Optimal Convergence Analysis of DDPM for General Distributions

Yuchen Jiao
Department of Statistics and Data Science
The Chinese University of Hong Kong
Shatin, N.T. Hong Kong SAR, China
yuchenjiao@cuhk.edu.hk

Yuchen Zhou
Department of Statistics
University of Illinois Urbana-Champaign
Champaign, IL 61820 USA
yuchenz@illinois.edu

Gen Li*
Department of Statistics and Data Science
The Chinese University of Hong Kong
Shatin, N.T. Hong Kong SAR, China
genli@cuhk.edu.hk

Abstract

Score-based diffusion models have achieved remarkable empirical success in generating high-quality samples from target data distributions. Among them, the Denoising Diffusion Probabilistic Model (DDPM) is one of the most widely used samplers, generating samples via estimated score functions. Despite its empirical success, a tight theoretical understanding of DDPM — especially its convergence properties — remains limited. In this paper, we provide a refined convergence analysis for the DDPM sampler and establish near-optimal convergence rates under general distributional assumptions. Specifically, we consider a relaxed smoothness condition parameterized by L, which is small for many practical distributions (e.g., Gaussian mixture models). We prove that, to approximate a target distribution on $\mathbb{R}^d$ to accuracy ε in total variation distance and ε^2 in KL-divergence, the DDPM sampler with accurate score estimates requires at most

$$T = \widetilde{O}\left(\frac{\sqrt{d}\min\{\sqrt{d}, L\}}{\varepsilon}\right)$$

iterations, where $\widetilde{O}$ hides polylogarithmic factors in d and $1/\varepsilon$. This result substantially improves upon the best-known $\widetilde{O}(d/\varepsilon)$ iteration complexity when $L < \sqrt{d}$. By establishing a matching lower bound, we show that our convergence analysis is tight for a wide array of target distributions. Moreover, it reveals that DDPM and Denoising Diffusion Implicit Models (DDIM) share the same dependence on d, raising an interesting question of why DDIM often appears empirically faster.

1 Introduction

The past few years have witnessed the emergence of diffusion models as a leading generative paradigm, achieving top performance across a wide range of applications, including images (Rombach et al., 2022; Saharia et al., 2022; Song and Ermon, 2019; Croitoru et al., 2023; Lugmayr et al., 2022; Nichol et al., 2021), audio (Villegas et al., 2022; Liu et al., 2023), and text (Ramesh et al., 2022; Li et al., 2022). In short, they consist of two components:

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of convergence rates (in TV) and data assumptions for different samplers.

Sampler	Convergence Rate (TV)	Data Assumption
DDPM (Chen et al., 2022)	$\tilde{L}\sqrt{d/T}$	$\tilde{L}$-Lipschitz $s_t^\star$; bounded $\mathbb{E}[\|X_0\|_2^2]$
DDPM (Chen et al., 2023)	$\sqrt{d^2/T}$	bounded $\mathbb{E}[\|X_0\|_2^2]$
DDPM (Benton et al., 2023)	$\sqrt{d/T}$	bounded $\mathbb{E}[\|X_0\|_2^2]$
DDPM (Li and Yan, 2024b)	d/T	bounded $\mathbb{E}[\|X_0\|_2^2]$
DDIM Chen et al. (2024)	$\tilde{L}^2\sqrt{d}/T$	$\tilde{L}$-Lipschitz $s_t^\star$; bounded $\mathbb{E}[\|X_0\|_2^2]$
DDIM (Li et al., 2023)	$d^2/T + d^6/T^2$	bounded support
DDIM (Li et al., 2024b)	d/T (when $T \gtrsim d^2$)	bounded support
DDPM **(this paper)**	$\sqrt{d}\min(\sqrt{d}, L)/T$	L non-uniform Lipschitz $s_t^\star$, bounded $\mathbb{E}[\|X_0\|_2^2]$

(1) a forward process

$$X_0 \sim p_{\mathsf{data}} \to X_1 \to \cdots \to X_T$$

that starts with X_0 drawn from the target data distribution and, by sequentially adding noise, yields $X_T \approx \mathcal{N}(0, I_d)$; and

(2) a backward process

$$Y_T \sim \mathcal{N}(0, I_d) \to Y_{T-1} \to \cdots \to Y_0$$

that successively transforms Gaussian noise into Y_0 whose distribution is close to p_{data} using learned scores $s_t \approx s_t^\star = \nabla \log p_{X_t}$ (Hyvärinen and Dayan, 2005; Ho et al., 2020; Hyvärinen, 2007; Vincent, 2011; Song and Ermon, 2019; Pang et al., 2020).

Along the trajectory, the backward marginals track the forward ones: for $0 \le t \le T$, $Y_t \overset{\mathrm{d}}{\approx} X_t$. To this end, there are two mainstream approaches for constructing the reverse process: SDE-based samplers and ODE-based samplers, prototyped by the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) and the Denoising Diffusion Implicit Model (DDIM) (Song et al., 2021), respectively. They differ in that the former adds an independent Gaussian perturbation when updating Y_{t-1} from Y_t, whereas the latter is deterministic.

Motivated by their remarkable empirical success, the convergence behavior of both sampler classes, particularly DDPM and DDIM, has attracted increasing theoretical attention. A common objective is to control the total variation (TV) distance between the distribution generated by the sampler and the target. Existing theoretical analyses can be broadly categorized into two directions. One line of work focuses on the role of the intrinsic dimension d (De Bortoli, 2022; Gao et al., 2023; Lee et al., 2023; Benton et al., 2023; Li et al., 2023, 2024b,a; Li and Yan, 2024a,b; Huang et al., 2024c; Liang et al., 2025; Holzmüller and Bach, 2023; Li et al., 2025b, 2024c; Li and Cai, 2024; Xu and Chi, 2024; Jiao et al., 2025a; Li et al., 2025a; Li and Cai, 2025; Li and Jiao, 2025), while another seeks weaker dependence on d under certain Lipchitz conditions on the score functions (Lee et al., 2022; Chen et al., 2022, 2023; Gupta et al., 2024; Chen et al., 2024; Li and Jiao, 2024; Huang et al., 2024a,b; Gao and Zhu, 2024). Along the first line, the best-known rates for both DDPM and DDIM are $\widetilde{O}(d/T)$ (Li and Yan, 2024b; Li et al., 2024b). Along the second line, for a variant of the DDIM sampler, Chen et al. (2024) established an $\widetilde{O}(\widetilde{L}^2\sqrt{d}/T)$[2] TV rate under a bounded second-moment assumption on the target distribution and $\widetilde{L}$-Lipschitz conditions on both the true score $s_t^\star(\cdot)$ and its estimator $s_t(\cdot)$. Notably, when $\widetilde{L} \ll d^{1/4}$, this bound is much sharper than the $\widetilde{O}(d/T)$ rate. Nevertheless, analogous results for DDPM are not yet available in the literature. This naturally leads to the following question:

> *Under comparable or even weaker assumptions, can DDPM match or improve upon the state-of-the-art TV rate $\widetilde{O}(\widetilde{L}^2\sqrt{d}/T)$ established for this DDIM variant? Can it further achieve a KL rate sharper than $\widetilde{O}(\widetilde{L}^4 d/T^2)$?*

[2]Throughout the paper, we write $f(L,d,T) = O(g(L,d,T))$ or $f(L,d,T) \lesssim g(L,d,T)$ if $|f(L,d,T)| \le Cg(L,d,T)$ for some universal constant $C > 0$; and $f(L,d,T) \asymp g(L,d,T)$ if both $f(L,d,T) = O(g(L,d,T))$ and $g(L,d,T) = O(f(L,d,T))$ hold. The notation $\widetilde{O}(g(L,d,T))$ denotes $O(g(L,d,T))$ up to polylogarithmic factors.

Main contributions. Our work contributes to the aforementioned line of research by showing more precisely how smoothness can improve the convergence rate of DDPM. We summarize our main results in the following informal theorem; a formal version is deferred to Section 3.2.

Theorem (Informal). Suppose that the score functions associated with the forward process satisfies a relaxed Lipschitz condition with parameter L (see Definition 3.2), a much smaller quantity than the global Lipschitz parameter commonly assumed in prior work. Then, with access to sufficiently accurate score estimates, the convergence rate of the DDPM sampler in total variation (resp. in KL divergence) is

$$\widetilde{O}\Big(\frac{\sqrt{d}\min\{L,\sqrt{d}\}}{T}\Big) \quad (\text{resp. } \widetilde{O}\Big(\frac{d\min\{L^2,d\}}{T^2}\Big)).$$

We next discuss the implications of this result.

1. **Optimal $\sqrt{d}$ dependence for common distributions**. We show that our relaxed Lipschitz condition is satisfied by many common target distributions with $L \le \mathrm{poly}(\log(dT))$, and consequently the convergence rate in total variation distance (resp. KL divergence) becomes

$$\widetilde{O}\Big(\frac{\sqrt{d}}{T}\Big) \quad (\text{resp. } \widetilde{O}\Big(\frac{d}{T^2}\Big)).$$

   Our results significantly reduce the dependence on d in DDPM convergence analyses. Given that the ambient dimension d often scales as $10^4 - 10^6$ in common image/video tasks, this improvement is substantial. Compared with the $\widetilde{O}(\widetilde{L}^2\sqrt{d}/T)$ rate for DDIM, our result shares the same d dependence while improving the dependence on the Lipschitz parameter L. More detailed comparison is in Table 1. We further established matching lower bounds, implying that our TV and KL rates are tight (up to logarithmic factors).
2. $\widetilde{O}(d^2/T^2)$ **KL rate under minimal assumptions**. Even without the non-uniform Lipschitz assumption (i.e., allowing $L = \infty$), our analysis yields upper bounds of order $\widetilde{O}(d/T)$ in TV distance and $\widetilde{O}(d^2/T^2)$ in KL divergence under the minimal assumption on the target distribution (Assumption 3.1). In this regime ($L > \sqrt{d}$), our bound is the same as prior works in Li and Yan (2024b) and Jain and Zhang (2025) for the TV- and KL-rate analysis, respectively.

Paper organization. The rest of the paper is organized as follows. In Section 2, we formally introduce diffusion models and the DDPM sampler. The main results are presented in Section 3. In Section 4, we outline the proof of our main results. The proofs of all technical lemmas are deferred to the appendix.

2 Preliminary

In this section, we introduce preliminaries on diffusion models and the DDPM sampler.

The forward process. Diffusion models comprise a forward (noising) process and the reverse (denoising) process. The forward process is Markovian: starting from $X_0 \sim p_{\mathsf{data}}$ on $\mathbb{R}^d$,

$$X_t = \sqrt{\alpha_t}X_{t-1} + \sqrt{1-\alpha_t}Z_t, \quad t = 1, \cdots, T, \tag{1}$$

where $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ are independent Gaussian random vectors, and the α_t's are the learning rates. Then it is straightforward to show that

$$X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{Z}_t,$$
$$\text{where} \quad \overline{\alpha}_t = \prod_{k=1}^{t}\alpha_k \text{ and } \overline{Z}_t \sim \mathcal{N}(0, I_d). \tag{2}$$

In the continuous-time limit, the forward process admits the following widely studied SDE:

$$\mathrm{d}\overline{X}_\tau = -\frac{1}{2(1-\tau)}\overline{X}_\tau \mathrm{d}\tau + \frac{1}{\sqrt{1-\tau}}\mathrm{d}B_\tau, \qquad \tau \in (0,1);$$

$$\overline{X}_0 \sim p_{\mathsf{data}}, \tag{3}$$

where B_τ is the standard Brownian motion. In fact, it has been shown that the distribution of $\overline{X}_\tau$ is given by

$$\overline{X}_\tau \overset{\mathrm{d}}{=} \sqrt{1-\tau}\,\overline{X}_0 + \sqrt{\tau} W, \quad W \sim \mathcal{N}(0, I_d). \tag{4}$$

Putting (2) and (4) together, we establish the following connection between the forward processes in discrete time and continuous time:

$$\overline{X}_{\tau_t} \overset{\mathrm{d}}{=} X_t, \quad \text{where} \quad \tau_t = 1 - \overline{\alpha}_t. \tag{5}$$

Throughout this paper, we denote by p_{X_t} and $p_{\overline{X}_\tau}$ the probablity density function of X_t and $\overline{X}_\tau$, respectively.

Score functions. A key ingredient in the sampling process is the (Stein) score function, i.e., the gradient of the log marginal density of the forward process (1):

$$\begin{aligned} s_t^\star(x) &\coloneqq \nabla \log p_{X_t}(x) \\ &= -\frac{1}{1-\overline{\alpha}_t} \int (x - \sqrt{\overline{\alpha}_t} x_0) p_{X_0|X_t}(x_0|x) \mathrm{d}x_0, \end{aligned} \tag{6}$$

where the last equality follows from Tweedie's formula (Efron, 2011). For notational convenience, we also introduce the score for the continuous-time process (3):

$$\begin{aligned} \overline{s}_\tau^\star(x) &\coloneqq \nabla \log p_{\overline{X}_\tau}(x) \\ &= -\frac{1}{\tau} \int (x - \sqrt{1-\tau} x_0) p_{\overline{X}_0|\overline{X}_\tau}(x_0|x) \mathrm{d}x_0. \end{aligned} \tag{7}$$

Recalling that $\overline{X}_{\tau_t} \overset{\mathrm{d}}{=} X_t$ (cf. (5)), we immediately have $\overline{s}_{\tau_t}^\star(x) = s_t^\star(x)$.

We usually do not have access to the exact score functions $s_t^\star(\cdot)$. Here, we assume that some estimates for the score functions, $\{s_t(\cdot)\}_{1\le t\le T}$, are available.

DDPM Sampler and learning rate schedule. The Denoising Diffusion Probabilistic Model (DDPM) constructs a reverse process for (2), with the goal of generating samples whose distribution is close to p_{data}. More specifically, each Y_{t-1} is a function of Y_t plus independent Gaussian noise $Z_t \sim \mathcal{N}(0, I_d)$:

$$\begin{aligned} Y_{t-1} &= \frac{1}{\sqrt{\alpha_t}} \left(Y_t + \eta_t s_t(Y_t) + \sigma_t Z_t \right), \quad t = T, \cdots, 2, \\ Y_T &\sim \mathcal{N}(0, I_d) \end{aligned} \tag{8}$$

where η_t and σ_t are parameters that play pivotal roles for achieving satisfactory performance. Following Li et al. (2024b), we choose

$$\eta_t = 1 - \alpha_t, \qquad \sigma_t^2 = 1 - \alpha_t. \tag{9}$$

We assume that the learning rates satisfy the following conditions: (i) $\beta_t := 1 - \alpha_t$ is small for every $1 < t < T$; and (ii) $\overline{\alpha}_T = \prod_{t=1}^T \alpha_t$ is vanishingly small, ensuring that the distribution of X_T, is exceedingly close to $\mathcal{N}(0, I_d)$. More specifically, we assume that the learning rates $\{\alpha_t\}_{1\le t\le T}$ satisfy

$$\begin{aligned} \overline{\alpha}_T \le \frac{1}{T^{c_0}}, \qquad 1 - \alpha_1 \le \frac{1}{T^{c_0}}, \\ \overline{\alpha}_{t-1} - \overline{\alpha}_t \le \frac{c \log T}{T} \overline{\alpha}_t (1 - \overline{\alpha}_t). \end{aligned} \tag{10}$$

An example of learning rate schedule that obeys (10) is:

$$\begin{aligned} \beta_1 &= \frac{1}{T^{c_0}}, \\ \beta_{t+1} &= \frac{c_1 \log T}{T} \min\left\{ \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^t, 1 \right\}, \\ & \qquad t = 1, \cdots, T-1. \end{aligned} \tag{11}$$

Notation For two probability measures P and Q, the total-variation distance between them is defined as $\mathsf{TV}(P,Q) \coloneqq \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$. If P is absolutely continuous with respect to Q, the Kullbeck-Leibler (KL) divergence of P from Q is $\mathsf{KL}(P,Q) \coloneqq \int \log(\frac{\mathrm{d}P}{\mathrm{d}Q})\mathrm{d}P$. For any random vector X, we let p_X denote its probability density function. For any matrix A, we denote by $\|A\|_{\mathsf{op}}$ its spectral norm.

3 Main theory

In this section, we develop an optimal convergence rate theory for the DDPM sampler, showing that it has the same $O(\sqrt{d})$ dependence as DDIM. Before proceeding, we first introduce the assumptions used in our analysis.

3.1 Assumptions

The first assumption allows the second-order moment of the target distribution to scale at most polynomially in T, which encompasses a wide array of applications.

Assumption 3.1. The target distribution p_{data} has a bounded second-order moment:

$$\mathbb{E}_{X_0\sim p_{\mathsf{data}}}[\|X_0\|_2^2] \le d^{c_R}, \tag{12}$$

where $c_R > 0$ is an arbitrarily large constant.

Our analysis makes use of a relaxed smoothness condition. Specifically, we consider a non-uniform Lipschitz parameter for the normalized score functions $\tau\overline{s}_\tau^\star$ as follows:

Definition 3.2 (Non-uniform Lipschitz property)**.** Let $L \ge 1$ denote the smallest quantity such that, for every $\tau \in (0,1)$,

$$\mathbb{P}\left(\tau\|\nabla\overline{s}_\tau^\star(\overline{X}_\tau)\|_{\mathsf{op}} \le L\right) \ge 1 - \frac{1}{d^4}.$$

The global smoothness parameter $\widetilde{L}$, which is widely used in previous diffusion-model analyses, is defined as the smallest quantity such that

$$\|\nabla s_t^\star(x)\|_{\mathsf{op}} \le \widetilde{L}, \qquad \forall x \in \mathbb{R}^d, \quad \forall t. \tag{13}$$

Compared to this global smoothness condition, our assumption is milder and applies to a broader range of data distributions: (1) it only requires a high-probability bound, and (2) it bounds the scaled quantity $\tau\|\nabla\overline{s}_\tau^\star(\overline{X}_\tau)\|_{\mathsf{op}}$, thereby permitting $\|\nabla\overline{s}_\tau^\star(\overline{X}_\tau)\|_{\mathsf{op}}$ to be much larger when τ is small, i.e., when the distribution of X_τ is closer to the target distribution p_{data}. In practice, data distributions are often discrete, as in image data, or exhibit multiple modes, both of which can lead to an extremely large global Lipschitz parameter $\widetilde{L}$. The following examples show that the non-uniform Lipschitz property holds for many common distributions with $L \lesssim \log(dT)$, whereas the global Lipschitz parameter can be much larger, leading to substantially worse convergence rates. The proofs of these examples are postponed to Appendix G.

Example 1. Suppose that $X_0 \sim \mathcal{N}(\mu,\Sigma)$, where $\mu \in \mathbb{R}^d$ and Σ is a covariance matrix. Then the non-uniform Lipschitz parameter satisfies $L = 1$. In comparison, there exist μ and Σ such that $\widetilde{L} = 1/(1-\overline{\alpha}_1)$, which becomes very large as $\overline{\alpha}_1 \to 1$.

Example 2. Suppose that X_0 follows a d-dimensional Gaussian mixture model

$$\sum_{h=1}^{H}\pi_h\mathcal{N}(\mu_h,\sigma_h^2 I_d), \qquad \pi_h \ge 0, \qquad \sum_{h=1}^{H}\pi_h = 1.$$

Then the non-uniform Lipschitz parameter obeys $L \lesssim \log(d)\log(H)$. On the other hand, there exists a two-component Gaussian mixture with variances $\sigma_1 = \sigma_2 = 1$ such that $\widetilde{L} \gtrsim d$.

Example 3. Suppose that the entries of X_0 are independent and satisfy $\mathbb{E}[|X_{0,i}|] \le d^{c_R}$, where $X_{0,i}$ is the i-th coordinate and c_R is an arbitrarily large constant. Then the non-uniform Lipschitz parameter satisfies $L \lesssim \log d$. However, if $X_{0,i} \sim \mathsf{Unif}(\{-1,1\})$ independently (which satisfies $c_R = 0$), then the global Lipschitz parameter satisfies $\widetilde{L} \gtrsim \frac{1}{(1-\overline{\alpha}_1)^2}$.

In fact, we conjecture that the non-uniform Lipschitz condition holds with $L \le \mathrm{poly}(\log(dT))$ for any absolutely continuous target distribution. We leave further investigation on this for future work.

In addition, the following assumption captures the quality of the score estimates.

