag{78a}$$

$$\mathbb{P}\big(\|X_0 - x_0\|_2 \leq \epsilon\big) \geq \left(\frac{\epsilon}{2T^{c_6 + c_R}}\right)^d \qquad \text{with} \quad \epsilon = \frac{1}{T^{c_0/2}} \tag{78b}$$

hold simultaneously, where c_0 is defined in (20) and c_R is defined in (38).

With the above claim in place, we are ready to prove Lemma 1. For notational simplicity, we let X represent a random vector whose distribution $p_X(\cdot)$ obeys

$$p_X(x) = p_{X_0 | X_t}(x \,|\, y). \tag{79}$$

Let us look at a set:

$$\mathcal{E} := \left\{ x : \sqrt{\overline{\alpha}_t} \|x - x_0\|_2 > 4c_5 \sqrt{d(1 - \overline{\alpha}_t) \log T} \right\},$$

where $c_5 \geq 5$ (see Claim 1). Combining this with (78a) results in

$$\mathbb{P}\Big(\|\sqrt{\overline{\alpha}_t} X - y\|_2 > 5c_5 \sqrt{d(1 - \overline{\alpha}_t) \log T}\Big) \leq \mathbb{P}(X \in \mathcal{E}). \tag{80}$$

Consequently, everything boils down to bounding $\mathbb{P}(X \in \mathcal{E})$. Towards this, we first invoke the Bayes rule $p_{X_0 \,|\, X_t}(x \,|\, y) \propto p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x)$ to derive

$$\begin{aligned}
\mathbb{P}(X_0 \in \mathcal{E} \,|\, X_t = y) &= \frac{\int_{x \in \mathcal{E}} p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x}{\int_x p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x} \\
&\leq \frac{\int_{x \in \mathcal{E}} p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x}{\int_{x : \|x - x_0\|_2 \leq \epsilon} p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x}
\end{aligned}$$

$$\leq \frac{\sup_{x\in\mathcal{E}} p_{X_t \,|\, X_0}(y \,|\, x)}{\inf_{x:\|x-x_0\|_2\leq\epsilon} p_{X_t \,|\, X_0}(y \,|\, x)} \cdot \frac{\mathbb{P}(X_0 \in \mathcal{E})}{\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon)}. \tag{81}$$

To further bound this quantity, note that: in view of the definition of $\mathcal{E}$ and expression (78a), one has

$$\begin{aligned}
\sup_{x\in\mathcal{E}} p_{X_t \,|\, X_0}(y \,|\, x) &= \sup_{x:\|\sqrt{\overline{\alpha}_t}x-\sqrt{\overline{\alpha}_t}x_0\|_2>4c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_t \,|\, X_0}(y \,|\, x) \\
&\leq \sup_{x:\|\sqrt{\overline{\alpha}_t}x-y\|_2>3c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_t \,|\, X_0}(y \,|\, x) \\
&\leq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\Big(-\frac{9c_5^2 d\log T}{2}\Big)
\end{aligned}$$

and

$$\begin{aligned}
\inf_{x:\|x-x_0\|_2\leq\epsilon} p_{X_t \,|\, X_0}(y \,|\, x) &\geq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \inf_{x:\|x-x_0\|_2\leq\epsilon} \exp\Big(-\frac{\|y-\sqrt{\overline{\alpha}_t}x\|_2^2}{2(1-\overline{\alpha}_t)}\Big) \\
&\geq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \inf_{x:\|x-x_0\|_2\leq\epsilon} \exp\Big(-\frac{\|y-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{1-\overline{\alpha}_t} - \frac{\|\sqrt{\overline{\alpha}_t}x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{1-\overline{\alpha}_t}\Big) \\
&\geq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\Big(-\frac{\|y-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{1-\overline{\alpha}_t} - \frac{\epsilon^2}{1-\overline{\alpha}_t}\Big) \\
&\geq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\Big(-c_5^2 d\log T - \frac{1}{T^{c_0}}\frac{1}{1-\overline{\alpha}_t}\Big) \\
&\geq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\big(-2c_5^2 d\log T\big),
\end{aligned}$$

where the penultimate line relies on (78a), and the last line holds true since $1-\overline{\alpha}_t \geq 1-\alpha_1 = 1/T^{c_0}$ (see (20)). Substitution of the above two displays into (81), we arrive at

$$\begin{aligned}
\mathbb{P}(X_0 \in \mathcal{E} \,|\, X_t = y) &\leq \exp(-2.5c_5^2 d\log T) \cdot \frac{1}{\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon)} \\
&\leq \exp(-2.5c_5^2 d\log T) \cdot \Big(2T^{c_6+c_0/2+c_R}\Big)^d \\
&\leq \exp\big(-(2c_5^2 - c_6 - c_0/2 - c_R)d\log T\big).
\end{aligned} \tag{82}$$

Substituting this into (80) and recalling the distribution (79) of X, we arrive at

$$\mathbb{P}\Big(\|\sqrt{\overline{\alpha}_t}X - y\|_2 > 5c_5\sqrt{d\log T(1-\overline{\alpha}_t)}\Big) \leq \exp\big(-(2c_5^2 - c_6 - c_0/2 - c_R)d\log T\big).$$

This concludes the proof of the advertised result (40) for sufficiently large c_5, as long as Claim 1 can be justified.

With the above result in place, it then follows that

$$\begin{aligned}
&\mathbb{E}\left[\big\|x_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2 \,|\, X_t = x_t\right] \\
&\quad\leq 5\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T} + \mathbb{E}\left[\big\|x_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2 \mathbb{1}\left\{\|x_t - \sqrt{\overline{\alpha}_t}X_0\|_2 \geq 5\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T}\right\} \,\Big|\, X_t = x_t\right] \\
&\quad\leq 5\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T} + \int_{5\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T}}^{\infty} \mathbb{P}\big(\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2 \geq \tau \,|\, X_t = x_t\big)\mathrm{d}\tau \\
&\quad\leq 5\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T} + \int_{5\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T}}^{\infty} \exp\Big(-\frac{\tau^2}{25(1-\overline{\alpha}_t)}\Big)\mathrm{d}\tau \\
&\quad\leq 5\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T} + \exp\big(-\overline{c}_5^2 d\log T\big)
\end{aligned}$$

$$\leq 6\overline{c}_5 \sqrt{d(1-\overline{\alpha}_t)\log T},$$

as claimed in (41a). The proofs for (41b), (41c) and (41d) follow from similar aguments and are hence omitted for the sake of brevity.

Proof of Claim 1. We prove this claim by contradiction. Specifically, suppose instead that: for every x obeying $\|\sqrt{\overline{\alpha}_t}x - y\|_2 \leq c_5\sqrt{d(1-\overline{\alpha}_t)\log T}$, we have

$$\mathbb{P}(\|X_0 - x\|_2 \leq \epsilon) \leq \left(\frac{\epsilon}{2T^{c_6}R}\right)^d \qquad \text{with } \epsilon = \frac{1}{T^{c_0/2}}. \tag{83}$$

Clearly, the choice of ϵ ensures that $\epsilon < \frac{1}{2}\sqrt{d(1-\overline{\alpha}_t)\log T}$. In the following, we would like to show that this assumption leads to contradiction.

First of all, let us look at p_{X_t}, which obeys

$$\begin{aligned}
p_{X_t}(y) &= \int_x p_{X_0}(x) p_{X_t \mid X_0}(y \mid x) \mathrm{d}x \\
&= \int_{x:\|\sqrt{\overline{\alpha}_t}x - y\|_2 \geq c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x) p_{X_t \mid X_0}(y \mid x) \mathrm{d}x \\
&\quad + \int_{x:\|\sqrt{\overline{\alpha}_t}x - y\|_2 < c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x) p_{X_t \mid X_0}(y \mid x) \mathrm{d}x \\
&\leq \int_{x:\|\sqrt{\overline{\alpha}_t}x - y\|_2 \geq c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_t \mid X_0}(y \mid x) \mathrm{d}x + \int_{x:\|\sqrt{\overline{\alpha}_t}x - y\|_2 < c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x) \mathrm{d}x.
\end{aligned} \tag{84}$$

To further control (84), we make two observations:

1) The first term on the right-hand side of (84) can be bounded by

$$\begin{aligned}
&\int_{x:\|\sqrt{\overline{\alpha}_t}x - y\|_2 \geq c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_t \mid X_0}(y \mid x) \mathrm{d}x \\
&\quad = \int_{z:\|z\|_2 \geq c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\left(-\frac{\|z\|_2^2}{2(1-\overline{\alpha}_t)}\right) \mathrm{d}z \\
&\quad \leq \frac{1}{2}\exp\big(-c_6 d \log T\big),
\end{aligned} \tag{85}$$

   for some constant $c_6 > 0$, provided that c_5 is sufficiently large. Here, we have used $X_t \overset{\text{(i)}}{=} \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W$ with $W \sim \mathcal{N}(0, I_d)$ as well as standard properties about Gaussian distributions.

2) Regarding the second term on the right-hand side of (84), let us construct an epsilon-net $\mathcal{N}_\epsilon = \{z_i\}$ for the set

$$\big\{x : \|\sqrt{\overline{\alpha}_t}x - y\|_2 \leq c_5\sqrt{d(1-\overline{\alpha}_t)\log T} \text{ and } \|x\|_2 \leq R\big\},$$

   so that for each x in this set, one can find a vector $z_i \in \mathcal{N}_\epsilon$ such that $\|x - z_i\|_2 \leq \epsilon$. Define $\mathcal{B}_i \coloneqq \{x \mid \|x - z_i\|_2 \leq \epsilon\}$ for each $z_i \in \mathcal{N}_\epsilon$. Armed with these sets, we can derive

$$\begin{aligned}
\int_{x:\|\sqrt{\overline{\alpha}_t}x - y\|_2 < 5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x)\mathrm{d}x &\leq \sum_{i=1}^{|\mathcal{N}_\epsilon|} \mathbb{P}(X_0 \in \mathcal{B}_i) \\
&\leq \left(\frac{\epsilon}{2T^{c_6}R}\right)^d \left(\frac{R}{\epsilon}\right)^d \\
&\leq \frac{1}{2}\exp(-c_6 d \log T),
\end{aligned}$$

   where the penultimate step comes from the assumption (83).

The above results taken collectively lead to

$$p_{X_t}(y) \leq \exp(-c_6 d \log T), \tag{86}$$

thus contradicting the assumption (39). This in turn validates this claim. □

A.4 Proof of Lemma 2

For notational convenience, let us denote $\widehat{x}_t = x_t/\sqrt{\alpha_t}$ throughout the proof. As a key step of the proof, we note that for any $x \in \mathbb{R}^d$,

$$\begin{aligned} p_{X_{t-1}}(x) &= \int_{x_0} p_{X_0}(x_0) p_{X_{t-1} \,|\, X_0}(x \,|\, x_0) \mathrm{d}x_0 \\ &= \int_{x_0} p_{X_0}(x_0) p_{X_t \,|\, X_0}(x_t \,|\, x_0) \cdot \frac{p_{X_{t-1} \,|\, X_0}(x \,|\, x_0)}{p_{X_t \,|\, X_0}(x_t \,|\, x_0)} \mathrm{d}x_0 \\ &= p_{X_t}(x_t) \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \cdot \frac{p_{X_{t-1} \,|\, X_0}(x \,|\, x_0)}{p_{X_t \,|\, X_0}(x_t \,|\, x_0)} \mathrm{d}x_0, \end{aligned} \tag{87}$$

thus establishing a link between $p_{X_{t-1}}(x)$ and $p_{X_t}(x_t)$. Consequently, in order to control $p_{X_{t-1}}(x)$, it is helpful to first look at the density ratio $\frac{p_{X_{t-1} \,|\, X_0}(x \,|\, x_0)}{p_{X_t \,|\, X_0}(x_t \,|\, x_0)}$, which we accomplish in the sequel.

Recall that $X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}}\, X_0 + \sqrt{1-\overline{\alpha}}\, W$ with $W \sim \mathcal{N}(0, I_d)$. In what follows, let us consider any $x_t, x_0 \in \mathbb{R}^d$ and any x obeying

$$\|\widehat{x}_t - x\|_2 \leq c_3 \sqrt{d(1-\alpha_t)\log T} \tag{88}$$

for some constant $c_3 > 0$. The density ratio of interest satisfies

$$\begin{aligned} \frac{p_{X_{t-1} \,|\, X_0}(x \,|\, x_0)}{p_{X_t \,|\, X_0}(x_t \,|\, x_0)} &= \left(\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}\right)^{d/2} \exp\left(\frac{\big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\big\|x - \sqrt{\overline{\alpha}_{t-1}} x_0\big\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right) \\ &\leq \exp\left(\frac{d(1-\alpha_t) + \big\|\widehat{x}_t - x\big\|_2^2 + 2\big\|\widehat{x}_t - x\big\|_2 \big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2}{2(1-\overline{\alpha}_{t-1})} + \frac{(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2}\right), \end{aligned} \tag{89}$$

where the last inequality follows from the two relations below:

$$\log \frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} = \log\left(1 + \frac{\overline{\alpha}_{t-1}(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\right) \leq \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}$$

and

$$\begin{aligned} &\left|\frac{\big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\big\|x - \sqrt{\overline{\alpha}_{t-1}} x_0\big\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right| \\ &= \left|\frac{\big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\big\|\widehat{x}_t - x\big\|_2^2 + \big\|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}} x_0\big\|_2^2 - 2\big\langle \widehat{x}_t - x, \widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}} x_0\big\rangle}{2(1-\overline{\alpha}_{t-1})}\right| \\ &\leq \frac{\big\|\widehat{x}_t - x\big\|_2^2}{2(1-\overline{\alpha}_{t-1})} + \frac{\big\|\widehat{x}_t - x\big\|_2 \big\|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}} x_0\big\|_2}{1-\overline{\alpha}_{t-1}} + \frac{(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{2\alpha_t(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t-1})} \\ &\leq \frac{\big\|\widehat{x}_t - x\big\|_2^2 + 2\big\|\widehat{x}_t - x\big\|_2 \big\|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}} x_0\big\|_2}{2(1-\overline{\alpha}_{t-1})} + \frac{(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2}. \end{aligned}$$

Next, let us define the following set given x_t:

$$\widetilde{\mathcal{E}} := \left\{x_0 : \big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2 \leq c_4 \sqrt{d(1-\overline{\alpha}_t)\log T}\right\} \tag{90}$$

for some large enough numerical constant $c_4 > 0$, and we shall look at $\widetilde{\mathcal{E}}$ and $\widetilde{\mathcal{E}}^{\mathrm{c}}$ separately. Towards this, we make the following observations:

- For any $x_0 \in \widetilde{\mathcal{E}}$, one can utilize (90) and (88) to deduce that

$$\frac{d(1-\alpha_t) + \big\|\widehat{x}_t - x\big\|_2^2 + 2\big\|\widehat{x}_t - x\big\|_2 \big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2}{2(1-\overline{\alpha}_{t-1})} + \frac{(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2}$$

$$
\begin{aligned}
&\lesssim \frac{d(1-\alpha_t)}{1-\overline{\alpha}_{t-1}} + \frac{d(1-\alpha_t)\log T + d\sqrt{(1-\overline{\alpha}_t)(1-\alpha_t)}\log T}{1-\overline{\alpha}_{t-1}} + \frac{(1-\alpha_t)^2 d(1-\overline{\alpha}_t)\log T}{(1-\overline{\alpha}_{t-1})^2} \\
&\lesssim (d\log T)\left\{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} + \sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\sqrt{\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}} + \left(\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}\right)^2\right\} \\
&\lesssim d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T, \qquad (91)
\end{aligned}
$$

where the last inequality makes use of the facts $\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \leq 1$ (cf. (37b)) and

$$
\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} = \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} + \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} = \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} + \alpha_t \leq 2. \qquad (92)
$$

Moreover, the properties (37) of the stepsizes tell us that

$$
d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T \lesssim d\sqrt{\frac{\log^3 T}{T}} \leq c_{10}
$$

for some small enough constant $c_{10} > 0$, as long as $T \geq c_{11} d^2 \log^3 T$ for some sufficiently large constant $c_{11} > 0$. Taking this together with (89) and (91) reveals that

$$
\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} = 1 + O\bigg(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T\bigg), \qquad (93)
$$

with the proviso that (88) holds and $x_0 \in \widetilde{\mathcal{E}}$.

