

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

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

June 2023; Revised: February 2024

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 ℓ_2 -accurate 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), we derive a convergence rate proportional to $1/\sqrt{T}$, matching the state-of-the-art theory. Imposing only minimal assumptions on the target data distribution (e.g., no smoothness assumption is imposed), our results characterize how ℓ_2 score estimation errors affect the quality of the data generation processes. In contrast to prior works, our theory 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, probability flow ODE, reverse SDE, denoising diffusion probabilistic model

Contents

1	Introduction	2
2	Preliminaries	4
2.1	Diffusion generative models	4
2.2	Deterministic vs. stochastic samplers: a continuous-time interpretation	5
3	Algorithms and main results	6
3.1	Assumptions and learning rates	6
3.2	Deterministic samplers	8
3.3	Stochastic samplers	10
4	Other related works	12
5	Analysis	13
5.1	Preliminary facts	13
5.2	Analysis for the sampler based on probability flow ODE (Theorem 1)	15
5.3	Analysis for the DDPM-type sampler (Theorem 3)	20

*Department of Statistics, The Chinese University of Hong Kong, Hong Kong.

†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.

6 Discussion	24
A Proof for several preliminary facts	24
A.1 Proof of properties (38)	24
A.2 Proof of properties (39) regarding the learning rates	25
A.3 Proof of Lemma 1	26
A.4 Proof of Lemma 2	29
A.5 Proof of Lemma 3	32
B Proof of auxiliary lemmas for the ODE-based sampler	32
B.1 Proof of Lemma 4	32
B.2 Proof of Lemma 5	37
B.3 Proof of Lemma 6	38
B.4 Proof of Lemma 7	39
C Proofs of auxiliary lemmas for the DDPM-type sampler	42
C.1 Proof of Lemma 8	42
C.2 Proof of Lemma 9	48
C.3 Proof of Lemma 10	49
D Analysis for the accelerated deterministic sampler (Theorem 2)	52
D.1 Proof of Theorem 2	52
D.2 Proof of Lemma 11	55
E Analysis for the accelerated stochastic sampler (Theorems 4)	58
E.1 Proof of Theorem 4	59
E.2 Proof of Lemma 12	60
E.3 Proof of Lemma 13	65

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; Jolicœur-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

$$X_0 \rightarrow X_1 \rightarrow \dots \rightarrow X_T \tag{1}$$

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

$$Y_T \rightarrow Y_{T-1} \rightarrow \dots \rightarrow Y_0 \tag{2}$$

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 proximity¹

$$Y_t \stackrel{d}{\approx} X_t, \quad t = T, \dots, 1 \quad (3)$$

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; Haussmann and Pardoux, 1986) (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); Koehler et al. (2023)).

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; Tang, 2023; Tang and Zhao, 2024; Li et al., 2024a). For instance, polynomial-time convergence guarantees have been established for stochastic samplers (e.g., Chen et al. (2022b,a); Benton et al. (2023a); Tang (2023)) and deterministic samplers (e.g., Chen et al. (2023c); Benton et al. (2023b)), 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 sampler); and (ii) a DDPM-type stochastic sampler motivated by reverse-time SDEs. We impose only minimal assumptions on the target data distribution (e.g., no smoothness condition is needed), and would like to quantify the impact of ℓ_2 score estimation errors. 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 variation (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). This improves upon prior convergence guarantees considerably (Chen et al., 2023c) and does not exhibit exponential dependency on the smoothness or regularity conditions as in Chen et al. (2023c); Benton et al. (2023b) (e.g., the regularity parameter used in Benton et al. (2023b) might even scale with the dimension d). For another DDPM-type stochastic sampler, we establish an iteration complexity proportional to $1/\varepsilon^2$ via a new non-asymptotic analysis framework, matching existing theory Chen et al. (2022b,a); Benton et al. (2023a) in terms of the ε -dependency.
- *ℓ_2 score estimation errors for the deterministic sampler.* Our theory for the deterministic sampler reveals that the TV distance between X_1 and Y_1 are shown to be proportional to the ℓ_2 score estimation error as well as the associated mean Jacobian errors. As far as we know, this is the first result for this deterministic sampler that accounts for score estimation errors in discrete time. In comparison, existing theoretical results that accommodate score errors for the probability flow ODE approach either study

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

stochastic variations of this deterministic sampler (Chen et al., 2023b) (so that the samplers are no longer the original deterministic sampler) or fall short of accommodating discretization errors (Benton et al., 2023b).

- *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.
- *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).

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 $\tilde{O}(\cdot)$ is defined similar to $O(\cdot)$ except that it hides the logarithmic dependency. Additionally, the notation $f(d, T) = o(g(d, T))$ means that $f(d, T)/g(d, T) \rightarrow 0$ 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 $\text{TV}(P, Q) := \frac{1}{2} \int |dP - dQ|$. 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. $\|A\|_{\text{F}}$) the spectral norm (resp. Frobenius norm) of A . Also, for any vector-valued function f , we let J_f or $\frac{\partial f}{\partial x}$ represent the Jacobian matrix of f .

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 $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_T$ as follows:

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

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

where $\{W_t\}_{1 \leq t \leq T}$ indicates a sequence of independent noise vectors drawn from $W_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. 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

$$\alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t := \prod_{k=1}^t \alpha_k, \quad 1 \leq t \leq T, \tag{5}$$

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

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

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

$$q_t := \text{law}(X_t) \quad (7)$$

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

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

The reverse process. The reverse chain $Y_T \rightarrow Y_{T-1} \rightarrow \dots \rightarrow Y_1$ 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

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

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

$$\text{(goal)} \quad Y_t \stackrel{d}{\approx} X_t, \quad t = T, T-1, \dots, 1. \quad (9b)$$

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

$$p_t := \text{law}(Y_t). \quad (10)$$

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:

$$Y_{t-1} = \Phi_t(Y_t), \quad t = T, \dots, 1. \quad (11)$$

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:

$$Y_{t-1} = \Psi_t(Y_t, Z_t), \quad t = T, \dots, 1, \quad (12)$$

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

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

$$dX_t = f(X_t, t)dt + g(t)dW_t \quad (0 \leq t \leq T), \quad X_0 \sim p_{\text{data}} \quad (13)$$

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 form² (Song et al., 2021b)

$$dX_t = -\frac{1}{2}\beta(t)X_t dt + \sqrt{\beta(t)}dW_t \quad (0 \leq t \leq T), \quad X_0 \sim p_{\text{data}} \quad (14)$$

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

²To see its connection with (4), it suffices to derive from (4) that $X_t - X_{t-dt} = \sqrt{1 - \beta_t}X_{t-dt} - X_{t-dt} + \sqrt{\beta_t}W_t \approx -\frac{1}{2}\beta_t X_{t-dt} + \sqrt{\beta_t}W_t$.

- *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)

$$dY_t^{\text{ode}} = \left(-f(Y_t^{\text{ode}}, T-t) + \frac{1}{2}g(T-t)^2 \nabla \log q_{T-t}(Y_t^{\text{ode}}) \right) dt \quad (0 \leq t \leq T), \quad Y_0^{\text{ode}} \sim q_T, \quad (15)$$

which exhibits matching distributions as follows:

$$Y_{T-t}^{\text{ode}} \stackrel{d}{=} X_t, \quad 0 \leq t \leq T.$$

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 classical results Anderson (1982); Haussmann and Pardoux (1986), one can also construct a “reverse-time” SDE

$$dY_t^{\text{sde}} = \left(-f(Y_t^{\text{sde}}, T-t) + g(T-t)^2 \nabla \log q_{T-t}(Y_t^{\text{sde}}) \right) dt + g(T-t) dZ_t^{\text{sde}} \quad (0 \leq t \leq T) \quad (16)$$

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

$$Y_{T-t}^{\text{sde}} \stackrel{d}{=} X_t, \quad 0 \leq t \leq T.$$

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. For instance, prior ODE-based theory (e.g., Chen et al. (2023b,c)) encountered certain technical challenges when analyzing the deterministic sampler directly, and our elementary approach is able to shed new light on the convergence of this important sampler.

3.1 Assumptions and learning rates

Before proceeding, we impose some assumptions on the score estimates and the target data distributions, and specify the hyper-parameters $\{\alpha_t\}$ that 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. Towards this end, let us first formally introduce the true score function as follows.

Definition 1 (Score function). *The score function, denoted by $s_t^* : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ($1 \leq t \leq T$), is defined as*

$$s_t^*(X) := \nabla \log q_t(X), \quad 1 \leq t \leq T. \quad (17)$$

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

$$s_t^* := \arg \min_{s: \mathbb{R}^d \rightarrow \mathbb{R}^d} \mathbb{E}_{W \sim \mathcal{N}(0, I_d), X_0 \sim p_{\text{data}}} \left[\left\| s(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} W) + \frac{1}{\sqrt{1 - \bar{\alpha}_t}} W \right\|_2^2 \right], \quad (18)$$

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

With Definition 1 in place, we can readily introduce the following assumptions that capture the quality of the score estimate $\{s_t\}_{1 \leq t \leq T}$ we have available.

Assumption 1. *Suppose that the score function estimate $\{s_t\}_{1 \leq t \leq T}$ obeys*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{X \sim q_t} \left[\|s_t(X) - s_t^*(X)\|_2^2 \right] \leq \varepsilon_{\text{score}}^2. \quad (19)$$

Assumption 2. *For each $1 \leq t \leq T$, assume that $s_t(\cdot)$ is continuously differentiable, and denote by $J_{s_t^*} = \frac{\partial s_t^*}{\partial x}$ and $J_{s_t} = \frac{\partial s_t}{\partial x}$ the Jacobian matrices of $s_t^*(\cdot)$ and $s_t(\cdot)$, respectively. Assume that the score function estimate $\{s_t\}_{1 \leq t \leq T}$ obeys*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{X \sim q_t} \left[\|J_{s_t}(X) - J_{s_t^*}(X)\| \right] \leq \varepsilon_{\text{Jacobi}}. \quad (20)$$

In a nutshell, Assumption 1 reflects the ℓ_2 score estimation error, whereas Assumption 2 is concerned with the estimation error in terms of the corresponding Jacobian matrix (so as to ensure certain continuity of the score estimator). Both assumptions consider the *average* estimation errors over all T steps. As we shall see momentarily, our theory for the deterministic sampler relies on both Assumptions 1 and 2, while the theory for the stochastic sampler requires only Assumption 1. We shall discuss in Section 3.2.1 the insufficiency of Assumption 1 alone for the deterministic sampler.

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 an absolutely continuous random vector, and

$$\mathbb{P}(\|X_0\|_2 \leq R = T^{c_R} \mid X_0 \sim p_{\text{data}}) = 1 \quad (21)$$

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_R 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

$$\beta_1 = 1 - \alpha_1 = \frac{1}{T^{c_0}}; \quad (22a)$$

$$\beta_t = 1 - \alpha_t = \frac{c_1 \log T}{T} \min \left\{ \beta_1 \left(1 + \frac{c_1 \log T}{T} \right)^t, 1 \right\}. \quad (22b)$$

Remark 1. *As will be seen in the analysis, in general the discretization error depends crucially on the quantity $\frac{1 - \alpha_t}{1 - \bar{\alpha}_t}$, whereas the initialization error relies on $\bar{\alpha}_1$ and $\bar{\alpha}_T$. Our learning rates (22) are designed to make $\frac{1 - \alpha_t}{1 - \bar{\alpha}_t}$ as small as possible, while guaranteeing $\bar{\alpha}_1$ (resp. $\bar{\alpha}_T$) is close to 1 (resp. 0); see the properties in (39). In addition, note that our theoretical framework can readily accommodate much broader choices of learning rates, although the resultant convergence rates might vary.*

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 $\{s_t\}_{1 \leq t \leq T}$, a discrete-time version of the probability flow ODE approach (cf. (15)) adopts the following update rule:

$$Y_T \sim \mathcal{N}(0, I_d), \quad Y_{t-1} = \Phi_t(Y_t) \quad \text{for } t = T, \dots, 1, \quad (23a)$$

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

$$\Phi_t(x) := \frac{1}{\sqrt{\alpha_t}} \left(x + \frac{1 - \alpha_t}{2} s_t(x) \right). \quad (23b)$$

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 variation distance between the forward and the reverse processes — for the above deterministic sampler (23). The proof of this result is postponed to Section 5.2.

Theorem 1. *Suppose that (21) holds true. Assume that the score estimates $s_t(\cdot)$ ($1 \leq t \leq T$) satisfy Assumptions 1 and 2. Then the sampling process (23) with the learning rate schedule (22) satisfies*

$$\text{TV}(q_1, p_1) \leq C_1 \frac{d^2 \log^4 T}{T} + C_1 \frac{d^6 \log^6 T}{T^2} + C_1 \sqrt{d \log^3 T} \varepsilon_{\text{score}} + C_1 d(\log T) \varepsilon_{\text{Jacobi}} \quad (24)$$

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).

Remark 2. *Note that our theory is concerned with convergence to q_1 (the first step of the forward process). Given that $X_1 \sim q_1$ and $X_0 \sim q_0$ are exceedingly close due to the choice of α_1 , focusing on the convergence w.r.t. q_1 instead of q_0 remains practically relevant.*

Let us remark on the main implications of Theorem 1, as well as several points worth discussing. Before proceeding, we shall note that our theory is concerned with convergence to q_1 . Given that $X_1 \sim q_1$ and $X_0 \sim q_0$ are very close due to the choice of α_1 , focusing on the convergence w.r.t. q_1 instead of q_0 remains practically relevant.

Iteration complexity. Consider first the scenario that has access to perfect score estimates (i.e., $\varepsilon_{\text{score}} = 0$). In order to achieve ε -accuracy (in the sense that $\text{TV}(q_1, p_1) \leq \varepsilon$), the number of steps T only needs to exceed³

$$\tilde{O} \left(\frac{d^2}{\varepsilon} + \frac{d^3}{\sqrt{\varepsilon}} \right). \quad (25)$$

Stability. Turning to the more general case with imperfect score estimates (i.e., $\varepsilon_{\text{score}} > 0$), the deterministic sampler (23) yields a distribution whose distance to the target distribution (measured again by the TV distance) scales proportionally with $\varepsilon_{\text{score}}$ and $\varepsilon_{\text{Jacobi}}$. It is noteworthy that in addition to the ℓ_2 score estimation errors, we are in need of an assumption on the stability of the associated Jacobian matrices, which plays a pivotal in ensuring that the reverse-time deterministic process does not deviate considerably from the desired process.

³As a technical note, the suboptimal d -dependency in our theory for the deterministic sampler comes mainly from Lemma 4; the main difficulty to improve Lemma 4 lies in obtaining tighter control of some quantities regarding the conditional distribution of x_0 given x_t (e.g., the Jacobian matrix of s_t).

Insufficiency of the score estimation error assumption alone. The careful reader might wonder why we are in need of additional assumptions beyond the ℓ_2 score error stated in Assumption 1. To answer this question, we find it helpful to look at a simple example below.

- **Example.** Consider the case where $X_0 \sim \mathcal{N}(0, 1)$, and hence $X_1 \sim \mathcal{N}(0, 1)$. Suppose that the reverse process for time $t = 2$ can lead to the desired distribution if exact score function is employed, namely,

$$Y_1^* := \frac{1}{\sqrt{\alpha_2}} \left(Y_2 - \frac{1 - \alpha_2}{2} s_2^*(Y_2) \right) \sim \mathcal{N}(0, 1).$$

Now, suppose that the score estimate $s_2(\cdot)$ we have available obeys

$$s_2(y_2) = s_2^*(y_2) + \frac{2\sqrt{\alpha_2}}{1 - \alpha_2} \left\{ y_1^* - L \left\lfloor \frac{y_1^*}{L} \right\rfloor \right\} \quad \text{with } y_1^* := \frac{1}{\sqrt{\alpha_2}} \left(y_2 - \frac{1 - \alpha_2}{2} s_2^*(y_2) \right)$$

for some $L > 0$, where $\lfloor z \rfloor$ is the greatest integer not exceeding z . It follows that

$$Y_1 = Y_1^* + \frac{1 - \alpha_2}{2\sqrt{\alpha_2}} [s_2^*(Y_2) - s_2(Y_2)] = L \left\lfloor \frac{Y_1^*}{L} \right\rfloor.$$

Clearly, the score estimation error $\mathbb{E}_{X_2 \sim \mathcal{N}(0,1)} [|s_2(X_2) - s_2^*(X_2)|^2]$ can be made arbitrarily small by taking L to be sufficiently small. However, the discrete nature of Y_1 forces the TV distance to be

$$\text{TV}(Y_1, X_1) = 1.$$

The above example demonstrates that, for the deterministic sampler, the TV distance between Y_1 and X_1 might not improve as the score error decreases. As we shall see in Section 3.3.1, this is in stark contrast to the stochastic sampler. If we wish to eliminate the need of imposing Assumption 2, one potential way is to resort to other metrics (e.g., the Wasserstein distance) instead of the TV distance between Y_1 and X_1 .

Relaxing the boundedness assumption on X_0 . As it turns out, the assumption (21) can also be relaxed. Supposing that $\mathbb{P}(\|X_0\|_2 \leq B \mid X_0 \sim p_{\text{data}}) = 1$ for some quantity $B > 0$ (which is allowed to grow faster than a polynomial in T), we can readily extend our analysis to obtain

$$\text{TV}(q_1, p_1) \leq C_1 \frac{d^2 \log^4 T \log^2 B}{T} + C_1 \frac{d^6 \log^6 T \log^3 B}{T^2} + C_1 \sqrt{d \log^3 T \log B} \varepsilon_{\text{score}} + C_1 d (\log T) \varepsilon_{\text{Jacobi}}.$$

Importantly, the convergence rate depends only logarithmically in B .

Comparisons to previous works. Next, let us compare our results with past works. To the best of our knowledge, the only non-asymptotic analysis for the discretized 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 $1/\varepsilon$, suffered from exponential dependency in the Lipschitz constant of the score function, and relied on exact score estimates. 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 on the target data distribution, and makes explicit the effect of ℓ_2 score estimation errors, both of which were previously unavailable for such discrete-time deterministic samplers. Another recent work Benton et al. (2023b) studied the convergence of the probability flow ODE approach without accounting for the discretization error; the result therein also exhibited exponential dependency on a certain Lipschitz constant w.r.t. the forward flow and a regularity parameter (denoted by λ therein, which might scale with the dimension d). 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 (so strictly speaking, these variations are no longer deterministic samplers), Chen et al. (2023b) showed that $\tilde{O}(L^3 d/\varepsilon^2)$ (resp. $\tilde{O}(L^2 \sqrt{d}/\varepsilon)$) 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; one limitation of our result, however, is the sub-optimal d -dependency compared to the variants studied in Chen et al. (2023b).

3.2.2 An accelerated deterministic sampler

Thus far, we have demonstrated that the iteration complexity of the deterministic sampler (23) 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 perfect estimates of two additional quantities in addition to exact score estimates, then a modified version of the sampler (23) is able to achieve much improved convergence guarantees. These estimates are made precise in the following assumption.

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

$$w_t := \arg \min_{w: \mathbb{R}^d \rightarrow \mathbb{R}^d} \mathbb{E} \left[\left\| \frac{1}{\sqrt{1-\bar{\alpha}_t}} W + s_t^*(X_t) \right\|_2^2 + \left\| WW^\top s_t^*(X_t) - w(X_t) \right\|_2^2 \right], \quad (26)$$

where $X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t} W$. Here, the expectation is with respect to $W \sim \mathcal{N}(0, I_d)$ and $X_0 \sim p_{\text{data}}$.

Armed with the score estimate in Assumption 1 and the additional estimates in Assumption 3, we are ready to introduce an accelerated variant of (23) as follows:

$$Y_T \sim \mathcal{N}(0, I_d), \quad Y_{t-1} = \Phi_t(Y_t) \quad \text{for } t = T, \dots, 1, \quad (27a)$$

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

$$\Phi_t(x) = \frac{1}{\sqrt{\alpha_t}} \left(x + \left(\frac{1-\alpha_t}{2} + \frac{(1-\alpha_t)^2}{8(1-\bar{\alpha}_t)} - \frac{(1-\alpha_t)^2}{8} \|s_t^*(x)\|_2^2 \right) s_t^*(x) + \frac{(1-\alpha_t)^2}{8(1-\bar{\alpha}_t)} w_t(x) \right). \quad (27b)$$

Notably, this new variant (27) is closely related to the original sampler (23); 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 (27), assuming access to exact score functions. The proof of our convergence result below is postponed to Section D.1.

Theorem 2. *Suppose that (21) holds true and that the score estimates are perfect (i.e., $s_t = s_t^*$). Equipped with the estimates in Assumptions 3 and the learning rate schedule (22), the sampling process (27) obeys*

$$\text{TV}(q_1, p_1) \leq C_1 \frac{d^6 \log^6 T}{T^2} \quad (28)$$

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 $\text{TV}(q_1, p_1) \leq \varepsilon$, the accelerated deterministic sampler (27) only requires the number of steps T to be on the order of

$$\tilde{O}\left(\frac{d^3}{\sqrt{\varepsilon}}\right), \quad (29)$$

thus improving the dependency on ε from $\tilde{O}(1/\varepsilon)$ (cf. (25)) to $\tilde{O}(1/\sqrt{\varepsilon})$ 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 $\{s_t\}$, we can readily introduce the following stochastic sampler that operates in discrete time, motivated by the reverse-time SDE (16):

$$Y_T \sim \mathcal{N}(0, I_d), \quad Y_{t-1} = \Psi_t(Y_t, Z_t) \quad \text{for } t = T, \dots, 1 \quad (30a)$$

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

$$\Psi_t(y, z) = \frac{1}{\sqrt{\alpha_t}} \left(y + (1 - \alpha_t) s_t(y) \right) + \sigma_t z \quad \text{with } \sigma_t^2 = \frac{1}{\alpha_t} - 1. \quad (30b)$$

The key difference between this sampler and the deterministic sampler (23) 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) as well as a very recent paper Benton et al. (2023a). 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 (21) holds true. Equipped with the estimates in Assumption 1 and the learning rate schedule (22), the stochastic sampler (30) achieves*

$$\text{TV}(q_1, p_1) \leq \sqrt{\frac{1}{2} \text{KL}(q_1 \parallel p_1)} \leq C_1 \frac{d^2 \log^3 T}{\sqrt{T}} + C_1 \sqrt{d} \varepsilon_{\text{score}} \log^2 T \quad (31)$$

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

Theorem 3 establishes non-asymptotic convergence guarantees for the stochastic sampler (30). As asserted by the theorem, if we have access to perfect score estimates, then 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. In addition, in the presence of score estimation error, the sampler achieves a TV distance proportional to $\varepsilon_{\text{score}}$, again consistent with prior results. 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 (30), assuming access to reliable estimates of additional objects as described below.

Assumption 4. *Suppose that we have access to the estimates $v_t : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ ($1 \leq t \leq T$) as follows:*

$$v_t := \arg \min_{v: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d} \mathbb{E} \left[\left\| WW^\top Z - v(\sqrt{\alpha_t} X + \sqrt{1 - \alpha_t} W, Z) \right\|_2^2 \right], \quad 1 \leq t \leq T, \quad (32)$$

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

With perfect score estimates as well as the additional estimates in Assumption 1 and Assumption 4 in place, we are positioned to introduce the proposed accelerated sampler as follows:

$$Y_T \sim \mathcal{N}(0, I_d), \quad Y_{t-1} = \Psi_t(Y_t, Z_t) \quad \text{for } t = T, \dots, 1, \quad (33a)$$

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

$$\Psi_t(y, z) = \frac{1}{\sqrt{\alpha_t}} \left(y + (1 - \alpha_t) s_t(y) \right) + \sigma_t \left\{ z - \frac{1 - \alpha_t}{2(1 - \bar{\alpha}_t)} \left[z + (1 - \bar{\alpha}_t) s_t(y) s_t(y)^\top z - v_t(y, z) \right] \right\} \quad (33b)$$

with

$$\sigma_t^2 = \frac{1}{\alpha_t} - 1. \quad (33c)$$

Clearly, the modified update mapping (33b) 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 (32)) 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 (21) holds true and that the score estimates are exact (i.e., $s_t = s_t^*$). Equipped with the estimates in Assumption 1, 4 and the learning rate schedule (22), the sampling process (33) satisfies*

$$\text{TV}(q_1, p_1) \leq \sqrt{\frac{1}{2} \text{KL}(q_1 \parallel p_1)} \leq C_1 \frac{d^3 \log^{4.5} T}{T} \quad (34)$$

for some universal constants $C_1 > 0$, provided that $T \geq C_2 d^3 \log^{4.5} T$ 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 (30), Theorem 4 asserts that the iteration complexity of the sampler (33) is at most

$$\tilde{O}\left(\frac{d^3}{\varepsilon}\right), \quad (35)$$

thus significantly reducing the scaling $\tilde{O}(1/\varepsilon^2)$ for the original sampler (30) to $\tilde{O}(1/\varepsilon)$ 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

Convergence theory for diffusion models. 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-Sobolev inequality. 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-Sobolev 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 interspersed 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). Moreover, convergence results based on the Wasserstein distance have recently been derived as well (e.g., Tang (2023); Benton et al. (2023b)), although these results typically exhibit exponential dependency on the Lipschitz constants of the score functions. Theoretical guarantees have also recently been extended to accommodate other popular methods like consistency models (Song et al., 2023; Li et al., 2024b) and diffusion guidance (Ho and Salimans, 2022; Wu et al., 2024).

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. Furthermore, the recent work Koehler et al. (2023) studied the statistical efficiency of score matching by connecting it with the isoperimetric properties of the target data 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., (23) and (30)). 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 function and the property of the minimum mean square error (MMSE) estimator (e.g., [Hajek \(2015, Section 3.3.1\)](#)), we know that the true score function s_t^* is given by the conditional expectation

$$\begin{aligned} s_t^*(x) &= \mathbb{E} \left[-\frac{1}{\sqrt{1-\bar{\alpha}_t}} W \mid \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t} W = x \right] = \frac{1}{1-\bar{\alpha}_t} \mathbb{E} [\sqrt{\bar{\alpha}_t} X_0 - x \mid \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t} W = x] \\ &= -\frac{1}{1-\bar{\alpha}_t} \underbrace{\int_{x_0} (x - \sqrt{\bar{\alpha}_t} x_0) p_{X_0|X_t}(x_0|x) dx_0}_{=: g_t(x)}. \end{aligned} \quad (36)$$

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

$$J_t(x) := \frac{\partial g_t(x)}{\partial x}, \quad (37)$$

which can be equivalently rewritten as

$$\begin{aligned} J_t(x) &= I_d + \frac{1}{1-\bar{\alpha}_t} \left\{ \mathbb{E}[X_t - \sqrt{\bar{\alpha}_t} X_0 \mid X_t = x] \left(\mathbb{E}[X_t - \sqrt{\bar{\alpha}_t} X_0 \mid X_t = x] \right)^\top \right. \\ &\quad \left. - \mathbb{E} \left[(X_t - \sqrt{\bar{\alpha}_t} X_0)(X_t - \sqrt{\bar{\alpha}_t} X_0)^\top \mid X_t = x \right] \right\}. \end{aligned} \quad (38)$$

Properties about the learning rates. Next, we isolate a few useful properties about the learning rates as specified by $\{\alpha_t\}$ in (22):

$$\alpha_t \geq 1 - \frac{c_1 \log T}{T} \geq \frac{1}{2}, \quad 1 \leq t \leq T \quad (39a)$$

$$\frac{1}{2} \frac{1-\alpha_t}{1-\bar{\alpha}_t} \leq \frac{1}{2} \frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t} \leq \frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}} \leq \frac{4c_1 \log T}{T}, \quad 2 \leq t \leq T \quad (39b)$$

$$1 \leq \frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}} \leq 1 + \frac{4c_1 \log T}{T}, \quad 2 \leq t \leq T \quad (39c)$$

$$\bar{\alpha}_T \leq \frac{1}{T^{c_2}}, \quad (39d)$$

provided that T is large enough. Here, c_1 is defined in (22), and $c_2 \geq 1000$ is some large numerical constant. In addition, if $\frac{d(1-\alpha_t)}{\alpha_t - \bar{\alpha}_t} \lesssim 1$, then one has

$$\left(\frac{1 - \bar{\alpha}_t}{\alpha_t - \bar{\alpha}_t}\right)^{d/2} = 1 + \frac{d(1 - \alpha_t)}{2(\alpha_t - \bar{\alpha}_t)} + \frac{d(d-2)(1 - \alpha_t)^2}{8(\alpha_t - \bar{\alpha}_t)^2} + O\left(d^3 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^3\right). \quad (39e)$$

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

Properties of the forward process. Additionally, recall that the forward process satisfies $X_t \stackrel{d}{=} \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} W$ 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*

$$\mathbb{P}(\|X_0\|_2 \leq R) = 1 \quad \text{and} \quad R = T^{c_R}. \quad (40)$$

Consider any $y \in \mathbb{R}$, and let

$$\theta(y) := \max \left\{ \frac{-\log p_{X_t}(y)}{d \log T}, c_6 \right\} \quad (41)$$

for some large enough constant $c_6 \geq 2c_R + c_0$. Then for any quantity $c_5 \geq 2$, conditioned on $X_t = y$ one has

$$\|\sqrt{\bar{\alpha}_t} X_0 - y\|_2 \leq 5c_5 \sqrt{\theta(y) d(1 - \bar{\alpha}_t) \log T} \quad (42)$$

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

$$\mathbb{E} [\|\sqrt{\bar{\alpha}_t} X_0 - y\|_2 \mid X_t = y] \leq 12 \sqrt{\theta(y) d(1 - \bar{\alpha}_t) \log T}, \quad (43a)$$

$$\mathbb{E} [\|\sqrt{\bar{\alpha}_t} X_0 - y\|_2^2 \mid X_t = y] \leq 120 \theta(y) d(1 - \bar{\alpha}_t) \log T, \quad (43b)$$

$$\mathbb{E} [\|\sqrt{\bar{\alpha}_t} X_0 - y\|_2^3 \mid X_t = y] \leq 1040 (\theta(y) d(1 - \bar{\alpha}_t) \log T)^{3/2}, \quad (43c)$$

$$\mathbb{E} [\|\sqrt{\bar{\alpha}_t} X_0 - y\|_2^4 \mid X_t = y] \leq 10080 (\theta(y) d(1 - \bar{\alpha}_t) \log T)^2. \quad (43d)$$

In order to interpret Lemma 1, let us look at the case with $\theta(y) = c_6$, corresponding to the scenario where $p_{X_t}(y) \geq \exp(-c_6 d \log T)$ (so that $p_{X_t}(y)$ is not exceedingly small). In this case, Lemma 1 implies that conditional on $X_t = y$ taking on a ‘‘typical’’ value, the vector $\sqrt{\bar{\alpha}_t} X_0 - X_t = \sqrt{1 - \bar{\alpha}_t} \bar{W}_t$ (see (6)) might still follow a sub-Gaussian tail, whose expected norm remains on the same order of that of an unconditional Gaussian vector $\mathcal{N}(0, (1 - \bar{\alpha}_t) I_d)$.