Assumption 3.3. We assume access to an estimate $s_t(\cdot)$ for each $s_t^\star(\cdot)$, with the averaged ℓ_2 score estimation error

$$\varepsilon_{\mathsf{score}}^2 := \frac{1}{T}\sum_{t=1}^{T-1}(1-\overline{\alpha}_t)\mathbb{E}_{x_t\sim p_{X_t}}\left[\|s_t(x_t)-s_t^\star(x_t)\|_2^2\right], \tag{14}$$

where $\overline{\alpha}_t$ satisfies (10).

Note that $\varepsilon_{\mathsf{score}}^2$ is no larger than the commonly used unweighted error (Gupta et al., 2024; Huang et al., 2024a,b; Lee et al., 2022; Li and Cai, 2024; Li et al., 2024a)

$$\widetilde{\varepsilon}_{\mathsf{score}}^2 := \frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}_{x_t\sim p_{X_t}}\left[\|s_t(x_t)-s_t^\star(x_t)\|_2^2\right] \tag{15}$$

since $0 < 1-\overline{\alpha}_t \le 1$ for all t. In particular, under the learning-rate schedule (10), the weights $1-\overline{\alpha}_t$ can be very small for early steps, so Assumption 3.3 allows larger per-step errors $\mathbb{E}\|s_t(X_t)-s_t^\star(X_t)\|_2^2$ when t is small.

3.2 Theory

The following theorem establishes a sharp convergence rate for the classical DDPM sampler.

Theorem 3.4. *Suppose that Assumptions 3.1 and 3.3 hold. Assume $T \gtrsim d^\delta$, where $\delta > 0$ is an arbitrarily small constant. Then the DDPM sampler* (8) *with the learning rate schedule* (10) *satisfies*

$$\begin{aligned}\mathsf{TV}\left(p_{X_1}, p_{Y_1}\right) &\le \sqrt{\frac{1}{2}\mathsf{KL}\left(p_{X_1}\|p_{Y_1}\right)} \\ &\le \frac{Cd^{1/2}\min\{d^{1/2}, L\}\log^2 T}{T} + C\varepsilon_{\mathsf{score}}\log^{1/2} T,\end{aligned} \tag{16}$$

for some constant $C > 0$ large enough, where L is defined in Definition 3.2.

Assume that the perfect scores are available, i.e., $\varepsilon_{\mathsf{score}} = 0$. Then the convergence rate in total variation distance (resp. KL divergence) becomes

$$\widetilde{O}\left(\frac{d^{1/2}\min\{d^{1/2}, L\}}{T}\right) \qquad \left(\text{resp. } \widetilde{O}\Big(\frac{d\min\{d, L^2\}}{T^2}\Big)\right).$$

When $L < \sqrt{d}$, our results substantially improve upon the state-of-the-art $\widetilde{O}(d/T)$ TV rate (Li and Yan, 2024b) and $\widetilde{O}(d^2/T^2)$ KL rate (Jain and Zhang, 2025). As demonstrated in Examples 1 - 3, this scenario is frequently encountered in practice. Moreover, our analysis shows that DDPM achieves an $O(\sqrt{d})$ dependence on the dimension, matching that of a variant of DDIM, while enjoying a better dependence on the Lipschitz parameter. Specifically, our bound depends linearly on the non-uniform Lipschitz parameter L, whereas the DDIM bound in Chen et al. (2024) depends quadratically on the global smoothness parameter $\widetilde{L}$. Recall that $\widetilde{L}$ is the global smoothness parameter and can be much larger than our non-uniform parameter L. For the score-error term, prior work incurs a dependence of order $O(\widetilde{\varepsilon}_{\mathrm{score}}\log^{1/2} T)$, where $\widetilde{\varepsilon}_{\mathrm{score}}$ is defined in (15). By contrast, our TV bound depends on $\varepsilon_{\mathrm{score}}\log^{1/2} T$, which is tighter since $\widetilde{\varepsilon}_{\mathrm{score}} \ge \varepsilon_{\mathrm{score}}$.

Our results apply to a broad class of schedules satisfying the conditions in Equation (10), which are designed to control both discretization and initialization errors. Specifically, the discretization error depends on $\frac{1-\alpha_t}{1-\overline{\alpha}_t}$, whereas the initialization error is governed by $\overline{\alpha}_1$ and $\overline{\alpha}_T$. Common schedules, including linear, cosine, quadratic, and variance-preserving schedules, can be accommodated as long as they satisfy, or closely approximate, these conditions.

Remark 3.5. In practice, the global Lipschitz parameter $\widetilde{L}$ can be very large, leading to substantially worse convergence rates. As demonstrated in Examples 1 and 3, replacing the non-uniform Lipschitz

condition with the global Lipschitz condition can make the TV convergence rate in Chen et al. (2024) as large as $\widetilde{L}^2\sqrt{d}/T \geq \sqrt{d}/T/(1-\alpha_1)^2 \gg 1$, as $1/(1-\alpha_1) \gg T$. Furthermore, for the two-component isotropic GMM considered in Example 2, the global Lipschitz parameter satisfies $\widetilde{L} = \Omega(d)$, yielding a TV convergence rate of $\frac{\widetilde{L}^2\sqrt{d}}{T} \gtrsim \frac{d^{2.5}}{T}$ under the framework of Chen et al. (2024). This rate is considerably worse than existing results derived without any Lipschitz condition.

In addition, we can prove the following lower bound, which confirms the tightness of Theorem 3.4 when $L \leq \mathrm{poly}(\log(dT))$. The proof is postponed to Section F. We also discuss other variance schedules and discretizations at the end of Appendix F.

Theorem 3.6. *Assume that the learning rates $\{\alpha_t\}_{t=1}^T$ satisfy* (10) *and $\eta_t = 1-\alpha_t$ for all t, and assume $\{\sigma_t^2\}_{t=1}^T$ satisfy one of the following: (i) $\sigma_t^2 = 1-\alpha_t$, (ii) $\sigma_t^2 = (1-\alpha_t)\alpha_t$, and (iii) $\sigma_t^2 = \frac{(\alpha_t-\overline{\alpha}_t)(1-\alpha_t)}{1-\overline{\alpha}_t}$. If $p_{data} = \mathcal{N}(0, \lambda I_d)$ with some constant $\lambda \geq 2$, the output of sampler* (8) *with the oracle scores $s_t(\cdot) = s_t^\star(\cdot)$ obeys:*

$$\mathsf{KL}(p_{X_1}||p_{Y_1}) \geq \frac{c_{\mathsf{low}} d}{T^2}$$

for some universal constant $c_{\mathsf{low}} > 0$.

Remark 3.7. Theorem 3.6 focuses on the learning-rate schedules most widely used in the literature (Ho et al., 2020; Li and Yan, 2024b; Li et al., 2023; Huang et al., 2024c). Extending the analysis to arbitrary schedules is left for future investigation.

4 Proof of Theorem 3.4

Step 1. constructing an auxiliary reverse process. To begin with, we introduce an operator

$$\Phi_{\tau_1\to\tau_2}(x) \coloneqq x_{\tau_2}|_{x_{\tau_1}=x},$$

which is defined via the following ODE:

$$\mathrm{d}\frac{x_\tau}{\sqrt{1-\tau}} = -\frac{\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{3/2}}\mathrm{d}\tau. \tag{17}$$

By virtue of Li et al. (2024a, Eqn. (20)), we know that

$$\Phi_{\tau_1\to\tau_2}(\overline{X}_{\tau_1}) \overset{\mathrm{d}}{=} \overline{X}_{\tau_2}. \tag{18}$$

Motivated by this fact, we construct the following auxiliary reverse process with initialization $\widehat{X}_T \sim p_{X_T}$:

$$\begin{aligned}\widehat{X}_{t-1} &\coloneqq \frac{\sqrt{\overline{\alpha}_{t-1}}}{\sqrt{1-\widehat{\tau}_{t-1}}}\Phi_{\tau_t\to\widehat{\tau}_{t-1}}(\widehat{X}_t) + \frac{\sigma_t}{\sqrt{\alpha_t}}Z_t \\ &= \frac{1}{\sqrt{\alpha_t}}\widehat{X}_t + \int_{\widehat{\tau}_{t-1}}^{\tau_t}\frac{\sqrt{\overline{\alpha}_{t-1}}\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{\frac{3}{2}}}\mathrm{d}\tau + \frac{\sigma_t}{\sqrt{\alpha_t}}Z_t, \qquad t = T, \cdots, 2.\end{aligned} \tag{19}$$

Here, $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ are independent Gaussian random vectors, $\tau_t = 1-\overline{\alpha}_t$ is defined in (5), and the x_τ's are defined via the ODE (17) with $x_{\tau_t} = \widehat{X}_t$, and

$$\widehat{\tau}_{t-1} = 1-\widehat{\alpha}_{t-1}, \quad \text{where} \quad \widehat{\alpha}_{t-1} \coloneqq \frac{\overline{\alpha}_t}{2\alpha_t - 1}. \tag{20}$$

The following lemma shows that $\widehat{X}_t$ and the forward process X_t share the same marginal distribution, with the proof deferred to Appendix B.

Lemma 4.1. *For all $1 \leq t \leq T$, we have $\widehat{X}_t \overset{\mathrm{d}}{=} X_t$,where $\widehat{X}_t$ and X_t are defined in* (20) *and* (2)*, respectively.*

In view of Pinsker's inequality and Lemma 4.1, we have

$$\begin{aligned}\mathsf{TV}^2(p_{Y_1}, p_{X_1}) &\leq \frac{1}{2}\mathsf{KL}\left(p_{X_1}\|p_{Y_1}\right) = \frac{1}{2}\mathsf{KL}\left(p_{\widehat{X}_1}\|p_{Y_1}\right) \\ &\leq \frac{1}{2}\mathsf{KL}\left(p_{\widehat{X}_T}\|p_{Y_T}\right) + \frac{1}{2}\sum_{t=2}^T \mathbb{E}_{x_t\sim\widehat{X}_t}\mathsf{KL}(p_{\widehat{X}_{t-1}\,|\,\widehat{X}_t}(\cdot\,|\,x_t)\|p_{Y_{t-1}\,|\,Y_t}(\cdot\,|\,x_t)).\end{aligned} \tag{21}$$

Here, the last inequality makes use of the chain rule of KL divergence.

Step 2. controlling discretization and estimation error. This step focuses on bounding the term

$$\sum_{t=2}^{T} \mathbb{E}_{x_t \sim \widehat{X}_t} \mathsf{KL}(p_{\widehat{X}_{t-1} \,|\, \widehat{X}_t}(\cdot \,|\, x_t) \| p_{Y_{t-1} \,|\, Y_t}(\cdot \,|\, x_t))$$

appearing on the right-hand side of (21). Eqn. (19) tells us that the conditional density of $\widehat{X}_{t-1}$ given $\widehat{X}_t = x_t$ is

$$p_{\widehat{X}_{t-1} \,|\, \widehat{X}_t}(x \,|\, x_t) = \phi\bigg(x \;\bigg|\; \frac{\sqrt{\overline{\alpha}_{t-1}}}{\sqrt{1-\widehat{\tau}_{t-1}}} \Phi_{\tau_t \to \widehat{\tau}_{t-1}}(x_t), \frac{\sigma_t^2}{\alpha_t}\bigg), \tag{22}$$

where $\phi(x|\mu, \sigma^2)$ denotes the probability density function of the d-dimensional Gaussian distribution with mean vector μ and covariance matrix $\sigma^2 I_d$. Moreover, from (8), the conditional density of Y_{t-1} given $Y_t = y_t$, for $t = 0, \cdots, T-1$, is

$$P_{Y_{t-1}|Y_t}(y|y_t) = \phi\left(y \;\middle|\; \frac{1}{\sqrt{\alpha_t}}(y_t + \eta_t s_t(y_t)), \frac{\sigma_t^2}{\alpha_t}\right). \tag{23}$$

In other words, both $\widehat{X}_{t-1}$ and Y_{t-1} are conditionally Gaussian given $\widehat{X}_t = Y_t = x_t$, with the same variance σ_t^2/α_t but different mean vectors. As a consequence, the KL divergence between $p_{\widehat{X}_{t-1} \,|\, \widehat{X}_t}(\cdot \,|\, x_t)$ and $p_{Y_{t-1} \,|\, Y_t}(\cdot \,|\, x_t)$ can be calculated as

$$\mathsf{KL}(p_{\widehat{X}_{t-1} \,|\, \widehat{X}_t}(\cdot \,|\, x_t) \| p_{Y_{t-1} \,|\, Y_t}(\cdot \,|\, x_t)) = \frac{\alpha_t}{2\sigma_t^2} \|\mu_{\widehat{X}_{t-1}|\widehat{X}_t}(x_t) - \mu_{Y_{t-1}|Y_t}(x_t)\|_2^2, \tag{24}$$

where the conditional mean vectors are

$$\mu_{\widehat{X}_{t-1}|\widehat{X}_t}(x_t) = \frac{\sqrt{\overline{\alpha}_{t-1}}}{\sqrt{1-\widehat{\tau}_{t-1}}} \Phi_{\tau_t \to \widehat{\tau}_{t-1}}(x_t), \tag{25a}$$

$$\mu_{Y_{t-1}|Y_t}(x_t) = \frac{1}{\sqrt{\alpha_t}}(x_t + \eta_t s_t(x_t)) = \frac{1}{\sqrt{\alpha_t}}(x_t + \eta_t s_t^\star(x_t)) + \frac{\eta_t}{\sqrt{\alpha_t}}(s_t(x_t) - s_t^\star(x_t)). \tag{25b}$$

Substituting (25) into (24) and applying the inequality $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$, we obtain

$$\begin{aligned}
&\mathsf{KL}(p_{\widehat{X}_{t-1} \,|\, \widehat{X}_t}(\cdot \,|\, x_t) \| p_{Y_{t-1} \,|\, Y_t}(\cdot \,|\, x_t)) \\
&\le \frac{\alpha_t}{\sigma_t^2} \left\| \mu_{\widehat{X}_{t-1}|\widehat{X}_t}(x_t) - \frac{1}{\sqrt{\alpha_t}}(x_t + \eta_t s_t^\star(x_t)) \right\|_2^2 + \frac{\eta_t^2}{\sigma_t^2} \|s_t(x_t) - s_t^\star(x_t)\|_2^2 \\
&=: \|\xi_{t-1}(x_t)\|_2^2 + \frac{\eta_t^2}{\sigma_t^2} \|s_t(x_t) - s_t^\star(x_t)\|_2^2.
\end{aligned} \tag{26}$$

Here, the first term $\xi_{t-1}(x_t)$ represents the discretization error, and the second term represents the score estimation error. We now proceed to analyze these two error components separately. The following lemma gives a sharp bound on the discretization error.

Lemma 4.2. *Let $\Sigma_\tau(x)$ denote the conditional covariance matrix of the standard Gaussian noise $Z \sim N(0, I_d)$ conditioned on $\sqrt{1-\tau}X_0 + \sqrt{\tau}Z = x$, i.e.,*

$$\Sigma_\tau(x) = \mathsf{Cov}(Z | \sqrt{1-\tau}X_0 + \sqrt{\tau}Z = x). \tag{27}$$

Then the sum of the expected discretization error $\|\xi_{t-1}(\widehat{X}_t)\|_2^2$ defined in (26) *satisfies*

$$\sum_{t=2}^{T} \mathbb{E}[\|\xi_{t-1}(\widehat{X}_t)\|_2^2] \lesssim \sum_{t=2}^{T} \frac{d\overline{\alpha}_t \log^3 T}{T^2(1-\overline{\alpha}_t)} \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \frac{\mathbb{E}\big[\, \big\|\Sigma_\tau(\overline{X}_\tau)\big\|_{\mathsf{op}}^2 \,\big]}{(1-\tau)^2} \mathrm{d}\tau + \frac{d \min\{d \log T, L\} \log^3 T}{T^2},$$

where

$$\widetilde{\tau}_{t-1} = 1 - \widetilde{\alpha}_{t-1}, \quad \text{with} \quad \widetilde{\alpha}_{t-1} \coloneqq \frac{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\alpha_t) + \alpha_t(1-\overline{\alpha}_{t-1})}. \tag{28}$$

Recalling the definition of η_t and σ_t in (9), the estimation error can be written as

$$\frac{\eta_t^2}{\sigma_t^2}\|s_t(x_t)-s_t^\star(x_t)\|_2^2=(1-\alpha_t)\|s_t(x_t)-s_t^\star(x_t)\|_2^2=\frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{\overline{\alpha}_{t-1}}\|s_t(x_t)-s_t^\star(x_t)\|_2^2. \quad (29)$$

Applying the learning rate relationship (10), it follows that $\frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{\overline{\alpha}_{t-1}}\le\frac{c\log T}{T}(1-\overline{\alpha}_t)$, provided that $T\ge 2c\log T$. Putting the previous inequality and (29) together, one has

$$\frac{\eta_t^2}{\sigma_t^2}\|s_t(x_t)-s_t^\star(x_t)\|_2^2\le\frac{c\log T}{T}(1-\overline{\alpha}_t)\|s_t(x_t)-s_t^\star(x_t)\|_2^2.$$

By virtue of Lemma 4.1, Assumption 3.3 and the previous inequality, we arrive at

$$\begin{aligned}\sum_{t=2}^{T}\mathbb{E}_{x_t\sim\widehat{X}_t}\left[\frac{\eta_t^2}{\sigma_t^2}\|s_t(x_t)-s_t^\star(x_t)\|_2^2\right]&\le\frac{c\log T}{T}\sum_{t=2}^{T}(1-\overline{\alpha}_t)\mathbb{E}_{x_t\sim\widehat{X}_t}[\|s_t(x_t)-s_t^\star(x_t)\|_2^2]\\&=\frac{c\log T}{T}\sum_{t=2}^{T}(1-\overline{\alpha}_t)\mathbb{E}_{x_t\sim X_t}[\|s_t(x_t)-s_t^\star(x_t)\|_2^2]\\&=c\log T\varepsilon_{\mathsf{score}}^2. \quad (30)\end{aligned}$$

Step 3. putting everything together. Combining (21), (26), (30) and Lemma 4.2, we obtain an upper bound on the expectation of the KL divergence:

$$\begin{aligned}\mathsf{TV}^2(p_{Y_1},p_{X_1})&\le\frac{1}{2}\mathsf{KL}\left(p_{X_1}\|p_{Y_1}\right)\\&\lesssim\mathsf{KL}(p_{\widehat{X}_T}\|p_{Y_T})+\sum_{t=2}^{T}\frac{d\overline{\alpha}_t\log^3 T}{T^2(1-\overline{\alpha}_t)}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_2^2}{(1-\tau)^2}\mathrm{d}\tau\\&\quad+\frac{d\min\{d\log T,L\}\log^3 T}{T^2}+\log T\varepsilon_{\mathsf{score}}^2. \quad (31)\end{aligned}$$

To prove the desired error rate, one still needs to bound $\mathsf{KL}(p_{\widehat{X}_T}\|p_{Y_T})$ and $\sum_{t=2}^{T}\frac{d\overline{\alpha}_t\log^3 T}{T^2(1-\overline{\alpha}_t)}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_{\mathsf{op}}^2}{(1-\tau)^2}\mathrm{d}\tau$. The following lemma gives the required bounds.

Lemma 4.3. *Suppose that the learning rates satisfy* (10)*. Then we have*

$$\begin{aligned}\mathsf{KL}\left(p_{\widehat{X}_T}\|p_{Y_T}\right)&\lesssim\frac{1}{T^{10}},\\\sum_{t=2}^{T}\frac{\overline{\alpha}_t}{(1-\overline{\alpha}_t)}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_{\mathsf{op}}^2}{(1-\tau)^2}\mathrm{d}\tau&\lesssim\min\{L^2,d\}\log T. \quad (32)\end{aligned}$$

The proof of Lemma 4.3 is deferred to Appendix D. Lemma 4.3 together with (31) implies

$$\begin{aligned}\mathsf{TV}(p_{X_1},p_{Y_1})&\le\sqrt{\frac{1}{2}\mathsf{KL}(p_{X_1}\|p_{Y_1})}\lesssim\frac{1}{T^5}+\frac{d^{1/2}\min\{d^{1/2},L\}\log^2 T}{T}\\&\quad+\frac{d^{1/2}\min\{d^{1/2}\log^{1/2}T,L^{1/2}\}\log^{3/2}T}{T}+\varepsilon_{\mathsf{score}}\log^{1/2}T\\&\lesssim\frac{d^{1/2}\min\{d^{1/2},L\}\log^2 T}{T}+\varepsilon_{\mathsf{score}}\log^{1/2}T. \quad (33)\end{aligned}$$

5 Discussion

This paper provides an optimal convergence analysis of the DDPM sampler. Under a minimal assumption on the second moment of the target distribution and a mild non-uniform Lipschitz assumption, we show that the DDPM sampler's convergence rate scales as $O(\sqrt{d})$, which significantly improves the state-of-the-art linear dependence on d. Our convergence guarantees in both total variation and KL divergence are tight for a broad class of distributions. These results also challenge

the prevailing view that DDPM is inherently slower than DDIM due to linear dependence on d: we prove that DDPM and DDIM share the same d-dependence, and the empirical advantage of DDIM remains an open question.