- Instead, if $x_0 \notin \widetilde{\mathcal{E}}$, then one can obtain

$$
\begin{aligned}
&\frac{d(1-\alpha_t) + \big\|\widehat{x}_t - x\big\|_2^2 + 2\big\|\widehat{x}_t - x\big\|_2\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2}{2(1-\overline{\alpha}_{t-1})} + \frac{(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2} \\
&\overset{\text{(i)}}{\leq} \frac{d(1-\alpha_t) + \big(1+\frac{1-\overline{\alpha}_{t-1}}{1-\alpha_t}\big)\big\|\widehat{x}_t - x\big\|_2^2 + \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_{t-1})} + \frac{(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2} \\
&\overset{\text{(ii)}}{\leq} \frac{d(1-\alpha_t) + c_3\big(1+\frac{1-\overline{\alpha}_{t-1}}{1-\alpha_t}\big)d(1-\alpha_t)\log T}{2(1-\overline{\alpha}_{t-1})} + \frac{2(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2} \\
&\overset{\text{(iii)}}{\leq} \frac{c_3 d\big(1-\alpha_t+1-\overline{\alpha}_{t-1}\big)\log T}{1-\overline{\alpha}_{t-1}} \cdot \frac{\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{c_4 d(1-\overline{\alpha}_{t-1})\log T} + \frac{2(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2} \\
&\leq \frac{2c_3}{c_4} \cdot \frac{\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{1-\overline{\alpha}_{t-1}} + \frac{8c_1\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\log T}{T(1-\overline{\alpha}_{t-1})} \\
&\overset{\text{(iv)}}{\leq} \left(\frac{2c_3}{c_4} + \frac{8c_1\log T}{T}\right)\frac{2\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{1-\overline{\alpha}_t} \leq \frac{8c_3}{c_4}\frac{\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{1-\overline{\alpha}_t},
\end{aligned}
$$

where (i) comes from the Cauchy-Schwarz inequality, (ii) is valid due to the assumption on $\|\widehat{x}_t - x\|_2$, (iii) follows from the definition of $\widetilde{\mathcal{E}}$, and (iv) is a consequence of (92). Substitution into (89) leads to

$$
\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} \leq \exp\left(\frac{8c_3}{c_4}\frac{\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{1-\overline{\alpha}_t}\right), \qquad (94)
$$

provided that (88) holds and $x_0 \notin \widetilde{\mathcal{E}}$.

In light of the above calculations, we can invoke (95) to demonstrate that: for any x obeying (88),

$$
\begin{aligned}
p_{X_{t-1}}(x) &= p_{X_t}(x_t)\left(\int_{x_0\in\widetilde{\mathcal{E}}}+\int_{x_0\notin\widetilde{\mathcal{E}}}\right)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\cdot\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)}\mathrm{d}x_0\\
&= p_{X_t}(x_t)\int_{x_0\in\widetilde{\mathcal{E}}}\left(1+O\left(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T\right)\right)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\\
&\quad + p_{X_t}(x_t)\int_{x_0\notin\widetilde{\mathcal{E}}}O\left(\exp\left(\frac{8c_3}{c_4}\frac{\left\|x_t-\sqrt{\overline{\alpha}_t}x_0\right\|_2^2}{1-\overline{\alpha}_t}\right)\right)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0. \qquad (95)
\end{aligned}
$$

By virtue of Lemma 1, if $-\log p_{X_t}(x_t)\le\frac{1}{2}c_6 d\log T$ for some large constant $c_6>0$, then it holds that

$$
\mathbb{P}\left\{\left\|\sqrt{\overline{\alpha}_t}X_0-x_t\right\|_2\ge 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T}\mid X_t=x_t\right\}\le\exp(-c_5^2 d\log T) \qquad (96)
$$

for any large enough $c_5>0$. Some elementary calculation then reveals that

$$
\int_{x_0\notin\widetilde{\mathcal{E}}}\exp\left(\frac{8c_3}{c_4}\frac{\left\|x_t-\sqrt{\overline{\alpha}_t}x_0\right\|_2^2}{1-\overline{\alpha}_t}\right)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\lesssim\frac{1}{T^{c_0}} \qquad (97)
$$

with c_0 defined in (20), provided that c_4/c_3 is sufficiently large. Hence, substituting it into (95) demonstrates that, for any x satisfying (88),

$$
\begin{aligned}
p_{X_{t-1}}(x) &= \left(1+O\left(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T\right)\right)(1-o(1))p_{X_t}(x_t)+O\left(\frac{1}{T^{c_0}}\right)p_{X_t}(x_t) \qquad (98)\\
&= \left(1+O\left(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T+\frac{1}{T^{c_0}}\right)\right)p_{X_t}(x_t)\\
&= \left(1+O\left(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T\right)\right)p_{X_t}(x_t) \qquad (99)\\
&\in\left[\frac{1}{2}p_{X_t}(x_t),\ \frac{3}{2}p_{X_t}(x_t)\right], \qquad (100)
\end{aligned}
$$

where the penultimate inequality holds since (according to (20))

$$
\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\ge\sqrt{1-\alpha_t}=\sqrt{\beta_t}\ge\sqrt{\frac{c_1\log T}{T^{c_0+1}}},
$$

and the last inequality holds due to (37b) and the condition that $T/(d^2\log^3 T)$ is sufficiently large. In other words, (100) reveals that $p_{X_{t-1}}(x)$ is sufficiently close to $p_{X_t}(x_t)$.

We are now ready to establish our claim. In view of the assumption (42), we have

$$
\left\|x_t(\gamma)-\widehat{x}_t\right\|_2=\left\|\gamma x_{t-1}+(1-\gamma)\widehat{x}_t-\widehat{x}_t\right\|_2=\gamma\left\|x_{t-1}-\widehat{x}_t\right\|_2\le c_3\sqrt{d(1-\alpha_t)\log T}.
$$

Therefore, taking x to be $x_t(\gamma)$ in (100) tells us that: if $-\log p_{X_t}(x_t)\le\frac{1}{2}c_6 d\log T$, then

$$
-\log p_{X_{t-1}}\big(x_t(\gamma)\big)\le-\log p_{X_t}(x_t)+\log 2\le c_6 d\log T
$$

as claimed.

B Proof of auxiliary lemmas for the ODE-based sampler

B.1 Proof of Lemma 3

B.1.1 Proof of relation (49a)

Recall the definition of ϕ_t in (46), and introduce the following vector:

$$u := x - \phi_t(x) = \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \int_{x_0} \big(x - \sqrt{\overline{\alpha}_t} x_0\big) p_{X_0 \,|\, X_t}(x_0 \,|\, x) \mathrm{d}x_0. \tag{101}$$

The proof consists of the following steps.

Step 1: decomposing $p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)/p_{X_t}(x)$. Recognizing that

$$X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\, W \qquad \text{with } W \sim \mathcal{N}(0, I_d) \tag{102}$$

and making use of the Bayes rule, we can express the conditional distribution $p_{X_0 \,|\, X_t}\big(\phi_t(x)\big)$ as

$$p_{X_0 \,|\, X_t}(x_0 \,|\, x) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} p_{X_t \,|\, X_0}(x \,|\, x_0) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} \cdot \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\bigg(-\frac{\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)}\bigg). \tag{103}$$

Moreover, it follows from (102) that

$$\sqrt{\alpha_t} X_{t-1} \overset{\mathrm{d}}{=} \sqrt{\alpha_t}\big(\sqrt{\overline{\alpha}_{t-1}} X_0 + \sqrt{1-\overline{\alpha}_{t-1}}\, W\big) = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{\alpha_t - \overline{\alpha}_t}\, W. \tag{104}$$

These taken together allow one to rewrite $p_{\sqrt{\alpha_t}X_{t-1}}$ such that:

$$\begin{aligned}
\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} &\overset{\text{(i)}}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{\big(2\pi(\alpha_t - \overline{\alpha}_t)\big)^{d/2}} \exp\bigg(-\frac{\big\|\phi_t(x) - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)}\bigg) \mathrm{d}x_0 \\
&\overset{\text{(ii)}}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{\big(2\pi(\alpha_t - \overline{\alpha}_t)\big)^{d/2}} \exp\bigg(-\frac{\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)}\bigg) \\
&\qquad \cdot \exp\bigg(-\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)}\bigg) \mathrm{d}x_0 \\
&\overset{\text{(iii)}}{=} \Big(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\Big)^{d/2} \cdot \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \cdot \\
&\qquad \exp\bigg(-\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)}\bigg) \mathrm{d}x_0 \\
&\overset{\text{(iv)}}{=} \bigg\{1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2\bigg)\bigg\} \cdot \\
&\qquad \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \exp\bigg(-\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)}\bigg) \mathrm{d}x_0.
\end{aligned} \tag{105}$$

Here, identity (i) holds due to (104) and hence

$$p_{\sqrt{\alpha_t}X_{t-1}}(x) = \int_{x_0} p_{X_0}(x_0) p_{\sqrt{\alpha_t - \overline{\alpha}_t}W}\big(x - \sqrt{\overline{\alpha}_t}x_0\big) \mathrm{d}x_0;$$

identity (ii) follows from (101) and elementary algebra; relation (iii) is a consequence of the Bayes rule (103); and relation (iv) results from (37e).

Step 2: controlling the integral in the decomposition (105). In order to further control the right-hand side of expression (105), we need to evaluate the integral in (105). To this end, we make a few observations.

- To begin with, Lemma 1 tells us that

$$\mathbb{P}\Big(\big\|\sqrt{\overline{\alpha}_t}X_0 - x\big\|_2 > 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T} \mid X_t = x\Big) \le \exp\big(-c_5^2 d\log T\big) \tag{106a}$$

  for any large enough c_5, provided that x satisfies $-\log p_{X_t}(x) \le 2d\log T$.

- A little algebra based on this relation allows one to bound u (cf. (101)) as follows:

$$\begin{aligned}
\|u\|_2 &\le \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\int_{x_0:\|x-\sqrt{\overline{\alpha}_t}x_0\|_2\le c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0\\
&\quad+\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\int_{x_0:\|x-\sqrt{\overline{\alpha}_t}x_0\|_2> c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0\\
&\le \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\cdot c_5\sqrt{d(1-\overline{\alpha}_t)\log T}+\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\cdot\int_{c_5}^{\infty}\Big(d(1-\overline{\alpha}_t)\log T\Big)\tau\exp\Big(-\frac{1}{25}\tau^2 d\log T\Big)\mathrm{d}\tau\\
&\le \frac{2c_5(1-\alpha_t)\sqrt{d\log T}}{3\sqrt{1-\overline{\alpha}_t}}\le\frac{2c_5}{3}\sqrt{d(1-\alpha_t)\log T},
\end{aligned}\tag{106b}$$

  with the proviso that $-\log p_{X_t}(x) \le 2d\log T$.

Equipped with the above properties, let us define

$$\mathcal{E} \coloneqq \Big\{x : \big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2 \le 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T}\Big\}. \tag{107}$$

For any $x\in\mathcal{E}$, the Taylor expansion $e^{-x} = 1-x+O(x^2)$ (for all $|x|<1$) gives

$$\begin{aligned}
&\exp\Bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|u\|_2^2-2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\Bigg)\\
&=1-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}+\frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t}+O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg)\\
&=1-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}+\frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t}+O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg),
\end{aligned}\tag{108}$$

where the penultimate line invokes (106), and the last line holds true since, according to (106b),

$$\frac{\|u\|_2^2}{|\alpha_t-\overline{\alpha}_t|}\le\frac{1}{\alpha_t-\overline{\alpha}_t}\cdot\frac{9c_5^2(1-\alpha_t)^2 d\log T}{(1-\overline{\alpha}_t)}\le\frac{9c_5^2(1-\alpha_t)^2 d\log T}{(\alpha_t-\overline{\alpha}_t)^2}.$$

In contrast, for any $x\notin\mathcal{E}$, we invoke the crude bound

$$\begin{aligned}
&-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}+\frac{2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\\
&\quad\le-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}+\frac{\frac{1-\alpha_t}{1-\overline{\alpha}_t}\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2+\frac{1-\overline{\alpha}_t}{1-\alpha_t}\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}\\
&\quad\le\frac{\Big(\frac{1-\overline{\alpha}_t}{1-\alpha_t}-1\Big)\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}=\frac{\|u\|_2^2}{2(1-\alpha_t)},
\end{aligned}$$

and as a result,

$$\exp\Bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|u\|_2^2-2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\Bigg)\le\exp\bigg(\frac{\|u\|_2^2}{2(1-\alpha_t)}\bigg)\le\exp\bigg(\frac{2c_5^2 d\log T}{9}\bigg).\tag{109}$$

Combine (108) and (109) to show that: if $\frac{d(1-\alpha_t)\log T}{\alpha_t-\overline{\alpha}_t} \lesssim 1$, then one has

$$\begin{aligned}
&\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \exp\bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&= \bigg(\int_{x_0\in\mathcal{E}} + \int_{x_0\notin\mathcal{E}}\bigg) p_{X_0 \,|\, X_t}(x_0 \,|\, x) \exp\bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&= \bigg(1+O\Big(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\Big)\bigg)\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\bigg(1-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} + \frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t}\bigg)\mathrm{d}x_0 \\
&\quad + O\bigg(\exp\bigg(\frac{2c_5^2 d\log T}{9}\bigg)\int_{x_0\notin\mathcal{E}} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\mathrm{d}x_0\bigg) \\
&= 1 - \frac{(1-\alpha_t)\big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_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)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} \\
&\quad + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg) + O\bigg(\exp\bigg(-\frac{7c_5^2 d\log T}{9}\bigg)\bigg) \\
&= 1 - \frac{(1-\alpha_t)\big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_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)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} \\
&\quad + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg),
\end{aligned} \tag{110}$$

where the last relation makes use of the definition (101) of u.

Step 3: putting everything together. Substitution of (110) into (105) yields

$$\begin{aligned}
\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} &= 1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg) - \\
&\qquad \frac{(1-\alpha_t)\big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0 - \big\|\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2\big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}
\end{aligned}$$

as claimed.

