Towards Faster Non-Asymptotic Convergence for Diffusion Generative Models

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

May 23, 2023

Abstract

Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain far from mature. In this work, we develop a suite of non-asymptotic theory towards understanding the data generation process of diffusion models, assuming access to reliable estimates of the (Stein) score functions. For a popular deterministic sampler (based on the probability flow ODE), we establish a convergence rate proportional to $1/T$ (with T the total number of steps), improving upon past results; for another mainstream stochastic sampler (i.e., a type of the denoising diffusion probabilistic model (DDPM)), we derive a convergence rate proportional to $1/\sqrt{T}$, matching the state-of-the-art theory. Our theory imposes only minimal assumptions on the target data distribution, and is developed based on an elementary yet versatile non-asymptotic approach without resorting to toolboxes for SDEs and ODEs. Further, we design two accelerated variants, improving the convergence to $1/T^2$ for the ODE-based sampler and $1/T$ for the DDPM-type sampler, which might be of independent theoretical and empirical interest.

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

Contents

1 Introduction **2**

2 Preliminaries **3**
2.1 Diffusion generative models 4
2.2 Deterministic vs. stochastic samplers: a continuous-time interpretation 4

3 Algorithms and main results **6**
3.1 Assumptions and learning rates 6
3.2 Deterministic samplers 7
3.3 Stochastic samplers 9

4 Other related works **10**

5 Discussion **11**

*Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.
†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

A Analysis for the deterministic samplers (Theorems 1 and 2) **15**
A.1 Preliminary facts . . . 15
A.2 Proof of Theorem 1 . . . 15
A.3 Proof of Theorem 2 . . . 18
A.4 Proof of properties (41) . . . 19
A.5 Proof of Lemma 1 . . . 20
A.6 Proof of Lemma 2 . . . 22
A.7 Proof of Lemma 3 . . . 26

B Analysis for the stochastic samplers (Theorems 3 and 4) **28**
B.1 Step 1: approximation of $p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)$ . . . 28
B.2 Step 2: uniform control of the density ratios . . . 29
B.3 Step 3: computing the KL divergence . . . 31
B.4 In summary . . . 32
B.5 Proof of Claim (94) . . . 32

1 Introduction

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

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

1) a forward process
$$X_0 \to X_1 \to \cdots \to X_T \tag{1}$$
eq:forward-process-informal
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 Gaussian);

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

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

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

through proper learning based on how the training data propagate in the forward process. Encouragingly, there often exist feasible strategies to achieve this goal as long as faithful estimates about the (Stein) score functions — the gradients of the log marginal density of the forward process — are available, an intriguing fact that can be illuminated by the existence and construction of reverse-time stochastic differential equations

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

(SDEs) (Anderson, 1982) (see Section 2.2 for more precise discussions). Viewed in this light, a diverse array of diffusion models are frequently referred to as *score-based generative modeling (SGM)*. The popularity of SGM was initially motivated by, and has since further inspired, numerous recent studies on the problem of learning score functions, a subroutine that also goes by the name of score matching (e.g., Hyvärinen (2005, 2007); Vincent (2011); Song et al. (2020b)).

Nonetheless, despite the mind-blowing empirical advances, a mathematical theory for diffusion generative models is still in its infancy. Given the complexity of developing a full-fledged end-to-end theory, a divide-and-conquer approach has been advertised, decoupling the score learning phase (i.e., how to estimate score functions reliably from training data) and the generative sampling phase (i.e., how to generate new data instances given the score estimates). In particular, the past two years have witnessed growing interest and remarkable progress from the theoretical community towards understanding the generative sampling phase (Block et al., 2020; De Bortoli et al., 2021; Liu et al., 2022; De Bortoli, 2022; Lee et al., 2023; Pidstrigach, 2022; Chen et al., 2022b,a, 2023b). For instance, polynomial-time convergence guarantees have been established for stochastic samplers (e.g., Chen et al. (2022b,a)) and deterministic samplers (e.g., Chen et al. (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 DDIM); and (ii) a DDPM-type stochastic sampler motivated by reverse-time SDEs. In comparisons to past works, our main contributions are three-fold.

- *Non-asymptotic convergence guarantees.* For a popular deterministic sampler, we demonstrate that the number of steps needed to yield ε-accuracy — meaning that the total variance (TV) distance between the distribution of X_1 and that of Y_1 is no larger than ε — is proportional to $1/\varepsilon$ (in addition to other polynomial dimension dependency), which improves upon prior results Chen et al. (2023b). For another DDPM-type stochastic sampler, we establish an iteration complexity proportional to $1/\varepsilon^2$, matching existing theory Chen et al. (2022b,a) in terms of the ε-dependency.
- *Accelerating data generation processes.* In order to further speed up the sampling processes, we develop an accelerated variant for each of the above two samplers, taking advantage of estimates of a small number of additional quantities. As it turns out, these variants achieve more rapid convergence, with the deterministic (resp. stochastic) variant exhibiting a $1/\sqrt{\varepsilon}$ (resp. $1/\varepsilon$) scaling in the accuracy level ε.
- *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.

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 $f(d,T) \le C_1 g(d,T)$ (resp. $f(d,T) \ge C_1 g(d,T)$) for all d and T; moreover, the notation $f(d,T) \asymp g(d,T)$ indicates that $f(d,T) \lesssim g(d,T)$ and $f(d,T) \gtrsim g(d,T)$ hold at once. The notation $\widetilde{O}(\cdot)$ is defined similar to $O(\cdot)$ except that it hides the logarithmic dependency. We shall often use capital letters to denote random variables/vectors/processes, and lowercase letters for deterministic variables. For any two probability measures P and Q, the total variation (TV) distance between them is defined to be $\mathsf{TV}(P,Q) \coloneqq \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$.

2 Preliminaries

sec:preliminaries

In this section, we introduce the basics of diffusion generative models. The ultimate goal of a generative model can be concisely stated: given data samples drawn from an unknown distribution of interest p_{data} in

$\mathbb{R}^d$, we wish to generate new samples whose distributions closely resemble p_{data}.

2.1 Diffusion generative models

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

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

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

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

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

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

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

$$X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\, \overline{W}_t \qquad \text{for some } \overline{W}_t \sim \mathcal{N}(0, I_d). \tag{6}$$ eqn:Xt-X0

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

$$q_t \coloneqq \mathsf{law}\big(X_t\big) \tag{7}$$ eq:defn-qt

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

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

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

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

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

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

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

$$p_t \coloneqq \mathsf{law}\big(Y_t\big). \tag{10}$$ eq:defn-pt

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

sec:det-stochastic-samplers

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 $\{\Phi_t(\cdot)\}_{1\le t\le T}$ and computes:
$$Y_{t-1} = \Phi_t\big(Y_t\big), \qquad t = T, \cdots, 1. \tag{11}$$
eq:deterministic-sampler
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 $\{\Psi_t(\cdot,\cdot)\}_{1\le t\le T}$ and performs the updates:
$$Y_{t-1} = \Psi_t\big(Y_t, Z_t\big), \qquad t = T, \cdots, 1, \tag{12}$$
eq:stochastic-sampler
where the Z_t's are independent noise vectors obeying $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$.

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

- *The forward process.* A continuous-time analog of the forward diffusion process can be modeled as
$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t \quad (0 \le t \le T), \qquad X_0 \sim p_{\mathsf{data}} \tag{13}$$
eq:forward-SDE-general
for some functions $f(\cdot,\cdot)$ and $g(\cdot)$ (denoting respectively the drift and diffusion coefficient), where W_t denotes a d-dimensional standard Brownian motion. As a special example, the continuum limit of (4) takes the following form (Song et al., 2021b)
$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}W_t \quad (0 \le t \le T), \qquad X_0 \sim p_{\mathsf{data}} \tag{14}$$
eq:forward-SDE
for some function $\beta(t)$. As before, we denote by q_t the distribution of X_t in (13).

- *The reverse process.* As it turns out, the following two reverse processes are both capable of reconstructing the distribution of the forward process, motivating the design of two distinctive samplers. Here and throughout, we use $\nabla \log q_t(X)$ to abbreviate $\nabla_X \log q_t(X)$ for notational simplicity.
  - One feasible approach is to resort to the so-called *probability flow ODE* (Song et al., 2021b)
$$\mathrm{d}Y_t^{\mathsf{ode}} = \Big(f\big(Y_t^{\mathsf{ode}}, t\big) - \frac{1}{2}g(t)^2\nabla \log q_t\big(Y_t^{\mathsf{ode}}\big)\Big)\mathrm{d}t \quad (0 \le t \le T), \qquad Y_0^{\mathsf{ode}} \sim q_T, \tag{15}$$
eq:prob-flow-ODE
which exhibits matching distributions as follows:
$$Y_{T-t}^{\mathsf{ode}} \overset{\mathrm{d}}{=} X_t, \qquad 0 \le t \le T.$$
The deterministic nature of this approach often enables faster sampling. It has been shown that this family of deterministic samplers is closely related to the DDIM sampler (Karras et al., 2022; Song et al., 2021b).
  - In view of the seminal result by Anderson (1982), one can also construct a "reverse-time" SDE
$$\mathrm{d}Y_t^{\mathsf{sde}} = \Big(f\big(Y_t^{\mathsf{sde}}, t\big) - g(t)^2\nabla \log q_t\big(Y_t^{\mathsf{sde}}\big)\Big)\mathrm{d}t + g(t)\mathrm{d}Z_t^{\mathsf{sde}} \quad (0 \le t \le T), \qquad Y_0^{\mathsf{sde}} \sim q_T \tag{16}$$
eq:reverse-SDE
with Z_t^{sde} being a standard Brownian motion, which also satisfies
$$Y_{T-t}^{\mathsf{sde}} \overset{\mathrm{d}}{=} X_t, \qquad 0 \le t \le T.$$
The popular DDPM sampler (Ho et al., 2020; Nichol and Dhariwal, 2021) falls under this category.

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

3 Algorithms and main results

sec: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 reverse process without resorting to any toolbox of SDEs and ODEs tailored to the continuous-time limits. This elementary approach might potentially be versatile for analyzing a broad class of variations of these samplers.

3.1 Assumptions and learning rates

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

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

assumption:score-estimate

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

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

eqn:training-score

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

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

eqn:training-score-equiv

which is often more amenable to training.

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

$$\mathbb{P}\big(\|X_0\|_2 \le T^{c_3} \mid X_0 \sim p_{\mathsf{data}}\big) = 1 \tag{19}$$

eq:assumption-data-bounded

for some arbitrarily large constant $c_3 > 0$. This assumption allows the radius of the support of p_{data} to be exceedingly large (given that the exponent c_3 can be arbitrarily large).

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

eqn:alpha-t

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

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

We immediately single out several properties about the choice (20) that prove useful throughout:

eqn:properties-alpha

$$\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \le \frac{4c_1\log T}{T}, \qquad 1 \le t \le T \tag{21a}$$

$$\overline{\alpha}_T \le \frac{1}{T^{c_2}} \tag{21b}$$

for some large enough numerical constant $c_2 > 0$. The proof of these properties can be found in Appendix (A.4).

3.2 Deterministic samplers

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

3.2.1 An ODE-based deterministic sampler

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

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \Phi_t\big(Y_t\big) \quad \text{for } t = T, \cdots, 1, \tag{22a}$$ eqn:ode-sampling-Y

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

$$\Phi_t(x) := \frac{1}{\sqrt{\alpha_t}}\bigg(x + \frac{1-\alpha_t}{2} s_t(x)\bigg). \tag{22b}$$ eqn:phi-func

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

We first derive non-asymptotic convergence guarantees — measured by the total variance distance between the forward and the reverse processes — for the above deterministic sampler (22).

thm:main-ODE

Theorem 1. *Suppose that (19) holds true. Equipped with the score estimates in Assumption 1 and the learning rate schedule (20), the sampling process (22) satisfies*

$$\mathsf{TV}\big(q_1, p_1\big) \le C_1 \frac{d^2 \log^4 T}{T} + C_1 \frac{d^6 \log^6 T}{T^2} \tag{23}$$ eq:ratio-ODE

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

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

$$\widetilde{O}\bigg(\frac{d^2}{\varepsilon} + \frac{d^3}{\sqrt{\varepsilon}}\bigg). \tag{24}$$ eq:iteration-complexity-ODE

To the best of our knowledge, the only non-asymptotic analysis for such deterministic samplers in prior literature was derived by a very recent work Chen et al. (2023b), which established the first polynomial-time convergence guarantees (see, e.g., Chen et al. (2023b, Theorem 4.1)). However, it fell short of providing concrete polynomial dependency. In contrast, our result in Theorem 1 uncovers a concrete d^2/ε scaling (ignoring lower-order and logarithmic terms), which was previously unavailable.

