A PDE PERSPECTIVE ON GENERATIVE DIFFUSION MODELS

KANG LIU[1] AND ENRIQUE ZUAZUA[234]

Abstract. Score-based diffusion models have emerged as a powerful class of generative methods, achieving state-of-the-art performance across diverse domains. Despite their empirical success, the mathematical foundations of those models remain only partially understood, particularly regarding the stability and consistency of the underlying stochastic and partial differential equations governing their dynamics.

In this work, we develop a rigorous partial differential equation (PDE) framework for score-based diffusion processes. Building on the Li–Yau differential inequality for the heat flow, we prove well-posedness and derive sharp L^p-stability estimates for the associated score-based Fokker–Planck dynamics, providing a mathematically consistent description of their temporal evolution. Through entropy stability methods, we further show that the reverse-time dynamics of diffusion models concentrate on the data manifold for compactly supported data distributions and a broad class of initialization schemes, with a concentration rate of order $\sqrt{t}$ as $t \to 0$.

These results yield a theoretical guarantee that, under exact score guidance, diffusion trajectories return to the data manifold while preserving imitation fidelity. Our findings also provide practical insights for designing diffusion models, including principled criteria for score-function construction, loss formulation, and stopping-time selection. Altogether, this framework provides a quantitative understanding of the trade-off between generative capacity and imitation fidelity, bridging rigorous analysis and model design within a unified mathematical perspective.

1. Introduction

The development of generative models, one of the most dynamic areas of contemporary artificial intelligence (AI), aims to endow machines with the ability to create new, realistic samples that are statistically consistent with a given dataset drawn from an unknown distribution. In the context of image generation, this corresponds to producing novel images that belong to the same underlying class as those in the training set. More broadly, generative models seek to approximate and sample from complex, high-dimensional data distributions across diverse domains, including text, audio, video, molecular structures, physical fields, and even solutions of partial differential equations

[1]Université Bourgogne Europe, CNRS, Institut de Mathematiques de Bourgogne, 21000 Dijon, France.

[2]Friedrich - Alexander - Universität Erlangen - Nürnberg, Department of Mathematics, Chair for Dynamics, Control, Machine Learning, and Numerics (Alexander von Humboldt Professorship), 91058 Erlangen, Germany.

[3]Universidad Autónoma de Madrid, Departamento de Matemáticas, 28049 Madrid, Spain.

[4]Chair of Computational Mathematics, Fundación Deusto, 48007 Bilbao, Basque Country, Spain.

E-mail addresses: kang.liu@u-bourgogne.fr, enrique.zuazua@fau.de.

1991 *Mathematics Subject Classification.* 34D05, 35B35, 35Q68, 35Q84, 68T99.

Key words and phrases. Fokker–Planck equation, energy estimate, entropy method, asymptotic behavior, diffusion models, generative AI.

E. Zuazua was funded by the European Research Council (ERC) under the European Union's Horizon 2030 research and innovation programme (grant agreement NO: 101096251-CoDeFeL), the Alexander von Humboldt-Professorship program, the ModConFlex Marie Curie Action, HORIZON-MSCA-2021-dN-01, the COST Action MAT-DYNNET, the Transregio 154 Project of the DFG, AFOSR Proposal 24IOE027 and grants PID2020-112617GB-C22/AEI/10.13039/501100011033 and TED2021131390B-I00/ AEI/10.13039/501100011033 of MINECO (Spain), and Madrid GovernmentUAM Agreement for the Excellence of the University Research Staff in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

(PDEs). A central challenge lies in achieving a balance between *imitation fidelity* and *generative diversity*.

Formally, we consider a collection of samples drawn from an unknown data distribution $u_0 \in \mathcal{P}(\mathbb{R}^d)$, where d denotes the data dimensionality and $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures. For instance, $d = 256 \times 256$ for grayscale images, $d = 256 \times 256 \times 3$ for RGB images, and much higher for multimodal data such as audio signals, molecular coordinates, or discretized physical fields in scientific simulations. The learning objective is to construct a generator that produces new samples statistically consistent with u_0. Classical nonparametric density estimation methods, such as *kernel density estimation* [25], face severe challenges in high-dimensional settings due to the *curse of dimensionality* [32, Chp. 4], which renders them impractical for modern data domains.

To overcome these limitations, deep generative models have been developed over the past decade. Among them, *diffusion models* [15, 28] have emerged as a particularly powerful class. These models construct generators capable of sampling from u_0 by learning an approximation of its *score function* and modeling the reverse-time dynamics of an underlying diffusion process [1]. The analytical objective of this work is to rigorously characterize, within a PDE framework, the *well-posedness*, *stability*, and *concentration* properties of these generative dynamics.

More concretely, diffusion models consist of two complementary processes. The forward noising process is governed by a *Fokker–Planck* (FP) or heat equation, while the reverse-time generative process corresponds to its backward evolution. In this framework, the learned score function serves as a velocity field driving the backward FP equation, or equivalently, as the drift term in the corresponding reverse-time *stochastic differential equation* (SDE). The generation procedure thus begins with samples drawn from a Gaussian prior and evolves backward under the influence of the learned score function, gradually transforming random noise into structured data supported on the target manifold.

This intrinsic mathematical structure reveals a deep and previously unexplored connection between modern generative AI and classical analysis, and it motivates the PDE-based theoretical framework developed in this work.

Our main contribution is to establish a rigorous analytical framework for score-based diffusion models by leveraging the *Li–Yau* differential inequality [22] for the heat flow, an essential result in geometric analysis that provides sharp pointwise differential estimates for positive solutions of the heat equation on Riemannian manifolds. In our context, this inequality yields unilateral bounds on the divergence of the score field, ensuring the well-posedness and sharp L^p-stability of the backward FP dynamics that govern data generation. Building on this foundation, we employ entropy stability methods [31] to demonstrate that, when guided by the exact score, the diffusion model trajectories concentrate on the original data manifold for a wide class of initial generation distributions and every finite terminal time. These results provide a mathematically rigorous explanation for the generative consistency of diffusion models.

Beyond their theoretical significance, our findings offer practical guidelines for the design and training of diffusion-based architectures, including the formulation of loss functions, the choice of stopping time, and strategies to balance imitation fidelity with generative capacity. More broadly, this work opens a dialogue between PDE theory and machine learning, demonstrating how tools from classical analysis can illuminate and guide the design of modern AI systems. The proposed framework thus not only advances the mathematical understanding of diffusion models but also outlines a broader analytical foundation for generative AI.

1.1. **Organization.**

Section 2 provides the preliminaries of diffusion models and summarizes the key ideas underlying our contributions. The main stability and concentration results for the backward FP dynamics, together with comparisons to related works, are presented in Section 3. In Section 4, we discuss the interplay between imitation fidelity and generative capacity. Section 5

contains the detailed proofs of the main results stated in Section 3. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Preliminaries on score-based diffusion models and main results

2.1. **Overview of diffusion models and main results.** Given a time horizon $0 < T < \infty$ and the ambient space $\mathbb{R}^d$, we first introduce the *forward heat equation*:

$$\begin{cases} \partial_t u(x,t) - \Delta u(x,t) = 0, & (x,t) \in Q, \\ u(\cdot,0) = u_0 \in \mathcal{P}(\mathbb{R}^d). \end{cases} \tag{2.1}$$

Here, Q denotes the space-time cylinder

$$Q = \mathbb{R}^d \times (0,T], \qquad \mathring{Q} = \mathbb{R}^d \times (0,T),$$

and $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures in $\mathbb{R}^d$.

The solution of this forward heat equation is given by the convolution of the initial datum with the heat kernel:

$$u(x,t) = (G_t * u_0)(x), \qquad (x,t) \in Q, \tag{2.2}$$

where for $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$,

$$G_t(x) = (4\pi t)^{-d/2} \exp\Big(-\frac{\|x\|^2}{4t}\Big).$$

In the context of diffusion models, u_0 represents the (unknown) density of probability of the data distribution.

Obviously, since $u_0 \in \mathcal{P}(\mathbb{R}^d)$ (and is therefore positive), it follows that $u > 0$ everywhere for any $t > 0$. This allows us to introduce the *score function* associated with the heat flow (2.1):

$$s(x,t) = \nabla \log u(x,t) = \frac{\nabla u(x,t)}{u(x,t)}, \qquad (x,t) \in Q. \tag{2.3}$$

Since u is smooth and strictly positive for $t > 0$, the score is well-defined and smooth $s \in \mathcal{C}^\infty(Q;\mathbb{R}^d)$.

This score function allows us to redefine the *backward heat equation* in a way that its intrinsic instability can be better understood, a fact that will be employed to analyze the generative capacity of diffusion models.

For this purpose, it is sufficient to observe that the heat equation in (2.1), when one of its solutions $u = u(x,t)$ is given, can be equivalently rewritten as

$$\partial_t u + \epsilon\, \Delta u - (1+\epsilon)\, \mathrm{div}\big(s\, u\big) = 0, \tag{2.4}$$

with the score velocity field $s = s(x,t)$ as in (2.3), and $\epsilon \geq 0$.

Note that in this new formulation, the sense of the diffusion has been reversed ($\epsilon > 0$) or it was simply suppressed $\epsilon = 0$. Therefore, (2.4) is expected to be a well-posed FP or convection-diffusion model (when $\epsilon > 0$) or a hyperbolic transport model (when $\epsilon = 0$) in the backward sense of time. As mentioned above, the score velocity field $s(x,t)$ is smooth. However, the actual dynamic of the backward model (2.4) depends sensitively on the available bounds on its divergence.

The first main contribution of this paper is to show that the necessary bounds follow from the classical Li–Yau inequality [22] for positive solutions of the heat equation, ensuring that

$$\Delta \log(u) \;\geq\; -\frac{d}{2t}, \qquad (x,t) \in Q, \tag{2.5}$$

which, in terms of the score field, can be rewritten as :

$$\mathrm{div}\, s \;\geq\; -\frac{d}{2t}, \qquad (x,t) \in Q. \tag{2.6}$$

Our second contribution, building on Li–Yau's inequality, is the derivation of sharp L^p estimates for the backward flow (2.4), as stated in Theorem 3.1. Moreover, we establish an entropy stability analysis based on the Kullback–Leibler (KL) divergence between solutions with different terminal conditions.

As we shall see, these estimates play a fundamental role in understanding the generative process of diffusion models. In such models, new data samples are obtained by solving the backward SDE associated with the learned score function with (2.4):

$$\begin{cases} dX_t = -(1+\epsilon)\, s(X_t, t)\, dt + \sqrt{2\epsilon}\, dW_t, \quad t \in (0, T], \\ X_T \sim v_T, \end{cases} \tag{2.7}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion, and v_T is a prescribed probability measure, typically Gaussian. The generated sample is obtained from the trace at $t = 0$ of the solution to (2.7), which is computed using a suitable numerical integration scheme.

Our analysis establishes sharp bounds on the obtained traces. It elucidates their connection to the data manifold encoded by the initial probability density u_0, thereby allowing a rigorous quantification of the imitation and generation capacities of diffusion models.

In particular, the imitation capacity is quantified through the concentration of the solutions of the SDE (2.7) as $t \to 0$. We prove that the probability of the generated flow X_t approaching the data manifold $\mathrm{supp}(u_0)$ is equal to 1, see Theorem 3.4. In the deterministic ODE case (i.e., $\epsilon = 0$ in (2.7)), when the data measure u_0 is a finite sum of Dirac masses, we further quantify the convergence rate to be proportional to $\sqrt{t}$, see Theorem 3.11.