B.1.2 Proof of relation (49b)

Consider any random vector Y. To understand the density ratio $p_{\phi_t(Y)}(\phi_t(x))/p_Y(x)$, we make note of the transformation

$$p_{\phi_t(Y)}(\phi_t(x)) = \mathsf{det}\Big(\frac{\partial\phi_t(x)}{\partial x}\Big)^{-1} p_Y(x), \tag{111}$$

where $\frac{\partial\phi_t(x)}{\partial x}$ denotes the Jacobian matrix. It thus suffices to control the quantity $\mathsf{det}\Big(\frac{\partial\phi_t(x)}{\partial x}\Big)^{-1}$.

To begin with, recall from (46) and (34) that

$$\phi_t(x) = x - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}g_t(x).$$

As a result, one can use (35) and (36) derive

$$\begin{aligned}
I - \frac{\partial\phi_t(x)}{\partial x} = \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x) = \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\bigg\{I + \frac{1}{1-\overline{\alpha}_t}\Big\{\mathbb{E}\big[X_t-\sqrt{\overline{\alpha}_t}X_0 \mid X_t=x\big]\Big(\mathbb{E}\big[X_t-\sqrt{\overline{\alpha}_t}X_0 \mid X_t=x\big]\Big)^\top \\
- \mathbb{E}\Big[\big(X_t-\sqrt{\overline{\alpha}_t}X_0\big)\big(X_t-\sqrt{\overline{\alpha}_t}X_0\big)^\top \mid X_t=x\Big]\Big\}\bigg\}
\end{aligned}$$

$$=: \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\Big\{I + \frac{1}{1-\overline{\alpha}_t}B\Big\}. \tag{112}$$

This allows one to show that

$$\mathsf{Tr}\Big(I - \frac{\partial \phi_t(x)}{\partial x}\Big) = \frac{d(1-\alpha_t)}{2(1-\overline{\alpha}_t)} + \\ \frac{(1-\alpha_t)\big(\big\|\int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2 - \int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\big)}{2(1-\overline{\alpha}_t)^2}. \tag{113a}$$

Moreover, the matrix B defined in (112) satisfies

$$\|B\|_{\mathrm{F}} \leq \Big\|\mathbb{E}\Big[\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)^\top \mid X_t = x\Big]\Big\|_{\mathrm{F}} \leq \int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0$$

due to Jensen's inequality. Taking this together with (112) and Lemma 1 reveals that

$$\Big\|\frac{\partial \phi_t(x)}{\partial x} - I\Big\| \leq \Big\|\frac{\partial \phi_t(x)}{\partial x} - I\Big\|_{\mathrm{F}} \lesssim \frac{1-\alpha_t}{1-\overline{\alpha}_t}\bigg(\sqrt{d} + \frac{\int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0}{1-\overline{\alpha}_t}\bigg) \\ \lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t}. \tag{113b}$$

Additionally, the Taylor expansion guarantees that

$$\mathsf{det}\big(I + A\big) = 1 + \mathsf{Tr}(A) + O\big((\mathsf{Tr}(A))^2 + \|A\|_{\mathrm{F}}^2 + d^3\|A\|^3\big) \tag{114}$$

as long as $d\|A\| \lesssim 1$. The above properties taken collectively allow us to demonstrate that

$$\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)} = \mathsf{det}\Big(\frac{\partial \phi_t(x)}{\partial x}\Big)^{-1} \\ = 1 - \mathsf{Tr}\Big(\frac{\partial \phi_t(x)}{\partial x} - I\Big) + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2\log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\log^3 T\bigg) \\ = 1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + \frac{(1-\alpha_t)\Big(\big\|\int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2 - \int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\Big)}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} \\ + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2\log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\log^3 T\bigg), \tag{115}$$

with the fact that $\big|\frac{1-\alpha_t}{1-\overline{\alpha}_t} - \frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\big| \lesssim \big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\big)^2$and the proviso that $\frac{d^2(1-\alpha_t)\log T}{\alpha_t - \overline{\alpha}_t} \lesssim 1$.

C Proofs of auxiliary lemmas for the DDPM-type sampler

C.1 Proof of Lemma 4

For notational simplicity, we find it helpful to define, for any constant $\gamma \in [0,1]$,

$$x_t(\gamma) \coloneqq \gamma x_{t-1} + (1-\gamma)\widehat{x}_t \qquad \text{and} \qquad \widehat{x}_t \coloneqq \frac{1}{\sqrt{\alpha_t}}x_t. \tag{116}$$

Step 1: decomposing the target distribution $p_{X_{t-1} \mid X_t}(x_{t-1} \mid x_t)$. With this piece of notation in mind, we can recall the forward process (4) and calculate: for any $x_{t-1}, x_t \in \mathbb{R}^d$,

$$p_{X_{t-1} \mid X_t}(x_{t-1} \mid x_t) \\ = \frac{1}{p_{X_t}(x_t)}p_{X_{t-1},X_t}(x_{t-1}, x_t) = \frac{1}{p_{X_t}(x_t)}\exp\Big(\log p_{X_{t-1}}(x_{t-1}) + \log p_{X_t \mid X_{t-1}}(x_t \mid x_{t-1})\Big)$$

$$
\begin{aligned}
&= \frac{1}{p_{X_t}(x_t)} \exp\bigg(\log p_{X_{t-1}}(\widehat{x}_t) + \int_0^1 \Big[\nabla \log p_{X_{t-1}}\big(x_t(\gamma)\big)\Big]^\top (x_{t-1} - \widehat{x}_t)\mathrm{d}\gamma + \log p_{X_t \,|\, X_{t-1}}(x_t \,|\, x_{t-1}) \bigg) \\
&= \frac{p_{X_{t-1}}(\widehat{x}_t)}{p_{X_t}(x_t)} \exp\bigg((x_{t-1} - \widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1} \,|\, X_0}\big(x_t(\gamma) \,|\, x_0\big) p_{X_0}(x_0)}{p_{X_{t-1}}\big(x_t(\gamma)\big)} \mathrm{d}x_0 + \log p_{X_t \,|\, X_{t-1}}(x_t \,|\, x_{t-1}) \bigg),
\end{aligned} \tag{117}
$$

where the penultimate line comes from the fundamental theorem of calculus. In particular, the exponent in (117) consists of a term that satisfies

$$
\begin{aligned}
&(x_{t-1} - \widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1} \,|\, X_0}\big(x_t(\gamma) \,|\, x_0\big) p_{X_0}(x_0)}{p_{X_{t-1}}\big(x_t(\gamma)\big)} \mathrm{d}x_0 \\
&= (x_{t-1} - \widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1} \,|\, X_0}\big(x_t(\gamma) \,|\, x_0\big)}{p_{X_{t-1} \,|\, X_0}\big(x_t(\gamma) \,|\, x_0\big)} p_{X_0 \,|\, X_{t-1}}\big(x_0 \,|\, x_t(\gamma)\big) \mathrm{d}x_0 \\
&= -(x_{t-1} - \widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}} x_0}{1 - \overline{\alpha}_{t-1}} p_{X_0 \,|\, X_{t-1}}\big(x_0 \,|\, x_t(\gamma)\big) \mathrm{d}x_0 \\
&=: -\frac{1}{1 - \overline{\alpha}_{t-1}} (x_{t-1} - \widehat{x}_t)^\top \int_0^1 g_{t-1}\big(x_t(\gamma)\big) \mathrm{d}\gamma,
\end{aligned} \tag{118}
$$

where the second line holds since $p_{X_0 \,|\, X_{t-1}}\big(x_0 \,|\, x_t(\gamma)\big) p_{X_{t-1}}\big(x_t(\gamma)\big) = p_{X_{t-1} \,|\, X_0}\big(x_t(\gamma) \,|\, x_0\big) p_{X_0}(x_0)$, and we remind the reader of the definition of $g_t(\cdot)$ in (34).

To continue, it is then seen from the fundamental theorem of calculus that

$$
g_{t-1}\big(x_t(\gamma)\big) = g_{t-1}\big(\widehat{x}_t\big) + \int_0^1 J_{t-1}\Big((1-\tau)\widehat{x}_t + \tau x_t(\gamma)\Big)\big(x_t(\gamma) - \widehat{x}_t\big)\mathrm{d}\tau,
$$

where $J_{t-1}(x) := \frac{\partial g_{t-1}(x)}{\partial x} \in \mathbb{R}^{d \times d}$ is the associated Jacobian matrix. As a consequence, we can show that

$$
\begin{aligned}
&(x_{t-1} - \widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1} \,|\, X_0}\big(x_t(\gamma) \,|\, x_0\big) p_{X_0}(x_0)}{p_{X_{t-1}}\big(x_t(\gamma)\big)} \mathrm{d}x_0 \\
&= -\frac{1}{1 - \overline{\alpha}_{t-1}} \bigg\{ (x_{t-1} - \widehat{x}_t)^\top g_{t-1}\big(\widehat{x}_t\big) + (x_{t-1} - \widehat{x}_t)^\top \int_0^1 \int_0^1 J_{t-1}\Big((1-\tau)\widehat{x}_t + \tau x_t(\gamma)\Big)\big(x_t(\gamma) - \widehat{x}_t\big)\mathrm{d}\tau \mathrm{d}\gamma \bigg\} \\
&= -\frac{1}{1 - \overline{\alpha}_{t-1}} \bigg\{ (x_{t-1} - \widehat{x}_t)^\top g_{t-1}\big(\widehat{x}_t\big) + (x_{t-1} - \widehat{x}_t)^\top \bigg[\int_0^1 \int_0^1 \gamma J_{t-1}\Big((1-\tau)\widehat{x}_t + \tau x_t(\gamma)\Big)\mathrm{d}\tau \mathrm{d}\gamma\bigg] \big(x_{t-1} - \widehat{x}_t\big) \bigg\}.
\end{aligned} \tag{119}
$$

Combining (117) and (119) allows us to rewrite the target quantity $p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t)$ as:

$$
\begin{aligned}
&p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t) \\
&= \frac{p_{X_{t-1}}(\widehat{x}_t)}{p_{X_t}(x_t)} \exp\bigg((x_{t-1} - \widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1} \,|\, X_0}(\widetilde{x}_t \,|\, x_0) p_{X_0}(x_0)}{p_{X_{t-1}}(\widetilde{x}_t)} \mathrm{d}x_0 + \log p_{X_t \,|\, X_{t-1}}(x_t \,|\, x_{t-1}) \bigg) \\
&= \frac{p_{X_{t-1}}(\widehat{x}_t)}{p_{X_t}(x_t)} \exp\Bigg(-\frac{(x_{t-1} - \widehat{x}_t)^\top g_{t-1}\big(\widehat{x}_t\big) + (x_{t-1} - \widehat{x}_t)^\top \Big[\int_0^1 \int_0^1 \gamma J_{t-1}\Big((1-\tau)\widehat{x}_t + \tau x_t(\gamma)\Big)\mathrm{d}\tau \mathrm{d}\gamma\Big] \big(x_{t-1} - \widehat{x}_t\big)}{1 - \overline{\alpha}_{t-1}} \\
&\qquad - \frac{\alpha_t \|x_{t-1} - \widehat{x}_t\|_2^2}{2(1-\alpha_t)} - \frac{d}{2}\log\big(2\pi(1-\alpha)\big) \Bigg),
\end{aligned} \tag{120}
$$

where we have also used the fact that conditional on $X_t \,|\, X_{t-1} = x_{t-1} \sim \mathcal{N}\big(\sqrt{\overline{\alpha}_t} x_{t-1}, (1-\overline{\alpha}_t) I_d\big)$. Note that the pre-factor $\frac{p_{X_{t-1}}(\widehat{x}_t)}{p_{X_t}(x_t)}$ in the above display is independent from the specific value of x_{t-1}.

Step 2: controlling the exponent in (120). Consider now any $(x_t, x_{t-1}) \in \mathcal{E}$ (cf. (62)). In order to further simplify the exponent in the display (120), we make the following claims:

(a) for any x that can be written as $x = wx_{t-1} + (1-w)x_t/\sqrt{\alpha_t}$ for some $w \in [0,1]$, the Jacobian matrix $J_{t-1}(x) = \frac{\partial g_{t-1}(x)}{\partial x}$ obeys

$$\big\|J_{t-1}(x) - I\big\| \lesssim d\log T; \tag{121a}$$

(b) in addition, one has

$$\frac{1}{1-\overline{\alpha}_t}\Big\|\big(x_{t-1}-\widehat{x}_t\big)g_t(x_t) - \sqrt{\alpha_t}\big(x_{t-1}-\widehat{x}_t\big)g_t(x_t)\Big\|_2 \lesssim \frac{d\log^2 T}{T^{3/2}}. \tag{121b}$$

$$\left\|\frac{g_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{g_t(x_t)}{1-\overline{\alpha}_t}\right\|_2 \lesssim (1-\alpha_t)\Big(\frac{d\log T}{\alpha_t - \overline{\alpha}_t}\Big)^{3/2}, \tag{121c}$$

Assuming the validity of these claims (which will be established in Appendix C.1.1) and recalling the definition of $\mu_t(\cdot)$ in (56), we can use (120) together with a little algebra to obtain

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = f_0(x_t)\exp\left(-\frac{\alpha_t\|x_{t-1}-\mu_t(x_t)\|_2^2}{2(1-\alpha_t)} + \zeta_t(x_{t-1},x_t)\right) \tag{122}$$

for some function $f_0(\cdot)$ and some residual term $\zeta_t(x_{t-1},x_t)$ obeying

$$\begin{aligned}
|\zeta_t(x_{t-1},x_t)| &\lesssim \big\|x_{t-1}-\widehat{x}_t\big\|_2^2 \sup_{x:\,\exists w\in[0,1]\text{ s.t. } x=(1-w)\widehat{x}_t+wx_{t-1}} \big\|J_{t-1}(x)\big\| \\
&\quad + \big\|x_{t-1}-\widehat{x}_t\big\|_2\left\|\frac{\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(\widehat{x}_t-\sqrt{\overline{\alpha}_{t-1}}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_{t-1}} - \frac{\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_t}\right\|_2 \\
&\lesssim \frac{\big(d(1-\alpha_t)\log T\big)d\log T}{1-\overline{\alpha}_{t-1}} + \sqrt{d(1-\alpha_t)\log T}\,(1-\alpha_t)\Big(\frac{d\log T}{\alpha_t-\overline{\alpha}_t}\Big)^{3/2} \\
&\lesssim \frac{(1-\alpha_t)d^2\log^2 T}{1-\overline{\alpha}_{t-1}} + d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^{3/2}\log^2 T \asymp \frac{(1-\alpha_t)d^2\log^2 T}{\alpha_t-\overline{\alpha}_t},
\end{aligned}$$

where the penultimate line makes use of the assumption $(x_t, x_{t-1}) \in \mathcal{E}$ (cf. (62)), and the last inequality holds since $\alpha_t \geq 1/2$ (cf. (20)).