[Yuxin: might need a discussion with the new paper]

3.2.2 An accelerated deterministic sampler

Thus far, we have demonstrated that the iteration complexity of the deterministic sampler (22) is proportional to $1/\varepsilon$ (for small enough ε). A natural question is whether this convergence rate can be further improved.

As it turns out, if we have access to reliable estimates of a few additional scalar quantities, then a modified version of the sampler (22) is able to achieve much improved convergence guarantees. These estimates are made precise in the following assumption.

assumption:ODE-estimate

Assumption 2. *Suppose that we have access to the estimates $A_t, \ldots, E_t : \mathbb{R}^d \to \mathbb{R}$ and $F_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ $(1 \le t \le T)$ defined as follows:* eqn:training-ODE

$$A_t := \arg\min_{A:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}\Big[\Big\{\|W\|_2^2 - A\big(\sqrt{\alpha_t}X_0 + \sqrt{1-\alpha_t}W\big)\Big\}^2\Big], \tag{25a}$$

$$B_t := \arg\min_{B:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}\Big[\Big\{\sqrt{1-\alpha_t}W^\top s_t\big(\sqrt{\alpha_t}X_0 + \sqrt{1-\alpha_t}W\big) + B\big(\sqrt{\alpha_t}X_0 + \sqrt{1-\alpha_t}W\big)\Big\}^2\Big], \tag{25b}$$

$$C_t \coloneqq \arg\min_{C:\mathbb{R}^d\to\mathbb{R}} \mathbb{E}\Big[\Big\{\|W\|_2^4 - C\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W\big)\Big\}^2\Big], \tag{25c}$$

$$D_t \coloneqq \arg\min_{D:\mathbb{R}^d\to\mathbb{R}} \mathbb{E}\Big[\Big\{\Big(\sqrt{1-\overline{\alpha}_t}W^\top s_t\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W\big)\Big)^2 - D\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W\big)\Big\}^2\Big], \tag{25d}$$

$$E_t \coloneqq \arg\min_{E:\mathbb{R}^d\to\mathbb{R}} \mathbb{E}\Big[\Big\{\sqrt{1-\overline{\alpha}_t}\|W\|_2^2 W^\top s_t\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W\big) + E\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W\big)\Big\}^2\Big], \tag{25e}$$

$$F_t \coloneqq \arg\min_{F:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}} \mathbb{E}\Big[\Big\{\langle M, M'\rangle - F\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W, \sqrt{\overline{\alpha}_t}X_0' + \sqrt{1-\overline{\alpha}_t}W'\big)\Big\}^2\Big], \tag{25f}$$

where $s_t(\cdot)$ is defined in Assumption 1, and eq:defn-M-Mprime

$$M \coloneqq (1-\overline{\alpha}_t)s_t\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W\big)s_t\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W\big)^\top - WW^\top; \tag{26a}$$

$$M' \coloneqq (1-\overline{\alpha}_t)s_t\big(\sqrt{\overline{\alpha}_t}X_0' + \sqrt{1-\overline{\alpha}_t}W'\big)s_t\big(\sqrt{\overline{\alpha}_t}X_0' + \sqrt{1-\overline{\alpha}_t}W'\big)^\top - W'W'^\top. \tag{26b}$$

Here, the expectation is with respect to $W, W' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and $X_0, X_0' \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}$.

Armed with the score estimate in Assumption 1 and the additional scalar estimates in Assumption 2, we are ready to introduce an accelerated variant of (22) as follows: eqn:ode-sampling-R

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \Phi_t\big(Y_t\big) \quad \text{for } t = T, \cdots, 1, \tag{27a}$$ eqn:ode-sampling-R-Y

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

$$\Phi_t(x) = \frac{1}{\sqrt{\alpha_t}}\left(x + \frac{\lambda_t(x)(1-\alpha_t)}{2}s_t(x)\right), \tag{27b}$$

Here, $\lambda_t : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\lambda_t(x) = 1 + \frac{(1-\alpha_t)w_t(x)}{4(\alpha_t - \overline{\alpha}_t)}, \tag{27c}$$

where $w_t : \mathbb{R}^d \mapsto \mathbb{R}$ is chosen to be a solution to the following equation:

$$(1-\overline{\alpha}_t)s_t^\top \nabla w + (d - A_t)w + (d + B_t - A_t)^2 - 2d^2 + 7d - 6A_t + 7B_t - C_t - D_t + 2E_t + \widetilde{F}_t = 0 \tag{27d}$$

with $\widetilde{F}_t(x) \coloneqq F_t(x, x)$. Notably, this new variant (27) is closely related to the original sampler (22); in fact, they both move along the direction specified by the score estimate s_t, except that the stepsize of the accelerated variant (mainly the quantity λ_t) is adaptively chosen based on additional information.

Encouragingly, our non-asymptotic analysis framework can be extended to derive enhanced convergence guarantees for the sampler (27), as stated below.

thm:main-ODE-fast

Theorem 2. *Suppose that (19) holds true. Equipped with the estimates in Assumptions 1-2 and the learning rate schedule (20), the sampling process (27) obeys*

$$\mathsf{TV}\big(q_1, p_1\big) \le C_1 \frac{d^6 \log^6 T}{T^2} \tag{28}$$ eq:ratio-ODE-R

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

Theorem 2 reveals that: in order to achieve $\mathsf{TV}(q_1, p_1) \le \varepsilon$, the accelerated deterministic sampler (27) only requires the number of steps T to be on the order of

$$\widetilde{O}\left(\frac{d^3}{\sqrt{\varepsilon}}\right), \tag{29}$$ eq:iteration-complexity-ODE-fast

thus improving the dependency on ε from $\widetilde{O}(1/\varepsilon)$ (cf. (24)) to $\widetilde{O}(1/\sqrt{\varepsilon})$ for small enough ε. Consequently, the improved convergence result underscores the crucial role of stepsize selection even when the search direction has been determined.

3.3 Stochastic samplers

3.3.1 A DDPM-type stochastic sampler

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

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

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

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

The key difference between this sampler and the deterministic sampler (22) is that: (i) there exists an additional pre-factor of $1/2$ on s_t in the deterministic sampler; and (ii) the stochastic sampler injects additional noise Z_t in each step.

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

thm:main-SDE

Theorem 3. *Suppose that (19) holds true. Equipped with the estimates in Assumption 1 and the learning rate schedule (20), the stochastic sampler (30) achieves*

$$\mathsf{TV}\big(q_1, p_1\big) \le C_1 \frac{d^2 \log^3 T}{\sqrt{T}} \tag{31}$$

eq:ratio-SDE

for some universal constants $C_1 > 0$.

Theorem 3 establishes non-asymptotic convergence guarantees for the stochastic sampler (30). As asserted by the theorem, the number of steps needed to attain ε-accuracy (measured by the TV distance between p_1 and q_1) is proportional to $1/\varepsilon^2$, matching the state-of-the-art ε-dependency derived in Chen et al. (2022a), albeit exhibiting a worse dimensional dependency. Our analysis follows a completely different path compared with the SDE-based approach in Chen et al. (2022a), thus offering complementary interpretations for this important sampler. In order to further illustrate the versatility of our analysis approach, we shall demonstrate how it can be applied to study an accelerated version in the next subsection.

3.3.2 An accelerated stochastic sampler

In this subsection, we come up with a potential strategy to speed up the stochastic sampler (30), assuming access to reliable estimates of additional objects as described below.

assumption:variance-estimate

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

$$v_t \coloneqq \arg\min_{v:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d} \mathbb{E}\left[\Big\|WW^\top Z - v\big(\sqrt{\overline{\alpha}_t}X + \sqrt{1-\overline{\alpha}_t}W, Z\big)\Big\|_2^2\right], \qquad 1 \le t \le T, \tag{32}$$

eqn:training-variance

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

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

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

eqn:sde-sampling-R-Y

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

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

eqn:sde-sampling-R-Psi

with

$$\sigma_t^2 = \frac{1}{\alpha_t} - 1. \tag{33c}$$

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.

thm:main-SDE-R

Theorem 4. *Suppose that (19) holds true. Equipped with the estimates in Assumption 1, 3 and the learning rate schedule (20), the sampling process (33) satisfies*

$$\mathsf{TV}\big(q_1, p_1\big) \le C_1 \frac{d^3 \log^{4.5} T}{T} \tag{34}$$

eq:ratio-SDE-R

for some universal constants $C_1 > 0$.

In comparison to the stochastic sampler (30), Theorem 4 asserts that the iteration complexity of the sampler (33) is at most

$$\widetilde{O}\Big(\frac{d^3}{\varepsilon}\Big), \tag{35}$$

eq:iteration-complexity-SDE-fast

thus significantly reducing the scaling $\widetilde{O}(1/\varepsilon^2)$ for the original sampler (30) to $\widetilde{O}(1/\varepsilon)$ regarding the ε-dependency. All in all, our theory reveals that having information about a small number of additional objects might substantially speed up the data generation process.

4 Other related works

sec:related-works

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

Score matching. Hyvärinen (2005) showed that the score function can be estimated via integration by parts, a result that was further extended in Hyvärinen (2007). Song et al. (2020b) proposed sliced score matching to tame the computational complexity in high dimension. The consistency of the score matching estimator was studied in Hyvärinen (2005), with asymptotic normality established in Forbes and Lauritzen (2015). Optimizing the score matching loss has been shown to be intimately connected to minimizing upper bounds on the Kullback-Leibler divergence (Song et al., 2021a) and Wasserstein distance (Kwon et al., 2022) between the generated distribution and the target data distribution. From a non-asymptotic perspective, Koehler et al. (2022) studied the statistical efficiency of score matching by connecting it with the isoperimetric properties of the distribution.

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

5 Discussion

sec: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. In addition to demonstrating the non-asymptotic iteration complexities of two mainstream discrete-time samplers — a deterministic sampler based on the probability flow ODE, and a DDPM-type stochastic sampler — we have discovered potential strategies to further accelerate the sampling process, taking advantage of estimates of a small number of additional objects. The analysis framework laid out in the current paper might shed light on how to analyze other variants of score-based generative models as well.

Moving forward, there are plenty of questions that require in-depth theoretical understanding. For instance, the dimension dependency in our convergence results remains sub-optimal; can we further refine our theory in order to reveal tight dependency in this regard? To what extent can we further accelerate the sampling process, without requiring much more information than the score functions? 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.

References

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

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

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

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

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

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

Chen, S., Chewi, S., 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. (2023b). Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. *arXiv preprint arXiv:2303.03384*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Analysis for the deterministic samplers (Theorems 1 and 2)

In this section, we present the proofs for our main results tailored to the two deterministic samplers.

A.1 Preliminary facts

Before proceeding, we gather a couple of useful facts for the proof. First of all, in view of the alternative expression (18) for the score estimate s_t and the property of the minimum mean square error estimator (e.g., Hajek (2015, Section 3.3.1)), we know that s_t is given by the conditional expectation

$$s_t(x) = \mathbb{E}\left[-\frac{1}{\sqrt{1-\overline{\alpha}_t}}W \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right] = \frac{1}{1-\overline{\alpha}_t}\mathbb{E}\left[\sqrt{\overline{\alpha}_t}X_0 - x \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right]$$
$$= \frac{1}{1-\overline{\alpha}_t}\int_{x_0}(\sqrt{\overline{\alpha}_t}x_0 - x)p_{X_0|X_t}(x_0\,|\,x)\mathrm{d}x_0. \tag{36}$$

eq:st-MMSE-expression

Next, given that $X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W$ with $W \sim \mathcal{N}(0, I_d)$, we have the following tail bound for the random vector X_0 conditional on X_t.

lem:x0

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

$$\mathbb{P}\big(\|X_0\|_2 \le R\big) = 1 \qquad \text{and} \qquad R = T^{c_R}. \tag{37}$$

eq:cR-defn-lem

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

$$-\log p_{X_t}(y) \le c_6 d\log T \tag{38}$$

eqn:choice-y

for some large enough constant $c_6 > 0$, and let X be a random vector whose distribution $p_X(\cdot)$ obeys

$$p_X(x) = p_{X_0|X_t}(x\,|\,y). \tag{39}$$

eq:pX-properties

Then for any $c_5 \ge \sqrt{2 + c_0/2 + c_R}$ (with c_0 defined in (20)), one has

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

with probability at least $1 - \exp\big(-c_5^2 d\log T\big)$.