The main tool we employ to prove this concentration is an entropy stability analysis for the FP equation, inspired by the hypocoercivity framework for kinetic equations [31, Chap. 1.6]. The derivation of explicit convergence rates further relies on a detailed analysis of the non-autonomous and singular gradient flow system (2.7).

In this article, we focus on the heat equation, without including an additional transport term (such as $-\operatorname{div}(x\, u(x,t))$) that appears in the forward dynamics of certain diffusion models [15, 27]. However, our analysis extends naturally to that setting, since such dynamics are equivalent to the heat equation under a self-similar change of variables, see [2, 34] and Section 3.4.

2.2. **Novelty of the present work.** The main contributions of this work are summarized as follows:

- *PDE-based theoretical foundation.* Our work provides a rigorous PDE framework for score-based diffusion models. Using the Li–Yau estimate for the heat flow, we establish a well-posed and L^p-stable formulation of the score-based FP equation, valid for any strictly positive time.
- *Entropy-based concentration.* Through an entropy stability analysis, we prove that the reverse-time solution concentrates on the original data manifold under weak assumptions (in particular, compactly supported data distributions), thus ensuring convergence for any finite terminal horizon.
- *Empirical measures and explicit rates.* In the deterministic and empirical setting, we derive an explicit $\sqrt{t}$ convergence rate of the generated flow toward the true data samples, which is particularly sharp for empirical measures.
- *Learned versus empirical score analysis.* We extend our concentration estimates to the empirical score, showing that the resulting diffusion model exhibits a high imitation capacity but limited generative ability. We then quantify the deviation between the samples generated through the empirical score and those obtained via the learned score function s_θ, offering quantitative insight into the model's generative capability.

- *Training guidelines and hyperparameter tuning.* Our stability results provide practical guidance for the training process, clarifying how the score field should be adjusted and how hyperparameters (such as the time horizon and viscosity coefficient ϵ) should be tuned to achieve an optimal balance between imitation and generation.

3. Stability and concentration

In this section, we present the main results we achieve in terms of the stability of the diffusion model and its concentration/imitation capacity.

We consider the *backward Fokker–Planck equation* associated with the generative process (2.7):

$$\begin{cases} \partial_t v(x,t) + \epsilon\,\Delta v(x,t) - (1+\epsilon)\,\mathrm{div}\big(s(x,t)\,v(x,t)\big) = 0, & (x,t) \in \mathring{Q}, \\ v(x,T) = v_T(x), & x \in \mathbb{R}^d, \end{cases} \tag{3.1}$$

where $\epsilon \geq 0$ is a hyperparameter, $v_T \in \mathcal{P}(\mathbb{R}^d)$, and $s = s(x,t)$ denotes the score function (2.3) associated with the solution $u = u(x,t)$ of the heat equation (2.1), with initial datum u_0, the probability density of the unknown distribution under consideration.

Note that in practical applications, v_T is often taken to be a Gaussian. In particular, $v_T \geq 0$ and $\int_{\mathbb{R}^d} v_T(x)dx = 1$. Its role is to lead, by sampling, to the initialization at time $t = T$ of the backward SDE dynamics (2.7) aimed at dynamically generating new samples at $t = 0$. So, we emphasize that v_T is not related to the solution $u(T)$ of the forward heat equation (2.1) starting from the unknown distribution u_0 or its empirical approximation. However, according to the classical results on the asymptotic behavior of the (2.1), we know that as $T \to \infty$, $u(T)$ in the forward heat equation approximates G_T, see [34].

3.1. L^p stability.

As noted above, the Li–Yau inequality [22] implies that the score vector field $s = \nabla \log u$ in (2.3) associated with any nonnegative initial condition $u_0 \geq 0$ satisfies (2.6). Combining this fact with an energy estimate for (3.1) yields the following stability bound.

Theorem 3.1 (Energy estimate of the score-based FP equation)**.** *Let v be the solution of (3.1). Then,*

$$\|v(t)\|_{L^p} \;\leq\; \left(\frac{T}{t}\right)^{\frac{d(1+\epsilon)(p-1)}{2p}} \|v_T\|_{L^p}, \qquad \forall\, t \in (0,T],\;\; p \in [1,\infty), \tag{3.2}$$

for any fixed $\epsilon \geq 0$.

Remark 3.2 (Backward well-posedness)**.** The backward heat equation is a prototypical ill-posed problem, as it lacks continuous dependence on the terminal condition. In fact, applying the Fourier transform to the backward heat equation shows that the solution exhibits extremely rapid growth in the high-frequency regime [10]. Nevertheless, by the Li–Yau estimate and Theorem 3.1, the backward score-based FP equation (3.1), although originated as a reinterpretation of the heat equation, is well-posed for any strictly positive initial time $t > 0$. In particular, the solution at time $t > 0$ depends continuously on the terminal condition in the L^p sense and blows up when t approaches $t \sim 0$ only.

Remark 3.3 (Sharpness of the estimate)**.** We first note that the blow-up rate of the order of $1/t$ in the Li–Yau estimate is optimal. This can be readily verified, for instance, when the initial data u_0 is a finite sum of Dirac measures, so that u is the corresponding solution of the heat equation, a linear combination of finitely many Gaussian kernels (see Figure 1). The estimate in (3.2) is also sharp in this setting for $\epsilon = 0$. Indeed, for the terminal distribution $v_T = u(T)$, the uniqueness of solutions to the backward FP equation implies that $v \equiv u$. As $t \to 0$, the solution $v(t)$ thus converges to u_0, a finite linear combination of Dirac measures. Consequently, the blow-up in (3.2) manifests in any L^p-norm with $p > 1$, following precisely the rate given in (3.2) for $\epsilon = 0$.

Note that the blow-up estimate in (3.2) deteriorates as ϵ increases, even though the diffusion term regularizes the flow in the backward sense of time. Indeed, when ϵ increases, the impact of the singularity of the score function at $t \sim 0$ is enhanced. Thus, the L^p estimate near $t \sim 0$ deteriorates as well. This fact is related to the enhanced generative power of the diffusion model as the viscosity parameter ϵ increases.

For $p = 1$, the positivity and total mass of v_T are preserved; hence, the L^1-norm remains constant, excluding any blow-up as $t \to 0$. However, the potential convergence of $v(t)$ does not occur in the L^1 sense, but rather in the Wasserstein sense.

From a numerical standpoint, however, the blow-up described in (3.2) is naturally mitigated by time discretization, which regularizes the singular behavior near $t = 0$.

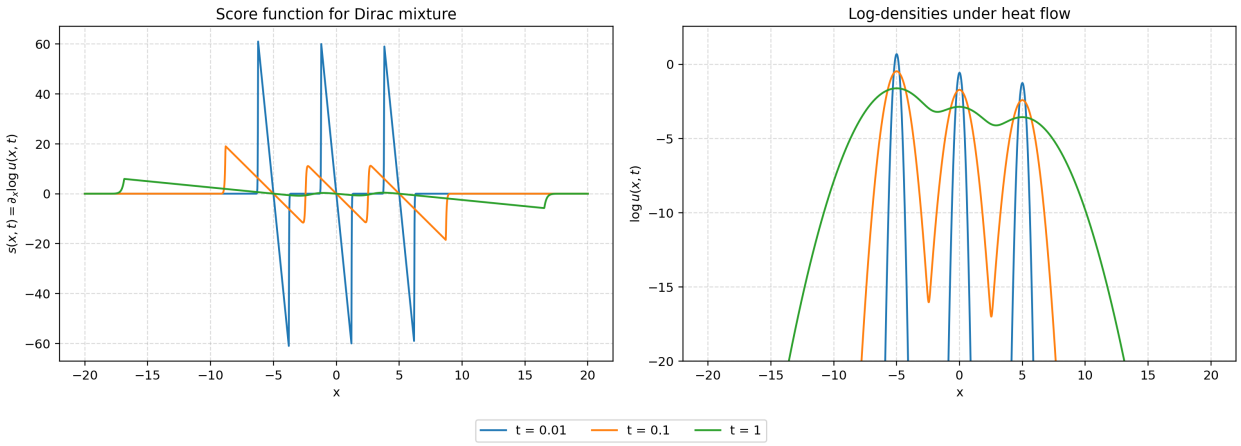


FIGURE 1. Score function (left) and log-density (right) of the heat flow originating from the initial distribution $u_0 = 0.7\,\delta_{-5} + 0.3\,\delta_0 + 0.1\,\delta_5$. The singular behavior predicted by the right-hand side of the Li–Yau estimate as $t \to 0$ is evident from the steep slopes of the score function near the Dirac locations of the initial data. The right panel also shows that the local maxima of the log-densities occur at these Dirac positions as $t \to 0$.

3.2. **Imitation capacity of the diffusion model.** The solution of (3.1) yields the probability density of the solutions of the SDE (2.7) and, in particular, the probability density of the generated samples at $t = 0$.

To analyze the imitation capacity of the diffusion model, it is important to analyze whether the solution of (2.7) converges to the data manifold, i.e. the support of the initial distribution u_0, denoted by $\mathrm{supp}(u_0)$. We address this critical issue by performing an entropy stability analysis of the FP equation, drawing inspiration from hypocoercivity methods [31, Chp. 1.6].

Recall that the KL divergence (or relative entropy) between two probability measures $m_1, m_2 \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\mathrm{KL}(m_1 \,\|\, m_2) \;=\; \begin{cases} \displaystyle\int_{\mathbb{R}^d} \log\left(\frac{m_1(x)}{m_2(x)}\right) dm_1(x), & \text{if } m_1 \ll m_2, \\ +\infty, & \text{otherwise,} \end{cases} \tag{3.3}$$

where $m_1 \ll m_2$ means that m_1 is absolutely continuous with respect to m_2.

As we shall see in Lemma 5.3, the KL divergence between v (the solution of (3.1)) and u (the reference solution of the heat equation defining the score function) decays when viewed backward

in time:

$$\mathrm{KL}\big(v(t_1)\,\|\,u(t_1)\big) \;\le\; \mathrm{KL}\big(v(t_2)\,\|\,u(t_2)\big), \qquad \forall\, 0 < t_1 \le t_2 \le T. \tag{3.4}$$

This estimate is stronger than the classical L^1-contraction property, since the L^1-distance can be bounded in terms of the KL divergence (see Pinsker's inequality [30, Rem. 22.12]), while the converse implication does not hold.

Based on the previous uniform bound on the KL divergence, we can establish the concentration of the support of $v(t)$ towards that of u_0, under the following assumption:

Assumption A. Recall that u_0 is the initial distribution of the heat equation (2.1) and v_T is the initial distribution of the backward generative process (2.7). Denote by $\mathcal{L}^d$ the Lebesgue measure on $\mathbb{R}^d$. Assume that:

(1) The initial distribution u_0 has compact support.
(2) The measure v_T is absolutely continuous with respect to $\mathcal{L}^d$, with density $v_T > 0$ almost everywhere, and satisfies

$$v_T \log v_T \in L^1(\mathbb{R}^d), \qquad |x|^2 v_T \in L^1(\mathbb{R}^d).$$

Theorem 3.4 (Concentration of support)**.** *Let Assumption A hold. Let $v(t)$ and X_t be the solutions of the FP equation (3.1) and the corresponding SDE (2.7), respectively. Then, the following holds:*

(i) *For any sequence $t_n \to 0^+$, the family $\{v(t_n)\}_{n\ge1}$ is precompact in the weak-* topology on $\mathcal{P}(\mathbb{R}^d)$ and any weak-* limit point v^* satisfies*

$$\mathrm{supp}(v^*) \subset \mathrm{supp}(u_0). \tag{3.5}$$

Furthermore, this inclusion becomes an equality when $|\log v_T(x)|$ grows polynomially at infinity.