Beyond this work, there are many future directions that are worth investigating. For example, our theory focuses on the original DDPM sampler; it would be interesting to see whether our analysis framework can sharpen the convergence rates for accelerated samplers (Luhman and Luhman, 2021; Lu et al., 2022; Zheng et al., 2023; Zhao et al., 2024; Zhang and Chen, 2023; Jolicoeur-Martineau et al., 2021; Xue et al., 2024; Li et al., 2024a; Huang et al., 2024a; Wu et al., 2024; Huang et al., 2025; Li et al., 2025b; Jiao et al., 2025b). In addition, it remains unclear if we can obtain sharper convergence rates when the target distribution exhibits low-dimensional structure (e.g., low intrinsic dimension (Li and Yan, 2024a,b; Huang et al., 2024c)). Furthermore, it would be worthwhile to explore if the non-uniform Lipschitz property holds for all distributions on $\mathbb{R}^d$ with $L \lesssim \mathrm{poly}(\log(dT))$.

Acknowledgments and Disclosure of Funding

G. Li is supported in part by the Chinese University of Hong Kong Direct Grant for Research and the Hong Kong Research Grants Council ECS 24305724 and GRF 14307525. The authors declare no competing interests.

References

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. (2023). Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*.

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

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2024). The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36.

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

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869.

De Bortoli, V. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

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

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

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

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Holzmüller, D. and Bach, F. (2023). Convergence rates for non-log-concave sampling and log-partition estimation. *arXiv preprint arXiv:2303.03237*.

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

Huang, D. Z., Huang, J., and Lin, Z. (2025). Fast convergence for high-order ode solvers in diffusion probabilistic models. *arXiv preprint arXiv:2506.13061*.

Huang, X., Zou, D., Dong, H., Zhang, Y., Ma, Y.-A., and Zhang, T. (2024b). Reverse transition kernel: A flexible framework to accelerate diffusion inference. *arXiv preprint arXiv:2405.16387*.

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

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

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

Jain, N. and Zhang, T. (2025). A sharp kl-convergence analysis for diffusion models under minimal assumptions. *arXiv preprint arXiv:2508.16306*.

Jiao, Y., Chen, Y., and Li, G. (2025a). Towards a unified framework for guided diffusion models. *arXiv preprint arXiv:2512.04985*.

Jiao, Y., Li, N., Cai, C., and Li, G. (2025b). Are first-order diffusion samplers really slower? a fast forward-value approach. *arXiv preprint arXiv:2512.24927*.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., Kachman, T., and Mitliagkas, I. (2021). Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338.

Lee, H., Lu, J., and Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882.

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. PMLR.

Li, G. and Cai, C. (2024). Provable acceleration for diffusion models under minimal assumptions. *arXiv preprint arXiv:2410.23285*.

Li, G. and Cai, C. (2025). A convergence theory for diffusion language models: An information-theoretic perspective. *arXiv preprint arXiv:2505.21400*.

Li, G., Cai, C., and Wei, Y. (2025a). Dimension-free convergence of diffusion models for approximate gaussian mixtures. *arXiv preprint arXiv:2504.05300*.

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

Li, G. and Jiao, Y. (2024). Improved convergence rate for diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*.

Li, G. and Jiao, Y. (2025). Provable efficiency of guidance in diffusion models for general data distribution. In *Forty-second International Conference on Machine Learning*.

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

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2024b). A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*.

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

Li, G. and Yan, Y. (2024b). $O(d/T)$ convergence theory for diffusion probabilistic models under minimal assumptions. *arXiv preprint arXiv:2409.18959*.

Li, G., Zhou, Y., Wei, Y., and Chen, Y. (2025b). Faster diffusion models via higher-order approximation. *arXiv preprint arXiv:2506.24042*.

Li, R., Di, Q., and Gu, Q. (2024c). Unified convergence analysis for score-based diffusion models with deterministic samplers. *arXiv preprint arXiv:2410.14237*.

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. (2022). Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Liang, J., Huang, Z., and Chen, Y. (2025). Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. (2023). Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

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

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471.

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

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

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

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*, 1(2):3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

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

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

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

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. (2022). Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*.

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

Wu, Y., Chen, Y., and Wei, Y. (2024). Stochastic runge-kutta methods: Provable acceleration of diffusion models. *arXiv preprint arXiv:2410.04760*.

Xu, X. and Chi, Y. (2024). Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *Advances in Neural Information Processing Systems*, 37:36148–36184.

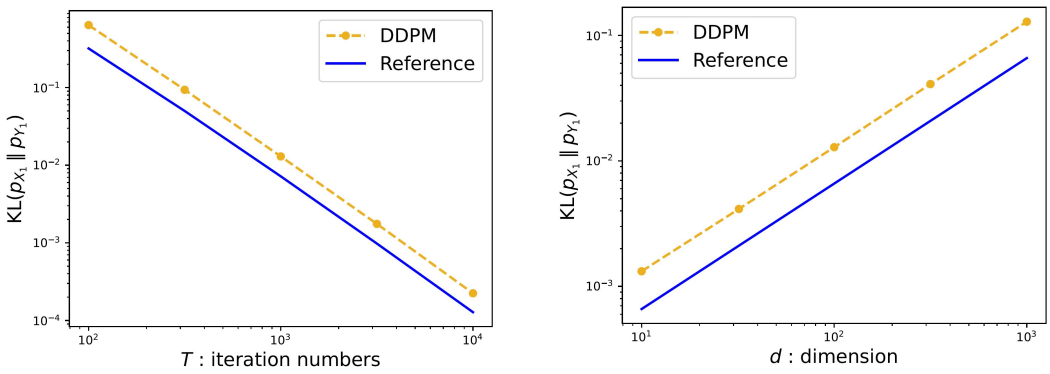


Figure 1: KL divergence of DDPM and fitted rate $\Theta(d \log^4 T/T^2)$: (a) $d = 100$; (b) $T = 1000$.

Xue, S., Yi, M., Luo, W., Zhang, S., Sun, J., Li, Z., and Ma, Z.-M. (2024). Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Zhang, Q. and Chen, Y. (2023). Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*.

Zhao, W., Bai, L., Rao, Y., Zhou, J., and Lu, J. (2024). Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Zheng, K., Lu, C., Chen, J., and Zhu, J. (2023). Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36:55502–55542.

A Numerical experiments

A.1 Evaluation of convergence rate

In this section, we conduct numerical experiments to verify our theoretical findings. Specifically, we take the target distribution p_{data} to be a d-dimensional Gaussian distribution with zero mean and independent coordinates: its covariance is diagonal with variances $\{\sigma_i^2\}_{i=1}^d$ drawn i.i.d. from Unif$[0, 10]$. We assume access to the exact score, i.e., $s_t(\cdot) = s_t^\star(\cdot)$. Parameters η_t and σ_t^2 are chosen as (9), and $\{\overline{\alpha}_t\}_{t=1}^T$ (or $\{\beta_t\}_{t=1}^T$) are set as (11) with $c_0 = 2$ and $c_1 = 4$. Under this choice, each Y_t in (8) is Gaussian, and hence the KL divergence between X_1 and Y_1 has a closed-form expression.

To show the dependence on the number of iterations T and data dimension d, we implement DDPM (8) with d fixed at 100 and the number of iterations T varying from 100 to 10000, and with T fixed at 1000 and d varying from 10 to 1000, and compute the KL divergence between distributions p_{X_1} and the output Y_1.

The results are presented in Figure 1. The orange line represents the empirical results, and the blue line corresponds to the theoretical rate $O(\log^2 T/T^2)$. According to Theorem 3.4, our theoretical analysis predicts a convergence rate of $O(d \log^4 T/T^2)$ in terms of KL divergence in case of Gaussian distribution, which is consistent with empirical observations ignoring dependence on logarithmic factors. Refining this dependency requires further effort and is left for future work.

A.2 Empirical evaluation of the non-uniform Lipschitz parameter

We next assess the practical magnitude of the non-uniform Lipschitz parameter L in Definition 3.2 using pretrained diffusion models. We consider three settings: CIFAR-10 with EDM, ImageNet at a resolution of 512×512 with EDM2-XXL, and LSUN Church with a latent diffusion model. In these experiments, s_t denotes the score estimated by the pretrained model.

Experimental setup. We evaluate the models at $T = 100$ time steps following a practical sampling schedule. Specifically, among the $N = 1000$ noise levels indexed by n, we use the levels with indices $n_t = 10t$ for $t = 1, \ldots, T$; below, $\overline{\alpha}_t$ denotes the cumulative signal coefficient at the selected noise level n_t. At each time step t, we draw 10 samples x_t from the forward process (2). For each sample, we estimate the operator norm of the score Jacobian using 10 iterations of a power method:

$$\widehat{J}_t(x_t) = \frac{1}{\delta} \left\| s_t(x_t + \delta \Delta_{10}) - s_t(x_t) \right\|_2 \approx \left\| \nabla s_t(x_t) \right\|_{\mathsf{op}},$$

where δ is a sufficiently small perturbation parameter and, starting from a unit vector Δ_0,

$$\Delta_k = \frac{s_t(x_t + \delta \Delta_{k-1}) - s_t(x_t)}{\left\| s_t(x_t + \delta \Delta_{k-1}) - s_t(x_t) \right\|_2}, \qquad k = 1, \ldots, 10.$$

We then compute the empirical estimate

$$\widehat{L} = \max_{(t, x_t) \text{ evaluated}} (1 - \overline{\alpha}_t) \widehat{J}_t(x_t).$$

Results. The estimated values $\widehat{L}$ for CIFAR-10, ImageNet, and LSUN Church are 6.4391, 8.8578, and 7.0160, respectively. These values are substantially smaller than the corresponding values of $\sqrt{d}$, which are approximately 55, 128, and 111, respectively, and are on the order of one-tenth of $\sqrt{d}$.

Sensitivity to the number of samples. To examine the effect of sample size, we repeat the CIFAR-10 experiment with 100 samples x_t at each time step, obtaining $\widehat{L} = 7.2149$. In this experiment, values of $(1 - \overline{\alpha}_t)\widehat{J}_t(x_t) > 5$ occur only at $t \leq 4$. We therefore draw an additional 10^4 samples at each of $t = 1, 2, 3, 4$ to examine the upper tail more closely. The upper 1% quantiles of $(1-\overline{\alpha}_t)\widehat{J}_t(x_t)$ at these four time steps are 8.8586, 6.1125, 5.1590, and 4.4391, respectively; the corresponding upper 0.1% quantiles are 9.8144, 6.9655, 5.8734, and 5.0647. These tail estimates remain modest relative to $\sqrt{d}$, providing further empirical evidence that the non-uniform Lipschitz parameter can be small in practical diffusion models.

B Proof of Lemma 4.1

The proof is similar to Lemma 2 in Li and Jiao (2024); we include it here for completeness. We shall establish Lemma 4.1 via induction. From the definition of $\widehat{X}_T$, we have $\widehat{X}_T \overset{\mathrm{d}}{=} X_T$. Assume that $\widehat{X}_t \overset{\mathrm{d}}{=} X_t \overset{\mathrm{d}}{=} \overline{X}_{\tau_t}$. We intend to prove that $\widehat{X}_{t-1} \overset{\mathrm{d}}{=} X_{t-1}$. In view of (19), we have

$$\begin{aligned}
\frac{\widehat{X}_{t-1}}{\sqrt{\overline{\alpha}_{t-1}}} &= \frac{1}{\sqrt{\overline{\alpha}_t}} \widehat{X}_t + \int_{\widehat{\tau}_{t-1}}^{\tau_t} \frac{\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{\frac{3}{2}}} \mathrm{d}\tau + \frac{\sigma_t}{\sqrt{\overline{\alpha}_t}} Z_t \\
&= \frac{1}{\sqrt{1-\tau_t}} \widehat{X}_t + \int_{\widehat{\tau}_{t-1}}^{\tau_t} \frac{\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{\frac{3}{2}}} \mathrm{d}\tau + \frac{\sigma_t}{\sqrt{\overline{\alpha}_t}} Z_t \\
&= \frac{1}{\sqrt{1-\widehat{\tau}_{t-1}}} \Phi_{\tau_t \to \widehat{\tau}_{t-1}}(\widehat{X}_t) + \frac{\sigma_t}{\sqrt{\overline{\alpha}_t}} Z_t \\
&\overset{\mathrm{d}}{=} \frac{\overline{X}_{\widehat{\tau}_{t-1}}}{\sqrt{1-\widehat{\tau}_{t-1}}} + \frac{\sigma_t}{\sqrt{\overline{\alpha}_t}} Z_t \qquad (34) \\
&\overset{\mathrm{d}}{=} X_0 + \left(\frac{\widehat{\tau}_{t-1}}{1-\widehat{\tau}_{t-1}} + \frac{\sigma_t^2}{\overline{\alpha}_t} \right)^{1/2} Z_t',
\end{aligned}$$

where $Z_t' \sim \mathcal{N}(0, I_d)$. Here, the penultimate line is due to (18), and the last line makes use of (4). It suffices to prove that

$$\frac{\widehat{\tau}_{t-1}}{1-\widehat{\tau}_{t-1}} + \frac{\sigma_t^2}{\overline{\alpha}_t} = \frac{1-\overline{\alpha}_{t-1}}{\overline{\alpha}_{t-1}},$$

which can be completed by direct calculation:

$$\frac{\widehat{\tau}_{t-1}}{1-\widehat{\tau}_{t-1}} + \frac{\sigma_t^2}{\overline{\alpha}_t} = \frac{2\alpha_t - 1 - \overline{\alpha}_t}{\overline{\alpha}_t} + \frac{1-\alpha_t}{\overline{\alpha}_t} = \frac{1-\overline{\alpha}_{t-1}}{\overline{\alpha}_{t-1}}.$$

C Proof of Lemma 4.2

Before diving into the proof details, we make the observation that

$$\int_{\widetilde{\tau}_{t-1}}^{\tau_t} \frac{1}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\mathrm{d}\tau \overset{\text{(i)}}{=} \sqrt{\frac{\tau_t}{1-\tau_t}} - \sqrt{\frac{\widetilde{\tau}_{t-1}}{1-\widetilde{\tau}_{t-1}}} \overset{\text{(ii)}}{=} \sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}} - \sqrt{\frac{1-\widetilde{\alpha}_{t-1}}{\widetilde{\alpha}_{t-1}}}, \tag{35}$$

where (i) holds since

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\sqrt{\frac{\tau}{1-\tau}} = \frac{1}{2}\sqrt{\frac{1-\tau}{\tau}}\frac{1}{(1-\tau)^2} = \frac{1}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}, \tag{36}$$

and (ii) arises from the definition of $\widetilde{\tau}_{t-1}$ in (28). Recalling the definition of $\widetilde{\alpha}_{t-1}$ (cf. (28)), one has

$$\begin{aligned}\frac{1-\widetilde{\alpha}_{t-1}}{\widetilde{\alpha}_{t-1}} &= \frac{\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})-\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)} \\ &= \frac{\alpha_t - 2\overline{\alpha}_t + \overline{\alpha}_t\overline{\alpha}_{t-1}}{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)} = \frac{\alpha_t(1-\overline{\alpha}_{t-1})^2}{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)} = \frac{\alpha_t^2(1-\overline{\alpha}_{t-1})^2}{\overline{\alpha}_t(1-\overline{\alpha}_t)} = \frac{(\alpha_t-\overline{\alpha}_t)^2}{\overline{\alpha}_t(1-\overline{\alpha}_t)}.\end{aligned}$$

Combining (35) and the previous equation, we obtain

$$\int_{\widetilde{\tau}_{t-1}}^{\tau_t} \frac{1}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\mathrm{d}\tau = \sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}} - \sqrt{\frac{1-\widetilde{\alpha}_{t-1}}{\widetilde{\alpha}_{t-1}}} = \sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}} - \frac{\alpha_t-\overline{\alpha}_t}{\sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)}} = \frac{1-\alpha_t}{\sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)}},$$

Consequently, $\eta_t s_t^\star(x)$ can be expressed as the integral

$$\begin{aligned}\eta_t s_t^\star(x) = \eta_t \overline{s}_{\tau_t}^\star(x) &= \frac{\eta_t}{\sqrt{\overline{\alpha}_t(1-\overline{\alpha}_t)}}\sqrt{\overline{\alpha}_t\tau_t}\overline{s}_{\tau_t}^\star(x) \\ &= \sqrt{\overline{\alpha}_t}\int_{\widetilde{\tau}_{t-1}}^{\tau_t} \frac{\sqrt{\tau_t}\overline{s}_{\tau_t}^\star(x)}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\mathrm{d}\tau.\end{aligned} \tag{37}$$

Now we are ready to bound the expectation of $\|\xi_{t-1}(x_t)\|_2^2$. Recalling the definition of $\mu_{\widehat{X}_{t-1}|\widehat{X}_t}(x_t)$, we have

$$\begin{aligned}\xi_{t-1}(x_t) &= \frac{\sqrt{\alpha_t}}{\sigma_t}\left(\frac{\sqrt{\overline{\alpha}_{t-1}}}{\sqrt{1-\widehat{\tau}_{t-1}}}\Phi_{\tau_t\to\widehat{\tau}_{t-1}}(x_t) - \frac{1}{\sqrt{\alpha_t}}(x_t+\eta_t s_t^\star(x_t))\right) \\ &= \frac{1}{\sigma_t}\left(x_t + \sqrt{\alpha_t}\int_{\widehat{\tau}_{t-1}}^{\tau_t}\frac{\sqrt{\overline{\alpha}_{t-1}}\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{\frac{3}{2}}}\mathrm{d}\tau - x_t - \eta_t s_t^\star(x_t)\right) \\ &= \frac{1}{\sigma_t}\left(\sqrt{\overline{\alpha}_t}\int_{\widehat{\tau}_{t-1}}^{\tau_t}\frac{\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{\frac{3}{2}}}\mathrm{d}\tau - \eta_t s_t^\star(x_t)\right).\end{aligned} \tag{38}$$

We introduce the following auxiliary variable:

$$\zeta_t(x_t) \coloneqq \frac{\sqrt{\overline{\alpha}_t}}{\sigma_t}\int_{\widehat{\tau}_{t-1}}^{\widetilde{\tau}_{t-1}}\frac{\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{\frac{3}{2}}}\mathrm{d}\tau. \tag{39}$$

Putting (37) and (38) together, it follows that

$$\begin{aligned}\xi_{t-1}(x_t) &= \frac{\sqrt{\overline{\alpha}_t}}{\sigma_t}\left(\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\overline{s}_\tau^\star(x_\tau)}{2(1-\tau)^{\frac{3}{2}}}\mathrm{d}\tau - \int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\sqrt{\tau_t}\overline{s}_{\tau_t}^\star(x_t)}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\mathrm{d}\tau\right) + \zeta_t(x_t) \\ &= \frac{\sqrt{\overline{\alpha}_t}}{\sigma_t}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{1}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\left(\sqrt{\tau}\overline{s}_\tau^\star(x_\tau) - \sqrt{\tau_t}\overline{s}_{\tau_t}^\star(x_t)\right)\mathrm{d}\tau + \zeta_t(x_t) \\ &= -\frac{\sqrt{\overline{\alpha}_t}}{\sigma_t}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{1}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\int_{\tau}^{\tau_t}\frac{\partial\sqrt{\tau'}\overline{s}_{\tau'}^\star(x'_\tau)}{\partial\tau'}\mathrm{d}\tau'\mathrm{d}\tau + \zeta_t(x_t) \\ &= -\frac{\sqrt{\overline{\alpha}_t}}{\sigma_t}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\partial\sqrt{\tau'}\overline{s}_{\tau'}^\star(x'_\tau)}{\partial\tau'}\int_{\widetilde{\tau}_{t-1}}^{\tau'}\frac{1}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\mathrm{d}\tau\mathrm{d}\tau' + \zeta_t(x_t)\end{aligned}$$

$$= \frac{\sqrt{\overline{\alpha}_t}}{\sigma_t} \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \left(\sqrt{\frac{\widetilde{\tau}_{t-1}}{1-\widetilde{\tau}_{t-1}}} - \sqrt{\frac{\tau'}{1-\tau'}} \right) \frac{\partial \sqrt{\tau'} \overline{s}^\star_{\tau'}(x'_\tau)}{\partial \tau'} \mathrm{d}\tau' + \zeta_t(x_t), \tag{40}$$

where the third line makes use of fundamental theorem of calculus, the fourth line invokes follows by exchanging the order of integration, and the last equation holds due to (36). Therefore, one can bound the norm of $\xi_{t-1}(x_t) - \zeta_t(x_t)$ as follows

$$\|\xi_{t-1}(x_t) - \zeta_t(x_t)\|_2 \le \left| \sqrt{\frac{\widetilde{\tau}_{t-1}}{1-\widetilde{\tau}_{t-1}}} - \sqrt{\frac{\tau_t}{1-\tau_t}} \right| \frac{\sqrt{\overline{\alpha}_t}}{\sigma_t} \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \left\| \frac{\partial \sqrt{\tau'} \overline{s}^\star_{\tau'}(x'_\tau)}{\partial \tau'} \right\|_2 \mathrm{d}\tau'. \tag{41}$$

In the rest of the proof, we bound $\|\xi_{t-1}(x_t)\|_2^2$ using the following lemma, which establishes key properties of $\widetilde{\tau}_{t-1}$. Its proof is postponed to Appendix C.1.

