Towards a mathematical theory for consistency training in diffusion models

Gen Li*
The Chinese University of Hong Kong

Zhihan Huang*
University of Pennsylvania

Yuting Wei
University of Pennsylvania

Abstract

Consistency models, which were proposed to mitigate the high computational overhead during the sampling phase of diffusion models, facilitate single-step sampling while attaining state-of-the-art empirical performance. When integrated into the training phase, consistency models attempt to train a sequence of consistency functions capable of mapping any point at any time step of the diffusion process to its starting point. Despite the empirical success, a comprehensive theoretical understanding of consistency training remains elusive. This paper takes a first step towards establishing theoretical underpinnings for consistency models. We demonstrate that, in order to generate samples within ε proximity to the target in distribution (measured by some Wasserstein metric), it suffices for the number of steps in consistency learning to exceed the order of $d^{5/2}/\varepsilon$, with d the data dimension. Our theory offers rigorous insights into the validity and efficacy of consistency models, illuminating their utility in downstream inference tasks.

1 Introduction

Diffusion models (Sohl-Dickstein et al. (2015); Song and Ermon (2019); Ho et al. (2020)) have garnered growing interest in recent years due to their impressive capabilities in a wide swath of generative modeling tasks, such as image synthesis, video generation, and audio synthesis (Dhariwal and Nichol (2021); Ramesh et al. (2022); Rombach et al. (2022); Kong et al. (2020); Ho et al. (2022); Popov et al. (2021)). In comparison with other deep generative models, such as Generative Adversarial Networks, which oftentimes suffer from training instability and mode collapse, diffusion models are capable of generating high-fidelity samples based on learning the gradient of the log-density function or the score function. On a high level, diffusion models concentrate on two processes: a forward Markov process that gradually degrades data into noise, and a reverse-time stochastic or deterministic process that starts from pure noise, performs iterative denoising to generate new data that resemble true data samples in distribution. Interestingly, while the forward process is often straightforwardly designed by progressively injecting more noise into the data samples, it is feasible to revert the process and ensure (almost) matching marginals as the forward process, as long as faithful score function estimates are obtainable (Anderson (1982); Haussmann and Pardoux (1986)).

Nevertheless, given that diffusion models generate new data by implementing a sequence of steps in the reverse process (with each step computing the score function by evaluating a large neural network), they often incur substantially higher computational cost compared to other single-step generative modeling algorithms, thereby limiting their sampling efficiency in real-time applications. To remedy this issue, there has been an explosion of efforts in developing acceleration procedures to speed up the sampling process in diffusion generative modeling (e.g. Song and Ermon (2020); Lu et al. (2022a,b); Zhao et al. (2023); Zhang and Chen (2022); Xue et al. (2023); Luhman and Luhman (2021); Salimans and Ho (2022); Song et al. (2023); Li et al. (2024)). Among these efforts, training-based methods, exemplified by progressive distillation and consistency models hold great promises in producing samplers that are computationally efficient and ready for real-time implementation without sacrificing sampling fidelity (Salimans and Ho (2022); Meng et al. (2023); Sun et al. (2023); Song et al. (2023)).

In this paper, our focal point is the consistency model, which was originally proposed by Song et al. (2023) and claims the state-of-the-art performance. In a nut-

*The first two authors contributed equally. Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

shell, the consistency model seeks to learn a function that is able to map any point at any time step of the diffusion process to the process' starting point (the end corresponding to the data distribution). In the sampling phase, the consistency model enables sample generation with only a single evaluation of the neural network. The surprising efficacy of consistency models has been demonstrated in various image datasets, including CIFAR-10, ImageNet 64×64, LSUN 256×256, and also video generation (Song et al. (2023); Wang et al. (2023)), to name just a few. This approach has received considerable recent attention, covering various extensions (e.g. Song and Dhariwal (2023); Kim et al. (2023)) as well as applications beyond generative models (e.g. reinforcement learning Ding and Jin (2023)).

Despite the aforementioned mind-blowing empirical successes, however, a theoretical understanding of consistency models remains elusive even in the most basic setting. In light of the flexibility and versatility of the consistency model idea (which only requires enforcing some self-consistency conditions), establishing theoretical underpinnings for these models not only provides rigorous justifications for their validity, but also yields practical implications in downstream inference tasks by providing theoretical benchmarks to compare different training strategies. However, the challenge in establishing theoretical performance guarantees lies in understanding the role of consistency enforcement in preserving the sampling fidelity.

An overview of our contributions. In this paper, we take a first step towards establishing theoretical support for consistency models, focusing on consistency training (namely, applying the consistency model idea from the training stage). More specifically, we consider a consistency training paradigm that recursively learns a sequence of functions $\{f_t\}_{1\leq t\leq T}$, in the hope that the ultimate sampling process can be readily completed by evaluating $f_T(X_T)$ with $X_T \sim \mathcal{N}(0, I_d)$. Our theory reveals that: it is sufficient for consistency training to take a number of steps exceeding the order of

$$\frac{d^{5/2}}{\varepsilon} \tag{1}$$

up to some logarithmic factor in order to generate samples that are 2ε close in distribution to the target data distribution (measured by the Wasserstein metric). Here, d denotes the dimension of the target distribution, and we omit the logarithm factors and dependence on other universal constants. In other words, it tells us how many steps need to be included in the training stage in order to enable one-shot sampling that achieves the desirable sampling fidelity.

Notation. We introduce a couple of notation to be used throughout this paper. Given two probability measures μ and ν on $\mathbb{R}^d$, we denote by $\mathcal{C}(\mu,\nu)$ the set of all couplings of μ and ν (i.e., all joint distributions $\gamma(x,y)$ whose marginals coincide with μ and ν, respectively). The Wasserstein distance of order q between these two distributions is defined as

$$W_q(\mu,\nu) := \left(\inf_{\gamma\in\mathcal{C}(\mu,\nu)} \mathop{\mathbb{E}}_{(x,y)\sim\gamma} \left[\|x-y\|_2^q \right] \right)^{1/q}, \tag{2}$$

and we often employ $W_q(X,Y)$ for random variables X and Y to denote the Wasserstein distance between distributions of X and Y. In addition, given any two functions $f(d,T)$ and $g(d,T)$, we write $f(d,T) \lesssim g(d,T)$ or $f(d,T) = O(g(d,T))$ (resp. $f(d,T) \gtrsim g(d,T)$) if there exists some universal constant $C_1 > 0$ such that $f(d,T) \leq C_1 g(d,T)$ (resp. $f(d,T) \geq C_1 g(d,T)$) for all d and T. Furthermore, the notation $\widetilde{O}(\cdot)$ is defined analogously to $O(\cdot)$ except that the logarithmic dependency is hidden. Given a matrix $M \in \mathbb{R}^{d\times d}$, we denote $\|M\|$ as the operator norm of M.

2 Preliminaries

In this section, we introduce the basics of diffusion generative modeling and consistency models. While the consistency model was originally motivated to accelerate the probability flow ODE sampler and distill information from a pre-trained model, the idea of promoting consistency along the trajectory can be incorporated directly into the training stage, which we focus on in this paper.

2.1 Diffusion-based generative models

Forward process. As briefly mentioned above, in diffusion generative models, one starts from a forward process and progressively perturbs the data into pure noise, where the noise distribution is often chosen to be Gaussian. The forward process is often modeled as solution to an Itô stochastic differential equation (SDE)

$$\mathrm{d}X_t = f(X_t,t)\mathrm{d}t + g(t)\mathrm{d}W_t \qquad (0 \leq t \leq T), \tag{3}$$

where W_t corresponds to a standard Brownian motion, $f(\cdot,t) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector-valued function that determines the drift of this process, and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a function that adjusts the variance of the injected noise. We shall adopt the notation $q_t := \mathsf{Law}(X_t)$ throughout to represent the distribution of X_t in this forward process. In particular, $q_0 := \mathsf{Law}(X_0)$ is our target distribution to generate samples from, and it is also frequently denoted by p_{data}. A popular special case that motivates DDPM and DDIM algorithms (Song et al.

(2020); Ho et al. (2020); Nichol and Dhariwal (2021)) is to take $f(X_t, t) = -\frac{1}{2}\beta(t)X_t$ and $g(t) = \sqrt{\beta(t)}$ for some function $\beta(\cdot)$ (which can be interpreted as determining the learning rate schedule). The SDE defined above then reduces to

$$\begin{aligned} &X_0 \sim p_{\mathsf{data}}, \\ &\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}W_t \end{aligned} \tag{4}$$

for any $0 \leq t \leq T$.

Given the continuous-time nature of the above forward process, it would oftentimes be helpful to look at the discrete-time counterpart instead. More specifically, consider the following discrete-time random process:

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

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

with T representing the total number of steps. Here, we denote by $\{\beta_t\} \subseteq (0, 1)$ the prescribed learning rates that control the strength of the noise injected at each step, and $\{W_t\}_{1\leq t\leq T}$ a sequence of independent noise vectors drawn from $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. If we further define α_t and $\overline{\alpha}_t$ such that

$$\alpha_t = 1 - \beta_t, \qquad \overline{\alpha}_t/\overline{\alpha}_{t-1} = \alpha_t, \qquad 1 \leq t \leq T, \tag{6}$$

one can write

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

In practice, $\overline{\alpha}_T$ is oftentimes chosen to be vanishingly small (as long as T is large enough), so as to make sure that the distribution q_T of X_T is approximately $\mathcal{N}(0, I_d)$.

Reverse process. Reversing the above process in time leads to a process that transforms noise into samples with distribution approximately equal to p_{data}, which is how diffusion models generate data.

A popular sampler, called the Denoising Diffusion Implicit Model (DDIM) Karras et al. (2022); Song et al. (2020, 2021), leverages upon the so-called probability flow ODE. More precisely, consider the following ODE famliy

$$\begin{aligned} &Y_0^{\mathsf{ode}} \sim q_T, \\ &\mathrm{d}Y_t^{\mathsf{ode}} = \frac{1}{2}\beta(T-t)\Big(Y_t^{\mathsf{ode}} + \nabla \log q_{T-t}(Y_t^{\mathsf{ode}})\Big)\mathrm{d}t \end{aligned} \tag{8}$$

for all $0 \leq t \leq T$, which again yields matching marginal distributions for X_t as

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

Evidently, to implement such a process, it requires obtaining faithful estimates of the score function [1]

$$s_t(X) := \nabla_X \log q_t(X). \tag{9}$$

It is noteworthy that this deterministic ODE-based approach is often faster than the SDE-based approach (Song et al. (2021)), which has also been justified in theory (Li et al. (2023)).

We note that the probability flow ODE considered here in (8) is slightly different from the one in Song et al. (2023), the latter of which takes the form

$$\mathrm{d}Y_t = -t\nabla \log q_t(Y_t)\mathrm{d}t \tag{10}$$

and corresponds to the forward process $\mathrm{d}X_t = \sqrt{2t}\mathrm{d}W_t$. In particular, if the covariance of X_0 is equal to I_d, then q_T is close to a Gaussian distribution $\mathcal{N}(0, T^2 I_d)$ (so that the covariance explodes), whereas in the process (7), the covariance for X_t is preserved and equals I_d throughout the trajectory.

Finally, we note that recent years have witnessed remarkable theoretical advances towards understanding the sampling performance of diffusion models. A highly incomplete list includes Block et al. (2020); De Bortoli et al. (2021); Liu et al. (2022); De Bortoli (2022); Lee et al. (2023); Pidstrigach (2022); Chen et al. (2022b); Tang (2023); Benton et al. (2023a); Chen et al. (2022a, 2023b); Tang and Zhao (2024); Li et al. (2023). In particular, the recent works Chen et al. (2022b,a); Benton et al. (2023a); Chen et al. (2023b,a); Li et al. (2023) have established the convergence rates of both the DDPM and DDIM samplers, as well as their stability against ℓ_2 score estimation errors.

2.2 Consistency training

While the probability flow ODE approach already achieves much faster sampling compared to the DDPM sampler, it still requires a large number of steps (or equivalently, a large number of neural network evaluations) and does not yet meet the demand for real-time sample generation. This motivates the development of the consistency model as a means to accomplish sampling in one step (Song et al. (2023)).

Specifically, given a solution trajectory $\{x_t\}_{t\in[\epsilon,T]}$ of the probability ODE in (8), a consistency function is a parametrized function (parameterized by θ) designed to achieve

$$f_\theta : (x_t, t) \xrightarrow{\text{ideally}} x_\epsilon \qquad \text{for all } t \in [\epsilon, T], \tag{11}$$

[1] For notational convenience, we also adopt the shorthand notation $\nabla \log q_t(X)$ to denote the score function (by suppressing the dependency on X).

which maps a point x_t at time t back to the desired sample x_ϵ. Therefore, given a well-trained consistency model f_θ, in the sampling phase, instead of recursively applying denoising function $p_\theta(x_{t-1} \mid x_t)$ as the reverse diffusion process in diffusion model, it suffices to evaluate $f_\theta(\widehat{x}_T, T)$ once to produce an approximation of $\widehat{x}_\epsilon$. By doing so, one forward pass through the consistency model (or one evaluation of the neural network) suffices to generate a sample that mimics the target distribution.

When the consistency approach is integrated into the training phase, it entails an iterative procedure to find suitable parameterization θ. More specifically, the idea put forward by Song et al. (2023) is to minimize a certain consistency training objective over the parameter θ where

$$\begin{aligned} \underset{\theta}{\text{minimize}} \;\; & \mathcal{L}_{\theta^-}(\theta) \\ := \mathbb{E}\Big[& \lambda(t_n) \cdot \mathsf{dist}\big(f_\theta(X_0 + t_{n+1}Z, t_{n+1}), \\ & f_{\theta^-}(X_0 + t_n Z, t_n)\big)\Big], \end{aligned} \tag{12}$$

where the time horizon $[\epsilon, T]$ is discretized into $N-1$ sub-intervals, with boundaries $t_1 = \epsilon < t_2 < \ldots < t_N = T$.[2] Here, $\mathsf{dist}(\cdot,\cdot)$ is some distance measure between two vectors in d dimension, $\lambda(\cdot)$ is some weighting function, and θ^- is some moving average of θ during the course of training. The expectation is taken over $X_0 \sim p_{\mathsf{data}}$, $Z \sim \mathcal{N}(0, I_d)$ and n drawn uniformly from $\{0, 1, \ldots, N\}$.

3 A non-asymptotic convergence theory for consistency training

3.1 Assumptions and setup

Before delving into our main results, let us introduce several notation and terminologies. To simplify the analysis, in the following maniscript, we use $\overline{\alpha}_t = \exp\left(-\int_0^t \beta(t)\mathrm{d}t\right)$ to be the noise level in the forawrd process at time t, where $\beta(t)$ is coefficient in SDE (4)(Ho et al., 2020). For any $0 < \overline{\alpha} < 1$, we denote

$$X(\overline{\alpha}) = \sqrt{\overline{\alpha}}X_0 + \sqrt{1-\overline{\alpha}}Z, \tag{13}$$

$$s_{\overline{\alpha}}(x) = \nabla_x \log p_{X(\overline{\alpha})}(x), \tag{14}$$

to be primal variable and score function at noise level $\overline{\alpha}$, where $X_0 \sim p_{\mathsf{data}}$ and $Z \sim \mathcal{N}(0, I_d)$. By this change of variables, compared with X_t in SDE (4), we have

$$X_t \overset{d}{=} X(\overline{\alpha}_t), \quad s_t(x) := s_{\overline{\alpha}_t}(x),$$

which enable us to use $\overline{\alpha}$ as the index variable which amounts to use t.

Also, we denote $\Phi_{t\to k}(x)$ to be the trajectory of the probability flow ODE (8) from time t to k with the initial condition $X_t = x$. Thus, by the property of the probability flow ODE, we have $\Phi_{t\to k}(X_t) \overset{\mathsf{d}}{=} X_k$. The schedule of $\beta(t)$ in (8) is to be determined. We refer our readers to (25) for the specific schedule used in the analysis.

With these definitions in mind, the probability flow ODE can be written in the following equivalent form:

$$\Phi_{t\to k}(x) = g_t(x, \overline{\alpha}_k), \tag{15}$$

where g_t is defined as the solution to the following PDE

$$\frac{\partial g_t(x,\overline{\alpha})}{\partial \overline{\alpha}} = \frac{1}{2\overline{\alpha}}\Big(g_t(x,\overline{\alpha}) + s_{\overline{\alpha}}(g_t(x,\overline{\alpha}))\Big), \tag{16}$$

and the boundary condition $g_t(x, \overline{\alpha}_t) = x$. In fact, the relatioship (15) can be directly derived from $\overline{\alpha}_k = \exp\left(-\int_0^k \beta(t)\mathrm{d}t\right)$ and change of varibles formula. Further, property (16) holds true for the entire ODE trajectory in the sense that for any $0 < \overline{\alpha} \le \overline{\alpha}_t$, we have $g_t(X_t, \overline{\alpha}) \overset{\mathsf{d}}{=} X(\overline{\alpha})$.

For notational simplicity, we shall denote

$$\phi_t := \Phi_{t\to t-1} \quad \text{and} \quad \Phi_t := \Phi_{t\to 1}, \tag{17}$$

and it satisfies

$$\begin{aligned} \Phi_{t\to k}(x) &= \Phi_{k+1\to k} \circ \cdots \Phi_{t-1\to t-2} \circ \Phi_{t\to t-1}(x) \\ &= \phi_{k+1}(\phi_{k+2}(\cdots \phi_t(x)\cdots)). \end{aligned}$$

The parameter θ are trained according to (12) without a pre-trained backward process. Our main result is established under the following two assumptions.