Step 3: approximating the function $f_0(x_t)$. To finish up, it remains to quantify the function $f_0(\cdot)$ in (122). Note that for any x_t obeying $p_{X_t}(x_t) \geq \exp\big(-\frac{1}{2}c_6 d\log T\big)$, it is easily seen that

$$\begin{aligned}
\int_{x_{t-1}:(x_t,x_{t-1})\notin\mathcal{E}} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\mathrm{d}x_{t-1} &= \frac{\int_{x_{t-1}:(x_t,x_{t-1})\notin\mathcal{E}} p_{X_t\,|\,X_{t-1}}(x_t\,|\,x_{t-1})p_{X_{t-1}}(x_{t-1})\mathrm{d}x_{t-1}}{p_{X_t}(x_t)} \\
&\leq \frac{\frac{1}{(2\pi(1-\alpha_t))^{d/2}}\int_{x_{t-1}:(x_t,x_{t-1})\notin\mathcal{E}} \exp\big(-\frac{\|x_t-\sqrt{\alpha_t}x_{t-1}\|_2^2}{2(1-\alpha_t)}\big)\mathrm{d}x_{t-1}}{\exp\big(-\frac{1}{2}c_6 d\log T\big)} \\
&\leq \frac{\exp\big(-c_3 d\log T\big)}{\exp\big(-\frac{1}{2}c_6 d\log T\big)} \leq \exp\Big(-\frac{1}{4}c_6 d\log T\Big),
\end{aligned}$$

provided that $c_3 \geq 3c_6/4$. This means that

$$1 \geq \int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\mathrm{d}x_{t-1} \geq 1 - \exp\left(-\frac{1}{4}c_6 d\log T\right). \tag{123}$$

Moreover, for any $(x_t, x_{t-1}) \in \mathcal{E}$, one has

$$\sqrt{\alpha_t}\big\|x_{t-1}-\mu_t(x_t)\big\|_2 = \sqrt{\alpha_t}\left\|x_{t-1}-\widehat{x}_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\mathbb{E}\Big[x_t - \sqrt{\overline{\alpha}_t}X_0 \mid X_t = x_t\Big]\right\|_2$$

$$\geq \sqrt{\alpha_t}\big\|x_{t-1} - \widehat{x}_t\big\|_2 - \frac{1-\alpha_t}{(1-\overline{\alpha}_t)}\mathbb{E}\Big[\big\|x_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2 \mid X_t = x_t\Big]$$
$$\geq c_3\sqrt{\alpha_t}\cdot\sqrt{d(1-\alpha_t)\log T} - \frac{1-\alpha_t}{1-\overline{\alpha}_t}6\overline{c}_5\sqrt{d(1-\overline{\alpha}_t)\log T}$$
$$\geq \frac{1}{2}c_3\sqrt{d(1-\alpha_t)\log T},$$

where the first identity comes from the definition (56) of $\mu_t(x_t)$, the penultimate line makes use of the result (41a) in Lemma 1, and the last inequality is valid as long as c_3 is sufficiently large. This in turn allows one to derive

$$\frac{1}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{d/2}}\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}}\exp\bigg(-\frac{\alpha_t\|x_{t-1}-\mu_t(x_t)\|_2^2}{2(1-\alpha_t)}\bigg)\mathrm{d}x_{t-1} \geq 1-\exp\big(-c_3 d\log T\big). \tag{124}$$

In addition, by virtue of (122), the integral in (123) can be respectively bounded from above and from below as follows:

$$\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\mathrm{d}x_{t-1}$$
$$= \frac{f_0(x_t)\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}}\exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1}-\mu_t(x_t)\big\|_2^2+\zeta_t(x_{t-1},x_t)\Big)\mathrm{d}x_{t-1}}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{-d/2}\int_{x_{t-1}}\exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1}-\mu_t(x_t)\big\|_2^2\Big)\mathrm{d}x_{t-1}}$$
$$\leq \frac{f_0(x_t)\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}}\exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1}-\mu_t(x_t)\big\|_2^2+\zeta_t(x_{t-1},x_t)\Big)\mathrm{d}x_{t-1}}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{-d/2}\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}}\exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1}-\mu_t(x_t)\big\|_2^2\Big)\mathrm{d}x_{t-1}}$$
$$\lesssim \frac{f_0(x_t)}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{-d/2}}\exp\bigg\{O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\log^2 T\bigg)\bigg\}$$

and

$$\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\mathrm{d}x_{t-1}$$
$$\geq \big(1-\exp(-c_3 d\log T)\big)\frac{f_0(x_t)\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}}\exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1}-\mu_t(x_t)\big\|_2^2+\zeta_t(x_{t-1},x_t)\Big)\mathrm{d}x_{t-1}}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{-d/2}\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}}\exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1}-\mu_t(x_t)\big\|_2^2\Big)\mathrm{d}x_{t-1}}$$
$$\geq \big(1-\exp(-c_3 d\log T)\big)\frac{f_0(x_t)}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{-d/2}}\exp\bigg\{-O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\log^2 T\bigg)\bigg\}.$$

These taken collectively with (123) allow one to demonstrate that

$$\max\bigg\{\frac{f_0(x_t)}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{-d/2}},\frac{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{-d/2}}{f_0(x_t)}\bigg\} = \exp\bigg\{O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\log^2 T\bigg)\bigg\} = 1+O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\log^2 T\bigg),$$

with the proviso that $d^2\big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\big)\log^2 T \lesssim 1$.

Combining this with (122) concludes the proof of Lemma 4, as long as the two claims in (121) are valid (to be justified in Appendix C.1.1).

C.1.1 Proof of auxiliary claims (121) in Lemma 4

Proof of relation (121a). Recall from (36) that

$$J_{t-1}(x) - I = \frac{1}{1-\overline{\alpha}_{t-1}}\mathbb{E}\Big[x-\sqrt{\overline{\alpha}_{t-1}}X_0 \mid X_{t-1}=x\Big]\Big(\mathbb{E}\Big[x-\sqrt{\overline{\alpha}_{t-1}}X_0 \mid X_{t-1}=x\Big]\Big)^\top$$

$$
- \frac{1}{1-\overline{\alpha}_{t-1}}\mathbb{E}\left[\big(x-\sqrt{\overline{\alpha}_{t-1}}X_0\big)\big(x-\sqrt{\overline{\alpha}_{t-1}}X_0\big)^\top \mid X_{t-1}=x\right].
$$

Recognizing that

$$
\big\|\mathbb{E}\big[ZZ^\top\big]-\mathbb{E}[Z]\mathbb{E}[Z]^\top\big\| = \Big\|\mathbb{E}\Big[\big(Z-\mathbb{E}[Z]\big)\big(Z-\mathbb{E}[Z]\big)^\top\Big]\Big\| \le \big\|\mathbb{E}\big[ZZ^\top\big]\big\| \le \mathbb{E}\big[\,\big\|ZZ^\top\big\|\,\big] = \mathbb{E}\big[\|Z\|_2^2\big]
$$

for any random vector Z, we can readily obtain

$$
\big\|J_{t-1}(x)-I\big\| \le \frac{1}{1-\overline{\alpha}_{t-1}}\mathbb{E}\left[\big\|x-\sqrt{\overline{\alpha}_{t-1}}X_0\big\|_2^2 \mid X_{t-1}=x\right]. \tag{125}
$$

When $(x_t, x_{t-1}) \in \mathcal{E}$, it follows from Lemma 2 that

$$
-\log p_{X_{t-1}}(x) \le c_6 d \log T \tag{126}
$$

for any x lying in the line segment connecting x_{t-1} and $\widehat{x}_t$. With this result in place, taking (125) together with the bound (41b) in Lemma 1 immediately leads to

$$
\big\|J_{t-1}(x)-I\big\| \lesssim \frac{1}{1-\overline{\alpha}_{t-1}}\cdot\big\{d\big(1-\overline{\alpha}_{t-1}\big)\log T\big\} \asymp d\log T
$$

for any x lying within the line segment between x_{t-1} and $x_t/\sqrt{\alpha_t}$, as claimed.

Proof of relation (121b). To prove this result, we observe that

$$
\begin{aligned}
\left|(1-\sqrt{\alpha_t})\,\frac{(x_{t-1}-\widehat{x}_t)^\top g_t(x_t)}{1-\overline{\alpha}_t}\right| &\le \frac{1-\alpha_t}{1+\sqrt{\alpha_t}}\,\frac{\|x_{t-1}-\widehat{x}_t\|_2\mathbb{E}\big[\|x_t-\sqrt{\overline{\alpha}_t}X_0\|_2 \mid X_t=x_t\big]}{1-\overline{\alpha}_t} \\
&\lesssim \frac{\log T}{T}\cdot\sqrt{d(1-\alpha_t)\log T}\cdot\sqrt{d(1-\overline{\alpha}_t)\log T} \lesssim \frac{d\log^2 T}{T^{3/2}},
\end{aligned}
$$

where the last line comes from Lemma 1 as well as the basic property (37a) about the learning rates.

Proof of relation (121c). To begin with, the triangle inequality together with the fact $\overline{\alpha}_t=\prod_{k=1}^t \alpha_k$ gives

$$
\begin{aligned}
&\left\|\frac{\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(\widehat{x}_t-\sqrt{\overline{\alpha}_{t-1}}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_{t-1}} - \frac{\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_t}\right\|_2 \\
&\le \left\|\frac{\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}\right\|_2 \\
&\quad + \left\|\left(\frac{1}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}-\frac{1}{1-\overline{\alpha}_t}\right)\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0\right\|_2.
\end{aligned} \tag{127}
$$

Let us first consider the last term in (127). According to Lemma 1, given that $-\log p_{X_t}(x_t) \le \frac{1}{2}c_6 d\log T$ for some constant $c_6>0$, one has

$$
\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0 \lesssim \sqrt{d(1-\overline{\alpha}_t)\log T}. \tag{128}
$$

This in turn reveals that

$$
\begin{aligned}
&\left\|\left(\frac{1}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}-\frac{1}{1-\overline{\alpha}_t}\right)\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0\right\|_2 \\
&\lesssim \left|\frac{1}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}-\frac{1}{1-\overline{\alpha}_t}\right|\cdot\sqrt{d(1-\overline{\alpha}_t)\log T} \\
&\asymp \frac{(1-\sqrt{\alpha_t})(1+\overline{\alpha}_{t-1}\sqrt{\alpha_t})}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)}\sqrt{d(1-\overline{\alpha}_t)\log T}
\end{aligned}
$$

$$\asymp \frac{1-\alpha_t}{(\alpha_t-\overline{\alpha}_t)^{3/2}}\sqrt{d\log T}, \tag{129}$$

where the last inequality makes use of the properties (37).

Next, we turn attention to the first term in (127), which relies on the following claim.

Claim 2. *Consider any point x_t obeying $-\log p_{X_t}(x_t) \le c_6 d\log T$ for some large constant $c_6 > 0$. One has*

$$p_{X_{t-1}}(\widehat{x}_t) = \Big(1+O\Big(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\Big)\Big)p_{X_t}(x_t), \tag{130a}$$

In addition, by defining the following set

$$\mathcal{E}_1 \coloneqq \big\{x : \|x_t-\sqrt{\overline{\alpha}_t}x\|_2 \le c_4\sqrt{d(1-\overline{\alpha}_t)\log T}\big\}$$

for some large enough constant $c_4 > 0$, we have

$$\frac{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} = 1+O\Big(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\Big), \qquad \text{if } x_0\in\mathcal{E}_1, \tag{130b}$$

$$\frac{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} \le \exp\left(\frac{16c_1\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\log T}{(1-\overline{\alpha}_t)T}\right), \qquad \text{if } x_0\notin\mathcal{E}_1. \tag{130c}$$

As an immediate consequence of this claim, one has

$$\frac{p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)}{p_{X_0\,|\,X_t}(x_0\,|\,x_t)} = \frac{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)}\cdot\frac{p_{X_t}(x_t)}{p_{X_{t-1}}(\widehat{x}_t)} = \Big(1+O\Big(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\Big)\Big)\frac{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)}$$
$$\begin{cases} = 1+O\Big(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\Big), & \text{if } x_0\in\mathcal{E}_1,\\ \le \exp\Big(\frac{16c_1\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2\log T}{(1-\overline{\alpha}_t)T}\Big), & \text{if } x_0\notin\mathcal{E}_1.\end{cases} \tag{131}$$

In turn, this allows one to deduce that

$$\begin{aligned}
&\Big\|\int_{x_0}p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0-\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0\Big\|_2\\
&\quad\le\int_{x_0\in\mathcal{E}_1}\big|p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)-p_{X_0\,|\,X_t}(x_0\,|\,x_t)\big|\cdot\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0\\
&\qquad\quad+\int_{x_0\notin\mathcal{E}_1}\big|p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)-p_{X_0\,|\,X_t}(x_0\,|\,x_t)\big|\cdot\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0\\
&\quad\le\int_{x_0}O\Big(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\Big)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0\\
&\qquad\quad+\int_{x_0\notin\mathcal{E}_1}\exp\left(\frac{16c_1\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\log T}{(1-\overline{\alpha}_t)T}\right)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0\\
&\quad\le O\Big(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\Big)\mathbb{E}\left[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2\mid X_t=x_0\right]\\
&\qquad\quad+\int_{x_0\notin\mathcal{E}_1}\exp\left(\frac{20c_1\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\log T}{(1-\overline{\alpha}_t)T}\right)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0.
\end{aligned}$$

Invoking (41a) in Lemma 1 as well as similar arugments for (97) (assuming that c_4 is sufficiently large), we arrive at

$$\begin{aligned}
&\frac{1}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}\Big\|\int_{x_0}p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0-\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0\Big\|_2\\
&\quad\lesssim\frac{d^{3/2}(1-\alpha_t)\sqrt{1-\overline{\alpha}_t}\log^{3/2}T}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})^2}+\frac{1}{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}\cdot\frac{1}{T^{c_0}}
\end{aligned}$$

$$\asymp \frac{d^{3/2}(1-\alpha_t)\log^{3/2}T}{(1-\overline{\alpha}_{t-1})^{3/2}}, \tag{132}$$

where c_0 is some large enough constant, and we have also made use of the properties in (37).

Substituting (129) and (132) into (127) readily concludes the proof.

