Dimension-free convergence of diffusion models for approximate Gaussian mixtures

Gen Li[*†] Changxiao Cai[*‡] Yuting Wei[§]

April 3, 2025

Abstract

Diffusion models are distinguished by their exceptional generative performance, particularly in producing high-quality samples through iterative denoising. While current theory suggests that the number of denoising steps required for accurate sample generation should scale linearly with data dimension, this does not reflect the practical efficiency of widely used algorithms like Denoising Diffusion Probabilistic Models (DDPMs). This paper investigates the effectiveness of diffusion models in sampling from complex high-dimensional distributions that can be well-approximated by Gaussian Mixture Models (GMMs). For these distributions, our main result shows that DDPM takes at most $\widetilde{O}(1/\varepsilon)$ iterations to attain an ε-accurate distribution in total variation (TV) distance, independent of both the ambient dimension d and the number of components K, up to logarithmic factors. Furthermore, this result remains robust to score estimation errors. These findings highlight the remarkable effectiveness of diffusion models in high-dimensional settings given the universal approximation capability of GMMs, and provide theoretical insights into their practical success.

Keywords: diffusion models, DDPM, Gaussian mixture models, iteration complexity, generative modeling

Contents

[*]The authors contributed equally. Corresponding author: Yuting Wei.
[†]Department of Statistics, The Chinese University of Hong Kong, genli@cuhk.edu.hk.
[‡]Department of Industrial and Operations Engineering, University of Michigan, cxcai@umich.edu.
[§]Department of Statistics and Data Science, University of Pennsylvania, ytwei@wharton.upenn.edu.

1 Introduction

Diffusion models have garnered significant attention for their remarkable generative capabilities, producing high-quality samples with enhanced stability (Diakonikolas et al., 2018; Dhariwal and Nichol, 2021; Song et al., 2020c; Ramesh et al., 2022). Compared to methods like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which generate samples in a single forward pass, diffusion models are designed to iteratively denoise samples over hundreds or thousands of steps. A prominent example is the widely used Denoising Diffusion Probabilistic Models (DDPM) sampler (Ho et al., 2020). The current theory suggests the number of denoising steps required for accurate sample generation should scale at least linearly with the data dimension (Chen et al., 2022; Benton et al., 2024) in order to learn the distribution accurately. While various acceleration schemes have been proposed in literature (see, e.g. Li and Cai (2024); Li et al. (2024a); Li and Jiao (2024); Wu et al. (2024b); Huang et al. (2024b,a); Taheri and Lederer (2025)), in practical applications such as high-resolution image synthesis, where the dimensionality of the data can be extremely large, DDPM often requires far fewer steps than predicted by theory while maintaining excellent sample quality.

This gap between theoretical complexity bounds and empirical performance has inspired a strand of recent research, investigating whether diffusion models have implicitly exploited structural properties of real-world data to circumvent worst-case complexity bounds. A growing line of works have shed light on this question by showing that diffusion models, in its original form, can automatically adapt to the intrinsic dimension of the target distribution without explicitly modeling its low-dimensional structure. Notably, prior work has examined cases where the data lies in low-dimensional linear spaces, low-dimensional manifolds, or distributions whose support have small covering number (Li and Yan, 2024a; Tang and Yang, 2024; Huang et al., 2024c; Potaptchik et al., 2024; Liang et al., 2025). In this work, we take a different perspective, and explore this question by focusing on a fundamental and well-studied statistical model: Gaussian Mixture Models (GMMs). GMMs serve as a cornerstone of statistical modeling and have been widely used to approximate complex distributions. Formally, we consider the setting where the target distribution is or can be approximated well by a mixture of isotropic Gaussians:

$$X_0 \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma^2 I_d), \tag{1}$$

where $\{\pi_k\}$ are mixture weights satisfying $\pi_k \in (0,1)$ and $\sum_{k=1}^{K} \pi_k = 1$. The study of Gaussian Mixture Models (GMMs) dates back to Pearson (1894), and a vast body of literature has since explored various aspects of GMMs, including parameter estimation, distribution learning, information-theoretic limits, computational efficiency and etc. This paper studies the performance of diffusion models in their original form when they are used to generate samples from a distribution that is close to GMMs. We refer readers to a more detailed exposition of related work in Section 1.3.

1.1 Diffusion models and sampling efficiency

In a nutshell, diffusion models consist of two processes: a forward process and a backward process. In the forward process, noise is gradually added to the data, transforming it into a noise-like distribution chosen *a priori* (e.g., a Gaussian distribution). Mathematically, given an initial sample $X_0 \in \mathbb{R}^d$ from the target distribution p_{data}, this transformation follows

$$X_t = \sqrt{\alpha_t} X_{t-1} + \sqrt{1-\alpha_t} W_t, \quad t = 1, 2, \ldots, T, \tag{2}$$

where $\alpha_t \in (0,1), t \geq 1$ denote the learning rates and $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), t \geq 1$ are i.i.d. standard d-dimensional Gaussian vectors independent of $(X_t)_{t=0}^T$. In the backward process, starting from $Y_T \sim \mathcal{N}(0, I_d)$, diffusion models iteratively denoise Y_T to approximate p_{data}. Classical results from stochastic differential equations (SDE) theory (e.g. Anderson (1982); Haussmann and Pardoux (1986)) show that under mild conditions, recovering p_{data} is possible provided access to the (Stein) score function $s_t^\star(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ for all $1 \leq t \leq T$, defined as

$$s_t^\star(x) := \nabla \log p_{X_t}(x), \quad \forall x \in \mathbb{R}^d. \tag{3}$$

Given the complexity of developing a comprehensive end-to-end theory, a divide-and-conquer approach — pioneered by (Chen et al., 2022) — has become standard, separating the score learning phase (i.e., estimating score functions reliably from training data) from the generative sampling phase (i.e., generating new data instances based on the estimated scores). The quality of the sampler in terms of its discrepancy to the target distribution depends on the errors from both phases. Over the past several years, the theoretical community has made significant process in understanding both phases. Notably, for the sampling phase, convergence theory has been established for various samplers (Liu et al., 2022; Lee et al., 2023; Chen et al., 2023a; Li et al., 2023; Chen et al., 2023c; Tang and Zhao, 2024; Liang et al., 2024a; Huang et al., 2024a; Gao and Zhu, 2024), especially DDPM and Denoising Diffusion Implicit Models (DDIM) which are widely adopted in practice (Ho et al., 2020; Song et al., 2020a). For DDPM, Benton et al. (2024) establishes an iteration complexity of $\widetilde{O}(d/\varepsilon^2)$ [1] in Kullback Leibler (KL) divergence, and Li and Yan (2024b) shows a complexity of $\widetilde{O}(d/\varepsilon)$ in total variation (TV) distance. When it comes to the DDIM sampler or the probability flow ODE, notably, an $\widetilde{O}(d/\varepsilon)$ iteration complexity has been established in Li et al. (2024b).

1.2 Learning GMMs using diffusion models

In the context of GMMs, several recent works have contributed towards unraveling the capabilities of diffusion models. In particular, inspired by diffusion models, Shah et al. (2023) introduced an algorithm designed for GMMs that achieves polynomial time complexity in d, provided the component centers are well-separated. Liang et al. (2024b) established an iteration complexity of $\widetilde{O}(d/\varepsilon^2)$ for obtaining an ε-accurate distribution measured in TV distance by analyzing the Lipschitz and second moments of GMMs. Additionally, Wu et al. (2024a); Chidambaram et al. (2024) investigated the role of guidance in diffusion models. Two exciting recent works (Chen et al., 2024; Gatmiry et al., 2024) proposed using piecewise polynomial regression to estimate the score functions, and they combined this with existing convergence result for DDPM to develop an end-to-end theory for DDPM. Notably, in these works, the number of diffusion steps scales also linearly with d. Further, Wang et al. (2024) explored diffusion models for mixtures of low-rank Gaussians. Despite these advancements, a fundamental question remains open:

Can diffusion models achieve efficient sampling when the target distribution is close to a GMM?

A glimpse of our main contributions. This paper investigates sampling from target distributions that admit faithful approximations by general isotropic GMMs, without imposing constraints on component separation or mixture weights. Our main result provides a non-asymptotic characterization of DDPM's iteration complexity for learning an ε-accurate distribution in TV distance. We prove that, given perfect score estimates and small GMM approximation error, DDPM requires at most

$$\widetilde{O}\Big(\frac{1}{\varepsilon}\Big),$$

number of iterations. Remarkably, this iteration complexity is independent of both the ambient dimension d and the number of components K, up to some logarithmic factors. Moreover, our result is robust to score estimation errors: the TV distance between the learned distribution and the target distribution scales proportionally to the score estimation error, modulo logarithmic factor. This leads to a surprising insight:

Even in ultra-high-dimensional settings, diffusion models remain highly effective in sampling distributions that are close to GMMs.

[1] The definition for $O(\cdot)$ and $\widetilde{O}(\cdot)$ notation can be found in Section 1.4.

1.3 Other related works

Learning GMMs. GMMs are fundamental statistical models that bear a well-established body of research from both statistics and computer science communities. One major line of research focuses on parameter estimation with some separation conditions. Partial examples include Dasgupta (1999); Vempala and Wang (2004); Arora and Kannan (2005); Kalai et al. (2010); Hsu and Kakade (2013); Diakonikolas et al. (2018); Hopkins and Li (2018); Kothari et al. (2018); Liu and Li (2022).

Our work is more closely related to the density estimation perspective, where no separation conditions are imposed (e.g. Diakonikolas and Kane (2020); Moitra and Valiant (2010); Dwivedi et al. (2020); Bakshi et al. (2022); Ho and Nguyen (2016)). In this setting, parameter estimation is information-theoretically infeasbile, yet accurate density estimation is still possible. The information theoretical limit of this problem is first characterized in Ashtiani et al. (2018) up to logarithmic factors, with a brute-forth algorithm that scales exponentially in both d and K. For one-dimensional Gaussian mixtures, Chen (1995); Heinrich and Kahn (2018); Wu and Yang (2020) obtained optimal estimation rates and practical algorithms, which were generalized to the high-dimensional case for mixtures of spherical Gaussians with a computationally efficient algorithm in Doss et al. (2023). Beyond finite mixtures, when mixing distribution is an arbitrary probability measure (e.g. Genovese and Wasserman (2000); Ghosal and Van Der Vaart (2001)), Saha and Guntuboyina (2020); Polyanskiy and Wu (2020); Kim and Guntuboyina (2022) established convergence rates and adaptivity, regarding the non-parametric maximum likelihood estimatior, generalizing the one-dimension results in Zhang (2009).

Score estimation. As mentioned earlier, score estimation plays a crucial role in diffusion models. Hyvärinen (2005) introduced an integration-by-parts-based approach to simplify score estimation. More recently, Song et al. (2020b) proposed training neural networks to learn score functions by minimizing the score matching objective. The theoretical guarantees for score estimation using neural networks have been analyzed across various distributional settings, including sub-Gaussian distributions (Cole and Lu, 2024), graphical models (Mei and Wu, 2023), low-dimensional structured distributions (Chen et al., 2023b; Kwon et al., 2025; De Bortoli, 2022), and Besov function space (Oko et al., 2023). These guarantees are often achieved by designing neural architectures that well approximate the true score function. Other than neural networks, classical methods such as kernel-based approaches and empirical Bayes smoothing have also been studied for score estimation (Cai and Li, 2025; Wibisono et al., 2024; Zhang et al., 2024; Dou et al., 2024). These methods have been shown to achieve minimax-optimal rates under some smoothness assumptions. Furthermore, Feng et al. (2024) demonstrated that statistical procedures based on score matching can achieve minimal asymptotic covariance for convex M-estimation.

1.4 Notation

For any a, the Dirac delta function $\delta_a(x)$ is defined as $\delta_a(x) = \infty$ if $x = a$ and $\delta_a(x) = 0$ otherwise. For positive integer $N > 0$, let $[N] \coloneqq \{1, \cdots, N\}$. In addition, given any matrix A, we use $\|A\|$, $\mathsf{tr}(A)$, and $\mathsf{det}(A)$ to denote the spectral norm, trace, and determinant of the matrix, respectively. Next, we recall the definitions of the KL divergence and TV distance to measure the discrepancies between two distributions. Specifically, for random vectors X and Y with probability density functions p_X and p_Y, let

$$\mathsf{KL}(X \,\|\, Y) \equiv \mathsf{KL}(p_X \,\|\, p_Y) = \int p_X(x) \log \frac{p_X(x)}{p_Y(x)} \, \mathrm{d}x,$$
$$\mathsf{TV}(X, Y) \equiv \mathsf{TV}(p_X, p_Y) = \frac{1}{2} \int |p_X(x) - p_Y(x)| \, \mathrm{d}x.$$

For any two functions $f(T)$, $g(T) > 0$, we write $f(T) \lesssim g(T)$ or $f(T) = O\big(g(T)\big)$ to indicate $f(T) \le Cg(T)$ for some absolute constant $C > 0$. We say $f(T) \asymp g(T)$ when $Cf(T) \le g(T) \le C'f(T)$ for some absolute constants $C' > C > 0$. The notation $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ represent the respective bounds up to logarithmic factors. Finally, we write $f(T) = o(g(T))$ to denote that $\limsup_{T\to\infty} f(T)/g(T) = 0$.

2 Preliminaries for diffusion models

Given training samples from a target distribution p_{data} on $\mathbb{R}^d$, diffusion models aim to generate new samples from p_{data}. Recall the forward process (2). If we define

$$\overline{\alpha}_t := \prod_{k=1}^{t} \alpha_k, \quad t = 1, 2, \ldots, T, \tag{4}$$

the forward process can be expressed as a linear combination of the initial distribution and a Gaussian noise

$$X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1 - \overline{\alpha}_t}\, \overline{W}_t, \quad t = 1, 2, \ldots, T, \tag{5}$$

where $\overline{W}_t \sim \mathcal{N}(0, I_d)$ denotes a d-dimensional standard Gaussian random vector independent of X_0. When $\overline{\alpha}_T$ is sufficiently small, X_T is well-approximated by a standard Gaussian distribution. Taking the continuum limit of (2), the process satisfies the stochastic differential equation (SDE):

$$\mathrm{d}x_t = -\frac{1}{2}\beta_t X_t \,\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}B_t \quad X_0 \sim p_{\mathsf{data}}; \quad t \in [0, T] \tag{6}$$

for some function $\beta_t : [0, T] \to \mathbb{R}$, where $(B_t)_{t \in [0,T]}$ is a standard Brownian motion in $\mathbb{R}^d$.

Diffusion models seek to reverse the above process by iteratively denoising noisy samples genearated from $\mathcal{N}(0, I_d)$, reconstructing data samples from p_{data}. From a continuous perspective, given a solution $(X_t)_{t \in [0,T]}$ to (6), classical SDE theory (Anderson, 1982; Haussmann and Pardoux, 1986) ensures that its time reversal $Y_t^{\mathsf{SDE}} := X_{T-t}$ satisfies:

$$\mathrm{d}Y_t^{\mathsf{SDE}} = \frac{1}{2}\beta_{T-t}\Big(Y_t^{\mathsf{SDE}} + 2\nabla \log p_{X_{T-t}}\big(Y_t^{\mathsf{SDE}}\big)\Big)\,\mathrm{d}t + \sqrt{\beta_{T-t}}\,\mathrm{d}B_t, \quad Y_0^{\mathsf{SDE}} \sim p_{X_T}; \quad t \in [0, T]. \tag{7}$$

Here, p_{X_t} denotes the marginal distribution of X_t in the forward SDE (6).

Score learning/matching. It is clear from the continuous perspective, that the score function $s_t^\star(x) := \nabla \log p_{X_t}(x)$ plays an important role in characterizing the reverse process. In fact, if $s_t^\star(x)$ were known exactly, the reverse process would be uniquely identified. In practice, however, score functions must be learned from training samples. A natural approach is to estimate $s_t^\star(x)$ within a pre-selected function class $\mathcal{F}$ by minimizing the expected squared error:

$$\min_{\widehat{s}_t \in \mathcal{F}} \mathbb{E}_{X \sim p_{X_t}}\Big[\big\|\widehat{s}_t(X) - \nabla \log p_{X_t}(X)\big\|^2\Big].$$

For Gaussian distributions, integration by parts allows reformulating this objective as (e.g., Hyvärinen (2005); Vincent (2011); Chen et al. (2022))

$$\min_{\widehat{s}_t : \mathbb{R}^d \to \mathbb{R}^d} \mathop{\mathbb{E}}_{W \sim \mathcal{N}(0, I_d), X_0 \sim p_{\mathsf{data}}}\left[\left\|\widehat{s}_t(X_t) + \frac{1}{\sqrt{1 - \overline{\alpha}_t}} W\right\|_2^2\right]. \tag{8}$$

Here, given the observed $X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1 - \overline{\alpha}_t}\, W$, one seeks to predict the independent noise W, a strategy known as score matching. This formulation is particularly useful for practical training since it does not require explicit knowledge of the score function $\nabla \log p_{X_t}$. Instead, it can be approximated using finite samples, making it more feasible for learning the score function from data.

The DDPM sampling procedure. To implement the sampling process, we must discretize the continuous dynamics and obtain score estimates at discrete time steps. Suppose that one obtains score estimates $\{\widehat{s}_t(\cdot)\}$ at $t = 1, \ldots, T$. Equipped with these score estimates, the renown DDPM algorithm Ho et al. (2020) is a stochastic sampler that recursively generates samples using the following update rule. Starting from $Y_T \sim \mathcal{N}(0, I_d)$, DDPM computes Y_{t-1} by

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + (1 - \alpha_t)\widehat{s}_t(Y_t)\big) + \sqrt{1 - \alpha_t}\, Z_t, \quad t = T, \ldots, 2. \tag{9}$$

Here, $Z_2, \ldots, Z_T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ is a sequence of i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$ independent of $(Y_t)_{t=2}^T$. In words, at each step, Y_{t-1} is a weighted sum of Y_t and its estimated score, plus an independent Gaussian noise.

3 Main results

In this section, we state our main results on the performances of DDPM when applied to distributions that can be well-approximated by GMMs (1) and discuss their consequences. Without loss of generality, we focus on the case where $\sigma = 1$ and therefore the covariance of each component is the identity matrix. Otherwise, our algorithm and analysis framework are readily extended to the general case by either rescaling the data or adjusting the learning rates accordingly.

We start by introducing some assumptions on the target distribution and the quality of our score estimates.