Lemma C.1. *For any $1 \le t \le T$, the $\widetilde{\tau}_{t-1}$ and $\widetilde{\alpha}_{t-1}$ defined in* (20) *satisfy*

$$\widetilde{\tau}_{t-1} \le \tau_{t-1}, \qquad \tau_t - \widetilde{\tau}_{t-1} = \widetilde{\alpha}_{t-1} - \overline{\alpha}_t \lesssim \frac{\log T}{T} \overline{\alpha}_{t-1}(1-\overline{\alpha}_t), \tag{42a}$$

$$\frac{1-\overline{\alpha}_t}{1-\widetilde{\alpha}_{t-1}} \lesssim 1, \tag{42b}$$

$$\left| \sqrt{\frac{\tau_t}{1-\tau_t}} - \sqrt{\frac{\widetilde{\tau}_{t-1}}{1-\widetilde{\tau}_{t-1}}} \right| \lesssim \frac{\log T}{T} \sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}}, \tag{42c}$$

$$\mathbb{E}[\|\zeta_t(x_t)\|_2^2] \lesssim \frac{d \log^3 T}{T^3}. \tag{42d}$$

Inserting (42c) into (41) yields

$$\|\xi_{t-1}(x_t) - \zeta_t(x_t)\|_2 \le \frac{2c \log T}{T} \frac{\sqrt{1-\overline{\alpha}_t}}{\sigma_t} \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \left\| \frac{\partial \sqrt{\tau'} \overline{s}^\star_{\tau'}(x'_\tau)}{\partial \tau'} \right\|_2 \mathrm{d}\tau'.$$

Applying Cauchy-Schwarz inequality leads to

$$\begin{aligned}
\|\xi_{t-1}(x_t) - \zeta_t(x_t)\|_2^2 &\le \frac{4c^2 \log^2 T}{T^2} \frac{1-\overline{\alpha}_t}{\sigma_t^2} (\tau_t - \widetilde{\tau}_{t-1}) \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \left\| \frac{\partial \sqrt{\tau} \overline{s}^\star_{\tau}(x_\tau)}{\partial \tau} \right\|_2^2 \mathrm{d}\tau \\
&\overset{\text{(i)}}{=} \frac{4c^2 \log^2 T}{T^2} \frac{1-\overline{\alpha}_t}{1-\alpha_t} (\tau_t - \widetilde{\tau}_{t-1}) \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \left\| \frac{\partial \sqrt{\tau} \overline{s}^\star_{\tau}(x_\tau)}{\partial \tau} \right\|_2^2 \mathrm{d}\tau \\
&\overset{\text{(ii)}}{\lesssim} \frac{\log^2 T}{T^2} \overline{\alpha}_t (1-\overline{\alpha}_t) \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \left\| \frac{\partial \sqrt{\tau} \overline{s}^\star_{\tau}(x_\tau)}{\partial \tau} \right\|_2^2 \mathrm{d}\tau,
\end{aligned} \tag{43}$$

where (i) invokes the definition of σ_t^2 in (9), and (ii) holds since

$$\begin{aligned}
\frac{\tau_t - \widetilde{\tau}_{t-1}}{1-\alpha_t} &\overset{(5)}{=} \frac{\widetilde{\alpha}_{t-1} - \overline{\alpha}_t}{1-\alpha_t} \overset{(28)}{=} \frac{1}{1-\alpha_t} \Big(\frac{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\alpha_t) + \alpha_t(1-\overline{\alpha}_{t-1})} - \overline{\alpha}_t \Big) \\
&= \frac{\overline{\alpha}_{t-1}}{1-\alpha_t} \frac{1 - \overline{\alpha}_t - \overline{\alpha}_t(1-\alpha_t) - \alpha_t^2(1-\overline{\alpha}_{t-1})}{\overline{\alpha}_{t-1}(1-\alpha_t) + \alpha_t(1-\overline{\alpha}_{t-1})} \\
&= \frac{\overline{\alpha}_{t-1}}{1-\alpha_t} \frac{(1-\alpha_t)(1+\alpha_t - 2\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\alpha_t) + \alpha_t(1-\overline{\alpha}_{t-1})} \\
&= \frac{\overline{\alpha}_{t-1}(1+\alpha_t - 2\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\alpha_t) + \alpha_t(1-\overline{\alpha}_{t-1})} \\
&\le \frac{2\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)}{\alpha_t(1-\overline{\alpha}_{t-1})} \lesssim \overline{\alpha}_t,
\end{aligned} \tag{44}$$

where the last inequality follows from (49) and (50). Taking expectation with respect to x_t in (43) and applying Lemma 4.1 together with (18), we obtain

$$\mathbb{E}_{x_t \sim \widehat{X}_t} \left[\|\xi_{t-1}(x_t) - \zeta_t(x_t)\|_2^2 \right]$$

$$\lesssim \frac{c^2\log^2 T}{T^2}\overline{\alpha}_t(1-\overline{\alpha}_t)\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\left\|\frac{\partial\sqrt{\tau}\overline{s}_\tau^\star(x_\tau)}{\partial\tau}\right\|_2^2\right]\mathrm{d}\tau. \tag{45}$$

Furthermore, the following lemma bounds the right-hand side of (45); its proof is deferred to Appendix C.2.

Lemma C.2. *For any $0<\tau<1$, we have*

$$\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\left\|\frac{\partial\sqrt{\tau}\overline{s}_\tau^\star(x_\tau)}{\partial\tau}\right\|_2^2\right]\lesssim\frac{d}{\tau^2(1-\tau)^2}\left(\log T\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left\|\Sigma_\tau(x_\tau)\right\|_{\mathsf{op}}^2+1\right)+\frac{d\min\left\{d\log T,L\right\}}{\tau^2(1-\tau)^2}. \tag{46}$$

By virtue of Lemma C.2, (42c) and (45), one has

$$\begin{aligned}
\mathbb{E}_{x_t\sim\widehat{X}_t}\left[\|\xi_{t-1}(x_t)\|_2^2\right]&\le 2\left(\mathbb{E}_{x_t\sim\widehat{X}_t}\left[\|\xi_{t-1}(x_t)-\|\zeta_t(x_t)\|_2^2\|_2^2\right]+\mathbb{E}_{x_t\sim\widehat{X}_t}\left[\|\|\zeta_t(x_t)\|_2^2\|_2^2\right]\right)\\
&\lesssim\frac{d\overline{\alpha}_t(1-\overline{\alpha}_t)\log^3 T}{T^2\widetilde{\tau}_{t-1}^2}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left\|\Sigma_\tau(x_\tau)\right\|_{\mathsf{op}}^2}{(1-\tau)^2}\mathrm{d}\tau\\
&\quad+\frac{d\min\left\{d\log T,L\right\}\overline{\alpha}_t(1-\overline{\alpha}_t)\log^2 T}{T^2\widetilde{\tau}_{t-1}^2(1-\tau_t)^2}(\tau_t-\widetilde{\tau}_{t-1})+\frac{d\log^3 T}{T^3}\\
&\lesssim\frac{d\overline{\alpha}_t\log^3 T}{T^2(1-\overline{\alpha}_t)}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left\|\Sigma_\tau(x_\tau)\right\|_{\mathsf{op}}^2}{(1-\tau)^2}\mathrm{d}\tau+\frac{d\min\left\{d\log T,L\right\}\log^3 T}{T^3},
\end{aligned}$$

where the last inequality makes use of the following facts:

$$\begin{aligned}
\frac{\overline{\alpha}_t(1-\overline{\alpha}_t)}{T^2\widetilde{\tau}_{t-1}^2}&\overset{\text{(42a)}}{\le}\frac{\overline{\alpha}_t}{T^2(1-\overline{\alpha}_t)}\frac{(1-\overline{\alpha}_t)^2}{(1-\overline{\alpha}_{t-2})^2}\lesssim\frac{\overline{\alpha}_t}{T^2(1-\overline{\alpha}_t)},\\
\frac{\overline{\alpha}_t(1-\overline{\alpha}_t)}{T^2\widetilde{\tau}_{t-1}^2(1-\tau_t)^2}(\tau_t-\widetilde{\tau}_{t-1})&\overset{\text{(42a)}}{\lesssim}\frac{\log T}{T^3}\frac{\overline{\alpha}_{t-1}}{\overline{\alpha}_t}\frac{(1-\overline{\alpha}_t)^2}{(1-\overline{\alpha}_{t-2})^2}\lesssim\frac{\log T}{T^2}.
\end{aligned}$$

C.1 Proof of Lemma C.1

Recalling the definitions of $\widetilde{\tau}_{t-1}$ and $\widetilde{\alpha}_{t-1}$ (cf. (28)), we have

$$\begin{aligned}
\widetilde{\tau}_{t-1}-\tau_{t-1}=\overline{\alpha}_{t-1}-\widetilde{\alpha}_{t-1}&=\overline{\alpha}_{t-1}-\frac{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})}\\
&=\overline{\alpha}_{t-1}\frac{\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})-(1-\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})}\le 0,
\end{aligned} \tag{47}$$

where the last inequality arises from

$$\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})-(1-\overline{\alpha}_t)=\overline{\alpha}_{t-1}+\alpha_t-\overline{\alpha}_t-1=(\overline{\alpha}_{t-1}-1)(1-\alpha_t)\le 0.$$

Furthermore, one can show that

$$\tau_t-\widetilde{\tau}_{t-1}\overset{(1)}{\lesssim}\overline{\alpha}_t(1-\alpha_t)=\alpha_t(\overline{\alpha}_{t-1}-\overline{\alpha}_t)\overset{\text{(ii)}}{\le}\frac{c\log T}{T}\overline{\alpha}_t(1-\overline{\alpha}_t)\le\frac{c\log T}{T}\overline{\alpha}_{t-1}(1-\overline{\alpha}_t),$$

where (i) was proved in (44), and (ii) arises from assumption (10). This has finished the proof of the last inequality in (42a).

Next, we verify (42b) via a direct calculation:

$$\begin{aligned}
\frac{1-\overline{\alpha}_t}{1-\widetilde{\alpha}_{t-1}}&\overset{(20)}{=}(1-\overline{\alpha}_t)\frac{\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})}{\alpha_t(1-\overline{\alpha}_{t-1})^2}=\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}\frac{\overline{\alpha}_{t-1}(\overline{\alpha}_{t-1}-\overline{\alpha}_t)+\overline{\alpha}_t(1-\overline{\alpha}_{t-1})}{\overline{\alpha}_t(1-\overline{\alpha}_{t-1})}\\
&\overset{(10)}{\le}\frac{1}{1-\frac{c\overline{\alpha}_t\log T}{T}}\left(1+\frac{c\overline{\alpha}_{t-1}\log T}{T}\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}\right)\le\frac{1}{1-\frac{c\log T}{T}}\left(1+\frac{2c\log T}{T}\right)\le 2,
\end{aligned}$$

as long as $T \geq 4c\log T$, where the last inequality makes use of the fact that

$$\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} = 1 + \frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} \leq 1 + \frac{c\overline{\alpha}_t \log T}{T}\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}} \leq \frac{1}{1-\frac{c\overline{\alpha}_t\log T}{T}} \leq \frac{4}{3}. \tag{48}$$

Now we turn to establishing (42c). In view of (36), we have

$$\begin{aligned}
\left|\sqrt{\frac{\tau_t}{1-\tau_t}} - \sqrt{\frac{\widetilde{\tau}_{t-1}}{1-\widetilde{\tau}_{t-1}}}\right| &= \int_{\widetilde{\tau}_{t-1}}^{\tau_t} \frac{1}{2(1-\tau)^{\frac{3}{2}}\tau^{\frac{1}{2}}}\mathrm{d}\tau \\
&\leq \frac{\tau_t - \widetilde{\tau}_{t-1}}{2(1-\tau_t)^{\frac{3}{2}}\widetilde{\tau}_{t-1}^{\frac{1}{2}}} = \frac{\widetilde{\alpha}_{t-1}-\overline{\alpha}_t}{2\overline{\alpha}_t^{\frac{3}{2}}(1-\widetilde{\alpha}_{t-1})^{\frac{1}{2}}} \\
&\overset{\text{(i)}}{\leq} \frac{c\log T}{T}\frac{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)}{\overline{\alpha}_t^{\frac{3}{2}}(1-\widetilde{\alpha}_{t-1})^{\frac{1}{2}}} \overset{\text{(ii)}}{\leq} \frac{2c\log T}{T}\sqrt{\frac{1-\overline{\alpha}_t}{\overline{\alpha}_t}},
\end{aligned}$$

where (i) invokes (42a), and (ii) arises from (48) and

$$\frac{\overline{\alpha}_{t-1}}{\overline{\alpha}_t}\sqrt{\frac{1-\overline{\alpha}_t}{1-\widetilde{\alpha}_{t-1}}} \overset{\text{(10) and (42b)}}{\leq} \left(1+\frac{c\log T}{T}\right)\sqrt{2} \leq 2,$$

provided that $T \geq 4c\log T$.

Finally, we show that (42d) holds. Eqn. (10) tells us that

$$1-\alpha_t \leq \frac{c\log T}{T}\alpha_t(1-\overline{\alpha}_t) \leq \frac{c\log T}{T} \leq \frac{1}{4}, \tag{49}$$

$$1-\overline{\alpha}_{t-1} \geq 1-\overline{\alpha}_t - \frac{c\log T}{T}\overline{\alpha}_t(1-\overline{\alpha}_t) \geq \left(1-\frac{c\log T}{T}\right)(1-\overline{\alpha}_t) \geq \frac{1}{2}(1-\overline{\alpha}_t). \tag{50}$$

We make the observation that

$$\begin{aligned}
\widetilde{\tau}_{t-1} - \widehat{\tau}_{t-1} &= \widehat{\alpha}_{t-1} - \widetilde{\alpha}_{t-1} \\
&= \frac{\overline{\alpha}_t}{2\alpha_t - 1} - \frac{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})} \\
&= \frac{\overline{\alpha}_{t-1}(1-\alpha_t)^2}{(2\alpha_t-1)(\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1}))} \\
&\overset{(49)}{\geq} 0.
\end{aligned} \tag{51}$$

Applying Cauchy-Schwarz inequality yields

$$\begin{aligned}
\|\zeta_t(x_t)\|_2^2 &\leq \frac{\overline{\alpha}_t}{\sigma_t^2}\int_{\widehat{\tau}_{t-1}}^{\widetilde{\tau}_{t-1}} \frac{1}{4\tau(1-\tau)^3}\mathrm{d}\tau \int_{\widehat{\tau}_{t-1}}^{\widetilde{\tau}_{t-1}} \tau\|\overline{s}_\tau^\star(x_\tau)\|_2^2\mathrm{d}\tau \\
&\leq \frac{\overline{\alpha}_t}{1-\alpha_t}\frac{\widetilde{\tau}_{t-1}-\widehat{\tau}_{t-1}}{4\widehat{\tau}_{t-1}(1-\widetilde{\tau}_{t-1})^3}\int_{\widehat{\tau}_{t-1}}^{\widetilde{\tau}_{t-1}} \tau\|\overline{s}_\tau^\star(x_\tau)\|_2^2\mathrm{d}\tau.
\end{aligned}$$

As a consequence, we have

$$\begin{aligned}
\mathbb{E}_{x_t\sim X_t}[\|\zeta_t(x_t)\|_2^2] &\leq \frac{\overline{\alpha}_t}{1-\alpha_t}\frac{\widetilde{\tau}_{t-1}-\widehat{\tau}_{t-1}}{4\widehat{\tau}_{t-1}(1-\widetilde{\tau}_{t-1})^3}\int_{\widehat{\tau}_{t-1}}^{\widetilde{\tau}_{t-1}} \tau\mathbb{E}_{x_\tau\sim\overline{X}_\tau}[\|\overline{s}_\tau^\star(x_\tau)\|_2^2]\mathrm{d}\tau \\
&\overset{(91)}{\leq} \frac{\overline{\alpha}_t}{1-\alpha_t}\frac{\widetilde{\tau}_{t-1}-\widehat{\tau}_{t-1}}{4\widehat{\tau}_{t-1}(1-\widetilde{\tau}_{t-1})^3}\int_{\widehat{\tau}_{t-1}}^{\widetilde{\tau}_{t-1}} \tau\left(\frac{d^4}{\tau^4}\right)^{1/4}\mathrm{d}\tau \\
&= \frac{\overline{\alpha}_t}{1-\alpha_t}\frac{(\widetilde{\tau}_{t-1}-\widehat{\tau}_{t-1})^2 d}{4\widehat{\tau}_{t-1}(1-\widetilde{\tau}_{t-1})^3} \\
&= \frac{\overline{\alpha}_t}{1-\alpha_t}\frac{(\widetilde{\alpha}_{t-1}-\widehat{\alpha}_{t-1})^2 d}{4\widetilde{\alpha}_{t-1}^3(1-\widehat{\alpha}_{t-1})}.
\end{aligned} \tag{52}$$

In view of (51), one has

$$\begin{aligned}\widehat{\alpha}_{t-1}-\widetilde{\alpha}_{t-1} &= \frac{\overline{\alpha}_{t-1}(1-\alpha_t)^2}{(2\alpha_t-1)(\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1}))} \\ &\le \frac{c\frac{\log T}{T}\overline{\alpha}_t(1-\overline{\alpha}_t)}{\frac{1}{2}\alpha_t(1-\overline{\alpha}_{t-1})}(1-\alpha_t) \\ &\lesssim \frac{\log T}{T}\overline{\alpha}_{t-1}(1-\alpha_t), \end{aligned} \tag{53}$$

where the second line arises from (10) and (49) and the last line makes use (50). Moreover, we can verify that

$$\begin{aligned}1-\widehat{\alpha}_{t-1} &= 1-\frac{\overline{\alpha}_t}{2\alpha_t-1} = 1-\frac{\overline{\alpha}_t\cdot\overline{\alpha}_{t-1}}{\overline{\alpha}_t-(\overline{\alpha}_{t-1}-\overline{\alpha}_t)} \overset{(10)}{\ge} 1-\frac{\overline{\alpha}_t\cdot\overline{\alpha}_{t-1}}{\overline{\alpha}_t-\frac{c\log T}{T}\overline{\alpha}_t(1-\overline{\alpha}_t)} = 1-\frac{\overline{\alpha}_{t-1}}{1-\frac{c\log T}{T}(1-\overline{\alpha}_t)} \\ &\ge 1-\frac{\overline{\alpha}_{t-1}}{1-\frac{c\log T}{T}(1-\overline{\alpha}_{t-1})} = (1-\overline{\alpha}_{t-1})\left(1-\frac{c\log T}{T}\right)\frac{1}{1-\frac{c\log T}{T}(1-\overline{\alpha}_{t-1})} \\ &\ge \frac{1-\overline{\alpha}_{t-1}}{4} \ge \frac{1-\overline{\alpha}_t}{8}. \end{aligned} \tag{54}$$

and

$$\widetilde{\alpha}_{t-1} \overset{(47)}{\ge} \overline{\alpha}_{t-1}. \tag{55}$$

