# Accelerating Convergence of Score-Based Diffusion Models, Provably

Gen Li*† CUHK
Yu Huang*‡ UPenn
Timofey Efimov§ CMU
Yuting Wei‡ UPenn
Yuejie Chi§ CMU
Yuxin Chen‡ UPenn

March 6, 2024

**Abstract**

Score-based diffusion models, while achieving remarkable empirical performance, often suffer from low sampling speed, due to extensive function evaluations needed during the sampling phase. Despite a flurry of recent activities towards speeding up diffusion generative modeling in practice, theoretical underpinnings for acceleration techniques remain severely limited. In this paper, we design novel training-free algorithms to accelerate popular deterministic (i.e., DDIM) and stochastic (i.e., DDPM) samplers. Our accelerated deterministic sampler converges at a rate $O(\frac{1}{T^2})$ with $T$ the number of steps, improving upon the $O(\frac{1}{T})$ rate for the DDIM sampler; and our accelerated stochastic sampler converges at a rate $O(\frac{1}{T})$, outperforming the rate $O(\frac{1}{\sqrt{T}})$ for the DDPM sampler. The design of our algorithms leverages insights from higher-order approximation, and shares similar intuitions as popular high-order ODE solvers like the DPM-Solver-2. Our theory accommodates $\ell_2$-accurate score estimates, and does not require log-concavity or smoothness on the target distribution.

**Keywords:** diffusion models, training-free samplers, DDPM, DDIM, probability flow ODE, higher-order ODE

## Contents



*The first two authors contributed equally.
†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.

# 1 Introduction

Initially introduced by Sohl-Dickstein et al. (2015) and subsequently gaining momentum through the works Ho et al. (2020); Song et al. (2021), diffusion models have risen to the forefront of generative modeling. Remarkably, score-based diffusion models have demonstrated superior performance across various domains like computer vision, natural language processing, medical imaging, and bioinformatics (Croitoru et al., 2023; Yang et al., 2023; Kazerouni et al., 2023; Guo et al., 2023), outperforming earlier generative methods such as GANs (Goodfellow et al., 2020) and VAEs (Kingma and Welling, 2014) on multiple fronts (Dhariwal and Nichol, 2021).

## 1.1 Score-based diffusion models

On a high level, diffusion-based generative modeling begins by considering a forward Markov diffusion process that progressively diffuses a data distribution into noise:

$$X_0 \overset{\text{add noise}}{\longrightarrow} X_1 \overset{\text{add noise}}{\longrightarrow} X_2 \overset{\text{add noise}}{\longrightarrow} \cdots \overset{\text{add noise}}{\longrightarrow} X_T, \tag{1}$$

where $X_0 \sim p_{\mathsf{data}}$ is drawn from the target data distribution in $\mathbb{R}^d$, and $X_T$ resembles pure noise (e.g., with a distribution close to $\mathcal{N}(0, I_d)$). The pivotal step then lies in learning to construct a reverse Markov process

$$Y_0 \overset{\text{use scores}}{\longleftarrow} Y_1 \overset{\text{use scores}}{\longleftarrow} Y_2 \overset{\text{use scores}}{\longleftarrow} \cdots \overset{\text{use scores}}{\longleftarrow} Y_T, \tag{2}$$

which starts from purse noise $Y_T \sim \mathcal{N}(0, I_d)$ and maintains distributional proximity throughout in the sense that $Y_t \overset{\mathrm{d}}{\approx} X_t$ $(t \leq T)$. To accomplish this goal, $Y_{t-1}$ in each step is typically obtained from $Y_t$ with the aid of (Stein) score functions — namely, $\nabla_X \log p_{X_t}(X)$, with $p_{X_t}$ denoting the distribution of $X_t$ — where the score functions are pre-trained by means of score matching techniques (e.g., Hyvärinen (2005); Ho et al. (2020); Hyvärinen (2007); Vincent (2011); Song and Ermon (2019); Pang et al. (2020)).

The mainstream approaches for constructing the reverse-time process (2) can roughly be divided into two categories, as described below.

- *Stochastic (or SDE-based) samplers.* A widely adopted strategy involves exploiting both the score function and some injected random noise when generating each $Y_{t-1}$; that is, $Y_{t-1}$ is taken to be a function of $Y_t$ and some independent noise $Z_t$. A prominent example of this kind is the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), to be detailed in Section 2. Notably, this approach has intimate connections with certain stochastic differential equations (SDEs), which can be elucidated via celebrated SDE results concerning the existence of reverse-time diffusion processes (Anderson, 1982; Haussmann and Pardoux, 1986).

- *Deterministic (or ODE-based) samplers.* In contrast, another approach is purely deterministic (except for the generation of $Y_T$), constructing $Y_{t-1}$ as a function of the previously computed steps (e.g., $Y_t$) without injecting any additional noise. This approach was introduced by Song et al. (2021), as inspired by the existence of ordinary differential equations (ODEs) — termed *probability flow ODEs* — exhibiting the same marginal distributions as the above-mentioned reverse-time diffusion process. A notable example in this category is often referred to as the Denoising Diffusion Implicit Model (DDIM) (Song et al., 2020).

In practice, it is often observed that DDIM converges more rapidly than DDPM, although the final data instances produced by DDPM (given sufficient runtime) enjoy better diversity compared to the output of DDIM.

### 1.2 Non-asymptotic convergence theory and acceleration

Despite the astounding empirical success, theoretical analysis for diffusion-based generative modeling is still in its early stages of development. Treating the score matching step as a blackbox and exploiting only (crude) information about the score estimation error, a recent strand of works have explored the convergence rate of the data generating process (i.e., the reverse Markov process) in a non-asymptotic fashion, in an attempt to uncover how fast sampling can be performed (e.g., Lee et al. (2022, 2023); Chen et al. (2022, 2023a,c,b); Li et al. (2023); Benton et al. (2023b,a); Liang et al. (2024)). In what follows, let us give a brief overview of the state-of-the-art results in this direction. Here and throughout, the iteration complexity of a sampler refers to the number of steps $T$ needed to attain $\varepsilon$ accuracy in the sense that $\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \varepsilon$, where $\mathsf{TV}(\cdot,\cdot)$ represents the total-variation (TV) distance between two distributions, and $p_{X_1}$ (resp. $p_{Y_1}$) stands for the distribution of $X_1$ (resp. $Y_1$).

- *Convergence rate of stochastic samplers.* Assuming Lipschitz continuity (or smoothness) of the score functions across all steps, Chen et al. (2022) proved that the iteration complexity of the DDPM sampler is proportional to $1/\varepsilon^2$. The Lipschitz assumption is then relaxed by Chen et al. (2023a); Benton et al. (2023a); Li et al. (2023), revealing that the scaling $1/\varepsilon^2$ is achievable for a fairly general family of data distributions.

- *Convergence rate of deterministic samplers.* As alluded to previously, deterministic samplers often exhibit faster convergence in both practice and theory. For instance, Chen et al. (2023c) provided the first polynomial convergence guarantees for the probability flow ODE sampler under exact scores, whereas Li et al. (2023) demonstrated that its iteration complexity scales proportionally to $1/\varepsilon$ allowing score estimation error. Additionally, it is noteworthy that an iteration complexity proportional to $1/\varepsilon$ has also been established by Chen et al. (2023b) for a variant of the probability flow ODE sampler, although the sampler studied therein incorporates a stochastic corrector step in each iteration.

**Acceleration?** While the theoretical studies outlined above have offered non-asymptotic convergence guarantees for both the stochastic and deterministic samplers, one might naturally wonder whether there is potential for achieving faster rates. In practice, the evaluation of Stein scores in each step often entails computing the output of a large neural network, thereby calling for new solutions to reduce the number of score evaluations without compromising sampling fidelity. Indeed, this has inspired a large strand of recent works focused on speeding up diffusion generative modeling. Towards this end, one prominent approach is *distillation*, which attempts to distill a pre-trained diffusion model into another model (e.g., progressive distillation, consistency model) that can be executed in significantly fewer steps (Luhman and Luhman, 2021; Salimans and Ho, 2021; Meng et al., 2023; Song et al., 2023). However, while distillation-based techniques have

achieved outstanding empirical performance, they often necessitate additional training processes, imposing high computational burdens beyond score matching. In contrast, an alternative route towards acceleration is "training-free," which directly invokes the pre-trained diffusion model (particularly the pre-trained score functions) for sampling without requiring additional training processes. Examples of training-free accelerated samplers include DPM-Solver (Lu et al., 2022a), DPM-Solver++ (Lu et al., 2022b), DEIS (Zhang and Chen, 2022), UniPC (Zhao et al., 2023), the SA-Solver (Xue et al., 2023), among others, which leverage faster solvers for ODE and SDE using only the pre-trained score functions. Nevertheless, non-asymptotic convergence analyses for these methods remain largely absent, making it challenging to rigorize the degrees of acceleration compared to the non-accelerated results (Lee et al., 2023; Chen et al., 2022, 2023a; Li et al., 2023; Benton et al., 2023a). All of this leads to the following question that we aim to explore in this work:

> *Can we design a training-free deterministic (resp. stochastic) sampler that converges provably faster than the DDIM (resp. DDPM)?*

### 1.3 Our contributions

In this paper, we answer the above question in the affirmative. Our main contributions can be summarized as follows.

- In the deterministic setting, we demonstrate how to speed up the ODE-based sampler (i.e., the DDIM-type sampler). The proposed sampler, which exploits some sort of momentum term to adjust the update rule, leverages insights from higher-order ODE approximation in discrete time and shares similar intuitions with the fast ODE-based sampler DPM-Solver-2 (Lu et al., 2022a). We establish non-asymptotic convergence guarantees for the accelerated DDIM-type sampler, showing that its iteration complexity scales proportionally to $1/\sqrt{\varepsilon}$ (up to log factor). This substantially improves upon the prior convergence theory for the original DDIM sampler (Li et al., 2023) (which has an iteration complexity proportional to $1/\varepsilon$).
- In the stochastic setting, we propose a novel sampling procedure to accelarate the SDE-based sampler (i.e., the DDPM-type sampler). For this new sampler, we establish an iteration complexity bound proportional to $1/\varepsilon$ (modulo some log factor), thus unveiling the superiority of the proposed sampler compared to the original DDPM sampler (recall that the original DDPM sampler has an iteration complexity proportional to $1/\varepsilon^2$ (Li et al., 2023; Chen et al., 2023a, 2022)).

In addition, two aspects of our theory are worth emphasizing: (i) our theory accommodates $\ell_2$-accurate score estimates, rather than requiring $\ell_\infty$ score estimation accuracy; (ii) our theory covers a fairly general family of target data distributions, without imposing stringent assumptions like log-concavity and smoothness on the target distributions.

### 1.4 Other related works

We now briefly discuss additional related works in the prior art.

**Convergence of score-based generative models (SGMs).** For stochastic samplers of SGMs, the convergence guarantees were initially provided by early works including but not limited to De Bortoli et al. (2021); Liu et al. (2022b); Pidstrigach (2022); Block et al. (2020); De Bortoli (2022); Wibisono and Yang (2022); Gao et al. (2023), which often faced issues of either being not quantitative or suffering from the curse of dimensionality. More recent research has advanced this field by relaxing the assumptions on the score function and achieving polynomial convergence rates (Lee et al., 2022, 2023; Chen et al., 2022, 2023a,b; Li et al., 2023; Benton et al., 2023a; Liang et al., 2024; Tang and Zhao, 2024b). Furthermore, theoretical insights into probability flow-based ODE samplers, though less abundant, have been explored in recent works (Chen et al., 2023c; Li et al., 2023; Chen et al., 2023b; Benton et al., 2023b; Gao and Zhu, 2024). Additionally, Tang and Zhao (2024a) provided a continuous-time sampling error guarantee for a novel class of contraction diffusion models. Gao and Zhu (2024) studies the convergence properties for general probability flow ODEs w.r.t. Wasserstein distances. Most recently, Chen and Ying (2024) makes a step towards the convergence analysis of discrete state space diffusion model. Note that this body of research primarily aims

to quantify the proximity between distributions generated by SGMs and the ground truth distributions, assuming availability of an accurate score estimation oracle. Interestingly, a very recent research by Li et al. (2024b) reveals that even SGMs with empirically optimized score functions might underperform due to strong memorization effects. Moreover, some works delve into other aspects of the theoretical understanding of diffusion models. Furthermore, Wu et al. (2024) investigated how diffusion guidance combined with DDPM and DDIM samplers influences the conditional sampling quality.

**Fast sampling in diffusion models.** A recent strand of works to achieve few-step sampling — or even one-step sampling — falls under the category of training-based samplers, primarily focused on knowledge distillation (Meng et al., 2023; Salimans and Ho, 2021; Song et al., 2023). This method aims to distill a pre-trained diffusion model into another model that can be executed in significantly fewer steps. The recent work Li et al. (2024a) provided a first attempt towards theoretically understanding the sampling efficiency of consistency models. Another line of works aims to design training-free samplers (Lu et al., 2022a,b; Zhao et al., 2023; Zhang and Chen, 2022; Liu et al., 2022a; Zhang et al., 2022), which addresses the efficiency issue by developing faster solvers for the reverse-time SDE or ODE without requiring other information beyond the pre-trained SGMs. In addition, Li et al. (2023); Liang et al. (2024) introduced accelerated samplers that require additional training pertaining to estimating Hessian information at each step. Furthermore, combining GAN with diffusion has shown to be an effective strategy to speed up the sampling process (Wang et al., 2022; Xiao et al., 2021).

### 1.5 Notation

Before continuing, we find it helpful to introduce some notational conventions to be used throughout this paper. Capital letters are often used to represent random variables/vectors/processes, while lowercase letters denote deterministic variables. When considering two probability measures $P$ and $Q$, we define their total-variation (TV) distance as $\mathsf{TV}(P,Q) \coloneqq \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$, and the Kullback-Leibler (KL) divergence as $\mathsf{KL}(P \,\|\, Q) \coloneqq \int \big(\log \frac{\mathrm{d}P}{\mathrm{d}Q}\big)\mathrm{d}P$. We use $p_X(\cdot)$ and $p_{X\,|\,Y}(\cdot\,|\,\cdot)$ to denote the probability density function of a random vector $X$, and the conditional probability of $X$ given $Y$, respectively. For matrices, $\|A\|$ and $\|A\|_{\mathrm{F}}$ refer to the spectral norm and Frobenius norm of a matrix $A$, respectively. For vector-valued functions $f$, we use $J_f$ or $\frac{\partial f}{\partial x}$ to represent the Jacobian matrix of $f$. Given two functions $f(d,T)$ and $g(d,T)$, we employ 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 indicate the existence of a universal constant $C_1 > 0$ such that for all $d$ and $T$, $f(d,T) \le C_1 g(d,T)$ (resp. $f(d,T) \ge C_1 g(d,T)$). The notation $f(d,T) \asymp g(d,T)$ indicates that both $f(d,T) \lesssim g(d,T)$ and $f(d,T) \gtrsim g(d,T)$ hold at once.

## 2 Problem settings

In this section, we formulate the problem, and introduce a couple of key assumptions.

### 2.1 Model and sampling process

**Forward process.** Consider the forward Markov process (1) in discrete time that starts from the target data distribution $X_0 \sim p_{\mathsf{data}}$ in $\mathbb{R}^d$ and proceeds as follows:

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\,W_t, \qquad t = 1, \cdots, T, \tag{3}$$

where the $W_t$'s are independently drawn from $\mathcal{N}(0, I_d)$. This process is said to be "variance-preserving," in the sense that the covariance $\mathsf{Cov}(X_t) = I_d$ holds throughout if $\mathsf{Cov}(X_0) = I_d$. Taking

$$\overline{\alpha}_t \coloneqq \prod_{k=1}^{t} \alpha_k \qquad \text{with } \alpha_t \coloneqq 1-\beta_t \tag{4}$$

for every $1 \le t \le T$, one can write

$$X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t \quad \text{for } \overline{W}_t \sim \mathcal{N}(0, I_d). \tag{5}$$

Throughout the paper, we shall use $q_t(\cdot)$ or $p_{X_t}(\cdot)$ interchangeably to denote the probability density function (PDF) of $X_t$. While we shall concentrate on the discrete-time process in the current paper, we shall note that the forward process has also been commonly studied in the continuous-time limit through the following diffusion process

$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}W_t \quad (0 \le t \le T), \quad X_0 \sim p_{\text{data}} \tag{6}$$

for some function $\beta(t)$ related to the learning rate, where $W_t$ is the standard Brownian motion.

**Score functions and score estimates.** A key ingredient that plays a pivotal role in the sampling process is the (Stein) score function, defined as the log marginal density of the forward process.

**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

$$s_t^\star(X) \coloneqq \nabla \log q_t(X) = -\frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0 \,|\, X_t}(x_0 \,|\, x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0. \tag{7}$$

Here, the last identity follows from standard properties about Gaussians; see, e.g., Chen et al. (2022). In most applications, we have no access to perfect score functions; instead, what we have available are certain estimates for the score functions, to be denoted by $\{s_t(\cdot)\}_{1\le t\le T}$ throughout.

**Data generation process.** The sampling process is performed via careful construction of the reverse process (2) to ensure distributional proximity. Working backward from $t = T, \ldots, 1$, we assume throughout that $Y_T \sim \mathcal{N}(0, I_d)$.

- *Deterministic sampler.* A deterministic sampler typically chooses $Y_{t-1}$ for each $t$ to be a function of $\{Y_t, \ldots, Y_T\}$. For instance, the following construction

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(Y_t + \frac{1-\alpha_t}{2}s_t(Y_t)\right), \qquad t = T, \ldots, 1 \tag{8}$$

  can be viewed as a DDIM-type sampler in discrete time. Note that the DDIM sampler is intimately connected with the following ODE — called the probability flow ODE or the diffusion ODE — in the continuous-time limit:

$$\mathrm{d}\widetilde{Y}_t = -\frac{1}{2}\beta(t)\left(\widetilde{Y}_t + \nabla \log q_t\big(\widetilde{Y}_t\big)\right)\mathrm{d}t, \quad \widetilde{Y}_T \sim q_T, \tag{9}$$

  which enjoys matching marginal distributions as the forward diffusion process (6) in the sense that $\widetilde{Y}_t \overset{\mathrm{d}}{=} X_t$ for all $0 \le t \le T$ (Song et al., 2021).

- *Stochastic sampler.* In contrast to the deterministic case, each $Y_{t-1}$ is a function of not only $\{Y_t, \ldots, Y_T\}$ but also an additional independent noise $Z_t \sim \mathcal{N}(0, I_d)$. One example is the following sampler:

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(Y_t + (1-\alpha_t)s_t(Y_t) + \sqrt{1-\alpha_t}Z_t\Big), \quad t = T, \ldots, 1 \tag{10}$$

  which is closely related to the DDPM sampler in discrete time. The design of DDPM draws inspiration from a well-renowned result in the SDE literature (Anderson, 1982; Haussmann and Pardoux, 1986); namely, there exists a reverse-time SDE

$$\mathrm{d}\widehat{Y}_t = -\frac{1}{2}\beta(t)\left(\widehat{Y}_t + 2\nabla \log q_t\big(\widehat{Y}_t\big)\right)\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\widehat{Z}_t \tag{11}$$

  with $\widehat{Y}_T \sim q_T$ that exhibits the same marginals — $\widehat{Y}_t \overset{\mathrm{d}}{=} X_t$ for all $t$ — as the forward diffusion process (6). Here, $\widehat{Z}_t$ indicates a backward standard Brownian motion.

## 2.2 Assumptions

Before moving on to our algorithms and theory, let us introduce several assumptions that shall be used multiple times in this paper. To begin with, we impose the following assumption on the target data distribution.

**Assumption 1.** *Suppose that $X_0$ is a continuous random vector, and obeys*

$$\mathbb{P}\big(\|X_0\|_2 \le R = T^{c_R} \mid X_0 \sim p_{\mathsf{data}}\big) = 1 \tag{12}$$

*for some arbitrarily large constant $c_R > 0$.*

In words, the size of $X_0$ is allowed to grow polynomially (with arbitrarily large constant degree) in the number of steps, which suffices to accommodate the vast majority of practical applications.

Next, we specify the learning rates $\{\beta_t\}$ (or $\{\alpha_t\}$) employed in the forward process (3). Throughout this paper, we select the same learning rate schedule as in Li et al. (2023), namely,

$$\beta_1 = 1 - \alpha_1 = \frac{1}{T^{c_0}}, \tag{13a}$$

$$\beta_t = 1 - \alpha_t = \frac{c_1 \log T}{T} \min\left\{\beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^t, 1\right\}, \quad t > 1 \tag{13b}$$

for some large enough numerical constants $c_0, c_1 > 0$. In short, there are two phases here: at first $\beta_t$ grows exponentially fast, and then stays unchanged after surpassing some threshold. This also resembles the learning rate choices recommended by Benton et al. (2023a).

Moreover, let us also introduce two assumptions regarding the accuracy of the score estimates $\{s_t\}$, which are adopted in Li et al. (2023). Here and throughout, we denote by

$$J_{s_t^\star} = \frac{\partial s_t^\star}{\partial x} \qquad \text{and} \qquad J_{s_t} = \frac{\partial s_t}{\partial x} \tag{14}$$

the Jacobian matrices of $s_t^\star(\cdot)$ and $s_t(\cdot)$, respectively.