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*

$$-\log p_{X_t}(x_t) \leq \frac{1}{2} c_6 d \log T, \quad \text{and} \quad \left\| x_{t-1} - \frac{x_t}{\sqrt{\bar{\alpha}_t}} \right\|_2 \leq c_3 \sqrt{d(1 - \bar{\alpha}_t) \log T} \quad (44)$$

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

$$-\log p_{X_{t-1}}(x_t(\gamma)) \leq c_6 d \log T, \quad \forall \gamma \in [0, 1]. \quad (45)$$

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 $x_t / \sqrt{\bar{\alpha}_t}$, 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 (44) holds, then conditional on $X_{t-1} = x_t(\gamma)$ for any $\gamma \in [0, 1]$, one has

$$\|\sqrt{\bar{\alpha}_{t-1}} X_0 - x_t(\gamma)\|_2 \leq 5c_5 \sqrt{c_6 d(1 - \bar{\alpha}_{t-1}) \log T} \quad (46a)$$

$$\left\| \sqrt{\bar{\alpha}_{t-1}} X_0 - \frac{x_t}{\sqrt{\bar{\alpha}_t}} \right\|_2 \leq \|\sqrt{\bar{\alpha}_{t-1}} X_0 - x_t(\gamma)\|_2 + \left\| x_t(\gamma) - \frac{x_t}{\sqrt{\bar{\alpha}_t}} \right\|_2 \leq (5c_5 + c_3) \sqrt{c_6 d(1 - \bar{\alpha}_{t-1}) \log T} \quad (46b)$$

with probability exceeding $1 - \exp(-c_5^2 c_6 d \log T)$, where the last inequality invokes the property (39b).

Distance between p_T and q_T . We now record a simple result that demonstrates the proximity between p_T and q_T , whose proof is provided in Appendix A.5.

Lemma 3. *For any large enough T , it holds that*

$$\left(\text{TV}(p_{X_T} \parallel p_{Y_T})\right)^2 \leq \frac{1}{2} \text{KL}(p_{X_T} \parallel p_{Y_T}) \lesssim \frac{1}{T^{200}}. \quad (47)$$

Additional notation about score errors. For any vector $x \in \mathbb{R}^d$ and any $1 < t \leq T$, let us define

$$\varepsilon_{\text{score},t}(x) := \|s_t(x) - s_t^*(x)\|_2 \quad \text{and} \quad \varepsilon_{\text{Jacobi},t}(x) := \|J_{s_t}(x) - J_{s_t^*}(x)\|, \quad (48)$$

with J_{s_t} and $J_{s_t^*}$ the Jacobian matrices of $s_t(\cdot)$ and $s_t^*(\cdot)$, respectively. Under Assumption 1, we have

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{X \sim q_t} [\varepsilon_{\text{score},t}(X)] \leq \left(\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{X \sim q_t} [\varepsilon_{\text{score},t}(X)^2] \right)^{1/2} \leq \varepsilon_{\text{score}}. \quad (49a)$$

Also, Assumption 2 says that

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{X \sim q_t} [\varepsilon_{\text{Jacobi},t}(X)] \leq \varepsilon_{\text{Jacobi}}. \quad (49b)$$

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 (23) based on the probability flow ODE. Given that the total variation distance is always bounded above by 1, it suffices to assume

$$T \geq C_1 d^2 \log^4 T + \sqrt{C_1} d^3 \log^3 T \quad (50a)$$

$$\varepsilon_{\text{score}} \leq \frac{1}{C_1 \sqrt{d} \log^2 T} \quad (50b)$$

$$\varepsilon_{\text{Jacobi}} \leq \frac{1}{C_1 d \log^2 T} \quad (50c)$$

throughout the proof; otherwise the claimed result (24) becomes trivial.

Preparation. Before proceeding, we find it convenient to introduce a function

$$\phi_t^*(x) = x + \frac{1 - \alpha_t}{2} s_t^*(x) = x - \frac{1 - \alpha_t}{2(1 - \bar{\alpha}_t)} \int_{x_0}^x (x - \sqrt{\bar{\alpha}_t} x_0) p_{X_0|X_t}(x_0 | x) dx_0, \quad (51a)$$

$$\phi_t(x) = x + \frac{1 - \alpha_t}{2} s_t(x), \quad (51b)$$

where the first line follows from (36). The update rule (23) can then be expressed as follows:

$$Y_{t-1} = \Phi_t(Y_t) = \frac{1}{\sqrt{\alpha_t}} \phi_t(Y_t). \quad (52)$$

Moreover, for any point $y_T \in \mathbb{R}^d$ (resp. $y'_T \in \mathbb{R}^d$), let us define the corresponding deterministic sequence

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \phi_t(y_t), \quad y'_{t-1} = \frac{1}{\sqrt{\alpha_t}} \phi_t(y'_t), \quad t = T, T-1, \dots \quad (53)$$

In other words, $\{y_{T-1}, \dots, y_1\}$ (resp. $\{y'_{T-1}, \dots, y'_1\}$) is the (reverse-time) sequence generated by the probability flow ODE (cf. (52)) when initialized to $Y_T = y_T$ (resp. $Y_T = y'_T$). We also define the following quantities for any point $y_T \in \mathbb{R}^d$ and its associated sequence $\{y_{T-1}, \dots, y_1\}$:

$$\xi_t(y_T) := \frac{\log T}{T} (d \varepsilon_{\text{Jacobi},t}(y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(y_t)); \quad (54a)$$

$$S_t(y_T) := \sum_{1 < k \leq t} \xi_k(y_k), \quad \text{for } t \geq 2, \quad \text{and} \quad S_1(y_T) = 0. \quad (54b)$$

In words, for any given starting point y_T , $\xi_t(y_t)$ captures the (properly weighted) score error incurred in the t -th iteration, whereas $S_t(y_T)$ quantifies the aggregate weighted score error up to the t -th iteration.

With the above notation in place, we can readily proceed to our proof, which consists of several steps.

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

$$\begin{aligned} \frac{p_{Y_{t-1}}(y_{t-1})}{p_{X_{t-1}}(y_{t-1})} &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})} \\ &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{Y_t}(y_t)} \cdot \left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{X_t}(y_t)} \right)^{-1} \cdot \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)}, \end{aligned} \quad (55)$$

thus converting the density ratio of interest into the product of three other density ratios. Noteworthy, this observation (55) 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}). In light of 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 lemma is postponed to Appendix B.1.

Lemma 4. For any $x \in \mathbb{R}^d$, let

$$\theta_t(x) := \max \left\{ -\frac{\log p_{X_t}(x)}{d \log T}, c_6 \right\} \quad (56)$$

for some large enough constant $c_6 \geq 2c_R + c_0$, and suppose that $\frac{40c_1 \varepsilon_{\text{score},t}(x) \log^{\frac{3}{2}} T}{T} \leq \sqrt{\theta_t(x)d}$. Then one has

$$\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} \leq 2 \exp \left(\left(5\varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x)d \log T} + 60\theta_t(x)d \log T \right) \frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right). \quad (57)$$

If, in addition, we have $C_{10} \frac{\theta_t(x)d \log^2 T + \varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x)d \log^3 T}}{T} \leq 1$ for some large enough constant $C_{10} > 0$, then it holds that

$$\begin{aligned} &\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} \\ &= 1 + \frac{d(1 - \alpha_t)}{2(\alpha_t - \bar{\alpha}_t)} + \frac{(1 - \alpha_t) \left(\left\| \int (x - \sqrt{\bar{\alpha}_t}x_0) p_{X_0|X_t}(x_0|x) dx_0 \right\|_2^2 - \int \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 p_{X_0|X_t}(x_0|x) dx_0 \right)}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} \\ &\quad + O \left(\theta_t(x)^2 d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^2 \log^2 T + \varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x)d \log T} \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right) \right). \end{aligned} \quad (58a)$$

Moreover, for any random vector Y , one has

$$\begin{aligned} &\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)} \\ &= 1 + \frac{d(1 - \alpha_t)}{2(\alpha_t - \bar{\alpha}_t)} + \frac{(1 - \alpha_t) \left(\left\| \int (x - \sqrt{\bar{\alpha}_t}x_0) p_{X_0|X_t}(x_0|x) dx_0 \right\|_2^2 - \int \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 p_{X_0|X_t}(x_0|x) dx_0 \right)}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} \\ &\quad + O \left(\theta_t(x)^2 d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^2 \log^2 T + \theta_t(x)^3 d^6 \log^3 T \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^3 + (1 - \alpha_t) d \varepsilon_{\text{Jacobi},t}(x) \right), \end{aligned} \quad (58b)$$

provided that $C_{11} \frac{d^2 \log^2 T + d \varepsilon_{\text{Jacobi},t}(x) \log T}{T} \leq 1$ for some large enough constant $C_{11} > 0$.

Remark 3. Combining Lemma 4 with Lemma 1 and (39), gives: if $C_{10} \frac{\theta_t(x)d \log^2 T + \varepsilon_{\text{score},t}(x)\sqrt{\theta_t(x)d \log^3 T}}{T} \leq 1$ and if $\theta_t(x) \lesssim 1$, then (58a) tells us that

$$\log \frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} \leq \frac{4c_1 d \log T}{T} + C_{10} \left\{ \frac{d^2 \log^4 T}{T^2} + \frac{d^6 \log^6 T}{T^3} + \frac{\varepsilon_{\text{score},t}(x)\sqrt{d \log^3 T}}{T} \right\} \quad (59)$$

under our sample size assumption (50), where $C_{10} > 0$ is some large enough constant. Here, we have made use of the fact that the second-to-last term in (58a) is non-positive due to Jensen's inequality.

Step 2: decomposing the TV distance based on ‘‘typical’’ points. To bound the TV distance of interest, it is helpful to isolate the following sets

$$\mathcal{E} := \left\{ y : q_1(y) > \max \{ p_1(y), \exp(-c_6 d \log T) \} \right\}, \quad (60)$$

where $c_6 > 0$ is some large enough universal constant introduced in Lemma 4. In words, this set \mathcal{E} contains all y that can be viewed as ‘‘typical’’ values under the distribution q_1 (meaning that $q_1(y)$ is not exceedingly small), while at the same time obeying $q_1(y) > p_1(y)$.

In view of the basic properties about the TV distance, we can derive

$$\begin{aligned} \text{TV}(q_1, p_1) &= \int_{y:q_1(y)>p_1(y)} (q_1(y) - p_1(y)) dy \\ &= \int_{y \in \mathcal{E}} (q_1(y) - p_1(y)) dy + \int_{y:p_1(y)<q_1(y) \leq \exp(-c_6 d \log T)} (q_1(y) - p_1(y)) dy. \end{aligned} \quad (61)$$

In order to bound the second term on the right-hand side of (61), we make note of a basic fact: since $X_t \stackrel{(d)}{=} \sqrt{\alpha_t}X_0 + \sqrt{1-\alpha_t}W$ with $W \sim \mathcal{N}(0, I_d)$ and $\mathbb{P}(\|X_0\|_2 \leq T^{c_R}) = 1$, it holds that

$$\mathbb{P} \{ \|X_t\|_2 \geq T^{c_R+2} \} \leq \mathbb{P} \{ \|W\|_2 \geq T^2 \} < \exp(-c_6 d \log T) \quad (62)$$

under our assumption (50) on T , thereby indicating that

$$\int_{y:\|y\|_2 \geq T^{c_R+2}} q_t(y) dy < \exp(-c_6 d \log T). \quad (63)$$

This basic fact in turn reveals that

$$\begin{aligned} \int_{y:p_1(y)<q_1(y) \leq \exp(-c_{12} d \log T)} (q_1(y) - p_1(y)) dy &\leq \int_{y:q_1(y) \leq \exp(-c_6 d \log T)} q_1(y) dy \\ &\leq \exp(-c_6 d \log T) \int_{y:\|y\|_2 \leq T^{c_R+2}} dy + \exp(-c_6 d \log T) \\ &\leq \exp(-c_6 d \log T) (2T^{c_R+2})^d + \exp(-c_6 d \log T) \\ &\leq \exp(-0.5c_6 d \log T), \end{aligned}$$

provided that $c_6 \geq 4(c_R + 2)$. Substitution into (61) then yields

$$\text{TV}(q_1, p_1) \leq \mathbb{E}_{Y_1 \sim p_1} \left[\left(\frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right) \mathbb{1} \{ Y_1 \in \mathcal{E} \} \right] + \exp(-c_6 d \log T), \quad (64)$$

with the proviso that $c_6 \geq 4(c_R + 2)$.

To proceed, let us isolate the following set

$$\mathcal{I}_1 := \left\{ y_T \mid S_T(y_T) \leq c_{14} \right\} \quad (65)$$

for some small enough constant $c_{14} > 0$. In words, \mathcal{I}_1 is composed of a set of points whose aggregate score error along the backward trajectory is well-controlled; in fact, these are points that exhibit ‘‘typical’’ behavior

under the assumptions (50b) and (50c). As a result, we can decompose the first term of (64) into the influence of “typical” points and that of the remaining points as follows:

$$\begin{aligned} \mathbb{E}_{Y_1 \sim p_1} \left[\left(\frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right) \mathbb{1}\{Y_1 \in \mathcal{E}\} \right] &= \mathbb{E}_{Y_T \sim p_T} \left[\left(\frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right) \mathbb{1}\{Y_1 \in \mathcal{E}\} \right] \\ &= \mathbb{E}_{Y_T \sim p_T} \left[\left(\frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right) \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1\} \right] + \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \notin \mathcal{I}_1\} \right], \end{aligned} \quad (66)$$

where the first identity holds since Y_1 is determined purely by Y_T via deterministic update rules. The decomposition (66) leaves us with two terms to control, which we accomplish in the next two steps.

Step 3: controlling the first term on the right-hand side of (66). This step analyzes the first term on the right-hand side of (66). We would like to make the analysis in this step slightly more general than needed, given that it will be useful for the subsequent analysis as well.

To begin with, let us introduce the following quantity:

$$\tau(y_T) := \max \left\{ 2 \leq t \leq T + 1 : S_{t-1}(y_T) \leq c_{14} \right\}, \quad (67)$$

meaning that the score errors exhibit “typical” behavior up to the $(\tau(y_T) - 1)$ -th iteration. As can be clearly seen from the definition (65) of \mathcal{I}_1 ,

$$\tau(y_T) = T + 1, \quad \forall y_T \in \mathcal{I}_1. \quad (68)$$

In the sequel, we first single out the following lemma, whose proof is deferred to Appendix B.2.

Lemma 5. *Consider any y_T and its associated sequence $\{y_{T-1}, \dots, y_1\}$ (see (53)). If $-\log q_1(y_1) \leq c_6 d \log T$, then one has*

$$-\log q_k(y_k) \leq 2c_6 d \log T \quad (69)$$

for any $1 \leq k < \tau(y_T)$ (cf. (67)), provided that $c_6 \geq 3c_1$.

As a consequence of Lemma 5, we are able to control the density ratio q_t/p_t up to the $(\tau(y_T) - 1)$ -th iteration, as stated in the following lemma. The proof can be found in Appendix B.3.

Lemma 6. *Consider any y_T , along with the deterministic sequence $\{y_{T-1}, \dots, y_1\}$ (cf. (53)), and set $\tau = \tau(y_T)$ (cf. (67)). Then one has*

$$\frac{q_1(y_1)}{p_1(y_1)} = \left\{ 1 + O \left(\frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + S_{\tau-1}(y_{\tau-1}) \right) \right\} \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})}, \quad (70a)$$

$$\text{and} \quad \frac{q_k(y_k)}{2p_k(y_k)} \leq \frac{q_1(y_1)}{p_1(y_1)} \leq 2 \frac{q_k(y_k)}{p_k(y_k)}, \quad \forall k < \tau. \quad (70b)$$

Now let us look at the set \mathcal{I}_1 . Taking $\tau(y_T) = T + 1$ (cf. (68)) in Lemma 6 yields

$$\begin{aligned} &\mathbb{E}_{Y_T \sim p_T} \left[\left(\frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right) \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1\} \right] \\ &= \mathbb{E}_{Y_T \sim p_T} \left[\left(\left\{ 1 + O \left(\frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + S_T(y_T) \right) \right\} \frac{q_T(Y_T)}{p_T(Y_T)} - 1 \right) \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1\} \right] \\ &= \int \left\{ \left(1 + O \left(\frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + S_T(y_T) \right) \right) q_T(y_T) - p_T(y_T) \right\} \mathbb{1}\{y_1 \in \mathcal{E}, y_T \in \mathcal{I}_1\} dy_T \\ &\leq \int |q_T(y_T) - p_T(y_T)| dy_T + O \left(\frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} \right) \int q_T(y_T) dy_T + O \left(\sqrt{d \log^3 T} \varepsilon_{\text{score}} + (d \log T) \varepsilon_{\text{Jacobi}} \right) \end{aligned}$$

$$\lesssim \frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + \sqrt{d \log^3 T \varepsilon_{\text{score}}} + (d \log T) \varepsilon_{\text{Jacobi}}. \quad (71)$$

Here, the last line holds since $\text{TV}(p_T, q_T) \lesssim T^{-100}$ (according to Lemma 4), and the penultimate line follows from the observation below:

$$\begin{aligned} & \int S_T(y_T) q_T(y_T) \mathbb{1}\{y_1 \in \mathcal{E}, y_T \in \mathcal{I}_1\} dy_T \\ &= \frac{\log T}{T} \sum_{t=1}^T \int \left(d\varepsilon_{\text{Jacobi},t}(y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(y_t) \right) q_T(y_T) \mathbb{1}\{y_1 \in \mathcal{E}, y_T \in \mathcal{I}_1\} dy_T \\ &\leq \frac{4 \log T}{T} \sum_{t=1}^T \int \left(d\varepsilon_{\text{Jacobi},t}(y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(y_t) \right) \frac{q_t(y_t)}{p_t(y_t)} p_T(y_T) \mathbb{1}\{y_1 \in \mathcal{E}, y_T \in \mathcal{I}_1\} dy_T \\ &\leq \frac{4 \log T}{T} \sum_{t=1}^T \mathbb{E}_{Y_T \sim p_T} \left[\left(d\varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \frac{q_t(Y_t)}{p_t(Y_t)} \right] \\ &= \frac{4 \log T}{T} \sum_{t=1}^T \mathbb{E}_{Y_t \sim p_t} \left[\left(d\varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \frac{q_t(Y_t)}{p_t(Y_t)} \right] \\ &= \frac{4 \log T}{T} \sum_{t=1}^T \mathbb{E}_{Y_t \sim q_t} \left[d\varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right] \\ &\lesssim (d \log T) \varepsilon_{\text{Jacobi}} + \sqrt{d \log^3 T \varepsilon_{\text{score}}}, \end{aligned}$$

where the first inequality is due to (70), and the last relation comes from (49).

Step 4: controlling the second term on the right-hand side of (66). In this step, we find it helpful to introduce the following sets (in addition to \mathcal{I}_1 defined in (65)), where we again abbreviate $\tau = \tau(y_T)$ as long as it is clear from the context:

$$\mathcal{I}_2 := \left\{ y_T : c_{14} \leq S_\tau(y_T) \leq 2c_{14} \right\}, \quad (72a)$$

$$\mathcal{I}_3 := \left\{ y_T : S_{\tau-1}(y_T) \leq c_{14}, \xi_\tau(y_T) \geq c_{14}, \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})} \leq \frac{8q_\tau(y_\tau)}{p_\tau(y_\tau)} \right\}, \quad (72b)$$

$$\mathcal{I}_4 := \left\{ y_T : S_{\tau-1}(y_T) \leq c_{14}, \xi_\tau(y_T) \geq c_{14}, \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})} > \frac{8q_\tau(y_\tau)}{p_\tau(y_\tau)} \right\}. \quad (72c)$$

It follows immediately from the definition that $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4 = \mathbb{R}^d$. In words, for any point y_T in \mathcal{I}_2 , the resulting score error remains well-controlled in the τ -th iteration; in comparison, the points in \mathcal{I}_3 and \mathcal{I}_4 might incur large score errors in the τ -th iteration. The difference between \mathcal{I}_3 and \mathcal{I}_4 then lies in the comparison between the density ratios q_t/p_t in the $(\tau-1)$ -th and the τ -th iteration.

We shall tackle each of these sets separately, with the combined result summarized in the lemma below.

Lemma 7. *It holds that*

$$\mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4\} \right] \lesssim \frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + \sqrt{d \log^3 T \varepsilon_{\text{score}}} + (d \log T) \varepsilon_{\text{Jacobi}}. \quad (73)$$

See Appendix B.4 for the proof of this lemma.

Step 5: putting all pieces together. Recall that $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4 = \mathbb{R}^d$. Taking (193), (66), (71) and (73) collectively, we conclude that

$$\text{TV}(p_1, q_1) \leq \mathbb{E}_{Y_T \sim p_T} \left[\left(\frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right) \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1\} \right]$$

$$\begin{aligned}
& + \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1} \{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4\} \right] + \exp(-c_6 d \log T) \\
& \lesssim \frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + \sqrt{d \log^3 T} \varepsilon_{\text{score}} + d \varepsilon_{\text{Jacobi}} \log T
\end{aligned}$$

as claimed.

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

Turning attention to the DDPM-type stochastic sampler (30), we now present the main steps for the proof of Theorem 3.

Preparation. Let us first introduce the following mapping

$$\mu_t^*(x_t) := \frac{1}{\sqrt{\alpha_t}} \left(x_t + (1 - \alpha_t) s_t^*(x_t) \right) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}(1 - \bar{\alpha}_t)} \int_{x_0} p_{X_0 | X_t}(x_0 | x_t) (x_t - \sqrt{\bar{\alpha}_t} x_0) dx_0, \quad (74a)$$

$$\mu_t(x_t) := \frac{1}{\sqrt{\alpha_t}} \left(x_t + (1 - \alpha_t) s_t(x_t) \right), \quad (74b)$$

where the first line relies on the expression (36). For any t , let us also introduce an auxiliary vector

$$Y_{t-1}^* := \frac{1}{\sqrt{\alpha_t}} \left(Y_t + (1 - \alpha_t) s_t^*(Y_t) \right) + \sigma_t Z_t, \quad (75)$$

which applies the DDPM update rule from Y_t using the true score function s_t^* . From the update rule (30) and (75), we can write

$$p_{Y_{t-1} | Y_t}(x_{t-1} | x_t) = \frac{1}{(2\pi \frac{1-\alpha_t}{\alpha_t})^{d/2}} \exp \left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t(x_t)\|_2^2 \right) \quad (76a)$$

$$p_{Y_{t-1}^* | Y_t}(x_{t-1} | x_t) = \frac{1}{(2\pi \frac{1-\alpha_t}{\alpha_t})^{d/2}} \exp \left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t^*(x_t)\|_2^2 \right) \quad (76b)$$

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

$$\hat{x}_t := \frac{1}{\sqrt{\alpha_t}} x_t. \quad (77)$$

Armed with this set of notation, we are ready to present the proof of Theorem 3, which consists of 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

$$\text{TV}(p_{X_1}, p_{Y_1}) \leq \sqrt{\frac{1}{2} \text{KL}(p_{X_1} \parallel p_{Y_1})}, \quad (78)$$

and hence it suffices to work with the KL divergence. Recall that $X_1 \rightarrow \dots \rightarrow X_T$ and $Y_T \rightarrow \dots \rightarrow Y_1$ 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:

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

where we recall that q_t (resp. p_t) denotes the distribution of X_t (resp. Y_t). This combined with the non-negativity of the KL divergence indicates that

$$\text{KL}(p_{X_1} \parallel p_{Y_1}) \leq \text{KL}(p_{X_T} \parallel p_{Y_T}) + \sum_{t=2}^T \mathbb{E}_{x \sim q_t} \left[\text{KL} \left(p_{X_{t-1} | X_t}(\cdot | x) \parallel p_{Y_{t-1} | Y_t}(\cdot | x) \right) \right]. \quad (79)$$

This allows us to focus attention on the transition probabilities at each time instance t . On the right-hand side of (79), the term that is the easiest to bound is $\text{KL}(p_{X_T} \parallel p_{Y_T})$. It has been shown in Lemma 3 that

$$\text{KL}(p_{X_T} \parallel p_{Y_T}) \lesssim \frac{1}{T^{200}}.$$

Thus, it suffices to focus attention on bounding $\text{KL}(p_{X_{t-1} | X_t}(\cdot | x) \parallel p_{Y_{t-1} | Y_t}(\cdot | x))$ for each $1 < t \leq T$, 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 (79), 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

$$\mathcal{E} := \left\{ (x_t, x_{t-1}) \mid -\log p_{X_t}(x_t) \leq \frac{1}{2} c_6 d \log T, \|x_{t-1} - \hat{x}_t\|_2 \leq c_3 \sqrt{d(1 - \alpha_t) \log T} \right\}, \quad (80)$$

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 (76a)). The proof can be found in Appendix C.1.

Lemma 8. *There exists some large enough numerical constant $c_\zeta > 0$ such that: for every $(x_t, x_{t-1}) \in \mathcal{E}$,*

$$p_{X_{t-1} | X_t}(x_{t-1} | x_t) = \frac{1}{(2\pi \frac{1-\alpha_t}{\alpha_t})^{d/2}} \exp \left(-\frac{\alpha_t \|x_{t-1} - \mu_t^*(x_t)\|_2^2}{2(1-\alpha_t)} + \zeta_t(x_{t-1}, x_t) \right) \quad (81)$$

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

$$|\zeta_t(x_{t-1}, x_t)| \leq c_\zeta d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right) \log^2 T. \quad (82)$$

Here, we recall the definition of $\mu_t^*(x_t)$ in (74b).