Assumption 3.1. Assume that for $1 \le k < t \le T$, $\Phi_{t\to k}$ is L_f-Lipschitz continuous such that

$$\big\|\Phi_{t\to k}(x) - \Phi_{t\to k}(y)\big\|_2 \le L_f \|x - y\|_2. \tag{18}$$

Assumption 3.2. Suppose there exists $\varepsilon := \sum_t \varepsilon_t > 0$ such that the parameter θ trained according to (12)[3] converges to some $\widehat{\theta}$ satisfying

$$\begin{aligned} \mathbb{E}\Big[\big\| & f_{\widehat{\theta}}(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z, t) \\ & - f^\star_{\widehat{\theta}}(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z, t)\big\|_2^2\Big] \le \varepsilon_t^2, \end{aligned} \tag{19}$$

where

$$\begin{aligned} f^\star_{\widehat{\theta}} = \arg\min_f \;\; & \mathbb{E}\Big[\big\|f(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z, t) \\ & - f_{\widehat{\theta}}(\sqrt{\overline{\alpha}_{t-1}}X_0 + \sqrt{1-\overline{\alpha}_{t-1}}Z, t-1)\big\|_2^2\Big]. \end{aligned} \tag{20}$$

[2] The exact formulas for $\{t_i\}_{1\le i\le N}$ can be found in Song et al. (2023).

[3] Here, we consider the ℓ_2 distance and an equivalent scaling regime, i.e., $t_n = \sqrt{\overline{\alpha}_t^{-1} - 1}$.

In words, Assumption 3.1 requires the mappings from X_t to X_k to be Lipschitz continuous for every k and t. By the definition of $\Phi_{t\to k}$, this assumption also ensures that the score function s_t remains Lipschitz continuous throughout the forward process. Intuitively, this Lipschitz condition helps to control how error propagates with time as we learn the consistency functions. Without this condition, a one-step sampler might fail to generate high-quality samples. We can further relax this assumption to be held in the averaged sense instead of point-wisely. More specifically, if we replace Assumption 3.1 with the following condition

$$\mathbb{E}\big[\|\nabla\Phi_{t\to k}(x)\|^3\big] \le L_f^3, \tag{21}$$

and all the analysis in Appendix A still hold.

Assumption 3.2 is concerned with the estimation error of $f_{\widehat{\theta}}$ relative to $f^\star_{\widehat{\theta}}$. In the following, for ease of presentation, we shall let

$$f_t(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z) := f_{\widehat{\theta}}(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z, t).$$

Then the above assumption means that the functions $\{f_t\}_{1\le t\le T}$ satisfy

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big\|f_t(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z) - f_t^\star(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z)\big\|_2\Big] \le \varepsilon,$$

where

$$f_t^\star := \mathbb{E}\Big[f_{t-1}\big(\sqrt{\overline{\alpha}_{t-1}}X_0+\sqrt{1-\overline{\alpha}_{t-1}}Z\big)\,\Big|\,\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z\Big].$$

More generally, if one is only allowed to optimize over a specific function class $\mathcal{F}$, let the optimal solution within that class as

$$f_t^{\mathcal{F}} := \arg\min_{f\in\mathcal{F}}\mathbb{E}\Big[\big\|f(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z) - f_{t-1}(\sqrt{\overline{\alpha}_{t-1}}X_0+\sqrt{1-\overline{\alpha}_{t-1}}Z)\big\|_2^2\Big]. \tag{22}$$

Assumption 3.2 can be further relaxed to asking

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big\|f_t(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z) - f_t^{\mathcal{F}}(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z)\big\|_2\Big] \le \varepsilon,$$

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big\|f_t^{\mathcal{F}}(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z) - f_t^\star(\sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z)\big\|_2\Big] \le \varepsilon_{\mathcal{F}},$$

for quantities $\varepsilon, \varepsilon_{\mathcal{F}} \ge 0$. We will establish our main results based on this more general assumption. When $\varepsilon_{\mathcal{F}} = 0$, our results reduces to the case when the original form of Assumption 3.2 holds. In addition, it can be seen that Assumption 3.2 involves two sources of errors in the training process: (i) ε controls the estimation error of the consistency functions $\{f_t\}_{1\le t\le T}$ we have obtained. As the training sample size n increases, suitable optimization methods converge to the best approximation in $\mathcal{F}$, thus $\varepsilon \to 0$; (ii) $\varepsilon_{\mathcal{F}}$ corresponds to the approximation error of restricting the consistency functions to lie within some fixed function class $\mathcal{F}$, where $\varepsilon_{\mathcal{F}} = 0$ for function classes with large capacity (or representation power) like neural networks. We remark that the optimization step (12) for obtaining $\{f_t\}_{1\le t\le T}$ is typically accomplished through proper training of large neural networks. Given the complexity of developing an end-to-end theory, we adopt the common divide-and-conquer strategy and decouple the training phase with the sampling phase. In the sequel, we shall focus our attention on quantifying the sampling fidelity, assuming small estimation/optimization errors in the training phase.

Target data distribution. To streamline our main proof, we impose an additional constraint on the target data distribution, namely,

$$\mathbb{P}(\|X_0\|_2 \le T^{c_R}) = 1, \tag{24}$$

where $X_0 \sim p_{\mathsf{data}}$ and $c_R > 0$ is some arbitrarily large constant. This assumption covers a broad family of data distribution with polynomially large support size. We remark that this constraint can be replaced by some careful assumptions on the tail probability of the target data distribution, and the resulting proof is expected to be similar.

Learning rate schedule. Finally, let us specify the choice of $\beta(t)$ and corresponding learning rate schedule $\{\overline{\alpha}_t\}_{1\le t\le T}$ we would like to employ during consistency training (12). We choose $\beta(t)$ in SDE (4) and (8) such that they satisfies for some large enough numerical constants $c_0, c_1 > 0$,

$$\overline{\alpha}_1 = \alpha_1 = 1 - \frac{1}{T^{c_0}}; \tag{25a}$$

$$\alpha_t = 1 - \frac{c_1\log T}{T}\min\Big\{\beta_1\Big(1+\frac{c_1\log T}{T}\Big)^t, 1\Big\}; \tag{25b}$$

$$\overline{\alpha}_t = \prod_{i=1}^{t}\alpha_i, \qquad t = 2,\dots,T. \tag{25c}$$

Note that we only need to specify $\overline{\alpha}_t$ for $t \in \mathbb{Z}^+$ (since the consistency function is trained on samples from dicretized time points), and $\beta(t)$ can be chosen to be any continuous function that satisfies the above conditions.

Such scheduled learning rates have been employed in the prior work Li et al. (2023) to achieve the desired convergence guarantees and are similar to what is used in practice, as suggested by Song et al. (2023). A couple of other useful properties about these learning rates are provided in Appendix A.

3.2 Main results

We are now positioned to state our main theoretical guarantees for consistency training.

Theorem 3.3. *Suppose the learning rates are selected according to* (6) *and the target distribution satisfies property* (24). *Under Assumptions 3.1 and 3.2, it obeys*

$$\mathcal{W}_1\big(f_T(X_T), X_1\big) < C_1 \frac{L_f^3 d^{5/2} \log^5 T}{T} + \varepsilon + \varepsilon_{\mathcal{F}} \quad (26)$$

for some universal constant $C_1 > 0$.

In Theorem 3.3, we characterize the convergence of backward process starting from X_T, the noisy version of the initial distribution, which corresponds to the multistep sampling procedure in Song et al. (2023). When running the backward process from the pure noise, we have the following result by similar analysis. Let $\overline{Z}$ be a truncated Gaussian random noise in $\mathbb{R}^d$, where $p_{\overline{Z}}(x) \propto p_Z(x) 1(\|x\|_2 \le T^c)$ for $Z \sim \mathcal{N}(0, I_d)$ and some constant $c > 0$ large enough. Notice that since c is large enough, this truncation makes exponentially small difference in the sampling process (By Gaussian tail probability, we have $\mathbb{P}(\|Z\|_2 > T^c) \le \exp(-\Omega(T^{2c}))$. Based on our main theorem above, we have the following result immediately:

Corollary 3.4. *Under the same setting as Theorem 3.3, the backward process starting from* $\overline{Z}$ *shares the same convergence rate as in Theorem 3.3:*

$$\mathcal{W}_1(f_T(\overline{Z}), X_1) < C_2 \frac{L_f^3 d^{5/2} \log^5 T}{T} + \varepsilon + \varepsilon_{\mathcal{F}}. \quad (27)$$

for some universal constant $C_2 > 0$.

We note the high probability statement used here is mainly from the analysis in Wasserstein metric, and can be removed, i.e. replace $\overline{Z}$ by Z in the above corollary by adopting more assumptions, e.g., the Lipchitzness assumption in Theorem 1 in Song et al. (2023). To illustrate our result, we demonstrate in Figure 1 the convergence of the consistency model when the target distribution is a 10-dimensional Gaussian distribution.[4]

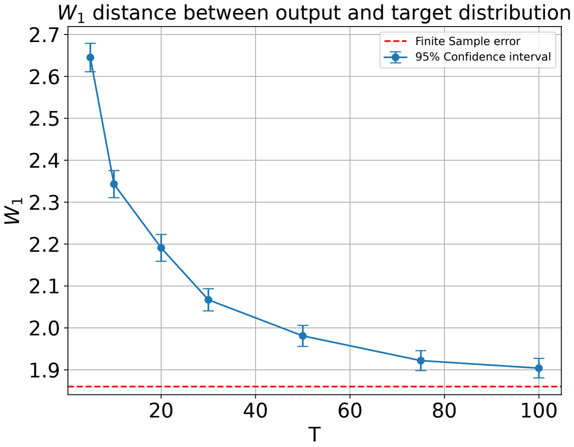

Figure 1: The convergence of consistency models for learning a 10-dimensional heterogeneous Gaussian distribution. The W_1 distance between our output and the target distribution is estimated by 2000 samples from both distributions and the confidence intervals are obtained from 200 repetitions. The red dash line indicates that finite-sample error from consistency function training and W_1 distance estimation.

To better understand this result, we provide some remarks below in order.

- In our results, we consider the convergence towards X_1 instead of X_0 to avoid the explosion of the score functions, which is a standard practice in the theory and practical training of diffusion models and appears across diffusion models literature (see, e.g. Benton et al. (2023a); Li et al. (2023) among others). Notice that since the step size is chosen to be exponentially small, the distribution of X_1 is exponentially close to X_0. More specifically, it can be easily shown that
$$\mathcal{W}_1(X_1, X_0) < \frac{1}{\mathsf{poly}(T)},$$
which implies the convergence towards target data distribution X_0 combining with the triangle inequality.
- Theorem 3.3 implies that, in order to achieve a sampling error of $2(\varepsilon + \varepsilon_{\mathcal{F}})$—in the sense that $\mathcal{W}_1\big(f_T(X_T), X_1\big) \le 2(\varepsilon + \varepsilon_{\mathcal{F}})$—it is sufficient for the number of steps in consistency training to exceed
$$\widetilde{O}\Big(\frac{L_f^3 d^{5/2}}{\varepsilon + \varepsilon_{\mathcal{F}}}\Big), \quad (28)$$
where d denotes the dimension of the target distribution. In particular, if the function class $\mathcal{F}$

[4] We estimate the W_1 distance between our output and the target distribution based on samples, as there is no close-form formula to compute the W_1 distance for high-dimensional distributions.

is rich enough and the approximation error $\varepsilon_{\mathcal{F}}$ equals zero, then the number of steps required is about the order of $\frac{L_f^3 d^{5/2}}{\varepsilon}$. This result offers an explicit characterization of the dependence of the Lipschitz constant as well as the dimension of the problem in the worst case scenario. In real applications, the target distribution often enjoys some low-dimensional structures, which can lead to a better performance than the worst-case guarantee (see, e.g. Li and Yan (2024); Huang et al. (2024)). As far as we are aware, this is the first result that theoretically measures the sampling fidelity of consistency models, which serve as a theoretical justification for consistency models as a family of generative models. As alluded to previously, compared to the popular diffusion models, consistency models bear the benefit of one-step sampling, requiring only a single function evaluation at the sampling stage instead of undergoing recursive denoising. Consequently, our theoretical result provides insights into when one-step sampling is reliable.

- We remark that prior results concerning convergence guarantees for diffusion models mostly consider weaker metrics such as the TV distance or KL divergence (e.g., Chen et al. (2022a); Benton et al. (2023a)). In terms of the Wasserstein metric, the existing results often encounter an exponential dependence on the smoothness parameter of the score function (e.g., Benton et al. (2023b); Tang and Zhao (2024)). This is mainly due to a direct use of Grönwall's inequality, which provides comparisons to the solution to the initial value problem. Tackling this exponential dependence is regarded as a challenging open problem. Our result is, however, not directly comparable with these results as the smoothness assumption is imposed instead on the mapping between random variables along the forward trajectory.

- The learning schedule (cf. (6)) adopted in this paper decays exponentially when T is close to 0 and remains constant when T far from 0. This type of schedule is commonly used in the training of diffusion models, as seen in Li et al. (2023); Chen et al. (2022a); Li et al. (2024), and plays an essential role in deriving our convergence results. We note that our analysis framework can accommodate other decaying schedules as well, although alternative schedules may lead to slower convergence guarantees. The design of this learning schedule aims to balance two key factors: the discretization error, which depends crucially upon the quantity $\frac{1-\alpha_t}{1-\overline{\alpha}_t}$, and the initialization error, which depends on $\overline{\alpha}_1$. By managing this balance, this schedule ensures that errors are minimized effectively throughout the learning process.

Comparisons with consistency distillation. In this work, we concentrate our attention on the consistency training method as proposed in Song et al. (2023). Unlike the consistency distillation method (see, Song et al. (2023); Lyu et al. (2023)), consistency training integrates the learning of consistency functions directly in the training phase. Therefore, analyzing its performance requires studying the joint distribution of X_{t-1} and X_t, as well as investigating how training errors propagate over time. In contrast, consistency distillation constructs the consistency functions based on a pre-trained score estimate. This pre-trained score estimate serves as an effective initialization for estimating consistency functions. Given an accurate score estimate, our proof technique can be readily extended to study the distillation-based approach by incorporating the error of the score estimation into our error decomposition (see, eq. (29)).

A brief proof outline. Before concluding, let us take a moment to provide a brief proof outline for this result; the full technical details are postponed to Appendix A and C. In order to prove Theorem 3.3, we find it helpful to study how the error $\|f_t(X_t) - \Phi_t(X_t)\|_2$ propagates along the probability flow ODE path. Specifically, we establish the following recursive relation for each t, where

$$\begin{aligned}
&\mathbb{E}\big[\|f_t(X_t) - \Phi_t(X_t)\|_2\big] - \mathbb{E}\big[\|f_{t-1}(X_{t-1}) - \Phi_{t-1}(X_{t-1})\|_2\big] \\
&\leq \mathbb{E}\big[\|f_t(X_t) - f_t^{\mathcal{F}}(X_t)\|_2\big] + \mathbb{E}\big[\|f_t^{\mathcal{F}}(X_t) - f_t^{\star}(X_t)\|_2\big] \\
&\quad + \mathbb{E}\Big[\Big\|\frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\big(\mathbb{E}\big[X_{t-1}\,|\,X_t\big] - \phi_t(X_t)\big)\Big\|_2\Big] \\
&\quad + \mathbb{E}\Big[\int_0^1 \Big(\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\Big) \\
&\qquad\qquad \big(X_{t-1} - \phi_t(X_t)\big)\mathrm{d}\gamma\Big].
\end{aligned} \tag{29}$$

Here, we denote $X_{t-1}(\gamma) := \gamma X_{t-1} + (1-\gamma)\phi_t(X_t)$. If the right-hand side of (29) can be properly controlled, then Theorem 3.3 can be easily established by applying this relation recursively. Consequently, it boils down to bounding each term on the right-hand side separately. Towards this, the first two terms are concerned with the optimization error and approximation error in training the consistency function, which can be controlled in view of Assumption 3.2.

When it comes to the last two terms, in view of the Taylor expansion, we make the following observation

$$\mathbb{E}\Big\|\frac{\partial \Phi_{T-1}}{\partial x}\big(\phi_T(X_T)\big)\big(\mathbb{E}\big[X_{T-1}\,|\,X_T\big] - \phi_T(X_T)\big)\Big\|_2$$

$$
\begin{aligned}
&+\mathbb{E}\Big[\Big\|\int_0^1\Big(\frac{\partial\Phi_{T-1}}{\partial x}(X_{T-1}(\gamma))-\frac{\partial\Phi_{T-1}}{\partial x}\big(\phi_T(X_T)\big)\Big) \\
&\qquad\qquad\big(X_{T-1}-\phi_T(X_T)\big)\mathrm{d}\gamma\Big\|_2\Big] \\
\le{}& L_f\mathbb{E}\Big[\big\|\mathbb{E}\big[X_{t-1}\,|\,X_t\big]-\phi_t(X_t)\big\|_2\Big] \\
&+\sup_{\gamma}\mathbb{E}\Big[\Big\|\frac{\partial\Phi_{t-1}}{\partial x}(X_{t-1}(\gamma))-\frac{\partial\Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\Big\| \\
&\qquad\qquad\big\|X_{t-1}-\phi_t(X_t)\big\|_2\Big].
\end{aligned}
\tag{30}
$$

While the Lipschitz property of $\Phi_{t\to k}$ allows us to control terms involving derivatives, the main difficulty lies in controlling $\mathbb{E}[X_{t-1}\,|\,X_t]-\phi_t(X_t)$ as well as $X_{t-1}-\phi_t(X_t)$. Accomplishing this requires a careful study of the probability flow ODE in (16). We would also like to point out that the analyses of the probability flow ODE are inspired by the framework established in Li et al. (2023). The two terms resulting from the decomposition in (30) are then controlled separately.