(ii) *For any open set $U \supset \mathrm{supp}(u_0)$,*

$$\lim_{t\to0^+} v(t)(U) = 1, \quad \lim_{t\to0^+} \mathbb{P}(X_t \in U) = 1, \tag{3.6}$$

where $\mathbb{P}$ denotes the probability law induced by the random variable X_t.

Remark 3.5 (Gradient flows)**.** Given that the score function (2.3), which defines the deterministic ($\epsilon = 0$) or stochastic ($\epsilon > 0$) dynamics in (2.7), is a pure gradient field associated with the potential $\log(u)$, the corresponding particle dynamics naturally drive trajectories to concentrate (backward in time) around the local maximizers of this potential, that is, precisely on the support of u_0 (as illustrated in the right panel of Figure 1). This is exactly what the theorem above establishes.

Remark 3.6 (The inviscid case)**.** As we shall see below in the next subsection, in the deterministic case ($\epsilon = 0$), one can quantify the convergence rate.

Remark 3.7 (Concentration of support)**.** This result shows that the solution of (3.1) becomes progressively concentrated on the support of u_0 as $t \to 0^+$. This concentration manifests both at the level of individual trajectories of the backward generative SDE and in the density function governed by the backward FP equation.

This behavior highlights the imitation capacity of the diffusion model, namely, its ability to generate realizations that tend to remain within the support of the initial distribution u_0.

Remark 3.8 (Numerical experiment in Figure 2)**.** A visualization of this concentration phenomenon is presented in Figure 2, which illustrates the dynamics in two spatial dimensions. We consider a true data distribution supported on a one-dimensional curve, represented by the lemniscate.

Panels (A)-(B) show trajectories of the generation flow (2.7) driven by the exact score drift corresponding to this full distribution. In panel (A), we observe how the trajectories of the generative SDE return to the support of u_0. In panel (B), we display the locations of the trajectories when the dynamics are stopped prematurely, before reaching $t = 0$. As expected, the distance to the support

of u_0 increases with t, clearly illustrating how early stopping enhances the generation capacity of the model.

Panels (C)-(D) depict analogous trajectories when the generation flow (2.7) is driven by the empirical score drift, obtained from an approximation of the full distribution u_0 by the average of several Dirac realizations. In panel (C), the trajectories of the generative SDE return to the support of this empirical approximation, while panel (D) shows the state of the trajectories under early stopping. Again, we observe that the distance to the support of the initial empirical measure increases with t.

The contrast between the two experiments, (A)-(B) and (C)-(D), is striking. In the first case, the entire support of u_0 acts as an attractor of the diffusion model, faithfully capturing the geometry of the underlying distribution. In the second case, by contrast, only those realizations of u_0 used to estimate the empirical score serve as effective attractors. This highlights the intrinsic dependence of the generation process on the quality and completeness of the score approximation.

Remark 3.9 (Empirical score function)**.** If the initial distribution is an empirical measure supported on the finite set $\{y_1, \dots, y_N\}$,

$$u_0 \;=\; \frac{1}{N}\sum_{k=1}^{N} \delta_{y_k},$$

then the exact (empirical) score admits the explicit formula

$$s(x,t) = \frac{1}{2t}\left(\frac{\sum_{k=1}^{N} e^{-\frac{\|x-y_k\|^2}{4t}}\,(y_k - x)}{\sum_{k=1}^{N} e^{-\frac{\|x-y_k\|^2}{4t}}}\right) = \frac{1}{2t}\left(\frac{\sum_{k=1}^{N} e^{-\frac{\|x-y_k\|^2}{4t}}\,y_k}{\sum_{k=1}^{N} e^{-\frac{\|x-y_k\|^2}{4t}}} - x\right), \tag{3.7}$$

for all $(x,t) \in \mathbb{R}^d \times (0,T]$.

By Theorem 3.4, the trajectories of the backward stochastic generation SDE (2.7) concentrate on the finite set $\{y_1, \dots, y_N\}$ as $t \to 0^+$; in particular, there is no genuine "novel" generation in this setting. This corresponds to pure imitation.

This result helps to clarify the central role of the score function in determining the performance of diffusion models. It also offers insight into how the score can be strategically manipulated for design purposes, since a suitable modification of the score function can deliberately perturb the samples generated by the model.

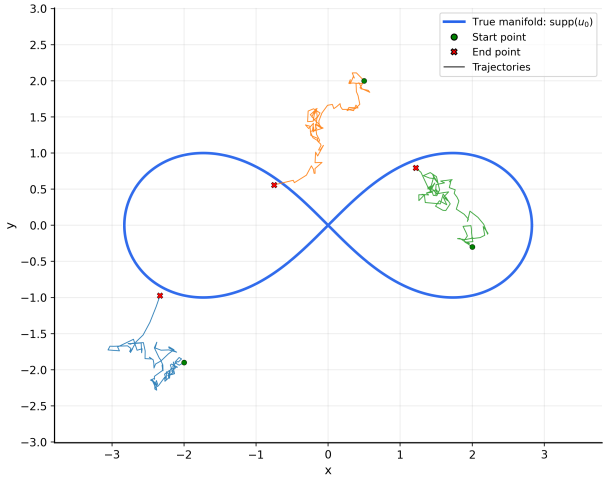
In particular, different samplings of the original distribution correspond to different score functions, and therefore to different attractors in the diffusion process. This can be viewed as a limitation of the standard diffusion model, which tends to reproduce samples of the original distribution unless an early-stopping procedure is applied, as discussed above.

This observation suggests the need to adjust or redesign the score function, for instance, through neural network-based approximations that reduce its complexity and guide the diffusion dynamics toward alternative attractors, roughly corresponding to the critical points of the surrogate score function. In this way, one can mitigate the model's natural tendency toward imitation and enhance its generative capability. More details are presented in Section 4.

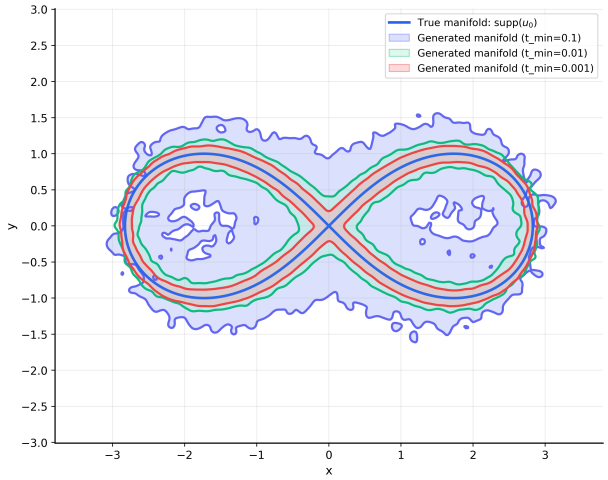
Remark 3.10 (Necessity of structural conditions on v_T)**.** The conditions imposed on v_T in Assumption A(2) are rather mild. Indeed, any Gaussian distribution satisfies these requirements, which are standard in the implementation of diffusion models. More precisely, Assumption A(2) requires that v_T admits a density that is strictly positive almost everywhere, and that $v_T \log v_T$ and $\|x\|^2 v_T$ belong to $L^1(\mathbb{R}^d)$. These assumptions ensure the key property, established in Lemma 5.2,

$$\mathrm{KL}(v_T \,\|\, u(T)) < \infty,$$

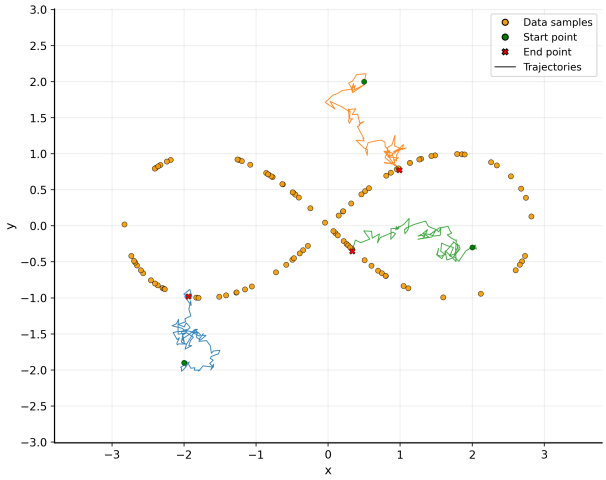
which, together with the contraction property of the KL divergence, yields a uniform bound on the divergence for all $t \in (0,T]$.



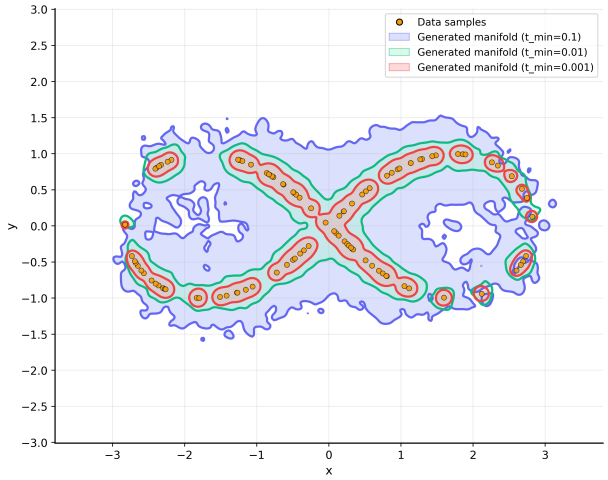
(A) Three backward stochastic trajectories with the true score function.



(B) Early-stopping data manifolds with the true score function.



(C) Three backward stochastic trajectories with the empirical score function of data samples.



(D) Early-stopping data manifolds with the empirical score function of data samples.

FIGURE 2. Score-based generation on the lemniscate dataset: comparison between the true ((A)-(B)) and empirical scores ((C)-(D)), and the effect of the choice of the stopping time $t_{\min} \in \{0.1, 0.01, 0.001\}$. The true score ((A)-(B)) corresponds to the solution of the heat equation, given by the convolution with the Gaussian heat kernel, computed by means of numerical quadrature, while the empirical score ((C)-(D)) is obtained from a finite sample of data points via the explicit expression (3.7). In all experiments, we set the backward diffusion parameter to be $\epsilon = 0.2$. The generated manifold is obtained as the region encompassing 95% of the 10,000 reverse-time samples, reducing the effect of extreme stochastic trajectories.

The necessity of these conditions can be seen from counterexamples where the generative flow fails to recover the data manifold or only recovers part of it. Let

$$u_0 = \tfrac{1}{2}\big(\delta_{(-1,0)} + \delta_{(1,0)}\big),$$

and consider the deterministic flow case ($\epsilon = 0$):

(1) *Case 1.* Suppose that $\mathrm{supp}(v_T) \subset \{(0,y) : y \in \mathbb{R}\}$. Then v_T does not possess a density in $\mathbb{R}^2$. From (3.7), one verifies that

$$-s((0,y),t) = (0, -y/(2t)), \qquad \forall\, y \in \mathbb{R}.$$

Consequently, the generative flow collapses every point in $\mathrm{supp}(v_T)$ to the single point $(0,0)$, which lies outside $\mathrm{supp}(u_0)$, see the middle panel of Figure 3.