By comparing Lemma 8 with expression (76a), 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 9. *For all $(x_t, x_{t-1}) \in \mathbb{R}^d \times \mathbb{R}^d$, it holds that*

$$\log \frac{p_{X_{t-1} | X_t}(x_{t-1} | x_t)}{p_{Y_{t-1} | Y_t}(x_{t-1} | x_t)} \leq 2T (\|x_{t-1} - \hat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}). \quad (83)$$

The proof of Lemma 9 is provided in Appendix C.2.

Step 3: bounding the KL divergence between $p_{X_{t-1}|X_t}$ and $p_{Y_{t-1}^*|Y_t}$. Equipped with the above two lemmas, we are positioned to first control the KL divergence between $p_{X_{t-1}|X_t}$ and $p_{Y_{t-1}^*|Y_t}$; by doing so we are neglecting the impact of the score estimation error for the moment (see the definition (75) of Y_{t-1}^*). It is first seen from Lemma 8 and (76a) that: for any $(x_t, x_{t-1}) \in \mathcal{E}$,

$$\begin{aligned} \frac{p_{X_{t-1}|X_t}(x_{t-1}|x_t)}{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)} &= \exp\left(O\left(d^2\left(\frac{1-\alpha_t}{\alpha_t-\bar{\alpha}_t}\right)\log^2 T\right)\right) = 1 + O\left(d^2\left(\frac{1-\alpha_t}{\alpha_t-\bar{\alpha}_t}\right)\log^2 T\right) \\ &= 1 + O\left(d^2\frac{\log^3 T}{T}\right) \in \left[\frac{1}{2}, 2\right], \end{aligned} \quad (84)$$

where the last line results from (39b) and the assumption that $T \geq c_{10}d^2\log^3 T$ for some large enough constant $c_{10} > 0$. We can then calculate

$$\begin{aligned} &\mathbb{E}_{x_t \sim q_t} \left[\text{KL}\left(p_{X_{t-1}|X_t}(\cdot|x_t) \parallel p_{Y_{t-1}^*|Y_t}(\cdot|x_t)\right) \right] \\ &= \left(\int_{\mathcal{E}} + \int_{\mathcal{E}^c} \right) p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \log \frac{p_{X_{t-1}|X_t}(x_{t-1}|x_t)}{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)} dx_{t-1} dx_t, \\ &\stackrel{(i)}{=} \int_{\mathcal{E}} p_{X_t}(x_t) \left\{ p_{X_{t-1}|X_t}(x_{t-1}|x_t) - p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t) \right. \\ &\quad \left. + p_{X_{t-1}|X_t}(x_{t-1}|x_t) \cdot O\left(\left(\frac{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)}{p_{X_{t-1}|X_t}(x_{t-1}|x_t)} - 1\right)^2\right) \right\} dx_{t-1} dx_t \\ &\quad + \int_{\mathcal{E}^c} p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \log \frac{p_{X_{t-1}|X_t}(x_{t-1}|x_t)}{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)} dx_{t-1} dx_t \\ &\stackrel{(ii)}{=} \int_{\mathcal{E}} p_{X_t}(x_t) \left\{ p_{X_{t-1}|X_t}(x_{t-1}|x_t) - p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t) + p_{X_{t-1}|X_t}(x_{t-1}|x_t) O\left(d^4\left(\frac{1-\alpha_t}{\alpha_t-\bar{\alpha}_t}\right)^2 \log^4 T\right) \right\} dx_{t-1} dx_t \\ &\quad + \int_{\mathcal{E}^c} p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \left\{ 2T(\|x_t\|_2^2 + \|x_{t-1} - \hat{x}_t\|_2^2 + T^{2c_R}) \right\} dx_{t-1} dx_t. \end{aligned} \quad (85)$$

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

$$\begin{aligned} p_X(x) \log \frac{p_X(x)}{p_Y(x)} &= -p_X(x) \log \left(1 + \frac{p_Y(x) - p_X(x)}{p_X(x)}\right) \\ &= p_X(x) - p_Y(x) + p_X(x) O\left(\left(\frac{p_Y(x)}{p_X(x)} - 1\right)^2\right); \end{aligned}$$

regarding (ii), we invoke (84) and Lemma 9.

To continue, let us bound each term on the right-hand side of (85) separately. From the definition of the set \mathcal{E} (cf. (80)), direct calculations yield

$$\begin{aligned} \mathbb{P}((X_t, X_{t-1}) \notin \mathcal{E}) &= \int_{(x_t, x_{t-1}) \notin \mathcal{E}} p_{X_{t-1}}(x_{t-1}) p_{X_t|X_{t-1}}(x_t|x_{t-1}) dx_{t-1} dx_t \\ &= \int_{(x_t, x_{t-1}) \notin \mathcal{E}} p_{X_{t-1}}(x_{t-1}) \frac{1}{(2\pi(1-\alpha_t))^{d/2}} \exp\left(-\frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|_2^2}{2(1-\alpha_t)}\right) dx_{t-1} dx_t \\ &\leq \exp(-c_3 d \log T), \end{aligned} \quad (86)$$

and similarly,

$$\int_{(x_{t-1}, x_t) \notin \mathcal{E}} p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \left(2T(\|x_t\|_2^2 + \|x_{t-1} - \hat{x}_t\|_2^2 + T^{2c_R})\right) dx_{t-1} dx_t \leq \exp(-c_3 d \log T). \quad (87)$$

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 $-\log p_{X_t}(x_t) \leq \frac{1}{2}c_6d\log T$, one can use the definition (74b) of $\mu_t^*(\cdot)$ to obtain

$$\begin{aligned} \|x_{t-1} - \mu_t^*(x_t)\|_2 &= \left\| x_{t-1} - \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\mathbb{E}\left[x_t - \sqrt{\bar{\alpha}_t}X_0 \mid X_t = x_t\right] \right\|_2 \\ &\geq \left\| x_{t-1} - \frac{1}{\sqrt{\alpha_t}}x_t \right\|_2 - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\mathbb{E}\left[\|x_t - \sqrt{\bar{\alpha}_t}X_0\|_2 \mid X_t = x_t\right] \\ &\geq c_3\sqrt{d(1-\alpha_t)\log T} - 6\bar{c}_5\frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\sqrt{d\log T} \\ &= \left(c_3 - 6\bar{c}_5\frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\right)\sqrt{d(1-\alpha_t)\log T} \geq \frac{c_3}{2}\sqrt{d(1-\alpha_t)\log T}, \end{aligned} \quad (88)$$

where the third line results from (43a) in Lemma 1, and the last line applies (39) and holds true as long as c_3 is large enough. In turn, this combined with (76a) indicates that: for any x_t obeying $-\log p_{X_t}(x_t) \leq \frac{1}{2}c_6d\log T$,

$$\int_{x_{t-1}: \|x_{t-1} - x_t/\sqrt{\alpha_t}\|_2 > c_3\sqrt{d(1-\alpha_t)\log T}} p_{Y_{t-1}^*|Y_t}(x_{t-1} \mid x_t) dx_{t-1} \leq \exp\left(-\frac{c_3}{2}d\log T\right). \quad (90)$$

As a result, (86) and (90) taken collectively demonstrate that

$$\begin{aligned} &\left| \int_{\mathcal{E}} p_{X_t}(x_t) \left\{ p_{X_{t-1}|X_t}(x_{t-1} \mid x_t) - p_{Y_{t-1}^*|Y_t}(x_{t-1} \mid x_t) \right\} dx_{t-1} dx_t \right| \\ &= \left| 1 - \mathbb{P}((X_t, X_{t-1}) \notin \mathcal{E}) - \int_{(x_t, x_{t-1}) \notin \mathcal{E}} p_{X_t}(x_t) \left\{ 1 - p_{Y_{t-1}^*|Y_t}(x_{t-1} \mid x_t) \right\} dx_{t-1} dx_t \right| \\ &\leq 2 \exp\left(-\frac{c_3}{2}d\log T\right). \end{aligned} \quad (91)$$

Substituting (87) and (91) into (85) yields: for each $t \geq 2$,

$$\mathbb{E}_{x_t \sim q_t} \left[\text{KL}\left(p_{X_{t-1}|X_t}(\cdot \mid x_t) \parallel p_{Y_{t-1}^*|Y_t}(\cdot \mid x_t)\right) \right] \lesssim d^4 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^2 \log^4 T + 3 \exp\left(-\frac{c_3}{2}d\log T\right) \lesssim \frac{d^4 \log^6 T}{T^2}, \quad (92)$$

where the last inequality utilizes the properties (39) of the learning rates.

Step 4: quantifying the effect of score estimation errors. Thus far, we have quantified the KL divergence between $p_{X_{t-1}|X_t}$ and $p_{Y_{t-1}^*|Y_t}$. Given that Y_{t-1}^* is obtained from Y_t using the true score function, we still need to control the effect of the score estimation error, which is accomplished by means of the following lemma. The proof of this lemma can be found in Appendix C.3.

Lemma 10. *For any $1 < t \leq T$, one has*

$$\begin{aligned} &\mathbb{E}_{x_t \sim q_t} \left[\text{KL}\left(p_{X_{t-1}|X_t}(\cdot \mid x_t) \parallel p_{Y_{t-1}|Y_t}(\cdot \mid x_t)\right) \right] - \mathbb{E}_{x_t \sim q_t} \left[\text{KL}\left(p_{X_{t-1}|X_t}(\cdot \mid x_t) \parallel p_{Y_{t-1}^*|Y_t}(\cdot \mid x_t)\right) \right] \\ &\lesssim \exp\left(-c_{20}d\log T\right) + \frac{d\log^3 T}{T} \mathbb{E}_{X_t \sim q_t} \left[\varepsilon_{\text{score},t}(X_t)^2 \right] \end{aligned} \quad (93)$$

for some universal constant $c_{20} > 0$.

Step 5: putting all this together. To finish up, substitute (92), (179) and (179) into the decomposition (79) to obtain

$$\text{KL}(p_{X_1} \parallel p_{Y_1}) \lesssim \text{KL}(p_{X_T} \parallel p_{Y_T}) + \sum_{t=1}^{T-1} \mathbb{E}_{x_t \sim q_t} \left[\text{KL}\left(p_{X_{t-1}|X_t}(\cdot \mid x_t) \parallel p_{Y_{t-1}^*|Y_t}(\cdot \mid x_t)\right) \right]$$

$$\begin{aligned}
& + \sum_{t=1}^{T-1} \left\{ \mathbb{E}_{x_t \sim q_t} \left[\text{KL} \left(p_{X_{t-1} | X_t}(\cdot | x_t) \parallel p_{Y_{t-1} | Y_t}(\cdot | x_t) \right) \right] - \mathbb{E}_{x_t \sim q_t} \left[\text{KL} \left(p_{X_{t-1} | X_t}(\cdot | x_t) \parallel p_{Y_{t-1}^* | Y_t}(\cdot | x_t) \right) \right] \right\} \\
& \lesssim \text{KL}(p_{X_T} \parallel p_{Y_T}) + \sum_{2 \leq t \leq T} \frac{d^4 \log^6 T}{T^2} + (d \log^3 T) \sum_{t=2}^T \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score}, t}(X_t)^2] \\
& \asymp \frac{d^4 \log^6 T}{T} + d \varepsilon_{\text{score}}^2 \log^3 T,
\end{aligned}$$

where the last relation applies the bound on $\text{KL}(p_{X_T} \parallel p_{Y_T})$ as in (122). 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 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 processes, 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? Can we establish sharp convergence results in terms of the Wasserstein distance, which could sometimes be “closer” to how humans differentiate pictures and might potentially help relax Assumption 2 in the case of deterministic samplers? 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 acceleration 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 (38)

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

$$\begin{aligned}
[J_t(x)]_{i,j} &= \mathbf{1}\{i = j\} + \frac{1}{1 - \bar{\alpha}_t} \left\{ \left(\int_{x_0} p_{X_0 | X_t}(x_0 | x) (x_i - \sqrt{\bar{\alpha}_t} x_{0,i}) dx_0 \right) \left(\int_{x_0} p_{X_0 | X_t}(x_0 | x) (x_j - \sqrt{\bar{\alpha}_t} x_{0,j}) dx_0 \right) \right. \\
&\quad \left. - \int_{x_0} p_{X_0 | X_t}(x_0 | x) (x_i - \sqrt{\bar{\alpha}_t} x_{0,i}) (x_j - \sqrt{\bar{\alpha}_t} x_{0,j}) dx_0 \right\}. \tag{94}
\end{aligned}$$

This immediately establishes the matrix expression (38).

A.2 Proof of properties (39) regarding the learning rates

Proof of property (39a). From the choice of β_t in (22), we have

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

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 (39b) and (39c). We start by proving (39b). Let τ be an integer obeying

$$\beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^\tau \leq 1 < \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{\tau+1}, \quad (95)$$

and we divide into two cases based on τ .

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

$$1 - \bar{\alpha}_{t-1} \geq \frac{1}{3} \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^t. \quad (96)$$

Clearly, if (96) is valid, then any $t \leq \tau$ obeys

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

as claimed. Towards proving (96), first note that the base case with $t = 2$ holds true trivially since $1 - \bar{\alpha}_1 = 1 - \alpha_1 = \beta_1 \geq \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^2 / 3$. Next, let $t_0 > 2$ be the first time that Condition (96) fails to hold and suppose that $t_0 \leq \tau$. It then follows that

$$1 - \bar{\alpha}_{t_0-2} = 1 - \frac{\bar{\alpha}_{t_0-1}}{\alpha_{t_0-1}} \leq 1 - \bar{\alpha}_{t_0-1} < \frac{1}{3} \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{t_0} \leq \frac{1}{2} \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{t_0-1} < \frac{1}{2}, \quad (97)$$

where the last inequality result from (95) and the assumption $t_0 \leq \tau$. This taken together with the assumptions (96) and $t_0 \leq \tau$ implies that

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

As a result, we can further derive

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

where the penultimate line holds since (96) 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 (96) 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

$$1 - \bar{\alpha}_{t-1} \geq 1 - \bar{\alpha}_{\tau-1} \geq \frac{1}{3} \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^\tau = \frac{\frac{1}{3} \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{\tau+1}}{1 + \frac{c_1 \log T}{T}} \geq \frac{1}{4}, \quad (98)$$

where the second and the third inequalities come from (96). Therefore, one obtains

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

The above arguments taken together establish property (39b).

In addition, it comes immediately from (39b) that

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

thereby justifying property (39c).

Proof of property (39d). Turning attention to the second claim (39d), we note that for any t obeying $t \geq \frac{T}{2} \gtrsim \frac{T}{\log T}$, one has

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

This in turn allows one to deduce that

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

for an arbitrarily large constant $c_2 > 0$.

Proof of property (39e). Finally, it is easily seen from the Taylor expansion that the learning rates $\{\alpha_t\}$ satisfy

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

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

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 $c_5 \geq 2$ and suppose that $c_6 \geq 2c_R$. There exists some $x_0 \in \mathbb{R}^d$ such that

$$\|\sqrt{\bar{\alpha}_t}x_0 - y\|_2 \leq c_5 \sqrt{\theta(y)d(1 - \bar{\alpha}_t) \log T} \quad \text{and} \quad (99a)$$

$$\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon) \geq \left(\frac{\epsilon}{T^{2\theta(y)}} \right)^d \quad \text{with} \quad \epsilon = \frac{1}{T^{c_0/2}} \quad (99b)$$

hold simultaneously, where c_0 is defined in (22).

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

$$p_X(x) = p_{X_0|X_t}(x|y). \quad (100)$$

Consider the point x_0 in Claim 1, and let us look at a set:

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

where $c_5 \geq 2$ (see Claim 1). Combining this with property (99a) about x_0 results in

$$\mathbb{P}\left(\|\sqrt{\bar{\alpha}_t}X - y\|_2 > 5c_5\sqrt{\theta(y)d(1 - \bar{\alpha}_t)\log T}\right) \leq \mathbb{P}(X \in \mathcal{E}). \quad (101)$$

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

$$\begin{aligned} \mathbb{P}(X_0 \in \mathcal{E} | X_t = y) &= \frac{\int_{x \in \mathcal{E}} p_{X_0}(x)p_{X_t | X_0}(y | x)dx}{\int_x p_{X_0}(x)p_{X_t | X_0}(y | x)dx} \\ &\leq \frac{\int_{x \in \mathcal{E}} p_{X_0}(x)p_{X_t | X_0}(y | x)dx}{\int_{x: \|x - x_0\|_2 \leq \epsilon} p_{X_0}(x)p_{X_t | X_0}(y | x)dx} \\ &\leq \frac{\sup_{x \in \mathcal{E}} p_{X_t | X_0}(y | x)}{\inf_{x: \|x - x_0\|_2 \leq \epsilon} p_{X_t | X_0}(y | x)} \cdot \frac{\mathbb{P}(X_0 \in \mathcal{E})}{\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon)}. \end{aligned} \quad (102)$$

To further bound this quantity, note that: in view of the definition of \mathcal{E} and expression (99a), one has

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

and

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

where the second line is due to the elementary inequality $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, the penultimate line relies on (99), and the last line holds true since $1 - \bar{\alpha}_t \geq 1 - \alpha_1 = 1/T^{c_0}$ (see (22)). Substitution of the above two displays into (102), we arrive at

$$\begin{aligned} \mathbb{P}(X_0 \in \mathcal{E} | X_t = y) &\leq \exp(-2.5c_5^2\theta(y)d\log T) \cdot \frac{1}{\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon)} \\ &\leq \exp(-2.5c_5^2\theta d\log T) \cdot \left(T^{2\theta(y) + c_0/2}\right)^d \\ &\leq \exp(-(2.5c_5^2\theta(y) - 2\theta(y) - c_0/2)d\log T), \end{aligned} \quad (103)$$

where the second inequality invokes (99b). Substituting this into (101) and recalling the distribution (100) of X , we arrive at

$$\mathbb{P}\left(\|\sqrt{\bar{\alpha}_t}X - y\|_2 > 5c_5\sqrt{\theta(y)d(1 - \bar{\alpha}_t)\log T}\right) \leq \exp(-(2.5c_5^2\theta(y) - 2\theta(y) - c_0/2)d\log T)$$

$$\leq \exp(-c_5^2 \theta(y) d \log T),$$

with the proviso that $c_5 \geq 2$ and $c_6 \geq c_0$ (so that $\theta(y) \geq c_6 \geq c_0$). This concludes the proof of the advertised result (42) when $c_5 \geq 2$ and $c_6 \geq 2c_R + c_0$, as long as Claim 1 can be justified.

With the above result in place, it then follows that

$$\begin{aligned} & \mathbb{E} [\|x_t - \sqrt{\bar{\alpha}_t} X_0\|_2 \mid X_t = x_t] \\ & \leq 5c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T} + \mathbb{E} \left[\|x_t - \sqrt{\bar{\alpha}_t} X_0\|_2 \mathbf{1} \left\{ \|x_t - \sqrt{\bar{\alpha}_t} X_0\|_2 \geq 5c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T} \right\} \mid X_t = x_t \right] \\ & \leq 5c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T} + \int_{5c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}}^{\infty} \mathbb{P}(\|x_t - \sqrt{\bar{\alpha}_t} x_0\|_2 \geq \tau \mid X_t = x_t) d\tau \\ & \leq 5c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T} + \int_{5c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}}^{\infty} \exp\left(-\frac{\tau^2}{25(1 - \bar{\alpha}_t)}\right) d\tau \\ & \leq 5c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T} + \exp(-c_5^2 \theta(y) d \log T) \\ & \leq 6c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}, \end{aligned}$$

as claimed in (43a) by taking $c_5 = 2$. The proofs for (43b), (43c) and (43d) follow from similar arguments 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 $\|\sqrt{\bar{\alpha}_t} x - y\|_2 \leq c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}$, we have

$$\mathbb{P}(\|X_0 - x\|_2 \leq \epsilon) \leq \left(\frac{\epsilon}{2T^{\theta(y)R}}\right)^d \quad \text{with } \epsilon = \frac{1}{T^{c_0/2}}. \quad (104)$$

Clearly, the choice of ϵ ensures that $\epsilon < \frac{1}{2} \sqrt{d(1 - \bar{\alpha}_t) \log T}$. 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

$$\begin{aligned} p_{X_t}(y) &= \int_x p_{X_0}(x) p_{X_t \mid X_0}(y \mid x) dx \\ &= \int_{x: \|\sqrt{\bar{\alpha}_t} x - y\|_2 \geq c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}} p_{X_0}(x) p_{X_t \mid X_0}(y \mid x) dx \\ &\quad + \int_{x: \|\sqrt{\bar{\alpha}_t} x - y\|_2 < c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}} p_{X_0}(x) p_{X_t \mid X_0}(y \mid x) dx. \end{aligned} \quad (105)$$

To further control (105), we make two observations:

- 1) The first term on the right-hand side of (105) can be bounded by

$$\begin{aligned} & \int_{x: \|\sqrt{\bar{\alpha}_t} x - y\|_2 \geq c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}} p_{X_0}(x) p_{X_t \mid X_0}(y \mid x) dx \\ & \leq \sup_{z: \|z\|_2 \geq c_5 \sqrt{\theta(y) d (1 - \bar{\alpha}_t) \log T}} \frac{1}{(2\pi(1 - \bar{\alpha}_t))^{d/2}} \exp\left(-\frac{\|z\|_2^2}{2(1 - \bar{\alpha}_t)}\right) \\ & < \frac{1}{2} \exp(-\theta(y) d \log T), \end{aligned} \quad (106)$$

provided that $c_5 \geq 2$ and $c_6 > 0$ is large enough (note that $\theta(y) \geq c_6$). Here, we have used $X_t \stackrel{(i)}{\sim} \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} W$ 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 (105), let us construct an epsilon-net $\mathcal{N}_\epsilon = \{z_i\}$ for the following set

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

so that for each x in this set, one can find a vector $z_i \in \mathcal{N}_\epsilon$ such that $\|x - z_i\|_2 \leq \epsilon$. Clearly, we can choose \mathcal{N}_ϵ so that its cardinality obeys $|\mathcal{N}_\epsilon| \leq (2R/\epsilon)^d$. Define $\mathcal{B}_i := \{x \mid \|x - z_i\|_2 \leq \epsilon\}$ for each $z_i \in \mathcal{N}_\epsilon$. Armed with these sets, we can derive

$$\begin{aligned} \int_{x: \|\sqrt{\bar{\alpha}_t}x - y\|_2 < c_5 \sqrt{\theta(y)d(1-\bar{\alpha}_t)\log T}} p_{X_0}(x)p_{X_t|X_0}(y|x)dx &\leq (2\pi(1-\bar{\alpha}_t))^{-d/2} \sum_{i=1}^{|\mathcal{N}_\epsilon|} \mathbb{P}(X_0 \in \mathcal{B}_i) \\ &\leq (2\pi(1-\bar{\alpha}_t))^{-d/2} \left(\frac{\epsilon}{2T^{2\theta(y)}R}\right)^d \left(\frac{2R}{\epsilon}\right)^d \\ &< \frac{1}{2} \exp(-\theta(y)d\log T), \end{aligned}$$

where the penultimate step comes from the assumption (104).

The above results taken collectively lead to

$$p_{X_t}(y) < \exp(-\theta(y)d\log T), \quad (107)$$

thus contradicting the definition (41) of $\theta(y)$.

Consequently, we have proven the existence of x obeying $\|\sqrt{\bar{\alpha}_t}x - y\|_2 \leq c_5 \sqrt{\theta(y)d(1-\bar{\alpha}_t)\log T}$ and

$$\mathbb{P}(\|X_0 - x\|_2 \leq \epsilon) > \left(\frac{\epsilon}{2T^{2\theta(y)}R}\right)^d \geq \left(\frac{\epsilon}{T^{2\theta(y)}}\right)^d,$$

provided that $\theta(y) \geq c_6 \geq 2c_R$. This completes the proof of Claim 1.

A.4 Proof of Lemma 2

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

$$\begin{aligned} p_{X_{t-1}}(x) &= \int_{x_0} p_{X_0}(x_0)p_{X_{t-1}|X_0}(x|x_0)dx_0 \\ &= \int_{x_0} p_{X_0}(x_0)p_{X_t|X_0}(x_t|x_0) \cdot \frac{p_{X_{t-1}|X_0}(x|x_0)}{p_{X_t|X_0}(x_t|x_0)} dx_0 \\ &= p_{X_t}(x_t) \int_{x_0} p_{X_0|X_t}(x_0|x_t) \cdot \frac{p_{X_{t-1}|X_0}(x|x_0)}{p_{X_t|X_0}(x_t|x_0)} dx_0, \end{aligned} \quad (108)$$

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 $X_t \stackrel{d}{=} \sqrt{\bar{\alpha}_t}X_0 + \sqrt{1-\bar{\alpha}_t}W$ 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

$$\|\hat{x}_t - x\|_2 \leq c_3 \sqrt{d(1-\alpha_t)\log T} \quad (109)$$

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

$$\begin{aligned} \frac{p_{X_{t-1}|X_0}(x|x_0)}{p_{X_t|X_0}(x_t|x_0)} &= \left(\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}\right)^{d/2} \exp\left(\frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(1-\bar{\alpha}_t)} - \frac{\|x - \sqrt{\bar{\alpha}_{t-1}}x_0\|_2^2}{2(1-\bar{\alpha}_{t-1})}\right) \\ &\leq \exp\left(\frac{d(1-\alpha_t) + \|\hat{x}_t - x\|_2^2 + 2\|\hat{x}_t - x\|_2\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2}{2(1-\bar{\alpha}_{t-1})} + \frac{(1-\alpha_t)\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{(1-\bar{\alpha}_{t-1})^2}\right), \end{aligned} \quad (110)$$

where the last inequality follows from the two relations below:

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

and

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

Next, let us define the following set given x_t :

$$\tilde{\mathcal{E}} := \left\{ x_0 : \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 \leq c_4 \sqrt{d(1 - \bar{\alpha}_t) \log T} \right\} \quad (111)$$

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

- For any $x_0 \in \tilde{\mathcal{E}}$, one can utilize (111) and (109) to deduce that

$$\begin{aligned}
& \frac{d(1 - \alpha_t) + \|\hat{x}_t - x\|_2^2 + 2\|\hat{x}_t - x\|_2 \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2}{2(1 - \bar{\alpha}_{t-1})} + \frac{(1 - \alpha_t)\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{(1 - \bar{\alpha}_{t-1})^2} \\
&\lesssim \frac{d(1 - \alpha_t)}{1 - \bar{\alpha}_{t-1}} + \frac{d(1 - \alpha_t) \log T + d\sqrt{(1 - \bar{\alpha}_t)(1 - \alpha_t)} \log T}{1 - \bar{\alpha}_{t-1}} + \frac{(1 - \alpha_t)^2 d(1 - \bar{\alpha}_t) \log T}{(1 - \bar{\alpha}_{t-1})^2} \\
&\lesssim (d \log T) \left\{ \frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}} + \sqrt{\frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}}} \sqrt{\frac{1 - \bar{\alpha}_t}{1 - \bar{\alpha}_{t-1}}} + \left(\frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}} \right)^2 \right\} \\
&\lesssim d \sqrt{\frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}}} \log T,
\end{aligned} \quad (112)$$

where the last inequality makes use of the facts $\frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}} \leq 1$ (cf. (39b)) and

$$\frac{1 - \bar{\alpha}_t}{1 - \bar{\alpha}_{t-1}} = \frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}} + \frac{\alpha_t - \bar{\alpha}_t}{1 - \bar{\alpha}_{t-1}} = \frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}} + \alpha_t \leq 2. \quad (113)$$

Moreover, the properties (39) of the stepsizes tell us that

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

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 (110) and (112) reveals that

$$\frac{p_{X_{t-1}|X_0}(x|x_0)}{p_{X_t|X_0}(x_t|x_0)} = 1 + O\left(d \sqrt{\frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}}} \log T\right), \quad (114)$$

with the proviso that (109) holds and $x_0 \in \tilde{\mathcal{E}}$.