Assumption 1. There exists a Gaussian mixture model with K components such that that target distribution is close to a GMM in TV distance, namely,

$$\mathsf{TV}(p_{X_0}, p_{X_0^{\mathsf{GMM}}}) \leq \varepsilon_{\mathsf{apprx}} \qquad \text{with} \qquad X_0^{\mathsf{GMM}} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}\big(\mu_k, I_d\big). \tag{10}$$

In addition, the components of the GMM satisfies

$$\max_{k \in [K]} \|\mu_k\|_2 \leq T^{c_R}, \tag{11}$$

for some absolute constant $c_R > 0$.

Here, the ℓ_2 norm of the mean of each component is required to grow at most polynomially with the iteration number T. Expressing the boundedness condition in terms of T allows for cleaner and more concise convergence guarantees. Given that the constant c_R can be chosen arbitrarily large, this assumption allows each component to have exceedingly large mean value. Therefore, it holds true for most distributions that are encountered in practice. In addition, we do not impose any assumptions on the component separation $\min_{i \neq j} \|\mu_i - \mu_j\|_2$ or mixture weights $\{\pi_k\}$.

Remark 1. It is important to emphasize that our results stated below rely solely on the existence of such a GMM approximation X_0^{GMM} to the target distribution X_0. Crucially, there is no requirement to explicitly construct or estimate X_0^{GMM} in practice. Given the universal approximation capabilities of GMMs, this assumption is mild and widely applicable.

Next, we assess the quality of score estimates by their averaged ℓ_2 accuracy. This form of estimation error matches naturally with training procedures such as the score matching mentioned above.

Assumption 2. We assume there exist an score estimator $\{s_t\}_{t \in [T]}$ satisfy

$$\frac{1}{T} \sum_{t=1}^{T} \varepsilon_{\mathsf{score},t}^2 \leq \varepsilon_{\mathsf{score}}^2, \tag{12a}$$

where we define

$$\varepsilon_{\mathsf{score},t}^2 := \mathbb{E}_{X_t \sim p_{X_t}} \Big[\big\| s_t(X_t) - s_t^\star(X_t) \big\|_2^2 \Big], \qquad t \in [T]. \tag{12b}$$

Notably, this assumption requires the mean squared estimation error averaged over time steps is bounded, rather than the error at any individual step. This is commonly assumed in the literature of diffusion models (e.g., Chen et al. (2022); Benton et al. (2024); Li and Yan (2024b)).

When implementing DDPM (9), we use the score estimates $\{\widehat{s}_t\}_{t=1}^T$ given by $\widehat{s}_t := \mathsf{clip}\{s_t\}$ for $t \in [T]$, where the truncation function $\mathsf{clip}\{x\} : \mathbb{R}^d \to \mathbb{R}^d$ is defined as follows:

$$\mathsf{clip}\{x\} = \begin{cases} x, & \text{if } \|x\|_2 \leq C_{\mathsf{clip}} \sqrt{\frac{d \log(dT)}{1 - \overline{\alpha}_t}}, \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

with $C_{\mathsf{clip}} > 0$ being a sufficiently large absolute constant. We remark that this truncation procedure is introduced purely for technical reasons to simplify the analysis, which can be removed with a more refined analysis.

Convergence theory for DDPM. Before stating our main result, we introduce the learning rate schedule $\{\alpha_t\}_{t\in[T]}$. As adopted in previous works on diffusion models (e.g. Li and Cai (2024)), the learning rate sequence is defined iteratively using the cumulative products $\overline{\alpha}_t = \prod_{k=1}^t \alpha_k$ in Eq. (4). More specifically, define

$$\overline{\alpha}_T = \frac{1}{T^{c_0}}, \qquad \text{and} \qquad \overline{\alpha}_{t-1} = \overline{\alpha}_t + c_1 \frac{\log T}{T} \overline{\alpha}_t (1 - \overline{\alpha}_t), \quad t = T, \ldots, 2, \tag{14}$$

where $c_0, c_1 > 0$ are absolute constants such that c_0, c_1 are sufficiently large and $c_1/c_0 > 4$. As shown in Lemma 1, this choice of the learning rates yields that property that

$$1 - \alpha_t \leq \frac{1-\alpha_t}{1-\overline{\alpha}_t} \lesssim \frac{\log T}{T} \;\; \text{for } t \geq 2, \qquad \text{and} \qquad 1 - \alpha_1 \leq T^{-c_1/4}.$$

With these assumptions and preparations, we are positioned to state our main result below. The proof of this result is provided in Section 4, with the proofs of auxiliary lemmas postponed to Appendix A.

Theorem 1. *Under Assumptions 1–2, the output Y_1 of the DDPM sampler (9) with the learning rate selected according to (14) satisfies*

$$\mathsf{TV}(X_0, Y_1) \lesssim \frac{\log^2(KT)\log^2 T}{T} + \varepsilon_{\mathsf{score}}\sqrt{\log T} + \sqrt{d\varepsilon_{\mathsf{apprx}}}\log^{3/2}(dT). \tag{15}$$

As a special example, when the target distribution is exactly a GMM (i.e., $\varepsilon_{\mathsf{apprx}} = 0$), we have the following guarantees:

Corollary 1. *Suppose the target distribution is a GMM with its components satisfying (11) and the score estimates satisfy Assumption 2. The output Y_1 of the DDPM sampler (9) with the learning rate selected according to (14) obeys*

$$\mathsf{TV}(X_0, Y_1) \lesssim \frac{\log^2(KT)\log^2 T}{T} + \varepsilon_{\mathsf{score}}\sqrt{\log T}. \tag{16}$$

In a nutshell, Theorem 1 guarantees that for a target distribution that is close to an isotropic GMM, the sampling quality of DDPM, measured in TV distance, is governed by three components: the first accounts for the time discretization error arising from approximating the continuous SDE in Eq. (7) with a discrete procedure; the second component results from the score estimation error; the third component controls the approximation error of GMMs.

As a result, given access to perfect score estimates, and small enough ε, it only takes DDPM no larger than

$$\widetilde{O}\Big(\frac{1}{\varepsilon}\Big),$$

number of iterations to yield a sampler that is ε-close to the target distribution in terms of TV distance, provided that the target distribution is close to some GMM with $\varepsilon_{\mathsf{apprx}} = \widetilde{O}(\varepsilon^2/d)$. Notably, this iteration complexity is independent of both the ambient dimension d and the number of components K up to some logarithmic factors. In addition, our result is robust to score estimation error: the TV distance between our output distribution and the target distribution scales proportionally to the $\varepsilon_{\mathsf{score}}$, modulo logarithmic factor. This finding stands in sharp contrast to the common belief that diffusion models inherently require iteration complexity that scales with the dimension d.

Theorem 1 is established with respect to the TV distance between X_0 and Y_1, whereas, most theoretical results in diffusion models fail to directly handle the TV distance due to technical reasons. More specifically, most prior works consider the KL divergence which is a natural choice if Girsanov's theorem is invoked to handle the discrepancy between the forward process and the process when imperfect score functions are concerned. Noteworthily, a recent line of literature (e.g. Li et al. (2023); Li and Yan (2024b); Li et al. (2024b)) enriches the toolbox of analyzing diffusion models by providing a framework of directly working with the TV distance.

Intuition for efficient sampling of GMMs. Let us first provide some intuition on why one should expect an iteration number independent of both d and K up to logarithmic factors for GMMs. Consider an auxiliary forward process $(X_t^{\mathsf{GMM}})_{t=0}^T$ starting with the GMM $X_0^{\mathsf{GMM}} \sim \sum_{k=1}^K \pi_k \mathcal{N}\big(\mu_k, I_d\big)$ and

$$X_t^{\mathsf{GMM}} = \sqrt{\alpha_t} X_{t-1}^{\mathsf{GMM}} + \sqrt{1-\alpha_t} W_t, \quad t = 1, 2, \ldots, T, \tag{17}$$

where the learning rates $(\alpha_t)_{t=1}^T$ and Gaussian noise $(W_t)_{t=1}^T$ are the same as those in the original forward process Eq. (2). At time t, the distribution of this auxiliary forward process obeys

$$X_t^{\mathsf{GMM}} \sim \sum_{k=1}^K \pi_k \mathcal{N}\big(\sqrt{\overline{\alpha}_t}\mu_k, I_d\big). \tag{18}$$

Regarding this process, define the Jacobian matrix $J_t(x) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ of the score function $s_t^{\star\mathsf{GMM}}(x) \coloneqq \nabla \log p_{X_t^{\mathsf{GMM}}}(x)$ with $J_t(x) \coloneqq \frac{\partial}{\partial x} s_t^{\star\mathsf{GMM}}(x)$. By direct computation (as in Eq. (25)), it satisfies

$$J_t(x) = -I_d + \overline{\alpha}_t \Bigg\{ \sum_{k=1}^K \pi_k^{(t)}(x) \mu_k \mu_k^\top - \bigg(\sum_{k=1}^K \pi_k^{(t)}(x) \mu_k \bigg) \bigg(\sum_{k=1}^K \pi_k^{(t)}(x) \mu_k \bigg)^\top \Bigg\}, \quad \forall x \in \mathbb{R}^d. \tag{19}$$

Here, $\pi_k^{(t)}(x)$ denotes the probability of x lying in the cluster k at time t of the forward process

$$\pi_k^{(t)}(x) \coloneqq \frac{\pi_k \exp\big(-\frac{1}{2}\|x - \sqrt{\overline{\alpha}_t}\mu_k\|_2^2\big)}{\sum_{i=1}^K \pi_i \exp\big(-\frac{1}{2}\|x - \sqrt{\overline{\alpha}_t}\mu_i\|_2^2\big)}, \quad \forall k \in [K],\ t \in [T]. \tag{20}$$

We prove in Lemma 6 that with high probability

$$\mathsf{tr}\Big(I_d + J_t\big(X_t^{\mathsf{GMM}}\big)\Big) \le C_1 \log(KT), \tag{21}$$

for some absolute constant C_1 independent of the problem parameters. This relation, which does not generally hold, is the key to our dimension-free iteration complexity for GMMs. Since the target distribution is well-approximated by the Gaussian mixture satisfying Eq. (21), it is reasonable to expect a dimension-independent iteration complexity for this general distribution as well.

However, generalizing this result to distributions that are close to GMMs without imposing stronger assumptions on the score estimation quality is highly non-trivial. The challenge lies in the fact that Assumption 2 concerns the forward process starting from the target distribution, not the auxiliary GMMs. We carefully control the score estimation error regarding GMMs in terms of $\varepsilon_{\mathsf{score}}^2$ in our Lemma 4, which relies heavily on the statistical properties of GMMs, and may be of independent interest.

Comparisons to prior literature. Alongside the seminar work (Chen et al., 2022) and the follow-up works (e.g. Lee et al. (2023); Chen et al. (2023a); Benton et al. (2024)), Theorem 1 investigates the algorithmic aspect of learning GMMs assuming efficient score estimation/matching. Among existing works, the most closely related to ours is Liang et al. (2024b), which established an iteration complexity of $\widetilde{O}(d/\varepsilon^2)$ by analyzing the Lipschitz properties and second moments of GMMs. Compared to Li and Yan (2024b), which studied DDPM for general distributions under mild assumptions, and attained an $\widetilde{O}(d/\varepsilon)$ iteration complexity, Liang et al. (2024b) did not demonstrate any adaptation of DDPM to GMMs. Our result, however, highlights the surprising adaptive property of diffusion models in this setting.

Beyond the algorithmic aspect, Chen et al. (2024); Gatmiry et al. (2024) developed an end-to-end theory by leveraging piecewise polynomial regression for score estimation and integrating it with existing convergence results on DDPM. The runtime and sample complexity of the resulting algorithms scale quasi-polynomially with K/ε or $\log(K/\varepsilon)$ depending on the covariance assumptions. Notably, the number of diffusion steps used in these two works still scales linearly with d. Our result serves as a complementary contribution to Chen et al. (2024); Gatmiry et al. (2024) by isolating the component of the iteration complexity that is independent of both d and K, up to a logarithmic factor.

4 Analysis

In this section, we describe our proof strategies for deriving Theorem 1. The proofs for auxiliary lemmas and facts are deferred to the appendix.

4.1 Preliminaries

Before proceeding with the main analysis, let us collect several key properties that shall be used in our later analysis.

To begin with, Lemma 1 below characterizes the behavior of the learning rates $(\alpha_t)_{t\in[T]}$ chosen in (14). The proof of this result can be found in Li and Cai (2024, Appendix B.1).

Lemma 1. *The learning rates $(\alpha_t)_{t\in[T]}$ specified in (14) satisfy that*

$$1-\alpha_t \le c_1\frac{\log T}{T}, \quad t=2,\dots,T, \tag{22a}$$

$$1-\alpha_1 \le \frac{1}{T^{c_1/4}}, \tag{22b}$$

where c_1 is defined in (14).

Next, in light of Assumption 1 on the GMM and the forward process (2), it is straightforward to verify each X_t^{GMM} is another GMM. Concretely, we have

$$X_t^{\mathsf{GMM}} \sim \sum_{k=1}^K \pi_k \mathcal{N}\big(\sqrt{\overline{\alpha}_t}\mu_k, I_d\big),$$

with the density function given by

$$p_{X_t^{\mathsf{GMM}}}(x) = \sum_{k=1}^K \pi_k \frac{1}{(2\pi)^{d/2}}\exp\Big(-\frac{1}{2}\big\|x-\sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2\Big). \tag{23}$$

Through direct computation, we can derive that its score function takes the following explicit form:

$$s_t^{\star\mathsf{GMM}}(x) \coloneqq \nabla \log p_{X_t^{\mathsf{GMM}}}(x) = -\sum_{k=1}^K \pi_k^{(t)}(x)\big(x-\sqrt{\overline{\alpha}_t}\mu_k\big) = -x + \sqrt{\overline{\alpha}_t}\sum_{k=1}^K \pi_k^{(t)}(x)\mu_k, \tag{24}$$

where we recall in Eq. (20) that $\pi_k^{(t)}(x):\mathbb{R}^d\to[0,1]$ is defined as

$$\pi_k^{(t)}(x) \coloneqq \frac{\pi_k \exp\big(-\frac{1}{2}\|x-\sqrt{\overline{\alpha}_t}\mu_k\|_2^2\big)}{\sum_{i=1}^K \pi_i \exp\big(-\frac{1}{2}\|x-\sqrt{\overline{\alpha}_t}\mu_i\|_2^2\big)}, \quad \forall k\in[K],\ t\in[T].$$

In addition, the Jacobian matrix $J_t(x):\mathbb{R}^d\to\mathbb{R}^{d\times d}$ of $s_t^{\star\mathsf{GMM}}(x)$ can be computed as

$$\begin{aligned} J_t(x) \coloneqq \frac{\partial s_t^{\star\mathsf{GMM}}(x)}{\partial x} &= -I_d + \overline{\alpha}_t\sum_{k=1}^K \pi_k^{(t)}\Big(\mu_k - \sum_{i=1}^K \pi_i^{(t)}\mu_i\Big)\Big(\mu_k - \sum_{i=1}^K \pi_i^{(t)}\mu_i\Big)^\top \\ &= -I_d + \overline{\alpha}_t\Bigg\{\sum_{k=1}^K \pi_k^{(t)}\mu_k\mu_k^\top - \Big(\sum_{k=1}^K \pi_k^{(t)}\mu_k\Big)\Big(\sum_{k=1}^K \pi_k^{(t)}\mu_k\Big)^\top\Bigg\}. \end{aligned} \tag{25}$$

As a remark, we note that $I_d + J_t(x) \succeq 0$ for any $t\in[T]$ and $x\in\mathbb{R}^d$.

Next, we introduce the event $\mathcal{E}_t$ for each $t\in[T]$ as follows:

$$\mathcal{E}_t \coloneqq \Big\{ x\in\mathbb{R}^d : \mathsf{tr}\big(I_d + J_t(x)\big) \le C_1\log(KT) \quad \text{and}$$

$$\sum_{k=1}^{K} \pi_k^{(t)} \exp\Big(-\zeta_k^{(t)}(x)\Big) \le \exp\Big(C_2(1-\alpha_t)^2 \log^2(KT)\Big)\Big\}, \tag{26}$$

for some absolute constants $C_1, C_2 > 0$, where we further define $\zeta_k^{(t)}(x) : \mathbb{R}^d \to \mathbb{R}$ for each $k \in [K]$:

$$\zeta_k^{(t)}(x) \coloneqq \frac{1-\alpha_t^2}{2\alpha_t^2}\left(\big\|x - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 - \sum_{i=1}^{K} \pi_i^{(t)} \big\|x - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2\right) + \frac{1-\alpha_t}{\alpha_t^2} s_t^\star(x)^\top \sum_{i=1}^{K} \pi_i^{(t)} \sqrt{\overline{\alpha}_t}(\mu_i - \mu_k). \tag{27}$$

Finally, we extend the d-dimensional Euclidean space $\mathbb{R}^d$ by adding a single point ∞, to obtain $\mathbb{R}^d \cup \{\infty\}$. Intuitively, this set $\{\infty\}$ serves as a convenient way to capture all atypical points in the reverse process.

4.2 Step 1: Constructing auxiliary processes

To facilitate our main analysis, we introduce several auxiliary processes below. These processes are constructed only for analysis purpose and are not used in our sampling algorithm.

Sequence $(Y_t^\star)_{t=0}^T$ using true scores. We begin by constructing an auxiliary reverse process $(Y_t^\star)_{t=0}^T$ using the true score functions $\{s_t^{\star\mathsf{GMM}}(\cdot)\}_{t=1}^T$ of the diffused GMMs $(X_t^{\mathsf{GMM}})_{t=1}^T$:

$$Y_T^\star \sim \mathcal{N}(0, I_d), \quad Y_{t-1}^\star \coloneqq \frac{1}{\sqrt{\alpha_t}}\big(Y_t^\star + (1-\alpha_t) s_t^{\star\mathsf{GMM}}(Y_t^\star)\big) + \sqrt{1-\alpha_t} Z_t; \quad t = T, \dots, 1. \tag{28}$$

where $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ is a sequence of i.i.d. standard d-dimensional Gaussian random vectors independent of $(Y_t^\star)_{t=0}^T$.