(2) *Case 2.* Assume that $v_T \ll \mathcal{L}^d$ and that $\mathrm{supp}(v_T) \subset \mathbb{R}_+ \times \mathbb{R}$, where we lose the strict positivity of v_T. According to the formula of the score function (3.7), for points in the right half-plane, the attractive influence originating from $(1,0)$ dominates that from $(-1,0)$. As a result, the generative flow drives all particles within the support of v_T toward the right-hand source at $(1,0)$, thereby preventing a full reconstruction of the data manifold, as illustrated in the right panel of Figure 3.

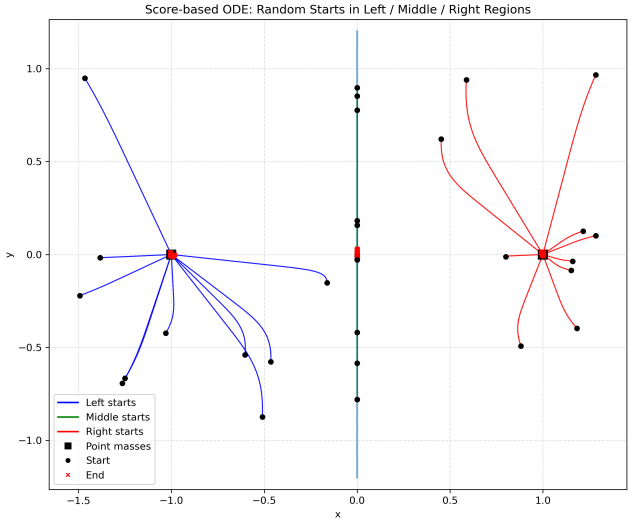


FIGURE 3. Trajectories of the score-based ODE ($\epsilon = 0$) with $u_0 = \frac{1}{2}\big(\delta_{(-1,0)} + \delta_{(1,0)}\big)$. Initial points with $x < 0$ converge to $(-1,0)$, those with $x > 0$ converge to $(1,0)$, and points with $x = 0$ converge to the barycenter $(0,0)$.

3.3. **Imitation rate in ODE setting.** Theorem 3.4 provides qualitative concentration results under fairly weak assumptions. Here, we study the *quantitative* concentration rate for the deterministic case of the generative flow (2.7), i.e., $\epsilon = 0$. The resulting ODE is

$$\begin{cases} \dfrac{dX_t}{dt} = -s(X_t,\, t), \quad t \in (0,T), \\ X_T = x_T, \end{cases} \tag{3.8}$$

where s is the score function (2.3) and x_T is any point in $\mathbb{R}^d$.

Theorem 3.11 (Imitation rate)**.** *Assume that u_0 has compact support, and let X_t denote the solution of (3.8) with initial condition $X_T = x_T \in \mathbb{R}^d$. Then the following statements hold:*

(i) *Let $S = \mathrm{supp}(u_0)$. Then, for Lebesgue-almost every $x_T \in \mathbb{R}^d$,*

$$\liminf_{t \to 0^+} d\big(X_t, S\big) = 0. \tag{3.9}$$

(ii) *Let $K = \mathrm{conv}(S)$. Then, for any $x_T \in \mathbb{R}^d$,*

$$d\big(X_t, K\big) \;\le\; d\big(x_T, K\big)\, T^{-1/2}\, \sqrt{t}, \qquad t \in (0,T]. \tag{3.10}$$

(iii) *Suppose that u_0 is a finite sum of Dirac measures:*

$$u_0 = \sum_{k=1}^{N} w_k \, \delta_{y_k},$$

with weights $w_k > 0$ and locations $y_k \in \mathbb{R}^d$. Then, for Lebesgue-almost every $x_T \in \mathbb{R}^d$, there exists an index $i \in \{1, \dots, N\}$, depending on x_T, such that

$$\lim_{t \to 0^+} X_t = y_i.$$

Moreover, there exist a constant $C > 0$, depending only on x_T and the time horizon T, such that

$$\|X_t - y_i\| \;\le\; C\sqrt{t}, \qquad \forall\, t \in (0, T]. \tag{3.11}$$

Remark 3.12. We provide several remarks on Theorem 3.11.

- Point (i) establishes a qualitative convergence result for the lower limit of $d(X_t, S)$, for almost every x_T. The result may be expected to hold as a full limit, as a consequence of the a priori estimates of the score function from Lemma 5.1. But this is an open problem.
- Point (ii) presents a quantitative convergence result toward the convex hull of the support of u_0, which holds for any x_T. And this result is sharp, in some sense.

  Indeed, as illustrated in Figure 3, when x_T lies on the y-axis, the trajectories converge to $(0, 0)$, a point contained in the convex hull of $(-1, 0)$ and $(1, 0)$, but not in the initial support of the empirical measure.

  This result does not contradict Point (i), since $S \subset K$. Thus, once a trajectory enters K (making the distance to K vanish) at some $t > 0$, it may continue to evolve toward certain points in S.
- Finally, Point (iii) provides a quantitative convergence result to S, but only in the empirical setting. Extending point (iii) to the non-empirical case would require additional geometric assumptions on S, which we leave for future work.

Remark 3.13 (Concentration rate in the stochastic case)**.** The convergence rate established in Theorem 3.11 addresses the deterministic generation flow associated with the empirical score function ($\epsilon = 0$). A natural next step is to extend this analysis to the stochastic generation flow with $\epsilon > 0$. Near $t = 0$, the diffusion term becomes negligible compared to the drift term, which is of order $\mathcal{O}(1/t)$ near $t = 0$; hence, the stochastic dynamics can be interpreted as a small random perturbation of the deterministic system. In this regime, one would expect large deviation principles (see, e.g., [12]) to provide a quantitative description of the probabilistic concentration rate.

For instance, for the stochastic generation process (2.7), where $\sqrt{2\epsilon} > 0$ is the standard deviation of the Brownian motion, one would anticipate estimates of the form

$$\mathbb{P}\big(\|X_t - y_i\| \le C\sqrt{t}\big) \;\ge\; 1 - \varepsilon(\epsilon, t),$$

where $\varepsilon(\epsilon, t)$ represents the probability of large deviations, vanishing as $\epsilon \to 0$ or $t \to 0$.

This direction remains largely unexplored and clearly warrants further investigation.

3.4. **Related work and our contribution.** We begin by briefly reviewing the state of the art in deep learning–based generative models, and then compare our results with existing theoretical analyses of diffusion models.

3.4.1. *State of the Art.* The earliest successful deep generative models include the *variational autoencoder* (VAE) [18] and the *generative adversarial network* (GAN) [13]. The VAE introduces a probabilistic latent-variable framework that learns an approximate likelihood of the data distribution via variational inference, whereas the GAN formulates data synthesis as a two-player minimax game between a generator and a discriminator. While both frameworks have achieved empirical

success, VAEs typically yield blurred reconstructions due to their reliance on tractable variational bounds, and GANs often exhibit instability and mode collapse, capturing only a subset of the data manifold.

In recent years, diffusion-based generative models have emerged as a mathematically principled and empirically effective alternative to earlier approaches, exhibiting remarkable performance in practical applications. Among these, a widely studied class is that of score-based models [27, 28], which constitutes the main focus of our work. Another prominent framework is the *denoising diffusion probabilistic model* (DDPM) [15], which has been shown to be mathematically equivalent to the score-based formulation, as discussed in Section 4.

3.4.2. *Comparison and Novelty.* Theoretical analyzes of score-based diffusion models have been developed in, for example, [21, 7, 3, 8] and the references therein. We compare our results with these works.

In [21, 7, 3, 8], the authors prove the imitation property for diffusion models governed by an *Ornstein–Uhlenbeck* FP equation, where the heat equation is perturbed by a confining transport term:

$$\begin{cases} \partial_t w - \Delta w - \operatorname{div}(xw) = 0, & (x,t) \in Q, \\ w(\cdot,0) = w_0. \end{cases} \tag{3.12}$$

It is important to note that, applying the standard self-similar change of variables [34],

$$u(x,t) = (2t+1)^{-d/2}\, w\left(\frac{x}{\sqrt{2t+1}}, \frac{\log(2t+1)}{2}\right), \quad (x,t) \in Q,$$

u satisfies the heat equation (2.1) with the initial condition $u_0 = w_0$.

Therefore, our results can be easily adapted to that setting as well, providing some interesting generalizations. In particular, in [21, 7, 3, 8], the corresponding backward FP dynamics is considered with $\varepsilon = 1$ and

$$v^T(T) = \mathcal{N}(0, I_d),$$

since the standard Gaussian distribution is the *stationary* solution of (3.12).

Our analysis yields more general imitation results, valid for any $\varepsilon \geq 0$ and for a broad class of terminal distributions $v^T(T)$ satisfying Assumption A(2); in particular, all Gaussian measures.

The works [7, 3, 8] provide upper bounds for $\mathrm{KL}(v^T(t)\|w(t))$ in terms of $\mathrm{KL}(v^T(T)\|w(T))$ and the mismatch between the score functions, under various structural assumptions. In the case of an exact score (no mismatch), these results yield the contraction of the KL divergence, consistent with our Lemma 5.3 and Remark 5.4. The work of [21] focuses instead on the χ^2-divergence, leading to analogous conclusions.

Despite this shared contraction property, the interpretation and form of the imitation results differ substantially between the existing literature and our work. In [7, 3, 8], imitation is understood in a strict sense; namely, $v^T(0) = w_0$. This holds in the asymptotic regime $T \to \infty$, which ensures that $\mathrm{KL}(\mathcal{N}(0, I_d)\|w(T)) \to 0$. The main assumptions and conclusions of these references can be summarized as follows:

(1) *Exact imitation at $t = 0$.* The work of [7] assumes that the score function in (3.12) has a globally Lipschitz gradient, while [8] assumes that the initial distribution w_0 (which coincides with u_0 in our notation) has finite *relative Fisher information* with respect to $\mathcal{N}(0, I_d)$. In both cases,

$$\lim_{T\to\infty} \mathrm{KL}(v^T(0)\|w_0) = 0.$$

Under these assumptions, there is no blow-up of the score function near $t = 0$, and convergence holds even at $t = 0$.

(2) *Early stopping.* In [3], the only assumption is that w_0 has a finite second moment. Under this weaker condition,

$$\lim_{T\to\infty} \mathrm{KL}(v^T(t)\|w(t)) = 0, \quad \forall\, t > 0.$$

A similar result is also proved in the second part of [8]. This asymptotic equality holds for any fixed positive time, corresponding to the "early stopping" strategy in diffusion models. However, this equality cannot, in general, be extended to $t = 0$ without stronger regularity assumptions on w_0, due to the blow-up behavior of $v^T(t)$ as $t \to 0^+$.

(3) *Finite-sample case.* When w_0 is a finite sum of Diracs, it automatically falls into the previous category. Building on this, [23] leverages the result of [3] to prove the convergence of the generated distribution toward the empirical measure without early stopping, i.e., at $t = 0$. Again, this convergence is understood in the limit $T \to \infty$. Empirical evidence of this imitative behavior in the finite-sample case is reported in [6, 26].

