A Sharp Convergence Theory for The Probability Flow ODEs of Diffusion Models

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

August 6, 2024

Abstract

Diffusion models, which convert noise into new data instances by learning to reverse a diffusion process, have become a cornerstone in contemporary generative modeling. In this work, we develop non-asymptotic convergence theory for a popular diffusion-based sampler (i.e., the probability flow ODE sampler) in discrete time, assuming access to ℓ_2-accurate estimates of the (Stein) score functions. For distributions in $\mathbb{R}^d$, we prove that d/ε iterations — modulo some logarithmic and lower-order terms — are sufficient to approximate the target distribution to within ε total-variation distance. This is the first result establishing nearly linear dimension-dependency (in d) for the probability flow ODE sampler. Imposing only minimal assumptions on the target data distribution (e.g., no smoothness assumption is imposed), our results also characterize how ℓ_2 score estimation errors affect the quality of the data generation processes. In contrast to prior works, our theory is developed based on an elementary yet versatile non-asymptotic approach without the need of resorting to SDE and ODE toolboxes.

Keywords: diffusion models, score-based generative modeling, non-asymptotic theory, probability flow ODE

Contents

This manuscript presents improved theory for probability flow ODEs compared to its earlier version Li et al. (2023).

*Department of Statistics, The Chinese University of Hong Kong, Hong Kong.
†Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.
‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
§Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.

1 Introduction

Diffusion models have emerged as a cornerstone in contemporary generative modeling, a task that learns to generate new data instances (e.g., images, text, audio) that look similar in distribution to the training data (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Dhariwal and Nichol, 2021; Jolicoeur-Martineau et al., 2021; Chen et al., 2021; Kong et al., 2021; Austin et al., 2021). Originally proposed by Sohl-Dickstein et al. (2015) and later popularized by Song and Ermon (2019); Ho et al. (2020), the mainstream diffusion generative models — e.g., denoising diffusion implicit models (DDIMs) (Song et al., 2020a) and denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) — 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); Chen et al. (2024b) for overviews of recent development.

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

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

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

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

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

through proper learning based on how the training data propagate in the forward process. Encouragingly, there often exist feasible strategies to achieve this goal as long as faithful estimates about the (Stein) score functions — the gradients of the log marginal density of the forward process — are available (Anderson, 1982; Haussmann and Pardoux, 1986). 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

[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.

has since further inspired, numerous recent studies on the problem of learning score functions, a subroutine that also goes by the name of score matching (e.g., Hyvärinen (2005, 2007); Vincent (2011); Song et al. (2020b); Koehler et al. (2023)).

Nonetheless, despite the mind-blowing empirical advances, a mathematical theory for diffusion generative models is still in its infancy. Given the complexity of developing a full-fledged end-to-end theory, a divide-and-conquer approach has been advertised, decoupling the score learning phase (i.e., how to estimate score functions reliably from training data) and the generative sampling phase (i.e., how to generate new data instances given the score estimates). In particular, the past few years have witnessed growing interest and remarkable progress from the theoretical community towards understanding the generative sampling phase (Block et al., 2020; De Bortoli et al., 2021; Liu et al., 2022; De Bortoli, 2022; Lee et al., 2023; Pidstrigach, 2022; Chen et al., 2022b,a, 2023c; Tang and Zhao, 2024a,b; Pedrotti et al., 2023; Liang et al., 2024; Li and Yan, 2024). For instance, polynomial-time convergence guarantees have been established for stochastic samplers (e.g., Chen et al. (2022b,a); Benton et al. (2024); Li et al. (2024c); Tang and Zhao (2024a); Li et al. (2024a); Mbacke and Rivasplata (2023); Liang et al. (2024); Li and Yan (2024)) and deterministic samplers (e.g., Chen et al. (2023c); Benton et al. (2023); Li et al. (2024c); Gao and Zhu (2024); Li et al. (2024a); Huang et al. (2024)), 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 non-asymptotic convergence theory for a popular deterministic sampler (Song et al., 2021b) — originally proposed based on a sort of ordinary differential equations (ODEs) for the reverse process called probability flow ODEs or diffusion ODEs, closely related to the DDIM sampler (Song et al., 2020a). For concreteness, we prove that the iteration complexity is no larger than the order of

$$\text{(iteration complexity)}: \quad d/\varepsilon \tag{4}$$

(up to some logarithmic factor and lower-order term), with d the data dimension and ε the target accuracy level in total-variation (TV) distance. We impose only minimal assumptions on the target distribution (e.g., no smoothness condition is needed), and quantify the impact of ℓ_2 score estimation errors upon convergence. In comparisons to past works, our main contributions are as follows.

- *Linear d-dependency.* Our iteration complexity scales nearly linearly in the dimension d, which improves upon all prior theoretical guarantees for deterministic samplers (Li et al., 2023; Chen et al., 2023c; Huang et al., 2024); in fact, the state-of-the-art d-dependency before our work scales with d^2 (Li et al., 2023; Huang et al., 2024). Note that d-linear convergence theory was established for the stochastic sampler DDPM (Benton et al., 2024); the theoretical framework for DDPM is not applicable for analyzing probability flow ODEs, but the use of a stochastic localization result in Benton et al. (2024) motivates our approach in sharpening the d dependency. Additionally, our result does not exhibit exponential dependency on the smoothness or regularity conditions as in Chen et al. (2023c); Benton et al. (2023) (e.g., the regularity parameter used in Benton et al. (2023) might even scale with the dimension d).

- *Linear dependency on $1/\varepsilon$.* We derive an iteration complexity upper bound that is proportional to $1/\varepsilon$. Note that this was already accomplished in an earlier version of this work (Li et al., 2023), strengthening prior convergence guarantees considerably (Chen et al., 2023c). This scaling $1/\varepsilon$ was also proven by a recent work Huang et al. (2024) via a completely different ODE-based approach.

- *ℓ_2 score estimation errors for the deterministic sampler.* Our theory reveals that the TV distance between X_1 and Y_1 is proportional to the ℓ_2 score estimation error as well as the associated mean Jacobian errors, an appealing property already established in an earlier version of this work (Li et al., 2023). In comparison, prior theoretical results either study stochastic variations of this deterministic sampler (Chen et al., 2023b) (so that the samplers are no longer the original deterministic sampler) or fall short of accommodating discretization errors (Benton et al., 2023), with the only exception being the recent work Huang et al. (2024) that also accounts for score errors for deterministic samplers.

- *An elementary 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 on SDEs or ODEs is needed for establishing our theory, resulting in a versatile framework and sometimes lowering the technical barrier towards understanding diffusion models (for those with no background in SDEs/ODEs).

It is worth emphasizing that our analysis for the probability flow ODE differs drastically from the analysis for DDPM (Chen et al., 2022b,a; Benton et al., 2024). More concretely, the state-of-the-art analysis for DDPM (Benton et al., 2024) is built upon the Girsanov theorem, a hammer that provides a powerful way to control the Kullback-Leibler (KL) divergence between the forward process and the sampling process. This approach, however, is known to be inapplicable to ODE-based deterministic samplers, given that the aforementioned KL divergence might even approach infinity. Working backward, our proof attempts to track the proximity of p_{X_t} and p_{Y_t} by iteratively computing how p_{X_t}/p_{Y_t} evolves from $p_{X_{t+1}}/p_{Y_{t+1}}$.

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. Additionally, the notation $f(d,T) = o\big(g(d,T)\big)$ means that $f(d,T)/g(d,T) \to 0$ as d, T tend to infinity. We shall often use capital letters to denote random variables/vectors/processes, and lowercase letters for deterministic variables. For any two probability measures P and Q, the TV distance between them is defined to be $\mathsf{TV}(P,Q) \coloneqq \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$. Throughout the paper, $p_X(\cdot)$ (resp. $p_{X\,|\,Y}(\cdot\,|\,\cdot)$) denotes the probability density function of X (resp. X given Y). For any matrix A, we denote by $\|A\|$ (resp. $\|A\|_{\mathrm{F}}$) the spectral norm (resp. Frobenius norm) of A. Also, for any vector-valued function f, we let J_f or $\frac{\partial f}{\partial x}$ represent the Jacobian matrix of f.

2 Preliminaries

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

2.1 Diffusion generative models

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

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

$$X_0 \sim p_{\mathsf{data}}, \tag{5a}$$

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\,W_t, \qquad 1 \le t \le T, \tag{5b}$$

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

$$\alpha_t \coloneqq 1-\beta_t, \qquad \overline{\alpha}_t \coloneqq \prod_{k=1}^{t}\alpha_k, \qquad 1 \le t \le T, \tag{6}$$

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

$$X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t \qquad \text{for some } \overline{W}_t \sim \mathcal{N}(0, I_d). \tag{7}$$

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 (5) is sometimes referred to as variance-preserving (Song et al., 2021b). Throughout this paper, we employ the notation

$$q_t \coloneqq \mathsf{distribution}\big(X_t\big) \tag{8}$$

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

$$q_T \approx \mathcal{N}(0, I_d). \tag{9}$$

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

$$Y_T \sim \mathcal{N}(0, I_d), \tag{10a}$$

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

$$\text{(goal)} \qquad Y_t \overset{\mathrm{d}}{\approx} X_t, \qquad t = T, T-1, \cdots, 1. \tag{10b}$$

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

$$p_t \coloneqq \mathsf{distribution}\big(Y_t\big). \tag{11}$$

2.2 The probability flow ODE

Evidently, the most crucial step of the diffusion model lies in effective design of the reverse process. The data-generation process of a deterministic sampler typically proceeds as follows: starting from $Y_T \sim \mathcal{N}(0, I_d)$, one selects a set of functions $\{\Phi_t(\cdot)\}_{1\le t\le T}$ and computes:

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \Phi_t\big(Y_t\big) \quad \text{for } t = T, \cdots, 1. \tag{12a}$$

Clearly, the sampling process is fully deterministic except for the initialization Y_T. Suppose now that we are armed with the estimates $\{s_t(\cdot)\}_{1\le t\le T}$ for the log density functions $\{s_t^\star(\cdot) \coloneqq \nabla \log q_t(\cdot)\}_{1\le t\le T}$ — often referred to as the (Stein) score functions. Then a discrete-time version of the probability flow ODE approach (cf. (15)) adopts the following mapping:

$$\Phi_t(x) \coloneqq \frac{1}{\sqrt{\alpha_t}}\left(x + \frac{1-\alpha_t}{2} s_t(x)\right). \tag{12b}$$

This approach, based on the probability flow ODE (15), often achieves faster sampling compared to the stochastic counterpart like DDPM (Song et al., 2021b).

In order to elucidate the plausibility of a deterministic approach, 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 knowledge of SDEs and ODEs.

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

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

  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 (5) takes the following form[2] (Song et al., 2021b)

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

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

[2] To see its connection with (5), it suffices to derive from (5) that $X_t - X_{t-\mathrm{d}t} = \sqrt{1-\beta_t}X_{t-\mathrm{d}t} - X_{t-\mathrm{d}t} + \sqrt{\beta_t}W_t \approx -\frac{1}{2}\beta_t X_{t-\mathrm{d}t} + \sqrt{\beta_t}W_t$.

- *The reverse process.* As it turns out, there exist reverse processes capable of reconstructing the marginal distribution of the forward process. In particular, the *probability flow ODE* is a reverse process taking the following form (Song et al., 2021b)

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

  where we use $\nabla \log q_t(X)$ to abbreviate $\nabla_X \log q_t(X)$ for notational simplicity. This ODE exhibits matching distributions with the forward process in that

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

  As can be easily shown, the continuous-time limit of (12) falls under this category. Note that this family of deterministic samplers is closely related to the DDIM sampler (Karras et al., 2022; Song et al., 2021b).

Interestingly, in addition to the functions f and g that define the forward process, construction of (15) relies only upon knowledge of the (Stein) score function $\nabla \log q_t(\cdot)$ of the intermediate steps of the forward diffusion process, an intriguing fact that also holds when designing stochastic samplers like DDPM. Consequently, a key enabler of diffusion models lies in reliable learning of the score function, and hence the name *score-based generative modeling*.

3 Convergence theory for the probability flow ODE sampler

In this section, we analyze the probability flow ODE sampler in discrete time. While the proofs for our main theory are all postponed to the appendix, it is worth emphasizing upfront that our analysis framework directly tackles the discrete-time processes without the need of 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 hyper-parameters $\{\alpha_t\}$ that shall be adopted throughout all cases.

Score estimates. Given that the score functions are an essential component in score-based generative modeling, we assume access to faithful estimates of the score functions $\nabla \log q_t(\cdot)$ across all intermediate steps t, thus disentangling the score learning phase and the data generation phase. Towards this end, let us first formally introduce the true score function as follows.

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

$$s_t^\star(X) \coloneqq \nabla \log q_t(X), \qquad 1 \le t \le T. \tag{16}$$

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

$$s_t^\star \coloneqq \arg\min_{s:\mathbb{R}^d \to \mathbb{R}^d} \mathop{\mathbb{E}}_{W \sim \mathcal{N}(0, I_d), X_0 \sim p_{\mathsf{data}}} \left[\left\| s\big(\sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} W\big) + \frac{1}{\sqrt{1-\overline{\alpha}_t}} W \right\|_2^2 \right], \tag{17}$$

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

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{X\sim q_t}\left[\left\|s_t(X)-s_t^\star(X)\right\|_2^2\right]\le \varepsilon_{\mathsf{score}}^2. \tag{18}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{X\sim q_t}\left[\left\|J_{s_t}(X)-J_{s_t^\star}(X)\right\|\right]\le \varepsilon_{\mathsf{Jacobi}}. \tag{19}$$

In a nutshell, Assumption 1 reflects the ℓ_2 score estimation error, whereas Assumption 2 is concerned with the estimation error in terms of the corresponding Jacobian matrix (so as to ensure certain continuity of the score estimator). Both assumptions consider the *average* estimation errors over all T steps. As we shall see momentarily, our theory for the deterministic sampler relies on both Assumptions 1 and 2, while the theory for the stochastic sampler requires only Assumption 1. We shall discuss in Section 3.2 the insufficiency of Assumption 1 alone for the probability flow ODE sampler.

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

- X_0 is an absolutely continuous random vector, and

$$\mathbb{P}\big(\|X_0\|_2\le R=T^{c_R}\big)=1,\qquad X_0\sim p_{\mathsf{data}} \tag{20}$$

  for some arbitrarily large constant $c_R>0$.

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

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

$$\beta_1=1-\alpha_1=\frac{1}{T^{c_0}}; \tag{21a}$$

$$\beta_t=1-\alpha_t=\frac{c_1\log T}{T}\min\left\{\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^t,\,1\right\}. \tag{21b}$$

In words, our choice of $\{\beta_t\}$ undergoes two phases: at the beginning (when t is small), β_t exhibits exponential increase; once it reaches the level of $\frac{c_1\log T}{T}$, it stays flat for the remaining steps. This two-phase choice shares similarity with the choice adopted in prior diffusion model theory like Benton et al. (2024).

3.2 Main results

We are now ready to present our non-asymptotic convergence guarantee — measured by the total variation distance between the forward and the reverse processes — for the discrete-time version (12) of the probability flow ODE. The proof of our theory is postponed to Section 5.2.

Theorem 1. *Suppose that* (20) *holds true. Assume that the score estimates* $s_t(\cdot)$ $(1\le t\le T)$ *satisfy Assumptions 1 and 2. Then the sampling process* (12) *with the learning rate schedule* (21) *satisfies*

$$\mathsf{TV}\big(q_1,p_1\big)\le C_1\frac{d\log^4 T}{T}+C_1\sqrt{d\log^4 T}\,\varepsilon_{\mathsf{score}}+C_1 d(\log^2 T)\varepsilon_{\mathsf{Jacobi}} \tag{22}$$

for some universal constants $C_1>0$*, provided that* $T\ge C_2 d^2\log^5 T$ *for some large enough constant* $C_2>0$*. Here, we recall that* p_1 *(resp.* q_1*) represents the distribution of* Y_1 *(resp.* X_1*).*

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

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

$$\widetilde{O}\left(\frac{d}{\varepsilon}\right) \tag{23}$$

for small enough accuracy level ε. As far as we know, this is the first result that unveils linear dimension dependency for the probability flow ODE sampler. Note that our theory is established without assuming any sort of smoothness or log-concavity on the target data distribution.

Stability. Turning to the more general case with imperfect score estimates (i.e., $\varepsilon_{\mathsf{score}} > 0$), the deterministic sampler (12) yields a distribution whose distance to the target distribution (measured again by the TV distance) scales proportionally with $\varepsilon_{\mathsf{score}}$ and $\varepsilon_{\mathsf{Jacobi}}$. It is noteworthy that in addition to the ℓ_2 score estimation errors, we are in need of an assumption on the stability of the associated Jacobian matrices, which plays a pivotal in ensuring that the reverse-time deterministic process does not deviate considerably from the desired process.

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

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

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

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

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

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

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

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

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

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

Support size of p_{data}. It is noteworthy that our theory holds true even when the support size of the target distribution is polynomially large (see (20)). This implies that careful normalization of the target data is often unnecessary. Furthermore, we note that the assumption (20) can also be relaxed. Supposing that $\mathbb{P}\big(\|X_0\|_2 \le B \mid X_0 \sim p_{\mathsf{data}}\big) = 1$ for some quantity $B > 0$ (which is allowed to grow faster than a polynomial in T), we can readily extend our analysis to obtain

$$\mathsf{TV}\big(q_1, p_1\big) \le C_1 \left(\frac{d}{T} + \sqrt{d}\,\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}\right) \mathrm{polylog}(T, B).$$

Importantly, the convergence rate depends only logarithmically in B.

Comparisons to previous works. Next, let us compare our results with past works.

- The first analysis for the discretized probability flow ODE approach in prior literature was derived by a recent work Chen et al. (2023c), which established non-asymptotic convergence guarantees that exhibit polynomial dependency in both d and $1/\varepsilon$ (see, e.g., Chen et al. (2023c, Theorem 4.1)). However, it fell short of providing concrete polynomial dependency in d and $1/\varepsilon$, suffered from exponential dependency in the Lipschitz constant of the score function, and relied on exact score estimates. In contrast, our result in Theorem 1 uncovers a concrete $\widetilde{O}(d/\varepsilon)$ scaling (ignoring lower-order and logarithmic terms) without imposing any smoothness assumption on the target data distribution, and makes explicit the effect of ℓ_2 score estimation errors, both of which were previously unavailable for such discrete-time deterministic samplers.

- Benton et al. (2023) studied the convergence of the probability flow ODE approach without accounting for the discretization error. The result therein also exhibited exponential dependency on a certain Lipschitz constant w.r.t. the forward flow and a regularity parameter (denoted by λ therein, which might scale with the dimension d).

- Chen et al. (2023b) studied two variants of the probability flow ODE. By inserting an additional stochastic corrector step — based on overdamped (resp. underdamped) Langevin diffusion — in each iteration of the probability flow ODE (so strictly speaking, these variations are no longer deterministic samplers), Chen et al. (2023b) showed that $\widetilde{O}(L^3 d/\varepsilon^2)$ (resp. $\widetilde{O}(L^2\sqrt{d}/\varepsilon)$) steps are sufficient, where L denotes the Lipschitz constant of the score function. In comparison, our result demonstrates for the first time that the plain probability flow ODE already achieves the $\widetilde{O}(d/\varepsilon)$ scaling without requiring either corrector steps or smoothness assumptions.