Proof of Claim 2. Let us make the following observations: for any $x_0 \in \mathcal{E}_1$, one has

$$\begin{aligned}
\frac{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} &= \left(\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}\right)^{d/2}\exp\left(\frac{\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)}-\frac{\big\|\widehat{x}_t-\sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right)\\
&= \exp\left\{\big(1+o(1)\big)\frac{\overline{\alpha}_{t-1}(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\cdot\frac{d}{2}\right\}\exp\left(-\frac{(1-\alpha_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)(\alpha_t-\overline{\alpha}_t)}\right)\\
&= \exp\left\{O\left(\frac{d(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}+\frac{(1-\alpha_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)}\right)\right\}\\
&= \exp\left\{O\left(\frac{d(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}+\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\right)\right\}\\
&= 1+O\left(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\right),
\end{aligned} \tag{133}$$

where the second line holds since

$$\log\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} = \log\left(1+\frac{\overline{\alpha}_{t-1}(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\right) = \big(1+o(1)\big)\frac{\overline{\alpha}_{t-1}(1-\alpha_t)}{1-\overline{\alpha}_{t-1}},$$

and we have also made use of (37) (so that $\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}} = o(1)$ under our assumption on T). Additionally, for any $x_0 \notin \mathcal{E}_1$, it follows from the argument for (89) that

$$\begin{aligned}
\frac{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} &\le \exp\left(\frac{d(1-\alpha_t)}{2(1-\overline{\alpha}_{t-1})}+\frac{(1-\alpha_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2}\right)\\
&\le \exp\left(\frac{(1-\alpha_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2c_4^2(1-\overline{\alpha}_{t-1})^2\log T}+\frac{(1-\alpha_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{(1-\overline{\alpha}_{t-1})^2}\right)\\
&\le \exp\left(\frac{8c_1\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\log T}{(1-\overline{\alpha}_{t-1})T}\right) \le \exp\left(\frac{16c_1\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\log T}{(1-\overline{\alpha}_t)T}\right),
\end{aligned}$$

where the last line comes from (37b) and (37c). Taking the above bounds together and invoking the same calculation as in (95), (97) and (100), we reach

$$p_{X_{t-1}}(\widehat{x}_t) = \left(1+O\Big(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\Big)\right)p_{X_t}(x_t),$$

thus concluding the proof of Claim 2. □

C.2 Proof of Lemma 5

First, recognizing that X_t follows a Gaussian distribution when conditioned on $X_{t-1} = x_{t-1}$, we can derive

$$\begin{aligned}
\log p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) &= \log\frac{p_{X_t\,|\,X_{t-1}}(x_t\,|\,x_{t-1})p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)}\\
&= \log\frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)}+\frac{\|x_t-\sqrt{\alpha_t}x_{t-1}\|_2^2}{2(1-\alpha_t)}-\frac{d}{2}\log\big(2\pi(1-\alpha_t)\big)
\end{aligned}$$

$$\le \log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} + T\big(\|x_{t-1} - \widehat{x}_t\|_2^2 + 1\big),$$

where the last inequality makes use of the properties (37) about α_t. Some direct calculations then yield

$$\begin{aligned}
\log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} &= \log \frac{\int_{x_0} p_{X_0}(x_0) \exp\Big(-\frac{\|x_{t-1}-\sqrt{\overline{\alpha}_{t-1}}x_0\|_2^2}{2(1-\overline{\alpha}_{t-1})}\Big)\mathrm{d}x_0}{\int_{x_0} p_{X_0}(x_0) \exp\Big(-\frac{\|x_{t}-\sqrt{\overline{\alpha}_{t}}x_0\|_2^2}{2(1-\overline{\alpha}_{t})}\Big)\mathrm{d}x_0} - \frac{d}{2}\log\Big(\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\Big) \\
&\le \sup_{x_0:\|x_0\|_2\le T^{c_R}} \left\{\frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\|x_{t-1} - \sqrt{\overline{\alpha}_{t-1}}x_0\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right\} + \frac{d}{2}\log 2 \\
&\le \sup_{x_0:\|x_0\|_2\le T^{c_R}} \frac{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)} + \frac{d}{2}\log 2 \\
&\le \sup_{x_0:\|x_0\|_2\le T^{c_R}} \frac{\|x_t\|_2^2 + \|x_0\|_2^2}{1-\overline{\alpha}_t} + \frac{d}{2}\log 2 \\
&\le 2T\left(\|x_t\|_2^2 + T^{2c_R}\right),
\end{aligned}$$

where the second line makes use of the properties in (37). Combining the above two relations, we arrive at

$$\log p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \le 2T\big(\|x_t\|_2^2 + \|x_{t-1} - \widehat{x}_t\|_2^2 + T^{2c_R}\big). \tag{134}$$

We then turn attention to the conditional distribution $p_{Y_{t-1}\,|\,Y_t}$. By recalling from (57) that

$$p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t) = \frac{1}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{d/2}} \exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\big\|x_{t-1} - \mu_t(x_t)\big\|_2^2\Big) \tag{135}$$

with $\mu_t(x_t)$ defined in (56), we can deduce that

$$\begin{aligned}
\log \frac{1}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} &= \frac{\alpha_t\|x_{t-1} - \mu_t(x_t)\|_2^2}{2(1-\alpha_t)} + \frac{d}{2}\log\left(2\pi\frac{1-\alpha_t}{\alpha_t}\right) &(136)\\
&\le T\left(\|x_{t-1} - \widehat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}\right) + d\log T. &(137)
\end{aligned}$$

To see why the last inequality follows, we recall from (56) that

$$\begin{aligned}
\|x_{t-1} - \mu_t(x_t)\|_2^2 &\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + 2\|\widehat{x}_t - \mu_t(x_t)\|_2^2 \\
&= 2\|x_{t-1} - \widehat{x}_t\|_2^2 + 2\Big(\frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\Big)^2 \Big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big\|_2^2 \\
&\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + \frac{2(1-\alpha_t)^2}{\alpha_t(1-\overline{\alpha}_{t-1})^2} \sup_{x_0:\|x_0\|_2\le T^{c_R}} \|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \\
&\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + \frac{64c_1^2\log^2 T}{T^2}\left(2\|x_t\|_2^2 + 2\overline{\alpha}_t T^{2c_R}\right) \\
&\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}, &(138)
\end{aligned}$$

where we have invoked the properties (37).

Putting everything together, we arrive at the advertised crude bound:

$$\log \frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} \le 2T\left(\|x_{t-1} - \widehat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}\right).$$

D Analysis for the accelerated deterministic sampler (Theorem 2)

The aim of this section is to establish Theorem 2. The proof follows similar arguments as that of Theorem 1.

D.1 Proof of Theorem 2

Auxiliary vectors and their properties. Before embarking on the proof, let us introduce several pieces of notation:

$$\varphi_t(x) \coloneqq x - u_t(x), \tag{139a}$$

where

$$\begin{aligned} u_t(x) &\coloneqq \left\{ \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2} - \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^3} \|g_t(x)\|_2^2 \right\} g_t(x) \\ &\quad + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^3} \mathbb{E}\left[\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)^\top \mid X_t = x \right] g_t(x) \\ &\quad - \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^3} \mathbb{E}\left[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \big(X_t - \sqrt{\overline{\alpha}_t}X_0 - g_t(X_t)\big) \mid X_t = x \right] \qquad (139b) \\ &= -\left\{ \frac{1-\alpha_t}{2} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)} - \frac{(1-\alpha_t)^2}{8} \|s_t(x)\|_2^2 \right\} s_t(x) - \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)} \mathbb{E}\left[\overline{W}_t \overline{W}_t^\top s_t(X_t) \mid X_t = x \right] \\ &\quad - \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)} \mathbb{E}\left[\|\overline{W}_t\|_2^2 \left(\frac{1}{\sqrt{1-\overline{\alpha}_t}} \overline{W}_t + (1-\overline{\alpha}_t) s_t(X_t) \right) \mid X_t = x \right]. \qquad (139c) \end{aligned}$$

Here, we recall that $g_t(x) = -(1-\overline{\alpha}_t)s_t(x)$ has been defined in (34) and that $X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t$ with $\overline{W}_t \sim \mathcal{N}(0, I_d)$ (see (6)). In view of Assumption 2 and the fact that the MMSE estimator is the conditional expectation (Hajek, 2015, Section 3.3.1), we have

$$w_t(x) = \mathbb{E}\left[\|\overline{W}_t\|_2^2 \left(\frac{1}{\sqrt{1-\overline{\alpha}_t}} \overline{W}_t + (1-\overline{\alpha}_t) s_t(X_t) \right) + \frac{1}{1-\overline{\alpha}_t} \overline{W}_t \overline{W}_t^\top s_t(X_t) \mid X_t = x \right],$$

which taken together with (139) and (25) confirms the following equivalent expression for the sampler (25):

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \varphi_t(Y_t) = \frac{1}{\sqrt{\alpha_t}} Y_t - \frac{1}{\sqrt{\alpha_t}} u_t(Y_t). \tag{140}$$

We now single out a useful property about $u_t(\cdot)$. For any point $x_t \in \mathbb{R}^d$ obeying $-\log p_{X_t}(x_t) \le c_6 d \log T$ for some large constant $c_6 > 0$ (see Lemma 1), one has

$$\big\|u_t(x_t)\big\|_2 \lesssim (1-\alpha_t)\left(\frac{d \log T}{1-\overline{\alpha}_t} \right)^{1/2}, \tag{141a}$$

and one can also write

$$u_t(x_t) = \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} g_t(x_t) + \xi_t(x_t) \tag{141b}$$

for some residual term $\xi_t(x_t)$ obeying

$$\big\|\xi_t(x_t)\big\|_2 \lesssim (1-\alpha_t)^2 \left(\frac{d \log T}{1-\overline{\alpha}_t} \right)^{3/2}. \tag{141c}$$

To streamline presentation, we leave the proof of (141) to the end of this subsection.

Main steps of the proof. Akin to the proof of Theorem 1, the key idea lies in understanding the transformation Φ_t (cf. (25)), or equivalently, φ_t (cf. (139)). There are several objects that play an important role in the analysis, which we single out as follows:

$$A_t \coloneqq \frac{1}{1-\overline{\alpha}_t} \int p_{X_0 \mid X_t}(x_0 \mid x) \big\|x - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2 \mathrm{d}x_0; \tag{142a}$$

$$B_t := \frac{1}{1-\overline{\alpha}_t}\Big\| \int p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big\|_2^2; \tag{142b}$$

$$C_t := \frac{1}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^4\mathrm{d}x_0; \tag{142c}$$

$$D_t := \frac{1}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\Big(\big\langle g_t(x),\, x-\sqrt{\overline{\alpha}_t}x_0\big\rangle\Big)^2\mathrm{d}x_0; \tag{142d}$$

$$E_t := \frac{1}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big\langle g_t(x),\, x-\sqrt{\overline{\alpha}_t}x_0\big\rangle\mathrm{d}x_0. \tag{142e}$$

Here, we suppress the dependency on x in the above five objects to simplify notation whenever it is clear from the context. In view of Lemma 1 and the properties (37), we have the following bounds:

$$\begin{aligned}
\big|B_t\big| \le \big|A_t\big| &\lesssim \frac{1}{1-\overline{\alpha}_t}\cdot d(1-\overline{\alpha}_t)\log T \asymp d\log T;\\
\big|C_t\big| &\lesssim \frac{1}{(1-\overline{\alpha}_t)^2}d^2(1-\overline{\alpha}_t)^2\log^2 T \asymp d^2\log^2 T;\\
\big|D_t\big| &\le \frac{\|g_t(x)\|_2^2}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0 \lesssim d^2\log^2 T;\\
\big|E_t\big| &\le \frac{\|g_t(x)\|_2^2}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^3\mathrm{d}x_0 \lesssim d^2\log^2 T.
\end{aligned}$$

As it turns out, Theorem 2 can be established in a very similar way as in the proof of Theorem 1. In essence, the only step that needs to be changed is to replace (49) in Lemma 3 with (143) in the lemma below.

Lemma 6. *Suppose that* $\frac{d^2(1-\alpha_t)\log T}{\alpha_t-\overline{\alpha}_t}\lesssim 1$. *For every* $x\in\mathbb{R}$ *obeying* $-\log p_{X_t}(x)\le c_6 d\log T$ *for some large enough constant* $c_6>0$, *we have*

$$\begin{aligned}
\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\varphi_t(x))}{p_{X_t}(x)} &= 1+\frac{(1-\alpha_t)(d+B_t-A_t)}{2(1-\overline{\alpha}_t)}+O\Big(d^3\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big)\\
&\quad+\frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[d(d+2)+(4+2d)(B_t-A_t)-B_t^2+C_t+2D_t-3E_t+A_tB_t\big].
\end{aligned}\tag{143a}$$

Moreover, for any random vector Y, *one has*

$$\begin{aligned}
\frac{p_{\varphi_t(Y)}(\varphi_t(x))}{p_Y(x)} &= 1+\frac{(1-\alpha_t)(d+B_t-A_t)}{2(1-\overline{\alpha}_t)}+O\Big(d^6\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big)\\
&\quad+\frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[d(d+2)+(4+2d)(B_t-A_t)-B_t^2+C_t+2D_t-3E_t+A_tB_t\big].
\end{aligned}\tag{143b}$$

Here, the quantities $A_t,\dots,E_t$ *are defined in* (142).

The proof of this lemma can be found in Appendix D.2. Crucially, the terms in (143a) and those in (143b) coincide (except for the residual terms $O\big(d^3\big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\big)^3\log^3 T\big)$ and $O\big(d^6\big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\big)^3\log^3 T\big)$), which will cancel each other during the subsequent proof.

With Lemma 6 in place, we are able to prove — in the same way as in the proof of Theorem 1 — that for every $t\ge 1$,

$$\mathbb{P}\big(Y_t\in\widetilde{\mathcal{E}}_t\big)\ge 1-(T-t+1)\exp(-c_3 d\log T) \tag{144}$$

for some constant $c_3>0$, where the set $\mathcal{E}_t$ is defined as

$$\widetilde{\mathcal{E}}_t := \Big\{y:\Big|\frac{p_{Y_t}(y)}{p_{X_t}(y)}-1\Big|\le c_5(T-t+1)\frac{d^6\log^6 T}{T^3}\Big\}$$

for some constant $c_5>0$. In comparison to the set $\mathcal{E}_t$ (cf. (50)) defined in the proof of Theorem 1, the condition that defines $\widetilde{\mathcal{E}}_t$ does not include a term $O\big(\frac{d^2\log^4 T}{T^2}\big)$, which plays a crucial role in enabling faster convergence. Repeating the remaining arguments that justify (52) in the proof of Theorem 1 (which we omit for the sake of brevity), we have established the claimed result in Theorem 2.

Proof of properties (141). To justify the above results (141), note that Lemma 1 implies that

$$\big\|g_t(x_t)\big\|_2 \le \mathbb{E}\left[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2 \,|\, X_t = x_t\right] \lesssim \sqrt{d(1-\overline{\alpha}_t)\log T}, \tag{145a}$$

$$\frac{1-\alpha_t}{1-\overline{\alpha}_t}\big\|g_t(x_t)\big\|_2 \lesssim \frac{1-\alpha_t}{1-\overline{\alpha}_t}\sqrt{d(1-\overline{\alpha}_t)\log T} \asymp (1-\alpha_t)^2\sqrt{\frac{d\log T}{1-\overline{\alpha}_t}}, \tag{145b}$$

$$\frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^2}\big\|g_t(x_t)\big\|_2 \lesssim \frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^2}\sqrt{d(1-\overline{\alpha}_t)\log T} \asymp (1-\alpha_t)^2\frac{\sqrt{d\log T}}{(1-\overline{\alpha}_t)^{3/2}}, \tag{145c}$$

$$\frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^3}\|g_t(x_t)\|_2^3 \lesssim \frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^3}\left(d(1-\overline{\alpha}_t)\log T\right)^{3/2} \asymp (1-\alpha_t)^2\left(\frac{d\log T}{1-\overline{\alpha}_t}\right)^{3/2}, \tag{145d}$$

$$\begin{aligned}\frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^3}\Big\|\mathbb{E}\Big[\big(x_t-\sqrt{\overline{\alpha}_t}X_0\big)\big(x_t-\sqrt{\overline{\alpha}_t}X_0\big)^\top \,|\, X_t=x_t\Big]\,g_t(x_t)\Big\|_2 &\lesssim \frac{(1-\alpha_t)^2\big\|g_t(x_t)\big\|_2}{(1-\overline{\alpha}_t)^3}\mathbb{E}\Big[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \,|\, X_t=x_t\Big]\\ &\lesssim (1-\alpha_t)^2\left(\frac{d\log T}{1-\overline{\alpha}_t}\right)^{3/2},\end{aligned} \tag{145e}$$

and

$$\begin{aligned}&\frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^3}\Big\|\mathbb{E}\Big[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2^2\big(x_t-\sqrt{\overline{\alpha}_t}x_0-g_t(X_t)\big) \,|\, X_t=x_t\Big]\Big\|_2\\ &\le \frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^3}\Big\{\Big\|\mathbb{E}\Big[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2^3 \,|\, X_t=x_t\Big]\Big\|_2 + \big\|g_t(x_t)\big\|_2\Big\|\mathbb{E}\Big[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \,|\, X_t=x_t\Big]\Big\|_2\Big\}\\ &\lesssim (1-\alpha_t)^2\left(\frac{d\log T}{1-\overline{\alpha}_t}\right)^{3/2}.\end{aligned} \tag{145f}$$

Substituting the bounds (145) into (139b) immediately establishes (141). □

D.2 Proof of Lemma 6

The proof of Lemma 6 is derived in a very similar way to the proof of Lemma 3, as detailed below. For notational simplicity, we shall abbreviate

$$u = u_t(x) \qquad \text{and} \qquad z = g_t(x) \tag{146}$$

whenever it is clear from the context. [YC: z is also used to denote noise in the stochastic sampler.]