Regarding the first term, the main challenge is to track the dynamics of $\phi_t(X_t)$. It is equivalent to studying the ODE flow, for which the evolution of the score function plays a crucial role. We characterize the properties of the score function by means of the following two lemmas, which will be proven in Appendix C.

Lemma 3.5. *For $X_t \sim \sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z$, where $X_0\sim p_{\mathsf{data}}$ and $Z\sim\mathcal{N}(0,I_d)$, the second moment of the score function satisfies*

$$\mathbb{E}\big[\|s_t(X_t)\|_2^2\big]\le\frac{d}{1-\overline{\alpha}_t}.$$

Lemma 3.6. *Let X_t be defined in the same way as in Lemma 3.5. For the pre-selected $\{\alpha_i\}_{1\le i\le t}$ and then corresponding $\overline{\alpha}_t$, $\overline{\alpha}_{t-1}$, we can deduce, as T grows, that*

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha}))-s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2\Big] \\
&\qquad\qquad\lesssim\frac{(1-\alpha_t)^4d^3\log^3 T}{(1-\overline{\alpha}_t)^3}.
\end{aligned}
$$

In words, Lemma 3.5 provides a bound on the second moment of the score function during the forward process. By virtue of the definition of the ODE flow in (16), this lemma ensures that the backward process will not change too fast. In addition, Lemma 3.6 is a key technical result that allows us to estimate the ODE flow reliably. It tells us that, for the step size schedule we select, the score function only moves a little bit from its original value during one step, which implies that we can discretize the continuously varying score function with a small cost. With these useful properties at hand, we can readily describe the conditional expectation via the score functions, and cope with the remaining term $\mathbb{E}[X_{t-1}\,|\,X_t]$ in the similar spirit. The two parts of analyses, taken collectively, lead to the following desired result:

$$\mathbb{E}\Big[\big\|\mathbb{E}\big[X_{t-1}\,|\,X_t\big]\big\|_2-\phi_t(X_t)\Big]\lesssim\frac{d^{3/2}\log^{7/2}T}{T^2},$$

when $T\to\infty$.

When it comes to the second term, the primary task is to tackle the following quantity:

$$\Big\|\frac{\partial\Phi_{t-1}}{\partial x}(X_{t-1}(\gamma))-\frac{\partial\Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\Big\|^2.$$

Notably, this is not easy to estimate since it involves Φ_{t-1}, which represents a long backward trajectory from $t-1$ to 1. To cope with this issue, we once again invoke the discretization strategy: attempting to estimate the effect in a single step, and accumulating the effect along the trajectory in a careful manner. The one-step effect is controlled by means of the lemma below.

Lemma 3.7. *For $2\le k<t\le T$, X_t and $X_{t-1}(\gamma):=\gamma X_{t-1}+(1-\gamma)\phi_t(X_t)$, when $T\to\infty$, it holds that*

$$
\begin{aligned}
&\mathbb{E}\Big\|\frac{\partial\phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big)-\frac{\partial\phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\Big\|^2 \\
\lesssim{}&\frac{(1-\alpha_k)^2(1-\alpha_t)^2L_f^2d^4\log^4T}{(1-\overline{\alpha}_k)^2(1-\overline{\alpha}_t)}+\frac{(1-\alpha_t)^4d^4\log^4T}{(1-\overline{\alpha}_t)^2}.
\end{aligned}
$$

The proof of Lemma 3.7 and the detailed derivations for the accumulated error are provided in Appendix C. The core of the proof lies in studying the stability of the backward ODE flow (16). We aim to show that, with high probability, the trajectory of the backward ODE flow is stable when the starting point x_t moves a little bit towards some directions. Therefore, the derivative function (e.g. $\partial\phi_t/\partial x$) can also be proven to be stable with high probability. With stability of $\Phi_{t\to k}$ resulting from Assumption 3.1, this deduction eventually leads to our proof of Lemma 3.7. Applying the Cauchy-Schwartz inequality would then result in

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{\partial\Phi_{t-1}}{\partial x}(X_{t-1}(\gamma))-\frac{\partial\Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\Big\| \\
&\qquad\qquad\big\|X_{t-1}-\phi_t(X_t)\big\|_2\Big]\lesssim\frac{L_f^3d^{5/2}\log^5T}{T^2}
\end{aligned}
$$

when $T\to\infty$.

It is worth noting that the complete proofs of the results mentioned above are carried out in a more delicate way. We truncate these expectations based on some pre-selected typical events, and obtain high-probability bounds. Then, we attempt to analyze samples outside the event directly with sufficient knowledge of p_{data}. Future research could consider how different kinds of "knowledge", such as different forms of tail probability, affect out analysis and convergence rates. See Appendix A for more details.

4 Discussion

In this work, we have developed a rigorous mathematical framework for analyzing consistency training in diffusion models. Given a set of consistency functions with sufficiently small training error, we have pinned down the finite-sample performance for the consistency model in terms of the Wasserstein metric, with explicit dependencies on the problem parameters. The analysis framework laid out in the current paper might potentially be applicable to other generative and distillation models, such as the progressive training procedure in Salimans and Ho (2022).

Moving forward, we highlight several possible directions worthy of future investigation. For instance, it remains unclear whether our theory offers optimal dependencies on the Lipschitz constant of the $\Phi_{t\to k}$ mappings and the ambient dimension d. Can we further refine our theory in order to obtain tighter dependencies or establish matching lower bounds? In addition, our theory decouples the training phase from the sampling phase by assuming a small optimization/estimation error. It would be of great interest to consider whether one can establish end-to-end results that combine these two phases. Moving beyond consistency models, it would also be interesting to compare our theory—in terms of sampling efficiency—with other generative sampling methods, such as accelerated ODE and SDE methods (Song and Ermon (2020); Lu et al. (2022a)).

5 Acknowledgements

Gen Li is supported in part by the Chinese University of Hong Kong Direct Grant for Research and the Hong Kong Research Grants Council ECS 2191363. Y. Wei is supported in part by the NSF grants CCF-2106778, CCF-2418156 and CAREER award DMS-2143215.

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 autoencoders and Langevin sampling. *arXiv preprint arXiv:2002.00107*.

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

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

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

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

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.

Ding, Z. and Jin, C. (2023). Consistency models as a rich and efficient policy class for reinforcement learning. *arXiv preprint arXiv:2309.16984*.

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

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. (2022). Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

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

Huang, Z., Wei, Y., and Chen, Y. (2024). Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*.

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

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

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. (2023). Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

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

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. (2024). Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*.

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

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

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

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*.

Lyu, J., Chen, Z., and Feng, S. (2023). Convergence guarantee for consistency models. *arXiv preprint arXiv:2308.11449*.

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

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

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

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. (2021). Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR.

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

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

Salimans, T. and Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

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. and Dhariwal, P. (2023). Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency models. *arXiv preprint arXiv:2303.01469*.

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

Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448.

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*.

Sun, W., Chen, D., Wang, C., Ye, D., Feng, Y., and Chen, C. (2023). Accelerating diffusion sampling with classifier-based feature distillation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 810–815. IEEE.

Tang, W. (2023). Diffusion probabilistic models. *preprint*.

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

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

Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., and Sang, N. (2023). Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*.

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*.

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

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*.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proof of Theorem 3.3

A.1 Preliminary properties

Before diving into our main analysis, we collect several auxiliary facts and properties that shall be used frequently throughout this proof.

Properties of learning rates. First, we enumerate some of useful properties about the learning rates as specified by $\{\alpha_t\}$ in (6).

$$\alpha_t \geq 1 - \frac{c_1 \log T}{T} \geq \frac{1}{2}, \qquad 1 \leq t \leq T \tag{31a}$$

$$\frac{1}{2}\frac{1-\alpha_t}{1-\overline{\alpha}_t} \leq \frac{1}{2}\frac{1-\alpha_t}{\alpha_t-\overline{\alpha}_t} \leq \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \leq \frac{4c_1 \log T}{T}, \qquad 2 \leq t \leq T \tag{31b}$$

$$1 \leq \frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} \leq 1 + \frac{4c_1 \log T}{T}, \qquad 2 \leq t \leq T \tag{31c}$$

$$\overline{\alpha}_T \leq \frac{1}{T^{c_2}}. \tag{31d}$$

In the last line, $c_2 \geq 1000$ is some large numerical constant. All the properties hold provided that T is large enough. The proof of these properties can be found in (Li et al., 2023, Appendix A.2)

Truncation on typical events. Next, let us introduce the following event:

$$\mathcal{E}_t \coloneqq \bigg\{ (x_t, x_{t-1}) \in \mathbb{R}^d \times \mathbb{R}^d \,\bigg|\, -\log p_{X_t}(x_t) \leq c_3 d \log T, \\ \|x_{t-1} - x_t/\sqrt{\alpha_t}\|_2 \leq c_4\sqrt{d(1-\alpha_t)\log T} \bigg\}, \tag{32}$$

where c_3 and c_4 are some numerical constants to be specified later. Generally speaking, $\mathcal{E}$ encompasses a typical range of the values of (X_t, X_{t-1}), and some part of our analysis proceed by seperately considering the points in $\mathcal{E}$ and those outside $\mathcal{E}$. While truncated on $\mathcal{E}$, there are some nice continuity properties on the trajectories, and for $(x_t, x_{t-1}) \in \mathcal{E}^c$, we have

$$\begin{aligned}
\mathbb{P}\big((X_t, X_{t-1}) \notin \mathcal{E}\big) &= \int_{(x_t,x_{t-1})\notin\mathcal{E}} p_{X_{t-1}}(x_{t-1}) p_{X_t \,|\, X_{t-1}}(x_t \,|\, x_{t-1}) \mathrm{d}x_{t-1}\mathrm{d}x_t \\
&= \int_{(x_t,x_{t-1})\notin\mathcal{E}} p_{X_t-1}(x_{t-1}) \frac{1}{\big(2\pi(1-\alpha_t)\big)^{d/2}} \\
&\qquad\qquad \exp\bigg(-\frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|_2^2}{2(1-\alpha_t)}\bigg)\mathrm{d}x_{t-1}\mathrm{d}x_t \\
&\leq \exp\big(-c_4 d \log T\big),
\end{aligned} \tag{33}$$

which can be a high order term in T when c_4 is large enough.

On the typical event $\mathcal{E}$, the score and density functions behave regularly, which are clarified by the following two lemmas from Li et al. (2023).

Lemma A.1 (Li et al. (2023), Lemma 1)**.** *Consider any $x_t \in \mathbb{R}^d$ satisfying $-\log p_{X_t}(x_t) \leq c_3 d \log T$ for some large enough constant c_3. Then it holds that*

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - x_t\big\|_2 \,\big|\, X_t = x_t\right] \lesssim \sqrt{d(1-\overline{\alpha}_t)\log T}, \tag{34a}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - x_t\big\|_2^2 \,\big|\, X_t = x_t\right] \lesssim d(1-\overline{\alpha}_t)\log T, \tag{34b}$$

$$\mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - x_t\big\|_2^3 \,\big|\, X_t = x_t\right] \lesssim \big(d(1-\overline{\alpha}_t)\log T\big)^{3/2}. \tag{34c}$$

Lemma A.1 implies that if X_t taking on a "typical" value, then condition on it, the vector $\sqrt{\overline{\alpha}_t}X_0 - X_t = \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t$ might still follow a sub-Gaussian tail, whose expected norm remains on the same order of that of an unconditional Gaussian vector $\mathcal{N}(0, (1-\overline{\alpha}_t)I_d)$.

Lemma A.2 (Li et al. (2023), Lemma 2)**.** *Consider any two points $x_t, x_{t-1} \in \mathbb{R}^d$ obeying*

$$-\log p_{X_t}(x_t) \le \frac{1}{2}c_3 d \log T, \quad \text{and} \quad \left\|x_{t-1} - \frac{x_t}{\sqrt{\alpha_t}}\right\|_2 \le c_4\sqrt{d(1-\alpha_t)\log T} \tag{35}$$

for some large constants $c_3, c_4 > 0$. Then we have

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

and for all $\gamma \in [0,1]$,

$$-\log p_{X_{t-1}}\big(x_t(\gamma)\big) \le c_6 d \log T. \tag{36}$$

In other words, Lemma A.2 ensures that if x_t falls within a typical set of X_t and the point x_{t-1} is not too far away from $x_t/\sqrt{\alpha_t}$, then x_{t-1} is also a typical value of X_{t-1}. Lemma A.2 here is in a slightly different form from the original version in Li et al. (2023) due to a different definition of $x_{t-1}(\gamma)$. Notice that using the inequality (59), the proof of Lemma 2 in Li et al. (2023) remains valid with the new definition of $x_{t-1}(\gamma)$, so we keep the original statement of this lemma.

A.2 Main analysis

Throughout this proof, we shall use capital letters to denote random vectors, and lower case letters to denote their corresponding realizations, i.e. for some specific point in the sample space $\omega \in \Omega$, we could write $x_t := X_t(\omega)$ and $x_{t-1} := X_{t-1}(\omega)$.

First, notice that $X_1 \overset{\mathsf{d}}{=} \Phi_T(X_T)$, which gives

$$\mathcal{W}_1\big(f_T(X_T), X_1\big) = \mathcal{W}_1\big(f_T(X_T), \Phi_T(X_T)\big) \le \mathbb{E}\big[\|f_T(X_T) - \Phi_T(X_T)\|_2\big].$$

To control the right hand side above, let us introduce a piece of notation

$$\xi_t(x) := f_t(x) - \Phi_t(x), \tag{37}$$

and we claim that ξ_t satisfies the following recursive relation with $\xi_1 = 0$:

$$\begin{aligned}\xi_t(x_t) = \mathbb{E}\big[\xi_{t-1}(X_{t-1}) \,|\, X_t = x_t\big] + \big(f_t(x_t) - f_t^{\mathcal{F}}(x_t)\big) + \big(f_t^{\mathcal{F}}(x_t) - f_t^{\star}(x_t)\big) \\ + \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\big(\mathbb{E}\big[X_{t-1} \,|\, X_t = x_t\big] - \phi_t(x_t)\big) \\ + \mathbb{E}\bigg[\int_0^1 \bigg(\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\bigg)\big(X_{t-1} - \phi_t(x_t)\big)\mathrm{d}\gamma \,\bigg|\, X_t = x_t\bigg],\end{aligned} \tag{38}$$

where we let

$$x_{t-1}(\gamma) := \gamma x_{t-1} + (1-\gamma)\phi_t(x_t) \tag{39}$$

for $\gamma \in [0,1]$. We leave its derivation to Section C.1. In addition, let us denote $X_{t-1}(\gamma) = \gamma X_{t-1} + (1-\gamma)\phi_t(X_t)$, and the above relation implies that

$$\begin{aligned}\mathbb{E}\big[\big\|\xi_T(X_T)\big\|_2\big] \le \mathbb{E}\big[\big\|\xi_{T-1}(X_{T-1})\big\|_2\big] + \mathbb{E}\big[\big\|f_T(X_T) - f_T^{\mathcal{F}}(X_T)\big\|_2\big] + \mathbb{E}\big[\big\|f_T^{\mathcal{F}}(X_T) - f_T^{\star}(X_T)\big\|_2\big] \\ + \mathbb{E}\bigg\{\bigg\|\frac{\partial \Phi_{T-1}}{\partial x}\big(\phi_T(X_T)\big)\big(\mathbb{E}\big[X_{T-1} \,|\, X_T\big] - \phi_T(X_T)\big)\bigg\|_2 \\ + \mathbb{E}\bigg[\bigg\|\int_0^1 \bigg(\frac{\partial \Phi_{T-1}}{\partial x}(X_{T-1}(\gamma)) - \frac{\partial \Phi_{T-1}}{\partial x}\big(\phi_T(X_T)\big)\bigg) \\ \big(X_{T-1} - \phi_T(X_T)\big)\mathrm{d}\gamma\bigg\|_2\bigg]\bigg\}\end{aligned} \tag{40}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \sum_{t=1}^{T} \mathbb{E}\big[\big\|f_t(X_t) - f_t^{\mathcal{F}}(X_t)\big\|_2\big] + \mathbb{E}\big[\big\|f_t^{\mathcal{F}}(X_t) - f_t^{\star}(X_t)\big\|_2\big] \\
&\qquad + \sum_{t=2}^{T} \bigg\{ \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\mathbb{E}\Big[\big\|\mathbb{E}\big[X_{t-1} \,|\, X_t\big] - \phi_t(X_t)\big\|_2\Big] \\
&\qquad\qquad + \int_0^1 \mathbb{E}\bigg[\bigg\|\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\bigg\| \\
&\qquad\qquad\qquad\qquad \big\|X_{t-1} - \phi_t(X_t)\big\|_2 \mathrm{d}\lambda\bigg]\bigg\} \\
&\overset{\text{(ii)}}{\leq} \varepsilon + \varepsilon_{\mathcal{F}} + \sum_{t=2}^{T} \bigg\{ L_f \mathbb{E}\Big[\big\|\mathbb{E}\big[X_{t-1} \,|\, X_t\big] - \phi_t(X_t)\big\|_2\Big] \\
&\qquad + \sup_{\gamma} \mathbb{E}\bigg[\bigg\|\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\bigg\|\big\|X_{t-1} - \phi_t(X_t)\big\|_2\bigg]\bigg\}, \\
&= \varepsilon + \varepsilon_{\mathcal{F}} + T_1 + T_2,
\end{aligned}
\tag{41}
$$

where relation (i) applies inequality (40) recursively and relation (ii) invokes the triangle inequality and Assumption 3.1. In the following, we proceed to bound the latter two terms separately.