**Assumption 2.** *Suppose that the mean squared estimation error of the score estimates $\{s_t\}_{1\le t\le T}$ obeys*

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

**Assumption 3.** *Suppose that $s_t(\cdot)$ is continuously differentiable for each $1 \le t \le T$, and that the Jacobian matrices associated with the score estimates $\{s_t\}_{1\le t\le T}$ satisfy*

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

In short, Assumption 2 is concerned with the $\ell_2$ score estimation error averaged across all steps, whereas Assumption 3 is about the average discrepancy in the associated Jacobian matrices. It is worth noting that Assumption 3 will only be imposed when analyzing the convergence of deterministic samplers, and is completely unnecessary for the stochastic counterpart.

# 3 Algorithm and main theory

In this section, we put forward two accelerated samplers — an ODE-based algorithm and an SDE-based algorithm — and present convergence theory to confirm the acceleration compared with prior DDIM and DDPM approaches.

### 3.1 Accelerated ODE-based sampler

The first algorithm we propose is an accelerated variant of the ODE-based deterministic sampler. Specifically, starting from $Y_T \sim \mathcal{N}(0, I_d)$, the proposed discrete-time sampler adopts the following update rule:

$$Y_t^- = \Phi_t(Y_t), \qquad Y_{t-1} = \Psi_t(Y_t, Y_t^-) \quad \text{for } t = T, \cdots, 1 \tag{15a}$$

where the mappings $\Phi_t(\cdot)$ and $\Psi_t(\cdot,\cdot)$ are chosen to be

$$\Phi_t(x) = \sqrt{\alpha_{t+1}}\left(x - \frac{1-\alpha_{t+1}}{2} s_t(x)\right), \tag{15b}$$

$$\Psi_t(x,y) = \frac{1}{\sqrt{\alpha_t}}\left(x + \frac{1-\alpha_t}{2} s_t(x) + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\big(s_t(x) - \sqrt{\alpha_{t+1}} s_{t+1}(y)\big)\right), \tag{15c}$$

and we remind the reader that $s_t$ is the score estimate. In contrast to the original DDIM-type solver (8), the proposed accelerated sampler enjoys two distinguishing features:

- In each iteration $t$, the proposed sampler computes a mid-point $Y_t^- = \Phi_t(Y_t)$ (cf. (15b)). As it turns out, this mid-point is designed as a prediction of the probability flow ODE at time $t+1$ using $Y_t$.
- In contrast to (8), the proposed update rule $Y_{t-1} = \Psi_t(Y_t, Y_t^-)$ (see (15c)) includes an additional term that is a properly scaled version of $s_t(Y_t) - \sqrt{\alpha_{t+1}} s_{t+1}(Y_t^-)$. In some sense, this term can be roughly viewed as exploiting "momentum" in adjusting the original sampling rule.

**Theoretical guarantees.** Let us proceed to present our convergence theory and its implications for the proposed deterministic sampler.

**Theorem 1.** *Suppose that Assumptions 1, 2 and 3 hold. Then the proposed sampler (15) with the learning rate schedule (13b) satisfies*

$$\mathsf{TV}\big(q_1, p_1\big) \le C_1 \frac{d^6 \log^6 T}{T^2} + C_1 \sqrt{d \log^3 T}\, \varepsilon_{\mathsf{score}} + C_1 (d \log T) \varepsilon_{\mathsf{Jacobi}} \tag{16}$$

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

We now take a moment to discuss the implications about this theorem.

- *Iteration complexity.* When the target accuracy level $\varepsilon$ is small enough, the number of iterations needed to yield $\mathsf{TV}\big(q_1, p_1\big) \le \varepsilon$ is no larger than

$$\text{(iteration complexity)} \qquad \frac{\mathrm{poly}(d)}{\sqrt{\varepsilon}}, \tag{17}$$

  ignoring any logarithmic factor in $1/\varepsilon$. Clearly, the dependency on $1/\varepsilon$ substantially improves upon the vanilla DDIM sampler, the latter of which has an iteration complexity proportional to $1/\varepsilon$ (Li et al., 2023).

- *Stability vis-a-vis score errors.* The discrepancy between the distribution of $Y_1$ and the target distribution of $X_1$ is proportional to the $\ell_2$ score estimation error $\varepsilon_{\mathsf{score}}$ defined in Assumption 2, as well as the Jacobian error $\varepsilon_{\mathsf{Jacobi}}$ defined in Assumption 3. It is worth noting, however, that the same result might not hold if we remove Assumption 3. More specifically, when only score estimation accuracy is assumed, the deterministic sampler is not guaranteed to achieve small TV error; see Li et al. (2023) for an illustrative example.

**Interpretation via second-order ODE.** In order to help elucidate the rationale of the proposed sampler, we make note of an intimate connection between (15) and high-order ODE, the latter of which has facilitated the design of fast deterministic samplers (e.g., DPM-Solver (Lu et al., 2022a)).

In view of the relation (5), for any $0 < \gamma < 1$, let us first abuse the notation and introduce

$$X(\gamma) \overset{\mathrm{d}}{=} \sqrt{\gamma} X_0 + \sqrt{1-\gamma} Z, \quad Z \sim \mathcal{N}(0, I_d) \tag{18a}$$

$$s_\gamma^\star(X) \coloneqq \nabla_X \log p_{X(\gamma)}(X). \tag{18b}$$

We further consider the following continuous-time analog $\overline{\alpha}(t)$ of the discrete learning rate $\overline{\alpha}_t$ (cf. (4)):

$$\frac{\mathrm{d}\overline{\alpha}(t)}{\mathrm{d}t} = -\beta(t)\overline{\alpha}(t), \quad \overline{\alpha}(T) = \overline{\alpha}_T. \tag{18c}$$

Given that the probability flow ODE (9) yields identical marginal distributions as the forward process $X_t$ (cf. (6)) for every $t$, invoking (18c), we can easily see that $X(\overline{\alpha}(t)) \overset{\mathrm{d}}{=} X_t$ can be generated as follows:

$$\frac{\mathrm{d}X\big(\overline{\alpha}(t)\big)}{\mathrm{d}\overline{\alpha}(t)} = \frac{1}{2\overline{\alpha}(t)}\Big(X\big(\overline{\alpha}(t)\big) + s^\star_{\overline{\alpha}(t)}\Big(X\big(\overline{\alpha}(t)\big)\Big)\Big), \quad X\big(\overline{\alpha}(T)\big) \sim q_T, \tag{19}$$

By taking $f\big(\gamma\big) = \frac{1}{\sqrt{\gamma}} X\big(\gamma\big)$, we can apply (19) to derive

$$\frac{\mathrm{d}f\big(\gamma\big)}{\mathrm{d}\gamma} = -\frac{1}{2\sqrt{\gamma^3}} X\big(\gamma\big) + \frac{1}{\sqrt{\gamma}} \frac{\mathrm{d}X\big(\gamma\big)}{\mathrm{d}\gamma} = \frac{1}{2\sqrt{\gamma^3}} s_\gamma^\star\Big(X\big(\gamma\big)\Big).$$

This taken together with $\overline{\alpha}_t = \overline{\alpha}_{t-1}\alpha_t$ (cf. (4)) immediately implies that

$$\begin{aligned} \frac{1}{\sqrt{\overline{\alpha}_{t-1}}} X(\overline{\alpha}_{t-1}) &= \frac{1}{\sqrt{\overline{\alpha}_t}} X(\overline{\alpha}_t) + \frac{1}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} s_\gamma^\star\big(X(\gamma)\big)\mathrm{d}\gamma, \\ \Longrightarrow \quad X(\overline{\alpha}_{t-1}) &= \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} s_\gamma^\star\big(X(\gamma)\big)\mathrm{d}\gamma. \end{aligned} \tag{20}$$

With this relation in mind, we are ready to discuss the following approximation in discrete time:

- *Scheme 1:* If we approximate $s_\gamma^\star\big(X(\gamma)\big)$ for $\gamma \in [\overline{\alpha}_t, \overline{\alpha}_{t-1}]$ by $s_\gamma^\star\big(X(\gamma)\big) \approx s^\star_{\overline{\alpha}_t}\big(X(\overline{\alpha}_t)\big) \approx s_t(X_t)$, then we arrive at

$$\begin{aligned} X(\overline{\alpha}_{t-1}) &\approx \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \left(\frac{\sqrt{\overline{\alpha}_{t-1}}}{\sqrt{\overline{\alpha}_t}} - 1\right) s_t(X_t) \\ &\approx \frac{1}{\sqrt{\alpha_t}}\left\{X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t)\right\}, \end{aligned}$$

  where we use the facts that $\overline{\alpha}_t/\overline{\alpha}_{t-1} = \alpha_t$ and $\alpha_t \approx 1$. This coincides with the deterministic sampler (8).

- *Scheme 2:* If we invoke a more refined approximation for $s_\gamma^\star\big(X(\gamma)\big)$ as

$$\begin{aligned} s_\gamma^\star\big(X(\gamma)\big) &\approx s^\star_{\overline{\alpha}_t}\big(X(\overline{\alpha}_t)\big) + \frac{\mathrm{d}s_\gamma^\star\big(X(\gamma)\big)}{\mathrm{d}\gamma}(\gamma - \overline{\alpha}_t) \qquad (21) \\ &\approx s^\star_{\overline{\alpha}_t}(X(\overline{\alpha}_t)) + \frac{\gamma - \overline{\alpha}_t}{\overline{\alpha}_t - \overline{\alpha}_{t+1}}\Big(s^\star_{\overline{\alpha}_t}(X(\overline{\alpha}_t)) - s^\star_{\overline{\alpha}_{t+1}}(X(\overline{\alpha}_{t+1}))\Big) \\ &\approx s_t(X_t) + \frac{\gamma - \overline{\alpha}_t}{\overline{\alpha}_t - \overline{\alpha}_{t+1}}\big(s_t(X_t) - s_{t+1}(X_{t+1})\big), \qquad (22) \end{aligned}$$

  then (20) can be approximated by

  $X(\overline{\alpha}_{t-1})$

$$\approx \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}} s_t(X_t)}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \mathrm{d}\gamma + \frac{\sqrt{\overline{\alpha}_{t-1}} \big(s_t(X_t) - s_{t+1}(X_{t+1})\big)}{2(\overline{\alpha}_t - \overline{\alpha}_{t+1})} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{\gamma - \overline{\alpha}_t}{\sqrt{\overline{\alpha}^3}} \mathrm{d}\gamma$$

$$\approx \frac{1}{\sqrt{\alpha_t}} \left\{ X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})} \big(s_t(X_t) - \sqrt{\alpha_{t+1}} s_{t+1}(X_{t+1})\big) \right\}, \tag{23}$$

which resembles the proposed sampler (15), and is computationally more appealing since it reuses the previous score function evaluation.

It is worth noting that similar approximation as in Scheme 2 has been invoked previously in Lu et al. (2022a, Eqn (3.6)) to construct high-order ODE solvers (e.g., the DPM-Solver-2, with 2 indicating second-order ODEs). Consequently, the acceleration achieved by our sampler is achieved through ideas akin to the second-order ODE; in turn, our convergence guarantees shed light on the effectiveness of high-order ODE solvers like the popular DPM-Solver.

### 3.2 Accelerated SDE-based sampler

Next, we turn to stochastic samplers, and propose a new stochastic sampling procedure that enjoys improved convergence guarantees compared to the DDPM-type sampler (10). To be precise, the proposed sampler begins by drawing $Y_T \sim \mathcal{N}(0, I_d)$ and adopts the following update rule:

$$Y_t^+ = \Phi_t(Y_t, Z_t), \quad Y_{t-1} = \Psi_t(Y_t^+, Z_t^+) \tag{24a}$$

for $t = T, \ldots, 1$, where $Z_t, Z_t^+ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and

$$\Phi_t(x, z) = x + \sqrt{\frac{1-\alpha_t}{2}} z, \tag{24b}$$

$$\Psi_t(y, z) = \frac{1}{\sqrt{\alpha_t}} \left( y + (1-\alpha_t) s_t(y) + \sqrt{\frac{1-\alpha_t}{2}} z \right). \tag{24c}$$

The key difference between the proposed sampler and the original DDPM-type sampler lies in the additional operation $\Phi_t(\cdot, \cdot)$. In this step, a random noise $Z_t$ is injected into the current sample $Y_t$ to obtain an intermediate point $Y_t^+$, which together with another random noise $Z_t^+$ is subsequently fed into $\Psi_t(\cdot, \cdot)$ — a mapping identical to (10).

**Theoretical guarantees.** Let us present the convergence guarantees of the proposed stochastic sampler and their implications, followed by some interpretation of the design rationale of the algorithm.

**Theorem 2.** *Suppose that Assumptions 1 and 2 hold. Then the proposed stochastic sampler (24) with the learning rate schedule (13b) achieves*

$$\mathsf{TV}(q_1, p_1) \le \sqrt{\frac{1}{2} \mathsf{KL}(q_1 \parallel p_1)} \le C_1 \frac{d^3 \log^{4.5} T}{T} + C_1 \sqrt{d} \varepsilon_{\mathsf{score}} \log^{1.5} T \tag{25}$$

*for some universal constant $C_1 > 0$.*

Theorem 2 provides non-asymptotic characterizations for the data generation quality of the accelerated stochastic sampler. In comparison with the convergence theory for the DDPM-type sampler — which has a convergence rate proportional to $1/\sqrt{T}$ (Chen et al., 2022, 2023a; Li et al., 2023; Benton et al., 2023a) — Theorem 2 asserts that the proposed accelerated sampler achieves a faster convergence rate proportional to $1/T$. In contrast to Theorem 1 for the ODE-based sampler, the SDE-based sampler does not require continuity of the Jacobian matrix (i.e., Assumption 3). As before, the total-variation distance between $X_1$ and $Y_1$ is proportional to the $\ell_2$ score estimation error when $T$ is sufficiently large, which covers a broad range of target data distributions with no requirement on the smoothness or log-concavity of the data distribution.

**Interpretation via higher-order approximation.** Now we provide some insights into the motivation of the proposed sampler. We start with the characterizations of conditional density $p_{X_{t-1}|X_t}$. Denoting $\mu_t^\star(x_t) := \frac{1}{\sqrt{\alpha_t}}\left(x_t + (1-\alpha_t)\, s_t^\star(x_t)\right)$ and $J_t(x_t) = -(1-\overline{\alpha}_t) J_{s_t^*}(x_t)$, we can approximate $p_{X_{t-1}|X_t}$ by

$$p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) \approx \exp\Big(-\frac{\alpha_t}{2(1-\alpha_t)}\cdot\Big\|\Big(I-\frac{1-\alpha_t}{2(\alpha_t-\overline{\alpha}_t)}J_t(x_t)\Big)^{-1}(x_{t-1}-\mu_t^\star(x_t))\Big\|^2\Big). \tag{26}$$

which is tighter than the one used in analysis of the original SDE-based sampler (Li et al., 2023) by adopting a higher-order expansion. This in turn motivates us to consider the following sequence

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\underbrace{Y_t + \sqrt{\frac{1-\alpha_t}{2}}Z_t}_{\Phi(Y_t,Z_t)} + \sqrt{\frac{1-\alpha_t}{2}}Z_t^+ + \underbrace{(1-\alpha_t)s_t^\star(Y_t) - \frac{(1-\alpha_t)^{3/2}}{\sqrt{2}(\alpha_t-\overline{\alpha}_t)}J_t(Y_t)Z_t}_{\approx s_t^*\left(\Phi(Y_t,Z_t)\right)}\Big)$$

with $Z_t, Z_t^+ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and $p_{Y_{t-1}|Y_t}(x_{t-1}\mid x_t)$ follows

$$\mathcal{N}\left(\mu_t^\star(x_t), \frac{1-\alpha_t}{\alpha_t}\left(I-\frac{1-\alpha_t}{2\,(1-\bar{\alpha}_t)}J_t\,(x_t)\right)\left(I-\frac{1-\alpha_t}{2\,(1-\bar{\alpha}_t)}J_t\,(x_t)\right)^\top\right)$$

which aligns with (26). In addition, if we further utilize $(1-\alpha_t)s_t^\star(Y_t) - \frac{(1-\alpha_t)^{3/2}}{\sqrt{2}(\alpha_t-\overline{\alpha}_t)}J_t(Y_t)Z_t$ as a first-order approximation of $s_t^\star\big(Y_t + \sqrt{\frac{1-\alpha_t}{2}}Z_t\big)$, we can then arrive at the update rule of the proposed sampler in (24).

# 4 Experiments

In this section, we illustrate the performance of the proposed accelerated samplers, focusing on highlighting the relative comparisons with respect to the original DDIM/DDPM ones using the same pre-trained score functions. As an initial step, we focus on reporting result for the deterministic samplers, leaving the stochastic samplers to future work.

## 4.1 Practical implementation

In practice, the pre-trained score functions are often available in the form of noise-prediction networks $\epsilon_t(\cdot)$, which are connected via the following relationship in view of (7):

$$s_t^\star(X) := -\frac{1}{\sqrt{1-\overline{\alpha}_t}}\epsilon_t^\star(X), \tag{27}$$

and $\epsilon_t(\cdot)$ is the estimate of $\epsilon_t^\star(\cdot)$. To better align with the empirical practice, it is judicious that the integration in (20) be approximated in terms of $\epsilon_t^\star(X)$, leading to an equivalent rewrite as

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}}X(\overline{\alpha}_t) - \frac{\sqrt{\overline{\alpha}_{t-1}}}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\frac{1}{\sqrt{\gamma^3}\sqrt{1-\gamma}}\epsilon_\gamma^\star\big(X(\gamma)\big)\mathrm{d}\gamma.$$

Following similar discussions in Section 3.1, we discuss its first-order and second-order approximations in discrete time.

- *Scheme 1:* If we approximate $\epsilon_\gamma^\star\big(X(\gamma)\big)$ for $\gamma \in [\overline{\alpha}_t, \overline{\alpha}_{t-1}]$ by $\epsilon_\gamma^\star\big(X(\gamma)\big) \approx \epsilon_{\overline{\alpha}_t}^\star\big(X(\overline{\alpha}_t)\big) \approx \epsilon_t(X_t)$, then we arrive at

$$X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}}X(\overline{\alpha}_t) + \left(\sqrt{1-\overline{\alpha}_{t-1}} - \frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{\alpha_t}}\right)\epsilon_t(X_t), \tag{28}$$

  which matches exactly with the DDIM sampler in Song et al. (2020).

- *Scheme 2:* If we invoke the refined approximation (22) in terms of $\epsilon^\star_\gamma\big(X(\gamma)\big)$, we have

$$X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) - \frac{\sqrt{\overline{\alpha}_{t-1}}\epsilon_t\big(X_t\big)}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3(1-\gamma)}} \mathrm{d}\gamma \\ - \frac{\sqrt{\overline{\alpha}_{t-1}}\left(\epsilon_t\big(X_t\big) - \epsilon_{t+1}\big(X_{t+1}\big)\right)}{2(\overline{\alpha}_t - \overline{\alpha}_{t+1})} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{(\gamma - \overline{\alpha}_t)}{\sqrt{\gamma^3(1-\gamma)}} \mathrm{d}\gamma,$$

which after integration becomes:

$$X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \left(\sqrt{1-\overline{\alpha}_{t-1}} - \frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{\alpha}_t}\right) \epsilon_t(X_t) \\ + \left(\frac{\sqrt{\overline{\alpha}_{t-1}}}{\overline{\alpha}_t - \overline{\alpha}_{t+1}}\right) \left(\overline{\alpha}_t \frac{\sqrt{1-\overline{\alpha}_{t-1}}}{\sqrt{\overline{\alpha}_{t-1}}} + \arcsin\sqrt{\overline{\alpha}_{t-1}} - \overline{\alpha}_t \frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{\overline{\alpha}_t}} - \arcsin\sqrt{\overline{\alpha}_t}\right) (\epsilon_{t+1}(X_{t+1}) - \epsilon_t(X_t)). \tag{29}$$

This is our new sampler for implementation.

### 4.2 Experimental results

We use pre-trained score functions from Huggingface (von Platen et al., 2022) for three datasets: CelebA-HQ, LSUN-Bedroom and LSUN-Churches. The same score functions are used in all the samplers. Note that we have not attempted to optimize the speed nor the performance using additional tricks, e.g., employing better score functions, but aim to corroborate our theoretical findings regarding the acceleration of the new samplers without training additional functions when the implementations are otherwise kept the same.

We first compare the vanilla DDIM-type sampler (cf. (28)) and the accelerated DDIM-type sampler (cf. (29)). To begin, Figure 1 illustrates the progress of the generated samples over different numbers of function evaluation (NFEs) (between 4 and 50) from the same random seed, using pre-trained scores from the LSUN-Churches dataset. Here, the NFE is the same as the number of diffusion steps since each step takes one score evaluation.


Figure 1: The progress of the generated samples over different numbers of NFEs (from 4 to 50), using pre-trained scores from the LSUN-Churches dataset. Top row: the vanilla DDIM-type sampler. Bottom row: the accelerated DDIM-type sampler (ours).

To further demonstrate the quality of the sampled images, Figure 2 provides examples of sampled images from the DDIM-type samplers, using pre-trained scores from CelebA-HQ, LSUN-Bedroom and LSUN-Churches datasets, respectively. It can be seen that the sampled images are crisper and less noisy from the accelerated DDIM-type sampler, compared with from the original one, indicating the effectiveness of our method.