Putting (52) - (55) together, one arrives at

$$\begin{aligned}\mathbb{E}[\|\zeta_t(x_t)\|_2^2] &\lesssim \frac{\overline{\alpha}_t}{1-\alpha_t}\frac{\left(\frac{\log T}{T}\overline{\alpha}_{t-1}(1-\alpha_t)\right)^2 d}{\overline{\alpha}_{t-1}^3(1-\overline{\alpha}_t)} \\ &= d\frac{\log^2 T}{T^2}\alpha_t\frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)} \\ &\overset{(10)}{\lesssim} d\frac{\log^2 T}{T^2}\alpha_t\frac{\frac{\log T}{T}\overline{\alpha}_t(1-\overline{\alpha}_t)}{\overline{\alpha}_{t-1}(1-\overline{\alpha}_t)} \\ &\le \frac{d\log^3 T}{T^3}. \end{aligned}$$

C.2 Proof of Lemma C.2

To begin with, let us write

$$\frac{\partial\sqrt{\tau}\overline{s}_\tau^\star(x_\tau)}{\partial\tau} = \frac{\partial}{\partial y}\sqrt{\tau}\overline{s}_\tau^\star(\sqrt{1-\tau}y)\big|_{y=x_\tau/\sqrt{1-\tau}}\frac{\partial x_\tau/\sqrt{1-\tau}}{\partial\tau} + \frac{\partial}{\partial\tau}\sqrt{\tau}\overline{s}_\tau^\star(\sqrt{1-\tau}y)\big|_{y=x_\tau/\sqrt{1-\tau}}. \tag{56}$$

Direct calculation reveals that the Jacobian matrix of $\overline{s}_\tau^\star(x)$ is (Li et al., 2024b)

$$J_\tau(x) \coloneqq \nabla_x\overline{s}_\tau^\star(x) = -\frac{1}{\tau}I_d + \frac{1}{\tau}\Sigma_\tau(x), \tag{57}$$

where $\Sigma_\tau(x)$ is the conditional covariance matrix of the standard Gaussian noise $Z\sim N(0,I_d)$ conditioned on $\sqrt{1-\tau}X_0+\sqrt{\tau}Z=x$ (cf. (27)). Then the two terms on the right-hand side of (56) can be rewritten as follows:

$$\begin{aligned}\frac{\partial}{\partial y}\sqrt{\tau}\overline{s}_\tau^\star(\sqrt{1-\tau}y)\big|_{y=x_\tau/\sqrt{1-\tau}}\frac{\partial x_\tau/\sqrt{1-\tau}}{\partial\tau} &\overset{(17)\text{ and }(57)}{=} -\frac{\sqrt{\tau}}{2(1-\tau)}J_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau) \\ &= \frac{1}{2\sqrt{\tau}(1-\tau)}\overline{s}_\tau^\star(x_\tau) - \frac{1}{2\sqrt{\tau}(1-\tau)}\Sigma_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau), \end{aligned} \tag{58a}$$

$$\frac{\partial}{\partial\tau}\sqrt{\tau}\overline{s}_\tau^\star(\sqrt{1-\tau}y) \overset{(7)}{=} -\frac{\partial}{\partial\tau}\sqrt{\frac{1-\tau}{\tau}}\int(y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0$$

$$= -\frac{\overline{s}_\tau^\star(\sqrt{1-\tau}y)}{2\sqrt{\tau}(1-\tau)} - \sqrt{\frac{1-\tau}{\tau}}\frac{\partial}{\partial\tau}\int (y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0. \tag{58b}$$

Combining (56), (58a) and (58b), we obtain

$$\frac{\partial\sqrt{\tau}\overline{s}_\tau^\star(x_\tau)}{\partial\tau} = -\frac{1}{2\sqrt{\tau}(1-\tau)}\Sigma_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau) - \sqrt{\frac{1-\tau}{\tau}}\frac{\partial}{\partial\tau}\int (y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0.$$

which, together with the inequality $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$, further leads to

$$\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\left\|\frac{\partial\sqrt{\tau}\overline{s}_\tau^\star(x_\tau)}{\partial\tau}\right\|_2^2\right] \le \frac{1}{2\tau(1-\tau)^2}\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau)\|_2^2\right] + \frac{2(1-\tau)}{\tau}\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\frac{\partial}{\partial\tau}\int (y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0\right\|_2^2\right]. \tag{59}$$

In the following, we proceed to control these two terms separately.

- **Bounding $\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau)\|_2^2\right]$.** We define the following set:

$$\mathcal{S}_\tau := \{x : -\log p_{\overline{X}_\tau}(x) \le \theta d\log T\}, \tag{60}$$

where θ is a constant satisfying $\theta \ge c_R + 17$. Then we can decompose $\mathbb{E}_{x_\tau\sim\overline{X}_\tau}[\|\Sigma_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau)\|_2^2]$ into two terms:

$$\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau)\|_2^2\right] = \int_{\mathcal{S}_\tau}\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^2 p_{\overline{X}_\tau}(x)\mathrm{d}x + \int_{\mathcal{S}_\tau^{\mathrm{c}}}\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^2 p_{\overline{X}_\tau}(x)\mathrm{d}x. \tag{61}$$

The following lemma allows us to bound $\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2$ for $x\in\mathcal{S}_\tau$ and the expectation $\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^4\right]$. The proof is deferred to Section E.

Lemma C.3. *The following inequalities hold:*

$$\|\overline{s}_\tau^\star(x)\|_2^2 \le \frac{25(\theta+c_0)d\log T}{\tau} \quad\text{and}\quad \|\Sigma_\tau(x)\|_{\mathsf{op}} \le 12(\theta+c_0)d\log T, \qquad \forall x\in\mathcal{S}_\tau, \tag{62a}$$

$$\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^4\right] \le \frac{d^6}{\tau^2}. \tag{62b}$$

Here, c_0 is defined in (10)*.*

By virtue of (62a), the integral over the set $\mathcal{S}_\tau$ can be bounded by

$$\int_{\mathcal{S}_\tau}\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^2 p_{\overline{X}_\tau}(x)\mathrm{d}x \lesssim \frac{(\theta+c_0)d\log T}{\tau}\int\|\Sigma_\tau(x)\|_{\mathsf{op}}^2 p_{\overline{X}_\tau}(x)\mathrm{d}x. \tag{63}$$

For the integral over $\mathcal{S}_\tau^{\mathrm{c}}$, one can apply Cauchy-Schwarz inequality to obtain

$$\int_{\mathcal{S}_\tau^{\mathrm{c}}}\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^2 p_{\overline{x}_\tau}(x)\mathrm{d}x \le \sqrt{\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^4\right]\mathbb{P}\left\{\overline{X}_\tau\in\mathcal{S}_\tau^{\mathrm{c}}\right\}}. \tag{64}$$

To control the right-hand side of (64), we prove that $\mathcal{S}_\tau$ is a high-probability set, meaning that the probability $\mathbb{P}\left\{\overline{X}_\tau\in\mathcal{S}_\tau^{\mathrm{c}}\right\}$ is small. To this end, we define $\mathcal{B} = \{x : \|x\|_2 \le \sqrt{1-\tau}T^{c_R+8} + 10\sqrt{\tau d\log T}\}$. Then we have

$$\mathbb{P}\left\{\overline{X}_\tau\in\mathcal{S}_\tau^{\mathrm{c}}\right\} \le \int_{\mathcal{S}_\tau^{\mathrm{c}}\cap\mathcal{B}} p_{\overline{X}_\tau}(x)\mathrm{d}x + \int_{\mathcal{B}^{\mathrm{c}}} p_{\overline{X}_\tau}(x)\mathrm{d}x$$

$$
\begin{aligned}
&\leq \Big(2\sqrt{1-\tau}T^{c_R+8} + 20\sqrt{\tau d\log T}\Big)^d \exp(-\theta d\log T) \\
&\quad + \mathbb{P}\left\{\|X_0\|_2 \geq T^{c_R+8}\right\} + \mathbb{P}\left\{\|\sqrt{\tau}W\|_2 \geq 10\sqrt{\tau d\log T}\right\} \\
&\leq \exp(-(\theta - c_R - 9)d\log T) + \frac{\mathbb{E}[\|X_0\|_2]}{T^{c_R+8}} + 2\exp\left(-8\log T\right) \lesssim \frac{1}{T^8},
\end{aligned} \tag{65}
$$

provided that $T \geq 20\sqrt{d\log T}$ and $\theta \geq c_R + 17$. In addition, Eqn. (64) together with (62b) and (65) yields

$$
\int_{\mathcal{S}_\tau^{\mathrm{c}}} \|\Sigma_\tau(x)\overline{s}_\tau^\star(x)\|_2^2\, p_{\overline{X}_\tau}(x)\mathrm{d}x \lesssim \frac{d^3}{\tau T^4} \leq \frac{d}{\tau}, \tag{66}
$$

as long as $T \geq \sqrt{d}$. Putting (61), (63) and (66) together, we obtain the desired bound

$$
\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x_\tau)\overline{s}_\tau^\star(x_\tau)\|_2^2\right] \lesssim \frac{(\theta+c_0)d\log T}{\tau}\int \|\Sigma_\tau(x)\|_{\mathsf{op}}^2\, p_{\overline{X}_\tau}(x)\mathrm{d}x + \frac{d}{\tau}. \tag{67}
$$

- **Controlling $\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\frac{\partial}{\partial\tau}\int(y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0\right\|_2^2\right]$.** We define

$$
t \coloneqq \frac{\tau}{1-\tau} \qquad \text{and} \qquad \widetilde{X}_t = X_0 + \sqrt{t}Z \qquad \text{where} \qquad Z\sim\mathcal{N}(0, I_d). \tag{68}
$$

Then it is straightforward to verify that

$$
\begin{aligned}
\frac{\partial}{\partial t}\Big((2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)\Big) &= \frac{\partial}{\partial t}\exp\Big(-\frac{\|y-x_0\|_2^2}{2t}\Big) \\
&= \frac{\|y-x_0\|_2^2}{2t^2}\exp\Big(-\frac{\|y-x_0\|_2^2}{2t}\Big) = \frac{\|y-x_0\|_2^2}{2t^2}(2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0),
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial}{\partial t}\int(2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0 &= \int\left(\frac{\partial}{\partial t}\Big((2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)\Big)\right)p_{X_0}(x_0)\mathrm{d}x_0 \\
&= \int\frac{\|y-x_0\|_2^2}{2t^2}(2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0.
\end{aligned}
$$

Therefore, one can write

$$
\begin{aligned}
&\frac{\partial}{\partial\tau}\int_{x_0}(y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0 \\
&\quad= \frac{\partial t}{\partial\tau}\cdot\frac{\partial}{\partial t}\int(y-x_0)p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0 \\
&\quad= \frac{1}{(1-\tau)^2}\frac{\partial}{\partial t}\frac{\int(2\pi t)^{d/2}(y-x_0)p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0}{\int(2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0} \\
&\quad= \frac{1}{(1-\tau)^2}\Bigg(\frac{\frac{\partial}{\partial t}\int(2\pi t)^{d/2}(y-x_0)p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0}{\int(2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0} \\
&\qquad\qquad - \frac{\int(2\pi t)^{d/2}(y-x_0)p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0\frac{\partial}{\partial t}\int(2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0}{\Big(\int(2\pi t)^{d/2}p_{\widetilde{X}_t|X_0}(y|x_0)p_{X_0}(x_0)\mathrm{d}x_0\Big)^2}\Bigg) \\
&\quad= \frac{1}{(1-\tau)^2}\Big(\int\frac{\|y-x_0\|_2^2}{2t^2}(y-x_0)p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0 \\
&\qquad\qquad - \int(y-x_0)p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0\int\frac{\|y-x_0\|_2^2}{2t^2}p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0\Big).
\end{aligned} \tag{69}
$$

Recognizing that $p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0 = p_{Z|X_0+\sqrt{t}Z}(\frac{y-x_0}{\sqrt{t}}|y)\mathrm{d}(\frac{y-x_0}{\sqrt{t}})$, we have

$$\begin{aligned}\int \frac{\|y-x_0\|_2^2}{2t^2}(y-x_0)p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0 &\overset{z=\frac{y-x_0}{\sqrt{t}}}{=} \int \frac{\|z\|_2^2}{2t}\sqrt{t}zp_{Z|X_0+\sqrt{t}Z}(z|y)\mathrm{d}z \\ &= \frac{1}{2\sqrt{t}}\mathbb{E}\big[\|Z\|_2^2 Z|X_0+\sqrt{t}Z=y\big]. \end{aligned} \tag{70}$$

Similarly, one has

$$\begin{aligned}&\int (y-x_0)p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0 \int \frac{\|y-x_0\|_2^2}{2t^2}p_{X_0|\widetilde{X}_t}(x_0|y)\mathrm{d}x_0 \\ &= \sqrt{t}\mathbb{E}\big[Z|X_0+\sqrt{t}Z=y\big]\cdot\frac{1}{2t}\mathbb{E}\big[\|Z\|_2^2|X_0+\sqrt{t}Z=y\big] \\ &= \frac{1}{2\sqrt{t}}\mathbb{E}\big[Z|X_0+\sqrt{t}Z=y\big]\mathbb{E}\big[\|Z\|_2^2|X_0+\sqrt{t}Z=y\big]. \end{aligned} \tag{71}$$

For notational convenience, we write

$$\mathbb{E}_{Z|y}[\cdot] := \mathbb{E}[\cdot|X_0+\sqrt{t}Z=y],$$

i.e., the expectation conditioned on $X_0+\sqrt{t}Z=y$. Combining (69) - (71) and (68) yields

$$\begin{aligned}&\frac{\partial}{\partial\tau}\int_{x_0}(y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0 \\ &= \frac{1}{2\tau^{1/2}(1-\tau)^{3/2}}\mathbb{E}_{Z|y}\left[\|Z\|_2^2 Z - \mathbb{E}_{Z|y}[Z]\mathbb{E}_{Z|y}[\|Z\|_2^2]\right] \\ &= \frac{1}{2\tau^{1/2}(1-\tau)^{3/2}}\mathbb{E}_{Z|y}\left[\big(Z-\mathbb{E}_{Z|y}[Z]\big)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]. \end{aligned}$$

Consequently, one has

$$\begin{aligned}&\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\frac{\partial}{\partial\tau}\int(y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0\right\|_2^2\right] \\ &= \frac{1}{\tau(1-\tau)^3}\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\big(Z-\mathbb{E}_{Z|y}[Z]\big)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\right]. \end{aligned} \tag{72}$$

We claim for the moment that

$$\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\big(Z-\mathbb{E}_{Z|y}[Z]\big)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\right] \lesssim d\min\left\{d\log T, L\right\}; \tag{73}$$

we defer the proof to end of the section. Combining (72) and (73), we have

$$\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\frac{\partial}{\partial\tau}\int(y-x_0)p_{X_0|\overline{X}_\tau}(x_0|\sqrt{1-\tau}y)\mathrm{d}x_0\right\|_2^2\right] \le \frac{d\min\left\{d\log T, L\right\}}{\tau(1-\tau)^3}. \tag{74}$$

Putting (59), (67) and (74) together, we arrive at

$$\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\left[\left\|\frac{\partial\sqrt{\tau}\overline{s}_\tau^\star(x_\tau)}{\partial\tau}\right\|_2^2\right] \lesssim \frac{(\theta+c_0)d}{\tau^2(1-\tau)^2}\left(\log T\int\|\Sigma_\tau(x)\|_2^2 p_{\overline{X}_\tau}(x)\mathrm{d}x+1\right)+\frac{d\min\left\{d\log T, L\right\}}{\tau^2(1-\tau)^2}. \tag{75}$$

Proof of Claim (73)**.** It remains to validate (73). By virtue of Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left\|\mathbb{E}_{Z|y}\left[\big(Z-\mathbb{E}_{Z|y}[Z]\big)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2 \\ &= \sup_{u\in\mathbb{S}^{d-1}}\left(u^\top\mathbb{E}_{Z|y}\left[\big(Z-\mathbb{E}_{Z|y}[Z]\big)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right)^2 \end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}_{Z|y}\left[\left(u^\top\left(Z-\mathbb{E}_{Z|y}[Z]\right)\right)^2\right]\mathbb{E}_{Z|y}\left[\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right]\\
&= \sup_{u\in\mathbb{S}^{d-1}} u^\top\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(Z-\mathbb{E}_{Z|y}[Z]\right)^\top\right]u\cdot\left(\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right)\\
&= \sup_{u\in\mathbb{S}^{d-1}} u^\top\mathsf{Cov}\left(Z|X_0+\sqrt{t}Z=y\right)u\cdot\left(\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right)\\
&\overset{(27)}{=} \sup_{u\in\mathbb{S}^{d-1}} u^\top\left(\Sigma_\tau(\sqrt{1-\tau}y)\right)u\cdot\left(\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right)\\
&\leq \left\|\Sigma_\tau(\sqrt{1-\tau}y)\right\|_2\left(\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right),
\end{aligned}\tag{76}
$$

where $\mathbb{S}^{d-1}=\{x|x\in\mathbb{R}^d,\|x\|_2=1\}$. We consider two scenarios: $L<d\log T$ and $L\geq d\log T$.

Case 1: $L<d\log T$**.** In view of (76), we have

$$
\begin{aligned}
&\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\mathbb{1}(\|\Sigma_\tau(\sqrt{1-\tau}y)\|_2\leq L+1)\right]\\
&\quad\leq \mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x_\tau)\|_2\left(\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right)\mathbb{1}(\|\Sigma_\tau(\sqrt{1-\tau}y)\|_2\leq L+1)\right]\\
&\quad\leq (L+1)\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right]\\
&\quad= (L+1)\left(\mathbb{E}[\|Z\|_2^4]-\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right]\right)\\
&\quad\overset{\text{Jensen's inequality}}{\leq} (L+1)\left(\mathbb{E}[\|Z\|_2^4]-\left(\mathbb{E}[\|Z\|_2^2]\right)^2\right)\\
&\quad= (L+1)(d^2+2d-d^2)=2d(L+1).
\end{aligned}\tag{77}
$$

Furthermore, using Cauchy-Schwarz inequality, one can show that

$$
\begin{aligned}
&\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\mathbb{1}(\|\Sigma_\tau(\sqrt{1-\tau}y)\|_2>L)\right]\\
&\quad\leq\left(\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^4\right]\right)^{1/2}\sqrt{\mathbb{P}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}(\|\Sigma_\tau(\sqrt{1-\tau}y)\|_2>L)}\\
&\quad\leq\left(\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\|Z-\mathbb{E}_{Z|y}[Z]\|_2^8\right]\cdot\mathbb{E}_{Z|y}\left[\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^8\right]\right]\right)^{1/4}\sqrt{\mathbb{P}_{x_\tau\sim\overline{X}_\tau}(\|\Sigma_\tau(x_\tau)\|_2>L)}\\
&\quad\leq\left(\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\|Z-\mathbb{E}_{Z|y}[Z]\|_2^{16}\right]\right]\cdot\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^{16}\right]\right]\right)^{1/8}\\
&\qquad\cdot\sqrt{\mathbb{P}_{x_\tau\sim\overline{X}_\tau}(\|\Sigma_\tau(x_\tau)\|_2>L+1)}.
\end{aligned}\tag{78}
$$

Applying Jensen's inequality yields

$$
\begin{aligned}
&\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\|Z-\mathbb{E}_{Z|y}[Z]\|_2^{16}\right]\right]\\
&\quad\lesssim\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\|Z\|_2^{16}\right]+\mathbb{E}_{Z|y}\left[\|\mathbb{E}_{Z|y}[Z]\|_2^{16}\right]\right]\\
&\quad\leq\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\|Z\|_2^{16}\right]+\mathbb{E}_{Z|y}\left[\mathbb{E}_{Z|y}\left[\|Z\|_2^{16}\right]\right]\right]\\
&\quad=2\mathbb{E}\left[\|Z\|_2^{16}\right]\lesssim d^8.
\end{aligned}\tag{79}
$$

Similarly, one has

$$
\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^{16}\right]\right]\lesssim\mathbb{E}\left[\|Z\|_2^{32}\right]\leq d^{16}.\tag{80}
$$

By virtue of Eqn. (25b) in Li et al. (2024b), we have

$$
\Sigma_\tau(x_\tau)=\tau\nabla\overline{s}_\tau^\star(x_\tau)+I_d,
$$

which, together with Definition 3.2, implies

$$
\mathbb{P}_{x_\tau\sim\overline{X}_\tau}(\|\Sigma_\tau(x_\tau)\|_{\mathsf{op}}>L+1)\leq\mathbb{P}_{x_\tau\sim\overline{X}_\tau}\left(\|\tau\nabla\overline{s}_\tau^\star(x_\tau)\|_{\mathsf{op}}>L\right)\leq d^{-4}.\tag{81}
$$