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

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

$$\begin{aligned}
&\stackrel{(i)}{\leq} \frac{d(1-\alpha_t) + \left(1 + \frac{1-\bar{\alpha}_{t-1}}{1-\alpha_t}\right) \|\hat{x}_t - x\|_2^2 + \frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}} \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(1-\bar{\alpha}_{t-1})} + \frac{(1-\alpha_t) \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{(1-\bar{\alpha}_{t-1})^2} \\
&\stackrel{(ii)}{\leq} \frac{d(1-\alpha_t) + c_3\left(1 + \frac{1-\bar{\alpha}_{t-1}}{1-\alpha_t}\right)d(1-\alpha_t) \log T}{2(1-\bar{\alpha}_{t-1})} + \frac{2(1-\alpha_t) \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{(1-\bar{\alpha}_{t-1})^2} \\
&\stackrel{(iii)}{\leq} \frac{c_3d(1-\alpha_t + 1 - \bar{\alpha}_{t-1}) \log T}{1-\bar{\alpha}_{t-1}} \cdot \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{c_4d(1-\bar{\alpha}_{t-1}) \log T} + \frac{2(1-\alpha_t) \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{(1-\bar{\alpha}_{t-1})^2} \\
&\leq \frac{2c_3}{c_4} \cdot \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{1-\bar{\alpha}_{t-1}} + \frac{8c_1 \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2 \log T}{T(1-\bar{\alpha}_{t-1})} \\
&\stackrel{(iv)}{\leq} \left(\frac{2c_3}{c_4} + \frac{8c_1 \log T}{T}\right) \frac{2 \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{1-\bar{\alpha}_t} \leq \frac{8c_3}{c_4} \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{1-\bar{\alpha}_t},
\end{aligned}$$

where (i) comes from the Cauchy-Schwarz inequality, (ii) is valid due to the assumption on $\|\hat{x}_t - x\|_2$, (iii) follows from the definition of $\tilde{\mathcal{E}}$, and (iv) is a consequence of (113). Substitution into (110) leads to

$$\frac{p_{X_{t-1}|X_0}(x|x_0)}{p_{X_t|X_0}(x_t|x_0)} \leq \exp\left(\frac{8c_3}{c_4} \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{1-\bar{\alpha}_t}\right), \quad (115)$$

provided that (109) holds and $x_0 \notin \tilde{\mathcal{E}}$.

In light of the above calculations, we can invoke (116) to demonstrate that: for any x obeying (109),

$$\begin{aligned}
p_{X_{t-1}}(x) &= p_{X_t}(x_t) \left(\int_{x_0 \in \tilde{\mathcal{E}}} + \int_{x_0 \notin \tilde{\mathcal{E}}} \right) p_{X_0|X_t}(x_0|x_t) \cdot \frac{p_{X_{t-1}|X_0}(x|x_0)}{p_{X_t|X_0}(x_t|x_0)} dx_0 \\
&= p_{X_t}(x_t) \int_{x_0 \in \tilde{\mathcal{E}}} \left(1 + O\left(d\sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}} \log T\right) \right) p_{X_0|X_t}(x_0|x_t) dx_0 \\
&\quad + p_{X_t}(x_t) \int_{x_0 \notin \tilde{\mathcal{E}}} O\left(\exp\left(\frac{8c_3}{c_4} \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{1-\bar{\alpha}_t}\right)\right) p_{X_0|X_t}(x_0|x_t) dx_0. \quad (116)
\end{aligned}$$

By virtue of Lemma 1, if $-\log p_{X_t}(x_t) \leq \frac{1}{2}c_6d \log T$ for some large constant $c_6 > 0$, then it holds that

$$\mathbb{P}\left\{\|\sqrt{\bar{\alpha}_t}X_0 - x_t\|_2 \geq 5c_5\sqrt{d(1-\bar{\alpha}_t) \log T} \mid X_t = x_t\right\} \leq \exp(-c_5^2c_6d \log T) \quad (117)$$

for any $c_5 \geq 2$. Some elementary calculation then reveals that

$$\int_{x_0 \notin \tilde{\mathcal{E}}} \exp\left(\frac{8c_3}{c_4} \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{1-\bar{\alpha}_t}\right) p_{X_0|X_t}(x_0|x_t) dx_0 \lesssim \frac{1}{T^{c_0}} \quad (118)$$

with c_0 defined in (22), provided that c_4/c_3 is sufficiently large. Hence, substituting it into (116) demonstrates that, for any x satisfying (109),

$$p_{X_{t-1}}(x) = \left(1 + O\left(d\sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}} \log T\right)\right) (1 - o(1))p_{X_t}(x_t) + O\left(\frac{1}{T^{c_0}}\right)p_{X_t}(x_t) \quad (119)$$

$$= \left(1 + O\left(d\sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}} \log T + \frac{1}{T^{c_0}}\right)\right) p_{X_t}(x_t)$$

$$= \left(1 + O\left(d\sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}} \log T\right)\right) p_{X_t}(x_t) \quad (120)$$

$$\in \left[\frac{1}{2}p_{X_t}(x_t), \frac{3}{2}p_{X_t}(x_t)\right], \quad (121)$$

where the penultimate inequality holds since (according to (22))

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

and the last inequality holds due to (39b) and the condition that $T/(d^2 \log^3 T)$ is sufficiently large. In other words, (121) 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 (44), we have

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

Therefore, taking x to be $x_t(\gamma)$ in (121) tells us that: if $-\log p_{X_t}(x_t) \leq \frac{1}{2}c_6 d \log T$, then

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

as claimed.

A.5 Proof of Lemma 3

Recognizing that $Y_T \sim \mathcal{N}(0, I_d)$ and that $X_T \stackrel{d}{=} \sqrt{\bar{\alpha}_T} X_0 + \sqrt{1 - \bar{\alpha}_T} \bar{W}_t$ with $\bar{W}_t \sim \mathcal{N}(0, I_d)$ (independent from X_0), one has

$$\begin{aligned} \text{KL}(p_{X_T} \parallel p_{Y_T}) &= \int p_{X_T}(x) \log \frac{p_{X_T}(x)}{p_{Y_T}(x)} dx \\ &\stackrel{(i)}{=} \int p_{X_T}(x) \log \frac{\int_{y: \|y\|_2 \leq \sqrt{\bar{\alpha}_T} T^{c_R}} p_{\sqrt{\bar{\alpha}_T} X_0}(y) p_{\sqrt{1 - \bar{\alpha}_T} \bar{W}_t}(x - y) dy}{p_{Y_T}(x)} dx \\ &\leq \int p_{X_T}(x) \log \frac{\sup_{y: \|y\|_2 \leq \sqrt{\bar{\alpha}_T} T^{c_R}} p_{\sqrt{1 - \bar{\alpha}_T} \bar{W}_t}(x - y)}{p_{Y_T}(x)} dx \\ &= \int p_{X_T}(x) \left(-d/2 \log(1 - \bar{\alpha}_T) + \sup_{y: \|y\|_2 \leq \sqrt{\bar{\alpha}_T} T^{c_R}} \left(-\frac{\|x - y\|_2^2}{2(1 - \bar{\alpha}_T)} + \frac{\|x\|_2^2}{2} \right) \right) dx \\ &\stackrel{(ii)}{\leq} \int p_{X_T}(x) \left(-d/2 \log(1 - \bar{\alpha}_T) + \|x\|_2 \sup_{y: \|y\|_2 \leq \sqrt{\bar{\alpha}_T} T^{c_R}} \frac{\|y\|_2}{1 - \bar{\alpha}_T} \right) dx \\ &\leq -d/2 \log(1 - \bar{\alpha}_T) + \frac{\sqrt{\bar{\alpha}_T} T^{c_R}}{2(1 - \bar{\alpha}_T)} \mathbb{E}[\|X_T\|_2] \\ &\stackrel{(iii)}{\lesssim} \bar{\alpha}_T d + \frac{\sqrt{\bar{\alpha}_T} T^{c_R}}{2(1 - \bar{\alpha}_T)} \left(\sqrt{\bar{\alpha}_T} T^{c_R} + \sqrt{d} \right) \stackrel{(iv)}{\lesssim} \frac{1}{T^{200}}, \end{aligned} \tag{122}$$

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

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

and (iv) makes use of (39d) given that $c_2 \geq 1000$. The proof is thus completed by invoking the Pinsker inequality (Tsybakov, 2009, Lemma 2.5).

B Proof of auxiliary lemmas for the ODE-based sampler

B.1 Proof of Lemma 4

B.1.1 Proof of relations (57) and (58a)

Recall the definition of ϕ_t and ϕ_t^* in (51), and introduce the following vector:

$$u := x - \phi_t(x) = x - \phi_t^*(x) + \phi_t^*(x) - \phi_t(x)$$

$$= \frac{1 - \alpha_t}{2(1 - \bar{\alpha}_t)} \int_{x_0} (x - \sqrt{\bar{\alpha}_t}x_0) p_{X_0 | X_t}(x_0 | x) dx_0 - \frac{1 - \alpha_t}{2} (s_t(x) - s_t^*(x)). \quad (123)$$

The proof is composed of the following steps.

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

$$X_t \stackrel{d}{=} \sqrt{\bar{\alpha}_t}X_0 + \sqrt{1 - \bar{\alpha}_t}W \quad \text{with } W \sim \mathcal{N}(0, I_d) \quad (124)$$

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

$$p_{X_0 | X_t}(x_0 | x) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} p_{X_t | X_0}(x | x_0) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} \cdot \frac{1}{(2\pi(1 - \bar{\alpha}_t))^{d/2}} \exp\left(-\frac{\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(1 - \bar{\alpha}_t)}\right). \quad (125)$$

Moreover, it follows from (124) that

$$\sqrt{\alpha_t}X_{t-1} \stackrel{d}{=} \sqrt{\alpha_t}(\sqrt{\bar{\alpha}_{t-1}}X_0 + \sqrt{1 - \bar{\alpha}_{t-1}}W) = \sqrt{\bar{\alpha}_t}X_0 + \sqrt{\alpha_t - \bar{\alpha}_t}W. \quad (126)$$

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

$$\begin{aligned} \frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} &\stackrel{(i)}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{(2\pi(\alpha_t - \bar{\alpha}_t))^{d/2}} \exp\left(-\frac{\|\phi_t(x) - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)}\right) dx_0 \\ &\stackrel{(ii)}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{(2\pi(\alpha_t - \bar{\alpha}_t))^{d/2}} \exp\left(-\frac{\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(1 - \bar{\alpha}_t)}\right) \\ &\quad \cdot \exp\left(-\frac{(1 - \alpha_t)\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{2(\alpha_t - \bar{\alpha}_t)}\right) dx_0 \\ &\stackrel{(iii)}{=} \left(\frac{1 - \bar{\alpha}_t}{\alpha_t - \bar{\alpha}_t}\right)^{d/2} \cdot \int_{x_0} p_{X_0 | X_t}(x_0 | x) \cdot \\ &\quad \exp\left(-\frac{(1 - \alpha_t)\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{2(\alpha_t - \bar{\alpha}_t)}\right) dx_0 \\ &\stackrel{(iv)}{=} \left\{1 + \frac{d(1 - \alpha_t)}{2(\alpha_t - \bar{\alpha}_t)} + O\left(d^2\left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^2\right)\right\} \cdot \\ &\quad \int_{x_0} p_{X_0 | X_t}(x_0 | x) \exp\left(-\frac{(1 - \alpha_t)\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{2(\alpha_t - \bar{\alpha}_t)}\right) dx_0. \end{aligned} \quad (127)$$

Here, identity (i) holds due to (126) and hence

$$p_{\sqrt{\alpha_t}X_{t-1}}(x) = \int_{x_0} p_{X_0}(x_0) p_{\sqrt{\alpha_t - \bar{\alpha}_t}W}(x - \sqrt{\bar{\alpha}_t}x_0) dx_0;$$

identity (ii) follows from (123) and elementary algebra; relation (iii) is a consequence of the Bayes rule (125); and relation (iv) results from (39e).

Step 2: controlling the integral in the decomposition (127). In order to further control the right-hand side of expression (127), we need to evaluate the integral in (127). To this end, we make a few observations.

- To begin with, Lemma 1 tells us that

$$\mathbb{P}\left(\|\sqrt{\bar{\alpha}_t}X_0 - x\|_2 > 5c_5\sqrt{\theta_t(x)d(1 - \bar{\alpha}_t)\log T} \mid X_t = x\right) \leq \exp(-c_5^2\theta_t(x)d\log T) \quad (128a)$$

for any quantity $c_5 \geq 2$, provided that $c_6 \geq 2c_R + c_0$.

- A little algebra based on this relation allows one to bound u (cf. (123)) as follows:

$$\begin{aligned} \|u\|_2 &\leq \frac{1-\alpha_t}{2}\varepsilon_{\text{score},t}(x) + \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)}\mathbb{E}[\|\sqrt{\bar{\alpha}_t}X_0 - x\|_2 \mid X_t = x] \\ &\leq \frac{1-\alpha_t}{2}\varepsilon_{\text{score},t}(x) + \frac{6(1-\alpha_t)}{1-\bar{\alpha}_t}\sqrt{\theta_t(x)d(1-\bar{\alpha}_t)\log T}, \end{aligned} \quad (128b)$$

where the last inequality arises from Lemma 1.

Next, let us define

$$\mathcal{E}_c^{\text{typical}} := \left\{ x_0 : \|x - \sqrt{\bar{\alpha}_t}x_0\|_2 \leq 5c\sqrt{\theta_t(x)d(1-\bar{\alpha}_t)\log T} \right\} \quad (129)$$

for any quantity $c > 0$. Then for any $x_0 \in \mathcal{E}_c^{\text{typical}}$, it is clearly seen from (128) and (39) that

$$\frac{(1-\alpha_t)\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1-\bar{\alpha}_t)} \leq \frac{25c^2(1-\alpha_t)\theta_t(x)d\log T}{2(\alpha_t - \bar{\alpha}_t)} \leq \frac{100c_1c^2\theta_t(x)d\log^2 T}{T}; \quad (130a)$$

$$\begin{aligned} \frac{\|u\|_2^2}{2(\alpha_t - \bar{\alpha}_t)} &\leq \frac{(1-\alpha_t)^2}{4(\alpha_t - \bar{\alpha}_t)}\varepsilon_{\text{score},t}(x)^2 + \frac{36(1-\alpha_t)^2}{(1-\bar{\alpha}_t)(\alpha_t - \bar{\alpha}_t)}\theta_t(x)d\log T \\ &\leq \frac{2c_1^2\log^2 T}{T^2}\varepsilon_{\text{score},t}(x)^2 + \frac{2304c_1^2}{T^2}\theta_t(x)d\log^3 T, \end{aligned} \quad (130b)$$

$$\begin{aligned} \left| \frac{u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{\alpha_t - \bar{\alpha}_t} \right| &\leq \frac{\|u\|_2\|x - \sqrt{\bar{\alpha}_t}x_0\|_2}{\alpha_t - \bar{\alpha}_t} \\ &\leq \frac{5c(1-\alpha_t)}{2(\alpha_t - \bar{\alpha}_t)}\varepsilon_{\text{score},t}(x)\sqrt{\theta_t(x)d(1-\bar{\alpha}_t)\log T} + \frac{30c(1-\alpha_t)\theta_t(x)d\log T}{\alpha_t - \bar{\alpha}_t} \end{aligned} \quad (130c)$$

$$\leq \frac{20cc_1}{T}\varepsilon_{\text{score},t}(x)\sqrt{\theta_t(x)d(1-\bar{\alpha}_t)\log^3 T} + \frac{240cc_1\theta_t(x)d\log^2 T}{T}. \quad (130d)$$

As a consequence, for any $x_0 \in \mathcal{E}_c^{\text{typical}}$ for $c \geq 2$, we have seen from (130d) and (39) that

$$\begin{aligned} -\frac{(1-\alpha_t)\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1-\bar{\alpha}_t)} - \frac{\|u\|_2^2}{2(\alpha_t - \bar{\alpha}_t)} + \frac{u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{\alpha_t - \bar{\alpha}_t} &\leq \frac{u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{\alpha_t - \bar{\alpha}_t} \\ &\leq \frac{5c(1-\alpha_t)}{2(\alpha_t - \bar{\alpha}_t)}\varepsilon_{\text{score},t}(x)\sqrt{\theta_t(x)d(1-\bar{\alpha}_t)\log T} + \frac{30c(1-\alpha_t)\theta_t(x)d\log T}{\alpha_t - \bar{\alpha}_t} \end{aligned} \quad (131)$$

$$\leq \frac{20cc_1}{T}\varepsilon_{\text{score},t}(x)\sqrt{\theta_t(x)d\log^3 T} + \frac{240cc_1}{T}\theta_t(x)d\log^2 T \leq c\theta_t(x)d, \quad (132)$$

provided that

$$\frac{40c_1\varepsilon_{\text{score},t}(x)\log^{\frac{3}{2}}T}{T} \leq \sqrt{\theta_t(x)d} \quad \text{and} \quad T \geq 480c_1\log^2 T.$$

Step 2(a): proof of relation (57). Substituting (131) into (127) and making use of (39) under our assumption on T yield

$$\begin{aligned} \frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} &\leq 2 \exp\left(\frac{5c(1-\alpha_t)}{2(\alpha_t - \bar{\alpha}_t)}\varepsilon_{\text{score},t}(x)\sqrt{\theta_t(x)d\log T} + \frac{30c(1-\alpha_t)}{\alpha_t - \bar{\alpha}_t}\theta_t(x)d\log T\right) \int_{x_0} p_{X_0|X_t}(x_0|x)dx_0 \\ &\leq 2 \exp\left(\frac{5c(1-\alpha_t)}{2(\alpha_t - \bar{\alpha}_t)}\varepsilon_{\text{score},t}(x)\sqrt{\theta_t(x)d\log T} + \frac{30c(1-\alpha_t)}{\alpha_t - \bar{\alpha}_t}\theta_t(x)d\log T\right), \end{aligned}$$

thus establishing (57) by taking $c = 2$.

Step 2(b): proof of relation (58a). Suppose now that

$$C_{10} \frac{\theta_t(x) d \log^2 T + \varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x) d \log^3 T}}{T} \leq 1 \quad (133)$$

holds for some large enough constant $C_{10} > 0$. Under this additional condition, it can be easily verified that

$$\begin{aligned} & \left| -\frac{(1-\alpha_t) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2}{2(\alpha_t - \bar{\alpha}_t)} + \frac{u^\top (x - \sqrt{\bar{\alpha}_t} x_0)}{\alpha_t - \bar{\alpha}_t} \right| \\ & \leq c_{10} \left(\theta_t(x) d \log T + \varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x) d \log T} \right) \frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \end{aligned} \quad (134)$$

for any $x_0 \in \mathcal{E}_2^{\text{typical}}$ (with $c = 2$), where $c_{10} > 0$ is some sufficiently small constant. Therefore, the Taylor expansion $e^{-z} = 1 - z + O(z^2)$ (for all $|z| < 1$) gives

$$\begin{aligned} & \exp \left(-\frac{(1-\alpha_t) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top (x - \sqrt{\bar{\alpha}_t} x_0)}{2(\alpha_t - \bar{\alpha}_t)} \right) \\ & = 1 - \frac{(1-\alpha_t) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} + \frac{u^\top (x - \sqrt{\bar{\alpha}_t} x_0)}{\alpha_t - \bar{\alpha}_t} + O \left((\theta_t(x)^2 d^2 \log^2 T + \varepsilon_{\text{score},t}(x)^2 \theta_t(x) d \log T) \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^2 \right) \end{aligned} \quad (135)$$

for any $x_0 \in \mathcal{E}_2^{\text{typical}}$, which invokes (134) and (130b) (under the assumption (133)). Combine (135) and (132) to show that

$$\begin{aligned} & \int_{x_0} p_{X_0 | X_t}(x_0 | x) \exp \left(-\frac{(1-\alpha_t) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top (x - \sqrt{\bar{\alpha}_t} x_0)}{2(\alpha_t - \bar{\alpha}_t)} \right) dx_0 \\ & = \left(\int_{x_0 \in \mathcal{E}_2} + \int_{x_0 \notin \mathcal{E}_2} \right) p_{X_0 | X_t}(x_0 | x) \exp \left(-\frac{(1-\alpha_t) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top (x - \sqrt{\bar{\alpha}_t} x_0)}{2(\alpha_t - \bar{\alpha}_t)} \right) dx_0 \\ & = \int_{x_0 \in \mathcal{E}_2} p_{X_0 | X_t}(x_0 | x) \left(1 - \frac{(1-\alpha_t) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} + \frac{u^\top (x - \sqrt{\bar{\alpha}_t} x_0)}{\alpha_t - \bar{\alpha}_t} \right) dx_0 \\ & \quad + O \left((\theta_t(x)^2 d^2 \log^2 T + \varepsilon_{\text{score},t}(x)^2 \theta_t(x) d \log T) \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^2 \right) + O \left(\sum_{c=3}^{\infty} \int_{x_0 \in \mathcal{E}_c \setminus \mathcal{E}_{c-1}} p_{X_0 | X_t}(x_0 | x) \exp(c\theta_t(x)d) dx_0 \right) \\ & = 1 - \frac{(1-\alpha_t) \left(\int_{x_0} p_{X_0 | X_t}(x_0 | x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 dx_0 - \left\| \int_{x_0} p_{X_0 | X_t}(x_0 | x) (x - \sqrt{\bar{\alpha}_t} x_0) dx_0 \right\|_2^2 \right)}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} \\ & \quad + O \left(\theta_t(x)^2 d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^2 \log^2 T + \varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x) d \log T} \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right) \right) + O \left(\exp(-\theta_t(x) d \log T) \right) \\ & = 1 - \frac{(1-\alpha_t) \left(\int_{x_0} p_{X_0 | X_t}(x_0 | x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 dx_0 - \left\| \int_{x_0} p_{X_0 | X_t}(x_0 | x) (x - \sqrt{\bar{\alpha}_t} x_0) dx_0 \right\|_2^2 \right)}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} \\ & \quad + O \left(\theta_t(x)^2 d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^2 \log^2 T + \varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x) d \log T} \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right) \right), \end{aligned} \quad (136)$$

where the penultimate relation holds since, according to (128a),

$$\begin{aligned} \sum_{c=3}^{\infty} \int_{x_0 \in \mathcal{E}_c \setminus \mathcal{E}_{c-1}} p_{X_0 | X_t}(x_0 | x) \exp(c\theta_t(x)d) dx_0 & \leq \sum_{c=3}^{\infty} \exp(-c^2 \theta_t(x) d \log T) \exp(c\theta_t(x)d) \\ & \leq \sum_{c=3}^{\infty} \exp \left(-\frac{1}{2} c^2 \theta_t(x) d \log T \right) \leq \exp(-\theta_t(x) d \log T), \end{aligned}$$

and the last line in (136) again utilizes (39) and the fact that $\theta_t(x) \geq c_6$ for some large enough constant $c_6 > 0$.

Putting (136) and (127) together yields

$$\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} = 1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \bar{\alpha}_t)} + O\left(\theta_t(x)^2 d^2 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^2 \log^2 T + \varepsilon_{\text{score},t}(x) \sqrt{\theta_t(x)} d \log T \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right)\right) - \frac{(1-\alpha_t) \left(\int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0 - \left\| \int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t}x_0) dx_0 \right\|_2^2 \right)}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)}$$

as claimed.

B.1.2 Proof of relation (58b)

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

$$p_{\phi_t(Y)}(\phi_t(x)) = \det\left(\frac{\partial\phi_t(x)}{\partial x}\right)^{-1} p_Y(x), \quad (137a)$$

$$p_{\phi_t^*(Y)}(\phi_t^*(x)) = \det\left(\frac{\partial\phi_t^*(x)}{\partial x}\right)^{-1} p_Y(x), \quad (137b)$$

where $\frac{\partial\phi_t(x)}{\partial x}$ and $\frac{\partial\phi_t^*(x)}{\partial x}$ denote the Jacobian matrices. It thus suffices to control the quantity $\det\left(\frac{\partial\phi_t(x)}{\partial x}\right)^{-1}$.

To begin with, recall from (51) and (36) that

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

As a result, one can use (37) and (38) to derive

$$\begin{aligned} I - \frac{\partial\phi_t^*(x)}{\partial x} &= \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)} J_t(x) = \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)} \left\{ I + \frac{1}{1-\bar{\alpha}_t} \left\{ \mathbb{E}[X_t - \sqrt{\bar{\alpha}_t}X_0 | X_t = x] \left(\mathbb{E}[X_t - \sqrt{\bar{\alpha}_t}X_0 | X_t = x] \right)^\top \right. \right. \\ &\quad \left. \left. - \mathbb{E}\left[(X_t - \sqrt{\bar{\alpha}_t}X_0)(X_t - \sqrt{\bar{\alpha}_t}X_0)^\top | X_t = x \right] \right\} \right\} \\ &=: \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)} \left\{ I + \frac{1}{1-\bar{\alpha}_t} B \right\}. \end{aligned} \quad (138)$$

This allows one to show that

$$\begin{aligned} \text{Tr}\left(I - \frac{\partial\phi_t^*(x)}{\partial x}\right) &= \frac{d(1-\alpha_t)}{2(1-\bar{\alpha}_t)} + \\ &\quad \frac{(1-\alpha_t) \left(\left\| \int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t}x_0) dx_0 \right\|_2^2 - \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0 \right)}{2(1-\bar{\alpha}_t)^2}. \end{aligned} \quad (139a)$$

Moreover, the matrix B defined in (138) satisfies

$$\|B\|_{\text{F}} \leq \left\| \mathbb{E}\left[(X_t - \sqrt{\bar{\alpha}_t}X_0)(X_t - \sqrt{\bar{\alpha}_t}X_0)^\top | X_t = x \right] \right\|_{\text{F}} \leq \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0$$

due to Jensen's inequality. Taking this together with (138) and Lemma 1 reveals that

$$\begin{aligned} \left\| \frac{\partial\phi_t^*(x)}{\partial x} - I \right\| &\leq \left\| \frac{\partial\phi_t^*(x)}{\partial x} - I \right\|_{\text{F}} \lesssim \frac{1-\alpha_t}{1-\bar{\alpha}_t} \left(\sqrt{d} + \frac{\int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0}{1-\bar{\alpha}_t} \right) \\ &\lesssim \frac{\theta_t(x) d(1-\alpha_t) \log T}{1-\bar{\alpha}_t}. \end{aligned} \quad (139b)$$

Additionally, the Taylor expansion guarantees that for any A and Δ ,

$$\det(I + A + \Delta) = 1 + \text{Tr}(A) + O((\text{Tr}(A))^2 + \|A\|_{\mathbb{F}}^2 + d^3\|A\|^3 + d\|\Delta\|) \quad (140a)$$

$$\det(I + A + \Delta)^{-1} = 1 - \text{Tr}(A) + O((\text{Tr}(A))^2 + \|A\|_{\mathbb{F}}^2 + d^3\|A\|^3 + d\|\Delta\|) \quad (140b)$$

hold as long as $d\|A\| + d\|\Delta\| \leq c_{11}$ for some small enough constant $c_{11} > 0$. The above properties taken collectively with (51) and (48) allow us to demonstrate that

$$\begin{aligned} \frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)} &= \det\left(\frac{\partial\phi_t(x)}{\partial x}\right)^{-1} = \left(\det\left(\frac{\partial\phi_t^*(x)}{\partial x} + \frac{1-\alpha_t}{2}[J_{s_t}(x) - J_{s_t^*}(x)]\right)\right)^{-1} \\ &= \left(\det\left(I + \frac{\partial\phi_t^*(x)}{\partial x} - I + \frac{1-\alpha_t}{2}[J_{s_t}(x) - J_{s_t^*}(x)]\right)\right)^{-1} \\ &= 1 - \text{Tr}\left(\frac{\partial\phi_t^*(x)}{\partial x} - I\right) + O\left(\theta_t(x)^2 d^2 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^2 \log^2 T + \theta^3 d^6 \log^3 T \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^3 + (1-\alpha_t)d\varepsilon_{\text{Jacobi},t}(x)\right) \\ &= 1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \bar{\alpha}_t)} + \frac{(1-\alpha_t)\left(\left\|\int_{x_0} p_{X_0|X_t}(x_0|x)(x - \sqrt{\bar{\alpha}_t}x_0)dx_0\right\|_2^2 - \int_{x_0} p_{X_0|X_t}(x_0|x)\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0\right)}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} \\ &\quad + O\left(\theta_t(x)^2 d^2 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^2 \log^2 T + \theta^3 d^6 \log^3 T \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^3 + (1-\alpha_t)d\varepsilon_{\text{Jacobi},t}(x)\right), \end{aligned} \quad (141)$$

with the proviso that

$$\frac{d^2(1-\alpha_t)\log T}{\alpha_t - \bar{\alpha}_t} \leq \frac{8c_1 d^2 \log^2 T}{T} \leq c_{12} \quad \text{and} \quad (1-\alpha_t)d\varepsilon_{\text{Jacobi},t}(x) \leq \frac{c_1 d\varepsilon_{\text{Jacobi},t}(x)\log T}{T} \leq c_{12}$$

for some sufficiently small constant $c_{12} > 0$ (see (39)).