In contrast, our results reveal a fundamentally different imitation property. Rather than aiming to exactly recover the initial distribution w_0, we focus on the concentration behavior of $v(t)$ as $t \to 0^+$, which captures its support convergence toward the data manifold. This concentration holds for any $T > 0$ and any v_T satisfying mild assumptions. Our results imply that:

(1) *Support imitation at $t = 0$ for any finite T.* We assume that w_0 has compact support. This is a more realistic assumption than in [21, 7, 23], since, for instance, for image data, each pixel's RGB value lies within a bounded cube. Moreover, $v^T(T)$ can be any probability measure satisfying Assumption A(2), which is much weaker than the Gaussian terminal condition used in previous works. In Theorem 3.4, we prove that for any $T > 0$,

$$\lim_{t\to 0^+} \mathbb{P}(X_t^T \in U) = 1,$$

where X_t^T is a trajectory associated with $v^T(t)$, and U is any open neighborhood of $\mathrm{supp}(w_0)$. In our setting, the terminal time T does not need to tend to infinity; it can be any positive finite horizon. This shows that the blow-up of the score function indeed drives the flow toward the data manifold at an increasingly fast rate as $t \to 0^+$.

(2) *Finite-sample case.* For the finite-sample setting, Theorem 3.11 provides an explicit concentration rate of order $\sqrt{t}$ for the above convergence, which reproduces the qualitative results of [23] and provides theoretical support for the findings of [14, 33].

4. From imitation to generation

As we have seen, in diffusion models, stability guaranties that the backward denoising dynamics can reliably invert the forward diffusion process and recover the original data manifold. However, enforcing strict stability often limits the expressive capacity of the learned dynamics, thereby constraining the model's generative diversity. An effective diffusion model must, therefore, strike a balance between accurate reconstruction of the training data and flexible generation of novel, realistic samples.

In what follows, we discuss several strategies to mitigate over-imitation during the training of diffusion models and enhance their generative capability.

4.1. **Implicit regularization by neural networks.**

We now turn to the training procedure of diffusion models and highlight how neural networks implicitly regularize the learned score function.

Assume we are given samples $\{y_i\}_{i=1}^N$ drawn from an unknown data distribution u_0. Rather than directly constructing the empirical score function associated with these samples, one typically trains a neural network to approximate the (unknown) true score function $s(x,t) = \nabla_x \log u(x,t)$.

This leads to the standard score-matching formulation [29]:

$$\inf_{\theta\in\mathbb{R}^P}\int_0^T\int_{\mathbb{R}^d} u(x,t)\left\|s(x,t)-s_{\mathrm{NN}}(x,t;\theta)\right\|^2 dx\,dt, \tag{4.1}$$

where $s_{\mathrm{NN}}(\,\cdot\,;\theta)$ denotes a neural network parameterized by θ. A motivation for this score-matching problem is provided in Remark 5.4, which arises from the KL divergence stability analysis in Lemma 5.3.

However, the optimization problem (4.1) is not directly tractable in practice, as the true data distribution u_0, and hence the associated solution u and score function s, are unknown.

Using the following identity:

$$u(x,t)s(x,t)=\nabla u(x,t)=\nabla G_t * u_0(x)=\int_{\mathbb{R}^d}\frac{y-x}{2t}G_t(x-y)du_0(y),$$

where G_t is the Gaussian heat kernel, (4.1) turns out to be equivalent to

$$\inf_{\theta\in\mathbb{R}^P}\int_{\mathbb{R}^d}\int_0^T\int_{\mathbb{R}^d}\left\|s_{\mathrm{NN}}(x,t;\theta)-\frac{y-x}{2t}\right\|^2 G_t(x-y)\,dx\,dt\,du_0(y). \tag{4.2}$$

In practice, we only have access to a finite collection of samples $\{y_k\}_{k=1}^N$ drawn from the initial distribution u_0.

By replacing the integral with respect to u_0 by the empirical average over the observed samples, we recover the classical empirical score-matching objective [29, 28]:

$$\inf_{\theta\in\mathbb{R}^P}\frac{1}{N}\sum_{k=1}^N\int_0^T\int_{\mathbb{R}^d}\left\|s_{\mathrm{NN}}(x,t;\theta)-\frac{y_k-x}{2t}\right\|^2 G_t(x-y_k)\,dx\,dt. \tag{4.3}$$

Moreover, by performing the change of variables $z=(x-y)/(2t)$, one recovers the DDPM [15] type loss:

$$\inf_{\theta\in\mathbb{R}^P}\frac{1}{N}\sum_{k=1}^N\int_0^T\int_{\mathbb{R}^d}\left\|s_{\mathrm{NN}}(2tz+y_k,t;\theta)+z\right\|^2 G_{(4t)^{-1}}(z)\,dz\,dt. \tag{4.4}$$

Here, each clean data sample y_k is perturbed by Gaussian noise z to produce a noisy sample $x=2tz+y_k$. The neural network $s_{\mathrm{NN}}(x,t;\theta)$ is trained to predict the negative of this added noise, which serves as the continuous-time analogue of the discrete DDPM training objective introduced by [15].

Another classical reformulation of the score-matching objective is the *Hyvärinen* loss introduced in [16]. It can be derived by applying integration by parts to the original score-matching problem (4.1). In the finite-sample setting, the Hyvärinen loss function takes the form

$$\inf_{\theta\in\mathbb{R}^P}\frac{1}{N}\sum_{k=1}^N\int_0^T\int_{\mathbb{R}^d}\Big(\left\|s_{\mathrm{NN}}(x,t;\theta)\right\|^2+2\,\mathrm{div}_x\, s_{\mathrm{NN}}(x,t;\theta)\Big)\,G_t(x-y_k)\,dx\,dt. \tag{4.5}$$

If the expressivity of neural networks were unbounded, i.e., if there were no constraints on s_{NN}, then the empirical score function (3.7) would naturally minimize (4.3) by satisfying its first-order optimality condition, see also [14, 33]. However, as shown by our main results, such a score function drives the generative flow to concentrate precisely on the training samples $\{y_k\}_{k=1}^N$, leading to pure imitation rather than genuine sample generation.

This observation highlights the fundamental role played by neural network parameterization. By constraining the class of admissible score functions, the network architecture introduces an implicit regularization effect that smooths the learned score and prevents it from collapsing onto the empirical distribution. As a result, the trained score exhibits improved generalization and supports more meaningful generative behavior, even though the training objective itself does not

explicitly include any regularization term. This observation is consistent with the explanation proposed in [33], which attributes such improved generalization to the optimization bias inherent in neural network training. To see this more clearly, consider training the score function using a multilayer perceptron with a ReLU activation. Once the parameters are fixed, the resulting neural network is globally Lipschitz continuous with respect to its inputs (x,t). Moreover, penalizing the network parameters implicitly regularizes this Lipschitz constant, thereby preventing the Li–Yau type blow-up as $t \to 0$.

4.2. **Other regularization techniques.** As shown by our main results, the key mechanism driving the imitation behavior of the generative flow is the $-1/t$ blow-up in the divergence of the true score function as $t \to 0$. Despite the implicit regularization provided by neural networks, additional strategies to improve the generative capacity of diffusion models, by mitigating this singularity, are presented below:

(1) *Early-stopping strategy.* When using the empirical score function to generate new data, one can halt the reverse process slightly before $t = 0$, at some positive time $t > 0$. In this case, the divergence of the empirical score is uniformly bounded from below by $-1/t$, thereby preventing the generated flow from collapsing strictly onto the training samples. This approach has been mentioned in [28, 3, 23]. Moreover, Theorem 3.11 provides an upper bound $\mathcal{O}(\sqrt{t})$ on the distance between generated data and the nearest training sample, offering insight into the trade-off between robustness and generative diversity.

(2) *Regularization in the loss function.* Another approach is to add a penalty term based on the divergence of the learned score function to the score-matching loss (4.3), yielding:

$$\inf_{\theta\in\mathbb{R}^P} \frac{1}{N}\sum_{k=1}^{N}\int_0^T\int_{\mathbb{R}^d}\left(\left\|s_{\mathrm{NN}}(x,t;\theta)-\frac{y_k-x}{2t}\right\|^2+\lambda\left(\mathrm{div}_x\, s_{\mathrm{NN}}(x,t;\theta)\right)^2\right)G_t(x-y_k)\,dx\,dt, \tag{4.6}$$

where $\lambda \geq 0$ is a hyperparameter controlling the strength of the regularization. This penalty was first proposed in the static setting by Kingma and LeCun [17] for generative models, and is closely related to *curvature-driven* smoothing technique in [4].

5. Proofs of main results

Throughout this section, we denote by u the solution of the heat equation (2.1), and by v the solution of the backward FP equation (3.1).

5.1. **Proof of Theorem 3.1.** The proof consists of the following three steps.

Step 1 (Regularity of the score function). Recall that $Q = \mathbb{R}^d \times (0,T]$ and that

$$s(x,t) = \nabla \log u(x,t) = \frac{\nabla u(x,t)}{u(x,t)}, \qquad \forall\,(x,t)\in Q.$$

By the convolution representation of u, the score function s is well-defined and belongs to $\mathcal{C}^\infty(Q)$. Therefore, the classical solution of the backward FP equation (3.1), denoted by v, exists and is unique in Q. By the monotonicity property of the FP equation (and of the continuity equation), we have

$$v(x,t) \geq 0, \quad \forall\,(x,t)\in Q.$$

Moreover, the Li–Yau inequality for the heat equation [22, Thm. 1.3] implies that

$$\mathrm{div}(s(x,t)) \geq -\frac{d}{2t}, \qquad \forall\,(x,t)\in Q. \tag{5.1}$$

Step 2 (Energy identity). Fix any $p \in [1,\infty)$. Assume that $v_T \in L^p(\mathbb{R}^d)$. Multiply equation (3.1) by v^{p-1} and integrate over $\mathbb{R}^d$:

$$\int_{\mathbb{R}^d} \left(v^{p-1}\partial_t v + \epsilon\, v^{p-1}\Delta v - (1+\epsilon)\, v^{p-1}\,\mathrm{div}(s\,v)\right) dx = 0.$$

Each term is handled as follows:

- Time derivative:
$$\int_{\mathbb{R}^d} v^{p-1}\partial_t v\, dx = \frac{1}{p}\frac{d}{dt}\|v\|_{L^p}^p.$$
- Diffusion term:
$$\int_{\mathbb{R}^d} v^{p-1}\Delta v\, dx = -(p-1)\int_{\mathbb{R}^d} v^{p-2}\|\nabla v\|^2 dx.$$
- Advection term:
$$\begin{aligned}\int_{\mathbb{R}^d} v^{p-1}\,\mathrm{div}(s\,v)\,dx &= -\int_{\mathbb{R}^d}\big\langle \nabla(v^{p-1}),\, v\,s\big\rangle\,dx\\ &= -(p-1)\int_{\mathbb{R}^d} v^{p-2}\big\langle \nabla v,\, v\,s\big\rangle\,dx\\ &= -(p-1)\int_{\mathbb{R}^d} v^{p-1}\big\langle \nabla v,\, s\big\rangle\,dx\\ &= -\frac{p-1}{p}\int_{\mathbb{R}^d}\big\langle \nabla(v^p),\, s\big\rangle\,dx = \frac{p-1}{p}\int_{\mathbb{R}^d} v^p\,\mathrm{div}(s)\,dx.\end{aligned}$$

Putting all terms together yields the energy identity:

$$\frac{1}{p}\frac{d}{dt}\|v\|_{L^p}^p - \epsilon(p-1)\int_{\mathbb{R}^d} v^{p-2}\|\nabla v\|^2 dx - (1+\epsilon)\frac{p-1}{p}\int_{\mathbb{R}^d} v^p\,\mathrm{div}(s)\,dx = 0. \tag{5.2}$$

Step 3 (Energy estimate). Now, using the lower bound on $\mathrm{div}(s)$ from (5.1), we deduce from (5.2) the following differential inequality:

$$\begin{cases}\dfrac{d}{dt}\|v(t)\|_{L^p}^p \;\geq\; -\dfrac{d(1+\epsilon)(p-1)}{2t}\,\|v(t)\|_{L^p}^p, \quad \forall\, t\in(0,T),\\ \|v(T)\|_{L^p}^p = \|v_T\|_{L^p}^p.\end{cases}$$

Dividing both sides by $\|v(t)\|_{L^p}^p$ yields

$$\frac{d}{dt}\log\big(\|v(t)\|_{L^p}^p\big) \;\geq\; -\frac{C}{2t}, \qquad C := d(1+\epsilon)(p-1).$$

Consequently,

$$\frac{d}{dt}\log\big(\|v(t)\|_{L^p}\big) \;\geq\; -\frac{C}{2pt}.$$

Integrating this differential inequality from t to T gives

$$\log\big(\|v_T\|_{L^p}\big) - \log\big(\|v(t)\|_{L^p}\big) \;\geq\; -\frac{C}{2p}\log\left(\frac{T}{t}\right).$$

Exponentiating both sides, we obtain

$$\|v(t)\|_{L^p} \;\leq\; \left(\frac{T}{t}\right)^{\frac{C}{2p}}\,\|v_T\|_{L^p}.$$

This establishes the desired estimate (3.2).