Control quantity T_1. Let us start with the term T_1, where the goal is to control each quantity in the summation, which is $\mathbb{E}\Big[\big\|\mathbb{E}\big[X_{t-1} \,|\, X_t\big] - \phi_t(X_t)\big\|_2^2\Big]$. Recalling the backward ODE flow (16) that $\Phi_{t \to k}(x) := g_t(x, \overline{\alpha}_k)$ and

$$
\frac{\partial g_t(x, \overline{\alpha})}{\partial \overline{\alpha}} = \frac{1}{2\overline{\alpha}}\Big(g_t(x, \overline{\alpha}) + s_{\overline{\alpha}}(g_t(x, \overline{\alpha}))\Big), \qquad \text{and } g_t(x, \overline{\alpha}_t) = x,
$$

it is easy to check that

$$
\frac{\partial\Big(\frac{1}{\sqrt{\overline{\alpha}}} g_t(x, \overline{\alpha})\Big)}{\partial \overline{\alpha}} = \frac{1}{2\overline{\alpha}^{\frac{3}{2}}} s_{\overline{\alpha}}(g_t(x, \overline{\alpha})).
\tag{42}
$$

As a result, we can track the backward process with the score function as:

$$
\begin{aligned}
\sqrt{\alpha_t}\phi_t(X_t) &= X_t + (\sqrt{\alpha_t}\Phi_{t \to t-1}(X_t) - \Phi_{t \to t}(X_t)) \\
&= X_t + \sqrt{\overline{\alpha}_t}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{2\overline{\alpha}^{3/2}} s_{\overline{\alpha}}(g_t(X_t, \overline{\alpha}))\mathrm{d}\overline{\alpha} \\
&= X_t + (1 - \sqrt{\alpha_t}) s_t(X_t) + \frac{1}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t, \overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}.
\end{aligned}
\tag{43}
$$

For the remaining term, we first apply the definition of the forward process:

$$
\begin{aligned}
\sqrt{\alpha_t}\mathbb{E}\big[X_{t-1} \,|\, X_t\big] &= \sqrt{\alpha_t}\mathbb{E}\Big[\sqrt{\overline{\alpha}_{t-1}} X_0 + \sqrt{1 - \overline{\alpha}_{t-1}} Z \,|\, X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1 - \overline{\alpha}_t} Z\Big] \\
&= X_t + \mathbb{E}\big[\big(\sqrt{\alpha_t - \overline{\alpha}_t} - \sqrt{1 - \overline{\alpha}_t}\big) Z \,|\, X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1 - \overline{\alpha}_t} Z\big].
\end{aligned}
\tag{44}
$$

The previous work on score matching admits a minimum mean square error (MMSE) form for the score function (e.g. Hyvärinen (2005); Vincent (2011); Chen et al. (2022b)):

$$
s_{\overline{\alpha}} := \arg \min_{s: \mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}\bigg[\bigg\|s(\sqrt{\overline{\alpha}} X_0 + \sqrt{1 - \overline{\alpha}} Z) + \frac{1}{\sqrt{1 - \overline{\alpha}}} Z\bigg\|_2^2\bigg],
$$

which leads to an alternative expression by the change of variables:

$$
s_{\overline{\alpha}}(x) = \mathbb{E}\bigg[-\frac{1}{\sqrt{1 - \overline{\alpha}}} Z \;\bigg|\; \sqrt{\overline{\alpha}} X_0 + \sqrt{1 - \overline{\alpha}} Z = x\bigg].
\tag{45}
$$

Plugging equation (45) into (44), we obtain

$$\sqrt{\alpha_t}\mathbb{E}\big[X_{t-1}\,|\,X_t\big] = X_t + \big(1-\overline{\alpha}_t - \sqrt{(1-\overline{\alpha}_t)(\alpha_t-\overline{\alpha}_t)}\big)s_t(X_t), \tag{46}$$

which when combined with (44) yields

$$\begin{aligned}
&\sqrt{\alpha_t}\phi_t(X_t) - \sqrt{\alpha_t}\mathbb{E}\big[X_{t-1}\,|\,X_t\big] \\
&\quad = \Big((1-\sqrt{\alpha_t}) - (1-\overline{\alpha}_t) + \sqrt{(1-\overline{\alpha}_t)(\alpha_t-\overline{\alpha}_t)}\Big)s_t(X_t) \\
&\qquad\qquad + \frac{1}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}.
\end{aligned} \tag{47}$$

With equations (43) and (46) in place, we arrive at

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\phi_t(X_t) - \mathbb{E}\big[X_{t-1}\,|\,X_t\big]\big\|_2^2\Big] \\
&= \frac{1}{\alpha_t}\mathbb{E}\Big[\big\|\sqrt{\alpha_t}\phi_t(X_t) - \sqrt{\alpha_t}\mathbb{E}\big[X_{t-1}\,|\,X_t\big]\big\|_2^2\Big] \\
&= \frac{1}{\alpha_t}\mathbb{E}\Bigg[\bigg\|\Big(-(1-\overline{\alpha}_t) + \sqrt{(1-\overline{\alpha}_t)(\alpha_t-\overline{\alpha}_t)} + (1-\sqrt{\alpha_t})\Big)s_t(X_t) \\
&\qquad\qquad + \frac{1}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}\bigg\|_2^2\Bigg] \\
&\le \frac{2}{\alpha_t}\Big(-\sqrt{1-\overline{\alpha}_t}\,\big(\sqrt{1-\overline{\alpha}_t}-\sqrt{\alpha_t-\overline{\alpha}_t}\big) + (1-\sqrt{\alpha_t})\Big)^2\mathbb{E}\big[\|s_t(X_t)\|_2^2\big] \\
&\qquad + \frac{1}{2\alpha_t}\mathbb{E}\Big[\Big\|\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2\Big].
\end{aligned}$$

In view of the Taylor expansion, we can further control the right hand side above as

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\phi_t(X_t) - \mathbb{E}\big[X_{t-1}\,|\,X_t\big]\big\|_2^2\Big] \\
&\le \frac{2}{\alpha_t}\bigg(-\Big(\frac{1-\alpha_t}{2} - \frac{(1-\alpha_t)^2}{8(1-\overline{\alpha}_t)}\Big) + \Big(\frac{1-\alpha_t}{2} + \frac{(1-\alpha_t)^2}{8}\Big)\bigg)^2\mathbb{E}\big[\|s_t(X_t)\|_2^2\big] \\
&\qquad + O\Big(\frac{(1-\alpha_t)^5}{\alpha_t(1-\overline{\alpha}_t)^{5/2}}\Big)\mathbb{E}\big[\|s_t(X_t)\|_2^2\big] \\
&\qquad\quad + \frac{1}{2\alpha_t}\mathbb{E}\Big[\Big\|\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2\Big] \\
&\lesssim \frac{(1-\alpha_t)^4}{(1-\overline{\alpha}_t)^2}\mathbb{E}\big[\|s_t(X_t)\|_2^2\big] + \mathbb{E}\Big[\Big\|\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2\Big].
\end{aligned} \tag{48}$$

To further control the right hand side of expression (48), we introduce the following Lemma A.3 and Lemma A.4, which provide upper bounds for the two expectations in (48) respectively. The proofs of these lemmas can be found in Sections C.2 and C.3 respectively.

Lemma A.3. *For $X_t \sim \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\ Z$, where $X_0 \sim p_{\mathsf{data}}$ and $Z \sim \mathcal{N}(0, I_d)$, the second moment of the score function satisfies*

$$\mathbb{E}\big[\|s_t(X_t)\|_2^2\big] \le \frac{d}{1-\overline{\alpha}_t}.$$

Moreover, for any $0 < \overline{\alpha} < 1$, the lemma still holds when replace $\overline{\alpha}_t$ with $\overline{\alpha}$.

Lemma A.4. *For X_t defined the same as in Lemma A.3, pre-selected $\{\alpha_i\}_{1\le i\le t}$ and corresponding $\overline{\alpha}_t$, $\overline{\alpha}_{t-1}$, we deduce that*

$$\mathbb{E}\Big[\Big\|\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2\Big] \lesssim \frac{(1-\alpha_t)^4 d^3\log^3 T}{(1-\overline{\alpha}_t)^3}.$$

In view of Lemma A.3 and Lemma A.4, the right hand side of (48) is further controlled as

$$\begin{aligned}\mathbb{E}\Big[\big\|\phi_t(X_t) - \mathbb{E}\big[X_{t-1}\,|\,X_t\big]\big\|_2^2\Big] &\lesssim \frac{(1-\alpha_t)^4 d}{(1-\overline{\alpha}_t)^3} + \frac{(1-\alpha_t)^4 d^3 \log^3 T}{(1-\overline{\alpha}_t)^3} \\ &\lesssim \frac{(1-\alpha_t)^4 d^3 \log^3 T}{(1-\overline{\alpha}_t)^3}. \end{aligned} \tag{49}$$

Now by properties of the step sizes mentioned in (31b), this upper bound can be simplified as

$$\mathbb{E}\Big[\big\|\phi_t(X_t) - \mathbb{E}\big[X_{t-1}\,|\,X_t\big]\big\|_2\Big] \le \frac{C_2 d^{3/2}\log^{7/2} T}{T^2}, \tag{50}$$

where C_2 denotes some universal constant.

Control quantity T_2. Now, let us turn our attention to control the term T_2. We first decompose this term by the Cauchy-Schwartz inequality:

$$\begin{aligned}&\mathbb{E}\bigg[\bigg\|\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\bigg\|\big\|X_{t-1} - \phi_t(X_t)\big\|_2\bigg] \\ \le&\frac{1}{2}\mathbb{E}\big\|X_{t-1} - \phi_t(X_t)\big\|_2^2 + \frac{1}{2}\mathbb{E}\bigg\|\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\bigg\|^2, \end{aligned} \tag{51}$$

and we aim to handle the two components respectively.

- Towards bounding the first term in (51), in view of relation (43), we make the observation that

$$\begin{aligned}&\mathbb{E}\big\|X_{t-1} - \phi_t(X_t)\big\|_2^2 \\ =&\frac{1}{\alpha_t}\mathbb{E}\big\|(\sqrt{\alpha_t}X_{t-1} - X_t) - (\sqrt{\alpha_t}\phi_t(X_t) - X_t)\big\|_2^2 \\ =&\frac{1}{\alpha_t}\mathbb{E}\Big\|\big(\sqrt{\alpha_t - \overline{\alpha}_t} - \sqrt{1-\overline{\alpha}_t}\big)Z + (1-\sqrt{\alpha_t})s_t(X_t) \\ &\qquad + \frac{1}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2 \\ \le&\frac{3}{\alpha_t}\mathbb{E}\Big[\big\|\big(\sqrt{\alpha_t - \overline{\alpha}_t} - \sqrt{1-\overline{\alpha}_t}\big)Z\big\|_2^2 + \big\|(1-\sqrt{\alpha_t})s_t(X_t)\big\|_2^2 \\ &\qquad + \frac{1}{4}\Big\|\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha})) - s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2\Big] \\ \lesssim&\frac{(1-\alpha_t)^2}{1-\overline{\alpha}_t} + \frac{(1-\alpha_t)^2 d}{1-\overline{\alpha}_t} + \frac{(1-\alpha_t)^4 d^3\log^3 T}{(1-\overline{\alpha}_t)^3} \\ \lesssim&\frac{d\log^2 T}{T^2}. \end{aligned} \tag{52}$$

  Here, we recall the properties of the learning rates as in (31a) and (31b).

- When it comes to the second term in (51) , we claim that for any (x_t, x_{t-1}) pair, it can be decomposed as

$$\begin{aligned}&\bigg\|\frac{\partial \Phi_{t-1}}{\partial x}(x_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\bigg\| \\ &\qquad \le L_f^2\sum_{k=1}^{t}\bigg\|\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(x_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(x_t)\big)\bigg\|. \end{aligned} \tag{53}$$

  The proof of claim (53) is provided in our Section C.5. We proceed to control the right hand side above with the aid of the following lemma.

Lemma A.5. *For $2 \leq k < t \leq T$, X_t and $X_{t-1}(\gamma)$ defined as above, it holds that*

$$\mathbb{E}\bigg\|\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\bigg\|^2 \lesssim \frac{(1-\alpha_k)^2(1-\alpha_t)^2 L_f^2 d^4 \log^3 T}{(1-\overline{\alpha}_k)^2(1-\overline{\alpha}_t)^2} + \frac{(1-\alpha_t)^4 d^4 \log^4 T}{(1-\overline{\alpha}_t)^4}. \tag{54}$$

We defer the proof of this result to Section C.5. With Lemma A.5 in place, we can further derive that

$$\begin{aligned}\mathbb{E}\bigg\|\frac{\partial \Phi_{t-1}}{\partial x}(x_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\bigg\|^2 &\lesssim L_f^4 T^2\bigg(\frac{(1-\alpha_k)^2(1-\alpha_t)^2 L_f^2 d^4 \log^3 T}{(1-\overline{\alpha}_k)^2(1-\overline{\alpha}_t)^2} + \frac{(1-\alpha_t)^4 d^4 \log^4 T}{(1-\overline{\alpha}_t)^4}\bigg) \\ &\lesssim \frac{L_f^6 d^4 \log^8 T}{T^2}.\end{aligned} \tag{55}$$

Here, again we use the properties of step size in (31b).

Putting expressions (52) and (55) together leads to

$$\mathbb{E}\bigg[\bigg\|\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(X_t)\big)\bigg\|\big\|X_{t-1} - \phi_t(X_t)\big\|_2\bigg] \leq \frac{C_1 L_f^3 d^{5/2}\log^5 T}{T^2}. \tag{56}$$

In conclusion, taking relations (50) and (56) collectively with relation (41), we arrive at

$$\begin{aligned}\mathcal{W}_1\big(f_T(X_T), X_1\big) &\leq \mathbb{E}\big[\big\|f_T(X_T) - \Phi_T(X_T)\big\|_2\big] \\ &\leq \frac{C_2 d^{3/2}\log^{7/2} T}{T^2}\cdot T + \frac{C_1 L_f^3 d^{5/2}\log^5 T}{T^2}\cdot T + \varepsilon + \varepsilon_{\mathcal{F}} \\ &\leq \frac{C_1 L_f^3 d^{5/2}\log^5 T}{T} + \varepsilon + \varepsilon_{\mathcal{F}}.\end{aligned}$$

This thus completes the proof of our advertised result.

B Proof of Corollary 3.4

For the sampling process starting from $\overline{Z}$, one can easily write

$$\mathcal{W}_1(f_T(\overline{Z}), X_1) \leq \mathbb{E}\big[\|f_T(\overline{Z}) - \Phi_T(X_T)\|_2\big] \leq \mathbb{E}\big[\|f_T(\overline{Z}) - \Phi_T(\overline{Z})\|_2\big] + \mathbb{E}\big[\|\Phi_T(\overline{Z}) - \Phi_T(X_T)\|_2\big].$$

In view of the Lipschitz property, one has

$$\mathbb{E}\big[\|\Phi_T(\overline{Z}) - \Phi_T(X_T)\|_2\big] \leq L_f \mathbb{E}\big[\|\overline{Z} - X_T\|_2\big] \leq \frac{L_f}{T}.$$

In addition, for the first term, we have

$$\begin{aligned}\mathbb{E}[\|f_T(\overline{Z}) - \Phi_T(\overline{Z})\|_2] &\lesssim \int_{\|x\|_2 \leq T^c} p_Z(x)\|f_T(x) - \Phi_T(x)\|_2 \mathrm{d}x \\ &\lesssim \int_{\|x\|_2 \leq T^c} p_{X_T}(x)\|f_T(x) - \Phi_T(x)\|_2 \mathrm{d}x \\ &\lesssim \mathbb{E}[\|f_T(X_T) - \Phi_T(X_T)\|_2],\end{aligned}$$

where the second line holds since for $\|x\|_2 \leq T^c$,

$$p_{X_T}(x) = \int_{x_0} p_{X_0}(x_0)(2\pi(1-\overline{\alpha}_T))^{-d/2}\exp\Big(-\frac{\|x - \sqrt{\overline{\alpha}_T}x_0\|_2^2}{2(1-\overline{\alpha}_T)}\Big)\mathrm{d}x_0$$

$$
\begin{aligned}
&\geq \inf_{x_0 \in \mathsf{supp}(X_0)} (2\pi(1-\overline{\alpha}_T))^{-d/2} \exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_T}x_0\|_2^2}{2(1-\overline{\alpha}_T)}\Big) \\
&\geq (2\pi)^{-d/2} \exp\Big(-\frac{\|x\|_2^2}{2}\Big)(1-\overline{\alpha}_T)^{-d/2} \exp\Big(\inf_{x_0 \in \mathsf{supp}(X_0)} \Big[\frac{\|x\|_2^2}{2} - \frac{\|x-\sqrt{\overline{\alpha}_T}x_0\|_2^2}{2(1-\overline{\alpha}_T)}\Big]\Big) \\
&\gtrsim p_Z(x).
\end{aligned}
$$

Here, the last line holds since

$$
\begin{aligned}
\inf_{x_0 \in \mathsf{supp}(X_0)} \Big[\frac{\|x\|_2^2}{2} - \frac{\|x-\sqrt{\overline{\alpha}_T}x_0\|_2^2}{2(1-\overline{\alpha}_T)}\Big] &\geq -\frac{2\sqrt{\overline{\alpha}_T}\|x\|_2\|x_0\|_2 + \overline{\alpha}_T\|x_0\|_2^2}{2(1-\overline{\alpha}_T)} \\
&\geq -\frac{2\sqrt{T^{-c_2}}T^{c}T^{c_R} + T^{-c_2}T^{2c_R}}{2(1-T^{-c_2})} \geq -1,
\end{aligned}
$$

where the second line comes from the relations $\|x\|_2 \leq T^c$ (the threshold of $\overline{Z}$), Eq. (26) for $\|x_0\|_2$ and Eq. (32d) for $\overline{\alpha}_T$, provided that $c_2 > 2c + 2c_R + 1$. (This is where $\sqrt{\overline{\alpha}_T}\|x\|_2 \sup\|X_0\|_2 \lesssim 1$ is used previously.) Putting everything together leads to our desired result.