- The very recent work Huang et al. (2024) developed a novel suite of theory for the probability flow ODE, accounting for p-th ($p \ge 1$) order Runge-Kutta integrators (so as to demonstrate the degree of acceleration based on higher-order ODEs). When $p = 1$, the algorithm resembles what we analyze herein; let us make comparisons for this case in the following. The iteration complexity derived by Huang et al. (2024) scales as $\widetilde{O}(d^2/\varepsilon)$, whereas we obtain a sharper bound $\widetilde{O}(d/\varepsilon)$. In addition, the iteration complexity in Huang et al. (2024) scales quadratically in the support size of the target distribution, while our theory allows the support size to be polynomially large without affecting the iteration complexity. Moreover, the TV distance bound in Huang et al. (2024) scales proportionally to $d^{3/4}\varepsilon_{\mathsf{score}}$ (in addition to other multiplicative factors like the support size and Lipschitz constants), which is weaker than our result $\sqrt{d}\varepsilon_{\mathsf{score}}$ in terms of the d-dependency.

Another recent work Gao and Zhu (2024) established the first non-asymptotic theory for the probability flow ODE in 2-Wasserstein distance. The results therein require the target data distribution to satisfy strong log-concavity though.

Finally, let us briefly compare our result with the theory for the popular stochastic sampler: DDPM. The state-of-the-art convergence theory Benton et al. (2024) reveals that the iteration complexity for DDPM scales as $\widetilde{O}(d/\varepsilon^2)$, which exhibits worse ε-dependency compared to our theory for the probability flow ODE.

4 Other related works

Convergence theory for diffusion models. Early theoretical efforts in understanding the convergence of score-based stochastic samplers suffered from being either not quantitative (De Bortoli et al., 2021; Liu et al., 2022; Pidstrigach, 2022), or the curse of dimensionality (e.g., exponential dependencies in the convergence guarantees) (Block et al., 2020; De Bortoli, 2022). The recent work Lee et al. (2022) provided the first polynomial convergence guarantee in the presence of ℓ_2-accurate score estimates, for any smooth distribution satisfying the log-Sobelev inequality. Chen et al. (2022b); Lee et al. (2023); Chen et al. (2022a) subsequently lifted such a stringent data distribution assumption. More concretely, Chen et al. (2022b) accommodated a broad family of data distributions under the premise that the score functions over the entire trajectory of the forward process are Lipschitz; Lee et al. (2023) only required certain smoothness assumptions but came with worse dependence on the problem parameters; and more recent results in Chen et al. (2022a); Benton et al. (2024) 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. The recent work Li and Yan (2024) further showed that DDPM can automatically adapt to intrinsic low dimensionality of the target distribution and converge faster. Turning attention to samplers based on the probability flow ODE, Chen et al. (2023c) derived the first non-asymptotic bounds for this type of samplers. Improved convergence guarantees have recently been provided by a concurrent work Chen et al. (2023b), with the assistance of additional corrector steps inerspersed in each iteration of the probability flow ODE. It is worth noting that the corrector steps proposed therein are based on Langevin-type diffusion and inject additive noise, and hence the resulting sampling processes are not deterministic. Additionally, theoretical justifications for DDPM in the context of image in-painting have been developed by Rout et al. (2023). Moreover, convergence results based on the Wasserstein distance have recently been derived as well (e.g., Tang and Zhao (2024a); Benton et al. (2023)), although these results typically exhibit exponential dependency on the Lipschitz constants of the score functions. While the vast majority of past theory has been devoted to accommodating general distributions in $\mathbb{R}^d$, acceleration is shown to be possible if we restrict attention to discrete-valued distributions (Chen and Ying, 2024). Another strand of recent works (e.g., Chen et al. (2024a); Gupta et al. (2024)) explored how to exploit parallel sampling to achieve considerable speed-up. Theoretical guarantees have also recently been extended to cover other popular methods like consistency models (Song et al., 2023; Li et al., 2024b; Dou et al., 2024) and diffusion guidance (Ho and Salimans, 2022; Wu et al., 2024; Fu et al., 2024).

Score matching. Hyvärinen (2005) showed that the score function can be estimated via integration by parts, a result that was further extended in Hyvärinen (2007). Song et al. (2020b) proposed sliced score matching to tame the computational complexity in high dimension. The consistency of the score matching estimator was studied in Hyvärinen (2005), with asymptotic normality established in Forbes and Lauritzen (2015). Optimizing the score matching loss has been shown to be intimately connected to minimizing upper bounds on the Kullback-Leibler divergence (Song et al., 2021a) and Wasserstein distance (Kwon et al., 2022) between the generated distribution and the target data distribution. The recent work Koehler et al. (2023) studied the statistical efficiency of score matching by connecting it with the isoperimetric properties of the target data distribution. Furthermore, Feng et al. (2024) showed that statistical procedures based on score matching can achieve minimal asymptotic covariance for convex M-estimation.

Other theory for diffusion models. The development of diffusion model theory is certainly beyond the above two strand of works. For instance, Oko et al. (2023) studied the approximation and generalization capabilities of diffusion modeling for distribution estimation; Kadkhodaie et al. (2024); Zhang et al. (2024); Biroli et al. (2024) investigated the phase transition between the memorization regime and the generalization regime in diffusion models; Chen et al. (2023a); Wang et al. (2024) studied how diffusion models can adapt to low-dimensional structure. 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)); some results discovered in the stochastic localization literature (e.g., Eldan (2020)) have also paved the way to sharpening of dimension dependency (Benton et al., 2024). In addition to the DDPM and DDIM type samplers discussed herein, convergence of other flow-based generative modeling has also been established in recent works (e.g., Gao et al. (2024); Cheng et al. (2024); Xu et al. (2024)). Xu and Chi (2024) developed provably robust methods for posterior sampling with diffusion priors for general nonlinear inverse problems, whereas Montanari and Wu (2024) exploited the idea of measure decomposition to improve posterior sampling for linear inverse problems. There have also been a couple of recent works that delve into various properties of diffusion models for Gaussian mixture models (Wu et al., 2024; Chen et al., 2024c; Cui et al., 2023; Li and Chen, 2024).

5 Analysis

In this section, we describe our non-asymptotic proof strategies for establishing Theorem 1.

5.1 Preliminary facts

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

Properties related to the score function. First of all, in view of the alternative expression (17) for the score function and the property of the minimum mean square error (MMSE) estimator (e.g., Hajek (2015, Section 3.3.1)), we know that the true score function $s_t^\star$ is given by the conditional expectation

$$\begin{aligned} s_t^\star(x) &= \mathbb{E}\left[-\frac{1}{\sqrt{1-\overline{\alpha}_t}}W \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right] = \frac{1}{1-\overline{\alpha}_t}\mathbb{E}\left[\sqrt{\overline{\alpha}_t}X_0 - x \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right] \\ &= -\frac{1}{1-\overline{\alpha}_t}\underbrace{\int_{x_0}(x - \sqrt{\overline{\alpha}_t}x_0)p_{X_0|X_t}(x_0\,|\,x)\mathrm{d}x_0}_{=:\,g_t(x)}. \end{aligned} \tag{24}$$

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

$$J_t(x) \coloneqq \frac{\partial g_t(x)}{\partial x}, \tag{25a}$$

which can be equivalently rewritten as

$$J_t(x) = I_d - \frac{1}{1-\overline{\alpha}_t}\mathsf{Cov}\Big(X_t - \sqrt{\overline{\alpha}_t}X_0 \mid X_t = x\Big). \tag{25b}$$

Properties about the learning rates. Next, we isolate a few useful properties about the learning rates as specified by $\{\alpha_t\}$ in (21):

$$\alpha_t \geq 1 - \frac{c_1\log T}{T} \geq \frac{1}{2}, \qquad 1 \leq t \leq T \tag{26a}$$

$$\frac{1}{2}\frac{1-\alpha_t}{1-\overline{\alpha}_t} \leq \frac{1}{2}\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t} \leq \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \leq \frac{4c_1\log T}{T}, \qquad 2 \leq t \leq T \tag{26b}$$

$$1 \leq \frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} \leq 1 + \frac{4c_1\log T}{T}, \qquad 2 \leq t \leq T \tag{26c}$$

$$\overline{\alpha}_T \leq \frac{1}{T^{c_2}}, \tag{26d}$$

$$\frac{\overline{\alpha}_{t+1}}{1-\overline{\alpha}_{t+1}} \leq \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t} \leq \frac{4\overline{\alpha}_{t+1}}{1-\overline{\alpha}_{t+1}}, \qquad 1 \leq t < T \tag{26e}$$

provided that T is large enough. Here, c_1 is defined in (21), and $c_2 \geq 1000$ is some large numerical constant. In addition, if $\frac{d(1-\alpha_t)}{\alpha_t-\overline{\alpha}_t} \ll 1$, then one has

$$\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} = 1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{d(d-2)(1-\alpha_t)^2}{8(\alpha_t-\overline{\alpha}_t)^2} + O\bigg(d^3\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\bigg), \tag{26f}$$

$$\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} = \exp\bigg(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\cdot\frac{d}{2}\bigg)\cdot\bigg(1+O\Big(d\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\Big)\bigg). \tag{26g}$$

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

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

$$\theta_t(x) \coloneqq \max\bigg\{-\frac{\log p_{X_t}(x)}{d\log T}, c_6\bigg\} \tag{27}$$

for any $x \in \mathbb{R}^d$, where $c_6 > 0$ is some large enough constant obeying $c_6 \geq 2c_R + c_0$.

Lemma 1. *Suppose that (20) holds true. Then for any quantity $c_5 \geq 2$, conditioned on $X_t = y$ one has*

$$\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2 \leq 5c_5\sqrt{\theta_t(y)d(1-\overline{\alpha}_t)\log T} \tag{28}$$

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

$$\mathbb{E}\Big[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2 \,\big|\, X_t = y\Big] \leq 12\sqrt{\theta_t(y)d(1-\overline{\alpha}_t)\log T}, \tag{29a}$$

$$\mathbb{E}\Big[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^2 \,\big|\, X_t = y\Big] \leq 120\theta_t(y)d(1-\overline{\alpha}_t)\log T, \tag{29b}$$

$$\mathbb{E}\Big[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^3 \,\big|\, X_t = y\Big] \leq 1040\big(\theta_t(y)d(1-\overline{\alpha}_t)\log T\big)^{3/2}, \tag{29c}$$

$$\mathbb{E}\Big[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^4 \,\big|\, X_t = y\Big] \leq 10080\big(\theta_t(y)d(1-\overline{\alpha}_t)\log T\big)^2. \tag{29d}$$

In order to interpret Lemma 1, let us look at the case with $\theta_t(y) = c_6$, corresponding to the scenario where $p_{X_t}(y) \geq \exp(-c_6 d\log T)$ (so that $p_{X_t}(y)$ is not exceedingly small). In this case, Lemma 1 implies that conditional on $X_t = y$ taking on a "typical" value, the vector $\sqrt{\overline{\alpha}_t}X_0 - X_t = \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t$ (see (7)) might still follow a sub-Gaussian tail, whose expected norm remains on the same order of that of an unconditional Gaussian vector $\mathcal{N}(0, (1-\overline{\alpha}_t)I_d)$.

Properties about the conditional covariance matrices. We shall also single out two basic properties about certain conditional covariances as follows. To be precise, generate

$$X_0 \sim p_{\mathsf{data}} \qquad \text{and} \qquad Z \sim \mathcal{N}(0, I_d) \tag{30}$$

independently. Define, for any $\overline{\alpha} \in (0,1)$ and any $x \in \mathbb{R}^d$, the following conditional covariance matrix

$$\Sigma_{\overline{\alpha}}(x) \coloneqq \mathsf{Cov}\Big(Z \,\big|\, \sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z = x\Big). \tag{31}$$

The lemma below reveals two properties about $\Sigma_{\overline{\alpha}}(\cdot)$ that play a crucial role in our analysis; the proof is postponed to Appendix A.4.

Lemma 2. *The conditional covariance matrix defined in (31) satisfies the following properties.*

(a) For any $\overline{\alpha}, \overline{\alpha}' \in (0,1)$ obeying $\frac{|\overline{\alpha}'-\overline{\alpha}|}{\overline{\alpha}(1-\overline{\alpha})} \lesssim \frac{1}{d\log T}$ and $1-\overline{\alpha} \geq T^{-c_0}$ (with c_0 the constant defined in (21)), it holds that

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

for some universal constants $C_3, C_8, C_9 > 0$.

(b) For the learning rates (21), one has

$$\sum_{t=2}^{T} \frac{1-\alpha_t}{1-\overline{\alpha}_t} \mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) \lesssim d\log T.$$

Remark 1. *This lemma, which plays a pivotal role in achieving linear d-dependency, is inspired by the analysis of Benton et al. (2024) for DDPM, exploiting an intriguing property (see (88)) originally discovered in the stochastic localization literature (Eldan, 2020). Note, however, that this property can also be established using elementary analysis without resorting to any sort of SDE toolboxes (El Alaoui and Montanari, 2022).*

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

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

$$\big(\mathsf{TV}(p_{X_T} \,\|\, p_{Y_T})\big)^2 \le \frac{1}{2}\mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) \lesssim \frac{1}{T^{200}}. \tag{32}$$

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

$$\varepsilon_{\mathsf{score},t}(x) \coloneqq \big\|s_t(x) - s_t^\star(x)\big\|_2 \qquad \text{and} \qquad \varepsilon_{\mathsf{Jacobi},t}(x) \coloneqq \big\|J_{s_t}(x) - J_{s_t^\star}(x)\big\|, \tag{33}$$

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

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X\sim q_t}\big[\varepsilon_{\mathsf{score},t}(X)\big] \le \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X\sim q_t}\left[\varepsilon_{\mathsf{score},t}(X)^2\right]\right)^{1/2} \le \varepsilon_{\mathsf{score}}. \tag{34a}$$

Also, Assumption 2 says that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X\sim q_t}\big[\varepsilon_{\mathsf{Jacobi},t}(X)\big] \le \varepsilon_{\mathsf{Jacobi}}. \tag{34b}$$

5.2 Main steps for the proof of Theorem 1

We now present the proof for our main result (i.e., Theorem 1) for the discrete-time sampler (12) based on the probability flow ODE. Given that the TV distance is always bounded above by 1, it suffices to assume

$$\varepsilon_{\mathsf{score}} \le \frac{1}{C_1\sqrt{d}\log^2 T} \tag{35a}$$

$$\varepsilon_{\mathsf{Jacobi}} \le \frac{1}{C_1 d\log^2 T} \tag{35b}$$

throughout the proof; otherwise the claimed result (22) becomes trivial.

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

$$\phi_t^\star(x) = x + \frac{1-\alpha_t}{2}s_t^\star(x) = x - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\int_{x_0}\big(x - \sqrt{\overline{\alpha}_t}x_0\big)p_{X_0|X_t}(x_0\,|\,x)\mathrm{d}x_0, \tag{36a}$$

$$\phi_t(x) = x + \frac{1-\alpha_t}{2}s_t(x), \tag{36b}$$

where the first line follows from (24). The update rule (12) can then be expressed as follows:

$$Y_{t-1} = \Phi_t(Y_t) = \frac{1}{\sqrt{\alpha_t}}\phi_t(Y_t). \tag{37}$$

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

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\phi_t(y_t), \qquad y'_{t-1} = \frac{1}{\sqrt{\alpha_t}}\phi_t(y'_t), \qquad t = T, T-1, \cdots \tag{38}$$

In other words, $\{y_{T-1}, \ldots, y_1\}$ (resp. $\{y'_{T-1}, \ldots, y'_1\}$) is the (reverse-time) sequence generated by the probability flow ODE (cf. (37)) when initialized to $Y_T = y_T$ (resp. $Y_T = y'_T$). We also define the following quantities for any point $y_T \in \mathbb{R}^d$ and its associated sequence $\{y_{T-1}, \ldots, y_1\}$:

$$\xi_t(y_t) \coloneqq \frac{\log T}{T}\big(d\varepsilon_{\mathsf{Jacobi},t}(y_t) + \sqrt{d\log T}\,\varepsilon_{\mathsf{score},t}(y_t)\big); \tag{39a}$$

$$S_t(y_T) \coloneqq \sum_{1<k\le t} \xi_k(y_k), \quad \text{for } t \ge 2, \qquad \text{and} \qquad S_1(y_T) = 0. \tag{39b}$$

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

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

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

$$\begin{aligned}\frac{p_{Y_{t-1}}(y_{t-1})}{p_{X_{t-1}}(y_{t-1})} &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})} \\ &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{Y_t}(y_t)} \cdot \left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\sqrt{\alpha_t}y_{t-1})}{p_{X_t}(y_t)}\right)^{-1} \cdot \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)},\end{aligned} \tag{40}$$

thus converting the density ratio of interest into the product of three other density ratios. Noteworthily, this observation (40) connects the target density ratio $\frac{p_{Y_{t-1}}}{p_{X_{t-1}}}$ at the $(t-1)$-th step with its counterpart $\frac{p_{Y_t}}{p_{X_t}}$ at the t-th step, motivating us to look at the density changes within adjacent steps in both the forward and the reverse processes (i.e., $p_{X_{t-1}}$ vs. p_{X_t} and $p_{Y_{t-1}}$ vs. p_{Y_t}). In light of this expression, we develop a key lemma related to some of these density ratios.

Lemma 4. *Recall the definition of $\theta_t(x)$ in (27). Consider any $x \in \mathbb{R}^d$ obeying $\frac{40c_1\varepsilon_{\mathsf{score},t}(x)\log^{\frac{3}{2}} T}{T} \le \sqrt{\theta_t(x)d}$. Then one has*

$$\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} \le 2\exp\left(\left(5\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log T} + 60\theta_t(x)d\log T\right)\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\right). \tag{41}$$

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

$$\begin{aligned}&\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)} \\ &= 1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\Big(\big\|\int\big(x-\sqrt{\overline{\alpha}_t}x_0\big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\big\|_2^2 - \int\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} \\ &\quad + O\left(\theta_t(x)^2d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T + \varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log T}\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\right).\end{aligned} \tag{42a}$$

Moreover, for any random vector Y, one has

$$\begin{aligned}&\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)} \\ &= 1 + \frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\Big(\big\|\int\big(x-\sqrt{\overline{\alpha}_t}x_0\big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\big\|_2^2 - \int\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\end{aligned}$$

$$+O\bigg(\theta_t(x)^2 d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2 T+\frac{d\log T\varepsilon_{\mathsf{Jacobi},t}(x)}{T}\bigg), \tag{42b}$$

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

Proof. The proof of this lemma is postponed to Appendix B.1. □

Remark 2. *Combining Lemma 4 with Lemma 1 and (26) gives: if* $C_{10}\frac{\theta_t(x)d\log^2 T+\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log^3 T}}{T}\le 1$ *and if* $\theta_t(x)\lesssim 1$, *then (42a) tells us that*

$$\log\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)}\le\frac{4c_1 d\log T}{T}+C_{10}\left\{\frac{d^2\log^4 T}{T^2}+\frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T}\right\} \tag{43}$$

under our sample size assumption (35), where $C_{10}>0$ *is some large enough constant. Here, we have made use of the fact that the penultimate term in (42a) is non-positive due to Jensen's inequality.*

Informally, the result in (42) already tells us that

$$\frac{p_{\phi_t(Y_t)}(\phi_t(x))}{p_{Y_t}(x)}\Big/\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)}\approx 1$$

for many points x if we ignore the residual terms, which combined with (40) shows that

$$\frac{p_{Y_{t-1}}(y_{t-1})}{p_{X_{t-1}}(y_{t-1})}\approx\frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)}$$

for many points y_t. However, it is worth pointing out that: while Lemma 4 already provides useful estimates for the density ratios of interest, these results alone are not sufficient to yield the desired d-dependency. For instance, the residual term in (42) scales quadratically in d, thereby precluding one from obtaining linear d-dependency.