5.2. **Proof of Theorem 3.4.** To prove Theorem 3.4, we first establish three auxiliary lemmas. The first one provides *a priori* bounds for the solution u of the heat equation under the assumption of a compactly supported initial distribution. The second shows that the terminal KL divergences are finite under our assumptions. The third concerns the contraction of the KL divergence in the backward time direction, known in information theory as the *data-processing inequality* [19].

Lemma 5.1 (A priori estimates for the heat equation)**.** *Let u_0 be an initial density with compact support contained in the ball $B_R(0)$ for some $R > 0$, i.e., $\operatorname{supp}(u_0) \subset B_R(0)$. Then, the corresponding solution $u(x,t) = (G_t * u_0)(x)$ satisfies the following estimates for all $(x,t) \in Q$:*

(1) ***Gaussian bounds:***

$$(4\pi t)^{-d/2} \exp\left(-\frac{(\|x\|+R)^2}{4t}\right) \le u(x,t) \le (4\pi t)^{-d/2} \exp\left(-\frac{(\|x\|-R)^2}{4t}\right). \tag{5.3}$$

(2) ***Hessian bounds:***

$$-\frac{1}{2t} I_d \preceq \operatorname{Hess}\big(\log u(x,t)\big) \preceq \left(-\frac{1}{2t} + \frac{R^2}{4t^2}\right) I_d, \tag{5.4}$$

where I_d denotes the $d \times d$ identity matrix.

(3) ***Exponential moment bound:*** *There exists a constant $C_{d,R,T} > 0$, depending only on d, R, and T, such that*

$$\sup_{0<t\le T} \int_{\mathbb{R}^d} e^{\|x\|}\, u(x,t)\, dx \le C_{d,R,T}. \tag{5.5}$$

Proof. The proof is presented in Section 5.4. □

Lemma 5.2 (Terminal bound)**.** *Under Assumption A, the following holds:*

$$\mathrm{KL}\big(v_T \,\|\, u(T)\big) < \infty.$$

Moreover, if we further assume that $|\log v_T(x)|$ is bounded by a polynomial function of $\|x\|$, then

$$\mathrm{KL}\big(u(T) \,\|\, v_T\big) < \infty.$$

Proof. The proof is presented in Section 5.4. □

Lemma 5.3 (Data-processing and strong data-processing inequalities)**.** *Under Assumption A, for every $t \in (0,T)$, we have*

$$\mathrm{KL}\big(v(t) \,\|\, u(t)\big) \le \mathrm{KL}\big(v_T \,\|\, u(T)\big) < \infty.$$

Moreover, the relative entropy satisfies the differential identity

$$\frac{d}{dt} \mathrm{KL}(v(t) \,\|\, u(t)) = \epsilon \int_{\mathbb{R}^d} \left\| \nabla \log \frac{v(x,t)}{u(x,t)} \right\|^2 v(x,t)\, dx. \tag{5.6}$$

If, in addition, $|\log v_T(x)|$ is bounded by a polynomial function of $\|x\|$, then for every $t \in (0,T)$,

$$\mathrm{KL}\big(u(t) \,\|\, v(t)\big) \le \mathrm{KL}\big(u(T) \,\|\, v_T\big) < \infty.$$

Furthermore, if v_T has a finite fourth moment, then the following differential identity holds:

$$\frac{d}{dt} \mathrm{KL}(u(t) \,\|\, v(t)) = \epsilon \int_{\mathbb{R}^d} \left\| \nabla \log \frac{u(x,t)}{v(x,t)} \right\|^2 u(x,t)\, dx. \tag{5.7}$$

The right-hand sides of (5.6) and (5.7) are referred to as the relative Fisher information between $v(t)$ and $u(t)$.

Proof. The proof is presented in Section 5.4. □

Remark 5.4. The dynamical identities (5.6) and (5.7) can be extended to the case where the transport terms are not identical. Let v_θ denote the solution of the backward FP equation (3.1) with transport term $\mathrm{div}(s_\theta v)$, which differs from $\mathrm{div}(sv)$. Then, by an argument analogous to that of Lemma 5.3, we obtain

$$\frac{d}{dt}\,\mathrm{KL}(u(t)\,\|\,v_\theta(t)) \;=\; \epsilon\int_{\mathbb{R}^d}\left\|\nabla\log\frac{u}{v_\theta}\right\|^2 u\,dx \;-\; (1+\epsilon)\int_{\mathbb{R}^d} u\left\langle s-s_\theta,\;\nabla\log\frac{u}{v_\theta}\right\rangle dx.$$

Applying Young's inequality to the second term, for any $\eta\in(0,\epsilon]$, yields

$$\frac{d}{dt}\,\mathrm{KL}(u(t)\,\|\,v_\theta(t)) \;\geq\; (\epsilon-\eta)\int_{\mathbb{R}^d}\left\|\nabla\log\frac{u}{v_\theta}\right\|^2 u\,dx \;-\; \frac{(1+\epsilon)^2}{4\eta}\underbrace{\int_{\mathbb{R}^d}\|s-s_\theta\|^2\,u\,dx}_{\text{Score-matching term}}.$$

This differential inequality highlights the natural appearance of the score-matching loss function discussed in (4.1). Moreover, integrating the inequality over (t,T) yields a quantitative stability estimate that connects the evolution of the KL divergence with the score discrepancy. This allows proving results similar to those in [7, 3], where the authors used the *Girsanov* Theorem to bound the error.

We are now in a position to present the proof of Theorem 3.4.

Proof of Theorem 3.4. The argument proceeds in several steps.

Step 1 (Pre-compactness). By Lemma 5.3, for any $t\in(0,T)$, we have

$$\mathrm{KL}(v(t)\|u(t)) \leq \mathrm{KL}(v_T\|u(T)) < \infty. \tag{5.8}$$

Let $\{t_n\}_{n\geq1}$ be any sequence with $t_n\to0^+$. We claim that $\{v(t_n)\}_{n\geq1}$ is pre-compact in the weak-* topology on $\mathcal{P}(\mathbb{R}^d)$. By the *Prokhorov* theorem, it suffices to show that the family $\{v(t_n)\}$ is tight, which follows from a uniform bound on the first moment:

$$\sup_{n\geq1}\int_{\mathbb{R}^d}\|x\|\,v(x,t_n)\,dx<\infty.$$

The *Donsker–Varadhan variational formula* for the KL divergence [9] gives, for any measurable function ϕ and any $\lambda>0$,

$$\int_{\mathbb{R}^d}\phi(x)\,v(x,t_n)\,dx \;\leq\; \frac{1}{\lambda}\mathrm{KL}(v(t_n)\,\|\,u(t_n)) + \frac{1}{\lambda}\log\int_{\mathbb{R}^d} e^{\lambda\phi(x)}u(x,t_n)\,dx.$$

Choosing $\phi(x)=\|x\|$ and $\lambda=1$, we obtain

$$\int_{\mathbb{R}^d}\|x\|\,v(x,t_n)\,dx \;\leq\; \underbrace{\mathrm{KL}(v(t_n)\,\|\,u(t_n))}_{=:\gamma_1} + \underbrace{\log\int_{\mathbb{R}^d} e^{\|x\|}u(x,t_n)\,dx}_{=:\gamma_2}.$$

The term γ_1 is uniformly bounded by (5.8). The term γ_2 is uniformly bounded by (5.5). Thus, the first moment is uniformly bounded, implying tightness and hence pre-compactness.

Step 2 (Absolute continuity and support of limit measures). By Step 1, passing to a subsequence if necessary, assume $v(t_n)\rightharpoonup^* v^*$ as $n\to\infty$. Since u_0 has compact support, we have $u(t_n)\to u_0$ in the W_2-Wasserstein distance [30, Chp. 6], and thus also weak-*.

By the lower semicontinuity (in the weak-* topology) of the KL divergence,

$$\mathrm{KL}(v^*\,\|\,u_0)\leq\liminf_{n\to\infty}\mathrm{KL}(v(t_n)\,\|\,u(t_n))<\infty.$$

Finite relative entropy implies absolute continuity, hence $v^*\ll u_0$. Consequently,

$$\mathrm{supp}(v^*)\subset\mathrm{supp}(u_0).$$

The first part of assertion (i) follows.

In the case where $|\log v_T(x)|$ is bounded by a polynomial function of $\|x\|$, the KL divergence $\mathrm{KL}\big(u(t)\,\|\,v(t)\big)$ remains uniformly bounded by $\mathrm{KL}\big(u(T)\,\|\,v_T\big)$, which is finite by Lemma 5.2. Consequently, we have

$$\mathrm{KL}(u_0\,\|\,v^*) \le \liminf_{n\to\infty} \mathrm{KL}(u(t_n)\,\|\,v(t_n)) < \infty.$$

It follows that

$$\mathrm{supp}(u_0) \subset \mathrm{supp}(v^*).$$

Hence, under this additional assumption, the two support sets coincide.

