Unraveling the Smoothness Properties of Diffusion Models: A Gaussian Mixture Perspective

Yingyu Liang[*] Zhizhou Sha[†] Zhenmei Shi[‡] Zhao Song[§] Mingda Wan[¶] Yufa Zhou[‖]

Abstract

Diffusion models have made rapid progress in generating high-quality samples across various domains. However, a theoretical understanding of the Lipschitz continuity and second momentum properties of the diffusion process is still lacking. In this paper, we bridge this gap by providing a detailed examination of these smoothness properties for the case where the target data distribution is a mixture of Gaussians, which serves as a universal approximator for smooth densities such as image data. We prove that if the target distribution is a k-mixture of Gaussians, the density of the entire diffusion process will also be a k-mixture of Gaussians. We then derive tight upper bounds on the Lipschitz constant and second momentum that are independent of the number of mixture components k. Finally, we apply our analysis to various diffusion solvers, both SDE and ODE based, to establish concrete error guarantees in terms of the total variation distance and KL divergence between the target and learned distributions. Furthermore, our preliminary experiments support our theoretical analysis. Our results provide deeper theoretical insights into the dynamics of the diffusion process under common data distributions.

1. Introduction

Diffusion models, a prominent generative modeling framework, have rapid progress and garnered significant attention in recent years, due to their potential powerful ability to generate high-quality samples across various domains and diverse applications. Score-based generative diffusion models [30, 63] can generate high-quality image samples comparable to GANs, which require adversarial optimization. Based on the U-Net [50], stable diffusion [49], a conditional multi-modality generation model, can successfully generate business-used images. Based on the Transformer (DiT) [45], OpenAI released a video diffusion model, SORA [44], with a surprising performance.

However, despite these technological strides, a critical gap persists in the theoretical understanding of diffusion processes, especially concerning the Lipschitz continuity and second momentum properties of these models. Many existing studies ([10, 13, 14, 21, 35] and many more) make simplifying assumptions about these key smoothness properties but lack rigorous formulation or comprehensive analysis. This paper aims to bridge this theoretical gap by providing a detailed examination of the Lipschitz and second momentum characteristics. To make the data distribution analyzable, we consider the mixture of Gaussian data distribution (see illustration in Figure 1), which is a universal approximation (Fact 1.1) for any smooth density, such as complex image and video data distributions.

Fact 1.1. *A Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific nonzero amount of error by a Gaussian mixture model with enough components [52].*

Furthermore, we prove that if the target data distribution is a k-mixture of Gaussian, the density of the whole diffusion process will be a k-mixture of Gaussian (Lemma 3.2). Thus, our focus is studying a mixture of Gaussian target data distribution in the diffusion process, where we can gain a concrete Lipschitz constant and second momentum (Lemma 3.5 and Lemma 3.6). Moreover, we explore the implications of these properties through the lens of various solvers, both Stochastic Differential Equation (SDE) and Ordinary Differential Equation (ODE) based, providing a deeper insight into the dynamic behavior of diffusion processes and concrete guarantees in Table 1. Moreover, our preliminary experiments support our theoretical analysis.

Our main contribution are summarized as follows:

- As the Gaussian mixture model is a universal approximator of densities (Fact 1.1), we assume the target/image data distribution as a k-mixture of Gaussian. Then, we show that the density of the whole diffusion process is a k-mixture of Gaussian (Lemma 3.2).
- We analyze the Lipschitz and second momentum of k-

[*] yingyul@hku.hk. The University of Hong Kong. yliang@cs.wisc.edu. University of Wisconsin-Madison.
[†] shazz20@mails.tsinghua.edu.cn. Tsinghua University.
[‡] zhmeishi@cs.wisc.edu. University of Wisconsin-Madison.
[§] magic.linuxkde@gmail.com. The Simons Institute for the Theory of Computing at the University of California, Berkeley.
[¶] dylan.r.mathison@gmail.com. Anhui University.
[‖] yufazhou@seas.upenn.edu. University of Pennsylvania.

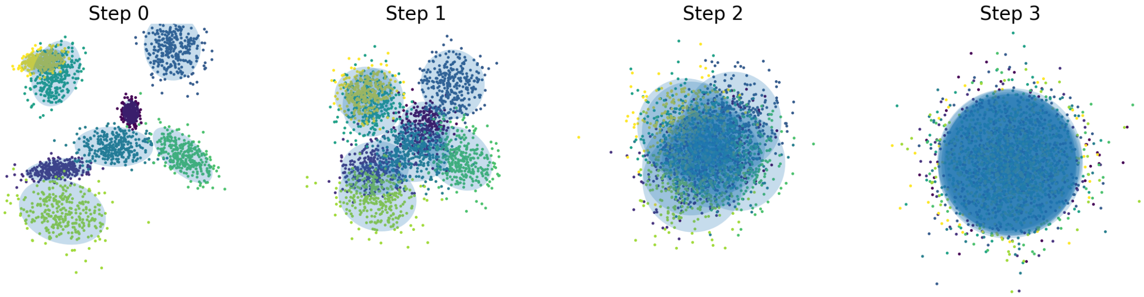

Figure 1. An illustration of discrete diffusion process for 8 mixture of Gaussian as shown in Eq. (2). The left figure represents the target 2-dimensional data distribution p_0. The right figure represents the standard normal distribution $p_T = \mathcal{N}(0, I_{2\times 2})$, where $T = 3$. This experimental setup is also used in [13].

mixture of Gaussian data distribution and provide a tight upper bound, which is independent of k, (Lemma 3.5 and Lemma 3.6), even when k goes to infinity (see more discussion in Section 3.1).

- After applying our analysis to DDPM, which is the SDE version of reverse process, we prove the dynamic of the diffusion process satisfies some concrete total variation bound (Theorem 5.4) and concrete KL divergence bound (Theorem 5.5 and Theorem 5.6) under choosing some total discretization steps. See a summary in Table 1.
- After applying our analysis to DPOM and DPUM, which is the ODE version of the reverse process, we prove the dynamic of the diffusion process satisfies some concrete total variation bound (Theorem 5.7 and Theorem 5.8) under choosing some total discretization steps.

Other studies of the mixture of Gaussian under diffusion. There is a rich line of work that studies a mixture of Gaussian data distribution under the diffusion process, which shares a similar popular setting. However, none focus on the Lipschitz smoothness and second momentum property as ours. We provide a short summary here for convenience. [71] analyses the effect of diffusion guidance and provides theoretical insights on the instillation of task-specific information into the score function and conditional generation, e.g., text-to-image, by studying the mixture of Gaussian data distribution as a case study. [27] propose a new SDE-solver for diffusion models under a mixture of Gaussian data. [26, 53] learn mixtures of Gaussian data using the DDPM objective and gives bound for sample complexity by assuming all covariance matrices are identity but does not use Lipschitz, while our work does not need identity assumptions. [16] mainly focuses on solving k-mixture of Gaussian in diffusion model by leveraging the property of the mixture of Gaussian and polynomial approximation methods and gives bound for sample complexity by assuming the covariance matrices have bounded condition number and that the mean vectors and covariance matrices lie in a bounded ball.

2. Related Work

Diffusion models and score-based models. [30] introduced the concept of denoising diffusion probabilistic models (DDPM), which learn to reverse a gradual noising process to generate samples and have been applied to many area [68–70]. Then, [63] used Stochastic Differential Equations (SDE) to build the diffusion model and explored the equivalence between score-based generative models (SGMs) and diffusion models, generalizing the diffusion model to continuous time. There is a line of works [58] studying the connection between diffusion models and non-equilibrium thermodynamics, providing a theoretical foundation for these models. Another line of work has focused on improving the efficiency and quality of diffusion models [42]. Particularly, [59] leverages score matching to train diffusion models more efficiently; [43] improved architectures and training techniques for diffusion models. Diffusion models have been applied to various domains beyond image generation, achieving impressive results, e.g., text-to-image synthesis [12], audio synthesis [33], image super-resolution [51], and so on.

Mixture of Gaussian. Mixtures of Gaussian are among the most fundamental and widely used statistical models [1, 3–6, 8, 9, 19, 20, 22–25, 37, 41, 55, 56, 66, 72] and studied in neural networks learning [39, 47, 54, 57]. Recently, they have also been widely studied in diffusion models [16, 26, 27, 53, 71] (see detailed discussion in Section 1).

Lipschitz and second momentum in score estimation. There is a line of work using bounded Lipschitz and second momentum as their key smoothness assumptions for the whole diffusion process without giving any concrete formulation [7, 10, 11, 13–15, 32, 34, 35, 40, 74], while our work gives a “close-form” formulation of Lipschitz and second momentum. There is another rich line of work studying how to train the diffusion models to have a better theoretical guarantee [11, 14–16, 26–28, 31, 36, 40, 48, 53, 59–62, 64, 65, 73] and many more.

Type	Error guarantee	Steps for $\widetilde{O}(\epsilon_0^2)$ error	Reference
DDPM [13]	$\mathrm{TV}(p_0, \widehat{q})^2$	$\widetilde{\Theta}(d/\epsilon_0^2)$	Theorem 5.4
DDPM [10]	$\mathrm{KL}(p_0, \widehat{q})$	$\widetilde{\Theta}(d/\epsilon_0^2)$	Theorem 5.5
DDPM [10]	$\mathrm{KL}(p_0, \widehat{q})$	$\widetilde{\Theta}(d^2/\epsilon_0^2)$	Theorem 5.6
DPOM [14]	$\mathrm{TV}(p_0, \widehat{q})^2$	$\widetilde{\Theta}(d/\epsilon_0^2)$	Theorem 5.7
DPUM [14]	$\mathrm{TV}(p_0, \widehat{q})^2$	$\widetilde{\Theta}(\sqrt{d}/\epsilon_0)$	Theorem 5.8

Table 1. A summary of our applications using our Lipschitz and second momentum bound, when we assume $\sigma_{\min(p_t)}$ as a constant (defined in Condition 3.4). The third column means the number of discretization points required to guarantee a small total variation/KL divergence distance between the target data distribution p_0 and the learned distribution $\widehat{q}$.

Roadmap. Section 3 provides the notation we use, several definitions, and lemmas related to k mixtures of Gaussian. In Section 4, we provide preliminary knowledge on score-based generative models (SGMs) and diffusion models. Section 5 presents our main results. Section 6 provides some empirical evaluation that as evidence for our theoretical results. In Section 7, we conclude the paper.

3. Lipschitz and Second Momentum of Mixture of Gaussian

In this section, we discuss the notations used, several key definitions for k mixtures of Gaussian distributions, and lemmas concerning the Lipschitz continuity and second momentum of these mixtures. We begin by presenting the notations that are used throughout the paper.

Notations. For two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, we use $\langle x, y \rangle$ to denote the inner product between x, y, i.e., $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$. We use $\Pr[]$ to denote the probability, we use $\mathbb{E}[]$ to denote the expectation. We use e_i to denote a vector where only i-th coordinate is 1, and other entries are 0. For each $a, b \in \mathbb{R}^n$, we use $a \circ b \in \mathbb{R}^n$ to denote the vector where i-th entry is $(a \circ b)_i = a_i b_i$ for all $i \in [n]$, and this is the Hardamard product. We use $\mathbf{1}_n$ to denote a length-n vector where all the entries are ones. We use $x_{i,j}$ to denote the j-th coordinate of $x_i \in \mathbb{R}^n$. We use $\|x\|_p$ to denote the ℓ_p norm of a vector $x \in \mathbb{R}^n$, i.e. $\|x\|_1 := \sum_{i=1}^n |x_i|$, $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$, and $\|x\|_\infty := \max_{i \in [n]} |x_i|$. For $k > n$, for any matrix $A \in \mathbb{R}^{k \times n}$, we use $\|A\|$ to denote the spectral norm of A, i.e. $\|A\| := \sup_{x \in \mathbb{R}^n} \|Ax\|_2 / \|x\|_2$. We use $\sigma_{\min}(A), \sigma_{\max}(A)$ to denote the minimum/maximum singular value of matrix A. For a square matrix A, we use $\mathrm{tr}[A]$ to denote the trace of A, i.e., $\mathrm{tr}[A] = \sum_{i=1}^n A_{i,i}$. We use $\det(A)$ to denote the determinant of matrix A. We use $f * g$ to denote the convolution of 2 functions f, g. In addition to $O(\cdot)$ notation, for two functions f, g, we use the shorthand $f \lesssim g$ (resp. $\gtrsim$) to indicate that $f \leq Cg$ (resp. $\geq$) for an absolute constant C.

k mixtures of Gaussian. Now, we are ready to introduce k mixtures of Gaussian. Formally, we can have the following definition of the pdf for k mixtures of Gaussian.

Definition 3.1 (k mixtures of Gaussian pdf). *Let $x \in \mathbb{R}^d$, $i \in [k]$, $t \geq 0 \in \mathbb{R}$. For a fixed timestamp t, (1) let $\alpha_i(t) \in (0, 1)$ be the weight for the i-th Gaussian component at time t and $\sum_{i=1}^k \alpha_i(t) = 1$; (2) let $\mu_i(t) \in \mathbb{R}^d$ and $\Sigma_i(t) \in \mathbb{R}^{d \times d}$ be the mean vector and covariance matrix for the i-th Gaussian component at time t. Then, we define*

$$p_t(x) := \sum_{i=1}^k \frac{\alpha_i(t)}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1} (x - \mu_i(t))).$$

Then, the following lemma shows that the linear combination between k mixtures of Gaussian and a single standard Gaussian is still a k mixtures of Gaussian. The proof is in Appendix H.

Lemma 3.2 (Informal version of Lemma H.1). *Let $a, b \in \mathbb{R}$. Let $\mathcal{D}$ be a k-mixture of Gaussian distribution, and let p be its* pdf *defined in Definition 3.1, i.e.,*

$$p(x) := \sum_{i=1}^k \frac{\alpha_i}{(2\pi)^{d/2} \det(\Sigma_i)^{1/2}} \cdot \exp(-\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i))$$

Let $x \in \mathbb{R}^d$ sample from $\mathcal{D}$. Let $z \in \mathbb{R}^d$ and $z \sim \mathcal{N}(0, I)$, which is independent from x. Then, we have a new random variable $y = ax + bz$ which is also a k-mixture of Gaussian distribution $\widetilde{\mathcal{D}}$, whose pdf *is*

$$\widetilde{p}(x) := \sum_{i=1}^k \frac{\alpha_i}{(2\pi)^{d/2} \det(\widetilde{\Sigma}_i)^{1/2}} \cdot \exp(-\frac{1}{2}(x - \widetilde{\mu}_i)^\top \widetilde{\Sigma}_i^{-1} (x - \widetilde{\mu}_i)),$$

where $\widetilde{\mu}_i = a\mu_i, \widetilde{\Sigma}_i = a^2 \Sigma_i + b^2 I$.

Note that the Gaussian mixture model is a universal approximator of densities (Fact 1.1). Then, it is natural to assume that the target/image (p_0) data distribution as the k

mixtures of Gaussian. Lemma 3.2 tell us that if the target data distribution is k mixtures of Gaussian, then by Eq. (2), the pdf of the whole diffusion process is k mixtures of Gaussian, i.e., the p_t for any $t \in [0, T]$. See details in Section 4.

Main properties. Thus, we only need to analyze the property, i.e., Lipschitz constant and second momentum, of k mixtures of Gaussian, to have a clear guarantee for the whole diffusion process. In the following two lemmas, we will give concrete bounds of the Lipschitz constant and second momentum of k mixtures of Gaussian. Both proofs can be found in Appendix H.

First, we define the following conditions of the k mixtures of Gaussian.

Condition 3.3. *All conditions here are related to the beginning time density p_0 in Eq. (9).*

- *Let* $\sigma_{\min(p_0)} = \min_{i\in[k]}\{\sigma_{\min}(\Sigma_i)\}$ *and* $\sigma_{\max(p_0)} = \max_{i\in[k]}\{\sigma_{\max}(\Sigma_i)\}$.
- *Let* $\mu_{\max(p_0)} = \max_{i\in[k]}\{\|\mu_i\|_2^2\}$ *and* $\det_{\min(p_0)} = \min_{i\in[k]}\{\det(\Sigma_i)\}$.

Condition 3.3 denotes $\sigma_{\min(p_0)}, \sigma_{\max(p_0)}$ as the minimum/maximum singular value between the covariance matrices of k-components at the original distribution p_0. $\mu_{\max(p_0)}$ is the maximum ℓ_2^2-norm between mean vectors of the k-components at p_0. $\det_{\min(p_0)}$ is the minimum determinant between the covariance matrices of k-components at p_0.

Condition 3.4. *All conditions here are related to the all time density p_t in Eq. (10), where $t \in [0, T]$. Let $x \in \mathbb{R}^d$ and a_t, b_t is defined by Definition 4.1. Assume Assumption 5.1.*

- *Let $\beta \le \|x - a_t\mu_i\|_2 \le R$, where $R \ge 1$ and $\beta \in (0, 0.1)$, for each $i \in [k]$.*
- *Let $p_t(x)$ be defined as Eq. (10) and $p_t(x) \ge \gamma$, where $\gamma \in (0, 0.1)$.*
- *Let* $\sigma_{\min(p_t)} := \min_{i\in[k]}\{\sigma_{\min}(a_t^2\Sigma_i + b_t^2 I)\}$, $\sigma_{\max(p_t)} := \max_{i\in[k]}\{\sigma_{\max}(a_t^2\Sigma_i + b_t^2 I)\}$.
- *Let* $\det_{\min(p_t)} := \min_{i\in[k]}\{\det(a_t^2\Sigma_i + b_t^2 I)\}$.

In Condition 3.4, we denote $\sigma_{\min(p_t)}, \sigma_{\max(p_t)}$ as the minimum/maximum singular value between the covariance matrices of k-components at p_t. β, R be the lower/upper bound of the ℓ_2 distance between the input data x and all k scaled mean vectors $a_t\mu_i$ at timestep t. Let γ be the lower bound of $p_t(x)$, the pdf at time t and $\det_{\min(p_t)}$ be the minimum determinant at p_t. We assume β and γ are lower bounded/constant, implying that each Gaussian component in the mixture is well-conditioned. This is a standard assumption in diffusion theory and is consistent with other theoretical findings [16, 17, 26, 38, 53, 71].

Clearly, we have $\sigma_{\max(p_t)} \ge \sigma_{\max(p_0)}$, so Condition 3.4 is a stronger condition.

Lemma 3.5 (Lipschitz, informal version of Lemma H.3)**.** *Assume Condition 3.4. The Lipschitz constant for the score function $\frac{\mathrm{d}\log(p_t(x))}{\mathrm{d}x}$ is given by:*

$$L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2\sigma_{\min(p_t)}^2} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}}) \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2}\det_{\min(p_t)}^{1/2}}).$$

In Lemma 3.5, we can clearly see that roughly $L = O(1/\operatorname{poly}(\sigma_{\min(p_t)}))$, which means the Lipschitz is only conditioned on the smallest singular value of all Gaussian component but independent with k. We provide additional empirical study in Section 6 to support our analysis.

Lemma 3.6 (Second momentum, informal version of Lemma H.2)**.** *Let $x_0 \sim p_0$, where p_0 is defined by Eq. (9). Then, we have*

$$m_2^2 := \mathop{\mathbb{E}}_{x_0\sim p_0}[\|x_0\|_2^2] = \sum_{i=1}^k \alpha_i(\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]) \le \max_{i\in[k]}\{\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]\}.$$

The proof idea of Lemma 3.5 and Lemma 3.6 is that we first consider the case when $k = 1$ in Appendix D and then we extend to 2 mixtures of Gaussian setting in Appendix E. Finally, we can generalize to k setting in Appendix F and summarize it in Appendix H.

In Lemma 3.6, we can see that $m_2^2 = O(1)$ roughly, which is independent of k as well. Later, we will apply our Lemma 3.5 and Lemma 3.6 in Section 5 to get concrete bound for each diffusion process, e.g., DDPM, DPOM and DPUM.

3.1. Practical insights

We believe our results in Lemma 3.5 is non-trivial. First, the Lipschitz bound depends inversely exponential on the dimension d. If d grows larger, the Lipschitz constant L becomes much smaller. Second, in practice, the k will be super large for complicated data distribution, e.g., millions of Gaussian components for an image distribution. Many studies need to overcome the large k hardness in learning Gaussian mixture models [5, 8, 47, 57], while our Lipschitz upper-bound is independent with k. These theoretical insights can help practitioners make more informed decisions about model architecture, optimization strategies, and computational resources needed for different applications.