C Proof of auxiliary results

C.1 Proof of the recursion (38)

Recalling the definitions of $f_t(x)$ and $f_t^\star(x)$ yields

$$
\begin{aligned}
f_t(x_t) &= \mathbb{E}\big[\Phi_{t-1}(X_{t-1}) \,|\, X_t = x_t\big] + \mathbb{E}\big[\xi_{t-1}(X_{t-1}) \,|\, X_t = x_t\big] \\
&\qquad + \big(f_t(x_t) - f_t^{\mathcal{F}}(x_t)\big) + \big(f_t^{\mathcal{F}}(x_t) - f_t^\star(x_t)\big) \\
&= \Phi_t(x_t) + \mathbb{E}\big[\Phi_{t-1}(X_{t-1}) - \Phi_{t-1}\big(\phi_t(x_t)\big) \,|\, X_t = x_t\big] + \mathbb{E}\big[\xi_{t-1}(x_{t-1}) \,|\, X_t = x_t\big] \\
&\qquad + \big(f_t(x_t) - f_t^{\mathcal{F}}(x_t)\big) + \big(f_t^{\mathcal{F}}(x_t) - f_t^\star(x_t)\big).
\end{aligned}
$$

Invoking the Taylor expansion to obtain

$$
\begin{aligned}
&\Phi_{t-1}(x_{t-1}) - \Phi_{t-1}\big(\phi_t(x_t)\big) \\
&= \int_\gamma \frac{\partial \Phi_{t-1}}{\partial x}(x(\gamma))\big(x_{t-1} - \phi_t(x_t)\big)\mathrm{d}\gamma \\
&= \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\big(x_{t-1} - \phi_t(x_t)\big) \\
&\qquad + \int_\gamma \Big(\frac{\partial \Phi_{t-1}}{\partial x}(x_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\Big)\big(x_{t-1} - \phi_t(x_t)\big)\mathrm{d}\gamma,
\end{aligned}
$$

further leads to

$$
\begin{aligned}
f_t(x_t) &= \Phi_t(x_t) + \mathbb{E}\big[\xi_{t-1}(X_{t-1}) \,|\, X_t = x_t\big] + \big(f_t(x_t) - f_t^{\mathcal{F}}(x_t)\big) + \big(f_t^{\mathcal{F}}(x_t) - f_t^\star(x_t)\big) \\
&\quad + \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\big(\mathbb{E}\big[X_{t-1} \,|\, X_t = x_t\big] - \phi_t(x_t)\big) \\
&\quad + \mathbb{E}\Big[\int_0^1 \Big(\frac{\partial \Phi_{t-1}}{\partial x}(X_{t-1}(\gamma)) - \frac{\partial \Phi_{t-1}}{\partial x}\big(\phi_t(x_t)\big)\Big)\big(X_{t-1} - \phi_t(x_t)\big)\mathrm{d}\gamma \,\Big|\, X_t = x_t\Big].
\end{aligned} \tag{57}
$$

This thus establishes relation (38).

C.2 Proof of Lemma A.3

We first recall the definition of $s_t(x)$, which is the score function of $X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z$. If we let $P_{\sqrt{\overline{\alpha}_t}}$ be the probability measure of $\sqrt{\overline{\alpha}_t}X_0$, and $p_{\sqrt{1-\overline{\alpha}_t}Z}$ be the density of $\sqrt{1-\overline{\alpha}_t}Z$, by definition of the score function, we can write

$$
s_t(x) = -\frac{\nabla_x \int_{x_0} p_{\sqrt{1-\overline{\alpha}_t}Z}(x - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}P_{\sqrt{\overline{\alpha}_t}X_0}(\sqrt{\overline{\alpha}_t}x_0)}{\int_{x_0} p_{\sqrt{1-\overline{\alpha}_t}Z}(x - \sqrt{\overline{\alpha}_t}x_0)\mathrm{d}P_{\sqrt{\overline{\alpha}_t}X_0}(\sqrt{\overline{\alpha}_t}x_0)}
$$

$$= -\frac{\int_{x_0} \frac{\sqrt{\overline{\alpha}_t}x_0 - x}{1-\overline{\alpha}_t} \exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}P_{\sqrt{\overline{\alpha}_t}X_0}(\sqrt{\overline{\alpha}_t}x_0)}{\int_{x_0} \exp\Big(-\frac{\|x-\sqrt{\overline{\alpha}_t}x_0\|^2}{2(1-\overline{\alpha}_t)}\Big)\mathrm{d}P_{\sqrt{\overline{\alpha}_t}X_0}(\sqrt{\overline{\alpha}_t}x_0)}$$

$$= \mathbb{E}_{X_0 \mid X_t = x}\Big[\frac{\sqrt{\overline{\alpha}_t}X_0 - x}{1-\overline{\alpha}_t}\Big]. \tag{58}$$

The second moment of score function thus can be written as

$$\begin{aligned}
\mathbb{E}\|s_t(X_t)\|^2 &= \mathbb{E}_{X_t}\Big\|\mathbb{E}_{X_0 \mid X_t}\Big[\frac{\sqrt{\overline{\alpha}_t}X_0 - X_t}{1-\overline{\alpha}_t}\Big]\Big\|_2^2 \\
&\le \mathbb{E}_{X_t}\Big[\mathbb{E}_{X_0 \mid X_t}\Big\|\frac{\sqrt{\overline{\alpha}_t}X_0 - X_t}{1-\overline{\alpha}_t}\Big\|_2^2\Big] \\
&= \mathbb{E}_{X_t}\Big[\frac{1}{(1-\overline{\alpha}_t)^2}\mathbb{E}_{X_0 \mid X_t}\|\sqrt{\overline{\alpha}_t}X_0 - X_t\|_2^2\Big] \\
&= \frac{d}{1-\overline{\alpha}_t},
\end{aligned}$$

where the last line makes use of the expression (7).

C.3 Proof of Lemma A.4

Throughout this proof, we adopt the truncation strategy onto the typical event $\mathcal{E}_t$ (defined in expression (32)). The targeted expectation is then calculated by considering the typical event and its complement separately.

On the typical event $\mathcal{E}_t$. Let us first consider the case when $(x_t, x_{t-1}) \in \mathcal{E}_t$. We claim that

$$\|s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha}))\|_2^2 \le c_5\frac{d\log T}{1-\overline{\alpha}_t} \quad \text{and} \quad \Big\|g_t(x_t,\overline{\alpha}) - \sqrt{\frac{\overline{\alpha}}{\overline{\alpha}_t}}x_t\Big\|_2 \le c_6\sqrt{d(1-\alpha_t)\log T} \tag{59}$$

hold for all $\overline{\alpha}_t \le \overline{\alpha} \le \overline{\alpha}_{t-1}$. This claim essentially means that every $(x_t, x_{t-1}) \in \mathcal{E}_t$ induces a trajectory on which all the points share similar properties as the definition of $\mathcal{E}_t$. In the following proof, we shall use $\widetilde{\alpha}$ as the variable of integration to differentiate from $\overline{\alpha}$, which serves as an argument.

Before proceeding, we isolate some properties obtained with the help of this claim. In particular, if relation (59) holds, then dynamic (42) implies that

$$\begin{aligned}
g_t(x_t,\overline{\alpha}) =&\sqrt{\overline{\alpha}}\Big(\frac{x_t}{\sqrt{\overline{\alpha}_t}} + \frac{1}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}}\sqrt{\frac{1}{\widetilde{\alpha}^3}}\big(s_{\widetilde{\alpha}}(g_t(x_t,\widetilde{\alpha})) - s_t(x_t)\big)\mathrm{d}\widetilde{\alpha}\Big) \\
=&\sqrt{\overline{\alpha}}\Big(\frac{x_t}{\sqrt{\overline{\alpha}_t}} + \frac{1}{2}\int_{\overline{\alpha}_t}^{\overline{\alpha}}\sqrt{\frac{1}{\widetilde{\alpha}^3}}\mathrm{d}\widetilde{\alpha}\cdot O\big(\sup_{\overline{\alpha}_t<\widetilde{\alpha}<\overline{\alpha}}\|s_{\widetilde{\alpha}}(g_t(x_t,\widetilde{\alpha})) - s_t(x_t)\|_2\big)\Big) \\
\le&\sqrt{\frac{\overline{\alpha}}{\overline{\alpha}_t}}x_t + O\bigg(\sqrt{\overline{\alpha}_{t-1}}\Big(\frac{1}{\sqrt{\overline{\alpha}_t}} - \frac{1}{\overline{\alpha}_{t-1}}\Big)\sup_{\overline{\alpha}_t<\widetilde{\alpha}<\overline{\alpha}}\|s_{\widetilde{\alpha}}(g_t(x_t,\widetilde{\alpha}))\|_2\bigg) \\
=&\sqrt{\frac{\overline{\alpha}}{\overline{\alpha}_t}}x_t + O\bigg((1-\alpha_t)\sup_{\overline{\alpha}_t<\widetilde{\alpha}<\overline{\alpha}}\|s_{\widetilde{\alpha}}(g_t(x_t,\widetilde{\alpha}))\|_2\bigg) \\
=&\sqrt{\frac{\overline{\alpha}}{\overline{\alpha}_t}}x_t + O\bigg(\sqrt{\frac{d(1-\alpha_t)^2\log T}{1-\overline{\alpha}_t}}\bigg),
\end{aligned} \tag{60}$$

where the last line holds using the bound (59). In addition, given the claim (59), according to (161c) in (Li et al., 2023, Appendix C.1), the following inequality holds:

$$\|s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha})) - s_t(x_t)\|_2 \lesssim (1-\alpha_t)\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{3/2}. \tag{61}$$

Proof of relation (59)**.** We establish the relation (59) by contradiction. If the condition does not hold along the trajectory, let us define

$$\widehat{\alpha} := \min\Big\{\overline{\alpha} : \|s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha}))\|_2^2 > \frac{c_5 d\log T}{1-\overline{\alpha}_t} \text{ or } \|g_t(x_t,\overline{\alpha}) - \sqrt{\overline{\alpha}/\overline{\alpha}_t}x_t\|_2 > c_6\sqrt{d(1-\alpha_t)\log T}\Big\}.$$

The contradiction appears if we show both scenarios in the definition of $\widehat{\alpha}$ cannot happen. By virtue of this definition, it satisfies that for $\overline{\alpha}_t \le \widehat{\alpha} < \overline{\alpha}$, inequalities (60) and (61) still hold true.

- If for the defined $\widehat{\alpha}$, we have $\|g_t(x_t,\overline{\alpha}) - \sqrt{\overline{\alpha}/\overline{\alpha}_t}x_t\|_2 > c_6\sqrt{d(1-\alpha_t)\log T}$, Then, by calculations in expression (60), $g_t(x_t,\widehat{\alpha})$ can be written as

$$\begin{aligned} g_t(x_t,\widehat{\alpha}) =& \sqrt{\widehat{\alpha}}\Big(\frac{x_t}{\sqrt{\overline{\alpha}_t}} + \frac{1}{2}\int_{\overline{\alpha}_t}^{\widehat{\alpha}}\sqrt{\frac{1}{\widetilde{\alpha}^3}}\big(s_{\widetilde{\alpha}}(g_t(x_t,\widetilde{\alpha})) - s_t(x_t)\big)\mathrm{d}\widetilde{\alpha}\Big)\\ =& \sqrt{\frac{\widehat{\alpha}}{\overline{\alpha}_t}}x_t + O\bigg(\sqrt{\frac{d(1-\alpha_t)^2\log T}{1-\overline{\alpha}_t}}\bigg)\\ \le& \sqrt{\frac{\widehat{\alpha}}{\overline{\alpha}_t}}x_t + O(\sqrt{d(1-\alpha_t)\log T}). \end{aligned}$$

  which is contradicted with the assumption $\|g_t(x_t,\overline{\alpha}) - \sqrt{\overline{\alpha}/\overline{\alpha}_t}x_t\|_2 > c_5\sqrt{d(1-\alpha_t)\log T}$.

- Otherwise, consider the case that $\|s_{\widehat{\alpha}}(g_t(x_t,\widehat{\alpha}))\|_2^2 > \frac{c_5 d\log T}{1-\overline{\alpha}_t}$. For $\overline{\alpha}_t \le \widehat{\alpha} < \overline{\alpha}$, by inequality (61), we directly obtain

$$\|s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha}))\|_2 \le O\bigg((1-\alpha_t)\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{3/2}\bigg) + O\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{1/2} = O\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{1/2},$$

  where we use the fact that

$$\|s_t(x_t)\|^2 \lesssim \frac{d\log T}{1-\overline{\alpha}_t}, \tag{62}$$

  whose proof can be found as in (128b) of (Li et al., 2023, Appendix B.1.1). We can then make use of the continuity of $s_\alpha(x)$ and trajectory to obtain $\|s_{\widehat{\alpha}}(g_t(x_t,\widehat{\alpha}))\|_2 \lesssim (\frac{d\log T}{1-\overline{\alpha}_t})^{1/2}$. This result is also contradicted with the definition of $\widehat{\alpha}$.

Putting everything together, we conclude that $\widehat{\alpha} \in [\overline{\alpha}_t, \overline{\alpha}_{t-1}]$ does not exist, which thus validates the claim (59).

On the complement of the typical event $\mathcal{E}_t^c$. Let us now turn to the case when $(x_t, x_{t-1}) \in \mathcal{E}_t^c$. Using the upper bound in Lemma A.3, we integrate over the tail event of X_t and X_{t-1} as inequality (33) to obtain

$$\begin{aligned} \mathbb{E}_{X_t,X_{t-1}}\Big[\|s_t(X_t)\|_2^2\mathbf{1}\big((X_t,X_{t-1}) \in \mathcal{E}_t^c\big)\Big] &\lesssim \int_{\mathcal{E}_t^c}\|s_t(X_t)\|_2^2 p_{X_{t-1},X_t}(x_{t-1},x_t)\mathrm{d}x_{t-1}\mathrm{d}x_t\\ &\lesssim \int_{\mathcal{E}_t^c}\|s_t(X_t)\|_2^2 p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)p_{X_t}(x_t)\mathrm{d}x_{t-1}\mathrm{d}x_t\\ &\lesssim \frac{d}{1-\overline{\alpha}_t}\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}_t^c} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\mathrm{d}x_{t-1}. \end{aligned} \tag{63}$$

It has been shown in (Li et al., 2023, Step 3, Appendix C.1) that

$$\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}_t^c} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)dx_{t-1} \lesssim \exp(-c_4 d\log T).$$

By virtue of this relation, we can conclude that

$$\mathbb{E}_{X_t,X_{t-1}}\big[\|s_t(X_t)\|_2^2\mathbf{1}\big((X_t,X_{t-1}) \in \mathcal{E}_t^c\big)\big] \lesssim \exp(-c_4 d\log T). \tag{64}$$

Here, similar to the proof of Lemma A.3, it holds that

$$\mathbb{E}_{X_t}\|s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha}))\|_2^2 = \mathbb{E}_{X(\overline{\alpha})}\|s_{\overline{\alpha}}(X(\overline{\alpha}))\|_2^2 \le \frac{d}{1-\overline{\alpha}},$$

where use the fact that $g_t(X_t,\overline{\alpha}) \overset{\mathsf{d}}{=} X(\overline{\alpha})$. As a result, this inequality enables us to bound the expectation of the truncation error in a similar way as in inequality (64):

$$\begin{aligned}
&\mathbb{E}_{X_t,X_{t-1}}\Big[\|s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha}))\|_2^2\mathbf{1}\big((X_t,X_{t-1})\in\mathcal{E}_t^c\big)\Big]\\
&\lesssim \int_{\mathcal{E}_t^c}\|s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha}))\|_2^2 p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)p_{X_t}(x_t)\mathrm{d}x_{t-1}\mathrm{d}x_t\\
&\lesssim \frac{d}{1-\overline{\alpha}_t}\int_{x_{t-1}:(x_t,x_{t-1})\in\mathcal{E}_t^c} p_{X_{t-1}\,|\,X_t}(x_{t-1}\,|\,x_t)\mathrm{d}x_{t-1}\\
&\lesssim \exp(-c_4 d\log T).
\end{aligned}$$

In summary. Combining the two cases above, we conclude that

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha}))-s_t(X_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2^2\Big]\\
&\le \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\mathbb{E}\big[\|s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha}))-s_t(X_t)\|_2^2\big]\mathrm{d}\overline{\alpha}\\
&= \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\mathbb{E}\Big[\|s_{\overline{\alpha}}(g_t(X_t,\overline{\alpha}))-s_t(X_t)\|_2^2\big(\mathbf{1}\big((X_t,X_{t-1})\in\mathcal{E}_t\big)+\mathbf{1}\big((X_t,X_{t-1})\in\mathcal{E}_t^c\big)\big)\Big]\mathrm{d}\overline{\alpha}\\
&\lesssim \frac{(1-\alpha_t)^4d^3\log^3 T}{(1-\overline{\alpha}_t)^3}+\exp(-c_4 d\log T)\\
&\lesssim \frac{(1-\alpha_t)^4d^3\log^3 T}{(1-\overline{\alpha}_t)^3},
\end{aligned}$$

which thus validates the claimed result.