Sequence $(\overline{Y}_t^-, \overline{Y}_t)_{t=0}^T$. Next, we introduce two auxiliary sequences $(\overline{Y}_t^-)_{t=0}^T$ and $(\overline{Y}_t)_{t=0}^T$ to capture the discretization error modulo some low probability event. These sequences, together with Y_T implemented practically, form a Markov chain with the following transition structure:

$$Y_T \to \overline{Y}_T^- \to \overline{Y}_T \to \overline{Y}_{T-1}^- \to \overline{Y}_{T-1} \to \cdots \to \overline{Y}_1^- \to \overline{Y}_1 \to \overline{Y}_0^- \to \overline{Y}_0. \tag{29}$$

- *Initialization.* For $t = T$, we define

$$\overline{Y}_T^- \coloneqq \begin{cases} Y_T, & \text{if } Y_T \in \mathcal{E}_T, \\ \infty, & \text{otherwise.} \end{cases} \tag{30a}$$

  The density of $\overline{Y}_T^-$ satisfies

$$p_{\overline{Y}_T^-}(y) = p_{Y_T}(y) \mathbb{1}\big\{y \in \mathcal{E}_T\big\} + \mathbb{P}\big\{Y_T \notin \mathcal{E}_T\big\}\, \delta_\infty(y). \tag{30b}$$

- *Transition from $\overline{Y}_t^-$ to $\overline{Y}_t$.* For $t = T, \dots, 0$, we define $\overline{Y}_t$ as follows: conditional on $\overline{Y}_t^- = y$,

$$\overline{Y}_t \coloneqq \begin{cases} y, & \text{with prob. } \overline{p}_t(y)/p_{\overline{Y}_t^-}(y) \wedge 1, \\ \infty, & \text{with prob. } 1 - \big\{\overline{p}_t(y)/p_{\overline{Y}_t^-}(y) \wedge 1\big\}, \end{cases} \tag{31a}$$

  where we denote

$$\overline{p}_t \coloneqq p_{X_t^{\mathsf{GMM}}}, \quad \forall t \ge 0. \tag{31b}$$

  The conditional density of $\overline{Y}_t$ given $\overline{Y}_t^- = y$ obeys

$$p_{\overline{Y}_t | \overline{Y}_t^-}(x \mid y) = \big\{\overline{p}_t(y)/p_{\overline{Y}_t^-}(y) \wedge 1\big\}\delta_y(x) + \big(1 - \big\{\overline{p}_t(y)/p_{\overline{Y}_t^-}(y) \wedge 1\big\}\big)\delta_\infty(x). \tag{31c}$$

  We make a critical implication of the above construction: for any $t \ge 0$, the density of $\overline{Y}_t$ satisfies

$$p_{\overline{Y}_t}(y) = \big\{\overline{p}_t(y)/p_{\overline{Y}_t^-}(y) \wedge 1\big\} p_{\overline{Y}_t^-}(y) = p_{X_t^{\mathsf{GMM}}}(y) \wedge p_{\overline{Y}_t^-}(y), \quad \forall y \in \mathbb{R}^d. \tag{32}$$

- *Transition from $\overline{Y}_t$ to $\overline{Y}_{t-1}^-$.* For each $t = T, \dots, 1$, we first draw a candidate sample

$$\widetilde{Y}_{t-1} \coloneqq \frac{1}{\sqrt{\alpha_t}}\big(\overline{Y}_t + (1-\alpha_t)s_t^{\star\mathsf{GMM}}(\overline{Y}_t)\big) + \sqrt{1-\alpha_t}W_t, \tag{33a}$$

  where $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), t \geq 1$ is a sequence of i.i.d. standard Gaussian random vectors independent of $(Z_t)_{t=1}^T$, and then define

$$\overline{Y}_{t-1}^- \coloneqq \begin{cases} \widetilde{Y}_{t-1}, & \text{if } \overline{Y}_t \in \mathcal{E}_t \text{ and } \widetilde{Y}_{t-1} \in \mathcal{E}_t, \\ \infty, & \text{otherwise.} \end{cases} \tag{33b}$$

  The conditional density of $\overline{Y}_{t-1}^-$ given $\overline{Y}_t = y$ satisfies: if $y \in \mathcal{E}_t$, then

$$p_{\overline{Y}_{t-1}^-|\overline{Y}_t}(x \mid y) = p_{Y_{t-1}^\star|Y_t^\star}(x \mid y)\mathbb{1}\big\{x \in \mathcal{E}_t\big\} + \mathbb{P}\big\{Y_{t-1}^\star \notin \mathcal{E}_t \mid Y_t^\star = y\big\}\delta_\infty(x); \tag{33c}$$

  otherwise,

$$p_{\overline{Y}_{t-1}^-|\overline{Y}_t}(x \mid y) = \delta_\infty(x). \tag{33d}$$

Sequence $(\widehat{Y}_t^-, \widehat{Y}_t)_{t=0}^T$. Finally, we introduce two additional auxiliary sequences $(\widehat{Y}_t^-)_{t=0}^T$ and $(\widehat{Y}_t)_{t=0}^T$, which forms the following Markov chain together with Y_T:

$$Y_T \to \widehat{Y}_T^- \to \widehat{Y}_T \to \widehat{Y}_{T-1}^- \to \widehat{Y}_{T-1} \to \cdots \to \widehat{Y}_1^- \to \widehat{Y}_1 \to \widehat{Y}_0^- \to \widehat{Y}_0. \tag{34}$$

- *Initialization.* For $t = T$, we initialize $\widehat{Y}_T^- = \overline{Y}_T^-$.
- *Transition from $\widehat{Y}_t^-$ to $\widehat{Y}_t$.* For $t = T, \dots, 0$, the conditional density of $\widehat{Y}_t$ given $\widehat{Y}_t^- = y$ obeys

$$p_{\widehat{Y}_t|\widehat{Y}_t^-}(x \mid y) = p_{\overline{Y}_t|\overline{Y}_t^-}(x \mid y). \tag{35}$$

- *Transition from $\widehat{Y}_t$ to $\widehat{Y}_{t-1}^-$.* For $t = T, \dots, 1$, the conditional density of $\overline{Y}_{t-1}^-$ given $\overline{Y}_t = y$ satisfies: if $y \in \mathcal{E}_t$, then

$$p_{\widehat{Y}_{t-1}^-|\widehat{Y}_t}(x \mid y) = p_{Y_{t-1}|Y_t}(x \mid y)\mathbb{1}\big\{x \in \mathcal{E}_t\big\} + \mathbb{P}\Big\{Y_{t-1} \notin \mathcal{E}_t \mid Y_t = y\Big\}\,\delta_\infty(x); \tag{36a}$$

  otherwise,

$$p_{\widehat{Y}_{t-1}^-|\widehat{Y}_t}(x \mid y) = \delta_\infty(x). \tag{36b}$$

The sequences $(\widehat{Y}_t^-)_{t=0}^T$ and $(\widehat{Y}_t)_{t=0}^T$ are constructed following the same principles as $(\overline{Y}_t^-)_{t=0}^T$ and $(\overline{Y}_t)_{t=0}^T$, with one key difference: the transition from $\widehat{Y}_t$ to $\widehat{Y}_t^-$ is computed using estimated score functions rather than the true score functions.

A crucial property. It is noteworthy that for any $t \geq 0$, the density of $\widehat{Y}_t$ satisfies

$$p_{\widehat{Y}_t}(x) \leq p_{Y_t}(x), \quad \forall x \in \mathbb{R}^d, \tag{37}$$

and consequently, $p_{\widehat{Y}_t}(x) \geq p_{Y_t}(x)$ for $x = \infty$. To see this, we first note that the base case $t = T$ holds since $\widehat{Y}_t \overset{\text{d}}{=} Y_t$, which arises from $\widehat{Y}_T^- = \overline{Y}_T^-$ and $p_{\widehat{Y}_t|\widehat{Y}_t^-} = p_{\overline{Y}_t|\overline{Y}_t^-}$ by (35). Next, suppose that (37) holds for $t+1$. Then for any $x \in \mathbb{R}^d$, one has

$$\begin{aligned} p_{\widehat{Y}_t}(x) &\overset{\text{(i)}}{=} \big\{p_{X_t^{\mathsf{GMM}}}(x)/p_{\overline{Y}_t^-}(x) \wedge 1\big\}p_{\widehat{Y}_t^-}(x) \leq p_{\widehat{Y}_t^-}(x) = \int_{\mathbb{R}^d} p_{\widehat{Y}_t^-|\widehat{Y}_{t+1}}(x \mid y)p_{\widehat{Y}_{t+1}}(y)\,\mathrm{d}y \\ &\overset{\text{(ii)}}{\leq} \int_{\mathbb{R}^d} p_{Y_t|Y_{t+1}}(x \mid y)p_{Y_{t+1}}(y)\,\mathrm{d}y = p_{Y_t}(x), \end{aligned}$$

where (i) uses (35) and (31c); (ii) is true due to the induction hypothesis and (36a).

Error decomposition. In view of triangle's inequality, we can upper bound the TV distance between p_{Y_1} and p_{X_0} by

$$
\begin{aligned}
\mathsf{TV}\big(p_{Y_1}, p_{X_0}\big) &\le \mathsf{TV}\big(p_{Y_1}, p_{X_1^{\mathsf{GMM}}}\big) + \mathsf{TV}\big(p_{X_1^{\mathsf{GMM}}}, p_{X_0^{\mathsf{GMM}}}\big) + \mathsf{TV}\big(p_{X_0^{\mathsf{GMM}}}, p_{X_0}\big) \\
&\le \mathsf{TV}\big(p_{\overline{Y}_1}, p_{X_1^{\mathsf{GMM}}}\big) + \mathsf{TV}\big(p_{\overline{Y}_1}, p_{Y_1}\big) + \mathsf{TV}\big(p_{X_1^{\mathsf{GMM}}}, p_{X_0^{\mathsf{GMM}}}\big) + \mathsf{TV}\big(p_{X_0^{\mathsf{GMM}}}, p_{X_0}\big) \\
&= \mathsf{TV}\big(p_{\overline{Y}_1}, p_{X_1^{\mathsf{GMM}}}\big) + \mathsf{TV}\big(p_{\overline{Y}_1}, p_{Y_1}\big) + \mathsf{TV}\big(p_{X_1^{\mathsf{GMM}}}, p_{X_0^{\mathsf{GMM}}}\big) + \varepsilon_{\mathsf{apprx}} \qquad (38)
\end{aligned}
$$

where the last line leverages Assumption 1. Here, the first term acts in the role of discretization error modulo some low probability event, as $\overline{Y}_t$ is defined using the true scores; the second term captures the error caused by imperfect score estimation; the last two terms arise from the GMM approximation. In the sequel, we control each term separately.

4.3 Step 2: Bounding discretization error

In this section, we proceed to bound $\mathsf{TV}(p_{X_1^{\mathsf{GMM}}}, p_{\overline{Y}_1})$. Let us first define function $\Delta_t(x) : \mathbb{R}^d \to \mathbb{R}$, where for each $t = 1, \dots, T$:

$$
\Delta_t(x) := p_{X_t^{\mathsf{GMM}}}(x) - p_{\overline{Y}_t}(x), \quad \forall x \in \mathbb{R}^d. \qquad (39)
$$

In view of relation (32), one has $\Delta_t(x) \ge 0$ for all $t \ge 0$ and $x \in \mathbb{R}^d$. Applying the formula for the total variation $\mathsf{TV}(p, q) = \int_{x:\, p(x) > q(x)} \big(p(x) - q(x)\big)\,\mathrm{d}x$, we find that

$$
\mathsf{TV}\big(p_{X_1^{\mathsf{GMM}}}, p_{\overline{Y}_1}\big) = \int_{\mathbb{R}^d \cup \{\infty\}} \big(p_{X_1^{\mathsf{GMM}}}(x) - p_{\overline{Y}_1}(x)\big) \mathbb{1}\big\{p_{X_1^{\mathsf{GMM}}}(x) > p_{\overline{Y}_1}(x)\big\}\,\mathrm{d}x = \int_{\mathbb{R}^d} \Delta_1(x)\,\mathrm{d}x. \qquad (40)
$$

Thus, it is sufficient to bound $\int \Delta_1(x)\,\mathrm{d}x$, which shall be done using an inductive argument. We start with the base case, which is characterized by the Lemma 2 below.

Lemma 2. *It satisfies that*

$$
\int_{\mathbb{R}^d} \Delta_T(x)\,\mathrm{d}x \lesssim T^{-3}. \qquad (41)
$$

Proof. See Appendix A.1. □

In addition, Lemma 3 below establishes the inductive relationship between t and $t-1$.

Lemma 3. *For all $t = T, \dots, 1$, one has*

$$
\int_{\mathbb{R}^d} \Delta_{t-1}(x)\,\mathrm{d}x - \int_{\mathbb{R}^d} \Delta_t(x)\,\mathrm{d}x \lesssim (1-\alpha_t)^2 \log^2(KT) + T^{-3}, \qquad (42)
$$

Proof. See Appendix A.2. □

Consequently, combining (41)–(42) with (40) leads to

$$
\begin{aligned}
\mathsf{TV}\big(p_{X_1^{\mathsf{GMM}}}, p_{\overline{Y}_1}\big) &= \int_{\mathbb{R}^d} \Delta_1(x)\,\mathrm{d}x \le \int_{\mathbb{R}^d} \Delta_T(x)\,\mathrm{d}x + T \cdot O\big((1-\alpha_t)^2 \log^2(KT)\big) + T \cdot O\big(T^{-3}\big) \\
&\lesssim \frac{1}{T^3} + \frac{\log^2(KT)\log^2 T}{T} + \frac{1}{T^2} \\
&\asymp \frac{\log^2(KT)\log^2 T}{T}, \qquad (43)
\end{aligned}
$$

where the penultimate step uses $1 - \alpha_t \lesssim \log T / T$ by (22).

4.4 Step 3: Relating to score estimation error

Next, we control the term $\mathsf{TV}\big(p_{\overline{Y}_1}, p_{Y_1}\big)$. First, in view of basic calculations, we can write

$$\begin{aligned}
\mathsf{TV}\big(p_{\overline{Y}_1},\ p_{Y_1}\big) &= \int_{\mathbb{R}^d} \big(p_{\overline{Y}_1}(x) - p_{Y_1}(x)\big)\mathbb{1}\big\{p_{\overline{Y}_1}(x) > p_{Y_1}(x)\big\}\,\mathrm{d}x + \mathbb{P}\big\{\overline{Y}_1 = \infty\big\} \\
&\overset{\text{(i)}}{\leq} \int_{\mathbb{R}^d} \big(p_{\overline{Y}_1}(x) - p_{\widehat{Y}_1}(x)\big)\mathbb{1}\big\{p_{\overline{Y}_1}(x) > p_{\widehat{Y}_1}(x)\big\}\,\mathrm{d}x + \mathbb{P}\big\{\overline{Y}_1 = \infty\big\} \\
&\overset{\text{(ii)}}{\leq} \mathsf{TV}\big(p_{\overline{Y}_1}, p_{\widehat{Y}_1}\big) + \mathsf{TV}\big(p_{X_1^{\mathsf{GMM}}}, p_{\overline{Y}_1}\big) \\
&\overset{\text{(iii)}}{\leq} \sqrt{\mathsf{KL}\big(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1}\big)} + O\bigg(\frac{\log^2(KT)\log^2 T}{T}\bigg).
\end{aligned} \tag{44}$$

where (i) arises from (37) that $p_{Y_1}(x) \geq p_{\widehat{Y}_1}(x)$ for any $x \in \mathbb{R}^d$; (ii) uses $\mathbb{P}\{\overline{Y}_1 = \infty\} \leq \mathsf{TV}(p_{X_1^{\mathsf{GMM}}}, p_{\overline{Y}_1})$ since $X_1^{\mathsf{GMM}} \in \mathbb{R}^d$; (iii) applies Pinsker's inequality and (43).

To further control the right hand side of (44), it suffices to bound $\mathsf{KL}\big(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1}\big)$ from above. Towards this, notice that

$$\begin{aligned}
\mathsf{KL}\big(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1}\big) &\overset{\text{(i)}}{\leq} \mathsf{KL}\big(p_{\overline{Y}_T^-, \overline{Y}_T, \dots, \overline{Y}_1^-, \overline{Y}_1} \parallel p_{\widehat{Y}_T^-, \widehat{Y}_T, \dots, \widehat{Y}_1^-, \widehat{Y}_1}\big) \\
&\overset{\text{(ii)}}{=} \mathsf{KL}\big(p_{\overline{Y}_T^-} \parallel p_{\widehat{Y}_T^-}\big) + \sum_{t=1}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t^-}}\Big[\mathsf{KL}\big(p_{\overline{Y}_t | \overline{Y}_t^- = x_t} \parallel p_{\widehat{Y}_t | \widehat{Y}_t^- = x_t}\big)\Big] \\
&\quad + \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t}}\Big[\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big)\Big] \\
&\overset{\text{(iii)}}{=} \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t}}\Big[\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big)\Big].
\end{aligned} \tag{45}$$

Here, (i) applies the data-processing inequality; (ii) uses the chain rule of KL divergence and the Markov property; (iii) is true since we initialize $\widehat{Y}_T^- = \overline{Y}_T^-$ and the transition kernels from $\widehat{Y}_t^-$ to $\widehat{Y}_t$ are the same as those from $\overline{Y}_t^-$ to $\overline{Y}_t$ for all $t \geq 1$.