4. Score Based Model and Diffusion Model

In Section 4.1, we first briefly introduce DDPM [30], a stochastic differential equations (SDE) version of the reverse diffusion process, and score-based generative models

(SGMs) [63], which is a generalization of DDPM. Then, in Section 4.2, we introduce multiple solvers for the reverse process.

4.1. Background on score based model and diffusion model

First, we denote the input data as $x \in \mathbb{R}^d$ and the target original data distribution as $p_0(x)$. Assuming that the noisy latent state of score-based generative models (SGMs) at time t is a stochastic process x_t for $0 \leq t \leq T$, we have the forward SDE is defined by:

$$\mathrm{d}x_t = f(x_t, t)\mathrm{d}t + g(t)\mathrm{d}w_t, \;\; x_0 \sim p_0, \;\; 0 \leq t \leq T \quad (1)$$

where $w_t \in \mathbb{R}^d$ is the standard Brownian motion, $f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ which is called drift coefficient, and $g(t) : \mathbb{R} \to \mathbb{R}$ which is called diffusion coefficient. We use $p_t(x)$ to denote the marginal probability density function of x_t. The pdf at last time step p_T is the prior distribution which often defined as standard Gaussian $p_T(x) = \mathcal{N}(0, I)$.

The continuous forward SDE Eq. (1) also has the discrete Markov chain form as

$$x_t = a_t x_0 + b_t z, \;\; x_0 \sim p_0 \quad (2)$$

where $x_0 \sim p_0(x)$ and $z \sim \mathcal{N}(0, I)$, and $a_t, b_t \in \mathbb{R}$ are functions of time. Additionally, we assume as $t \to T$, $a_t \to 0$ and $b_t \to 1$. Thus, $x_T \sim \mathcal{N}(0, I)$. Also, clearly when $t \to 0$, $a_t \to 1$ and $b_t \to 0$ for this is the boundary condition. More specifically, the Eq. (2) can be viewed as the iterative equation in DDPM [30].

From [2], we know x_t also satisfy the reverse SDE:

$$\mathrm{d}x_t = \big(f(x_t, t) - g(t)^2 \nabla \log p_t(x_t)\big)\,\mathrm{d}t + g(t)\,\mathrm{d}\widetilde{w}_t, \quad (3)$$

where $\widetilde{w}_t \in \mathbb{R}^d$ is backward Brownian motion [2], and $\nabla \log p_t(x_t)$ is the score function. For convenience, we rewrite the reverse SDE Eq. (3) in a forward version by switching the time direction, replacing t with $T - t$. Let $\widetilde{x}_t := x_{T-t}$. The law of $(\widetilde{x}_t)_{0 \leq t \leq T}$ is identical to the law of $(x_{T-t})_{0 \leq t \leq T}$. We use q_t to denote the density of $\widetilde{x}_t$:

$$\begin{aligned}\mathrm{d}\widetilde{x}_t = {} & (-f(\widetilde{x}_t, T-t) + g(T-t)^2 \nabla \log p_{T-t}(\widetilde{x}_t))\mathrm{d}t \\ & + g(T-t)\mathrm{d}w_t. \end{aligned} \quad (4)$$

The process $(\widetilde{x}_t)_{0 \leq t \leq T}$ converts noise into samples from p_0, thereby achieving the objective of generative modeling.

Since we can not obtain the score function $\nabla \log p_t$ directly, we can use a neural network to approximate it, and we denote the estimated score function by $s_t(x)$. By replacing the score function $\nabla \log p_t$ with our approximated score function $s_t(x)$, we can rewrite Eq. (4) as:

$$\begin{aligned}\mathrm{d}y_t = {} & (-f(y_t, T-t) + g(T-t)^2 s_{T-t}(y_t))\mathrm{d}t \\ & + g(T-t)\,\mathrm{d}w_t, \;\; y_0 \sim p_T, \;\; 0 \leq t \leq T. \end{aligned} \quad (5)$$

where y_t is the process we approximate by our SGM s_t. For clarity, we mainly focus on the Ornstein-Uhlenbeck (OU) process, which is a diffusion process with $\exp$ coefficient:

Definition 4.1. *The forward SDE of OU process ($f(x_t, t) = -x_t$, and $g(t) \equiv \sqrt{2}$) is: $\mathrm{d}x_t = -x_t \mathrm{d}t + \sqrt{2}\mathrm{d}w_t$. The corresponding discrete Markov chain form of OU process is given by: $x_t = e^{-t}x_0 + \sqrt{1 - e^{-2t}}z$, where we can see that $a_t = e^{-t}$, $b_t = \sqrt{1 - e^{-2t}}$ in Eq. (2).*

From Eq. (5) under the OU process, we can have:

$$\mathrm{d}y_t = (x_t + 2s_{T-t}(y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}w_t. \quad (6)$$

4.2. Definitions of different solvers

In practical applications, it's necessary to adopt a discrete-time approximation for sampling dynamics. We have the following definition.

Definition 4.2 (Time discretization). *Define the N discretization points as $\delta = t_0 \leq t_1 \leq \cdots \leq t_N = T$, where $\delta \geq 0$ is the early stopping parameter, with $\delta = 0$ presenting the normal setting. For each discretization step k, where $1 \leq k \leq N$, the step size is denoted by $h_k := t_k - t_{k-1}$.*

Let $\widehat{y}_t$ be the discrete approximation of y_t in Eq. (6) starting from $\widehat{y}_0 \sim \mathcal{N}(0, I)$. We use $\widehat{q}_t$ to denote the density of $\widehat{y}_t$. Let N be defined in Definition 4.2 and $t'_k = T - t_{N-k}$. To solve Eq. (6), we have the following two numerical solvers:

Definition 4.3. *We define* $\mathsf{EulerMaruyama}$ *as the numerical solver satisfying $\widehat{y}_{T-\delta} = \mathsf{EulerMaruyama}(T, s)$, where $\widehat{y}_{T-\delta} \in \mathbb{R}^d$ is the output, $T \in \mathbb{R}^+$ is the total time, and s is the score estimates. The Euler-Maruyama scheme is, (Equation 5 in [10]), for t $\in [t'_k, t'_{k+1}]$*

$$\mathrm{d}\widehat{y}_t = (\widehat{y}_{t'_k} + 2s_{T-t'_k}(\widehat{y}_{t'_k}))\mathrm{d}t + \sqrt{2}\mathrm{d}w_t. \quad (7)$$

Definition 4.4. *We define* $\mathsf{ExponentialIntegrator}$ *as the numerical solver satisfying $\widehat{y}_{T-\delta} = \mathsf{ExponentialIntegrator}(T, s)$, where $\widehat{y}_{T-\delta} \in \mathbb{R}^d$ is the output, $T \in \mathbb{R}^+$ is the total time, and s is the score estimates. The Exponential Integrator scheme is, (Equation 6 in [10]), for t $\in [t'_k, t'_{k+1}]$*

$$\mathrm{d}\widehat{y}_t = (\widehat{y}_t + 2s_{T-t'_k}(\widehat{y}_{t'_k}))\mathrm{d}t + \sqrt{2}\mathrm{d}w_t. \quad (8)$$

The two methods above are used for solving SDE. The difference is that in the first term of RHS of Euler-Maruyama uses $\widehat{y}_{t'_k}$, while Exponential Integrator uses $\widehat{y}_t$. The Exponential Integrator scheme has a closed-form solution (see detail in Section 1.1 of [10]).

We now introduce two ordinary differential equation (ODE) solvers, DPOM and DPUM, which omit the Brownian motion term in Eq. (6). These solvers, defined in

Definitions 4.5 and 4.6, represent distinct approaches using predictor-corrector steps to address the reverse process via probability flow ODEs. Specifically, DPOM employs overdamped Langevin Monte Carlo (LMC), while DPUM uses underdamped LMC, enabling more efficient sampling. We will present their concrete step bounds in Section 5.

Definition 4.5. *We define* DPOM *(Diffusion Predictor with Overdamped Modeling) as Algorithm 1 in [14] satisfying* $\widehat{y}_{T-\delta} = \mathsf{DPOM}(T, h_{\mathrm{pred}}, h_{\mathrm{corr}}, s)$*, where* δ *is the early stopping parameter,* $\widehat{y}_{T-\delta} \in \mathbb{R}^d$ *is the output,* $T \in \mathbb{R}^+$ *is the total steps,* h_{pred} *is the predictor step size,* h_{corr} *is the corrector step size, (see detailed definition in Algorithm 1 of [14]), and* s *is the score estimates.*

Definition 4.6. *We define* DPUM *(Diffusion Predictor with Underdamped Modeling) as Algorithm 2 in [14], satisfying* $\widehat{y}_{T-\delta} = \mathsf{DPUM}(T, h_{\mathrm{pred}}, h_{\mathrm{corr}}, s)$*, where* δ *is the early stopping parameter,* $\widehat{y}_{T-\delta} \in \mathbb{R}^d$ *is the output,* $T \in \mathbb{R}^+$ *is the total steps,* h_{pred} *is the predictor step size,* h_{corr} *is the corrector step size, (see detailed definition in Algorithm 2 of [14]), and* s *is the score estimates.*

5. Main Result for Application

In this section, we will provide the main results of applications. In Section 5.1, we provide our key definitions and assumptions used. In Section 5.2, we provide our results for the total variation bound. In Section 5.3, we provide our results for the KL divergence bound. In Section 5.4, we provide our results for the probability flow ODE method, including DPOM and DPUM.

5.1. Key definitions and assumptions

We first assume that the loss of the learned score estimator is upper bounded by ϵ_0^2 in Assumption 5.1. Then, we can show that we can recover the target/image data distribution under a small total variation or KL divergence gap later.

Assumption 5.1 (Score estimation error, Assumption 1 in [10], page 6, and Assumption 3 in [13], page 6)**.** *The learned score function* $s_t(x)$ *satisfies for any* $1 \le k \le N$*,*

$$\frac{1}{T} \sum_{k=1}^{N} h_k \mathop{\mathbb{E}}_{p_{t_k}} [\|\nabla \log p_{t_k}(x) - s_{t_k}(x)\|_2^2] \le \epsilon_0^2.$$

where h_k *is the step size defined in Definition 4.2 for step* k*,* N *is the total steps, and* $\sum_{k=1}^N h_k = T$*.*

To avoid ill-distribution q_T, we have the below definition follows [10, 14].

Definition 5.2. *We define* ϵ *as total variation distance between* $q_{T-\delta}$ *and* q_T*.*

Let the data distribution $p_0(x)$ be a k-mixture of Gaussians:

$$p_0(x) := \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \Sigma_i). \tag{9}$$

Using the Lemma 3.2, we know the pdf of x_t at the any time t is given by:

$$p_t(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(a_t \mu_i, a_t^2 \Sigma_i + b_t^2 I). \tag{10}$$

We introduce Pinsker's inequality, which relates total variation distance to the Kullback-Leibler divergence.

Lemma 5.3 (Pinsker's inequality [18])**.** *Let* p, q *are two probability distributions, then we have* $\mathrm{TV(p, q)} \le \sqrt{\frac{1}{2}\mathrm{KL}(p\|q)}$*.*

From Lemma 5.3, we can see the total variation is a weaker bound than KL divergence.

5.2. Total variation

Now, we are ready to present our result for total variation bound assuming data distribution is k-mixture of Gaussian (p_0 in Eq. (9)). Recall ϵ_0 is defined in Assumption 5.1, ϵ is defined in Definition 5.2 and h_k is defined in Definition 4.2. See the proof in Appendix I.

Theorem 5.4 (DDPM, total variation, informal version of Theorem I.1)**.** *Assume Condition 3.3 and 3.4 hold. The step size* $h_k := T/N$ *satisfies* $h_k = O(1/L)$ *and* $L \ge 1$ *for* $k \in [N]$*. Let* $\widehat{q}$ *denote the density of the output of the* EulerMaruyama *defined by Definition 4.3. Then, we have*

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim \underbrace{\sqrt{\mathrm{KL}(p_0 \| \mathcal{N}(0, I))} \exp(-T)}_{\text{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L m_2 h)\sqrt{T}}_{\text{discretization error}} + \underbrace{\epsilon_0 \sqrt{T}}_{\text{score estimation error}}.$$

where $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma^2_{\min(p_t)}} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det^{1/2}_{\min(p_t)}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}})$*,* $m_2 = (\sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]))^{1/2}$*, and* $\mathrm{KL}(p_0(x) \| \mathcal{N}(0, I)) \le \frac{1}{2}(-\log(\det_{\min(p_0)}) + d\sigma_{\max(p_0)} + \mu_{\max(p_0)} - d)$*.*

In Theorem 5.4, suppose that $m_2 \le d$ and choose $T = \Theta(\log(\mathrm{KL}(p_0 \| \mathcal{N}(0, I))/\epsilon))$, $h_k = \Theta(\frac{\epsilon^2}{L^2 d})$, then we have $\mathrm{TV}(\widehat{q}, p_0) \le \widetilde{O}(\epsilon + \epsilon_0)$, for $N = \widetilde{O}(L^2 d/\epsilon^2)$. In particular, in order to have $\mathrm{TV}(\widehat{q}, p_0) \le \epsilon$, it suffices to have score error $\epsilon_0 \le \widetilde{O}(\epsilon)$, where $\widetilde{O}(\cdot)$ hides T which is a log term. Thus, Theorem 5.4 provides a guarantee for total variation bound between target data distribution p_0 and learned output distribution $\widehat{q}$ with concrete L and m_2.

5.3. KL divergence

Similarly, we can present our result for the KL divergence bound in the following two theorems, assuming data distribution is k-mixture of Gaussian. Recall ϵ_0 is defined in Assumption 5.1, ϵ is denied in Definition 5.2 and h_k is defined in Definition 4.2. See the proof in Appendix I.

Theorem 5.5 (DDPM, KL divergence, informal version of Theorem I.2)**.** *Assume Condition 3.4. We use uniform discretization points.*

(1) Let $\widehat{q}$ denote the density of the output of the ExponentialIntegrator *(Definition 4.4), we have*

$$\mathrm{KL}(p_0 \| \widehat{q}) \lesssim (M_2 + d) \exp(-T) + T\epsilon_0^2 + \frac{dT^2L^2}{N}.$$

In particular, choosing $T = \Theta(\log(M_2 d/\epsilon_0))$ and $N = \Theta(dT^2L^2/\epsilon_0^2)$, then $\mathrm{KL}(p_0 \| \widehat{q}) = \widetilde{O}(\epsilon_0^2)$.

(2) Let $\widehat{q}$ denote the density of the output of the EulerMaruyama *(Definition 4.3), we have*

$$\mathrm{KL}(p_0 \| \widehat{q}) \lesssim (M_2 + d) \exp(-T) + T\epsilon_0^2 + \frac{dT^2L^2}{N} + \frac{T^3 M_2}{N^2}.$$

where $M_2 = \sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i])$ and $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma^2_{\min(p_t)}} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det^{1/2}_{\min(p_t)}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}})$.

Theorem 5.6 (DDPM, KL divergence for smooth data distribution, informal version of Theorem I.3)**.** *Assume Condition 3.3 and 3.4. We use the exponentially decreasing (then constant) step size $h_k = c \min\{\max\{t_k, 1/L\}, 1\}, c = \frac{T + \log L}{N} \leq \frac{1}{Kd}$. Let $\widehat{q}$ denote the density of the output of the* ExponentialIntegrator *defined by Definition 4.4. Then, we have*

$$\mathrm{KL}(p_0 \| \widehat{q}) \lesssim (M_2 + d) \exp(-T) + T\epsilon_0^2 + \frac{d^2(T + \log L)^2}{N},$$

where $M_2 = \sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i])$ and $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma^2_{\min(p_t)}} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det^{1/2}_{\min(p_t)}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}})$. In particular, if we set $T = \Theta(\log(M_2 d/\epsilon_0))$ and $N = \Theta(d^2(T + \log L)^2/\epsilon_0^2)$, then $\mathrm{KL}(p_0 \| \widehat{q}) \leq \widetilde{O}(\epsilon_0^2)$. In addition, for Euler-Maruyama scheme defined in Definition 4.3, the same bounds hold with an additional $M_2 \sum_{k=1}^N h_k^3$ term.

Theorem 5.6 and Theorem 5.5 provide a guarantee for KL divergence bound between target data distribution p_0 and learned output distribution $\widehat{q}$ with concrete L and M_2. Theorem 5.6 has $\log^2 L$ instead of L^2 in the bound for total number of discretization points N, but includes an additional d compared to Theorem 5.5. On the other hand, Theorem 5.6 requires the data distribution p_0 is Lipschitz and second-order differentiable [10], allowing a better bound in terms of L, while all other theorems [13, 14] require conditions about Lipschitz on p_t for any $0 \leq t \leq T$. However, our assumption p_0 is k-mixture of Gaussians satisfies all conditions they use.

From Lemma 5.3, we can compare KL divergence results with total variation results. Notice that $M_2 = m_2^2$ when comparing theorems for total variation and KL divergence. According to Lemma 5.3, the square of the TV distance is comparable to the KL divergence, which explains the squared relationship of the second momentum term.

5.4. Probability flow ODE

Notice that in the previous results we are considering SDE based models. However from [63], we know that we can also use ODE to run the reverse process, which has the same marginal distribution with SDE reverse process but is thereby deterministic. In this section, we provide results of DPOM and DPUM (Algorithm 1 and 2 in [14]) algorithms which are based on ODE reverse process.

Theorem 5.7 (DPOM, informal version of Theorem I.4)**.** *Assume Condition 3.4. We use the* DPOM *algorithm defined in Definition 4.5, and let $\widehat{q}$ be the output density of it. Then, we have*

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim (\sqrt{d} + m_2) \exp(-T) + L^2 T d^{1/2} h_{\mathrm{pred}} \\ + L^{3/2} T d^{1/2} h_{\mathrm{corr}}^{1/2} + L^{1/2} T \epsilon_0 + \epsilon.$$

where $m_2 = (\sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]))^{1/2}$ and $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma^2_{\min(p_t)}} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det^{1/2}_{\min(p_t)}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}})$. In particular, if we set $T = \Theta(\log(dm_2/\epsilon))$, $h_{\mathrm{pred}} = \widetilde{\Theta}(\frac{\epsilon}{L^2 d^{1/2}})$, $h_{\mathrm{corr}} = \widetilde{\Theta}(\frac{\epsilon}{L^3 d})$, and if the score estimation error satisfies $\epsilon_0 \leq \widetilde{O}(\frac{\epsilon}{\sqrt{L}})$, then we can obtain TV error ϵ with a total iteration complexity of $\widetilde{\Theta}(L^3 d/\epsilon^2)$ steps.

Theorem 5.8 (DPUM, informal version of Theorem I.5)**.** *Assume Condition 3.4. We use the* DPUM *algorithm defined in Definition 4.6, and let $\widehat{q}$ be the output density of it. Then, we have*

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim (\sqrt{d} + m_2) \exp(-T) + L^2 T d^{1/2} h_{\mathrm{pred}} \\ + L^{3/2} T d^{1/2} h_{\mathrm{corr}}^{1/2} + L^{1/2} T \epsilon_0 + \epsilon.$$

where $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma^2_{\min(p_t)}} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det^{1/2}_{\min(p_t)}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}})$ and $m_2 = (\sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]))^{1/2}$. In particular, if we set $T = \Theta(\log(dm_2/\epsilon))$, $h_{\mathrm{pred}} = \widetilde{\Theta}(\frac{\epsilon}{L^2 d^{1/2}})$, $h_{\mathrm{corr}} = \widetilde{\Theta}(\frac{\epsilon}{L^{3/2} d^{1/2}})$, and if the score estimation error *satisfies* $\epsilon_0 \leq \widetilde{O}(\frac{\epsilon}{\sqrt{L}})$*, then we can obtain TV error* ϵ *with a total iteration complexity of* $\widetilde{\Theta}(L^2 d^{1/2}/\epsilon)$ *steps.*

Theorem 5.7 and Theorem 5.8 provide a guarantee for total variation bound between target data distribution p_0 and learned output distribution $\widehat{q}$ with concrete L and M_2 for ODE reverse process. The difference between the DPOM (Theorem 5.7) and DPUM (Theorem 5.8) is the complexity of $h_{\rm corr}$ and the final iteration complexity term. Using DPUM algorithm, we can reduce the total iteration complexity from $\widetilde{\Theta}(L^3 d/\epsilon^2)$ to $\widetilde{\Theta}(L^2 \sqrt{d}/\epsilon)$.