D.2.1 Proof of relation (143a)

Through direct calculations as shown in (105), we can obtain

$$\begin{aligned}p_{\sqrt{\alpha_t}X_{t-1}}\big(\varphi_t(x)\big) &= p_{X_t}(x)\left(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\right)^{d/2}\\ &\cdot\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\Bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\Bigg)\mathrm{d}x_0.\end{aligned} \tag{147}$$

Moreover, similar to the analysis of Lemma 3, we focus our attention on the set given x:

$$\mathcal{E} \coloneqq \Big\{x_0 : \big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2 \le 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T}\Big\}, \tag{148}$$

which allows us to derive, for some numerical constant $c_8 > 0$,

$$\begin{aligned}&\int p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\Bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\Bigg)\mathrm{d}x_0\\ &= O\big(\exp(-c_8c_5^2 d\log T)\big) + \int_{x_0\in\mathcal{E}} p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\Bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\Bigg)\mathrm{d}x_0\end{aligned}$$

$$=: \text{RHS} \tag{149}$$

To further control the right-hand side above, recall that the learning rates are selected such that $\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \le \frac{4c_1 \log T}{T}$ for $1 < t \le T$ (see (37b)). In view of the Taylor expansion $e^{-x} = 1 - x + \frac{1}{2}x^2 + O(x^3)$ for $x \le 1/2$, we can derive

$$\begin{aligned}\text{RHS} &= O\big(\exp(-c_8 c_5^2 d \log T)\big) \\ &+ \int_{x_0 \in \mathcal{E}} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\bigg\{1 - \frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\|z\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)} \\ &+ \frac{(1-\alpha_t)^2}{8(\alpha_t - \overline{\alpha}_t)^2(1-\overline{\alpha}_t)^2}\Big(\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2 - z^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\Big)^2 + O\Big(d^3\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3 \log^3 T\Big)\bigg\}\mathrm{d}x_0. \end{aligned} \tag{150}$$

In order to see this, we recall the property of u (cf. (141)) as

$$\Big\|u - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}z\Big\|_2 \le O\Big((1-\alpha_t)^2\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{3/2}\Big). \tag{151}$$

As a consequence, for any $x_0 \in \mathcal{E}$ we have

$$\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} = O\Big(d\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)\log T\Big)$$

and

$$\begin{aligned}\frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)} &= \frac{\frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\|z\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)} + O\Big(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3 \log^2 T\Big) \\ &= \frac{z^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} + O\Big(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \log^2 T\Big) \\ &= O\Big(d\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)\log T\Big),\end{aligned}$$

where we have invoked the properties (37). Taking the above results together and using the following basic properties regarding quantities $A_t, \ldots, E_t$ (defined in (142))

$$\begin{aligned}&\int p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2 \mathrm{d}x_0 = (1-\overline{\alpha}_t)A_t, \\ &\int p_{X_0 \,|\, X_t}(x_0 \,|\, x)\|z\|_2^2 \mathrm{d}x_0 = (1-\overline{\alpha}_t)B_t, \\ &\int p_{X_0 \,|\, X_t}(x_0 \,|\, x)u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0 = \frac{1-\alpha_t}{2}B_t + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)}\big[B_t - B_t^2 + D_t - E_t + A_tB_t\big], \\ &\int p_{X_0 \,|\, X_t}(x_0 \,|\, x)\Big(\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2 - z^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\Big)^2 \mathrm{d}x_0 = (1-\overline{\alpha}_t)^2\big[C_t + D_t - 2E_t\big],\end{aligned}$$

we arrive at

$$(150) = 1 - \frac{(1-\alpha_t)(A_t - B_t)}{2(\alpha_t - \overline{\alpha}_t)} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[-B_t^2 + C_t + 2D_t - 3E_t + A_tB_t\big] + O\Big(d^3\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3 \log^3 T\Big).$$

Once again, we note that integrating over set $\mathcal{E}$ and over all possible x_0 only incurs a difference at most as large as $O\big(\exp(-c_8 c_5^2 d \log T)\big)$. Putting all this together establishes the advertised result (143a).

D.2.2 Proof of relation (143b)

Consider any random vector Y, and let us invoke again the basic transformation

$$p_{\varphi_t(Y)}(\varphi_t(x)) = \mathsf{det}\Big(\frac{\partial \varphi_t(x)}{\partial x}\Big)^{-1} p_Y(x),$$

where $\frac{\partial \varphi_t(x)}{\partial x}$ denotes the Jacobian matrix. We are then left with controlling the quantity $\mathsf{det}\Big(\frac{\partial \varphi_t(x)}{\partial x}\Big)^{-1}$.

Towards this, let us again recall that the determinant of a matrix satisfies

$$\mathsf{det}(I+A)^{-1} = 1 - \mathsf{Tr}(A) + \frac{1}{2}\big[\mathsf{Tr}(A)^2 + \|A\|_{\mathrm{F}}^2\big] + O\big(d^3\|A\|^3\big),$$

provided that $d\|A\| \le c_{20}$ for some small enough constant $c_{20} > 0$. This relation leads to

$$\begin{aligned} p_{\varphi_t(Y)}(\varphi_t(x)) &= \mathsf{det}\Big(\frac{\partial \varphi_t(x)}{\partial x}\Big)^{-1} p_Y(x) \\ &= \Big\{1 + \mathsf{Tr}\Big(\frac{\partial u}{\partial x}\Big) + \frac{1}{2}\Big[\mathsf{Tr}\Big(\frac{\partial u}{\partial x}\Big)^2 + \Big\|\frac{\partial u}{\partial x}\Big\|_{\mathrm{F}}^2\Big] + O\Big(d^3\Big\|\frac{\partial u}{\partial x}\Big\|\Big)\Big\} p_Y(x), \end{aligned} \tag{152}$$

where we invoke the definition in (139) that

$$\begin{aligned} \varphi_t(x) = x - u = x &- \Big(\frac{(1-\alpha_t)}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2} - \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^3}\|z\|_2^2\Big) z \\ &- \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^3}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top z \mathrm{d}x_0 \\ &+ \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^3}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x - \sqrt{\overline{\alpha}_t}x_0 - z\big)\mathrm{d}x_0. \end{aligned}$$

To further control the right-hand side above of the above display, let us first make note of several identities. Proving these identities only requires elementary calculation regarding Gaussian integration and derivatives, which is omitted here for brevity. Specifically, one has

$$\frac{\partial z}{\partial x} = J_t, \tag{153a}$$

$$\frac{\partial}{\partial x}\|z\|_2^2 z = \|z\|_2^2 J_t + 2zz^\top J_t, \tag{153b}$$

$$\begin{aligned} &\frac{\partial}{\partial x}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0 \\ &= \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0 I + 2\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0 \\ &\quad + \frac{1}{1-\overline{\alpha}_t}\Big(\Big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\Big)\Big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)^\top \\ &\quad - \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0\Big), \end{aligned} \tag{153c}$$

$$\begin{aligned} &\frac{\partial}{\partial x}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2 z = \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0 J_t + 2zz^\top \\ &\quad + \frac{1}{1-\overline{\alpha}_t}\Big(\Big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\Big)zz^\top - z\Big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)^\top\Big), \end{aligned} \tag{153d}$$

$$\begin{aligned} \frac{\partial}{\partial x}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top z\mathrm{d}x_0 &= \|z\|_2^2 I + zz^\top \\ &+ \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top J_t\mathrm{d}x_0 \\ &+ \frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(z^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)z^\top\mathrm{d}x_0 \\ &- \frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(z^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0. \end{aligned} \tag{153e}$$

Equipped with the above relations, we can easily verify that

$$\Big\|\frac{\partial u}{\partial x}\Big\| \lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t}, \tag{154a}$$

$$\begin{aligned}\mathsf{Tr}\Big(\frac{\partial u}{\partial x}\Big) &= \frac{(1-\alpha_t)\big(d+B_t-A_t\big)}{2(1-\overline{\alpha}_t)} \\ &\quad + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big(d-2A_t-A_t^2+3A_tB_t+2B_t-3B_t^2+C_t+4D_t-3E_t-F_t\big),\end{aligned} \tag{154b}$$

$$\begin{aligned}\Big\|\frac{\partial u}{\partial x}\Big\|_{\mathrm{F}}^2 &= \frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\Big\|\frac{\partial z}{\partial x}\Big\|_{\mathrm{F}}^2 + O\Big(d^5\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big) \\ &= \frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\big(d+2(B_t-A_t)+B_t^2+F_t-2D_t\big) + O\Big(d^5\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big),\end{aligned} \tag{154c}$$

as long as $d^2\big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\big)\log T \lesssim 1$, where we recall the definition of the quantities A_t to E_t in (142), and

$$F_t(x) \coloneqq \Big\|\frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top \mathrm{d}x_0\Big\|_{\mathrm{F}}^2. \tag{154d}$$

Plugging these results into inequality (152) leads to

$$\begin{aligned}p_{\varphi_t(Y)}(\varphi_t(x)) = p_Y(x)\Big\{&1+\frac{(1-\alpha_t)(d+B_t-A_t)}{2(1-\overline{\alpha}_t)} + O\Big(d^6\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big) \\ &+ \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[d(d+2)+(4+2d)(B_t-A_t)-B_t^2+C_t+2D_t-3E_t+A_tB_t\big]\Big\}.\end{aligned} \tag{154e}$$

We have thus completed the proof of Lemma 6.

E Analysis for the accelerated stochastic sampler (Theorems 4)

The proof of Theorem 4 follows similar structure as the proof of Theorem 3. Throughout the proof, we shall employ the notation $\widehat{x}_t \coloneqq x_t/\sqrt{\alpha_t}$ as before.

E.1 Proof of Theorem 4

Step 1: expressing the update rule in terms of a Jacobian matrix. Let us recall the Jacobian matrix $J_t(x) = \frac{\partial g_t(x)}{\partial x}$ defined in (35). In view of the expression (36) as well as (6) (i.e., $X_t - \sqrt{\overline{\alpha}_t}X_0 = \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t$ with $\overline{W}_t \sim \mathcal{N}(0, I_d)$), one can write

$$\begin{aligned}J_t(x) &= I_d + \mathbb{E}\big[\overline{W}_t \mid X_t = x\big]\Big(\mathbb{E}\big[\overline{W}_t \mid X_t = x\big]\Big)^\top - \mathbb{E}\big[\overline{W}_t\overline{W}_t^\top \mid X_t = x\big] \\ &= I_d + \big(1-\overline{\alpha}_t\big)s_t(x)s_t(x)^\top - \mathbb{E}\big[\overline{W}_t\overline{W}_t^\top \mid X_t = x\big],\end{aligned} \tag{155}$$

where the last line makes use of the relation (34). Additionally, recall that (i) $v_t(x,z)$ (cf. (30)) is the MMSE estimator for estimating $\overline{W}_t\overline{W}_t z$ given $\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t = x$ and $Z_t = z$, and (ii) Z_t is independent from X_0 and $\overline{W}_t$. Then this MMSE estimator admits the following expression (Hajek, 2015, Section 3.3.1):

$$v_t(x,z) = \mathbb{E}\big[\overline{W}_t\overline{W}_t^\top z \mid X_t = x\big] = \mathbb{E}\big[\overline{W}_t\overline{W}_t^\top \mid X_t = x\big]z. \tag{156}$$

As a result, the mapping introduced in (31b) can be alternatively expressed as:

$$\begin{aligned}\Psi_t(x,z) &= \frac{1}{\sqrt{\alpha_t}}\Big(x+(1-\alpha_t)s_t(x)\Big) + \sigma_t\Big\{z - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\big[z+(1-\overline{\alpha}_t)s_t(x)s_t(x)^\top z - v_t(x,z)\big]\Big\} \\ &= \mu_t(x) + \sigma_t\Big(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x)\Big)z,\end{aligned} \tag{157}$$

where we have also used the definition (56) of $\mu_t(\cdot)$. In comparison to the plain DDPM-type sampler (cf. (28)), the key correction term is the second component on the right-hand side of (157), which adjusts the covariance of the additive Gaussian noise.

Equipped with the above expression (157), we can readily express the conditional distribution of Y_{t-1} (cf. (31a)) given Y_t such that: for any points $x_t, x_{t-1} \in \mathbb{R}^d$,

$$p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t) = \frac{1}{\left(2\pi\frac{1-\alpha_t}{\alpha_t}\right)^{d/2}\left|\det\left(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\right)\right|} \cdot \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)}\left\|\left(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\right)^{-1}\left(x_{t-1} - \mu_t(x_t)\right)\right\|_2^2\right). \tag{158}$$

Step 2: controlling the conditional distributions $p_{X_{t-1}\,|\,X_t}$ and $p_{Y_{t-1}\,|\,Y_t}$. Akin to Step 2 in the proof of Theorem 3, we need to look at the conditional distribution $p_{X_{t-1}\,|\,X_t}$ when restricted to points from the following set:

$$\mathcal{E} := \left\{(x_t, x_{t-1}) \,|\, -\log p_{X_t}(x_t) \le \frac{1}{2}c_6 d\log T,\ \|x_{t-1} - \widehat{x}_t\|_2 \le c_3\sqrt{d(1-\alpha_t)\log T}\right\}, \tag{159}$$

with the numerical constants $c_3, c_6 > 0$ introduced in Lemma 2. The following lemma, which is a counterpart of Lemma 4 for the accelerated sampler, characterizes $p_{X_{t-1}\,|\,X_t}$ in a fairly tight manner over the set $\mathcal{E}$. The proof is deferred to Appendix E.2.