To further make improvements, we develop a more refined bound below when $\theta_t(x)\lesssim 1$, whose proof can be found in Appendix B.2.

Lemma 5. *Recall the definition of* $\theta_t(\cdot)$ *in (27). There exists some function* $\zeta_t(\cdot)$ *such that: for any* x *obeying* $\theta_t(x)\lesssim 1$, $C_{10}\frac{\theta_t(x)d\log^2 T+\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log^3 T}}{T}\le 1$ *and* $C_{11}\frac{d\varepsilon_{\mathsf{Jacobi},t}(x)\log T}{T}\le 1$ *(with the constants* C_{10},C_{11} *defined in Lemma 4), one has*

$$\begin{aligned}&\frac{p_{\phi_t(Y_t)}(\phi_t(x))}{p_{Y_t}(x)}\Big/\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)}\\&\quad=1+\zeta_t(x)+O\bigg(\Big\|\frac{\partial\phi_t^\star(x)}{\partial x}-I\Big\|_{\mathrm{F}}^2+\frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T}+\frac{d\log T\varepsilon_{\mathsf{Jacobi},t}(x)}{T}+\frac{d\log^3 T}{T^2}\bigg)\end{aligned} \tag{44}$$

with $\zeta_t(x)\le 0$. *In addition, this function* $\zeta_t(\cdot)$ *satisfies*

$$\mathop{\mathbb{E}}_{X\sim q_t}\big[|\zeta_t(X)|\big]\lesssim\mathop{\mathbb{E}}_{X\sim q_t}\bigg[\Big\|\frac{\partial\phi_t^\star}{\partial x}(X)-I\Big\|_{\mathrm{F}}^2\bigg]+\frac{d\log^3 T}{T^2}, \tag{45}$$

provided that $T\gtrsim d^2\log^5 T$.

In words, Lemma 5 makes apparent that a key quantity to control when bounding the density ratios of interest is

$$\Big\|\frac{\partial\phi_t^\star}{\partial x}(X)-I\Big\|_{\mathrm{F}}^2 \tag{46}$$

While we are unable to obtain the desired control of (46) in a pointwise manner, the expected sum of this quantity (46) over all t can be bounded in a fairly tight manner (we shall demonstrate this momentarily in (58)), which forms a crucial step towards sharpening the dimension dependency.

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

$$\mathcal{E} \coloneqq \Big\{y : q_1(y) > \max\big\{p_1(y),\, \exp\big(-c_6 d\log T\big)\big\}\Big\}, \tag{47}$$

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

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

$$\begin{aligned}\mathsf{TV}\big(q_1, p_1\big) &= \int_{y: q_1(y) > p_1(y)} \big(q_1(y) - p_1(y)\big)\mathrm{d}y \\ &= \int_{y\in\mathcal{E}} \big(q_1(y) - p_1(y)\big)\mathrm{d}y + \int_{y: p_1(y) < q_1(y) \le \exp(-c_6 d\log T)} \big(q_1(y) - p_1(y)\big)\mathrm{d}y. \end{aligned} \tag{48}$$

In order to bound the second term on the right-hand side of (48), we make note of a basic fact: since $X_t \overset{\text{(d)}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} W$ with $W \sim \mathcal{N}(0, I_d)$ and $\mathbb{P}(\|X_0\|_2 \le T^{c_R}) = 1$, it holds that

$$\mathbb{P}\left\{\|X_t\|_2 \ge T^{c_R+2}\right\} \le \mathbb{P}\left\{\|W\|_2 \ge T^2\right\} < \exp\left(-c_6 d\log T\right) \tag{49}$$

under our assumption (35) on T, thereby indicating that

$$\int_{y: \|y\|_2 \ge T^{c_R+2}} q_t(y)\mathrm{d}y < \exp\left(-c_6 d\log T\right). \tag{50}$$

This basic fact in turn reveals that

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

provided that $c_6 \ge 4(c_R+2)$. Substitution into (48) then yields

$$\mathsf{TV}\big(q_1, p_1\big) \le \mathbb{E}_{Y_1\sim p_1}\left[\Big(\frac{q_1(Y_1)}{p_1(Y_1)} - 1\Big)\mathbb{1}\left\{Y_1 \in \mathcal{E}\right\}\right] + \exp\big(-c_6 d\log T\big), \tag{51}$$

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

To proceed, let us isolate the following set

$$\mathcal{I}_1 \coloneqq \Big\{y_T \mid S_T\big(y_T\big) \le c_{14}\Big\} \tag{52}$$

for some small enough constant $c_{14} > 0$. In words, $\mathcal{I}_1$ is composed of a set of points whose aggregated score error along the backward trajectory is well-controlled; in fact, these are points that exhibit "typical" behavior under the assumptions (35a) and (35b). As a result, we can decompose the first term of (51) into the influence of "typical" points and that of the remaining points as follows:

$$\begin{aligned}&\mathop{\mathbb{E}}_{Y_1\sim p_1}\left[\Big(\frac{q_1(Y_1)}{p_1(Y_1)} - 1\Big)\mathbb{1}\left\{Y_1 \in \mathcal{E}\right\}\right] = \mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\Big(\frac{q_1(Y_1)}{p_1(Y_1)} - 1\Big)\mathbb{1}\left\{Y_1 \in \mathcal{E}\right\}\right] \\ &= \mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\Big(\frac{q_1(Y_1)}{p_1(Y_1)} - 1\Big)\mathbb{1}\left\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1\right\}\right] + \mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1 \in \mathcal{E}, Y_T \notin \mathcal{I}_1\right\}\right], \end{aligned} \tag{53}$$

where the first identity holds since Y_1 is determined purely by Y_T via deterministic update rules. The decomposition (53) leaves us with two terms to control, which we accomplish in the next two steps.

Step 3: controlling the first term on the right-hand side of (53). This step analyzes the first term on the right-hand side of (53). We would like to make the analysis in this step slightly more general than needed, given that it will be useful for the subsequent analysis as well.

To begin with, let us introduce the following quantity:

$$\tau(y_T) \coloneqq \max\Big\{2 \le t \le T+1 : S_{t-1}\big(y_T\big) \le c_{14}\Big\}, \tag{54}$$

meaning that the score errors exhibit "typical" behavior up to the $\big(\tau(y_T)-1\big)$-th iteration. As can be clearly seen from the definition (52) of $\mathcal{I}_1$,

$$\tau(y_T) = T+1, \qquad \forall y_T \in \mathcal{I}_1. \tag{55}$$

In the sequel, we first single out the following lemma, whose proof is deferred to Appendix B.3.

Lemma 6. *Consider any y_T and its associated sequence $\{y_{T-1}, \cdots, y_1\}$ (see (38)). If $-\log q_1(y_1) \le c_6 d \log T$, then one has*

$$-\log q_k(y_k) \le 2c_6 d \log T \tag{56}$$

for any $1 \le k < \tau(y_T)$ (cf. (54)), provided that $c_6 \ge 3c_1$.

As a consequence of Lemma 6, we are able to control the density ratio q_t/p_t up to the $\big(\tau(y_T)-1\big)$-th iteration, as stated in the following lemma. The proof can be found in Appendix B.4.

Lemma 7. *Consider any y_T, along with the deterministic sequence $\{y_{T-1}, \cdots, y_1\}$ (cf. (38))), and set $\tau = \tau(y_T)$ (cf. (54)). Then one has*

$$\frac{q_1(y_1)}{p_1(y_1)} = \left\{1 + O\left(\frac{d\log^4 T}{T} + \sum_{t<\tau}\left(\zeta_t(y_t) + \left\|\frac{\partial \phi_t^\star(y_t)}{\partial x} - I\right\|_{\mathrm{F}}^2\right) + S_{\tau-1}(y_{\tau-1})\right)\right\}\frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})}, \tag{57a}$$

$$\text{and} \qquad \frac{q_k(y_k)}{2p_k(y_k)} \le \frac{q_1(y_1)}{p_1(y_1)} \le 2\frac{q_k(y_k)}{p_k(y_k)}, \qquad \forall k < \tau, \tag{57b}$$

where the function $\zeta_t(\cdot)$ is defined in Lemma 5.

Moreover, according to the definition in (36), we can invoke the properties (25) to obtain

$$\frac{\partial \phi_t^\star}{\partial x}(x) - I_d = -\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x) = \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\mathsf{Cov}\left(\frac{X_t - \sqrt{\overline{\alpha}_t}X_0}{\sqrt{1-\overline{\alpha}_t}} \,|\, X_t = x\right) - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} I_d,$$

which combined with Lemma 2(b) and the property (26b) leads to

$$\begin{aligned}
&\sum_{t=2}^{T} \mathop{\mathbb{E}}_{X_t\sim q_t}\left[\left\|\frac{\partial \phi_t^\star}{\partial x}(X_t) - I\right\|_{\mathrm{F}}^2\right] \le \sum_{t=2}^{T} \mathop{\mathbb{E}}_{X_t\sim q_t}\left[\left\|\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\mathsf{Cov}\left(\frac{X_t - \sqrt{\overline{\alpha}_t}X_0}{\sqrt{1-\overline{\alpha}_t}} \,|\, X_t\right)\right\|_{\mathrm{F}}^2\right] + \sum_{t=2}^{T}\left\|\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} I_d\right\|_{\mathrm{F}}^2 \\
&= \sum_{t=2}^{T}\left(\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\right)^2 \mathop{\mathbb{E}}_{X_t\sim q_t}\left[\mathsf{Tr}\left(\left(\mathsf{Cov}\left(\frac{X_t - \sqrt{\overline{\alpha}_t}X_0}{\sqrt{1-\overline{\alpha}_t}} \,|\, X_t\right)\right)^2\right)\right] + \sum_{t=2}^{T}\left\|\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} I_d\right\|_{\mathrm{F}}^2 \\
&\lesssim \frac{\log T}{T}\sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\mathsf{Tr}\left(\mathop{\mathbb{E}}_{X_0\sim p_{\mathsf{data}}, Z\sim\mathcal{N}(0,I_d)}\left[\left(\Sigma_{\overline{\alpha}_t}\left(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\right)\right)^2\right]\right) + \sum_{t=2}^{T}\frac{d\log^2 T}{T^2} \\
&\asymp \frac{d\log^2 T}{T}.
\end{aligned} \tag{58}$$

Now let us look at the set $\mathcal{I}_1$. Taking $\tau(y_T) = T+1$ (cf. (55)) in Lemma 7 yields

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

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\left\{1+O\left(\frac{d\log^4 T}{T}+\sum_t\left(\zeta_t(y_t)+\left\|\frac{\partial\phi_t^\star(y_t)}{\partial x}-I\right\|_{\mathrm{F}}^2\right)+S_T(y_T)\right)\right\}\frac{q_T(Y_T)}{p_T(Y_T)}-1\right)\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_1\right\}\right]\\
&= \int\left\{\left(1+O\left(\frac{d\log^4 T}{T}+\sum_t\left(\zeta_t(y_t)+\left\|\frac{\partial\phi_t^\star(y_t)}{\partial x}-I\right\|_{\mathrm{F}}^2\right)+S_T(y_T)\right)\right)q_T(y_T)-p_T(y_T)\right\}\mathbb{1}\left\{y_1\in\mathcal{E},y_T\in\mathcal{I}_1\right\}\mathrm{d}y_T\\
&\le \int\left|q_T(y_T)-p_T(y_T)\right|\mathrm{d}y_T+O\left(\frac{d\log^4 T}{T}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+(d\log T)\varepsilon_{\mathsf{Jacobi}}\right)\\
&\lesssim \frac{d\log^4 T}{T}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+(d\log T)\varepsilon_{\mathsf{Jacobi}}. \qquad (59)
\end{aligned}
$$

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

$$
\begin{aligned}
&\int\left(S_T(y_T)+\sum_t\left(|\zeta_t(y_t)|+\left\|\frac{\partial\phi_t^\star(y_t)}{\partial x}-I\right\|_{\mathrm{F}}^2\right)\right)q_T(y_T)\,\mathbb{1}\left\{y_1\in\mathcal{E},y_T\in\mathcal{I}_1\right\}\mathrm{d}y_T\\
&\quad=\sum_{t=1}^T\int\left(\frac{\log T}{T}\left(d\varepsilon_{\mathsf{Jacobi},t}(y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(y_t)\right)+|\zeta_t(y_t)|+\left\|\frac{\partial\phi_t^\star(y_t)}{\partial x}-I\right\|_{\mathrm{F}}^2\right)q_T(y_T)\,\mathbb{1}\left\{y_1\in\mathcal{E},y_T\in\mathcal{I}_1\right\}\mathrm{d}y_T\\
&\quad\le 4\sum_{t=1}^T\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{\log T}{T}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)+|\zeta_t(Y_t)|+\left\|\frac{\partial\phi_t^\star(Y_t)}{\partial x}-I\right\|_{\mathrm{F}}^2\right)\frac{q_t(Y_t)}{p_t(Y_t)}\right]\\
&\quad= 4\sum_{t=1}^T\mathop{\mathbb{E}}_{Y_t\sim q_t}\left[\frac{\log T}{T}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)+|\zeta_t(Y_t)|+\left\|\frac{\partial\phi_t^\star(Y_t)}{\partial x}-I\right\|_{\mathrm{F}}^2\right]\\
&\quad\lesssim \frac{d\log^4 T}{T}+(d\log T)\varepsilon_{\mathsf{Jacobi}}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}},
\end{aligned}
$$

where the first inequality is due to (57), and the last relation comes from (34) and (58).

Step 4: controlling the second term on the right-hand side of (53). In this step, we find it helpful to introduce the following sets (in addition to $\mathcal{I}_1$ defined in (52)), where we again abbreviate $\tau=\tau(y_T)$ as long as it is clear from the context:

$$
\mathcal{I}_2 \coloneqq \left\{y_T : c_{14}\le S_\tau(y_T)\le 2c_{14}\right\}, \qquad (60a)
$$

$$
\mathcal{I}_3 \coloneqq \left\{y_T : S_{\tau-1}(y_T)\le c_{14}, \xi_\tau(y_T)\ge c_{14}, \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})}\le\frac{8q_\tau(y_\tau)}{p_\tau(y_\tau)}\right\}, \qquad (60b)
$$

$$
\mathcal{I}_4 \coloneqq \left\{y_T : S_{\tau-1}(y_T)\le c_{14}, \xi_\tau(y_T)\ge c_{14}, \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})}>\frac{8q_\tau(y_\tau)}{p_\tau(y_\tau)}\right\}. \qquad (60c)
$$

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

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

Lemma 8. *It holds that*

$$
\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_2\cup\mathcal{I}_3\cup\mathcal{I}_4\right\}\right]\lesssim\frac{d\log^4 T}{T}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+(d\log T)\varepsilon_{\mathsf{Jacobi}}. \qquad (61)
$$

See Appendix B.5 for the proof of this lemma.

Step 5: putting all pieces together. Recall that $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4 = \mathbb{R}^d$. Taking (51), (53), (59) and (61) collectively, we conclude that

$$\begin{aligned}\mathsf{TV}(p_1, q_1) &\le \mathop{\mathbb{E}}_{Y_T \sim p_T}\left[\Big(\frac{q_1(Y_1)}{p_1(Y_1)} - 1\Big)\mathbb{1}\left\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1\right\}\right] \\ &\quad + \mathop{\mathbb{E}}_{Y_T \sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4\right\}\right] + \exp(-c_6 d\log T) \\ &\lesssim \frac{d\log^4 T}{T} + \sqrt{d\log^3 T}\,\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}\log T\end{aligned}$$

as claimed.

6 Discussion

In this paper, we have developed a new suite of non-asymptotic theory for establishing the convergence and faithfulness of the probability flow ODE based sampler, 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. Our result is the first to establish nearly linear dimension dependency for the iteration complexity of this sampler, where only very minimal assumptions on the target data distribution are imposed. 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, can we establish sharp convergence results in terms of the Wasserstein distance for general non-strongly-log-concave data distributions, which could sometimes be "closer" to how humans differentiate pictures and might potentially help relax Assumption 2 in the case of deterministic samplers? To what extent can we further accelerate the sampling process, without requiring much more information than the score functions? Ideally, one would hope to achieve acceleration with the aid of the score functions only. It would also be of paramount interest to establish end-to-end performance guarantees that take into account both the score learning phase and the sampling phase.

Acknowledgements

G. Li is supported in part by the Chinese University of Hong Kong Direct Grant for Research. Y. Wei is supported in part by the the NSF grants CAREER award DMS-2143215, CCF-2106778, CCF-2418156, and the Google Research Scholar Award. Y. Chi is supported in part by the grants ONR N00014-19-1-2404, NSF CCF-2106778, DMS-2134080 and ECCS-2126634. 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.

A Proof for several preliminary facts

A.1 Proof of properties (25b)

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

$$\begin{aligned}\big[J_t(x)\big]_{i,j} = \mathbb{1}\{i=j\} + \frac{1}{1-\overline{\alpha}_t}\bigg\{&\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_i - \sqrt{\overline{\alpha}_t}x_{0,i}\big)\mathrm{d}x_0\Big)\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_j - \sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\Big) \\ &- \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x_i - \sqrt{\overline{\alpha}_t}x_{0,i}\big)\big(x_j - \sqrt{\overline{\alpha}_t}x_{0,j}\big)\mathrm{d}x_0\bigg\}. \qquad (62)\end{aligned}$$

This immediately establishes the matrix expression (25b).

A.2 Proof of properties (26) regarding the learning rates

Proof of property (26a). From the choice of β_t in (21), we have

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

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

Proof of properties (26b) and (26c). We start by proving (26b). Let τ be an integer obeying

$$\beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{\tau} \leq 1 < \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{\tau+1}, \tag{63}$$

and we divide into two cases based on τ.

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

$$1 - \overline{\alpha}_{t-1} \geq \frac{1}{3}\beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{t}. \tag{64}$$

  Clearly, if (64) 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 (64), 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 (64) fails to hold and suppose that $t_0 \leq \tau$. It then follows that

$$1 - \overline{\alpha}_{t_0-2} = 1 - \frac{\overline{\alpha}_{t_0-1}}{\alpha_{t_0-1}} \leq 1 - \overline{\alpha}_{t_0-1} < \frac{1}{3}\beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{t_0} \leq \frac{1}{2}\beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{t_0-1} < \frac{1}{2}, \tag{65}$$

  where the last inequality result from (63) and the assumption $t_0 \leq \tau$. This taken together with the assumptions (64) 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 (64) 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 (64) 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

$$1 - \overline{\alpha}_{t-1} \geq 1 - \overline{\alpha}_{\tau-1} \geq \frac{1}{3}\beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^{\tau} = \frac{\frac{1}{3}\beta_1 \big(1 + \frac{c_1 \log T}{T}\big)^{\tau+1}}{1 + \frac{c_1 \log T}{T}} \geq \frac{1}{4}, \tag{66}$$

where the second and the third inequalities come from (64). Therefore, one obtains

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

The above arguments taken together establish property (26b).