## 5 Discussion

In this paper, we have developed novel strategies to achieve provable acceleration in score-based generative modeling. The proposed deterministic sampler achieves a convergence rate $1/T^2$ that substantially improves

(a) LSUN-Churches (b) LSUN-Bedroom (c) CelebA-HQ

Figure 2: Examples of sampled images from the DDIM-type samplers with 5 NFEs, using pre-trained scores from the LSUN-Churches, LSUN-Bedroom, and CelebA-HQ datasets. For each dataset, the top image is the original DDIM-type sampler, and the bottom image is the accelerated DDIM-type sampler (ours).

upon prior theory for the probability flow ODE approach, whereas the proposed stochastic sampler enjoys a converge rate $1/T$ that also significantly outperforms the convergence theory for the DDPM-type sampler. We have demonstrated the stability of these samplers, establishing non-asymptotic theoretical guarantees that hold in the presence of $\ell_2$-accurate score estimates. Our algorithm development for the deterministic case draws inspiration from higher-order ODE approximations in discrete time, which might shed light on understanding popular ODE-based samplers like the DPM-Solver. In comparison, the accelerated stochastic sampler is designed based on higher-order expansions of the conditional density.

Our findings further suggest multiple directions that are worthy of future exploration. For instance, our convergence theory remains sub-optimal in terms of the dependency on the problem dimension $d$, which calls for a more refined theory to sharpen dimension dependency. Additionally, given the conceptual similarity between our accelerated deterministic sampler and second-order ODE, it would be interesting to extend the algorithm and theory using ideas arising from third-order or even higher-order ODE. In particular, third-order ODE has been implemented in DPM-Solver-3, which is among the most effective DPM-Solvers in practice. Finally, it would be important to design higher-order solvers for SDE-based samplers, in order to unveil the degree of acceleration that can be achieved through high-order SDE.

## Acknowledgements

Y. Wei is supported in part by the NSF grants DMS-2147546/2015447, CAREER award DMS-2143215, CCF-2106778, and the Google Research Scholar Award. The work of T. Efimov and Y. Chi is supported in part by the grants ONR N00014-19-1-2404, NSF DMS-2134080, ECCS-2126634 and FHWA 693JJ321C000013. 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.

## References

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

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. (2023a). Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*.

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

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

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

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

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2023b). The probability flow ODE is provably fast. *Neural Information Processing Systems*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2022). 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. (2023c). Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. In *International Conference on Machine Learning*, pages 4462–4484.

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

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

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

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Gao, X., Nguyen, H. M., and Zhu, L. (2023). Wasserstein convergence guarantees for a general class of score-based generative models. *arXiv preprint arXiv:2311.11003*.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Guo, Z., Liu, J., Wang, Y., Chen, M., Wang, D., Xu, D., and Cheng, J. (2023). Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering*, pages 1–19.

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.

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.

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. *International Conference on Learning Representations*.

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

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, S., Chen, S., and Li, Q. (2024b). A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*.

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, L., Ren, Y., Lin, Z., and Zhao, Z. (2022a). Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.

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

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022a). DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022b). DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Luhman, E. and Luhman, T. (2021). Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*.

Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., and Salimans, T. (2023). On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306.

Pang, T., Xu, K., Li, C., Song, Y., Ermon, S., and Zhu, J. (2020). Efficient learning of generative models via finite-difference score matching. *Advances in Neural Information Processing Systems*, 33:19175–19188.

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

Salimans, T. and Ho, J. (2021). Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*.

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. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency models. In *International Conference on Machine Learning*.

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). 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*.

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

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. (2022). Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Wang, Z., Zheng, H., He, P., Chen, W., and Zhou, M. (2022). Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*.

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

Wu, Y., Chen, M., Li, Z., Wang, M., and Wei, Y. (2024). Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *arXiv preprint arXiv:arXiv:2403.01639*.

Xiao, Z., Kreis, K., and Vahdat, A. (2021). Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*.

Xue, S., Yi, M., Luo, W., Zhang, S., Sun, J., Li, Z., and Ma, Z.-M. (2023). SA-Solver: Stochastic Adams solver for fast sampling of diffusion models. *arXiv preprint arXiv:2309.05019*.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39.

Zhang, Q. and Chen, Y. (2022). Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*.

Zhang, Q., Tao, M., and Chen, Y. (2022). gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*.

Zhao, W., Bai, L., Rao, Y., Zhou, J., and Lu, J. (2023). UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. *arXiv preprint arXiv:2302.04867*.

# A Preliminaries

Before delving into the proof, we make note of a couple of preliminary facts, primarily from Li et al. (2023).

## A.1 Basic facts

**Score functions.** We first give some characterizations of the score function, which follow from Li et al. (2023, properties (38)).

**Lemma 1.** *The true score function $s_t^\star$ is given by the conditional expectation below:*

$$s_t^\star(x) = \mathbb{E}\left[-\frac{1}{\sqrt{1-\overline{\alpha}_t}}W \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right] = \frac{1}{1-\overline{\alpha}_t}\mathbb{E}\left[\sqrt{\overline{\alpha}_t}X_0 - x \,\middle|\, \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}W = x\right]$$
$$= -\frac{1}{1-\overline{\alpha}_t}\underbrace{\int_{x_0}(x - \sqrt{\overline{\alpha}_t}x_0)p_{X_0|X_t}(x_0\,|\,x)\mathrm{d}x_0}_{=:\,g_t(x)}. \tag{30}$$

*Also, the Jacobian matrix*

$$J_t(x) \coloneqq \frac{\partial g_t(x)}{\partial x} \tag{31}$$

*associated with the function $g_t(\cdot)$ (defined in (30)) satisfies*

$$J_t(x) = I_d + \frac{1}{1-\overline{\alpha}_t}\Big\{\mathbb{E}\big[X_t - \sqrt{\overline{\alpha}_t}X_0 \mid X_t = x\big]\Big(\mathbb{E}\big[X_t - \sqrt{\overline{\alpha}_t}X_0 \mid X_t = x\big]\Big)^\top - \mathbb{E}\Big[\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)^\top \mid X_t = x\Big]\Big\}. \tag{32}$$

**Learning rates.** The learning rates $\{\alpha_t\}$ as specified in (13b) enjoy several properties that will be used multiple times in the analysis. We record several of these properties below, which have been proven in Li et al. (2023, properties (39)).

**Lemma 2.** *The learning rates specified in (13b) obey*

$$\alpha_t \geq 1 - \frac{c_1 \log T}{T} \geq \frac{1}{2}, \qquad 1 \leq t \leq T \tag{33a}$$

$$\frac{1}{2}\frac{1-\alpha_t}{1-\overline{\alpha}_t} \leq \frac{1}{2}\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t} \leq \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \leq \frac{4c_1 \log T}{T}, \qquad 2 \leq t \leq T \tag{33b}$$

$$1 \leq \frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} \leq 1 + \frac{4c_1 \log T}{T}, \qquad 2 \leq t \leq T \tag{33c}$$

$$\overline{\alpha}_T \leq \frac{1}{T^{c_2}}, \tag{33d}$$

*provided that $T$ is large enough. Here, $c_1$ is defined in (13b), and $c_2 \geq 1000$ is some large numerical constant. In addition, if $\frac{d(1-\alpha_t)}{\alpha_t - \overline{\alpha}_t} \lesssim 1$, then one has*

$$\Big(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\Big)^{d/2} = 1 + \frac{d(1-\alpha_t)}{2(\alpha_t - \overline{\alpha}_t)} + \frac{d(d-2)(1-\alpha_t)^2}{8(\alpha_t - \overline{\alpha}_t)^2} + O\Big(d^3\Big(\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t}\Big)^3\Big). \tag{33e}$$

**The forward process.** Next, we gather several conditional tail bounds for the random vector $X_0$ of the forward process, which have been established in Li et al. (2023, Lemmas 1 and 2).

**Lemma 3.** *Suppose that there exists some numerical constant $c_R > 0$ obeying*

$$\mathbb{P}\big(\|X_0\|_2 \leq R\big) = 1 \qquad \text{and} \qquad R = T^{c_R}. \tag{34}$$

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

$$\theta(y) \coloneqq \max\Big\{\frac{-\log p_{X_t}(y)}{d \log T}, c_6\Big\} \tag{35}$$

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

$$\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2 \leq 5c_5\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T} \tag{36}$$

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

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2 \,\middle|\, X_t = y\right] \leq 12\sqrt{\theta(y)d(1-\overline{\alpha}_t)\log T}, \tag{37a}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^2 \,\middle|\, X_t = y\right] \leq 120\theta(y)d(1-\overline{\alpha}_t)\log T, \tag{37b}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^3 \,\middle|\, X_t = y\right] \leq 1040\big(\theta(y)d(1-\overline{\alpha}_t)\log T\big)^{3/2}, \tag{37c}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - y\big\|_2^4 \,\middle|\, X_t = y\right] \leq 10080\big(\theta(y)d(1-\overline{\alpha}_t)\log T\big)^2. \tag{37d}$$

**Lemma 4.** *For some $\theta > c_6$, assume that $\|y - x\|_2 \lesssim (1-\alpha_{t+1})\sqrt{\frac{\theta d \log T}{1-\overline{\alpha}_t}}$ and $\log p_{X_t}(x) \geq -\theta d \log T$. Then we have*

$$p_{X_0 \mid X_{t+1}}(x_0 \mid \sqrt{\alpha_{t+1}}y) = p_{X_0 \mid X_t}(x_0 \mid x)\Big\{1 + \frac{(1-\alpha_{t+1})\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)^2} - \frac{(x - \sqrt{\overline{\alpha}_t}x_0)^\top(y-x)}{1-\overline{\alpha}_t} -$$

$$\int_{x_0}\Bigg(\frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)^2}-\frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(y-x)}{1-\overline{\alpha}_t}\Bigg)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0+O\Big(\theta d\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}\Big)^{3/2}\log T\Big)\Bigg\}. \tag{38}$$

*Proof.* See Section A.2. □

**Proximity of $p_{X_T}$ and $q_{Y_T}$.** When the number of steps $T$ is sufficiently large, the distribution of $p_{X_T}$ and that of $p_{Y_T}$ become exceedingly close, as asserted by the following lemma (see Li et al. (2023, Lemma 3)).

**Lemma 5.** *For any large enough $T$, one has*

$$\mathsf{TV}(p_{X_T}\,\|\,p_{Y_T})^2\le\frac{1}{2}\mathsf{KL}(p_{X_T}\,\|\,p_{Y_T})\lesssim\frac{1}{T^{200}}. \tag{39}$$

**Score estimation errors.** Consider any vector $x\in\mathbb{R}^d$. For any $1<t\le T$, define

$$\varepsilon_{\mathsf{score},t}(x):=\big\|s_t(x)-s_t^\star(x)\big\|_2\qquad\text{and}\qquad\varepsilon_{\mathsf{Jacobi},t}(x):=\big\|J_{s_t}(x)-J_{s_t^\star}(x)\big\|, \tag{40}$$

where we use $J_{s_t}$ and $J_{s_t^\star}$ to represent the Jacobian matrices of $s_t(\cdot)$ and $s_t^\star(\cdot)$, respectively. We also have, under Assumptions 2 and 3, that

$$\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{X\sim q_t}\big[\varepsilon_{\mathsf{score},t}(X)\big]\le\bigg(\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{X\sim q_t}\big[\varepsilon_{\mathsf{score},t}(X)^2\big]\bigg)^{1/2}\le\varepsilon_{\mathsf{score}}, \tag{41a}$$

$$\frac{1}{T}\sum_{t=1}^T\mathop{\mathbb{E}}_{X\sim q_t}\big[\varepsilon_{\mathsf{Jacobi},t}(X)\big]\le\varepsilon_{\mathsf{Jacobi}}. \tag{41b}$$

## A.2 Proof of Lemma 4

To establish this lemma, we observe that

$$\begin{aligned}
p_{X_0\,|\,X_{t+1}}(x_0\,|\,\sqrt{\alpha_{t+1}}y)&=\frac{p_{X_0}(x_0)p_{X_{t+1}\,|\,X_0}(y\,|\,x_0)}{\int_{x_0}p_{X_0}(x_0)p_{X_{t+1}\,|\,X_0}(y\,|\,x_0)\mathrm{d}x_0}\\
&=\frac{p_{X_0}(x_0)\exp\Big(-\frac{\|y-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}\Big)}{\int_{x_0}p_{X_0}(x_0)\exp\Big(-\frac{\|y-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}\Big)\mathrm{d}x_0}\\
&=\frac{p_{X_0}(x_0)\exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\Big)\Big(1-\frac{(1-\alpha_{t+1}^{-1})\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)^2}-\frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(y-x)}{1-\overline{\alpha}_t}+O\Big(\theta d\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}\Big)^{3/2}\log T\Big)\Big)}{\int_{x_0}p_{X_0}(x_0)\exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\Big)\Big(1-\frac{(1-\alpha_{t+1}^{-1})\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)^2}-\frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(y-x)}{1-\overline{\alpha}_t}+O\Big(\theta d\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}\Big)^{3/2}\log T\Big)\Big)\mathrm{d}x_0}\\
&=p_{X_0\,|\,X_t}(x_0\,|\,x)\Bigg\{1+\frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)^2}-\frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(y-x)}{1-\overline{\alpha}_t}\\
&\quad-\int_{x_0}\Bigg(\frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_t)^2}-\frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(y-x)}{1-\overline{\alpha}_t}\Bigg)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\\
&\quad+O\Big(\theta d\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}\Big)^{3/2}\log T\Big)\Bigg\},
\end{aligned}$$

which follows from the following property:

$$\begin{aligned}
&\frac{\|y-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\\
&\quad=\frac{(1-\alpha_{t+1}^{-1})\|y-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}+\frac{\|y-\sqrt{\overline{\alpha}_t}x_0\|_2^2-\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)}\\
&\quad=-\frac{(1-\alpha_{t+1})\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)^2}+\frac{2(x-\sqrt{\overline{\alpha}_t}x_0)^\top(y-x)}{2(1-\overline{\alpha}_t)}+O\Big(\theta d\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}\Big)^{3/2}\log T\Big).
\end{aligned}$$

# B Analysis for the accelerated ODE sampler (proof of Theorem 1)

In this section, we present our non-asymptotic analysis for the accelerated ODE sampler. Considering the total variation distance is always upper bounded by 1, we can reasonably assume the following conditions throughout the proof, which are necessary for the claimed result eq. (16) to be non-trivial.

$$T \geq \sqrt{C_1} d^3 \log^3 T, \tag{42a}$$

$$\varepsilon_{\text{score}} \leq \frac{1}{C_1 \sqrt{d} \log^2 T}, \tag{42b}$$

$$\varepsilon_{\text{Jacobi}} \leq \frac{1}{C_1 d \log^2 T}. \tag{42c}$$

## B.1 Main steps of the proof

**Preparation.** To begin with, let us introduce the following functions that help ease presentation:

$$\phi_t^\star(x) \coloneqq x - \frac{1-\alpha_{t+1}}{2} s_t^\star(x), \tag{43a}$$

$$\phi_t(x) \coloneqq x - \frac{1-\alpha_{t+1}}{2} s_t(x), \tag{43b}$$

$$\psi_t^\star(x) \coloneqq x + \frac{1-\alpha_t}{2} s_t^\star(x) + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})} \Big( s_t^\star(x) - \sqrt{\alpha_{t+1}} s_{t+1}^\star\big(\phi_t^\star(x)\big) \Big), \tag{43c}$$

$$\psi_t(x) \coloneqq x + \frac{1-\alpha_t}{2} s_t(x) + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})} \Big( s_t(x) - \sqrt{\alpha_{t+1}} s_{t+1}\big(\phi_t(x)\big) \Big). \tag{43d}$$

Armed with these functions, one can equivalently rewrite our update rule (15) as follows:

$$Y_{t-1} = \Psi_t\big(Y_t, \Phi_t(Y_t)\big) = \frac{1}{\sqrt{\alpha_t}} \psi_t\left(Y_t\right). \tag{43e}$$

Additionally, for any point $y_T \in \mathbb{R}^d$, we introduce the corresponding sequence

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \psi_t\left(y_t\right), \quad y_t^- = \sqrt{\alpha_{t+1}} \phi_t(y_t), \quad t = T, T-1, \cdots \tag{44}$$

Furthermore, it is also useful to single out the following error-related quantities for any point $y_T \in \mathbb{R}^d$ and its associated sequence $\{y_t\}_{t=1}^{T-1}$ and $\{y_t^-\}_{t=2}^{T}$:

$$\xi_t(y_T) \coloneqq \frac{\log T}{T} \Big\{ d\big(\varepsilon_{\mathsf{Jacobi},t}(y_t) + \varepsilon_{\mathsf{Jacobi},t+1}(y_t^-)\big) + \sqrt{d \log T}\big(\varepsilon_{\mathsf{score},t}(y_t) + \varepsilon_{\mathsf{score},t+1}(y_t^-)\big) \Big\}; \tag{45a}$$

$$S_t(y_T) \coloneqq \sum_{1<k\leq t} \xi_k(y_k), \quad \text{for } t \geq 2, \qquad \text{and} \qquad S_1(y_T) = 0, \tag{45b}$$

where we recall the definitions of $\varepsilon_{\mathsf{Jacobi},t}(\cdot)$ and $\varepsilon_{\mathsf{score},t}$ in (40). To understand these quantities, note that if we start from a point $y_T$, then $\xi_t(y_t)$ reflects a certain weighted score estimation error in the $t$-th step, while $S_t(y_T)$ aggregates these weighted score estimation errors from the very beginning to the $t$-th iteration.

In addition, there are several objects that play crucial roles in the subsequent analysis, which we single out as follows; here and throughout, we suppress their dependency on $x$ to streamline presentation.

$$A_t \coloneqq \frac{1}{1-\overline{\alpha}_t} \int p_{X_0 \,|\, X_t}(x_0 \,|\, x) \big\| x - \sqrt{\overline{\alpha}_t} x_0 \big\|_2^2 \mathrm{d}x_0; \tag{46a}$$

$$B_t \coloneqq \frac{1}{1-\overline{\alpha}_t} \left\| \int p_{X_0 \,|\, X_t}(x_0 \,|\, x) \big(x - \sqrt{\overline{\alpha}_t} x_0\big) \mathrm{d}x_0 \right\|_2^2; \tag{46b}$$

$$C_t \coloneqq \frac{1}{(1-\overline{\alpha}_t)^2} \int p_{X_0 \,|\, X_t}(x_0 \,|\, x) \big\| x - \sqrt{\overline{\alpha}_t} x_0 \big\|_2^4 \mathrm{d}x_0; \tag{46c}$$

$$D_t \coloneqq \frac{1}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\Big(\big\langle g_t(x),\, x-\sqrt{\overline{\alpha}_t}x_0\big\rangle\Big)^2\mathrm{d}x_0; \tag{46d}$$

$$E_t \coloneqq \frac{1}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big\langle g_t(x),\, x-\sqrt{\overline{\alpha}_t}x_0\big\rangle\mathrm{d}x_0. \tag{46e}$$

With the above preparation in place, we can now readily proceed to our proof.

**Step 1: bounding the density ratios.** To begin with, we make the observation that

$$\begin{aligned}\frac{p_{Y_{t-1}}\left(y_{t-1}\right)}{p_{X_{t-1}}\left(y_{t-1}\right)} &= \frac{p_{\sqrt{\alpha_t}Y_{t-1}}\left(\sqrt{\alpha_t}y_{t-1}\right)}{p_{\sqrt{\alpha_t}X_{t-1}}\left(\sqrt{\alpha_t}y_{t-1}\right)}\\ &= \frac{p_{\psi_t(Y_t)}\left(\psi_t(y_t)\right)}{p_{Y_t}\left(y_t\right)}\cdot\left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}\left(\psi_t(y_t)\right)}{p_{X_t}\left(y_t\right)}\right)^{-1}\cdot\frac{p_{Y_t}\left(y_t\right)}{p_{X_t}\left(y_t\right)},\end{aligned} \tag{47}$$

which is an elementary identity that allows one to link the target density ratio $\frac{p_{Y_{t-1}}}{p_{X_{t-1}}}$ at the $(t-1)$-th step with the density ratio $\frac{p_{Y_t}}{p_{X_t}}$ at the $t$-th step. This relation reveals the importance of bounding $\frac{p_{\psi_t(Y_t)}(\psi_t(y_t))}{p_{Y_t}(y_t)}$ and $\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\psi_t(y_t))}{p_{X_t}(y_t)}$, towards which we resort to the following lemma.