Proof. See Appendix A.5. □

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

eqn:properties-alpha-

$$\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \le \frac{4c_1\log T}{T}, \qquad 1 \le t \le T \tag{41a}$$

eqn:properties-alpha-proof-1

$$\overline{\alpha}_T \le \frac{1}{T^{c_2}} \tag{41b}$$

eqn:properties-alpha-proof-alphaT

for some large enough numerical constant $c_2 > 0$. In addition, if $\frac{d(1-\alpha_t)}{\alpha_t - \overline{\alpha}_t} \lesssim 1$, then one has

$$\Big(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\Big)^{d/2} = 1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{d(d-2)(1-\alpha_t)^2}{8(\alpha_t-\overline{\alpha}_t)^2} + O\bigg(d^3\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\bigg). \tag{41c}$$

eq:expansion-ratio-1-alpha

The proof of these properties is postponed to Appendix A.4.

A.2 Proof of Theorem 1

Before proceeding, we find it convenient to introduce a function

$$\phi_t(x) = x + \frac{1-\alpha_t}{2}s_t(x) = x - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\int_{x_0}(x - \sqrt{\overline{\alpha}_t}x_0)p_{X_0|X_t}(x_0\,|\,x)\mathrm{d}x_0, \tag{42}$$

defn:phit-x

where the second identity follows from (36). This allows us to express the update rule (22) as follows:

$$Y_{t-1} = \Phi_t(Y_t) = \frac{1}{\sqrt{\alpha_t}}\phi_t(Y_t). \tag{43}$$

eq:Yt-phi-ODE

Our proof consists of three steps below.

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

$$\begin{aligned}\frac{p_{Y_{t-1}}(y_{t-1})}{p_{X_{t-1}}(y_{t-1})} &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})} \\ &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{Y_t}(y_t)} \cdot \left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{X_t}(y_t)}\right)^{-1} \cdot \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)}, \end{aligned} \tag{44}$$

eq:recursion

thus converting the density ratio of interest into the product of three other density ratios. Motivated by this expression, we develop a lemma related to some of these density ratios, which plays a crucial role in establishing Theorem 1. The proof of this result is postponed to Appendix A.6.

lem:main-ODE

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

eq:ODE

$$\begin{aligned}\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} &= 1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg) \\ &\quad + \frac{(1-\alpha_t)\Big(\big\|\int_{x_0}(x-\sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\big\|_2^2 - \int_{x_0}\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}. \end{aligned} \tag{45a}$$

eq:xt

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

$$\begin{aligned}\frac{p_{\phi_t(X)}(\phi_t(x))}{p_X(x)} &= 1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\bigg) \\ &\quad + \frac{(1-\alpha_t)\Big(\big\|\int_{x_0}(x-\sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\big\|_2^2 - \int_{x_0}\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}. \end{aligned} \tag{45b}$$

eq:yt

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

$$\mathcal{E}_t \coloneqq \left\{y : \left|\frac{q_t(y)}{p_t(y)} - 1\right| \le c_5\left(\frac{d^2\log^4 T}{T^2} + \frac{d^6\log^6 T}{T^3}\right)(T-t+1)\right\}$$

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

$$\mathbb{P}(Y_t \in \mathcal{E}_t) \ge 1 - (T-t+1)\exp\big(-c_3 d\log T\big), \qquad t \ge 1 \tag{46}$$

eq:claim-Py

for some constant $c_3 > 0$. Suppose that this claim is valid for the moment. Then taking $t = 1$ leads to

$$\mathbb{P}(Y_1 \in \mathcal{E}_1) = \mathbb{P}\left(\left|\frac{q_1(Y_1)}{p_1(Y_1)} - 1\right| \le c_5\left(\frac{d^2\log^4 T}{T} + \frac{d^6\log^6 T}{T^2}\right)\right) \ge 1 - \exp\big(-c_3 d\log T\big), \tag{47}$$

which also reveals that

$$\begin{aligned}\int_{y\notin\mathcal{E}_1}\big|p_1(y)-q_1(y)\big|\mathrm{d}y &\le c_5\left(\frac{d^2\log^4 T}{T} + \frac{d^6\log^6 T}{T^2}\right)\int_{y\notin\mathcal{E}_1} p_1(y)\mathrm{d}y = c_5\left(\frac{d^2\log^4 T}{T} + \frac{d^6\log^6 T}{T^2}\right)\mathbb{P}(Y_1\notin\mathcal{E}_1) \\ &\le c_5\left(\frac{d^2\log^4 T}{T} + \frac{d^6\log^6 T}{T^2}\right)\exp\big(-c_3 d\log T\big) \le \exp\big(-c_3 d\log T\big).\end{aligned}$$

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

$$\mathsf{TV}\big(q_1,p_1\big) = \frac{1}{2}\int_x\big|q_1(x)-p_1(x)\big|\mathrm{d}x = \frac{1}{2}\mathbb{E}_{Y_1\sim p_1}\left[\left|\frac{q_1(Y_1)}{p_1(Y_1)}-1\right|\cdot\mathbb{1}\left\{Y_1\in\mathcal{E}_1\right\}\right] + \frac{1}{2}\int_{y\notin\mathcal{E}_1}\big|p_1(y)-q_1(y)\big|\mathrm{d}y$$

$$\leq c_5\left(\frac{d^2\log^4 T}{T}+\frac{d^6\log^6 T}{T^2}\right)+\frac{1}{2}\exp\big(-c_3 d\log T\big)$$
$$\asymp \frac{d^2\log^4 T}{T}+\frac{d^6\log^6 T}{T^2}.$$

This establishes the advertised result in Theorem 1, provided that Claim (46) can be verified.

Step 3: justifying the claim (46). We would like to prove this claim by induction, for which we start with the base case with $t=T$. Recall that $X_T \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_T}X_0+\sqrt{1-\overline{\alpha}_T}B$ and $Y_T \overset{\mathrm{d}}{=} B$ with $B\sim\mathcal{N}(0,I_d)$ independent of X_0, and that $\|X_0\|_2\leq R$ with $R=T^{c_R}$ for some constant $c_R>0$. For large enough T, it immediately follows from (41b) that

$$\mathbb{P}(Y_T\in\mathcal{E}_T)\geq 1-\exp(-c_3 d\log T) \tag{48}$$

for some constant $c_3>0$ large enough.

Suppose now that the claim (46) 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

$$q_t(Y_t)\geq \exp\big(-c_6 d\log T\big). \tag{49}$$

With (49) in place, one sees that Lemma 2 is applicable to $x=Y_t$ with high probability.

Next, observe that for any y obeying $-\log p_{X_t}(y)\leq c_6\log T$, applying relations (45a) and (45b) in Lemma 2 leads to

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

Replacing y with Y_t in the above display, using the fact that $Y_{t-1}=\frac{1}{\sqrt{\alpha_t}}\phi_t(Y_t)$, and invoking the relation (44), we immediately arrive at

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

with probability exceeding $1-(T-t)\exp(-c_3 d\log T)$. This concludes the proof of Claim (46) via standard induction arguments.

Proof of property (49). First, for any y obeying $y\in\mathcal{E}_t$ and $p_t(y)\geq\exp(-c_6 d\log T)$, one has

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

and, as a result,

$$\Big\{y\,|\,p_t(y)\geq 2\exp\big(-c_6 d\log T\big)\Big\}\cap\mathcal{E}_t\subseteq\Big\{y\,|\,q_t(y)\geq\exp\big(-c_6 d\log T\big)\Big\}. \tag{50}$$

We can then deduce that

$$\begin{aligned}&\mathbb{P}\Big(q_t(Y_t)\geq\exp\big(-c_6 d\log T\big)\Big)\geq\mathbb{P}\Big(p_t(Y_t)\geq 2\exp\big(-c_6 d\log T\big)\text{ and }Y_t\in\mathcal{E}_t\Big)\\&\quad\geq\mathbb{P}\Big(Y_t\in\big\{y:p_t(y)\geq 2\exp\big(-c_6 d\log T\big)\big\}\cap\mathcal{E}_t\cap\big\{y:\|y\|_2\leq T^{c_y}\big\}\Big)\\&\quad\geq 1-\mathbb{P}\{Y_t\notin\mathcal{E}_t\}-\mathbb{P}\Big(Y_t\in\big\{y:p_t(y)<2\exp\big(-c_6 d\log T\big)\big\}\cap\big\{y:\|y\|_2\leq T^{c_y}\big\}\Big)-\mathbb{P}\big\{\|Y_t\|_2>T^{c_y}\big\}\end{aligned}$$

$$\geq 1-(T-t-1)\exp\big(-c_3 d\log T\big)-\int_{y:\|y\|_2\leq T^{c_y}} 2\exp\big(-c_6 d\log T\big)\mathrm{d}y-\mathbb{P}\big\{\|Y_T\|_2>T^{c_y/2}\big\}$$
$$\geq 1-(T-t)\exp\big(-c_3 d\log T\big).$$

Here, the last line makes use of the fact $y_T\sim\mathcal{N}(0,I_d)$ and holds as long as c_6 is large enough; regarding the penultimate line, it suffices to recognize that (cf. (42) and (21))

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

and hence $\|y_t\|_2\lesssim \mathsf{poly}(T)$ unless $\|y_T\|_2\geq T^{c_y/2}$ (assuming that c_y is large). □

A.3 Proof of Theorem 2

Before continuing, we introduce the following additional notation

$$\varphi_t(x):=x+\frac{\lambda_t(1-\alpha_t)}{2}s_t(x)=x-\frac{\lambda_t(1-\alpha_t)}{2(1-\overline{\alpha}_t)}\int_{x_0}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0, \tag{51}$$

where the second identity follows from (36). Here, it is worth noting that $|\lambda_t-1|\lesssim d^{3/2}\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}$. With this notation in place, we can write

$$Y_{t-1}=\frac{1}{\sqrt{\alpha_t}}\varphi_t\big(Y_t\big). \tag{52}$$ eq:Yt-recursion-accelerate-proof

As it turns out, Theorem 2 can be established in a very similar way as in the proof of Theorem 1. The only step that needs to be changed is to replace (45) with (53) in the following lemma. The proof of this lemma can be found in Appendix A.7.

Lemma 3. *Consider every $x\in\mathbb{R}$ such that $-\log p_{X_t}(x)\lesssim d\log T$. Suppose $\frac{d^2\log T(1-\alpha_t)}{\alpha_t-\overline{\alpha}_t}\lesssim 1$, it satisfies that* eq:ODE-fast lem:main-ODE-fast

$$\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\varphi_t(x))}{p_{X_t}(x)}=1+\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{d(d-2)(1-\alpha_t)^2}{8(\alpha_t-\overline{\alpha}_t)^2}+O\Big(d^3\log^3 T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\Big)$$
$$-\frac{(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}\Big(A_t-\lambda_t B_t-\frac{1-\alpha_t}{4(\alpha_t-\overline{\alpha}_t)}(-B_t+C_t+D_t-2E_t)\Big) \tag{53a}$$ eq:xt-higher

and

$$\frac{p_{\varphi_t(X)}(\varphi_t(x))}{p_X(x)}=1+\frac{\lambda_t d(1-\alpha_t)}{2(1-\overline{\alpha}_t)}+\frac{\lambda_t(1-\alpha_t)(B_t-A_t)}{2(1-\overline{\alpha}_t)}+O\Big(d^6\log^3 T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\Big)$$
$$+\frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\Big[\Big(d+B_t-A_t\Big)^2+d+2(B_t-A_t)+F_t\Big]-\frac{(1-\alpha_t)}{2}s_t^\top\nabla\lambda_t, \tag{53b}$$ eq:yt-higher

where $A_t,\ldots,F_t$ are defined in (25).