6. Experiments

We provide empirical evidence supporting our theoretical results. In Section 6.1, we describe the experimental settings and data configurations used in our study. In Section 6.2, we analyze the correlation between the Lipschitz constant and the performance of the diffusion model.

6.1. Experiment setup

We used a Gaussian dataset with three modes (clusters) in two-dimensional space, each having the same covariance matrix. By adjusting the covariance, we examine the impact of the Lipschitz constant on model performance. The model we use is a one hidden layer MLP with 32 hidden dimensions. We use DDPM algorithm for our forward diffusion process and reverse denoising process. The number of inference steps we use is 200. It is optimized by MSE loss, and the performance metrics are log likelihood and Mahalanobis squared distance. For the training, we use AdamW optimizer with 0.001 learning rate and training epochs 100.

For the evaluation, we sample 1000 dots from central Gaussian distribution with variance 1. Then calculate the two losses of output and ground truth.

6.2. Correlation study

This section explores the relationship between the Lipschitz constant and key performance indicators of the diffusion model. Table 2 shows the average log likelihoods with Lipschitz constants, and Table 3 presents the average Mahalanobis squared distances.

Figure 2 portrays two line charts: the upper panel maps the log likelihood against the Lipschitz constant, and the lower panel charts the Mahalanobis squared distance.

From Table 2, Table 3 and Figure 2, we can conclude the message that a larger covariance leads to a smaller Lipschitz (more smooth) target function, which is easier to learn. This is aligned with our theoretical analysis, i.e., Lemma 3.5, Theorem 5.4, Theorem 5.5, and Theorem 5.6.

Covariance	Lipschitz	Avg Log Likelihood
1	3.69×10^7	-9901.2324
10	3.23×10^4	-1001.9962
100	31.8916	-106.1019
200	3.9870	-57.2244
500	0.2567	-28.0872
600	0.1491	-24.9577
700	0.0943	-22.6524
800	0.0634	-21.0397
900	0.0448	-19.7688
1000	0.0328	-18.7532

Table 2. Log Likelihood Loss for Different Lipschitz. For the Covariance column, the number being z means that the covariance matrix $\Sigma = zI_{2\times 2}$. We compute the corresponding Lipschitz L and report the evaluation loss (Avg Log Likelihood).

Covariance	Lipschitz	Avg Mahalanobis^2
1	3.69×10^7	19798.7871
10	3.23×10^4	1995.7114
100	31.8916	199.3178
200	3.9870	100.1765
500	0.2567	40.0695
600	0.1491	33.4458
700	0.0943	28.5269
800	0.0634	25.0344
900	0.0448	22.2571
1000	0.0328	20.0152

Table 3. Mahalanobis Squared Loss for Different Lipschitz. For the Covariance column, the number being z means that the covariance matrix $\Sigma = zI_{2\times 2}$. We compute the corresponding Lipschitz L and report the evaluation loss (Avg Mahalanobis).

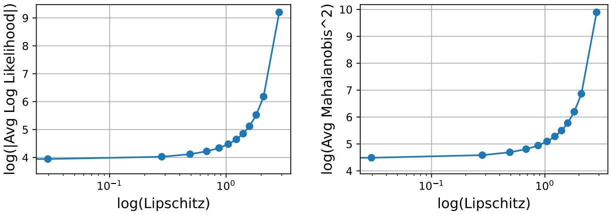

Figure 2. **Left:** Log likelihood vs. Lipschitz constant. **Right:** Mahalanobis squared distance vs. Lipschitz constant.

7. Conclusion

We have presented a theoretical analysis of the Lipschitz continuity and second momentum properties of diffusion models. Our results provide concrete bounds on key properties of the diffusion process and establish error guarantees for various diffusion solvers. These findings contribute to a deeper understanding of diffusion models and spur further theoretical and practical advancements in this field.

References

[1] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017. 2

[2] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 5

[3] Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[4] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *arXiv preprint arXiv:2005.02970*, 2020.

[5] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022. 4

[6] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE, 2010. 2

[7] Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023. 2

[8] Rares-Darius Buhai and David Steurer. Beyond parallel pancakes: Quasi-polynomial time guarantees for non-spherical gaussian mixtures. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 548–611. PMLR, 2023. 2, 4

[9] Siu-On Chan, Ilias Diakonikolas, Xiaorui Sun, and Rocco A Servedio. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1380–1394. SIAM, 2013. 2

[10] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023. 1, 2, 3, 5, 6, 7, 42

[11] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023. 2

[12] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 2

[13] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022. 1, 2, 3, 6, 7, 42

[14] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 2, 3, 6, 7, 42

[15] Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023. 2

[16] Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024. 2, 4

[17] Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 4

[18] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011. 6

[19] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999. 2

[20] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory*, pages 1183–1213. PMLR, 2014. 2

[21] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022. 1

[22] Ilias Diakonikolas and Daniel M Kane. 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, 2020. 2

[23] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[24] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020.

[25] Jon Feldman, Ryan O'Donnell, and Rocco A Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008. 2

[26] Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024. 2, 4

[27] Hanzhong Guo, Cheng Lu, Fan Bao, Tianyu Pang, Shuicheng Yan, Chao Du, and Chongxuan Li. Gaussian mixture solvers for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[28] Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[29] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, pages IV–317. IEEE, 2007. 45

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 4, 5

[31] Jerry Yao-Chieh Hu, Weimin Wu, Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (dits). *arXiv preprint arXiv:2407.01079*, 2024. 2

[32] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[33] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2

[34] Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the wasserstein distance. *Advances in Neural Information Processing Systems*, 35:20205–20217, 2022. 2

[35] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35: 22870–22882, 2022. 1, 2

[36] Chenyang Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Exploring the frontiers of softmax: Provable optimization, applications in diffusion model, and beyond. *arXiv preprint arXiv:2405.03251*, 2024. 2

[37] Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, pages 1302–1382. PMLR, 2017. 2

[38] Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models. In *Forty-first International Conference on Machine Learning*, 2024. 4

[39] Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Differential privacy mechanisms in neural tangent kernel regression. *arXiv preprint arXiv:2407.13621*, 2024. 2

[40] Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space, 2023. *URL https://arxiv. org/abs/2302.07400*, 2023. 2

[41] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010. 2

[42] Eliya Nachmani, Robin San Roman, and Lior Wolf. Non gaussian denoising diffusion models. *arXiv preprint arXiv:2106.07582*, 2021. 2

[43] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[44] OpenAI. Video generation models as world simulators, 2024. https://openai.com/research/video-generation-models-as-world-simulators. 1

[45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[46] João M Pereira, Joe Kileel, and Tamara G Kolda. Tensor moments of gaussian mixture models: Theory and applications. *arXiv preprint arXiv:2202.06930*, 2022. 44

[47] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *International Conference on Machine Learning*, pages 8936–8947. PMLR, 2021. 2, 4

[48] Alex Reneau, Jerry Yao-Chieh Hu, Chenwei Xu, Weijian Li, Ammar Gilani, and Han Liu. Feature programming for multivariate time series prediction. In *Fortieth International Conference on Machine Learning (ICML)*, 2023. 2

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[51] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 2

[52] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015. 1

[53] Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023. 2, 4

[54] Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2021. 2

[55] Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[56] Zhenmei Shi, Yifei Ming, Ying Fan, Frederic Sala, and Yingyu Liang. Domain generalization via nuclear norm regularization. In *Conference on Parsimony and Learning*, pages 179–201. PMLR, 2024. 2

[57] Zhenmei Shi, Junyi Wei, and Yingyu Liang. Provable guarantees for neural networks via gradient feature learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[58] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[59] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[60] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[61] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

[62] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020. 2

[63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2, 5, 7

[64] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021. 2

[65] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32211–32252, 2023. 2

[66] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[67] Susana Vinga. Convolution integrals of normal distribution functions. *Supplementary material to Vinga and Almeida (2004)"Rényi continuous entropy of DNA sequences*, 2004. 43

[68] Yilin Wang, Zeyuan Chen, Liangjun Zhong, Zheng Ding, Zhizhou Sha, and Zhuowen Tu. Dolfin: Diffusion layout transformers without autoencoder. *arXiv preprint arXiv:2310.16305*, 2023. 2

[69] Yilin Wang, Haiyang Xu, Xiang Zhang, Zeyuan Chen, Zhizhou Sha, Zirui Wang, and Zhuowen Tu. Omnicontrolnet: Dual-stage integration for conditional image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7436–7448, 2024.

[70] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Grounding diffusion with token-level supervision. *arXiv preprint arXiv:2312.03626*, 2023. 2

[71] Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. In *Forty-first International Conference on Machine Learning*, 2024. 2, 4

[72] Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, and Yingyu Liang. Towards few-shot adaptation of foundation models via multitask finetuning. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[73] Zhantao Yang, Ruili Feng, Han Zhang, Yujun Shen, Kai Zhu, Lianghua Huang, Yifei Zhang, Yu Liu, Deli Zhao, Jingren Zhou, et al. Lipschitz singularities in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[74] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation. *arXiv preprint arXiv:2402.19159*, 2024. 2

Appendix

Roadmap.

- Section A discusses the limitations of the paper.
- Section B discusses the societal impacts of the paper.
- Section C provides the preliminary for the paper.
- Section D provides the case when we consider a continuous time score function, specifically a single Gaussian.
- Section E provides the case when we consider the score function to be 2 mixtures of Gaussians.
- Section F provides the case when we consider the score function to be k mixture of Gaussians.
- Section G provides the tools that we use from previous papers.
- Section H provides lemmas that we use for a more concrete calculation for theorems in Section G.
- Section I provides our main results when we consider the data distribution is k mixture of Gaussians.

A. Limitations

This work has not directly addressed the practical applications of our results. Additionally, we did not provide a sample complexity bound for our settings. Future research could explore how these findings might be implemented in real-world scenarios and work on improving these limitations.

B. Societal Impacts

We explore and provide a deeper understanding of the diffusion models and also explicitly give the Lipschitz constant for k-mixture of Gaussians, which may inspire a better algorithm design.

On the other hand, our paper is purely theoretical in nature, so we foresee no immediate negative ethical impact.

C. Preliminary

This section provides some preliminary knowledge and is organized as below:

- Section C.1 provides the facts we use.
- Section C.2 provides the property of exp function we use.
- Section C.3 provides the Lipschitz multiplication property we use.

C.1. Facts

We provide several basic facts from calculus and linear algebra that are used in the proofs.

Fact C.1 (Calculus)**.** *For $x \in \mathbb{R}$, $y \in \mathbb{R}$, $t \in \mathbb{R}$, $u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, it is well-known that*

- $\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}t}$ *(chain rule)*
- $\frac{\mathrm{d}xy}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t}y + \frac{\mathrm{d}y}{\mathrm{d}t}x$ *(product rule)*
- $\frac{\mathrm{d}x^n}{\mathrm{d}x} = nx^{n-1}$ *(power rule)*
- $\frac{\mathrm{d}\langle u,v\rangle}{\mathrm{d}u} = v$ *(derivative of the inner product)*
- $\frac{\mathrm{d}\exp(x)}{\mathrm{d}x} = \exp(x)$ *(derivative of exponential function)*
- $\frac{\mathrm{d}\log x}{\mathrm{d}x} = \frac{1}{x}$ *(derivative of logarithm function)*
- $\frac{\mathrm{d}}{\mathrm{d}u}\|u\|_2^2 = 2u$ *(derivative of ℓ_2 norm)*
- $\frac{\mathrm{d}y}{\mathrm{d}x} = 0$ *if y is independent from x. (derivative of independent variables)*

Fact C.2 (Norm Bounds)**.** *For $a \in \mathbb{R}$, $b \in \mathbb{R}$, $u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $A \in \mathbb{R}^{k\times n}$, $W \in \mathbb{R}^{n\times n}$ is symmetric and p.s.d., we have*

- $\|au\|_2 = |a| \cdot \|u\|_2$ *(absolute homogeneity)*
- $\|u + v\|_2 \leq \|u\|_2 + \|v\|_2$ *(triangle inequality)*
- $|u^\top v| \leq \|u\|_2 \cdot \|v\|_2$ *(Cauchy–Schwarz inequality)*
- $\|u^\top\|_2 = \|u\|_2$
- $\|Au\|_2 \leq \|A\| \cdot \|u\|_2$
- $\|aA\| = |a| \cdot \|A\|$
- $\|A\| = \sigma_{\max}(A)$
- $\|A^{-1}\| = \frac{1}{\sigma_{\min}(A)}$

- $u^\top W u \geq \|u\|_2^2 \cdot \sigma_{\min}(W)$.
- $\sigma_{\min}(W^{-1}) = \frac{1}{\sigma_{\max}(W)}$.

Fact C.3 (Matrix Calculus)**.** *Let $W \in \mathbb{R}^{n \times n}$ denote a symmetric matrix. Let $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^n$. Suppose that s is independent of x. Then, we know*
- $\frac{\mathrm{d}}{\mathrm{d}x}(x-s)^\top W(x-s) = 2W(x-s)$

C.2. Properties of $\exp$ functions

During the course of proving the Lipschitz continuity for mixtures of Gaussians, we found that we need to use the following bound for the $\exp$ function.

Fact C.4. *For $|a-b| \leq 0.1$, where $a \in \mathbb{R}$, $b \in \mathbb{R}$, we have*

$$|\exp(a) - \exp(b)| \leq |\exp(a)| \cdot 2|a-b|$$

Proof. We have

$$\begin{aligned} |\exp(a) - \exp(b)| &= |\exp(a) \cdot (1 - \exp(b-a))| \\ &= |\exp(a)| \cdot |(1 - \exp(b-a))| \\ &\leq |\exp(a)| \cdot 2|a-b| \end{aligned}$$

where the first step follows from simple algebra, the second step follows from $|a \cdot b| = |a| \cdot |b|$, and the last step follows from $|\exp(x) - 1| \leq 2x$ for all $x \in (0, 0.1)$. □

Fact C.5. *For $\|u-v\|_\infty \leq 0.1$, where $u, v \in \mathbb{R}^n$, we have*

$$\|\exp(u) - \exp(v)\|_2 \leq \|\exp(u)\|_2 \cdot 2\|u-v\|_\infty$$

Proof. We have

$$\begin{aligned} \|\exp(u) - \exp(v)\|_2 &= \|\exp(u) \circ (\mathbf{1}_n - \exp(v-u))\|_2 \\ &\leq \|\exp(u)\|_2 \cdot \|\mathbf{1}_n - \exp(v-u)\|_\infty \\ &\leq \|\exp(u)\|_2 \cdot 2\|u-v\|_\infty \end{aligned}$$

where the first step follows from notation of Hardamard product, the second step follows from $\|u \circ v\|_2 \leq \|u\|_\infty \cdot \|v\|_2$, and the last step follows from $|\exp(x) - 1| \leq 2x$ for all $x \in (0, 0.1)$. □

Fact C.6 (Mean value theorem for vector function)**.** *For vector $x, y \in C \subset \mathbb{R}^n$, vector function $f(x) : C \to \mathbb{R}$, $g(x) : C \to \mathbb{R}^m$, let f, g be differentiable on open convex domain C, we have*
- *Part 1: $f(y) - f(x) = \nabla f(x + t(y-x))^\top (y-x)$*
- *Part 2: $\|g(y) - g(x)\|_2 \leq \|g'(x + t(y-x))\| \cdot \|y-x\|_2$ for some $t \in (0,1)$, where $g'(a)$ denotes a matrix which its (i,j)-th term is $\frac{\mathrm{d}g(a)_j}{\mathrm{d}a_i}$.*
- *Part 3: If $\|g'(a)\| \leq M$ for all $a \in C$, then $\|g(y) - g(x)\|_2 \leq M\|y-x\|_2$ for all $x, y \in C$.*

Proof. **Proof of Part 1**

Part 1 can be verified by applying Mean Value Theorem of 1-variable function on $\gamma(c) = f(x + c(y-x))$.

$$f(y) - f(x) = \gamma(1) - \gamma(0) = \gamma'(t)(1-0) = \nabla f(x + t(y-x))^\top (y-x)$$

where $t \in (0,1)$.

Proof of Part 2

Let $G(c) := (g(y) - g(x))^\top g(c)$, we have

$$\begin{aligned} \|g(y) - g(x)\|_2^2 &= G(y) - G(x) \\ &= \nabla G(x + t(y-x))^\top (y-x) \end{aligned}$$

$$
\begin{aligned}
&= (\underbrace{g'(x+t(y-x))}_{n\times m} \cdot \underbrace{(g(y)-g(x))}_{m\times 1})^\top \cdot \underbrace{(y-x)}_{n\times 1} \\
&\leq \|g'(x+t(y-x))\| \cdot \|g(y)-g(x)\|_2 \cdot \|y-x\|_2
\end{aligned}
$$

the initial step is by basic calculation, the second step is from **Part 1**, the third step uses chain rule, the 4th step is due to Cauchy-Schwartz inequality. Removing $\|g(y)-g(x)\|_2$ on both sides gives the result.

Proof of Part 3

Part 3 directly follows from **Part 2**. □

We show the upper bound of $\exp'$ below, assuming input is bounded.

Fact C.7. *Let $g'(a)$ denotes a matrix whose (i,j)-th term is $\frac{\mathrm{d}g(a)_j}{\mathrm{d}a_i}$. For $u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\|u\|_2, \|v\|_2 \leq R$, where $R \geq 0$, $t \in (0,1)$, we have*

$$
\|\exp'(u+t(v-u))\| \leq \exp(R)
$$

Proof. We can show

$$
\begin{aligned}
\|\exp'(u+t(v-u))\| &= \|\mathrm{diag}(\exp(u+t(v-u)))\| \\
&\leq \sigma_{\max}(\mathrm{diag}(\exp(u+t(v-u)))) \\
&= \max_{i\in[n]} \exp(u_i+t(v_i-u_i)) \\
&\leq \max_{i\in[n]} \max\{\exp(v_i),\exp(u_i)\} \\
&\leq \exp(R)
\end{aligned}
$$

where the first step follows from $\frac{\mathrm{d}\exp(x)}{\mathrm{d}x} = \mathrm{diag}(\exp(x))$, the second step follows from Fact C.2, the third step follows from spectral norm of a diagonal matrix is the absolute value of its largest entry, the fourth step follows from $t \in (0,1)$, and the last step follows from $\|\exp(v)\|_\infty \leq \exp(\|v\|_\infty) \leq \exp(\|v\|_2)$. □

Fact C.8. *For $u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\|u\|_2, \|v\|_2 \leq R$, where $R \geq 0$, we have*

$$
\|\exp(u)-\exp(v)\|_2 \leq \exp(R)\|u-v\|_2
$$

Proof. We can show, for $t \in (0,1)$,

$$
\begin{aligned}
\|\exp(u)-\exp(v)\|_2 &\leq \|\exp'(u+t(v-u))\| \cdot \|u-v\|_2 \\
&\leq \exp(R)\|u-v\|_2
\end{aligned}
$$

where the first step follows from Fact C.6, the second step follows from Fact C.7. □

Fact C.9. *For $a \in \mathbb{R}$, $b \in \mathbb{R}$, $a, b \leq R$, where $R \geq 0$, we have*

$$
|\exp(a)-\exp(b)| \leq \exp(R)|a-b|
$$

Proof. We can show, for $t \in (0,1)$,

$$
\begin{aligned}
|\exp(a)-\exp(b)| &= |\exp'(a+t(b-a))| \cdot |a-b| \\
&= |\exp(a+t(b-a))| \cdot |a-b| \\
&\leq \max\{\exp(a),\exp(b)\} \cdot |a-b| \\
&\leq \exp(R)|u-v|
\end{aligned}
$$

where the first step follows from Mean Value Theorem, the second step follows from Fact C.1, the third step follows from $t \in (0,1)$, and the last step follows from $a, b \leq R$.

□

C.3. Lipschitz multiplication property

Our overall proofs of Lipschitz constant for k-mixture of Gaussians follow the idea from Fact below.