Lemma 7. *There exists some large enough numerical constant $c_\zeta > 0$ such that: for every $(x_t, x_{t-1}) \in \mathcal{E}$,*

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = \frac{1}{\left(2\pi\frac{1-\alpha_t}{\alpha_t}\right)^{d/2}\left|\det\left(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\right)\right|} \cdot \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)}\left\|\left(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\right)^{-1}\left(x_{t-1} - \mu_t(x_t)\right)\right\|_2^2 + \zeta_t(x_{t-1}, x_t)\right) \tag{160}$$

holds for some residual term $\zeta_t(x_{t-1}, x_t)$ obeying

$$\left|\zeta_t(x_{t-1}, x_t)\right| \le c_\zeta \frac{d^3\log^{4.5} T}{T^{3/2}}. \tag{161}$$

Here, we recall the definition of $\mu_t(\cdot)$ (resp. $J_t(\cdot)$) in (56) (resp. (35)).

Moving beyond the set $\mathcal{E}$, we are still in need of bounding the log density ratio $\log\frac{p_{X_{t-1}\,|\,X_t}}{p_{Y_{t-1}\,|\,Y_t}}$ for all pairs (x_t, x_{t-1}) outside $\mathcal{E}$, in a way similar to Lemma 5 in the proof of Theorem 3. Our crude bound towards this end is stated as follows, whose proof is postponed to Appendix E.3.

Lemma 8. *For all $(x_t, x_{t-1}) \in \mathbb{R}^d \times \mathbb{R}^d$, we have*

$$\log\frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} \le T^{c_0+2c_R+2}\left\{\left\|x_{t-1} - \widehat{x}_t\right\|_2^2 + \|x_t\|_2^2 + 1\right\}, \tag{162}$$

where c_0 is defined in (20).

Step 3: bounding the KL divergence of interest. With Lemmas 7-8 in place, one can repeat the arguments in Step 3 in the proof of Theorem 3 to arrive at

$$\mathbb{E}_{x_t\sim X_t}\left[\mathsf{KL}\left(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\,\|\,p_{Y_{t-1}\,|\,Y_t}(\cdot\,|\,x_t)\right)\right] \lesssim \left(\frac{d^3\log^{4.5} T}{T^{3/2}}\right)^2.$$

Substitution into (60) and (59) then yields

$$2\mathsf{TV}(p_{X_1}, p_{Y_1})^2 \le \mathsf{KL}(p_{X_1}\,\|\,p_{Y_1}) \lesssim \mathsf{KL}(p_{X_T}\,\|\,p_{Y_T}) + \sum_{t\ge 2}^{T}\frac{d^6\log^9 T}{T^3} \asymp \frac{d^6\log^9 T}{T^2},$$

where the last relation results from (61). This completes the proof of Theorem 4.

E.2 Proof of Lemma 7

Recall that an explicit expression for $p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)$ has already been established in Lemma 4 (see (120)) A little algebra then allows one to write

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = f_1(x_t)\exp\Big(-f_2(x_t,x_{t-1}) + \zeta_{t,1}(x_t,x_{t-1})\Big), \tag{163}$$

for some function $f_1(\cdot)$, where

$$g_{t-1}(x) \coloneqq \int_{x_0}\big(x-\sqrt{\overline{\alpha}_{t-1}}x_0\big)p_{X_0|X_{t-1}}(x_0\,|\,x)\mathrm{d}x_0, \tag{164a}$$

$$f_2(x_t,x_{t-1}) = \frac{\|x_t-\sqrt{\alpha_t}x_{t-1}\|_2^2}{2(1-\alpha_t)} + \frac{(x_{t-1}-\widehat{x}_t)^\top g_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} + \frac{\frac{1}{2}(x_{t-1}-\widehat{x}_t)^\top J_{t-1}(\widehat{x}_t)(x_{t-1}-\widehat{x}_t)}{1-\overline{\alpha}_{t-1}}, \tag{164b}$$

$$\zeta_{t,1}(x_t,x_{t-1}) = (x_{t-1}-\widehat{x}_t)^\top\frac{\int_0^1\int_0^1\gamma\Big[J_{t-1}\Big((1-\tau)\widehat{x}_t+\tau x_t(\gamma)\Big)-J_{t-1}(\widehat{x}_t)\Big]\mathrm{d}\tau\mathrm{d}\gamma}{1-\overline{\alpha}_{t-1}}(x_{t-1}-\widehat{x}_t), \tag{164c}$$

and we remind the readers that $x_t(\gamma) = \gamma x_{t-1} + (1-\gamma)\widehat{x}_t$ with $\widehat{x}_t = x_t/\sqrt{\alpha_t}$.

In order to control (163), we single out two useful facts: for every $(x_t,x_{t-1})\in\mathcal{E}$ (cf. (159)),

$$\big\|J_{t-1}\big(x_t(\gamma)\big)-J_{t-1}(\widehat{x}_t)\big\| \lesssim d^2\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log^2 T, \qquad \forall\gamma\in[0,1] \tag{165}$$

and

$$\left\|\frac{J_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{J_t(x_t)}{1-\overline{\alpha}_t}\right\| \lesssim \frac{d^2(1-\alpha_t)\log^2 T}{(\alpha_t-\overline{\alpha}_t)^2}. \tag{166}$$

The proofs of these two facts are postponed to Appendix E.2.1. These facts in turn allow us to bound

$$\big|\zeta_{t,1}(x_t,x_{t-1})\big| \le \frac{\big\|x_{t-1}-\widehat{x}_t\big\|_2^2}{2(1-\overline{\alpha}_{t-1})}\sup_{\gamma\in[0,1]}\big\|J_{t-1}\big(x_t(\gamma)\big)-J_{t-1}(\widehat{x}_t)\big\| \lesssim d^3\left(\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}\right)^{3/2}\log^3 T \tag{167}$$

and

$$\begin{aligned}&\left|\frac{(x_{t-1}-\widehat{x}_t)^\top J_{t-1}(\widehat{x}_t)(x_{t-1}-\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{(x_{t-1}-\widehat{x}_t)^\top J_t(x_t)(x_{t-1}-\widehat{x}_t)}{1-\overline{\alpha}_t}\right|\\ &\qquad\le \big\|x_{t-1}-\widehat{x}_t\big\|_2^2\left\|\frac{J_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{J_t(x_t)}{1-\overline{\alpha}_t}\right\| \lesssim \frac{d^3(1-\alpha_t)^2\log^3 T}{(\alpha_t-\overline{\alpha}_t)^2} \asymp \frac{d^3(1-\alpha_t)^2\log^3 T}{(1-\overline{\alpha}_{t-1})^2}\end{aligned} \tag{168}$$

for any $(x_t,x_{t-1})\in\mathcal{E}$. Consequently, there exists some function $f_3(\cdot)$ such that

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = f_3(x_t)\exp\Big(-f_4(x_t,x_{t-1}) + \zeta_{t,2}(x_t,x_{t-1})\Big), \tag{169a}$$

where

$$f_4(x_t,x_{t-1}) = \frac{\alpha_t\|\widehat{x}_t-x_{t-1}\|_2^2}{2(1-\alpha_t)} + \frac{(x_{t-1}-\widehat{x}_t)^\top g_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} + \frac{\frac{1}{2}(x_{t-1}-\widehat{x}_t)^\top J_t(x_t)(x_{t-1}-\widehat{x}_t)}{1-\overline{\alpha}_t}, \tag{169b}$$

$$\big|\zeta_{t,2}(x_t,x_{t-1})\big| \lesssim d^3\left(\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}\right)^{3/2}\log^3 T \lesssim \frac{d^3\log^{4.5}T}{T^{3/2}}. \tag{169c}$$

To continue, we further observe that

$$\left|\frac{(x_{t-1}-\widehat{x}_t)^\top g_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{\sqrt{\alpha_t}(x_{t-1}-\widehat{x}_t)^\top g_t(x_t)}{1-\overline{\alpha}_t}\right| \lesssim \frac{d^3\log^{3.5}T}{T^{3/2}}, \tag{170}$$

which is an immediate consequence of the following two bounds (obtained using (121c), (121b) and (37b)):

$$\left|\frac{(x_{t-1}-\widehat{x}_t)^\top g_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}}-\frac{(x_{t-1}-\widehat{x}_t)^\top g_t(x_t)}{1-\overline{\alpha}_t}\right| \lesssim d^2\sqrt{\frac{(1-\alpha_t)^3}{(\alpha_t-\overline{\alpha}_t)^3}}\log^2 T \lesssim \frac{d^3\log^{3.5}T}{T^{3/2}}$$
$$\left|(1-\sqrt{\alpha_t})\,\frac{(x_{t-1}-\widehat{x}_t)^\top g_t(x_t)}{1-\overline{\alpha}_t}\right| \lesssim \frac{d\log^2 T}{T^{3/2}}.$$

This bound (170) allows us to replace the second term on the right-hand side of (169b) with $\frac{\sqrt{\alpha_t}(x_{t-1}-\widehat{x}_t)^\top g_t(x_t)}{1-\overline{\alpha}_t}$. It is also seen from (121a) and the properties (37) that

$$\left\|\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\right\| \lesssim \frac{\log T}{T}\cdot d\log T \asymp \frac{d\log^2 T}{T}=o(1), \tag{171}$$

and therefore,

$$\left\|(1-\alpha_t)\,\frac{(x_{t-1}-\widehat{x}_t)^\top J_t(x_t)\big(x_{t-1}-\widehat{x}_t\big)}{2(1-\overline{\alpha}_t)}\right\| \lesssim \frac{d\log^2 T}{T}\cdot\big\|x_{t-1}-\widehat{x}_t\big\|_2^2 \lesssim \frac{(1-\alpha_t)\,d^2\log^3 T}{T} \lesssim \frac{d^2\log^4 T}{T^2}.$$

These combined with (169) allow us to show that: there exist some functions $f_5(\cdot)$ and $\widetilde{f}_5(\cdot)$ such that

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = f_5(x_t)\exp\Big(-f_6(x_t,x_{t-1})+\zeta_{t,3}(x_t,x_{t-1})\Big) \tag{172a}$$

where

$$\begin{aligned} f_6(x_t,x_{t-1}) &= \frac{\alpha_t\|x_{t-1}-\widehat{x}_t\|_2^2}{2(1-\alpha_t)}+\frac{\sqrt{\alpha_t}(x_{t-1}-\widehat{x}_t)^\top g_t(x_t)}{1-\overline{\alpha}_t}+\frac{\frac{1}{2}\alpha_t(x_{t-1}-\widehat{x}_t)^\top J_t(x_t)\big(x_{t-1}-\widehat{x}_t\big)}{1-\overline{\alpha}_t}\\ &= \frac{\alpha_t}{2(1-\alpha_t)}\left\{\|x_{t-1}-\mu_t(x_t)\|_2^2+\frac{1-\alpha_t}{1-\overline{\alpha}_t}(x_{t-1}-\widehat{x}_t)^\top J_t(x_t)\big(x_{t-1}-\widehat{x}_t\big)\right\}+\widetilde{f}_t(x_t), \qquad (172b)\\ \big|\zeta_{t,3}(x_t,x_{t-1})\big| &\lesssim \frac{d^3\log^{4.5}T}{T^{3/2}}. \qquad (172c)\end{aligned}$$

To further proceed, we make note of another useful fact:

$$\begin{aligned} &\frac{\alpha_t}{1-\alpha_t}\left|\big(x_{t-1}-\widehat{x}_t\big)\Big(\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(x_t)\Big)\Big(\frac{1}{\sqrt{\alpha_t}}\frac{1-\alpha_t}{1-\overline{\alpha}_t}g_t(x_t)\Big)\right|\\ &\quad\le \frac{\alpha_t}{1-\alpha_t}\big\|x_{t-1}-\widehat{x}_t\big\|_2\cdot\left\|\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\right\|\cdot\frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\mathbb{E}\left[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2\,|\,X_t=x_t\right]\\ &\quad\lesssim \sqrt{d(1-\alpha_t)\log T}\cdot\frac{d\log^2 T}{T}\cdot\frac{\log T}{T}\cdot\sqrt{d(1-\overline{\alpha}_t)\log T}\\ &\quad\asymp \frac{d^2\log^4 T}{T^2}\sqrt{1-\alpha_t}\lesssim\frac{d^2\log^{4.5}T}{T^{3/2}},\end{aligned}$$

where we have made use of the crude bound (171) in conjunction with the properties (37). Taking this observation together with (172) and (56), we can apply a little algebra to derive

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = f_7(x_t)\exp\Big(-f_8(x_t,x_{t-1})+\zeta_{t,4}(x_t,x_{t-1})\Big) \tag{173a}$$

for some function $f_7(\cdot)$, where

$$\begin{aligned} f_8(x_t,x_{t-1}) &= \frac{\alpha_t}{2(1-\alpha_t)}\left\{\big(x_{t-1}-\mu_t(x_t)\big)^\top\Big(I+\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\Big)\big(x_{t-1}-\mu_t(x_t)\big)\right\}, \qquad (173b)\\ \big|\zeta_{t,4}(x_t,x_{t-1})\big| &\lesssim \frac{d^3\log^{4.5}T}{T^{3/2}}. \qquad (173c)\end{aligned}$$

Note, however, that the covariance matrix $I + \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x_t)$ still differs from the desired one $\big(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x_t)\big)^{-2}$. As it turns out, these two matrices are fairly close to each other. To see this, we write

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

where A is a matrix obeying (see (171))

$$\|A\| \lesssim \left\| \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x_t) \right\|^2 \lesssim \frac{d^2 \log^4 T}{T^2}.$$

Consequently, we can demonstrate that

$$\begin{aligned}
&\frac{\alpha_t}{2(1-\alpha_t)} \left\| \left(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x_t)\right)^{-1} \big(x_{t-1} - \mu_t(x_t)\big) \right\|_2^2 \\
&= \frac{\alpha_t}{2(1-\alpha_t)} \left\{ \big(x_{t-1} - \mu_t(x_t)\big)^\top \left(I + \frac{1-\alpha_t}{1-\overline{\alpha}_t} J_t(x_t)\right) \big(x_{t-1} - \mu_t(x_t)\big) \right\} + O\bigg(\frac{\alpha_t}{2(1-\alpha_t)} \|A\| \, \big\|x_{t-1} - \mu_t(x_t)\big\|_2^2 \bigg) \\
&= \frac{\alpha_t}{2(1-\alpha_t)} \big(x_{t-1} - \mu_t(x_t)\big)^\top \left(I + \frac{1-\alpha_t}{1-\overline{\alpha}_t} J_t(x_t)\right) \big(x_{t-1} - \mu_t(x_t)\big) + O\bigg(\frac{d^3 \log^5 T}{T^2} \bigg).
\end{aligned}$$

To see why the last line holds, note that (according to Lemma 1 and the properties (37))

$$\begin{aligned}
\big\|x_{t-1} - \mu_t(x_t)\big\|_2 &\le \big\|x_{t-1} - \widehat{x}_t\big\|_2 + \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)} \big\| \mathbb{E}\left[\big\|x_t - \sqrt{\overline{\alpha}_t} X_0\big\|_2 \,|\, X_t = x_t \right] \big\|_2 \\
&\lesssim \sqrt{d(1-\alpha_t)\log T} + \sqrt{\frac{d \log T}{1-\overline{\alpha}_t}} (1-\alpha_t) \asymp \sqrt{d(1-\alpha_t)\log T},
\end{aligned}$$

and hence

$$\frac{\alpha_t}{1-\alpha_t} \|A\| \, \big\|x_{t-1} - \mu_t(x_t)\big\|_2^2 \lesssim \frac{d^3 \log^5 T}{T^2}.$$

Combining the above bound with (173), we arrive at

$$p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t) = f_9(x_t) \exp\Big(- f_{10}(x_t, x_{t-1}) + \zeta_{t,5}(x_t, x_{t-1}) \Big) \tag{174a}$$

for some function $f_9(\cdot)$, where

$$f_{10}(x_t, x_{t-1}) = \frac{\alpha_t}{2(1-\alpha_t)} \left\| \left(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x_t)\right)^{-1} \big(x_{t-1} - \mu_t(x_t)\big) \right\|_2^2, \tag{174b}$$

$$\big|\zeta_{t,5}(x_t, x_{t-1})\big| \lesssim \frac{d^3 \log^{4.5} T}{T^{3/2}}. \tag{174c}$$

To finish up, repeat Step 3 in the proof of Lemma 4 to yield

$$f_7(x_t) = \left(1 + O\left(\frac{d^3 \log^{4.5} T}{T^{3/2}} \right)\right) \frac{1}{\big(2\pi \frac{1-\alpha_t}{\alpha_t}\big)^{d/2} \big| \det\big(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x_t)\big) \big|}$$

as claimed. This combined with (174) concludes the proof.