Step 3 (Concentration on the support). Suppose, for contradiction, that the first equality of assertion (ii) fails. Then there exist $\delta > 0$, an open set $U \supset \mathrm{supp}(u_0)$, and a sequence $t_n \to 0^+$ such that

$$v(t_n)(U) \le 1 - \delta, \qquad \forall n \in \mathbb{N}.$$

By Step 1, passing to a subsequence (still denoted t_n) we may assume $v(t_n) \rightharpoonup^* v^*$. By Step 2, $\mathrm{supp}(v^*) \subset \mathrm{supp}(u_0) \subset U$. Thus,

$$v^*(U) = 1.$$

Applying the *Portmanteau theorem* [20, Thm. 13.16] for the open set U, we have

$$\liminf_{n\to\infty} v(t_n)(U) \ge v^*(U) = 1,$$

contradicting $v(t_n)(U) \le 1 - \delta$. Hence, for any open neighborhood U of $\mathrm{supp}(u_0)$,

$$\lim_{t\to 0^+} v(t)(U) = 1.$$

Finally, let X_t denote the solution to the generative SDE (2.7). Then $v(t)$ is precisely the law of X_t, i.e.,

$$\mathbb{P}(X_t \in A) = v(t)(A), \qquad \forall \text{ Borel sets } A.$$

Taking $A = U$ in the above convergence result yields

$$\lim_{t\to 0^+} \mathbb{P}(X_t \in U) = 1,$$

establishing assertion (ii) of Theorem 3.4. □

5.3. **Proof of Theorem 3.11.** For convenience, we recall the generation ODE (3.8):

$$\begin{cases} \dfrac{dX_t}{dt} = -\,s(X_t,\,t), \quad t \in (0,T), \\ X_T = x_T \in \mathbb{R}^d. \end{cases} \tag{5.9}$$

The overall structure of the proof follows a Lyapunov stability argument for the non-autonomous case. It proceeds in three steps, corresponding respectively to assertions (i)–(iii) stated in the theorem.

Step 1 (Subsequence convergence). Consider the ODE (5.9) with x_T distributed as $\mathcal{N}(0, I_d)$, which satisfies Assumption A(ii). Recall that $S = \mathrm{supp}(u_0)$ and, for $t \in (0,T)$, define

$$Y_t :- d(X_t, S).$$

By Theorem 3.4 (ii) (in the case $\epsilon = 0$), for every $\delta > 0$,

$$\mathbb{P}\big(Y_t \ge \delta\big) = 1 - \mathbb{P}\big(Y_t < \delta\big) \xrightarrow[t\to 0^+]{} 0.$$

Hence, we can choose a deterministic sequence $t_n \to 0^+$ such that

$$\mathbb{P}\big(Y_{t_n} > \tfrac{1}{n}\big) \le 2^{-n}.$$

By the *Borel–Cantelli* lemma, it follows that $Y_{t_n} \to 0$ for $\mathbb{P}$-almost every x_T. Since $\mathcal{N}(0, I_d)$ and $\mathcal{L}^d$ are mutually absolutely continuous, the exceptional set is Lebesgue-null. Consequently,

$$d\big(X_{t_n},\, S\big) \to 0$$

for Lebesgue-a.e. x_T. As a consequence, the lower limit convergence (3.9) follows.

Step 2 (Convergence rate to the convex hull). Recall that

$$K = \mathrm{conv}(S) = \mathrm{conv}(\mathrm{supp}(u_0)).$$

Since S is compact, the convex hull K is also compact. Let

$$V(x) := \frac{1}{2}\, d(x, K)^2 \quad \text{and} \quad p(x) := \mathrm{proj}_K(x).$$

Then $V \in C^1(\mathbb{R}^d)$ with

$$\nabla V(x) = x - p(x).$$

Along any trajectory of (5.9),

$$\frac{d}{dt} V\big(X_t\big) = \big\langle \nabla V\big(X_t\big),\, \dot{X}_t \big\rangle = \big\langle X_t - p\big(X_t\big),\, -\nabla \log u\big(X_t, t\big) \big\rangle. \tag{5.10}$$

The score function s can be written as the Gaussian mean-shift formula: for any $(x,t) \in Q$,

$$s(x,t) = \nabla \log u(x,t) = \frac{m(x,t) - x}{2t}, \qquad \text{where} \quad m(x,t) := \frac{\int y\, e^{-\frac{|x-y|^2}{4t}}\, du_0(y)}{\int e^{-\frac{|x-y|^2}{4t}}\, du_0(y)}.$$

Thus, using (5.10), we have

$$\frac{d}{dt} V\big(X_t\big) = \frac{1}{2t} \big\langle X_t - m\big(X_t, t\big),\, X_t - p\big(X_t\big) \big\rangle.$$

Since $m(x,t)$ is a barycenter with positive weights supported on $\mathrm{supp}(u_0)$, it follows that $m(x,t) \in K$ for all $(x,t) \in Q$.

For any nonempty closed convex set $A \subset \mathbb{R}^d$, and for any $x \in \mathbb{R}^d$ and $y \in A$, the following inequality holds:

$$\langle x - y,\; x - \mathrm{proj}_A(x) \rangle \;\geq\; \|x - \mathrm{proj}_A(x)\|^2.$$

Applying this inequality with $A = K$, $y = m(X_t, t)$, and $x = X_t$, we obtain

$$\big\langle X_t - m\big(X_t, t\big),\, X_t - p\big(X_t\big) \big\rangle \;\geq\; \big\|X_t - p\big(X_t\big)\big\|^2.$$

Therefore,

$$\frac{d}{dt} V\big(X_t\big) \;\geq\; \frac{1}{2t} \|X_t - p(X_t)\|^2 = \frac{1}{t}\, V\big(X_t\big).$$

This implies that $\big(V(X_t)/t\big)' \geq 0$. Hence,

$$\frac{V(X_t)}{t} \leq \frac{V(X_T)}{T} \quad \forall\, t \in (0, T).$$

The desired estimate (3.10) follows immediately from the definition of V.

Step 3 (Convergence rate in the sum of Dirac case). We now assume that the initial datum u_0 of the heat equation is in the form of a sum of Dirac measures:

$$u_0 = \sum_{k=1}^{N} w_k\, \delta_{y_k},$$

with weights $w_k > 0$ and locations $y_k \in \mathbb{R}^d$. By Step 1, there exists a deterministic sequence $t_n \to 0^+$ such that

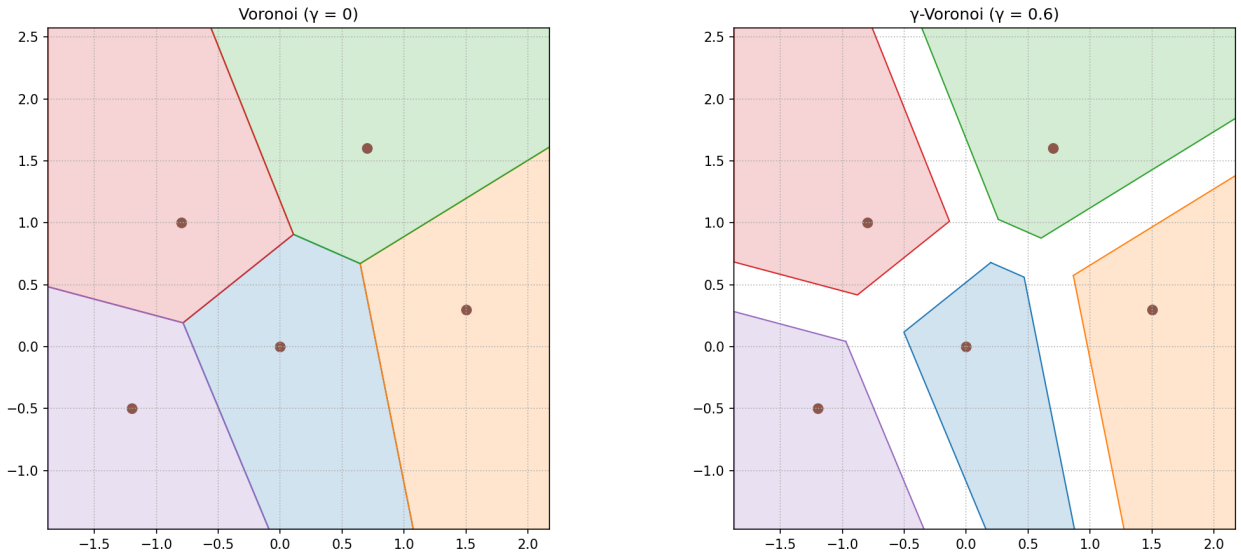
$$d\big(X_{t_n}, \operatorname{supp}(u_0)\big) \longrightarrow 0 \qquad \text{for Lebesgue-a.e. } x_T \in \mathbb{R}^d.$$

Fix any $x_T \in \mathbb{R}^d$ for which this convergence holds. Since $\operatorname{supp}(u_0)$ consists of finitely many points $\{y_k\}_{k=1}^N$, there exists an index $i \in \{1, \dots, N\}$ such that a subsequence of X_{t_n} converges to y_i. Define

$$\gamma := \frac{1}{2} \min_{j \neq i} \|y_i - y_j\|^2 \; > \; 0, \tag{5.11}$$

and introduce the corresponding γ-*Voronoi core* of y_i (An illustrative diagram of the γ-Voronoi core is shown in Figure 4):

$$V_i(\gamma) := \Big\{ x \in \mathbb{R}^d : \|x - y_j\|^2 - \|x - y_i\|^2 \geq \gamma \;\; \forall\, j \neq i \Big\}.$$



(A) Voronoi diagram ($\gamma = 0$) (B) γ–Voronoi core ($\gamma = 0.6$)

FIGURE 4. Examples of γ–Voronoi core of 5 points in $\mathbb{R}^2$. When $\gamma = 0$, we recover the classical Voronoi diagram.

We now make the following two claims (proved in the subsequent proofs):

(1) **Claim 1 (Varadhan-type concentration for the Gaussian mean shift).** For any $x \in V_i(\gamma)$ and any $\tau > 0$,

$$\big\|m(x,t) - y_i\big\| \;\leq\; \Big(\sum_{j \neq i} \frac{w_j}{w_i} \|y_j - y_i\|\Big)\, e^{-\frac{\gamma}{4t}},$$

where

$$m(x,t) = \frac{\sum_{k=1}^N w_k\, e^{-\frac{|x-y_k|^2}{4t}}\, y_k}{\sum_{k=1}^N w_k\, e^{-\frac{|x-y_k|^2}{4t}}}.$$

(2) **Claim 2 (Invariance of the γ–Voronoi core at late generation times).** There exists $t_* \in (0, T)$ such that

$$X_t \in V_i(\gamma) \quad \text{for all } t \in (0, t_*).$$

Assuming the two claims above, we proceed with the proof. Let

$$Z_t := X_t - y_i.$$

Using $\nabla \log u(x,t) = (m(x,t)-x)/2t$, we compute

$$\frac{d}{dt}\frac{1}{2}\|Z_t\|^2 = \frac{1}{2t}\big(\langle Z_t,\, y_i - m(X_t,t)\rangle + \|Z_t\|^2\big).$$

By Claims 1-2, for all $t \in (0,t_*)$,

$$\big\|m(X_t,t)-y_i\big\| \;\le\; \underbrace{\Big(\sum_{j\neq i}\frac{w_j}{w_i}\,\|y_j-y_i\|\Big)}_{=:C_i}\, e^{-\frac{\gamma}{4t}}.$$

Hence, for all $t \in (0,t_*)$,

$$\frac{d}{dt}\frac{1}{2}\|Z_t\|^2 \ge \frac{1}{2t}\|Z_t\|^2 - \frac{1}{2t}\|Z_t\|\, C_i\, e^{-\frac{\gamma}{4t}}.$$

Introduce the similarity time variable $s := -\log t$ (so that $ds/dt = -1/t$) and define $\rho(s) := \|Z_{t(s)}\|$. Using the chain rule,

$$\frac{d\rho}{ds} = \frac{d\rho}{dt}\frac{dt}{ds} = -\,t\,\frac{d\rho}{dt}.$$

Substituting the inequality above, we obtain

$$\frac{d\rho}{ds} \;\le\; -\frac{1}{2}\rho(s) + \frac{C_i}{2}\,e^{-\frac{\gamma}{4}e^s}, \qquad s \ge s_* := -\log t_*.$$