In addition, it comes immediately from (26b) that

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

thereby justifying property (26c).

Proof of property (26d). Turning attention to the second claim (26d), 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 (26e). It follows that

$$\frac{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}}{\frac{\overline{\alpha}_{t+1}}{1-\overline{\alpha}_{t+1}}} = \frac{\frac{1-\overline{\alpha}_{t+1}}{1-\overline{\alpha}_t}}{\alpha_{t+1}} \in [1,4],$$

where the last inequality makes use of (26a) and (26c).

Proof of property (26f). It is easily seen from the Taylor expansion that the learning rates $\{\alpha_t\}$ satisfy

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

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

Proof of property (26g). Finally, recognizing that

$$\frac{\exp(dx)-(1+x)^d}{\exp(dx)} = 1 - \Big(\frac{1+x}{\exp(x)}\Big)^d = 1-\big(1-O(x^2)\big)^d = O(dx^2)$$

for any x obeying $|dx| < 1/4$, one can deduce that

$$\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} = \Big(1+\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} = \exp\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\cdot\frac{d}{2}\Big)\cdot\bigg(1+O\Big(d\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\Big)\bigg),$$

given the fact that $\frac{d(1-\alpha_t)}{\alpha_t-\overline{\alpha}_t} \ll 1$.

A.3 Proof of Lemma 1

For notational simplicity, we drop the subscript t and denote $\theta(y) := \theta_t(y)$ throughout this subsection. To establish this lemma, we first make the following claim, whose proof is deferred to the end of this subsection.

Claim 1. *Consider any $c_5 \geq 2$ and suppose that $c_6 \geq 2c_R$. There exists some $x_0 \in \mathbb{R}^d$ such that*

$$\|\sqrt{\overline{\alpha}_t}x_0 - y\|_2 \leq c_5\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T} \qquad \text{and} \tag{67a}$$

$$\mathbb{P}\big(\|X_0 - x_0\|_2 \leq \epsilon\big) \geq \Big(\frac{\epsilon}{T^{2\theta(y)}}\Big)^d \qquad \text{with} \quad \epsilon = \frac{1}{T^{c_0/2}} \tag{67b}$$

hold simultaneously, where c_0 is defined in (21).

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

$$p_X(x) = p_{X_0|X_t}(x\,|\,y). \tag{68}$$

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

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

where $c_5 \geq 2$ (see Claim 1). Combining this with property (67a) about x_0 results in

$$\mathbb{P}\Big(\|\sqrt{\overline{\alpha}_t}X - y\|_2 > 5c_5\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T}\Big) \leq \mathbb{P}(X \in \mathcal{E}). \tag{69}$$

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

$$\begin{aligned}
\mathbb{P}(X_0 \in \mathcal{E}\,|\,X_t = y) &= \frac{\int_{x\in\mathcal{E}} p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x}{\int_x p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x} \\
&\leq \frac{\int_{x\in\mathcal{E}} p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x}{\int_{x:\|x-x_0\|_2\leq\epsilon} p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x} \\
&\leq \frac{\sup_{x\in\mathcal{E}} p_{X_t\,|\,X_0}(y\,|\,x)}{\inf_{x:\|x-x_0\|_2\leq\epsilon} p_{X_t\,|\,X_0}(y\,|\,x)} \cdot \frac{\mathbb{P}(X_0 \in \mathcal{E})}{\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon)}.
\end{aligned} \tag{70}$$

To further bound this quantity, note that: in view of the definition of $\mathcal{E}$ and expression (67a), one has

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

and

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

$$\geq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\bigg(- c_5^2\theta(y) d\log T - \frac{1}{T^{c_0}}\frac{1}{1-\overline{\alpha}_t}\bigg)$$
$$\geq \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\big(- 2c_5^2\theta(y) d\log T\big),$$

where the second line is due to the elementary inequality $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, the penultimate line relies on (67), and the last line holds true since $1-\overline{\alpha}_t \geq 1-\alpha_1 = 1/T^{c_0}$ (see (21)). Substitution of the above two displays into (70), we arrive at

$$\begin{aligned}\mathbb{P}(X_0 \in \mathcal{E} \,|\, X_t = y) &\leq \exp\big(- 2.5c_5^2\theta(y) d\log T\big) \cdot \frac{1}{\mathbb{P}(\|X_0 - x_0\|_2 \leq \epsilon)} \\ &\leq \exp\big(- 2.5c_5^2\theta d\log T\big) \cdot \Big(T^{2\theta(y)+c_0/2}\Big)^d \\ &\leq \exp\big(- (2.5c_5^2\theta(y) - 2\theta(y) - c_0/2) d\log T\big), \end{aligned} \tag{71}$$

where the second inequality invokes (67b). Substituting this into (69) and recalling the distribution (68) of X, we arrive at

$$\begin{aligned}\mathbb{P}\Big(\|\sqrt{\overline{\alpha}_t}X - y\|_2 > 5c_5\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T}\Big) &\leq \exp\big(- (2.5c_5^2\theta(y) - 2\theta(y) - c_0/2) d\log T\big) \\ &\leq \exp\big(- c_5^2\theta(y) d\log T\big),\end{aligned}$$

with the proviso that $c_5 \geq 2$ and $c_6 \geq c_0$ (so that $\theta(y) \geq c_6 \geq c_0$). This concludes the proof of the advertised result (28) when $c_5 \geq 2$ and $c_6 \geq 2c_R + c_0$, as long as Claim 1 can be justified.

With the above result in place, it then follows that

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

as claimed in (29a) by taking $c_5 = 2$. The proofs for (29b), (29c) and (29d) follow from similar aguments and are hence omitted for the sake of brevity.

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

$$\mathbb{P}(\|X_0 - x\|_2 \leq \epsilon) \leq \bigg(\frac{\epsilon}{2T^{\theta(y)}R}\bigg)^d \qquad \text{with } \epsilon = \frac{1}{T^{c_0/2}}. \tag{72}$$

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

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

$$+\int_{x:\,\|\sqrt{\overline{\alpha}_t}x-y\|_2<c_5\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x. \tag{73}$$

To further control (73), we make two observations:

1) The first term on the right-hand side of (73) can be bounded by

$$\begin{aligned}
&\int_{x:\,\|\sqrt{\overline{\alpha}_t}x-y\|_2\geq c_5\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T}} p_{X_0}(x)p_{X_t\,|\,X_0}(y\,|\,x)\mathrm{d}x \\
&\quad\leq \sup_{z:\,\|z\|_2\geq c_5\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T}} \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}}\exp\bigg(-\frac{\|z\|_2^2}{2(1-\overline{\alpha}_t)}\bigg) \\
&\quad< \frac{1}{2}\exp\big(-\theta(y)d\log T\big),
\end{aligned} \tag{74}$$

provided that $c_5\geq 2$ and $c_6>0$ is large enough (note that $\theta(y)\geq c_6$). Here, we have used $X_t \overset{\text{(i)}}{=} \sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}W$ 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 (73), let us construct an epsilon-net $\mathcal{N}_\epsilon=\{z_i\}$ for the following set

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

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

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

where the penultimate step comes from the assumption (72).

The above results taken collectively lead to

$$p_{X_t}(y)<\exp\big(-\theta(y)d\log T\big), \tag{75}$$

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

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

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

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

A.4 Proof of Lemma 2

Part (a). Before proceeding, we abuse the notation by introducing the following convenient notation:

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

where we recall that $X_0\sim p_{\mathsf{data}}$ and $Z\sim\mathcal{N}(0,I_d)$ are independently generated. Also, when $\frac{|\overline{\alpha}'-\overline{\alpha}|}{\overline{\alpha}(1-\overline{\alpha})}\lesssim\frac{1}{d\log T}$, we make note of several properties that can be easily verified:

$$\alpha\asymp\alpha'\qquad\text{and}\qquad 1-\alpha\asymp 1-\alpha';\tag{76a}$$

$$\left(\frac{1-\overline{\alpha}'}{1-\overline{\alpha}}\right)^{d/2} = \left(1+\frac{\overline{\alpha}-\overline{\alpha}'}{1-\overline{\alpha}}\right)^{d/2} \lesssim \left(1+O\Big(\frac{1}{d\log T}\Big)\right)^{d/2} \lesssim 1, \quad \text{and} \quad \left(\frac{1-\overline{\alpha}}{1-\overline{\alpha}'}\right)^{d/2} \lesssim 1; \tag{76b}$$

$$\frac{|\overline{\alpha}'-\overline{\alpha}|}{\overline{\alpha}(1-\overline{\alpha})(1-\overline{\alpha}')} = \frac{1}{1-\overline{\alpha}} O\left(\frac{1}{d\log T}\right). \tag{76c}$$

Consider any x' and let

$$x = \sqrt{\overline{\alpha}/\overline{\alpha}'}x'.$$

Our first step is to demonstrate a certain equivalence result between $p_{X_{\overline{\alpha}'}}(x')$ and $p_{X_{\overline{\alpha}}}(x)$. Towards this end, a little algebra reveals that

$$\frac{\|x'-\sqrt{\overline{\alpha}'}x_0\|_2^2}{2(1-\overline{\alpha}')} = \frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})} + \frac{\overline{\alpha}'-\overline{\alpha}}{\overline{\alpha}(1-\overline{\alpha}')(1-\overline{\alpha})}\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2}, \tag{77}$$

and as a result,

$$\begin{aligned} p_{X_{\overline{\alpha}'}}(x') &= \int p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha}')\big)^{d/2}}\exp\left(-\frac{\|x'-\sqrt{\overline{\alpha}'}x_0\|_2^2}{2(1-\overline{\alpha}')}\right)\mathrm{d}x_0 \\ &= \left(\frac{1-\overline{\alpha}}{1-\overline{\alpha}'}\right)^{d/2}\int p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})} - \frac{(\overline{\alpha}'-\overline{\alpha})\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2\overline{\alpha}(1-\overline{\alpha})(1-\overline{\alpha}')}\right)\mathrm{d}x_0. \end{aligned} \tag{78}$$

Combine this with the assumption $\frac{|\overline{\alpha}'-\overline{\alpha}|}{\overline{\alpha}(1-\overline{\alpha})} \lesssim \frac{1}{d\log T}$ and the properties (76) to yield

$$\begin{aligned} p_{X_{\overline{\alpha}'}}(x') &\asymp \int p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\left(1+O\Big(\frac{1}{d\log T}\Big)\right)\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\right)\mathrm{d}x_0 \\ &= \left(\int_{x_0\in\mathcal{E}} + \int_{x_0\notin\mathcal{E}}\right) p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\left(1+O\Big(\frac{1}{d\log T}\Big)\right)\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\right)\mathrm{d}x_0, \end{aligned} \tag{79}$$

where

$$\mathcal{E} \coloneqq \left\{x_0 \mid \frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\left(1+O\Big(\frac{1}{d\log T}\Big)\right)\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\right) \geq \exp\left(-4c_6 d\log T\right)\right\}$$

with the constant $c_6 > 0$ defined in Lemma 1. Given our assumption that $1-\overline{\alpha} \geq \frac{1}{T^{c_0}}$ and the fact $c_0 \leq c_6$, a little algebra leads to

$$\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})} \leq 12c_6 d\log T, \qquad \forall x_0 \in \mathcal{E},$$

and as a consequence,

$$\begin{aligned} &\int_{x_0\in\mathcal{E}} p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\left(1+O\Big(\frac{1}{d\log T}\Big)\right)\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\right)\mathrm{d}x_0 \\ &\quad\asymp \int_{x_0\in\mathcal{E}} p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\right)\mathrm{d}x_0 \\ &\quad= p_{X_{\overline{\alpha}}}(x) - \int_{x_0\notin\mathcal{E}} p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\right)\mathrm{d}x_0. \end{aligned}$$

Regarding those $x_0 \notin \mathcal{E}$, one can easily derive

$$\int_{x_0\notin\mathcal{E}} p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\left(-\left(1+O\Big(\frac{1}{d\log T}\Big)\right)\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\right)\mathrm{d}x_0$$

$$\leq \exp\left(-4c_6 d\log T\right)\int p_{\mathsf{data}}(x_0)\mathrm{d}x_0 = \exp\left(-4c_6 d\log T\right);$$

similarly, it can also be easily verified that (which we omit here for conciseness)

$$\int_{x_0\notin\mathcal{E}} p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\bigg(-\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\bigg)\mathrm{d}x_0 \leq \exp\left(-1.5c_6 d\log T\right).$$

Combine the above results with (79) to deduce that

$$\begin{aligned} p_{X_{\overline{\alpha}'}}(x') &\asymp \int p_{\mathsf{data}}(x_0)\frac{1}{\big(2\pi(1-\overline{\alpha})\big)^{d/2}}\exp\bigg(-\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\bigg)\mathrm{d}x_0 + O\big(\exp\left(-c_6 d\log T\right)\big) \\ &= p_{X_{\overline{\alpha}}}(x) + O\big(\exp\left(-1.5c_6 d\log T\right)\big) \qquad (80)\\ &\asymp p_{X_{\overline{\alpha}}}(x), \qquad (81)\end{aligned}$$

where the last line is valid provided that $-\log p_{X_{\overline{\alpha}}}(x) \leq c_6 d\log T$.

Based on the above results, we can further demonstrate another equivalence result concerning $p_{X_0|X_{\overline{\alpha}'}}$ and $p_{X_0|X_{\overline{\alpha}}}$: if $-\log p_{X_{\overline{\alpha}}}(x) \leq c_6 d\log T$ holds and

$$\|x'-\sqrt{\overline{\alpha}'}x_0\|_2^2 \asymp \|x-\sqrt{\overline{\alpha}}x_0\|_2^2 \lesssim d(1-\overline{\alpha})\log T,$$

then one has

$$\begin{aligned} p_{X_0|X_{\overline{\alpha}'}}(x_0\,|\,x') &= \frac{p_{\mathsf{data}}(x_0)\frac{1}{(2\pi(1-\overline{\alpha}'))^{d/2}}\exp\big(-\frac{\|x'-\sqrt{\overline{\alpha}'}x_0\|_2^2}{2(1-\overline{\alpha}')}\big)}{p_{X_{\overline{\alpha}'}}(x')} \\ &\asymp \frac{p_{\mathsf{data}}(x_0)\frac{1}{(2\pi(1-\overline{\alpha}))^{d/2}}\exp\big(-\big(1+O\big(\frac{1}{d\log T}\big)\big)\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\big)}{p_{X_{\overline{\alpha}}}(x)} \\ &\asymp \frac{p_{\mathsf{data}}(x_0)\frac{1}{(2\pi(1-\overline{\alpha}))^{d/2}}\exp\big(-\frac{\|x-\sqrt{\overline{\alpha}}x_0\|_2^2}{2(1-\overline{\alpha})}\big)}{p_{X_{\overline{\alpha}}}(x)} = p_{X_0|X_{\overline{\alpha}}}(x_0\,|\,x). \qquad (82)\end{aligned}$$

Now, we are ready to analyze the conditional covariance matrices of interest. Recalling that

$$\big(x'-\sqrt{\overline{\alpha}'}x_0\big)\big(x'-\sqrt{\overline{\alpha}'}x_0\big)^\top = \frac{\overline{\alpha}'(1-\overline{\alpha})}{\overline{\alpha}(1-\overline{\alpha}')}\big(x-\sqrt{\overline{\alpha}}x_0\big)\big(x-\sqrt{\overline{\alpha}}x_0\big)^\top,$$

we can deduce that

$$\begin{aligned} \Sigma_{\overline{\alpha}'}(x') &= \mathsf{Cov}\left(Z\,|\,\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z=x'\right) = \mathsf{Cov}\left(\frac{x'-\sqrt{\overline{\alpha}'}X_0}{\sqrt{1-\overline{\alpha}'}}\,|\,\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z=x'\right) \\ &= \frac{\overline{\alpha}'(1-\overline{\alpha})}{\overline{\alpha}(1-\overline{\alpha}')}\mathsf{Cov}\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}\,|\,\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z=x'\right) \\ &\overset{\text{(i)}}{\preceq} C_0\mathsf{Cov}\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}\,|\,\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z=x'\right) \\ &\overset{\text{(ii)}}{=} C_0\inf_{\mu}\mathbb{E}\left[\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}-\mu(x')\right)\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}-\mu(x')\right)^\top\Big\}\,|\,\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z=x'\right] \\ &\overset{\text{(iii)}}{\preceq} C_0\mathbb{E}\left[\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}-\mu'(x)\right)\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}-\mu'(x)\right)^\top\mathbb{1}\left\{\frac{\|x-\sqrt{\overline{\alpha}}X_0\|_2^2}{d(1-\overline{\alpha})\log T}\lesssim 1\right\}\,|\,\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z=x'\right] \\ &\quad + C_2\exp\big(-C_1 d\log T\big)I_d\end{aligned}$$

$$\overset{\text{(iv)}}{\preceq} \underbrace{C_3\mathbb{E}\left[\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}-\mu'(x)\right)\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}-\mu'(x)\right)^\top \mathbb{1}\left\{\frac{\left\|x-\sqrt{\overline{\alpha}}X_0\right\|_2^2}{d(1-\overline{\alpha})\log T}\lesssim 1\right\} \mid \sqrt{\overline{\alpha}}X_0+\sqrt{1-\overline{\alpha}}Z=x\right]}_{=:\widetilde{\Sigma}_{\overline{\alpha}}(x)}$$
$$+ C_2\exp\big(-C_1 d\log T\big)I_d$$

for some universal constants $C_0, C_1, C_2, C_3 > 0$, where

$$\mu'(x) \coloneqq \mathbb{E}\left[Z \mid \sqrt{\overline{\alpha}}X_0+\sqrt{1-\overline{\alpha}}Z=x\right].$$

Here, (i) follows since $\alpha \asymp \alpha'$ and $1-\alpha \asymp 1-\alpha'$ (cf. (76)); (ii) holds since $\mathsf{Cov}\big(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}} \mid \sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z = x'\big)$ represents the error covariance associated with the minimum mean square error (MMSE) estimator for Z given $\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z=x'$; (iii) arises from Lemma 1 (particularly the high-probability bound (28) stating that the probability of the event $\frac{\|x-\sqrt{\overline{\alpha}}X_0\|_2^2}{d(1-\overline{\alpha})\log T} \gg 1$ is exponentially small); and (iv) is an immediate consequence of (82). In particular, the matrix $\widetilde{\Sigma}_{\overline{\alpha}}(x)$ defined in the step (iv) obeys

$$\widetilde{\Sigma}_{\overline{\alpha}}(x) \preceq \mathsf{Cov}\left(\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}} \mid \sqrt{\overline{\alpha}}X_0+\sqrt{1-\overline{\alpha}}Z=x\right) = \Sigma_{\overline{\alpha}}(x),$$
$$\big\|\widetilde{\Sigma}_{\overline{\alpha}}(x)\big\| \le \mathbb{E}\left(\left\|\frac{x-\sqrt{\overline{\alpha}}X_0}{\sqrt{1-\overline{\alpha}}}\right\|_2^2 \mathbb{1}\left\{\frac{\left\|x-\sqrt{\overline{\alpha}}X_0\right\|_2^2}{d(1-\overline{\alpha})\log T}\lesssim 1\right\} \mid \sqrt{\overline{\alpha}}X_0+\sqrt{1-\overline{\alpha}}Z=x\right) \lesssim d\log T,$$

provided that $-\log p_{X_{\overline{\alpha}}}(x) \le c_6 d\log T$. These results in turn imply that

$$\big(\Sigma_{\overline{\alpha}'}(x')\big)^2 \preceq \Big(C_3\widetilde{\Sigma}_{\overline{\alpha}}(x)+C_2\exp\big(-C_1 d\log T\big)I_d\Big)^2 \preceq C_3^2\big(\Sigma_{\overline{\alpha}}(x)\big)^2 + C_4\exp\big(-C_5 d\log T\big)I_d \tag{83}$$

for some universal constants $C_4, C_5 > 0$, as long as $-\log p_{X_{\overline{\alpha}}}(x) \le c_6 d\log T$.