Combining (78) - (81), one has

$$\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\mathbb{1}(\|\Sigma_\tau(\sqrt{1-\tau}y)\|_{\mathsf{op}}>L)\right]\lesssim d. \tag{82}$$

Eqn. (77) and (82) together show that

$$\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\right]\lesssim dL. \tag{83}$$

Case 2: $L\geq d\log T$**.** We set $\theta=c_R+17$, where c_R is defined in (12). Recalling the definition of $\mathcal{S}_\tau$ in (60) and applying Lemma C.3 and (76), one has

$$\begin{aligned}
&\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\mathbb{1}(\sqrt{1-\tau}y\in\mathcal{S}_\tau)\right]\\
&\leq\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\Sigma_\tau(\sqrt{1-\tau}y)\right\|_{\mathsf{op}}\left(\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right)\mathbb{1}(\sqrt{1-\tau}y\in\mathcal{S}_\tau)\right]\\
&\leq 12(\theta+c_0)d\log T\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left(\mathbb{E}_{Z|y}\|Z\|_2^4-\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right)\right]\\
&=12(\theta+c_0)d\log T\left(\mathbb{E}\left[\|Z\|_2^4\right]-\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left(\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^2\right]\right)\\
&\overset{\text{Jensen's inequality}}{\leq}12(\theta+c_0)d\log T\left(\mathbb{E}[\|Z\|_2^4]-\left(\mathbb{E}[\|Z\|_2^2]\right)^2\right)\\
&\leq 12(\theta+c_0)d\log T(d^2+2d-d^2)\\
&\lesssim d^2\log T.
\end{aligned}\tag{84}$$

Repeating similar arguments as in (82) yields

$$\begin{aligned}
&\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\mathbb{1}(\sqrt{1-\tau}y\in\mathcal{S}_\tau^{\mathrm{c}})\right]\\
&\leq\left(\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\|Z-\mathbb{E}_{Z|y}[Z]\|_2^{16}\right]\right]\cdot\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\mathbb{E}_{Z|y}\left[\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)^{16}\right]\right]\right)^{1/8}\\
&\qquad\cdot\sqrt{\mathbb{P}_{x_\tau\sim\overline{X}_\tau}\left(x_\tau\in\mathcal{S}_\tau^{\mathrm{c}}\right)}\\
&\lesssim d^3\sqrt{\mathbb{P}_{x_\tau\sim\overline{X}_\tau}\left(x_\tau\in\mathcal{S}_\tau^{\mathrm{c}}\right)}.
\end{aligned}$$

By virtue of (65), we have

$$\begin{aligned}
&\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\mathbb{1}(\sqrt{1-\tau}y\in\mathcal{S}_\tau^{\mathrm{c}})\right]\\
&\lesssim d^3\cdot T^{-4}\leq d,
\end{aligned}\tag{85}$$

provided that $T\geq\sqrt{d}$. Putting (84) and (85) together leads to

$$\mathbb{E}_{\sqrt{1-\tau}y\sim\overline{X}_\tau}\left[\left\|\mathbb{E}_{Z|y}\left[\left(Z-\mathbb{E}_{Z|y}[Z]\right)\left(\|Z\|_2^2-\mathbb{E}_{Z|y}[\|Z\|_2^2]\right)\right]\right\|_2^2\mathbb{1}(\sqrt{1-\tau}y\in\mathcal{S}_\tau)\right]\lesssim d^2\log T. \tag{86}$$

Combining (83) and (86) finishes the proof of Claim (73).

D Proof of Lemma 4.3

The first inequality comes from Lemma 3 in Li et al. (2024b). It suffices to prove the second one. To this end, we consider two scenarios: $d>L^2$ and $d\leq L^2$.

- *Case 1:* $d>L^2$. Let $\mathcal{L}_\tau$ denote the set of x such that the operator norm of the covariance matrix $\Sigma_\tau(x)$ is at most L, i.e.,

  $$\mathcal{L}_\tau\coloneqq\{x:\|\Sigma_\tau(x)\|_{\mathsf{op}}\leq L+1\}.$$

  By virtue of (81), we know that $\mathbb{P}(\overline{X}_\tau\in\mathcal{L}_\tau^{\mathrm{c}})\leq d^{-4}$, and consequently

  $$\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_{\mathsf{op}}^2\leq(L+1)^2+\int_{\mathcal{L}_\tau^{\mathrm{c}}}\|\Sigma_\tau(x)\|_{\mathsf{op}}^2\,p_{\overline{X}_\tau}(x)\mathrm{d}x$$

$$\overset{\text{(i)}}{\leq} (L+1)^2 + \sqrt{\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\|_{\mathsf{op}}^4\right]}\sqrt{\mathbb{P}(\overline{X}_\tau \in \mathcal{L}_\tau^{\mathrm{c}})}$$
$$\overset{\text{(ii)}}{\lesssim} L^2 + \sqrt{\frac{d^4}{d^4}} \lesssim L^2,$$

where (i) holds due to the Cauchy-Schwarz inequality, and (ii) makes use of the fact that

$$\begin{aligned}\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\|_{\mathsf{op}}^4\right] &\leq \mathbb{E}_{x\sim\overline{X}_\tau}\left[\mathbb{E}\left[\|Z\|_2^8 | \sqrt{1-\tau}X_0 + \sqrt{\tau}Z = x\right]\right] \\ &= \mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\left[\|Z\|_2^8\right] \lesssim d^4.\end{aligned}$$

Thus we have

$$\begin{aligned}&\sum_{t=2}^{T}\frac{\overline{\alpha}_t}{(1-\overline{\alpha}_t)}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_{\mathsf{op}}^2}{(1-\tau)^2}\mathrm{d}\tau \\ &\lesssim \sum_{t=2}^{T}\frac{L^2(\tau_t-\widetilde{\tau}_{t-1})}{\overline{\alpha}_t(1-\overline{\alpha}_t)} \\ &\overset{\text{(42a)}}{\lesssim} \sum_{t=1}^{T}\frac{L^2\overline{\alpha}_{t-1}\log T}{T\overline{\alpha}_t} \\ &\overset{\text{(10)}}{\lesssim} L^2\log T. \end{aligned} \tag{87}$$

- *Case 2: $d \leq L^2$.* Now we move on to the second case. By virtue of Claim (90) in Li et al. (2024b), one has, for any $0 < \tau_1 < \tau_2 < 1$,

$$\int_{\tau_1}^{\tau_2}\frac{\mathbb{E}[\mathsf{Tr}(\Sigma_\tau^2(x_\tau))]}{(1-\tau)^2}\mathrm{d}\tau = \frac{\tau_1}{1-\tau_1}\mathbb{E}[\mathsf{Tr}(\Sigma_{\tau_1}(x_{\tau_1}))] - \frac{\tau_2}{1-\tau_2}\mathbb{E}[\mathsf{Tr}(\Sigma_{\tau_2}(x_{\tau_2}))]. \tag{88}$$

We make the observation that

$$\begin{aligned}\tau_{t-1}-\widetilde{\tau}_{t-1} = \widetilde{\alpha}_{t-1}-\overline{\alpha}_{t-1} &= \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})\overline{\alpha}_{t-1}}{\overline{\alpha}_{t-1}(1-\alpha_t)+\alpha_t(1-\overline{\alpha}_{t-1})} \\ &\leq \frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{\alpha_t} \leq 2(\tau_t-\tau_{t-1}).\end{aligned}$$

In view of Lemma 2 Part (a) in Li et al. (2024b) and the fact that

$$|\tau_t-\widetilde{\tau}_{t-1}|/\tau_t(1-\tau_t) \lesssim \log T/T \lesssim 1/(d\log T),$$

we have

$$\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\mathsf{Tr}\left(\Sigma_\tau^2(x_\tau)\right) \lesssim \mathbb{E}_{x_\tau\sim X_{\tau_t}}\mathsf{Tr}\left(\Sigma_{\tau_t}^2(x_\tau)\right) + \frac{1}{T^{c_2 d}}.$$

Combining (88) and the previous inequality yields

$$\begin{aligned}&\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\mathsf{Tr}\left(\Sigma_\tau^2(x_\tau)\right)}{(1-\tau)^2}\mathrm{d}\tau \\ &\lesssim \frac{\tau_t-\widetilde{\tau}_{t-1}}{(1-\tau_t)^2}\left(\mathbb{E}_{x_\tau\sim\overline{X}_{\tau_t}}\left[\mathsf{Tr}\left(\Sigma_{\tau_t}^2(x_\tau)\right)\right] + T^{-c_2 d}\right),\end{aligned}$$

which further implies

$$\begin{aligned}&\frac{\overline{\alpha}_{t-1}}{1-\overline{\alpha}_{t-1}}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_{\mathsf{op}}^2}{(1-\tau)^2}\mathrm{d}\tau \\ &\leq \frac{\overline{\alpha}_{t-1}}{1-\overline{\alpha}_{t-1}}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_{\mathrm{F}}^2}{(1-\tau)^2}\mathrm{d}\tau \\ &= \frac{\overline{\alpha}_{t-1}}{1-\overline{\alpha}_{t-1}}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\mathsf{Tr}\left(\Sigma_\tau^2(x_\tau)\right)}{(1-\tau)^2}\mathrm{d}\tau\end{aligned}$$

$$\lesssim \frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{\overline{\alpha}_{t-1}(1-\overline{\alpha}_{t-1})}\left(\mathbb{E}_{x_\tau\sim\overline{X}_{\tau_t}}\left[\mathsf{Tr}\left(\Sigma^2_{\tau_t}(x_\tau)\right)\right]+T^{-c_2 d}\right)$$
$$\asymp \frac{1-\alpha_t}{1-\overline{\alpha}_t}\left(\mathbb{E}_{x_\tau\sim\overline{X}_{\tau_t}}\left[\mathsf{Tr}\left(\Sigma^2_{\tau_t}(x_\tau)\right)\right]+T^{-c_2 d}\right),$$

where the second line arises from $\mathsf{Tr}\left(\Sigma^2_\tau(x_\tau)\right)=\|\Sigma_\tau(x_\tau)\|_{\mathrm{F}}^2$. Summing over $t=2,\ldots,T$, we obtain

$$\sum_{t=2}^{T}\frac{\overline{\alpha}_{t-1}}{1-\overline{\alpha}_{t-1}}\int_{\widetilde{\tau}_{t-1}}^{\tau_t}\frac{\mathbb{E}_{x_\tau\sim\overline{X}_\tau}\|\Sigma_\tau(x_\tau)\|_{\mathsf{op}}^2}{(1-\tau)^2}\mathrm{d}\tau$$
$$\lesssim \sum_{t=2}^{T}\frac{1-\alpha_t}{1-\overline{\alpha}_t}\left(\mathbb{E}_{x_\tau\sim\overline{X}_{\tau_t}}\left[\mathsf{Tr}\left(\Sigma^2_{\tau_t}(x_\tau)\right)\right]+T^{-c_2 d}\right)$$
$$\lesssim d\log T+T^{-c_2 d}\log T, \tag{89}$$

where the second line holds due to Lemma 2 in Li et al. (2024b) and (10).

Putting (87) and (89) for the above two cases together, we know that Lemma 4.3 always holds.

E Proof of Lemma C.3

The inequality (62a) comes from Lemma 1 in Li and Yan (2024b). The inequality (62b) makes use of the fact that

$$\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\overline{s}^\star_\tau(x)\|_2^4\right]\overset{\text{Cauchy-Schwarz}}{\leq}\sqrt{\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\|_{\mathsf{op}}^8\right]\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\overline{s}^\star_\tau(x)\|_2^8\right]}\leq\frac{d^6}{\tau^2},$$

where the last inequality holds due to the facts

$$\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\Sigma_\tau(x)\|_{\mathsf{op}}^8\right]=\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\mathsf{Cov}(Z|\overline{X}_\tau=x)\|_{\mathsf{op}}^8\right]\leq\mathbb{E}_{x\sim\overline{X}_\tau}\left[\left(\mathbb{E}[\|Z\|_2^2|\overline{X}_\tau=x]\right)^8\right]$$
$$\overset{\text{Jensen's inequality}}{\leq}\mathbb{E}_{x\sim\overline{X}_\tau}\left[\mathbb{E}[\|Z\|_2^{16}|\overline{X}_\tau=x]\right]=\mathbb{E}[\|Z\|_2^{16}]\asymp d^8, \tag{90}$$

$$\mathbb{E}_{x\sim\overline{X}_\tau}\left[\|\overline{s}^\star_\tau(x)\|_2^8\right]=\mathbb{E}_{x\sim\overline{X}_\tau}\left[\left\|\mathbb{E}\left[\frac{Z}{\sqrt{\tau}}|\overline{X}_\tau=x\right]\right\|_2^8\right]\overset{\text{Jensen's inequality}}{\leq}\frac{1}{\tau^4}\mathbb{E}_{x\sim\overline{X}_\tau}\left[\mathbb{E}\left[\|Z\|_2^8|\overline{X}_\tau=x\right]\right]$$
$$=\frac{1}{\tau^4}\mathbb{E}[\|Z\|_2^8]\asymp\frac{d^4}{\tau^4}. \tag{91}$$

F Proof of Theorem 3.6 (lower bound)

We know from (2) that $X_t\sim N(0,\lambda_t I_d)$, where

$$\lambda_t=\alpha_t\lambda_{t-1}+1-\alpha_t,\qquad\forall 1\leq t\leq T,\qquad\lambda_0=\lambda. \tag{92}$$

Solving (92) gives us

$$\lambda_t=(1-\overline{\alpha}_t)+\overline{\alpha}_t\lambda,\qquad\forall 1\leq t\leq T. \tag{93}$$

By virtue of (8) and the fact $s_t^\star(x)=\nabla\log p_{X_t}(x)=-\frac{x}{\lambda_t}$, it is straightforward to check that

$$Y_t\sim N(0,\widehat{\lambda}_t I_d), \tag{94}$$

where the variances $\{\widehat{\lambda}_t\}_{t=1}^T$ satisfy $\widehat{\lambda}_T=1$ and

$$\widehat{\lambda}_{t-1}=\frac{\widehat{\lambda}_t}{\alpha_t}\left(1-\frac{\eta_t}{\lambda_t}\right)^2+\frac{\sigma_t^2}{\alpha_t},\qquad\forall 2\leq t\leq T.$$

Unrolling the recursion gives

$$\widehat{\lambda}_1=\frac{\overline{\alpha}_T\lambda_1^2}{\alpha_1\lambda_T^2}\widehat{\lambda}_T+\sum_{t=2}^{T}\prod_{i=2}^{t-1}\frac{1}{\alpha_i}\left(1-\frac{\eta_i}{\lambda_i}\right)^2\frac{\sigma_t^2}{\alpha_t}. \tag{95}$$

Then we can write the close form of the KL divergence between X_1 and Y_1 admits the following closed-form expression:

$$\mathsf{KL}(p_{X_1}||p_{Y_1}) = \frac{d}{2}\left(\log\frac{\widehat{\lambda}_1}{\lambda_1} + \frac{\lambda_1}{\widehat{\lambda}_1} - 1\right).$$

For $\lambda_1/\widehat{\lambda}_1 \in (0,2)$, using the inequality $-\log x \geq (1-x) + \frac{1}{4}(x-1)^2$ for $0 < x < 2$, we have

$$\mathsf{KL}(p_{X_1}||p_{Y_1}) \geq \frac{d}{8}\left(1 - \frac{\lambda_1}{\widehat{\lambda}_1}\right)^2.$$

For $\lambda_1/\widehat{\lambda}_1 \geq 2$, the inequality $\log x \leq (x-1)\log 2$ for $x \geq 2$ yields

$$\begin{aligned}\mathsf{KL}(p_{X_1}||p_{Y_1}) &\geq \frac{d}{2}\left(\left(1 - \frac{\lambda_1}{\widehat{\lambda}_1}\right)\log 2 + \frac{\lambda_1}{\widehat{\lambda}_1} - 1\right) \\ &= \frac{d(1-\log 2)}{2}\left(\frac{\lambda_1}{\widehat{\lambda}_1} - 1\right). \end{aligned} \tag{96}$$

Therefore, it suffices to prove that $\left|\frac{\lambda_1}{\widehat{\lambda}_1} - 1\right| \geq \frac{C}{T}$ for some small constant C, which can be guaranteed by

$$\left|\frac{\widehat{\lambda}_1}{\lambda_1} - 1\right| \geq \frac{2C}{T}$$

as long as $T \gg C$.

We first focus on the case $\eta_t = 1 - \alpha_t$ and $\sigma_t^2 = 1 - \alpha_t$. Note that when $\eta_t = 1 - \alpha_t$, we have

$$1 - \frac{\eta_t}{\lambda_t} = 1 - \frac{1-\alpha_t}{\lambda_t} = \frac{\lambda_t - 1 + \alpha_t}{\lambda_t} \overset{(92)}{=} \frac{\alpha_t\lambda_{t-1} + 1 - \alpha_t - 1 + \alpha_t}{\lambda_t} = \frac{\alpha_t\lambda_{t-1}}{\lambda_t}.$$

Putting the previous equation and (95) together, one has

$$\begin{aligned}\widehat{\lambda}_1 &= \frac{\overline{\alpha}_T\lambda_1^2}{\alpha_1\lambda_T^2}\widehat{\lambda}_T + \sum_{t=2}^{T}\prod_{i=2}^{t-1}\frac{1}{\alpha_i}\left(\frac{\alpha_i\lambda_{i-1}}{\lambda_i}\right)^2\frac{\sigma_t^2}{\alpha_t} \\ &= \frac{\overline{\alpha}_T\lambda_1^2}{\alpha_1\lambda_T^2}\widehat{\lambda}_T + \sum_{t=2}^{T}\frac{\lambda_1^2}{\alpha_1}\frac{\overline{\alpha}_{t-1}\sigma_t^2}{\alpha_t\lambda_{t-1}^2} \\ &= \frac{\overline{\alpha}_T\lambda_1^2}{\alpha_1\lambda_T^2}\widehat{\lambda}_T + \frac{\lambda_1^2}{\alpha_1}\sum_{t=1}^{T-1}\frac{\overline{\alpha}_t(1-\alpha_{t+1})}{\lambda_t^2\alpha_{t+1}} \\ &= \frac{\overline{\alpha}_T\lambda_1^2}{\alpha_1\lambda_T^2}\widehat{\lambda}_T + \frac{\lambda_1^2}{\alpha_1}\sum_{t=1}^{T-1}\frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\lambda_t^2\alpha_{t+1}}.\end{aligned}$$

In addition, we can decompose λ_1 as

$$\begin{aligned}\lambda_1 = \frac{\lambda_1^2}{\alpha_1}\cdot\frac{\overline{\alpha}_1}{\lambda_1} &= \frac{\lambda_1^2\overline{\alpha}_T}{\alpha_1\lambda_T} + \frac{\lambda_1^2}{\alpha_1}\sum_{t=1}^{T-1}\left(\frac{\overline{\alpha}_t}{\lambda_t} - \frac{\overline{\alpha}_{t+1}}{\lambda_{t+1}}\right) \\ &\overset{(92)}{=} \frac{\overline{\alpha}_T\lambda_1^2}{\alpha_1\lambda_T^2}\lambda_T + \frac{\lambda_1^2}{\alpha_1}\sum_{t=1}^{T-1}\frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\lambda_t\lambda_{t+1}}.\end{aligned}$$