**Lemma 6.** *For any $x\in\mathbb{R}^d$, suppose that*

$$-\frac{\log p_{X_t}(x)}{d\log T}\le c_6 \tag{48}$$

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

$$\frac{p_{X_{t+1}/\sqrt{\alpha_{t+1}}}\big(\phi_t(x)\big)}{p_{X_t}(x)}\ge\exp\bigg(-\Big(\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log T}+d\log T\Big)\frac{c_7\log T}{T}\bigg) \tag{49}$$

*for some universal constant $c_7>0$. If, in addition, we have $C_{10}\frac{d\log^2 T+(\varepsilon_{\mathsf{score},t}(x)+\varepsilon_{\mathsf{score},t+1}(\Phi_t(x)))\sqrt{d\log^3 T}}{T}\le 1$ for some large enough constant $C_{10}>0$, then it holds that*

$$\begin{aligned}\frac{p_{\sqrt{\alpha_t}X_{t-1}}(\psi_t(x))}{p_{X_t}(x)} &= 1+\frac{(1-\alpha_t)(d+B_t-A_t)}{2(1-\overline{\alpha}_t)}\\ &\quad+\frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[d(d+2)+(4+2d)(B_t-A_t)-B_t^2+C_t+2D_t-3E_t+A_tB_t\big]\\ &\quad+O\bigg(\frac{d^3\log^6 T}{T^3}+\frac{\sqrt{d\log^3 T}}{T}\Big(\varepsilon_{\mathsf{score},t}(x)+\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)\Big)\bigg).\end{aligned} \tag{50a}$$

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

$$\begin{aligned}\frac{p_{\psi_t(Y)}(\psi_t(x))}{p_Y(x)} &= 1+\frac{(1-\alpha_t)(d+B_t-A_t)}{2(1-\overline{\alpha}_t)}\\ &\quad+\frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[d(d+2)+(4+2d)(B_t-A_t)-B_t^2+C_t+2D_t-3E_t+A_tB_t\big]\\ &\quad+O\bigg(\frac{d^6\log^6 T}{T^3}+\frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t}(x)+\frac{d\log T}{T}\Big(\varepsilon_{\mathsf{Jacobi},t}(x)+\varepsilon_{\mathsf{Jacobi},t+1}\big(\Phi_t(x)\big)\Big)\bigg),\end{aligned} \tag{50b}$$

*provided that $C_{11}\frac{d^2\log^2 T+\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log^3 T}+d(\varepsilon_{\mathsf{Jacobi},t}(x)+\varepsilon_{\mathsf{Jacobi},t+1}(\Phi_t(x)))\log T}{T}\le 1$ holds for some large enough constant $C_{11}>0$.*

*Additionally, if $\frac{d^2\log^2 T+\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(x)+d\varepsilon_{\mathsf{Jacobi},t}(x)\log T}{T}\lesssim 1$, then we have*

$$\frac{p_{\Phi_t(X_t)}(\Phi_t(x))}{p_{X_{t+1}}(\Phi_t(x))}=1+O\bigg(\frac{d^2\log^4 T}{T^2}+\frac{d^6\log^6 T}{T^3}+\frac{\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(x)+d\varepsilon_{\mathsf{Jacobi},t}(x)\log T}{T}\bigg). \tag{51}$$

The proof of this lemma is postponed to Section B.2. Notheworthily, the main terms in (50a) and (50b) coincide, a crucial fact that allows one to focus on the lower-order term later on. Moreover, the relation (51) captures the effect of performing one iteration of the probability flow ODE sampler (as captured by the mapping $\Phi_t(\cdot)$).

**Step 2: decomposing the TV distance of interest.** We now move on to look at the TV distance of interest. Akin to Li et al. (2023), we first single out the following set:

$$\mathcal{E} \coloneqq \Big\{ y : q_1(y) > \max\big\{p_1(y),\, \exp\big(-c_6 d \log T\big)\big\}\Big\}, \tag{52}$$

where $c_6 > 0$ is some large enough numerical constant introduced in Lemma 6. The points in $\mathcal{E}$ satisfy two properties: (i) $q_1(y) > p_1(y)$, and (ii) $q_1(y)$ is not too small, so that $y$ falls within a more normal range (w.r.t. $p_{X_1}(\cdot)$).

Following similar calculations as in Li et al. (2023, Step 2 of Section 5.2) and invoking the properties of the forward process in Lemma 3, we can demonstrate that

$$\mathsf{TV}\big(q_1, p_1\big) \le \mathop{\mathbb{E}}_{Y_1\sim p_1}\left[\Big(\frac{q_1(Y_1)}{p_1(Y_1)} - 1\Big)\,\mathbb{1}\left\{Y_1 \in \mathcal{E}\right\}\right] + \exp\big(-c_6 d\log T\big), \tag{53}$$

and hence it suffices to focus attention on what happens on the set $\mathcal{E}$. To proceed, for any point $y_T$, we define

$$\tau(y_T) \coloneqq \max\Big\{2 \le t \le T+1 : S_{t-1}\big(y_T\big) \le c_{14} \text{ and } -\log q_k(y_k) \le c_\tau d\log T, \text{ for all } k < t\Big\} \tag{54}$$

for some universal constant $c_\tau > 0$. We shall often abbreviate $\tau(y_T)$ as $\tau$ as long as it is clear from the context. Taking $\{y_t\}_{t=1}^{T-1}$ to be the associated sequence of our deterministic sampler initialized at $y_T$ (cf. (44)), we can further define

$$\mathcal{I}_0 \coloneqq \Big\{y_T : \tau(y_T) = T+1\Big\}, \tag{55a}$$

$$\mathcal{I}_1 \coloneqq \Big\{y_T : S_{\tau-1}(y_T) \le c_{14} \text{ and } -\log q_\tau(y_\tau) > c_\tau d\log T\Big\}, \tag{55b}$$

$$\mathcal{I}_2 \coloneqq \Big\{y_T : c_{14} \le S_\tau\big(y_T\big) \le 2c_{14}\Big\} \cap \mathcal{I}_1^{\mathrm{c}}, \tag{55c}$$

$$\mathcal{I}_3 \coloneqq \left\{y_T : S_{\tau-1}\big(y_T\big) \le c_{14}, \xi_\tau\big(y_T\big) \ge c_{14}, \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})} \le \frac{8q_\tau(y_\tau)}{p_\tau(y_\tau)}\right\} \cap \mathcal{I}_1^{\mathrm{c}}, \tag{55d}$$

$$\mathcal{I}_4 \coloneqq \left\{y_T : S_{\tau-1}\big(y_T\big) \le c_{14}, \xi_\tau\big(y_T\big) \ge c_{14}, \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})} > \frac{8q_\tau(y_\tau)}{p_\tau(y_\tau)}\right\} \cap \mathcal{I}_1^{\mathrm{c}}. \tag{55e}$$

As an immediate consequence of the above definitions, one has

$$\mathcal{I}_0 \cup \mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4 = \mathbb{R}^d.$$

In the following, we shall look at each of these sets separately, and combine the respective bounds to control the first term on the right-hand side of (53).

**Step 3: coping with the set $\mathcal{I}_0$.** In order to obtain a useful bound when restricting attention to $\mathcal{I}_0$ (cf. (55a)), we resort to the following key lemma, whose proof is provided in Section B.3.

**Lemma 7.** *Consider any $y_T$, along with the deterministic sequences $\{y_{T-1}, \cdots, y_1\}$ and $\{y_T^-, \cdots, y_2^-\}$. Set $\tau = \tau(y_T)$ (cf. (54)). Then one has*

$$\frac{q_1(y_1)}{p_1(y_1)} = \left\{1 + O\left(\frac{d^6\log^6 T}{T^2} + S_{\tau-1}(y_{\tau-1})\right)\right\}\frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})}, \tag{56a}$$

$$\text{and} \qquad \frac{q_k(y_k)}{2p_k(y_k)} \le \frac{q_1(y_1)}{p_1(y_1)} \le 2\frac{q_k(y_k)}{p_k(y_k)}, \qquad \forall k < \tau. \tag{56b}$$

With this lemma in mind, we are ready to cope with the set $\mathcal{I}_0$. Taking $\tau(y_T) = T+1$ in Lemma 7 yields

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{q_1(Y_1)}{p_1(Y_1)}-1\right)\mathbb{1}\left\{Y_1\in\mathcal{E}, Y_T\in\mathcal{I}_0\right\}\right] \\
&\overset{\text{(i)}}{=} \mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\left\{1+O\left(\frac{d^6\log^6 T}{T^2}+S_T(y_T)\right)\right\}\frac{q_T(Y_T)}{p_T(Y_T)}-1\right)\mathbb{1}\left\{Y_1\in\mathcal{E}, Y_T\in\mathcal{I}_0\right\}\right] \\
&= \int\left\{\left(1+O\left(\frac{d^6\log^6 T}{T^2}+S_T(y_T)\right)\right)q_T(y_T)-p_T(y_T)\right\}\mathbb{1}\left\{y_1\in\mathcal{E}, y_T\in\mathcal{I}_0\right\}\mathrm{d}y_T \\
&\overset{\text{(ii)}}{\le} \int\left|q_T(y_T)-p_T(y_T)\right|\mathrm{d}y_T + O\left(\frac{d^6\log^6 T}{T^2}\right)\int q_T(y_T)\mathrm{d}y_T + O\left(\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+(d\log T)\varepsilon_{\mathsf{Jacobi}}\right) \\
&\overset{\text{(iii)}}{\lesssim} \frac{d^6\log^6 T}{T^2}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+(d\log T)\varepsilon_{\mathsf{Jacobi}}. \qquad (57)
\end{aligned}
$$

Here, (i) invokes Lemma 7, whereas (iii) holds since $\mathsf{TV}(p_T, q_T)\lesssim T^{-100}$ (according to Lemma 6). To see why (ii) is valid, it suffices to make the following observation:

$$
\begin{aligned}
&\int S_T(y_T)q_T(y_T)\,\mathbb{1}\left\{y_1\in\mathcal{E}, y_T\in\mathcal{I}_0\right\}\mathrm{d}y_T = \\
&\quad= \frac{\log T}{T}\sum_{t=1}^{T}\int\left\{d\big(\varepsilon_{\mathsf{Jacobi},t}(y_t)+\varepsilon_{\mathsf{Jacobi},t+1}(y_t^-)\big)+\sqrt{d\log T}\big(\varepsilon_{\mathsf{score},t}(y_t)+\varepsilon_{\mathsf{score},t+1}(y_t^-)\big)\right\} \\
&\qquad\qquad \cdot q_T(y_T)\,\mathbb{1}\left\{y_1\in\mathcal{E}, y_T\in\mathcal{I}_0\right\}\mathrm{d}y_T \\
&\quad\overset{\text{(iv)}}{\le} \frac{4\log T}{T}\sum_{t=1}^{T}\int\left\{d\big(\varepsilon_{\mathsf{Jacobi},t}(y_t)+\varepsilon_{\mathsf{Jacobi},t+1}(y_t^-)\big)+\sqrt{d\log T}\big(\varepsilon_{\mathsf{score},t}(y_t)+\varepsilon_{\mathsf{score},t+1}(y_t^-)\big)\right\} \\
&\qquad\qquad \cdot\frac{q_t(y_t)}{p_t(y_t)}p_T(y_T)\,\mathbb{1}\left\{y_1\in\mathcal{E}, y_T\in\mathcal{I}_0\right\}\mathrm{d}y_T \\
&\quad\le \frac{4\log T}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left\{d\big(\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\varepsilon_{\mathsf{Jacobi},t+1}(Y_t^-)\big)+\sqrt{d\log T}\big(\varepsilon_{\mathsf{score},t}(Y_t)+\varepsilon_{\mathsf{score},t+1}(Y_t^-)\big)\right\}\right. \\
&\qquad\qquad \left.\cdot\frac{q_t(Y_t)}{p_t(Y_t)}\,\mathbb{1}\left\{\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)+d\varepsilon_{\mathsf{Jacobi},t}(Y_t)\log T\lesssim T\right\}\right] \\
&\quad= \frac{4\log T}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{Y_t\sim p_t}\left[\left\{d\big(\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\varepsilon_{\mathsf{Jacobi},t+1}(Y_t^-)\big)+\sqrt{d\log T}\big(\varepsilon_{\mathsf{score},t}(Y_t)+\varepsilon_{\mathsf{score},t+1}(Y_t^-)\big)\right\}\right. \\
&\qquad\qquad \left.\cdot\frac{q_t(Y_t)}{p_t(Y_t)}\,\mathbb{1}\left\{\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)+d\varepsilon_{\mathsf{Jacobi},t}(Y_t)\log T\lesssim T\right\}\right] \\
&\quad= \frac{4\log T}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{Y_t\sim q_t}\left[\left\{d\big(\varepsilon_{\mathsf{Jacobi},t}(Y_t)+\varepsilon_{\mathsf{Jacobi},t+1}(Y_t^-)\big)+\sqrt{d\log T}\big(\varepsilon_{\mathsf{score},t}(Y_t)+\varepsilon_{\mathsf{score},t+1}(Y_t^-)\big)\right\}\right. \\
&\qquad\qquad \left.\cdot\,\mathbb{1}\left\{\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)+d\varepsilon_{\mathsf{Jacobi},t}(Y_t)\log T\lesssim T\right\}\right] \\
&\quad\overset{\text{(v)}}{\lesssim} \frac{\log T}{T}\sum_{t=1}^{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] \\
&\quad\overset{\text{(vi)}}{\lesssim} (d\log T)\varepsilon_{\mathsf{Jacobi}}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}.
\end{aligned}
$$

Here, (iv) follows from Lemma 7, while (vi) comes from (41). To understand why (v) is valid, let us denote the probability density of $\Phi(X_t)$ by $q_t^-$ and, by referring to (51), we derive that

$$
\mathop{\mathbb{E}}_{Y_t\sim q_t}\left[\varepsilon_{\mathsf{score},t+1}(Y_t^-)\,\mathbb{1}\left\{\sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t)+d\varepsilon_{\mathsf{Jacobi},t}(Y_t)\log T\lesssim T\right\}\right]
$$

$$= \mathop{\mathbb{E}}_{Y_t^- \sim q_t^-} \left[ \varepsilon_{\mathsf{score},t+1}(Y_t^-) \mathbb{1}\left\{ \sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t) + d\varepsilon_{\mathsf{Jacobi},t}(Y_t)\log T \lesssim T \right\} \right]$$
$$\lesssim \mathop{\mathbb{E}}_{Y_t^- \sim q_{t+1}} \left[ \varepsilon_{\mathsf{score},t+1}(Y_t^-) \right]. \tag{58a}$$

Similarly, we have

$$\mathop{\mathbb{E}}_{Y_t \sim q_t} \left[ \varepsilon_{\mathsf{Jacobi},t+1}(Y_t^-) \mathbb{1}\left\{ \sqrt{d\log T}\varepsilon_{\mathsf{score},t}(Y_t) + d\varepsilon_{\mathsf{Jacobi},t}(Y_t)\log T \lesssim T \right\} \right] \lesssim \mathop{\mathbb{E}}_{Y_t^- \sim q_{t+1}} \left[ \varepsilon_{\mathsf{Jacobi},t+1}(Y_t^-) \right]. \tag{58b}$$

**Step 4: coping with the set $\mathcal{I}_1$.** In view of Lemma 7, the condition $S_{\tau-1}(y_{\tau-1}) \le c_{14}$ implies that

$$\frac{q_1(y_1)}{p_1(y_1)} \le \frac{2q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})}.$$

This in turn allows one to obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_1(Y_1)}{p_1(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1 \right\} \right] &\le 2 \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{\tau-1}(Y_1)}{p_{\tau-1}(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1 \right\} \right] \\
&= 2\sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{t-1}(Y_1)}{p_{t-1}(Y_1)} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_T \in \mathcal{I}_1 \right\} \mathbb{1}\{\tau = t\} \right] \\
&\le 2\sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\left\{ Y_1 \in \mathcal{E}, Y_t \in \mathcal{J}_t \right\} \right],
\end{aligned} \tag{59}$$

where the last line comes from the definition of $\mathcal{I}_1$ (cf. (55b)), with $\mathcal{J}_t$ defined as

$$\mathcal{J}_t \coloneqq \left\{ y_t : -\log q_t(y_t) \ge c_\tau d \log T \right\}. \tag{60}$$

To bound the right-hand side of (59), we make note of the following identities:

$$1 = \mathop{\mathbb{E}}_{Y_t \sim p_t} \left[ \frac{q_t(Y_t)}{p_t(Y_t)} \right] = \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_t(Y_t)}{p_t(Y_t)} \right] = \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_t(Y_t)}{p_t(Y_t)} \big( \mathbb{1}\{Y_t \in \mathcal{J}_t\} + \mathbb{1}\{Y_t \in \mathcal{J}_t^c\} \big) \right], \tag{61a}$$

$$1 = \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \right] = \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \big( \mathbb{1}\{Y_t \in \mathcal{J}_t\} + \mathbb{1}\{Y_t \in \mathcal{J}_t^c\} \big) \right], \tag{61b}$$

which in turn imply that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\{Y_t \in \mathcal{J}_t\} \right] \\
&= 1 - \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\{Y_t \in \mathcal{J}_t^c\} \right] \\
&= \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \left( \frac{q_t(Y_t)}{p_t(Y_t)} - \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \right) \mathbb{1}\{Y_t \in \mathcal{J}_t^c\} \right] + \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_t(Y_t)}{p_t(Y_t)} \mathbb{1}\{Y_t \in \mathcal{J}_t\} \right] \\
&\le \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \left( \frac{q_t(Y_t)}{p_t(Y_t)} - \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \right) \mathbb{1}\{Y_t \in \mathcal{J}_t^c\} \right] + O\Big( \exp\big( -c_6 d \log T \big) \Big).
\end{aligned}$$

Here, the last line follows since

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_t(Y_t)}{p_t(Y_t)} \mathbb{1}\{Y_t \in \mathcal{J}_t\} \right] = \mathop{\mathbb{E}}_{Y_t \sim p_t} \left[ \frac{q_t(Y_t)}{p_t(Y_t)} \mathbb{1}\{Y_t \in \mathcal{J}_t\} \right] &= \mathop{\mathbb{E}}_{Y_t \sim q_t} \left[ \mathbb{1}\{Y_t \in \mathcal{J}_t\} \right] \\
&\lesssim \exp\big( -c_6 d \log T \big),
\end{aligned}$$

provided that $c_{20} > 0$ is large enough. As a consequence, the right-hand side of (59) can be bounded by

$$\sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T \sim p_T} \left[ \frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})} \mathbb{1}\{Y_1 \in \mathcal{E}, Y_t \in \mathcal{J}_t\} \right]$$

$$\leq \sum_{t=2}^{T} \mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{q_t(Y_t)}{p_t(Y_t)}-\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})}\right)\mathbb{1}\left\{Y_t\in\mathcal{J}_t^{\mathrm{c}}\right\}\right]+O\Big(T\exp\big(-c_6 d\log T\big)\Big). \tag{62}$$

Moreover, the first term in the last line (62) can be decomposed as follows

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{q_t(Y_t)}{p_t(Y_t)}-\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})}\right)\mathbb{1}\left\{Y_t\in\mathcal{J}_t^{\mathrm{c}}\right\}\right]\\
&\quad=\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{q_t(Y_t)}{p_t(Y_t)}-\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})}\right)\mathbb{1}\left\{Y_t\in\mathcal{J}_t^{\mathrm{c}},\xi_t(Y_t)<c_{14}\right\}\right]\\
&\qquad+\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{q_t(Y_t)}{p_t(Y_t)}-\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})}\right)\mathbb{1}\left\{Y_t\in\mathcal{J}_t^{\mathrm{c}},\xi_t(Y_t)\geq c_{14}\right\}\right],
\end{aligned} \tag{63}$$

leaving us with two terms to control.

- With regards to the first term on the right-hand side of (63), for $Y_t$ satisfies $\log q_t(Y_t)\leq -c_\tau d\log T$ and $\xi_t(Y_t)<c_{14}$, we can directly apply Lemma 6 to obtain a one-step version of Lemma 7 to control the difference between the density ratio $\frac{q_t(Y_t)}{p_t(Y_t)}$ and $\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})}$ as follows

$$\begin{aligned}
\frac{p_{t-1}(y_{t-1})}{q_{t-1}(y_{t-1})}=\frac{p_t(y_t)}{q_t(y_t)}\cdot\Bigg\{1+O\left(\frac{d^6\log^6 T}{T^3}\right)+\\
O\left(\frac{\big(\varepsilon_{\mathsf{score},t}(y_t)+\varepsilon_{\mathsf{score},t}(\Phi_t(y_t))\big)\sqrt{d\log^3 T}}{T}+\frac{d\log T\big(\varepsilon_{\mathsf{Jacobi},t}(y_t)+\varepsilon_{\mathsf{Jacobi},t}(\Phi_t(y_t))\big)}{T}\right)\Bigg\}
\end{aligned}$$

which is an intermediate step in the proof of Lemma 7 (referring (94) in Section B.3 for details). The above relation yields:

$$\begin{aligned}
&\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{q_t(Y_t)}{p_t(Y_t)}-\frac{q_{t-1}(Y_{t-1})}{p_{t-1}(Y_{t-1})}\right)\mathbb{1}\left\{Y_t\in\mathcal{J}_t^{\mathrm{c}},\xi_t(Y_t)<c_{14}\right\}\right]\\
&\qquad\lesssim\frac{d^6\log^6 T}{T^2}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+d\log T\varepsilon_{\mathsf{Jacobi}}.
\end{aligned}$$

- When it comes to the second term on the right-hand side of (63), we can invoke similar arguments as in Li et al. (2023) (i.e., the arguments therein to bound $\mathcal{I}_3$), as well as the relation (58) for the score error of $Y_t^-$, to obtain

$$\sum_{t=2}^{T}\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_t(Y_t)}{p_t(Y_t)}\mathbb{1}\left\{Y_t\in\mathcal{J}_t^{\mathrm{c}},\xi_t(Y_t)\geq c_{14}\right\}\right]\lesssim\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+d\log T\varepsilon_{\mathsf{Jacobi}}.$$

Putting all this together, we arrive at

$$\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_1\right\}\right]\lesssim\frac{d^6\log^6 T}{T^2}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+d\log T\varepsilon_{\mathsf{Jacobi}}. \tag{64}$$

**Step 5: coping with the remaining sets.** The analyses for $\mathcal{I}_2,\mathcal{I}_3$ and $\mathcal{I}_4$ are similar to Li et al. (2023). For the sake of brevity, we state the combined result in the lemma below and omit the proof.

