

A Good Score Does not Lead to A Good Generative Model

Sixu Li¹ Shi Chen² Qin Li²

Abstract

Score-based Generative Models (SGMs) is one leading method in generative modeling, renowned for their ability to generate high-quality samples from complex, high-dimensional data distributions. The method enjoys empirical success and is supported by rigorous theoretical convergence properties. In particular, it has been shown that SGMs can generate samples from a distribution that is close to the ground-truth if the underlying score function is learned well, suggesting the success of SGM as a generative model. We provide a counter-example in this paper. Through the sample complexity argument, we provide one specific setting where the score function is learned well. Yet, SGMs in this setting can only output samples that are Gaussian blurring of training data points, mimicking the effects of kernel density estimation. The finding resonates a series of recent finding that reveal that SGMs can demonstrate strong memorization effect and fail to generate.

1. Introduction

Generative modeling aims to understand the dataset structure so to generate similar examples. It has been widely used in image and text generation (Wang et al., 2018; Huang et al., 2018; Rombach et al., 2022; Li et al., 2022; Gong et al., 2022), speech and audio synthesis (Donahue et al., 2018; Kong et al., 2020a;b; Huang et al., 2022), and even the discovery of protein structures (Watson et al., 2023; Ni et al., 2023).

Among the various types of generative models, Score-based Generative Models (SGMs) (Song et al., 2020; Ho et al., 2020; Karras et al., 2022) have recently emerged as a forefront method, and achieved state-of-the-art empirical results across diverse domains. It views the data structure of existing examples coded in a probability distribution, that we

call the target distribution. Once SGM learns the target distribution from the dataset, it generates a new sample from it.

Despite their empirical successes, a thorough theoretical understanding of why SGMs perform well remains elusive. A more fundamental question is:

What are the criteria to evaluate the performance of a generative model?

Heuristically, two key components of generative models are “imitating” and “generating”. The “imitating” is about learning from the existences, while generating calls for creativity to produce new. A successful generative model should exhibit both *imitation ability*, so to produce samples that resemble the training data, and at the same time, manifest *creativity*, and generate samples that are not mere replicas of existing ones.

In the past few years, significance theoretical progresses have been made on assessing the *imitation ability* of SGMs. In particular, recently made available theory provides a very nice collection of error bounds to evaluate the difference between the learned distribution and the ground-truth distribution. Such discussion has been made available in various statistical distances, including total variation, KL divergence, Wasserstein distance and others. These discoveries suggest that SGMs have strong imitation ability, i.e. can approximate the ground-truth distribution well if the score function (gradient of log-density) of the target distribution along the diffusion process can be effectively learned.

We would like to discuss the other side of the story: Relying solely on these upper error bounds might be misleading in assessing the overall performance of SGMs. In particular, this criterion does not adequately address the issue of memorization – the possibility that the produced samples are simply replicas of the training data. In other words, SGMs with strong imitation ability can be lack of creativity.

1.1. A toy model argument

At the heart of our argument is that a simple Kernel Density Estimation (KDE) of the target ground-truth distribution can be arbitrarily close. Yet, drawing a sample from the ground-truth and drawing one from a KDE presents very different features. The latter fails on the task of “generation.”

¹Department of Statistics, University of Wisconsin-Madison, Madison WI, USA ²Department of Mathematics, University of Wisconsin-Madison, Madison WI, USA. Correspondence to: Qin Li <qinli@math.wisc.edu>.

To be mathematically more precise, let $p_*(x)$ be the ground-truth distribution, and $\{y_i\}_{i=1}^N$ be a set of i.i.d samples drawn from it. The empirical distribution is $p_* := \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$. We denote $p_*^\gamma := p_* * \mathcal{N}_\gamma$ the distribution obtained by smoothing p_* with a Gaussian kernel $\mathcal{N}_\gamma := \mathcal{N}(0, \gamma^2 I_{d \times d})$. Such definition naturally puts p_*^γ as one kind of Kernel Density Estimation (KDE) of p_* with the bandwidth γ .

It is intuitive that when the sample size N is large, and when the bandwidth γ is properly chosen, the KDE p_*^γ approximates the true distribution q . In the most extreme case, when the bandwidth $\gamma \rightarrow 0$, the kernel density estimate p_*^γ degenerates to the empirical distribution p_* . Throughout the paper we view the empirical distribution as a special case of KDE.

Though p_* and p_*^γ are close, generating samples from p_* and from p_*^γ are drastically different stories. Drawing from p_* amounts to generate a completely new sample, independent of the dataset, while generating from p_*^γ essentially means selecting a sample uniformly from the set $\{y_i\}_{i=1}^N$ and then applying a Gaussian blurring. Regardless of how close p_*^γ approximates the ground-truth p_* , sampling from KDE ultimately gives replicas of the existing samples.

Would SGM be different from KDE? SGM is built on a complicated procedure, incorporating forward noise injection, score matching, and backward sampling processes. The machinery is significantly more convoluted than the straightforward KDE approach. Would it be able to generate new samples?

We are to show in this paper that the perfect SGM is actually a KDE itself. The mathematical statement is presented in Theorem 4.3. The “perfect” means the minimizer of the empirical score matching objective is achieved during the score-matching procedure. We term the learned score function the *empirical optimal score function*. Since SGM equipped with the empirical optimal score function is effectively a KDE, it sees the limitation of KDE and fails to “generate.” This phenomenon is clearly demonstrated in Figure 1 with the test conducted over the CIFAR10 dataset.

It is important to note that this observation does not contradict existing theories that suggest SGMs can approximate the target distribution q when the score function is accurately learned. Indeed, in Theorem 3.1 we provide a sample complexity estimate and derive a lower bound of the sample size N . When the sample size is sufficiently large, the empirical optimal score function approximates the ground-truth score function. Consequently, according to the existing theories, the output of SGM is a distribution close to the ground-truth target distribution. Yet, two distribution being close is not sufficient for the task of generation.



Figure 1. Images generated based on CIFAR10 dataset. The first row shows the original images, the second row presents the images blurred according to the Gaussian KDE, and the third row shows images generated by SGM equipped with the perfect score function learned from samples. Both KDE and SGM present simple replica (with Gaussian blurring) of the original images.

1.2. Contributions

The primary contribution of this paper is presenting a counter-example of score-based generative models (SGMs) with accurate approximated score function, yet producing unoriginal, replicated samples. Our findings are substantiated through the following two steps:

- We establish in Theorem 3.1 the score-matching error of the empirical optimal score function, and present an explicit non-asymptotic error bound with the sample complexity. This result illustrates that the empirical optimal score function satisfies the standard L^2 bound on the score estimation error used in the convergence analysis in the existing literature (Chen et al., 2022; 2023c;d; Benton et al., 2023a), which presumably should lead to the conclusion that SGMs equipped with the empirical optimal score function produces a distribution close to the target distribution.
- We show in Theorem 4.3 that SGMs equipped with empirical optimal score function resembles a Gaussian KDE, and thus presents strong memorization effects and fails to produce novel samples.

These results combined rigorously demonstrates that the SGM with precise empirical score matching is capable to produce a distribution close to the target, but the procedure does not ensure the efficacy of an SGM in its ability to generate innovative and diverse samples. This observation underscores the limitation of current upper bound type guarantees and highlights the need for new theoretical criteria to assess the performance of generative models.

Notations: Let \mathbb{R}^d to be the d -dimensional Euclidean space and $T > 0$ is the time horizon. Denote $x = (x_1, x_2, \dots, x_d)^\top \in \mathbb{R}^d$ and $t \in [0, T]$ to be the spatial variable and time variable respectively. We denote p_* as the target data distribution supported on a subset of \mathbb{R}^d , and indicate the empirical distribution by p_* . The Gaussian ker-

nel with bandwidth γ is denoted by $\mathcal{N}_\gamma := \mathcal{N}(0, \gamma^2 I_{d \times d})$. For the special case $\gamma = 1$, i.e. standard Gaussian, we use notation $\pi^d := \mathcal{N}(0, I_{d \times d})$. We denote the Gaussian KDE with bandwidth γ as $p_\star^\gamma := p_\star * \mathcal{N}_\gamma$. In general, we use p_t and q_t (or p_t and q_t) to represent the laws of forward and backward SDEs at time t respectively (a thorough summary of PDEs and SDEs' notations used in this paper is provided in Appendix A). We denote $\delta \in [0, T)$ to be the early stopping time for running SDEs.

1.3. Literature review

We are mainly concerned of three distinct lines of research related to SGM performance, as summarized below.

Convergence of SGMs. The first line of research concerns theoretical convergence properties of SGMs. This addresses the most fundamental performance of the algorithm: What elements are needed for SGM to perform well? In this context, a good performance amounts to generating a new sample from the learned distribution that is close to the ground-truth. This line of research has garnered a large amount of interests, drawing its relation to sampling. For most studies, the analysis becomes quantifying the deviation between distributions generated by SGMs and the ground-truth distributions. This includes the earlier studies such as (Lee et al., 2022; Wibisono & Yingxi Yang, 2022; De Bortoli et al., 2021; De Bortoli, 2022; Kwon et al., 2022; Block et al., 2022), and later (Chen et al., 2022; 2023a;b; Benton et al., 2023a; Li et al., 2023) that significantly relaxed the Lipschitz condition of the score function and achieved polynomial convergence rate. In these discoveries, Girsanov's theorem turns out to be a crucial proof strategy. Parallel to these findings, convergence properties of ODE-based SGMs have also been explored (Chen et al., 2023c;d; Benton et al., 2023b; Albergo et al., 2023; Li et al., 2023), and comparison to SDE-based SGMs have been drawn.

Sample complexity studies of SGMs. Another line of research focuses on sample complexity. How many samples/training data points are needed to learn the score? In line with convergence rate analysis, the sample complexity study has been conducted with the criteria set to be L^2 -approximation of the score function (Block et al., 2022; Cui et al., 2023; Chen et al., 2023b; Oko et al., 2023). The involved techniques range from deploying Rademacher complexity for certain hypothesis classes, to utilizing specific neural network structures. Often in times, there are also assumptions made on the structure of data.

Memorization effect of SGMs. The third line of research on SGM concerns its memorizing effect. This line of research was triggered by some experimental discovery that SGMs, when trained well, tend to produce replicas of training samples (Somepalli et al., 2022; 2023; Carlini et al., 2023). This phenomenon draws serious privacy concerns,

and motivates studies on the fundamental nature of SGMs: Are SGMs memorizers or generalizers? In (Yoon et al., 2023), the authors presented a dichotomy, showing through numerical experiments that SGMs can generate novel samples when they fail to memorize training data. Furthermore, when confined to a basis of harmonic functions adapted to the geometry of image features, (Kadkhodaie et al., 2023) suggest that neural network denoisers in SGMs might have an inductive bias, aiming the generation. In (Gu et al., 2023; Yi et al., 2023), the authors derive the optimal solution to the empirical score-matching problem and show that the SGMs equipped with this score function exhibit a strong memorization effect. This suggests that with limited amount of training data and a large neural network capacity, SGMs tend to memorize rather than generalize.

To summarize: the convergence results of SGMs suggest a well-learned score function can be called to produce a sample drawn from a distribution close to the ground-truth, and the studies on the memorization effect of SGMs suggest the new drawings are simple replicas of the training dataset. It is worth noting that the two sets of results do not contradict. In particular, the convergence results do not rule out the explicit dependence of new generated samples on the training data. The connection between the two aspects of SGM performance is yet to be developed, and this is our main task of the current paper. We show that SGMs, despite having favorable convergence properties, can still resort to memorization, in the form of kernel density estimation. The finding underscores the need for a new theoretical framework to evaluate SGMs' performance, taking into account both imitation ability and creativity of SGMs.

2. Score-based Generative Models

We provide a brief expository to the Score-based Generative models (SGM) (Song et al., 2020) in this section. Mathematically, SGM is equivalent to denoising diffusion probabilistic modeling (DDPM) (Ho et al., 2020), so we use the two terms interchangeably.

2.1. Mathematical foundation for DDPM

The foundation for SGM stems from two mathematical observations. Firstly, a diffusion type partial differential equation (PDE) drives an arbitrary distribution to a Gaussian distribution, forming a bridge between the complex target distribution to the standard Gaussian, an easy-to-sample distribution. Secondly, such diffusion process can be simulated by its samples, translating the complicated PDE to a set of stochastic differential equations (SDEs) that are computationally easy to manipulate.

More precisely, denote $p_t(x)$ the solution to the PDE:

$$\partial_t p_t = \nabla \cdot (x p_t) + \Delta p_t. \quad (1)$$

It can be shown that, for *arbitrary* initial data p_0 , when T is big enough,

$$p_T \approx \lim_{t \rightarrow \infty} p_t = \pi^d,$$

and the convergence is exponentially fast (Bakry et al., 2014). In our context, we set the initial data $p_0 = p_*$, the to-be-learned target distribution.

This PDE can be run backward in time. Denote $q_t = p_{T-t}$, a quick calculation shows

$$\partial_t q_t = -\nabla \cdot ((x + 2\nabla \ln p_{T-t})q_t) + \Delta q_t. \quad (2)$$

This means with the full knowledge of $\nabla \ln p_{T-t}$, the flow field $x + 2\nabla \ln p_{T-t}(x) = x + 2u(T-t, x)$ drives the standard Gaussian ($q_0 = p_T \approx \pi^d$) back to its original distribution, the target $q_T = p_0 = p_*$. The term $u(t, x) = \nabla \ln p_t(x)$ is called the *score function*.

Simulating these two PDEs (1) and (2) directly is computationally infeasible, especially when dimension $d \gg 1$, but both equations can be represented by samples whose dynamics satisfy the corresponding SDEs. In particular, letting

$$dX_t^\rightarrow = -X_t^\rightarrow dt + \sqrt{2}dB_t, \quad (3)$$

the standard OU process, and

$$dX_t^\leftarrow = [X_t^\leftarrow + 2u(T-t, X_t^\leftarrow)]dt + \sqrt{2}dB_t', \quad (4)$$

where B_t and B_t' are two Brownian motions, then, with proper initial conditions:

$$\text{Law}(X_t^\leftarrow) = q_t = p_{T-t} = \text{Law}(X_{T-t}^\rightarrow).$$

This relation translates directly simulating two PDEs (1) and (2) to running its representative samples governed by SDEs (3)-(4), significantly reducing the computational complexity. It is worth noting that if one draws $X_{t=0}^\leftarrow \sim p_T$ and runs (4), then:

$$\text{Law}(X_T^\leftarrow) = p_*,$$

meaning the dynamics of (4) returns a sample from the target distribution p_* , achieving the task of sampling. Here the notation \sim stands for drawing an i.i.d. sample from.

2.2. Score-function, explicit solution and score matching

It is clear the success of SGM, being able to draw a sample from the target distribution p_* , lies in finding a good approximation of the score function $u(t, x)$. In the idealized setting, this score function can be explicitly expressed. In the practical computation, this function is learned from existing dataset through the score-matching procedure.