Note that $Y_{t-1}^\star \mid Y_t^\star = x_t \sim \mathcal{N}\big(\frac{1}{\sqrt{\alpha_t}}\big(x_t + (1-\alpha_t)s_t^{\star\mathsf{GMM}}(x_t)\big), (1-\alpha_t)I_d\big)$ and $Y_{t-1} \mid Y_t = x_t \sim \mathcal{N}\big(\frac{1}{\sqrt{\alpha_t}}\big(x_t + (1-\alpha_t)\widehat{s}_t(x_t)\big), (1-\alpha_t)I_d\big)$. For any $x_t \in \mathcal{E}_t$, write $p(\cdot) = p_{Y_{t-1}^\star | Y_t^\star = x_t}(\cdot)$ and $q(\cdot) = p_{Y_{t-1} | Y_t = x_t}(\cdot)$. In view of the definitions (36a) and (31c), one has

$$\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big) = \int_{\mathcal{E}_t} p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x + \log\frac{\int_{\mathcal{E}_t^{\mathrm{c}}} p(x)\,\mathrm{d}x}{\int_{\mathcal{E}_t^{\mathrm{c}}} q(x)\,\mathrm{d}x}\int_{\mathcal{E}_t^{\mathrm{c}}} p(x)\,\mathrm{d}x. \tag{46}$$

Now invoke Li and Yan (2024b, Lemma 6) to derive

$$\begin{aligned}
\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big) &\leq \int_{\mathbb{R}^d} p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x \\
&= \mathsf{KL}\big(p_{Y_{t-1}^\star | Y_t^\star = x_t} \parallel p_{Y_{t-1} | Y_t = x_t}\big) \\
&\overset{\text{(i)}}{=} \frac{1-\alpha_t}{2\alpha_t}\big\|\widehat{s}_t(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\big\|_2^2 \\
&\overset{\text{(ii)}}{\lesssim} \frac{\log T}{T}(1-\overline{\alpha}_t)\big\|\widehat{s}_t(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\big\|_2^2.
\end{aligned} \tag{47}$$

Here, (i) uses the formula of the KL divergence for two normal distributions; (ii) uses $1 - \alpha_t \leq (1-\alpha_t)/(1-\overline{\alpha}_t) \lesssim \log T/T$ by (22). Meanwhile, for any $x_t \in \mathcal{E}_t^{\mathrm{c}}$, we know from (36b) that

$$\mathsf{KL}\big(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\big) = 0. \tag{48}$$

Combined with (45), the above bounds give

$$
\begin{aligned}
\mathsf{KL}\big(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1}\big) &\overset{\text{(i)}}{\leq} \sum_{t=2}^{T} \mathbb{E}_{x_t \sim \overline{p}_t}\Big[\mathsf{KL}\big(p_{\overline{Y}_{t-1}^{-}|\overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^{-}|\widehat{Y}_t = x_t}\big)\Big] \\
&\overset{\text{(ii)}}{\lesssim} \frac{\log T}{T} \sum_{t=2}^{T} \int_{\mathbb{R}^d} (1-\overline{\alpha}_t)\|\widehat{s}_t(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\|_2^2 \, p_{X_t^{\mathsf{GMM}}}(x_t)\, \mathrm{d}x_t \\
&\overset{\text{(iii)}}{=} \big(\varepsilon_{\mathsf{score}}^{\mathsf{GMM}}\big)^2 \log T,
\end{aligned}
\tag{49}
$$

where (i) arises from (48) and (32) that $p_{\overline{Y}_t}(x) \leq \overline{p}_t(x)$ for all $x \in \mathbb{R}^d$; (ii) uses (47); in (iii) we define

$$
\varepsilon_{\mathsf{score},t}^{\mathsf{GMM}} \coloneqq \sqrt{\int_{\mathbb{R}^d} \big\|\widehat{s}_t(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\big\|_2^2 \, p_{X_t^{\mathsf{GMM}}}(x_t)\, \mathrm{d}x_t}; \tag{50a}
$$

$$
\varepsilon_{\mathsf{score}}^{\mathsf{GMM}} \coloneqq \sqrt{\frac{1}{T}\sum_{t=2}^{T}(1-\overline{\alpha}_t)\big(\varepsilon_{\mathsf{score},t}^{\mathsf{GMM}}\big)^2}. \tag{50b}
$$

Substituting (49) into (44) leads to

$$
\mathsf{TV}\big(p_{Y_1},\, p_{\overline{Y}_1}\big) \lesssim \frac{\log^2(KT)\log^2 T}{T} + \varepsilon_{\mathsf{score}}^{\mathsf{GMM}}\sqrt{\log T}. \tag{51}
$$

4.5 Step 4: Controlling GMM approximation error

Putting relations (43) and (51) together with (38) yields

$$
\mathsf{TV}\big(p_{X_0}, p_{Y_1}\big) \lesssim \frac{\log^2(KT)\log^2 T}{T} + \varepsilon_{\mathsf{score}}^{\mathsf{GMM}}\sqrt{\log T} + \mathsf{TV}\big(p_{X_0^{\mathsf{GMM}}}, p_{X_1^{\mathsf{GMM}}}\big) + \varepsilon_{\mathsf{apprx}}. \tag{52}
$$

Hence, it remains to control $\mathsf{TV}\big(p_{X_0^{\mathsf{GMM}}}, p_{X_1^{\mathsf{GMM}}}\big)$ and the term $\varepsilon_{\mathsf{score}}^{\mathsf{GMM}}$ defined in (50).

We begin with $\mathsf{TV}\big(p_{X_0^{\mathsf{GMM}}}, p_{X_1^{\mathsf{GMM}}}\big)$. Since $X_1^{\mathsf{GMM}} = \sqrt{\alpha_1}X_0^{\mathsf{GMM}} + \sqrt{1-\alpha_1}\,W_1$ with $W_1 \sim \mathcal{N}(0, I_d)$ independent of X_0^{GMM}, one knows

$$
X_1^{\mathsf{GMM}} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\sqrt{\alpha_1}\mu_k, I_d).
$$

For each component, we can use Pinsker's inequality and the KL divergence formula for Gaussian distributions to get

$$
\mathsf{TV}^2\big(\mathcal{N}(\mu_k, I_d), \mathcal{N}(\sqrt{\alpha_1}\mu_k, I_d)\big) \leq \frac{1}{2}\mathsf{KL}\big(\mathcal{N}(\mu_k, I_d) \parallel \mathcal{N}(\sqrt{\alpha_1}\mu_k, I_d)\big) \leq \frac{1}{2}(1-\sqrt{\alpha_1})^2\|\mu_k\|_2^2 \leq \frac{1}{2}(1-\alpha_1)^2\|\mu_k\|_2^2,
$$

where the last step holds as $\alpha_1 \in (0,1)$. It follow from the convexity of the TV distance and Jensen's inequality that

$$
\begin{aligned}
\mathsf{TV}\big(p_{X_0^{\mathsf{GMM}}}, p_{X_1^{\mathsf{GMM}}}\big) &\leq \sum_{k=1}^{K} \pi_k \mathsf{TV}\big(\mathcal{N}(\mu_k, I_d), \mathcal{N}(\sqrt{\alpha_1}\mu_k, I_d)\big) \\
&\lesssim (1-\alpha_1)\sum_{k=1}^{K}\pi_k\|\mu_k\|_2 \lesssim T^{-c_1/4}T^{c_R} \leq T^{-1},
\end{aligned}
\tag{53}
$$

where the penultimate step arises from (22) in Lemma 1 and Assumption 1, and the last step holds as long as c_1 is large enough.

Now, it is only left for us to control the term each $\varepsilon_{\mathsf{score}}^{\mathsf{GMM}}$. Note that $\varepsilon_{\mathsf{score},t}^{\mathsf{GMM}}$ for $t \in [T]$ is defined with respect to the truncated score estimator $\widehat{s}_t = \mathsf{clip}\{s_t\}$ (see (13)) and $p_{X_t^{\mathsf{GMM}}}$ whereas Assumption 2 on the score matching only guarantees the estimation error of s_t with respect to p_{X_t}. To handle this issue, we will apply similar analysis as in Li and Cai (2024, Lemma 4) to bound it in terms of the score estimation error $\varepsilon_{\mathsf{score},t}$ specified in Assumption 2. The resulting bound is presented in the following lemma; the proof is deferred to Appendix A.3.

Lemma 4. *The score error term $\varepsilon_{\mathsf{score}}^{\mathsf{GMM}}$ defined in (50) satisfies:*

$$\varepsilon_{\mathsf{score}}^{\mathsf{GMM}} \lesssim \varepsilon_{\mathsf{score}} + \sqrt{d\varepsilon_{\mathsf{apprx}}}\,\log(dT) + T^{-1}. \tag{54}$$

Substituting (53) and (54) into (52), we conclude that

$$\mathsf{TV}(p_{X_0}, p_{Y_1}) \lesssim \frac{\log^2(KT)\log^2 T}{T} + \sqrt{d\varepsilon_{\mathsf{apprx}}}\,\log^{3/2}(dT) + \varepsilon_{\mathsf{score}}\sqrt{\log T}.$$

This completes the proof of Theorem 1.

5 Discussion

In summary, this paper explores the effectiveness of diffusion models in learning distributions that can be well-approximiated by GMMs and presents new theoretical insights on how diffusion models implicitly exploit data structure to achieve efficient sampling. While DDPM requires a number of iterations that scale linearly with data dimension in the worst case, our main result unveils a surprising efficiency of DDPM: it only requires an iteration complexity of $\widetilde{O}(1/\varepsilon)$ to learn an ε-accurate distribution, independent of both the data dimension d and the number of mixture components K, up to logarithmic factors. This result suggests that diffusion models can efficiently learn structured distributions even in ultra-high-dimensional settings.

Before concluding, we highlight several promising directions for future investigation. First, while this paper focuses on mixtures of spherical Gaussians, an important next step is to analyze the iteration complexity of DDPM when applied to more general cases, such as mixtures with well-conditioned but arbitrary covariances, as considered in Chen et al. (2024). Additionally, in order to learn an ε-accurate distribution, it takes DDPM an iteration complexity of order $\widetilde{O}(1/\varepsilon)$. While we suspect that the ε dependence can not be further improved, it would be interesting to derive a matching lower bound to rigorously confirm the sharpness of our result. Finally, our analysis primarily addresses the sampling phase, leaving open the question of how score estimation efficiency is affected by the structure of GMMs. It remains a crucial direction to establish an end-to-end theory that integrates both score learning and sampling and fully unleashes the potential of diffusion models adapting to low-dimensional structure.

Ackowledgement

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

References

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

Arora, S. and Kannan, R. (2005). Learning mixtures of separated nonspherical Gaussians. *The Annals of Applied Probability*, 15(1A):69 – 92.

Ashtiani, H., Ben-David, S., Harvey, N., Liaw, C., Mehrabian, A., and Plan, Y. (2018). Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. *Advances in Neural Information Processing Systems*, 31.

Bakshi, A., Diakonikolas, I., Jia, H., Kane, D. M., Kothari, P. K., and Vempala, S. S. (2022). Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247.

Benton, J., Bortoli, V., Doucet, A., and Deligiannidis, G. (2024). Nearly d-linear convergence bounds for diffusion models via stochastic localization.

Cai, C. and Li, G. (2025). Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*.

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

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, pages 221–233.

Chen, M., Huang, K., Zhao, T., and Wang, M. (2023b). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2023c). The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36:68552–68575.

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

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

Chidambaram, M., Gatmiry, K., Chen, S., Lee, H., and Lu, J. (2024). What does guidance do? a fine-grained analysis in a simple setting. *arXiv preprint arXiv:2409.13074*.

Cole, F. and Lu, Y. (2024). Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian distributions. In *The Twelfth International Conference on Learning Representations*.

Dasgupta, S. (1999). Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE.

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

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

Diakonikolas, I. and Kane, D. M. (2020). Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195. IEEE.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060.

Doss, N., Wu, Y., Yang, P., and Zhou, H. H. (2023). Optimal estimation of high-dimensional gaussian location mixtures. *The Annals of Statistics*, 51(1):62–95.

Dou, Z., Kotekal, S., Xu, Z., and Zhou, H. H. (2024). From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*.

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., and Yu, B. (2020). Singularity, misspecification and the convergence rate of em. *The Annals of Statistics*, 48(6):3161–3182.

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

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

Gatmiry, K., Kelner, J., and Lee, H. (2024). Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127.

Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263.

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

Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation.

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

Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures.

Hopkins, S. B. and Li, J. (2018). Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034.

Hsu, D. and Kakade, S. M. (2013). Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20.

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

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647 – 1684.

Kalai, A. T., Moitra, A., and Valiant, G. (2010). Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562.

Kim, A. K. and Guntuboyina, A. (2022). Minimax bounds for estimating multivariate gaussian location mixtures. *Electronic Journal of Statistics*, 16(1):1461–1484.

Kothari, P. K., Steinhardt, J., and Steurer, D. (2018). Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046.

Kwon, H. K., Kim, D., Ohn, I., and Chae, M. (2025). Nonparametric estimation of a factorizable density using diffusion models. *arXiv preprint arXiv:2501.01783*.

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. (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., 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. *arXiv preprint arXiv:2410.13738*.

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*.

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

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

Liang, Y., Shi, Z., Song, Z., and Zhou, Y. (2024b). Unraveling the smoothness properties of diffusion models: A gaussian mixture perspective. *arXiv preprint arXiv:2405.16418*.

Liu, A. and Li, J. (2022). Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261.

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

Mei, S. and Wu, Y. (2023). Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*.

Moitra, A. and Valiant, G. (2010). Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE.

Oko, K., Akiyama, S., and Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Polyanskiy, Y. and Wu, Y. (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

Potaptchik, P., Azangulov, I., and Deligiannidis, G. (2024). Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*.

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.

Saha, S. and Guntuboyina, A. (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics*, 48(2):738–762.

Shah, K., Chen, S., and Klivans, A. (2023). Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649.

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

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020c). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Taheri, M. and Lederer, J. (2025). Regularization can make diffusion models more efficient. *arXiv preprint arXiv:2502.09151*.

Tang, R. and Yang, Y. (2024). Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR.

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

Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860.

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

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. (2024). Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*.

Wibisono, A., Wu, Y., and Yang, K. Y. (2024). Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*.

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

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

Wu, Y. and Yang, P. (2020). Optimal estimation of gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48(4):1981–2007.

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318.

Zhang, K., Yin, H., Liang, F., and Liu, J. (2024). Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*.

A Proof of technical lemmas

A.1 Proof of Lemma 2

Recalling that $\Delta_T(x) \coloneqq p_{X_T^{\mathsf{GMM}}}(x) - p_{\overline{Y}_T}(x) \geq 0$ for any $x \in \mathbb{R}^d$, we can derive

$$\int_{\mathbb{R}^d} \Delta_T(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \big(p_{X_T^{\mathsf{GMM}}}(x) - p_{\overline{Y}_T}(x)\big)\mathbb{1}\big\{p_{X_T^{\mathsf{GMM}}}(x) > p_{\overline{Y}_T}(x)\big\}\,\mathrm{d}x$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} \int_{\mathbb{R}^d} \big(p_{X_T^{\mathsf{GMM}}}(x) - p_{\overline{Y}_T^-}(x)\big) \mathbb{1}\big\{p_{X_T^{\mathsf{GMM}}}(x) > p_{\overline{Y}_T^-}(x)\big\}\, \mathrm{d}x \\
&\overset{\text{(ii)}}{=} \mathsf{TV}\big(p_{X_T^{\mathsf{GMM}}}, p_{\overline{Y}_T^-}\big) \\
&\le \mathsf{TV}\big(p_{X_T^{\mathsf{GMM}}}, p_{Y_T}\big) + \mathsf{TV}\big(p_{Y_T}, p_{\overline{Y}_T^-}\big),
\end{aligned} \tag{55}
$$

where (i) arises from (32) that $p_{\overline{Y}_T}(x) = p_{X_T^{\mathsf{GMM}}}(x) \wedge p_{\overline{Y}_T^-}(x)$ for any $x \in \mathbb{R}^d$, (ii) uses the formula of the total variation $\mathsf{TV}(p,q) = \int_{x:\, p(x)>q(x)} \big(p(x) - q(x)\big)\, \mathrm{d}x$ and $X_T^{\mathsf{GMM}} \in \mathbb{R}^d$.

Consequently, it suffices to control the two quantities in (55) respectively.

- For the first term $\mathsf{TV}\big(p_{X_T^{\mathsf{GMM}}}, p_{Y_T}\big)$ corresponding to the initialization error, we can derive

$$
\begin{aligned}
\mathsf{KL}\big(p_{X_T^{\mathsf{GMM}}} \parallel p_{Y_T}\big) &\overset{\text{(i)}}{\le} \mathbb{E}\Big[\mathsf{KL}\big(p_{X_T^{\mathsf{GMM}}}(\cdot \mid X_0^{\mathsf{GMM}}) \parallel p_{Y_T}(\cdot)\big)\Big] \\
&\overset{\text{(ii)}}{=} \frac{1}{2}\mathbb{E}\Big[d(1-\overline{\alpha}_T) - d + \big\|\sqrt{\overline{\alpha}_T} X_0^{\mathsf{GMM}}\big\|_2^2 - d\log(1-\overline{\alpha}_T)\Big] \\
&\overset{\text{(iii)}}{\lesssim} T^{-c_0}\mathbb{E}\Big[\big\|X_0^{\mathsf{GMM}}\big\|_2^2\Big] \overset{\text{(iv)}}{\lesssim} T^{-c_0}(T^{c_R} + d) \overset{\text{(v)}}{\lesssim} T^{-8},
\end{aligned} \tag{56}
$$

  where (i) arises from the convexity of the KL divergence; (ii) applies the KL divergence formula for two normal distributions; (iii) is due to the choice of the learning rate $\overline{\alpha}_T = T^{-c_0} = o(1)$ in (14) and $\log(1-x) \ge -x$ for any $x \in [0, 1/2]$; (iv) holds due to Assumption 1 that

$$
\mathbb{E}\Big[\big\|X_0^{\mathsf{GMM}}\big\|_2^2\Big] = \sum_{k=1}^{K} \pi_k \big(\|\mu_k\|_2^2 + d\big) \le T^{c_R} + d;
$$

  and (v) holds as long as T and c_0 are sufficiently large. It then follows from Pinsker's inequality that

$$
\mathsf{TV}\big(p_{X_T^{\mathsf{GMM}}} \parallel p_{Y_T}\big) \le \sqrt{\mathsf{KL}\big(p_{X_T^{\mathsf{GMM}}} \parallel p_{Y_T}\big)} \lesssim T^{-4}. \tag{57}
$$

- We proceed to consider the second term $\mathsf{TV}\big(p_{Y_T}, p_{\overline{Y}_T^-}\big)$. By the construction of $\overline{Y}_T^-$ (see (30)), one can write

$$
\begin{aligned}
\mathsf{TV}\big(p_{Y_T}, p_{\overline{Y}_T^-}\big) &\overset{\text{(i)}}{=} \int_{\mathbb{R}^d} \big(p_{Y_T}(x) - p_{\overline{Y}_T^-}(x)\big)\mathbb{1}\big\{p_{Y_T}(x) > p_{\overline{Y}_T^-}(x)\big\}\, \mathrm{d}x \\
&\overset{\text{(ii)}}{=} \int_{\mathcal{E}_t^{\mathrm{c}}} p_{Y_T}(x)\, \mathrm{d}x \\
&\overset{\text{(iii)}}{\le} \int_{\mathcal{E}_t^{\mathrm{c}}} p_{X_T^{\mathsf{GMM}}}(x)\, \mathrm{d}x + \mathsf{TV}\big(p_{X_T^{\mathsf{GMM}}}, p_{Y_T}\big).
\end{aligned}
$$

  Here, (i) uses the formula of the total variation $\mathsf{TV}(p,q) = \int_{x:\, p(x)>q(x)} \big(p(x) - q(x)\big)\, \mathrm{d}x$; (ii) follows from $Y_T \in \mathbb{R}^d$ and (30b) that $p_{\overline{Y}_T^-}(x) = p_{Y_T}(x)$ if $x \in \mathcal{E}_T$ and $p_{\overline{Y}_T^-}(x) = 0$ if $x \in \mathbb{R}^d \setminus \mathcal{E}_T$; (iii) arises from the definition of total variation distance that $\mathsf{TV}(p,q) = \sup_B |p(B) - q(B)|$. As we shall see momentarily in Lemma 6 in Appendix A.5, one has

$$
\int_{\mathcal{E}_T^{\mathrm{c}}} p_{X_T^{\mathsf{GMM}}}(x) \lesssim T^{-3}.
$$

  Combined with (57), this leads to

$$
\mathsf{TV}\big(p_{Y_T}, p_{\overline{Y}_T^-}\big) \lesssim T^{-3} + T^{-4} \asymp T^{-3}. \tag{58}
$$

In conclusion, substituting (57) and (58) into (55) completes the proof of Lemma 2.