Treating x' as a random vector with the same distribution as $\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z$ — so that x is a random vector with the same distribution as $\sqrt{\overline{\alpha}}X_0+\sqrt{1-\overline{\alpha}}Z$ — and taking expectation over x' (and hence x) on both sides of (83), we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\Big(\Sigma_{\overline{\alpha}'}\big(\sqrt{\overline{\alpha}'}X_0+\sqrt{1-\overline{\alpha}'}Z\big)\Big)^2\right] = \mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\big(\Sigma_{\overline{\alpha}'}(x')\big)^2\right] \\
&= \mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\big(\Sigma_{\overline{\alpha}'}(x')\big)^2 \mathbb{1}\left\{-\log p_{X_{\overline{\alpha}}}(x) \le c_6 d\log T\right\}\right] + \mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\big(\Sigma_{\overline{\alpha}'}(x')\big)^2 \mathbb{1}\left\{-\log p_{X_{\overline{\alpha}}}(x) > c_6 d\log T\right\}\right] \\
&\preceq C_3^2 \mathop{\mathbb{E}}_{x\sim p_{X_{\overline{\alpha}}}}\left[\big(\Sigma_{\overline{\alpha}}(x)\big)^2\right] + C_4\exp\big(-C_5 d\log T\big)I_d + \mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\big(\Sigma_{\overline{\alpha}'}(x')\big)^2 \mathbb{1}\left\{-\log p_{X_{\overline{\alpha}}}(x) > c_6 d\log T\right\}\right],
\end{aligned} \tag{84}$$

where we use $p_{X_{\overline{\alpha}'}}$ and $p_{X_{\overline{\alpha}}}$ to denote the distribution of $X_{\overline{\alpha}'}$ and $X_{\overline{\alpha}}$, respectively. To bound the last term in the last line of (84), note that for any x obeying $-\log p_{X_{\overline{\alpha}}}(x) > c_6 d\log T$, it follows from (80) that

$$p_{X_{\overline{\alpha}'}}(x') \asymp p_{X_{\overline{\alpha}}}(x) + O\big(\exp(-1.5c_6 d\log T)\big) = o\big(\exp(-c_6 d\log T)\big),$$

and hence

$$\mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\big(\Sigma_{\overline{\alpha}'}(x')\big)^2 \mathbb{1}\left\{\frac{-\log p_{X_{\overline{\alpha}}}(x)}{d\log T} > c_6\right\}\right] \preceq \mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\big\|\Sigma_{\overline{\alpha}'}(x')\big\|^2 \mathbb{1}\left\{\frac{-\log p_{X_{\overline{\alpha}'}}(x')}{d\log T} \ge c_6\right\}\right] I_d.$$

Defining

$$\theta(x') = \max\left\{\frac{-\log p_{X_{\overline{\alpha}'}}(x')}{d\log T}, c_6\right\},$$

we can invoke Lemma 1 with a little algebra to derive (details are omitted for brevity)

$$\mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\left\|\Sigma_{\overline{\alpha}'}(x')\right\|^2 \mathbb{1}\left\{\frac{-\log p_{X_{\overline{\alpha}'}}(x')}{d\log T} > c_0\right\}\right] \le \sum_{k=1}^{\infty} \mathop{\mathbb{E}}_{x'\sim p_{X_{\overline{\alpha}'}}}\left[\left\|\Sigma_{\overline{\alpha}'}(x')\right\|^2 \mathbb{1}\left\{2^{k-1}c_6 \le \theta(x') \le 2^k c_6\right\}\right]$$
$$\le C_6 \exp\left(-C_7 d\log T\right)$$

for some universal constants $C_6, C_7 > 0$, where we have made use of the basic fact that

$$\left\|\Sigma_{\overline{\alpha}'}(x')\right\| \le \mathbb{E}\left[\|ZZ^\top\| \,|\, \sqrt{\overline{\alpha}'}X_0 + \sqrt{1-\overline{\alpha}'}Z = x'\right] = \mathbb{E}\left[\left\|\frac{x' - \sqrt{\overline{\alpha}'}X_0}{\sqrt{1-\overline{\alpha}'}}\right\|_2^2 \,|\, \sqrt{\overline{\alpha}'}X_0 + \sqrt{1-\overline{\alpha}'}Z = x'\right].$$

Putting the preceding resutls together, we can conclude that

$$\mathbb{E}\left[\left(\Sigma_{\overline{\alpha}'}\left(\sqrt{\overline{\alpha}'}X_0 + \sqrt{1-\overline{\alpha}'}Z\right)\right)^2\right] \preceq C_3^2 \mathop{\mathbb{E}}_{x\sim p_{X_{\overline{\alpha}}}}\left[\left(\Sigma_{\overline{\alpha}}(x)\right)^2\right] + \left\{C_4\exp\left(-C_5 d\log T\right) + C_6\exp\left(-C_7 d\log T\right)\right\} I_d$$
$$\preceq C_3^2 \mathbb{E}\left[\left(\Sigma_{\overline{\alpha}}\left(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\right)\right)^2\right] + C_8\exp\left(-C_9 d\log T\right) I_d$$

for some universal constants $C_8, C_9 > 0$, as claimed.

Part (b). First, we find it convenient to introduce another conditional covariance, defined as follows:

$$A_s(x) \coloneqq \mathsf{Cov}\left(X_0 \,|\, sX_0 + \sqrt{s}Z = x\right), \tag{85}$$

which clearly satisfies

$$\mathsf{Cov}\left(Z \,|\, sX_0 + \sqrt{s}Z = x\right) = \mathsf{Cov}\left(\frac{1}{\sqrt{s}}x - \sqrt{s}X_0 \,|\, sX_0 + \sqrt{s}Z = x\right) = sA_s(x). \tag{86}$$

It is easily seen that (by taking $s = \frac{\overline{\alpha}}{1-\overline{\alpha}}$)

$$\Sigma_{\overline{\alpha}}(x) = \mathsf{Cov}\Big(Z \,|\, \sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z = x\Big) = \frac{\overline{\alpha}}{1-\overline{\alpha}} A_{\frac{\overline{\alpha}}{1-\overline{\alpha}}}\Big(\frac{\sqrt{\overline{\alpha}}}{1-\overline{\alpha}}x\Big). \tag{87}$$

Let us single out a basic property about A_s and $\Sigma_{\overline{\alpha}}$ that plays an important role in the subsequent proof. First of all, it has been shown in previous work (see, e.g., Eldan (2020); El Alaoui and Montanari (2022)) that the time-differential of (85) admits a simple form[3]

$$\mathrm{d}\mathbb{E}\big[A_s(sX_0 + \sqrt{s}Z)\big] = -\mathbb{E}\big[\big(A_s(sX_0 + \sqrt{s}Z)\big)^2\big]\mathrm{d}s. \tag{88}$$

Replacing s with $\frac{\overline{\alpha}}{1-\overline{\alpha}}$ and using (31), we have

$$A_s\big(sX_0 + \sqrt{s}Z\big) = \frac{1-\overline{\alpha}}{\overline{\alpha}}\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big),$$

and hence (88) immediately tells us that

$$\mathrm{d}\Big(\frac{1-\overline{\alpha}}{\overline{\alpha}}\mathbb{E}\Big[\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\Big]\Big) = -\frac{(1-\overline{\alpha})^2}{\overline{\alpha}^2}\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\Big)^2\Big]\mathrm{d}\Big(\frac{\overline{\alpha}}{1-\overline{\alpha}}\Big). \tag{89}$$

From now on, let us consider any $0 < \overline{\alpha}_{\mathsf{l}} < \overline{\alpha}_{\mathsf{u}} < 1$ obeying $\frac{\overline{\alpha}_{\mathsf{l}}}{1-\overline{\alpha}_{\mathsf{l}}} \le \frac{\overline{\alpha}_{\mathsf{u}}}{1-\overline{\alpha}_{\mathsf{u}}} \le \frac{4\overline{\alpha}_{\mathsf{l}}}{1-\overline{\alpha}_{\mathsf{l}}}$, and the monotonicity of $f(x) = \frac{x}{1-x}$ in x gives

$$\frac{\overline{\alpha}_{\mathsf{l}}}{1-\overline{\alpha}_{\mathsf{l}}} \le \frac{\overline{\alpha}}{1-\overline{\alpha}} \le \frac{4\overline{\alpha}_{\mathsf{l}}}{1-\overline{\alpha}_{\mathsf{l}}} \qquad \text{for any } \overline{\alpha} \in [\overline{\alpha}_{\mathsf{l}}, \overline{\alpha}_{\mathsf{u}}].$$

[3]While this result was originally established by Eldan (2020) using stochastic localization, it can also be derived using an elementary estimation-theoretic approach without introducing any SDEs (see El Alaoui and Montanari (2022)).

Use the positive semidefiniteness of the covariance matrix and the fact $\mathrm{d}\big(\frac{\overline{\alpha}}{1-\overline{\alpha}}\big) = \frac{\mathrm{d}\overline{\alpha}}{(1-\overline{\alpha})^2}$ to derive

$$\begin{aligned}
&\int_{\overline{\alpha}_{\mathsf{l}}}^{\overline{\alpha}_{\mathsf{u}}} \frac{(1-\overline{\alpha})^2}{\overline{\alpha}^2} \mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\Big)^2\Big]\Big) \frac{1}{(1-\overline{\alpha})^2}\mathrm{d}\overline{\alpha} \\
&\quad = -\int_{\overline{\alpha}_{\mathsf{l}}}^{\overline{\alpha}_{\mathsf{u}}} \frac{\mathrm{d}\Big(\frac{1-\overline{\alpha}}{\overline{\alpha}}\mathsf{Tr}\big(\mathbb{E}\big[\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\big]\big)\Big)}{\mathrm{d}\overline{\alpha}}\mathrm{d}\overline{\alpha} \\
&\quad = \frac{1-\overline{\alpha}_{\mathsf{l}}}{\overline{\alpha}_{\mathsf{l}}}\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{\mathsf{l}}}\big(\sqrt{\overline{\alpha}_{\mathsf{l}}}X_0 + \sqrt{1-\overline{\alpha}_{\mathsf{l}}}Z\big)\big]\Big) - \frac{1-\overline{\alpha}_{\mathsf{u}}}{\overline{\alpha}_{\mathsf{u}}}\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{\mathsf{u}}}\big(\sqrt{\overline{\alpha}_{\mathsf{u}}}X_0 + \sqrt{1-\overline{\alpha}_{\mathsf{u}}}Z\big)\big]\Big) \geq 0, \qquad (90)
\end{aligned}$$

where the penultimate line arises from (89).

Moreover, recalling that $\overline{\alpha}_t = \prod_{i=1}^t \alpha_i \geq \overline{\alpha}_{t+1}$, we have

$$\begin{aligned}
&\Big(\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}\Big)^2 \mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) \cdot \int_{\overline{\alpha}_{t+1}}^{\overline{\alpha}_t} \frac{1}{(1-\overline{\alpha})^2}\mathrm{d}\overline{\alpha} \\
&\quad \overset{\text{(i)}}{\geq} \frac{1-\overline{\alpha}_{t+1}}{4\overline{\alpha}_{t+1}} \cdot \frac{1-\overline{\alpha}_t}{\overline{\alpha}_t} \cdot \mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) \cdot \frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t+1})} \\
&\quad \overset{\text{(ii)}}{=} \frac{1-\alpha_{t+1}}{4\overline{\alpha}_{t+1}}\mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) \\
&\quad \overset{\text{(iii)}}{\geq} \frac{1-\alpha_t}{4\overline{\alpha}_t}\mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) \qquad (91)
\end{aligned}$$

for any $t \geq 2$, where (i) results from (26), (ii) is valid since $\overline{\alpha}_t - \overline{\alpha}_{t+1} = \overline{\alpha}_t(1-\alpha_{t+1})$, and (iii) uses the property $\alpha_t \leq 1$ (cf. (26a)) and the fact that $1-\alpha_t \leq 1-\alpha_{t+1}$ for $t \geq 2$. Recall that (cf. (26b))

$$0 \leq \frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\overline{\alpha}_t(1-\overline{\alpha}_t)} = \frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t} \lesssim \frac{\log T}{T} \lesssim \frac{1}{d\log T}$$

and $1-\overline{\alpha}_{t+1} \geq 1-\overline{\alpha}_t \geq 1-\overline{\alpha}_1 = T^{-c_0}$. Taking inequality (91) together with Lemma 2(a) yields

$$\begin{aligned}
&\int_{\overline{\alpha}_{t+1}}^{\overline{\alpha}_t} \Big(\frac{1-\overline{\alpha}}{\overline{\alpha}}\Big)^2 \mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z\big)\Big)^2\Big]\Big) \frac{1}{(1-\overline{\alpha})^2}\mathrm{d}\overline{\alpha} \\
&\quad \geq C_8\Big(\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}\Big)^2 \mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) \cdot \int_{\overline{\alpha}_{t+1}}^{\overline{\alpha}_t} \frac{1}{(1-\overline{\alpha})^2}\mathrm{d}\overline{\alpha} - C_{10}\exp(-C_9 d\log T)\int_{\overline{\alpha}_{t+1}}^{\overline{\alpha}_t} \frac{1}{\overline{\alpha}^2}\mathrm{d}\overline{\alpha} \\
&\quad \geq C_8\frac{1-\alpha_t}{4\overline{\alpha}_t}\mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) - C_{10}\exp(-C_9 d\log T)\frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\overline{\alpha}_{t+1}\overline{\alpha}_t} \\
&\quad \geq C_8\frac{1-\alpha_t}{4\overline{\alpha}_t}\Big\{\mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) - C_{12}\exp(-C_{11}d\log T)\Big\} \qquad (92)
\end{aligned}$$

for some universal constants $C_8, C_9, C_{10}, C_{11}, C_{12} > 0$, where the first inequality invokes Lemma 2(a), and the last inequality is valid since $\frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\overline{\alpha}_{t+1}\overline{\alpha}_t} = \frac{1-\alpha_{t+1}}{\overline{\alpha}_{t+1}} \asymp \frac{1-\alpha_t}{\overline{\alpha}_t}$.

Combine inequality (92) with (90) (with $\overline{\alpha}_{\mathsf{l}} = \overline{\alpha}_{t+1}$ and $\overline{\alpha}_{\mathsf{u}} = \overline{\alpha}_t$) to reach

$$\begin{aligned}
&C_8\frac{1-\alpha_t}{4\overline{\alpha}_t}\Big\{\mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) - C_{12}\exp(-C_{11}d\log T)\Big\} \\
&\leq \frac{1-\overline{\alpha}_{t+1}}{\overline{\alpha}_{t+1}}\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{t+1}}\big(\sqrt{\overline{\alpha}_{t+1}}X_0 + \sqrt{1-\overline{\alpha}_{t+1}}Z\big)\big]\Big) - \frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\big]\Big).
\end{aligned}$$

Multiplying both sides by $\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}$, we are left with

$$C_8\frac{1-\alpha_t}{4(1-\overline{\alpha}_t)}\Big\{\mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big) - C_{12}\exp(-C_{11}d\log T)\Big\}$$

$$
\begin{aligned}
&\le \frac{1-\overline{\alpha}_{t+1}}{\alpha_{t+1}(1-\overline{\alpha}_t)}\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{t+1}}\big(\sqrt{\overline{\alpha}_{t+1}}X_0+\sqrt{1-\overline{\alpha}_{t+1}}Z\big)\big]\Big)-\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z\big)\big]\Big)\\
&\le \mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{t+1}}\big(\sqrt{\overline{\alpha}_{t+1}}X_0+\sqrt{1-\overline{\alpha}_{t+1}}Z\big)\big]\Big)-\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z\big)\big]\Big)\\
&\quad+\frac{32c_1\log T}{T}\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{t+1}}\big(\sqrt{\overline{\alpha}_{t+1}}X_0+\sqrt{1-\overline{\alpha}_{t+1}}Z\big)\big]\Big)\\
&\le \mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{t+1}}\big(\sqrt{\overline{\alpha}_{t+1}}X_0+\sqrt{1-\overline{\alpha}_{t+1}}Z\big)\big]\Big)-\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z\big)\big]\Big)+\frac{32c_1d\log T}{T} \qquad (93)
\end{aligned}
$$

for any $t\ge 2$. Here, the penultimate inequality in (93) holds since (according to (26a), (26b) and (26c))

$$
\frac{1-\overline{\alpha}_{t+1}}{\alpha_{t+1}(1-\overline{\alpha}_t)}-1=\frac{1-\alpha_{t+1}}{\alpha_{t+1}(1-\overline{\alpha}_t)}\le\frac{4(1-\alpha_{t+1})}{1-\overline{\alpha}_{t+1}}\le\frac{32c_1\log T}{T};
$$

and the last inequality in (93) follows since, for any $\overline{\alpha}\in(0,1)$,

$$
\mathbb{E}\big[\Sigma_{\overline{\alpha}}\big(\sqrt{\overline{\alpha}}X_0+\sqrt{1-\overline{\alpha}}Z\big)\big]=\mathbb{E}\Big[\mathsf{Cov}\big(Z\mid\sqrt{\overline{\alpha}}X_0+\sqrt{1-\overline{\alpha}}Z\big)\Big]\preceq\mathsf{Cov}(Z)=I_d. \qquad (94)
$$

Consequently, sum over $t=2,\ldots,T$ to form a telescopic sum and derive

$$
\begin{aligned}
&C_8\sum_{t=2}^{T}\frac{1-\alpha_t}{4(1-\overline{\alpha}_t)}\mathsf{Tr}\Big(\mathbb{E}\Big[\Big(\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z\big)\Big)^2\Big]\Big)\\
&\le\sum_{t=2}^{T}\Big\{\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{t+1}}\big(\sqrt{\overline{\alpha}_{t+1}}X_0+\sqrt{1-\overline{\alpha}_{t+1}}Z\big)\big]\Big)-\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_t}\big(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z\big)\big]\Big)\Big\}+32c_1d\log T\\
&\quad+C_8\sum_{t=2}^{T}\frac{1-\alpha_t}{4(1-\overline{\alpha}_t)}C_{12}\exp(-C_{11}d\log T)\\
&\le\mathsf{Tr}\Big(\mathbb{E}\big[\Sigma_{\overline{\alpha}_{T+1}}\big(\sqrt{\overline{\alpha}_{T+1}}X_0+\sqrt{1-\overline{\alpha}_{T+1}}Z\big)\big]\Big)+32c_1d\log T+2c_1C_8C_{12}\exp(-C_{11}d\log T)\log T\\
&\le 34c_1d\log T,
\end{aligned}
$$

where the last inequality uses (94) and property (26b). This concludes the proof.