To explicitly express the score function amounts to solving (1), or equivalently (3). Taking the SDE perspective,

we analyze the OU process in (3) and obtain an explicit solution:

$$X_t^\rightarrow := \mu(t)y + \sigma(t)Z \quad \text{with} \quad \begin{cases} \mu(t) := e^{-t} \\ \sigma(t) := \sqrt{1 - e^{-2t}}, \end{cases} \quad (5)$$

where y is the initial data and $Z \sim \pi^d$. Equivalently, using the PDE perspective, one sets $p_0 = \delta_y$ as the initial condition to run (1) to form a set of Green's functions:

$$p_t(x|y) := \mathcal{N}(x; \mu(t)y, \sigma(t)^2 I_{d \times d}). \quad (6)$$

These functions are Gaussian functions of x centered at $\mu(t)y$ with isotropic variance $\sigma(t)^2$. This set of functions is also referred to as the transition kernel from time 0 conditioned on $X_0^\rightarrow = y$ to time t with $X_t^\rightarrow = x$.

In the idealized setting with the target distribution p_* fully known, then with $p_0 = p_*$, the solution of (1) becomes the superposition of Green's functions weighted by p_* , namely:

$$p_t(x) = \int p_t(x|y)p_*(y)dy, \quad (7)$$

thus by definition, the score function is explicit:

$$\begin{aligned} u(t, x) &= \nabla \ln p_t(x) = \frac{\nabla p_t(x)}{p_t(x)} \\ &= \frac{\int u(t, x|y)p_t(x|y)p_*(y)dy}{\int p_t(x|y)p_*(y)dy}, \end{aligned} \quad (8)$$

where we called (7) and used the notation $u(t, x|y) = \nabla \ln p_t(x|y)$ to denote the conditional flow field. This function maps $\mathbb{R}_+ \times \mathbb{R}^d$ to \mathbb{R}^d . Using the explicit formula (6), we have the explicit solution for the conditional flow field:

$$u(t, x|y) = -\frac{x - \mu(t)y}{\sigma(t)^2}. \quad (9)$$

It is a linear function on x with Lipschitz constant $\frac{1}{\sigma(t)^2}$ that blows up at $t = 0$.

Score matching. The practical setting is not idealized: The lack of explicit formulation p_* prevents direct computation of (8). Algorithmically, one needs to learn $u(t, x)$ from existing samples. A neural network (NN) is then deployed.

Intuitively, the NN should provide a function as close as possible to the true score function, meaning it solves:

$$\min_{s \in \mathcal{F}} \mathcal{L}_{\text{SM}}(s) := \mathbb{E}_{t,x} [\|s(t, x) - u(t, x)\|^2],$$

where $t \sim U[0, T]$, the uniform distribution over the time interval, and $x \sim p_t(x)$. \mathcal{F} is a hypothesis space, and in this context, the function space representable by a class of neural networks. However, neither p_t nor $u(t, x)$ is known in the formulation, so we turn to an equivalent problem:

$$\min_{s \in \mathcal{F}} \mathcal{L}_{\text{CSM}}(s) := \mathbb{E}_{t,y,x} [\|s(t, x) - u(t, x|y)\|^2],$$

where $t \sim U[0, T]$, $y \sim p_*$ and $x \sim p_t(x|y)$. The subindex CSM stands for conditional-score-matching. The two problems can be shown to be mathematically equivalent, see Lemma B.1. Practically, however, this new problem is much more tractable, now with both $p_t(x|y)$ and $u(t, x|y)$ explicit, see (6) and (8).

The target distribution p_* is still unknown. At hands, we have many samples drawn from it: $\{y_i\}_{i=1}^N$. This allows us to reformulate the problem into an empirical risk minimization (ERM) problem:

$$\min_{s \in \mathcal{F}} \mathcal{L}_{\text{CSM}}^N(s) := \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{t,x} \left[\|s(t, x) - u(t, x|y_i)\|^2 \right] \quad (10)$$

with $t \sim U[0, T]$ and $x \sim p_t(x|y_i)$.

In the execution of a practical DDPM algorithm, (10) is first run to find an NN serving as a good approximation to the score function, termed $s(t, x)$, and the user end then deploys this $s(t, x)$ in (4) in place of $u(t, x)$ for generating a new sample from p_* . Sample $\bar{X}_0^{\leftarrow} \sim \pi^d$ and run:

$$d\bar{X}_t^{\leftarrow} = (\bar{X}_t^{\leftarrow} + 2s(T - t, \bar{X}_t^{\leftarrow})) dt + \sqrt{2d}B_t. \quad (11)$$

The law is denoted to be $\bar{q}_t := \text{Law}(\bar{X}_t^{\leftarrow})$. We note two differences comparing (4) and (11): the initial data p_T is replaced by π^d and the score function $u(t, x)$ is replaced by the empirically learned score function $s(t, x)$. If both approximations are accurate, we expect $\bar{q}_t \approx q_t$ for all t .

When minimizing the objective (10), noting the singularity at $t = 0$ of $u(t, x|y_i)$ as in (9), it is a standard practice to conduct “early stopping” (Song et al., 2020). This is to take out a small fraction around the origin of time in the training (10) and learn the score with $t \sim U[\delta, T]$. Consequently, the sampling is also only ran up to $T - \delta$. The algorithm returns samples $\bar{X}_{T-\delta}^{\leftarrow}$ drawn from $\bar{q}_{T-\delta}$. The hope is $\bar{q}_{T-\delta}$ approximates the target p_* using the following approximation chain:

$$\underbrace{\bar{q}_{T-\delta} \approx q_{T-\delta}}_{\text{if } s \approx u, \pi^d \approx p_T} = \underbrace{p_\delta \approx p_0}_{\text{if } \delta \rightarrow 0} = p_*.$$

2.3. Error analysis for DDPM

In the idealized setting, $T \rightarrow \infty$, $s(t, x) = u(t, x)$, $\delta \rightarrow 0$, and backward SDE (11) is run perfectly, then the sample initially drawn from Gaussian π^d will represents the target distribution p_* at T . Computationally, these assumptions all break: all four factors, finite T , nontrivial δ , imperfect $s(t, x)$ and discretization error of (11) induce error. These errors were beautifully analyzed in (Chen et al., 2022; Benton et al., 2023a). We summarize their results briefly.

All analysis require the target distribution to have bounded second moment.

Assumption 2.1 (bounded second moment). We assume that $m_2^2 := \mathbb{E}_{y \sim p_*} [\|y\|^2] < \infty$.

The learned score function is also assumed to be close to the ground-truth in $L_2(dt, p_t dx)$:

Assumption 2.2 (score estimation error). The score estimate $s(x, t)$ satisfies

$$\mathbb{E}_{t \sim U[\delta, T], x \sim p_t} [\|s(t, x) - u(t, x)\|^2] \leq \varepsilon_{\text{score}}^2.$$

Under these assumptions, it was concluded DDPM samples well:

Theorem 2.3 (Modified version of Theorem 1 in (Benton et al., 2023a)). Suppose the Assumptions 2.1 and 2.2 hold and $T \geq 1$, $\delta > 0$. Let $\bar{q}_{T-\delta}$ be the output of the DDPM algorithm (11) at time $T - \delta$. Then it holds that

$$\text{TV}(\bar{q}_{T-\delta}, p_\delta) \lesssim \varepsilon_{\text{score}} + \sqrt{d} \exp(-T)$$

The discretization error in the original result is irrelevant to the discussion here and is omitted. This upper error bound consists of two parts. The first term $\varepsilon_{\text{score}}$ comes from the score approximation error, while the second term $\sqrt{d} \exp(-T)$ comes from the finite truncation, where we forcefully replace p_T by π^d .

The theorem states that, when T is large enough and the score function is approximated well in $L_2(dt, p_t dx)$ sense, the TV distance between the law of generated samples $\bar{q}_{T-\delta}$ and $p_\delta \approx p_*$ is very small, concluding that DDPM is a good sampling strategy.

It is tempting to further this statement and claim that DDPM is also a good generative model. Indeed, on the surface, it is typically claimed that generative models are equivalent to drawing samples from a target distribution p_* . However, we should note a stark difference between sampling and generation: A meaningful generative model should be able to produce samples that are not mere replica of known ones. The error bound in Theorem 2.3 does not exclude this possibility. As will be shown in Section 3, it is possible to design a DDPM whose score function is learned well, so according to Theorem 2.3 produces a distribution close to the target. Yet in Section 4, we demonstrate that this model fails to be produce new samples. These two sections combined suggest DDPM with a well-learned score function does not necessarily produce a meaningful generative model.

3. A good score estimate: sample complexity analysis

Inspired by Theorem 2.3, we are to design a DDPM whose learned score function satisfies Assumption 2.2. Throughout the section, we assume the hypothesis space is large enough

($\mathcal{F} \supseteq L^2([0, T] \times \mathbb{R}^d)$, for example), and the learned score estimate achieves the global minimum of the ERM (10). In practical training, the error heavily depends on the specific NN structure utilized in the optimization. The approximation error of the NN training is beyond the discussion point of the current paper.

Noting the objective $\mathcal{L}_{\text{CSM}}^N(s)$ is a convex functional of s , the optimizer has a closed-form. As derived in Proposition B.2, for $(t, x) \in [0, T] \times \mathbb{R}^d$, the *empirical optimal score function* is:

$$s_{\{y_i\}}^N(t, x) := \frac{\sum_{i=1}^N u(t, x|y_i) p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)}, \quad (12)$$

where $u(t, x|y)$ is the conditional flow field, see (9).

Accordingly, the DDPM draws an initial data from $\hat{X}_0^\leftarrow \sim \pi^d$ and evolves the following SDE:

$$d\hat{X}_t^\leftarrow = (\hat{X}_t^\leftarrow + 2s_{\{y_i\}}^N(T - t, \hat{X}_t^\leftarrow))dt + \sqrt{2}dB_t. \quad (13)$$

We denote the law of samples $\hat{q}_t := \text{Law}(\hat{X}_t^\leftarrow)$. The choice of the font indicates the law is produced by a finite dimensional object $s_{\{y_i\}}^N$.

To understand the empirical optimal score function, we compare (12) with the ground-truth score function (8). It is clear $s_{\{y_i\}}^N$ can be interpreted as a Monte-Carlo (MC) sampling of $u(t, x)$, replacing both integrals in the numerator and the denominator in (8) by empirical means. The law of large number suggests the empirical mean should converge to the true mean when the number of samples is big. Therefore, it is expected $s_{\{y_i\}}^N$ approximates u well with a very high probability when $N \gg 1$. We formulate this result in the following theorem.

Theorem 3.1 (Approximation error of empirical optimal score function). *Let $\{y_i\}_{i=1}^N$ be N i.i.d samples drawn from the target data distribution p_* . Denote $u(t, x)$ and $s_{\{y_i\}}^N(t, x)$ the true and empirical optimal score function, respectively, as defined in (8) and (12). Then for any fixed $0 < \delta < T < \infty$, $\varepsilon_{\text{score}} > 0$ and $\tau > 0$, we have*

$$\mathbb{E}_{t \sim U[\delta, T], x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \leq \varepsilon_{\text{score}}^2,$$

with probability at least $1 - \tau$ provided that the number of training samples $N \geq N(\varepsilon_{\text{score}}, \delta, \tau)$, in particular

- *Case 1: If p_* is an isotropic Gaussian, i.e. $p_*(y) = \mathcal{N}(y; \mu_{p_*}, \sigma_{p_*}^2 I_{d \times d})$, with second moment $m_2^2 = O(d)$, then $N(\varepsilon_{\text{score}}, \delta, \tau) = \frac{1}{\tau \varepsilon_{\text{score}}^2} \frac{O(d)}{\delta^{5/2}}$;*
- *Case 2: If p_* is supported on the Euclidean ball of radius R such that $R^2 = O(d)$, then $N(\varepsilon_{\text{score}}, \delta, \tau) = \frac{1}{\tau \varepsilon_{\text{score}}^2} \exp\left(\frac{O(d)}{\delta}\right)$.*

The theorem implies that when the sample size is large with $N \geq N(\varepsilon_{\text{score}}, \delta, \tau)$, we have high confidence, $1 - \tau$, to state that the empirical optimal score function $s_{\{y_i\}}^N$, computed using the i.i.d. samples $\{y_i\}$, is within $\varepsilon_{\text{score}}$ distance from the true score function $u(t, x)$ in $L_2(dt, p_t dx)$.

Remark 3.2. A few comments are in line:

- (a) Second moment $m_2^2 = O(d)$ and support radius $R^2 = O(d)$: The second moment and support radius being the same order as d is only for notational convenience. In the proof, the assumption can be relaxed. When we do so, the success rate needs to be adjusted accordingly (see the discussions in Appendix C).
- (b) Implication on DDPM performance: Combining Theorem 3.1 with Theorem 2.3, it is straightforward to draw a conclusion on the performance of DDPM in terms of sample complexity. Under the same assumptions in Theorem 3.1, for any tolerance error $\varepsilon > 0$, by choosing $T = \log \frac{\sqrt{d}}{\varepsilon}$, $N \geq N(\varepsilon, \delta, \tau)$, then it holds that, the DDPM algorithm ran according to (13) with the empirical optimal score function s^N computed from (12) gives:

$$\text{TV}(\hat{q}_{T-\delta}, p_\delta) \lesssim \varepsilon$$

with probability at least $1 - \tau$.

- (c) Error dependence on parameters: Both the confidence level parameter τ and the accuracy parameter $\varepsilon_{\text{score}}$ appears algebraically in $N(\varepsilon_{\text{score}}, \delta, \tau)$. The rate of $\varepsilon_{\text{score}}^{-2}$ comes from MC sampling convergence of $\frac{1}{\sqrt{N}}$ and is expected to be the optimal one. The rate of τ^{-1} reflects the fact that the proof uses the simple Markov inequality.

We leave the main proof to Appendix C and only briefly discuss the proof strategy using Case 2 as an example.

Sketch of proof. Denote the error term

$$\left| E_{\{y_i\}}^t \right|^2 = \mathbb{E}_{x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \quad (14)$$

and

$$\left| E_{\{y_i\}} \right|^2 = \mathbb{E}_{t \sim U[\delta, T]} \left| E_{\{y_i\}}^t \right|^2 = \frac{1}{T - \delta} \int_\delta^T \left| E_{\{y_i\}}^t \right|^2 dt.$$

$E_{\{y_i\}}$ defines a function that maps $\{y_i\} \in \mathbb{R}^{Nd}$ to \mathbb{R}^+ , and is a random variable itself. According to the Markov's inequality:

$$\mathbb{P}(E_{\{y_i\}} > \varepsilon_{\text{score}}) \leq \frac{\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2}{\varepsilon_{\text{score}}^2}. \quad (15)$$

To compute the right hand side, we note

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2 = \mathbb{E}_{t, \{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}^t|^2, \quad (16)$$

and for fixed $t \in [\delta, T]$, according to the definition (14), one can show:

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}^t|^2 \lesssim \frac{1}{N} \frac{1}{t} \exp\left(\frac{O(d)}{t}\right). \quad (17)$$

Taking expectation with respect to t in $[\delta, T]$, we have

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2 \lesssim \frac{1}{N} \exp\left(\frac{O(d)}{\delta}\right),$$

finishing the proof when combined with (15). \square

It is clear the entire proof is built upon a direct use of the Markov inequality, and the most technical component of the proof is to give an estimate to the mean of the error term $|E_{\{y_i\}}^t|^2$ in (17). We provide this estimate in Lemma C.2.