B.2 Proof of Lemma 5

In view of the definition (67), one has

$$S_k(y_T) \leq c_{14}, \quad \text{for any } k < \tau(y_T). \quad (142)$$

Suppose instead that (69) does not hold true, namely, $-\log q_k(y_k) > 2c_6 d \log T$ for some $k < \tau(y_T)$, and we would like to show that this leads to contradiction.

Towards this, let $1 < t \leq k$ be the smallest time step obeying

$$\theta_t(y_t) = \max\left\{-\frac{\log q_t(y_t)}{d \log T}, c_6\right\} > 2c_6 = 2\theta_1(y_1), \quad (143)$$

where the last identity holds since $-\log q_1(y_1) \leq c_6 d \log T$ and hence $\theta_1(y_1) = \max\left\{-\frac{\log q_1(y_1)}{d \log T}, c_6\right\} = c_6$. We claim that t necessarily obeys

$$2c_6 < \theta_t(y_t) \leq 4c_6. \quad (144)$$

Assuming the validity of Claim (144) for the moment, it necessarily satisfies

$$\theta_1(y_1), \theta_2(y_2), \dots, \theta_t(y_t) \in [c_6, 4c_6].$$

According to the relations (59) and (142), we derive

$$\begin{aligned} c_6 = \theta_1(y_1) &\leq \theta_t(y_t) - \theta_1(y_1) = -\frac{\log q_t(y_t)}{d \log T} - \theta_1(y_1) \leq \frac{-\log q_t(y_t) + \log q_1(y_1)}{d \log T} \\ &= \frac{1}{d \log T} \sum_{j=1}^{t-1} (\log q_j(y_j) - \log q_{j+1}(y_{j+1})) \end{aligned}$$

$$\leq 2c_1 + C_{10} \left\{ \frac{d \log^3 T}{T} + \frac{S_{\tau(y_T)-1}(y_T)}{d \log T} \right\} < 3c_1$$

under our sample size condition. This, however, cannot possibly hold if $c_6 \geq 3c_1$ as assumed for Lemma 5.

To finish up, it suffices to justify Claim (144). In order to see this, suppose instead that $\theta_t(y_t) > 4c_6$. Given relation (142) that $S_k(y_T) \leq c_{14}$, it can be readily seen from (57), (142) as well as the learning rate properties (39) that

$$\begin{aligned} \theta_{t-1}(y_{t-1}) &= \theta_t(y_t) + \theta_{t-1}(y_{t-1}) - \theta_t(y_t) \\ &= \theta_t(y_t) + \theta_{t-1}(y_{t-1}) + \frac{\log p_t(y_t)}{d \log T} \geq \theta_t(y_t) - \frac{\log p_{t-1}(y_{t-1}) - \log p_t(y_t)}{d \log T} \\ &\geq \theta_t(y_t) - \frac{4c_1 \left(5\varepsilon_{\text{score},t}(y_t) \sqrt{\theta_t(y_t) d \log T} + 60\theta_t(y_t) d \log T \right)}{dT} - \frac{\log 2}{d \log T} \\ &\geq \theta_t(y_t) - \frac{4c_1 \left(5\varepsilon_{\text{score},t}(y_t) \sqrt{d \log T} + 60d \log T \right)}{dT} \theta_t(y_t) - \frac{\log 2}{d \log T} \\ &> \frac{1}{2} \theta_t(y_t) > 2c_6, \end{aligned}$$

which is contradictory with the assumption that t is the smallest step obeying $\theta_t(y_t) > 2c_6$. Thus, we complete the proof of relation (69) as required.

B.3 Proof of Lemma 6

Next, consider any y_T , with $\{y_{T-1}, \dots, y_1\}$ being the associated deterministic sequence (cf. (53)). As an immediate consequence of Lemma 5 and the definition (56) of $\theta_t(\cdot)$, one has

$$\theta_t(y_t) \leq 2c_6, \quad \forall t < \tau(y_T) \quad (145)$$

We then intend to invoke Lemma 4 to control the term of interest. To do so, note that Lemma 1, (39) and the definition (67) of $\tau(y_T)$ taken together reveal that: for all $t < \tau(y_T)$ one has

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

$$\begin{aligned} &\theta_t(y_t)^2 d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^2 \log^2 T + \varepsilon_{\text{score},t}(y_t) \sqrt{\theta_t(y_t) d \log T} \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right) + \theta_t(y_t)^3 d^6 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^3 \log^3 T + (1 - \alpha_t) d \varepsilon_{\text{Jacobi},t}(y_t) \\ &\lesssim \frac{d^2 \log^4 T}{T^2} + \frac{d^6 \log^6 T}{T^3} + \frac{\varepsilon_{\text{score},t}(y_t) \sqrt{d \log^3 T}}{T} + \frac{d \varepsilon_{\text{Jacobi},t}(y_t) \log T}{T} = o(1), \end{aligned}$$

and

$$\begin{aligned} &\left| \frac{(1 - \alpha_t) \left(\left\| \int \mathbb{E}[X_t - \sqrt{\bar{\alpha}_t} X_0 | X_t = y_t] \right\|_2^2 - \int \mathbb{E}[\|X_t - \sqrt{\bar{\alpha}_t} X_0\|_2^2 | X_t = y_t] \right)}{(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} \right| \\ &\leq \left| \frac{(1 - \alpha_t) \int \mathbb{E}[\|X_t - \sqrt{\bar{\alpha}_t} X_0\|_2^2 | X_t = y_t]}{(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} \right| \lesssim \frac{(1 - \alpha_t) d \log T}{\alpha_t - \bar{\alpha}_t} \lesssim \frac{d \log^2 T}{T} = o(1). \end{aligned}$$

With these bounds in mind, applying relations (58a) and (58b) in Lemma 4 leads to

$$\begin{aligned} &\frac{p_{\sqrt{\alpha_t} Y_{t-1}}(\phi_t(y_t))}{p_{Y_t}(y_t)} \left(\frac{p_{\sqrt{\alpha_t} X_{t-1}}(\phi_t(y_t))}{p_{X_t}(y_t)} \right)^{-1} = \frac{p_{\phi_t(Y_t)}(\phi_t(y_t))}{p_{Y_t}(y_t)} \left(\frac{p_{\sqrt{\alpha_t} X_{t-1}}(\phi_t(y_t))}{p_{X_t}(y_t)} \right)^{-1} \\ &= 1 + O \left(\frac{d^2 \log^4 T}{T^2} + \frac{d^6 \log^6 T}{T^3} + \frac{\varepsilon_{\text{score},t}(y_t) \sqrt{d \log^3 T}}{T} + \frac{d \varepsilon_{\text{Jacobi},t}(y_t) \log T}{T} \right) \end{aligned}$$

for all $t < \tau(y_T)$. Using the fact that $y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \phi_t(y_t)$ and invoking the relation (55), we arrive at

$$\frac{p_{t-1}(y_{t-1})}{q_{t-1}(y_{t-1})} = \left\{ 1 + O\left(\frac{d^2 \log^4 T}{T^2} + \frac{d^6 \log^6 T}{T^3} + \frac{\varepsilon_{\text{score},t}(y_t) \sqrt{d \log^3 T}}{T} + \frac{d \varepsilon_{\text{Jacobi},t}(y_t) \log T}{T} \right) \right\} \frac{p_t(y_t)}{q_t(y_t)}$$

for any $t < \tau(y_T)$. By abbreviating $\tau = \tau(y_T)$ for notational simplicity, we reach

$$\begin{aligned} \frac{q_1(y_1)}{p_1(y_1)} &= \left\{ 1 + O\left(\frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + S_{\tau-1}(y_{\tau-1}) \right) \right\} \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})} \\ &\in \left[\frac{p_{\tau-1}(y_{\tau-1})}{2q_{\tau-1}(y_{\tau-1})}, \frac{2p_{\tau-1}(y_{\tau-1})}{q_{\tau-1}(y_{\tau-1})} \right], \end{aligned} \quad (146a)$$

and similarly,

$$\frac{q_k(y_k)}{2p_k(y_k)} \leq \frac{q_1(y_1)}{p_1(y_1)} \leq 2 \frac{q_k(y_k)}{p_k(y_k)}, \quad \forall k < \tau. \quad (146b)$$

B.4 Proof of Lemma 7

In the following, we shall tackle \mathcal{I}_2 , \mathcal{I}_3 and \mathcal{I}_4 separately. Throughout this proof, we shall abbreviate $\tau = \tau(Y_T)$ (cf. (67)) whenever it is clear from the context.

The sub-collection in \mathcal{I}_2 . By virtue of the definition (72a) of \mathcal{I}_2 , we make the observation that

$$\begin{aligned} &\mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2\} \right] \stackrel{(i)}{\leq} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2\} \frac{S_\tau(Y_T)}{c_{14}} \right] \\ &\stackrel{(ii)}{=} \frac{\log T}{c_{14} T} \sum_{t=2}^{\tau} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2\} \left(d \varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \right] \\ &\stackrel{(iii)}{\leq} \frac{2 \log T}{c_{14} T} \sum_{t=2}^{\tau} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2\} \left(d \varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \right] \\ &= \frac{2 \log T}{c_{14} T} \sum_{t=2}^T \sum_{i \in \mathcal{I}_2, \tau \geq t} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2\} \left(d \varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \right] \\ &\leq \frac{2 \log T}{c_{14} T} \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \left(d \varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \right] \\ &= \frac{2 \log T}{c_{14} T} \sum_{t=2}^T \mathbb{E}_{Y_t \sim p_t} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \left(d \varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \right] \\ &= \frac{2 \log T}{c_{14} T} \sum_{t=2}^T \mathbb{E}_{Y_t \sim q_t} \left[d \varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right] \\ &\stackrel{(iv)}{\lesssim} (d \varepsilon_{\text{Jacobi}} + \sqrt{d \log T} \varepsilon_{\text{score}}) \log T. \end{aligned} \quad (147)$$

Here, (i) follows since $S_\tau(y_T) \geq c_{14}$ in \mathcal{I}_2 (see (72a)); (ii) comes from the definition of $S_t(\cdot)$ (see (54)); (iii) holds since (by repeating the same proof arguments as for (70) as long as $2c_{14}$ is small enough)

$$\frac{p_1(y_1)}{q_1(y_1)} \leq \frac{2p_t(y_t)}{q_t(y_t)}, \quad \forall t \leq \tau;$$

and (iv) arises from (49).

The sub-collection in \mathcal{I}_3 . With regards to \mathcal{I}_3 (cf. (72b)), we can derive the following bound in a way similar to (147):

$$\begin{aligned}
& \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_3\} \right] \stackrel{(i)}{\leq} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_3\} \frac{\xi_\tau(Y_T)}{c_{14}} \right] \\
&= \frac{\log T}{c_{14}T} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_3\} \left(d\varepsilon_{\text{Jacobi},\tau}(Y_\tau) + \sqrt{d \log T} \varepsilon_{\text{score},\tau}(Y_\tau) \right) \right] \\
&\stackrel{(ii)}{\leq} \frac{2 \log T}{c_{14}T} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_{\tau-1}(Y_{\tau-1})}{p_{\tau-1}(Y_{\tau-1})} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_3\} \left(d\varepsilon_{\text{Jacobi},\tau}(Y_\tau) + \sqrt{d \log T} \varepsilon_{\text{score},\tau}(Y_\tau) \right) \right] \\
&= \frac{2 \log T}{c_{14}T} \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_3\} \left(d\varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \mathbb{1}\{\tau = t\} \right] \quad (148) \\
&\stackrel{(iii)}{\leq} \frac{16 \log T}{c_{14}T} \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_\tau(Y_\tau)}{p_\tau(Y_\tau)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_3\} \left(d\varepsilon_{\text{Jacobi},\tau}(Y_\tau) + \sqrt{d \log T} \varepsilon_{\text{score},\tau}(Y_\tau) \right) \right] \\
&\leq \frac{16 \log T}{c_{14}T} \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \left(d\varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \right] \\
&= \frac{16 \log T}{c_{14}T} \sum_{t=2}^T \mathbb{E}_{Y_t \sim p_t} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \left(d\varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right) \right] \\
&= \frac{16 \log T}{c_{14}T} \sum_{t=2}^T \mathbb{E}_{Y_t \sim q_t} \left[d\varepsilon_{\text{Jacobi},t}(Y_t) + \sqrt{d \log T} \varepsilon_{\text{score},t}(Y_t) \right] \\
&\lesssim \left(d\varepsilon_{\text{Jacobi}} + \sqrt{d \log T} \varepsilon_{\text{score}} \right) \log T. \quad (149)
\end{aligned}$$

Here, (i) comes from (72b), (ii) arises from (70b), whereas (iii) is a consequence of (72b).

The sub-collection in \mathcal{I}_4 . We now turn attention to \mathcal{I}_4 (cf. (72c)), towards which we find it helpful to define

$$\mathcal{J}_{1,t} := \left\{ y_T : \xi_t(y_T) < c_{14} \right\} \quad (150a)$$

$$\mathcal{J}_{2,t} := \left\{ y_T : \xi_t(y_T) \geq c_{14}, \frac{q_{t-1}(y_{t-1})}{p_{t-1}(y_{t-1})} \leq \frac{8q_t(y_t)}{p_t(y_t)} \right\} \quad (150b)$$

$$\mathcal{J}_{3,t} := \left\{ y_T : \xi_t(y_T) \geq c_{14}, \frac{q_{t-1}(y_{t-1})}{p_{t-1}(y_{t-1})} > \frac{8q_t(y_t)}{p_t(y_t)} \right\} \quad (150c)$$

for each $2 \leq t \leq T$. Equipped with the above definitions, we first make the observation that

$$\begin{aligned}
\mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4\} \right] &\leq 2 \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_{\tau-1}(Y_{\tau-1})}{p_{\tau-1}(Y_{\tau-1})} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4\} \right] \\
&= 2 \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4\} \mathbb{1}\{\tau = t\} \right] \\
&\leq 2 \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{J}_{3,t}\} \right], \quad (151)
\end{aligned}$$

where the first inequality follows from (70b), and the last line comes from the definition of \mathcal{I}_4 (cf. (72c)) and $\mathcal{J}_{3,t}$ (cf. (150c)). For notational simplicity, let us define, for $2 \leq t \leq T$,

$$h_t := \frac{q_t(Y_t)}{p_t(Y_t)}.$$

In view of the second inequality in (150c), one has $h_{t-1} > 8h_t$ as long as $y_T \in \mathcal{J}_{3,t}$. Consequently,

$$\begin{aligned}
& \sum_{t=2}^T h_{t-1} \mathbb{1}\{Y_T \in \mathcal{J}_{3,t}\} \\
& < \sum_{t=2}^T h_{t-1} \mathbb{1}\{Y_T \in \mathcal{J}_{3,t}\} + \frac{1}{7} \sum_{t=2}^T h_{t-1} \mathbb{1}\{Y_T \in \mathcal{J}_{3,t}\} - \frac{8}{7} \sum_{t=2}^T h_t \mathbb{1}\{Y_T \in \mathcal{J}_{3,t}\} \\
& = \frac{8}{7} \sum_{t=2}^T \left(\left(h_{t-1} - h_{t-1} \mathbb{1}\{Y_T \in \mathcal{J}_{1,t}\} - h_{t-1} \mathbb{1}\{Y_T \in \mathcal{J}_{2,t}\} \right) - \left(h_t - h_t \mathbb{1}\{Y_T \in \mathcal{J}_{1,t}\} - h_t \mathbb{1}\{Y_T \in \mathcal{J}_{2,t}\} \right) \right) \\
& = \frac{8}{7} \sum_{t=2}^T (h_t - h_{t-1}) \mathbb{1}\{Y_T \in \mathcal{J}_{1,t} \cup \mathcal{J}_{2,t}\} + \frac{8}{7} \sum_{t=2}^T (h_{t-1} - h_t).
\end{aligned}$$

Here, the second line holds true since, for all t , one has (i) $\mathcal{J}_{1,t} \cup \mathcal{J}_{2,t} \cup \mathcal{J}_{3,t} = \mathbb{R}^d$, and (ii) $\mathcal{J}_{1,t}$, $\mathcal{J}_{2,t}$ and $\mathcal{J}_{3,t}$ are disjoint. Substituting this into (151), we arrive at

$$\begin{aligned}
& \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{J}_{3,t}\} \right] \leq 2 \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} [h_{t-1} \mathbb{1}\{Y_T \in \mathcal{J}_{3,t}\}] \\
& \leq \frac{8}{7} \sum_{t=2}^T \left(\mathbb{E}_{Y_T \sim p_T} [h_t \mathbb{1}\{Y_T \in \mathcal{J}_{1,t} \cup \mathcal{J}_{2,t}\}] - \mathbb{E}_{Y_T \sim p_T} [h_{t-1} \mathbb{1}\{Y_T \in \mathcal{J}_{1,t} \cup \mathcal{J}_{2,t}\}] \right) \\
& \quad + \frac{8}{7} \sum_{t=2}^T \left(\mathbb{E}_{Y_T \sim p_T} [h_{t-1}] - \mathbb{E}_{Y_T \sim p_T} [h_t] \right). \tag{152}
\end{aligned}$$

In order to further bound (152), we make note of a few basic facts. Firstly, the identity below holds:

$$\mathbb{E}_{Y_T \sim p_T} [h_t] = \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \right] = \mathbb{E}_{Y_t \sim p_t} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \right] = 1, \quad 2 \leq t \leq T.$$

Secondly, by defining the set

$$\mathcal{E}_t := \left\{ y : q_t(y) > \exp(-c_6 d \log T) \right\}, \quad 2 \leq t \leq T, \tag{153}$$

we can show that

$$\begin{aligned}
& \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} [h_t \mathbb{1}\{Y_t \notin \mathcal{E}_t, Y_T \in \mathcal{J}_{1,t}\}] \leq \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \mathbb{1}\{Y_t \notin \mathcal{E}_t\} \right] = \sum_{t=2}^T \mathbb{E}_{Y_t \sim p_t} \left[\frac{q_t(Y_t)}{p_t(Y_t)} \mathbb{1}\{Y_t \notin \mathcal{E}_t\} \right] \\
& = \sum_{t=2}^T \mathbb{P}_{Y_t \sim q_t} \{Y_t \notin \mathcal{E}_t\} = \sum_{t=2}^T \mathbb{P}_{X_t \sim q_t} \{X_t \notin \mathcal{E}_t\} \\
& \leq \sum_{t=2}^T \mathbb{P}_{X_t \sim q_t} \{X_t \notin \mathcal{E}_t \text{ and } \|X_t\|_2 \leq T^{2c_R+2}\} + \sum_{t=2}^T \mathbb{P}_{X_t \sim q_t} \{\|X_t\|_2 > T^{2c_R+2}\} \\
& \leq \sum_{t=2}^T \int_{x_t: q_t(x_t) \leq \exp(-c_6 d \log T), \|x_t\|_2 \leq T^{2c_R+2}} q_t(x_t) dx_t + T \exp(-c_6 d \log T) \\
& \leq T(2T^{2c_R+2})^d \exp(-c_6 d \log T) + T \exp(-c_6 d \log T) \leq \exp\left(-\frac{c_6}{2} d \log T\right),
\end{aligned}$$

where the penultimate line comes from (62), and the last inequality holds true as long as c_6 is large enough. Plugging the preceding two results into (152), we reach

$$\mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4\} \right] \leq \frac{8}{7} \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} [(h_t - h_{t-1}) \mathbb{1}\{y_t \in \mathcal{E}_t, Y_T \in \mathcal{J}_{1,t}\}]$$

$$+ \frac{8}{7} \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} [h_t \mathbb{1}\{Y_T \in \mathcal{J}_{2,t}\}] + \exp\left(-\frac{c_6}{2} d \log T\right). \quad (154)$$

As it turns out, the sum w.r.t. the set $\mathcal{J}_{1,t}$ and the sum w.r.t. the set $\mathcal{J}_{2,t}$ in (154) can be controlled respectively using the same arguments as for \mathcal{I}_1 and \mathcal{I}_3 to derive

$$\begin{aligned} \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} [(h_t - h_{t-1}) \mathbb{1}\{y_t \in \mathcal{E}_t, Y_T \in \mathcal{J}_{1,t}\}] &\lesssim \frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + \sqrt{d \log^3 T} \varepsilon_{\text{score}} + (d \log T) \varepsilon_{\text{Jacobi}}, \\ \sum_{t=2}^T \mathbb{E}_{Y_T \sim p_T} [h_t \mathbb{1}\{Y_T \in \mathcal{J}_{2,t}\}] &\lesssim \sqrt{d \log^3 T} \varepsilon_{\text{score}} + (d \log T) \varepsilon_{\text{Jacobi}}; \end{aligned}$$

we omit the arguments here for the sake of brevity. Therefore, we have proven that

$$\mathbb{E}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4\} \right] \lesssim \frac{d^2 \log^4 T}{T} + \frac{d^6 \log^6 T}{T^2} + \sqrt{d \log^3 T} \varepsilon_{\text{score}} + (d \log T) \varepsilon_{\text{Jacobi}}. \quad (155)$$

Putting all this together. Taking (147), (149) and (155) together, we establish the advertised result.

C Proofs of auxiliary lemmas for the DDPM-type sampler

C.1 Proof of Lemma 8

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

$$x_t(\gamma) := \gamma x_{t-1} + (1 - \gamma) \hat{x}_t \quad \text{and} \quad \hat{x}_t := \frac{1}{\sqrt{\alpha_t}} x_t. \quad (156)$$

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

$$\begin{aligned} &p_{X_{t-1}|X_t}(x_{t-1}|x_t) \\ &= \frac{1}{p_{X_t}(x_t)} p_{X_{t-1}, X_t}(x_{t-1}, x_t) = \frac{1}{p_{X_t}(x_t)} \exp\left(\log p_{X_{t-1}}(x_{t-1}) + \log p_{X_t|X_{t-1}}(x_t|x_{t-1})\right) \\ &= \frac{1}{p_{X_t}(x_t)} \exp\left(\log p_{X_{t-1}}(\hat{x}_t) + \int_0^1 \left[\nabla \log p_{X_{t-1}}(x_t(\gamma))\right]^\top (x_{t-1} - \hat{x}_t) d\gamma + \log p_{X_t|X_{t-1}}(x_t|x_{t-1})\right) \\ &= \frac{p_{X_{t-1}}(\hat{x}_t)}{p_{X_t}(x_t)} \exp\left((x_{t-1} - \hat{x}_t)^\top \int_0^1 d\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}|X_0}(x_t(\gamma)|x_0) p_{X_0}(x_0)}{p_{X_{t-1}}(x_t(\gamma))} dx_0 + \log p_{X_t|X_{t-1}}(x_t|x_{t-1})\right), \end{aligned} \quad (157)$$

where the penultimate line comes from the fundamental theorem of calculus. In particular, the exponent in (157) consists of a term that satisfies

$$\begin{aligned} &(x_{t-1} - \hat{x}_t)^\top \int_0^1 d\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}|X_0}(x_t(\gamma)|x_0) p_{X_0}(x_0)}{p_{X_{t-1}}(x_t(\gamma))} dx_0 \\ &= (x_{t-1} - \hat{x}_t)^\top \int_0^1 d\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}|X_0}(x_t(\gamma)|x_0)}{p_{X_{t-1}|X_0}(x_t(\gamma)|x_0)} p_{X_0|X_{t-1}}(x_0|x_t(\gamma)) dx_0 \\ &= -(x_{t-1} - \hat{x}_t)^\top \int_0^1 d\gamma \int_{x_0} \frac{x_t(\gamma) - \sqrt{\bar{\alpha}_{t-1}} x_0}{1 - \bar{\alpha}_{t-1}} p_{X_0|X_{t-1}}(x_0|x_t(\gamma)) dx_0 \\ &=: -\frac{1}{1 - \bar{\alpha}_{t-1}} (x_{t-1} - \hat{x}_t)^\top \int_0^1 g_{t-1}(x_t(\gamma)) d\gamma, \end{aligned} \quad (158)$$

where the second line holds since $p_{X_0|X_{t-1}}(x_0|x_t(\gamma))p_{X_{t-1}}(x_t(\gamma)) = p_{X_{t-1}|X_0}(x_t(\gamma)|x_0)p_{X_0}(x_0)$, and we remind the reader of the definition of $g_t(\cdot)$ in (36).

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

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

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

$$\begin{aligned} & (x_{t-1} - \widehat{x}_t)^\top \int_0^1 d\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}|X_0}(x_t(\gamma)|x_0)p_{X_0}(x_0)}{p_{X_{t-1}}(x_t(\gamma))} dx_0 \\ &= -\frac{1}{1 - \bar{\alpha}_{t-1}} \left\{ (x_{t-1} - \widehat{x}_t)^\top g_{t-1}(\widehat{x}_t) + (x_{t-1} - \widehat{x}_t)^\top \int_0^1 \int_0^1 J_{t-1}\left((1-\tau)\widehat{x}_t + \tau x_t(\gamma)\right)(x_t(\gamma) - \widehat{x}_t) d\tau d\gamma \right\} \\ &= -\frac{1}{1 - \bar{\alpha}_{t-1}} \left\{ (x_{t-1} - \widehat{x}_t)^\top g_{t-1}(\widehat{x}_t) + (x_{t-1} - \widehat{x}_t)^\top \left[\int_0^1 \int_0^1 \gamma J_{t-1}\left((1-\tau)\widehat{x}_t + \tau x_t(\gamma)\right) d\tau d\gamma \right] (x_{t-1} - \widehat{x}_t) \right\}. \end{aligned} \quad (159)$$

Combining (157) and (159) allows us to rewrite the target quantity $p_{X_{t-1}|X_t}(x_{t-1}|x_t)$ as:

$$\begin{aligned} & p_{X_{t-1}|X_t}(x_{t-1}|x_t) \\ &= \frac{p_{X_{t-1}}(\widehat{x}_t)}{p_{X_t}(x_t)} \exp \left((x_{t-1} - \widehat{x}_t)^\top \int_0^1 d\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}|X_0}(\widehat{x}_t|x_0)p_{X_0}(x_0)}{p_{X_{t-1}}(\widehat{x}_t)} dx_0 + \log p_{X_t|X_{t-1}}(x_t|x_{t-1}) \right) \\ &= \frac{p_{X_{t-1}}(\widehat{x}_t)}{p_{X_t}(x_t)} \exp \left(-\frac{(x_{t-1} - \widehat{x}_t)^\top g_{t-1}(\widehat{x}_t) + (x_{t-1} - \widehat{x}_t)^\top \left[\int_0^1 \int_0^1 \gamma J_{t-1}\left((1-\tau)\widehat{x}_t + \tau x_t(\gamma)\right) d\tau d\gamma \right] (x_{t-1} - \widehat{x}_t)}{1 - \bar{\alpha}_{t-1}} \right. \\ & \quad \left. - \frac{\alpha_t \|x_{t-1} - \widehat{x}_t\|_2^2}{2(1 - \alpha_t)} - \frac{d}{2} \log(2\pi(1 - \alpha_t)) \right), \end{aligned} \quad (160)$$

where we have also used the fact that conditional on $X_t | X_{t-1} = x_{t-1} \sim \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I_d)$. 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 (160). Consider now any $(x_t, x_{t-1}) \in \mathcal{E}$ (cf. (80)). In order to further simplify the exponent in the display (160), 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

$$\|J_{t-1}(x) - I\| \lesssim d \log T; \quad (161a)$$

- (b) in addition, one has

$$\frac{1}{1 - \bar{\alpha}_t} \left\| (x_{t-1} - \widehat{x}_t)g_t(x_t) - \sqrt{\alpha_t}(x_{t-1} - \widehat{x}_t)g_t(x_t) \right\|_2 \lesssim \frac{d \log^2 T}{T^{3/2}}. \quad (161b)$$

$$\left\| \frac{g_{t-1}(\widehat{x}_t)}{1 - \bar{\alpha}_{t-1}} - \frac{g_t(x_t)}{1 - \bar{\alpha}_t} \right\|_2 \lesssim (1 - \alpha_t) \left(\frac{d \log T}{\alpha_t - \bar{\alpha}_t} \right)^{3/2}, \quad (161c)$$

Assuming the validity of these claims (which will be established in Appendix C.1.1) and recalling the definition of $\mu_t(\cdot)$ in (74b), we can use (160) together with a little algebra to obtain

$$p_{X_{t-1}|X_t}(x_{t-1}|x_t) = f_0(x_t) \exp \left(-\frac{\alpha_t \|x_{t-1} - \mu_t^*(x_t)\|_2^2}{2(1 - \alpha_t)} + \zeta_t(x_{t-1}, x_t) \right) \quad (162)$$

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

$$\begin{aligned}
|\zeta_t(x_{t-1}, x_t)| &\lesssim \|x_{t-1} - \hat{x}_t\|_2^2 \sup_{x: \exists w \in [0,1] \text{ s.t. } x = (1-w)\hat{x}_t + wx_{t-1}} \|J_{t-1}(x)\| \\
&\quad + \|x_{t-1} - \hat{x}_t\|_2 \left\| \frac{\int_{x_0} p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)dx_0}{1 - \bar{\alpha}_{t-1}} - \frac{\int_{x_0} p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0}{1 - \bar{\alpha}_t} \right\|_2 \\
&\lesssim \frac{(d(1 - \alpha_t) \log T) d \log T}{1 - \bar{\alpha}_{t-1}} + \sqrt{d(1 - \alpha_t) \log T} (1 - \alpha_t) \left(\frac{d \log T}{\alpha_t - \bar{\alpha}_t} \right)^{3/2} \\
&\lesssim \frac{(1 - \alpha_t) d^2 \log^2 T}{1 - \bar{\alpha}_{t-1}} + d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^{3/2} \log^2 T \asymp \frac{(1 - \alpha_t) d^2 \log^2 T}{\alpha_t - \bar{\alpha}_t},
\end{aligned}$$

where the penultimate line makes use of the assumption $(x_t, x_{t-1}) \in \mathcal{E}$ (cf. (80)), and the last inequality holds since $\alpha_t \geq 1/2$ (cf. (22)).