C.4 Proof of Claim (53)

Towards this, let us first make the observation that

$$\begin{aligned}
\frac{\partial\Phi_{t\to k}}{\partial x}(x) &= \frac{\partial\Phi_{t-1\to k}}{\partial x}\big(\phi_t(x)\big)\frac{\partial\phi_t}{\partial x}(x)\\
&= \frac{\partial\Phi_{t-2\to k}}{\partial x}\big(\Phi_{t\to t-2}(x)\big)\frac{\partial\phi_{t-1}}{\partial x}\big(\Phi_{t\to t-1}(x)\big)\frac{\partial\phi_t}{\partial x}(x)\\
&= \frac{\partial\Phi_{k'\to k}}{\partial x}\big(\Phi_{t\to k'}(x)\big)\prod_{i=k}^{k'}\frac{\partial\phi_i}{\partial x}\big(\Phi_{t\to i}(x)\big)\\
&= \prod_{i=k+1}^{t}\frac{\partial\phi_i}{\partial x}\big(\Phi_{t\to i}(x)\big),
\end{aligned}$$

where we recursively apply the definition of $\Phi_{k'\to k}=\phi_{k'}\circ\Phi_{k'-1\to k}$. In view of the relation above, by some direct algebra, we deduce

$$\begin{aligned}
&\left\|\frac{\partial\Phi_t}{\partial x}(x)-\frac{\partial\Phi_t}{\partial x}(y)\right\|\\
&= \left\|\prod_{i=2}^{t}\frac{\partial\phi_i}{\partial x}\big(\Phi_{t\to i}(x)\big)-\prod_{i=2}^{t}\frac{\partial\phi_i}{\partial x}\big(\Phi_{t\to i}(y)\big)\right\|
\end{aligned}$$

$$
\begin{aligned}
&= \Big\| \sum_{k=3}^{t} \Big(\prod_{i=2}^{k-1} \frac{\partial \phi_i}{\partial x}\big(\Phi_{t\to i}(x)\big) \Big(\frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(x)\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(y)\big) \Big) \prod_{i=k+1}^{t} \frac{\partial \phi_i}{\partial x}\big(\Phi_{t\to i}(y)\big) \Big) \Big\| \\
&\le \sum_{k=2}^{t} \Big\| \prod_{i=2}^{k-1} \frac{\partial \phi_i}{\partial x}\big(\Phi_{t\to i}(x)\big) \Big\| \Big\| \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(x)\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(y)\big) \Big\| \Big\| \prod_{i=k+1}^{t} \frac{\partial \phi_i}{\partial x}\big(\Phi_{t\to i}(y)\big) \Big\| \\
&= \sum_{k=2}^{t} \Big\| \frac{\partial \Phi_{k-1}}{\partial x}\big(\Phi_{t\to k-1}(x)\big) \Big\| \Big\| \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(x)\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(y)\big) \Big\| \Big\| \frac{\partial \Phi_{t\to k}}{\partial x}(y) \Big\| \\
&\le L_f^2 \sum_{k=2}^{t} \Big\| \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(x)\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(y)\big) \Big\|.
\end{aligned}
$$

where we denote $\prod_i^{i-1}(\partial\phi_i/\partial x) := 1$ for saimplicity, and the last invokes the Assumption 3.1 again.

C.5 Proof of Lemma A.5

To begin with, let us first provide a more succinct expression for quantity $\frac{\partial \phi_k(x)}{\partial x}$. Recall that $\phi_k(x) := g_k(x, \overline{\alpha}_{k-1})$. In view of relation (42), we can write

$$
\begin{aligned}
\phi_k(x) &= \sqrt{\overline{\alpha}_{k-1}} \left(\frac{g_k(x,\overline{\alpha}_k)}{\sqrt{\overline{\alpha}_k}} + \int_{\overline{\alpha}_k}^{\overline{\alpha}_{k-1}} \frac{1}{2\overline{\alpha}^{\frac{3}{2}}} s_{\overline{\alpha}}(g_k(x,\overline{\alpha})) \mathrm{d}\overline{\alpha} \right) \\
&= \frac{1}{\sqrt{\alpha_k}} \left(x + \frac{1}{2} \int_{\overline{\alpha}_k}^{\overline{\alpha}_{k-1}} \sqrt{\frac{\overline{\alpha}_k}{\overline{\alpha}^3}} s_{\overline{\alpha}}(g_k(x,\overline{\alpha})) \mathrm{d}\overline{\alpha} \right). \qquad (65)
\end{aligned}
$$

By some direct calculations, we arrive at

$$
\frac{\partial \phi_k(x)}{\partial x} = \frac{1}{\sqrt{\alpha_k}} \left(I + \frac{1}{2} \int_{\overline{\alpha}_k}^{\overline{\alpha}_{k-1}} \sqrt{\frac{\overline{\alpha}_k}{\overline{\alpha}^3}} \nabla s_{\overline{\alpha}}(g_k(x,\overline{\alpha})) \frac{\partial g_k(x,\overline{\alpha})}{\partial x} \mathrm{d}\overline{\alpha} \right) \qquad (66)
$$

where we write $\nabla s_{\overline{\alpha}}(g_k(x,\overline{\alpha})) := \nabla_y s_{\overline{\alpha}}(y)\big|_{y=g_k(x,\overline{\alpha})}$. We then proceed to control each term in the above expression. To do so, let us introduce the following two lemmas whose proofs are provided in Section C.7 and C.8 respectively.

Lemma C.1. *For $2 \le t \le T$, $\overline{\alpha}_t \le \overline{\alpha} \le \overline{\alpha}_{t-1}$ and $(x_t, x_{t-1}) \in \mathcal{E}_t$, the derivative of the score function satisfies*

$$
\big\| \nabla s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha})) - \nabla s_t(x_t) \big\|_2 \lesssim \frac{d^2(1-\alpha_t)\log^2 T}{(1-\overline{\alpha}_t)^2}.
$$

Lemma C.2. *For $2 \le t \le T$ and $(x_t, x_{t-1}) \in \mathcal{E}_t$, the stability of the backward ODE (42) starting at x_t can be bounded as follows:*

$$
\left\| \frac{\partial g_t(x_t,\overline{\alpha})}{\partial x} - I \right\| \lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t}.
$$

Plugging in the bounds from Lemma C.1 and Lemma C.2 to equation (66), we obtain

$$
\begin{aligned}
\frac{\partial \phi_k(x)}{\partial x} &= \frac{1}{\sqrt{\alpha_k}} \Bigg(I + \frac{1}{2} \int_{\overline{\alpha}_k}^{\overline{\alpha}_{k-1}} \sqrt{\frac{\overline{\alpha}_k}{\overline{\alpha}^3}} \Big(\frac{\partial s_k(x)}{\partial x} + O\Big(\frac{d^2(1-\alpha_k)\log^2 T}{(1-\overline{\alpha}_k)^2} \Big) \Big) \\
&\qquad\qquad\qquad\qquad \Big(I_d + O\Big(\frac{d(1-\alpha_k)\log T}{1-\overline{\alpha}_t} \Big) \Big) \mathrm{d}\overline{\alpha} \Bigg) \\
&= \frac{1}{\sqrt{\alpha_k}} \left(I + \frac{1}{2} \int_{\overline{\alpha}_k}^{\overline{\alpha}_{k-1}} \sqrt{\frac{\overline{\alpha}_k}{\overline{\alpha}^3}} \frac{\partial s_k(x)}{\partial x} \mathrm{d}\overline{\alpha} \right) + O\left(\frac{d^2(1-\alpha_k)^2\log^2 T}{(1-\overline{\alpha}_k)^2} \right) \\
&= \frac{1}{\sqrt{\alpha_k}} \left(I - \frac{1-\sqrt{\alpha_k}}{1-\overline{\alpha}_k} J_k(x) \right) + O\left(\frac{d^2(1-\alpha_k)^2\log^2 T}{(1-\overline{\alpha}_k)^2} \right), \qquad (67)
\end{aligned}
$$

where we denote

$$J_k(x) := I_d + \frac{1}{1-\overline{\alpha}_k}\Big\{\mathbb{E}\big[X_k - \sqrt{\overline{\alpha}_k}X_0 \mid X_k = x\big]\Big(\mathbb{E}\big[X_k - \sqrt{\overline{\alpha}_k}X_0 \mid X_k = x\big]\Big)^\top$$
$$- \mathbb{E}\Big[\big(X_k - \sqrt{\overline{\alpha}_k}X_0\big)\big(X_k - \sqrt{\overline{\alpha}_k}X_0\big)^\top \mid X_k = x\Big]\Big\}. \tag{68}$$

The details for deriving expression (67) are included in Section C.6.

In order to prove Lemma A.5 and cope with the difference $\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)$, inequality (67) suggests to study the Lipschitz property of function J_k. For this purpose, we introduce our final auxiliary result, whose proof is provided in Section C.9.

Lemma C.3. *For $2 \le t \le T$ and $(x_t, x_{t-1}) \in \mathcal{E}_t$, $J_t(x)$ is locally Lipschitz continuous with respect to x:*

$$\|J_t(x_{t-1}) - J_t(\phi(x_t))\| \lesssim \frac{1}{\sqrt{1-\overline{\alpha}_t}} d^{3/2}\log^{3/2} T\|x_{t-1} - \phi(x_t)\|_2. \tag{69a}$$

In addition, for $1 \le k \le t-1$, the Lipschitz constant along the backward trajectory satisfies

$$\Big\|J_k\big(\Phi_{t-1\to k}(x_{t-1}(\gamma))\big) - J_k\big(\Phi_{t\to k}(x_t)\big)\Big\|$$
$$\lesssim \frac{1}{\sqrt{1-\overline{\alpha}_t}} d^{3/2}\log^{3/2} T\Big\|\Phi_{t-1\to k}(x_{t-1}(\gamma)) - \Phi_{t\to k}(x_t)\Big\|_2. \tag{69b}$$

To proceed, let us again decompose the quantity of interest as

$$\mathbb{E}\Bigg[\bigg\|\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\bigg\|^2\Bigg]$$
$$= \mathbb{E}\Bigg[\bigg\|\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\bigg\|^2 \mathbf{1}\big((X_t, X_{t-1}) \in \mathcal{E}_t\big)\Bigg]$$
$$+ \mathbb{E}\Bigg[\bigg\|\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\bigg\|^2 \mathbf{1}\big((X_t, X_{t-1}) \in \mathcal{E}_t^c\big)\Bigg]. \tag{70}$$

We shall control each term respectively.

The first term. Taking Lemma C.3 collectively with expression (67), we obtain

$$\mathbb{E}\Bigg[\bigg\|\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\bigg\|^2 \mathbf{1}\big((X_t, X_{t-1}) \in \mathcal{E}_t\big)\Bigg]$$
$$\overset{\text{(i)}}{\lesssim} \frac{(1-\alpha_k)^2}{(1-\overline{\alpha}_k)^2}\mathbb{E}\Big[\Big\|J_k\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - J_k\big(\Phi_{t\to k}(X_t)\big)\Big\|^2 \mathbf{1}\big((X_t, X_{t-1}) \in \mathcal{E}_t\big)\Big]$$
$$+ \frac{d^4(1-\alpha_k)^4\log^4 T}{(1-\overline{\alpha}_k)^4}$$
$$\overset{\text{(ii)}}{\lesssim} \frac{(1-\alpha_k)^2 d^3\log^3 T}{(1-\overline{\alpha}_k)^2(1-\overline{\alpha}_t)}\mathbb{E}\Big[\Big\|\Phi_{t-1\to k}(X_{t-1}(\gamma)) - \Phi_{t\to k}(X_t)\Big\|_2^2 \mathbf{1}\big((X_t, X_{t-1}) \in \mathcal{E}_t\big)\Big]$$
$$+ \frac{d^4(1-\alpha_k)^4\log^4 T}{(1-\overline{\alpha}_k)^4}$$
$$\overset{\text{(iii)}}{\lesssim} \frac{(1-\alpha_k)^2 d^3\log^3 T}{(1-\overline{\alpha}_k)^2(1-\overline{\alpha}_t)}\mathbb{E}\Big[L_f^2\big\|X_{t-1}(\gamma) - \phi(X_t)\big\|_2^2 \mathbf{1}\big((X_t, X_{t-1}) \in \mathcal{E}_t\big)\Big] + \frac{d^4(1-\alpha_k)^4\log^4 T}{(1-\overline{\alpha}_k)^4}. \tag{71}$$

Note that, to ensure inequalities (i) and (ii), one invokes Lemma C.3 which requires $(x_k, x_{k-1}) \in \mathcal{E}_k$. We shall verify this relation momentarily. In (ii) we invoke the Lipschitz continuity of $\Phi_{t-1\to k}$ and $\Phi_{t\to k}$ and the property that $\Phi_{t\to k}(X_t) \overset{d}{=} X_{t-1}$. To further control the right hand side above, recall that we have established the inequality (52) when (x_t, x_{t-1}) in $\mathcal{E}_t$. As a result, we conclude that

$$\mathbb{E}\Bigg[\bigg\|\frac{\partial \phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big) - \frac{\partial \phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\bigg\|^2 \mathbf{1}\big((X_t, X_{t-1}) \in \mathcal{E}_t\big)\Bigg]$$

$$\lesssim \frac{(1-\alpha_k)^2(1-\alpha_t)^2L_f^2d^4\log^3 T}{(1-\overline{\alpha}_k)^2(1-\overline{\alpha}_t)^2}+\frac{(1-\alpha_k)^4d^4\log^4 T}{(1-\overline{\alpha}_k)^4}. \tag{72}$$

It is therefore only left for us to show that (x_k, x_{k-1}) in $\mathcal{E}_k$, which holds true owing to the Lipschitz property of $\Phi_{t-1\to k}(x)$ and $\Phi_{t\to k}(x)$. Specifically, for every (x_t, x_{t-1}) in $\mathcal{E}_t$, by definition, it holds for large enough constant c_4 that

$$\|x_{t-1}-x_t/\sqrt{\alpha_t}\|_2\le c_4\sqrt{d(1-\alpha_t)\log T}.$$

The Lipschitz continuity of $\Phi_{t\to k}$ also implies that $-\log p_{X_k}(x_k)\le c_3 d\log T$ as $X_k \overset{\mathrm{d}}{=} \Phi_{t\to k}(X_t)$. As a result, if we define

$$\mathcal{E}_k' := \bigg\{(x_k, x_{k-1})\in\mathbb{R}^d\times\mathbb{R}^d \ \Big|\ -\log p_{X_k}(x_k)\le c_3 d\log T,$$
$$\|x_{k-1}-x_k/\sqrt{\alpha_k}\|_2\le c_4L_f\sqrt{d(1-\alpha_t)\log T}\bigg\},$$

then one can check that $(\Phi_{t-1\to k}(x_{t-1}), \Phi_{t\to k}(x_t))\in\mathcal{E}_k'$. Notice that $\mathcal{E}_k$ and $\mathcal{E}_k'$ share the same form for every $2\le k<t\le T$, only with a different constant in the second condition, we conclude that Lemma C.1, C.2 and C.3 still hold true with slight different constants. Therefore, we have validated the relation (72).

The second term. When $(x_t, x_{t-1})\in\mathcal{E}^c$ holds true, it is sufficient to consider a crude upper bound for

$$\bigg\|\frac{\partial\phi_k}{\partial x}\big(\Phi_{t-1\to k}(x_{t-1}(\gamma))\big)-\frac{\partial\phi_k}{\partial x}\big(\Phi_{t\to k}(x_t)\big)\bigg\|^2\mathbf{1}\big((x_t,x_{t-1})\in\mathcal{E}_t^c\big).$$

Owing to the Lipschitz condition in Assumption 3.1, we know that $\frac{\partial}{\partial x}\Phi_{t\to k}(x)\le L_f$. Simply choosing $k=t-1$ gives us $\frac{\partial}{\partial x}\phi_t(x)\le L_f$, which in turn leads to

$$\begin{aligned}&\mathbb{E}\bigg[\bigg\|\frac{\partial\phi_k}{\partial x}\big(\Phi_{t-1\to k}(X_{t-1}(\gamma))\big)-\frac{\partial\phi_k}{\partial x}\big(\Phi_{t\to k}(X_t)\big)\bigg\|^2\mathbf{1}\big((X_t,X_{t-1})\in\mathcal{E}_t^c\big)\bigg]\\ \le&4L_f^2\mathbb{P}\big((X_t,X_{t-1})\in\mathcal{E}_t^c\big)\\ \lesssim&L_f^2\exp(-c_4 d\log T).\end{aligned} \tag{73}$$

Putting relations (72) and (73) together verifies the target result in Lemma A.5.