4. A bad SGM: memorization Effects

Results in Theorem 2.3 and Theorem 3.1 combined implies that the DDPM (11) ran with the empirical optimal score function $s_{\{y_i\}}^N$ provides a good sampling method with a high probability. It is tempting to further this statement and call it a good generative model. We are to show in this section that this is not the case. In particular, we claim DDPM ran by $s_{\{y_i\}}^N$ will lead to a kernel density estimation (KDE).

To be more precise, with $\{y_i\}_{i=1}^N$ i.i.d drawn from the target distribution p_* , DDPM (11) ran with $s_{\{y_i\}}^N$ produces a distribution that is a convolution of a Gaussian with $p_* = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$, and hence becomes a KDE of p_* . Since the context is clear, throughout the section we drop the lower index $\{y_i\}$ in $s_{\{y_i\}}^N$.

The statement above stems from the following two simple observations. Firstly, the solution to the system (1) with initial distribution set to be p_* is a simple Gaussian convolution with p_* ; and secondly, the exact score function for this new system (initialized at p_*) happens to be the empirical optimal score function (12).

To expand on it, we first set the initial data for (1) as p_* , the empirical distribution. Theory in Section 2.2 still applies. In particular, the solution to (1), denoted by p_t , and the solution to (2), denoted by q_t , still have explicit forms using the Green's functions:

$$\begin{aligned} p_t(x) &= q_{T-t}(x) = \int p_t(x|y) p_*(y) dy = \frac{1}{N} \sum_{i=1}^N p_t(x|y_i) \\ &= \frac{1}{N} \sum_{i=1}^N \mathcal{N}(x; \mu(t)y_i, \sigma(t)^2 I_{d \times d}). \end{aligned} \quad (18)$$

For small t , $\mu(t) \approx 1$ and $\sigma(t) \approx 0$, the PDE solution (18) presents a strong similarity to a KDE of p_* with parameter $\gamma = \sigma(t)$:

$$p_*^\gamma(x) := p_* * \mathcal{N}(0, \gamma^2) = \frac{1}{N} \sum_{i=1}^N \mathcal{N}(x; y_i, \gamma^2 I_{d \times d}),$$

where $*$ is the convolution operator. The resemblance can be characterized mathematically precisely:

Proposition 4.1. *Suppose the training samples $\{y_i\}_{i=1}^N$ satisfy $\|y_i\|_2 \leq d$, for $\delta \geq 0$, $\text{TV}(q_{T-\delta}, p_*^\gamma) \leq \frac{d\sqrt{\delta}}{2}$ with $\gamma = \sigma(\delta)$, where $\sigma(\cdot)$ is defined in (5).*

This means the forward and backward procedure described in (1)-(2) approximately provides a simple KDE to the target distribution when initialized with the empirical distribution.

We now further claim this forward and backward procedure is realized by running SGM using the empirical optimal score s^N . To see this, we follow the computation in (8), and call (18) to obtain:

$$\nabla \ln p_t(x) = \frac{\sum_{i=1}^N \nabla p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)} = \frac{\sum_{i=1}^N u(t, x|y_i) p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)}.$$

This means the exact score function for the KDE approximation $p_t = q_{T-t}$ exactly recovers s^N , the empirical optimal score for p_t , and thus SGM with empirical optimal score realizes the KDE approximation, as seen in the following proposition.

Proposition 4.2. *Under the same assumptions as in Proposition 4.1, on the time interval $t \in [0, T]$, the total variation between the output distribution of SGM algorithm (13) with the empirical optimal score function \hat{q}_t and the KDE approximation q_t is bounded by $\text{TV}(\hat{q}_t, q_t) \leq \frac{d}{2} \exp(-T)$.*

Combine Propositions 4.1 and 4.2 using triangle inequality, we see \hat{q}_t is essentially a kernel density estimation when t approaches T . Furthermore, if one pushes $t = T \rightarrow +\infty$, we obtain the finite-support result:

Theorem 4.3 (SGM with empirical optimal score function resembles KDE). *Under the same assumptions as Proposition 4.2, SGM algorithm (13) with the empirical optimal score function s^N returns a simple Gaussian convolution with the empirical distribution in the form of (18), and it presents the following behavior:*

- (with early stopping) for any $\varepsilon > 0$, set $T = \log \frac{d}{\varepsilon}$ and $\delta = \frac{\varepsilon^2}{d}$, we have

$$\text{TV}(\hat{q}_{T-\delta}, p_*^\gamma) \leq \varepsilon, \quad \text{with } \gamma = \sigma(\delta),$$

- (without early stopping) by taking the limit $T \rightarrow +\infty$ and $\delta = 0$, we have $\hat{q}_\infty = p_* = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$.

The theorem suggests DDPM with empirical optimal score function s^N is, in the end, simply a KDE of the target p_* . However close KDE p_*^γ is to the target p_* , it is nevertheless only an object with finite amount of information.

Unlike drawing from p_* where one can generate a completely new sample independent of the training samples, drawing from p_*^γ can only provide replicas of y_i (with a slight shift and polluted with Gaussian noise). As a summary, SGM ran by the empirical optimal score function fails the task of generation.

Some mathematical comments are in line. We first note that (4.3) does not contradict (3.2). Indeed, with high probability, $\hat{q}_{T-\delta}$ approximates both p_δ and the KDE p_*^γ . The second bullet point (without early stopping) was also discussed in (Gu et al., 2023). Our result generalize theirs to any small time $T - \delta$.

5. Numerical Experiments

This section is dedicated to providing numerical evidence for Theorem 3.1 and Theorem 4.3. Throughout the experiment, we choose the target data distribution p_* to be a 2-dimensional isotropic Gaussian, denoted by $p_*(x) = \mathcal{N}(x; \mu_{p_*}, \sigma_{p_*}^2 I_{2 \times 2})$. The implementation details are provided in Appendix E.

We first estimate the score approximation error of the empirical optimal score function, as delineated in (3), for various size of training sample N . Figure 2 shows that the error has decreasing rate approximately $O(\frac{1}{N})$, confirming the theoretical finding in Theorem 3.1, see also Remark 3.2(c).

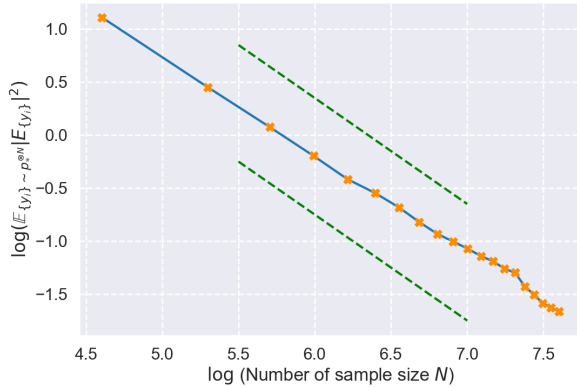


Figure 2. Score approximation error of the empirical optimal score function defined in (16) versus the number of training samples N . Both x -axis and y -axis are in the logarithmic scales. The orange crosses represent the score approximation error for varying values of N , with a fitted blue trend line. Reference lines with a slope of -1 are depicted by the green dashed lines, illustrating that the slope of the blue line is also approximately -1 . This observation corroborates the rate $O(\frac{1}{N})$ provided in Theorem 3.1.

Secondly, we showcase Theorem 4.3 and demonstrate that DDPM behaves as a KDE when equipped with empirical optimal score function. As seen in Figure 3, samples produced by DDPM ran with s^N exhibit a high concentration around the training samples. Conversely, while the samples generated by DDPM ran with the true score function $u(t, x)$ appear to be drawn from the same distribution as the training samples, they are not mere duplicates of the existing ones.

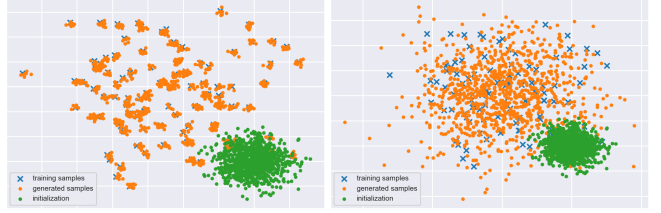


Figure 3. **Left:** Samples generated by DDPM with *empirical optimal score function* $s^N(t, x)$. **Right:** Samples generated by DDPM with *true score function* $u(t, x)$. In both plots, the blue crosses are the training samples, the green dots are the initialization positions and the orange dots are the outputs of DDPM with early stop of $\delta = 0.01$.

6. Discussion and Conclusion

The classical theory measures the success of score-based generative model based on the distance of the learned distribution and the ground-truth distribution. Under this criterion, SGM would be successful if the score function is learned well.

In this paper, we provide a counter-example of SGM that has a good score approximation while produces meaningless samples. On one hand, the application of Theorem 2.3 and Theorem 3.1 combined suggest SGM equipped with empirical optimal score function learns a distribution close to the ground-truth. On the other hand, Theorem 4.3 suggests this scenario resembles the Gaussian kernel density estimation and can only generate existing training samples with Gaussian blurring.

This apparent paradox between sound theoretical convergence and poor empirical new sample generations indicates that current theoretical criteria may not be sufficient to fully evaluate the performance of generative models. It strongly focuses on the “imitation” capability and losses out on quantifying “creativity”. Similar features were presented in other generative models like generative adversarial networks (Vardanyan et al., 2023), and different criteria have been proposed (Vardanyan et al., 2023; Yi et al., 2023), yet a comprehensive end-to-end convergence analysis for these criteria has not been done for SGMs. We leave this exploration to future research.

Acknowledgements

The three authors are supported in part by NSF-DMS 1750488, and NSF-DMS 2308440. S. Li is further supported by NSF-DMS 2236447.

References

- Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.
- Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.
- Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023a.
- Benton, J., Deligiannidis, G., and Doucet, A. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023b.
- Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling, 2022.
- Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models, 2023.
- Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023a.
- Chen, M., Huang, K., Zhao, T., and Wang, M. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*, 2023b.
- Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.
- Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023c.
- Chen, S., Daras, G., and Dimakis, A. G. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *arXiv preprint arXiv:2303.03384*, 2023d.
- Cui, H., Krzakala, F., Vanden-Eijnden, E., and Zdeborová, L. Analysis of learning a flow-based generative model from limited sample complexity, 2023.
- De Bortoli, V. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.
- De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.
- Donahue, C., McAuley, J., and Puckette, M. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.
- Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.
- Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020.
- Huang, H., He, R., Sun, Z., Tan, T., et al. Introvae: Introspective variational autoencoders for photographic image synthesis. *Advances in neural information processing systems*, 31, 2018.
- Huang, R., Lam, M. W., Wang, J., Su, D., Yu, D., Ren, Y., and Zhao, Z. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022.
- Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representation, 2023.
- Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.
- Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020a.
- Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020b.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

-
- Kwon, D., Fan, Y., and Lee, K. Score-based generative modeling secretly minimizes the wasserstein distance. *Advances in Neural Information Processing Systems*, 35: 20205–20217, 2022.
- Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35: 22870–22882, 2022.
- Li, G., Wei, Y., Chen, Y., and Chi, Y. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.
- Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.
- Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- Ni, B., Kaplan, D. L., and Buehler, M. J. Generative design of de novo proteins based on secondary-structure constraints using an attention-based diffusion model. *Chem*, 2023.
- Oko, K., Akiyama, S., and Suzuki, T. Diffusion models are minimax optimal distribution estimators, 2023.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models, 2022.
- Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Terrell, G. R. and Scott, D. W. Variable kernel density estimation. *The Annals of Statistics*, pp. 1236–1265, 1992.
- Vardanyan, E., Minasyan, A., Hunanyan, S., Galstyan, T., and Dalalyan, A. Guaranteed optimal generative modeling with maximum deviation from the empirical distribution. *arXiv preprint arXiv:2307.16422*, 2023.
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.
- Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.
- Wibisono, A. and Yingxi Yang, K. Convergence in kl divergence of the inexact langevin algorithm with application to score-based generative models. *arXiv e-prints*, pp. arXiv–2211, 2022.
- Yi, M., Sun, J., and Li, Z. On the generalization of diffusion model, 2023.
- Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference* & *Generative Modeling*, 2023.

A. Notations

Partial differential equations (PDEs). Let \mathbb{R}^d to be the d -dimensional Euclidean space and $T > 0$ is the time horizon. Denote $x = (x_1, x_2, \dots, x_d)^T \in \mathbb{R}^d$ and $t \in [0, T]$ to be the spatial variable and time variable respectively. The gradient of a real-valued function p with respect to the spatial variable and the time-derivative of p are denoted by $\nabla p = \left(\frac{\partial p}{\partial x_1}, \frac{\partial p}{\partial x_2}, \dots, \frac{\partial p}{\partial x_d} \right)$ and $\partial_t p$ respectively. The Laplacian of p is denoted by $\Delta p = \nabla \cdot (\nabla p)$. Here, $\nabla \cdot F = \sum_{i=1}^d \frac{\partial F_i}{\partial x_i}$ indicates the divergence of $F = (F_1, F_2, \dots, F_d)$ with respect to the spatial variable x .

Stochastic differential equations (SDEs) and their laws.

- The target data distribution is p_* .
- The forward process (3) initialized at the target distribution p_* is denoted $(X_t^{\rightarrow})_{t \in [0, T]}$, and $p_t := \text{Law}(X_t^{\rightarrow})$.
- The backward process (4) is denoted $(X_t^{\leftarrow})_{t \in [0, T]}$, where $\text{Law}(X_t^{\leftarrow}) := q_t = p_{T-t} = \text{Law}(X_{T-t}^{\rightarrow})$.
- The DDPM algorithm (11) with arbitrary learned score function is denoted $(\bar{X}_t^{\leftarrow})_{t \in [0, T]}$ and $\bar{q}_t := \text{Law}(\bar{X}_t^{\leftarrow})$. We initialize the process at $\bar{q}_0 = \pi^d$, the standard Gaussian distribution.
- The DDPM algorithm (13) with the empirical optimal score function s^N is denoted by $(\hat{X}_t^{\leftarrow})_{t \in [0, T]}$. We indicate the law at time t as $\hat{q}_t := \text{Law}(\hat{X}_t^{\leftarrow})$ and let $\hat{q}_0 = \pi^d$.
- The law of forward process (3) initialized at the empirical distribution p_* at time $t \in [0, T]$ is indicated by p_t . The law of corresponding backward process at time $t \in [0, T]$ is denoted by $q_t = p_{T-t}$.