Step 3: approximating the function $f_0(x_t)$. To finish up, it remains to quantify the function $f_0(\cdot)$ in (162). Note that for any x_t obeying $p_{X_t}(x_t) \geq \exp(-\frac{1}{2}c_6 d \log T)$, it is easily seen that

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

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

$$1 \geq \int_{x_{t-1}: (x_t, x_{t-1}) \in \mathcal{E}} p_{X_{t-1}|X_t}(x_{t-1}|x_t) dx_{t-1} \geq 1 - \exp\left(-\frac{1}{4}c_6 d \log T\right). \quad (163)$$

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

$$\begin{aligned}
\sqrt{\alpha_t} \|x_{t-1} - \mu_t^*(x_t)\|_2 &= \sqrt{\alpha_t} \left\| x_{t-1} - \hat{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}(1 - \bar{\alpha}_t)} \mathbb{E}[x_t - \sqrt{\bar{\alpha}_t}X_0 | X_t = x_t] \right\|_2 \\
&\geq \sqrt{\alpha_t} \|x_{t-1} - \hat{x}_t\|_2 - \frac{1 - \alpha_t}{(1 - \bar{\alpha}_t)} \mathbb{E}[\|x_t - \sqrt{\bar{\alpha}_t}X_0\|_2 | X_t = x_t] \\
&\geq c_3 \sqrt{\alpha_t} \cdot \sqrt{d(1 - \alpha_t) \log T} - \frac{1 - \alpha_t}{1 - \bar{\alpha}_t} 6\bar{c}_5 \sqrt{d(1 - \bar{\alpha}_t) \log T} \\
&\geq \frac{1}{2} c_3 \sqrt{d(1 - \alpha_t) \log T},
\end{aligned}$$

where the first identity comes from the definition (74b) of $\mu_t^*(x_t)$, the penultimate line makes use of the result (43a) in Lemma 1, and the last inequality is valid as long as c_3 is sufficiently large. This in turn allows one to derive

$$\frac{1}{(2\pi \frac{1-\alpha_t}{\alpha_t})^{d/2}} \int_{x_{t-1}: (x_t, x_{t-1}) \in \mathcal{E}} \exp\left(-\frac{\alpha_t \|x_{t-1} - \mu_t^*(x_t)\|_2^2}{2(1-\alpha_t)}\right) dx_{t-1} \geq 1 - \exp(-c_3 d \log T). \quad (164)$$

In addition, by virtue of (162), the integral in (163) can be respectively bounded from above and from below as follows:

$$\int_{x_{t-1}: (x_t, x_{t-1}) \in \mathcal{E}} p_{X_{t-1}|X_t}(x_{t-1}|x_t) dx_{t-1}$$

$$\begin{aligned}
&= \frac{f_0(x_t) \int_{x_{t-1}:(x_t, x_{t-1}) \in \mathcal{E}} \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t^*(x_t)\|_2^2 + \zeta_t(x_{t-1}, x_t)\right) dx_{t-1}}{\left(2\pi \frac{1-\alpha_t}{\alpha_t}\right)^{-d/2} \int_{x_{t-1}} \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t^*(x_t)\|_2^2\right) dx_{t-1}} \\
&\leq \frac{f_0(x_t) \int_{x_{t-1}:(x_t, x_{t-1}) \in \mathcal{E}} \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t^*(x_t)\|_2^2 + \zeta_t(x_{t-1}, x_t)\right) dx_{t-1}}{\left(2\pi \frac{1-\alpha_t}{\alpha_t}\right)^{-d/2} \int_{x_{t-1}:(x_t, x_{t-1}) \in \mathcal{E}} \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t^*(x_t)\|_2^2\right) dx_{t-1}} \\
&\lesssim \frac{f_0(x_t)}{\left(2\pi \frac{1-\alpha_t}{\alpha_t}\right)^{-d/2}} \exp\left\{O\left(d^2 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right) \log^2 T\right)\right\}
\end{aligned}$$

and

$$\begin{aligned}
&\int_{x_{t-1}:(x_t, x_{t-1}) \in \mathcal{E}} p_{X_{t-1} | X_t}(x_{t-1} | x_t) dx_{t-1} \\
&\geq (1 - \exp(-c_3 d \log T)) \frac{f_0(x_t) \int_{x_{t-1}:(x_t, x_{t-1}) \in \mathcal{E}} \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t^*(x_t)\|_2^2 + \zeta_t(x_{t-1}, x_t)\right) dx_{t-1}}{\left(2\pi \frac{1-\alpha_t}{\alpha_t}\right)^{-d/2} \int_{x_{t-1}:(x_t, x_{t-1}) \in \mathcal{E}} \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)} \|x_{t-1} - \mu_t^*(x_t)\|_2^2\right) dx_{t-1}} \\
&\geq (1 - \exp(-c_3 d \log T)) \frac{f_0(x_t)}{\left(2\pi \frac{1-\alpha_t}{\alpha_t}\right)^{-d/2}} \exp\left\{-O\left(d^2 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right) \log^2 T\right)\right\}.
\end{aligned}$$

These taken collectively with (163) allow one to demonstrate that

$$\max\left\{\frac{f_0(x_t)}{\left(2\pi \frac{1-\alpha_t}{\alpha_t}\right)^{-d/2}}, \frac{\left(2\pi \frac{1-\alpha_t}{\alpha_t}\right)^{-d/2}}{f_0(x_t)}\right\} = \exp\left\{O\left(d^2 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right) \log^2 T\right)\right\} = 1 + O\left(d^2 \left(\frac{1-\alpha_t}{\alpha_t - \bar{\alpha}_t}\right) \log^2 T\right),$$

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

Combining this with (162) concludes the proof of Lemma 8, as long as the two claims in (161) are valid (to be justified in Appendix C.1.1).

C.1.1 Proof of auxiliary claims (161) in Lemma 8

Proof of relation (161a). Recall from (38) that

$$\begin{aligned}
J_{t-1}(x) - I &= \frac{1}{1 - \bar{\alpha}_{t-1}} \mathbb{E}\left[x - \sqrt{\bar{\alpha}_{t-1}} X_0 \mid X_{t-1} = x\right] \left(\mathbb{E}\left[x - \sqrt{\bar{\alpha}_{t-1}} X_0 \mid X_{t-1} = x\right]\right)^\top \\
&\quad - \frac{1}{1 - \bar{\alpha}_{t-1}} \mathbb{E}\left[(x - \sqrt{\bar{\alpha}_{t-1}} X_0)(x - \sqrt{\bar{\alpha}_{t-1}} X_0)^\top \mid X_{t-1} = x\right].
\end{aligned}$$

Recognizing that

$$\left\|\mathbb{E}[ZZ^\top] - \mathbb{E}[Z]\mathbb{E}[Z]^\top\right\| = \left\|\mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^\top\right]\right\| \leq \mathbb{E}[ZZ^\top] \leq \mathbb{E}[\|ZZ^\top\|] = \mathbb{E}[\|Z\|_2^2]$$

for any random vector Z , we can readily obtain

$$\|J_{t-1}(x) - I\| \leq \frac{1}{1 - \bar{\alpha}_{t-1}} \mathbb{E}\left[\|x - \sqrt{\bar{\alpha}_{t-1}} X_0\|_2^2 \mid X_{t-1} = x\right]. \quad (165)$$

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

$$-\log p_{X_{t-1}}(x) \leq c_6 d \log T \quad (166)$$

for any x lying in the line segment connecting x_{t-1} and \hat{x}_t . With this result in place, taking (165) together with the bound (43b) in Lemma 1 immediately leads to

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

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

Proof of relation (161b). To prove this result, we observe that

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

where the last line comes from Lemma 1 as well as the basic property (39a) about the learning rates.

Proof of relation (161c). To begin with, the triangle inequality together with the fact $\bar{\alpha}_t = \prod_{k=1}^t \alpha_k$ gives

$$\begin{aligned} &\left\| \frac{\int_{x_0} p_{X_0 \mid X_{t-1}}(x_0 \mid \hat{x}_t) (\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}} x_0) dx_0}{1 - \bar{\alpha}_{t-1}} - \frac{\int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x_t) (x_t - \sqrt{\bar{\alpha}_t} x_0) dx_0}{1 - \bar{\alpha}_t} \right\|_2 \\ &\leq \left\| \frac{\int_{x_0} p_{X_0 \mid X_{t-1}}(x_0 \mid \hat{x}_t) (x_t - \sqrt{\bar{\alpha}_t} x_0) dx_0 - \int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x_t) (x_t - \sqrt{\bar{\alpha}_t} x_0) dx_0}{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})} \right\|_2 \\ &\quad + \left\| \left(\frac{1}{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})} - \frac{1}{1 - \bar{\alpha}_t} \right) \int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x_t) (x_t - \sqrt{\bar{\alpha}_t} x_0) dx_0 \right\|_2. \end{aligned} \quad (167)$$

Let us first consider the last term in (167). According to Lemma 1, given that $-\log p_{X_t}(x_t) \leq \frac{1}{2} c_6 d \log T$ for some constant $c_6 > 0$, one has

$$\int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x_t) \|x_t - \sqrt{\bar{\alpha}_t} x_0\|_2 dx_0 \lesssim \sqrt{d(1 - \bar{\alpha}_t) \log T}. \quad (168)$$

This in turn reveals that

$$\begin{aligned} &\left\| \left(\frac{1}{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})} - \frac{1}{1 - \bar{\alpha}_t} \right) \int_{x_0} p_{X_0 \mid X_t}(x_0 \mid x_t) (x_t - \sqrt{\bar{\alpha}_t} x_0) dx_0 \right\|_2 \\ &\lesssim \left| \frac{1}{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})} - \frac{1}{1 - \bar{\alpha}_t} \right| \cdot \sqrt{d(1 - \bar{\alpha}_t) \log T} \\ &\asymp \frac{(1 - \sqrt{\alpha_t})(1 + \bar{\alpha}_{t-1} \sqrt{\alpha_t})}{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1}) (1 - \bar{\alpha}_t)} \sqrt{d(1 - \bar{\alpha}_t) \log T} \\ &\asymp \frac{1 - \alpha_t}{(\alpha_t - \bar{\alpha}_t)^{3/2}} \sqrt{d \log T}, \end{aligned} \quad (169)$$

where the last inequality makes use of the properties (39).

Next, we turn attention to the first term in (167), which relies on the following claim.

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

$$p_{X_{t-1}}(\hat{x}_t) = \left(1 + O\left(\frac{d(1 - \alpha_t) \log T}{1 - \bar{\alpha}_{t-1}} \right) \right) p_{X_t}(x_t), \quad (170a)$$

In addition, by defining the following set

$$\mathcal{E}_1 := \{x : \|x_t - \sqrt{\bar{\alpha}_t} x\|_2 \leq c_4 \sqrt{d(1 - \bar{\alpha}_t) \log T}\}$$

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

$$\frac{p_{X_{t-1} \mid X_0}(\hat{x}_t \mid x_0)}{p_{X_t \mid X_0}(x_t \mid x_0)} = 1 + O\left(\frac{d(1 - \alpha_t) \log T}{1 - \bar{\alpha}_{t-1}} \right), \quad \text{if } x_0 \in \mathcal{E}_1, \quad (170b)$$

$$\frac{p_{X_{t-1} \mid X_0}(\hat{x}_t \mid x_0)}{p_{X_t \mid X_0}(x_t \mid x_0)} \leq \exp\left(\frac{16c_1 \|x_t - \sqrt{\bar{\alpha}_t} x_0\|_2^2 \log T}{(1 - \bar{\alpha}_t) T} \right), \quad \text{if } x_0 \notin \mathcal{E}_1. \quad (170c)$$

As an immediate consequence of this claim, one has

$$\begin{aligned} \frac{p_{X_0|X_{t-1}}(x_0|\hat{x}_t)}{p_{X_0|X_t}(x_0|x_t)} &= \frac{p_{X_{t-1}|X_0}(\hat{x}_t|x_0)}{p_{X_t|X_0}(x_t|x_0)} \cdot \frac{p_{X_t}(x_t)}{p_{X_{t-1}}(\hat{x}_t)} = \left(1 + O\left(\frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}}\right)\right) \frac{p_{X_{t-1}|X_0}(\hat{x}_t|x_0)}{p_{X_t|X_0}(x_t|x_0)} \\ &\begin{cases} = 1 + O\left(\frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}}\right), & \text{if } x_0 \in \mathcal{E}_1, \\ \leq \exp\left(\frac{16c_1\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2 \log T}{(1-\bar{\alpha}_t)T}\right), & \text{if } x_0 \notin \mathcal{E}_1. \end{cases} \end{aligned} \quad (171)$$

In turn, this allows one to deduce that

$$\begin{aligned} &\left\| \int_{x_0} p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 - \int_{x_0} p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 \right\|_2 \\ &\leq \int_{x_0 \in \mathcal{E}_1} |p_{X_0|X_{t-1}}(x_0|\hat{x}_t) - p_{X_0|X_t}(x_0|x_t)| \cdot \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 dx_0 \\ &\quad + \int_{x_0 \notin \mathcal{E}_1} |p_{X_0|X_{t-1}}(x_0|\hat{x}_t) - p_{X_0|X_t}(x_0|x_t)| \cdot \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 dx_0 \\ &\leq \int_{x_0} O\left(\frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}}\right) p_{X_0|X_t}(x_0|x_t) \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 dx_0 \\ &\quad + \int_{x_0 \notin \mathcal{E}_1} \exp\left(\frac{16c_1\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2 \log T}{(1-\bar{\alpha}_t)T}\right) p_{X_0|X_t}(x_0|x_t) \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 dx_0 \\ &\leq O\left(\frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}}\right) \mathbb{E}[\|x_t - \sqrt{\bar{\alpha}_t}X_0\|_2 | X_t = x_0] \\ &\quad + \int_{x_0 \notin \mathcal{E}_1} \exp\left(\frac{20c_1\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2 \log T}{(1-\bar{\alpha}_t)T}\right) p_{X_0|X_t}(x_0|x_t) dx_0. \end{aligned}$$

Invoking (43a) in Lemma 1 as well as similar arguments for (118) (assuming that c_4 is sufficiently large), we arrive at

$$\begin{aligned} &\frac{1}{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_{t-1})}} \left\| \int_{x_0} p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 - \int_{x_0} p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 \right\|_2 \\ &\lesssim \frac{d^{3/2}(1-\alpha_t)\sqrt{1-\bar{\alpha}_t}\log^{3/2}T}{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_{t-1})^2}} + \frac{1}{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_{t-1})}} \cdot \frac{1}{T^{c_0}} \\ &\asymp \frac{d^{3/2}(1-\alpha_t)\log^{3/2}T}{(1-\bar{\alpha}_{t-1})^{3/2}}, \end{aligned} \quad (172)$$

where c_0 is some large enough constant, and we have also made use of the properties in (39).

Substituting (169) and (172) into (167) readily concludes the proof.

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

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

$$= 1 + O\left(\frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}}\right), \quad (173)$$

where the second line holds since

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

and we have also made use of (39) (so that $\frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}} = o(1)$ under our assumption on T). Additionally, for any $x_0 \notin \mathcal{E}_1$, it follows from the argument for (110) that

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

where the last line comes from (39b) and (39c). Taking the above bounds together and invoking the same calculation as in (116), (118) and (121), we reach

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

thus concluding the proof of Claim 2. \square

C.2 Proof of Lemma 9

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

$$\begin{aligned} \log p_{X_{t-1}|X_t}(x_{t-1}|x_t) &= \log \frac{p_{X_t|X_{t-1}}(x_t|x_{t-1})p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} \\ &= \log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} + \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_{t-1}\|_2^2}{2(1-\alpha_t)} - \frac{d}{2} \log(2\pi(1-\alpha_t)) \\ &\leq \log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} + T(\|x_{t-1} - \hat{x}_t\|_2^2 + 1), \end{aligned}$$

where the last inequality makes use of the properties (39) about α_t (recall that \hat{x}_t has been defined in (77)). Some direct calculations then yield

$$\begin{aligned} \log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} &= \log \frac{\int_{x_0} p_{X_0}(x_0) \exp\left(-\frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\|_2^2}{2(1-\bar{\alpha}_{t-1})}\right) dx_0}{\int_{x_0} p_{X_0}(x_0) \exp\left(-\frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(1-\bar{\alpha}_t)}\right) dx_0} - \frac{d}{2} \log\left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\right) \\ &\leq \sup_{x_0: \|x_0\|_2 \leq T^{c_R}} \left\{ \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(1-\bar{\alpha}_t)} - \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\|_2^2}{2(1-\bar{\alpha}_{t-1})} \right\} + \frac{d}{2} \log 2 \\ &\leq \sup_{x_0: \|x_0\|_2 \leq T^{c_R}} \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(1-\bar{\alpha}_t)} + \frac{d}{2} \log 2 \\ &\leq \sup_{x_0: \|x_0\|_2 \leq T^{c_R}} \frac{\|x_t\|_2^2 + \|x_0\|_2^2}{1-\bar{\alpha}_t} + \frac{d}{2} \log 2 \\ &\leq 2T(\|x_t\|_2^2 + T^{2c_R}), \end{aligned}$$

where the second line makes use of the properties in (39). Combining the above two relations, we arrive at

$$\log p_{X_{t-1}|X_t}(x_{t-1}|x_t) \leq 2T(\|x_t\|_2^2 + \|x_{t-1} - \hat{x}_t\|_2^2 + T^{2c_R}). \quad (174)$$

We then turn attention to the conditional distribution $p_{Y_{t-1}^*|Y_t}$. First, it follows from (76b) that

$$\log \frac{1}{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)} = \frac{\alpha_t \|x_{t-1} - \mu_t^*(x_t)\|_2^2}{2(1-\alpha_t)} + \frac{d}{2} \log \left(2\pi \frac{1-\alpha_t}{\alpha_t} \right) \quad (175)$$

$$\leq T(\|x_{t-1} - \hat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}) + d \log T, \quad (176)$$

with $\mu_t(x_t)$ defined in (74b). To see why the last inequality follows, we recall from (74b) that

$$\begin{aligned} \|x_{t-1} - \mu_t^*(x_t)\|_2^2 &\leq 2\|x_{t-1} - \hat{x}_t\|_2^2 + 2\|\hat{x}_t - \mu_t^*(x_t)\|_2^2 \\ &= 2\|x_{t-1} - \hat{x}_t\|_2^2 + 2\left(\frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\right)^2 \left\| \int_{x_0} p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0) dx_0 \right\|_2^2 \\ &\leq 2\|x_{t-1} - \hat{x}_t\|_2^2 + \frac{2(1-\alpha_t)^2}{\alpha_t(1-\bar{\alpha}_{t-1})^2} \sup_{x_0: \|x_0\|_2 \leq T^{c_R}} \|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2^2 \\ &\leq 2\|x_{t-1} - \hat{x}_t\|_2^2 + \frac{64c_1^2 \log^2 T}{T^2} \left(2\|x_t\|_2^2 + 2\bar{\alpha}_t T^{2c_R} \right) \\ &\leq 2\|x_{t-1} - \hat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}, \end{aligned} \quad (177)$$

where we have invoked the properties (39).

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

$$\log \frac{p_{X_{t-1}|X_t}(x_{t-1}|x_t)}{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)} \leq 2T(\|x_{t-1} - \hat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}).$$

C.3 Proof of Lemma 10

To begin with, we make the observation that

$$\begin{aligned} &\mathbb{E}_{x_t \sim q_t} \left[\text{KL} \left(p_{X_{t-1}|X_t}(\cdot|x_t) \parallel p_{Y_{t-1}|Y_t}(\cdot|x_t) \right) \right] - \mathbb{E}_{x_t \sim q_t} \left[\text{KL} \left(p_{X_{t-1}|X_t}(\cdot|x_t) \parallel p_{Y_{t-1}^*|Y_t}(\cdot|x_t) \right) \right] \\ &= \int p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \log \frac{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)}{p_{Y_{t-1}|Y_t}(x_{t-1}|x_t)} dx_{t-1} dx_t \\ &= \int p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \frac{\alpha_t}{2(1-\alpha_t)} \left(\|x_{t-1} - \mu_t(x_t)\|_2^2 - \|x_{t-1} - \mu_t^*(x_t)\|_2^2 \right) dx_{t-1} dx_t \\ &= \underbrace{\int p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \left(\frac{1-\alpha_t}{2} \varepsilon_{\text{score},t}(x_t)^2 - \sqrt{\bar{\alpha}_t} (x_{t-1} - \mu_t^*(x_t))^\top (s_t(x) - s_t^*(x_t)) \right) dx_{t-1} dx_t}_{=: \mathcal{H}_t}, \end{aligned} \quad (178)$$

where we have used (76a) and (76b) as well as the definition (48) of $\varepsilon_{\text{score},t}(x)$. Everything comes down to proving that

$$\mathcal{H}_t \lesssim \exp \left(-\frac{\min\{c_3, c_6\}}{8} d \log T \right) + \frac{d \log^3 T}{T} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}(X_t)^2]. \quad (179)$$

Towards this end, one first notes that the first component of \mathcal{H}_t obeys

$$\begin{aligned} \int p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \frac{1-\alpha_t}{2} \varepsilon_{\text{score},t}(x_t)^2 dx_{t-1} dx_t &= \frac{1-\alpha_t}{2} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}(X_t)^2] \\ &\lesssim \frac{\log T}{T} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}(X_t)^2], \end{aligned} \quad (180)$$

where the last inequality comes from (39). We then switch attention to the second component of \mathcal{H}_t . In view of the distribution of $Y_{t-1}^* | Y_t$ in (76b), we obtain

$$\begin{aligned}
& \int p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) (x_{t-1} - \mu_t^*(x_t))^\top (s_t(x_t) - s_t^*(x_t)) dx_{t-1} dx_t \\
&= \int p_{X_t}(x_t) (p_{X_{t-1}|X_t}(x_{t-1}|x_t) - p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)) (x_{t-1} - \mu_t^*(x_t))^\top (s_t(x_t) - s_t^*(x_t)) dx_{t-1} dx_t \\
&\quad + \int_{x_t} p_{X_t}(x_t) \left\{ \int_{x_{t-1}} p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t) (x_{t-1} - \mu_t^*(x_t))^\top dx_{t-1} \right\} (s_t(x_t) - s_t^*(x_t)) dx_t \\
&\stackrel{(i)}{=} \int p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \left(1 - \frac{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)}{p_{X_{t-1}|X_t}(x_{t-1}|x_t)} \right) (x_{t-1} - \mu_t^*(x_t))^\top (s_t(x_t) - s_t^*(x_t)) dx_{t-1} dx_t \\
&\leq \underbrace{\int_{\mathcal{E}} p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \left(1 - \frac{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)}{p_{X_{t-1}|X_t}(x_{t-1}|x_t)} \right) (x_{t-1} - \mu_t^*(x_t))^\top (s_t(x_t) - s_t^*(x_t)) dx_{t-1} dx_t}_{=: \mathcal{K}_1} \\
&\quad + \underbrace{\int_{\mathcal{E}^c} p_{X_t}(x_t) \left\{ p_{X_{t-1}|X_t}(x_{t-1}|x_t) + p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t) \right\} \|x_{t-1} - \mu_t^*(x_t)\|_2 \|s_t(x_t) - s_t^*(x_t)\|_2 dx_{t-1} dx_t}_{=: \mathcal{K}_2}, \tag{181}
\end{aligned}$$

where (i) holds since $\mathbb{E}[Y_{t-1}^* - \mu_t^*(Y_t) | Y_t] = 0$. This leaves us with two terms to cope with.

- Regarding the term \mathcal{K}_1 , we can bound

$$\begin{aligned}
\mathcal{K}_1 &\stackrel{(i)}{\lesssim} \frac{d^2 \log^3 T}{T} \int_{\mathcal{E}} p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \|x_{t-1} - \mu_t^*(x_t)\|_2 \varepsilon_{\text{score},t}(x_t) dx_{t-1} dx_t \\
&\stackrel{(ii)}{\lesssim} \frac{d^3 \log^3 T}{T} \int_{\mathcal{E}} p_{X_t}(x_t) p_{X_{t-1}|X_t}(x_{t-1}|x_t) \|x_{t-1} - \mu_t^*(x_t)\|_2^2 dx_{t-1} dx_t + \frac{d \log^3 T}{T} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}^2(X_t)] \\
&= \frac{d^3 \log^3 T}{T} \int_{\mathcal{E}} p_{X_t}(x_t) \frac{p_{X_{t-1}|X_t}(x_{t-1}|x_t)}{p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t)} p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t) \|x_{t-1} - \mu_t^*(x_t)\|_2^2 dx_{t-1} dx_t \\
&\quad + \frac{d \log^3 T}{T} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}^2(X_t)] \\
&\stackrel{(iii)}{\lesssim} \frac{d^3 \log^3 T}{T} \left(1 + O\left(\frac{d^2 \log^3 T}{T}\right) \right) \frac{d(1 - \alpha_t)}{\alpha_t} + \frac{d \log^3 T}{T} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}^2(X_t)] + \exp\left(-\frac{c_3}{4} d \log T\right) \\
&\stackrel{(iv)}{\lesssim} \frac{d^4 \log^4 T}{T^2} + \frac{d \log^3 T}{T} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}(X_t)^2]. \tag{182}
\end{aligned}$$

Here, (i) comes from (84), (ii) is due to the elementary inequality $2ab \leq a^2 + b^2$, (iii) invokes the relation (84) as well as the Gaussian distribution of $Y_{t-1}^* | Y_t$ in (76b), whereas (iv) applies (39).