With Lemma 3 in place, we are able to prove, in the same way as in the proof of Theorem 1, that

$$\mathbb{P}(Y_t\in\mathcal{E}_t)\geq 1-(T-t+1)\exp(-\Omega(d\log T)) \tag{54}$$ eq:claim

for every $t\geq 1$, where the set $\mathcal{E}_t$ is defined as

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

This implies the desired result in Theorem 2.

A.4 Proof of properties (41)

sec:proof-properties-alpha

Proof of property (41a). We start by proving (41a). Let τ be an integer obeying

$$\beta_1\left(1+\frac{c_1\log T}{T}\right)^{\tau}\leq 1<\beta_1\left(1+\frac{c_1\log T}{T}\right)^{\tau+1}, \tag{55}$$

eq:defn-tau-proof-alpha

and we divide into two cases based on τ.

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

$$1-\overline{\alpha}_{t-1}\geq\frac{1}{3}\beta_1\left(1+\frac{c_1\log T}{T}\right)^{t}. \tag{56}$$

eq:induction-proof-alpha

Clearly, if (56) is valid, then any $t\leq\tau$ obeys

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

as claimed. Towards proving (56), first note that the base case with $t=2$ holds true trivially since $1-\overline{\alpha}_1=1-\alpha_1=\beta_1\geq\beta_1\big(1+\frac{c_1\log T}{T}\big)^2/3$. Next, let $t_0>2$ be *the first time* that Condition (56) fails to hold and suppose that $t_0\leq\tau$. It then follows that

$$1-\overline{\alpha}_{t_0-2}=1-\frac{\overline{\alpha}_{t_0-1}}{\alpha_{t_0-1}}\leq 1-\overline{\alpha}_{t_0-1}<\frac{1}{3}\beta_1\left(1+\frac{c_1\log T}{T}\right)^{t_0}\leq\frac{1}{2}\beta_1\left(1+\frac{c_1\log T}{T}\right)^{t_0-1}<\frac{1}{2}, \tag{57}$$

where the last inequality result from (55) and the assumption $t_0\leq\tau$. This taken together with the assumptions (56) and $t_0\leq\tau$ implies that

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

As a result, we can further derive

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

where the penultimate line holds since (56) 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 (56) holds for all $t\leq\tau$.

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

$$1-\overline{\alpha}_{t-1}\geq 1-\overline{\alpha}_{\tau-1}\geq\frac{1}{3}\beta_1\left(1+\frac{c_1\log T}{T}\right)^{\tau}=\frac{\frac{1}{3}\beta_1\big(1+\frac{c_1\log T}{T}\big)^{\tau+1}}{1+\frac{c_1\log T}{T}}\geq\frac{1}{4}, \tag{58}$$

where the second and the third inequalities come from (56). Therefore, one obtains

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

Proof of property (41b). Turning attention to the second claim (41b), we note that for any t obeying $t \geq \frac{T}{2} \gtrsim \frac{T}{\log T}$, one has

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

This in turn allows one to deduce that

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

for an arbitrarily large constant $c_2 > 0$.

Proof of property (41c). Finally, it is easily seen from the Taylor expansion that the learning rates $\{\alpha_t\}$ satisfy

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

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

A.5 Proof of Lemma 1

sec:proof-lem:x0

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

eq:claim-123

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

eqn:lemma-x0-claim

$$\|\sqrt{\overline{\alpha}_t} x_0 - y\|_2 \leq c_5 \sqrt{d(1-\overline{\alpha}_t)\log T} \qquad \text{and} \tag{59a}$$

eqn:lemma-x0-claim-1

$$\mathbb{P}\big(\|X_0 - x_0\|_2 \leq \epsilon\big) \geq \Big(\frac{\epsilon}{2T^{c_6 + c_R}}\Big)^d \qquad \text{with} \quad \epsilon = \frac{1}{T^{c_0/2}} \tag{59b}$$

eqn:lemma-x0-claim-2

hold simultaneously, where c_0 is defined in (20) and c_R is defined in (37).

With the above claim in place, we are ready to prove Lemma 1. Let us look at a set:

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

where $c_5 \geq 5$ (see Claim 1). Combining this with (59a) results in

$$\mathbb{P}\Big(\|\sqrt{\overline{\alpha}_t} X - y\|_2 > 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T}\Big) \leq \mathbb{P}(X \in \mathcal{E}). \tag{60}$$

eq:UB-P-set-E

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

$$\begin{aligned}\mathbb{P}(X_0 \in \mathcal{E} \,|\, X_t = y) &= \frac{\int_{x \in \mathcal{E}} p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x}{\int_x p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x} \\ &\leq \frac{\int_{x \in \mathcal{E}} p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x}{\int_{x: \|x - x_0\|_2 \leq \epsilon} p_{X_0}(x) p_{X_t \,|\, X_0}(y \,|\, x) \mathrm{d}x} \\ &\leq \frac{\sup_{x \in \mathcal{E}} p_{X_t \,|\, X_0}(y \,|\, x)}{\inf_{x: \|x - x_0\|_2 \leq \epsilon} p_{X_t \,|\, X_0}(y \,|\, x)} \cdot \frac{\mathbb{P}(X_0 \in \mathcal{E})}{\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon)}.\end{aligned} \tag{61}$$

eq:UB1-P-X0-Xt

To further bound this quantity, note that: in view of the definition of $\mathcal{E}$ and expression (59a), one has

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

and

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

where the penultimate line relies on (59a), and the last line holds true since $1-\overline{\alpha}_t\ge 1-\alpha_1 = 1/T^{c_0}$ (see (20)). Substitution of the above two displays into (61), we arrive at

$$\begin{aligned}\mathbb{P}(X_0\in\mathcal{E}\,|\,X_t=y) &\le \exp(-2.5c_5^2 d\log T)\cdot\frac{1}{\mathbb{P}(\|X_0-x_0\|_2\le\epsilon)}\\ &\le \exp(-2.5c_5^2 d\log T)\cdot\Big(2T^{c_6+c_0/2+c_R}\Big)^d\\ &\le \exp\big(-(2c_5^2-c_6-c_0/2-c_R)d\log T\big).\end{aligned}\tag{62}$$

Substituting this into (60) and recalling the distribution (39) of X, we arrive at

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

This concludes the proof of Lemma 1 for sufficiently large c_5, as long as Claim 1 can be justified.

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

$$\mathbb{P}(\|X_0-x\|_2\le\epsilon)\le\bigg(\frac{\epsilon}{2T^{c_6}R}\bigg)^d\qquad\text{with }\epsilon=\frac{1}{T^{c_0/2}}.\tag{63}$$

eq:contradition-x-y

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

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

$$\begin{aligned}p_{X_t}(y) &= \int_x p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x\\ &= \int_{x:\|\sqrt{\overline{\alpha}_t}x-y\|_2\ge c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x\end{aligned}$$

$$+\int_{x:\|\sqrt{\overline{\alpha}_t}x-y\|_2<c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x$$

$$\le \int_{x:\|\sqrt{\overline{\alpha}_t}x-y\|_2\ge c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x + \int_{x:\|\sqrt{\overline{\alpha}_t}x-y\|_2<c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x)\mathrm{d}x. \tag{64}$$

eqn:contradiction-12

To further control (64), we make two observations:

1) The first term on the right-hand side of (64) can be bounded by

$$\begin{aligned}
&\int_{x:\|\sqrt{\overline{\alpha}_t}x-y\|_2\ge c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x \\
&\quad= \int_{z:\|z\|_2\ge c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}}\exp\Big(-\frac{\|z\|_2^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}z \\
&\quad\le \frac{1}{2}\exp\big(-c_6 d\log T\big),
\end{aligned} \tag{65}$$

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

2) Regarding the second term on the right-hand side of (64), let us construct an epsilon-net $\mathcal{N}_\epsilon=\{z_i\}$ for the set

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

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

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

where the penultimate step comes from the assumption (63).

The above results taken collectively lead to

$$p_{X_t}(y)\le\exp(-c_6 d\log T), \tag{66}$$

eqn:contradiction

thus contradicting the assumption (38). This in turn validates this claim. □

A.6 Proof of Lemma 2

sec:proof-lem:main-ODE

A.6.1 Proof of relation (45a)

Recall the definition of ϕ_t in (42), and introduce the following vector:

$$u:=x-\phi_t(x)=\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\int_{x_0}(x-\sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0. \tag{67}$$

eq:defn-u-Lemma-main-ODE

The proof consists of the following steps.

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

$$X_t \stackrel{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W \qquad \text{with } W \sim \mathcal{N}(0, I_d) \tag{68}$$

eq:Xt-dist-proof1

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

$$p_{X_0\,|\,X_t}(x_0\,|\,x) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} p_{X_t\,|\,X_0}(x\,|\,x_0) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} \cdot \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\bigg(-\frac{\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)}\bigg). \tag{69}$$

eqn:bayes

In turn, this taken together with (68) allows one to rewrite $p_{\sqrt{\alpha_t}X_{t-1}}$ such that: for any set $\mathcal{E}$,

$$\begin{aligned}
\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} &\stackrel{\text{(i)}}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{\big(2\pi(\alpha_t-\overline{\alpha}_t)\big)^{d/2}} \exp\bigg(-\frac{\big\|\phi_t(x)-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&\stackrel{\text{(ii)}}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{\big(2\pi(\alpha_t-\overline{\alpha}_t)\big)^{d/2}} \exp\bigg(-\frac{\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)}\bigg) \\
&\qquad \cdot \exp\bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&\stackrel{\text{(iii)}}{=} \Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} \cdot \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x) \cdot \\
&\qquad\qquad \exp\bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&\stackrel{\text{(iv)}}{=} \bigg\{1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\bigg)\bigg\} \cdot \\
&\quad \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x) \exp\bigg(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0.
\end{aligned} \tag{70}$$

eqn:fei

Here, identity (i) holds true since $\sqrt{\alpha_t}X_{t-1} \stackrel{\mathrm{d}}{=} \sqrt{\alpha_t}\big(\sqrt{\overline{\alpha}_{t-1}}X_0 + \sqrt{1-\overline{\alpha}_{t-1}}W\big) = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{\alpha_t-\overline{\alpha}_t}W$ and hence

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

identity (ii) follows from elementary algebra; relation (iii) is a consequence of the Bayes rule (69); and relation (iv) results from (41c).

Step 2: controlling the integral in the decomposition (70). In order to further control the right-hand side of expression (70), we need to evaluate the integral in (70). To this end, we make a few observations.

- To begin with, Lemma 1 tells us that [eqn:BBB]

$$\mathbb{P}\Big(\big\|\sqrt{\overline{\alpha}_t}X_0 - x\big\|_2 > 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T}\,\big|\,X_t = x\Big) \le \exp\big(-c_5^2 d\log T\big) \tag{71a}$$

eqn:brahms

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

- A little algebra based on this relation allows one to bound u (cf. (67)) as follows:

$$\begin{aligned}
\|u\|_2 \le{}& \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \int_{x_0:\|x-\sqrt{\overline{\alpha}_t}x_0\|_2 \le c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0 \\
&+ \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \int_{x_0:\|x-\sqrt{\overline{\alpha}_t}x_0\|_2 > c_5\sqrt{d(1-\overline{\alpha}_t)\log T}} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2\mathrm{d}x_0
\end{aligned}$$

$$\le \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \cdot c_5\sqrt{d(1-\overline{\alpha}_t)\log T} + \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \cdot \int_{c_5}^{\infty} \Big(d(1-\overline{\alpha}_t)\log T\Big)\tau \exp\Big(-\frac{1}{25}\tau^2 d\log T\Big)\mathrm{d}\tau$$
$$\le \frac{2c_5(1-\alpha_t)\sqrt{d\log T}}{3\sqrt{1-\overline{\alpha}_t}} \le \frac{2c_5}{3}\sqrt{d(1-\alpha_t)\log T}, \tag{71b}$$

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

Equipped with the above properties, let us define

$$\mathcal{E} \coloneqq \Big\{x : \big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2 \le 5c_5\sqrt{d(1-\overline{\alpha}_t)\log T}\Big\}. \tag{72}$$

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

$$\exp\left(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\right)$$
$$= 1 - \frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)} + \frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg)$$
$$= 1 - \frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} + \frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t} + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg), \tag{73}$$

where the penultimate line invokes (71), and the last line holds true since, according to (71b),

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

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

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

and as a result,

$$\exp\left(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\right) \le \exp\left(\frac{\|u\|_2^2}{2(1-\alpha_t)}\right) \le \exp\left(\frac{2c_5^2 d\log T}{9}\right). \tag{74}$$

Combine (73) and (74) to show that: if $\frac{d(1-\alpha_t)\log T}{\alpha_t-\overline{\alpha}_t} \lesssim 1$, then one has

$$\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\left(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\right)\mathrm{d}x_0$$
$$= \left(\int_{x_0\in\mathcal{E}} + \int_{x_0\notin\mathcal{E}}\right) p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\left(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\right)\mathrm{d}x_0$$
$$= \left(1 + O\Big(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\Big)\right)\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\left(1 - \frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} + \frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t}\right)\mathrm{d}x_0$$
$$+ O\left(\exp\left(\frac{2c_5^2 d\log T}{9}\right)\int_{x_0\notin\mathcal{E}} p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\right)$$

$$
\begin{aligned}
&= 1 - \frac{(1-\alpha_t)\big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 \mathrm{d}x_0 - \big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2\big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} \\
&\quad + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg) + O\left(\exp\left(-\frac{7c_5^2 d\log T}{9}\right)\right) \\
&= 1 - \frac{(1-\alpha_t)\big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 \mathrm{d}x_0 - \big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2\big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} \\
&\quad + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T\bigg),
\end{aligned} \tag{75}
$$

where the last relation makes use of the definition (67) of u.