A.5 Proof of Lemma 3

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

$$
\begin{aligned}
\mathsf{KL}(p_{X_T}\,\|\,p_{Y_T})&=\int p_{X_T}(x)\log\frac{p_{X_T}(x)}{p_{Y_T}(x)}\mathrm{d}x\\
&\overset{\text{(i)}}{=}\int p_{X_T}(x)\log\frac{\int_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}}p_{\sqrt{\overline{\alpha}_T}X_0}(y)p_{\sqrt{1-\overline{\alpha}_T}\,\overline{W}_t}(x-y)\mathrm{d}y}{p_{Y_T}(x)}\mathrm{d}x\\
&\le\int p_{X_T}(x)\log\frac{\sup_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}}p_{\sqrt{1-\overline{\alpha}_T}\,\overline{W}_t}(x-y)}{p_{Y_T}(x)}\mathrm{d}x\\
&=\int p_{X_T}(x)\bigg(-d/2\log(1-\overline{\alpha}_T)+\sup_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}}\bigg(-\frac{\|x-y\|_2^2}{2(1-\overline{\alpha}_T)}+\frac{\|x\|_2^2}{2}\bigg)\mathrm{d}x\\
&\overset{\text{(ii)}}{\le}\int p_{X_T}(x)\bigg(-d/2\log(1-\overline{\alpha}_T)+\|x\|_2\sup_{y:\|y\|_2\le\sqrt{\overline{\alpha}_T}T^{c_R}}\frac{\|y\|_2}{1-\overline{\alpha}_T}\bigg)\mathrm{d}x\\
&\le-d/2\log(1-\overline{\alpha}_T)+\frac{\sqrt{\overline{\alpha}_T}T^{c_R}}{2(1-\overline{\alpha}_T)}\mathbb{E}\left[\|X_T\|_2\right]\\
&\overset{\text{(iii)}}{\lesssim}\overline{\alpha}_Td+\frac{\sqrt{\overline{\alpha}_T}T^{c_R}}{2(1-\overline{\alpha}_T)}\Big(\sqrt{\overline{\alpha}_T}T^{c_R}+\sqrt{d}\Big)\overset{\text{(iv)}}{\lesssim}\frac{1}{T^{200}}, \qquad (95)
\end{aligned}
$$

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

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

and (iv) makes use of (26d) given that $c_2 \ge 1000$. The proof is thus completed by invoking the Pinsker inequality (Tsybakov, 2009, Lemma 2.5).

B Proof of auxiliary lemmas

B.1 Proof of Lemma 4

B.1.1 Proof of relations (41) and (42a)

Recall the definition of ϕ_t and $\phi_t^\star$ in (36), and introduce the following vector:

$$\begin{aligned} u &:= x - \phi_t(x) = x - \phi_t^\star(x) + \phi_t^\star(x) - \phi_t(x) \\ &= \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \int_{x_0} (x - \sqrt{\overline{\alpha}_t}x_0) p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0 - \frac{1-\alpha_t}{2}\big(s_t(x) - s_t^\star(x)\big). \end{aligned} \tag{96}$$

The proof is composed of the following steps.

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

$$X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,W \qquad \text{with } W \sim \mathcal{N}(0, I_d) \tag{97}$$

and making use of the Bayes rule, we can express the conditional distribution as

$$p_{X_0\,|\,X_t}(x_0\,|\,x) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} p_{X_t\,|\,X_0}(x\,|\,x_0) = \frac{p_{X_0}(x_0)}{p_{X_t}(x)} \cdot \frac{1}{\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2}} \exp\left(-\frac{\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)}\right). \tag{98}$$

Moreover, it follows from (97) that

$$\sqrt{\alpha_t}X_{t-1} \overset{\mathrm{d}}{=} \sqrt{\alpha_t}\big(\sqrt{\overline{\alpha}_{t-1}}X_0 + \sqrt{1-\overline{\alpha}_{t-1}}\,W\big) = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{\alpha_t - \overline{\alpha}_t}\,W. \tag{99}$$

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

$$\begin{aligned} \frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} &\overset{\text{(i)}}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{\big(2\pi(\alpha_t - \overline{\alpha}_t)\big)^{d/2}} \exp\left(-\frac{\big\|\phi_t(x) - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)}\right)\mathrm{d}x_0 \\ &\overset{\text{(ii)}}{=} \frac{1}{p_{X_t}(x)} \int_{x_0} p_{X_0}(x_0) \frac{1}{\big(2\pi(\alpha_t - \overline{\alpha}_t)\big)^{d/2}} \exp\left(-\frac{\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)}\right) \\ &\qquad \cdot \exp\left(-\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)}\right)\mathrm{d}x_0 \\ &\overset{\text{(iii)}}{=} \left(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\right)^{d/2} \cdot \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x) \cdot \\ &\qquad \exp\left(-\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)}\right)\mathrm{d}x_0 \end{aligned} \tag{100}$$

$$\begin{aligned} &\overset{\text{(iv)}}{=} \left\{1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + O\left(d^2\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2\right)\right\} \cdot \\ &\qquad \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x) \exp\left(-\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t - \overline{\alpha}_t)}\right)\mathrm{d}x_0. \end{aligned} \tag{101}$$

Here, identity (i) holds due to (99) 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 (96) and elementary algebra; relation (iii) is a consequence of the Bayes rule (98); and relation (iv) results from (26f).

Step 2: controlling the integral in the decomposition (101). In order to further control the right-hand side of expression (101), we need to evaluate the integral in (101). To this end, we make a few observations.

- To begin with, Lemma 1 tells us that

$$\mathbb{P}\Big(\big\|\sqrt{\overline{\alpha}_t}X_0 - x\big\|_2 > 5c_5\sqrt{\theta_t(x)d(1-\overline{\alpha}_t)\log T}\,\big|\,X_t = x\Big) \le \exp\big(-c_5^2\theta_t(x)d\log T\big) \tag{102a}$$

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

- A little algebra based on this relation allows one to bound u (cf. (96)) as follows:

$$\begin{aligned}\|u\|_2 &\le \frac{1-\alpha_t}{2}\varepsilon_{\mathsf{score},t}(x) + \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - x\big\|_2\,\big|\,X_t = x\right] \\ &\le \frac{1-\alpha_t}{2}\varepsilon_{\mathsf{score},t}(x) + \frac{6(1-\alpha_t)}{1-\overline{\alpha}_t}\sqrt{\theta_t(x)d(1-\overline{\alpha}_t)\log T},\end{aligned} \tag{102b}$$

  where the last inequality arises from Lemma 1.

Next, let us define

$$\mathcal{E}_c^{\mathsf{typical}} \coloneqq \Big\{x_0 : \big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2 \le 5c\sqrt{\theta_t(x)d(1-\overline{\alpha}_t)\log T}\Big\} \tag{103}$$

for any quantity $c > 0$. Then for any $x_0 \in \mathcal{E}_c^{\mathsf{typical}}$, it is clearly seen from (102) and (26) that

$$\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} \le \frac{25c^2}{2}\frac{(1-\alpha_t)\theta_t(x)d\log T}{\alpha_t - \overline{\alpha}_t} \le \frac{100c_1c^2\theta_t(x)d\log^2 T}{T}; \tag{104a}$$

$$\begin{aligned}\frac{\|u\|_2^2}{2(\alpha_t - \overline{\alpha}_t)} &\le \frac{(1-\alpha_t)^2}{4(\alpha_t - \overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}(x)^2 + \frac{36(1-\alpha_t)^2}{(1-\overline{\alpha}_t)(\alpha_t - \overline{\alpha}_t)}\theta_t(x)d\log T \\ &\le \frac{2c_1^2\log^2 T}{T^2}\varepsilon_{\mathsf{score},t}(x)^2 + \frac{2304c_1^2}{T^2}\theta_t(x)d\log^3 T,\end{aligned} \tag{104b}$$

$$\begin{aligned}\left|\frac{u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t - \overline{\alpha}_t}\right| &\le \frac{\|u\|_2\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2}{\alpha_t - \overline{\alpha}_t} \\ &\le \frac{5c(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d(1-\overline{\alpha}_t)\log T} + \frac{30c(1-\alpha_t)\theta_t(x)d\log T}{\alpha_t - \overline{\alpha}_t} \qquad (104c) \\ &\le \frac{20cc_1}{T}\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d(1-\overline{\alpha}_t)\log^3 T} + \frac{240cc_1\theta_t(x)d\log^2 T}{T}. \qquad (104d)\end{aligned}$$

As a consequence, for any $x_0 \in \mathcal{E}_c^{\mathsf{typical}}$ for $c \ge 2$, we have seen from (104d) and (26) that

$$\begin{aligned}&-\frac{(1-\alpha_t)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2}{2(\alpha_t - \overline{\alpha}_t)} + \frac{u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t - \overline{\alpha}_t} \le \frac{u^\top\big(x - \sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t - \overline{\alpha}_t} \\ &\le \frac{5c(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d(1-\overline{\alpha}_t)\log T} + \frac{30c(1-\alpha_t)\theta_t(x)d\log T}{\alpha_t - \overline{\alpha}_t} \qquad (105) \\ &\le \frac{20cc_1}{T}\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log^3 T} + \frac{240cc_1}{T}\theta_t(x)d\log^2 T \le c\theta_t(x)d, \qquad (106)\end{aligned}$$

provided that

$$\frac{40c_1\varepsilon_{\mathsf{score},t}(x)\log^{\frac{3}{2}} T}{T} \le \sqrt{\theta_t(x)d} \qquad \text{and} \qquad T \ge 480c_1\log^2 T.$$

Step 2(a): proof of relation (41). Substituting (105) into (101) and making use of (26) under our assumption on T yield

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

thus establishing (41) by taking $c=2$.

Step 2(b): proof of relation (42a). Suppose now that

$$C_{10}\frac{\theta_t(x)d\log^2 T+\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log^3 T}}{T}\le 1 \tag{107}$$

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

$$\begin{aligned}&\left|-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}+\frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t}\right|\\ &\qquad\le c_{10}\left(\theta_t(x)d\log T+\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log T}\right)\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\end{aligned} \tag{108}$$

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

$$\begin{aligned}&\exp\left(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|u\|_2^2-2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\right)\\ &=1-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}+\frac{u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{\alpha_t-\overline{\alpha}_t}+O\left(\big(\theta_t(x)^2d^2\log^2T+\varepsilon_{\mathsf{score},t}(x)^2\theta_t(x)d\log T\big)\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\right)\end{aligned} \tag{109}$$

for any $x_0\in\mathcal{E}_2^{\mathsf{typical}}$, which invokes (108) and (104b) (under the assumption (107)). Combine (109) and (106) to show that

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

$$+ O\bigg(\theta_t(x)^2 d^2 \Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^2 \log^2 T + \varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x) d \log T}\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)\bigg), \tag{110}$$

where the penultimate relation holds since, according to (102a),

$$\begin{aligned}\sum_{c=3}^{\infty} \int_{x_0 \in \mathcal{E}_c^{\mathsf{typical}} \setminus \mathcal{E}_{c-1}^{\mathsf{typical}}} p_{X_0 \,|\, X_t}(x_0 \,|\, x) \exp\big(c\theta_t(x) d\big)\, \mathrm{d}x_0 &\le \sum_{c=3}^{\infty} \exp\big(-c^2 \theta_t(x) d \log T\big) \exp\big(c\theta_t(x) d\big) \\ &\le \sum_{c=3}^{\infty} \exp\bigg(-\frac{1}{2} c^2 \theta_t(x) d \log T\bigg) \le \exp\big(-\theta_t(x) d \log T\big),\end{aligned}$$

and the last line in (110) again utilizes (26) and the fact that $\theta_t(x) \ge c_6$ for some large enough constant $c_6 > 0$.

Putting (110) and (101) together yields

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

as claimed.

B.1.2 Proof of relation (42b)

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

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

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

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

To begin with, recall from (36) and (24) that

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

As a result, one can use (25a) and (25b) to derive

$$I - \frac{\partial \phi_t^\star(x)}{\partial x} = \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} J_t(x) = \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \bigg\{ I - \frac{1}{1-\overline{\alpha}_t} \underbrace{\mathsf{Cov}\big(X_t - \sqrt{\overline{\alpha}_t} X_0 \mid X_t = x\big)}_{=:\,B} \bigg\}. \tag{112}$$

This allows one to show that

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

Moreover, the matrix B defined in (112) satisfies

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

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

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

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

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

hold as long as $d\|A\|+d\|\Delta\|\le c_{11}$ for some small enough constant $c_{11}>0$. The above properties taken collectively with (36) and (33) allow us to demonstrate that

$$\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)} = \mathsf{det}\Big(\frac{\partial\phi_t(x)}{\partial x}\Big)^{-1} = \bigg(\mathsf{det}\bigg(\frac{\partial\phi_t^\star(x)}{\partial x}+\frac{1-\alpha_t}{2}\Big[J_{s_t}(x)-J_{s_t^\star}(x)\Big]\bigg)\bigg)^{-1}$$
$$= \bigg(\mathsf{det}\bigg(I+\frac{\partial\phi_t^\star(x)}{\partial x}-I+\frac{1-\alpha_t}{2}\Big[J_{s_t}(x)-J_{s_t^\star}(x)\Big]\bigg)\bigg)^{-1} \tag{115}$$
$$= 1-\mathsf{Tr}\Big(\frac{\partial\phi_t^\star(x)}{\partial x}-I\Big)+O\bigg(\theta_t(x)^2d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2T+\theta^3d^6\log^3T\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3+(1-\alpha_t)d\varepsilon_{\mathsf{Jacobi},t}(x)\bigg)$$
$$= 1+\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\Big(\big\|\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2-\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}$$
$$+O\bigg(\theta_t(x)^2d^2\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^2\log^2T+(1-\alpha_t)d\varepsilon_{\mathsf{Jacobi},t}(x)\bigg), \tag{116}$$

with the proviso that

$$\frac{d^2(1-\alpha_t)\log T}{\alpha_t-\overline{\alpha}_t}\le\frac{8c_1d^2\log^2T}{T}\le c_{12}\qquad\text{and}\qquad(1-\alpha_t)d\varepsilon_{\mathsf{Jacobi},t}(x)\le\frac{c_1d\varepsilon_{\mathsf{Jacobi},t}(x)\log T}{T}\le c_{12}$$

for some sufficiently small constant $c_{12}>0$ (see (26)).

B.2 Proof of Lemma 5

Before proceeding, let us make note of several basic facts: for any x with $\theta_t(x)\lesssim 1$, Lemma 1 and (26) taken together reveal that:

$$\bigg|\frac{1-\alpha_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\Big(\big\|\mathbb{E}\big[X_t-\sqrt{\overline{\alpha}_t}X_0\,|\,X_t=y_t\big]\big\|_2^2-\mathbb{E}\big[\big\|X_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2^2\,|\,X_t=y_t\big]\Big)\bigg|$$
$$\le\bigg|\frac{(1-\alpha_t)\mathbb{E}\big[\big\|X_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2^2\,|\,X_t=y_t\big]}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\bigg|\lesssim\frac{(1-\alpha_t)d\log T}{\alpha_t-\overline{\alpha}_t}\lesssim\frac{d\log^2T}{T}=o(1) \tag{117a}$$

$$\text{and}\qquad\frac{d(1-\alpha_t)}{\alpha_t-\overline{\alpha}_t}\lesssim\frac{d\log T}{T}=o(1). \tag{117b}$$

Our proof consists of several steps below.

Step 1: obtaining a refined approximation of $\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t(x))}{p_{X_t}(x)}$. To begin with, recalling the definition of $\mathcal{E}_c^{\mathsf{typical}}$ in (103), we can repeat the arguments in (100) and (110) to reach

$$\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)}$$

$$
\begin{aligned}
&= \Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} \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{2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)-\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&= \Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2}\bigg\{\int_{x_0\in\mathcal{E}_2^{\mathsf{typical}}} 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{2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)-\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&\qquad + O\Big(\exp\big(-\theta_t(x)d\log T\big)\Big)\bigg\} \\
&= \Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2}\bigg\{\int_{x_0\in\mathcal{E}_2^{\mathsf{typical}}} p_{X_0\,|\,X_t}(x_0\,|\,x) \exp\bigg(\frac{(1-\alpha_t)\big[\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathbb{E}\big[x-\sqrt{\overline{\alpha}_t}X_0\mid X_t=x\big]-\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big]}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\bigg) \\
&\qquad\cdot\exp\bigg(-\frac{(1-\alpha_t)\big(s_t(x)-s_t^\star(x)\big)^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)} - \frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 + O\Big(\exp\big(-\theta_t(x)d\log T\big)\Big)\bigg\} \\
&= O\Big(\exp\big(-\theta_t(x)d\log T\big)\Big) + \bigg(1+O\Big(\frac{d\log^2 T}{T^2}\Big)\bigg)\cdot \\
&\quad\int_{x_0\in\mathcal{E}_2^{\mathsf{typical}}} p_{X_0\,|\,X_t}(x_0\,|\,x) \exp\bigg(\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\big[\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathbb{E}\big[x-\sqrt{\overline{\alpha}_t}X_0\mid X_t=x\big]-\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big]}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\bigg) \\
&\qquad\cdot\exp\bigg(-\frac{(1-\alpha_t)\big(s_t(x)-s_t^\star(x)\big)^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)} - \frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0, \qquad (118)
\end{aligned}
$$

where we remind the reader of the definition of u in (96). Here, the last line in (118) follows since

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

a consequence of the property (26g) and the fact $\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t} \lesssim \frac{\log T}{T}$ (cf. (26b)).