Other notations. We denote p_* as the target data distribution supported on a subset of \mathbb{R}^d , and indicate the empirical distribution by p_* . The Gaussian kernel with bandwidth γ is denoted by $\mathcal{N}_\gamma := \mathcal{N}(0, \gamma^2 I_{d \times d})$. For the special case $\gamma = 1$, i.e. standard Gaussian, we use notation $\pi^d := \mathcal{N}(0, I_{d \times d})$. We denote the Gaussian KDE with bandwidth γ as $p_*^\gamma := p_* * \mathcal{N}_\gamma$. The early stopping time of running SDEs is indicated by $\delta \in [0, T]$. We use $i \in [N]$ to denote $i = 1, 2, \dots, N$.

B. Empirical optimal score function

Lemma B.1. *Assuming that $p_t(x) > 0$ for all $x \in \mathbb{R}^d$ and $t \in [0, T]$, then up to a constant independent of function $s \in L^2([0, T] \times \mathbb{R}^d)$, $\mathcal{L}_{SM}(s)$ and $\mathcal{L}_{CSM}(s)$ are equal.*

Proof. We follow the proof of Theorem 2 in (Lipman et al., 2022). We assume that $p_*(x)$ are decreasing to zero at a sufficient speed as $\|x\| \rightarrow \infty$, and $u(t, x)$, $s(t, x)$ are bounded in both time and space variables. These assumptions ensure the existence of all integrals and allow the changing of integration order (by Fubini's theorem).

To prove $\mathcal{L}_{SM}(s)$ and $\mathcal{L}_{CSM}(s)$ are equal up to a constant independent of function s , we only need to show that for any fixed $t \in [0, T]$,

$$\mathbb{E}_{x \sim p_t} [\|s(t, x) - u(t, x)\|^2] = \mathbb{E}_{y \sim p_*, x \sim p_t(x|y)} [\|s(t, x) - u(t, x|y)\|^2] + C,$$

where C is a constant function that independent of function s . We can compute that

$$\mathbb{E}_{x \sim p_t} [\|u(t, x)\|^2] = \int \|s(t, x)\|^2 p_t(x) dx = \int \int \|s(t, x)\|^2 p_t(x|y) p_*(y) dy = \mathbb{E}_{y \sim p_*, x \sim p_t(x|y)} [\|s(t, x)\|^2],$$

where the second equality we use the definition of $p_t(x)$, and in the third equality we change the order of integration.

$$\begin{aligned} \mathbb{E}_{x \sim p_t} [\langle s(t, x), u(t, x) \rangle] &= \int \langle s(t, x), \frac{\int u(t, x|y) p_t(x|y) p_*(y) dy}{p_t(x)} \rangle p_t(x) dx \\ &= \int \langle s(t, x), \int u(t, x|y) p_t(x|y) p_*(y) dy \rangle dx \\ &= \int \langle s(t, x), u(t, x|y) \rangle p_t(x|y) p_*(y) dy dx \\ &= \mathbb{E}_{y \sim p_*, x \sim p_t(x|y)} [\langle s(t, x), u(t, x|y) \rangle] \quad (\text{by Fubini's theorem}) \end{aligned}$$

Therefor we have

$$\begin{aligned}
\mathbb{E}_{x \sim p_t} [\|s(t, x) - u(t, x)\|^2] &= \mathbb{E}_{x \sim p_t} [\|s(t, x)\|^2] - 2\mathbb{E}_{x \sim p_t} [\langle s(t, x), u(t, x) \rangle] + \mathbb{E}_{x \sim p_t} [\|u(t, x)\|^2] \\
&= \mathbb{E}_{y \sim p_*, x \sim p_t(x|y)} [\|s(t, x)\|^2] - 2\mathbb{E}_{y \sim p_*, x \sim p_t(x|y)} [\langle s(t, x), u(t, x|y) \rangle] + \mathbb{E}_{x \sim p_t} [\|u(t, x)\|^2] \\
&= \mathbb{E}_{y \sim p_*, x \sim p_t(x|y)} [\|s(t, x) - u(t, x|y)\|^2] + C,
\end{aligned}$$

where the last inequality comes from the fact that $u(t, x)$ and $u(t, x|y)$ are independent of $s(t, x)$. \square

Lemma B.2. *The optimizer s^N of the objective function*

$$\min_{s \in L^2([0,1] \times \mathbb{R}^d)} \mathcal{L}_{\text{CSM}}^N(s) := \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{t \sim U[0,1], x \sim p_t(x|y_i)} [\|s(t, x) - u(t, x|y_i)\|^2]$$

has the form

$$s^N(t, x) := \frac{\sum_{i=1}^N u(t, x|y_i) p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)}, \quad t \in [0, T], x \in \mathbb{R}^d$$

Proof. Since the objective $\mathcal{L}_{\text{CSM}}(s)$ is a convex functional of s , by the first-order optimality condition, the optimizer s^N should satisfy

$$\left. \frac{\delta \mathcal{L}_{\text{CSM}}(s)}{\delta s} \right|_{s=s^N} = \frac{2}{N} \sum_{i=1}^N [s^N(t, x) - u(t, x|y_i)] p_t(x|y_i) = 0,$$

which implies that for $t \in [0, T], x \in \mathbb{R}^d$,

$$s^N(t, x) = \frac{\sum_{i=1}^N u(t, x|y_i) p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)}.$$

\square

C. Approximation error of empirical optimal score function

In this section, we provide the full proof of Theorem 3.1. For the completeness, we state the theorem again in the following:

Theorem C.1 (Approximation error of empirical optimal score function). *Let $\{y_i\}_{i=1}^N$ to be N i.i.d samples drawn from the target data distribution p_* . Denote $u(t, x)$ and $s_{\{y_i\}}^N(t, x)$ the true and empirical optimal score function respectively, as defined in (8) and (12). Then for any fixed $0 < \delta < T < \infty$, $\varepsilon_{\text{score}} > 0$ and $\tau > 0$, we have*

$$\mathbb{E}_{t \sim U[\delta, T], x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \leq \varepsilon_{\text{score}}^2,$$

with probability at least $1 - \tau$ provided that the number of training samples $N \geq N(\varepsilon_{\text{score}}, \delta, \tau)$, where $N(\varepsilon_{\text{score}}, \delta, \tau)$ is defined based on the nature of p_* :

- *Case 1: If p_* is an isotropic Gaussian, i.e. $p_*(y) = \mathcal{N}(y; \mu_{p_*}, \sigma_{p_*}^2 I_{d \times d})$, with second moment $\mathfrak{m}_2^2 = O(d)$, then $N(\varepsilon_{\text{score}}, \delta, \tau) = \frac{1}{\tau \varepsilon_{\text{score}}^2} \frac{O(d)}{\delta^{5/2}}$.*
- *Case 2: If p_* is supported on the Euclidean ball of radius R such that $R^2 = O(d)$, then $N(\varepsilon_{\text{score}}, \delta, \tau) = \frac{1}{\tau \varepsilon_{\text{score}}^2} \exp\left(\frac{O(d)}{\delta}\right)$.*

Proof of Theorem 3.1. Denote the error term

$$\left| E_{\{y_i\}}^t \right|^2 = \mathbb{E}_{x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \tag{19}$$

and

$$|E_{\{y_i\}}|^2 = \mathbb{E}_{t \sim U[\delta, T]} |E_{\{y_i\}}^t|^2 = \frac{1}{T - \delta} \int_{\delta}^T |E_{\{y_i\}}^t|^2 dt.$$

$E_{\{y_i\}}$ defines a function that maps $\{y_i\} \in \mathbb{R}^{Nd}$ to \mathbb{R}^+ , and is a random variable itself. According to the Markov's inequality:

$$\mathbb{P}(E_{\{y_i\}} > \varepsilon_{\text{score}}) \leq \frac{\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2}{\varepsilon_{\text{score}}^2}. \quad (20)$$

The final conclusions are mainly built on the upper bound of the right hand side in above Markov's inequality. We prove the results for Case 1 and Case 2 respectively.

- Case 1: Note that

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2 = \mathbb{E}_{t \sim U[\delta, T]} \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}, x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right], \quad (21)$$

and for fixed $t \in [\delta, T]$, according to the definition (19), one can show (Lemma C.2)

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}^t|^2 \lesssim \frac{O(d)}{N(1 - e^{-2t})^{7/2}} \quad (22)$$

Taking expectation with respect to t in $[\delta, T]$, we have

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2 \lesssim \frac{1}{T - \delta} \int_{\delta}^T \frac{O(d)}{N(1 - e^{-2t})^{7/2}} dt \lesssim \frac{O(d)}{N} \int_{\delta}^T \frac{1}{t^{7/2}} dt \lesssim \frac{1}{N} \frac{O(d)}{\delta^{5/2}}$$

By the Markov's inequality (20), we have

$$\mathbb{E}_{t \sim U[\delta, T], x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \leq \varepsilon_{\text{score}}^2,$$

with probability $1 - \frac{1}{N\varepsilon_{\text{score}}^2} \frac{O(d)}{\delta^{5/2}}$. Letting $\frac{1}{N\varepsilon_{\text{score}}^2} \frac{O(d)}{\delta^{5/2}} = \tau$, we compute the sample complexity $N(\varepsilon_{\text{score}}, \delta, \tau) = \frac{1}{\tau\varepsilon_{\text{score}}^2} \frac{O(d)}{\delta^{5/2}}$.

- Case 2: For fixed $t \in [\delta, T]$, according to the definition (19), one can show (Lemma C.2)

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}^t|^2 \lesssim \frac{1}{N} \frac{1}{t} \exp\left(\frac{O(d)}{t}\right) \quad (23)$$

Taking expectation with respect to t in $[\delta, T]$, we have (Lemma C.3)

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2 \lesssim \frac{1}{T - \delta} \int_{\delta}^T \frac{1}{N} \frac{1}{t} \exp\left(\frac{O(d)}{t}\right) dt \lesssim \frac{1}{N} \exp\left(\frac{O(d)}{\delta}\right)$$

Again by the Markov's inequality (20) and similar computations in Case 1, we have the sample complexity $N(\varepsilon_{\text{score}}, \delta, \tau) = \frac{1}{\tau\varepsilon_{\text{score}}^2} \exp\left(\frac{O(d)}{\delta}\right)$.

□

Lemma C.2. Under the same assumptions as in Theorem C.1, for fixed $t \in [\delta, T]$, we have

- Case 1: If p_* is an isotropic Gaussian with second moment $\mathbf{m}_2^2 = O(d)$, then

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}, x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \lesssim \frac{O(d)}{N(1 - e^{-2t})^{7/2}};$$

- Case 2: If p_* is supported on the Euclidean ball with radius $R > 0$ such that $R^2 = O(d)$, then

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}, x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \lesssim \frac{1}{N} \frac{1}{t} \exp \left(\frac{O(d)}{t} \right).$$

Proof. • Case 1: By the definitions of $u(t, x)$ and $s_{\{y_i\}}^N(t, x)$ in (8) and (12), we can rewrite them as

$$u(t, x) = \frac{\int u(t, x|y) p_t(x|y) p_*(y) dy}{\int p_t(x|y) p_*(y) dy} = -\frac{1}{\sigma(t)^2} x + \frac{\mu(t)}{\sigma(t)^2} \frac{\int y p_t(x|y) p_*(y) dy}{\int p_t(x|y) p_*(y) dy} := a_t x + b_t \frac{v_t(x)}{p_t(x)}, \quad (24)$$

$$s^N(t, x) = \frac{\sum_{i=1}^N u(t, x|y_i) p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)} = a_t x + b_t \frac{\frac{1}{N} \sum_{i=1}^N y_i p_t(x|y_i)}{\frac{1}{N} \sum_{j=1}^N p_t(x|y_j)} := a_t x + b_t \frac{v_t^N(x)}{p_t^N(x)}, \quad (25)$$

where we denote $a_t := -\frac{1}{\sigma(t)^2}$ and $b_t := \frac{\mu(t)}{\sigma(t)^2}$. Then we can compute

$$\begin{aligned} \left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 &= \left\| \left(a_t x + b_t \frac{v_t^N(x)}{p_t^N(x)} \right) - \left(a_t x + b_t \frac{v_t(x)}{p_t(x)} \right) \right\|^2 \\ &= b_t^2 \left\| \frac{1}{p_t(x)} (v_t^N(x) - v_t(x)) + \left(\frac{1}{p_t^N(x)} - \frac{1}{p_t(x)} \right) v_t(x) \right\|^2 \\ &\leq 2b_t^2 \left(\frac{1}{p_t(x)^2} \|v_t^N(x) - v_t(x)\|^2 + \left(\frac{p_t^N(x) - p_t(x)}{p_t(x)} \right)^2 \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} \right), \end{aligned}$$

where the last inequality is the Young's. Then we have:

$$\begin{aligned} &\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}, x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \\ &\leq 2b_t^2 \mathbb{E}_{x \sim p_t} \left[\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{p_t(x)^2} \|v_t^N(x) - v_t(x)\|^2 + \left(\frac{p_t^N(x) - p_t(x)}{p_t(x)} \right)^2 \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} \right] \right] \\ &\lesssim b_t^2 \mathbb{E}_{x \sim p_t} \left[\frac{\|x\|^2 + \mathbf{m}_2^2}{N \mu(t)^2} \exp \left(\frac{\|x - \mu(t) \mu_{p_*}\|^2}{2 \left(\frac{(\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2)(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)}{\mu(t)^2 \sigma_{p_*}^2} \right)} \right) \right] \quad (\text{by Lemma C.7}) \\ &\lesssim \frac{1}{N} \frac{b_t^2}{\mu(t)^2} \int (\|x\|^2 + \mathbf{m}_2^2) \exp \left(\frac{\|x - \mu(t) \mu_{p_*}\|^2}{2 \left(\frac{(\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2)(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)}{\mu(t)^2 \sigma_{p_*}^2} \right)} \right) \exp \left(-\frac{\|x - \mu(t) \mu_{p_*}\|^2}{2(\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2)} \right) dx \\ &= \frac{1}{N} \frac{b_t^2}{\mu(t)^2} \int (\|x\|^2 + \mathbf{m}_2^2) \exp \left(-\frac{\|x - \mu(t) \mu_{p_*}\|^2}{2 \left(\frac{(\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2)(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)}{\sigma(t)^2} \right)} \right) dx \\ &\propto \frac{1}{N} \frac{b_t^2}{\mu(t)^2} \frac{1}{\sigma(t)} \left[\left(\|\mu(t) \mu_{p_*}\|^2 + d \left(\frac{(\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2)(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)}{\sigma(t)^2} \right) \right) + \mathbf{m}_2^2 \right] \\ &\lesssim \frac{1}{N} \left(\frac{\mathbf{m}_2^2}{\sigma(t)^5} + \frac{d}{\sigma(t)^7} \right) \quad (\text{by the definition of } b_t = \frac{\mu(t)}{\sigma(t)^2}) \\ &\lesssim \frac{1}{N} \frac{O(d)}{\sigma(t)^7} = \frac{O(d)}{N(1 - e^{-2t})^{7/2}} \quad (\text{by } \mathbf{m}_2^2 = O(d)). \end{aligned}$$