Step 3: putting everything together. Substitution of (75) into (70) yields

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

as claimed.

A.6.2 Proof of relation (45b)

To understand the density ratio $p_{\phi_t(X)}(\phi_t(x))/p_X(x)$, we make note of the transformation

$$
p_{\phi_t(X)}(\phi_t(x)) = \mathsf{det}\Big(\frac{\partial\phi_t(x)}{\partial x}\Big)^{-1} p_X(x), \tag{76}
$$

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

To begin with, elementary calculations of the derivative — in conjunction with the fact that $X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W$ with $W\sim\mathcal{N}(0,I_d)$ — yield that: for any $1\le i,j\le d$,

$$
\begin{aligned}
&\frac{\partial\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_i-\sqrt{\overline{\alpha}_t}x_{0,i}\big)\mathrm{d}x_0}{\partial x_j} \\
&= \mathbb{1}\{i=j\} + \frac{1}{1-\overline{\alpha}_t}\bigg\{\bigg(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_i-\sqrt{\overline{\alpha}_t}x_{0,i}\big)\mathrm{d}x_0\bigg)\bigg(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_j-\sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\bigg) \\
&\qquad - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_i-\sqrt{\overline{\alpha}_t}x_{0,i}\big)\big(x_j-\sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\bigg\} \\
&= \mathbb{1}\{i=j\} + \frac{1}{1-\overline{\alpha}_t}\Big\{\mathbb{E}\big[g_i(X_0)\mid X_t=x\big]\mathbb{E}\big[g_j(X_0)\mid X_t=x\big] - \mathbb{E}\big[g_i(X_0)g_j(X_0)\mid X_t=x\big]\Big\},
\end{aligned} \tag{77}
$$

where $g_i(x_0) \coloneqq x_i - \sqrt{\overline{\alpha}_t}x_{0,i}$. Combining (77) with the definition of ϕ_t (cf. (42)), one can readily see that

$$
\begin{aligned}
\mathsf{Tr}\Big(I - \frac{\partial\phi_t(x)}{\partial x}\Big) &= \frac{d(1-\alpha_t)}{2(1-\overline{\alpha}_t)} + \\
&\quad \frac{(1-\alpha_t)\big(\big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2 - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\big)}{2(1-\overline{\alpha}_t)^2}.
\end{aligned} \tag{78a}
$$

Moreover, define a matrix $\boldsymbol{B} = [B_{i,j}]_{1\le i,j\le d}$ with

$$
B_{i,j} \coloneqq \mathbb{E}\big[g_i(X_0)g_j(X_0)\mid X_t=x\big] - \mathbb{E}\big[g_i(X_0)\mid X_t=x\big]\mathbb{E}\big[g_j(X_0)\mid X_t=x\big],
$$

which clearly satisfies (using Jensen's inequality)

$$\begin{aligned}\|\boldsymbol{B}\|_{\mathrm{F}} &= \left\| \mathbb{E}\left[\begin{bmatrix} g_1(X_0) \\ \vdots \\ g_d(X_0) \end{bmatrix} \begin{bmatrix} g_1(X_0) & \cdots & g_d(X_0) \end{bmatrix} \mid X_t = x \right] \right\|_{\mathrm{F}} \\ &\le \mathbb{E}\left[\left\| \begin{bmatrix} g_1(X_0) \\ \vdots \\ g_d(X_0) \end{bmatrix} \begin{bmatrix} g_1(X_0) & \cdots & g_d(X_0) \end{bmatrix} \mid X_t = x \right\|_{\mathrm{F}} \right] = \mathbb{E}\left[\sum_{i=1}^{d} \big(g_i(X_0)\big)^2 \mid X_t = x \right] \\ &= \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \big\| x - \sqrt{\overline{\alpha}_t} x_0 \big\|_2^2 \mathrm{d}x_0.\end{aligned}$$

Taking (77) and (42) together and using the above bound on $\|\boldsymbol{B}\|_{\mathrm{F}}$ also reveal that

$$\begin{aligned}\left\| \frac{\partial \phi_t(x)}{\partial x} - I \right\| \le \left\| \frac{\partial \phi_t(x)}{\partial x} - I \right\|_{\mathrm{F}} &\lesssim \frac{1-\alpha_t}{1-\overline{\alpha}_t}\left(\sqrt{d} + \frac{1}{1-\overline{\alpha}_t}\|\boldsymbol{B}\|_{\mathrm{F}} \right) \\ &\lesssim \frac{1-\alpha_t}{1-\overline{\alpha}_t}\left(\sqrt{d} + \frac{\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \big\| x - \sqrt{\overline{\alpha}_t} x_0 \big\|_2^2 \mathrm{d}x_0}{1-\overline{\alpha}_t} \right) \\ &\lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t}, \end{aligned} \tag{78b}$$

where the last line invokes the relation (71a). Additionally, the Taylor expansion guarantees that

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

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

$$\begin{aligned}&\frac{p_{\phi_t(X)}(\phi_t(x))}{p_X(x)} = \mathsf{det}\Big(\frac{\partial \phi_t(x)}{\partial x}\Big)^{-1} \\ &= 1 - \mathsf{Tr}\Big(\frac{\partial \phi_t(x)}{\partial x} - I\Big) + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3 \log^3 T \bigg) \\ &= 1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + \frac{(1-\alpha_t)\Big(\big\| \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t} x_0\big)\mathrm{d}x_0 \big\|_2^2 - \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \big\| x - \sqrt{\overline{\alpha}_t} x_0 \big\|_2^2 \mathrm{d}x_0 \Big)}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} \\ &\quad + O\bigg(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \log^2 T + d^6\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3 \log^3 T \bigg), \end{aligned} \tag{79}$$

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

A.7 Proof of Lemma 3

sec:proof-lem:main-ODE-fast

The proof of Lemma 3 is derived in the same way as the proof of Lemma 2.

A.7.1 Proof of relation (53a)

First let us recall that

$$\begin{aligned}p_{\sqrt{\alpha_t} X_{t-1}}(\varphi_t(x)) &= p_{X_t}(x)\Big(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\Big)^{d/2} \\ &\quad \cdot \int_{x_0} \mathrm{d}x_0 p_{X_0 \,|\, X_t}(x_0 \,|\, x) \exp\Big(- \frac{(1-\alpha_t)\big\| x - \sqrt{\overline{\alpha}_t} x_0 \big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t} x_0\big)}{2(\alpha_t - \overline{\alpha}_t)} \Big), \end{aligned} \tag{80}$$

where $u = x - \varphi_t(x) = \frac{\lambda_t(x)(1-\alpha_t)}{2(1-\overline{\alpha}_t)} \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t} x_0\big)\mathrm{d}x_0$. Moreover, according to Lemma 1, we have $\big\|x - \sqrt{\overline{\alpha}_t} x_0\big\| \lesssim \sqrt{d \log T(1-\overline{\alpha}_t)}$, and then

$$
\begin{aligned}
&\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \exp\Big(- \frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t} x_0\big)}{2(\alpha_t - \overline{\alpha}_t)}\Big)\mathrm{d}x_0 \\
&= \Big(1 + O\Big(d^3 \log^3 T\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\Big)\Big) \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\Big(1 - \frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t} x_0\big)}{2(\alpha_t - \overline{\alpha}_t)} \\
&\qquad + \frac{1}{2}\Big(\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t} x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{u^\top\big(x - \sqrt{\overline{\alpha}_t} x_0\big)}{\alpha_t - \overline{\alpha}_t}\Big)^2\Big)\mathrm{d}x_0 \\
&= 1 - \frac{(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)}\Big(A_t - \lambda_t B_t - \frac{1-\alpha_t}{4(\alpha_t - \overline{\alpha}_t)}\big[-B_t + C_t + D_t - 2E_t\big]\Big) + O\Big(d^3 \log^3 T\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\Big), \qquad (81)
\end{aligned}
$$

where in the last step, we plug in the definitions in Assumption 2. Therefore, the relation (53a) can be easily seen by recognizing the fact that

$$
\Big(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\Big)^{d/2} = 1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + \frac{d(d-2)(1-\alpha_t)^2}{4(\alpha_t - \overline{\alpha}_t)^2} + O\Big(d^3\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\Big).
$$

A.7.2 Proof of relation (53b)

Recall expression (77) that

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

which implies that

$$
\mathsf{Tr}\Big(\frac{\partial \phi_t(x)}{\partial x} - I\Big) = -\frac{\lambda_t(1-\alpha_t)\big(d + B_t - A_t\big)}{2(1-\overline{\alpha}_t)}, \qquad (82a)
$$

$$
\Big\|\frac{\partial \phi_t(x)}{\partial x} - I\Big\|_{\mathsf{F}}^2 = \frac{\lambda_t^2(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\Big(d + 2(B_t - A_t) + \sum F_{ij}^2\Big), \qquad (82b)
$$

$$
\Big\|\frac{\partial \phi_t(x)}{\partial x} - I\Big\|_{\mathsf{op}} \lesssim \frac{d \log T(1-\alpha_t)}{1-\overline{\alpha}_t}. \qquad (82c)
$$

Here, we denote

$$
\begin{aligned}
F_{ij} \coloneqq &\frac{1}{1-\overline{\alpha}_t}\Big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x_i - \sqrt{\overline{\alpha}_t} x_{0,i}\big)\mathrm{d}x_0\Big)\Big(\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x_j - \sqrt{\overline{\alpha}_t} x_{0,j}\big)\mathrm{d}x_0\Big) \\
&- \frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x_i - \sqrt{\overline{\alpha}_t} x_{0,i}\big)\big(x_j - \sqrt{\overline{\alpha}_t} x_{0,j}\big)\mathrm{d}x_0. \qquad (83)
\end{aligned}
$$

Now in view of the basic relation

$$
\mathsf{det}(I + A)^{-1} = 1 - \mathsf{Tr}(A) + \frac{1}{2}\big[\mathsf{Tr}(A)^2 + \|A\|_{\mathsf{F}}^2\big] + O\big(d^3 \|A\|_{\mathsf{op}}^3\big), \qquad (84)
$$

for any symmetric matrix A, we can derive

$$
p_{\phi_t(X)}(\phi_t(x)) = \mathsf{det}\Big(\frac{\partial \phi_t(x)}{\partial x}\Big)^{-1} p_X(x)
$$

$$
\begin{aligned}
&= \Big(1 - \mathsf{Tr}\Big(\frac{\partial \phi_t(x)}{\partial x} - I\Big) + \frac{1}{2}\Big[\mathsf{Tr}\Big(\frac{\partial \phi_t(x)}{\partial x} - I\Big)^2 + \Big\|\frac{\partial \phi_t(x)}{\partial x} - I\Big\|_F^2\Big] \\
&\qquad + \frac{(1-\alpha_t)}{2(1-\overline{\alpha}_t)} \nabla \lambda_t(x)^\top \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t} x_0\big) \mathrm{d}x_0 + O\Big(d^6 \log^3 T \Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\Big)\Big) p_X(x) \\
&= \Big(1 + \frac{\lambda_t d(1-\alpha_t)}{2(1-\overline{\alpha}_t)} + \frac{\lambda_t (1-\alpha_t)(B_t - A_t)}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\Big[\Big(d + B_t - A_t\Big)^2 + d + 2(B_t - A_t) + \sum F_{ij}^2\Big] \\
&\qquad + \frac{(1-\alpha_t)}{2(1-\overline{\alpha}_t)} \nabla \lambda_t(x)^\top \int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t} x_0\big) \mathrm{d}x_0 + O\Big(d^6 \log^3 T \Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\Big)\Big) p_X(x),
\end{aligned} \tag{85}
$$

provided that $d^2 \log T\big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\big) \lesssim 1$. We thus complete the proof of Lemma 3.