A.2 Proof of Lemma 3

Fix an arbitrary $t \in [T]$. To analyze $\int_{\mathbb{R}^d} \Delta_t(x_t)\,\mathrm{d}x_t$, let us first introduce a function $\Delta_{t\to t-1}(x) : \mathbb{R}^d \to \mathbb{R}$ where

$$\Delta_{t\to t-1}(x) \coloneqq \int_{x_t\in\mathcal{E}_t} p_{Y_{t-1}^\star|Y_t^\star}(x \mid x_t)\Delta_t(x_t)\,\mathrm{d}x_t, \quad \forall x \in \mathbb{R}^d. \tag{59}$$

Note that in view of relation (32), $\Delta_t(x) \geq 0$ for all $x \in \mathbb{R}^d$ and therefore $\Delta_{t\to t-1}(x) \geq 0$ for all $x \in \mathbb{R}^d$. It is easily seen that

$$\int_{\mathbb{R}^d} \Delta_{t\to t-1}(x_{t-1})\,\mathrm{d}x_{t-1} = \int_{x_t\in\mathcal{E}_t}\int_{x_{t-1}\in\mathbb{R}^d} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1} \mid x_t)\,\mathrm{d}x_{t-1}\Delta_t(x_t)\,\mathrm{d}x_t \leq \int_{\mathbb{R}^d} \Delta_t(x_t)\,\mathrm{d}x_t. \tag{60}$$

As a result, to upper bound $\int_{\mathbb{R}^d} \Delta_{t-1}(x)\,\mathrm{d}x - \int_{\mathbb{R}^d} \Delta_t(x)\,\mathrm{d}x$, it is sufficient to consider $\int_{\mathbb{R}^d} \Delta_{t-1}(x)\,\mathrm{d}x - \int_{\mathbb{R}^d} \Delta_{t\to t-1}(x_{t-1})\,\mathrm{d}x_{t-1}$.

Towards this, we find it helpful to first make the following observation. For any $x_{t-1} \in \mathbb{R}$ such that $\Delta_{t-1}(x_{t-1}) > 0$, or equivalently, $p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) > p_{\overline{Y}_{t-1}}(x_{t-1})$, we have

$$p_{X_{t-1}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) \geq p_{\overline{Y}_{t-1}^-}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}). \tag{61}$$

Here, we use the fact that $p_{\overline{Y}_{t-1}}(x_{t-1}) = p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) \wedge p_{\overline{Y}_{t-1}^-}(x_{t-1})$ since $p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) > p_{\overline{Y}_{t-1}}(x_{t-1})$. To further control the right hand side, recall the definition of $\Delta_t(x)$ in (39) and the constructed transition kernel of $p_{\overline{Y}_{t-1}^-|\overline{Y}_t}$ in (33c). For any $x_{t-1} \in \mathbb{R}^d$, we arrive at

$$\begin{aligned}
p_{\overline{Y}_{t-1}^-}(x_{t-1}) &\geq \int_{x_t\in\mathcal{E}_t} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1} \mid x_t)p_{\overline{Y}_t}(x_t)\,\mathrm{d}x_t\\
&= \int_{x_t\in\mathcal{E}_t} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1} \mid x_t)p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t - \Delta_{t\to t-1}(x_{t-1}).
\end{aligned} \tag{62}$$

As a result, we obtain

$$\begin{aligned}
&p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1})\\
&\quad\geq \int_{x_t\in\mathcal{E}_t} p_{Y_{t-1}^\star|Y_t^\star}(x_{t-1} \mid x_t)p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t\\
&\quad\overset{\text{(i)}}{=} \int_{\mathcal{E}_t}\left(\frac{1}{2\pi(1-\alpha_t)}\right)^{d/2}\exp\Big(-\frac{\|\sqrt{\alpha_t}x_{t-1} - x_t - (1-\alpha_t)s_t^\star(x_t)\|_2^2}{2\alpha_t(1-\alpha_t)}\Big)p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t\\
&\quad\overset{\text{(ii)}}{=} \int_{\mathcal{E}_t}\left(\frac{1}{2\pi(1-\alpha_t)}\right)^{d/2}\exp\Big(-\frac{\|\sqrt{\alpha_t}x_{t-1} - u_t\|_2^2}{2\alpha_t(1-\alpha_t)}\Big)\det\big(I_d + (1-\alpha_t)J_t(x_t)\big)^{-1}p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}u_t.
\end{aligned} \tag{63}$$

where (i) uses (62); (ii) arises from (28) that $Y_{t-1}^\star \mid Y_t^\star \sim \mathcal{N}\big(\alpha_t^{-1/2}[Y_t^\star + (1-\alpha_t)s_t^{\star\mathsf{GMM}}(Y_t^\star)], (1-\alpha_t)I_d\big)$; (ii) applies the change of variable

$$u_t \coloneqq x_t + (1-\alpha_t)s_t^{\star\mathsf{GMM}}(x_t).$$

To bound the integral in (63), we present Lemma 5 below.

Lemma 5. *For any $t \in [T]$, the following holds for any $x_t \in \mathcal{E}_t$:*

$$\begin{aligned}
&\det\big(I_d + (1-\alpha_t)J_t(x_t)\big)^{-1}p_{X_t^{\mathsf{GMM}}}(x_t)\\
&\quad= \left(\frac{1}{2\pi\alpha_t^2}\right)^{d/2}\exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big)\sum_{k=1}^K \pi_k\exp\Big(-\frac{\|u_t - \sqrt{\alpha_t}\mu_k\|^2}{2\alpha_t^2}\Big).
\end{aligned} \tag{64}$$

Proof. See Appendix A.4. □

Plugging (64) into (63) and invoking the inequality $\exp(x) \ge 1 + x$ leads to

$$
\begin{aligned}
&p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}^{\mathsf{GMM}}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) \\
&\quad\ge \exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big) \\
&\qquad\cdot \int_{\mathcal{E}_t}\bigg(\frac{1}{4\pi^2\alpha_t^2(1-\alpha_t)}\bigg)^{d/2}\exp\bigg(-\frac{\|\sqrt{\alpha_t}x_{t-1}-u_t\|_2^2}{2\alpha_t(1-\alpha_t)}\bigg)\sum_{k=1}^K\pi_k\exp\bigg(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}\mu_k\|_2^2}{2\alpha_t^2}\bigg)\,\mathrm{d}u_t. \qquad (65)
\end{aligned}
$$

To further control the right hand side, direct computations give that

$$
\begin{aligned}
&\int_{\mathbb{R}^d}\bigg(\frac{1}{4\pi^2\alpha_t^2(1-\alpha_t)}\bigg)^{d/2}\exp\bigg(-\frac{\|\sqrt{\alpha_t}x_{t-1}-u_t\|_2^2}{2\alpha_t(1-\alpha_t)}\bigg)\sum_{k=1}^K\pi_k\exp\bigg(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}\mu_k\|_2^2}{2\alpha_t^2}\bigg)\,\mathrm{d}u_t \\
&\quad= \int_{\mathbb{R}^d}\bigg(\frac{1}{4\pi^2\alpha_t^2(1-\alpha_t)}\bigg)^{d/2}\sum_{k=1}^K\pi_k\exp\bigg(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}(1-\alpha_t)\mu_k-\alpha_t^{3/2}x_{t-1}\|_2^2}{2\alpha_t^2(1-\alpha_t)}-\frac{\|x_{t-1}-\sqrt{\overline{\alpha}_t/\alpha_t}\mu_k\|_2^2}{2}\bigg)\,\mathrm{d}u_t \\
&\quad\overset{\text{(i)}}{=} \int_{\mathbb{R}^d}\bigg(\frac{1}{2\pi}\bigg)^{d/2}\sum_{k=1}^K\pi_k\exp\bigg(-\frac{1}{2}\|x_{t-1}-\sqrt{\overline{\alpha}_{t-1}}\mu_k\|_2^2\bigg)\,\mathrm{d}u_t \\
&\quad\overset{\text{(ii)}}{=} p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}). \qquad (66)
\end{aligned}
$$

Here (i) is true as $u_t \mapsto \big(2\pi\alpha_t^2(1-\alpha_t)\big)^{-d/2}\exp\big(-\big(2\alpha_t^2(1-\alpha_t)\big)^{-1}\|u_t-\sqrt{\overline{\alpha}_t}(1-\alpha_t)\mu_k-\alpha_t^{3/2}x_{t-1}\|_2^2\big)$ is a density function and $\overline{\alpha}_t \coloneqq \prod_{i=1}^t\alpha_t$, and (ii) arises from (23). Hence, if we define function $\delta_{t-1}(x) : \mathbb{R}^d \to \mathbb{R}$ to capture the integral on set $\mathcal{E}_t^{\mathrm{c}}$ where

$$
\delta_{t-1}(x) \coloneqq \int_{\mathcal{E}_t^{\mathrm{c}}}\bigg(\frac{1}{4\pi^2\alpha_t^2(1-\alpha_t)}\bigg)^{d/2}\exp\bigg(-\frac{\|\sqrt{\alpha_t}x-u_t\|_2^2}{2\alpha_t(1-\alpha_t)}\bigg)\sum_{k=1}^K\pi_k\exp\bigg(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}\mu_k\|^2}{2\alpha_t^2}\bigg)\,\mathrm{d}u_t,
$$

then it obeys

$$
\delta_{t-1}(x) = p_{X_{t-1}^{\mathsf{GMM}}}(x) - \int_{\mathcal{E}_t}\bigg(\frac{1}{4\pi^2\alpha_t^2(1-\alpha_t)}\bigg)^{d/2}\exp\bigg(-\frac{\|\sqrt{\alpha_t}x-u_t\|_2^2}{2\alpha_t(1-\alpha_t)}\bigg)\sum_{k=1}^K\pi_k\exp\bigg(-\frac{\|u_t-\sqrt{\overline{\alpha}_t}\mu_k\|^2}{2\alpha_t^2}\bigg)\,\mathrm{d}u_t.
$$

Combining this definition with relation (65), we obtain

$$
\begin{aligned}
&p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) - \Delta_{t-1}(x_{t-1}) + \Delta_{t\to t-1}(x_{t-1}) \\
&\quad\ge \exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big)\big(p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) - \delta_{t-1}(x_{t-1})\big) \\
&\quad\ge p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) + O\big((1-\alpha_t)^2\log^2(KT)\big)p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) - O(1)\delta_{t-1}(x_{t-1}),
\end{aligned}
$$

or equivalently,

$$
\Delta_{t-1}(x_{t-1}) \le \Delta_{t\to t-1}(x_{t-1}) + O\big((1-\alpha_t)^2\log^2(KT)\big)p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1}) + O(1)\delta_{t-1}(x_{t-1}). \qquad (67)
$$

Here, we use (22) that $(1-\alpha_t)^2\log^2(KT) \lesssim \log^2(KT)\log^2 T/T^2 = o(1)$ as long as T is large enough.

We claim that $\int_{\mathbb{R}^d}\delta_{t-1}(x)\,\mathrm{d}x$ satisfies

$$
\int_{\mathbb{R}^d}\delta_{t-1}(x)\,\mathrm{d}x \lesssim T^{-3} + (1-\alpha_t)^2\log^2(KT). \qquad (68)
$$

Therefore, substituting (68) and (65) into (67) and integrating over x_{t-1} yields

$$
\int_{\mathbb{R}^d}\Delta_{t-1}(x_{t-1})\,\mathrm{d}x_{t-1} \le \int_{\mathbb{R}^d}\Delta_{t\to t-1}(x_{t-1})\,\mathrm{d}x_{t-1} + O\big((1-\alpha_t)^2\log^2(KT)\big)\int_{\mathbb{R}^d}p_{X_{t-1}^{\mathsf{GMM}}}(x_{t-1})\,\mathrm{d}x_{t-1}
$$

$$
\begin{aligned}
&\quad + O(1)\int_{\mathbb{R}^d}\delta_{t-1}(x_{t-1})\,\mathrm{d}x_{t-1}\\
&\le \int_{\mathbb{R}^d}\Delta_t(x_{t-1})\,\mathrm{d}x_{t-1} + O\big((1-\alpha_t)^2\log^2(KT)\big) + O\big(T^{-3}\big),
\end{aligned}
$$

where the penultimate line uses (60)

This completes the proof of Lemma 3.

Proof of Claim (68). It remains to control $\int_{\mathbb{R}^d}\delta_{t-1}(x)\,\mathrm{d}x$. To this end, the expression above allows us to derive

$$
\begin{aligned}
&\int_{\mathbb{R}^d}\delta_{t-1}(x_{t-1})\,\mathrm{d}x_{t-1}\\
&= 1 - \int_{\mathbb{R}^d}\int_{\mathcal{E}_t}\left(\frac{1}{4\pi^2\alpha_t^2(1-\alpha_t)}\right)^{d/2}\sum_{k=1}^{K}\pi_k\exp\left(-\frac{\big\|u_t-\sqrt{\alpha_t}\mu_k\big\|_2^2}{2\alpha_t^2}\right)\exp\left(-\frac{\|\sqrt{\alpha_t}x_{t-1}-u_t\|_2^2}{2\alpha_t(1-\alpha_t)}\right)\mathrm{d}u_t\,\mathrm{d}x_{t-1}\\
&\overset{\text{(i)}}{=} 1 - \int_{\mathbb{R}^d}\int_{\mathcal{E}_t}\exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big)p_{X_t^{\mathsf{GMM}}}(x_t)\left(\frac{1}{2\pi(1-\alpha_t)}\right)^{d/2}\exp\left(-\frac{\|\sqrt{\alpha_t}x_{t-1}-u_t\|_2^2}{2\alpha_t(1-\alpha_t)}\right)\mathrm{d}x_t\,\mathrm{d}x_{t-1}\\
&= 1 - \int_{\mathcal{E}_t}\exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big)p_{X_t^{\mathsf{GMM}}}(x_t)\int_{\mathbb{R}^d}\left(\frac{1}{2\pi(1-\alpha_t)}\right)^{d/2}\exp\left(-\frac{\|x_{t-1}-u_t/\sqrt{\alpha_t}\|_2^2}{2(1-\alpha_t)}\right)\mathrm{d}x_{t-1}\,\mathrm{d}x_t\\
&\overset{\text{(ii)}}{=} 1 - \int_{\mathcal{E}_t}\exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big)p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t\\
&\overset{\text{(iii)}}{\le} 1 - \int_{\mathcal{E}_t}p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t + O\big((1-\alpha_t)^2\log^2(KT)\big)\int_{\mathcal{E}_t}p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t\\
&\le \int_{\mathcal{E}_t^{\mathrm{c}}}p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t + O\big((1-\alpha_t)^2\log^2(KT)\big). \qquad (69)
\end{aligned}
$$

Here, (i) invokes Lemma 5, (ii) is true as $x_{t-1}\mapsto\big(2\pi(1-\alpha_t)\big)^{-d/2}\exp\big(-\big(2(1-\alpha_t)\big)^{-1}\|x_{t-1}-u_t/\sqrt{\alpha_t}\|_2^2\big)$ is a density function, and (iii) holds as $\exp(x)\ge 1+x$ for all $x\in\mathbb{R}^d$.

Finally, the right-hand-side of the above bound is controlled by Lemma 6 below.

Lemma 6. *Recall the definition of $\mathcal{E}_t$ in (26). For any $t\in[T]$, one has*

$$
\int_{\mathcal{E}_t^{\mathrm{c}}}p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t \lesssim T^{-3}. \qquad (70)
$$

Proof. See Appendix A.5. $\square$

Putting everything together completes the proof of Claim (68).

A.3 Proof of Lemma 4

Recall the definition of $\varepsilon_{\mathsf{score}}^{\mathsf{GMM}}$ and $\varepsilon_{\mathsf{score},t}^{\mathsf{GMM}}$ in (50). For some sufficiently large absolute constants $C_7>0$, we define the sets

$$
\mathcal{A}_t \coloneqq \Big\{x\in\mathbb{R}^d\colon\ \log p_{X_t^{\mathsf{GMM}}}(x)\ge -C_7 d\log(dT),\ \|x\|_2\le\sqrt{2T^{2c_R}+16d\log T}\Big\}, \qquad (71)
$$

$$
\mathcal{B}_t \coloneqq \big\{x\in\mathbb{R}^d\colon p_{X_t^{\mathsf{GMM}}}(x)\le 2\,p_{X_t}(x)\big\}, \qquad (72)
$$

for each $t\in[T]$. We collect several important properties of the set $\mathcal{A}_t$ in the following lemma. The proof can be found in Appendix A.6.