**Lemma 8.** *It holds that*

$$\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1(Y_1)}{p_1(Y_1)}\mathbb{1}\left\{Y_1\in\mathcal{E},Y_T\in\mathcal{I}_2\cup\mathcal{I}_3\cup\mathcal{I}_4\right\}\right]\lesssim\frac{d^6\log^6 T}{T^2}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+(d\log T)\varepsilon_{\mathsf{Jacobi}}. \tag{65}$$

**Step 6: putting all pieces together.** Given the preceding results on $\mathcal{I}_0$ to $\mathcal{I}_4$, we can substitute the upper bounds derived in (57),(64) and (65) back into (53), which leads to the following conclusion:

$$\begin{aligned}\mathrm{TV}\left(p_1, q_1\right) &\leq \mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\left(\frac{q_1\left(Y_1\right)}{p_1\left(Y_1\right)}-1\right)\mathbb{1}\left\{Y_1\in\mathcal{E}, Y_T\in\mathcal{I}_0\right\}\right]+\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1\left(Y_1\right)}{p_1\left(Y_1\right)}\mathbb{1}\left\{Y_1\in\mathcal{E}, Y_T\in\mathcal{I}_1\right\}\right]\\ &\quad+\mathop{\mathbb{E}}_{Y_T\sim p_T}\left[\frac{q_1\left(Y_1\right)}{p_1\left(Y_1\right)}\mathbb{1}\left\{Y_1\in\mathcal{E}, Y_T\in\mathcal{I}_2\cup\mathcal{I}_3\cup\mathcal{I}_4\right\}\right]+O\Big(\exp\left(-c_6 d\log T\right)\Big)\\ &\lesssim \frac{d^6\log^6 T}{T^2}+\sqrt{d\log^3 T}\varepsilon_{\mathsf{score}}+(d\log T)\varepsilon_{\mathsf{Jacobi}}.\end{aligned}$$

## B.2 Proof of Lemma 6

### B.2.1 Proof of property (49)

This property can be established in a similar way to Li et al. (2023, Lemma 4). Before proceeding, let us introduce the following vector:

$$\begin{aligned}v_t(x) &\coloneqq x-\phi_t(x)=x-\phi_t^\star(x)+\phi_t^\star(x)-\phi_t(x)\\ &=-\frac{1-\alpha_{t+1}}{2\left(1-\overline{\alpha}_t\right)}\int_{x_0}\left(x-\sqrt{\overline{\alpha}_t}x_0\right)p_{X_0|X_t}\left(x_0\mid x\right)\mathrm{d}x_0+\frac{1-\alpha_{t+1}}{2}\big(s_t(x)-s_t^\star(x)\big),\end{aligned}$$

where we have invoked the definitions of $\phi_t^\star$ and $\phi_t$, as well as the property (30). For notational simplicity, we shall abbreviate $v=v_t(x)$ in the following analysis.

Recognizing that

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

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

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

Additionally, recalling that

$$\frac{1}{\sqrt{\alpha_{t+1}}}X_{t+1}\overset{\mathrm{d}}{=}\frac{1}{\sqrt{\alpha_{t+1}}}\left(\sqrt{\overline{\alpha}_{t+1}}X_0+\sqrt{1-\overline{\alpha}_{t+1}}W\right)=\sqrt{\overline{\alpha}_t}X_0+\sqrt{\frac{1}{\alpha_{t+1}}-\overline{\alpha}_t}W,$$

one can demonstrate that

$$\begin{aligned}\frac{p_{X_{t+1}/\sqrt{\alpha_{t+1}}}\left(\phi_t(x)\right)}{p_{X_t}(x)} &=\frac{1}{p_{X_t}(x)}\int_{x_0}p_{X_0}\left(x_0\right)\frac{1}{\left(2\pi\left(\alpha_{t+1}^{-1}-\overline{\alpha}_t\right)\right)^{d/2}}\exp\left(-\frac{\left\|\phi_t(x)-\sqrt{\overline{\alpha}_t}x_0\right\|_2^2}{2\left(\alpha_{t+1}^{-1}-\overline{\alpha}_t\right)}\right)\mathrm{d}x_0\\ &=\frac{1}{p_{X_t}(x)}\int_{x_0}p_{X_0}\left(x_0\right)\frac{1}{\left(2\pi\left(\alpha_{t+1}^{-1}-\overline{\alpha}_t\right)\right)^{d/2}}\exp\left(-\frac{\left\|x-\sqrt{\overline{\alpha}_t}x_0\right\|_2^2}{2\left(1-\overline{\alpha}_t\right)}\right)\\ &\qquad\cdot\exp\left(-\frac{(1-\alpha_{t+1}^{-1})\left\|x-\sqrt{\overline{\alpha}_t}x_0\right\|_2^2}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|v\|_2^2-2v^\top\left(x-\sqrt{\overline{\alpha}_t}x_0\right)}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}\right)\mathrm{d}x_0\\ &=\left(\frac{1-\overline{\alpha}_t}{\alpha_{t+1}^{-1}-\overline{\alpha}_t}\right)^{d/2}\cdot\int_{x_0}p_{X_0\,|\,X_t}(x_0\,|\,x)\cdot\\ &\qquad\exp\left(\frac{(\alpha_{t+1}^{-1}-1)\left\|x-\sqrt{\overline{\alpha}_t}x_0\right\|_2^2}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)(1-\overline{\alpha}_t)}-\frac{\|v\|_2^2-2v^\top\left(x-\sqrt{\overline{\alpha}_t}x_0\right)}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}\right)\mathrm{d}x_0.\end{aligned}$$

To lower bound the above expression, we can focus on controlling the second term within the exponential part of the integral over the set $\mathcal{G}_c^{\text{typical}}$ defined as follows, since the first term is always non-negative:

$$\mathcal{G}_c^{\text{typical}} \coloneqq \left\{x_0 : \left\|x - \sqrt{\overline{\alpha}_t}x_0\right\|_2 \le 5c\sqrt{c_6 d\,(1-\overline{\alpha}_t)\log T}\right\}. \tag{66}$$

Furthermore, we make the following observations:

- When (48) holds, Lemma 3 implies that

$$\mathbb{P}\left(\left\|\sqrt{\overline{\alpha}_t}X_0 - x\right\|_2 > 5c_5\sqrt{c_6 d\,(1-\overline{\alpha}_t)\log T} \mid X_t = x\right) \le \exp\left(-c_5^2 c_6 d\log T\right) \tag{67a}$$

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

- The we can makes use of the above relation to bound $v$ as follows:

$$\begin{aligned}\|v\|_2 &\le \frac{1-\alpha_{t+1}}{2}\varepsilon_{\text{score},t}(x) + \frac{1-\alpha_{t+1}}{2\,(1-\overline{\alpha}_t)}\mathbb{E}\left[\left\|\sqrt{\overline{\alpha}_t}X_0 - x\right\|_2 \mid X_t = x\right] \\ &\le \frac{1-\alpha_{t+1}}{2}\varepsilon_{\text{score},t}(x) + \frac{6\,(1-\alpha_{t+1})}{1-\overline{\alpha}_t}\sqrt{c_6 d\,(1-\overline{\alpha}_t)\log T}. \end{aligned} \tag{67b}$$

Therefore, we obtain that: for any $x_0 \in \mathcal{G}$,

$$\begin{aligned}\frac{\|v\|_2^2}{2\left(\alpha_{t+1}^{-1} - \overline{\alpha}_t\right)} &\overset{(i)}{\le} \frac{(1-\alpha_{t+1})^2}{4\left(\alpha_{t+1}^{-1} - \overline{\alpha}_t\right)}\varepsilon_{\text{score},t}(x)^2 + \frac{36\,(1-\alpha_{t+1})^2}{(1-\overline{\alpha}_t)^2\left(\alpha_{t+1}^{-1} - \overline{\alpha}_t\right)}c_6 d\log T \\ &\overset{(ii)}{\le} \frac{2c_1^2\log^2 T}{T^2}\varepsilon_{\text{score},t}(x)^2 + \frac{2304c_1^2}{T^2}c_6 d\log^3 T; \\ \left|\frac{v^\top\left(x - \sqrt{\overline{\alpha}_t}x_0\right)}{\alpha_{t+1}^{-1} - \overline{\alpha}_t}\right| &\overset{(iii)}{\le} \frac{\|v\|_2\left\|x - \sqrt{\overline{\alpha}_t}x_0\right\|_2}{\alpha_{t+1}^{-1} - \overline{\alpha}_t} \\ &\overset{(iv)}{\le} \frac{20cc_1}{T}\varepsilon_{\text{score},t}(x)\sqrt{c_6 d\,(1-\overline{\alpha}_t)\log^3 T} + \frac{240cc_1c_6 d\log^2 T}{T}\end{aligned}$$

Here, (i) is due to (67); (ii) holds by the choice of learning rates in (33); (iii) follows from the Cauchy-Schwarz inequality; and (iv) comes from the definition of $\mathcal{G}$ and (33). Moreover, (33) also guarantees that

$$\left(\frac{1-\overline{\alpha}_t}{\alpha_{t+1}^{-1} - \overline{\alpha}_t}\right)^{d/2} = \left(1 - \frac{1-\alpha_{t+1}}{1-\overline{\alpha}_{t+1}}\right)^{\frac{d}{2}} \ge \exp\left(-\frac{4c_1 d\log T}{T}\right).$$

Combine the above relations to yield

$$\begin{aligned}&\frac{p_{X_{t+1}/\sqrt{\alpha_{t+1}}}\left(\phi_t(x)\right)}{p_{X_t}(x)} \\ &\ge \left(\frac{1-\overline{\alpha}_t}{\alpha_{t+1}^{-1} - \overline{\alpha}_t}\right)^{d/2}\cdot\int_{x_0\in\mathcal{G}} p_{X_0\,|\,X_t}(x_0\,|\,x)\cdot\exp\left(-\frac{\|v\|_2^2 - 2v^\top\left(x - \sqrt{\overline{\alpha}_t}x_0\right)}{2(\alpha_{t+1}^{-1} - \overline{\alpha}_t)}\right)\mathrm{d}x_0 \\ &\ge \exp\left(-\left(20c_1\varepsilon_{\mathsf{score},t}(x)\sqrt{c_6 d\log T} + 240c_1 d\log T\right)\frac{c\log T}{T}\right) \\ &\quad\cdot\exp\left(-\frac{4c_1 d\log T}{T} - \frac{2304c_1^2}{T^2}c_6 d\log^3 T - \frac{2c_1\varepsilon_{\mathsf{score},t}(x)^2\log^2 T}{T^2}\right) \\ &\ge \exp\left(-\left(20\varepsilon_{\mathsf{score},t}(x)\sqrt{d\log T} + 300d\log T\right)\frac{cc_1\log T}{T}\right)\end{aligned}$$

provided that $T \ge \frac{386}{5}c_1c_6\log T$ and $\frac{40c_1\varepsilon_{\text{score},t}(x)\log^{\frac{3}{2}}T}{T} \le \sqrt{d}$. Taking any fixed $c \ge 2$ we obtain the desired result.

### B.2.2 Proof of property (50a)

Before embarking on the proof, we first single out some useful properties about $\psi_t^\star$ and $s_t^*$.

**Lemma 9.** *Under the same conditions as Lemma 6, it holds that*

$$\|\psi_t(x) - \psi_t^\star(x)\|_2 = O\bigg(\frac{\log T}{T}\Big\{\varepsilon_{\mathsf{score},t}(x) + \varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)\Big\}\bigg), \tag{68a}$$

*and*

$$\begin{aligned}\frac{\partial \psi_t(x,\Phi_t(x))}{\partial x} &= \frac{\partial \psi_t^\star(x,\Phi_t^\star(x))}{\partial x} + \widetilde{\zeta}_t \\ &= \left(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x) + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\frac{\partial\big(s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(\Phi_t^\star(x))\big)}{\partial x}\right) + \widetilde{\zeta}_t\end{aligned} \tag{68b}$$

*where the residual term $\widetilde{\zeta}_t$ satisfies*

$$\|\widetilde{\zeta}_t\| = O\bigg(\frac{\log T}{T}\Big\{\varepsilon_{\mathsf{score},t}(x)/\sqrt{d} + \varepsilon_{\mathsf{Jacobi},t}(x) + \varepsilon_{\mathsf{Jacobi},t+1}\big(\Phi_t(x)\big)\Big\}\bigg).$$

Additionally, we introduce the following notations for simplicity:

$$w = \Phi_t^\star(x) = \sqrt{\alpha_{t+1}}\bigg(x - \frac{1-\alpha_{t+1}}{2}s_t^\star(x)\bigg), \tag{69a}$$

$$z = -(1-\overline{\alpha}_t)s_t^\star(x) = g_t(x). \tag{69b}$$

Then the characterization of $s_t^*$ is summarized in the following lemma.

**Lemma 10.** *Under the same conditions as Lemma 6, equipped with the notation (69), we can write*

$$\begin{aligned}s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(w) = &-\bigg(\frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^2} - \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^3}\|z\|_2^2\bigg)z \\ &- \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^3}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top z\mathrm{d}x_0 \\ &+ \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^3}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x - \sqrt{\overline{\alpha}_t}x_0 - z\big)\mathrm{d}x_0 + \zeta_{s_t^*}\end{aligned} \tag{70}$$

*where $\|\zeta_{s_t^*}\|_2 = O\Big(\frac{(d(1-\alpha_{t+1})\log T)^{3/2}}{(1-\overline{\alpha}_t)^2}\Big)$, and*

$$\frac{\partial\big(s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(w)\big)}{\partial x} = -\frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^2}J_t(x) + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^3}(H_1 + H_4 + H_2 - H_3) + \zeta_{J_t} \tag{71}$$

*where $\|\zeta_{J_t}\| = O\Big(d^2\frac{(1-\alpha_{t+1})^{3/2}}{(1-\overline{\alpha}_{t+1})^{5/2}}\log^2 T\Big)$. Here, we denote $J_t = \frac{\partial z}{\partial x}$, and*

$$\begin{aligned}H_1 &:= \frac{\partial}{\partial x}\|z\|_2^2 z, \\ H_2 &:= \frac{\partial}{\partial x}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0, \\ H_3 &:= \frac{\partial}{\partial x}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2 z, \\ H_4 &:= \frac{\partial}{\partial x}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)\big(x - \sqrt{\overline{\alpha}_t}x_0\big)^\top z\mathrm{d}x_0.\end{aligned}$$

To streamline presentation, we leave the proofs of Lemma 9 and Lemma 10 to Section B.2.5.

Equipped with the relations in Lemma 10, we can derive that

$$\big\|s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(w)\big\|_2 \lesssim (1-\alpha_{t+1})\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{3/2}, \tag{73a}$$

$$\Big\|\frac{\partial\big(s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(w)\big)}{\partial x}\Big\| \lesssim \frac{d^2(1-\alpha_{t+1})\log^2 T}{(1-\overline{\alpha}_t)^2}. \tag{73b}$$

The proof of (73) is also deferred to Section B.2.5.

Next, let us introduce the following vectors:

$$u_t(x) := x - \psi_t(x),$$
$$u_t^\star(x) := x - \psi_t^\star(x).$$

For notational simplicity, we shall abbreviate $u = u_t(x)$ and $u^\star = u_t^\star(x)$ in the following analysis. Akin to the calculations in Appendix B.2.1, we can obtain

$$p_{\sqrt{\alpha_t}X_{t-1}}\big(\psi_t(x)\big) = p_{X_t}(x)\Big(\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\Big)^{d/2} \cdot \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\Big(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}x_0. \tag{74}$$

Similarly, by focusing mainly on the following set given $x$:

$$\mathcal{G} := \big\{x_0 : \big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2 \lesssim \sqrt{d(1-\overline{\alpha}_t)\log T}\big\}, \tag{75}$$

we can derive

$$\begin{aligned}&\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\exp\Big(-\frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\Big)\mathrm{d}x_0 = O\big(\exp(-c_8 d\log T)\big)\\ &+\int_{x_0\in\mathcal{E}} p_{X_0|X_t}\left(x_0\mid x\right)\exp\left(-\frac{(1-\alpha_t)\left\|x-\sqrt{\overline{\alpha}_t}x_0\right\|_2^2}{2\left(\alpha_t-\overline{\alpha}_t\right)\left(1-\overline{\alpha}_t\right)} - \frac{\|u\|_2^2 - 2u^\top\left(x-\sqrt{\overline{\alpha}_t}x_0\right)}{2\left(\alpha_t-\overline{\alpha}_t\right)}\right)\mathrm{d}x_0 =: \mathrm{RHS}\end{aligned} \tag{76}$$

for some numerical constant $c_8 > 0$. To further control the right-hand side above, recall that the learning rates are selected such that $\frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \le \frac{4c_1\log T}{T}$ for $1 < t \le T$ (see (33b)). In view of the Taylor expansion $e^{-x} = 1 - x + \frac{1}{2}x^2 + O\left(x^3\right)$ for $x \le 1/2$, we can derive

$$\begin{aligned}\mathrm{RHS} &= O\big(\exp(-c_8 d\log T)\big) + O\Big(\frac{d^3\log^6 T}{T^3} + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t}(x) + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)\Big)\\ &+\int_{x_0\in\mathcal{E}} p_{X_0\,|\,X_t}(x_0\,|\,x)\Big\{1 - \frac{(1-\alpha_t)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\|z\|_2^2 - 2u^{\star\top}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)}\\ &+\frac{(1-\alpha_t)^2}{8(\alpha_t-\overline{\alpha}_t)^2(1-\overline{\alpha}_t)^2}\Big(\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 - z^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\Big)^2\Big\}\mathrm{d}x_0.\end{aligned} \tag{77}$$

Here, we have made use of the following facts:

$$\begin{aligned}\Big\|u - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}z\Big\|_2 &= \Big\|x - \psi_t(x) + \frac{1-\alpha_t}{2}s_t^*(x)\Big\|_2\\ &\overset{\text{(i)}}{\le} \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\big\|s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star\big(\Phi_t^\star(x)\big)\big\|_2 + O\Big(\frac{\log T}{T}\Big\{\varepsilon_{\mathsf{score},t}(x) + \varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)\Big\}\Big)\\ &\overset{\text{(ii)}}{\lesssim} (1-\alpha_t)^2\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{3/2} + \frac{\log T}{T}\varepsilon_{\mathsf{score},t}(x) + \frac{\log T}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big),\end{aligned} \tag{78}$$

where (i) follows from (68a) in Lemma 9 and (ii) follows from (73).

Moreover, for any $x_0 \in \mathcal{E}$, using the definition of $\mathcal{E}$ (cf. (75)) and combining it with the properties (33) of the learning rates, we reach

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

As a result, we can derive

$$\begin{aligned}
&\frac{\|u\|_2^2 - 2u^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)} \\
&\overset{\text{(i)}}{=} \frac{\frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\|z\|_2^2 - 2u^{\star\top}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)} + O\bigg(\frac{d^2\log^5 T}{T^3} + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t}(x) + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)\bigg) \\
&= \frac{z^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} + O\bigg(\frac{d^2\log^4 T}{T^2} + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t}(x) + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)\bigg) \\
&= O\bigg(\frac{d\log^2 T}{T} + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t}(x) + \frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)\bigg),
\end{aligned}$$

where (i) follows from (78) and Lemma 10. Taking the above results together and using the following basic properties regarding quantities $A_t, \ldots, E_t$ (defined in (46))

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

we arrive at

$$(77) = 1 - \frac{(1-\alpha_t)(A_t-B_t)}{2(\alpha_t-\overline{\alpha}_t)} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[-B_t^2 + C_t + 2D_t - 3E_t + A_tB_t\big] + O\bigg(\frac{d^3\log^6 T}{T^3}\bigg).$$

Once again, we note that integrating over the set $\mathcal{E}$ and over all possible $x_0$ only incurs a difference at most as large as $O\big(\exp(-c_8 d\log T)\big)$.

Putting the preceding results together establishes the claimed property (50a).

### B.2.3 Proof of property (50b)

Consider any random vector $Y$, and recall the basic transformation

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

where $\frac{\partial\psi_t(x)}{\partial x}$ denotes the Jacobian matrix. It then comes down to controlling the quantity $\det\Big(\frac{\partial\psi_t(x)}{\partial x}\Big)^{-1}$.

Towards this end, note that the determinant of a matrix obeys

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

with the proviso that $d\|A\| \lesssim 1$. This relation taken together with $\frac{\partial\psi_t(x)}{\partial x} = I - \frac{\partial u^\star}{\partial x} + \frac{\partial(u^\star-u)}{\partial x}$ leads to

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

$$= \Big\{1 + \mathsf{Tr}\Big(\frac{\partial u^\star}{\partial x}\Big) + \frac{1}{2}\Big[\mathsf{Tr}\Big(\frac{\partial u^\star}{\partial x}\Big)^2 + \Big\|\frac{\partial u^\star}{\partial x}\Big\|_{\mathrm{F}}^2\Big] + O\Big(d^3\Big\|\frac{\partial u^\star}{\partial x}\Big\| + d\Big\|\frac{\partial (u-u^\star)}{\partial x}\Big\|\Big)\Big\} p_Y(x). \quad (79)$$

To further control the right-hand side of the above display, let us first make note of several identities introduced in Lemma 10:

$$J_t = I + \frac{1}{1-\overline{\alpha}_t}\Big\{\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)^\top - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top \mathrm{d}x_0\Big\}; \quad (80a)$$

$$H_1 = \|z\|_2^2 J_t + 2zz^\top J_t; \quad (80b)$$

$$H_2 = \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 \mathrm{d}x_0 I + 2\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top \mathrm{d}x_0 + \frac{1}{1-\overline{\alpha}_t}\Big(\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\Big)\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)^\top - \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top \mathrm{d}x_0\Big); \quad (80c)$$

$$H_3 = \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 \mathrm{d}x_0 J_t + 2zz^\top + \frac{1}{1-\overline{\alpha}_t}\Big(\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\Big)zz^\top - z\Big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\Big)^\top\Big); \quad (80d)$$

$$H_4 = \|z\|_2^2 I + zz^\top + \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top J_t \mathrm{d}x_0 + \frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(z^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)z^\top \mathrm{d}x_0 - \frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(z^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top \mathrm{d}x_0. \quad (80e)$$

The above identities can be directly verified through elementary calculation involving Gaussian integration and derivatives, which are omitted here for the sake of brevity.