B Analysis for the stochastic samplers (Theorems 3 and 4)

sec:analysis-stochastic-samplers

The proofs of Theorem 3 and Theorem 4 share the same structure, therefore in the following, we shall treat them simultaneously. We point out their differences as we move along.

In order to compute the KL divergence between p_{X_1} and p_{Y_1}. Let us recall the basic inequality that

$$
\begin{aligned}
\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1}) &\le \mathsf{KL}(p_{X_1,\ldots,X_T} \,\|\, p_{Y_1,\ldots,Y_T}) \\
&= \mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) + \sum_{t=2}^{T} \mathbb{E}_{X_t}\big[\mathsf{KL}(p_{X_{t-1} \,|\, X_t} \,\|\, p_{Y_{t-1} \,|\, Y_t})\big].
\end{aligned} \tag{86}
$$

eqn:kl-decomp

Recall that in the forward process (4), one has

$$
X_T = \sqrt{\overline{\alpha}_T} X_0 + \sqrt{1 - \overline{\alpha}_T}\, \overline{W}_t, \qquad \overline{W}_t \sim \mathcal{N}(0, I_d).
$$

By direct calculations, we find

$$
\begin{aligned}
\mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) &= \int p_{X_T}(x) \log \frac{p_{X_T}(x)}{p_{Y_T}(x)} \mathrm{d}x \\
&\overset{\text{(i)}}{=} \int p_{X_T}(x) \log \frac{\int_{y+z=x, |y| \le 1/T^c} p_{\sqrt{\overline{\alpha}_T} X_0}(y) p_{\sqrt{1-\overline{\alpha}_T}\, \overline{W}_t}(z) \mathrm{d}y \mathrm{d}z}{p_{Y_T}(x)} \mathrm{d}x \\
&\le \int p_{X_T}(x) \log \frac{\max_{y+z=x, |y| \le 1/T^c} p_{\sqrt{1-\overline{\alpha}_T}\, \overline{W}_t}(z)}{p_{Y_T}(x)} \mathrm{d}x \\
&\le \int p_{X_T}(x) \Big(- d/2 \log(1 - \overline{\alpha}_T) - \frac{\overline{\alpha}_T \|x\|_2^2 + \|x\|_2/T^c + 1/T^{2c}}{2(1-\overline{\alpha}_T)} \Big) \mathrm{d}x \\
&= O(\overline{\alpha}_T d).
\end{aligned} \tag{87}
$$

eqn:KL-T

where in (i), we recall that $\|X_0\|_2 \lesssim T^{c_3}$ while $\overline{\alpha}_T \lesssim T^{-c_2}$. It thus suffices for us to control $\mathsf{KL}(p_{X_{t-1} \,|\, X_t} \,\|\, p_{Y_{t-1} \,|\, Y_t})$ for each t.

B.1 Step 1: approximation of $p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t)$

Given any constant $\gamma \in [0, 1]$, let us write

$$
x_t(\gamma) = \gamma x_{t-1} + (1 - \gamma) \widehat{x}_t \qquad \text{and} \qquad \widehat{x}_t = x_t / \sqrt{\alpha_t}. \tag{88}
$$

With this piece of notation, we proceed to computing $p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t)$. Some direct calculations regarding the forward process (4) give

$$
\begin{aligned}
&p_{X_{t-1} \,|\, X_t}(x_{t-1} \,|\, x_t) \\
&\propto p_{X_{t-1}, X_t}(x_{t-1}, x_t) \propto \exp(\log p_{X_{t-1}}(x_{t-1}) + \log p_{X_t \,|\, X_{t-1}}(x_t \,|\, x_{t-1}))
\end{aligned}
$$

$$= \exp\Big(\log p_{X_{t-1}}(\widehat{x}_t) + \int_0^1 \nabla \log p_{X_{t-1}}(x_t(\gamma))^\top (x_{t-1}-\widehat{x}_t)\mathrm{d}\gamma + \log p_{X_t\,|\,X_{t-1}}(x_t\,|\,x_{t-1})\Big)$$

$$\propto \exp\Big((x_{t-1}-\widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}\,|\,X_0}(x_t(\gamma)\,|\,x_0)p_{X_0}(x_0)}{p_{X_{t-1}}(x_t(\gamma))}\mathrm{d}x_0 + \log p_{X_t\,|\,X_{t-1}}(x_t\,|\,x_{t-1})\Big). \tag{89}$$

eqn:sde-cond

To continue, in view of the mean value theorem, there exists some $\widetilde{x}_t = x_t(\widetilde{\gamma})$ such that

$$\begin{aligned}
&(x_{t-1}-\widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}\,|\,X_0}(x_t(\gamma)\,|\,x_0)p_{X_0}(x_0)}{p_{X_{t-1}}(x_t(\gamma))}\mathrm{d}x_0 \\
&= (x_{t-1}-\widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}\,|\,X_0}(x_t(\gamma)\,|\,x_0)}{p_{X_{t-1}\,|\,X_0}(x_t(\gamma)\,|\,x_0)} p_{X_0\,|\,X_{t-1}}(x_0\,|\,x_t(\gamma))\mathrm{d}x_0 \\
&= -(x_{t-1}-\widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{x_t(\gamma)-\sqrt{\overline{\alpha}_{t-1}}x_0}{1-\overline{\alpha}_{t-1}} p_{X_0\,|\,X_{t-1}}(x_0\,|\,x_t(\gamma))\mathrm{d}x_0 \\
&=: -(x_{t-1}-\widehat{x}_t)^\top \int_0^1 F(\gamma)\mathrm{d}\gamma \qquad (90) \\
&= -\frac{(x_{t-1}-\widehat{x}_t)^\top \int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(\widehat{x}_t-\sqrt{\overline{\alpha}_{t-1}}x_0)\mathrm{d}x_0 + \frac{1}{2}(x_{t-1}-\widehat{x}_t)^\top J_{t-1}(\widetilde{x}_t)(x_{t-1}-\widehat{x}_t)}{1-\overline{\alpha}_{t-1}}. \qquad (91)
\end{aligned}$$

eqn:schubert

Here, we compute the Jacobian matrix as in (77) where

$$\begin{aligned}
J_{t-1,ij} &:= \frac{\partial F(\gamma)}{\partial x_j} = \frac{\partial \int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,x)\big(x_i-\sqrt{\overline{\alpha}_t}x_{0,i}\big)\mathrm{d}x_0}{\partial x_j} \\
&= \delta_{ij} + \frac{1}{1-\overline{\alpha}_t}\Big(\Big(\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,x)\big(x_i-\sqrt{\overline{\alpha}_t}x_{0,i}\big)\mathrm{d}x_0\Big)\Big(\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,x)\big(x_j-\sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\Big) \\
&\qquad - \int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,x)\big(x_i-\sqrt{\overline{\alpha}_t}x_{0,i}\big)\big(x_j-\sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\Big) \\
&= \delta_{ij} - \frac{1}{1-\overline{\alpha}_t}\Big(\mathbb{E}_{X_0}[X_0X_0^\top\,|\,X_{t-1}=x] - \mathbb{E}_{X_0}[X_0\,|\,X_{t-1}=x]\cdot\mathbb{E}_{X_0}[X_0\,|\,X_{t-1}=x]^\top\Big). \qquad (92)
\end{aligned}$$

eqn:derivative-2

We prove at the end of this section that if we define set

$$\mathcal{E} := \Big\{(x_t,x_{t-1})\,|\,-\log p_{X_t}(x_t) \lesssim d\log T,\ \|x_{t-1}-x_t/\sqrt{\alpha_t}\| \lesssim \sqrt{d\log T(1-\alpha_t)}\Big\}, \tag{93}$$

eqn:eset

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

eqn:unfinished

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \propto \exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\Big\|\Big(I-\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(\widehat{x}_t)\Big)^{-1}(x_{t-1}-\mu_t)\Big\|^2 + O\Big(d^3\log^3 T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^{3/2}\Big)\Big) \tag{94a}$$

eq:cond-dist

$$\propto \exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\|x_{t-1}-\mu_t\|^2 + O\Big(d^2\log^2 T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\Big)\Big). \tag{94b}$$

eq:cond-dist-crude

Here, μ_t is given by

$$\mu_t := \widehat{x}_t - \frac{(1-\alpha_t)\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}. \tag{95}$$

eqn:nu-t-2

We remind the readers that the original DDPM (i.e. Theorem 3) approximates the conditional distribution of $X_{t-1}\,|\,X_t$ via the form (94b) while the accelerated DDPM (i.e. Theorem 4) approximates using the form as in (94a).

B.2 Step 2: uniform control of the density ratios

Recalling the distribution of Y_{t-1} (cf. (30)) to obtain

$$p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t) \propto \exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\Big\|\Big(I-\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(\widehat{x}_t)\Big)^{-1}(x_{t-1}-\mu_t)\Big\|^2\Big), \tag{96}$$

with μ and J defined as in expressions (95) and (92). In this section, we prove that

$$\log \frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} \lesssim \mathsf{poly}(T)(\|\widehat{x}_t\|^2 + \|x_{t-1}-\widehat{x}_t\|^2) + \mathsf{poly}(T). \tag{97}$$

According to Cauchy's inequality, one can easily check that $J_t(\widehat{x}_t) \preceq I$, and

$$J_t(\widehat{x}_t) \succ -\frac{1}{1-\overline{\alpha}_{t-1}} \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x) \big\|\widehat{x}_t - \sqrt{\overline{\alpha}_t}x_0\big\|^2 I \mathrm{d}x_0 \succ -\frac{2(\|\widehat{x}_t\|^2 + \mathsf{poly}(T))}{1-\overline{\alpha}_t} I, \tag{98}$$

where we recall that $\|X_0\|_2 \le \mathsf{poly}(T)$. If we write

$$Y_{t-1}\,|\,Y_t = x_t \sim \mathcal{N}(\mu_t, U_t M_t U_t^\top),$$

where $\{M_{t,i}\}$ denote eigenvalues of covariance matrix $\frac{1-\alpha_t}{\alpha_t}(I - (1-\alpha_t)/2(1-\overline{\alpha}_t)J_t(\widehat{x}_t))$, in view of the above properties of J_t, one has

$$M_{t,i} \gtrsim 1-\alpha_t, \qquad \text{and} \qquad \log M_{t,i} \lesssim \log(\|\widehat{x}_t\| + \mathsf{poly}(T)).$$

Given these relations, we are ready to bound the density function $p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)$. Recall that the basic relation that for $Z \sim \mathcal{N}(0,\Sigma)$

$$\log \frac{1}{p_Z(z)} = \frac{z^\top \Sigma^{-1} z}{2} + \frac{1}{2}\log \mathsf{det}(\Sigma) + \frac{d}{2}\log(2\pi),$$

which leads to

$$\begin{aligned}
\log \frac{1}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)} &\lesssim \frac{\|x_{t-1}-\mu_t\|^2}{1-\alpha_t} + d\log(\|\widehat{x}_t\| + \mathsf{poly}(T)) \\
&\lesssim \mathsf{poly}(T)(\|\widehat{x}_t\|^2 + \|x_{t-1}-\widehat{x}_t\|^2) + \mathsf{poly}(T),
\end{aligned} \tag{99}$$

Here, the last inequality uses

$$\begin{aligned}
\|x_{t-1}-\mu_t\|^2 &\le 2\|x_{t-1}-\widehat{x}_t\|^2 + 2\|\widehat{x}_t - \mu_t\|^2 \\
&\lesssim \|x_{t-1}-\widehat{x}_t\|^2 + \frac{(1-\alpha_t)}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)} \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t) \|\sqrt{\overline{\alpha}_t}\widehat{x}_t - \sqrt{\overline{\alpha}_t}x_0\|^2 \mathrm{d}x_0 \\
&\le \mathsf{poly}(T)(\|x_{t-1}-\widehat{x}_t\|^2 + \|\widehat{x}_t\|^2) + \mathsf{poly}(T)
\end{aligned}$$

for $x_t = \sqrt{\overline{\alpha}_t}\widehat{x}_t$ and $\|X_0\|_2 \le \mathsf{poly}(T)$.