Fact C.10. *If the following conditions hold*

- $\|f_i(x) - f_i(y)\|_2 \leq L \cdot \|x - y\|_2$
- $R := \max_{i \in [n], x} |f_i(x)|$

Then, we have

$$|\prod_{i=1}^{k} f_i(x) - \prod_{i=1}^{k} f_i(y)| \leq k \cdot R^{k-1} \cdot L \cdot \|x - y\|_2$$

Proof. We can show

$$\begin{aligned}
& |\prod_{i=1}^{k} f_i(x) - \prod_{i=1}^{k} f_i(y)| \\
= & |f_k(x) \prod_{i=1}^{k-1} f_i(x) - f_k(y) \prod_{i=1}^{k-1} f_i(y)| \\
\leq & |f_k(x) \prod_{i=1}^{k-1} f_i(x) - f_k(y) \prod_{i=1}^{k-1} f_i(x)| + |f_k(y) \prod_{i=1}^{k-1} f_i(x) - f_k(y) \prod_{i=1}^{k-1} f_i(y)| \\
= & |(f_k(x) - f_k(y)) \prod_{i=1}^{k-1} f_i(x)| + |f_k(y)(\prod_{i=1}^{k-1} f_i(x) - \prod_{i=1}^{k-1} f_i(y))| \\
\leq & L \cdot \|x - y\|_2 \cdot R^{k-1} + R \cdot |\prod_{i=1}^{k-1} f_i(x) - \prod_{i=1}^{k-1} f_i(y)| \\
\leq & L \cdot \|x - y\|_2 \cdot R^{k-1} + R \cdot (|L \cdot \|x - y\|_2 \cdot R^{k-2} + R \cdot |\prod_{i=1}^{k-2} f_i(x) - \prod_{i=1}^{k-2} f_i(y)|) \\
= & 2 \cdot L \cdot \|x - y\|_2 \cdot R^{k-1} + R^2 \cdot |\prod_{i=1}^{k-2} f_i(x) - \prod_{i=1}^{k-2} f_i(y)| \\
\leq & k \cdot R^{k-1} \cdot L \cdot \|x - y\|_2
\end{aligned}$$

where the first step follows from simple algebra, the second step follows from Fact C.2, the third step follows from rearranging terms, the fourth step follows from the assumptions of the lemma, the fifth step follows from the same logic of above, the sixth step follows from simple algebra, and the last step follows from the recursive process. □

D. Single Gaussian Case

In this section, we consider the continuous case of $p_t(x)$, which is the probability density function (pdf) of the input data x, and also a function of time t. More specifically, we consider the cases when $p_t(x)$ is: (1) a single Gaussian where either the mean is a function of time (Section D.1) or the covariance is a function of time (Section D.2), (2) a single Gaussian where both the mean and the covariance are a function of time (Section D.3). And then, we compute the upper bound and Lipschitz constant for the score function i.e. the gradient of log pdf $\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x}$.

D.1. Case when the mean of $p_t(x)$ is a function of time

We start our calculation by a simple case. Consider p_t such that

$$p_t(x) = \Pr_{x' \sim \mathcal{N}(t \mathbf{1}_d, I_d)}[x' = x]$$

Let pdf is $\mathbb{R}^d \to \mathbb{R}$ denote p_t. We have $\log(\mathsf{pdf}())$ is $\mathbb{R}^d \to \mathbb{R}$. Then, we can get gradient $\nabla \log(\mathsf{pdf}())$ is a function of t because of $\mathcal{N}(t\mathbf{1}_d, I_d)$. Inject x and y into the gradient function, then we are done.

Below we define the pdf for the continues case when the mean is a function of time.

Definition D.1. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

We define

$$p_t(x) := \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}\|x - t\mathbf{1}_d\|_2^2)$$

Further, we have

$$\log p_t(x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\|x - t\mathbf{1}_d\|_2^2$$

Below we calculate the score function of pdf for the continuous case when the mean is a function of time.

Lemma D.2. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

Then,

$$\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x} = t\mathbf{1}_d - x$$

Proof. We can show

$$\begin{aligned}\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x} &= \frac{\mathrm{d}}{\mathrm{d}x}(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\|x - t\mathbf{1}_d\|_2^2) \\ &= -\frac{1}{2} \cdot \frac{\mathrm{d}}{\mathrm{d}x}\|x - t\mathbf{1}_d\|_2^2 \\ &= -\frac{1}{2} \cdot 2(x - t\mathbf{1}_d) \\ &= t\mathbf{1}_d - x\end{aligned}$$

where the first step follows from Definition D.1, the second step follows from variables are independent, the third step follows from Fact C.1, and the last step follows from simple algebra. □

Below we calculate the upper bound for the score function of pdf for continuous case when the mean is a function of time.

Lemma D.3 (Linear growth)**.** *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

Then,

$$\|\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x}\|_2 \leq t + \|x\|_2$$

Proof. We can show

$$\begin{aligned}\|\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x}\|_2 &= \|t\mathbf{1}_d - x\|_2 \\ &\leq \|t\mathbf{1}_d\|_2 + \|-x\|_2 \\ &\leq t + \|x\|_2\end{aligned}$$

where the first step follows from Lemma D.2, the second step follows from Fact C.2, and the last step follows from simple algebra. □

Below we calculate the Lipschitz constant for the score function of pdf for continuous case when the mean is a function of time.

Lemma D.4 (Lipschitz)**.** *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.

Then,

$$\| \frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} - \frac{\mathrm{d} \log p_t(\widetilde{x})}{\mathrm{d} \widetilde{x}} \|_2 = \| \widetilde{x} - x \|_2$$

Proof. We can show

$$\begin{aligned} \| \frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} - \frac{\mathrm{d} \log p_t(\widetilde{x})}{\mathrm{d} \widetilde{x}} \|_2 &= \| t \mathbf{1}_d - x - (t \mathbf{1}_d - \widetilde{x}) \|_2 \\ &= \| \widetilde{x} - x \|_2 \end{aligned}$$

where the first step follows from Lemma D.2, and the last step follows from simple algebra. □

D.2. Case when the covariance of $p_t(x)$ is a function of time

Below we define the pdf for continuous case when the covariance is a function of time.

Definition D.5. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.

We define

$$p_t(x) := \frac{1}{t^{1/2} (2\pi)^{d/2}} \exp(-\frac{1}{2t} \| x - \mathbf{1}_d \|_2^2)$$

Further, we have

$$\log p_t(x) = -\frac{1}{2} \log t - \frac{d}{2} \log(2\pi) - \frac{1}{2t} \| x - \mathbf{1}_d \|_2^2$$

Below we calculate the score function of pdf for continuous case when the covariance is a function of time.

Lemma D.6. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.

Then,

$$\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} = \frac{1}{t} (\mathbf{1}_d - x)$$

Proof. We can show

$$\begin{aligned} \frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} &= \frac{\mathrm{d}}{\mathrm{d} x} (-\frac{1}{2} \log t - \frac{d}{2} \log(2\pi) - \frac{1}{2t} \| x - \mathbf{1}_d \|_2^2) \\ &= -\frac{1}{2t} \cdot \frac{\mathrm{d}}{\mathrm{d} x} \| x - \mathbf{1}_d \|_2^2 \\ &= -\frac{1}{2t} \cdot 2(x - \mathbf{1}_d) \\ &= \frac{1}{t} (\mathbf{1}_d - x) \end{aligned}$$

where the first step follows from Definition D.5, the second step follows from variables are independent, the third step follows from Fact C.1, and the last step follows from simple algebra. □

Below we calculate the upper bound of the score function of pdf for the continuous case when the covariance is a function of time.

Lemma D.7 (Linear growth)**.** *If the following conditions hold*
- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.

Then,

$$\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x}\|_2 \leq \frac{1}{t}(1 + \|x\|_2)$$

Proof. We can show

$$\begin{aligned}\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x}\|_2 &= \|\frac{1}{t}(\mathbf{1}_d - x)\|_2 \\ &= |\frac{1}{t}| \cdot \|\mathbf{1}_d - x\|_2 \\ &= \frac{1}{t}\|\mathbf{1}_d - x\|_2 \\ &\leq \frac{1}{t}(\|\mathbf{1}_d\|_2 + \| - x\|_2) \\ &= \frac{1}{t}(1 + \|x\|_2)\end{aligned}$$

where the first step follows from Lemma D.6, the second step follows from Fact C.2, the third step follows from $t \geq 0$, the fourth step follows from Fact C.2, and the last step follows from simple algebra. □

Below we calculate the Lipschitz constant of the score function of pdf for continuous case when the covariance is a function of time.

Lemma D.8 (Lipschitz)**.** *If the following conditions hold*
- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.

Then,

$$\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} - \frac{\mathrm{d} \log p_t(\widetilde{x})}{\mathrm{d} \widetilde{x}}\|_2 = \frac{1}{t}\|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned}\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} - \frac{\mathrm{d} \log p_t(\widetilde{x})}{\mathrm{d} \widetilde{x}}\|_2 &= \|\frac{1}{t}(\mathbf{1}_d - x) - \frac{1}{t}(\mathbf{1}_d - \widetilde{x})\|_2 \\ &= \|\frac{1}{t}(\widetilde{x} - x)\|_2 \\ &= \frac{1}{t}\|x - \widetilde{x}\|_2\end{aligned}$$

where the first step follows from Lemma D.6, the second step follows from simple algebra, the third step follows from Fact C.2. □

D.3. A general version for single Gaussian

Now we combine the previous results by calculate a slightly more complex case. Consider p_t such that

$$p_t(x) = \Pr_{x' \sim \mathcal{N}(\mu(t), \Sigma(t))}[x' = x]$$

where $\mu(t) \in \mathbb{R}^d$, $\Sigma(t) \in \mathbb{R}^{d \times d}$ and they are derivative to t and $\Sigma(t)$ is a symmetric p.s.d. matrix whose the smallest singular value is always larger than a fixed $\sigma_{\min} > 0$.

Definition D.9. *If the following conditions hold*
- *Let* $x \in \mathbb{R}^d$.

- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

  We define

$$p_t(x) := \frac{1}{(2\pi)^{d/2} \det(\Sigma(t))^{1/2}} \exp(-\frac{1}{2}(x - \mu(t))^\top \Sigma(t)^{-1}(x - \mu(t))).$$

Further, we have

$$\log p_t(x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma(t)) - \frac{1}{2}(x - \mu(t))^\top \Sigma(t)^{-1}(x - \mu(t))$$

Below we calculate the score function of pdf for continuous case when both the mean and covariance are a function of time.

Lemma D.10. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

Then,

$$\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x} = -\Sigma(t)^{-1}(x - \mu(t))$$

Proof. We can show

$$\begin{aligned}
\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x} = & \frac{\mathrm{d}}{\mathrm{d}x}(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma(t)) - \frac{1}{2}(x - \mu(t))^\top \Sigma(t)^{-1}(x - \mu(t))) \\
= & -\frac{1}{2} \cdot \frac{\mathrm{d}}{\mathrm{d}x}(x - \mu(t))^\top \Sigma(t)^{-1}(x - \mu(t)) \\
= & -\frac{1}{2} \cdot 2\Sigma(t)^{-1}(x - \mu(t)) \\
= & -\Sigma(t)^{-1}(x - \mu(t))
\end{aligned}$$

where the first step follows from Definition D.9, the second step follows from Fact C.1, the third step follows from Fact C.3, and the last step follows from simple algebra. □

Below we calculate the upper bound of the score function of pdf for continuous case when both the mean and covariance is a function of time.

Lemma D.11 (Linear growth)**.** *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

Then,

$$\|\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x}\|_2 \leq \frac{1}{\sigma_{\min}(\Sigma(t))} \cdot (\|\mu(t)\|_2 + \|x\|_2)$$

Proof. We can show

$$\begin{aligned}
\|\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x}\|_2 = & \| - \Sigma(t)^{-1}(x - \mu(t))\|_2 \\
\leq & \| - \Sigma(t)^{-1}\| \cdot \|x - \mu(t)\|_2 \\
= & \|\Sigma(t)^{-1}\| \cdot \|x - \mu(t)\|_2 \\
= & \frac{1}{\sigma_{\min}(\Sigma(t))} \cdot \|x - \mu(t)\|_2 \\
\leq & \frac{1}{\sigma_{\min}(\Sigma(t))} \cdot (\|x\|_2 + \| - \mu(t)\|_2) \\
= & \frac{1}{\sigma_{\min}(\Sigma(t))} \cdot (\|\mu(t)\|_2 + \|x\|_2)
\end{aligned}$$

where the first step follows from Lemma D.10, the second step follows from Fact C.2, the third step follows from Fact C.2, the fourth step follows from Fact C.2, the fifth step follows from Fact C.2, and the last step follows from Fact C.2. □

Below we calculate the Lipschitz constant of the score function of pdf for continuous case when both the mean and covariance are a function of time.

Lemma D.12 (Lipschitz)**.** *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.

Then,

$$\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d}x} - \frac{\mathrm{d} \log p_t(\widetilde{x})}{\mathrm{d}\widetilde{x}}\|_2 \leq \frac{1}{\sigma_{\min}(\Sigma(t))} \cdot \|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned}
\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d}x} - \frac{\mathrm{d} \log p_t(\widetilde{x})}{\mathrm{d}\widetilde{x}}\|_2 &= \| - \Sigma(t)^{-1}(x - \mu(t)) - (-\Sigma(t)^{-1}(\widetilde{x} - \mu(t)))\|_2 \\
&= \| - \Sigma(t)^{-1}(x - \widetilde{x})\|_2 \\
&\leq \| - \Sigma(t)^{-1}\| \cdot \|x - \widetilde{x}\|_2 \\
&= \frac{1}{\sigma_{\min}(\Sigma(t))} \cdot \|x - \widetilde{x}\|_2
\end{aligned}$$

where the first step follows from Lemma D.10, the second step follows from simple algebra, the third step follows from Fact C.2, and the last step follows from Fact C.2. □

E. A General Version for Two Gaussian

In this section, we compute the linear growth and Lipschitz constant for a mixture of 2 Gaussian where both the mean and covariance are a function of time. The organization of this section is as follows:

- Section E.1 defines the probability density function (pdf) $p_t(x)$ that we use, which is a mixture of 2 Gaussian.
- Section E.2 provides lemmas that are used for calculation of the score function i.e. gradient of the log pdf $\frac{\mathrm{d} \log p_t(x)}{\mathrm{d}x}$.
- Section E.3 provides the expression of the score function.
- Section E.4 provides lemmas that are used for calculation of the upper bound of the score function.
- Section E.5 provides the expression of the upper bound of the score function.
- Section E.6 provides lemmas of upper bound for some base functions that are used for calculation of the Lipschitz constant of the score function.
- Section E.7 provides lemmas of Lipschitz constant for some base functions that are used for calculation of the Lipschitz constant of the score function.
- Section E.8 provides lemmas of Lipschitz constant for $f(x)$ that are used for calculation of the Lipschitz constant of the score function.
- Section E.9 provides lemmas of Lipschitz constant for $g(x)$ that are used for calculation of the Lipschitz constant of the score function.
- Section E.10 provides the expression of the Lipschitz constant of the score function.

First, we define the following. Let $\alpha(t) \in (0, 1)$ and also is a function of time t. Consider p_t such that

$$p_t(x) = \Pr_{x' \sim \alpha(t)\mathcal{N}(\mu_1(t), \Sigma_1(t)) + (1-\alpha(t)))\mathcal{N}(\mu_2(t), \Sigma_2(t))}[x' = x]$$

where $\mu_1(t), \mu_2(t) \in \mathbb{R}^d$, $\Sigma_1(t), \Sigma_2(t) \in \mathbb{R}^{d \times d}$ and they are derivative to t and $\Sigma_1(t), \Sigma_2(t)$ is a symmetric p.s.d. matrix whose the smallest singular value is always larger than a fixed value $\sigma_{\min} > 0$.

For further simplicity of calculation, we denote $\alpha(t)$ to be α.

E.1. Definitions

Below we define function N_1 and N_2.

Definition E.1. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.

We define

$$N_1(x) := \frac{1}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}}\exp(-\frac{1}{2}(x-\mu_1(t))^\top\Sigma_1(t)^{-1}(x-\mu_1(t)))$$

and

$$N_2(x) := \frac{1}{(2\pi)^{d/2}\det(\Sigma_2(t))^{1/2}}\exp(-\frac{1}{2}(x-\mu_2(t))^\top\Sigma_2(t)^{-1}(x-\mu_2(t)))$$

It's clearly to see that $N_i \leq \frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}}$ *since* $N_i(x)$ *takes maximum when* $x = \mu_i$*.*

Below we define the pdf for 2 mixtures of Gaussians.

Definition E.2. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$*.*
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$*.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0,1)$*.*
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*

We define

$$\begin{aligned} p_t(x) :=& \frac{\alpha}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}}\exp(-\frac{1}{2}(x-\mu_1(t))^\top\Sigma_1(t)^{-1}(x-\mu_1(t)))+ \\ & \frac{1-\alpha}{(2\pi)^{d/2}\det(\Sigma_2(t))^{1/2}}\exp(-\frac{1}{2}(x-\mu_2(t))^\top\Sigma_2(t)^{-1}(x-\mu_2(t))). \end{aligned}$$

This can be further rewritten as follows:

$$p_t(x) = \alpha N_1(x) + (1-\alpha)N_2(x)$$

Further, we have

$$\log p_t(x) = \log(\alpha N_1(x) + (1-\alpha)N_2(x))$$

E.2. Lemmas for calculation of the score function

This subsection describes lemmas that are used for further calculation of the score function.

This lemma calculates the gradient of function N_i.

Lemma E.3. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$*.*
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$*.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0,1)$*.*
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*

Then, for $i \in \{1,2\}$*, we have*

$$\frac{\mathrm{d}N_i(x)}{\mathrm{d}x} = N_i(x)(-\Sigma_i(t)^{-1}(x-\mu_i(t)))$$

Proof. We can show

$$\begin{aligned} \frac{\mathrm{d}N_i(x)}{\mathrm{d}x} &= \frac{\mathrm{d}}{\mathrm{d}x}(\frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}}\exp(-\frac{1}{2}(x-\mu_i(t))^\top\Sigma_i(t)^{-1}(x-\mu_i(t)))) \\ &= N_i(x)\cdot\frac{\mathrm{d}}{\mathrm{d}x}(-\frac{1}{2}(x-\mu_i(t))^\top\Sigma_i(t)^{-1}(x-\mu_i(t))) \\ &= N_i(x)(-\frac{1}{2}\cdot 2\Sigma_i(t)^{-1}(x-\mu_i(t))) \\ &= N_i(x)(-\Sigma_i(t)^{-1}(x-\mu_i(t))) \end{aligned}$$

where the first step follows from Definition E.1, the second step follows from Fact C.1, the third step follows from Fact C.3, and the last step follows from simple algebra. □

This lemma calculates the gradient of function $p_t(x)$.

Lemma E.4. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0, 1)$.
- *Let* $p_t(x)$ *be defined as Definition E.2.*
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*

Then,

$$\frac{\mathrm{d} p_t(x)}{\mathrm{d} x} = \alpha N_1(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t))) + (1-\alpha) N_2(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t)))$$

Proof. We can show

$$\begin{aligned}\frac{\mathrm{d} p_t(x)}{\mathrm{d} x} &= \frac{\mathrm{d}}{\mathrm{d} x}(\alpha N_1(x) + (1-\alpha) N_2(x)) \\ &= \alpha \frac{\mathrm{d}}{\mathrm{d} x} N_1(x) + (1-\alpha) \frac{\mathrm{d}}{\mathrm{d} x} N_2(x) \\ &= \alpha N_1(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t))) + (1-\alpha) N_2(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t)))\end{aligned}$$

where the first step follows from Definition E.2, the second step follows from Fact C.1, and the last step follows from Lemma E.3. □

E.3. Calculation of the score function

Below we define $f(x)$ and $g(x)$ that simplify further calculation.

Definition E.5. *For further simplicity, we define the following functions:*

If the following conditions hold

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0, 1)$.
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*

We define

$$f(x) := \frac{\alpha N_1(x)}{\alpha N_1(x) + (1-\alpha) N_2(x)}$$

and

$$g(x) := \frac{(1-\alpha) N_2(x)}{\alpha N_1(x) + (1-\alpha) N_2(x)}$$

And it's clearly to see that $0 \leq f(x) \leq 1$, $0 \leq g(x) \leq 1$ *and* $f(x) + g(x) = 1$.

This lemma calculates the score function.