The previous two equations together imply

$$\begin{aligned}\frac{\widehat{\lambda}_1}{\lambda_1} - 1 &= \frac{\widehat{\lambda}_1 - \lambda_1}{\lambda_1} \\ &= \frac{\overline{\alpha}_T\lambda_1}{\alpha_1\lambda_T^2}\widehat{\lambda}_T + \frac{\lambda_1}{\alpha_1}\sum_{t=1}^{T-1}\frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\lambda_t^2\alpha_{t+1}}\end{aligned}$$

$$
\begin{aligned}
&\quad - \left(\frac{\overline{\alpha}_T \lambda_1}{\alpha_1 \lambda_T^2} \lambda_T + \frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\lambda_t \lambda_{t+1}} \right) \\
&= \frac{\overline{\alpha}_T \lambda_1}{\alpha_1 \lambda_T^2} (\widehat{\lambda}_T - \lambda_T) \\
&\quad + \frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\lambda_t} \left(\frac{1}{\lambda_t \alpha_{t+1}} - \frac{1}{\lambda_{t+1}} \right) \\
&\overset{(92)}{=} \frac{\overline{\alpha}_T \lambda_1}{\alpha_1 \lambda_T^2} (\widehat{\lambda}_T - \lambda_T) \\
&\quad + \frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{\overline{\alpha}_t - \overline{\alpha}_{t+1}}{\lambda_t} \frac{1 - \alpha_{t+1}}{\lambda_t \lambda_{t+1} \alpha_{t+1}} \\
&= \frac{\overline{\alpha}_T \lambda_1}{\alpha_1 \lambda_T^2} (\widehat{\lambda}_T - \lambda_T) + \frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{(\overline{\alpha}_t - \overline{\alpha}_{t+1})^2}{\lambda_t^2 \lambda_{t+1} \overline{\alpha}_{t+1}}.
\end{aligned} \tag{97}
$$

If $\lambda \asymp 1$, then we have $\lambda_t \asymp 1$ for all t. Applying Cauchy-Schwarz inequality and the fact $\overline{\alpha}_t \leq 1$ for all t yields

$$
\begin{aligned}
\frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{(\overline{\alpha}_t - \overline{\alpha}_{t+1})^2}{\lambda_t^2 \lambda_{t+1} \overline{\alpha}_{t+1}} &\asymp \frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{(\overline{\alpha}_t - \overline{\alpha}_{t+1})^2}{\overline{\alpha}_{t+1}} \\
&\geq \frac{\lambda_1}{\alpha_1} \frac{\left(\sum_{t=1}^{T-1} (\overline{\alpha}_t - \overline{\alpha}_{t+1}) \right)^2}{\sum_{t=1}^{T-1} \overline{\alpha}_{t+1}} \\
&\geq \frac{\lambda_1 (\overline{\alpha}_1 - \overline{\alpha}_T)^2}{T \alpha_1} \\
&= \frac{\lambda_1 (\alpha_1 - \overline{\alpha}_T)^2}{T \alpha_1}.
\end{aligned} \tag{98}
$$

This, together with (10), tells us that

$$
\frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{(\overline{\alpha}_t - \overline{\alpha}_{t+1})^2}{\lambda_t^2 \lambda_{t+1} \overline{\alpha}_{t+1}} \gtrsim \frac{\lambda_1 (1 - \frac{1}{T^{c_0}} - \frac{1}{T^{c_0}})^2}{T} \asymp \frac{1}{T}. \tag{99}
$$

Moreover, in view of (93) and the fact $\widehat{\lambda}_T = 1$, we have

$$
\left| \frac{\overline{\alpha}_T \lambda_1}{\alpha_1 \lambda_T^2} (\widehat{\lambda}_T - \lambda_T) \right| = \frac{\overline{\alpha}_T^2 \lambda_1}{\alpha_1 \lambda_T^2} (\lambda - 1) \asymp \frac{\overline{\alpha}_T^2}{\alpha_1} \overset{(10)}{\gtrsim} \frac{1}{T^{2c_0}}. \tag{100}
$$

Combining (97), (99) and (100), we have

$$
\begin{aligned}
\left| \frac{\widehat{\lambda}_1}{\lambda_1} - 1 \right| &\geq \frac{\lambda_1}{\alpha_1} \sum_{t=1}^{T-1} \frac{(\overline{\alpha}_t - \overline{\alpha}_{t+1})^2}{\lambda_t^2 \lambda_{t+1} \overline{\alpha}_{t+1}} \\
&\quad - \left| \frac{\overline{\alpha}_T \lambda_1}{\alpha_1 \lambda_T^2} (\widehat{\lambda}_T - \lambda_T) \right| \gtrsim \frac{1}{T}.
\end{aligned}
$$

This completes the proof.

For $\sigma_t^2 = (1 - \alpha_t)\alpha_t$ and $\frac{(\alpha_t - \overline{\alpha}_t)(1 - \alpha_t)}{1 - \overline{\alpha}_t}$, the desired bound follows by similar arguments. We omit the details here for the sake of brevity.

Discussion of other variance schedules and discretizations. Theorem 3.6 already covers a broad class of noise schedules, including $\sigma_t^2 = 1 - \alpha_t$, $\alpha_t(1 - \alpha_t)$, and $\frac{(\alpha_t - \overline{\alpha}_t)(1 - \alpha_t)}{1 - \overline{\alpha}_t}$. More generally, we can show that the iteration complexity of DDPM necessarily exhibits a $\sqrt{d}$ dependence, because at least one denoising step inevitably incurs an error with this dimensional dependence. Specifically,

consider a d-dimensional Gaussian distribution in which half of the coordinates follow $\mathcal{N}(0, \lambda_0)$ and the remaining coordinates follow $\mathcal{N}(0, \lambda_0')$, where $|\lambda_0 - \lambda_0'| = \Omega(1)$, $\lambda_0 = \Omega(1)$, and $\lambda_0' = \Omega(1)$. Under the forward diffusion process, the corresponding variances at time t are

$$\lambda_t = \overline{\alpha}_t \lambda_0 + 1 - \overline{\alpha}_t, \qquad \lambda_t' = \overline{\alpha}_t \lambda_0' + 1 - \overline{\alpha}_t.$$

Since the cumulative noise schedule satisfies $\overline{\alpha}_{t-1} - \overline{\alpha}_t \geq 1/T$ for at least one t, we focus on this step and analyze the corresponding single-step error. Assume that $Y_t \sim p_{X_t}$. Then the difference between the variance $\hat{\Sigma}_{t-1}$ of Y_{t-1} and variance Σ_{t-1} of X_{t-1} is

$$\begin{aligned}
\sum_{i=1}^{d} |\hat{\Sigma}_{t-1}(i,i) - \Sigma_{t-1}(i,i)| &= \sum_{i=1}^{d} \Big| \frac{\sigma_t^2}{\alpha_t} - \frac{(1-\alpha_t)(\Sigma_{t-1}(i,i) - 1 + \alpha_t)}{\alpha_t \Sigma_{t-1}(i,i)} \Big| \\
&\geq \frac{d}{2} \Big| \frac{(1-\alpha_t)(\lambda_t - 1 + \alpha_t)}{\alpha_t \lambda_t} - \frac{(1-\alpha_t)(\lambda_t' - 1 + \alpha_t)}{\alpha_t \lambda_t'} \Big| \\
&= \frac{d}{2} \frac{(\overline{\alpha}_{t-1} - \overline{\alpha}_t)^2 |\lambda_0 - \lambda_0'|}{\overline{\alpha}_{t-1} \lambda_t \lambda_t'} \gtrsim \frac{d}{T^2}. \qquad (101)
\end{aligned}$$

We define $\delta_i = \Sigma_{t-1}(i,i)/\hat{\Sigma}_{t-1}(i,i) - 1$, which obeys $\sum_{i=1}^{d} \delta_i^2 \gtrsim \frac{d}{T^4}$. The total variation distance between X_{t-1} and Y_{t-1} satisfies

$$\begin{aligned}
\mathsf{TV}(X_{t-1}, Y_{t-1}) &= \mathbb{E}_{x_i \sim X_{t-1,i}} \left[\left| \exp\left(\sum_{i=1}^{d} \log \frac{p_{Y_{t-1,i}}(x_i)}{p_{X_{t-1,i}}(x_i)} \right) - 1 \right| \right] \qquad (102) \\
&= \mathbb{E}_{z_i \sim \mathcal{N}(0,1)} \left[\left| \exp\left(\frac{1}{2} \sum_{i=1}^{d} \big(\log(1+\delta_i) - \delta_i z_i^2 \big) \right) - 1 \right| \right] \qquad (103) \\
&\gtrsim \sqrt{\sum_{i=1}^{d} \delta_i^2} = \Omega\Big(\frac{\sqrt{d}}{T^2} \Big). \qquad (104)
\end{aligned}$$

The above argument can be extended to multiple steps. First, suppose that there exists some $\overline{\alpha}_t \in (1/3, 2/3)$. Otherwise, we can use the above analysis to show that $\mathsf{TV}(X_{t-1}, Y_{t-1}) \gtrsim 1$ whenever $\overline{\alpha}_t \leq 1/3$ and $\overline{\alpha}_{t-1} \leq 2/3$. Next, suppose that there exists some $\overline{\alpha}_s \in (\overline{\alpha}_t + \epsilon/2, \overline{\alpha}_t + \epsilon)$ for a sufficiently small universal constant ϵ. Otherwise, we can again show that $\mathsf{TV}(X_{t-1}, Y_{t-1}) \gtrsim 1$.

Given the above time steps, we now turn to the variance coefficients σ_t^2. In particular, σ_t^2 determines the update from the t-th step to the $(t-1)$-th step and is a function of $\overline{\alpha}_{t-1}$ and $\overline{\alpha}_t$, or equivalently a function of $1 - \alpha_t$ and $\overline{\alpha}_t$ designed based on the discretization scheme. For all commonly used DDPM samplers, the variance coefficient admits the following expansion with respect to $1 - \alpha_t$:

$$\sigma_t^2 = 1 - \alpha_t + f(\overline{\alpha}_t)(1-\alpha_t)^2 + o\big((1-\alpha_t)^2\big), \qquad (105)$$

where f is a smooth function.

Assume again that $Y_t \sim p_{X_t}$. Under (105) and for the coordinates corresponding to $\mathcal{N}(0, \lambda_0)$, one can show that the variance mismatch between X_s and Y_s satisfies

$$\begin{aligned}
\hat{\Sigma}_s(j,j) - \Sigma_s(j,j) &= \sum_{i=s+1}^{t} \frac{\overline{\alpha}_t \lambda_i^2}{\overline{\alpha}_i \lambda_t^2} \frac{1}{\alpha_i} \left(\sigma_i^2 - \frac{(1-\alpha_i)(\lambda_i - 1 + \alpha_i)}{\lambda_i} \right) \\
&\gtrsim \frac{1}{T} \big(\frac{1}{\alpha_t} (\frac{1}{\lambda_t} + f(\overline{\alpha}_t)) + O(\epsilon) \big).
\end{aligned}$$

For sufficiently small ϵ, the above bound further tells us

$$\sum_{i=1}^{d} |\hat{\Sigma}_s(i,i) - \Sigma_s(i,i)| \gtrsim \frac{d}{T} \left\{ \Big| \frac{1}{\alpha_t} (\frac{1}{\lambda_t} + f(\overline{\alpha}_t)) + O(\epsilon) \Big| + \Big| \frac{1}{\alpha_t} (\frac{1}{\lambda_t'} + f(\overline{\alpha}_t)) + O(\epsilon) \Big| \right\} \asymp \frac{d}{T}.$$

Using an argument similar to (104), we obtain

$$\mathsf{TV}(X_s, Y_s) \gtrsim \frac{\sqrt{d}}{T}.$$

Therefore, this error demonstrates that a $\sqrt{d}$ dependence is unavoidable.

G Proof of examples

G.1 Proof of Example 1.

In this case, we can verify that

$$\tau\nabla\overline{s}_\tau^\star(x) = -\tau((1-\tau)\Sigma + \tau I)^{-1},$$

and thus $\|\tau\nabla\overline{s}_\tau^\star(x)\|_{\mathsf{op}} \le 1$. When the smallest singular value of Σ is zero, we have

$$\tilde{L} = \max_t \|\nabla^2 \log p_{X_t}(x)\|_{\mathsf{op}} = \frac{1}{1-\overline{\alpha}_1}.$$

G.2 Proof of Example 2

Proof of L. For ease of notation, we define the adjusted mean and variance by

$$\widetilde{\sigma}_h^2 \coloneqq (1-\tau)\sigma_h^2 + \tau, \qquad \widetilde{\mu}_h \coloneqq \sqrt{1-\tau}\mu_h.$$

Then the posterior probability of the h-th Gaussian component is given by

$$w_h(x) \coloneqq \frac{\pi_h \phi(x \mid \widetilde{\mu}_h, \widetilde{\sigma}_h^2 I_d)}{\sum_{h'=1}^H \pi_{h'}\phi(x \mid \widetilde{\mu}_{h'}, \widetilde{\sigma}_{h'}^2 I_d)},$$

where $\phi(\cdot \mid \widetilde{\mu}_h, \widetilde{\sigma}_h^2 I_d)$ denotes the density of a Gaussian distribution with mean vector $\widetilde{\mu}_h$ and covariance matrix $\widetilde{\sigma}_h^2 I_d$. For any $h \in [H]$, we define

$$z_h(x) := \frac{x - \widetilde{\mu}_h}{\widetilde{\sigma}_h}.$$

Then it is straightforward to show that

$$\overline{s}_\tau^\star(x) = \nabla \log p_{\overline{X}_\tau}(x) = \nabla \log \left(\sum_{h=1}^H \pi_h \phi\left(x \mid \widetilde{\mu}_h, \widetilde{\sigma}_h^2 I_d\right)\right) = -\sum_{h=1}^H w_h(x)\frac{x-\widetilde{\mu}_h}{\widetilde{\sigma}_h^2} = -\sum_{h=1}^H w_h(x)\frac{z_h(x)}{\widetilde{\sigma}_h}, \tag{106}$$

which further implies

$$\begin{aligned}
\nabla\overline{s}_\tau^\star(x) &= -\sum_{h=1}^H \frac{w_h(x)}{\widetilde{\sigma}_h^2} I_d + \sum_{h=1}^H w_h(x)\frac{x-\widetilde{\mu}_h}{\widetilde{\sigma}_h^2}\left(\frac{x-\widetilde{\mu}_h}{\widetilde{\sigma}_h^2}\right)^\top - \left(\sum_{h=1}^H w_h(x)\frac{x-\widetilde{\mu}_h}{\widetilde{\sigma}_h^2}\right)\left(\sum_{h=1}^H w_h(x)\frac{x-\widetilde{\mu}_h}{\widetilde{\sigma}_h^2}\right)^\top \\
&= -\sum_{h=1}^H \frac{w_h(x)}{\widetilde{\sigma}_h^2} I_d + \sum_{h=1}^H w_h(x)\left(\frac{z_h(x)}{\widetilde{\sigma}_h} - \sum_{h'=1}^H \frac{w_{h'}(x) z_{h'}(x)}{\widetilde{\sigma}_{h'}}\right)\left(\frac{z_h(x)}{\widetilde{\sigma}_h} - \sum_{h'=1}^H \frac{w_{h'}(x) z_{h'}(x)}{\widetilde{\sigma}_{h'}}\right)^\top.
\end{aligned} \tag{107}$$

We make the observation that for any vectors $u_1, \cdots, u_H \in \mathbb{R}^d$ and $w_1, \cdots, w_H \ge 0$ satisfying $\sum_{h=1}^H w_h = 1$,

$$\begin{aligned}
&\sum_{h=1}^H w_h (u_h - u)(u_h - u)^\top \\
&= \sum_{h=1}^H w_h \left(u_h - \sum_{i=1}^H w_i u_i + \sum_{i=1}^H w_i u_i - u\right)\left(u_h - \sum_{i=1}^H w_i u_i + \sum_{i=1}^H w_i u_i - u\right)^\top \\
&= \sum_{h=1}^H w_h \left(u_h - \sum_{i=1}^H w_i u_i\right)\left(u_h - \sum_{i=1}^H w_i u_i\right)^\top + \left[\sum_{h=1}^H w_h\left(u_h - \sum_{i=1}^H w_i u_i\right)\right]\left(\sum_{i=1}^H w_i u_i - u\right)^\top \\
&\quad + \left(\sum_{i=1}^H w_i u_i - u\right)\left[\sum_{h=1}^H w_h\left(u_h - \sum_{i=1}^H w_i u_i\right)^\top\right] + \sum_{h=1}^H w_h \left(\sum_{i=1}^H w_i u_i - u\right)\left(\sum_{i=1}^H w_i u_i - u\right)^\top
\end{aligned}$$

$$= \sum_{h=1}^{H} w_h \left(u_h - \sum_{i=1}^{H} w_i u_i\right)\left(u_h - \sum_{i=1}^{H} w_i u_i\right)^\top + \sum_{h=1}^{H} w_h \left(\sum_{i=1}^{H} w_i u_i - u\right)\left(\sum_{i=1}^{H} w_i u_i - u\right)^\top$$
$$\succeq \sum_{h=1}^{H} w_h \left(u_h - \sum_{i=1}^{H} w_i u_i\right)\left(u_h - \sum_{i=1}^{H} w_i u_i\right)^\top.$$

Here, $A \succeq B$ means that $A - B$ is positive semi-definite. By virtue of the previous inequality, we know that for any $k \in \{1, \ldots, H\}$,

$$\begin{aligned}
&\sum_{h=1}^{H} w_h(x) \left(\frac{z_h(x)}{\widetilde{\sigma}_h} - \sum_{h'=1}^{H} \frac{w_{h'}(x) z_{h'}(x)}{\widetilde{\sigma}_{h'}}\right)\left(\frac{z_h(x)}{\widetilde{\sigma}_h} - \sum_{h'=1}^{H} \frac{w_{h'}(x) z_{h'}(x)}{\widetilde{\sigma}_{h'}}\right)^\top \\
&\quad \preceq \sum_{h=1}^{H} w_h(x) \left(\frac{z_h(x)}{\widetilde{\sigma}_h} - \frac{z_k(x)}{\widetilde{\sigma}_k}\right)\left(\frac{z_h(x)}{\widetilde{\sigma}_h} - \frac{z_k(x)}{\widetilde{\sigma}_k}\right)^\top \\
&\quad = \widetilde{\sigma}_k^{-2} \sum_{h=1}^{H} w_h(x) \left(\frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x)\right)\left(\frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x)\right)^\top, \qquad (108)
\end{aligned}$$

where $\overline{\sigma}_{h,k} \coloneqq \widetilde{\sigma}_h / \widetilde{\sigma}_k$. Recalling that $\overline{X}_\tau$ follows the GMM $\sum_{k=1}^{H} \pi_k \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)$, we have

$$\begin{aligned}
&\mathbb{P}_{X \sim \overline{X}_\tau} \left(\|\tau \nabla \overline{s}_\tau^\star(X)\|_{\mathsf{op}} \geq C \log H \log d\right) \\
&\quad = \sum_{k=1}^{H} \pi_k \mathbb{P}_{X \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \left(\|\tau \nabla \overline{s}_\tau^\star(X)\|_{\mathsf{op}} \geq C \log H \log d\right) \\
&\quad = \sum_{k:\pi_k \geq 1/(2d^4 H)} \pi_k \mathbb{P}_{X \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \left(\|\tau \nabla \overline{s}_\tau^\star(X)\|_{\mathsf{op}} \geq C \log H \log d\right) \\
&\qquad + \sum_{k:\pi_k > 1/(2d^4 H)} \pi_k \mathbb{P}_{X \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \left(\|\tau \nabla \overline{s}_\tau^\star(X)\|_2 \geq C \log H \log d\right) \\
&\quad \leq \sum_{\pi_k \geq 1/(2d^4 H)} \pi_k \mathbb{P}_{X \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \left(\|\tau \nabla \overline{s}_\tau^\star(X)\|_{\mathsf{op}} \geq C \log H \log d\right) + \frac{1}{2d^4}.
\end{aligned}$$

Noting that $\widetilde{\sigma}_h^2 \geq \tau$ for any h, we obtain

$$\tau \sum_{h=1}^{H} \frac{w_h(x)}{\widetilde{\sigma}_h^2} I_d \preceq \sum_{h=1}^{H} w_h(x) I_d = I_d.$$

By virtue of (107), (108) and the previous two inequalities, we further have

$$\begin{aligned}
&\mathbb{P}_{X \sim \overline{X}_\tau} \left(\|\tau \nabla \overline{s}_\tau^\star(X)\|_{\mathsf{op}} \geq C \log H \log d\right) \\
&\leq \sum_{\pi_k \geq 1/(2d^4 H)} \pi_k \mathbb{P}_{X \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \Bigg(\left\| \sum_{h=1}^{H} w_h(X) \left(\frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X)\right)\left(\frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X)\right)^\top \right\|_{\mathsf{op}} \\
&\qquad \geq \frac{C}{2} \log H \log d \Bigg) + \frac{1}{2d^4} \\
&\leq \sum_{\pi_k \geq 1/(2d^4 H)} \pi_k \mathbb{P}_{X \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \left(\sum_{h=1}^{H} w_h(X) \left\| \frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X) \right\|_2^2 \geq \frac{C}{2} \log H \log d \right) + \frac{1}{2d^4},
\end{aligned}$$

where the last inequality uses the fact

$$\left\| \sum_{h=1}^{H} w_h(x) \left(\frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x)\right)\left(\frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x)\right)^\top \right\|_{\mathsf{op}} \leq \sum_{h=1}^{H} w_h(x) \left\| \frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x) \right\|_2^2.$$

To complete the proof, it suffices to show that, for any component k with prior probability $\pi_k \geq 1/(2d^4 H)$, if $X \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I)$ (so that $z_k(X) \sim \mathcal{N}(0, I_d)$), then the following holds

with probability at least $1 - 1/(2d^4)$:

$$\sum_{h=1}^{H} w_h(X) \left\| \frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X) \right\|_2^2 \lesssim \log(dH). \tag{109}$$

If $d \le C_0 \log(dH)$ for some sufficiently large constant $C > 0$, since $\log(d)\log(H) \ge d \log d$, Example 2 follows directly from Lemma 1 in Li and Yan (2024b). In the remaining analysis, we focus on the case $d \ge C_0 \log(dH)$. For convenience, we let

$$\gamma_{h,k} \coloneqq \overline{\sigma}_{h,k}^{-2} - 1 \in (-1, \infty).$$

We claim that, for any k satisfying $\pi_k \ge 1/(2d^4 H)$ and for each $h \in \{1, \cdots, H\}$,

$$\mathbb{P}_{X \sim N(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \left(\left\| \frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X) \right\|_2^2 \lesssim \max\left\{ \log(dH), \frac{w_k(X)}{w_h(X)} \right\} \right) \ge 1 - \frac{1}{2d^4 H}. \tag{110}$$

If (110) holds, then it together with the union bound shows that with probability exceeding $1 - 1/(2d^4)$,

$$\begin{aligned} \sum_{h=1}^{H} w_h(X) \left\| \frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X) \right\|_2^2 &\lesssim \sum_{h=1}^{H} w_h(X) \left(\log(dH) + \frac{w_k(X)}{w_h(X)} \right) \\ &= \log(dH) \sum_{h=1}^{H} w_h(X) + \sum_{h=1}^{H} w_k(X) = \log(dH) + 1, \end{aligned}$$

which confirms (109) and thus finishes the proof. Therefore, we only need to prove (110).