- Case 2: We use the same notations as in Case 1 and define

$$A_1 = \mathbb{E}_{x, \{y_i\}} \left[\frac{1}{p_t(x)^2} \|v_t^N(x) - v_t(x)\|^2 \right], \quad \text{and} \quad A_2 = \mathbb{E}_{x, \{y_i\}} \left[\left(\frac{p_t^N(x) - p_t(x)}{p_t(x)} \right)^2 \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} \right].$$

Then we have:

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}, x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] \leq 2b_t^2 (A_1 + A_2).$$

We now bound terms A_1 and A_2 respectively. For term A_1 , we have

$$\begin{aligned} A_1 &= \mathbb{E}_{x \sim p_t, \{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{p_t(x)^2} \left\| v_t^N(x) - v_t(x) \right\|^2 \right] \\ &= \mathbb{E}_{x \sim p_t} \left[\frac{1}{p_t(x)^2} \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left\| v_t^N(x) - v_t(x) \right\|^2 \right] \right] \\ &\leq \frac{1}{N} \mathbb{E}_{x \sim p_t} \left[\frac{1}{p_t(x)^2} \mathbb{E}_{y \sim p_*} \|y p_t(x|y)\|^2 \right] \quad (\text{by Lemma C.4}) \\ &= \frac{1}{N} \frac{1}{(2\pi\sigma(t)^2)^d} \int \int \frac{1}{p_t(x)} \|y\|^2 \exp\left(-\frac{2\|x - \mu(t)y\|^2}{2\sigma(t)^2}\right) p_*(y) dy dx \quad (\text{by the definition of } p_t(x|y)) \\ &\leq \frac{1}{N} \frac{K_t^{-1}}{(2\pi\sigma(t)^2)^{d/2}} \int \|y\|^2 \left(\int \exp\left(\frac{1 + \lambda\mu(t)}{2\sigma(t)^2} \|x\|^2\right) \exp\left(-\frac{2\|x - \mu(t)y\|^2}{2\sigma(t)^2}\right) dx \right) p_*(y) dy \quad (\text{by Lemma C.12}) \\ &= \frac{1}{N} \frac{K_t^{-1}}{(2\pi\sigma(t)^2)^{d/2}} \int \|y\|^2 \exp\left(\frac{\mu(t)^2(1 + \lambda\mu(t))}{\sigma(t)^2(1 - \lambda\mu(t))} \|y\|^2\right) \left(\int \exp\left(-\frac{1 - \lambda\mu(t)}{2\sigma(t)^2} \left\| x - \frac{2\mu(t)}{1 - \lambda\mu(t)} y \right\|^2\right) dx \right) p_*(y) dy \\ &= \frac{1}{N} \frac{K_t^{-1}}{(1 - \lambda\mu(t))^{d/2}} \int \|y\|^2 \exp\left(\frac{\mu(t)^2(1 + \lambda\mu(t))}{\sigma(t)^2(1 - \lambda\mu(t))} \|y\|^2\right) p_*(y) dy \\ &= \frac{1}{N} \frac{1}{(1 - \lambda\mu(t))^{d/2}} \frac{\int \|y\|^2 \exp\left(\frac{\mu(t)^2(1 + \lambda\mu(t))}{\sigma(t)^2(1 - \lambda\mu(t))} \|y\|^2\right) p_*(y) dy}{\int \exp\left(-\frac{\mu(t) + \lambda\mu(t)^2}{2\lambda\sigma(t)^2} \|y\|^2\right) p_*(y) dy} \quad (\text{by the definition of } K_t) \\ &\leq \frac{1}{N} \frac{1}{(1 - \lambda\mu(t))^{d/2}} R^2 \exp\left(\frac{\mu(t)^2(1 + \lambda\mu(t))}{\sigma(t)^2(1 - \lambda\mu(t))} R^2\right) \exp\left(\frac{\mu(t) + \lambda\mu(t)^2}{2\lambda\sigma(t)^2} R^2\right) \quad (\text{by } \text{supp}(p_*) \subseteq B(0, R)) \\ &= \frac{1}{N} \frac{R^2}{(1 - \lambda\mu(t))^{d/2}} \exp\left(\frac{\mu(t)(1 + \lambda\mu(t))^2}{2\lambda\sigma(t)^2(1 - \lambda\mu(t))} R^2\right) \\ &= \frac{1}{N} 2^{d/2} R^2 \exp\left(\frac{9\mu(t)^2}{2\sigma(t)^2} R^2\right) \quad (\text{by choosing } \lambda = \frac{1}{2\mu(t)}) \\ &= \frac{1}{N} \exp\left(\frac{\mu(t)^2}{\sigma(t)^2} O(d)\right), \end{aligned}$$

where we assume $R^2 = O(d)$. For term A_2 , we can calculate

$$\begin{aligned} A_2 &= \mathbb{E}_{x \sim p_t, \{y_i\} \sim p_*^{\otimes N}} \left[\left(\frac{p_t^N(x) - p_t(x)}{p_t(x)} \right)^2 \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} \right] \\ &\leq R^2 \mathbb{E}_{x \sim p_t} \left[\frac{1}{p_t(x)^2} \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} (p_t^N(x) - p_t(x))^2 \right] \quad (\text{by Lemma C.6}) \\ &\leq \frac{R^2}{N} \mathbb{E}_{x \sim p_t} \left[\frac{1}{p_t(x)^2} \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} [p_t(x|y)^2] \right] \quad (\text{by Lemma C.4}) \\ &\leq \frac{1}{N} \frac{K_t^{-1} R^2}{(2\pi\sigma(t)^2)^{d/2}} \int \left(\int \exp\left(\frac{1 + \lambda\mu(t)}{2\sigma(t)^2} \|x\|^2\right) \exp\left(-\frac{2\|x - \mu(t)y\|^2}{2\sigma(t)^2}\right) dx \right) p_*(y) dy \quad (\text{by Lemma C.12}) \\ &\leq \frac{1}{N} \exp\left(\frac{\mu(t)^2}{\sigma(t)^2} O(d)\right) \quad (\text{by the same computations as for term } A_1) \end{aligned}$$

Combining the upper bounds for terms A_1 and A_2 , we obtain

$$\begin{aligned}
\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}, x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right] &\lesssim \frac{1}{N} \frac{\mu(t)^2}{\sigma(t)^4} \exp \left(\frac{\mu(t)^2}{\sigma(t)^2} O(d) \right) \\
&= \frac{1}{N} \frac{\exp(-2t)}{(1 - \exp(-2t))^2} \exp \left(\frac{\exp(-2t)}{1 - \exp(-2t)} O(d) \right) \\
&\quad \text{(by the definitions of } \mu(t) \text{ and } \sigma(t)) \\
&\leq \frac{1}{N} \frac{1}{t} \exp \left(\frac{O(d)}{t} \right)
\end{aligned}$$

□

Lemma C.3. $\frac{1}{T-\delta} \int_{\delta}^T \frac{1}{N} \frac{1}{t} \exp \left(\frac{O(d)}{t} \right) dt \lesssim \frac{1}{N} \exp \left(\frac{O(d)}{\delta} \right).$

Proof.

$$\begin{aligned}
\frac{1}{T-\delta} \int_{\delta}^T \frac{1}{N} \frac{1}{t} \exp \left(\frac{O(d)}{t} \right) dt &= \frac{1}{N} \frac{1}{T-\delta} \int_{1/T}^{1/\delta} \frac{\exp(O(d)s)}{s} ds \\
&\leq \frac{1}{N} \frac{T}{T-\delta} \int_{1/T}^{1/\delta} \exp(O(d)s) ds \\
&\leq \frac{1}{N} \frac{1}{O(d)} \exp \left(\frac{O(d)}{\delta} \right) \lesssim \frac{1}{N} \exp \left(\frac{O(d)}{\delta} \right).
\end{aligned}$$

□

Lemma C.4. Suppose $\{y_i\}_{i=1}^N$ are i.i.d samples drawn from the distribution p_* . For $v_t(x)$, $p_t(x)$ and $v_t^N(x)$, $p_t^N(x)$ defined in (24) and (25) respectively, we have

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left\| v_t^N(x) - v_t(x) \right\|^2 \right] \leq \frac{1}{N} \mathbb{E}_{y \sim p_*} \left[\left\| y p_t(x|y) \right\|^2 \right]$$

and

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left\| p_t^N(x) - p_t(x) \right\|^2 \right] \leq \frac{1}{N} \mathbb{E}_{y \sim p_*} \left[p_t(x|y)^2 \right].$$

Remark C.5. Define $f_{t,x}(y) := y p_t(x|y)$. Due to the randomness in y , $f_{t,x}$ is also a random variable. According to the definition (24), $v_t(x) = \int y p_t(x|y) p_*(y) dy = \mathbb{E}_{p_*} [f_{t,x}(y)]$ is the mean of random variable $f_{t,x}(y)$, and $v_t^N(x) = \frac{1}{N} \sum_i f_{t,x}(y_i)$ is the ensemble average of N realizations of $f_{t,x}$. It is always true that the variance of the ensemble average is $\frac{1}{N}$ of the variance of the original random variable, so naturally:

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left\| v_t^N(x) - v_t(x) \right\|^2 \right] = \frac{1}{N} \text{Var}_{p_*} [f_{t,x}(y)] \leq \frac{1}{N} \mathbb{E}_{p_*} \|f_{t,x}\|^2.$$

Proof. We denote $f_{t,x}(y) := y p_t(x|y)$. By the definitions of $v_t^N(x)$ and $v_t(x)$, we can compute

$$\begin{aligned}
\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left\| v_t^N(x) - v_t(x) \right\|^2 \right] &= \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left\| \frac{1}{N} \sum_{i=1}^N (f_{t,x}(y_i) - \mathbb{E}_{y \sim p_*} [f_{t,x}(y)]) \right\|^2 \right] \\
&= \frac{1}{N} \mathbb{E}_{y \sim p_*} \left[\left\| f_{t,x}(y) - \mathbb{E}_{y \sim p_*} [f_{t,x}(y)] \right\|^2 \right] \\
&\leq \frac{1}{N} \mathbb{E}_{y \sim p_*} \left[\left\| f_{t,x}(y) \right\|^2 \right] = \frac{1}{N} \mathbb{E}_{y \sim p_*} \left[\left\| y p_t(x|y) \right\|^2 \right].
\end{aligned}$$

With similar computations, one can show

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left\| p_t^N(x) - p_t(x) \right\|^2 \right] \leq \frac{1}{N} \mathbb{E}_{y \sim p_*} \left[p_t(x|y)^2 \right].$$

□

Lemma C.6. Given a collection of vectors $\{y_i\}_{i=1}^N$, for any fixed $x \in \mathbb{R}^d$ and $t \in [0, T]$, the following inequality holds

$$\frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} = \left\| \sum_{i=1}^N \frac{p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)} y_i \right\|^2 \lesssim \frac{\sigma(t)^2}{\mu(t)^2} \|x\|^2 + \frac{1}{\mu(t)^2} \frac{1}{N} \sum_{i=1}^N \|x - \mu(t)y_i\|^2,$$

where v_t^N and p_t^N are defined in (25), $p_t(x|y)$ is the Green's function defined in (6) and $\mu(t) = e^{-t}$, $\sigma(t)^2 = 1 - e^{-2t}$ as defined in (5). If we further assume that $\|y_i\|_2^2 \leq R^2$ for all $i \in [N]$, then we have

$$\frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} = \left\| \sum_{i=1}^N \frac{p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)} y_i \right\|^2 \leq R^2.$$

Proof. We can compute that

$$\begin{aligned} \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} &= \left\| \sum_{i=1}^N \frac{p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)} y_i \right\|^2 \\ &= \left\| \sum_{i=1}^N \frac{\exp\left(-\frac{\|x - \mu(t)y_i\|^2}{2\sigma(t)^2}\right)}{\sum_{j=1}^N \exp\left(-\frac{\|x - \mu(t)y_j\|^2}{2\sigma(t)^2}\right)} \frac{\sigma(t)}{\mu(t)} \frac{(\mu(t)y_i - x + x)}{\sigma(t)} \right\|^2 \\ &\lesssim \frac{\sigma(t)^2}{\mu(t)^2} \left(\|x\|^2 + \left\| \sum_{i=1}^N \frac{\exp\left(-\frac{\|x - \mu(t)y_i\|^2}{2\sigma(t)^2}\right)}{\sum_{j=1}^N \exp\left(-\frac{\|x - \mu(t)y_j\|^2}{2\sigma(t)^2}\right)} \frac{(\mu(t)y_i - x)}{\sigma(t)} \right\|^2 \right) \\ &\leq \frac{\sigma(t)^2}{\mu(t)^2} \|x\|^2 + \frac{1}{\mu(t)^2} \frac{1}{N} \sum_{i=1}^N \|x - \mu(t)y_i\|^2 \quad (\text{by Lemma C.11}) \end{aligned}$$

If we assume that $\|y_i\|_2^2 \leq R^2$ for $i \in [N]$, then we have

$$\frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} = \left\| \sum_{i=1}^N \frac{p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)} y_i \right\|^2 \leq R^2 \left\| \frac{\sum_{i=1}^N p_t(x|y_i)}{\sum_{j=1}^N p_t(x|y_j)} \right\|^2 = R^2.$$

□

Lemma C.7. Under the same assumptions as in Theorem C.1 Case 1, for fixed $t \in [\delta, T]$ and $x \in \mathbb{R}^d$, we have

$$\begin{aligned} \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{p_t(x)^2} \|v_t^N(x) - v_t(x)\|^2 + \left(\frac{p_t^N(x) - p_t(x)}{p_t(x)} \right)^2 \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} \right] \\ \leq \frac{\|x\|^2 + \mathbf{m}_2^2}{N\mu(t)^2} \exp \left(\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 \left(\frac{(\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2)(\sigma(t)^2/2 + \mu(t)^2\sigma_{p_*}^2)}{\mu(t)^2\sigma_{p_*}^2} \right)} \right) \end{aligned}$$

Here we denote $\mathbf{m}_2^2 := \mathbb{E}_{y \sim p_*} [\|y\|^2] = \|\mu_{p_*}\|^2 + d\sigma_{p_*}^2$.