In addition, since $X_t\,|\,X_{t-1}$ follows a Gaussian distribution, we can write

$$\begin{aligned}
\log p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) &= \log \frac{p_{X_t\,|\,X_{t-1}}(x_t\,|\,x_{t-1}) p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} \\
&= \log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} + \frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|^2}{2(1-\alpha_t)} - d/2\log(2\pi(1-\alpha_t)) \\
&\le \log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} + \mathsf{poly}(T)(\|\widehat{x}_t\|^2 + \|x_{t-1}-\widehat{x}_t\|^2).
\end{aligned}$$

Some direct calculations yield

$$\begin{aligned}
\log \frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} &= \log \frac{\int_{x_0} p_{X_0}(x_0) \exp\Big(-\frac{\|x_{t-1}-\sqrt{\overline{\alpha}_{t-1}}x_0\|^2}{2(1-\overline{\alpha}_{t-1})}\Big)\mathrm{d}x_0}{\int_{x_0} p_{X_0}(x_0)\exp\Big(-\frac{\|x_t-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}x_0} - \frac{d}{2}\log\Big(\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\Big) \\
&\le \max_{x_0}\Big\{\|x_t - \sqrt{\overline{\alpha}_t}x_0\|^2 - \|x_{t-1}-\sqrt{\overline{\alpha}_t}x_0\|^2\Big\} - \frac{d}{2}\log\Big(\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\Big) \\
&\le \overline{\alpha}_t\|\widehat{x}_t - x_0\|^2 \le \|x_{t-1}\|^2 + \mathsf{poly}(T).
\end{aligned}$$

Combining the above two relations together, we arrive at

$$\log p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \lesssim \mathsf{poly}(T)(\|\widehat{x}_t\|^2 + \|x_{t-1}-\widehat{x}_t\|^2) + \mathsf{poly}(T), \tag{100}$$

Putting everything together, we complete the proof of (97).

B.3 Step 3: computing the KL divergence

Based on the analyses laid out previously, we are ready to compute the KL divergence between $p_{X_{t-1}\,|\,X_t}$ and $p_{Y_{t-1}\,|\,Y_t}$. In view of relations (94a), (97) and set (93), we decompose the integration as

$$\begin{aligned}
&\mathbb{E}_{X_t}\big[\mathsf{KL}(p_{X_{t-1}\,|\,X_t}\,\|\,p_{Y_{t-1}\,|\,Y_t})\big] \\
&= \Big(\int_{\mathcal{E}} + \int_{\mathcal{E}^{\mathrm{c}}}\Big) p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\log\frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)}\mathrm{d}x_{t-1}\mathrm{d}x_t, \\
&\overset{\text{(i)}}{=} \int_{\mathcal{E}} p_{X_t}(x_t)\Big(p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) - p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t) + p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\cdot O\Big(d^6\log^6 T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\Big)\Big)\mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\quad + \int_{\mathcal{E}^{\mathrm{c}}} p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\big(\mathsf{poly}(T)(\|\widehat{x}_t\|^2 + \|x_{t-1}-\widehat{x}_t\|^2) + \mathsf{poly}(T)\big)\mathrm{d}x_{t-1}\mathrm{d}x_t. \qquad \texttt{eqn:serenade} \quad (101)
\end{aligned}$$

Here, expression (i) makes use of the facts that if $\big|\frac{p_Y(x)}{p_X(x)}-1\big| < \frac{1}{2}$,

$$\begin{aligned}
p_X(x)\log\frac{p_X(x)}{p_Y(x)} &= -p_X(x)\log\Big(1+\frac{p_Y(x)-p_X(x)}{p_X(x)}\Big) \\
&= p_X(x) - p_Y(x) + p_X(x)O\Big(\Big(\frac{p_Y(x)}{p_X(x)}-1\Big)^2\Big). \qquad (102)
\end{aligned}$$

In particular, (101) holds true when we instantiate the above inequality to $p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)$ and $p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)$, and recall expression (94a) that

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)\cdot\exp\Big(O\big(d^3\log^3 T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^{3/2}\big)\Big).$$

To continue, let us bound each term on the right hand side of (101) respectively. Given set $\mathcal{E}$ (cf. 93), direct calculations yield

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

and similarly,

$$\int_{\mathcal{E}^{\mathrm{c}}} p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\big(\mathsf{poly}(T)(\|\widehat{x}_t\|^2 + \|x_{t-1}-\widehat{x}_t\|^2) + \mathsf{poly}(T)\big)\mathrm{d}x_{t-1}\mathrm{d}x_t \le \exp(-\Omega(d\log T)).$$

In addition, for every x_t obeying $-\log p_{X_t}(x_t) \lesssim d\log T$, it satisfies that

$$\int_{\|x_{t-1}-x_t/\sqrt{\alpha_t}\|^2 \gtrsim \sqrt{d\log T(1-\alpha_t)}} p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t)\mathrm{d}x_{t-1} \le \exp(-\Omega(d\log T)),$$

by recalling that

$$p_{Y_{t-1}\,|\,Y_t}(x_{t-1}\,|\,x_t) \propto \exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\Big\|\Big(I-\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(\widehat{x}_t)\Big)^{-1}(x_{t-1}-\mu_t)\Big\|^2\Big), \qquad (103)$$

where $\|J_t(\widehat{x}_t)\|_{\mathsf{op}} \lesssim d\log T$ and $\|\mu_t-\widehat{x}_t\|_2 \lesssim (1-\alpha_t)\sqrt{d\log T/(1-\overline{\alpha}_t)}$.

Taking everything collectively, for each $t\ge 2$, the right hand side of (101) satisfies

$$\mathbb{E}_{X_t}\big[\mathsf{KL}(p_{X_{t-1}\,|\,X_t}\,\|\,p_{Y_{t-1}\,|\,Y_t})\big] = O\Big(d^6\log^6 T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\Big). \qquad \texttt{eqn:each-kl} \quad (104)$$

B.4 In summary

To summarize, let us recall the decomposition (86) and the control of each term (104) to obtain

$$\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1}) \le \mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) + \sum_{2\le t\le T} d^6 \log^6 T \Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3$$

Putting the above together with our choice of learning rates (cf. (21)) and the bound of $\mathsf{KL}(p_{X_T} \,\|\, p_{Y_T})$ as in (87), yields

$$\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1}) \lesssim \frac{d^6 \log^9 T}{T^2}.$$

which completes the claimed bound (34) by virtue of Pinsker's inequality.

Similarly, if we instead consider the approximation (94b) concerning the original DDPM, the first term in (101) will be replaced by $O(d^2 \log^2 T(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}))$. As a consequence, we can derive

$$\begin{aligned}\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1}) &\lesssim \mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) + \sum_{t\ge 2}^{T} d^4 \log^4 T \Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \\ &\lesssim \frac{d^4 \log^6 T}{T}, \end{aligned} \tag{105}$$

which again leads to (31).

B.5 Proof of Claim (94)

In order to prove these relations, note that putting together (89) and (91) leads to

$$\begin{aligned} &p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \\ &\propto \exp\Big((x_{t-1}-\widehat{x}_t)^\top \int_0^1 \mathrm{d}\gamma \int_{x_0} \frac{\nabla p_{X_{t-1}\,|\,X_0}(\widetilde{x}_t\,|\,x_0) p_{X_0}(x_0)}{p_{X_{t-1}}(\widetilde{x}_t)} \mathrm{d}x_0 + \log p_{X_t\,|\,X_{t-1}}(x_t\,|\,x_{t-1})\Big) \\ &\propto \exp\Big(- \frac{(x_{t-1}-\widehat{x}_t)^\top \int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0)\mathrm{d}x_0 + \frac{1}{2}(x_{t-1}-\widehat{x}_t)^\top J_{t-1}(\widetilde{x}_t)(x_{t-1}-\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} \\ &\qquad - \frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|^2}{2(1-\alpha_t)}\Big). \end{aligned} \tag{106}$$

eqn:allegro

If we can show that for every $\gamma \in [0,1]$,

eq:Jacobi

$$\|J_{t-1}(x_t(\gamma)) - I\| \lesssim d\log T, \tag{107a}$$

eq:Jacobi-a

$$\|J_{t-1}(x_t(\gamma)) - J_{t-1}(\widehat{x}_t)\| \lesssim d^2 \log^2 T \sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}, \tag{107b}$$

eq:Jacobi-b

and

eq:approx-t

$$\Big\| \frac{\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_{t-1}} - \frac{\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_t} \Big\| \lesssim (1-\alpha_t)\Big(\frac{d\log T}{\alpha_t - \overline{\alpha}_t}\Big)^{3/2}, \tag{108a}$$

eq:approx-t-a

$$\Big\| \frac{J_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{J_t(\widehat{x}_t)}{1-\overline{\alpha}_t} \Big\| \lesssim \frac{d^2 \log^2 T(1-\alpha_t)}{(\alpha_t - \overline{\alpha}_t)^2}, \tag{108b}$$

eq:approx-t-b

the relation (94) shall follow immediately by recognizing that $\|x_{t-1} - \widehat{x}_t\|_2 \lesssim \sqrt{d\log T(1-\alpha_t)}$.

B.5.1 Preliminaries

For $(x_t, x_{t-1}) \in \mathcal{E}$, it holds that

$$-\log p_{X_t}(x_t) \lesssim d\log T, \quad \text{and} \quad \|x_{t-1} - \widehat{x}_t\| \lesssim \sqrt{d\log T(1-\alpha_t)}. \tag{109}$$

Based on this relation, we derive that

$$-\log p_{X_{t-1}}(x_t(\gamma)) \lesssim d\log T. \tag{110}$$

Proof of relation (110). For every x such that $\|\widehat{x}_t - x\| \lesssim \sqrt{d\log T(1-\alpha_t)}$, consider the density ratio

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

If we define set

$$\widetilde{\mathcal{E}} := \{x_0 : \|x_t - \sqrt{\overline{\alpha}_t}x_0\| \lesssim \sqrt{d\log T(1-\overline{\alpha}_t)}\},$$

for every $x_0 \in \widetilde{\mathcal{E}}$, we find

$$\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} = 1 + O\Big(d\log T\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\Big). \tag{111}$$

Here we remind the readers on our choice of step sizes in (21). Otherwise, for $x_0 \notin \widetilde{\mathcal{E}}$,

$$\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} = \exp\Big(O\Big(\frac{\log T\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|^2}{T(1-\overline{\alpha}_t)}\Big)\Big).$$

Based on the above bounds, some direct calculations yield

$$\begin{aligned}p_{X_{t-1}}(x) &= \int_{x_0} p_{X_0}(x_0)p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)\mathrm{d}x_0\\ &= \int_{x_0\in\widetilde{\mathcal{E}}} p_{X_0}(x_0)p_{X_t\,|\,X_0}(x_t\,|\,x_0)\cdot\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)}\mathrm{d}x_0\\ &\quad + \int_{x_0\in\widetilde{\mathcal{E}}^c} p_{X_0}(x_0)p_{X_t\,|\,X_0}(x_t\,|\,x_0)\cdot\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)}\mathrm{d}x_0\\ &= p_{X_t}(x_t)\int_{x_0\in\widetilde{\mathcal{E}}}\Big(1+O\Big(d\log T\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\Big)\Big)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\\ &\quad + p_{X_t}(x_t)\int_{x_0\in\widetilde{\mathcal{E}}^c}\exp\Big(O\Big(\frac{\log T\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|^2}{T(1-\overline{\alpha}_t)}\Big)\Big)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0.\end{aligned} \tag{112}$$

Recall that in Lemma 1, we showed that if $-\log p_{X_t}(x_t) \lesssim d\log T$, For random variable $X = X_0\,|\,X_t = x_t$, it holds with probability at least $1-\exp(-\Omega(d\log T)$ that

$$\|\sqrt{\overline{\alpha}_t}X - y\| \lesssim \sqrt{d\log T(1-\overline{\alpha}_t)}. \tag{113}$$

As a result, the right hand side of (112) equals to

$$p_{X_{t-1}}(x) = \Big(1+O\Big(d\log T\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\Big)\Big)\big(1+\exp(-\Omega(d\log T))\big)p_{X_t}(x_t) + \exp(-\Omega(d\log T))p_{X_t}(x_t)$$

$$= \Big(1 + O\Big(d \log T \sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\Big)\Big) p_{X_t}(x_t). \tag{114}$$ eqn:prokofiev

Therefore, relation (110) follows immediately as $-\log p_{X_t}(x_t) \lesssim d \log T$.