Recall the definition of $u^\star$ that

$$\begin{aligned}\frac{\partial u^\star}{\partial x} &= -\frac{1-\alpha_t}{2}J_{s_t^\star}(x) - \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\frac{\partial\big(s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(w)\big)}{\partial x} \\ &\overset{\text{(i)}}{=} \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x) - \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\cdot \\ &\qquad \left(-\frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^2}J_t(x) + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^3}(H_1+H_4+H_2-H_3) + \zeta_{J_t}\right),\end{aligned}$$

where $\|\zeta_{J_t}\| \lesssim d^2\frac{(1-\alpha_{t+1})^{3/2}}{(1-\overline{\alpha}_{t+1})^{5/2}}\log^2 T$. Here, (i) follows from Lemma 10. Then, invoking (80) and the definitions of $A_t$ to $E_t$ gives

$$\Big\|\frac{\partial u^\star}{\partial x}\Big\| \lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t}, \quad (81a)$$

$$\mathsf{Tr}\Big(\frac{\partial u^\star}{\partial x}\Big) = \frac{(1-\alpha_t)\big(d+B_t-A_t\big)}{2(1-\overline{\alpha}_t)} + \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big(d-2A_t-A_t^2+3A_tB_t+2B_t-3B_t^2+C_t+4D_t-3E_t-F_t\big), \quad (81b)$$

$$\Big\|\frac{\partial u^\star}{\partial x}\Big\|_{\mathrm{F}}^2 = \frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\Big\|\frac{\partial z}{\partial x}\Big\|_{\mathrm{F}}^2 + O\Big(d^5\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big)$$
$$= \frac{(1-\alpha_t)^2}{4(1-\overline{\alpha}_t)^2}\big(d+2(B_t-A_t)+B_t^2+F_t-2D_t\big) + O\Big(d^5\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big), \tag{81c}$$

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

$$F_t(x) \coloneqq \Big\|\frac{1}{1-\overline{\alpha}_t}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top \mathrm{d}x_0\Big\|_{\mathrm{F}}^2. \tag{81d}$$

Further note that $\frac{\partial(u^\star-u)}{\partial x} = \widetilde{\zeta}_t$, where $\widetilde{\zeta}_t$ is the residual term defined in (68b) from Lemma 9, and satisfies

$$\|\widetilde{\zeta}_t\| = O\Big(\frac{\log T}{T}\Big\{\varepsilon_{\mathsf{score},t}(x)/\sqrt{d} + \varepsilon_{\mathsf{Jacobi},t}(x) + \varepsilon_{\mathsf{Jacobi},t+1}\big(\Phi_t(x)\big)\Big\}\Big). \tag{81e}$$

Substituting these results into inequality (79) leads to

$$p_{\psi_t(Y)}(\psi_t(x)) = p_Y(x)\Big\{1 + \frac{(1-\alpha_t)(d+B_t-A_t)}{2(1-\overline{\alpha}_t)} + O\Big(d^6\Big(\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t}\Big)^3\log^3 T\Big)$$
$$+ O\Big(\frac{\sqrt{d\log^3 T}}{T}\varepsilon_{\mathsf{score},t}(x) + \frac{d\log T}{T}\Big(\varepsilon_{\mathsf{Jacobi},t}(x) + \varepsilon_{\mathsf{Jacobi},t+1}\big(\Phi_t(x)\big)\Big)\Big) \tag{82}$$
$$+ \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)^2}\big[d(d+2) + (4+2d)(B_t-A_t) - B_t^2 + C_t + 2D_t - 3E_t + A_tB_t\big]\Big\}. \tag{83}$$

### B.2.4 Proof of property (51)

Following similar arguments as in Li et al. (2023, relation (58a)), we can obtain

$$\frac{p_{X_{t+1}/\sqrt{\alpha_{t+1}}}\big(\phi_t(x)\big)}{p_{X_t}(x)}$$
$$= \Big(\frac{\alpha_{t+1}-\overline{\alpha}_{t+1}}{1-\overline{\alpha}_{t+1}}\Big)^{d/2}\cdot\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\cdot$$
$$\exp\Big(-\frac{(1-\alpha_{t+1}^{-1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\|v\|_2^2 - 2v^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}\Big)\mathrm{d}x_0$$
$$= \Big(1 - \frac{d(1-\alpha_{t+1})}{2(1-\overline{\alpha}_{t+1})} + O\big(\frac{d^2(1-\alpha_{t+1})^2}{(1-\overline{\alpha}_{t+1})^2}\big)\Big)\cdot\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\cdot$$
$$\exp\Big(\frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2(1-\overline{\alpha}_{t+1})(1-\overline{\alpha}_t)} - \frac{\|v\|_2^2 - 2v^\top\big(x-\sqrt{\overline{\alpha}_t}x_0\big)}{2(\alpha_{t+1}^{-1}-\overline{\alpha}_t)}\Big)\mathrm{d}x_0$$
$$= 1 - \frac{d(1-\alpha_{t+1})}{2(1-\overline{\alpha}_{t+1})} + O\Big(c_6^2d^2\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_{t+1}}\Big)^2\log^2 T + \varepsilon_{\mathsf{score},t}(x)\sqrt{c_6 d\log T}\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_{t+1}}\Big)\Big) +$$
$$\frac{(1-\alpha_{t+1})\big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0 + \alpha_{t+1}\big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2\big)}{2(1-\overline{\alpha}_{t+1})(1-\overline{\alpha}_t)}. \tag{84}$$

Similarly, using the arguments as in Li et al. (2023, relation (58b)), we can deduce that

$$\frac{p_{\phi_t(X_t)}\big(\phi_t(x)\big)}{p_{X_t}(x)} = 1 - \frac{d(1-\alpha_{t+1})}{2(1-\overline{\alpha}_{t+1})} +$$
$$\frac{(1-\alpha_{t+1})\big(\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathrm{d}x_0 + \alpha_{t+1}\big\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\big\|_2^2\big)}{2(1-\overline{\alpha}_{t+1})(1-\overline{\alpha}_t)} +$$
$$O\Big(c_6^2d^2\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_{t+1}}\Big)^2\log^2 T + c_6^3d^6\log^3 T\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_{t+1}}\Big)^3 + (1-\alpha_{t+1})d\varepsilon_{\mathsf{Jacobi},t}(x)\Big). \tag{85}$$

Consequently, it is readily seen that

$$\frac{p_{\Phi_t(X_t)}(\Phi_t(x))}{p_{X_{t+1}}(\Phi_t(x))} = \frac{p_{\phi_t(X_t)}\big(\phi_t(x)\big)}{p_{X_{t+1}/\sqrt{\alpha_{t+1}}}\big(\phi_t(x)\big)} = \frac{p_{\phi_t(X_t)}\big(\phi_t(x)\big)}{p_{X_t}(x)} \cdot \left(\frac{p_{X_{t+1}/\sqrt{\alpha_{t+1}}}\big(\phi_t(x)\big)}{p_{X_t}(x)}\right)^{-1}$$
$$= 1 + O\bigg(\frac{d^2\log^4 T}{T^2} + \frac{d^6\log^6 T}{T^3} + \frac{\sqrt{d\log T}\,\varepsilon_{\mathsf{score},t}(x) + d\varepsilon_{\mathsf{Jacobi},t}(x)\log T}{T}\bigg),$$

thus completing the proof of Lemma 6.

### B.2.5 Proof of additional lemmas

To establish Lemma 9 and Lemma 10, making use of Lemma 3, we first summarize the following norm properties of the score function $s_t^\star$ and the Jacobian matrix $J_{s_t^\star}$ for $x$ satisfying $\log p_{X_t}(x) \geq -c_6 d\log T$:

$$\|s_t^*(x_t)\|_2 \leq \frac{1}{(1-\bar\alpha_t)}\mathbb{E}\left[\left\|X_t - \sqrt{\bar\alpha_t}X_0\right\|_2 \mid X_t = x_t\right] \lesssim \sqrt{\frac{d\log T}{1-\bar\alpha_t}}, \tag{86a}$$

$$\left\|J_{s_t^\star}(x)\right\| \lesssim \frac{1}{(1-\bar\alpha_t)^2}\mathbb{E}\left[\left\|x - \sqrt{\bar\alpha_t}X_0\right\|_2^2 \mid X_t = x\right] \asymp d\log T, \tag{86b}$$

$$\|\nabla_x u^\top J_{s_t^\star}(x)u\|_2 \lesssim d^{3/2}\log^{3/2} T, \quad \text{for } u \in \mathbb{S}^{d-1}. \tag{86c}$$

The detailed calculation for the second property is presented in Li et al. (2023, Lemma 8). The third property follows a rationale akin to that for $J_{s_t^\star} = \frac{\partial s_t^\star(x)}{\partial x}$, and is therefore omitted here for the sake of brevity.

**Proof of Lemma 9.** To start with, it follows from the definitions of $\psi_t$ and $\psi_t^\star$ (cf. (43)) that

$$\psi_t(x) - \psi_t^\star(x) = \left(\frac{1-\alpha_t}{2} + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\right)\big(s_t(x) - s_t^\star(x)\big)$$
$$- \frac{(1-\alpha_t)^2\sqrt{\alpha_{t+1}}}{4(1-\alpha_{t+1})}\Big(s_{t+1}\big(\Phi_t(x)\big) - s_{t+1}^\star\big(\Phi_t(x)\big) + s_{t+1}^\star\big(\Phi_t(x)\big) - s_{t+1}^\star\big(\Phi_t^\star(x)\big)\Big).$$

Armed with this relation, to derive (68a), we only need to control the following term:

$$(1-\alpha_{t+1})\Big\|s_{t+1}^\star\big(\Phi_t(x)\big) - s_{t+1}^\star\big(\Phi_t^\star(x)\big)\Big\|_2 \overset{(i)}{\lesssim} \frac{\log T}{T}\Big\|s_{t+1}^\star\big(\Phi_t(x)\big) - s_{t+1}^\star\big(\Phi_t^\star(x)\big)\Big\|_2$$
$$\overset{(ii)}{\lesssim} \frac{\log T}{T}d(\log T)\big\|\Phi_t(x) - \Phi_t^\star(x)\big\|_2$$
$$\overset{(iii)}{\lesssim} \frac{d\log^3 T}{T^2}\varepsilon_{\mathsf{score},t}(x).$$

Here, (i) follows directly from the choice of learning rate in (33); (ii) holds by observing that both $\Phi_t(x)$ and $\Phi_t^\star(x)$ remain within the typical set with $\log p_{X_{t+1}}\Phi_t(x), \log p_{X_{t+1}}\Phi_t^\star(x) \geq -c_6 d\log T$ due to (49), and then invoking (86b); (iii) is due to the definition of $\varepsilon_{\mathsf{score},t}(x)$ (cf. (40)). Combining the above bound with (40) and (33), we arrive at

$$\|\psi_t(x) - \psi_t^\star(x)\| \lesssim \frac{\log T}{T}\varepsilon_{\mathsf{score},t}(x) + \frac{\log T}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi(x)\big) + \frac{d\log^3 T}{T^2}\varepsilon_{\mathsf{score},t}(x)$$
$$\lesssim \frac{\log T}{T}\Big(\varepsilon_{\mathsf{score},t}(x) + \varepsilon_{\mathsf{score},t+1}\big(\Phi(x)\big)\Big).$$

For (68b), by direct calculation, we have

$$\frac{\partial\psi_t(x,\Phi_t(x))}{\partial x} - \frac{\partial\psi_t^\star(x,\Phi_t^\star(x))}{\partial x} = \Big(\frac{1-\alpha_t}{2} + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\Big)\big(J_{s_t}(x) - J_{s_t^\star}(x)\big)$$

$$+\frac{\sqrt{\alpha_{t+1}}(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\Big(J_{s_{t+1}}\big(\Phi_t(x)\big)\Big(I-\frac{1-\alpha_{t+1}}{2}J_{s_t}(x)\Big)-J_{s^\star_{t+1}}\big(\Phi^\star_t(x)\big)\Big(I-\frac{1-\alpha_{t+1}}{2}J_{s^\star_t}(x)\Big)\Big).$$

The term in the first line can be directly bounded by the definition of $\varepsilon_{\mathsf{Jacobi},t}(x)$ and (33) as follows

$$\Big\|\Big(\frac{1-\alpha_t}{2}+\frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\Big)\big(J_{s_t}(x)-J_{s^\star_t}(x)\big)\Big\|\lesssim\frac{d\log T}{T}\varepsilon_{\mathsf{Jacobi},t}(x). \tag{87}$$

Turning to the second line, we have

$$\begin{aligned}&J_{s_{t+1}}\big(\Phi_t(x)\big)\Big(I-\frac{1-\alpha_{t+1}}{2}J_{s_t}(x)\Big)-J_{s^\star_{t+1}}\big(\Phi^\star_t(x)\big)\Big(I-\frac{1-\alpha_{t+1}}{2}J_{s^\star_t}(x)\Big)\\&=\Big(J_{s_{t+1}}\big(\Phi_t(x)\big)-J_{s^\star_{t+1}}\big(\Phi^\star_t(x)\big)\Big)\Big(I-\frac{1-\alpha_{t+1}}{2}J_{s^\star_t}(x)\Big)+\frac{1-\alpha_{t+1}}{2}J_{s_{t+1}}\big(\Phi_t(x)\big)\Big(J_{s^\star_t}(x)-J_{s_t}(x)\Big).\end{aligned} \tag{88}$$

To proceed, we further observe that

$$\big\|J_{s^\star_{t+1}}\big(\Phi_t(x)\big)-J_{s^\star_{t+1}}\big(\Phi^\star_t(x)\big)\big\|\lesssim(d\log T)^{\frac{3}{2}}\|\Phi_t(x)-\Phi^\star_t(x)\|_2\lesssim\frac{d^{\frac{3}{2}}\log^{\frac{5}{2}}T}{T}\varepsilon_{\mathsf{score},t}(x),$$

which is obtained by invoking (86c), (33) and (40). This bound together with (86b) allows us to control the first term in (88) as follows

$$\begin{aligned}&\Big\|\Big(J_{s_{t+1}}\big(\Phi_t(x)\big)-J_{s^\star_{t+1}}\big(\Phi^\star_t(x)\big)\Big)\Big(I-\frac{1-\alpha_{t+1}}{2}J_{s^\star_t}(x)\Big)\Big\|\\&\lesssim\Big\|\Big(J_{s_{t+1}}\big(\Phi_t(x)\big)-J_{s^\star_{t+1}}\big(\Phi_t(x)\big)\Big)\Big\|+\Big\|\Big(J_{s^\star_{t+1}}\big(\Phi_t(x)\big)-J_{s^\star_{t+1}}\big(\Phi^\star_t(x)\big)\Big)\Big\|\\&\lesssim\frac{d\log T}{T}\varepsilon_{\mathsf{Jacobi},t+1}\big(\Phi_t(x)\big)+\frac{d^{\frac{3}{2}}\log^{\frac{5}{2}}T}{T}\varepsilon_{\mathsf{score},t}(x).\end{aligned} \tag{89}$$

The second term in (88) can be controlled by (86b) and (40) as follows

$$\Big\|\frac{1-\alpha_{t+1}}{2}J_{s_{t+1}}\big(\Phi_t(x)\big)\Big(J_{s^\star_t}(x)-J_{s_t}(x)\Big)\Big\|\lesssim\frac{\log T}{T}\Big(\varepsilon_{\mathsf{Jacobi},t+1}\big(\Phi_t(x)\big)+d\log T\Big)\varepsilon_{\mathsf{Jacobi},t}(x). \tag{90}$$

Substituting (89) and (90) into (88), together with (87), we obtain

$$\begin{aligned}\Big\|\frac{\partial\psi_t(x,\Phi_t(x))}{\partial x}-\frac{\partial\psi^\star_t(x,\Phi^\star_t(x))}{\partial x}\Big\|&\lesssim\frac{d\log T}{T}\varepsilon_{\mathsf{Jacobi},t}(x)+\frac{d\log T}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)+\frac{d^{\frac{3}{2}}\log^{\frac{7}{2}}T}{T^2}\varepsilon_{\mathsf{score},t}(x)\\&\quad+\frac{\log^2T}{T^2}\Big(\varepsilon_{\mathsf{Jacobi},t+1}\big(\Phi_t(x)\big)+d\log T\Big)\varepsilon_{\mathsf{Jacobi},t}(x)\\&\lesssim\frac{d\log T}{T}\varepsilon_{\mathsf{Jacobi},t}(x)+\frac{d\log T}{T}\varepsilon_{\mathsf{score},t+1}\big(\Phi_t(x)\big)+\frac{\log T}{d^{\frac{1}{2}}T}\varepsilon_{\mathsf{score},t}(x)\end{aligned}$$

where the last inequality invokes conditions for $d$ and $T$ in (42). □

**Proof of Lemma 10.** To begin with, applying (86a) to $w$ leads to

$$\Big\|\frac{1}{\sqrt{\alpha_{t+1}}}w-x\Big\|_2=(1-\alpha_{t+1})\|s^*_t(x)\|_2\lesssim(1-\alpha_{t+1})\sqrt{\frac{d\log T}{1-\overline{\alpha}_t}}. \tag{91}$$

Then denoting $\widehat{w}\coloneqq w/\sqrt{\alpha_{t+1}}$, we can apply Lemma 4 to $(\widehat{w},x)$ to obtain

$$K_1\coloneqq\int_{x_0}p_{X_0\,|\,X_{t+1}}(x_0\,|\,w)\big(w/\sqrt{\alpha_{t+1}}-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0$$

$$
\begin{aligned}
&= \widehat{w} - x + \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\Big\{1 + \frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2\alpha_{t+1}(1-\overline{\alpha}_t)^2} - \frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(\widehat{w}-x)}{1-\overline{\alpha}_t} \\
&\quad - \int_{x_0}\Big(\frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2\alpha_{t+1}(1-\overline{\alpha}_t)^2} - \frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(\widehat{w}-x)}{1-\overline{\alpha}_t}\Big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0 \\
&\quad + O\Big(d\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}\Big)^{3/2}\log T\Big)\Big\}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0.
\end{aligned}
$$

Plugging $\widehat{w} - x = -\frac{1-\alpha_{t+1}}{2}s_t^\star(x) = \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)}z$ into the above equation and combining with the expression of $s_t^*(x)$ in (30), we can obtain

$$
\begin{aligned}
K_1 = \Big(1 + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)} + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^2}\|z\|_2^2\Big)z + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^2}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\Big\{\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 \\
- (x-\sqrt{\overline{\alpha}_t}x_0)^\top z - \int_{x_0}\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big\}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0 + \zeta_{K_1},
\end{aligned}
$$

where the residual term obeys

$$
\|\zeta_{K_1}\|_2 \lesssim \frac{\big(d(1-\alpha_{t+1})\log T\big)^{3/2}}{1-\overline{\alpha}_t}.
$$

Then we can immediately establish the first claim (70) by recognizing that

$$
s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(w) = \frac{1}{\alpha_{t+1}^{-1}-\overline{\alpha}_t}K_1 - \frac{1}{1-\overline{\alpha}_t}z.
$$

Similarly, one sees that

$$
\begin{aligned}
&K_2 \\
&\coloneqq \int_{x_0} p_{X_0\,|\,X_{t+1}}(x_0\,|\,w)\big(w/\sqrt{\alpha_{t+1}} - \sqrt{\overline{\alpha}_t}x_0\big)\big(w/\sqrt{\alpha_{t+1}} - \sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0 \\
&= \int_{x_0} p_{X_0\,|\,X_{t+1}/\sqrt{\alpha_{t+1}}}(x_0\,|\,\widehat{w})\Big[\big(\widehat{w}-x\big)\big(\widehat{w}-\sqrt{\overline{\alpha}_t}x_0\big)^\top + \big(\widehat{w}-\sqrt{\overline{\alpha}_t}x_0\big)\big(\widehat{w}-x\big)^\top - \big(\widehat{w}-x\big)\big(\widehat{w}-x\big)^\top\Big]\mathrm{d}x_0 \\
&\quad + \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\Big\{1 + \frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2\alpha_{t+1}(1-\overline{\alpha}_t)^2} - \frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(\widehat{w}-x)}{1-\overline{\alpha}_t} + O\Big(d\Big(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}\Big)^{3/2}\log T\Big) \\
&\qquad - \int_{x_0}\Big(\frac{(1-\alpha_{t+1})\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2}{2\alpha_{t+1}(1-\overline{\alpha}_t)^2} - \frac{(x-\sqrt{\overline{\alpha}_t}x_0)^\top(\widehat{w}-x)}{1-\overline{\alpha}_t}\Big)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big\} \\
&\qquad \cdot \big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0 \\
&= \int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0 \\
&\quad + \frac{1-\alpha_{t+1}}{1-\overline{\alpha}_t}zz^\top + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^2}\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x)\Big\{\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 - (x-\sqrt{\overline{\alpha}_t}x_0)^\top z \\
&\qquad - \int_{x_0}\big\|x-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2 p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0 - \|z\|_2^2\Big\}\big(x-\sqrt{\overline{\alpha}_t}x_0\big)\big(x-\sqrt{\overline{\alpha}_t}x_0\big)^\top\mathrm{d}x_0 + \zeta_{K_2},
\end{aligned}
$$

where the residual term $\zeta_{K_2}$ satisfies

$$
\|\zeta_{K_2}\| \lesssim d^2\frac{(1-\alpha_{t+1})^{3/2}}{(1-\overline{\alpha}_t)^{1/2}}\log^2 T.
$$