Lemma 7. *The set $\mathcal{A}_t$ satisfies*

$$
\mathbb{P}\big\{X_t^{\mathsf{GMM}}\notin\mathcal{A}_t\big\}\lesssim\exp(-2d\log T), \qquad (73)
$$

and on the set $\mathcal{A}_t$, the score function of X_t^{GMM} satisfies

$$\left\|s_t^{\star\mathsf{GMM}}(x)\right\|_2 \le C_{\mathsf{clip}}\sqrt{\frac{d\log(dT)}{1-\overline{\alpha}_t}} \tag{74}$$

where C_{clip} is defined in (13). In addition, one has

$$\mathbb{E}\Big[\left\|s_t^{\star\mathsf{GMM}}(X_t^{\mathsf{GMM}})\right\|_2^4\Big] \lesssim \left(\frac{d}{1-\alpha_t}\right)^2. \tag{75}$$

Now, we proceed to control $\varepsilon_{\mathsf{score}}^{\mathsf{GMM}}$. Recall that $\widehat{s}_t = \mathsf{clip}\{s_t\}$ defined in (13). Fixing an arbitrary $t \in [T]$, we first decompose

$$\begin{aligned}
\left(\varepsilon_{\mathsf{score},t}^{\mathsf{GMM}}\right)^2 &= \int_{\mathbb{R}^d} \left\|\mathsf{clip}\{s_t(x_t)\} - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t \\
&= \underbrace{\int_{\mathcal{A}_t\cap\mathcal{B}_t} \left\|\mathsf{clip}\{s_t(x_t)\} - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t}_{=:(\mathrm{I})} + \underbrace{\int_{\mathcal{A}_t\cap\mathcal{B}_t^{\mathrm{c}}} \left\|\mathsf{clip}\{s_t(x_t)\} - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t}_{=:(\mathrm{II})} \\
&\quad + \underbrace{\int_{\mathcal{A}_t^{\mathrm{c}}} \left\|\mathsf{clip}\{s_t(x_t)\} - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t}_{=:(\mathrm{III})}.
\end{aligned}$$

In what follows, we will bound these three quantities individually.

Controlling (I). Notice that on the set $\mathcal{A}_t$, the score function of X_t^{GMM} satisfies $\mathsf{clip}\{s_t^{\star\mathsf{GMM}}(x)\} = s_t^{\star\mathsf{GMM}}(x)$. Thus, we can use the Cauchy-Schwartz inequality to derive

$$\begin{aligned}
(\mathrm{I}) &\le \int_{\mathcal{A}_t\cap\mathcal{B}_t} \left\|s_t(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t \\
&\le 2\int_{\mathcal{A}_t\cap\mathcal{B}_t} \left\|s_t(x_t) - s_t^{\star}(x_t)\right\|_2^2 p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t + 2\int_{\mathcal{A}_t\cap\mathcal{B}_t} \left\|s_t^{\star}(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 p_{X_t^{\mathsf{GMM}}}(x_t)\,\mathrm{d}x_t \\
&\le 4\int_{\mathbb{R}^d} \left\|s_t(x_t) - s_t^{\star}(x_t)\right\|_2^2 p_{X_t}(x_t)\,\mathrm{d}x_t + 4\int_{\mathbb{R}^d} \left\|s_t^{\star}(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 \left\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\right\}\mathrm{d}x_t \\
&\asymp \varepsilon_{\mathsf{score},t}^2 + \int_{\mathbb{R}^d} \left\|s_t^{\star}(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 \left\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\right\}\mathrm{d}x_t,
\end{aligned}$$

where the last step uses Assumption 2. Next, we claim that

$$\int_{\mathbb{R}^d} \|s_t^{\star}(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\|_2^2 \{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\}\,\mathrm{d}x_t \lesssim \frac{\varepsilon_{\mathsf{apprx}} d\log T}{1-\overline{\alpha}_t} + \frac{1}{(1-\overline{\alpha}_t)T^d}, \tag{76}$$

with the proof deferred to the end of this section. As a result, one finds

$$(\mathrm{I}) \lesssim \varepsilon_{\mathsf{score},t}^2 + \frac{\varepsilon_{\mathsf{apprx}} d\log T}{1-\overline{\alpha}_t} + \frac{1}{(1-\overline{\alpha}_t)T^d}. \tag{77}$$

Controlling (II). By our construction of the truncation function $\mathsf{clip}\{\cdot\}$ in (13) and the set $\mathcal{A}_t$, one has

$$\left\|\mathsf{clip}\{s_t(x_t)\} - s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 \lesssim \left\|\mathsf{clip}\{s_t(x_t)\}\right\|_2^2 + \left\|s_t^{\star\mathsf{GMM}}(x_t)\right\|_2^2 \lesssim \frac{d\log(dT)}{1-\overline{\alpha}_t}, \tag{78}$$

where the last step follows from (74) in Lemma 7. In addition, on the set $\mathcal{B}_t^{\mathrm{c}} = \{2\,p_{X_t}(x) < p_{X_t^{\mathsf{GMM}}}(x)\}$, one has

$$p_{X_t^{\mathsf{GMM}}}(x_t) \le 2\left|p_{X_t^{\mathsf{GMM}}}(x_t) - p_{X_t}(x_t)\right|.$$

It follows that

$$\begin{aligned}\mathbb{P}\{X_t^{\mathsf{GMM}} \notin \mathcal{B}_t\} &= \int_{\mathbb{R}^d} p_{X_t^{\mathsf{GMM}}}(x_t)\mathbb{1}\{2\,p_{X_t}(x_t) < p_{X_t^{\mathsf{GMM}}}(x_t)\}\,\mathrm{d}x_t \\ &\leq 2\int_{\mathbb{R}^d} \big|p_{X_t^{\mathsf{GMM}}}(x_t) - p_{X_t}(x_t)\big|\,\mathrm{d}x_t \\ &= 4\,\mathsf{TV}(p_{X_t^{\mathsf{GMM}}}, p_{X_t}) \overset{\text{(i)}}{\lesssim} \mathsf{TV}(p_{X_0^{\mathsf{GMM}}}, p_{X_0}) \overset{\text{(ii)}}{=} \varepsilon_{\mathsf{apprx}},\end{aligned} \tag{79}$$

where (i) holds due to the data processing inequality, and (ii) arises from Assumption 1.

Collecting (78) and (79) together, we arrive at

$$\text{(II)} \lesssim \frac{\varepsilon_{\mathsf{apprx}} d\log(dT)}{1-\overline{\alpha}_t}. \tag{80}$$

Controlling (III). Using the definition of $\mathsf{clip}\{\cdot\}$ again, we apply the Cauchy-Schwartz inequality to obtain

$$\begin{aligned}\text{(III)} &\lesssim \frac{d\log(dT)}{1-\overline{\alpha}_t}\mathbb{P}\{X_t^{\mathsf{GMM}} \notin \mathcal{A}_t\} + \sqrt{\mathbb{E}\Big[\big\|s_t^{\star\mathsf{GMM}}(X_t^{\mathsf{GMM}})\big\|_2^4\Big]\mathbb{P}\{X_t^{\mathsf{GMM}} \notin \mathcal{A}_t\}} \\ &\overset{\text{(i)}}{\lesssim} \frac{d\log(dT)}{1-\overline{\alpha}_t}\mathbb{P}\{X_t^{\mathsf{GMM}} \notin \mathcal{A}_t\} + \frac{d}{1-\overline{\alpha}_t}\sqrt{\mathbb{P}\{X_t^{\mathsf{GMM}} \notin \mathcal{A}_t\}} \\ &\overset{\text{(ii)}}{\lesssim} \frac{d\log(dT)}{(1-\overline{\alpha}_t)}\exp(-2d\log T) + \frac{d}{(1-\overline{\alpha}_t)}\exp(-d\log T) \\ &\lesssim \frac{1}{(1-\overline{\alpha}_t)T}.\end{aligned} \tag{81}$$

where (i) invokes (75) from Lemma 7 and (ii) arises from (73) in Lemma 7.

Putting (I)–(III) together. Putting (77), (80), and (81) together, we obtain

$$(1-\overline{\alpha}_t)\big(\varepsilon_{\mathsf{score},t}^{\mathsf{GMM}}\big)^2 \lesssim (1-\overline{\alpha}_t)\varepsilon_{\mathsf{score},t}^2 + \varepsilon_{\mathsf{apprx}} d\log(dT) + T^{-1}.$$

Then the claim (54) in Lemma 4 immediately follows from the definition of $\varepsilon_{\mathsf{score}}^{\mathsf{GMM}}$ in (50).

Proof of Claim (76). Define the set

$$\mathcal{H}_x := \Big\{x_0 \in \mathbb{R}^d \colon \big\|x - \sqrt{\overline{\alpha}_t}x_0\big\|_2 \leq 3\sqrt{d(1-\overline{\alpha}_t)\log T}\Big\}.$$

Using Tweedie's formula, we can bound the discrepancy between the score functions of the target distribution and its GMM approximation as follows:

$$\begin{aligned}\big\|s_t^\star(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\big\|_2 &= \frac{1}{1-\overline{\alpha}_t}\Big\|\mathbb{E}\big[\sqrt{\overline{\alpha}_t}X_0 - X_t \mid X_t = x_t\big] - \mathbb{E}\big[\sqrt{\overline{\alpha}_t}X_0^{\mathsf{GMM}} - X_t^{\mathsf{GMM}} \mid X_t^{\mathsf{GMM}} = x_t\big]\Big\|_2 \\ &\leq \frac{1}{1-\overline{\alpha}_t}\int_{\mathbb{R}^d}\big|p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0\,|\,x_t) - p_{X_0|X_t}(x_0\,|\,x_t)\big|\,\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2\,\mathrm{d}x_0 \\ &\leq 2\,\mathsf{TV}(p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}=x_t}, p_{X_0|X_t=x_t})\sqrt{\frac{d\log T}{1-\overline{\alpha}_t}} \\ &\quad + \frac{1}{1-\overline{\alpha}_t}\int_{x_0\in\mathcal{H}_{x_t}^{\mathrm{c}}}\big(p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0\mid x_t) + p_{X_0|X_t}(x_0\mid x_t)\big)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2\,\mathrm{d}x_0,\end{aligned}$$

where the last uses the definition of $\mathcal{H}_x$ and the TV distance. By the Cauchy-Schwartz inequality, we can bound

$$\bigg(\int_{\mathcal{H}_{x_t}^{\mathrm{c}}} p_{X_0|X_t}(x_0\mid x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2\,\mathrm{d}x_0\bigg)^2 \leq \mathbb{P}\{X_0 \in \mathcal{H}_x^{\mathrm{c}} \mid X_t = x_t\}\int_{\mathcal{H}_{x_t}^{\mathrm{c}}} p_{X_0|X_t}(x_0\mid x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\,\mathrm{d}x_0$$

$$\leq \int_{\mathcal{H}_{x_t}^{\mathrm{c}}} p_{X_0|X_t}(x_0 \mid x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\,\mathrm{d}x_0.$$

Similarly, one also has

$$\bigg(\int_{\mathcal{H}_{x_t}^{\mathrm{c}}} p_{X_0|X_t}(x_0 \mid x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2\,\mathrm{d}x_0\bigg)^2 \leq \int_{\mathcal{H}_{x_t}^{\mathrm{c}}} p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0 \mid x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\,\mathrm{d}x_0.$$

As a result, the quantity we aim to bound can be controlled via

$$\begin{aligned}
&\int_{\mathbb{R}^d} \|s_t^\star(x_t) - s_t^{\star\mathsf{GMM}}(x_t)\|_2^2\left\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\right\}\mathrm{d}x_t\\
&\quad\lesssim \underbrace{\frac{d\log T}{1-\overline{\alpha}_t}\int_{\mathbb{R}^d}\mathsf{TV}(p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}=x_t}, p_{X_0|X_t=x_t})\left\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\right\}\mathrm{d}x_t}_{=:(\mathrm{IV})}\\
&\qquad+\underbrace{\frac{1}{(1-\overline{\alpha}_t)^2}\int_{x_t\in\mathbb{R}^d}\int_{x_0\in\mathcal{H}_{x_t}^{\mathrm{c}}}\big(p_{X_0^{\mathsf{GMM}},X_t^{\mathsf{GMM}}}(x_0,x_t)+p_{X_0,X_t}(x_0,x_t)\big)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\,\mathrm{d}x_0\,\mathrm{d}x_t}_{=:(\mathrm{V})}.
\end{aligned}\tag{82}$$

where we use the fact that $\mathsf{TV}(p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}=x_t}, p_{X_0|X_t=x_t}) \in [0,1]$.

To bound (IV) in (82), notice that

$$\begin{aligned}
&\big|p_{X_0|X_t}(x_0 \mid x_t) - p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0 \mid x_t)\big|\big\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\big\}\\
&\quad\leq p_{X_0|X_t}(x_0 \mid x_t)\cdot\big|\big\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\big\} - p_{X_t}(x_t)\big|\\
&\qquad+\big|p_{X_0|X_t}(x_0 \mid x_t)p_{X_t}(x_t) - p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0 \mid x_t)p_{X_t^{\mathsf{GMM}}}(x_t)\big|\\
&\qquad+p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0 \mid x_t)\cdot\big|p_{X_t^{\mathsf{GMM}}}(x_t) - \big\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\big\}\big|\\
&\quad\leq\big(p_{X_0|X_t}(x_0 \mid x_t) + p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0 \mid x_t)\big)\big|p_{X_t}(x_t) - p_{X_t^{\mathsf{GMM}}}(x_t)\big| + \big|p_{X_0,X_t}(x_0,x_t) - p_{X_0^{\mathsf{GMM}},X_t^{\mathsf{GMM}}}(x_0,x_t)\big|.
\end{aligned}$$

Consequently, we can bound

$$\begin{aligned}
&\int_{\mathbb{R}^d}\mathsf{TV}\big(p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}=x_t}, p_{X_0|X_t=x_t}\big)\big\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\big\}\mathrm{d}x_t\\
&\quad=\frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\big|p_{X_0|X_t}(x_0 \mid x_t) - p_{X_0^{\mathsf{GMM}}|X_t^{\mathsf{GMM}}}(x_0 \mid x_t)\big|\big\{p_{X_t}(x_t)\wedge p_{X_t^{\mathsf{GMM}}}(x_t)\big\}\mathrm{d}x_0\,\mathrm{d}x_t\\
&\quad\leq\int_{\mathbb{R}^d}\big|p_{X_t}(x_t) - p_{X_t^{\mathsf{GMM}}}(x_t)\big|\,\mathrm{d}x_t + \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\big|p_{X_0,X_t}(x_0,x_t) - p_{X_0^{\mathsf{GMM}},X_t^{\mathsf{GMM}}}(x_0,x_t)\big|\,\mathrm{d}x_0\,\mathrm{d}x_t\\
&\quad\leq 2\mathsf{TV}(p_{X_0^{\mathsf{GMM}}}, p_{X_0}) + \mathsf{TV}(p_{X_0^{\mathsf{GMM}},X_t^{\mathsf{GMM}}}, p_{X_0,X_t}) \leq 3\mathsf{TV}(p_{X_0^{\mathsf{GMM}}}, p_{X_0}) = 3\varepsilon_{\mathsf{apprx}},
\end{aligned}$$

where the last line follows from the data processing inequality and Assumption 1. Thus, we arrive

$$(\mathrm{IV}) \lesssim \frac{\varepsilon_{\mathsf{apprx}}d\log T}{1-\overline{\alpha}_t}.\tag{83}$$

Regarding (V) in (82), since $X_t \mid X_0 = x_0 \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I_d)$, we can derive

$$\begin{aligned}
&\int_{x_t\in\mathbb{R}^d}\int_{x_0\in\mathcal{H}_{x_t}^{\mathrm{c}}} p_{X_0,X_t}(x_0,x_t)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\,\mathrm{d}x_0\,\mathrm{d}x_t\\
&\quad=\int_{x_0\in\mathbb{R}^d} p_{X_0}(x_0)\int_{x_t\in\mathbb{R}^d} p_{X_t|X_0}(x_t \mid x_0)\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2^2\mathbb{1}\big\{\big\|x_t-\sqrt{\overline{\alpha}_t}x_0\big\|_2 > 3\sqrt{d(1-\overline{\alpha}_t)\log T}\big\}\,\mathrm{d}x_t\,\mathrm{d}x_0\\
&\quad\leq(1-\overline{\alpha}_t)\mathbb{E}\big[\|Z\|_2^2\mathbb{1}\big\{\|Z\|_2 > 3\sqrt{d\log T}\big\}\big]
\end{aligned}$$

where $Z \sim \mathcal{N}(0, I_d)$ is standard Gaussian random vector in $\mathbb{R}^d$. It follows that

$$\mathbb{E}\big[\|Z\|_2^2\mathbb{1}\big\{\|Z\|_2 > 3\sqrt{d\log T}\big\}\big] = \int_0^\infty \mathbb{P}\Big\{\|Z\|_2^2\mathbb{1}\big\{\|Z\|_2^2 > 9d\log T\big\} > x\Big\}\,\mathrm{d}x$$

$$
\begin{aligned}
&= \int_{9d\log T}^{\infty} \mathbb{P}\big\{\|Z\|_2^2 > x\big\}\,\mathrm{d}x \\
&\overset{\text{(i)}}{\le} \int_{9d\log T}^{\infty} \mathbb{P}\big\{\|Z\|_2^2 > 2d + x/2\big\}\,\mathrm{d}x \\
&\overset{\text{(ii)}}{\le} \int_{9d\log T}^{\infty} \exp\big(-x/6\big)\,\mathrm{d}x \\
&= 6\exp\big(-(3/2)d\log T\big) \lesssim T^{-d},
\end{aligned}
$$

where (i) holds as long as $9\log T > 4$; (ii) applies the Gaussian concentration inequality Laurent and Massart (2000, Lemma 1):

$$
\mathbb{P}\big\{\|Z\|_2^2 > 2d + 3x\big\} \le \mathbb{P}\big\{\|Z\|_2^2 - d > 2\sqrt{dx} + 2x\big\} \le \exp(-x), \qquad \forall x > 0. \tag{84}
$$

Clearly, the above bound is also valid for $\int_{x_t\in\mathbb{R}^d}\int_{x_0\in\mathcal{H}_{x_t}^{\mathrm{c}}} p_{X_0^{\mathsf{GMM}},X_t^{\mathsf{GMM}}}(x_0,x_t)\big\|x_t - \sqrt{\overline{\alpha}_t}x_0\big\|_2^2\,\mathrm{d}x_0\,\mathrm{d}x_t$. Hence, we conclude

$$
(\mathrm{V}) \lesssim \frac{1}{(1-\overline{\alpha}_t)T^d}. \tag{85}
$$

Plugging (83) and (85) into (82) completes the proof of Claim (76).