- With regards to the term \mathcal{K}_2 , it follows from the Cauchy-Schwarz inequality that

$$\mathcal{K}_2 \lesssim \sqrt{\mathcal{K}_3 \mathcal{K}_4},$$

where

$$\mathcal{K}_3 := \int_{\mathcal{E}^c} p_{X_t}(x_t) \left\{ p_{X_{t-1}|X_t}(x_{t-1}|x_t) + p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t) \right\} \|x_{t-1} - \mu_t^*(x_t)\|_2^2 dx_{t-1} dx_t; \tag{183a}$$

$$\begin{aligned}
\mathcal{K}_4 &:= \int_{\mathcal{E}^c} p_{X_t}(x_t) \left\{ p_{X_{t-1}|X_t}(x_{t-1}|x_t) + p_{Y_{t-1}^*|Y_t}(x_{t-1}|x_t) \right\} \|s_t(x_t) - s_t^*(x_t)\|_2^2 dx_{t-1} dx_t \\
&\leq 2 \int p_{X_t}(x_t) \|s_t(x_t) - s_t^*(x_t)\|_2^2 dx_t = 2 \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}(X_t)^2]. \tag{183b}
\end{aligned}$$

Consequently, it suffices to look at the term \mathcal{K}_3 . According to (88), we have

$$\begin{aligned} \|x_{t-1} - \mu_t^*(x_t)\|_2^2 &\lesssim \left\| x_{t-1} - \frac{1}{\sqrt{\alpha_t}} x_t \right\|_2^2 + \left(\frac{1 - \alpha_t}{\sqrt{\alpha_t}(1 - \bar{\alpha}_t)} \right)^2 \left(\mathbb{E} \left[\|x_t - \sqrt{\bar{\alpha}_t} X_0\|_2 \mid X_t = x_t \right] \right)^2 \\ &\lesssim \left\| x_{t-1} - \frac{1}{\sqrt{\alpha_t}} x_t \right\|_2^2 + \frac{\log^2 T}{T^2} (\|x_t\|_2^2 + T^{2c_R}), \end{aligned}$$

where the last inequality makes use of (39) and the assumption (21). Combining this with (87) gives

$$\begin{aligned} &\int_{\mathcal{E}^c} p_{X_t}(x_t) p_{X_{t-1} \mid X_t}(x_{t-1} \mid x_t) \|x_{t-1} - \mu_t^*(x_t)\|_2^2 dx_{t-1} dx_t \\ &\lesssim \int_{\mathcal{E}^c} p_{X_t}(x_t) p_{X_{t-1} \mid X_t}(x_{t-1} \mid x_t) \left\| x_{t-1} - \frac{1}{\sqrt{\alpha_t}} x_t \right\|_2^2 dx_{t-1} dx_t \\ &\quad + \frac{\log^2 T}{T^2} \int_{\mathcal{E}^c} p_{X_t}(x_t) p_{X_{t-1} \mid X_t}(x_{t-1} \mid x_t) (\|x_t\|_2^2 + T^{2c_R}) dx_{t-1} dx_t \\ &\lesssim \exp(-c_3 d \log T). \end{aligned}$$

Additionally, let us decompose

$$\begin{aligned} &\int_{\mathcal{E}^c} p_{X_t}(x_t) p_{Y_{t-1}^* \mid Y_t}(x_{t-1} \mid x_t) \|x_{t-1} - \mu_t^*(x_t)\|_2^2 dx_{t-1} dx_t \\ &\leq \underbrace{\int_{(x_t, x_{t-1}): p_{X_t}(x_t) < e^{-0.5c_6 d \log T}} p_{X_t}(x_t) p_{Y_{t-1}^* \mid Y_t}(x_{t-1} \mid x_t) \|x_{t-1} - \mu_t^*(x_t)\|_2^2 dx_{t-1} dx_t}_{=: \mathcal{K}_5} \\ &\quad + \underbrace{\int_{(x_t, x_{t-1}): p_{X_t}(x_t) \geq e^{-0.5c_6 d \log T}, \|x_t - \hat{x}_t\|_2 > c_3 \sqrt{d(1-\alpha_t) \log T}} p_{X_t}(x_t) p_{Y_{t-1}^* \mid Y_t}(x_{t-1} \mid x_t) \|x_{t-1} - \mu_t^*(x_t)\|_2^2 dx_{t-1} dx_t}_{=: \mathcal{K}_6}. \end{aligned}$$

Substitution into (183a) yields

$$\mathcal{K}_2 \lesssim \sqrt{\mathcal{K}_3 \mathcal{K}_4} \lesssim \sqrt{\{\exp(-c_3 d \log T) + \mathcal{K}_5 + \mathcal{K}_6\} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score}, t}(X_t)^2]}, \quad (184)$$

thus motivating us to bound \mathcal{K}_5 and \mathcal{K}_6 separately.

– Regarding \mathcal{K}_6 , we can use (89) to demonstrate that

$$\begin{aligned} \mathcal{K}_6 &\leq \int p_{X_t}(x_t) \mathbb{E} \left[\|Y_{t-1}^* - \mu_t^*(Y_t)\|_2^2 \mathbb{1} \left\{ \|Y_{t-1}^* - \mu_t^*(Y_t)\|_2 \geq \frac{c_3}{2} \sqrt{d(1-\alpha_t) \log T} \right\} \mid Y_t = x_t \right] dx_t \\ &\lesssim \exp\left(-\frac{c_3}{4} d \log T\right). \end{aligned}$$

– Turning to \mathcal{K}_5 , one can invoke (62) to derive

$$\begin{aligned} \mathcal{K}_5 &\leq \int_{x_t: p_{X_t}(x_t) < e^{-0.5c_6 d \log T}} p_{X_t}(x_t) \mathbb{E} \left[\|Y_{t-1}^* - \mu_t^*(Y_t)\|_2^2 \mid Y_t = x_t \right] dx_t \\ &\leq d \int_{x_t: p_{X_t}(x_t) < e^{-0.5c_6 d \log T}} p_{X_t}(x_t) dx_t \\ &\leq d \int_{x_t: p_{X_t}(x_t) < e^{-0.5c_6 d \log T}, \|x_t\|_2 \leq T^{c_R+2}} p_{X_t}(x_t) dx_t + \mathbb{P} \{ \|X_t\|_2 \geq T^{c_R+2} \} \\ &\leq d \exp\left(-\frac{1}{2} c_6 d \log T\right) \cdot (2T^{c_R+2})^d + \exp(-c_6 d \log T) \\ &\lesssim \exp\left(-\frac{1}{4} c_6 d \log T\right), \end{aligned}$$

provided that $c_6 > 0$ is large enough.

Combining the above results with (184), we arrive at

$$\begin{aligned} \mathcal{K}_2 &\lesssim \sqrt{\left\{ \exp\left(-\frac{c_3}{4}d\log T\right) + \exp\left(-\frac{c_6}{4}d\log T\right) \right\} \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}(X_t)^2]} \\ &\lesssim \exp\left(-\frac{\min\{c_3, c_6\}}{4}d\log T\right) + \exp\left(-\frac{\min\{c_3, c_6\}}{4}d\log T\right) \mathbb{E}_{X_t \sim q_t} [\varepsilon_{\text{score},t}(X_t)^2]. \end{aligned} \quad (185)$$

Taking the above bounds (180), (182) and (185) together with (178) immediately finishes the proof.

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:

$$\varphi_t(x) := x - u_t(x), \quad (186a)$$

where

$$\begin{aligned} u_t(x) &:= \left\{ \frac{1 - \alpha_t}{2(1 - \bar{\alpha}_t)} + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^2} - \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^3} \|g_t(x)\|_2^2 \right\} g_t(x) \\ &\quad + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^3} \mathbb{E} \left[(X_t - \sqrt{\bar{\alpha}_t} X_0)(X_t - \sqrt{\bar{\alpha}_t} X_0)^\top \mid X_t = x \right] g_t(x) \\ &\quad - \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^3} \mathbb{E} \left[\|X_t - \sqrt{\bar{\alpha}_t} X_0\|_2^2 (X_t - \sqrt{\bar{\alpha}_t} X_0 - g_t(X_t)) \mid X_t = x \right] \end{aligned} \quad (186b)$$

$$\begin{aligned} &= - \left\{ \frac{1 - \alpha_t}{2} + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)} - \frac{(1 - \alpha_t)^2}{8} \|s_t^*(x)\|_2^2 \right\} s_t^*(x) - \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)} \mathbb{E} \left[\bar{W}_t \bar{W}_t^\top s_t^*(X_t) \mid X_t = x \right] \\ &\quad - \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)} \mathbb{E} \left[\|\bar{W}_t\|_2^2 \left(\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \bar{W}_t + (1 - \bar{\alpha}_t) s_t^*(X_t) \right) \mid X_t = x \right]. \end{aligned} \quad (186c)$$

Here, we recall that $g_t(x) = -(1 - \bar{\alpha}_t)s_t^*(x)$ has been defined in (36) and that $X_t = \sqrt{\bar{\alpha}_t}X_0 + \sqrt{1 - \bar{\alpha}_t}\bar{W}_t$ with $\bar{W}_t \sim \mathcal{N}(0, I_d)$ (see (6)). In view of Assumption 3 and the fact that the MMSE estimator is the conditional expectation (Hajek, 2015, Section 3.3.1), we have

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

which taken together with (186) and (27) confirms the following equivalent expression for the sampler (27):

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \varphi_t(Y_t) = \frac{1}{\sqrt{\alpha_t}} Y_t - \frac{1}{\sqrt{\alpha_t}} u_t(Y_t). \quad (187)$$

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

$$\|u_t(x_t)\|_2 \lesssim (1 - \alpha_t) \left(\frac{d \log T}{1 - \bar{\alpha}_t} \right)^{1/2}, \quad (188a)$$

and one can also write

$$u_t(x_t) = \frac{1 - \alpha_t}{2(1 - \bar{\alpha}_t)} g_t(x_t) + \xi_t(x_t) \quad (188b)$$

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

$$\|\xi_t(x_t)\|_2 \lesssim (1 - \alpha_t)^2 \left(\frac{d \log T}{1 - \bar{\alpha}_t} \right)^{3/2}. \quad (188c)$$

To streamline presentation, we leave the proof of (188) 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. (27)), or equivalently, φ_t (cf. (186)). There are several objects that play an important role in the analysis, which we single out as follows:

$$A_t := \frac{1}{1 - \bar{\alpha}_t} \int p_{X_0 | X_t}(x_0 | x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 dx_0; \quad (189a)$$

$$B_t := \frac{1}{1 - \bar{\alpha}_t} \left\| \int p_{X_0 | X_t}(x_0 | x) (x - \sqrt{\bar{\alpha}_t} x_0) dx_0 \right\|_2^2; \quad (189b)$$

$$C_t := \frac{1}{(1 - \bar{\alpha}_t)^2} \int p_{X_0 | X_t}(x_0 | x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^4 dx_0; \quad (189c)$$

$$D_t := \frac{1}{(1 - \bar{\alpha}_t)^2} \int p_{X_0 | X_t}(x_0 | x) \left(\langle g_t(x), x - \sqrt{\bar{\alpha}_t} x_0 \rangle \right)^2 dx_0; \quad (189d)$$

$$E_t := \frac{1}{(1 - \bar{\alpha}_t)^2} \int p_{X_0 | X_t}(x_0 | x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 \langle g_t(x), x - \sqrt{\bar{\alpha}_t} x_0 \rangle dx_0. \quad (189e)$$

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 (39), we have the following bounds:

$$|B_t| \leq |A_t| \lesssim \frac{1}{1 - \bar{\alpha}_t} \cdot d(1 - \bar{\alpha}_t) \log T \asymp d \log T; \quad (190a)$$

$$|C_t| \lesssim \frac{1}{(1 - \bar{\alpha}_t)^2} d^2 (1 - \bar{\alpha}_t)^2 \log^2 T \asymp d^2 \log^2 T; \quad (190b)$$

$$|D_t| \leq \frac{\|g_t(x)\|_2^2}{(1 - \bar{\alpha}_t)^2} \int p_{X_0 | X_t}(x_0 | x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 dx_0 \lesssim d^2 \log^2 T; \quad (190c)$$

$$|E_t| \leq \frac{\|g_t(x)\|_2^2}{(1 - \bar{\alpha}_t)^2} \int p_{X_0 | X_t}(x_0 | x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^3 dx_0 \lesssim d^2 \log^2 T. \quad (190d)$$

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 (58) in Lemma 4 with (191) in the lemma below.

Lemma 11. *Suppose that $\frac{d^2(1-\alpha_t)\log T}{\alpha_t - \bar{\alpha}_t} \lesssim 1$. For every $x \in \mathbb{R}$ obeying $-\log p_{X_t}(x) \leq c_6 d \log T$ for some large enough constant $c_6 > 0$, we have*

$$\begin{aligned} \frac{p_{\sqrt{\bar{\alpha}_t} X_{t-1}}(\varphi_t(x))}{p_{X_t}(x)} &= 1 + \frac{(1 - \alpha_t)(d + B_t - A_t)}{2(1 - \bar{\alpha}_t)} + O\left(d^3 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^3 \log^3 T\right) \\ &\quad + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^2} [d(d + 2) + (4 + 2d)(B_t - A_t) - B_t^2 + C_t + 2D_t - 3E_t + A_t B_t]. \end{aligned} \quad (191a)$$

Moreover, for any random vector Y , one has

$$\begin{aligned} \frac{p_{\varphi_t(Y)}(\varphi_t(x))}{p_Y(x)} &= 1 + \frac{(1 - \alpha_t)(d + B_t - A_t)}{2(1 - \bar{\alpha}_t)} + O\left(d^6 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^3 \log^3 T\right) \\ &\quad + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^2} [d(d + 2) + (4 + 2d)(B_t - A_t) - B_t^2 + C_t + 2D_t - 3E_t + A_t B_t]. \end{aligned} \quad (191b)$$

Here, the quantities A_t, \dots, E_t are defined in (189).

The proof of this lemma can be found in Appendix D.2. Crucially, the terms in (191a) and those in (191b) coincide (except for the residual terms $O(d^3(\frac{1-\alpha_t}{\alpha_t-\bar{\alpha}_t})^3 \log^3 T)$ and $O(d^6(\frac{1-\alpha_t}{\alpha_t-\bar{\alpha}_t})^3 \log^3 T)$), which will cancel each other during the subsequent proof.

With Lemma 11 in place, we claim that for every $t \geq 1$,

$$\mathbb{P}(Y_t \in \mathcal{E}_t) \geq 1 - (T - t + 1) \exp(-c_3 d \log T) \quad (192)$$

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

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

for some constant $c_5 > 0$. We leave its proof to the end of this section.

Suppose for the moment that the above claim (192) is valid. Then taking $t = 1$ leads to

$$\mathbb{P}(Y_1 \in \mathcal{E}_1) = \mathbb{P} \left(\left| \frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right| \leq \frac{c_5 d^6 \log^6 T}{T^2} \right) \geq 1 - T \exp(-c_3 d \log T),$$

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

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

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

$$\begin{aligned} \text{TV}(q_1, p_1) &= \frac{1}{2} \int_{y \in \mathcal{E}_1} |q_1(y) - p_1(y)| dy + \frac{1}{2} \int_{y \notin \mathcal{E}_1} |q_1(y) - p_1(y)| dy \\ &\leq \int_{y \in \mathcal{E}_1} |q_1(y) - p_1(y)| dy + \int_{y \notin \mathcal{E}_1} p_1(y) dy \\ &= \mathbb{E}_{Y_1 \sim p_1} \left[\left| \frac{q_1(Y_1)}{p_1(Y_1)} - 1 \right| \cdot \mathbf{1}\{Y_1 \in \mathcal{E}_1\} \right] + \int_{y \notin \mathcal{E}_1} p_1(y) dy \\ &\leq \frac{c_5 d^6 \log^6 T}{T^2} + \exp(-c_3 d \log T) \\ &\asymp \frac{d^6 \log^6 T}{T^2}. \end{aligned} \quad (193)$$

This establishes the advertised result in Theorem 2, provided that Claim (192) can be verified.

Proof of properties (188). To justify the above results (188), note that Lemma 1 implies that

$$\|g_t(x_t)\|_2 \leq \mathbb{E} \left[\|X_t - \sqrt{\bar{\alpha}_t} X_0\|_2 \mid X_t = x_t \right] \lesssim \sqrt{d(1 - \bar{\alpha}_t) \log T}, \quad (194a)$$

$$\frac{1 - \alpha_t}{1 - \bar{\alpha}_t} \|g_t(x_t)\|_2 \lesssim \frac{1 - \alpha_t}{1 - \bar{\alpha}_t} \sqrt{d(1 - \bar{\alpha}_t) \log T} \asymp (1 - \alpha_t)^2 \sqrt{\frac{d \log T}{1 - \bar{\alpha}_t}}, \quad (194b)$$

$$\frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^2} \|g_t(x_t)\|_2 \lesssim \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^2} \sqrt{d(1 - \bar{\alpha}_t) \log T} \asymp (1 - \alpha_t)^2 \frac{\sqrt{d \log T}}{(1 - \bar{\alpha}_t)^{3/2}}, \quad (194c)$$

$$\frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^3} \|g_t(x_t)\|_2^3 \lesssim \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^3} (d(1 - \bar{\alpha}_t) \log T)^{3/2} \asymp (1 - \alpha_t)^2 \left(\frac{d \log T}{1 - \bar{\alpha}_t} \right)^{3/2}, \quad (194d)$$

$$\frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^3} \left\| \mathbb{E} \left[(x_t - \sqrt{\bar{\alpha}_t} X_0)(x_t - \sqrt{\bar{\alpha}_t} X_0)^\top \mid X_t = x_t \right] g_t(x_t) \right\|_2 \lesssim \frac{(1 - \alpha_t)^2 \|g_t(x_t)\|_2}{(1 - \bar{\alpha}_t)^3} \mathbb{E} \left[\|x_t - \sqrt{\bar{\alpha}_t} X_0\|_2^2 \mid X_t = x_t \right]$$

$$\lesssim (1 - \alpha_t)^2 \left(\frac{d \log T}{1 - \bar{\alpha}_t} \right)^{3/2}, \quad (194e)$$

and

$$\begin{aligned} & \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^3} \left\| \mathbb{E} \left[\left\| x_t - \sqrt{\bar{\alpha}_t} X_0 \right\|_2^2 (x_t - \sqrt{\bar{\alpha}_t} x_0 - g_t(X_t)) \mid X_t = x_t \right] \right\|_2 \\ & \leq \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^3} \left\{ \left\| \mathbb{E} \left[\left\| x_t - \sqrt{\bar{\alpha}_t} X_0 \right\|_2^3 \mid X_t = x_t \right] \right\|_2 + \|g_t(x_t)\|_2 \left\| \mathbb{E} \left[\left\| x_t - \sqrt{\bar{\alpha}_t} X_0 \right\|_2^2 \mid X_t = x_t \right] \right\|_2 \right\} \\ & \lesssim (1 - \alpha_t)^2 \left(\frac{d \log T}{1 - \bar{\alpha}_t} \right)^{3/2}. \end{aligned} \quad (194f)$$

Substituting the bounds (194) into (186b) immediately establishes (188). \square

Proof of the claim (192). We would like to prove this claim by induction, for which we start with the base case with $t = T$. Recall that $X_T \stackrel{d}{=} \sqrt{\bar{\alpha}_T} X_0 + \sqrt{1 - \bar{\alpha}_T} B$ and $Y_T \stackrel{d}{=} B$ with $B \sim \mathcal{N}(0, I_d)$ independent of X_0 , and that $\|X_0\|_2 \leq R$ with $R = T^{c_R}$ for some constant $c_R > 0$. For large enough T , it immediately follows from (39d) that

$$\mathbb{P}(Y_T \in \mathcal{E}_t) \geq 1 - \exp(-c_3 d \log T) \quad (195)$$

for some constant $c_3 > 0$ large enough.

Suppose now that the claim (192) holds for some $t \geq 2$, 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

$$q_t(Y_t) \geq \exp(-c_6 d \log T). \quad (196)$$

With (196) in place, one sees that Lemma 11 is applicable to $x = Y_t$ with high probability.

Next, consider any y obeying $-\log p_{X_t}(y) \leq c_6 d \log T$. With these bounds (190) in mind, applying relations (191a) and (191b) in Lemma 11 leads to

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

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 (55), we immediately arrive at

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

with probability exceeding $1 - (T - t) \exp(-c_3 d \log T)$. This concludes the proof of Claim (192) via standard induction arguments. \square

D.2 Proof of Lemma 11

The proof of Lemma 11 is derived in a very similar way to the proof of Lemma 4, as detailed below. For notational simplicity, we shall abbreviate

$$u = u_t(x) \quad \text{and} \quad z = g_t(x) \quad (197)$$

whenever it is clear from the context.

D.2.1 Proof of relation (191a)

Through direct calculations as shown in (127), we can obtain

$$p_{\sqrt{\alpha_t}X_{t-1}}(\varphi_t(x)) = p_{X_t}(x) \left(\frac{1 - \bar{\alpha}_t}{\alpha_t - \bar{\alpha}_t} \right)^{d/2} \cdot \int_{x_0} p_{X_0|X_t}(x_0|x) \exp \left(- \frac{(1 - \alpha_t) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{2(\alpha_t - \bar{\alpha}_t)} \right) dx_0. \quad (198)$$

Moreover, similar to the analysis of Lemma 4, we focus our attention on the set given x :

$$\mathcal{E} := \{x_0 : \|x - \sqrt{\bar{\alpha}_t}x_0\|_2 \leq 5c_5 \sqrt{d(1 - \bar{\alpha}_t) \log T}\}, \quad (199)$$

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

$$\begin{aligned} & \int p_{X_0|X_t}(x_0|x) \exp \left(- \frac{(1 - \alpha_t) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{2(\alpha_t - \bar{\alpha}_t)} \right) dx_0 \\ &= O(\exp(-c_8 c_5^2 d \log T)) + \int_{x_0 \in \mathcal{E}} p_{X_0|X_t}(x_0|x) \exp \left(- \frac{(1 - \alpha_t) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{2(\alpha_t - \bar{\alpha}_t)} \right) dx_0 \\ &=: \text{RHS} \end{aligned} \quad (200)$$

To further control the right-hand side above, recall that the learning rates are selected such that $\frac{1 - \alpha_t}{1 - \bar{\alpha}_t - 1} \leq \frac{4c_1 \log T}{T}$ for $1 < t \leq T$ (see (39b)). In view of the Taylor expansion $e^{-x} = 1 - x + \frac{1}{2}x^2 + O(x^3)$ for $x \leq 1/2$, we can derive

$$\begin{aligned} \text{RHS} &= O(\exp(-c_8 c_5^2 d \log T)) \\ &+ \int_{x_0 \in \mathcal{E}} p_{X_0|X_t}(x_0|x) \left\{ 1 - \frac{(1 - \alpha_t) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2}{2(\alpha_t - \bar{\alpha}_t)(1 - \bar{\alpha}_t)} - \frac{\frac{(1 - \alpha_t)^2}{4(1 - \bar{\alpha}_t)^2} \|z\|_2^2 - 2u^\top(x - \sqrt{\bar{\alpha}_t}x_0)}{2(\alpha_t - \bar{\alpha}_t)} \right. \\ &\left. + \frac{(1 - \alpha_t)^2}{8(\alpha_t - \bar{\alpha}_t)^2(1 - \bar{\alpha}_t)^2} \left(\|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 - z^\top(x - \sqrt{\bar{\alpha}_t}x_0) \right)^2 + O\left(d^3 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^3 \log^3 T \right) \right\} dx_0. \end{aligned} \quad (201)$$

In order to see this, we recall the property of u (cf. (188)) as

$$\left\| u - \frac{1 - \alpha_t}{2(1 - \bar{\alpha}_t)} z \right\|_2 \leq O\left((1 - \alpha_t)^2 \left(\frac{d \log T}{1 - \bar{\alpha}_t} \right)^{3/2} \right). \quad (202)$$

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

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

and

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

where we have invoked the properties (39). Taking the above results together and using the following basic properties regarding quantities A_t, \dots, E_t (defined in (189))

$$\int p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0 = (1 - \bar{\alpha}_t) A_t,$$

$$\begin{aligned}
\int p_{X_0|X_t}(x_0|x) \|z\|_2^2 dx_0 &= (1 - \bar{\alpha}_t) B_t, \\
\int p_{X_0|X_t}(x_0|x) u^\top (x - \sqrt{\bar{\alpha}_t} x_0) dx_0 &= \frac{1 - \alpha_t}{2} B_t + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)} [B_t - B_t^2 + D_t - E_t + A_t B_t], \\
\int p_{X_0|X_t}(x_0|x) \left(\|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 - z^\top (x - \sqrt{\bar{\alpha}_t} x_0) \right)^2 dx_0 &= (1 - \bar{\alpha}_t)^2 [C_t + D_t - 2E_t],
\end{aligned}$$

we arrive at

$$(201) = 1 - \frac{(1 - \alpha_t)(A_t - B_t)}{2(\alpha_t - \bar{\alpha}_t)} + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^2} [-B_t^2 + C_t + 2D_t - 3E_t + A_t B_t] + O\left(d^3 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t}\right)^3 \log^3 T\right).$$

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 $O(\exp(-c_8 c_5^2 d \log T))$. Putting all this together establishes the advertised result (191a).

D.2.2 Proof of relation (191b)

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

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

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

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

$$\det(I + A)^{-1} = 1 - \text{Tr}(A) + \frac{1}{2} [\text{Tr}(A)^2 + \|A\|_F^2] + O(d^3 \|A\|^3),$$

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

$$\begin{aligned}
p_{\varphi_t(Y)}(\varphi_t(x)) &= \det\left(\frac{\partial \varphi_t(x)}{\partial x}\right)^{-1} p_Y(x) \\
&= \left\{ 1 + \text{Tr}\left(\frac{\partial u}{\partial x}\right) + \frac{1}{2} \left[\text{Tr}\left(\frac{\partial u}{\partial x}\right)^2 + \left\| \frac{\partial u}{\partial x} \right\|_F^2 \right] + O\left(d^3 \left\| \frac{\partial u}{\partial x} \right\| \right) \right\} p_Y(x), \tag{203}
\end{aligned}$$

where we invoke the definition in (186) that

$$\begin{aligned}
\varphi_t(x) = x - u &= x - \left(\frac{(1 - \alpha_t)}{2(1 - \bar{\alpha}_t)} + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^2} - \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^3} \|z\|_2^2 \right) z \\
&\quad - \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^3} \int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t} x_0) (x - \sqrt{\bar{\alpha}_t} x_0)^\top z dx_0 \\
&\quad + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^3} \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 (x - \sqrt{\bar{\alpha}_t} x_0 - z) dx_0.
\end{aligned}$$

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

$$\frac{\partial z}{\partial x} = J_t, \tag{204a}$$

$$\frac{\partial}{\partial x} \|z\|_2^2 z = \|z\|_2^2 J_t + 2z z^\top J_t, \tag{204b}$$

$$\frac{\partial}{\partial x} \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t} x_0\|_2^2 (x - \sqrt{\bar{\alpha}_t} x_0) dx_0$$

$$\begin{aligned}
&= \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0 I + 2 \int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t}x_0)(x - \sqrt{\bar{\alpha}_t}x_0)^\top dx_0 \\
&\quad + \frac{1}{1 - \bar{\alpha}_t} \left(\left(\int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 (x - \sqrt{\bar{\alpha}_t}x_0) \right) \left(\int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t}x_0) dx_0 \right)^\top \right. \\
&\quad \left. - \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 (x - \sqrt{\bar{\alpha}_t}x_0)(x - \sqrt{\bar{\alpha}_t}x_0)^\top dx_0 \right), \tag{204c}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial x} \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 z &= \int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 dx_0 J_t + 2zz^\top \\
&+ \frac{1}{1 - \bar{\alpha}_t} \left(\left(\int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 \right) zz^\top - z \left(\int_{x_0} p_{X_0|X_t}(x_0|x) \|x - \sqrt{\bar{\alpha}_t}x_0\|_2^2 (x - \sqrt{\bar{\alpha}_t}x_0) dx_0 \right)^\top \right), \tag{204d}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial x} \int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t}x_0)(x - \sqrt{\bar{\alpha}_t}x_0)^\top z dx_0 &= \|z\|_2^2 I + zz^\top \\
&+ \int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t}x_0)(x - \sqrt{\bar{\alpha}_t}x_0)^\top J_t dx_0 \\
&+ \frac{1}{1 - \bar{\alpha}_t} \int_{x_0} p_{X_0|X_t}(x_0|x) (z^\top (x - \sqrt{\bar{\alpha}_t}x_0))(x - \sqrt{\bar{\alpha}_t}x_0) z^\top dx_0 \\
&- \frac{1}{1 - \bar{\alpha}_t} \int_{x_0} p_{X_0|X_t}(x_0|x) (z^\top (x - \sqrt{\bar{\alpha}_t}x_0))(x - \sqrt{\bar{\alpha}_t}x_0)(x - \sqrt{\bar{\alpha}_t}x_0)^\top dx_0. \tag{204e}
\end{aligned}$$

Equipped with the above relations, we can easily verify that

$$\left\| \frac{\partial u}{\partial x} \right\| \lesssim \frac{d(1 - \alpha_t) \log T}{1 - \bar{\alpha}_t}, \tag{205a}$$

$$\begin{aligned}
\text{Tr} \left(\frac{\partial u}{\partial x} \right) &= \frac{(1 - \alpha_t)(d + B_t - A_t)}{2(1 - \bar{\alpha}_t)} \\
&\quad + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^2} (d - 2A_t - A_t^2 + 3A_t B_t + 2B_t - 3B_t^2 + C_t + 4D_t - 3E_t - F_t), \tag{205b}
\end{aligned}$$

$$\begin{aligned}
\left\| \frac{\partial u}{\partial x} \right\|_F^2 &= \frac{(1 - \alpha_t)^2}{4(1 - \bar{\alpha}_t)^2} \left\| \frac{\partial z}{\partial x} \right\|_F^2 + O \left(d^5 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^3 \log^3 T \right) \\
&= \frac{(1 - \alpha_t)^2}{4(1 - \bar{\alpha}_t)^2} (d + 2(B_t - A_t) + B_t^2 + F_t - 2D_t) + O \left(d^5 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^3 \log^3 T \right), \tag{205c}
\end{aligned}$$

as long as $d^2 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right) \log T \lesssim 1$, where we recall the definition of the quantities A_t to E_t in (189), and

$$F_t(x) := \left\| \frac{1}{1 - \bar{\alpha}_t} \int_{x_0} p_{X_0|X_t}(x_0|x) (x - \sqrt{\bar{\alpha}_t}x_0)(x - \sqrt{\bar{\alpha}_t}x_0)^\top dx_0 \right\|_F^2. \tag{205d}$$

Plugging these results into inequality (203) leads to

$$\begin{aligned}
p_{\varphi_t(Y)}(\varphi_t(x)) &= p_Y(x) \left\{ 1 + \frac{(1 - \alpha_t)(d + B_t - A_t)}{2(1 - \bar{\alpha}_t)} + O \left(d^6 \left(\frac{1 - \alpha_t}{\alpha_t - \bar{\alpha}_t} \right)^3 \log^3 T \right) \right. \\
&\quad \left. + \frac{(1 - \alpha_t)^2}{8(1 - \bar{\alpha}_t)^2} [d(d + 2) + (4 + 2d)(B_t - A_t) - B_t^2 + C_t + 2D_t - 3E_t + A_t B_t] \right\}. \tag{205e}
\end{aligned}$$

We have thus completed the proof of Lemma 11.