C.6 Proof of Claim (67)

To establish this relation, we first find it useful to write the score function as

$$\begin{aligned}s_t(x)&=\mathbb{E}\Big[-\frac{1}{\sqrt{1-\overline{\alpha}_t}}Z\ \Big|\ \sqrt{\overline{\alpha}_t}X_0+\sqrt{1-\overline{\alpha}_t}Z=x\Big]\\ &=-\frac{1}{1-\overline{\alpha}_t}\mathbb{E}\Big[x-\sqrt{\overline{\alpha}_t}X_0\ \Big|\ \sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}z=x\Big]\\ &=-\frac{1}{1-\overline{\alpha}_t}\int_{x_0}(x-\sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0.\end{aligned} \tag{74}$$

As a result, the partial derivative is calculated as

$$\begin{aligned}\frac{\partial[s_t(x)]_i}{\partial x_j}&=-\frac{1}{1-\overline{\alpha}_t}\frac{\partial}{\partial x_j}\Big[\int_{x_0}(x_i-\sqrt{\overline{\alpha}_t}x_{0,i})p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0\Big]\\ &=-\frac{1}{1-\overline{\alpha}_t}\Big[\mathbf{1}_{\{i=j\}}+\int_{x_0}(x_i-\sqrt{\overline{\alpha}_t}x_{0,i})\frac{\partial}{\partial x_j}p_{X_0\,|\,X_t}(x_0\,|\,x)\Big]\mathrm{d}x_0\\ &=-\frac{1}{1-\overline{\alpha}_t}\Big[\mathbf{1}_{\{i=j\}}+\int_{x_0}(x_i-\sqrt{\overline{\alpha}_t}x_{0,i})\frac{\partial}{\partial x_j}\frac{p_{X_0}(x_0)p_{X_t\,|\,X_0}(x\,|\,x_0)}{p_{X_t}(x)}\Big]\mathrm{d}x_0.\end{aligned} \tag{75}$$

By noticing the fact that

$$\frac{\partial}{\partial x_j} p_{X_t \,|\, X_0}(x \,|\, x_0) = p_{X_t \,|\, X_0}(x \,|\, x_0) \cdot \frac{x_j - \sqrt{\overline{\alpha}_t} x_{0,j}}{1 - \overline{\alpha}_t}, \tag{76}$$

we can thus rewrite equation (75) as

$$\begin{aligned}
\frac{\partial [s_t(x)]_i}{\partial x_j} =& - \frac{1}{1 - \overline{\alpha}_t} \Big[\mathbf{1}_{\{i=j\}} + \frac{1}{1 - \overline{\alpha}_t} \Big[\int_{x_0} (x_i - \sqrt{\overline{\alpha}_t} x_{0,i}) p_{X_0 \,|\, X_t}(x_0 \,|\, x) \mathrm{d}x_0 \cdot \\
& \qquad \qquad \qquad \qquad \int_{x_0} (x_j - \sqrt{\overline{\alpha}_t} x_{0,j}) p_{X_0 \,|\, X_t}(x_0 \,|\, x) \mathrm{d}x_0 \\
& \qquad - \int_{x_0} (x_i - \sqrt{\overline{\alpha}_t} x_{0,i})(x_j - \sqrt{\overline{\alpha}_t} x_{0,j}) p_{X_0 \,|\, X_t}(x_0 \,|\, x) \mathrm{d}x_0 \Big] \Big] \qquad (77) \\
=& - \frac{1}{1 - \overline{\alpha}_t} [J_t(x)]_{ij}, \qquad (78)
\end{aligned}$$

which leads to equation (67).

C.7 Proof of Lemma C.1

The proof of Claim (67) provides an explicit expression of $\frac{\partial s_t(x)}{\partial x}$ via $J_t(x)$ as in expression (77). Organizing terms of expression (77) gives us

$$\begin{aligned}
& \nabla s_t(x_t) + \frac{1}{1 - \overline{\alpha}_t} I_d \\
=& - \frac{1}{(1 - \overline{\alpha}_t)^2} \Big[\underbrace{\int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0) p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \mathrm{d}x_0 \Big(\int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0) p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \mathrm{d}x_0 \Big)^\top}_{=: A_t} \\
& \qquad - \underbrace{\int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0)(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \mathrm{d}x_0}_{=: B_t} \Big]
\end{aligned}$$

and similarly, it holds that

$$\begin{aligned}
& \nabla s_{\overline{\alpha}}(g_t(x_t, \overline{\alpha})) + \frac{1}{1 - \overline{\alpha}} I_d \\
=& - \frac{1}{(1 - \overline{\alpha})^2} \Big[\underbrace{\int_{x_0} (g_t(x_t, \overline{\alpha}) - \sqrt{\overline{\alpha}} x_0) p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t, \overline{\alpha})) \mathrm{d}x_0 \Big(\int_{x_0} (g_t(x_t, \overline{\alpha}) - \sqrt{\overline{\alpha}} x_0) p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t, \overline{\alpha})) \mathrm{d}x_0 \Big)^\top}_{=: A_{\overline{\alpha}}} \\
& \qquad - \underbrace{\int_{x_0} (g_t(x_t, \overline{\alpha}) - \sqrt{\overline{\alpha}} x_0)(g_t(x_t, \overline{\alpha}) - \sqrt{\overline{\alpha}} x_0)^\top p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t, \overline{\alpha})) \mathrm{d}x_0}_{=: B_{\overline{\alpha}}} \Big].
\end{aligned}$$

In view of these two decompositions, we can bound

$$\Big\| \nabla s_{\overline{\alpha}}(g_t(x_t, \overline{\alpha})) - \nabla s_t(x_t) \Big\| \leq \Big\| \frac{1}{(1 - \overline{\alpha})^2} A_\alpha - \frac{1}{(1 - \overline{\alpha}_t)^2} A_t \Big\| + \Big\| \frac{1}{(1 - \overline{\alpha})^2} B_\alpha - \frac{1}{(1 - \overline{\alpha}_t)^2} B_t \Big\|. \tag{79}$$

We shall proceed by controlling each term on the right respectively.

Controlling the first term. Let us start by bounding the first term. By noticing the basic algebra fact that for vectors $z_1, z_2 \in \mathbb{R}^d$,

$$\| z_1 z_1^\top - z_2 z_2^\top \|_2 \leq \| z_1 - z_2 \|_2 \cdot \max\{ \| z_1 \|_2, \| z_2 \|_2 \},$$

we find

$$\begin{aligned}
&\Big\| \frac{1}{(1-\overline{\alpha}_t)^2} A_t - \frac{1}{(1-\overline{\alpha})^2} A_\alpha \Big\| \\
&\lesssim \Big\| \frac{1}{1-\overline{\alpha}_t} \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0) p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \mathrm{d}x_0 \\
&\qquad\qquad\qquad - \frac{1}{1-\overline{\alpha}} \int_{x_0} (g_t(x_t,\overline{\alpha}) - \sqrt{\overline{\alpha}} x_0) p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x,\overline{\alpha})) \mathrm{d}x_0 \Big\|_2 \cdot \\
&\qquad\qquad\qquad\qquad \Big(\max\{ \| s_{\overline{\alpha}}(g_t(x,\overline{\alpha})) \|_2, \| s_t(x_t) \|_2 \} \Big). \qquad (80)
\end{aligned}$$

By virtue of the bound (62), we can directly derive

$$\max\{ \| s_{\overline{\alpha}}(g_t(x,\overline{\alpha})) \|_2^2, \| s_t(x_t) \|_2^2 \} \lesssim \max\Big\{ \frac{d \log T}{1-\overline{\alpha}}, \frac{d \log T}{1-\overline{\alpha}_t} \Big\} = \frac{d \log T}{1-\overline{\alpha}_t}. \qquad (81)$$

It is then sufficient to control the first term on the right hand side of inequality (80), which shall be done as follows. To this end, let us define a set of interest by

$$\mathcal{E}_0 \coloneqq \big\{ x : \| x_t - \sqrt{\overline{\alpha}_t} x \|_2 \le c_6 \sqrt{d(1-\overline{\alpha}_t) \log T} \big\}.$$

We first consider the the following term

$$\begin{aligned}
&\Big\| \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0) p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \mathrm{d}x_0 - \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0) p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x,\overline{\alpha})) \mathrm{d}x_0 \Big\|_2 \\
&\le \int_{x_0 \in \mathcal{E}_0} \big| p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t,\overline{\alpha})) - p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \big| \cdot \big\| x_t - \sqrt{\overline{\alpha}_t} x_0 \big\|_2 \mathrm{d}x_0 \\
&\qquad + \int_{x_0 \in \mathcal{E}_0^c} \big| p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t,\overline{\alpha})) - p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \big| \cdot \big\| x_t - \sqrt{\overline{\alpha}_t} x_0 \big\|_2 \mathrm{d}x_0 \\
&= \int_{x_0 \in \mathcal{E}_0} \bigg| \frac{p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t,\overline{\alpha}))}{p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)} - 1 \bigg| \cdot p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \cdot \big\| x_t - \sqrt{\overline{\alpha}_t} x_0 \big\|_2 \mathrm{d}x_0 \\
&\qquad \int_{x_0 \in \mathcal{E}_0^c} \bigg| \frac{p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t,\overline{\alpha}))}{p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)} - 1 \bigg| \cdot p_{X_0 \,|\, X_t}(x_0 \,|\, x_t) \cdot \big\| x_t - \sqrt{\overline{\alpha}_t} x_0 \big\|_2 \mathrm{d}x_0. \qquad (82)
\end{aligned}$$

Next, we bound the right hand side above. Towards this, first recall that in Claim 2 in (Li et al., 2023, Appendix C.1), it has been shown by direct calculations that

$$\begin{aligned}
\frac{p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t,\overline{\alpha}))}{p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)} &= 1 + O\bigg(\frac{d(1-\alpha_t) \log T}{1-\overline{\alpha}_{t-1}} \bigg), && \text{if } x_0 \in \mathcal{E}_0, \qquad (83) \\
\frac{p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t,\overline{\alpha}))}{p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)} &\le \exp\left(\frac{16 c_1 \big\| x_t - \sqrt{\overline{\alpha}_t} x_0 \big\|_2^2 \log T}{(1-\overline{\alpha}_t) T} \right), && \text{if } x_0 \notin \mathcal{E}_0. \qquad (84)
\end{aligned}$$

Here, we remark that we replace $\sqrt{\overline{\alpha}/\overline{\alpha}_t}\, x_t$ in Li et al. (2023) by $g_t(x_t,\overline{\alpha})$. This is valid since for $(x_t, x_{t-1}) \in \mathcal{E}_t$, inequality (60) ensures

$$\Big\| \sqrt{\frac{\overline{\alpha}}{\overline{\alpha}_t}} x_t - g_t(x_t,\overline{\alpha}) \Big\|_2 = O\Big(\frac{d^{1/2}(1-\alpha_t) \log^{1/2} T}{(1-\overline{\alpha}_t)^{1/2}} \Big).$$

This approximation only leads to a lower order term in our final result.

Plugging the relations (83) and (84) into the right hand side of (82) and following the proof of (161c) in the proof in (Li et al., 2023, Appendix C.1), we can obtain

$$\begin{aligned}
&\Big\| \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0) p_{X_0 \,|\, X_t}(x_0 \,|\, x) \mathrm{d}x_0 - \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t} x_0) p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x,\overline{\alpha})) \mathrm{d}x_0 \Big\|_2 \\
&\lesssim \frac{d(1-\alpha_t) \log T}{1-\overline{\alpha}_{t-1}} \cdot \mathbb{E}\left[\big\| \sqrt{\overline{\alpha}_t} X_0 - x_t \big\|_2 \,\big|\, X_t = x_t \right]
\end{aligned}$$

$$\lesssim \frac{d^{3/2}(1-\alpha_t)\log^{3/2}T}{(1-\overline{\alpha}_t)^{1/2}}, \tag{85}$$

where we apply Lemma A.1 to deduce the last inequality. With the same calculations, we can similarly find

$$\begin{aligned}
&\Big\| \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t}x_0)(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\mathrm{d}x_0 \\
&\qquad\qquad - \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t}x_0)(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x, \overline{\alpha}))\mathrm{d}x_0 \Big\| \\
&\lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_{t-1}} \cdot \mathbb{E}\Big[\big\|\sqrt{\overline{\alpha}_t}X_0 - x_t\big\|_2^2 \,|\, X_t = x_t \Big] \\
&\lesssim \frac{d^2(1-\alpha_t)(1-\overline{\alpha}_t)\log^2 T}{(1-\overline{\alpha}_{t-1})} \lesssim d^2(1-\alpha_t)\log^2 T.
\end{aligned} \tag{86}$$

With these properties in place, we are ready to prove Lemma C.2 by studying the first term in (80). First, notice that

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

Now applying the inequality (85), Lemma A.1, and (60) on each term above separately, we achieve

$$\begin{aligned}
&\Big\| \frac{1}{1-\overline{\alpha}_t} \int_{x_0} (x_t - \sqrt{\overline{\alpha}_t}x_0) p_{X_0 \,|\, X_t}(x_0 \,|\, x_t)\mathrm{d}x_0 \\
&\qquad\qquad - \frac{1}{1-\overline{\alpha}} \int_{x_0} (g_t(x_t, \overline{\alpha}) - \sqrt{\overline{\alpha}}x_0) p_{X_0 \,|\, X_{\overline{\alpha}}}(x_0 \,|\, g_t(x_t, \overline{\alpha}))\mathrm{d}x_0 \Big\|_2 \\
&\lesssim \frac{1}{1-\overline{\alpha}_t} \cdot \frac{d^{3/2}(1-\alpha_t)\log^{3/2}T}{(1-\overline{\alpha}_t)^{1/2}} + \frac{(1-\alpha_t)}{\sqrt{\alpha_t}(1-\overline{\alpha}_t)^2} \mathbb{E}\left[\big\|\sqrt{\overline{\alpha}_t}X_0 - x_t\big\|_2 \,|\, X_t = x_t \right] \\
&\qquad\qquad\qquad + \frac{1}{1-\overline{\alpha}_t} \sup_{\overline{\alpha}_t < \overline{\alpha} < \overline{\alpha}_{t-1}} \Big\| \sqrt{\frac{\overline{\alpha}}{\overline{\alpha}_t}}x_t - g_t(x_t, \overline{\alpha}) \Big\|_2 \\
&\lesssim \frac{d^{3/2}(1-\alpha_t)\log^{3/2}T}{(1-\overline{\alpha}_t)^{3/2}} + \frac{d^{1/2}(1-\alpha_t)\log^{1/2}T}{(1-\overline{\alpha}_t)^{3/2}} + \frac{d^{1/2}(1-\alpha_t)\log^{1/2}T}{(1-\overline{\alpha}_t)^{3/2}}
\end{aligned}$$

$$\lesssim \frac{d^{3/2}(1-\alpha_t)\log^{3/2} T}{(1-\overline{\alpha}_t)^{3/2}}. \tag{87}$$

Finally, plugging inequalities (81) and (87) into expression (80) leads to

$$\begin{aligned}\Big\|\frac{1}{(1-\overline{\alpha})^2}A_\alpha - \frac{1}{(1-\overline{\alpha}_t)^2}A_t\Big\| &\lesssim \frac{d^{3/2}(1-\alpha_t)\log^{3/2} T}{(1-\overline{\alpha}_t)^{3/2}} \cdot \frac{d^{1/2}\log^{1/2} T}{(1-\overline{\alpha}_t)^{1/2}} \\ &\lesssim \frac{d^2(1-\alpha_t)\log^2 T}{(1-\overline{\alpha}_t)^2}. \end{aligned} \tag{88}$$

Controlling the second term. With expression (86), we can further control the quantity $\|\frac{1}{(1-\overline{\alpha})^2}B_\alpha - \frac{1}{(1-\overline{\alpha}_t)^2}B_t\|$. By similar analysis, we can obtain

$$\begin{aligned}&\Big\|\frac{1}{(1-\overline{\alpha})^2}B_\alpha - \frac{1}{(1-\overline{\alpha}_t)^2}B_t\Big\| \\ &\lesssim \frac{1}{\alpha_t(1-\overline{\alpha}_t)^2}\Big\|\int_{x_0}(x_t-\sqrt{\overline{\alpha}_t}x_0)(x_t-\sqrt{\overline{\alpha}_t}x_0)^\top p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0 \\ &\qquad - \int_{x_0}(x_t-\sqrt{\overline{\alpha}_t}x_0)(x_t-\sqrt{\overline{\alpha}_t}x_0)^\top p_{X_0\,|\,X_{\overline{\alpha}}}(x_0\,|\,g_t(x,\overline{\alpha}))\mathrm{d}x_0\Big\| \\ &\qquad\qquad + \frac{d(1-\alpha_t)\log T}{(1-\overline{\alpha}_t)^2} \\ &\lesssim \frac{d^2(1-\alpha_t)\log^2 T}{(1-\overline{\alpha}_t)^2}. \end{aligned} \tag{89}$$

In summary, taking the relations (88) and (89) collectively with inequality (79), we obtain the following bound

$$\begin{aligned}\Big\|\nabla s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha})) - \nabla s_t(x_t)\Big\| &\le \Big\|\frac{1}{(1-\overline{\alpha})^2}A_\alpha - \frac{1}{(1-\overline{\alpha}_t)^2}A_t\Big\| + \Big\|\frac{1}{(1-\overline{\alpha})^2}B_\alpha - \frac{1}{(1-\overline{\alpha}_t)^2}B_t\Big\| \\ &\lesssim \frac{d^2(1-\alpha_t)\log^2 T}{(1-\overline{\alpha}_t)^2}, \end{aligned}$$

which leads to the final result.