To validate (110), we proceed by considering the following two scenarios.

Case 1: both $\|\widetilde{\mu}_k - \widetilde{\mu}_h\|/\widetilde{\sigma}_k$ and $|\gamma_{h,k}|$ are small. We first consider the case that $\|\widetilde{\mu}_k - \widetilde{\mu}_h\|/\widetilde{\sigma}_k \le C_1\sqrt{\log(dH)}$ and $|\gamma_{h,k}| \le C_2\sqrt{\frac{\log(dH)}{d}}$ for some sufficiently large constants $C_1, C_2 > 0$. For any x satisfying $\|z_k(x)\|_2^2 \lesssim d$, we have

$$\begin{aligned} \left\| \frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x) \right\|_2^2 &= \left\| \frac{1}{\overline{\sigma}_{h,k}} \left(\frac{z_k(x)}{\overline{\sigma}_{h,k}} + \frac{1}{\overline{\sigma}_{h,k}} \frac{\widetilde{\mu}_k - \widetilde{\mu}_h}{\widetilde{\sigma}_k} \right) - z_k(x) \right\|_2^2 \\ &= \left\| \gamma_{h,k} z_k(x) + (\gamma_{h,k} + 1) \frac{\widetilde{\mu}_k - \widetilde{\mu}_h}{\widetilde{\sigma}_k} \right\|_2^2 \\ &\lesssim \gamma_{h,k}^2 \|z_k(x)\|_2^2 + (\gamma_{h,k}^2 + 1) \left\| \frac{\widetilde{\mu}_k - \widetilde{\mu}_h}{\widetilde{\sigma}_k} \right\|_2^2 \qquad (111) \\ &\lesssim \frac{\log(dH)}{d} \cdot d + \left(\frac{\log(dH)}{d} + 1 \right) \log(dH) \\ &\lesssim \log(dH), \qquad (112) \end{aligned}$$

provided that $d \ge \log(H)$. Moreover, by virtue of Lemma 1 in Laurent and Massart (2000), we have

$$\mathbb{P}_{X \sim N(\widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \left(\|z_k(X)\|_2^2 \le 13d \right) = \mathbb{P}_{z_k(X) \sim N(0, I_d)} \left(\|z_k(X)\|_2^2 \le 13d \right) \ge 1 - e^{-4d} \ge 1 - \frac{1}{2Hd^4},$$

provided that $d \ge \log(H)$. Putting (112) and the previous inequality together, we know that (110) holds.

Case 2: $\|\widetilde{\mu}_k - \widetilde{\mu}_h\|/\widetilde{\sigma}_k$ or $|\gamma_{h,k}|$ is large. Now, we turn to the scenario that $\|\widetilde{\mu}_k - \widetilde{\mu}_h\|/\widetilde{\sigma}_k \ge C_1\sqrt{\log(dH)}$ or $|\gamma_{h,k}| \ge C_2\sqrt{\frac{\log(dH)}{d}}$. It suffices to show that there exists a constant $C' > 0$ such that, with probability at least $1 - 1/(4d^4 H)$,

$$\log\left(\frac{w_h(X)}{C' w_k(X)} \left\| \frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X) \right\|_2^2 \right) \le 0. \tag{113}$$

To establish (113), we first compute

$$\begin{aligned}\frac{w_h(x)}{w_k(x)} &= \frac{\pi_h \phi(x \mid \widetilde{\mu}_h, \widetilde{\sigma}_h^2 I_d)}{\pi_k \phi(x \mid \widetilde{\mu}_k, \widetilde{\sigma}_k^2 I_d)} \\ &= \exp\left(-\frac{\|z_h(x)\|_2^2}{2} + \frac{\|z_k(x)\|_2^2}{2} + \frac{d}{2}\log\frac{\widetilde{\sigma}_k^2}{\widetilde{\sigma}_h^2} + \log\frac{\pi_h}{\pi_k}\right) \\ &= \exp\Big(-\frac{\gamma_{h,k}\|z_k(x)\|_2^2}{2} - \frac{(\gamma_{h,k}+1)}{2\widetilde{\sigma}_k^2}\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2 - \frac{z_k(x)^\top(\widetilde{\mu}_k - \widetilde{\mu}_h)}{\widetilde{\sigma}_k \overline{\sigma}_{h,k}^2} \\ &\qquad + \frac{d}{2}\log(\gamma_{h,k}+1) + \log\frac{\pi_h}{\pi_k}\Big).\end{aligned}$$

Since $\pi_k \geq 1/(2d^4H)$ and $H \geq 2$, we have

$$\log\frac{\pi_h}{\pi_k} \leq 4\log(dH).$$

In view of Lemma 1 in Laurent and Massart (2000), the standard Gaussian concentration inequality and the fact that $z_k(X) \sim N(0, I_d)$, we know that, with probability exceeding $1 - 1/(4d^4H)$,

$$\|z_k(X)\|_2^2 - d \lesssim \sqrt{d\log(dH)}, \qquad \text{and} \qquad |z_k(X)^\top(\widetilde{\mu}_k - \widetilde{\mu}_h)| \lesssim \|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2\sqrt{\log(dH)}. \tag{114}$$

We define the event $\mathcal{E} := \{(114) \text{ holds}\}$. On $\mathcal{E}$, one has

$$\begin{aligned}&\log\left(\frac{w_h(X)}{C' w_k(X)}\left\|\frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X)\right\|_2^2\right) \\ &\overset{(111)\text{ and }(114)}{\leq} -\frac{\gamma_{h,k}d - C_3|\gamma_{h,k}|\sqrt{d\log(dH)}}{2} - \frac{(\gamma_{h,k}+1)}{2\widetilde{\sigma}_k^2}\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2 \\ &\quad + \frac{C_4(\gamma_{h,k}+1)\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2\sqrt{\log(dH)}}{\widetilde{\sigma}_k} \\ &\quad + \frac{d}{2}\log(\gamma_{h,k}+1) + 4\log(dH) + \log\left(\gamma_{h,k}^2 d\right) + \log\left((\gamma_{h,k}+1)^2\frac{\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2}{C'\widetilde{\sigma}_k^2}\right) \\ &\overset{\text{(i)}}{\leq} -\frac{\gamma_{h,k}d - C_3'|\gamma_{h,k}|\sqrt{d\log(dH)}}{2} - \frac{(\gamma_{h,k}+1)}{2\widetilde{\sigma}_k^2}\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2 \\ &\quad + \frac{C_4'(\gamma_{h,k}+1)\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2\sqrt{\log(dH)}}{\widetilde{\sigma}_k} + \frac{d}{2}\log(\gamma_{h,k}+1) + 4\log(dH),\end{aligned} \tag{115}$$

where C_3 and C_4 are some universal constants, $C_3' = C_3 + 4$, $C_4' = C + 2/\sqrt{C'}$, and (i) comes from $\log(\gamma^2 d) = 2\log(|\gamma|\sqrt{d}) \leq 2|\gamma|\sqrt{d}$ and

$$\log\left((\gamma_{h,k}+1)^2\frac{\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2}{C'\widetilde{\sigma}_k^2}\right) \leq \frac{2}{\sqrt{C'}}(\gamma_{h,k}+1)^2\frac{\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2}{\widetilde{\sigma}_k}.$$

To prove (113), we consider three cases.

Case 2.1. If $\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2/\widetilde{\sigma}_k \geq C_1\sqrt{\log(dH)}$ and $|\gamma_{h,k}| \geq C_2\sqrt{\log(dH)/d}$, then we have

$$-\frac{(\gamma_{h,k}+1)}{2\widetilde{\sigma}_k^2}\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2 + \frac{C_4'(\gamma_{h,k}+1)\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2\sqrt{\log(dH)}}{\widetilde{\sigma}_k} \leq -\frac{(\gamma_{h,k}+1)}{4\widetilde{\sigma}_k^2}\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2, \tag{116}$$

In addition, by virtue of the fact that $\log(x+1) - x \leq -\frac{1}{4}|x|\min\{|x|, 1\}$ for any $x > -1$, we have

$$-\frac{\gamma_{h,k}d - C_3'|\gamma_{h,k}|\sqrt{d\log(dH)}}{2} + \frac{d}{2}\log(\gamma_{h,k}+1)$$

$$
\begin{aligned}
&\leq \frac{C_3'|\gamma_{h,k}|\sqrt{d\log(dH)}}{2} - \frac{d|\gamma_{h,k}|}{8}\min\{|\gamma_{h,k}|,1\} \\
&\leq \frac{d|\gamma_{h,k}|}{16}\sqrt{\frac{64C_3'^2\log(dH)}{d}} - \frac{d|\gamma_{h,k}|}{8}\min\{|\gamma_{h,k}|,1\} \\
&\leq \frac{d|\gamma_{h,k}|}{16}\min\left\{C_2\sqrt{\frac{\log(dH)}{d}},1\right\} - \frac{d|\gamma_{h,k}|}{8}\min\{|\gamma_{h,k}|,1\} \\
&\leq -\frac{d}{16}|\gamma_{h,k}|\min\{|\gamma_{h,k}|,1\}, \qquad (117)
\end{aligned}
$$

provided that $C_2 \geq 8C_3'$ and $C_0 \geq 64C_3'^2$. Combining (115), (116) and (117), we obtain that on the event $\mathcal{E}$,

$$
\begin{aligned}
&\log\left(\frac{w_h(X)}{C'w_k(X)}\left\|\frac{z_h(X)}{\overline{\sigma}_{h,k}} - z_k(X)\right\|_2^2\right) \\
&\quad\leq -\frac{d}{16}|\gamma_{h,k}|\min\{|\gamma_{h,k}|,1\} - \frac{(\gamma_{h,k}+1)}{4\widetilde{\sigma}_k^2}\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2^2 + 4\log(dH) \\
&\quad\leq -\frac{d}{16}C_2\sqrt{\log(dH)/d}\min\left\{C_2\sqrt{\log(dH)/d},1\right\} + 4\log(dH) \\
&\quad= \min\left\{-\frac{C_2^2}{16}\log(dH), -\frac{C_2}{16}\sqrt{d\log(dH)}\right\} + 4\log(dH) \leq 0.
\end{aligned}
$$

Case 2.2. If $\|\widetilde{\mu}_k - \widetilde{\mu}_h\|/\widetilde{\sigma}_k \geq C_1\sqrt{\log(dH)}$ and $|\gamma_{h,k}| \leq C_2\sqrt{\log(dH)/d}$, then (115) tells us that on the event $\mathcal{E}$,

$$
\begin{aligned}
&\log\left(\frac{w_h(x)}{C'w_k(x)}\left\|\frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x)\right\|_2^2\right) \\
&\leq \frac{C_3'|\gamma_{h,k}|\sqrt{d\log(dH)}}{2} - \frac{(\gamma_{h,k}+1)}{4\widetilde{\sigma}_k^2}\|\mu_k - \mu_h\|_2^2 + 4\log(dH) - \frac{d}{2}\left(\gamma_{h,k} - \log(\gamma_{h,k}+1)\right) \\
&\leq \frac{C_3'|\gamma_{h,k}|\sqrt{d\log(dH)}}{2} - \frac{(\gamma_{h,k}+1)}{4\widetilde{\sigma}_k^2}\|\mu_k - \mu_h\|_2^2 + 4\log(dH) \\
&\leq (C_2C_3'/2+4)\log(dH) - \frac{\|\mu_k - \mu_h\|_2^2}{4\widetilde{\sigma}_k^2} \leq 0,
\end{aligned}
$$

provided that $C_1^2 \geq 2C_2C_3' + 16$.

Case 2.3. If $|\gamma_{h,k}| \geq C_2\sqrt{\log(dH)/d}$ and $\|\widetilde{\mu}_k - \widetilde{\mu}_h\|/\widetilde{\sigma}_k \leq C_1\sqrt{\log(dH)}$, then (115) together with (117) shows that

$$
\begin{aligned}
&\log\left(\frac{w_h(x)}{C'w_k(x)}\left\|\frac{z_h(x)}{\overline{\sigma}_{h,k}} - z_k(x)\right\|_2^2\right) \\
&\quad\leq \frac{C_4'(\gamma_{h,k}+1)\|\widetilde{\mu}_k - \widetilde{\mu}_h\|_2\sqrt{\log(dH)}}{\widetilde{\sigma}_k} - \frac{d}{16}|\gamma_{h,k}|\min\{|\gamma_{h,k}|,1\} + 4\log(dH) \\
&\quad\leq (C_1C_4'\gamma_{h,k} + C_1C_4' + 4)\log(dH) - \frac{d}{16}|\gamma_{h,k}|\min\{|\gamma_{h,k}|,1\} \leq 0,
\end{aligned}
$$

provided that $C_2 \geq \max\{32(C_1C_4'+4), \sqrt{32(C_1C_4'+4)}\}$ and $C_0 \geq \max\{16C_1C_4', (32C_1C_4'/C_2)^2\}$.

Putting the results for all cases together, we conclude that (110) holds, which completes the proof.

Proof of $\widetilde{L}$. Consider a two-component Gaussian mixture with mean vectors $\mu_1 = \mathbf{1}$ and $\mu_2 = -\mathbf{1}$, $\sigma_1 = \sigma_2 = 1$, and $\pi_1 = \pi_2 = 0.5$. In this case, we have

$$
w_1(\mathbf{0}) = w_2(\mathbf{0}) = 0.5, \quad z_1(\mathbf{0}) = \sqrt{1-\tau}\mathbf{1}, \quad z_2(\mathbf{0}) = -\sqrt{1-\tau}\mathbf{1}, \quad \widetilde{\sigma}_1 = \widetilde{\sigma}_2 = 1,
$$

where $\mathbf{0}$ denotes the all-zero vector. Substituting into (107), and replacing $1-\tau$ with $\overline{\alpha}_t$, we have

$$\nabla s_t^\star(\mathbf{0}) = -I_d + \frac{\overline{\alpha}_t}{2}\mathbf{1}\mathbf{1}^\top + \frac{\overline{\alpha}_t}{2}\mathbf{1}\mathbf{1}^\top = -I_d + \overline{\alpha}_t \mathbf{1}\mathbf{1}^\top.$$

Thus we have

$$\widetilde{L} \geq \|\nabla \overline{s}_t^\star(\mathbf{0})\|_{\mathsf{op}} \geq \overline{\alpha}_t d - 1 = \Omega(d)$$

as long as $\overline{\alpha}_t \geq 1/2$ and $d \geq 4$.

G.3 Proof of Example 3.

In this case, $\nabla \overline{s}_\tau^\star(x)$ is a diagonal matrix with the i-th diagonal entry given by

$$\begin{aligned}[\nabla \overline{s}_\tau^\star(x)]_{i,i} = &-\frac{1}{\tau} + \frac{1}{\tau^2}\mathbb{E}[\|x_i - \sqrt{1-\tau}x_{0,i}\|_2^2 | \overline{X}_{\tau,i} = x_i] \\ &- \frac{1}{\tau^2}\|\mathbb{E}[x_i - \sqrt{1-\tau}x_{0,i}|\overline{X}_{\tau,i} = x_i]\|_2^2,\end{aligned}$$

where x_i, $\overline{X}_{\tau,i}$, and $x_{0,i}$ denote the i-th entry of x, $\overline{X}_\tau$, and x_0, respectively. Thus we have

$$\begin{aligned}&\|\nabla \overline{s}_\tau^\star(x)\|_{\mathsf{op}} \\ &\quad\leq \max_i |[\nabla \overline{s}_\tau^\star(x)]_{i,i}| \\ &\quad\leq \frac{1}{\tau} + \frac{1}{\tau^2}\max_i \mathbb{E}[\|x_i - \sqrt{1-\tau}x_{0,i}\|_2^2 | \overline{X}_{\tau,i} = x_i].\end{aligned} \tag{118}$$

Repeating similar arguments as in Eqn. (A.6) in Li and Yan (2024b), we know that for any $\delta \in (0,1)$ and x_i satisfying $\log p_{\overline{X}_{\tau,i}}(x_i) \geq -\theta \log \frac{1}{\delta}$ with some large constant $\theta \geq 1$,

$$\frac{1}{\tau^2}\mathbb{E}[\|x_i - \sqrt{1-\tau}x_{0,i}\|_2^2 | \overline{X}_{\tau,i} = x_i] \lesssim \frac{\log \frac{1}{\delta}}{\tau}.$$

Taking $\delta = d^{-1}$, we have

$$\frac{1}{\tau^2}\mathbb{E}[\|x_i - \sqrt{1-\tau}x_{0,i}\|_2^2 | \overline{X}_{\tau,i} = x_i] \lesssim \frac{\log d}{\tau}, \tag{119}$$

provided that $\log p_{\overline{X}_{\tau,i}}(x_i) \geq -\theta \log d$. Combining (118) and (119), we know that if $\log p_{\overline{X}_{\tau,i}}(x_i) \geq -\theta \log d$ for all $i \in d$

$$\tau \|\nabla \overline{s}_\tau^\star(x)\|_{\mathsf{op}} \lesssim \log d.$$

Repeating similar arguments as in proof of Eqn. (A.28b) in Li and Yan (2024b), we have

$$\mathbb{P}(\log p_{\overline{X}_{\tau,i}}(\overline{X}_{\tau,i}) \geq -\theta \log d) \leq \frac{1}{d^5},$$

which, together with the union bound, completes the proof.

For the global smoothness parameter $\widetilde{L}$, when $X_{0,i} \sim \mathsf{Unif}(\{-1,1\})$, we have

$$[\nabla s_t(0)]_{i,i} = \frac{2\overline{\alpha}_t - 1}{(1-\overline{\alpha}_t)^2},$$

and therefore

$$\|\nabla s_1(0)\|_{\mathsf{op}} = \frac{|2\overline{\alpha}_1 - 1|}{(1-\overline{\alpha}_1)^2} = \Omega\bigg(\frac{1}{(1-\overline{\alpha}_1)^2}\bigg).$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Some limitations have been discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are clearly stated in Theorems, and the proofs are in Section 4 and Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information needed to reproduce the experimental results in Figure 1 is provided in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification:

This is a theoretical work, and the experimental code is straightforward.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This is a theoretical paper and therefore does not involve any training or test procedures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Experiments in this paper does not require error bars or statistical significance measures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: As a theoretical paper, the experiments are simple to implement and do not rely on significant computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.