Proof. Denote

$$A_1 := \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{p_t(x)^2} \|v_t^N(x) - v_t(x)\|^2 \right] \quad \text{and} \quad A_2 := \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left(\frac{p_t^N(x) - p_t(x)}{p_t(x)} \right)^2 \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} \right]$$

We now bound terms A_1 and A_2 respectively. For term A_1 , we have

$$\begin{aligned}
A_1 &= \frac{1}{p_t(x)^2} \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\|v_t^N(x) - v_t(x)\|^2 \right] \\
&\leq \frac{1}{N} \frac{1}{p_t(x)^2} \mathbb{E}_{y \sim p_*} \left[\|y p_t(x|y)\|^2 \right] \quad (\text{by Lemma C.4}) \\
&\lesssim \frac{\|x\|^2 + \mathbf{m}_2^2}{N} \exp \left(\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 \left(\frac{\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2}{2} \right)} \right) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 (\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right) \quad (\text{by Lemma C.10}) \quad (26) \\
&= \frac{\|x\|^2 + \mathbf{m}_2^2}{N} \exp \left(\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 \left(\frac{(\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2)(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)}{\mu(t)^2 \sigma_{p_*}^2} \right)} \right)
\end{aligned}$$

For term A_2 , by Lemma (C.6) we obtain

$$\begin{aligned}
A_2 &= \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\left(\frac{p_t^N(x) - p_t(x)}{p_t(x)} \right)^2 \frac{\|v_t^N(x)\|^2}{p_t^N(x)^2} \right] \\
&\lesssim \frac{1}{\mu(t)^2} \frac{1}{p_t(x)^2} \left(\sigma(t)^2 \|x\|^2 \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[(p_t^N(x) - p_t(x))^2 \right] + \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{N} \sum_{i=1}^N \|x - \mu(t)y_i\|^2 (p_t^N(x) - p_t(x))^2 \right] \right) \\
&:= \frac{1}{\mu(t)^2} \frac{1}{p_t(x)^2} (A_{2,1} + A_{2,2})
\end{aligned}$$

By Lemma C.10, we have

$$A_{2,1} = \sigma(t)^2 \|x\|^2 \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[(p_t^N(x) - p_t(x))^2 \right] \lesssim \frac{\sigma(t)^2 \|x\|^2}{N} \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 (\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right)$$

By Lemma C.8, we know that

$$A_{2,2} = \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{N} \sum_{i=1}^N \|x - \mu(t)y_i\|^2 (p_t^N(x) - p_t(x))^2 \right] \lesssim \frac{\|x\|^2 + \mathbf{m}_2^2}{N} \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 (\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right)$$

Then we can obtain the upper bound for term A_2 , i.e.

$$\begin{aligned}
A_2 &\lesssim \frac{1}{\mu(t)^2} \frac{1}{p_t(x)^2} (A_{2,1} + A_{2,2}) \\
&\lesssim \frac{1}{\mu(t)^2} \frac{1}{p_t(x)^2} \frac{\|x\|^2 + \mathbf{m}_2^2}{N} \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 (\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right) \\
&\lesssim \frac{1}{\mu(t)^2} \frac{\|x\|^2 + \mathbf{m}_2^2}{N} \exp \left(\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 \left(\frac{(\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2)(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)}{\mu(t)^2 \sigma_{p_*}^2} \right)} \right) \quad (27)
\end{aligned}$$

We finish the proof by combining the upper bounds of terms A_1 and A_2 derived in (26) and (27). \square

Lemma C.8. *Under the same assumptions as in Lemma C.7, we have*

$$A_{2,2} := \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{N} \sum_{i=1}^N \|x - \mu(t)y_i\|^2 (p_t^N(x) - p_t(x))^2 \right] \lesssim \frac{\|x\|^2 + \mathbf{m}_2^2}{N} \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 (\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right)$$

Proof. For notation simplicity, we denote $g_{t,x}(y) := p_t(x|y)$ and use $\mathbb{E}_{\{y_i\}}$ as a short notation of $\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}}$ when the

context is clear. Then we have

$$\begin{aligned} A_{2,2} &= \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} \left[\frac{1}{N} \sum_{i=1}^N \|x - \mu(t)y_i\|^2 (p_t^N(x) - p_t(x))^2 \right] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\{y_i\}} \left[\|x - \mu(t)y_i\|^2 \left(\frac{1}{N} \sum_{j=1}^N (g_{t,x}(y_j) - \mathbb{E}_{y_j}[g_{t,x}(y_j)]) \right)^2 \right] \end{aligned}$$

For every $i \in [N]$, we can compute

$$\begin{aligned} &\mathbb{E}_{\{y_k\}_{k=1}^N} \left[\|x - \mu(t)y_i\|^2 \left(\frac{1}{N} \sum_{j=1}^N (g_{t,x}(y_j) - \mathbb{E}_{y_j}[g_{t,x}(y_j)]) \right)^2 \right] \\ &= \mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 \mathbb{E}_{\{y_j\}_{j \neq i}^N} \left[\frac{1}{N^2} \left((g_{t,x}(y_i) - \mathbb{E}_{y_i}[g_{t,x}(y_i)]) + \sum_{j \neq i}^N (g_{t,x}(y_j) - \mathbb{E}_{y_j}[g_{t,x}(y_j)]) \right)^2 \right] \right] \\ &\lesssim \frac{1}{N^2} \mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 \mathbb{E}_{\{y_j\}_{j \neq i}^N} \left[(g_{t,x}(y_i) - \mathbb{E}_{y_i}[g_{t,x}(y_i)])^2 + \left(\sum_{j \neq i}^N (g_{t,x}(y_j) - \mathbb{E}_{y_j}[g_{t,x}(y_j)]) \right)^2 \right] \right] \\ &= \frac{1}{N^2} \left(\mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 (g_{t,x}(y_i) - \mathbb{E}_{y_i}[g_{t,x}(y_i)])^2 \right] \right. \\ &\quad \left. + \mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 \right] \mathbb{E}_{\{y_j\}_{j \neq i}^N} \left[\left(\sum_{j \neq i}^N (g_{t,x}(y_j) - \mathbb{E}_{y_j}[g_{t,x}(y_j)]) \right)^2 \right] \right) \\ &\leq \frac{1}{N^2} \left(\mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 (g_{t,x}(y_i) - \mathbb{E}_{y_i}[g_{t,x}(y_i)])^2 \right] + \mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 \right] (N-1) \mathbb{E}_y \left[(g_{t,x}(y))^2 \right] \right) \end{aligned}$$

Therefore, we have

$$\begin{aligned} A_{2,2} &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\{y_i\}} \left[\|x - \mu(t)y_i\|^2 \left(\frac{1}{N} \sum_{j=1}^N (g_{t,x}(y_j) - \mathbb{E}_{y_j}[g_{t,x}(y_j)]) \right)^2 \right] \\ &\lesssim \frac{1}{N} \sum_{i=1}^N \frac{1}{N^2} \left(\mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 (g_{t,x}(y_i) - \mathbb{E}_{y_i}[g_{t,x}(y_i)])^2 \right] \right. \\ &\quad \left. + (N-1) \mathbb{E}_{y_i} \left[\|x - \mu(t)y_i\|^2 \right] \mathbb{E}_y \left[(g_{t,x}(y))^2 \right] \right) \\ &= \frac{1}{N^2} \mathbb{E}_y \left[\|x - \mu(t)y\|^2 (g_{t,x}(y) - \mathbb{E}_y[g_{t,x}(y)])^2 \right] + \frac{N-1}{N^2} \mathbb{E}_y \left[\|x - \mu(t)y\|^2 \right] \mathbb{E}_y \left[(g_{t,x}(y))^2 \right] \\ &:= A_{2,2,1} + A_{2,2,2} \end{aligned}$$

Note that

$$\begin{aligned} A_{2,2,1} &= \frac{1}{N^2} \mathbb{E}_y \left[\|x - \mu(t)y\|^2 (g_{t,x}(y) - \mathbb{E}_y[g_{t,x}(y)])^2 \right] \\ &\lesssim \frac{1}{N^2} \left(\mathbb{E}_y \left[\|x - \mu(t)y\|^2 g_{t,x}(y)^2 \right] + (\mathbb{E}_y[g_{t,x}(y)])^2 \mathbb{E}_y \left[\|x - \mu(t)y\|^2 \right] \right) \\ &:= \frac{1}{N^2} (B_1 + B_2) \end{aligned}$$

For term B_1 , we have

$$\begin{aligned}
B_1 &= \mathbb{E}_y \left[\|x - \mu(t)y\|^2 g_{t,x}(y) \right] \\
&= \mathbb{E}_y \left[\|x - \mu(t)y\|^2 \exp \left(-\frac{\|x - \mu(t)y\|^2}{2(\sigma(t)^2/2)} \right) \right] \\
&= \mathbb{E}_{\tilde{y}} \left[\|\tilde{y}\|^2 \exp \left(-\frac{\|\tilde{y}\|^2}{2(\sigma(t)^2/2)} \right) \right], \quad \text{where } \tilde{y} := x - \mu(t)y \sim \mathcal{N}(\tilde{y}; x - \mu(t)\mu_{p_*}, \mu(t)^2 \sigma_{p_*}^2 I_{d \times d}) \\
&\lesssim \int \|\tilde{y}\|^2 \exp \left(-\frac{\|\tilde{y}\|^2}{2(\sigma(t)^2/2)} \right) \exp \left(-\frac{\|\tilde{y} - (x - \mu(t)\mu_{p_*})\|^2}{2\mu(t)^2 \sigma_{p_*}^2} \right) dy \\
&\lesssim \mathcal{N} \left(x; \mu(t)\mu_{p_*}, \left(\frac{\sigma(t)^2}{2} + \mu(t)^2 \sigma_{p_*}^2 \right) I_{d \times d} \right) \mathbb{E}_{\hat{Y}} \left[\|\hat{Y}\|^2 \right], \quad (\text{similar to the computations in (28)}) \\
&\quad \text{where } \hat{Y} \sim \mathcal{N} \left(\hat{y}; \frac{\sigma(t)^2(x - \mu(t)\mu_{p_*})}{\sigma(t)^2 + 2\mu(t)^2 \sigma_{p_*}^2}, \frac{\sigma(t)^2 \mu(t)^2 \sigma_{p_*}^2}{\sigma(t)^2 + 2\mu(t)^2 \sigma_{p_*}^2} I_{d \times d} \right) \\
&\lesssim (\|x - \mu(t)\mu_{p_*}\|^2 + d\sigma_{p_*}^2) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right) \\
&\lesssim (\|x\|^2 + \|\mu_{p_*}\|^2 + d\sigma_{p_*}^2) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right) \\
&= (\|x\|^2 + \mathbf{m}_2^2) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right)
\end{aligned}$$

For term B_2 , we have

$$(\mathbb{E}_y [g_{t,x}(y)])^2 = \left(\mathbb{E}_y \left[\exp \left(-\frac{\|x - \mu(t)y\|^2}{2\sigma(t)^2} \right) \right] \right)^2 \lesssim \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 \left(\frac{\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2}{2} \right)} \right) \quad (\text{by Lemma C.10})$$

and

$$\begin{aligned}
\mathbb{E}_y [\|x - \mu(t)y\|^2] &= \mathbb{E}_{\tilde{y}} [\|\tilde{y}\|^2], \quad \text{where } \tilde{y} := x - \mu(t)y \sim \mathcal{N}(\tilde{y}; x - \mu(t)\mu_{p_*}, \mu(t)^2 \sigma_{p_*}^2 I_{d \times d}) \\
&= \|x - \mu(t)\mu_{p_*}\|^2 + d\mu(t)^2 \sigma_{p_*}^2 \\
&\lesssim \|x\|^2 + \mathbf{m}_2^2
\end{aligned}$$

Therefore, we know that

$$\begin{aligned}
B_2 &= (\mathbb{E}_y [g_{t,x}(y)])^2 \mathbb{E}_y [\|x - \mu(t)y\|^2] \lesssim (\|x\|^2 + \mathbf{m}_2^2) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2 \left(\frac{\sigma(t)^2 + \mu(t)^2 \sigma_{p_*}^2}{2} \right)} \right) \\
&\leq (\|x\|^2 + \mathbf{m}_2^2) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right)
\end{aligned}$$

Then, we can have the upper bound for $A_{2,2,1}$, i.e.

$$A_{2,2,1} = \frac{1}{N^2} (B_1 + B_2) \lesssim \frac{1}{N^2} (\|x\|^2 + \mathbf{m}_2^2) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right)$$

Similar to the computations for term $A_{2,2,1}$, we can compute the upper bound of $A_{2,2,2}$ as the following:

$$A_{2,2,2} = \frac{N-1}{N^2} \mathbb{E}_y [\|x - \mu(t)y\|^2] \mathbb{E}_y [(g_{t,x}(y))^2] \lesssim \frac{N-1}{N^2} (\|x\|^2 + \mathbf{m}_2^2) \exp \left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2(\sigma(t)^2/2 + \mu(t)^2 \sigma_{p_*}^2)} \right)$$

Therefore, for term $A_{2,2}$, we have

$$A_{2,2} \lesssim A_{2,2,1} + A_{2,2,2} \lesssim \frac{\|x\|^2 + \mathbf{m}_2^2}{N} \exp\left(-\frac{\|x - \mu(t)\mu_{p_*}\|^2}{2(\sigma(t)^2/2 + \mu(t)^2\sigma_{p_*}^2)}\right)$$

□

Lemma C.9 (Convolution of two Gaussian distributions). *Let $f_X(x) = \mathcal{N}(x; \mu_X, \sigma_X^2 I_{d \times d})$ and $f_Y(y) = \mathcal{N}(y; \mu_Y, \sigma_Y^2 I_{d \times d})$, then*

$$f_Z(z) := \int f_X(z - y) f_Y(y) dy = \mathcal{N}(z; \mu_X + \mu_Y, (\sigma_X^2 + \sigma_Y^2) I_{d \times d})$$

Proof. One can compute that

$$\begin{aligned} f_Z(z) &= \int f_X(z - y) f_Y(y) dy \\ &\propto \int \exp\left(-\frac{\|z - y - \mu_X\|^2}{2\sigma_X^2}\right) \exp\left(-\frac{\|y - \mu_Y\|^2}{2\sigma_Y^2}\right) dy \\ &= \int \exp\left(-\frac{1}{2\sigma_X^2\sigma_Y^2} \left[\sigma_Y^2 (\|z\|^2 + \|y\|^2 + \|\mu_X\|^2 - 2z^T y - 2z^T \mu_X + 2\mu_X^T y) \right. \right. \\ &\quad \left. \left. + \sigma_X^2 (\|y\|^2 - 2\mu_Y^T y + \|\mu_Y\|^2) \right] \right) dy \\ &\propto \int \exp\left(-\frac{1}{2\sigma_X^2\sigma_Y^2} \left[(\sigma_X^2 + \sigma_Y^2) \|y\|^2 - 2(\sigma_Y^2(z - \mu_X) + \sigma_X^2\mu_Y)^T y + \sigma_Y^2 \|z\|^2 \right] \right) dy \end{aligned}$$