Combining relation (110) with Lemma 1, with probability at least $1 - \exp(-\Omega(d \log T))$, it satisfies that

$$X = X_0 \,|\, X_{t-1} = x_t(\gamma), \qquad \sqrt{\overline{\alpha}_{t-1}} X = \widehat{x}_t + O(\sqrt{d \log T (1-\overline{\alpha}_{t-1})}). \tag{115}$$ eqn:condi-g

B.5.2 Main analysis

Proof of relation (107a). As a consequence of property (115), inequality (107a) holds true by noticing that

$$\|J_{t-1}(x_t(\gamma)) - I\|_{\mathsf{op}} \le \frac{2 \int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, x_t(\gamma)) \big\| x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}} x_0 \big\|^2 \mathrm{d}x_0}{1 - \overline{\alpha}_{t-1}} \lesssim d \log T,$$

where similar to the proof of (112), the integration is computed on the set $\|\sqrt{\overline{\alpha}_{t-1}} x_0 - \widehat{x}_t\| \lesssim \sqrt{d \log T (1-\overline{\alpha}_{t-1})}$ and its complement respectively.

Proof of relation (107b). Consider x_0 such that $\|\sqrt{\overline{\alpha}_{t-1}} x_0 - \widehat{x}_t\| \lesssim \sqrt{d \log T (1-\overline{\alpha}_{t-1})}$ and

$$\|x_t(\gamma) - \widehat{x}_t\| \le \|x_{t-1} - \widehat{x}_t\| \lesssim \sqrt{d \log T (1-\alpha_t)} \le \sqrt{\frac{1-\overline{\alpha}_t}{d \log T}}. \tag{116}$$

It guarantees that

$$\begin{aligned} \frac{p_{X_{t-1} \,|\, X_0}(x_t(\gamma) \,|\, x_0)}{p_{X_t(\gamma) \,|\, X_0}(\widehat{x}_t \,|\, x_0)} &= \exp\Big(\frac{\big\| \widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}} x_0 \big\|^2}{2(1-\overline{\alpha}_{t-1})} - \frac{\big\| x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}} x_0 \big\|^2}{2(1-\overline{\alpha}_{t-1})} \Big) \\ &= 1 + O\Big(\frac{\big\| \widehat{x}_t - x_t(\gamma) \big\|^2 + \big\| \widehat{x}_t - x_t(\gamma) \big\| \sqrt{d \log T (1-\overline{\alpha}_{t-1})}}{1-\overline{\alpha}_{t-1}} \Big). \end{aligned}$$

Similar to the proof of (112), we can also derive

$$\frac{p_{X_{t-1}}(x_t(\gamma))}{p_{X_{t-1}}(\widehat{x}_t)} = \frac{\int_{x_0} p_{X_0}(x_0) p_{X_{t-1} \,|\, X_0}(x_t(\gamma) \,|\, x_0) \mathrm{d}x_0}{\int_{x_0} p_{X_0}(x_0) p_{X_{t-1} \,|\, X_0}(\widehat{x}_t \,|\, x_0) \mathrm{d}x_0} \tag{117}$$ eqn:cond-xt-1

$$= 1 + O\Big(\frac{\big\| \widehat{x}_t - x_t(\gamma) \big\|^2 + \big\| \widehat{x}_t - x_t(\gamma) \big\| \sqrt{d \log T (1-\overline{\alpha}_{t-1})}}{1-\overline{\alpha}_{t-1}} \Big). \tag{118}$$

Putting everything together leads to

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

Equipped with the above relation, we find

$$\begin{aligned} &\Big\| \int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, x_t(\gamma)) \big(x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}} x_0 \big) \mathrm{d}x_0 - \int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t) \big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}} x_0 \big) \mathrm{d}x_0 \Big\| \\ &= O\Big(d \log T \sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}} \Big) \cdot \Big\| \int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, x_t(\gamma)) \big(x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}} x_0 \big) \mathrm{d}x_0 \Big\| \\ &\lesssim \sqrt{d^3 \log^3 T (1-\alpha_t)}, \end{aligned}$$

where the last step invokes the property (115) again. Similarly, we also arrive at

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

$$
\begin{aligned}
&- \int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)^\top \mathrm{d}x_0\Big\| \\
&= O\Big(d\log T\sqrt{\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}}}\Big)\cdot\Big\|\int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, x_t(\gamma))\big(x_t(\gamma) - \sqrt{\overline{\alpha}_{t-1}}x_0\big)\big(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0\big)^\top \mathrm{d}x_0\Big\| \\
&\lesssim d^2\log^2 T\sqrt{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}.
\end{aligned}
$$

Taking the above two above relations together with the definition (92) gives

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

Proof of relation (108a) To control the quantity on the left of (108a), by triangle's inequality, let us write

$$
\begin{aligned}
&\Big\|\frac{\int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_{t-1}} - \frac{\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_t}\Big\| \\
&\le \Big\|\frac{\int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 - \int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{\alpha_t - \overline{\alpha}_t}\Big\| \\
&\quad + \Big\|\Big(\frac{1}{\alpha_t - \overline{\alpha}_t} - \frac{1}{1-\overline{\alpha}_t}\Big)\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0\Big\|.
\end{aligned} \tag{119}
$$

eqn:adagio

As computed previously, given $-\log p_{X_t}(x_t) \lesssim d\log T$, Lemma 1 implies that

$$
\begin{aligned}
\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 &= \int_{\|x_t - \sqrt{\overline{\alpha}_t}x_0\| \le \sqrt{d\log T(1-\overline{\alpha}_t)}} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 \\
&\quad + \int_{\|x_t - \sqrt{\overline{\alpha}_t}x_0\| > \sqrt{d\log T(1-\overline{\alpha}_t)}} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 \\
&\lesssim \sqrt{d\log T(1-\overline{\alpha}_t)}.
\end{aligned} \tag{120}
$$

eqn:rondo

Additionally, similar argument gives

$$
\begin{aligned}
&\int_{x_0} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 \\
&= \int_{x_0 \in \|x_t - \sqrt{\overline{\alpha}_t}x_0\| \le \sqrt{d\log T(1-\overline{\alpha}_t)}} p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 + \exp(-\Omega(d\log T)) \\
&\overset{\text{(i)}}{=} \Big(1 + O\Big(\frac{d\log T(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\Big)\Big)\int_{x_0 \in \|x_t - \sqrt{\overline{\alpha}_t}x_0\| \le \sqrt{d\log T(1-\overline{\alpha}_t)}} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0 + \exp(-\Omega(d\log T)) \\
&= \Big(1 + O\Big(\frac{d\log T(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\Big)\Big)\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)(x_t - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0,
\end{aligned} \tag{121}
$$

eqn:schertzo

where in step (i), we make use of the following relation that

$$
p_{X_0 \,|\, X_{t-1}}(x_0 \,|\, \widehat{x}_t) = p_{X_0}(x_0)\frac{p_{X_{t-1} \,|\, X_0}(\widehat{x}_t \,|\, x_0)}{p_{X_{t-1}}(\widehat{x}_t)} = \Big(1 + O\Big(\frac{d\log T(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\Big)\Big)p_{X_0}(x_0)\frac{p_{X_t \,|\, X_0}(x_t \,|\, x_0)}{p_{X_t}(x_t)}. \tag{122}
$$

eqn:t-1-to-t

To see why (122) holds, we first observe that

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

$$= \exp\Big(O\Big(\frac{d(1-\alpha_t)}{1-\overline{\alpha}_{t-1}} + \frac{(1-\alpha_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|^2}{(1-\overline{\alpha}_{t-1})^2}\Big)\Big).$$

In addition, for $x_0 \in \mathcal{E} = \{x_0 \mid \|x_t - \sqrt{\overline{\alpha}_t}x_0\| \lesssim \sqrt{d\log T(1-\overline{\alpha}_t)}\|\}$, it satisfies

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

on the other hand, for $x_0 \notin \mathcal{E}$, we have

$$\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)} \lesssim \exp\Big(O\Big(\frac{\log T\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|^2}{T(1-\overline{\alpha}_t)}\Big)\Big).$$

Similiar to (112), integrate to obtain

$$\begin{aligned}
p_{X_t}(x) &= \int_{x_0} p_{X_0}(x_0)p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)\mathrm{d}x_0 \\
&= \int_{x_0\in\mathcal{E}} p_{X_0}(x_0)p_{X_t\,|\,X_0}(x_t\,|\,x_0)\cdot\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)}\mathrm{d}x_0 \\
&\quad + \int_{x_0\in\mathcal{E}^{\mathrm{c}}} p_{X_0}(x_0)p_{X_t\,|\,X_0}(x_t\,|\,x_0)\cdot\frac{p_{X_{t-1}\,|\,X_0}(x\,|\,x_0)}{p_{X_t\,|\,X_0}(x_t\,|\,x_0)}\mathrm{d}x_0 \\
&= p_{X_t}(x_t)\int_{x_0\in\mathcal{E}}\Big(1+O\Big(\frac{d\log T(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\Big)\Big)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0 \\
&\quad + p_{X_t}(x_t)\int_{x_0\in\mathcal{E}^{\mathrm{c}}}\exp\Big(O\Big(\frac{\log T\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|^2}{T(1-\overline{\alpha}_t)}\Big)\Big)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0 \\
&= \Big(1+O\Big(\frac{d\log T(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\Big)\Big)\big(1+\exp(-\Omega(d\log T))\big)p_{X_t}(x_t) + \exp(-\Omega(d\log T))p_{X_t}(x_t) \\
&= \Big(1+O\Big(\frac{d\log T(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\Big)\Big)p_{X_t}(x_t).
\end{aligned}$$

Plugging expressions (121) and (120) into (119) gives

$$\Big\|\frac{\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(\widehat{x}_t - \sqrt{\overline{\alpha}_{t-1}}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_{t-1}} - \frac{\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)\mathrm{d}x_0}{1-\overline{\alpha}_t}\Big\| \lesssim (1-\alpha_t)\Big(\frac{d\log T}{\alpha_t-\overline{\alpha}_t}\Big)^{3/2}.$$

Proof of relation (108b). First observe that

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

By virtue of relation (122), one can deduce

$$\begin{aligned}
&\int_{x_0} p_{X_0\,|\,X_{t-1}}(x_0\,|\,\widehat{x}_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)(x_t-\sqrt{\overline{\alpha}_t}x_0)^\top\mathrm{d}x_0 \\
&= \Big(1+O\Big(\frac{d\log T(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}\Big)\Big)\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)(x_t-\sqrt{\overline{\alpha}_t}x_0)(x_t-\sqrt{\overline{\alpha}_t}x_0)^\top\mathrm{d}x_0.
\end{aligned}$$

This property, combined with relation (121), implies that

$$\Big\|\frac{J_{t-1}(\widehat{x}_t)}{1-\overline{\alpha}_{t-1}} - \frac{J_t(\widehat{x}_t)}{1-\overline{\alpha}_t}\Big\| \le \frac{d^2\log^2 T(1-\alpha_t)}{(\alpha_t-\overline{\alpha}_t)^2}.$$