Lemma E.6. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0, 1)$.
- *Let* $p_t(x)$ *be defined as Definition E.2.*
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*
- *Let* $f(x), g(x)$ *be defined as Definition E.5.*

Then,

$$\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} = \frac{\alpha N_1(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t)))}{\alpha N_1(x) + (1-\alpha) N_2(x)} + \frac{(1-\alpha) N_2(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t)))}{\alpha N_1(x) + (1-\alpha) N_2(x)}$$

Proof. We can show

$$\begin{aligned}
\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} &= \frac{\mathrm{d} \log p_t(x)}{\mathrm{d} p_t(x)} \frac{\mathrm{d} p_t(x)}{\mathrm{d} x} \\
&= \frac{1}{p_t(x)} \frac{\mathrm{d} p_t(x)}{\mathrm{d} x} \\
&= \frac{1}{p_t(x)} (\alpha N_1(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t))) + (1-\alpha) N_2(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t)))) \\
&= \frac{\alpha N_1(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t)))}{\alpha N_1(x) + (1-\alpha) N_2(x)} + \frac{(1-\alpha) N_2(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t)))}{\alpha N_1(x) + (1-\alpha) N_2(x)} \\
&= f(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t)) + g(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t))
\end{aligned}$$

where the first step follows from Fact C.1, the second step follows from Fact C.1, the third step follows from Lemma E.4, the fourth step follows from Definition E.2 and the last step follows from Definition E.5.

□

E.4. Lemmas for the calculation of the upper bound of the score function

This section provides lemmas that are used in calculation of upper bound of the score function.

This lemma calculates the upper bound of function $\| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2$.

Lemma E.7. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

Then, for each $i \in \{1, 2\}$, we have

$$\| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2 \leq \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot (\|x\|_2 + \|\mu_i(t)\|_2)$$

Proof. We can show

$$\begin{aligned}
\| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2 &\leq \| - \Sigma_i(t)^{-1}\| \cdot \|x - \mu_i(t)\|_2 \\
&= \|\Sigma_i(t)^{-1}\| \cdot \|x - \mu_i(t)\|_2 \\
&= \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \mu(t)\|_2 \\
&\leq \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot (\|x\|_2 + \| - \mu_i(t)\|_2) \\
&= \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot (\|x\|_2 + \|\mu_i(t)\|_2)
\end{aligned}$$

where the first step follows from Fact C.2, the second step follows from Fact C.2, the third step follows from Fact C.2, the fourth step follows from Fact C.2, and the last step follows from simple algebra. □

E.5. Upper bound of the score function

This lemma calculates the upper bound of the score function.

Lemma E.8 (Linear growth)**.** *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $\alpha \in \mathbb{R}$ and $\alpha \in (0, 1)$.*
- *Let $p_t(x)$ be defined as Definition E.2.*
- *Let $f(x), g(x)$ be defined as Definition E.5.*
- *Let $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.*
- *Let $\mu_{\max} := \max\{\|\mu_1(t)\|_2, \|\mu_2(t)\|_2, 1\}$.*

Then,

$$\| \frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} \|_2 \leq \sigma_{\min}^{-1} \cdot \mu_{\max} \cdot (1 + \|x\|_2)$$

Proof. We can show

$$\begin{aligned}
\| \frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} \|_2 &= \| f(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t)) + g(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t))\|_2 \\
&\leq \| f(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t))\|_2 + \|g(x)(-\Sigma_2(t)^{-1}(x - \mu_2(t))_2\|_2 \\
&\leq \max_{i \in [2]} \| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2 \\
&\leq \max_{i \in [2]} (\frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot (\|x\|_2 + \|\mu_i(t)\|_2)) \\
&\leq \sigma_{\min}^{-1}(\mu_{\max} + \|x\|_2) \\
&\leq \sigma_{\min}^{-1} \cdot \mu_{\max} \cdot (1 + \|x\|_2)
\end{aligned}$$

where the first step follows from Lemma E.6, the second step follows from Fact C.2, the third step follows from $f(x)+g(x) = 1$ and $f(x), g(x) \geq 0$, the fourth step follows from Lemma E.7, the fifth step follows from definition of $\mu_{\max}$ and $\sigma_{\min}$, and the last step follows from $\mu_{\max} \geq 1$. □

E.6. Lemmas for Lipschitz calculation: upper bound of base functions

This section provides the lemmas of bounds of base functions that are used in calculation of Lipschitz of the score function.

This lemma calculate the upper bound of the function $\| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2$.

Lemma E.9. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $\|x - \mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in \{1, 2\}$.

Then, for each $i \in \{1, 2\}$, *we have*

$$\| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2 \leq \frac{R}{\sigma_{\min}(\Sigma_i(t))}$$

Proof. We can show

$$\begin{aligned}
\| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2 &\leq \| - \Sigma_i(t)^{-1}\| \cdot \|x - \mu_i(t)\|_2 \\
&= \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \mu_i(t)\|_2 \\
&\leq \frac{R}{\sigma_{\min}(\Sigma_i(t))}
\end{aligned}$$

where the first step follows from Fact C.2, the second step follows from Fact C.2, and the last step follows from $\|x - \mu_i(t)\|_2 \leq R$. □

This lemma calculate the lower bound of the function $(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_i(t))$.

Lemma E.10. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $\|x - \mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in \{1, 2\}$.
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in \{1, 2\}$.

Then,

$$(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_i(t)) \geq \frac{\beta^2}{\sigma_{\max}(\Sigma_i(t))}$$

Proof. We can show

$$\begin{aligned}\mathrm{LHS} \geq &\ \|x-\mu_i(t)\|_2^2 \cdot \sigma_{\min}(\Sigma_i(t)^{-1}) \\ = &\ \|x-\mu_i(t)\|_2^2 \cdot \frac{1}{\sigma_{\max}(\Sigma_i(t))} \\ \geq &\ \frac{\beta^2}{\sigma_{\max}(\Sigma_i(t))}\end{aligned}$$

where the first step follows from Fact C.2, the second step follows from Fact C.2, and the last step follows from $\|x-\mu_i(t)\|_2 \geq \beta$. □

This lemma calculate the upper bound of the function $\exp(-\frac{1}{2}(x-\mu_i(t))^\top \Sigma_i(t)^{-1}(x-\mu_i(t)))$.

Lemma E.11. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $\|x-\mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in \{1,2\}$.
- *Let* $\|x-\mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in \{1,2\}$.

Then,

$$\exp(-\frac{1}{2}(x-\mu_i(t))^\top \Sigma_i(t)^{-1}(x-\mu_i(t))) \leq \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))})$$

Proof. We can show

$$\begin{aligned}\mathrm{LHS} = &\ \exp(-\frac{1}{2}(x-\mu_i(t))^\top \Sigma_i(t)^{-1}(x-\mu_i(t))) \\ \leq &\ \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))})\end{aligned}$$

where the first step follows from Fact C.2, the second step follows from Lemma E.10. □

E.7. Lemmas for Lipschitz calculation: Lipschitz constant of base functions

This section provides the lemmas of Lipschitz constant of base functions that are used in calculation of Lipschitz of the score function.

This lemma calculates Lipschitz constant of function $\| - \Sigma_i(t)^{-1}(x-\mu_i(t)) - (-\Sigma_i(t)^{-1}(\widetilde{x}-\mu_i(t)))\|_2$.

Lemma E.12. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.

Then, for $i \in \{1,2\}$, *we have*

$$\| - \Sigma_i(t)^{-1}(x-\mu_i(t)) - (-\Sigma_i(t)^{-1}(\widetilde{x}-\mu_i(t)))\|_2 \leq \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x-\widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned}\mathrm{LHS} = &\ \| - \Sigma_i(t)^{-1}(x-\widetilde{x})\|_2 \\ \leq &\ \| - \Sigma_i(t)^{-1}\| \cdot \|x-\widetilde{x}\|_2 \\ = &\ \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x-\widetilde{x}\|_2\end{aligned}$$

where the first step follows from simple algebra, the second step follows from Fact C.2, and the last step follows from Fact C.2. □

This lemma calculates Lipschitz constant of function $| - \frac{1}{2}(x-\mu_i(t))^\top \Sigma_i(t)^{-1}(x-\mu_i(t)) - (-\frac{1}{2}(\widetilde{x}-\mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x}-\mu_i(t)))|$.

Lemma E.13. *If the following conditions hold*

- *Let $x, \widetilde{x} \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $\|x - \mu_i(t)\|_2 \leq R$, $\|\widetilde{x} - \mu_i(t)\|_2 \leq R$, where $R \geq 1$, for each $i \in \{1, 2\}$.*

Then, for each $i \in \{1, 2\}$, we have

$$|-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_i(t)) - (-\frac{1}{2}(\widetilde{x} - \mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))| \leq \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned}
\mathrm{LHS} \leq & \, |-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_i(t)) - (-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))| \\
& + |-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)) - (-\frac{1}{2}(\widetilde{x} - \mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))| \\
\leq & \, |-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \widetilde{x})| + |-\frac{1}{2}(x - \widetilde{x})^\top \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t))| \\
\leq & \, \frac{1}{2} \cdot \|\Sigma_i(t)^{-1}(x - \mu_i(t))\|_2 \cdot \|x - \widetilde{x}\|_2 + \frac{1}{2} \cdot \|\Sigma_i(t)^{-1}(x - \widetilde{x})\|_2 \cdot \|\widetilde{x} - \mu_i(t)\|_2 \\
\leq & \, \frac{1}{2} \cdot \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot R \cdot \|x - \widetilde{x}\|_2 + \frac{1}{2} \cdot \|\Sigma_i(t)^{-1}(x - \widetilde{x})\|_2 \cdot R \\
\leq & \, \frac{1}{2} \cdot \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot R \cdot \|x - \widetilde{x}\|_2 + \frac{1}{2} \cdot \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2 \cdot R \\
= & \, \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2
\end{aligned}$$

where the first step follows from Fact C.2, the second step follows from simple algebra, the third step follows from Fact C.2, the fourth step follows from $\|x - \mu_i(t)\|_2 \leq R$, $\|\widetilde{x} - \mu_i(t)\|_2 \leq R$, the fifth step follows from Lemma E.12, and the last step follows from simple algebra. □

This lemma calculates Lipschitz constant of function $|N_i(x) - N_i(\widetilde{x})|$.

Lemma E.14. *If the following conditions hold*

- *Let $x, \widetilde{x} \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $N_1(x), N_2(x)$ be defined as Definition E.1.*
- *Let $\|x - \mu_i(t)\|_2 \leq R$, where $R \geq 1$, for each $i \in \{1, 2\}$.*
- *Let $\|x - \mu_i(t)\|_2 \geq \beta$, where $\beta \in (0, 0.1)$, for each $i \in \{1, 2\}$.*

Then, for each $i \in \{1, 2\}$, we have

$$|N_i(x) - N_i(\widetilde{x})| \leq \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned}
|N_i(x) - N_i(\widetilde{x})| = & \, |\frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \exp(-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_i(t))) \\
& - \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \exp(-\frac{1}{2}(\widetilde{x} - \mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))| \\
= & \, \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \cdot |\exp(-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_i(t))) \\
& - \exp(-\frac{1}{2}(\widetilde{x} - \mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))| \\
\leq & \, \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))})
\end{aligned}$$

$$
\begin{aligned}
&\cdot |-\frac{1}{2}(x-\mu_i(t))^\top \Sigma_i(t)^{-1}(x-\mu_i(t)) - (-\frac{1}{2}(\widetilde{x}-\mu_i(t))^\top \Sigma_i(t)^{-1}(\widetilde{x}-\mu_i(t)))| \\
\leq & \frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x-\widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from Definition E.1, the second step follows from simple algebra, the third step follows from Fact C.9, and the last step follows from Lemma E.11. □

This lemma calculates Lipschitz constant of function $|\alpha N_1(x) + (1-\alpha)N_2(x) - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))|$.

Lemma E.15. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0,1)$.
- *Let* $\|x - \mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in \{1,2\}$.
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in \{1,2\}$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t))\}$.

Then, we have

$$
|\alpha N_1(x) + (1-\alpha)N_2(x) - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))| \leq \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2
$$

Proof. We can show

$$
\begin{aligned}
\mathrm{LHS} &= |\alpha N_1(x) - \alpha N_1(\widetilde{x}) + (1-\alpha)N_2(x) - (1-\alpha)N_2(\widetilde{x})| \\
&\leq \alpha|N_1(x) - N_1(\widetilde{x})| + (1-\alpha)|N_2(x) - N_2(\widetilde{x})| \\
&\leq \frac{\alpha}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_1(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_1(t))} \cdot \|x-\widetilde{x}\|_2 \\
&\quad + \frac{1-\alpha}{(2\pi)^{d/2}\det(\Sigma_2(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_2(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_2(t))} \cdot \|x-\widetilde{x}\|_2 \\
&\leq \frac{1}{(2\pi)^{d/2}} \max_{i\in[2]} \frac{1}{\det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x-\widetilde{x}\|_2 \\
&\leq \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from simple algebra, the second step follows from Fact C.2, the third step follows from Lemma E.14, the fourth step follows from $\alpha \in (0,1)$, and the last step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$. □

This lemma calculates Lipschitz constant of function $|(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}|$.

Lemma E.16. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0,1)$.
- *Let* $\|x - \mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in \{1,2\}$.
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in \{1,2\}$.
- *Let* $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$, *where* $\gamma \in (0, 0.1)$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t))\}$.

Then,

$$\begin{aligned}&|(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}|\\ \leq\ &\gamma^{-2} \cdot \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2\end{aligned}$$

Proof. We can show

$$\begin{aligned}\mathrm{LHS} \leq\ & (\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} \cdot (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}\\ & \cdot |\alpha N_1(x) + (1-\alpha)N_2(x) - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))|\\ \leq\ & \gamma^{-2} \cdot |\alpha N_1(x) + (1-\alpha)N_2(x) - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))|\\ \leq\ & \gamma^{-2} \cdot \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2\end{aligned}$$

where the first step follows from simple algebra, the second step follows from $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$, and the last step follows from Lemma E.15. □

E.8. Lemmas for Lipschitz calculation: $f(x)$

This lemma calculates Lipschitz constant of function $|f(x) - f(\widetilde{x})|$.

Lemma E.17. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*
- *Let* $f(x)$ *be defined as Definition E.5.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0, 1)$.
- *Let* $\|x - \mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in \{1, 2\}$.
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in \{1, 2\}$.
- *Let* $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$, *where* $\gamma \in (0, 0.1)$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t))\}$.

Then,

$$|f(x) - f(\widetilde{x})| \leq 2\alpha \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned}|f(x) - f(\widetilde{x})| =\ & |\frac{\alpha N_1(x)}{\alpha N_1(x) + (1-\alpha)N_2(x)} - \frac{\alpha N_1(\widetilde{x})}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})}|\\ \leq\ & |\frac{\alpha N_1(x)}{\alpha N_1(x) + (1-\alpha)N_2(x)} - \frac{\alpha N_1(x)}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})}|\\ & + |\frac{\alpha N_1(x)}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})} - \frac{\alpha N_1(\widetilde{x})}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})}|\\ =\ & \alpha \cdot |N_1(x)| \cdot |(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}|\\ & + \alpha \cdot |N_1(x) - N_1(\widetilde{x})| \cdot |(\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}|\end{aligned}$$

where the first step follows from Definition E.5, the second step follows from Fact C.2, and the last step follows from simple algebra.

For the first term in the above, we have

$$\alpha \cdot |N_1(x)| \cdot |(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}|$$

$$\begin{aligned}
&\leq \alpha \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}} \cdot |(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}| \\
&\leq \alpha \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}} \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
&\leq \alpha \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^{d}\det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \qquad (11)
\end{aligned}$$

where the first step follows from $N_1(x) \leq \frac{1}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}}$, the second step follows from Lemma E.16 and the last step follows from definition of $\det_{\min}$.

For the second term in the above, we have

$$\begin{aligned}
&\alpha \cdot |N_1(x) - N_1(\widetilde{x})| \cdot |(\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}| \\
&\leq \alpha \cdot \gamma^{-1} \cdot |N_1(x) - N_1(\widetilde{x})| \\
&\leq \alpha \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_1(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_1(t))} \cdot \|x - \widetilde{x}\|_2 \qquad (12)
\end{aligned}$$

where the first step follows from $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$, the second step follows from Lemma E.15.

Combining Eq. (11) and Eq. (12) together, we have

$$\begin{aligned}
|f(x) - f(\widetilde{x})| &\leq \alpha \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^{d}\det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
&\quad + \alpha \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_1(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_1(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_1(t))} \cdot \|x - \widetilde{x}\|_2 \\
&\leq \alpha \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^{d}\det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
&\quad + \alpha \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
&\leq 2\alpha \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^{d}\det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2
\end{aligned}$$

where the first step follows from the bound of the first term and the second term, the second step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$, and the third step follows from $\gamma < 0.1$. □

This lemma calculates Lipschitz constant of function $\|f(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t))) - f(\widetilde{x})(-\Sigma_1(t)^{-1}(\widetilde{x} - \mu_1(t)))\|_2$.

Lemma E.18. *If the following conditions hold*

- *Let $x, \widetilde{x} \in \mathbb{R}^d$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $N_1(x), N_2(x)$ be defined as Definition E.1.*
- *Let $f(x)$ be defined as Definition E.5.*
- *Let $\alpha \in \mathbb{R}$ and $\alpha \in (0,1)$.*
- *Let $\|x - \mu_i(t)\|_2 \leq R$, where $R \geq 1$, for each $i \in \{1,2\}$.*
- *Let $\|x - \mu_i(t)\|_2 \geq \beta$, where $\beta \in (0, 0.1)$, for each $i \in \{1,2\}$.*
- *Let $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$, where $\gamma \in (0, 0.1)$.*
- *Let $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.*
- *Let $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t))\}$.*
- *Let $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t))\}$.*

Then, we have

$$\begin{aligned}
&\|f(x)(-\Sigma_1(t)^{-1}(x - \mu_1(t))) - f(\widetilde{x})(-\Sigma_1(t)^{-1}(\widetilde{x} - \mu_1(t)))\|_2 \\
\leq &(\frac{1}{\sigma_{\min}} + 2\alpha \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^{d}\det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x - \widetilde{x}\|_2
\end{aligned}$$

Proof. We can show

$$
\begin{aligned}
\mathrm{LHS} \leq & \|f(x)(-\Sigma_1(t)^{-1}(x-\mu_1(t))) - f(x)(-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t)))\|_2 \\
& + \|f(x)(-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t))) - f(\widetilde{x})(-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t)))\|_2 \\
\leq & |f(x)| \cdot \|(-\Sigma_1(t)^{-1}(x-\mu_1(t))) - (-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t)))\|_2 \\
& + |f(x) - f(\widetilde{x})| \cdot \| - \Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t))\|_2
\end{aligned}
$$

where the first step follows from Fact C.2, the second step follows from Fact C.2.

For the first term in the above, we have

$$
\begin{aligned}
& |f(x)| \cdot \|(-\Sigma_1(t)^{-1}(x-\mu_1(t))) - (-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t)))\|_2 \\
\leq & \|(-\Sigma_1(t)^{-1}(x-\mu_1(t))) - (-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t)))\|_2 \\
\leq & \frac{1}{\sigma_{\min}(\Sigma_1(t))} \cdot \|x-\widetilde{x}\|_2
\end{aligned} \tag{13}
$$

where the first step follows from $f(x) \leq 1$, the second step follows from Lemma E.12.

For the second term in the above, we have

$$
\begin{aligned}
& |f(x) - f(\widetilde{x})| \cdot \| - \Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t))\|_2 \\
\leq & \frac{R}{\sigma_{\min}(\Sigma_1(t))} \cdot |f(x) - f(\widetilde{x})| \\
\leq & \frac{R}{\sigma_{\min}(\Sigma_1(t))} \cdot 2\alpha \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2
\end{aligned} \tag{14}
$$

where the first step follows from Lemma E.9, the second step follows from Lemma E.17.

Combining Eq. (13) and Eq. (14) together, we have

$$
\begin{aligned}
& \|f(x)(-\Sigma_1(t)^{-1}(x-\mu_1(t))) - f(\widetilde{x})(-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t)))\|_2 \\
\leq & \frac{1}{\sigma_{\min}(\Sigma_1(t))} \cdot \|x-\widetilde{x}\|_2 \\
& + \frac{R}{\sigma_{\min}(\Sigma_1(t))} \cdot 2\alpha \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2 \\
\leq & \frac{1}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2 \\
& + \frac{R}{\sigma_{\min}} \cdot 2\alpha \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2 \\
= & (\frac{1}{\sigma_{\min}} + 2\alpha \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x-\widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from the bound of the first term and the second term, the second step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$, and the last step follows from simple algebra. □

E.9. Lemmas for Lipschitz calculation: $g(x)$

This lemma calculates Lipschitz constant of function $|g(x) - g(\widetilde{x})|$.