A.4 Proof of Lemma 5

Let us first derive two relations that are key for this proof. To start with, fix an arbitrary $x_t \in \mathcal{E}_t$. Recalling the definition that $u_t \coloneqq x_t + (1-\alpha_t)s_t^\star(x_t)$, direct calculations yield

$$
\begin{aligned}
&\frac{1}{2\alpha_t^2}\big\|u_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 \\
&\quad= \frac{1}{2}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 + \frac{1-\alpha_t^2}{2\alpha_t^2}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 + \frac{(1-\alpha_t)}{\alpha_t^2}s_t^\star(x_t)^\top(x_t - \sqrt{\overline{\alpha}_t}\mu_k) + \frac{(1-\alpha_t)^2}{2\alpha_t^2}\big\|s_t^\star(x_t)\big\|_2^2 \\
&\quad= \frac{1}{2}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 + \frac{1-\alpha_t^2}{2\alpha_t^2}\sum_{i=1}^K \pi_i^{(t)}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2 + \left(\frac{(1-\alpha_t)^2}{2\alpha_t^2} - \frac{1-\alpha_t}{\alpha_t^2}\right)\big\|s_t^\star(x_t)\big\|_2^2 \\
&\qquad+ \frac{1-\alpha_t^2}{2\alpha_t^2}\left(\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 - \sum_{i=1}^K \pi_i^{(t)}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2\right) + \frac{1-\alpha_t}{\alpha_t^2}s_t^\star(x_t)^\top\big(s_t^\star(x_t) + x_t - \sqrt{\overline{\alpha}_t}\mu_k\big) \\
&\quad\overset{\text{(i)}}{=} \frac{1}{2}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 + \frac{1-\alpha_t^2}{2\alpha_t^2}\sum_{i=1}^K \pi_i^{(t)}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2 - \frac{1-\alpha_t^2}{2\alpha_t^2}\big\|s_t^\star(x_t)\big\|_2^2 + \zeta_k^{(t)}(x_t) \\
&\quad\overset{\text{(ii)}}{=} \frac{1}{2}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 + \frac{1-\alpha_t^2}{2\alpha_t^2}\left(\sum_{i=1}^K \pi_i^{(t)}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2 - \left\|\sum_{i=1}^K \pi_i^{(t)}\big(x_t - \sqrt{\overline{\alpha}_t}\mu_i\big)\right\|_2^2\right) + \zeta_k^{(t)}(x_t) \\
&\quad\overset{\text{(iii)}}{=} \frac{1}{2}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 + \frac{(1-\alpha_t^2)}{2\alpha_t^2}\mathsf{tr}\big(I_d + J_t(x_t)\big) + \zeta_k^{(t)}(x_t).
\end{aligned} \tag{86}
$$

Here, (i) uses the definition of $\zeta_k^{(t)}(x)$ in (27); (ii) arises from the expression of $s_t^\star(x) = -\sum_{k=1}^K \pi_k^{(t)}\big(x - \sqrt{\overline{\alpha}_t}\mu_k\big)$ in (24); (iii) uses the expression of $J_t(x)$ in (19) that

$$
\begin{aligned}
I_d + J_t(x) &= \sum_{k=1}^K \pi_k^{(t)}\left(\sqrt{\overline{\alpha}_t}\mu_k - \sum_{i=1}^K \pi_i^{(t)}\sqrt{\overline{\alpha}_t}\mu_i\right)\left(\sqrt{\overline{\alpha}_t}\mu_k - \sum_{i=1}^K \pi_i^{(t)}\sqrt{\overline{\alpha}_t}\mu_i\right)^\top \\
&= \sum_{k=1}^K \pi_k^{(t)}\left(x - \sqrt{\overline{\alpha}_t}\mu_k - \sum_{i=1}^K \pi_i^{(t)}\big(x - \sqrt{\overline{\alpha}_t}\mu_i\big)\right)\left(x - \sqrt{\overline{\alpha}_t}\mu_k - \sum_{i=1}^K \pi_i^{(t)}\big(x - \sqrt{\overline{\alpha}_t}\mu_i\big)\right)^\top
\end{aligned}
$$

$$= \sum_{k=1}^{K} \pi_k^{(t)} \big(x - \sqrt{\overline{\alpha}_t}\mu_k\big)\big(x - \sqrt{\overline{\alpha}_t}\mu_k\big)^\top - \bigg(\sum_{k=1}^{K} \pi_k^{(t)} \big(x - \sqrt{\overline{\alpha}_t}\mu_k\big)\bigg)\bigg(\sum_{k=1}^{K} \pi_k^{(t)} \big(x - \sqrt{\overline{\alpha}_t}\mu_k\big)\bigg)^\top.$$

In addition, recall the definition of $\mathcal{E}_t$ (cf. (26)). For any $x_t \in \mathcal{E}_t$, using (22) that $1-\alpha_t \lesssim \log T/T$, we know that

$$\begin{aligned}
\frac{1-\alpha_t}{\alpha_t}\mathsf{tr}\big(I_d + J_t(x_t)\big) &\lesssim (1-\alpha_t)\mathsf{tr}\big(I_d + J_t(x_t)\big) \lesssim \frac{\log(KT)}{T} = o(1),\\
\frac{1-\alpha_t^2}{2\alpha_t^2}\mathsf{tr}\big(I_d + J_t(x_t)\big) &\lesssim (1-\alpha_t)\mathsf{tr}\big(I_d + J_t(x_t)\big) \lesssim \frac{\log(KT)}{T} = o(1),
\end{aligned}$$

for large enough T. It follows that

$$\begin{aligned}
&\det\Big(I_d + \big(\alpha_t^{-1} - 1\big)\big(I_d + J_t(x_t)\big)\Big)\\
&\quad\overset{\text{(i)}}{=} 1 + \frac{1-\alpha_t}{\alpha_t}\mathsf{tr}\big(I_d + J_t(x_t)\big) + O\bigg(\frac{(1-\alpha_t)^2}{\alpha_t^2}\mathsf{tr}^2\big(I_d + J_t(x_t)\big)\bigg)\\
&\quad= 1 + \frac{1-\alpha_t^2}{2\alpha_t^2}\mathsf{tr}\big(I_d + J_t(x_t)\big) - \frac{(1-\alpha_t)^2}{2\alpha_t^2}\mathsf{tr}\big(I_d + J_t(x_t)\big) + O\bigg(\frac{(1-\alpha_t)^2}{\alpha_t^2}\mathsf{tr}^2\big(I_d + J_t(x_t)\big)\bigg)\\
&\quad\overset{\text{(ii)}}{=} 1 + \frac{1-\alpha_t^2}{2\alpha_t^2}\mathsf{tr}\big(I_d + J_t(x_t)\big) + O\big((1-\alpha_t)^2\log^2(KT)\big)\\
&\quad= \exp\bigg(\frac{1-\alpha_t^2}{2\alpha_t^2}\mathsf{tr}\big(I_d + J_t(x_t)\big) + O\big((1-\alpha_t)^2\log^2(KT)\big)\bigg).
\end{aligned}$$

where (i) holds as $I_d + J_t(x_t) \succeq 0$ and $\det(I + \varepsilon A) = 1 + \mathsf{tr}(A)\varepsilon + O\big(\varepsilon^2(\mathsf{tr}^2(A) - \mathsf{tr}(A^2))\big)$ for any matrix A and $\varepsilon > 0$; (ii) is true since $\alpha_t \gtrsim 1$ by (22) and $\mathsf{tr}\big(I_d + J_t(x_t)\big) \lesssim \log(KT)$ by the choice of $\mathcal{E}_t$ in (26). Consequently, we can derive

$$\begin{aligned}
\det\big(I_d + (1-\alpha_t)J_t(x_t)\big) &= \det\Big(\alpha_t I_d + (1-\alpha_t)\big(I_d + J_t(x_t)\big)\Big)\\
&= \alpha_t^d \det\Big(I_d + (\alpha_t^{-1} - 1)\big(I_d + J_t(x_t)\big)\Big)\\
&= \alpha_t^d \exp\bigg(\frac{1-\alpha_t^2}{2\alpha_t^2}\mathsf{tr}\big(I_d + J_t(x_t)\big) + O\big((1-\alpha_t)^2\log^2(KT)\big)\bigg), \qquad (87)
\end{aligned}$$

As a consequence of the above two relations, we move on to prove Lemma 5. In view of relation (86), we arrive at

$$\begin{aligned}
&\bigg(\frac{1}{2\pi\alpha_t^2}\bigg)^{d/2}\sum_{k=1}^{K}\pi_k \exp\Big(-\frac{1}{2\alpha_t^2}\big\|u_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2\Big)\\
&\quad= \bigg(\frac{1}{2\pi\alpha_t^2}\bigg)^{d/2}\exp\bigg(-\frac{(1-\alpha_t^2)}{2\alpha_t^2}\mathsf{tr}\big(I + J_t(x_t)\big)\bigg)\sum_{k=1}^{K}\pi_k \exp\bigg(-\frac{1}{2}\big\|x_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2\bigg)\exp\big(-\zeta_k^{(t)}(x_t)\big)\\
&\quad= \det\big(I + (1-\alpha_t)J_t(x_t)\big)^{-1}\exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big)p_{X_t^{\mathsf{GMM}}}(x_t)\sum_{k=1}^{K}\pi_k^t \exp\big(-\zeta_k^{(t)}(x_t)\big)
\end{aligned}$$

where the last equality uses (87) and $\pi_k \exp\big(-\|x - \sqrt{\overline{\alpha}_t}\mu_k\|_2^2/2\big) = \pi_k^{(t)}(2\pi)^{d/2}p_{X_t^{\mathsf{GMM}}}(x)$ due to (23) and (20). To further control the right hand side, by the definition of $\mathcal{E}_t$ in (26), it satisfies that

$$1 \le \sum_{k=1}^{K}\pi_k^t \exp\big(-\zeta_k^{(t)}(x_t)\big) \le \exp\big(C_2(1-\alpha_t)^2\log^2(KT)\big).$$

Therefore, we can conclude that

$$\bigg(\frac{1}{2\pi\alpha_t^2}\bigg)^{d/2}\sum_{k=1}^{K}\pi_k \exp\Big(-\frac{1}{2\alpha_t^2}\big\|u_t - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2\Big)$$

$$= \det\big(I + (1-\alpha_t)J_t(x_t)\big)^{-1}\exp\Big(O\big((1-\alpha_t)^2\log^2(KT)\big)\Big)p_{X_t^{\mathsf{GMM}}}(x_t),$$

which completes the proof of Lemma 5.

A.5 Proof of Lemma 6

Recalling the definition of $\mathcal{E}_t$ in expression (26), we have

$$\mathbb{P}\big\{X_t^{\mathsf{GMM}} \in \mathcal{E}_t^c\big\} \le \mathbb{P}\Big\{\mathsf{tr}\big(I_d + J_t(X_t^{\mathsf{GMM}})\big) \ge C_1\log(KT)\Big\} + \mathbb{P}\bigg\{\sum_{k=1}^K \pi_k^{(t)}\exp\big(-\zeta_k^{(t)}(X_t^{\mathsf{GMM}})\big) < 1\bigg\}$$
$$+ \mathbb{P}\bigg\{\sum_{k=1}^K \pi_k^{(t)}\exp\big(-\zeta_k^{(t)}(X_t^{\mathsf{GMM}})\big) > \exp\big(C_2(1-\alpha_t)^2\log^2(KT)\big)\bigg\}. \tag{88}$$

In the following, we bound the three terms on the right respectively.

Before proceeding, we make the following observation. Fix an arbitrary $t \ge 1$. For each $k \in [K]$, we define the event

$$\mathcal{T}_k := \Big\{x \in \mathbb{R}^d : \big|(x - \sqrt{\overline{\alpha}_t}\mu_k)^\top\sqrt{\overline{\alpha}_t}(\mu_i - \mu_k)\big| \le C_5\sqrt{\overline{\alpha}_t\log(KT)}\,\|\mu_i - \mu_k\|_2 \text{ for all } i \in [K]\Big\} \tag{89}$$

for some absolute constant $C_5 > 0$. Note that if we let $Z_k \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}\mu_k, I_d)$ be a Gaussian random vector in $\mathbb{R}^d$, which implies that $\big(Z_k - \sqrt{\overline{\alpha}_t}\mu_k\big)^\top\sqrt{\overline{\alpha}_t}(\mu_i - \mu_k) \sim \mathcal{N}\big(0, \overline{\alpha}_t\|\mu_i - \mu_k\|_2^2\big)$, the standard Gaussian concentration inequality guarantees that

$$\mathbb{P}\big\{Z_k \notin \mathcal{T}_k\big\} \lesssim T^{-3}, \tag{90}$$

provided C_5 is large enough.

Bounding the first term in Eq. (88)

Let us begin with the first event $\{\mathsf{tr}\big(I_d + J_t(X_t^{\mathsf{GMM}})\big) \le C_1\log(KT)\}$. As $X_t^{\mathsf{GMM}} \sim \sum_{k=1}^K \pi_k\mathcal{N}(\sqrt{\overline{\alpha}_t}\mu_k, I_d)$, it is easily seen that

$$\begin{aligned}
&\mathbb{P}\Big\{\mathsf{tr}\big(I_d + J_t(X_t^{\mathsf{GMM}})\big) > C_1\log(KT)\Big\} \\
&\quad= \sum_{k=1}^K \pi_k\mathbb{P}\Big\{\mathsf{tr}\big(I_d + J_t(Z_k)\big) > C_1\log(KT)\Big\} \\
&\quad\le \sum_{k=1}^k \pi_k\mathbb{P}\Big\{\mathsf{tr}\big(I_d + J_t(Z_k)\big) > C_1\log(KT)\Big\}\mathbb{1}\big\{\pi_k \ge 1/(KT^3)\big\} + \sum_{k=1}^K \pi_k\mathbb{1}\big\{\pi_k < 1/(KT^3)\big\} \\
&\quad\le \sum_{k=1}^k \pi_k\mathbb{P}\Big\{\mathsf{tr}\big(I_d + J_t(Z_k)\big) > C_1\log(KT)\Big\}\mathbb{1}\big\{\pi_k \ge 1/(KT^3)\big\} + T^{-3}.
\end{aligned}$$

We claim that for any $k \in [K]$ such that $\pi_k \ge 1/(KT^3)$, one has

$$\mathcal{T}_k \subset \Big\{x \in \mathbb{R}^d \colon \mathsf{tr}\big(I_d + J_t(x)\big) \le C_1\log(KT)\Big\}. \tag{91}$$

It then immediately follows from (90) that

$$\begin{aligned}
\mathbb{P}\Big\{\mathsf{tr}\big(I_d + J_t(X_t^{\mathsf{GMM}})\big) > C_1\log(KT)\Big\} &\le \sum_{k=1}^K \pi_k\mathbb{P}\big\{Z_k \notin \mathcal{T}_k\big\}\mathbb{1}\big\{\pi_k \ge 1/(KT^3)\big\} + T^{-3} \\
&\lesssim T^{-3}\sum_{k=1}^K \pi_k + T^{-3} \asymp T^{-3}.
\end{aligned} \tag{92}$$

Proof of Claim (91). Towards this, fix an arbitrary $k \in [K]$ such that $\pi_k \geq 1/(KT^3)$. For any $x \in \mathcal{T}_k$, we know that for all $i \in [K]$,

$$\begin{aligned}
\pi_i^{(t)} &\leq \frac{\pi_i}{\pi_k} \exp\Big(-\frac{1}{2}\big\|x - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2 + \frac{1}{2}\big\|x - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2\Big) \wedge 1 \\
&= \frac{\pi_i}{\pi_k} \exp\Big(-\frac{1}{2}\overline{\alpha}_t\|\mu_i - \mu_k\|_2^2 + (x - \sqrt{\overline{\alpha}_t}\mu_k)^\top \sqrt{\overline{\alpha}_t}(\mu_i - \mu_k)\Big) \wedge 1 \\
&\leq \exp\Big(-\frac{1}{2}\overline{\alpha}_t\|\mu_i - \mu_k\|_2^2 + C_5\sqrt{\overline{\alpha}_t \log(KT)}\,\|\mu_i - \mu_k\|_2 + 3\log(KT)\Big) \wedge 1,
\end{aligned}$$

where the last line holds due to the definition of $\mathcal{T}_k$. As a result, for any $i \in [K]$ satisfying $\sqrt{\overline{\alpha}_t}\|\mu_i - \mu_k\|_2 > 6C_5\sqrt{\log(KT)}$, one has

$$\pi_i^{(t)} \leq \exp\Big(-\frac{1}{6}\overline{\alpha}_t\|\mu_i - \mu_k\|_2^2\Big)$$

as long as $C_5 \geq \sqrt{2}/2$. This further implies that

$$\begin{aligned}
\pi_i^{(t)}\overline{\alpha}_t\|\mu_i - \mu_k\|_2^2 &\leq \overline{\alpha}_t\|\mu_i - \mu_k\|_2^2 \exp\Big(-\frac{1}{6}\overline{\alpha}_t\|\mu_i - \mu_k\|_2^2\Big) \\
&\leq \exp\Big(-\frac{1}{12}\overline{\alpha}_t\|\mu_i - \mu_k\|_2^2\Big) \leq \exp\big(-3C_5^2\log(KT)\big),
\end{aligned} \tag{93}$$

provided T is large enough. Meanwhile, for any $i \in [K]$ obeying $\sqrt{\overline{\alpha}_t}\|\mu_i - \mu_k\|_2 \leq 6C_5\sqrt{\log(KT)}$, we can simply upper bound

$$\pi_i^{(t)}\overline{\alpha}_t\|\mu_i - \mu_k\|_2^2 \leq \pi_i^{(t)} \cdot 36C_5^2\log(KT). \tag{94}$$

Denote the set $\mathcal{F}_k \coloneqq \big\{i \in [K] \colon \sqrt{\overline{\alpha}_t}\|\mu_i - \mu_k\|_2 \leq 6C_5\sqrt{\log(KT)}\big\}$. Using the expression of J_t (cf. (19)), we conclude that

$$\begin{aligned}
\mathsf{tr}\big(I_d + J_t(x)\big) &= \sum_{i=1}^K \pi_i^{(t)}\overline{\alpha}_t\Big\|\mu_i - \sum_{k=1}^K \pi_k^{(t)}\mu_i\Big\|_2^2 \\
&\overset{\text{(i)}}{\leq} \sum_{i=1}^K \pi_i^{(t)}\overline{\alpha}_t\big\|\mu_i - \mu_k\big\|_2^2 \\
&\overset{\text{(ii)}}{\leq} 36C_5^2\log(KT)\sum_{i\in\mathcal{F}_k}\pi_i^{(t)} + \sum_{i\in\mathcal{F}_k^{\mathrm{c}}}\exp\big(-3C_5^2\log(KT)\big) \\
&\leq 36C_5^2\log(KT)\log(KT) + K\exp\big(-3C_5^2\log(KT)\big) \\
&\leq C_1\log(KT)
\end{aligned} \tag{95}$$

provided C_5 and C_1/C_5^2 are large enough. Here, (i) is true since $\sum_{k=1}^K \pi_k^{(t)}\mu_k$ is the minimizer of the function $x \mapsto \sum_{i=1}^K \pi_i^{(t)}\|\mu_i - x\|_2^2$ and (ii) uses (93)–(94). This establishes the claim (91).