Then the second claim (71) immediately follows by recognizing

$$
\frac{\partial\big(s_t^\star(x) - \sqrt{\alpha_{t+1}}s_{t+1}^\star(w)\big)}{\partial x}
$$

$$
\begin{aligned}
&= \frac{\sqrt{\alpha_{t+1}}}{1-\overline{\alpha}_{t+1}} J_{t+1}(w)\frac{\partial w}{\partial x} - \frac{1}{1-\overline{\alpha}_t} J_t(x) \\
&= \frac{1}{\alpha_{t+1}^{-1}-\overline{\alpha}_t}\Big(I + \frac{1}{\alpha_{t+1}^{-1}-\overline{\alpha}_t}\Big(K_1K_1^\top - K_2\Big)\Big)\Big(I + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)} J_t(x)\Big) - \frac{1}{1-\overline{\alpha}_t} J_t(x) \\
&= -\frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^2} J_t(x) + \frac{1-\alpha_{t+1}}{2(1-\overline{\alpha}_t)^3}(H_1 + H_4 + H_2 - H_3) + \zeta_{J_t},
\end{aligned}
$$

where the residual term $\zeta_{J_t}$ satisfies

$$
\big\|\zeta_{J_t}\big\| \lesssim d^2 \frac{(1-\alpha_{t+1})^{3/2}}{(1-\overline{\alpha}_{t+1})^{5/2}} \log^2 T.
$$

□

**Proof of properties (73).** To prove these properties, we first note that Lemma 3 implies that

$$
\|z\|_2 \le \mathbb{E}\left[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2 \mid X_t = x\right] \lesssim \sqrt{d\,(1-\bar{\alpha}_t)\log T}, \tag{92a}
$$

$$
\begin{aligned}
&\left\|\int_{x_0} p_{X_0|X_t}(x_0 \mid x)\left(x - \sqrt{\overline{\alpha}_t}x_0\right)\left(X_t - \sqrt{\overline{\alpha}_t}x_0\right)^\top z \,\mathrm{d}x_0\right\|_2 \\
&\lesssim \|z\|_2\, \mathbb{E}\left[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \mid X_t = x\right] \lesssim \left(\frac{d\log T}{1-\bar{\alpha}_t}\right)^{3/2},
\end{aligned} \tag{92b}
$$

$$
\begin{aligned}
&\left\|\int_{x_0} p_{X_0|X_t}(x_0 \mid x)\big\|X_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\left(x - \sqrt{\overline{\alpha}_t}x_0 - z\right)\mathrm{d}x_0\right\|_2 \\
&\le \left\|\mathbb{E}\left[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2^3 \mid X_t = x\right]\right\|_2 + \|z\|_2\left\|\mathbb{E}\left[\big\|X_t - \sqrt{\overline{\alpha}_t}X_0\big\|_2^2 \mid X_t = x\right]\right\|_2 \\
&\lesssim \left(\frac{d\log T}{1-\bar{\alpha}_t}\right)^{3/2}.
\end{aligned} \tag{92c}
$$

Substituting the above bounds into (70) yields the first claim of (73). Similarly, the second claim follows by applying Lemma 3 and utilizing (80) in the context of (71). □

## B.3 Proof of Lemma 7

To begin with, it follows from the definition (54) of $\tau(y_T)$ that

$$
-\log q_t(y_t) \le c_\tau d\log T, \qquad \forall t < \tau(y_T).
$$

Our proof is mainly built upon Lemma 6. Specifically, combining Lemma 3, (33) and the definition (54) of $\tau(y_T)$ gives

$$
\big|B_t\big| \le \big|A_t\big| \lesssim \frac{1}{1-\overline{\alpha}_t}\cdot d(1-\overline{\alpha}_t)\log T \asymp d\log T \tag{93a}
$$

$$
\big|C_t\big| \lesssim \frac{1}{(1-\overline{\alpha}_t)^2} d^2(1-\overline{\alpha}_t)^2\log^2 T \asymp d^2\log^2 T \tag{93b}
$$

$$
\big|D_t\big| \le \frac{\|g_t(x)\|_2^2}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2 \mathrm{d}x_0 \lesssim d^2\log^2 T \tag{93c}
$$

$$
\big|E_t\big| \le \frac{\|g_t(x)\|_2^2}{(1-\overline{\alpha}_t)^2}\int p_{X_0\,|\,X_t}(x_0\,|\,x)\big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2^3 \mathrm{d}x_0 \lesssim d^2\log^2 T \tag{93d}
$$

for all $t < \tau(y_T)$. As a consequence, the properties (50a) and (50b) in Lemma 6 tell us that

$$
\frac{p_{\sqrt{\alpha_t}Y_{t-1}}\big(\psi_t(y_t)\big)}{p_{Y_t}(y_t)}\left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\psi_t(y_t)\big)}{p_{X_t}(y_t)}\right)^{-1} = \frac{p_{\psi_t(Y_t)}\big(\psi_t(y_t)\big)}{p_{Y_t}(y_t)}\left(\frac{p_{\sqrt{\alpha_t}X_{t-1}}\big(\psi_t(y_t)\big)}{p_{X_t}(y_t)}\right)^{-1}
$$

$$= 1 + O\left( \frac{d^6 \log^6 T}{T^3} + \frac{\big(\varepsilon_{\mathsf{score},t}(y_t) + \varepsilon_{\mathsf{score},t}(\Phi_t(y_t))\big)\sqrt{d\log^3 T}}{T} + \frac{d\log T\big(\varepsilon_{\mathsf{Jacobi},t}(y_t) + \varepsilon_{\mathsf{Jacobi},t}(\Phi_t(y_t))\big)}{T} \right)$$

for all $t < \tau(y_T)$. Give that $y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\psi_t(y_t)$, one can make use of the relation (47) and derive

$$\frac{p_{t-1}(y_{t-1})}{q_{t-1}(y_{t-1})} = \frac{p_t(y_t)}{q_t(y_t)} \cdot \tag{94}$$
$$\left\{1 + O\left( \frac{d^6 \log^6 T}{T^3} + \frac{\big(\varepsilon_{\mathsf{score},t}(y_t) + \varepsilon_{\mathsf{score},t}(\Phi_t(y_t))\big)\sqrt{d\log^3 T}}{T} + \frac{d\log T\big(\varepsilon_{\mathsf{Jacobi},t}(y_t) + \varepsilon_{\mathsf{Jacobi},t}(\Phi_t(y_t))\big)}{T} \right)\right\}$$

for any $t < \tau(y_T)$. If we employ the shorthand notation $\tau = \tau(y_T)$, then it can be seen that

$$\begin{aligned} \frac{q_1(y_1)}{p_1(y_1)} &= \left\{1 + O\left( \frac{d^6 \log^6 T}{T^2} + S_{\tau-1}(y_{\tau-1}) \right)\right\} \frac{q_{\tau-1}(y_{\tau-1})}{p_{\tau-1}(y_{\tau-1})} \\ &\in \left[ \frac{p_{\tau-1}(y_{\tau-1})}{2q_{\tau-1}(y_{\tau-1})}, \frac{2p_{\tau-1}(y_{\tau-1})}{q_{\tau-1}(y_{\tau-1})} \right]. \end{aligned} \tag{95a}$$

Repeating this argument also yields

$$\frac{q_t(y_t)}{2p_t(y_t)} \le \frac{q_1(y_1)}{p_1(y_1)} \le \frac{2q_t(y_t)}{p_t(y_t)}, \qquad \forall t < \tau. \tag{95b}$$

# C Analysis for the accelerated DDPM sampler (proof of Theorem 2)

In this section, we turn to the accelerated stochastic sampler and present the proof of Theorem 2.

## C.1 Main steps of the proof

**Preparation.** First, we find it convenient to introduce the following mapping

$$\widehat{\mu}_t^\star(x_t) = \frac{1}{\sqrt{\alpha_t}}\big(x_t + (1-\alpha_t)s_t^\star(x_t)\big). \tag{96}$$

For any $t$, introduce the following auxiliary sequences: $H_T \sim \mathcal{N}(0, I_d)$, and

$$H_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left\{ H_t + \sqrt{\frac{1-\alpha_t}{2}} Z_t + (1-\alpha_t)s_t^\star(H_t) - \frac{(1-\alpha_t)^{3/2}}{\sqrt{2}(1-\overline{\alpha}_t)} J_t(H_t) Z_t + \sqrt{\frac{1-\alpha_t}{2}} Z_t^+ \right\} \tag{97}$$

$$= \widehat{\mu}_t^\star(H_t) + \sqrt{\frac{1-\alpha_t}{2\alpha_t}}\Big( Z_t - \frac{1-\alpha_t}{1-\overline{\alpha}_t} J_t(H_t) Z_t + Z_t^+ \Big) \tag{98}$$

for $t = T, \cdots, 1$. We shall also adopt the following notation throughout for notational convenience:

$$\widehat{x}_t \coloneqq \frac{1}{\sqrt{\alpha_t}} x_t. \tag{99}$$

**Step 1: decomposing the KL divergence of interest.** Applying Pinsker's inequality and repeating the arguments as in Li et al. (2023, Section 5.3) lead to the following elementary decompositions:

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \sqrt{\frac{1}{2}\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1})}, \tag{100}$$

$$\mathsf{KL}(p_{X_1} \,\|\, p_{Y_1}) \le \mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) + \sum_{t=2}^{T} \mathop{\mathbb{E}}_{x\sim q_t}\Big[ \mathsf{KL}\Big( p_{X_{t-1}\,|\,X_t}(\cdot \,|\, x) \,\|\, p_{Y_{t-1}\,|\,Y_t}(\cdot \,|\, x) \Big)\Big]. \tag{101}$$

In particular, the term $\mathsf{KL}(p_{X_T} \,\|\, p_{Y_T})$ can be readily bounded by Lemma 5 as follows:

$$\mathsf{KL}(p_{X_T} \,\|\, p_{Y_T}) \lesssim \frac{1}{T^{200}}.$$

As a result, it suffices to bound $\mathsf{KL}\big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x)\,\|\,p_{Y_{t-1}\,|\,Y_t}(\cdot\,|\,x)\big)$ for each $1 < t \le T$ separately, which we shall accomplish next.

**Step 2: bounding the conditional distributions $p_{X_{t-1}\,|\,X_t}$ and $p_{H_{t-1}\,|\,H_t}$.** We now compare the two conditional distributions $p_{X_{t-1}\,|\,X_t}$ and $p_{H_{t-1}\,|\,H_t}$.

Towards this end, let us first introduce the set below:

$$\mathcal{E} \coloneqq \left\{(x_t, x_{t-1})\,|\, -\log p_{X_t}(x_t) \le \frac{1}{2}c_6 d\log T,\ \|x_{t-1} - \widehat{x}_t\|_2 \le c_3\sqrt{d(1-\alpha_t)\log T}\right\} \tag{102}$$

with $\widehat{x}_t$ defined in (99), and we would like to evaluate both $p_{H_{t-1}\,|\,H_t}$ and $p_{X_{t-1}\,|\,X_t}$ over the set $\mathcal{E}$. Regarding $p_{H_{t-1}\,|\,H_t}$, we have the following lemma.

**Lemma 11.** *For every $(x_t, x_{t-1}) \in \mathcal{E}$ as defined in (102), we have*

$$p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t) \propto \exp\left\{-\frac{\alpha_t}{2(1-\alpha_t)}\left\|\Big(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\Big)^{-1}\big(x_{t-1} - \widehat{\mu}_t^\star(x_t)\big)\right\|_2^2 + O\Big(\frac{d^3\log^5 T}{T^2}\Big)\right\}. \tag{103}$$

Turning to $p_{X_{t-1}\,|\,X_t}$ over the set $\mathcal{E}$, we can invoke Li et al. (2023, Lemma 12) to derive the following result.

**Lemma 12.** *There exists some large enough numerical constant $c_\zeta > 0$ such that: for every $(x_t, x_{t-1}) \in \mathcal{E}$,*

$$\begin{aligned} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) = {} & \frac{1}{\big(2\pi\frac{1-\alpha_t}{\alpha_t}\big)^{d/2}\big|\det\big(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\big)\big|} \\ & \cdot \exp\left(-\frac{\alpha_t}{2(1-\alpha_t)}\left\|\Big(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\Big)^{-1}\big(x_{t-1} - \widehat{\mu}_t^\star(x_t)\big)\right\|_2^2 + \zeta_t(x_{t-1}, x_t)\right) \end{aligned} \tag{104}$$

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

$$\big|\zeta_t(x_{t-1}, x_t)\big| \le c_\zeta\frac{d^3\log^{4.5} T}{T^{3/2}}. \tag{105}$$

Moving beyond the set $\mathcal{E}$, it suffices to bound the log density ratio $\log\frac{p_{X_{t-1}\,|\,X_t}}{p_{H_{t-1}\,|\,H_t}}$ for all pairs $(x_t, x_{t-1})$, which can be accomplished in a way similar to Li et al. (2023, Lemma 13).

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

$$\log\frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)} \le T^{c_0+2c_R+2}\left\{\big\|x_{t-1} - \widehat{x}_t\big\|_2^2 + \|x_t\|_2^2 + 1\right\}, \tag{106}$$

*where $c_0$ is defined in (13b).*

Equipped with Lemmas 11 to 13, one can readily repeat similar arguments as in Li et al. (2023, Step 3, Theorem 3) to derive the following result:

**Lemma 14.** *For any $1 < t \le T$, one has*

$$\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\,\|\,p_{H_{t-1}\,|\,H_t}(\cdot\,|\,x_t)\Big)\Big] \lesssim \left(\frac{d^3\log^{4.5} T}{T^{3/2}}\right)^2. \tag{107}$$

**Step 3: quantifying the KL divergence between $p_{H_{t-1}|H_t}$ and $p_{Y_{t-1}|Y_t}$.** In the previous step, we have quantified the KL divergence between $p_{X_{t-1}|X_t}$ and $p_{H_{t-1}|H_t}$. Recognizing that $H_{t-1}$ is a first-order approximation of $Y_{t-1}$ using the true score function, we still need to look at the influence of the score estimation errors, for which we resort to the lemma below.

**Lemma 15.** *For any $1 < t \le T$, one has*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\;\|\;p_{Y_{t-1}\,|\,Y_t}(\cdot\,|\,x_t)\Big)\Big] - \mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\;\|\;p_{H_{t-1}\,|\,H_t}(\cdot\,|\,x_t)\Big)\Big] \\
&\quad\lesssim \exp\big(-c_{20}d\log T\big) + \frac{d\log^3 T}{T}\mathop{\mathbb{E}}_{X_t\sim q_t}\big[\varepsilon_{\mathsf{score},t}(X_t)^2\big] + \frac{d^5\log^7 T}{T^3}. \qquad (108)
\end{aligned}$$

**Step 4: putting all this together.** We are now ready to complete the proof. Substituting (107) and (108) into the decomposition (101) yields

$$\begin{aligned}
\mathsf{KL}(p_{X_1}\,\|\,p_{Y_1}) &\le \mathsf{KL}(p_{X_T}\,\|\,p_{Y_T}) + \sum_{t=1}^{T-1}\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\;\|\;p_{H_{t-1}\,|\,H_t}(\cdot\,|\,x_t)\Big)\Big] \\
&\quad+\sum_{t=1}^{T-1}\Big\{\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\;\|\;p_{Y_{t-1}\,|\,Y_t}(\cdot\,|\,x_t)\Big)\Big] - \mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\;\|\;p_{H_{t-1}\,|\,H_t}(\cdot\,|\,x_t)\Big)\Big]\Big\} \\
&\lesssim \mathsf{KL}(p_{X_T}\,\|\,p_{Y_T}) + \sum_{2\le t\le T}\frac{d^6\log^9 T}{T^3} + \frac{d\log^3 T}{T}\sum_{t=2}^{T}\mathop{\mathbb{E}}_{X_t\sim q_t}\big[\varepsilon_{\mathsf{score},t}(X_t)^2\big] \\
&\asymp \frac{d^6\log^9 T}{T^2} + d\varepsilon_{\mathsf{score}}^2\log^3 T,
\end{aligned}$$

thereby concluding the proof of Theorem 2.

## C.2 Proof of Lemma 11

To begin with, we observe that

$$\begin{aligned}
&p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t) \propto \\
&\quad\exp\Big(-\frac{\alpha_t}{1-\alpha_t}\big(x_{t-1}-\widehat{\mu}_t^\star(x_t)\big)^\top\mathsf{Var}\Big(Z_t-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(H_t)Z_t+Z_t^+\mid H_t=x_t\Big)^{-1}\big(x_{t-1}-\widehat{\mu}_t^\star(x_t)\big)\Big).
\end{aligned}$$

It is easy to verify that

$$\begin{aligned}
&\mathsf{Var}\Big(Z_t-\frac{1-\alpha_t}{1-\overline{\alpha}_t}J_t(H_t)Z_t+Z_t^+\mid H_t=x_t\Big) \\
&\quad=2\Big(I-\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\Big)\Big(I-\frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\Big)^\top+\frac{(1-\alpha_t)^2}{2(1-\overline{\alpha}_t)^2}J_t(x_t)J_t(x_t)^\top.
\end{aligned}$$

For any $(x_{t-1},x_t)\in\mathcal{E}$, we can deduce that

$$\begin{aligned}
\|J_t(x_t)\| &\lesssim d\log T, \qquad (109a)\\
\|x_{t-1}-\widehat{\mu}_t^\star(x_t)\|_2 &\le \|x_{t-1}-\widehat{x}_t\|_2+\frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\mathbb{E}\big[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2\mid X_t=x_t\big] \\
&\overset{\text{(i)}}{\lesssim}\sqrt{d(1-\alpha_t)\log T}+\sqrt{\frac{d\log T}{1-\overline{\alpha}_t}}(1-\alpha_t)\asymp\sqrt{d(1-\alpha_t)\log T}, \qquad (109b)
\end{aligned}$$

where (109a) follows (86b) and (i) both arises from Lemma 3. Taking the above relations together and using the relation (13b), we arrive at

$$\frac{1-\alpha_t}{(1-\overline{\alpha}_t)^2}\|J_t(x_t)\|^2\|x_{t-1}-\widehat{\mu}_t^\star(x_t)\|_2^2\lesssim\frac{d^3\log^5 T}{T^2},$$

which completes the proof.

## C.3 Proof of Lemma 13

According to the expression (103), one has

$$H_{t-1} \mid H_t = x_t \;\sim\; \mathcal{N}\bigg(\widehat{\mu}_t^*(x_t),\, \underbrace{\frac{1-\alpha_t}{\alpha_t}\bigg(I - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}J_t(x_t)\bigg)^2 + \frac{(1-\alpha_t)^3}{4\alpha_t(1-\overline{\alpha}_t)^2}J_t(x_t)^2}_{=:\,\Sigma(\widehat{x}_t)}\bigg).$$

In order to quantify the above density of interest, we first bound the Jacobian matrix $J_t(x)$ defined in (31). On the one hand, the expression (32) tells us that $J_t(x) \preceq I_d$ for any $x$, given that the term within the curly bracket in (32) is a negative covariance matrix. On the other hand, $J_t(x)$ can be lower bounded by

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

where the second line applies the assumption that $\|X_0\|_2 \le T^{c_R}$, and the last line invokes the choice (13b). As a consequence, we obtain

$$\Sigma(\widehat{x}_t) \succeq \frac{1-\alpha_t}{\alpha_t}\bigg(1 - \frac{1-\alpha_t}{2(1-\overline{\alpha}_t)}\bigg)^2 I_d = \frac{1-\alpha_t}{4\alpha_t}\bigg(\frac{1-\overline{\alpha}_t + \alpha_t - \overline{\alpha}_t}{1-\overline{\alpha}_t}\bigg)^2 I_d \succeq \frac{1-\alpha_t}{4\alpha_t} I_d \succeq \frac{1-\alpha_t}{4} I_d; \tag{110a}$$

$$\Sigma(\widehat{x}_t) \preceq \frac{1-\alpha_t}{\alpha_t} T^{2c_0+2}\big(2\|\widehat{x}_t\|_2^4 + 2T^{4c_R}\big) I_d + \frac{(1-\alpha_t)^3}{4\alpha_t(1-\overline{\alpha}_t)^2} I_d \preceq 4T^{2c_0+2}\big(\|\widehat{x}_t\|_2^4 + T^{4c_R}\big) I_d. \tag{110b}$$

With the above relations in mind, we are ready to bound the density function $p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)$ for any $x_t, x_{t-1} \in \mathbb{R}^d$. It is seen from (103) that

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

where the second inequality results from (110), and the third inequality makes use of (13b) and the fact that

$$\begin{aligned}
\|x_{t-1} - \widehat{\mu}_t^\star(x_t)\|_2^2 &\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + 2\|\widehat{x}_t - \widehat{\mu}_t^\star(x_t)\|_2^2 \\
&= 2\|x_{t-1} - \widehat{x}_t\|_2^2 + 2\bigg(\frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\bigg)^2\bigg\|\int_{x_0} p_{X_0\,|\,X_t}(x_0\,|\,x_t)\big(x_t - \sqrt{\overline{\alpha}_t}x_0\big)\mathrm{d}x_0\bigg\|_2^2 \\
&\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + \frac{2(1-\alpha_t)^2}{\alpha_t(1-\overline{\alpha}_{t-1})^2}\sup_{x_0:\|x_0\|_2\le T^{c_R}}\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2 \\
&\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + \frac{64c_1^2\log^2 T}{T^2}\Big(2\|x_t\|_2^2 + 2\overline{\alpha}_t T^{2c_R}\Big) \\
&\le 2\|x_{t-1} - \widehat{x}_t\|_2^2 + \|x_t\|_2^2 + T^{2c_R}.
\end{aligned} \tag{111}$$

Given that $\log\frac{p_{X_{t-1}|X_t}(x_{t-1}\,|\,x_t)}{p_{H_{t-1}|H_t}(x_{t-1}\,|\,x_t)} \le \log\frac{1}{p_{H_{t-1}|H_t}(x_{t-1}\,|\,x_t)}$, we have concluded the proof.