Moreover, following the arguments in (104), we can easily derive that: for any $x_0 \in \mathcal{E}_2^{\mathsf{typical}}$,

$$
\begin{aligned}
&\exp\bigg(-\frac{(1-\alpha_t)\big(s_t(x)-s_t^\star(x)\big)^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)} - \frac{\|u\|_2^2}{2(\alpha_t-\overline{\alpha}_t)}\bigg) \\
&= 1 + O\bigg(\frac{(1-\alpha_t)\|s_t(x)-s_t^\star(x)\|_2\|x-\sqrt{\overline{\alpha}_t}x_0\|_2}{\alpha_t-\overline{\alpha}_t} + \frac{\|u\|_2^2}{\alpha_t-\overline{\alpha}_t}\bigg) \\
&= 1 + O\bigg(\frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{\theta_t(x)d\log^3 T}}{T} + \frac{\varepsilon_{\mathsf{score},t}(x)^2\log^2 T}{T^2} + \frac{\theta_t(x)d\log^3 T}{T^2}\bigg) \\
&= 1 + O\bigg(\frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T} + \frac{d\log^3 T}{T^2}\bigg), \qquad (119)
\end{aligned}
$$

where the last line invokes the assumptions $\theta_t(x) \lesssim 1$ and $\frac{\varepsilon_{\mathsf{score},t}(x)\log^{3/2}T}{T} \lesssim \sqrt{\theta_t(x)d} \lesssim \sqrt{d}$. With (118) and (119) in place, we obtain

$$
\begin{aligned}
&\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} = O\Big(\exp\big(-\theta_t(x)d\log T\big)\Big) + \bigg(1+O\Big(\frac{d\log^3 T}{T^2} + \frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T}\Big)\bigg)\cdot \\
&\int_{x_0\in\mathcal{E}_2^{\mathsf{typical}}} p_{X_0\,|\,X_t}(x_0\,|\,x) \exp\bigg(\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\big[\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathbb{E}\big[x-\sqrt{\overline{\alpha}_t}X_0\mid X_t=x\big]-\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big]}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\bigg)\mathrm{d}x_0 \\
&= O\Big(\frac{d\log^3 T}{T^2} + \frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T}\Big) +
\end{aligned}
$$

$$
\begin{aligned}
&\int_{x_0\in\mathcal{E}_2^{\text{typical}}} p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\left(\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\big[\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathbb{E}\big[x-\sqrt{\overline{\alpha}_t}X_0\mid X_t=x\big]-\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big]}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right)\mathrm{d}x_0\\
&=O\bigg(\frac{d\log^3 T}{T^2}+\frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T}\bigg)+\\
&\int p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\left(\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\big[\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathbb{E}\big[x-\sqrt{\overline{\alpha}_t}X_0\mid X_t=x\big]-\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big]}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right)\mathrm{d}x_0,
\end{aligned}
\tag{120}
$$

where the penultimate inequality invokes (117) (so that the integral above is at most $O(1)$), and the last inequality repeats the arguments in (110) once again to demonstrate that

$$
\begin{aligned}
&\int_{x_0\notin\mathcal{E}_2^{\text{typical}}} p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\left(\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\big[\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathbb{E}\big[x-\sqrt{\overline{\alpha}_t}X_0\mid X_t=x\big]-\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big]}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right)\mathrm{d}x_0\\
&\qquad\lesssim\exp\big(-\theta_t(x)d\log T\big)\lesssim\frac{d\log^3 T}{T^2}.
\end{aligned}
$$

Step 2: obtaining a refined approximation on $\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)}$. For any matrix $\Delta\in\mathbb{R}^{d\times d}$ and any symmetric matrix $A\in\mathbb{R}^{d\times d}$ obeying $\|A\|<1/2$ and $\|\Delta\|<1/2$, elementary linear algebra (e.g., Weyl's inequality) tells us that

$$
\begin{aligned}
\sum_{i=1}^d\sigma_i(I+A+\Delta)&=\sum_{i=1}^d\big(\sigma_i(I+A)+O(\|\Delta\|)\big)=\sum_{i=1}^d\lambda_i(I+A)+O(d\|\Delta\|)=d+\mathsf{Tr}(A)+O(d\|\Delta\|),\\
\sum_{i=1}^d\big(\sigma_i(I+A)-1\big)^2&=\sum_{i=1}^d\big(\lambda_i(I+A)-1\big)^2=\sum_{i=1}^d\big(\lambda_i(A)\big)^2=\|A\|_{\mathrm{F}}^2,\\
\sum_{i=1}^d\big(\sigma_i(I+A+\Delta)-1\big)^2&\le 2\sum_{i=1}^d\big(\sigma_i(I+A)-1\big)^2+2\sum_{i=1}^d\big(\sigma_i(I+A+\Delta)-\sigma_i(I+A)\big)^2\\
&\le 2\|A\|_{\mathrm{F}}^2+2d\|\Delta\|^2,
\end{aligned}
$$

with $\sigma_i(Z)$ (resp. $\lambda_i(Z)$) representing the i-th largest singular value (resp. eigenvalue) of a matrix Z. These properties in turn allow one to derive

$$
\begin{aligned}
\log\big|\det(I+A+\Delta)\big|&=\sum_{i=1}^d\log\big(\sigma_i(I+A+\Delta)\big)=\sum_{i=1}^d\big(\sigma_i(I+A+\Delta)-1\big)+O\bigg(\sum_{i=1}^d\big(\sigma_i(I+A+\Delta)-1\big)^2\bigg)\\
&=\mathsf{Tr}(A)+O\big(d\|\Delta\|+\|A\|_{\mathrm{F}}^2+d\|\Delta\|^2\big)=\mathsf{Tr}(A)+O\big(\|A\|_{\mathrm{F}}^2+d\|\Delta\|\big).
\end{aligned}
$$

With this approximation for the log-determinant function in mind, we can invoke (115) to obtain

$$
\begin{aligned}
&\log\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)}=-\log\left|\mathsf{det}\left(I_d+\frac{\partial\phi_t^\star(x)}{\partial x}-I+\frac{1-\alpha_t}{2}\Big[J_{s_t}(x)-J_{s_t^\star}(x)\Big]\right)\right|\\
&=-\mathsf{Tr}\left(\frac{\partial\phi_t^\star(x)}{\partial x}-I\right)+O\left(\left\|\frac{\partial\phi_t^\star(x)}{\partial x}-I\right\|_{\mathrm{F}}^2+d(1-\alpha_t)\varepsilon_{\mathsf{Jacobi},t}(x)\right)\\
&=O\left(\left\|\frac{\partial\phi_t^\star(x)}{\partial x}-I\right\|_{\mathrm{F}}^2+d(1-\alpha_t)\varepsilon_{\mathsf{Jacobi},t}(x)\right)+\left(1+O\left(\frac{\log T}{T}\right)\right)\\
&\quad\cdot\left\{\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\Big(\big\|\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2-\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right\}\\
&=\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)}+\frac{(1-\alpha_t)\Big(\big\|\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2-\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}
\end{aligned}
$$

$$+ O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + d(1-\alpha_t)\varepsilon_{\mathsf{Jacobi},t}(x) + \frac{d\log^2 T}{T^2}\bigg),$$

where the penultimate relation arises from (113a) and the following fact (which uses (26b))

$$\left|\frac{\frac{1}{\alpha_t-\overline{\alpha}_t} - \frac{1}{1-\overline{\alpha}_t}}{\frac{1}{1-\overline{\alpha}_t}}\right| = \frac{\frac{1-\alpha_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}}{\frac{1}{1-\overline{\alpha}_t}} = \frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t} = O\bigg(\frac{\log T}{T}\bigg),$$

and the last relation applies Lemma 1 (under the assumption $\theta_t(x) \lesssim 1$) and (26). It is then easily seen that

$$\begin{aligned}
&\frac{p_{\phi_t(Y)}(\phi_t(x))}{p_Y(x)} = \exp\left(O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + d(1-\alpha_t)\varepsilon_{\mathsf{Jacobi},t}(x) + \frac{d\log^2 T}{T^2}\bigg)\right) \\
&\cdot \exp\left(\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\Big(\big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2 - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right) \\
&= O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + d(1-\alpha_t)\varepsilon_{\mathsf{Jacobi},t}(x) + \frac{d\log^2 T}{T^2}\bigg) \\
&+ \exp\left(\frac{d(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{(1-\alpha_t)\Big(\big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2 - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right).
\end{aligned} \tag{121}$$

Step 3: computing the density ratio of interest. From relations (42a) and (42b) in Lemma 4 as well as (117), we see that

$$\frac{p_{\phi_t(Y_t)}(\phi_t(x))}{p_{Y_t}(x)} = 1 + o(1) \qquad \text{and} \qquad \frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} = 1 + o(1).$$

Then, compare the preceding two results (120) and (121) (with Y chosen to be Y_t) to arrive at

$$\begin{aligned}
&\frac{p_{\phi_t(Y_t)}(\phi_t(x))}{p_{Y_t}(x)} \Big/ \frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} \\
&\quad = \frac{g_1(x)}{g_2(x)} + O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + \frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T} + d(1-\alpha_t)\varepsilon_{\mathsf{Jacobi},t}(x) + \frac{d\log^3 T}{T^2}\bigg),
\end{aligned}$$

where the two functions $g_1(\cdot)$ and $g_2(\cdot)$ are defined as

$$\begin{aligned}
g_1(x) &\coloneqq \exp\left(\frac{(1-\alpha_t)\Big(\big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2 - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0\Big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right), \\
g_2(x) &\coloneqq \int p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\left(\frac{(1-\alpha_t)\big[\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top \mathbb{E}\big[x-\sqrt{\overline{\alpha}_t}X_0 \mid X_t = x\big] - \big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big]}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right)\mathrm{d}x_0.
\end{aligned}$$

Jensen's inequality tells us that $g_1(x) \le g_2(x)$, and hence we can write

$$\begin{aligned}
&\frac{p_{\phi_t(Y_t)}(\phi_t(x))}{p_{Y_t}(x)} \Big/ \frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t(x)\big)}{p_{X_t}(x)} \\
&\quad = 1 + \zeta_t(x) + O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + \frac{\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}}{T} + \frac{d\log T\varepsilon_{\mathsf{Jacobi},t}(x)}{T} + \frac{d\log^3 T}{T^2}\bigg)
\end{aligned}$$

for those x obeying the assumptions of this lemma, where $\zeta_t(\cdot) = g_1(\cdot)/g_2(\cdot) - 1$ is some function obeying $\zeta_t(x) \le 0$.

Similarly, replacing ϕ_t (resp. Y_t) with $\phi_t^\star$ (resp. X_t) in the above display and repeating the same arguments, we arrive at

$$\frac{p_{\phi_t^\star(X_t)}(\phi_t^\star(x))}{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))} = \frac{p_{\phi_t^\star(X_t)}(\phi_t^\star(x))}{p_{X_t}(x)} \Big/ \frac{p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))}{p_{X_t}(x)} = 1 + \zeta_t(x) + O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + \frac{d\log^3 T}{T^2}\bigg). \tag{122}$$

The careful reader would immediately note that we have not yet defined $\zeta_t(\cdot)$ for all x. To ease presentation, we shall simply take $\zeta_t(x) = 0$ for any x that does not satisfy the assumptions of this lemma.

Step 4: bounding the expectation of $\zeta_t(\cdot)$. Define the set

$$\mathcal{E}_\zeta \coloneqq \bigg\{x \mid \theta_t(x) \le 2C_{12}, \frac{C_{10}\theta_t(x) d\log^2 T}{T} \le 1\bigg\}$$

for some large enough constant $C_{12} > 0$. With (122) in place, we have

$$p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\zeta_t(x) = p_{\phi_t(X_t)}(\phi_t^\star(x)) - p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x)) - p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x)) O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + \frac{d\log^3 T}{T^2}\bigg)$$

for any $x \in \mathcal{E}_\zeta$. In addition, according to the properties (115), (117a), Lemma 1, and the assumption that $T \gtrsim d^2\log^5 T$, one can easily derive

$$\begin{aligned}
\bigg|\mathsf{det}\bigg(\frac{\partial \phi_t^\star(x)}{\partial x}\bigg)\bigg| &= \mathsf{det}\bigg(\bigg(1 - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\bigg) I_d + \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)} \mathsf{Cov}\bigg(\frac{X_t - \sqrt{\overline{\alpha}_t}X_0}{\sqrt{1-\overline{\alpha}_t}} \mid X_t = x\bigg)\bigg) \\
&\le \bigg(1 + O\Big(\frac{d\log^2 T}{T}\Big)\bigg)^d = 1 + O\Big(\frac{d^2\log^2 T}{T}\Big) \le 2 \qquad (123a) \\
\text{and} \qquad \bigg|\mathsf{det}\bigg(\frac{\partial \phi_t^\star(x)}{\partial x}\bigg)\bigg| &\ge \bigg(1 - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\bigg)^d = \bigg(1 - O\Big(\frac{\log T}{T}\Big)\bigg)^d \ge \frac{1}{2} \qquad (123b)
\end{aligned}$$

for any $x \in \mathcal{E}_\zeta$. These properties in turn allow one to derive

$$\begin{aligned}
0 &\le -\int p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\zeta_t(x)\mathrm{d}x = -\int_{x\in\mathcal{E}_\zeta} p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\zeta_t(x)\mathrm{d}x \\
&\asymp -\int_{x\in\mathcal{E}_\zeta} p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\zeta_t(x)\mathrm{d}x \\
&= -\int_{x\in\mathcal{E}_\zeta} p_{\phi_t^\star(X_t)}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\mathrm{d}x + \int_{x\in\mathcal{E}_\zeta} p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\mathrm{d}x \\
&\quad + \int_{x\in\mathcal{E}_\zeta} p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg| O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2 + \frac{d\log^3 T}{T^2}\bigg)\mathrm{d}x \\
&\le -\int_{x\in\mathcal{E}_\zeta} p_{\phi_t^\star(X_t)}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\mathrm{d}x + 1 + \int p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x)) O\bigg(\bigg\|\frac{\partial \phi_t^\star(x)}{\partial x} - I\bigg\|_{\mathrm{F}}^2\bigg)\mathrm{d}x + O\bigg(\frac{d\log^3 T}{T^2}\bigg),
\end{aligned} \tag{124}$$

where the last line is valid since

$$\int_{x\in\mathcal{E}_\zeta} p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\mathrm{d}x \le \int p_{\sqrt{\alpha_t}X_{t-1}}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\mathrm{d}x = \int p_{\sqrt{\alpha_t}X_{t-1}}(x)\mathrm{d}x = 1.$$

It then boils down to evaluating $\int_{x\in\mathcal{E}_\zeta} p_{\phi_t^\star(X_t)}(\phi_t^\star(x))\big|\det\frac{\partial \phi_t^\star(x)}{\partial x}\big|\mathrm{d}x$. Towards this end, we make the observation that

$$\int_{x\in\mathcal{E}_\zeta} p_{\phi_t^\star(X_t)}(\phi_t^\star(x))\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\mathrm{d}x = \int_{x\in\mathcal{E}_\zeta} p_{X_t}(x)\bigg|\mathsf{det}\bigg(\frac{\partial \phi_t^\star(x)}{\partial x}\bigg)^{-1}\bigg|\bigg|\det\Big(\frac{\partial \phi_t^\star(x)}{\partial x}\Big)\bigg|\mathrm{d}x$$

$$= \int_{x\in\mathcal{E}_\zeta} p_{X_t}(x)\mathrm{d}x = 1 - \int_{x\notin\mathcal{E}_\zeta} p_{X_t}(x)\mathrm{d}x. \tag{125}$$

Moreover, it is seen from (49) that

$$\mathbb{P}\left(\|X_t\|_2 > T^{c_R+2}\right) \le \exp(-c_6 d\log T),$$

thereby allowing us to derive that

$$\begin{aligned}\int_{x\notin\mathcal{E}_\zeta} p_{X_t}(x)\mathrm{d}x &\le \int_{x:\,\theta_t(x)\le C_{12},\|x\|_2\le T^{c_R+2}} p_{X_t}(x)\mathrm{d}x + \int_{\|x\|_2>T^{c_R+2}} p_{X_t}(x)\mathrm{d}x\\ &\le (T^{c_R+2})^d \exp\left(-2C_{12}d\log T\right) + \exp(-c_6 d\log T)\\ &\le 2\exp\big(-\min\{C_{12},c_6\}d\log T\big).\end{aligned}$$

Combine this with (125) to reach

$$\int_{x\in\mathcal{E}_\zeta} p_{\phi_t^\star(X_t)}\big(\phi_t^\star(x)\big)\left|\det\Big(\frac{\partial\phi_t^\star(x)}{\partial x}\Big)\right|\mathrm{d}x = 1 - O\big(\exp\left(-\min\{C_{12},c_6\}d\log T\right)\big).$$

Substitution into (124) then gives

$$\begin{aligned}0 &\le -\int p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t^\star(x)\big)\zeta_t(x)\mathrm{d}x\\ &\le -1 + O\big(\exp\left(-\min\{C_{12},c_6\}d\log T\right)\big) + 1 + \int p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t^\star(x)\big)O\bigg(\Big\|\frac{\partial\phi_t^\star(x)}{\partial x} - I\Big\|_{\mathrm{F}}^2\bigg)\mathrm{d}x + O\bigg(\frac{d\log^3 T}{T^2}\bigg)\\ &\asymp \int p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t^\star(x)\big)\Big\|\frac{\partial\phi_t^\star(x)}{\partial x} - I\Big\|_{\mathrm{F}}^2\mathrm{d}x + \frac{d\log^3 T}{T^2}.\end{aligned} \tag{126}$$

To finish up, note that Lemma 4 together with Lemma 1 and properties (26) tells us that, for any $x\in\mathcal{E}_\zeta$,

$$p_{X_t}(x) \asymp p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t^\star(x)\big).$$

This taken collectively with (126) leads to the advertised result

$$\begin{aligned}0 \le -\int p_{X_t}(x)\zeta_t(x)\mathrm{d}x &= -\int_{x\in\mathcal{E}_\zeta} p_{X_t}(x)\zeta_t(x)\mathrm{d}x \asymp -\int_{x\in\mathcal{E}_\zeta} p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t^\star(x)\big)\zeta_t(x)\mathrm{d}x\\ &\lesssim \int p_{\sqrt{\alpha_t}X_{t-1}}\big(\phi_t^\star(x)\big)\Big\|\frac{\partial\phi_t^\star(x)}{\partial x} - I\Big\|_{\mathrm{F}}^2\mathrm{d}x + \frac{d\log^3 T}{T^2} \asymp \int p_{X_t}(x)\Big\|\frac{\partial\phi_t^\star(x)}{\partial x} - I\Big\|_{\mathrm{F}}^2\mathrm{d}x + \frac{d\log^3 T}{T^2}.\end{aligned}$$

B.3 Proof of Lemma 6

In view of the definition (54), one has

$$S_k(y_T) \le c_{14}, \qquad \text{for any } k < \tau(y_T). \tag{127}$$

Suppose instead that (56) does not hold true, namely, $-\log q_k(y_k) > 2c_6 d\log T$ for some $k<\tau(y_T)$, and we would like to show that this leads to contradiction.

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

$$\theta_t(y_t) = \max\left\{-\frac{\log q_t(y_t)}{d\log T}, c_6\right\} > 2c_6 = 2\theta_1(y_1), \tag{128}$$

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

$$2c_6 < \theta_t(y_t) \le 4c_6. \tag{129}$$

Assuming the validity of Claim (129) for the moment, it necessarily satisfies

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

According to the relations (43) and (127), we derive

$$\begin{aligned} c_6 = \theta_1(y_1) \le \theta_t(y_t) - \theta_1(y_1) &= -\frac{\log q_t(y_t)}{d \log T} - \theta_1(y_1) \le \frac{-\log q_t(y_t) + \log q_1(y_1)}{d \log T} \\ &= \frac{1}{d \log T} \sum_{j=1}^{t-1} \big(\log q_j(y_j) - \log q_{j+1}(y_{j+1}) \big) \\ &\le 2c_1 + C_{10} \left\{ \frac{d \log^3 T}{T} + \frac{S_{\tau(y_T)-1}(y_T)}{d \log T} \right\} < 3c_1 \end{aligned}$$

under our sample size condition. This, however, cannot possibly hold if $c_6 \ge 3c_1$ as assumed for Lemma 6.