Bounding the remaining terms in Eq. (88)

Next, let analyze the second event $\big\{1 \leq \sum_{k=1}^K \pi_k^{(t)}\exp\big(-\zeta_k^{(t)}(X_t^{\mathsf{GMM}})\big) \leq \exp\big(C_2(1-\alpha_t)^2\log^2(KT)\big)\big\}$. We first establish the lower bound of 1. For any $x \in \mathbb{R}^d$, given $\sum_k \pi_k^{(t)} = 1$, direct calculation shows that

$$\begin{aligned}
&\sum_{k=1}^K \pi_k^{(t)}\bigg(\big\|x - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 - \sum_{i=1}^K \pi_i^{(t)}\big\|x - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2\bigg) + \sum_{k=1}^K \pi_k^{(t)}\sum_{i=1}^K \pi_i^{(t)}(\mu_i - \mu_k) \\
&\qquad = \sum_{k=1}^K \pi_k^{(t)}\big\|x - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 - \sum_{i=1}^K \pi_i^{(t)}\big\|x - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2 + \sum_{i=1}^K \pi_i^{(t)}\mu_i - \sum_{k=1}^K \pi_k^{(t)}\mu_k = 0.
\end{aligned}$$

Combined with the definition of $\zeta_k^{(t)}(x)$ in (27), this yields

$$\sum_{k=1}^{K} \pi_k^{(t)} \zeta_k^{(t)}(x) = 0.$$

We can then apply Jensen's inequality to obtain that for any $x \in \mathbb{R}^d$,

$$\sum_{k=1}^{K} \pi_k^{(t)} \exp\Big(-\zeta_k^{(t)}(x)\Big) \geq \exp\Big(-\sum_{k=1}^{K} \pi_k^{(t)} \zeta_k^{(t)}(x)\Big) = 1. \tag{96}$$

Recall the definition of $\mathcal{T}_k$ in expression (89). To bound the second term in Eq. (88), it suffices to prove that for any $k \in [K]$ such that $\pi_k \geq 1/(KT^3)$,

$$\mathcal{T}_k \subset \Big\{x \in \mathbb{R}^d \colon \sum_{k=1}^{K} \pi_k^{(t)} \exp\Big(-\zeta_k^{(t)}(x)\Big) \leq \exp\big(C_2(1-\alpha_t)^2 \log^2(KT)\big)\Big\}. \tag{97}$$

Indeed, assuming (97) holds, one can apply the same reasoning as that for (92) to obtain

$$\begin{aligned} &\mathbb{P}\Big\{\sum_{k=1}^{K} \pi_k^{(t)} \exp\Big(-\zeta_k^{(t)}\big(X_t^{\mathsf{GMM}}\big)\Big) > \exp\big(C_2(1-\alpha_t)^2 \log^2(KT)\big)\Big\} \\ &\qquad \leq \sum_{k=1}^{K} \pi_k \mathbb{P}\big\{Z_k \notin \mathcal{T}_k\big\} \mathbb{1}\big\{\pi_k \geq 1/(KT^3)\big\} + T^{-3} \lesssim T^{-3}, \end{aligned} \tag{98}$$

Taking this collectively with relations (92), and (88) completes the proof of Lemma 6. Now it is only left for us to prove inequality (97).

Proof of inequality (97). To this end, recall the definitions of $\zeta_k^{(t)}(x)$ and $s_t^\star(x)$ in (27) and (24), respectively. By some basic algebra, $\zeta_k^{(t)}(x)$ can be written as

$$\begin{aligned} \zeta_k^{(t)}(x) &= \frac{1-\alpha_t^2}{2\alpha_t^2} \sum_{i=1}^{K} \pi_i^{(t)} \Big(\big\|x - \sqrt{\overline{\alpha}_t}\mu_k\big\|_2^2 - \big\|x - \sqrt{\overline{\alpha}_t}\mu_i\big\|_2^2\Big) \\ &\quad + \frac{1-\alpha_t}{\alpha_t^2} \Big(-x + \sum_{i=1}^{K} \pi_i^{(t)} \sqrt{\overline{\alpha}_t}\mu_i\Big)^\top \Big(\sum_{i=1}^{K} \pi_i^{(t)} \sqrt{\overline{\alpha}_t}(\mu_i - \mu_k)\Big) \\ &= \frac{1-\alpha_t^2}{2\alpha_t^2} \sum_{i=1}^{K} \pi_i^{(t)} \Big(-\frac{1}{2}\overline{\alpha}_t \|\mu_i - \mu_k\|_2^2 + \big(x - \sqrt{\overline{\alpha}_t}\mu_k\big)^\top \sqrt{\overline{\alpha}_t}(\mu_i - \mu_k)\Big) \\ &\quad - \frac{1-\alpha_t}{\alpha_t^2} \sum_{i=1}^{K} \pi_i^{(t)} \big(x - \sqrt{\overline{\alpha}_t}\mu_k\big)^\top \sqrt{\overline{\alpha}_t}(\mu_i - \mu_k) + \frac{1-\alpha_t}{\alpha_t^2} \Big\|\sum_{i=1}^{K} \pi_i^{(t)} \sqrt{\overline{\alpha}_t}(\mu_i - \mu_k)\Big\|_2^2. \end{aligned}$$

For any $x \in \mathcal{T}_k$, one can obtain

$$\begin{aligned} \big|\zeta_k^{(t)}(x)\big| &\lesssim (1-\alpha_t) \sum_{i=1}^{K} \pi_i^{(t)} \Big(-\frac{1}{2}\overline{\alpha}_t \|\mu_i - \mu_k\|_2^2 + C_5 \sqrt{\overline{\alpha}_t \log(KT)}\, \|\mu_i - \mu_k\|_2\Big) \\ &\qquad + (1-\alpha_t)\sqrt{\log(KT)} \sum_{i=1}^{K} \pi_i^{(t)} \sqrt{\overline{\alpha}_t}\, \|\mu_i - \mu_k\|_2 + (1-\alpha_t) \sum_{i=1}^{K} \pi_i^{(t)} \overline{\alpha}_t \big\|\mu_i - \mu_k\big\|_2^2 \\ &\asymp (1-\alpha_t) \sum_{i=1}^{K} \pi_i^{(t)} \overline{\alpha}_t \big\|\mu_i - \mu_k\big\|_2^2 + (1-\alpha_t)\sqrt{\log(KT)} \sum_{i=1}^{K} \pi_i^{(t)} \sqrt{\overline{\alpha}_t}\, \|\mu_i - \mu_k\|_2. \end{aligned} \tag{99}$$

where the first inequality holds due to $1-\alpha_t \lesssim \log T/T$ in (22), the definition of $\mathcal{T}_k$ in (89), and Jensen's inequality. Using the same argument as that for (95), it can be easily seen that

$$\sum_{i=1}^{K} \pi_i^{(t)} \sqrt{\overline{\alpha}_t}\|\mu_i - \mu_k\|_2 \lesssim \sqrt{\log(KT)}. \tag{100}$$

Plugging (100) and (95) into (99) demonstrates that

$$\left|\zeta_k^{(t)}(x)\right| \lesssim (1-\alpha_t)\log(KT) = o(1), \tag{101}$$

since $1-\alpha_t \lesssim \log T/T$ as in (22).

As a consequence, for any $x \in \mathcal{T}_k$, we find that

$$\begin{aligned}
\sum_{k=1}^{K} \pi_k^{(t)} \exp\Big(-\zeta_k^{(t)}(x)\Big) &= \sum_{k=1}^{K} \pi_k^{(t)}\Big(1-\zeta_k^{(t)}(x)+\frac{1}{2}\big(\zeta_k^{(t)}(x)\big)^2 + o\Big(\big(\zeta_k^{(t)}(x)\big)^2\Big)\Big) \\
&= 1+\frac{1}{2}\sum_{k=1}^{K}\pi_k^{(t)}\big(\zeta_k^{(t)}(x)\big)^2 + \sum_{k=1}^{K}\pi_k^{(t)} o\Big(\big(\zeta_k^{(t)}(x)\big)^2\Big) \\
&= 1 + O\big((1-\alpha_t)^2\log^2(KT)\big) \\
&\le \exp\big(C_2(1-\alpha_t)^2\log^2(KT)\big)
\end{aligned}$$

as long as C_2 is sufficiently large. This establishes the claim (97), thereby leads to (98).

A.6 Proof of Lemma 7

Proof of Claim (73). In light of the definition of $\mathcal{A}_t$ in (71), one has

$$\begin{aligned}
\mathbb{P}\big\{X_t^{\mathsf{GMM}} \notin \mathcal{A}_t\big\} = &\underbrace{\int_{\mathbb{R}^d} p_{X_t^{\mathsf{GMM}}}(x)\mathbb{1}\Big\{\log p_{X_t^{\mathsf{GMM}}}(x) < -C_7 d\log(dT),\ \|x\|_2 \le \sqrt{2T^{2c_R}+16d\log T}\Big\}\,\mathrm{d}x}_{:=(\mathrm{I})} \\
&+ \underbrace{\int_{\mathbb{R}^d} p_{X_t^{\mathsf{GMM}}}(x)\mathbb{1}\Big\{\|x\|_2 > \sqrt{2T^{2c_R}+16d\log T}\Big\}\,\mathrm{d}x}_{:=(\mathrm{II})}.
\end{aligned}$$

Thus, it suffices to control these two terms individually.

For term (I), we can bound

$$\begin{aligned}
(\mathrm{I}) &\le \int_{\mathbb{R}^d} \exp\big(-C_7 d\log(dT)\big)\mathbb{1}\Big\{\|x\|_2 \le \sqrt{2T^{2c_R}+16d\log T}\Big\}\,\mathrm{d}x \\
&\le \exp\big(-C_7 d\log(dT)\big)2^d\big(2T^{2c_R}+16d\log T\big)^{d/2} \\
&\le \exp\big(-C_7 d\log(dT)\big)2^{3d/2}\big(2^{d/2}T^{c_R d}+16^{d/2}(d\log T)^{d/2}\big) \\
&\le \exp\big(-2d\log(dT)\big).
\end{aligned}$$

where the penultimate line holds as $(x+y)^{d/2} \le 2^{d/2}(x^{d/2}+y^{d/2})$ for any $d\ge 1$ and $x,y>0$; the last step is valid as long as C_7/c_R is large enough.

Regarding term (II), since $X_t^{\mathsf{GMM}} \sim \pi_k \sum_{k=1}^{K}\mathcal{N}(\sqrt{\overline{\alpha}_t}\mu_k, I_d)$, one can derive

$$\begin{aligned}
(\mathrm{II}) &= \sum_{k\in[K]} \pi_k \mathbb{P}\Big\{\big\|\sqrt{\overline{\alpha}_t}\mu_k + Z\big\|_2^2 > 2T^{2c_R}+16d\log T\Big\} \\
&\overset{(\mathrm{i})}{\le} \sum_{k\in[K]} \pi_k \mathbb{P}\Big\{\overline{\alpha}_t\|\mu_k\|_2^2 + \|Z\|_2^2 > T^{2c_R}+8d\log T\Big\}
\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \sum_{k\in[K]} \pi_k \mathbb{P}\big\{\|Z\|_2^2 > 2d + 6d\log T\big\}$$
$$\overset{\text{(iii)}}{\lesssim} \sum_{k\in[K]} \pi_k \exp\big(-2d\log T\big)$$
$$\leq \exp\big(-2d\log T\big).$$

where $Z \sim \mathcal{N}(0, I_d)$ is a standard Gaussian random vector in $\mathbb{R}^d$. Here, (i) applies the Cauchy-Schwartz inequality; (ii) holds due to $\overline{\alpha}_t \in (0,1)$ and Assumption 1 that $\max_k \|\mu_k\|_2 \leq T^{c_R}$; (iii) applies the Gaussian concentration inequality (84).

Putting the above bounds together yields Claim (73).

Proof of Claim (74). As $X_t^{\mathsf{GMM}} \mid X_0^{\mathsf{GMM}} = x_0 \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I_d)$, we can use Tweedie's formula to derive (see also Saha and Guntuboyina (2020, Lemma 4.3) and Jiang and Zhang (2009, Theorem 3)):

$$\begin{aligned}
\big\|s_t^{\star\mathsf{GMM}}(x)\big\|_2^2 &= \bigg\|\frac{1}{1-\overline{\alpha}_t}\mathbb{E}\Big[\sqrt{\overline{\alpha}_t}X_0^{\mathsf{GMM}} - x \mid X_t^{\mathsf{GMM}} = x\Big]\bigg\|_2^2 \\
&\leq \frac{1}{(1-\overline{\alpha}_t)^2}\mathbb{E}\Big[\big\|\sqrt{\overline{\alpha}_t}X_0^{\mathsf{GMM}} - x\big\|_2^2 \mid X_t^{\mathsf{GMM}} = x\Big] \\
&\overset{\text{(i)}}{\leq} \frac{2}{1-\overline{\alpha}_t}\log\mathbb{E}\bigg[\exp\Big(\frac{1}{2(1-\overline{\alpha}_t)}\big\|\sqrt{\overline{\alpha}_t}X_0^{\mathsf{GMM}} - x\big\|_2^2\Big) \mid X_t^{\mathsf{GMM}} = x\bigg] \\
&\overset{\text{(ii)}}{=} -\frac{2}{1-\overline{\alpha}_t}\log\Big(\big(2\pi(1-\overline{\alpha}_t)\big)^{d/2} p_{X_t^{\mathsf{GMM}}}(x)\Big).
\end{aligned} \tag{102}$$

Here, (i) applies Jensen's inequality; (ii) is true because

$$p_{\sqrt{\overline{\alpha}_t}X_0^{\mathsf{GMM}}\mid X_t^{\mathsf{GMM}}}(y \mid x) = \big(2\pi(1-\overline{\alpha})\big)^{-d/2}\exp\bigg(-\frac{1}{2(1-\overline{\alpha}_t)}\|y-x\|_2^2\bigg)\frac{p_{X_0^{\mathsf{GMM}}}(y)}{p_{X_t^{\mathsf{GMM}}}(x)},$$

which further leads to

$$\begin{aligned}
\mathbb{E}\bigg[\exp\Big(\frac{1}{2(1-\overline{\alpha}_t)}\big\|\sqrt{\overline{\alpha}_t}X_0^{\mathsf{GMM}} - x\big\|_2^2\Big) \mid X_t^{\mathsf{GMM}} = x\bigg] &= \int_{y\in\mathbb{R}^d}\exp\bigg(\frac{1}{2(1-\overline{\alpha}_t)}\|y-x\|_2^2\bigg)p_{\sqrt{\overline{\alpha}_t}X_0^{\mathsf{GMM}}\mid X_t^{\mathsf{GMM}}}(y\mid x)\,\mathrm{d}y \\
&= \big(2\pi(1-\overline{\alpha}_t)\big)^{-d/2}\frac{1}{p_{X_t^{\mathsf{GMM}}}(x)}.
\end{aligned}$$

Therefore, as $\log p_{X_t^{\mathsf{GMM}}}(x) \geq -C_7 d\log T$ on the set $\mathcal{A}_t$, we arrive at the claimed bound:

$$\begin{aligned}
\big\|s_t^{\star\mathsf{GMM}}(x)\big\|_2^2 &\leq \frac{d}{1-\overline{\alpha}_t}\log\frac{1}{2\pi(1-\overline{\alpha}_t)} - \frac{2}{1-\overline{\alpha}_t}\log p_{X_t^{\mathsf{GMM}}}(x) \\
&\lesssim \frac{d}{1-\overline{\alpha}_t}\log\frac{c_1 T}{\pi\log T} + \frac{2}{1-\overline{\alpha}_t}\log\big(C_7 d\log T\big) \\
&\leq C_{\mathsf{clip}}\frac{d\log T}{1-\overline{\alpha}_t}
\end{aligned} \tag{103}$$

where the penultimate step holds since $1-\overline{\alpha}_t \geq c_1(T/\log T)(1-\alpha_t) \geq (c_1/2)T/\log T$ due to (14) and (22) from Lemma 1; the last line holds provided C_{clip} is large enough.

Proof of Claim (75). Applying Tweedie's formula again, we can derive

$$\begin{aligned}
\mathbb{E}\Big[\big\|s_t^{\star\mathsf{GMM}}\big(X_t^{\mathsf{GMM}}\big)\big\|_2^4\Big] &= \frac{1}{(1-\overline{\alpha}_t)^4}\mathbb{E}\Big[\big\|\mathbb{E}\big[X_t^{\mathsf{GMM}} - \sqrt{\overline{\alpha}_t}\,X_0^{\mathsf{GMM}} \mid X_t^{\mathsf{GMM}}\big]\big\|_2^4\Big] \\
&\overset{\text{(i)}}{\leq} \frac{1}{(1-\overline{\alpha}_t)^4}\mathbb{E}\Big[\mathbb{E}\big[\big\|X_t^{\mathsf{GMM}} - \sqrt{\overline{\alpha}_t}\,X_0^{\mathsf{GMM}}\big\|_2^4 \mid X_t^{\mathsf{GMM}}\big]\Big]
\end{aligned}$$

$$\overset{\text{(ii)}}{=} \frac{1}{(1-\overline{\alpha}_t)^4}\mathbb{E}\Big[\big\|X_t^{\mathsf{GMM}} - \sqrt{\overline{\alpha}_t}\,X_0^{\mathsf{GMM}}\big\|_2^4\Big]$$
$$\overset{\text{(iii)}}{\lesssim} \left(\frac{d}{1-\overline{\alpha}_t}\right)^2.$$

Here, (i) applies the convexity of the function $x \mapsto \|x\|_2^4$ and Jensen's inequality; (ii) uses the tower property; and (iii) leverages the properties of the standard Gaussian distribution.