## C.4 Proof of Lemma 14

Firstly, it follows from Lemma 11 and Lemma 12 that: for any $(x_t, x_{t-1}) \in \mathcal{E}$,

$$\frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)} = \exp\Big(O\Big(\frac{d^3\log^{4.5}T}{T^{3/2}}\Big)\Big) \tag{112}$$

$$= 1 + O\bigg(\frac{d^3\log^{4.5}T}{T^{3/2}}\bigg) \in \Big[\frac{1}{2}, 2\Big], \tag{113}$$

which further allows one to derive

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\,|\,X_t}(\cdot\,|\,x_t)\;\|\;p_{H_{t-1}\,|\,H_t}(\cdot\,|\,x_t)\Big)\Big] \\
&= \Big(\int_{\mathcal{E}} + \int_{\mathcal{E}^{\mathrm{c}}}\Big)p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\log\frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)}\mathrm{d}x_{t-1}\mathrm{d}x_t, \\
&\overset{\text{(i)}}{=} \int_{\mathcal{E}} p_{X_t}(x_t)\bigg\{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) - p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t) \\
&\qquad\qquad + p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\cdot O\bigg(\Big(\frac{p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)}{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)} - 1\Big)^2\bigg)\bigg\}\mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\quad + \int_{\mathcal{E}^{\mathrm{c}}} p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\log\frac{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)}{p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)}\mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\overset{\text{(ii)}}{=} \int_{\mathcal{E}} p_{X_t}(x_t)\bigg\{p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t) - p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t) + p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)O\bigg(\frac{d^6\log^9 T}{T^3}\bigg)\bigg\}\mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\quad + \int_{\mathcal{E}^{\mathrm{c}}} p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\Big\{2T\big(\|x_t\|_2^2 + \|x_{t-1}-\widehat{x}_t\|_2^2 + T^{2c_R}\big)\Big\}\mathrm{d}x_{t-1}\mathrm{d}x_t.
\end{aligned} \tag{114}$$

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

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

and in (ii) we apply (113) and Lemma 13.

Next, we would like to bound each term on the right-hand side of (114) separately. In view of the definition of the set $\mathcal{E}$ (cf. (102)), one has

$$\begin{aligned}
\mathbb{P}\big((X_t, X_{t-1}) \notin \mathcal{E}\big) &= \int_{(x_t,x_{t-1})\notin\mathcal{E}} p_{X_{t-1}}(x_{t-1})p_{X_t\,|\,X_{t-1}}(x_t\,|\,x_{t-1})\mathrm{d}x_{t-1}\mathrm{d}x_t \\
&= \int_{(x_t,x_{t-1})\notin\mathcal{E}} p_{X_{t-1}}(x_{t-1})\frac{1}{\big(2\pi(1-\alpha_t)\big)^{d/2}}\exp\bigg(-\frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|_2^2}{2(1-\alpha_t)}\bigg)\mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\le \exp\big(-c_3 d\log T\big),
\end{aligned} \tag{115}$$

and similarly,

$$\int_{(x_{t-1},x_t)\notin\mathcal{E}} p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\Big(2T\big(\|x_t\|_2^2 + \|x_{t-1}-\widehat{x}_t\|_2^2\big) + T^{2c_R}\Big)\mathrm{d}x_{t-1}\mathrm{d}x_t \le \exp\big(-c_3 d\log T\big). \tag{116}$$

In addition, for every $(x_t, x_{t-1})$ obeying $\|x_{t-1} - x_t/\sqrt{\alpha_t}\|_2 > c_3\sqrt{d(1-\alpha_t)\log T}$ and $-\log p_{X_t}(x_t) \le \frac{1}{2}c_6 d\log T$, it follows from the definition (96) of $\widehat{\mu}_t^\star(\cdot)$ that

$$\|x_{t-1} - \widehat{\mu}_t^\star(x_t)\|_2 = \bigg\|x_{t-1} - \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\mathbb{E}\Big[x_t - \sqrt{\overline{\alpha}_t}X_0 \mid X_t = x_t\Big]\bigg\|_2 \tag{117}$$

$$
\begin{aligned}
&\geq \left\|x_{t-1}-\frac{1}{\sqrt{\alpha_t}}x_t\right\|_2-\frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\mathbb{E}\Big[\big\|x_t-\sqrt{\overline{\alpha}_t}X_0\big\|_2\mid X_t=x_t\Big]\\
&\geq c_3\sqrt{d(1-\alpha_t)\log T}-6\overline{c}_5\frac{1-\alpha_t}{\sqrt{\alpha_t(1-\overline{\alpha}_t)}}\sqrt{d\log T}\\
&=\left(c_3-6\overline{c}_5\frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t(1-\overline{\alpha}_t)}}\right)\sqrt{d(1-\alpha_t)\log T}\geq\frac{c_3}{2}\sqrt{d(1-\alpha_t)\log T},
\end{aligned}\tag{118}
$$

where the third line results from (37a) in Lemma 3, and the last line applies (33) and holds true as long as $c_3$ is large enough. Taking this result together with Lemma 11 reveals that: for any $x_t$ obeying $-\log p_{X_t}(x_t)\leq\frac{1}{2}c_6 d\log T$, one has

$$
\int_{x_{t-1}:\|x_{t-1}-x_t/\sqrt{\alpha_t}\|_2>c_3\sqrt{d(1-\alpha_t)\log T}}p_{H_{t-1}\mid H_t}(x_{t-1}\mid x_t)\mathrm{d}x_{t-1}\leq\exp\Big(-\frac{c_3}{2}d\log T\Big).\tag{119}
$$

Combine (115) and (119) to arrive at

$$
\begin{aligned}
&\left|\int_{\mathcal{E}}p_{X_t}(x_t)\Big\{p_{X_{t-1}\mid X_t}(x_{t-1}\mid x_t)-p_{H_{t-1}\mid H_t}(x_{t-1}\mid x_t)\Big\}\mathrm{d}x_{t-1}\mathrm{d}x_t\right|\leq\mathbb{P}\big((X_t,X_{t-1})\notin\mathcal{E}\big)\\
&\quad+\int_{\log p_{X_t}(x_t)\leq\frac{1}{2}c_6 d\log T,\|x_{t-1}-x_t/\sqrt{\alpha_t}\|_2>c_3\sqrt{d(1-\alpha_t)\log T}}p_{X_t}(x_t)p_{H_{t-1}\mid H_t}(x_{t-1}\mid x_t)\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&\quad\leq 2\exp\Big(-\frac{c_3}{2}d\log T\Big).
\end{aligned}\tag{120}
$$

To finish up, plugging (116) and (120) into (114) yields: for each $t\geq 2$,

$$
\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\mid X_t}(\cdot\mid x_t)\;\|\;p_{Y^\star_{t-1}\mid Y_t}(\cdot\mid x_t)\Big)\Big]\lesssim\frac{d^6\log^9 T}{T^3}+3\exp\Big(-\frac{c_3}{2}d\log T\Big)\lesssim\frac{d^6\log^9 T}{T^3}.\tag{121}
$$

## C.5 Proof of Lemma 15

We first introduce the following notation:

$$
\begin{aligned}
\mu_t(x_t,z_t)&:=\frac{1}{\sqrt{\alpha_t}}\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t+(1-\alpha_t)s_t\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)\Big);\\
\mu^\star_t(x_t,z_t)&:=\frac{1}{\sqrt{\alpha_t}}\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t+(1-\alpha_t)s^\star_t(x_t)-\frac{(1-\alpha_t)^{3/2}}{\sqrt{2}(1-\overline{\alpha}_t)}J_t(x_t)z_t\Big).
\end{aligned}
$$

In the sequel, we shall use $\mu_t$ and $\mu^\star_t$ to denote $\mu_t(x_t,z_t)$ and $\mu^\star_t(x_t,z_t)$, respectively, for simplicity, as long as it is clear from the context. It is observed that

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\mid X_t}(\cdot\mid x_t)\;\|\;p_{Y_{t-1}\mid Y_t}(\cdot\mid x_t)\Big)\Big]-\mathop{\mathbb{E}}_{x_t\sim q_t}\Big[\mathsf{KL}\Big(p_{X_{t-1}\mid X_t}(\cdot\mid x_t)\;\|\;p_{H_{t-1}\mid H_t}(\cdot\mid x_t)\Big)\Big]\\
&=\int p_{X_t}(x_t)p_{X_{t-1}\mid X_t}(x_{t-1}\mid x_t)\log\frac{p_{H_{t-1}\mid H_t}(x_{t-1}\mid x_t)}{p_{Y_{t-1}\mid Y_t}(x_{t-1}\mid x_t)}\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&=\int p_{X_t}(x_t)p_{X_{t-1}\mid X_t}(x_{t-1}\mid x_t)p_{Z_t\mid H_{t-1},H_t}(z_t\mid x_{t-1},x_t)\log\frac{p_{H_{t-1}\mid H_t}(x_{t-1}\mid x_t)}{p_{Y_{t-1}\mid Y_t}(x_{t-1}\mid x_t)}\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&\overset{\text{(i)}}{\leq}\int p_{X_t}(x_t)p_{X_{t-1}\mid X_t}(x_{t-1}\mid x_t)p_{Z_t\mid H_{t-1},H_t}(z_t\mid x_{t-1},x_t)\log\frac{p_{H_{t-1}\mid H_t,Z_t}(x_{t-1}\mid x_t,z_t)}{p_{Y_{t-1}\mid Y_t,Z_t}(x_{t-1}\mid x_t,z_t)}\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&\overset{\text{(ii)}}{=}\int p_{X_t}(x_t)p_{X_{t-1}\mid X_t}(x_{t-1}\mid x_t)p_{Z_t\mid H_{t-1},H_t}(z_t\mid x_{t-1},x_t)\frac{\alpha_t}{(1-\alpha_t)}\Big(\big\|x_{t-1}-\mu_t\big\|_2^2-\big\|x_{t-1}-\mu^\star_t\big\|_2^2\Big)\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&=\underbrace{\int p_{X_t}(x_t)p_{X_{t-1}\mid X_t}(x_{t-1}\mid x_t)p_{Z_t\mid H_{t-1},H_t}(z_t\mid x_{t-1},x_t)\frac{\alpha_t}{(1-\alpha_t)}\big\|\mu^\star_t-\mu_t\big\|_2^2\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t}_{\mathcal{H}_1}
\end{aligned}
$$

$$+\underbrace{\int p_{X_t}(x_t)p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)p_{Z_t\,|\,H_{t-1},H_t}(z_t\,|\,x_{t-1},x_t)\frac{2\alpha_t}{(1-\alpha_t)}(\mu_t^\star-\mu_t)^\top\big(x_{t-1}-\mu_t^\star\big)\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t}_{\mathcal{H}_2}.$$

Here, (i) follows the property of KL divergence that

$$\int p_{Z_t\,|\,H_{t-1},H_t}(z_t\,|\,x_{t-1},x_t)\log\frac{p_{Z_t\,|\,H_{t-1},H_t}(z_t\,|\,x_{t-1},x_t)}{p_{Z_t\,|\,Y_{t-1},Y_t}(z_t\,|\,x_{t-1},x_t)}\mathrm{d}z_t\geq 0,$$

whereas (ii) results from the following expressions:

$$p_{H_{t-1}\,|\,H_t,Z_t}(x_{t-1}\,|\,x_t,z_t)\propto\exp\Big(-\frac{\alpha_t}{(1-\alpha_t)}\Big\|(x_{t-1}-\mu_t^\star(x_t))\Big\|^2\Big)$$
$$p_{Y_{t-1}\,|\,Y_t,Z_t}(x_{t-1}\,|\,x_t,z_t)\propto\exp\Big(-\frac{\alpha_t}{(1-\alpha_t)}\Big\|(x_{t-1}-\mu_t(x_t))\Big\|^2\Big).$$

To bound $\mathcal{H}_1$, we first note that

$$\begin{aligned}
&\frac{1}{1-\alpha_t}\big\|\mu_t-\mu_t^\star\big\|_2^2=(1-\alpha_t)\cdot\\
&\quad\Big\|s_t\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)-s_t^\star\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)+s_t^\star\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)-s_t^\star(x_t)+\frac{(1-\alpha_t)^{1/2}}{\sqrt{2}(1-\overline{\alpha}_t)}J_t(x_t)z_t\Big\|_2^2\\
&\leq(1-\alpha_t)\Big\|s_t\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)-s_t^\star\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)\Big\|_2^2+\\
&\quad(1-\alpha_t)\Big\|s_t^\star\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)-s_t^\star(x_t)+\frac{(1-\alpha_t)^{1/2}}{\sqrt{2}(1-\overline{\alpha}_t)}J_t(x_t)z_t\Big\|_2^2\\
&\lesssim\frac{\log T}{T}\varepsilon_{\mathsf{score},t}\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)^2+\frac{d^5\log^7T}{T^3}
\end{aligned}$$

where the last inequality follows from the definition (40), the relation (13b), and the fact that

$$\begin{aligned}
&(1-\alpha_t)\Big\|s_t^\star\Big(x_t+\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)-s_t^\star(x_t)+\frac{(1-\alpha_t)^{1/2}}{\sqrt{2}(1-\overline{\alpha}_t)}J_t(x_t)z_t\Big\|_2^2\\
&=\frac{(1-\alpha_t)^2}{2(1-\overline{\alpha}_t)^2}\Big\|\int_0^1\Big(J_t(x_t)-J_t\Big(x_t+\gamma\sqrt{\frac{1-\alpha_t}{2}}z_t\Big)\Big)z_t\mathrm{d}\gamma\Big\|_2^2\lesssim\frac{d^5\log^7T}{T^3}.
\end{aligned}$$

Here, the last inequality holds by invoking the property (86c) that for $(x,x_{t-1})\in\mathcal{E}$,

$$\|J_t(x)-J_t(x_t)\|\leq\sup_{u\in\mathbb{S}^{d-1}}|u^\top(J_t(x)-J_t(x_t))u|\lesssim d^{3/2}\|x-x_t\|_2\log^{3/2}T. \tag{122}$$

For the case with $(x,x_{t-1})\notin\mathcal{E}$, this term will decay exponentially fast and can be bounded analogously. Furthermore, we observe that

$$p_{\Phi_t(X_t,Z_t)}(x)=\big(\pi(2(1-\overline{\alpha}_t)+1-\alpha_t)\big)^{-d/2}\int p_{X_0}(x_0)\exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|_2^2}{2(1-\overline{\alpha}_t)+1-\alpha_t}\Big)\mathrm{d}x_0\asymp p_{X_t}(x),$$

which in turn implies that

$$\begin{aligned}
\mathcal{H}_1&\leq\Big(1+O\Big(\frac{d^3\log^{4.5}T}{T^{3/2}}\Big)\Big)\int p_{X_t}(x_t)p_{H_{t-1}\,|\,H_t,Z_t}(x_{t-1}\,|\,x_t,z_t)p_{Z_t}(z_t)\frac{\alpha_t}{1-\alpha_t}\big\|\mu_t-\mu_t^\star\big\|_2^2\mathrm{d}x_t\mathrm{d}x_{t-1}\mathrm{d}z_t\\
&\lesssim\mathop{\mathbb{E}}_{x^+\sim\Phi_t(X_t,Z_t)}\Big[\frac{\log T}{T}\varepsilon_{\mathsf{score},t}(x^+)^2\Big]+\frac{d^5\log^7T}{T^3}
\end{aligned}$$

$$\asymp \frac{d\log^3 T}{T}\mathop{\mathbb{E}}_{X_t\sim q_t}\left[\varepsilon_{\mathsf{score},t}(X_t)^2\right]+\frac{d^5\log^7 T}{T^3}$$

We then decompose $\mathcal{H}_2$ as follows

$$\begin{aligned}
\mathcal{H}_2 &= \int p_{X_t}(x_t)\big(p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)-p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)\big)p_{Z_t\,|\,H_{t-1},H_t}(z_t\,|\,x_{t-1},x_t)\\
&\qquad\cdot\frac{2\alpha_t}{(1-\alpha_t)}(\mu_t^\star-\mu_t)^\top\big(x_{t-1}-\mu_t^\star\big)\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&\quad+\int p_{X_t}(x_t)p_{H_{t-1}\,|\,H_t,Z_t}(x_{t-1}\,|\,x_t,z_t)p_{Z_t}(z_t)\cdot\frac{2\alpha_t}{(1-\alpha_t)}(\mu_t^\star-\mu_t)^\top\big(x_{t-1}-\mu_t^\star\big)\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&\overset{\text{(i)}}{=}\int p_{X_t}(x_t)\big(p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)-p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)\big)p_{Z_t\,|\,H_{t-1},H_t}(z_t\,|\,x_{t-1},x_t)\\
&\qquad\cdot\frac{2\alpha_t}{(1-\alpha_t)}(\mu_t^\star-\mu_t)^\top\big(x_{t-1}-\mu_t^\star\big)\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&=\left(\int_{\mathcal{E}}+\int_{\mathcal{E}^c}\right)p_{X_t}(x_t)\big(p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)-p_{H_{t-1}\,|\,H_t}(x_{t-1}\,|\,x_t)\big)p_{Z_t\,|\,H_{t-1},H_t}(z_t\,|\,x_{t-1},x_t)\\
&\qquad\cdot\frac{2\alpha_t}{(1-\alpha_t)}(\mu_t^\star-\mu_t)^\top\big(x_{t-1}-\mu_t^\star\big)\mathrm{d}z_t\mathrm{d}x_{t-1}\mathrm{d}x_t,
\end{aligned}$$

where (i) follows the fact that $\mathbb{E}[H_{t-1}-\mu_t^*|H_t,Z_t]=0$. In the following, we mainly focus on the term $\int_{\mathcal{E}}$ denoted as $\mathcal{K}_1$, since the other term can be bounded similarly as (Li et al., 2023, Lemma 10) and is exponentially small.

$$\begin{aligned}
\mathcal{K}_1 &\overset{\text{(i)}}{\lesssim}\frac{d^3\log^{4.5}T}{T^{3/2}}\int_{\mathcal{E}}p_{X_t}(x_t)p_{H_{t-1}\,|\,H_t,Z_t}(x_{t-1}\,|\,x_t,z_t)P_{Z_t}(z_t)\big\|x_{t-1}-\mu_t^\star\big\|_2\frac{1}{1-\alpha_t}\big\|\mu_t-\mu_t^\star\big\|_2\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&\overset{\text{(ii)}}{\lesssim}\frac{d^3\log^{4.5}T}{T^{3/2}}\sqrt{\mathcal{K}_2\mathcal{K}_3}.
\end{aligned}\tag{123}$$

Here, we have

$$\begin{aligned}
\mathcal{K}_2&=\int_{\mathcal{E}}p_{X_t}(x_t)p_{H_{t-1}\,|\,H_t,Z_t}(x_{t-1}\,|\,x_t,z_t)P_{Z_t}(z_t)\big\|x_{t-1}-\mu_t^\star\big\|_2^2\mathrm{d}x_{t-1}\mathrm{d}x_t\mathrm{d}z_t\\
&\le\frac{d(1-\alpha_t)}{\alpha_t}\lesssim\frac{d\log T}{T};\\
\mathcal{K}_3&=\int_{\mathcal{E}}p_{X_t}(x_t)p_{H_{t-1}\,|\,H_t,Z_t}(x_{t-1}\,|\,x_t,z_t)P_{Z_t}(z_t)\frac{1}{(1-\alpha_t)^2}\big\|\mu_t-\mu_t^\star\big\|_2^2\mathrm{d}x_{t-1}\mathrm{d}x_t\mathrm{d}z_t\\
&\lesssim\mathop{\mathbb{E}}_{X_t\sim q_t}\left[\varepsilon_{\text{score},t}(X_t)^2\right]+\frac{d^5\log^6 T}{T^2}.
\end{aligned}$$

Therefore, we arrive at

$$\mathcal{K}_1\lesssim\frac{d^{3.5}\log^5 T}{T^2}\mathop{\mathbb{E}}_{X_t\sim q_t}\left[\varepsilon_{\text{score},t}(X_t)^2\right]+\frac{d^6\log^8 T}{T^3}.$$

Taking the above bounds on $\mathcal{H}_1$ and $\mathcal{K}_1$ together completes the proof.