E.2.1 Proof of auxiliary claims in Lemma 7

Proof of relation (165). For any $(x_t, x_{t-1}) \in \mathcal{E}$, one necessarily has

$$\|x_t(\gamma) - \widehat{x}_t\|_2 \le \|x_{t-1} - \widehat{x}_t\|_2 \le c_3 \sqrt{d(1-\alpha_t)\log T}. \tag{175}$$

Given x_t, we define the set

$$\mathcal{E}_1 \coloneqq \big\{x : \|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x\|_2 \le c_4\sqrt{d(1-\overline{\alpha}_{t-1})\log T}\big\}.$$

Then for any $x_0 \in \mathcal{E}_1$, one has

$$\begin{aligned}\big\|x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2 &\le \max\Big\{\|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\|_2, \|\widehat{x}_t - x_{t-1}\|_2\Big\} \le \max\Big\{c_4\sqrt{d(1-\overline{\alpha}_{t-1})\log T}, c_3\sqrt{d(1-\alpha_t)\log T}\Big\}\\ &= c_4\sqrt{d(1-\overline{\alpha}_{t-1})\log T},\end{aligned}$$

where the last inequality comes from (37b). This in turn reveals that

$$\begin{aligned}\left|\frac{p_{X_{t-1}\,|\,X_0}\big(x_t(\gamma)\,|\,x_0\big)}{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)} - 1\right| &= \left|\exp\left(\frac{\big\|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2^2}{2(1-\overline{\alpha}_{t-1})} - \frac{\big\|x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right) - 1\right|\\ &\le \left|\exp\left(\frac{\big\|\widehat{x}_t - x_t(\gamma)\big\|_2\big\{\big\|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2 + \big\|x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2\big\}}{2(1-\overline{\alpha}_{t-1})}\right) - 1\right|\\ &\lesssim \frac{\big\|\widehat{x}_t - x_t(\gamma)\big\|_2\big\{\big\|\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2 + \big\|x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2\big\}}{2(1-\overline{\alpha}_{t-1})}\\ &\lesssim d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T \lesssim d\sqrt{\frac{\log^3 T}{T}} = o(1),\end{aligned}$$

where the second line follows from the elementary relation

$$\Big|\|a\|_2^2 - \|b\|_2^2\Big| = \Big|\|a\|_2 - \|b\|_2\Big| \cdot (\|a\|_2 + \|b\|_2) \le \|a-b\|_2\left(\|a\|_2 + \|b\|_2\right),$$

and the last line relies on (37) and our assumption on T. Moreover, repeating the same argument as in (95) and (99), we arrive at

$$\frac{p_{X_{t-1}}\big(x_t(\gamma)\big)}{p_{X_{t-1}}(\widehat{x}_t)} = 1 + O\bigg(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T\bigg). \tag{176}$$

Putting the above results together leads to

$$\frac{p_{X_0\,|\,X_{t-1}}\big(x_0\,|\,x_t(\gamma)\big)}{p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)} = \frac{p_{X_{t-1}\,|\,X_0}\big(x_t(\gamma)\,|\,x_0\big)/p_{X_{t-1}}\big(x_t(\gamma)\big)}{p_{X_{t-1}\,|\,X_0}(\widehat{x}_t\,|\,x_0)/p_{X_{t-1}}(\widehat{x}_t)} = 1 + O\bigg(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T\bigg).$$

Equipped with the above relation, we can demonstrate that

$$\begin{aligned}&\Big\|\int p_{X_0\,|\,X_{t-1}}\big(x_0\,|\,x_t(\gamma)\big)\big(x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\mathrm{d}x_0 - \int p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\mathrm{d}x_0\Big\|_2\\ &\quad\le O\bigg(d\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log T\bigg)\cdot\bigg(\int p_{X_0\,|\,X_{t-1}}\big(x_0\,|\,x_t(\gamma)\big)\big\|x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big\|_2\mathrm{d}x_0\bigg)\\ &\qquad + \Big\|\int p_{X_0\,|\,X_{t-1}}\big(x_0\,|\,\widehat{x}_t\big)\big(x_t(\gamma) - \widehat{x}_t\big)\mathrm{d}x_0\Big\|_2 \qquad (177)\\ &\quad\lesssim \sqrt{d^3(1-\alpha_t)\log^3 T}, \qquad (178)\end{aligned}$$

where the last step invokes the property (41a). Following similar arguments (which we omit for brevity), we can also derive

$$\Big\|\int p_{X_0\,|\,X_{t-1}}\big(x_0\,|\,x_t(\gamma)\big)\big(x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\big(x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big)^\top\mathrm{d}x_0$$

$$
\qquad - \int p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)^\top \mathrm{d}x_0\Big\|
$$

$$
\lesssim d^2\sqrt{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}\log^2 T, \tag{179}
$$

where we have made use of the property (41b). Taking the above two above perturbation bounds together with the expression (73) and making use of (41a) immediately lead to the advertised result:

$$
\big\|J_{t-1}\big(x_t(\gamma)\big) - J_{t-1}(\widehat{x}_t)\big\| \lesssim d^2\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\log^2 T. \tag{180}
$$

Proof of relation (166). To establish this relation, we first apply the triangle inequality:

$$
\left\|\frac{J_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{J_t(x_t)}{1-\overline{\alpha}_t}\right\| \le \left\|\frac{J_{t-1}(\widehat{x}_t) - J_t(x_t)}{1-\overline{\alpha}_{t-1}}\right\| + \left\|\left(\frac{1}{1-\overline{\alpha}_{t-1}} - \frac{1}{1-\overline{\alpha}_t}\right)J_t(x_t)\right\|.
$$

Let us first consider the second term

$$
\left\|\left(\frac{1}{1-\overline{\alpha}_{t-1}} - \frac{1}{1-\overline{\alpha}_t}\right)J_t(x_t)\right\| = \frac{\overline{\alpha}_{t-1}(1-\alpha_t)}{(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)}\|J_t(x_t)\| \le \frac{\overline{\alpha}_{t-1}(1-\alpha_t)d\log T}{(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)} \lesssim \frac{(1-\alpha_t)d\log T}{(\alpha_t - \overline{\alpha}_t)^2}. \tag{181}
$$

where the last inequality uses (121a) and the properties (37).

Next, we move on to bound the difference $J_{t-1}(\widehat{x}_t) - J_t(x_t)$. By virtue of the relation (133), one can deduce that

$$
\begin{aligned}
&\left\|\int p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\mathrm{d}x_0 - \frac{1}{\sqrt{\alpha_t}}\int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\right\|_2 \\
&\quad = \frac{1}{\sqrt{\alpha_t}}\left\|\int p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0 - \int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\right\| \\
&\quad \le \frac{1}{\sqrt{\alpha_t}}O\left(\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\right)\int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0.
\end{aligned}
$$

In addition, recognizing that $\frac{1}{\sqrt{\alpha_t}} = 1 + \frac{1-\alpha_t}{\sqrt{\alpha_t}(1+\sqrt{\alpha_t})} = 1 + O(1-\alpha_t)$, we can further invoke the triangle inequality and (37a) to obtain

$$
\begin{aligned}
&\left\|\int p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\mathrm{d}x_0 - \int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\right\|_2 \\
&\quad \lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0.
\end{aligned}
$$

Repeating the same argument also reveals that

$$
\begin{aligned}
&\left\|\int p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)^\top\mathrm{d}x_0 - \int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0\right\| \\
&\lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}}\left\|\int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0\right\|.
\end{aligned}
$$

In view of the expression (73) for J_t, combining the preceding two bounds with a little algebra yields

$$
\begin{aligned}
&\frac{1}{1-\overline{\alpha}_{t-1}}\big\|J_{t-1}(\widehat{x}_t) - J_t(x_t)\big\| \\
&\lesssim \frac{1}{(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)}\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}} \\
&\quad \cdot \left\{\left\|\int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0\right\| + \left(\int p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0\right)^2\right\}
\end{aligned}
$$

$$\lesssim \frac{d^2(1-\alpha_t)\log^2 T}{(\alpha_t - \overline{\alpha}_t)^2},$$

where the last line follows from Lemma 1 and the properties (37).

Putting the above bounds together immediately establishes relation (166).

E.3 Proof of Lemma 8

According to the expression (158), one has

$$Y_{t-1} \mid Y_t = x_t \ \sim\ \mathcal{N}\Bigg(\mu_t(x_t), \underbrace{\frac{1-\alpha_t}{\alpha_t}\bigg(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\bigg)^2}_{=:\,\Sigma(\widehat{x}_t)}\Bigg).$$

In order to quantify the density, we first bound the Jacobian matrix $J_t(x)$ defined in (35). On the one hand, the expression (36) tells us that $J_t(x) \preceq I_d$ for any x, given that the term within the curly bracket in (36) is a negative covariance matrix. On the other hand, $J_t(x)$ can be lower bounded by

$$\begin{aligned}
J_t(x) &\succeq -\frac{1}{1-\overline{\alpha}_t}\mathbb{E}\Big[\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)^\top \mid X_t = x\Big] \\
&\succeq -\frac{\mathbb{E}\Big[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \mid X_t = x\Big]}{1-\overline{\alpha}_t}I_d \succeq -\frac{2\|x\|_2^2 + 2T^{2c_R}}{1-\overline{\alpha}_t}I_d \\
&\succeq -T^{c_0+1}\big(\|x\|_2^2 + T^{2c_R}\big)I_d,
\end{aligned}$$

where the second line applies the assumption that $\|X_0\|_2 \le T^{c_R}$, and the last line invokes the choice (20). As a consequence, we have

$$\Sigma(\widehat{x}_t) \succeq \frac{1-\alpha_t}{\alpha_t}\bigg(1 - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\bigg)^2 I_d = \frac{1-\alpha_t}{4\alpha_t}\bigg(\frac{1-\overline{\alpha}_t + \alpha_t - \overline{\alpha}_t}{1-\overline{\alpha}_t}\bigg)^2 I_d \succeq \frac{1-\alpha_t}{4\alpha_t}I_d \succeq \frac{1-\alpha_t}{4}I_d; \tag{182a}$$

$$\Sigma(\widehat{x}_t) \preceq \frac{1-\alpha_t}{\alpha_t}T^{2c_0+2}\big(2\|\widehat{x}_t\|_2^4 + 2T^{4c_R}\big)I_d \preceq 4T^{2c_0+2}\big(\|\widehat{x}_t\|_2^4 + T^{4c_R}\big)I_d. \tag{182b}$$

With the above relations in mind, we are ready to bound the density function $p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)$ for any $x_t, x_{t-1} \in \mathbb{R}^d$. It is seen from (158) that

$$\begin{aligned}
\log\frac{1}{p_{Y_{t-1}|Y_t}(x_{t-1}\,|\,x_t)} &= \frac{\big(x_{t-1} - \mu_t(x_t)\big)^\top\big(\Sigma(\widehat{x}_t)\big)^{-1}\big(x_{t-1} - \mu_t(x_t)\big)}{2} + \frac{1}{2}\log\mathsf{det}\big(\Sigma(\widehat{x}_t)\big) + \frac{d}{2}\log(2\pi) \\
&\le \frac{2\big\|x_{t-1} - \mu_t(x_t)\big\|_2^2}{1-\alpha_t} + \frac{d}{2}\log\Big(8\pi T^{2c_0+2}\big(\|\widehat{x}_t\|_2^4 + T^{4c_R}\big)\Big) \\
&\le 2T^{c_0+1}\Big\{2\big\|x_{t-1} - \widehat{x}_t\big\|_2^2 + \|x_t\|_2^2 + T^{2c_R}\Big\} + \frac{d}{2}\log\Big(8\pi T^{2c_0+2}\big(\|\widehat{x}_t\|_2^4 + T^{4c_R}\big)\Big) \\
&\le T^{c_0+2c_R+2}\Big\{\big\|x_{t-1} - \widehat{x}_t\big\|_2^2 + \|x_t\|_2^2 + 1\Big\},
\end{aligned}$$

where the second inequality results from (182), and the third inequality makes use of (138) and (20). Given that $\log\frac{p_{X_{t-1}|X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}|Y_t}(x_{t-1}\,|\,x_t)} \le \log\frac{1}{p_{Y_{t-1}|Y_t}(x_{t-1}\,|\,x_t)}$, we have concluded the proof.

References

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

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. (2021). Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

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

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

Chen, M., Huang, K., Zhao, T., and Wang, M. (2023a). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. (2021). WaveGrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*.

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

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

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

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709.

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

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

El Alaoui, A., Montanari, A., and Sellke, M. (2022). Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334.

Forbes, P. G. and Lauritzen, S. (2015). Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and its Applications*, 473:261–283.

Ghimire, S., Liu, J., Comas, A., Hill, D., Masoomi, A., Camps, O., and Dy, J. (2023). Geometry of score based generative models. *arXiv preprint arXiv:2302.04411*.

Hajek, B. (2015). *Random processes for engineers*. Cambridge university press.

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

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

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

Jolicoeur-Martineau, A., Piché-Taillefer, R., Mitliagkas, I., and des Combes, R. T. (2021). Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations*.

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pages 26565–26577.

Koehler, F., Heckett, A., and Risteski, A. (2022). Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021). DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*.

Kwon, D., Fan, Y., and Lee, K. (2022). Score-based generative modeling secretly minimizes the wasserstein distance. In *Advances in Neural Information Processing Systems*.

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

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

Li, G. and Wei, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*.

Liu, X., Wu, L., Ye, M., and Liu, Q. (2022). Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*.

Montanari, A. and Wu, Y. (2023). Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171.

Oko, K., Akiyama, S., and Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*.

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

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

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

Rout, L., Parulekar, A., Caramanis, C., and Shakkottai, S. (2023). A theoretical justification for image inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2302.01217*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

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

Song, J., Meng, C., and Ermon, S. (2020a). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021a). Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Y., Garg, S., Shi, J., and Ermon, S. (2020b). Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584.

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

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

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

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

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2022). Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*.