To finish up, it suffices to justify Claim (129). In order to see this, suppose instead that $\theta_t(y_t) > 4c_6$. Given relation (127) that $S_k(y_T) \le c_{14}$, it can be readily seen from (41), (127) as well as the learning rate properties (26) that

$$\begin{aligned} \theta_{t-1}(y_{t-1}) &= \theta_t(y_t) + \theta_{t-1}(y_{t-1}) - \theta_t(y_t) \\ &= \theta_t(y_t) + \theta_{t-1}(y_{t-1}) + \frac{\log q_t(y_t)}{d \log T} \ge \theta_t(y_t) - \frac{\log q_{t-1}(y_{t-1}) - \log q_t(y_t)}{d \log T} \\ &\ge \theta_t(y_t) - \frac{4c_1 \Big(5\varepsilon_{\mathsf{score},t}(y_t) \sqrt{\theta_t(y_t) d \log T} + 60 \theta_t(y_t) d \log T \Big)}{dT} - \frac{\log 2}{d \log T} \\ &\ge \theta_t(y_t) - \frac{4c_1 \Big(5\varepsilon_{\mathsf{score},t}(y_t) \sqrt{d \log T} + 60 d \log T \Big)}{dT} \theta_t(y_t) - \frac{\log 2}{d \log T} \\ &> \frac{1}{2} \theta_t(y_t) > 2c_6, \end{aligned}$$

which is contradictory with the assumption that t is the smallest step obeying $\theta_t(y_t) > 2c_6$. Thus, we complete the proof of relation (56) as required.

B.4 Proof of Lemma 7

Next, consider any y_T, with $\{y_{T-1}, \cdots, y_1\}$ being the associated deterministic sequence (cf. (38))). As an immediate consequence of Lemma 6 and the definition (27) of $\theta_t(\cdot)$, one has

$$\theta_t(y_t) \le 2c_6, \qquad \forall t < \tau(y_T) \tag{130}$$

We then intend to invoke Lemma 5 to control the term of interest. To do so, note that Lemma 1, (26) and the definition (54) of $\tau(y_T)$ taken together reveal that: for all $t < \tau(y_T)$ one has

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

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

and

$$\left| \frac{(1-\alpha_t) \Big(\big\| \mathbb{E}\big[X_t - \sqrt{\overline{\alpha}_t} X_0 \,|\, X_t = y_t \big] \big\|_2^2 - \mathbb{E}\big[\big\| X_t - \sqrt{\overline{\alpha}_t} X_0 \big\|_2^2 \,|\, X_t = y_t \big] \Big)}{(\alpha_t - \overline{\alpha}_t)(1 - \overline{\alpha}_t)} \right|$$

$$\leq \left| \frac{(1-\alpha_t)\mathbb{E}\big[\big\|X_t - \sqrt{\overline{\alpha}_t} X_0\big\|_2^2 \,|\, X_t = y_t\big]}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} \right| \lesssim \frac{(1-\alpha_t) d \log T}{\alpha_t - \overline{\alpha}_t} \lesssim \frac{d \log^2 T}{T} = o(1). \tag{131}$$

With these bounds in mind, applying relations (42a) and (42b) in Lemma 4 leads to

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

for all $t < \tau(y_T)$. Using the fact that $y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \phi_t(y_t)$ and invoking the relation (40), we arrive at

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

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

$$\begin{aligned}
\frac{p_1(y_1)}{q_1(y_1)} &= \left\{ 1 + O\left(\frac{d^2 \log^4 T}{T} + S_{\tau-1}(y_{\tau-1}) \right) \right\} \frac{p_{\tau-1}(y_{\tau-1})}{q_{\tau-1}(y_{\tau-1})} \\
&\in \left[\frac{p_{\tau-1}(y_{\tau-1})}{2 q_{\tau-1}(y_{\tau-1})}, \frac{2 p_{\tau-1}(y_{\tau-1})}{q_{\tau-1}(y_{\tau-1})} \right],
\end{aligned} \tag{132a}$$

and similarly,

$$\frac{q_k(y_k)}{2 p_k(y_k)} \leq \frac{q_1(y_1)}{p_1(y_1)} \leq 2 \frac{q_k(y_k)}{p_k(y_k)}, \qquad \forall k < \tau. \tag{132b}$$

This finishes the proof of the claim (57b).

Regarding the other claim (57a), we first observe from (113b) that

$$\left\| \frac{\partial \phi_t^\star(y_t)}{\partial x} - I \right\|_{\mathrm{F}}^2 \lesssim \left(\frac{\theta_t(y_t) d (1-\alpha_t) \log T}{1 - \overline{\alpha}_t} \right)^2 \lesssim \frac{d^2 \log^3 T}{T^2} \lesssim \frac{1}{T \log T},$$

given our assumption that $T \gtrsim d^2 \log^4 T$. Applying Lemma 4 leads to

$$\begin{aligned}
&\frac{p_{\sqrt{\alpha_t} Y_{t-1}}\big(\phi_t(y_t)\big)}{p_{Y_t}(y_t)} \left(\frac{p_{\sqrt{\alpha_t} X_{t-1}}\big(\phi_t(y_t)\big)}{p_{X_t}(y_t)} \right)^{-1} = \frac{p_{\phi_t(Y_t)}\big(\phi_t(y_t)\big)}{p_{Y_t}(y_t)} \left(\frac{p_{\sqrt{\alpha_t} X_{t-1}}\big(\phi_t(y_t)\big)}{p_{X_t}(y_t)} \right)^{-1} \\
&= 1 + \zeta_t(y_t) + O\left(\left\| \frac{\partial \phi_t^\star}{\partial x}(y_t) - I \right\|_{\mathrm{F}}^2 + \frac{\varepsilon_{\mathsf{score},t}(y_t) \sqrt{d \log^3 T}}{T} + \frac{d \log T \varepsilon_{\mathsf{Jacobi},t}(x)}{T} + \frac{d \log^3 T}{T^2} \right)
\end{aligned}$$

for all $t < \tau(y_T)$, where $\zeta_t(\cdot)$ is the function defined in Lemma 5. Recall the fact that $y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \phi_t(y_t)$ and invoke the relation (40) to arrive at

$$\frac{p_{t-1}(y_{t-1})}{q_{t-1}(y_{t-1})} = \left\{ 1 + \zeta_t(y_t) + O\left(\left\| \frac{\partial \phi_t^\star}{\partial x}(y_t) - I \right\|_{\mathrm{F}}^2 + \frac{d \log^3 T}{T^2} + \frac{\varepsilon_{\mathsf{score},t}(y_t) \sqrt{d \log^3 T}}{T} + \frac{d \varepsilon_{\mathsf{Jacobi},t}(y_t) \log T}{T} \right) \right\} \frac{p_t(y_t)}{q_t(y_t)}$$

for any $t < \tau(y_T)$. Apply this relation recursively over $1 < t < \tau$ to conclude the proof of the claim (57a).

B.5 Proof of Lemma 8

In the following, we shall tackle $\mathcal{I}_2$, $\mathcal{I}_3$ and $\mathcal{I}_4$ separately. Throughout this proof, we shall abbreviate $\tau = \tau(Y_T)$ (cf. (54)) whenever it is clear from the context.

The sub-collection in $\mathcal{I}_2$. By virtue of the definition (60a) of $\mathcal{I}_2$, we make the observation that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_2\right\}\right]\overset{\text{(i)}}{\le}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_2\right\}\frac{S_\tau(Y_T)}{c_{14}}\right]\\
&\overset{\text{(ii)}}{=}\frac{\log T}{c_{14}T}\sum_{t=2}^{\tau}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_2\right\}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)\right]\\
&\overset{\text{(iii)}}{\le}\frac{2\log T}{c_{14}T}\sum_{t=2}^{\tau}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_t(Y_t)}{p_t(Y_t)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_2\right\}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)\right]\\
&\le\frac{2\log T}{c_{14}T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_t(Y_t)}{p_t(Y_t)}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)\right]\\
&=\frac{2\log T}{c_{14}T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_t\sim p_t}\left[\frac{q_t(Y_t)}{p_t(Y_t)}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)\right]\\
&=\frac{2\log T}{c_{14}T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_t\sim q_t}\left[d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right]\\
&\overset{\text{(iv)}}{\lesssim}\left(d\varepsilon_{\mathsf{Jacobi}}+\sqrt{d\log T}\varepsilon_{\mathsf{score}}\right)\log T.
\end{aligned}\tag{133}$$

Here, (i) follows since $S_\tau(y_T)\ge c_{14}$ in $\mathcal{I}_2$ (see (60a)); (ii) comes from the definition of $S_t(\cdot)$ (see (39)); (iii) holds since (by repeating the same proof arguments as for (57) as long as $2c_{14}$ is small enough)

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

and (iv) arises from (34).

The sub-collection in $\mathcal{I}_3$. With regards to $\mathcal{I}_3$ (cf. (60b)), we can derive the following bound in a way similar to (133):

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_3\right\}\right]\overset{\text{(i)}}{\le}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_3\right\}\frac{\xi_\tau(Y_T)}{c_{14}}\right]\\
&=\frac{\log T}{c_{14}T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_3\right\}\left(d\varepsilon_{\mathsf{Jacobi},\tau}(Y_\tau)+\sqrt{d\log T}\varepsilon_{\mathsf{score},\tau}(Y_\tau)\right)\right]\\
&\overset{\text{(ii)}}{\le}\frac{2\log T}{c_{14}T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_{\tau-1}(Y_{\tau-1})}{p_{\tau-1}(Y_{\tau-1})}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_3\right\}\left(d\varepsilon_{\mathsf{Jacobi},\tau}(Y_\tau)+\sqrt{d\log T}\varepsilon_{\mathsf{score},\tau}(Y_\tau)\right)\right]\\
&=\frac{2\log T}{c_{14}T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_3\right\}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)\mathbb{1}\{\tau=t\}\right]\\
&\overset{\text{(iii)}}{\le}\frac{16\log T}{c_{14}T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_\tau(Y_\tau)}{p_\tau(Y_\tau)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_3\right\}\left(d\varepsilon_{\mathsf{Jacobi},\tau}(Y_\tau)+\sqrt{d\log T}\varepsilon_{\mathsf{score},\tau}(Y_\tau)\right)\right]\\
&\le\frac{16\log T}{c_{14}T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_t(Y_t)}{p_t(Y_t)}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)\right]\\
&=\frac{16\log T}{c_{14}T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_t\sim p_t}\left[\frac{q_t(Y_t)}{p_t(Y_t)}\left(d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right)\right]\\
&=\frac{16\log T}{c_{14}T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_t\sim q_t}\left[d\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)\right]\\
&\lesssim\left(d\varepsilon_{\mathsf{Jacobi}}+\sqrt{d\log T}\varepsilon_{\mathsf{score}}\right)\log T.
\end{aligned}\tag{134}\tag{135}$$

Here, (i) comes from (60b), (ii) arises from (57b), whereas (iii) is a consequence of (60b).

The sub-collection in $\mathcal{I}_4$. We now turn attention to $\mathcal{I}_4$ (cf. (60c)), towards which we find it helpful to define

$$\mathcal{J}_{1,t} \coloneqq \Big\{ y_T : \xi_t(y_T) < c_{14} \Big\} \tag{136a}$$

$$\mathcal{J}_{2,t} \coloneqq \left\{ y_T : \xi_t(y_T) \geq c_{14}, \frac{q_{t-1}(y_{t-1})}{p_{t-1}(y_{t-1})} \leq \frac{8q_t(y_t)}{p_t(y_t)} \right\} \tag{136b}$$

$$\mathcal{J}_{3,t} \coloneqq \left\{ y_T : \xi_t(y_T) \geq c_{14}, \frac{q_{t-1}(y_{t-1})}{p_{t-1}(y_{t-1})} > \frac{8q_t(y_t)}{p_t(y_t)} \right\} \tag{136c}$$

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

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4 \right\} \right] &\leq 2 \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[\frac{q_{\tau-1}(Y_1)}{p_{\tau-1}(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4 \right\} \right] \\
&= 2 \sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4 \right\} \mathbb{1}\{\tau = t\} \right] \\
&\leq 2 \sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{J}_{3,t} \right\} \right],
\end{aligned} \tag{137}$$

where the first inequality follows from (57b), and the last line comes from the definition of $\mathcal{I}_4$ (cf. (60c)) and $\mathcal{J}_{3,t}$ (cf. (136c)). For notational simplicity, let us define, for $2 \leq t \leq T$,

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

In view of the second inequality in (136c), one has $h_{t-1} > 8h_t$ as long as $y_T \in \mathcal{J}_{3,t}$. Consequently,

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

Here, the second line holds true since, for all t, one has (i) $\mathcal{J}_{1,t} \cup \mathcal{J}_{2,t} \cup \mathcal{J}_{3,t} = \mathbb{R}^d$, and (ii) $\mathcal{J}_{1,t}$, $\mathcal{J}_{2,t}$ and $\mathcal{J}_{3,t}$ are disjoint. Substituting this into (137), we arrive at

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{J}_{3,t} \right\} \right] &\leq 2 \sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[h_{t-1} \mathbb{1}\left\{ Y_T \in \mathcal{J}_{3,t} \right\} \right] \\
&\leq \frac{8}{7} \sum_{t=2}^{T} \Big(\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[h_t \mathbb{1}\left\{ Y_T \in \mathcal{J}_{1,t} \cup \mathcal{J}_{2,t} \right\} \right] - \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[h_{t-1} \mathbb{1}\left\{ Y_T \in \mathcal{J}_{1,t} \cup \mathcal{J}_{2,t} \right\} \right] \Big) \\
&\quad + \frac{8}{7} \sum_{t=2}^{T} \Big(\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[h_{t-1} \right] - \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[h_t \right] \Big).
\end{aligned} \tag{138}$$

In order to further bound (138), we make note of a few basic facts. Firstly, the identity below holds:

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

Secondly, by defining the set

$$\mathcal{E}_t \coloneqq \Big\{ y : q_t(y) > \exp\big(-c_6 d \log T\big) \Big\}, \qquad 2 \le t \le T, \tag{139}$$

we can show that

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

where the penultimate line comes from (49), and the last inequality holds true as long as c_6 is large enough. Plugging the preceding two results into (138), we reach

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4 \right\} \right] \le \frac{8}{7} \sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[(h_t - h_{t-1}) \mathbb{1}\left\{ y_t \in \mathcal{E}_t, Y_T \in \mathcal{J}_{1,t} \right\} \right] \\
+ \frac{8}{7} \sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[h_t \mathbb{1}\left\{ Y_T \in \mathcal{J}_{2,t} \right\} \right] + \exp\Big(-\frac{c_6}{2} d \log T\Big).
\end{aligned} \tag{140}$$

As it turns out, the sum w.r.t. the set $\mathcal{J}_{1,t}$ and the sum w.r.t. the set $\mathcal{J}_{2,t}$ in (140) can be controlled respectively using the same arguments as for $\mathcal{I}_1$ and $\mathcal{I}_3$ to derive

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

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

$$\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[\frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_4 \right\} \right] \lesssim \frac{d \log^4 T}{T} + \sqrt{d \log^3 T} \varepsilon_{\mathsf{score}} + (d \log T) \varepsilon_{\mathsf{Jacobi}}. \tag{141}$$

Putting all this together. Taking (133), (135) and (141) together, we establish the advertised result.

References

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

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

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

Benton, J., Deligiannidis, G., and Doucet, A. (2023). Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*.

Biroli, G., Bonnaire, T., De Bortoli, V., and Mézard, M. (2024). Dynamical regimes of diffusion models. *arXiv preprint arXiv:2402.18491*.

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, H., Ren, Y., Ying, L., and Rotskoff, G. M. (2024a). Accelerating diffusion models with parallel sampling: Inference at sub-linear time complexity. *arXiv preprint arXiv:2405.15986*.

Chen, H. and Ying, L. (2024). Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *arXiv preprint arXiv:2402.08095*.

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, M., Mei, S., Fan, J., and Wang, M. (2024b). An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*.

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

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

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

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

Chen, S., Kontonis, V., and Shah, K. (2024c). Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*.

Cheng, X., Lu, J., Tan, Y., and Xie, Y. (2024). Convergence of flow-based generative models via proximal gradient descent in Wasserstein space. *IEEE Transactions on Information Theory*.

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*.

Cui, H., Krzakala, F., Vanden-Eijnden, E., and Zdeborová, L. (2023). Analysis of learning a flow-based generative model from limited sample complexity. *arXiv preprint arXiv:2310.03575*.

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.

Dou, Z., Chen, M., Wang, M., and Yang, Z. (2024). Provable statistical rates for consistency diffusion models. *arXiv preprint arXiv:2406.16213*.

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

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.

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

Feng, O. Y., Kao, Y.-C., Xu, M., and Samworth, R. J. (2024). Optimal convex m-estimation via score matching. *arXiv preprint arXiv:2403.16688*.

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.

Fu, H., Yang, Z., Wang, M., and Chen, M. (2024). Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*.

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

Gao, Y., Huang, J., Jiao, Y., and Zheng, S. (2024). Convergence of continuous normalizing flows for learning probability distributions. *arXiv preprint arXiv:2404.00551*.

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*.

Gupta, S., Cai, L., and Chen, S. (2024). Faster diffusion-based sampling with randomized midpoints: Sequential and parallel. *arXiv preprint arXiv:2406.00924*.

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

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

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

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

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

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*.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. (2024). Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*.

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

Koehler, F., Heckett, A., and Risteski, A. (2023). Statistical efficiency of score matching: The view from isoperimetry. *International Conference on Learning Representations*.

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

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

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

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

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. (2024a). Accelerating convergence of score-based diffusion models, provably. *International Conference on Machine Learning*.

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

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

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

Li, G., Wei, Y., Chen, Y., and Chi, Y. (2024c). Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*.

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

Li, M. and Chen, S. (2024). Critical windows: non-asymptotic theory for feature emergence in diffusion models. *arXiv preprint arXiv:2403.01633*.

Liang, Y., Ju, P., Liang, Y., and Shroff, N. (2024). Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*.

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*.

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

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

Montanari, A. and Wu, Y. (2024). Provably efficient posterior sampling for sparse linear regression via measure decomposition. *arXiv preprint arXiv:2406.19550*.

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

Pedrotti, F., Maas, J., and Mondelli, M. (2023). Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*, 4.

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

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

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

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

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

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

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

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency models.

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

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

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

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

Tang, W. and Zhao, H. (2024a). Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*.

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

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

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

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. (2024). Diffusion model learns low-dimensional distributions via subspace clustering.

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

Wu, Y., Chen, M., Li, Z., Wang, M., and Wei, Y. (2024). Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *International Conference on Machine Learning*.

Xu, C., Cheng, X., and Xie, Y. (2024). Normalizing flow neural networks by JKO scheme. *Advances in Neural Information Processing Systems*, 36.

Xu, X. and Chi, Y. (2024). Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *arXiv preprint arXiv:2403.17042*.

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*.

Zhang, H., Zhou, J., Lu, Y., Guo, M., Wang, P., Shen, L., and Qu, Q. (2024). The emergence of reproducibility and consistency in diffusion models. In *Forty-first International Conference on Machine Learning*.