Define $\sigma_Z := \sqrt{\sigma_X^2 + \sigma_Y^2}$, and completing the square:

$$\begin{aligned} f_Z(z) &\propto \exp\left(-\frac{\|z\|^2}{2\sigma_X^2}\right) \int \exp\left(-\frac{1}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_Z}\right)^2} \left(\|y\|^2 - \frac{2}{\sigma_Z^2} (\sigma_Y^2(z - \mu_X) + \sigma_X^2\mu_Y)^T y \right) \right) dy \\ &\propto \exp\left(-\frac{\|z\|^2}{2\sigma_X^2} + \frac{\|\sigma_Y^2(z - \mu_X) + \sigma_X^2\mu_Y\|^2}{2\sigma_Z^2(\sigma_X\sigma_Y)^2}\right) \int \exp\left(-\frac{1}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_Z}\right)^2} \left\| y - \frac{\sigma_Y^2(z - \mu_X) + \sigma_X^2\mu_Y}{\sigma_Z^2} \right\|^2 \right) dy \\ &\propto \exp\left(-\frac{\|z - (\mu_X + \mu_Y)\|^2}{2(\sigma_X^2 + \sigma_Y^2)}\right) E_{\hat{Y}}[\mathbb{I}\{\hat{Y} \leq +\infty\}], \text{ where } \hat{Y} \sim \mathcal{N}\left(\hat{y}; \frac{\sigma_Y^2(z - \mu_X) + \sigma_X^2\mu_Y}{\sigma_Z^2}, \frac{\sigma_X^2\sigma_Y^2}{\sigma_Z^2} I_{d \times d}\right) \\ &\propto \mathcal{N}(z; \mu_X + \mu_Y, (\sigma_X^2 + \sigma_Y^2) I_{d \times d}) \end{aligned} \tag{28}$$

□

Lemma C.10. *Suppose $y \sim p_* = \mathcal{N}(y; \mu_{p_*}, \sigma_{p_*}^2 I_{d \times d})$, then one can compute the following quantities:*

1.

$$p_t(x) = \mathbb{E}_{y \sim p_*}[p_t(x|y)] = \mathcal{N}(x; \mu(t)\mu_{p_*}, (\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2) I_{d \times d}) := h(x)$$

2.

$$\mathbb{E}_{y \sim p_*} \left[y \exp\left(-\frac{\|x - \mu(t)y\|^2}{2\sigma(t)^2}\right) \right] \propto \left(\frac{\mu(t)\sigma_{p_*}^2 x + \sigma(t)^2 \mu_{p_*}}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2} \right) h(x)$$

3.

$$\mathbb{E}_{y \sim p_*} \left[\|y\|^2 \exp\left(-\frac{\|x - \mu(t)y\|^2}{2\sigma(t)^2}\right) \right] \lesssim (\|x\|^2 + \mathbf{m}_2^2) h(x),$$

where $\mathbf{m}_2^2 := \|\mu_{p_*}\|^2 + d\sigma_{p_*}^2$. Both \propto and \lesssim indicate ignoring the constants.

Proof. 1.

$$\begin{aligned}
\mathbb{E}_{y \sim p_*} [p_t(x|y)] &\propto \mathbb{E}_{y \sim p_*} \left[\exp \left(-\frac{\|x - \mu(t)y\|^2}{2\sigma(t)^2} \right) \right] \\
&\propto \int \exp \left(-\frac{\|x - \mu(t)y\|^2}{2\sigma(t)^2} \right) \exp \left(-\frac{\|y - \mu_{p_*}\|^2}{2\sigma_{p_*}^2} \right) dy \\
&= \int \exp \left(-\frac{\|x/\mu(t) - y\|^2}{2\sigma(t)^2/\mu(t)^2} \right) \exp \left(-\frac{\|y - \mu_{p_*}\|^2}{2\sigma_{p_*}^2} \right) dy \\
&= \mathcal{N} \left(\frac{x}{\mu(t)}; 0, \frac{\sigma(t)^2}{\mu(t)^2} I_{d \times d} \right) * \mathcal{N} (y; \mu_{p_*}, \sigma_{p_*}^2 I_{d \times d}) \\
&= \mathcal{N} (x; \mu(t)\mu_{p_*}, (\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2) I_{d \times d}) \quad (\text{by Lemma C.9})
\end{aligned} \tag{29}$$

2. Similar to the computations in (28) and (29), one can compute

$$\begin{aligned}
\mathbb{E}_{y \sim p_*} \left[y \exp \left(-\frac{\|x - \mu(t)y\|^2}{2\sigma(t)^2} \right) \right] &\propto \mathcal{N} (x; \mu(t)\mu_{p_*}, (\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2) I_{d \times d}) \mathbb{E}_{\hat{Y}} [\hat{Y}], \\
&\quad \left(\text{where } \hat{Y} \sim \mathcal{N} \left(\hat{y}; \frac{\mu(t)\sigma_{p_*}^2 x + \sigma(t)^2 \mu_{p_*}}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2}, \frac{\sigma(t)^2\sigma_{p_*}^2}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2} I_{d \times d} \right) \right) \\
&= \left(\frac{\mu(t)\sigma_{p_*}^2 x + \sigma(t)^2 \mu_{p_*}}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2} \right) h(x)
\end{aligned}$$

3.

$$\begin{aligned}
\mathbb{E}_{y \sim p_*} \left[\|y\|^2 \exp \left(-\frac{\|x - \mu(t)y\|^2}{2\sigma(t)^2} \right) \right] &\propto \mathcal{N} (x; \mu(t)\mu_{p_*}, (\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2) I_{d \times d}) \mathbb{E}_{\hat{Y}} [\|\hat{Y}\|^2], \\
&\quad \left(\text{where } \hat{Y} \sim \mathcal{N} \left(\hat{y}; \frac{\mu(t)\sigma_{p_*}^2 x + \sigma(t)^2 \mu_{p_*}}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2}, \frac{\sigma(t)^2\sigma_{p_*}^2}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2} I_{d \times d} \right) \right) \\
&= \left(\left\| \frac{\mu(t)\sigma_{p_*}^2 x + \sigma(t)^2 \mu_{p_*}}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2} \right\|^2 + d \frac{\sigma(t)^2\sigma_{p_*}^2}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2} \right) h(x) \\
&\lesssim (\|x\|^2 + \|\mu_{p_*}\|^2 + d\sigma_{p_*}^2) h(x) = (\|x\|^2 + \mathbf{m}_2^2) h(x)
\end{aligned}$$

□

Lemma C.11. *Given a collection of d -dimensional vectors $\{y_i\}_{i=1}^N$, the following inequality holds*

$$\left\| \sum_{i=1}^N \frac{\exp(-\|y_i\|^2)}{\sum_{j=1}^N \exp(-\|y_j\|^2)} y_i \right\|^2 \leq \frac{1}{N} \sum_{i=1}^N \|y_i\|^2$$

Proof. Denote $w_i^N := \frac{\exp(-\|y_i\|^2)}{\sum_{j=1}^N \exp(-\|y_j\|^2)}$ for all $i = 1, 2, \dots, n$, then we can compute

$$\begin{aligned}
\left\| \sum_{i=1}^N \frac{\exp(-\|y_i\|^2)}{\sum_{j=1}^N \exp(-\|y_j\|^2)} y_i \right\|^2 &= \left\| \sum_{i=1}^N w_i^N y_i \right\|^2 \\
&= \sum_{i=1}^N \sum_{j=1}^N w_i^N w_j^N y_i^T y_j \\
&\leq \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N w_i^N w_j^N (\|y_i\|^2 + \|y_j\|^2) \\
&= \sum_{i=1}^N w_i^N \|y_i\|^2 \quad (\text{by the fact that } \sum_{i=1}^N w_i^N = 1) \\
&\leq \frac{1}{N} \sum_{i=1}^N \|y_i\|^2.
\end{aligned}$$

□

Lemma C.12. For any $\lambda > 0$, with the Green's function $p_t(x|y)$ defined in (6), $p_t(x) := \int p_t(x|y)p_*(y)dy$ is lower bounded by

$$\begin{aligned}
p_t(x) &\geq \frac{1}{(2\pi\sigma(t)^2)^{d/2}} \exp\left(-\frac{1+\lambda\mu(t)}{2\sigma(t)^2}\|x\|^2\right) \int \exp\left(-\frac{\mu(t)+\lambda\mu(t)^2}{2\lambda\sigma(t)^2}\|y\|^2\right) p_*(y)dy \\
&:= \frac{1}{(2\pi\sigma(t)^2)^{d/2}} \exp\left(-\frac{1+\lambda\mu(t)}{2\sigma(t)^2}\|x\|^2\right) K_t
\end{aligned}$$

Proof. It comes from the direct computation:

$$\begin{aligned}
p_t(x) &= \int p_t(x|y)p_*(y)dy \\
&= \int \frac{1}{(2\pi\sigma(t)^2)^{d/2}} \exp\left(-\frac{\|x-\mu(t)y\|^2}{2\sigma(t)^2}\right) p_*(y)dy \\
&= \frac{1}{(2\pi\sigma(t)^2)^{d/2}} \int \exp\left(-\frac{1}{2\sigma(t)^2}(\|x\|^2 - 2\mu(t)x^T y + \mu(t)^2\|y\|^2)\right) p_*(y)dy \\
&\geq \frac{1}{(2\pi\sigma(t)^2)^{d/2}} \int \exp\left(-\frac{1}{2\sigma(t)^2}\left(\|x\|^2 + \lambda\mu(t)\|x\|^2 + \frac{1}{\lambda}\mu(t)\|y\|^2 + \mu(t)^2\|y\|^2\right)\right) p_*(y)dy \\
&= \frac{1}{(2\pi\sigma(t)^2)^{d/2}} \exp\left(-\frac{1+\lambda\mu(t)}{2\sigma(t)^2}\|x\|^2\right) \int \exp\left(-\frac{\mu(t)+\lambda\mu(t)^2}{2\lambda\sigma(t)^2}\|y\|^2\right) p_*(y)dy,
\end{aligned}$$

where the inequality comes from Young's inequality, i.e. $2a^T b \leq \lambda\|a\|^2 + \frac{1}{\lambda}\|b\|^2$ for any $\lambda > 0$. □

D. Memorization effects

In this section, we provide the proof of Propositions 4.1 and 4.2, and Theorem 4.3. For the completeness, we state all the propositions and theorem again before the proof.

Proposition D.1. Suppose the training samples $\{y_i\}_{i=1}^N$ satisfy $\|y_i\|_2 \leq d$, for $\delta \geq 0$, $\text{TV}(\mathbf{q}_{T-\delta}, \mathbf{p}_*^\gamma) \leq \frac{d\sqrt{\delta}}{2}$ with $\gamma = \sigma(\delta)$, where $\sigma(\cdot)$ is defined in (5).

Proof of Proposition 4.1. By the definition of total variation, we have

$$\begin{aligned}
\text{TV}(\mathbf{q}_{T-\delta}, \mathbf{p}_*^\gamma) &= \frac{1}{2} \int_{\mathbb{R}^d} \left| \frac{1}{N} \sum_{i=1}^N \mathcal{N}(x; \mu(\delta)y_i, \sigma(\delta)^2 I_{d \times d}) - \frac{1}{N} \sum_{i=1}^N \mathcal{N}(x; y_i, \sigma(\delta)^2 I_{d \times d}) \right| dx \\
&\leq \frac{1}{N} \sum_{i=1}^N \frac{1}{2} \int_{\mathbb{R}^d} |\mathcal{N}(x; \mu(\delta)y_i, \sigma(\delta)^2 I_{d \times d}) - \mathcal{N}(x; y_i, \sigma(\delta)^2 I_{d \times d})| dx \\
&= \frac{1}{N} \sum_{i=1}^N \text{TV}(\mathcal{N}(x; \mu(\delta)y_i, \sigma(\delta)^2 I_{d \times d}), \mathcal{N}(x; y_i, \sigma(\delta)^2 I_{d \times d})) \\
&\leq \frac{1}{N} \sum_{i=1}^N \sqrt{\frac{1}{2} \text{KL}(\mathcal{N}(x; \mu(\delta)y_i, \sigma(\delta)^2 I_{d \times d}) \parallel \mathcal{N}(x; y_i, \sigma(\delta)^2 I_{d \times d}))} \\
&= \frac{1}{N} \sum_{i=1}^N \|y_i\|_2 \frac{1 - \mu(\delta)}{2\sigma(\delta)} \quad (\text{by Lemma D.4}) \\
&\leq \frac{1 - \exp(-\delta)}{2\sqrt{1 - \exp(-2\delta)}} d \quad (\text{by the definitions of } \mu(\delta) \text{ and } \sigma(\delta), \text{ and } \|y_i\|_2 \leq d) \\
&\leq \frac{d\sqrt{\delta}}{2}.
\end{aligned}$$

□

Proposition D.2. *Under the same assumptions as in Proposition 4.1, on the time interval $t \in [0, T]$, the total variation between the output distribution of SGM algorithm (13) with the empirical optimal score function $\hat{\mathbf{q}}_t$ and the KDE approximation \mathbf{q}_t – is bounded by $\text{TV}(\hat{\mathbf{q}}_t, \mathbf{q}_t) \leq \frac{d}{2} \exp(-T)$.*

Proof of Proposition 4.2. By the data-processing inequality and Lemma D.5, we have

$$\text{TV}(\mathbf{q}_t, \hat{\mathbf{q}}_t) \leq \text{TV}(\mathbf{q}_0, \hat{\mathbf{q}}_0) = \text{TV}(\mathbf{p}_T, \pi^d) \leq \frac{d}{2} \exp(-T).$$

□

Theorem D.3 (SGM with empirical optimal score function resembles KDE). *Under the same assumptions as Proposition 4.2, SGM algorithm (13) with the empirical optimal score function s^N returns a simple Gaussian convolution with the empirical distribution in the form of (18), and it presents the following behavior:*

- *(with early stopping)* for any $\varepsilon > 0$, set $T = \log \frac{d}{\varepsilon}$ and $\delta = \frac{\varepsilon^2}{d}$, we have

$$\text{TV}(\hat{\mathbf{q}}_{T-\delta}, \mathbf{p}_*^\gamma) \leq \varepsilon, \quad \text{with } \gamma = \sigma(\delta),$$

- *(without early stopping)* by taking the limit $T \rightarrow +\infty$ and $\delta = 0$, we have $\hat{\mathbf{q}}_\infty = \mathbf{p}_* = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$.

Proof of Theorem 4.3.

(with early stopping) For $0 \leq \delta < T$, combining Proposition 4.1 and Proposition 4.2 using triangle inequality, we have

$$\text{TV}(\hat{\mathbf{q}}_{T-\delta}, \mathbf{p}_*^\gamma) \leq \text{TV}(\mathbf{q}_{T-\delta}, \mathbf{p}_*^\gamma) + \text{TV}(\mathbf{q}_{T-\delta}, \hat{\mathbf{q}}_{T-\delta}) \leq \frac{d}{2} (\sqrt{\delta} + \exp(-T)) \quad (30)$$

For any $\varepsilon > 0$, by choosing $T = \log \frac{d}{\varepsilon}$ and $\delta = \frac{\varepsilon^2}{d}$, we obtain $\text{TV}(\hat{\mathbf{q}}_{T-\delta}, \mathbf{p}_*^\gamma) \leq \varepsilon$.

(without early stopping) By taking the limit $T \rightarrow +\infty$ and $\delta = 0$ in inequality (30), we have $\text{TV}(\hat{\mathbf{q}}_\infty, \mathbf{p}_*) \leq 0$. This implies that $\hat{\mathbf{q}}_\infty$ equals to the empirical distribution $\mathbf{p}_* = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$.