Lemma E.19. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*
- *Let* $g(x)$ *be defined as Definition E.5.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0, 1)$.

- *Let* $\|x - \mu_i(t)\|_2 \leq R$*, where* $R \geq 1$*, for each* $i \in \{1, 2\}$.
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$*, where* $\beta \in (0, 0.1)$*, for each* $i \in \{1, 2\}$.
- *Let* $\alpha N_1(x) + (1 - \alpha)N_2(x) \geq \gamma$*, where* $\gamma \in (0, 0.1)$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t))\}$.

Then,

$$|g(x) - g(\widetilde{x})| \leq 2(1 - \alpha) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned} |g(x) - g(\widetilde{x})| &= |\frac{(1-\alpha)N_2(x)}{\alpha N_1(x) + (1-\alpha)N_2(x)} - \frac{(1-\alpha)N_2(\widetilde{x})}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})}| \\ &\leq |\frac{(1-\alpha)N_2(x)}{\alpha N_1(x) + (1-\alpha)N_2(x)} - \frac{(1-\alpha)N_2(x)}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})}| \\ &\quad + |\frac{(1-\alpha)N_2(x)}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})} - \frac{(1-\alpha)N_2(\widetilde{x})}{\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x})}| \\ &= (1-\alpha) \cdot |N_2(x)| \cdot |(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}| \\ &\quad + (1-\alpha) \cdot |N_2(x) - N_2(\widetilde{x})| \cdot |(\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}| \end{aligned}$$

where the first step follows from Definition E.5, the second step follows from Fact C.2, and the last step follows from simple algebra.

For the first term in the above, we have

$$\begin{aligned} &(1-\alpha) \cdot |N_2(x)| \cdot |(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}| \\ \leq &(1-\alpha) \cdot \frac{1}{(2\pi)^{d/2} \det(\Sigma_2(t))^{1/2}} \cdot |(\alpha N_1(x) + (1-\alpha)N_2(x))^{-1} - (\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}| \\ \leq &(1-\alpha) \cdot \frac{1}{(2\pi)^{d/2} \det(\Sigma_2(t))^{1/2}} \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\ \leq &(1-\alpha) \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^d \det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \end{aligned} \tag{15}$$

where the first step follows from $N_2(x) \leq \frac{1}{(2\pi)^{d/2} \det(\Sigma_2(t))^{1/2}}$, the second step follows from Lemma E.16.

For the second term in the above, we have

$$\begin{aligned} &(1-\alpha) \cdot |N_2(x) - N_2(\widetilde{x})| \cdot |(\alpha N_1(\widetilde{x}) + (1-\alpha)N_2(\widetilde{x}))^{-1}| \\ \leq &(1-\alpha) \cdot \gamma^{-1} \cdot |N_2(x) - N_2(\widetilde{x})| \\ \leq &(1-\alpha) \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2} \det(\Sigma_2(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_2(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_2(t))} \cdot \|x - \widetilde{x}\|_2 \end{aligned} \tag{16}$$

where the first step follows from $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$, the second step follows from Lemma E.15.

Combining Eq. (15) and Eq. (16) together, we have

$$\begin{aligned} |g(x) - g(\widetilde{x})| \leq &(1-\alpha) \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^d \det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\ &+ (1-\alpha) \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2} \det(\Sigma_2(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_2(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_2(t))} \cdot \|x - \widetilde{x}\|_2 \\ \leq &(1-\alpha) \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^d \det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \end{aligned}$$

$$
\begin{aligned}
&+ (1-\alpha)\cdot\gamma^{-1}\cdot\frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}\cdot\exp(-\frac{\beta^2}{2\sigma_{\max}})\cdot\frac{R}{\sigma_{\min}}\cdot\|x-\widetilde{x}\|_2\\
&\leq 2(1-\alpha)\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d\det_{\min}}+\frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}})\cdot\exp(-\frac{\beta^2}{2\sigma_{\max}})\cdot\frac{R}{\sigma_{\min}}\cdot\|x-\widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from the bound of the first term and the second term, the second step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$, and the last step follows from $\gamma < 0.1$. □

This lemma calculates Lipschitz constant of function $\|g(x)(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - g(\widetilde{x})(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2$.

Lemma E.20. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*
- *Let* $g(x)$ *be defined as Definition E.5.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0,1)$.
- *Let* $\|x-\mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in \{1,2\}$.
- *Let* $\|x-\mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in \{1,2\}$.
- *Let* $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$, *where* $\gamma \in (0, 0.1)$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t))\}$.

Then, we have

$$
\begin{aligned}
&\|g(x)(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - g(\widetilde{x})(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2\\
\leq&(\frac{1}{\sigma_{\min}}+2(1-\alpha)\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d\det_{\min}}+\frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}})\cdot\exp(-\frac{\beta^2}{2\sigma_{\max}})\cdot(\frac{R}{\sigma_{\min}})^2)\cdot\|x-\widetilde{x}\|_2
\end{aligned}
$$

Proof. We can show

$$
\begin{aligned}
\mathrm{LHS} \leq& \|g(x)(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - g(x)(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2\\
&+\|g(x)(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t))) - f(\widetilde{x})(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2\\
\leq& |g(x)|\cdot\|(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - (-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2\\
&+|g(x)-g(\widetilde{x})|\cdot\|-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t))\|_2
\end{aligned}
$$

where the first step follows from Fact C.2, the second step follows from Fact C.2.

For the first term in the above, we have

$$
\begin{aligned}
&|g(x)|\cdot\|(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - (-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2\\
\leq&\|(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - (-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2\\
\leq&\frac{1}{\sigma_{\min}(\Sigma_2(t))}\cdot\|x-\widetilde{x}\|_2
\end{aligned}
\tag{17}
$$

where the first step follows from $g(x) \leq 1$, the second step follows from Lemma E.12.

For the second term in the above, we have

$$
\begin{aligned}
&|g(x)-g(\widetilde{x})|\cdot\|-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t))\|_2\\
\leq&\frac{R}{\sigma_{\min}(\Sigma_2(t))}\cdot|g(x)-g(\widetilde{x})|\\
\leq&\frac{R}{\sigma_{\min}(\Sigma_2(t))}\cdot 2(1-\alpha)\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d\det_{\min}}+\frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}})\cdot\exp(-\frac{\beta^2}{2\sigma_{\max}})\cdot\frac{R}{\sigma_{\min}}\cdot\|x-\widetilde{x}\|_2
\end{aligned}
\tag{18}
$$

where the first step follows from Lemma E.9, the second step follows from Lemma E.19.

Combining Eq. (17) and Eq. (18) together, we have

$$\begin{aligned}
&\|g(x)(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - g(\widetilde{x})(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2 \\
\leq &\frac{1}{\sigma_{\min}(\Sigma_2(t))} \cdot \|x-\widetilde{x}\|_2 \\
&+ \frac{R}{\sigma_{\min}(\Sigma_2(t))} \cdot 2(1-\alpha)\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2 \\
\leq &\frac{1}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2 \\
&+ \frac{R}{\sigma_{\min}} \cdot 2(1-\alpha)\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x-\widetilde{x}\|_2 \\
= &(\frac{1}{\sigma_{\min}} + 2(1-\alpha)\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x-\widetilde{x}\|_2
\end{aligned}$$

where the first step follows from the bound of the first term and the second term, the second step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$, and the last step follows from simple algebra. □

E.10. Lipschitz constant of the score function

This lemma calculates the Lipschitz constant of the score funciton.

Lemma E.21 (Lipschitz). *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.
- *Let* $N_1(x), N_2(x)$ *be defined as Definition E.1.*
- *Let* $\alpha \in \mathbb{R}$ *and* $\alpha \in (0,1)$.
- *Let* $p_t(x)$ *be defined as Definition E.2.*
- *Let* $f(x), g(x)$ *be defined as Definition E.5.*
- *Let* $\|x-\mu_i(t)\|_2 \leq R$*, where* $R \geq 1$*, for each* $i \in \{1,2\}$.
- *Let* $\|x-\mu_i(t)\|_2 \geq \beta$*, where* $\beta \in (0, 0.1)$*, for each* $i \in \{1,2\}$.
- *Let* $\alpha N_1(x) + (1-\alpha)N_2(x) \geq \gamma$*, where* $\gamma \in (0, 0.1)$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t))\}$.

Then,

$$\|\frac{\mathrm{d}\log p_t(x)}{\mathrm{d}x} - \frac{\mathrm{d}\log p_t(\widetilde{x})}{\mathrm{d}\widetilde{x}}\|_2 \leq (\frac{2}{\sigma_{\min}} + \frac{2R^2}{\gamma^2\sigma_{\min}^2}\cdot(\frac{1}{(2\pi)^d\det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}})\cdot\exp(-\frac{\beta^2}{2\sigma_{\max}}))\cdot\|x-\widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned}
\mathrm{LHS} = &\|f(x)(-\Sigma_1(t)^{-1}(x-\mu_1(t))) + g(x)(-\Sigma_2(t)^{-1}(x-\mu_2(t))) \\
&- (f(\widetilde{x})(-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t))) + g(\widetilde{x})(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t))))\|_2 \\
\leq &\|f(x)(-\Sigma_1(t)^{-1}(x-\mu_1(t))) - f(\widetilde{x})(-\Sigma_1(t)^{-1}(\widetilde{x}-\mu_1(t)))\|_2 \\
&+ \|g(x)(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - g(\widetilde{x})(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2 \\
\leq &(\frac{1}{\sigma_{\min}} + 2\alpha\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x-\widetilde{x}\|_2 \\
&+ \|g(x)(-\Sigma_2(t)^{-1}(x-\mu_2(t))) - g(\widetilde{x})(-\Sigma_2(t)^{-1}(\widetilde{x}-\mu_2(t)))\|_2 \\
\leq &(\frac{1}{\sigma_{\min}} + 2\alpha\cdot\gamma^{-2}\cdot(\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x-\widetilde{x}\|_2
\end{aligned}$$

$$+ (\frac{1}{\sigma_{\min}} + 2(1-\alpha) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x - \widetilde{x}\|_2$$

$$= (\frac{2}{\sigma_{\min}} + \frac{2R^2}{\gamma^2 \sigma_{\min}^2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}})) \cdot \|x - \widetilde{x}\|_2$$

where the first step follows from Lemma E.6, the second step follows from Fact C.2, the third step follows from Lemma E.18, the fourth step follows from Lemma E.20, and the last step follows from simple algebra. □

F. A General Version for k Gaussian

In this section we consider a more general case of k mixture of Gaussians.

- Section F.1 provides the definition for k mixture of Gaussians.
- Section F.2 provides the expression of the score function.
- Section F.3 provides the upper bound of the score function.
- Section F.4 provides lemmas that are used in further calculation of Lipschitz constant.
- Section F.5 provides the Lipschitz constant for k mixture of Gaussians.

F.1. Definitions

Let $i \in [k]$. Let $\alpha_i(t) \in (0,1)$, $\sum_{i=1}^k \alpha_i(t) = 1$, and is a function of time t. Consider p_t such that

$$p_t(x) = \Pr_{x' \sim \sum_{i=1}^k \alpha_i(t) \mathcal{N}(\mu_i(t), \Sigma_i(t))} [x' = x]$$

where $\mu_i(t) \in \mathbb{R}^d$, $\Sigma_i(t) \in \mathbb{R}^{d \times d}$ and they are derivative to t and $\Sigma_i(t)$ is a symmetric p.s.d. matrix whose the smallest singular value is always larger than a fixed value $\sigma_{\min} > 0$.

Below we define the pdf for a single multivariate Gaussian.

Definition F.1. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $i \in [k]$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*

We define

$$N_i(x) := \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \exp(-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_1(t)))$$

This is the pdf *of a single Gaussian so it's clearly to see that $0 \leq N_i \leq \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}}$ since $N_i(x)$ takes maximum when $x = \mu_i$.*

Below we define the pdf for k mixtures of Gaussians.

Definition F.2. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $i \in [k]$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $\alpha_i(t) \in \mathbb{R}$, $\sum_{i=1}^k \alpha_i(t) = 1$, and $\alpha_i(t) \in (0,1)$.*
- *Let $N_i(x)$ be defined as Definition F.1.*

We define

$$p_t(x) := \sum_{i=1}^k \frac{\alpha_i(t)}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \exp(-\frac{1}{2}(x - \mu_i(t))^\top \Sigma_i(t)^{-1}(x - \mu_i(t)))$$

This can be further rewritten as follows:

$$p_t(x) = \sum_{i=1}^k \alpha_i(t) N_i(x)$$

Further, we have

$$\log p_t(x) = \log(\sum_{i=1}^k \alpha_i(t) N_i(x))$$

This lemma calculates the gradient of pdf for k mixture of Gaussians.

Lemma F.3. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $i \in [k]$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $\alpha_i(t) \in \mathbb{R}$, $\sum_{i=1}^k \alpha_i(t) = 1$, and $\alpha_i(t) \in (0,1)$.*
- *Let $N_i(x)$ be defined as Definition F.1.*
- *Let $p_t(x)$ be defined as Definition F.2*

We have

$$\frac{\mathrm{d} p_t(x)}{\mathrm{d} x} = \sum_{i=1}^k \alpha_i(t) N_i(x) (-\Sigma_i(t)^{-1}(x - \mu_i(t)))$$

Proof. We can show

$$\begin{aligned} \frac{\mathrm{d} p_t(x)}{\mathrm{d} x} &= \frac{\mathrm{d}}{\mathrm{d} x} \sum_{i=1}^k \alpha_i(t) N_i(x) \\ &= \sum_{i=1}^k \alpha_i(t) \frac{\mathrm{d} N_i(x)}{\mathrm{d} x} \\ &= \sum_{i=1}^k \alpha_i(t) N_i(x) (-\Sigma_i(t)^{-1}(x - \mu_i(t))) \end{aligned}$$

where the first step follows from Definition F.2, the second step follows from Fact C.1, and the last step follows from Lemma E.3.

□

Below we define f_i that simplifies further calculation.

Definition F.4. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $i \in [k]$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $\alpha_i(t) \in \mathbb{R}$, $\sum_{i=1}^k \alpha_i(t) = 1$, and $\alpha_i(t) \in (0,1)$.*
- *Let $N_i(x)$ be defined as Definition F.1.*

For further simplicity, we define

$$f_i(x) := \frac{\alpha_i(t) N_i(x)}{\sum_{i=1}^k \alpha_i(t) N_i(x)}$$

It's clearly to see that $0 \leq f_i(x) \leq 1$ and $\sum_{i=1}^k f_i(x) = 1$

F.2. Calculation of the score function

This lemma calculates the score function for k mixture of Gaussians.

Lemma F.5. *If the following conditions hold*

- *Let $x \in \mathbb{R}^d$.*
- *Let $i \in [k]$.*

- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.
- *Let* $\alpha_i(t) \in \mathbb{R}$*,* $\sum_{i=1}^k \alpha_i(t) = 1$*, and* $\alpha_i(t) \in (0, 1)$.
- *Let* $N_i(x)$ *be defined as Definition F.1.*
- *Let* $p_t(x)$ *be defined as Definition F.2.*
- *Let* $f_i(x)$ *be defined as Definition F.4.*

We have

$$\frac{\mathrm{d} \log p_t(x)}{\mathrm{d}x} = \sum_{i=1}^k f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t)))$$

Proof. We can show

$$\begin{aligned}
\frac{\mathrm{d} \log p_t(x)}{\mathrm{d}x} &= \frac{\mathrm{d} \log p_t(x)}{\mathrm{d}p_t(x)} \frac{\mathrm{d}p_t(x)}{\mathrm{d}x} \\
&= \frac{1}{p_t(x)} \frac{\mathrm{d}p_t(x)}{\mathrm{d}x} \\
&= \frac{1}{p_t(x)} \sum_{i=1}^k \alpha_i(t) N_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t))) \\
&= \frac{\sum_{i=1}^k \alpha_i(t) N_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t)))}{\sum_{i=1}^k \alpha_i(t) N_i(x)} \\
&= \sum_{i=1}^k f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t)))
\end{aligned}$$

where the first step follows from Fact C.1, the second step follows from Fact C.1, the third step follows from Lemma F.3, the fourth step follows from Definition F.2, and the last step follows from Definition F.4. □

F.3. Upper bound of the score function

This lemma calculates upper bound of the score function for k mixture of Gaussians.

Lemma F.6. *If the following conditions hold*

- *Let* $x \in \mathbb{R}^d$.
- *Let* $i \in [k]$.
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$.
- *Let* $\alpha_i(t) \in \mathbb{R}$*,* $\sum_{i=1}^k \alpha_i(t) = 1$*, and* $\alpha_i(t) \in (0, 1)$.
- *Let* $p_t(x)$ *be defined as Definition F.2.*
- *Let* $f_i(x)$ *be defined as Definition F.4.*
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t)), \ldots, \sigma_{\min}(\Sigma_k(t))\}$.
- *Let* $\mu_{\max} := \max\{1, \|\mu_1(t)\|_2, \|\mu_2(t)\|_2, \ldots, \|\mu_k(t)\|_2\}$.

Then, we have

$$\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d}x}\|_2 \leq \sigma_{\min}^{-1} \cdot \mu_{\max} \cdot (1 + \|x\|_2)$$

Proof. We can show

$$\begin{aligned}
\|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d}x}\|_2 &= \|\sum_{i=1}^k f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t)))\|_2 \\
&\leq \sum_{i=1}^k f_i(x) \| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2 \\
&\leq \max_{i \in [k]} \| - \Sigma_i(t)^{-1}(x - \mu_i(t))\|_2
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{i\in[k]} \frac{1}{\sigma_{\min}(\Sigma_i(t))} \cdot (\|x\|_2 + \|\mu_i(t)\|_2) \\
&\leq \sigma_{\min}^{-1}(\mu_{\max} + \|x\|_2) \\
&\leq \sigma_{\min}^{-1} \cdot \mu_{\max} \cdot (1 + \|x\|_2)
\end{aligned}
$$

where the first step follows from Lemma F.5, the second step follows from triangle inequality, the third step follows from $\sum_{i=1}^k f_i(x) = 1$ and $f_i(x) \geq 0$, the fourth step follows from Lemma E.7, the fifth step follows from definition of $\mu_{\max}$ and $\sigma_{\min}$, and the last step follows from $\mu_{\max} \geq 1$. □

F.4. Lemmas for Lipshitz calculation

This section provides lemmas for calculation of Lipschitz constant of the score function for k mixture of Gaussians.

This lemma calculates Lipschitz constant of function $|\sum_{i=1}^k \alpha_i(t) N_i(x) - \sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x})|$.

Lemma F.7. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $i \in [k]$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $\alpha_i(t) \in \mathbb{R}$, $\sum_{i=1}^k \alpha_i(t) = 1$, *and* $\alpha_i(t) \in (0, 1)$.
- *Let* $N_i(x)$ *be defined as Definition F.1.*
- *Let* $\|x - \mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in [k]$.
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in [k]$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t)), \ldots, \sigma_{\min}(\Sigma_k(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t)), \ldots, \sigma_{\max}(\Sigma_k(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t)), \ldots, \det(\Sigma_k(t))\}$.

Then, we have

$$
|\sum_{i=1}^k \alpha_i(t) N_i(x) - \sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x})| \leq \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2
$$

Proof. We can show

$$
\begin{aligned}
\mathrm{LHS} &= |\sum_{i=1}^k \alpha_i(t)(N_i(x) - N_i(\widetilde{x}))| \\
&\leq \sum_{i=1}^k \alpha_i(t) |N_i(x) - N_i(\widetilde{x})| \\
&\leq \sum_{i=1}^k \alpha_i(t) \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2 \\
&\leq \max_{i\in[k]} \frac{1}{(2\pi)^{d/2} \det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2 \\
&\leq \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from simple algebra, the second step follows from Fact C.2, the third step follows from Lemma E.14, the fourth step follows from $\sum_{i=1}^k \alpha_i(t) = 1$, and $\alpha_i(t) \in (0, 1)$, and the last step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$. □

This lemma calculates Lipschitz constant of function $|(\sum_{i=1}^k \alpha_i(t) N_i(x))^{-1} - (\sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x}))^{-1}|$.

Lemma F.8. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $i \in [k]$.

- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$*.*
- *Let* $\alpha_i(t) \in \mathbb{R}$*,* $\sum_{i=1}^k \alpha_i(t) = 1$*, and* $\alpha_i(t) \in (0, 1)$*.*
- *Let* $N_i(x)$ *be defined as Definition F.1.*
- *Let* $\|x - \mu_i(t)\|_2 \leq R$*, where* $R \geq 1$*, for each* $i \in [k]$*.*
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$*, where* $\beta \in (0, 0.1)$*, for each* $i \in [k]$*.*
- *Let* $\sum_{i=1}^k \alpha_i(t) N_i(x) \geq \gamma$*, where* $\gamma \in (0, 0.1)$*.*
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t)), \dots, \sigma_{\min}(\Sigma_k(t))\}$*.*
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t)), \dots, \sigma_{\max}(\Sigma_k(t))\}$*.*
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t)), \dots, \det(\Sigma_k(t))\}$*.*

Then, we have

$$|(\sum_{i=1}^k \alpha_i(t) N_i(x))^{-1} - (\sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x}))^{-1}| \leq \gamma^{-2} \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned} \mathrm{LHS} &= (\sum_{i=1}^k \alpha_i(t) N_i(x))^{-1} \cdot (\sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x}))^{-1} \cdot |\sum_{i=1}^k \alpha_i(t) N_i(x) - \sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x})| \\ &\leq \gamma^{-2} \cdot |\sum_{i=1}^k \alpha_i(t) N_i(x) - \sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x})| \\ &\leq \gamma^{-2} \cdot \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \end{aligned}$$

where the first step follows from simple algebra, the second step follows from $\sum_{i=1}^k \alpha_i(t) N_i(x) \geq \gamma$, and the last step follows from Lemma F.7. □

This lemma calculates Lipschitz constant of function $|f_i(x) - f_i(\widetilde{x})|$.

Lemma F.9. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$*.*
- *Let* $i \in [k]$*.*
- *Let* $t \in \mathbb{R}$*, and* $t \geq 0$*.*
- *Let* $\alpha_i(t) \in \mathbb{R}$*,* $\sum_{i=1}^k \alpha_i(t) = 1$*, and* $\alpha_i(t) \in (0, 1)$*.*
- *Let* $N_i(x)$ *be defined as Definition F.1.*
- *Let* $f_i(x)$ *be defined as Definition F.4.*
- *Let* $\|x - \mu_i(t)\|_2 \leq R$*, where* $R \geq 1$*, for each* $i \in [k]$*.*
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$*, where* $\beta \in (0, 0.1)$*, for each* $i \in [k]$*.*
- *Let* $\sum_{i=1}^k \alpha_i(t) N_i(x) \geq \gamma$*, where* $\gamma \in (0, 0.1)$*.*
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t)), \dots, \sigma_{\min}(\Sigma_k(t))\}$*.*
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t)), \dots, \sigma_{\max}(\Sigma_k(t))\}$*.*
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t)), \dots, \det(\Sigma_k(t))\}$*.*

Then, for each $i \in [k]$*, we have*

$$|f_i(x) - f_i(\widetilde{x})| \leq 2\alpha_i(t) \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2$$

Proof. We can show

$$\begin{aligned} |f_i(x) - f_i(\widetilde{x})| &= |\alpha_i(t) N_i(x) \cdot (\sum_{i=1}^k \alpha_i(t) N_i(x))^{-1} - \alpha_i(t) N_i(\widetilde{x}) \cdot (\sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x}))^{-1}| \\ &\leq |\alpha_i(t) N_i(x) \cdot (\sum_{i=1}^k \alpha_i(t) N_i(x))^{-1} - \alpha_i(t) N_i(x) \cdot (\sum_{i=1}^k \alpha_i(t) N_i(\widetilde{x}))^{-1}| \end{aligned}$$

$$
\begin{aligned}
& + |\alpha_i(t)N_i(x) \cdot (\sum_{i=1}^k \alpha_i(t)N_i(\widetilde{x}))^{-1} - \alpha_i(t)N_i(\widetilde{x}) \cdot (\sum_{i=1}^k \alpha_i(t)N_i(\widetilde{x}))^{-1}| \\
\leq & \alpha_i(t)N_i(x) \cdot |(\sum_{i=1}^k \alpha_i(t)N_i(x))^{-1} - (\sum_{i=1}^k \alpha_i(t)N_i(\widetilde{x}))^{-1}| \\
& + \alpha_i(t)(\sum_{i=1}^k \alpha_i(t)N_i(\widetilde{x}))^{-1}|N_i(x) - N_i(\widetilde{x})|
\end{aligned}
$$

where the first step follows from Definition F.4, the second step follows from Fact C.2, and the last step follows from simple algebra.

For the first term in the above, we have

$$
\begin{aligned}
& \alpha_i(t)N_i(x) \cdot |(\sum_{i=1}^k \alpha_i(t)N_i(x))^{-1} - (\sum_{i=1}^k \alpha_i(t)N_i(\widetilde{x}))^{-1}| \\
\leq & \alpha_i(t) \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}} \cdot |(\sum_{i=1}^k \alpha_i(t)N_i(x))^{-1} - (\sum_{i=1}^k \alpha_i(t)N_i(\widetilde{x}))^{-1}| \\
\leq & \alpha_i(t) \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}} \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
\leq & \alpha_i(t) \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^d \det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \qquad (19)
\end{aligned}
$$

where the first step follows from $N_i(x) \leq \frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}}$, the second step follows from Lemma F.8, and the last step follows from definition of $\det_{\min}$.

For the second term in the above, we have

$$
\begin{aligned}
& \alpha_i(t)(\sum_{i=1}^k \alpha_i(t)N_i(\widetilde{x}))^{-1}|N_i(x) - N_i(\widetilde{x})| \\
\leq & \alpha_i(t) \cdot \gamma^{-1} \cdot |N_i(x) - N_i(\widetilde{x})| \\
\leq & \alpha_i(t) \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2 \qquad (20)
\end{aligned}
$$

where the first step follows from $\sum_{i=1}^k \alpha_i(t)N_i(x) \geq \gamma$, the second step follows from Lemma F.7.

Combining Eq. (19) and Eq. (20) together, we have

$$
\begin{aligned}
|f_i(x) - f_i(\widetilde{x})| \leq & \alpha_i(t) \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^d \det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
& + \alpha_i(t) \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2}\det(\Sigma_i(t))^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}(\Sigma_i(t))}) \cdot \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2 \\
\leq & \alpha_i(t) \cdot \gamma^{-2} \cdot \frac{1}{(2\pi)^d \det_{\min}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
& + \alpha_i(t) \cdot \gamma^{-1} \cdot \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}} \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
\leq & 2\alpha_i(t) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from the bound of the first term and the second term, the second step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$, and the last step follows from $\gamma < 0.1$. □

This lemma calculates Lipschitz constant of function $\|f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - f_i(\widetilde{x})(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2$.

Lemma F.10. *If the following conditions hold*

- *Let* $x, \widetilde{x} \in \mathbb{R}^d$.
- *Let* $i \in [k]$.
- *Let* $t \in \mathbb{R}$, *and* $t \geq 0$.
- *Let* $\alpha_i(t) \in \mathbb{R}$, $\sum_{i=1}^k \alpha_i(t) = 1$, *and* $\alpha_i(t) \in (0,1)$.
- *Let* $N_i(x)$ *be defined as Definition F.1.*
- *Let* $f_i(x)$ *be defined as Definition F.4.*
- *Let* $\|x - \mu_i(t)\|_2 \leq R$, *where* $R \geq 1$, *for each* $i \in [k]$.
- *Let* $\|x - \mu_i(t)\|_2 \geq \beta$, *where* $\beta \in (0, 0.1)$, *for each* $i \in [k]$.
- *Let* $\sum_{i=1}^k \alpha_i(t) N_i(x) \geq \gamma$, *where* $\gamma \in (0, 0.1)$.
- *Let* $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t)), \dots, \sigma_{\min}(\Sigma_k(t))\}$.
- *Let* $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t)), \dots, \sigma_{\max}(\Sigma_k(t))\}$.
- *Let* $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t)), \dots, \det(\Sigma_k(t))\}$.

Then, for each $i \in [k]$, *we have*

$$\begin{aligned}
&\|f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - f_i(\widetilde{x})(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
\leq & (\frac{|f_i(x)|}{\sigma_{\min}} + 2\alpha_i(t) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x - \widetilde{x}\|_2
\end{aligned}$$

Proof. We can show

$$\begin{aligned}
\mathrm{LHS} \leq & \|f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - f_i(x)(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
& + \|f_i(x)(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t))) - f_i(\widetilde{x})(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
\leq & |f_i(x)| \cdot \|(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - (-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
& + |f_i(x) - f_i(\widetilde{x})| \cdot \| - \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t))\|_2
\end{aligned}$$

where the first step follows from Fact C.2, the second step follows from Fact C.2.

For the first term in the above, we have

$$\begin{aligned}
& |f_i(x)| \cdot \|(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - (-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
\leq & \frac{|f_i(x)|}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2 \qquad (21)
\end{aligned}$$

where the first step follows from Lemma E.12.

For the second term in the above, we have

$$\begin{aligned}
& |f_i(x) - f_i(\widetilde{x})| \cdot \| - \Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t))\|_2 \\
\leq & \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot |f_i(x) - f_i(\widetilde{x})| \\
\leq & \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot 2\alpha_i(t) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \qquad (22)
\end{aligned}$$

where the first step follows from Lemma E.9, the second step follows from Lemma F.9.

Combining Eq. (21) and Eq. (22) together, we have

$$\begin{aligned}
& \|f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - f_i(\widetilde{x})(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
\leq & \frac{|f_i(x)|}{\sigma_{\min}(\Sigma_i(t))} \cdot \|x - \widetilde{x}\|_2 \\
& + \frac{R}{\sigma_{\min}(\Sigma_i(t))} \cdot 2\alpha_i(t) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
\leq & \frac{|f_i(x)|}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2
\end{aligned}$$

$$
\begin{aligned}
& + \frac{R}{\sigma_{\min}} \cdot 2\alpha_i(t) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot \frac{R}{\sigma_{\min}} \cdot \|x - \widetilde{x}\|_2 \\
= & (\frac{|f_i(x)|}{\sigma_{\min}} + 2\alpha_i(t) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x - \widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from the bound of the first term and the second term, the second step follows from the definition of $\det_{\min}, \sigma_{\max}, \sigma_{\min}$, and the last step follows from simple algebra. □

F.5. Lipschitz constant of the score function

This lemma calculates Lipschitz constant of the score function for k mixture of Gaussians.

Lemma F.11. *If the following conditions hold*

- *Let $x, \widetilde{x} \in \mathbb{R}^d$.*
- *Let $i \in [k]$.*
- *Let $t \in \mathbb{R}$, and $t \geq 0$.*
- *Let $\alpha_i(t) \in \mathbb{R}$, $\sum_{i=1}^k \alpha_i(t) = 1$, and $\alpha_i(t) \in (0, 1)$.*
- *Let $N_i(x)$ be defined as Definition F.1.*
- *Let $p_t(x)$ be defined as Definition F.2.*
- *Let $f_i(x)$ be defined as Definition F.4.*
- *Let $\|x - \mu_i(t)\|_2 \leq R$, where $R \geq 1$, for each $i \in [k]$.*
- *Let $\|x - \mu_i(t)\|_2 \geq \beta$, where $\beta \in (0, 0.1)$, for each $i \in [k]$.*
- *Let $\sum_{i=1}^k \alpha_i(t) N_i(x) \geq \gamma$, where $\gamma \in (0, 0.1)$.*
- *Let $\sigma_{\min} := \min\{\sigma_{\min}(\Sigma_1(t)), \sigma_{\min}(\Sigma_2(t)), \dots, \sigma_{\min}(\Sigma_k(t))\}$.*
- *Let $\sigma_{\max} := \max\{\sigma_{\max}(\Sigma_1(t)), \sigma_{\max}(\Sigma_2(t)), \dots, \sigma_{\max}(\Sigma_k(t))\}$.*
- *Let $\det_{\min} := \min\{\det(\Sigma_1(t)), \det(\Sigma_2(t)), \dots, \det(\Sigma_k(t))\}$.*

Then, we have

$$
\begin{aligned}
& \|\frac{\mathrm{d} \log p_t(x)}{\mathrm{d} x} - \frac{\mathrm{d} \log p_t(\widetilde{x})}{\mathrm{d} \widetilde{x}}\|_2 \\
\leq & (\frac{1}{\sigma_{\min}} + \frac{2R^2}{\gamma^2 \sigma_{\min}^2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}})) \cdot \|x - \widetilde{x}\|_2
\end{aligned}
$$

Proof. We can show

$$
\begin{aligned}
& \mathrm{LHS} \\
= & \|\sum_{i=1}^k f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - \sum_{i=1}^k f_i(\widetilde{x})(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
\leq & \sum_{i=1}^k \|f_i(x)(-\Sigma_i(t)^{-1}(x - \mu_i(t))) - f_i(\widetilde{x})(-\Sigma_i(t)^{-1}(\widetilde{x} - \mu_i(t)))\|_2 \\
\leq & \sum_{i=1}^k (\frac{|f_i(x)|}{\sigma_{\min}} + 2\alpha_i(t) \cdot \gamma^{-2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}}) \cdot (\frac{R}{\sigma_{\min}})^2) \cdot \|x - \widetilde{x}\|_2 \\
= & (\frac{1}{\sigma_{\min}} + \frac{2R^2}{\gamma^2 \sigma_{\min}^2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2} \det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}})) \cdot \|x - \widetilde{x}\|_2
\end{aligned}
$$

where the first step follows from Lemma F.5, the second step follows from triangle inequality, the third step follows from Lemma F.10, and the last step follows from $\sum_{i=1}^k |f_i(x)| = \sum_{i=1}^k f_i(x) = 1$ and $\sum_{i=1}^k \alpha_i(t) = 1$. □

G. Tools From Previous Work

In this section, we present several key theoretical results from previous work that serve as building blocks for our analysis. We begin with important assumptions about score-based diffusion models, followed by theorems establishing error bounds for different numerical solvers.

Assumption G.1 (Lipschitz score, Assumption 1 in [13], page 6)**.** *For all $t \geq 0$, the score $\nabla \log p_t$ is L-Lipschitz.*

Assumption G.2 (Second momentum bound, Assumption 2 in [13], page 6 and Assumption 2 in [10], page 6)**.** *We assume that $m_2^2 := M_2 := \mathbb{E}_{p_0}[\|x\|_2^2] < \infty$.*

Assumption G.3 (Smooth data distributions, Assumption 4 in [10], page 10)**.** *The data distribution admits a density $p_0 \in C^2(\mathbb{R}^d)$ and $\nabla \log p_0$ is L-Lipschitz, where C^2 means second-order differentiable.*

Remark G.4. *We can notice $M_2 = m_2^2$. The theorems from [10] use M_2 for KL divergence. The theorems form [13] and [14] use $m_2 = \sqrt{m_2^2} = \sqrt{M_2}$ for total variance, because of Pinsker's inequality (Lemma 5.3).*

Each theorem presented below provides different perspectives on bounding the error between the learned distribution and the target distribution. Theorem G.5 focuses on total variation distance for DDPM, while Theorems G.6 and G.7 analyze KL divergence under different conditions. Theorems G.8 and G.9 establish bounds for the ODE-based solvers DPOM and DPUM respectively. We first state a tool from previous work [13].

Theorem G.5 (DDPM, Theorem 2 in [13], page 7)**.** *Suppose that Assumptions G.1, G.2 and 5.1 hold. Let $\widehat{q}_T$ be the output of DDPM algorithm at times T, and suppose that the step size $h := T/N$ satisfies $h \lesssim 1/L$, where $L \geq 1$. Then, it holds that*

$$\begin{aligned} \mathrm{TV}(\widehat{q}_T, p_0) \lesssim & \underbrace{\sqrt{\mathrm{KL}(p_0 \| \mathcal{N}(0, I))} \exp(-T)}_{\text{convergence of forward process}} \\ & + \underbrace{(L\sqrt{dh} + Lm_2 h)\sqrt{T}}_{\text{discretization error}} + \underbrace{\epsilon_0 \sqrt{T}}_{\text{score estimation error}} . \end{aligned}$$

Then, we state a tool from previous work [10].

Theorem G.6 (Theorem 1 in [10])**.** *Suppose that Assumptions G.1, G.2, 5.1 hold. If $L \geq 1$, $h_k \leq 1$ for $k \in [N]$ and $T \geq 1$, using uniform discretization points yields the following: (1) Using Exponential Integrator scheme (8), we have $\mathrm{KL}(p_0 \| \widehat{q}_T) \lesssim (M_2 + d)\exp(-T) + T\epsilon_0^2 + \frac{dT^2L^2}{N}$. In particular, choosing $T = \Theta(\log(dM_2/\epsilon_0^2))$ and $N = \Theta(dT^2L^2/\epsilon_0^2)$ makes this $\widetilde{O}(\epsilon_0^2)$. (2) Using the Euler-Maruyama scheme (7), we have $\mathrm{KL}(p_0 \| \widehat{q}_T) \lesssim (M_2 + d)\exp(-T) + T\epsilon_0^2 + \frac{dT^2L^2}{N} + \frac{T^3 M_2}{N^2}$.*

Theorem G.7 (Theorem 5 in [10], page 10)**.** *There is a universal constant K such that the following holds. Under Assumptions G.2, 5.1, and G.3 hold, by using the exponentially decreasing (then constant) step size $h_k = c \min\{\max\{t_k, \frac{1}{L}\}, 1\}, c = \frac{T + \log L}{N} \leq \frac{1}{Kd}$, the sampling dynamic (8) results in a distribution $\widehat{q}_T$ such that $\mathrm{KL}(p_0 \| \widehat{q}_T) \lesssim (M_2 + d)\exp(-T) + T\epsilon_0^2 + \frac{d^2(T + \log L)^2}{N}$. Choosing $T = \Theta(\log(dM_2/\epsilon_0^2))$ and $N = \Theta(d^2(T + \log L)^2/\epsilon_0^2)$ makes this $\widetilde{O}(\epsilon_0^2)$. In addition, for the Euler-Maruyama scheme (7), the same bounds hold with an additional $M_2 \sum_{k=1}^N h_k^3$ term.*

Finally, we state a tool from previous work [14].

Theorem G.8 (DPOM, Theorem 2 in [14], page 6)**.** *Suppose that Assumptions G.1, G.2 and 5.1 hold. If $\widehat{q}_T$ denotes the output of DPOM (see Algorithm 1 in [14]) with early stopping. Then, it holds that*

$$\begin{aligned} \mathrm{TV}(\widehat{q}_T, p_0) \lesssim & (\sqrt{d} + m_2)\exp(-T) + L^2 T d^{1/2} h_{\mathrm{pred}} \\ & + L^{3/2} T d^{1/2} h_{\mathrm{corr}}^{1/2} + L^{1/2} T \epsilon_0 + \epsilon. \end{aligned}$$

In particular, if we set $T = \Theta(\log(dm_2^2/\epsilon^2))$, $h_{\mathrm{pred}} = \widetilde{\Theta}(\frac{\epsilon}{L^2 d^{1/2}})$, $h_{\mathrm{corr}} = \widetilde{\Theta}(\frac{\epsilon}{L^3 d})$, and if the score estimation error satisfies $\epsilon_0 \leq \widetilde{O}(\frac{\epsilon}{\sqrt{L}})$, then we can obtain TV error ϵ with a total iteration complexity of $\widetilde{\Theta}(L^3 d/\epsilon^2)$ steps.

Theorem G.9 (DPUM, Theorem 3 in [14], page 7)**.** *Suppose that Assumptions G.1, G.2 and 5.1 hold. If $\widehat{q}_T$ denotes the output of DPUM (see Algorithm 2 in [14]) with early stopping. Then, it holds that*

$$\begin{aligned} \mathrm{TV}(\widehat{q}_T, p_0) \lesssim & (\sqrt{d} + m_2)\exp(-T) + L^2 T d^{1/2} h_{\mathrm{pred}} \\ & + L^{3/2} T d^{1/2} h_{\mathrm{corr}}^{1/2} + L^{1/2} T \epsilon_0 + \epsilon. \end{aligned}$$

In particular, if we set $T = \Theta(\log(dm_2^2/\epsilon^2))$, $h_{\mathrm{pred}} = \widetilde{\Theta}(\frac{\epsilon}{L^2 d^{1/2}})$, $h_{\mathrm{corr}} = \widetilde{\Theta}(\frac{\epsilon}{L^{3/2} d^{1/2}})$, and if the score estimation error satisfies $\epsilon_0 \leq \widetilde{O}(\frac{\epsilon}{\sqrt{L}})$, then we can obtain TV error ϵ with a total iteration complexity of $\widetilde{\Theta}(L^2 d^{1/2}/\epsilon)$ steps.