C.8 Proof of Lemma C.2

The proof of this lemma is similar to that of Lemma A.4. In particular, we shall prove this result by contradiction. Specifically, suppose that there exists $\overline{\alpha} \in [\overline{\alpha}_t, \overline{\alpha}_{t-1}]$ such that Lemma C.2 does not hold. Then, one can define

$$\widehat{\alpha} := \min\Big\{\overline{\alpha} \in [\overline{\alpha}_t, \overline{\alpha}_{t-1}] : \Big\|\frac{\partial g_t(x_t,\overline{\alpha})}{\partial x} - I\Big\| \gtrsim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t}\Big\}.$$

With this definition of $\widehat{\alpha}$, it holds that for all $\overline{\alpha}_{t-1} \ge \overline{\alpha} > \widehat{\alpha}$, one has

$$\Big\|\frac{\partial g_t(x_t,\overline{\alpha})}{\partial x}\Big\| = 1 + O(d(1-\alpha_t)\log T). \tag{90}$$

Now consider the partial derivative of $g_t(x,\widehat{\alpha})$ at $\widehat{\alpha}$ where

$$\frac{\partial g_t(x,\widehat{\alpha})}{\partial x} - I_d = \Big(\frac{1}{\sqrt{\alpha_t}} - 1\Big)I_d + \frac{1}{2\sqrt{\alpha_t}}\int_{\overline{\alpha}_t}^{\widehat{\alpha}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\nabla s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha}))\frac{\partial g_t(x_t,\overline{\alpha})}{\partial x}\mathrm{d}\overline{\alpha}.$$

The proof in Lemma A.4 ensures that

$$\Big\|g_t(x_t,\overline{\alpha}) - \sqrt{\frac{\overline{\alpha}}{\overline{\alpha}_t}}x_t\Big\| \le c_5\sqrt{d(1-\alpha_t)\log T}$$

for $(x_t, x_{t-1}) \in \mathcal{E}_t$. Thus, the analysis in the proof of (161a) in (Li et al., 2023, Appendix C.1) guarantees that

$$\big\|(1-\overline{\alpha})\nabla s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha})) - I_d\big\| \lesssim d\log T, \tag{91}$$

which directly implies that

$$\big\|\nabla s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha}))\big\| \lesssim \frac{d\log T}{1-\overline{\alpha}_t},$$

Combining these results together, we obtain

$$\begin{aligned}\Big\|\frac{\partial g_t(x,\widehat{\alpha})}{\partial x} - I\Big\| &\leq \Big|\frac{1}{\sqrt{\alpha}_t} - 1\Big| + \frac{1}{2\sqrt{\alpha_t}}\int_{\overline{\alpha}_t}^{\widehat{\alpha}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big\|\nabla s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha}))\big\|\Big\|\frac{\partial g_t(x_t,\overline{\alpha})}{\partial x}\Big\|\mathrm{d}\overline{\alpha} \\ &\leq \Big|\frac{1}{\sqrt{\alpha}_t} - 1\Big| + \Big\|\frac{d\log T}{2\sqrt{\alpha_t}(1-\overline{\alpha}_t)}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\mathrm{d}\overline{\alpha}\Big\| \\ &\lesssim \frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t},\end{aligned}$$

which contradicts the definition of $\widehat{\alpha}$.

C.9 Proof of Lemma C.3

Define $\mathbb{S} \coloneqq \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$. We first prove that for any $u \in \mathbb{S}^{d-1}$ and any $(x_t, x_{t-1}) \in \mathcal{E}_t$,

$$\big\|\nabla_x u^\top J_t(x_t)u\big\|_2 \lesssim \frac{1}{\sqrt{1-\overline{\alpha}_t}} d^{3/2}\log^{3/2} T, \tag{92}$$

where $\nabla_x u^\top J_t(x_t)u \coloneqq \nabla_x u^\top J_t(x)u \,\big|_{x=x_t}$.

Proof of relation (92)**.** Recall that in Section C.6, we have shown that

$$\begin{aligned}s_t(x) &= -\frac{1}{1-\overline{\alpha}_t}\int_{x_0}(x-\sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x)\mathrm{d}x_0, \\ J_t(x) &= -(1-\overline{\alpha}_t)\frac{\partial s_t(x)}{\partial x}.\end{aligned}$$

In view of these two relations and the definition of J_t, we can write $u^\top J_t(x_t)u$ as

$$\begin{aligned}u^\top J_t(x_t)u = 1 + \frac{1}{1-\overline{\alpha}_t}\bigg\{&\Big(\mathbb{E}\big[(X_t-\sqrt{\overline{\alpha}_t}X_0)^\top u \mid X_t = x_t\big]\Big)^2 \\ &- \mathbb{E}\Big[\big[\big(X_t-\sqrt{\overline{\alpha}_t}X_0\big)^\top u\big]_2^2 \mid X_t = x_t\Big]\bigg\}.\end{aligned}$$

To further control $\nabla_x u^\top J_t(x_t)u$, let us consider the two terms on the right hand side separately.

- For the first term, one has

$$\begin{aligned}&\Big\|\nabla_{x_t}\Big(\mathbb{E}\big[(X_t-\sqrt{\overline{\alpha}_t}X_0)^\top u \mid X_t = x_t\big]\Big)^2\Big\|_2 \\ =&\Big\|\nabla_{x_t}\Big(\int_{x_0}\big[(x_t-\sqrt{\overline{\alpha}_t}x_0)^\top u\big]p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big)^2\Big\|_2 \\ =&\Big\|2(1-\overline{\alpha}_t)(s_t(x_t)^\top u)\cdot(1-\overline{\alpha}_t)\frac{\partial s_t(x_t)}{\partial x}\Big\|_2 \\ \lesssim&(1-\overline{\alpha}_t)\|s_t(x_t)\|_2\cdot\Big\|(1-\overline{\alpha}_t)\frac{\partial s_t(x_t)}{\partial x}\Big\|.\end{aligned} \tag{93}$$

By equation (77), we can compute that

$$\begin{aligned}
&\Big\|(1-\overline{\alpha}_t)\frac{\partial s_t(x_t)}{\partial x}\Big\| \\
&\leq 1 + \frac{1}{1-\overline{\alpha}_t}\Big\|\int_{x_0}(x_t - \sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0 \\
&\qquad\qquad\qquad\qquad\Big(\int_{x_0}(x_t - \sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big)^\top \\
&\qquad\qquad - \int_{x_0}(x_t - \sqrt{\overline{\alpha}_t}x_0)(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\| \\
&\lesssim 1 + \frac{1}{1-\overline{\alpha}_t}\Big\|\int_{x_0}(x_t - \sqrt{\overline{\alpha}}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2^2 \\
&\leq 1 + \frac{1}{1-\overline{\alpha}_t}\mathbb{E}_{X_0}\Big[\big\|\sqrt{\overline{\alpha}_t}X_0 - x_t\big\|_2^2\,\big|\,X_t = x_t\Big] \lesssim d\log T.
\end{aligned} \tag{94}$$

Here, in the second inequality, we use the fact that for a column vector $Z \in \mathbb{R}^d$, we have

$$\begin{aligned}
\big\|\mathbb{E}\big[ZZ^\top\big] - \mathbb{E}[Z]\mathbb{E}[Z]^\top\big\| = \Big\|\mathbb{E}\Big[\big(Z - \mathbb{E}[Z]\big)\big(Z - \mathbb{E}[Z]\big)^\top\Big]\Big\| &\leq \big\|\mathbb{E}\big[ZZ^\top\big]\big\| \\
&\leq \mathbb{E}\big[\,\big\|ZZ^\top\big\|\,\big] = \mathbb{E}\big[\|Z\|_2^2\big],
\end{aligned}$$

and the last line invokes Lemma A.1. Now plugging the bounds in inequality (62) and (94) into inequality (93), we obtain

$$\Big\|\nabla_{x_t}\Big(\mathbb{E}\big[(X_t - \sqrt{\overline{\alpha}_t}X_0)^\top u \mid X_t = x_t\big]\Big)^2\Big\|_2 \lesssim (1-\overline{\alpha}_t)^{\frac{1}{2}}d^{3/2}\log^{3/2}T. \tag{95}$$

- When it comes to the second term, some direct calculations give

$$\begin{aligned}
&\Big\|\nabla_{x_t}\mathbb{E}\Big[\big[\big(X_t - \sqrt{\overline{\alpha}_t}X_0\big)^\top u\big]_2^2 \mid X_t = x_t\Big]\Big\|_2 \\
=&\Big\|\nabla_{x_t}\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]^2 p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2 \\
\leq&\Big\|2\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]u\cdot p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2 \\
&\qquad\qquad\qquad + \Big\|\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]^2\frac{\partial}{\partial x_t}p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2 \\
\leq&2\int_{x_0}\|(x_t - \sqrt{\overline{\alpha}_t}x_0)\|_2\cdot p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0 \\
&\qquad\qquad\qquad + \Big\|\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]^2\frac{\partial}{\partial x_t}p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2 \\
\lesssim&(1-\overline{\alpha}_t)^{\frac{1}{3}}d^{1/2}\log^{1/2}T + \Big\|\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]^2\frac{\partial}{\partial x_t}p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2,
\end{aligned} \tag{96}$$

where we use Lemma A.1 to obtain the last inequality. To further bound the second term in inequality (96), we repeat the calculations for equations (75) and (76) and deduce that

$$\begin{aligned}
&\Big\|\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]^2\frac{\partial}{\partial x_t}p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2 \\
\leq&\Big\|\frac{1}{1-\overline{\alpha}_t}\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]^2(x_t - \sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2 \\
&\quad + \Big\|\frac{1}{1-\overline{\alpha}_t}\int_{x_0}\big[(x_t - \sqrt{\overline{\alpha}_t}x_0)^\top u\big]^2 p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\cdot \\
&\qquad\qquad\qquad\qquad\qquad \int_{x_0}(x_t - \sqrt{\overline{\alpha}_t}x_0)p_{X_0\,|\,X_t}(x_0\,|\,x_t)\mathrm{d}x_0\Big\|_2 \\
\lesssim&\Big|\frac{1}{1-\overline{\alpha}_t}\mathbb{E}\big[\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^3\,|\,X_t = x_t\big]\Big| + \Big\|s_t(x_t)\mathbb{E}\big[\|x_t - \sqrt{\overline{\alpha}_t}x_0\|_2^2\,|\,X_t = x_t\big]\Big\|_2.
\end{aligned} \tag{97}$$

Taking colelctively the inequalities (96) and (97), we arrive at

$$\begin{aligned}
&\nabla_{x_t}\mathbb{E}\Big[\big[\big(X_t-\sqrt{\overline{\alpha}_t}X_0\big)^\top u\big]_2^2 \mid X_t=x_t\Big]\\
&\lesssim (1-\overline{\alpha}_t)^{\frac{1}{3}} d^{1/2}\log^{1/2}T + \Big|\frac{1}{1-\overline{\alpha}_t}\mathbb{E}\big[\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^3 \mid X_t=x_t\big]\Big|\\
&\qquad\qquad\qquad\qquad + \Big\|s_t(x_t)\mathbb{E}\big[\|x_t-\sqrt{\overline{\alpha}_t}x_0\|_2^2 \mid X_t=x_t\big]\Big\|_2\\
&\lesssim (1-\overline{\alpha}_t)^{\frac{1}{2}} d^{3/2}\log^{3/2}T.
\end{aligned}\tag{98}$$

where last inequality is a direct consequence of Lemma A.1. Therefore, combining the two relations (95) and (98) yields the claimed relation (92).

Next, we shall proceed to show that similar to the relation (92), one also has

$$\big\|\nabla_x u^\top J_t(x_{t-1}(\gamma))u\big\|_2 \lesssim \frac{1}{\sqrt{1-\overline{\alpha}_t}} d^{3/2}\log^{3/2}T, \tag{99}$$

which holds for every $0\leq\gamma\leq 1$.

Proof of inequality (99)**.** We make the observation that the derivations above to prove relation (92) only involves $X_t=x_t$ which satisfies the first condition in the definition of $\mathcal{E}_t$, namely, $-\log p_{X_t}(x_t)\leq c_3 d\log T$. Now, let us prove that $-\log p_{X_t}(x_{t-1}(\gamma))\leq 2c_3 d\log T$ for $x_{t-1}(\gamma)$. Similar as in deriving inequality (52), we can deduce

$$\begin{aligned}
\|x_{t-1}-\phi_t(x_t)\|_2 &\lesssim \Big\|x_{t-1}-\frac{x_t}{\sqrt{\alpha_t}}\Big\|_2 + \Big\|\Big(\frac{1}{\sqrt{\alpha}_t}-1\Big)s_t(x_t)\Big\|_2\\
&\qquad + \Big\|\frac{1}{2\sqrt{\alpha}_t}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\big(s_{\overline{\alpha}}(g_t(x_t,\overline{\alpha}))-s_t(x_t)\big)\mathrm{d}\overline{\alpha}\Big\|_2\\
&\lesssim \sqrt{d(1-\alpha_t)\log T}+\sqrt{d(1-\alpha_t)^2\log T}\\
&\qquad + \frac{1}{2\sqrt{\alpha_t}}\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}}\sqrt{\frac{\overline{\alpha}_t}{\overline{\alpha}^3}}\mathrm{d}\overline{\alpha}\Big(\sup_{\overline{\alpha}_t<\widetilde{\alpha}<\overline{\alpha}_{t-1}}\|s_{\widetilde{\alpha}}(g_t(x_t,\widetilde{\alpha}))-s_t(x_t)\|_2\Big)\\
&\lesssim \sqrt{d(1-\alpha_t)\log T}+\sqrt{d(1-\alpha_t)^2\log T}+(1-\alpha_t)^2\Big(\frac{d\log T}{1-\overline{\alpha}_t}\Big)^{3/2}\\
&\lesssim \sqrt{d(1-\alpha_t)\log T},
\end{aligned}\tag{100}$$

where we use inequality (61) in the third line. Since $x_{t-1}(\gamma):=\gamma x_{t-1}+(1-\gamma)\phi_t(x_t)$ and inequality (100), we can directly recognize that

$$\|x_{t-1}(\gamma)-\phi_t(x_t)\|_2\lesssim\sqrt{d(1-\alpha_t)\log T}. \tag{101}$$

Putting these two relations above together, it is easily seen that

$$\big\|x_{t-1}-x_{t-1}(\gamma)/\sqrt{\alpha_t}\big\|\leq c_4\sqrt{d(1-\alpha_t)\log T}. \tag{102}$$

In addition, in view of Lemma A.2, we know that for $(x_t,x_{t-1})\in\mathcal{E}_t$ and any $\gamma\in[0,1]$, it holds that

$$-\log p_{X_{t-1}}(x_{t-1}(\gamma))\leq 2c_3 d\log T \ \text{ and } \ p_{X_{t-1}}(x)=\bigg(1+O\Big(\sqrt{\frac{d(1-\alpha_t)\log T}{1-\overline{\alpha}_t}}\Big)\bigg)p_{X_t}(x). \tag{103}$$

From properties (102) and (103), we conclude $(x_{t-1}(\gamma),x_{t-1})\in\mathcal{E}_t$. It thus enables us to apply the same analysis as above on $J_t(x_{t-1}(\gamma))$, and draw the conclusion that

$$\big\|\nabla_x u^\top J_t(x_{t-1}(\gamma))u\big\|_2\lesssim\frac{1}{\sqrt{1-\overline{\alpha}_t}}d^{3/2}\log^{3/2}T$$

for $0\leq\gamma\leq 1$. We complete the proof of the inequality (99).

In Summary. Based on expression (92), some direct calculations yield

$$\begin{aligned}
&\|J_t(x_{t-1}) - J_t(\phi(x_t))\| \\
&\leq \sup_{u \in \mathbb{S}^{d-1}} \Big| u^\top \Big(J_t(x_{t-1}) - J_t\big(\phi(x_t)\big)\Big) u \Big| \\
&\lesssim \frac{1}{\sqrt{1-\overline{\alpha}_t}} d^{3/2} \log^{3/2} T \|x_{t-1} - \phi(x_t)\|_2, \qquad (104)
\end{aligned}$$

which concludes the proof of inequality (69a). In addition, as discussed after the inequality (72), the Lipschitz condition of $\Phi_{t-1\to k}(x)$ allows us to prove $(x_k, x_{k-1}) \in \mathcal{E}_k$. Repeating the analysis above, we can conclude that

$$\begin{aligned}
&\Big\| J_k\big(\Phi_{t-1\to k}(x_{t-1}(\gamma))\big) - J_k\big(\Phi_{t\to k}(x_t)\big) \Big\| \\
&\qquad\qquad \lesssim \frac{1}{\sqrt{1-\overline{\alpha}_t}} d^{3/2} \log^{3/2} T \Big\| \Phi_{t-1\to k}(x_{t-1}(\gamma)) - \Phi_{t\to k}(x_t) \Big\|_2,
\end{aligned}$$

which thus completes the proof of inequality (69b).