□

Lemma D.4 (KL divergence between two Gaussian distributions). *Let $p = \mathcal{N}(\mu_p, \Sigma_p)$ and $q = \mathcal{N}(\mu_q, \Sigma_q)$ be two Gaussian distributions on \mathbb{R}^d . Then the KL divergence between p and q is*

$$\text{KL}(p||q) = \frac{1}{2} \left[\log \frac{|\Sigma_q|}{|\Sigma_p|} - d + (\mu_p - \mu_q)^T \Sigma_q^{-1} (\mu_p - \mu_q) + \text{Tr} \{ \Sigma_q^{-1} \Sigma_p \} \right]$$

Lemma D.5 (Convergence of forward OU process). *Denote \mathbf{p}_T to be the distribution of forward OU process at time T initializing with the empirical distribution $\mathbf{p}_* = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$, where $\{y_i\}_{i=1}^N$ are i.i.d samples such that $\|y_i\|_2 \leq d$. Then for $T \geq 1$,*

$$\text{TV}(\mathbf{p}_T, \pi^d) \leq \frac{d}{2} \exp(-T).$$

Proof. Since $p_t(x|y) = \mathcal{N}(x; \exp(-t)y, \sigma(t)^2 I_{d \times d})$, by Lemma D.4 we have

$$\text{KL}(p_t(x|y)||\pi^d) = \frac{1}{2} \left[-d \log \sigma(t)^2 - d + d\sigma(t)^2 + \|\exp(-t)y\|^2 \right]$$

By the convexity of the KL divergence,

$$\begin{aligned} \text{KL}(\mathbf{p}_T||\pi^d) &= \text{KL} \left(\int_{\mathbb{R}^d} p_T(x|y) \mathbf{p}_*(y) dy \middle| \middle| \pi^d \right) \\ &\leq \int \text{KL}(p_T(x|y)||\pi^d) \mathbf{p}_*(y) dy \\ &= \frac{1}{2} \left[-d \log \sigma(T)^2 - d + d\sigma(T)^2 + \exp(-2T) \mathbb{E}_{y \sim \mathbf{p}_*} \|y\|^2 \right] \\ &= \frac{1}{2} \left[-d \log(1 - \exp(-2T)) - d + d(1 - \exp(-2T)) + \exp(-2T) \mathbb{E}_{y \sim \mathbf{p}_*} \|y\|^2 \right] \\ &\leq \frac{1}{2} \exp(-2T) \mathbb{E}_{y \sim \mathbf{p}_*} [\|y\|^2] \quad (\text{by the fact } \log(1-x) \geq -x \text{ for } x \geq 0) \\ &= \frac{1}{2} \exp(-2T) \frac{1}{N} \sum_{i=1}^N \|y_i\|_2^2 \leq \frac{d^2}{2} \exp(-2T) \end{aligned}$$

By the Pinsker's inequality, we have

$$\text{TV}(\mathbf{p}_T, \pi^d) \leq \sqrt{\frac{1}{2} \text{KL}(\mathbf{p}_T||\pi^d)} \leq \frac{d}{2} \exp(-T).$$

□

E. Numerical experiments

In this section, we provide the details of the numerical experiments. The code is available in https://github.com/SixuLi/DDPM_and_KDE.¹

E.1. Synthetic data distribution

We consider the target data distribution p_* is a 2-dimensional isotropic Gaussian, i.e. $p_*(x) := \mathcal{N}(x; \mu_{p_*}, \sigma_{p_*}^2 I_{2 \times 2})$. In this case, the law of forward OU process (3) p_t and the exact score function $u(t, x)$ defined in (8) have explicit formulations. Specifically, by Lemma C.10, we obtain

$$p_t(x) = \int p_t(x|y)p_*(y)dy = \mathcal{N}(x; \mu(t)\mu_{p_*}, (\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2) I_{2 \times 2}) \quad (31)$$

$$u(t, x) = \frac{\int u(t, x|y)p_t(x|y)p_*(y)dy}{\int p_t(x|y)p_*(y)dy} = -\frac{1}{\sigma(t)^2}x + \frac{\mu(t)}{\sigma(t)^2} \frac{\int yp_t(x|y)p_*(y)dy}{\int p_t(x|y)p_*(y)dy} = \frac{\mu(t)\mu_{p_*} - x}{\sigma(t)^2 + \mu(t)^2\sigma_{p_*}^2}, \quad (32)$$

where $\mu(t) = \exp(-t)$ and $\sigma(t)^2 = 1 - \exp(-2t)$ as defined in (5). We set choose $\mu_{p_*} = [-5, 5]$ and $\sigma_{p_*}^2 = 10$ in our experiments.

We first estimate the score approximation error of the empirical optimal score function $s_{\{y_i\}}^N$ across various training sample sizes N . Setting early stopping time $\delta = 0.02$, time interval length $T = 5$, and sample size N ranging from $N = 100$ to $N = 2000$, we numerically estimate

$$\mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}} |E_{\{y_i\}}|^2 = \mathbb{E}_{t \sim U[\delta, T]} \mathbb{E}_{\{y_i\} \sim p_*^{\otimes N}, x \sim p_t} \left[\left\| s_{\{y_i\}}^N(t, x) - u(t, x) \right\|^2 \right], \quad (33)$$

using the empirical average

$$\widehat{|E_{\{y_i\}}|}^2 := \frac{1}{K} \frac{1}{M} \sum_{k=1}^K \sum_{m=1}^M \left\| s_{\{y_i\}}^N(t_k, x_m^{t_k}) - u(t_k, x_m^{t_k}) \right\|^2 \quad (34)$$

This is achieved through repeating the following steps 10 times and computing the average output:

1. Randomly sample N training data $\{y_i\}_{i=1}^N$ from the target distribution p_* ;
2. Uniformly sample $\{t_k\}_{k=1}^K$ from time interval $[\delta, T]$ with step size $h = 0.02$ and total number of steps $K = \frac{T}{h}$;
3. For each t_k , sample $\{x_m^{t_k}\}_{m=1}^M$ (where $M = 1000$) from the distribution p_{t_k} as derived in (31);
4. Compute the empirical average $\widehat{|E_{\{y_i\}}|}^2$ (34) using $\{y_i\}$, $\{t_k\}$ and $\{x_m^{t_k}\}$.

The results (shown in Figure 2) align with the convergence rate $O(\frac{1}{N})$ as provided in Theorem 3.1.

In the second part of our experiments, we generate samples from DDPM using either the exact score function $u(t, x)$ or the empirical optimal score function $s_{\{y_i\}}^N(t, x)$. We discretize and simulate the SDEs (4) and (13) using Euler-maruyama method. The experiment parameters are: time interval length $T = 5$, discretization step $h = 0.0005$, early stopping times $\delta = 0$ or 0.01 , and number of training data $N = 100$. We generate 1000 new samples from DDPM with $u(t, x)$ and $s_{\{y_i\}}^N(t, x)$ respectively. Visualization results for $\delta = 0$ and $\delta = 0.01$ are shown in Figure 4 and Figure 5. The samples generated by DDPM with the empirical optimal score function $s_{\{y_i\}}^N(t, x)$ exhibit strong memorization effects, while those from DDPM with the exact score function $u(t, x)$ appear independent of the training data, yet maintain the same distribution. This numerical observation corroborates our theoretical findings in Theorem 4.3.

¹The implementation of KDE generation is built based on code <https://github.com/patrickphat/Generate-Handwritten-Digits-Kernel-Density-Estimation>;

The implementation of DDPM on CIFAR10 dataset follows the code <https://github.com/sail-sg/DiffMemorize/tree/main>.

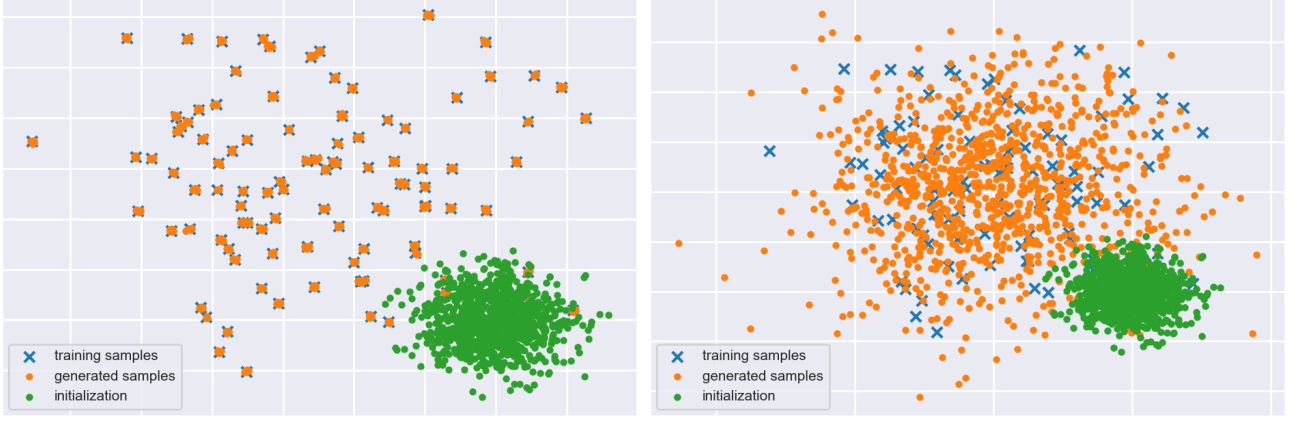


Figure 4. **Left:** Samples generated by DDPM with *empirical optimal score function* $s^N(t, x)$. **Right:** Samples generated by DDPM with *true score function* $u(t, x)$. Both two algorithms are ran up to time $T = 5$, i.e. early stopping time $\delta = 0$. The blue crosses are the training samples, the green dots are the initialization positions and the orange points are the generated samples.

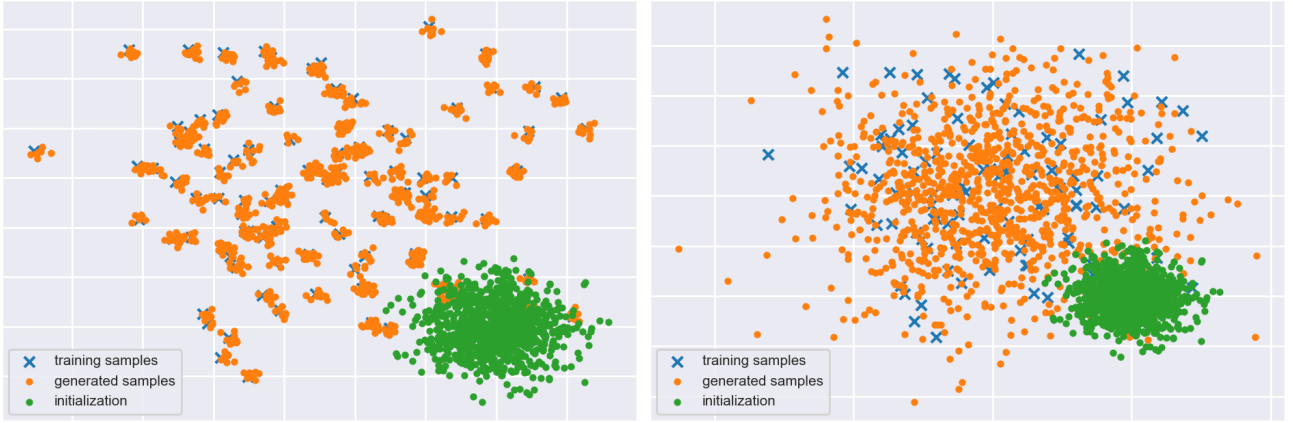


Figure 5. **Left:** Samples generated by DDPM with *empirical optimal score function* $s^N(t, x)$. **Right:** Samples generated by DDPM with *true score function* $u(t, x)$. Both two algorithms are early stopped with $\delta = 0.01$. The blue crosses are the training samples, the green dots are the initialization positions and the orange points are the generated samples.

E.2. Real-world data distribution

We consider p_* as the underlying distribution generating the CIFAR10 dataset images (Krizhevsky et al., 2009), comprising $N = 50000$ training samples of dimension $d = 32 \times 32 \times 3$. We denote $\{y_i\}_{i=1}^N$ as the 50000 images in the CIFAR10 dataset, and we use them to construct the following two generative models.

- The first one is simple Gaussian Kernel Density Estimation (KDE), i.e. $p_*^\gamma(x) = \frac{1}{N} \sum_{i=1}^N \mathcal{N}(x; y_i, \gamma^2 I_{d \times d})$, where γ is the Gaussian kernel’s bandwidth. To generate a sample, we first uniformly sample a data $y_{(j)}$ from $\{y_i\}_{i=1}^N$, then apply Gaussian blurring with bandwidth γ to $y_{(j)}$. The bandwidth γ is set to 0.1 times the optimal bandwidth $N^{-\frac{1}{d+4}} \sigma$, as per Scott’s rule (Terrell & Scott, 1992), where σ is the training data’s standard deviation. The sampling results are shown in the second row of Figure 1. Comparing with the training data (the first row in Figure 1), we can clearly see that the generated samples have strong dependence on existing ones.
- The second one is DDPM equipped with the empirical optimal score function as defined in (13). We follow the implementations in (Gu et al., 2023). To illustrate the details, we follow the notations used in (Gu et al., 2023). Recall the backward SDE (13)

$$d\hat{X}_t^\leftarrow = (\hat{X}_t^\leftarrow + 2s_{\{y_i\}}^N(T - t, \hat{X}_t^\leftarrow))dt + \sqrt{2}dB_t.$$

For sample generation, we discretize the time steps $0 = t_0 < t_1 < \dots < t_K = T$ with $T > 0$ being the time interval length and $K > 0$ being the total number of steps, and apply the Euler-maruyama solver. The update rule is as the following:

$$X_{t_n} = X_{t_{n-1}} + (t_n - t_{n-1}) \left(X_{t_{n-1}} + 2s_{\{y_i\}}^N(T - t_{n-1}, X_{t_{n-1}}) \right) + \sqrt{2(t_n - t_{n-1})}Z, \quad (35)$$

where $Z \sim \pi^d$. We terminate this update rule (35) at t_δ , where δ is the early stopping index². We set $T = 80$, $K = 18$, and vary δ . Figure 1’s third row shows the generated samples with $\delta = 5$. We can observe that the samples generated by DDPM equipped with the empirical optimal score function behave very similar to the samples generated by the Gaussian KDE (the second row in Figure 1). This aligns with our theoretical findings provided in Theorem 4.3. Additionally, Figure 6 displays samples $\delta = 3$ and $\delta = 5$, highlighting the strong memorization effect in DDPM with the empirical optimal score function, irrespective of the early stopping time.



Figure 6. Images generated by DDPM equipped with the empirical optimal score function based on CIFAR10 dataset. The first row is the original images from the CIFAR10 dataset. The second and third rows corresponding to the results of setting the early stopping index $\delta = 3$ and 5 respectively.

²Here we abuse the notation δ and refer it as the early stopping index. It is different from the δ we used in the main paper.