H. Putting It All Together

Our overall goal is that we want to provide a more concrete calculation for theorems in Section G by assuming the data distribution is a k mixture of Gaussian. Now we provide lemmas that are used in further calculation.

Now we provide the lemma for k-mixtue of Gaussians which states that if p_0 is mixture Gaussians, then all the pdfs in the diffusion process are also mixtures of Gaussians.

Lemma H.1 (Formal version of Lemma 3.2)**.** *Let $a, b \in \mathbb{R}$. Let $\mathcal{D}$ be a k-mixture of Gaussian distribution, and let p be its* pdf*, i.e.,*

$$p(x) := \sum_{i=1}^{k} \frac{\alpha_i}{(2\pi)^{d/2}\det(\Sigma_i)^{1/2}} \exp(-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i))$$

Let $x \in \mathbb{R}^d$ sample from $\mathcal{D}$. Let $z \in \mathbb{R}^d$ and $z \sim \mathcal{N}(0, I)$, which is independent from x. Then we have a new random variable $y = ax + bz$ which is also a k-mixture of Gaussian distribution $\widetilde{\mathcal{D}}$, whose pdf *is*

$$\widetilde{p}(x) := \sum_{i=1}^{k} \frac{\alpha_i}{(2\pi)^{d/2}\det(\widetilde{\Sigma}_i)^{1/2}} \exp(-\frac{1}{2}(x-\widetilde{\mu}_i)^\top \widetilde{\Sigma}_i^{-1}(x-\widetilde{\mu}_i)),$$

where $\widetilde{\mu}_i = a\mu_i, \widetilde{\Sigma}_i = a^2\Sigma_i + b^2 I$.

Proof. First, we know that the pdf of the sum of two independent random variables is the convolution of their pdf.

From [67] we know that the convolution of 2 Gaussians is another Gaussian, i.e. $\mathcal{N}(\mu_1, \Sigma_1) * \mathcal{N}(\mu_2, \Sigma_2) = \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$, where $*$ is the convolution operator.

And we know the pdf of a linear transformation of a random variable $x \in \mathbb{R}^d$, let's say $Ax + b$ where $A \in \mathbb{R}^{d\times d}, b \in \mathbb{R}^d$, is $\frac{1}{|\det(A)|}p(A^{-1}(x-b))$.

If we consider the transformation ax where $a \in \mathbb{R}, x \in \mathbb{R}^d$, this transformation can be written as aIx. Therefore the pdf of ax is $\frac{1}{|\det(aI)|}p((aI)^{-1}x) = \frac{1}{|a^d|}p(x/a)$.

Now we prove the lemma. To find the pdf of ax, where $x \sim p(x)$, we can show

$$\begin{aligned}
|\frac{1}{|a^d|}p(x/a)| &= \frac{1}{|a^d|}\sum_{i=1}^{k} \frac{\alpha_i}{(2\pi)^{d/2}\det(\Sigma_i)^{1/2}} \exp(-\frac{1}{2}(\frac{x}{a}-\mu_i)^\top \Sigma_i^{-1}(\frac{x}{a}-\mu_i)) \\
&= \sum_{i=1}^{k} \frac{\alpha_i}{(2\pi)^{d/2}\det(a^2\Sigma_i)^{1/2}} \exp(-\frac{1}{2}(x-a\mu_i)^\top a^{-2}\Sigma_i^{-1}(a-a\mu_i)) \\
&= \sum_{i=1}^{k} \alpha_i \mathcal{N}(a\mu_i, a^2\Sigma_i)
\end{aligned}$$

where the first step follows from the definition of $p(x)$, the second step follows from $a^{2d}\det(\Sigma_i) = \det(a^2\Sigma_i)$, and the last step follows from definition of Gaussian distribution.

For a single standard Gaussian random variable z, the pdf of bz will simply be $\mathcal{N}(0, b^2 I)$.

To find the pdf of $y = ax + bz$, we can show

$$\begin{aligned}
\widetilde{p}(x) &= \frac{1}{|a^d|}p(x/a) * \mathcal{N}(0, b^2 I) \\
&= (\sum_{i=1}^{k} \alpha_i \mathcal{N}(a\mu_i, a^2\Sigma_i)) * \mathcal{N}(0, b^2 I) \\
&= \sum_{i=1}^{k} (\alpha_i \mathcal{N}(a\mu_i, a^2\Sigma_i) * \mathcal{N}(0, b^2 I)) \\
&= \sum_{i=1}^{k} \alpha_i \mathcal{N}(a\mu_i, a^2\Sigma_i + b^2 I)
\end{aligned}$$

where the first step follows from the pdf of the sum of 2 independent random variables is the convolution of their pdf, the second step follows from the pdf of $\frac{1}{|a^d|}p(x/a)$, the third step follows from the distributive property of convolution, and the last step follows from $\mathcal{N}(a\mu_i, a^2\Sigma_i) * \mathcal{N}(0, b^2 I) = \mathcal{N}(a\mu_i, a^2\Sigma_i + b^2 I)$.

Thus, the pdf of y can be written as a mixture of k Gaussians:

$$\widetilde{p}(x) := \sum_{i=1}^{k} \frac{\alpha_i}{(2\pi)^{d/2}\det(\widetilde{\Sigma}_i)^{1/2}} \exp(-\frac{1}{2}(x-\widetilde{\mu}_i)^\top \widetilde{\Sigma}_i^{-1}(x-\widetilde{\mu}_i)),$$

where $\widetilde{\mu}_i = a\mu_i, \widetilde{\Sigma}_i = a^2\Sigma_i + b^2 I$. □

Now we provide the lemma for the second momentum of k-mixtue of Gaussians.

Lemma H.2 (Formal version of Lemma 3.6)**.** *If the following conditions hold:*

- *$x_0 \sim p_0$, where p_0 is defined by Eq.* (9).

Then, we have

$$m_2^2 := \mathop{\mathbb{E}}_{x_0 \sim p_0}[\|x_0\|_2^2] = \sum_{i=1}^{k} \alpha_i(\|\mu_i\|_2 + \mathrm{tr}[\Sigma_i])$$

Proof. From [46], we know the second momentum of data distribution $p_0(x)$ is given by:

$$\mathbb{E}[x_0 x_0^\top] = \sum_{i=1}^{k} \alpha_i \cdot (\|\mu_i\|_2^2 + \Sigma_i) \tag{23}$$

Then, we can show

$$\begin{aligned}\mathbb{E}[\|x_0\|_2^2] &= \mathbb{E}[x_0^\top x_0] \\ &= \mathbb{E}[\mathrm{tr}[x_0 x_0^\top]] \\ &= \mathrm{tr}[\mathbb{E}[x_0 x_0^\top]] \\ &= \mathrm{tr}[\sum_{i=1}^{k} \alpha_i(\mu_i\mu_i^\top + \Sigma_i)] \\ &= \sum_{i=1}^{k} \alpha_i(\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i])\end{aligned}$$

where the first step follows from definition of ℓ_2-norm, the second step follows from $\mathrm{tr}[aa^\top] = a^\top a$ where a is a vector, the third step follows from the linearity of the trace operator, the fourth step follows from Eq. (23), and the last step follows from $\mathrm{tr}[aa^\top] = \|a\|_2^2$. □

Now we give the Lipschitz constant explicitly.

Lemma H.3 (Formal version of Lemma 3.5)**.** *If the following conditions hold*

- *Let $\|x - a_t\mu_i\|_2 \le R$, where $R \ge 1$, for each $i \in [k]$.*
- *Let $\|x - a_t\mu_i\|_2 \ge \beta$, where $\beta \in (0, 0.1)$, for each $i \in [k]$.*
- *Let $p_t(x)$ be defined as Eq.* (10) *and $p_t(x) \ge \gamma$, where $\gamma \in (0, 0.1)$.*
- *Let $\sigma_{\min} := \min_{i\in[k]}\{\sigma_{\min}(a_t^2\Sigma_i + b_t^2 I)\}$.*
- *Let $\sigma_{\max} := \max_{i\in[k]}\{\sigma_{\max}(a_t^2\Sigma_i + b_t^2 I)\}$.*
- *Let $\det_{\min} := \min_{i\in[k]}\{\det(a_t^2\Sigma_i + b_t^2 I)\}$.*

The Lipschitz constant for the score function $\frac{\mathrm{d}\log(p_t(x))}{\mathrm{d}x}$ is given by:

$$L = \frac{1}{\sigma_{\min}} + \frac{2R^2}{\gamma^2\sigma_{\min}^2} \cdot (\frac{1}{(2\pi)^d \det_{\min}} + \frac{1}{(2\pi)^{d/2}\det_{\min}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max}})$$

Proof. Using Lemma F.11 and H.1, we can get the result. □

I. More Calculation for Application

In this section, we will provide a more concrete calculation for Theorem G.5, Theorem G.6, Theorem G.7, Theorem G.8 and Theorem G.9.

I.1. Concrete calculation of Theorem G.5

Theorem I.1 (DDPM, total variation, formal version of Theorem 5.4)**.** *If the following conditions hold:*

- *Condition 3.3 and 3.4.*
- *The step size $h_k := T/N$ satisfies $h_k = O(1/L)$ and $L \geq 1$ for $k \in [N]$.*
- *Let $\widehat{q}$ denote the density of the output of the* EulerMaruyama *defined by Definition 4.3.*

We have

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim \underbrace{\sqrt{\mathrm{KL}(p_0 \| \mathcal{N}(0,I))}\exp(-T)}_{\text{convergence of forward process}} + \underbrace{(L\sqrt{dh} + Lm_2 h)\sqrt{T}}_{\text{discretization error}} + \underbrace{\epsilon_0 \sqrt{T}}_{\text{score estimation error}}.$$

where

- $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma^2_{\min(p_t)}} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det^{1/2}_{\min(p_t)}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}})$,
- $m_2 = (\sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]))^{1/2}$,
- $\mathrm{KL}(p_0(x) \| \mathcal{N}(0,I)) \leq \frac{1}{2}(-\log(\det_{\min(p_0)}) + d\sigma_{\max(p_0)} + \mu_{\max(p_0)} - d)$.

Proof. Now we want to find a more concrete L in Assumption G.1. Notice that from Lemma H.1, we know that at any time between $0 \leq t \leq T$, p_t is also a k-mixture of gaussian, except that the mean and covariance change with time.

Using Lemma H.3, we can get L.

Now we want to find the second momentum in Assumption G.2. Using Lemma H.2, we know that $m_2 = (\sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]))^{1/2}$.

In Assumption 5.1, we also assume the same thing.

Now we want to have a more concrete setting for Theorem G.5 by calculating each term directly. Notice that now we have all the quantities except for the KL divergence term. Thus, we calculate $\mathrm{KL}(p_0 \| \mathcal{N}(0,I))$, which means the KL divergence of data distribution and standard Gaussian.

In our notation, we have

$$\begin{aligned}\mathrm{KL}(p_0(x) \| \mathcal{N}(0,I)) &= \mathrm{KL}(\sum_{i=1}^k \alpha_i \mathcal{N}(\mu_i, \Sigma_i) \| \mathcal{N}(0,I)) \\ &= \int \sum_{i=1}^k \alpha_i \mathcal{N}(\mu_i, \Sigma_i) \log(\frac{\sum_{i=1}^k \alpha_i \mathcal{N}(\mu_i, \Sigma_i)}{\mathcal{N}(0,I)}) \mathrm{d}x.\end{aligned}$$

However, this integral has no close form, but we can find an upper bound for this KL divergence instead.

We know the KL divergence of 2 normal distribution is given by:

$$\begin{aligned}&\mathrm{KL}(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2)) \\ = & -\frac{1}{2}\log(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}) + \frac{1}{2}\mathrm{tr}[(\Sigma_2)^{-1}\Sigma_1] + \frac{1}{2}(\mu_1 - \mu_2)^\top (\Sigma_2)^{-1}(\mu_1 - \mu_2) - \frac{d}{2}\end{aligned}$$

We define $\sigma_{\max(p_0)} = \max_{i \in [k]}\{\sigma_{\max}(\Sigma_i)\}$, $\det_{\min(p_0)} = \min_{i \in [k]}\{\det(\Sigma_i)\}$, $\mu_{\max(p_0)} = \max_{i \in [k]}\{\|\mu_i\|_2^2\}$. From [29], we can show

$$\begin{aligned}\mathrm{KL}(\sum_{i=1}^k \alpha_i \mathcal{N}(\mu_i, \Sigma_i) \| \mathcal{N}(0,I)) &\leq \sum_{i=1}^k \alpha_i \mathrm{KL}(\mathcal{N}(\mu_i, \Sigma_i) \| \mathcal{N}(0,I)) \\ &= \sum_{i=1}^k \frac{\alpha_i}{2}(-\log(\det(\Sigma_i)) + \mathrm{tr}[\Sigma_i] + \|\mu_i\|_2^2 - d) \\ &\leq \max_{i \in [k]} \frac{1}{2}(-\log(\det(\Sigma_i)) + \mathrm{tr}[\Sigma_i] + \|\mu_i\|_2^2 - d)\end{aligned}$$

$$\leq \frac{1}{2}(-\log(\det_{\min(p_0)}) + d\sigma_{\max(p_0)} + \mu_{\max(p_0)} - d)$$

where the first step follows from the convexity of KL divergence, the second step follows from KL divergence of 2 normal distribution, the third step follows from $\sum_{i=1}^k \alpha_i = 1$ and $0 \leq \alpha_i \leq 1$, and the last step follows from the definition of $\det_{\min(p_0)}$, $\sigma_{\max(p_0)}$, $\mu_{\max(p_0)}$.

Then we have all the quantities in Theorem G.5. After directly applying the theorem, we finish the proof. □

I.2. Concrete calculation for Theorem G.6

Theorem I.2 (DDPM, KL divergence, formal version of Theorem 5.5)**.** *If the following conditions hold:*

- *Condition 3.4.*
- *We use uniform discretization points.*

We have

- *Let $\widehat{q}$ denote the density of the output of the* ExponentialIntegrator *defined by Definition 4.4, we have*

$$\mathrm{KL}(p_0 \| \widehat{q}) \lesssim (M_2 + d)e^{-T} + T\epsilon_0^2 + \frac{dT^2L^2}{N}.$$

In particular, choosing $T = \Theta(\log(M_2 d/\epsilon_0))$ and $N = \Theta(dT^2L^2/\epsilon_0^2)$, then we can show that

$$\mathrm{KL}(p_0 \| \widehat{q}) = \widetilde{O}(\epsilon_0^2)$$

- *Let $\widehat{q}$ denote the density of the output of the* EulerMaruyama *defined by Definition 4.3, we have*

$$\mathrm{KL}(p_0 \| \widehat{q}) \lesssim (M_2 + d)e^{-T} + T\epsilon_0^2 + \frac{dT^2L^2}{N} + \frac{T^3M_2}{N^2}.$$

where

- $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2\sigma^2_{\min(p_t)}} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det^{1/2}_{\min(p_t)}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}})$,
- $M_2 = \sum_{i=1}^k \alpha_i(\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i])$.

Proof. Using Lemma H.3, we can get L. Using Lemma H.2, we can get m_2. Then we directly apply Theorem G.6. □

I.3. Concrete calculation for Theorem G.7

Theorem I.3 (DDPM, KL divergence for smooth data distribution, formal version of Theorem 5.6)**.** *If the following conditions hold:*

- *Condition 3.3 and 3.4.*
- *We use the exponentially decreasing (then constant) step size $h_k = c\min\{\max\{t_k, \frac{1}{L}\}, 1\}, c = \frac{T+\log L}{N} \leq \frac{1}{Kd}$.*
- *Let $\widehat{q}$ denote the density of the output of the* ExponentialIntegrator *defined by Definition 4.4.*

We have

$$\mathrm{KL}(p_0 \| \widehat{q}) \lesssim (M_2 + d)\exp(-T) + T\epsilon_0^2 + \frac{d^2(T + \log L)^2}{N},$$

where

- $L = \frac{1}{\sigma_{\min(p_0)}} + \frac{2R^2}{\gamma^2(\sigma_{\min(p_0)})^2} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_0)}} + \frac{1}{(2\pi)^{d/2}(\det_{\min(p_0)})^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_0)}})$,
- $M_2 = \sum_{i=1}^k \alpha_i(\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i])$.

Furthermore, if we choosing $T = \Theta(\log(M_2 d/\epsilon_0))$ and $N = \Theta(d^2(T + \log L)^2/\epsilon_0^2)$, then we can show

$$\mathrm{KL}(p_0 \| \widehat{q}) \leq \widetilde{O}(\epsilon_0^2)$$

In addition, for Euler-Maruyama scheme defined in Definition 4.3, the same bounds hold with an additional $M_2 \sum_{k=1}^N h_k^3$ term.

Proof. Clearly, p_0 is second-order differentiable. Using Lemma H.3, we can get L. Using Lemma H.2, we can get m_2. Then we directly apply Theorem G.7. □

I.4. Concrete calculation for Theorem G.8

Theorem I.4 (DPOM, formal version of Theorem 5.7)**.** *If the following conditions hold:*

- *Condition 3.4.*
- *We use the* DPOM *algorithm defined in Definition 4.5, and let $\widehat{q}$ be the output density of it.*

We have

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim (\sqrt{d} + m_2)\exp(-T) + L^2 T d^{1/2} h_{\mathrm{pred}} + L^{3/2} T d^{1/2} h_{\mathrm{corr}}^{1/2} + L^{1/2} T \epsilon_0 + \epsilon.$$

where

- $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma_{\min(p_t)}^2} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det_{\min(p_t)}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}}),$
- $m_2 = (\sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]))^{1/2}.$

In particular, if we set $T = \Theta(\log(dm_2/\epsilon))$, $h_{\mathrm{pred}} = \widetilde{\Theta}(\frac{\epsilon}{L^2 d^{1/2}})$, $h_{\mathrm{corr}} = \widetilde{\Theta}(\frac{\epsilon}{L^3 d})$, and if the score estimation error satisfies $\epsilon_0 \le \widetilde{O}(\frac{\epsilon}{\sqrt{L}})$, then we can obtain TV error ϵ with a total iteration complexity of $\widetilde{\Theta}(L^3 d/\epsilon^2)$ steps.

Proof. Using Lemma H.3, we can get L. Using Lemma H.2, we can get m_2. Then we directly apply Theorem G.8. □

I.5. Concrete calculation for Theorem G.9

Theorem I.5 (DPUM, formal version of Theorem 5.8)**.** *If the following conditions hold:*

- *Condition 3.4.*
- *We use the* DPUM *algorithm defined in Definition 4.6, and let $\widehat{q}$ be the output density of it.*

We have

$$\mathrm{TV}(\widehat{q}, p_0) \lesssim (\sqrt{d} + m_2)\exp(-T) + L^2 T d^{1/2} h_{\mathrm{pred}} + L^{3/2} T d^{1/2} h_{\mathrm{corr}}^{1/2} + L^{1/2} T \epsilon_0 + \epsilon.$$

where

- $L = \frac{1}{\sigma_{\min(p_t)}} + \frac{2R^2}{\gamma^2 \sigma_{\min(p_t)}^2} \cdot (\frac{1}{(2\pi)^d \det_{\min(p_t)}} + \frac{1}{(2\pi)^{d/2} \det_{\min(p_t)}^{1/2}}) \cdot \exp(-\frac{\beta^2}{2\sigma_{\max(p_t)}}),$
- $m_2 = (\sum_{i=1}^k \alpha_i (\|\mu_i\|_2^2 + \mathrm{tr}[\Sigma_i]))^{1/2}.$

In particular, if we set $T = \Theta(\log(dm_2/\epsilon))$, $h_{\mathrm{pred}} = \widetilde{\Theta}(\frac{\epsilon}{L^2 d^{1/2}})$, $h_{\mathrm{corr}} = \widetilde{\Theta}(\frac{\epsilon}{L^{3/2} d^{1/2}})$, and if the score estimation error satisfies $\epsilon_0 \le \widetilde{O}(\frac{\epsilon}{\sqrt{L}})$, then we can obtain TV error ϵ with a total iteration complexity of $\widetilde{\Theta}(L^2 d^{1/2}/\epsilon)$ steps.

Proof. Using Lemma H.3, we can get L. Using Lemma H.2, we can get m_2. Then we directly apply Theorem G.9. □