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 $\hat{x}_t := x_t / \sqrt{\alpha_t}$ 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 (37). In view of the expression (38) as well as (6) (i.e., $X_t - \sqrt{\alpha_t}X_0 = \sqrt{1 - \alpha_t}\bar{W}_t$ with $\bar{W}_t \sim \mathcal{N}(0, I_d)$), one can write

$$\begin{aligned} J_t(x) &= I_d + \mathbb{E}[\bar{W}_t | X_t = x] \left(\mathbb{E}[\bar{W}_t | X_t = x] \right)^\top - \mathbb{E}[\bar{W}_t \bar{W}_t^\top | X_t = x] \\ &= I_d + (1 - \alpha_t) s_t^*(x) s_t^*(x)^\top - \mathbb{E}[\bar{W}_t \bar{W}_t^\top | X_t = x], \end{aligned} \quad (206)$$

where the last line makes use of the relation (36). Additionally, recall that (i) $v_t(x, z)$ (cf. (32)) is the MMSE estimator for estimating $\bar{W}_t \bar{W}_t^\top z$ given $\sqrt{\alpha_t}X_0 + \sqrt{1 - \alpha_t}\bar{W}_t = x$ and $Z_t = z$, and (ii) Z_t is independent from X_0 and \bar{W}_t . Then this MMSE estimator admits the following expression (Hajek, 2015, Section 3.3.1):

$$v_t(x, z) = \mathbb{E}[\bar{W}_t \bar{W}_t^\top z | X_t = x] = \mathbb{E}[\bar{W}_t \bar{W}_t^\top | X_t = x] z. \quad (207)$$

As a result, the mapping introduced in (33b) can be alternatively expressed as:

$$\begin{aligned} \Psi_t(x, z) &= \frac{1}{\sqrt{\alpha_t}} \left(x + (1 - \alpha_t) s_t^*(x) \right) + \sigma_t \left\{ z - \frac{1 - \alpha_t}{2(1 - \alpha_t)} \left[z + (1 - \alpha_t) s_t^*(x) s_t^*(x)^\top z - v_t(x, z) \right] \right\} \\ &= \mu_t(x) + \sigma_t \left(I - \frac{1 - \alpha_t}{2(1 - \alpha_t)} J_t(x) \right) z, \end{aligned} \quad (208)$$

where we have also used the definition (74b) of $\mu_t(\cdot)$. In comparison to the plain DDPM-type sampler (cf. (30)), the key correction term is the second component on the right-hand side of (208), which adjusts the covariance of the additive Gaussian noise.

Equipped with the above expression (208), we can readily express the conditional distribution of Y_{t-1} (cf. (33a)) given Y_t such that: for any points $x_t, x_{t-1} \in \mathbb{R}^d$,

$$\begin{aligned} p_{Y_{t-1} | Y_t}(x_{t-1} | x_t) &= \frac{1}{(2\pi \frac{1 - \alpha_t}{\alpha_t})^{d/2} \left| \det \left(I - \frac{1 - \alpha_t}{2(1 - \alpha_t)} J_t(x_t) \right) \right|} \\ &\quad \cdot \exp \left(- \frac{\alpha_t}{2(1 - \alpha_t)} \left\| \left(I - \frac{1 - \alpha_t}{2(1 - \alpha_t)} J_t(x_t) \right)^{-1} (x_{t-1} - \mu_t(x_t)) \right\|_2^2 \right). \end{aligned} \quad (209)$$

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:

$$\mathcal{E} := \left\{ (x_t, x_{t-1}) \mid -\log p_{X_t}(x_t) \leq \frac{1}{2} c_6 d \log T, \|x_{t-1} - \hat{x}_t\|_2 \leq c_3 \sqrt{d(1 - \alpha_t) \log T} \right\}, \quad (210)$$

with the numerical constants $c_3, c_6 > 0$ introduced in Lemma 2. The following lemma, which is a counterpart of Lemma 8 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 12. *There exists some large enough numerical constant $c_\zeta > 0$ such that: for every $(x_t, x_{t-1}) \in \mathcal{E}$,*

$$\begin{aligned} p_{X_{t-1} | X_t}(x_{t-1} | x_t) &= \frac{1}{(2\pi \frac{1 - \alpha_t}{\alpha_t})^{d/2} \left| \det \left(I - \frac{1 - \alpha_t}{2(1 - \alpha_t)} J_t(x_t) \right) \right|} \\ &\quad \cdot \exp \left(- \frac{\alpha_t}{2(1 - \alpha_t)} \left\| \left(I - \frac{1 - \alpha_t}{2(1 - \alpha_t)} J_t(x_t) \right)^{-1} (x_{t-1} - \mu_t(x_t)) \right\|_2^2 + \zeta_t(x_{t-1}, x_t) \right) \end{aligned} \quad (211)$$

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

$$|\zeta_t(x_{t-1}, x_t)| \leq c_\zeta \frac{d^3 \log^{4.5} T}{T^{3/2}}. \quad (212)$$

Here, we recall the definition of $\mu_t(\cdot)$ (resp. $J_t(\cdot)$) in (74b) (resp. (37)).

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

$$\log \frac{p_{X_{t-1}|X_t}(x_{t-1}|x_t)}{p_{Y_{t-1}|Y_t}(x_{t-1}|x_t)} \leq T^{c_0+2c_R+2} \left\{ \|x_{t-1} - \hat{x}_t\|_2^2 + \|x_t\|_2^2 + 1 \right\}, \quad (213)$$

where c_0 is defined in (22).

Step 3: bounding the KL divergence of interest. With Lemmas 12-13 in place, one can repeat the arguments in Step 3 in the proof of Theorem 3 to arrive at

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

Substitution into (79) and (78) then yields

$$2\text{TV}(p_{X_1}, p_{Y_1})^2 \leq \text{KL}(p_{X_1} \parallel p_{Y_1}) \lesssim \text{KL}(p_{X_T} \parallel p_{Y_T}) + \sum_{t \geq 2}^T \frac{d^6 \log^9 T}{T^3} \asymp \frac{d^6 \log^9 T}{T^2},$$

where the last relation results from (122). This completes the proof of Theorem 4.

E.2 Proof of Lemma 12

Recall that an explicit expression for $p_{X_{t-1}|X_t}(x_{t-1}|x_t)$ has already been established in Lemma 8 (see (160)) A little algebra then allows one to write

$$p_{X_{t-1}|X_t}(x_{t-1}|x_t) = f_1(x_t) \exp \left(-f_2(x_t, x_{t-1}) + \zeta_{t,1}(x_t, x_{t-1}) \right), \quad (214)$$

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

$$g_{t-1}(x) := \int_{x_0} (x - \sqrt{\bar{\alpha}_{t-1}}x_0) p_{X_0|X_{t-1}}(x_0|x) dx_0, \quad (215a)$$

$$f_2(x_t, x_{t-1}) = \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_{t-1}\|_2^2}{2(1-\alpha_t)} + \frac{(x_{t-1} - \hat{x}_t)^\top g_{t-1}(\hat{x}_t)}{1 - \bar{\alpha}_{t-1}} + \frac{\frac{1}{2}(x_{t-1} - \hat{x}_t)^\top J_{t-1}(\hat{x}_t)(x_{t-1} - \hat{x}_t)}{1 - \bar{\alpha}_{t-1}}, \quad (215b)$$

$$\zeta_{t,1}(x_t, x_{t-1}) = (x_{t-1} - \hat{x}_t)^\top \frac{\int_0^1 \int_0^1 \gamma \left[J_{t-1} \left((1-\tau)\hat{x}_t + \tau x_t(\gamma) \right) - J_{t-1}(\hat{x}_t) \right] d\tau d\gamma}{1 - \bar{\alpha}_{t-1}} (x_{t-1} - \hat{x}_t), \quad (215c)$$

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

In order to control (214), we single out two useful facts: for every $(x_t, x_{t-1}) \in \mathcal{E}$ (cf. (210)),

$$\|J_{t-1}(x_t(\gamma)) - J_{t-1}(\hat{x}_t)\| \lesssim d^2 \sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}} \log^2 T, \quad \forall \gamma \in [0, 1] \quad (216)$$

and

$$\left\| \frac{J_{t-1}(\hat{x}_t)}{1 - \bar{\alpha}_{t-1}} - \frac{J_t(x_t)}{1 - \bar{\alpha}_t} \right\| \lesssim \frac{d^2(1-\alpha_t) \log^2 T}{(\alpha_t - \bar{\alpha}_t)^2}. \quad (217)$$

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

$$|\zeta_{t,1}(x_t, x_{t-1})| \leq \frac{\|x_{t-1} - \hat{x}_t\|_2^2}{2(1-\bar{\alpha}_{t-1})} \sup_{\gamma \in [0,1]} \|J_{t-1}(x_t(\gamma)) - J_{t-1}(\hat{x}_t)\| \lesssim d^3 \left(\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}} \right)^{3/2} \log^3 T \quad (218)$$

and

$$\begin{aligned} & \left| \frac{(x_{t-1} - \hat{x}_t)^\top J_{t-1}(\hat{x}_t)(x_{t-1} - \hat{x}_t)}{1 - \bar{\alpha}_{t-1}} - \frac{(x_{t-1} - \hat{x}_t)^\top J_t(x_t)(x_{t-1} - \hat{x}_t)}{1 - \bar{\alpha}_t} \right| \\ & \leq \|x_{t-1} - \hat{x}_t\|_2^2 \left\| \frac{J_{t-1}(\hat{x}_t)}{1 - \bar{\alpha}_{t-1}} - \frac{J_t(x_t)}{1 - \bar{\alpha}_t} \right\| \lesssim \frac{d^3(1 - \alpha_t)^2 \log^3 T}{(\alpha_t - \bar{\alpha}_t)^2} \asymp \frac{d^3(1 - \alpha_t)^2 \log^3 T}{(1 - \bar{\alpha}_{t-1})^2} \end{aligned} \quad (219)$$

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

$$p_{X_{t-1}|X_t}(x_{t-1}|x_t) = f_3(x_t) \exp\left(-f_4(x_t, x_{t-1}) + \zeta_{t,2}(x_t, x_{t-1})\right), \quad (220a)$$

where

$$f_4(x_t, x_{t-1}) = \frac{\alpha_t \|\hat{x}_t - x_{t-1}\|_2^2}{2(1 - \alpha_t)} + \frac{(x_{t-1} - \hat{x}_t)^\top g_{t-1}(\hat{x}_t)}{1 - \bar{\alpha}_{t-1}} + \frac{\frac{1}{2}(x_{t-1} - \hat{x}_t)^\top J_t(x_t)(x_{t-1} - \hat{x}_t)}{1 - \bar{\alpha}_t}, \quad (220b)$$

$$|\zeta_{t,2}(x_t, x_{t-1})| \lesssim d^3 \left(\frac{1 - \alpha_t}{1 - \bar{\alpha}_{t-1}} \right)^{3/2} \log^3 T \lesssim \frac{d^3 \log^{4.5} T}{T^{3/2}}. \quad (220c)$$

To continue, we further observe that

$$\left| \frac{(x_{t-1} - \hat{x}_t)^\top g_{t-1}(\hat{x}_t)}{1 - \bar{\alpha}_{t-1}} - \frac{\sqrt{\alpha_t}(x_{t-1} - \hat{x}_t)^\top g_t(x_t)}{1 - \bar{\alpha}_t} \right| \lesssim \frac{d^3 \log^{3.5} T}{T^{3/2}}, \quad (221)$$

which is an immediate consequence of the following two bounds (obtained using (161c), (161b) and (39b)):

$$\begin{aligned} & \left| \frac{(x_{t-1} - \hat{x}_t)^\top g_{t-1}(\hat{x}_t)}{1 - \bar{\alpha}_{t-1}} - \frac{(x_{t-1} - \hat{x}_t)^\top g_t(x_t)}{1 - \bar{\alpha}_t} \right| \lesssim d^2 \sqrt{\frac{(1 - \alpha_t)^3}{(\alpha_t - \bar{\alpha}_t)^3}} \log^2 T \lesssim \frac{d^3 \log^{3.5} T}{T^{3/2}} \\ & \left| (1 - \sqrt{\alpha_t}) \frac{(x_{t-1} - \hat{x}_t)^\top g_t(x_t)}{1 - \bar{\alpha}_t} \right| \lesssim \frac{d \log^2 T}{T^{3/2}}. \end{aligned}$$

This bound (221) allows us to replace the second term on the right-hand side of (220b) with $\frac{\sqrt{\alpha_t}(x_{t-1} - \hat{x}_t)^\top g_t(x_t)}{1 - \bar{\alpha}_t}$. It is also seen from (161a) and the properties (39) that

$$\left\| \frac{1 - \alpha_t}{2(1 - \bar{\alpha}_t)} J_t(x_t) \right\| \lesssim \frac{\log T}{T} \cdot d \log T \asymp \frac{d \log^2 T}{T} = o(1), \quad (222)$$

and therefore,

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

These combined with (220) allow us to show that: there exist some functions $f_5(\cdot)$ and $\tilde{f}_5(\cdot)$ such that

$$p_{X_{t-1}|X_t}(x_{t-1}|x_t) = f_5(x_t) \exp\left(-f_6(x_t, x_{t-1}) + \zeta_{t,3}(x_t, x_{t-1})\right) \quad (223a)$$

where

$$\begin{aligned} f_6(x_t, x_{t-1}) &= \frac{\alpha_t \|x_{t-1} - \hat{x}_t\|_2^2}{2(1 - \alpha_t)} + \frac{\sqrt{\alpha_t}(x_{t-1} - \hat{x}_t)^\top g_t(x_t)}{1 - \bar{\alpha}_t} + \frac{\frac{1}{2}\alpha_t(x_{t-1} - \hat{x}_t)^\top J_t(x_t)(x_{t-1} - \hat{x}_t)}{1 - \bar{\alpha}_t} \\ &= \frac{\alpha_t}{2(1 - \alpha_t)} \left\{ \|x_{t-1} - \mu_t(x_t)\|_2^2 + \frac{1 - \alpha_t}{1 - \bar{\alpha}_t} (x_{t-1} - \hat{x}_t)^\top J_t(x_t)(x_{t-1} - \hat{x}_t) \right\} + \tilde{f}_t(x_t), \end{aligned} \quad (223b)$$

$$|\zeta_{t,3}(x_t, x_{t-1})| \lesssim \frac{d^3 \log^{4.5} T}{T^{3/2}}. \quad (223c)$$

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

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

where we have made use of the crude bound (222) in conjunction with the properties (39). Taking this observation together with (223) and (74b), we can apply a little algebra to derive

$$p_{X_{t-1}|X_t}(x_{t-1}|x_t) = f_7(x_t) \exp\left(-f_8(x_t, x_{t-1}) + \zeta_{t,4}(x_t, x_{t-1})\right) \quad (224a)$$

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

$$f_8(x_t, x_{t-1}) = \frac{\alpha_t}{2(1-\alpha_t)} \left\{ (x_{t-1} - \mu_t(x_t))^\top \left(I + \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)} J_t(x_t) \right) (x_{t-1} - \mu_t(x_t)) \right\}, \quad (224b)$$

$$|\zeta_{t,4}(x_t, x_{t-1})| \lesssim \frac{d^3 \log^{4.5} T}{T^{3/2}}. \quad (224c)$$

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

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

where A is a matrix obeying (see (222))

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

Consequently, we can demonstrate that

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

To see why the last line holds, note that (according to Lemma 1 and the properties (39))

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

and hence

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

Combining the above bound with (224), we arrive at

$$p_{X_{t-1}|X_t}(x_{t-1}|x_t) = f_9(x_t) \exp\left(-f_{10}(x_t, x_{t-1}) + \zeta_{t,5}(x_t, x_{t-1})\right) \quad (225a)$$

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

$$f_{10}(x_t, x_{t-1}) = \frac{\alpha_t}{2(1-\alpha_t)} \left\| \left(I - \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)} J_t(x_t) \right)^{-1} (x_{t-1} - \mu_t(x_t)) \right\|_2^2, \quad (225b)$$

$$|\zeta_{t,5}(x_t, x_{t-1})| \lesssim \frac{d^3 \log^{4.5} T}{T^{3/2}}. \quad (225c)$$

To finish up, repeat Step 3 in the proof of Lemma 8 to yield

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

as claimed. This combined with (225) concludes the proof.

E.2.1 Proof of auxiliary claims in Lemma 12

Proof of relation (216). For any $(x_t, x_{t-1}) \in \mathcal{E}$, one necessarily has

$$\|x_t(\gamma) - \hat{x}_t\|_2 \leq \|x_{t-1} - \hat{x}_t\|_2 \leq c_3 \sqrt{d(1-\alpha_t) \log T}. \quad (226)$$

Given x_t , we define the set

$$\mathcal{E}_1 := \{x : \|\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x\|_2 \leq c_4 \sqrt{d(1-\bar{\alpha}_{t-1}) \log T}\}.$$

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

$$\begin{aligned} \|x_t(\gamma) - \sqrt{\bar{\alpha}_{t-1}}x_0\|_2 &\leq \max\left\{\|\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0\|_2, \|\hat{x}_t - x_{t-1}\|_2\right\} \leq \max\left\{c_4 \sqrt{d(1-\bar{\alpha}_{t-1}) \log T}, c_3 \sqrt{d(1-\alpha_t) \log T}\right\} \\ &= c_4 \sqrt{d(1-\bar{\alpha}_{t-1}) \log T}, \end{aligned}$$

where the last inequality comes from (39b). This in turn reveals that

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

where the second line follows from the elementary relation

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

and the last line relies on (39) and our assumption on T . Moreover, repeating the same argument as in (116) and (120), we arrive at

$$\frac{p_{X_{t-1}}(x_t(\gamma))}{p_{X_{t-1}}(\hat{x}_t)} = 1 + O\left(d \sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}} \log T\right). \quad (227)$$

Putting the above results together leads to

$$\frac{p_{X_0|X_{t-1}}(x_0|x_t(\gamma))}{p_{X_0|X_{t-1}}(x_0|\hat{x}_t)} = \frac{p_{X_{t-1}|X_0}(x_t(\gamma)|x_0)/p_{X_{t-1}}(x_t(\gamma))}{p_{X_{t-1}|X_0}(\hat{x}_t|x_0)/p_{X_{t-1}}(\hat{x}_t)} = 1 + O\left(d\sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}}\log T\right).$$

Equipped with the above relation, we can demonstrate that

$$\begin{aligned} & \left\| \int p_{X_0|X_{t-1}}(x_0|x_t(\gamma))(x_t(\gamma) - \sqrt{\bar{\alpha}_{t-1}}x_0)dx_0 - \int p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)dx_0 \right\|_2 \\ & \leq O\left(d\sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}}\log T\right) \cdot \left(\int p_{X_0|X_{t-1}}(x_0|x_t(\gamma))\|x_t(\gamma) - \sqrt{\bar{\alpha}_{t-1}}x_0\|_2 dx_0 \right) \\ & \quad + \left\| \int p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(x_t(\gamma) - \hat{x}_t)dx_0 \right\|_2 \end{aligned} \quad (228)$$

$$\lesssim \sqrt{d^3(1-\alpha_t)\log^3 T}, \quad (229)$$

where the last step invokes the property (43a). Following similar arguments (which we omit for brevity), we can also derive

$$\begin{aligned} & \left\| \int p_{X_0|X_{t-1}}(x_0|x_t(\gamma))(x_t(\gamma) - \sqrt{\bar{\alpha}_{t-1}}x_0)(x_t(\gamma) - \sqrt{\bar{\alpha}_{t-1}}x_0)^\top dx_0 \right. \\ & \quad \left. - \int p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)^\top dx_0 \right\| \\ & \lesssim d^2\sqrt{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\log^2 T, \end{aligned} \quad (230)$$

where we have made use of the property (43b). Taking the above two above perturbation bounds together with the expression (94) and making use of (43a) immediately lead to the advertised result:

$$\|J_{t-1}(x_t(\gamma)) - J_{t-1}(\hat{x}_t)\| \lesssim d^2\sqrt{\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}}}\log^2 T. \quad (231)$$

Proof of relation (217). To establish this relation, we first apply the triangle inequality:

$$\left\| \frac{J_{t-1}(\hat{x}_t)}{1-\bar{\alpha}_{t-1}} - \frac{J_t(x_t)}{1-\bar{\alpha}_t} \right\| \leq \left\| \frac{J_{t-1}(\hat{x}_t) - J_t(x_t)}{1-\bar{\alpha}_{t-1}} \right\| + \left\| \left(\frac{1}{1-\bar{\alpha}_{t-1}} - \frac{1}{1-\bar{\alpha}_t} \right) J_t(x_t) \right\|.$$

Let us first consider the second term

$$\left\| \left(\frac{1}{1-\bar{\alpha}_{t-1}} - \frac{1}{1-\bar{\alpha}_t} \right) J_t(x_t) \right\| = \frac{\bar{\alpha}_{t-1}(1-\alpha_t)}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)} \|J_t(x_t)\| \leq \frac{\bar{\alpha}_{t-1}(1-\alpha_t)d\log T}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)} \lesssim \frac{(1-\alpha_t)d\log T}{(\alpha_t - \bar{\alpha}_t)^2}. \quad (232)$$

where the last inequality uses (161a) and the properties (39).

Next, we move on to bound the difference $J_{t-1}(\hat{x}_t) - J_t(x_t)$. By virtue of the relation (173), one can deduce that

$$\begin{aligned} & \left\| \int p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)dx_0 - \frac{1}{\sqrt{\alpha_t}} \int p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 \right\|_2 \\ & = \frac{1}{\sqrt{\alpha_t}} \left\| \int p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 - \int p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 \right\| \\ & \leq \frac{1}{\sqrt{\alpha_t}} O\left(\frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}}\right) \int p_{X_0|X_t}(x_0|x_t)\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 dx_0. \end{aligned}$$

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 (39a) to obtain

$$\begin{aligned} & \left\| \int p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)dx_0 - \int p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)dx_0 \right\|_2 \\ & \lesssim \frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}} \int p_{X_0|X_t}(x_0|x_t)\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 dx_0. \end{aligned}$$

Repeating the same argument also reveals that

$$\begin{aligned} & \left\| \int p_{X_0|X_{t-1}}(x_0|\hat{x}_t)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)(\hat{x}_t - \sqrt{\bar{\alpha}_{t-1}}x_0)^\top dx_0 - \int p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)(x_t - \sqrt{\bar{\alpha}_t}x_0)^\top dx_0 \right\| \\ & \lesssim \frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}} \left\| \int p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)(x_t - \sqrt{\bar{\alpha}_t}x_0)^\top dx_0 \right\|. \end{aligned}$$

In view of the expression (94) for J_t , combining the preceding two bounds with a little algebra yields

$$\begin{aligned} & \frac{1}{1-\bar{\alpha}_{t-1}} \|J_{t-1}(\hat{x}_t) - J_t(x_t)\| \\ & \lesssim \frac{1}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)} \frac{d(1-\alpha_t)\log T}{1-\bar{\alpha}_{t-1}} \\ & \quad \cdot \left\{ \left\| \int p_{X_0|X_t}(x_0|x_t)(x_t - \sqrt{\bar{\alpha}_t}x_0)(x_t - \sqrt{\bar{\alpha}_t}x_0)^\top dx_0 \right\| + \left(\int p_{X_0|X_t}(x_0|x_t)\|x_t - \sqrt{\bar{\alpha}_t}x_0\|_2 dx_0 \right)^2 \right\} \\ & \lesssim \frac{d^2(1-\alpha_t)\log^2 T}{(\alpha_t - \bar{\alpha}_t)^2}, \end{aligned}$$

where the last line follows from Lemma 1 and the properties (39).

Putting the above bounds together immediately establishes relation (217).

E.3 Proof of Lemma 13

According to the expression (209), one has

$$Y_{t-1} | Y_t = x_t \sim \mathcal{N}\left(\mu_t(x_t), \underbrace{\frac{1-\alpha_t}{\alpha_t} \left(I - \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)} J_t(x_t) \right)^2}_{=: \Sigma(\hat{x}_t)}\right).$$

In order to quantify the density, we first bound the Jacobian matrix $J_t(x)$ defined in (37). On the one hand, the expression (38) tells us that $J_t(x) \preceq I_d$ for any x , given that the term within the curly bracket in (38) is a negative covariance matrix. On the other hand, $J_t(x)$ can be lower bounded by

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

where the second line applies the assumption that $\|X_0\|_2 \leq T^{c_R}$, and the last line invokes the choice (22). As a consequence, we have

$$\Sigma(\hat{x}_t) \succeq \frac{1-\alpha_t}{\alpha_t} \left(1 - \frac{1-\alpha_t}{2(1-\bar{\alpha}_t)}\right)^2 I_d = \frac{1-\alpha_t}{4\alpha_t} \left(\frac{1-\bar{\alpha}_t + \alpha_t - \bar{\alpha}_t}{1-\bar{\alpha}_t}\right)^2 I_d \succeq \frac{1-\alpha_t}{4\alpha_t} I_d \succeq \frac{1-\alpha_t}{4} I_d; \quad (233a)$$

$$\Sigma(\hat{x}_t) \preceq \frac{1 - \alpha_t}{\alpha_t} T^{2c_0+2} (2\|\hat{x}_t\|_2^4 + 2T^{4c_R}) I_d \preceq 4T^{2c_0+2} (\|\hat{x}_t\|_2^4 + T^{4c_R}) I_d. \quad (233b)$$

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 (209) that

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

where the second inequality results from (233), and the third inequality makes use of (177) and (22). Given that $\log \frac{p_{X_{t-1}|X_t}(x_{t-1}|x_t)}{p_{Y_{t-1}|Y_t}(x_{t-1}|x_t)} \leq \log \frac{1}{p_{Y_{t-1}|Y_t}(x_{t-1}|x_t)}$, 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.
- Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. (2023a). Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*.
- Benton, J., Deligiannidis, G., and Doucet, A. (2023b). Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*.
- 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.
- Hausmann, U. G. and Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.
- Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- 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. (2023). Statistical efficiency of score matching: The view from isoperimetry. *International Conference on Learning Representations*.
- 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., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. (2024a). Accelerating convergence of score-based diffusion models, provably. *preprint*.
- Li, G., Huang, Z., and Wei, Y. (2024b). Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*.

- Li, G. and 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., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency models.
- 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*.
- Tang, W. (2023). Diffusion probabilistic models. *preprint*.
- Tang, W. and Zhao, H. (2024). Score-based diffusion models via stochastic differential equations—a technical tutorial. *arXiv preprint arXiv:2402.07487*.
- 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*.
- Wu, Y., Chen, M., Li, Z., Wang, M., and Wei, Y. (2024). Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *arXiv preprint arXiv:2403.01639*.
- 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*.