Multiplying both sides by $e^{s/2}$, we get

$$\frac{d}{ds}\Big(e^{\frac{s}{2}}\rho(s)\Big) \;\le\; \frac{C_i}{2}\,e^{\frac{s}{2}}e^{-\frac{\gamma}{4}e^s}.$$

Integrating this differential inequality from s_* to $s \ge s_*$ yields

$$e^{\frac{s}{2}}\rho(s) - e^{\frac{s_*}{2}}\rho(s_*) \;\le\; \frac{C_i}{2}\int_{s_*}^{s} e^{\frac{\sigma}{2}}e^{-\frac{\gamma}{4}e^\sigma}\,d\sigma.$$

Thus,

$$\rho(s) \;\le\; e^{-\frac{s}{2}}e^{\frac{s_*}{2}}\rho(s_*) + \frac{C_i}{2}e^{-\frac{s}{2}}\int_{s_*}^{s} e^{\frac{\sigma}{2}}e^{-\frac{\gamma}{4}e^\sigma}\,d\sigma.$$

Returning to the original variable $t = e^{-s}$, we obtain the bound

$$\|Z_t\| \;\le\; \left(\frac{t}{t_*}\right)^{\frac{1}{2}}\|Z_{t_*}\| + \frac{C_i}{2}t^{\frac{1}{2}}\int_{-\log t_*}^{-\log t} e^{\frac{\sigma}{2}}e^{-\frac{\gamma}{4}e^\sigma}\,d\sigma.$$

The integral term is bounded by

$$\int_{-\log t_*}^{-\log t} e^{\frac{\sigma}{2}}e^{-\frac{\gamma}{4}e^\sigma}\,d\sigma \le \int_{\mathbb{R}} e^{\frac{\sigma}{2}}e^{-\frac{\gamma}{4}e^\sigma}\,d\sigma = \frac{2\sqrt{\pi}}{\sqrt{\gamma}},$$

where the last equality is by a change of variable and the Gamma integral. Therefore,

$$\|Z_t\| \;\le\; \Big((t_*)^{-1/2}\|Z_{t_*}\| + C_i\sqrt{\tfrac{\pi}{\gamma}}\Big)\,t^{1/2}, \qquad \forall\, t \in (0,t_*).$$

We conclude that $X_t \to y_i$ as $t \to 0^+$. Here, the constants t_*, C_i, and γ depend only on the initial point x_T and the time horizon T. Since Z_t is continuous, the desired convergence rate (3.11) holds over the entire interval $(0,T)$ upon choosing a sufficiently large constant C.

Proof of Claim 1. Fix i the index in Step 3 of the proof above. For any $x \in \mathbb{R}^d$, let us define

$$\psi_j(x) := \|x - y_j\|^2 - \|x - y_i\|^2, \quad \forall j \neq i. \tag{5.12}$$

By the definition of $V_i(\gamma)$, we have that for any $x \in V_i(\gamma)$,

$$\psi_j(x) \geq \gamma, \quad \forall j \neq i.$$

Factor the i-th term in the numerator and the denominator of $m(x,t)$:

$$m(x,t) = \frac{w_i y_i + \sum_{j\neq i} w_j e^{-\frac{\psi_j(x)}{4t}}\, y_j}{w_i + \sum_{j\neq i} w_j e^{-\frac{\psi_j(x)}{4t}}} = y_i \; + \; \frac{\sum_{j\neq i} w_j e^{-\frac{\psi_j(x)}{4t}} (y_j - y_i)}{w_i + \sum_{j\neq i} w_j e^{-\frac{\psi_j(x)}{4t}}}.$$

Since the denominator is larger than w_i, we obtain

$$\|m(x,t) - y_i\| \leq \frac{1}{w_i} \sum_{j\neq i} w_j e^{-\frac{\psi_j(x)}{4t}} \, |y_j - y_i| \leq \Big(\sum_{j\neq i} \frac{w_j}{w_i} \, |y_j - y_i| \Big) e^{-\frac{\gamma}{4t}},$$

because $e^{-\psi_j(x)/(4t)} \leq e^{-\gamma/(4t)}$ for all $j \neq i$. □

Proof of Claim 2. Let $d_{ij} := \|y_i - y_j\|$ and recall

$$C_i = \sum_{j\neq i} \frac{w_j}{w_i} \, \|y_j - y_i\|.$$

Since $\gamma > 0$, there exists $\tau_0 \in (0,T)$ such that for all $t \in (0, \tau_0)$,

$$C_i \, e^{-\frac{\gamma}{4t}} \; \leq \; \frac{1}{8} \, d_{ij} \qquad \forall \, j \neq i. \tag{5.13}$$

Since $y_i \in \mathring{V}_i(\gamma)$ and the sequence $\{X_{t_n}\}_{n\geq 1}$ converge to y_i, there exists $t_* < \tau_0$ such that

$$X_{t_*} \in V_i(\gamma).$$

We now prove that $X_t \in V_i(\gamma)$ for all $t < t_*$. Suppose not. Then, there exists an *entrance time* $\tau_1 < t_*$, the time at which the trajectory enters $V_i(\gamma)$ (and thus exits in the backward-time sense), such that

$$X_{\tau_1} \in \partial V_i(\gamma).$$

This implies that there exists some $j \neq i$ such that

$$\psi_j\big(X_{\tau_1}\big) = \gamma \quad \text{and} \quad \frac{d}{dt}\psi_j\big(X_{\tau_1}\big) \geq 0, \tag{5.14}$$

where ψ_j denotes the Lyapunov-type function defined in (5.12).

Differentiating along the trajectory gives

$$\frac{d}{dt}\psi_j\big(X_{\tau_1}\big) = \frac{1}{\tau_1} \, \big\langle y_i - y_j, \, X_{\tau_1} - m\big(X_{\tau_1}, \tau_1\big) \big\rangle.$$

Decompose the scalar product:

$$\begin{aligned} \big\langle y_i - y_j, \, X_{\tau_1} - m\big(X_{\tau_1}, \tau_1\big) \big\rangle &= \big\langle y_i - y_j, \, y_i - m\big(X_{\tau_1}, \tau_1\big) \big\rangle + \big\langle y_i - y_j, \, X_{\tau_1} - y_i \big\rangle \\ &= \underbrace{\big\langle y_i - y_j, \, y_i - m\big(X_{\tau_1}, \tau_1\big) \big\rangle}_{\text{error}} \; + \; \underbrace{\tfrac{1}{2}\big(\psi_j(X_{\tau_1}) - d_{ij}^2\big)}_{\text{geometric}}. \end{aligned}$$

Since $X_{\tau_1} \in \partial V_i(\gamma) \subset V_i(\gamma)$, Claim 1 yields

$$\big| \big\langle y_i - y_j, \, y_i - m\big(X_{\tau_1}, \tau_1\big) \big\rangle \big| \; \leq \; d_{ij} \, C_i \, e^{-\frac{\gamma}{4\tau_1}} \; \leq \; \frac{1}{8} \, d_{ij}^2, \tag{5.15}$$

where the last inequality uses (5.13) and $\tau_1 < t_* < \tau_0$.

For the geometric term, since $\psi_j(X_{\tau_1}) = \gamma$,

$$\frac{1}{2}\big(\psi_j(X_{\tau_1}) - d_{ij}^2\big) = \frac{1}{2}\big(\gamma - d_{ij}^2\big) \;\le\; -\frac{1}{4}\, d_{ij}^2, \tag{5.16}$$

where the final bound follows from (5.11).

Combining (5.15) to (5.16) gives

$$\frac{d}{dt}\psi_j\big(X_{\tau_1}\big) \;\le\; \frac{1}{\tau_1}\left(\frac{1}{8}\, d_{ij}^2 - \frac{1}{4}\, d_{ij}^2\right) = -\frac{1}{8\tau_1}\, d_{ij}^2 \;<\; 0.$$

We obtain a contradiction regarding (5.14) for the sign of the derivative. Hence, $X_t \in V_i(\gamma)$ for all $t < t_*$. □

5.4. **Proof of technical lemmas.**

Proof of Lemma 5.1. We split the proof in three steps for the three bounds in Lemma 5.1.

Step 1 (Proof of Gaussian bounds). Since $\mathrm{supp}(u_0) \subset B_R(0)$, we have that for any $x \in \mathbb{R}^d$,

$$u(x,t) = \int_{B_R(0)} G_t(x-y) du_0(y).$$

For all $y \in B_R(0)$, the triangle inequality implies

$$\|x\| - R \le \|x\| - \|y\| \le \|x-y\| \le \|x\| + \|y\| \le \|x\| + R,$$

Therefore,

$$\exp\left(-\frac{(\|x\|+R)^2}{4t}\right) \le \exp\left(-\frac{\|x-y\|^2}{4t}\right) \le \exp\left(-\frac{(\|x\|-R)^2}{4t}\right).$$

By the formulation of G_t and the fact that $\int_{B_R(0)} du_0(y) = 1$, the estimate (5.3) follows.

Step 2 (Proof of the Hessian bounds). We now establish (5.4). A direct computation yields

$$\mathrm{Hess}\big(\log u(x,t)\big) = -\frac{1}{2t}\, I_d + \frac{1}{4t^2}\, \mathrm{Cov}(Y_x),$$

where

$$\mathrm{Cov}(Y_x) = \int_{\mathbb{R}^d} y\, y^\top \frac{G_t(x-y)}{u(x,t)}\, du_0(y) - \left(\int_{\mathbb{R}^d} y\, \frac{G_t(x-y)}{u(x,t)}\, du_0(y)\right)\left(\int_{\mathbb{R}^d} y\, \frac{G_t(x-y)}{u(x,t)}\, du_0(y)\right)^\top.$$

This is precisely the covariance matrix of the random variable $Y_x \sim (G_t(x-\cdot)/u(x,t))\, du_0(\cdot)$.

For any unit vector $a \in \mathbb{R}^d$, since $\mathrm{supp}(u_0) \subset B_R(0)$, we have

$$0 \;\le\; \mathrm{Var}(a^\top Y_x) \;=\; a^\top \mathrm{Cov}(Y_x) a \;\le\; \mathbb{E}\big[(a^\top Y_x)^2\big] \;\le\; R^2.$$

Hence,

$$0 \;\preceq\; \mathrm{Cov}(Y_x) \;\preceq\; R^2 I_d.$$

Substituting this estimate into the previous expression gives

$$-\frac{1}{2t}\, I_d \;\preceq\; \mathrm{Hess}\big(\log u(x,t)\big) \;\preceq\; \left(-\frac{1}{2t} + \frac{R^2}{4t^2}\right) I_d,$$

which proves (5.4).

Step 3 (Proof of the exponential moment bound). By definition,

$$\int_{\mathbb{R}^d} e^{\|x\|}\, u(x,t)\, dx = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{\|x\|}\, G_t(x-y)\, dx\, du_0(y).$$

Here, the *Fubini* theorem applies because all terms are nonnegative. Using the triangle inequality $\|x\| \le \|y\| + \|x-y\|$, we have

$$e^{\|x\|} \le e^{\|y\|}\, e^{\|x-y\|}.$$

With the change of variables $z = x - y$, this gives

$$\int_{\mathbb{R}^d} e^{\|x\|}\, u(x,t)\, dx \le \left(\int_{\mathbb{R}^d} e^{\|y\|}\, du_0(y)\right)\left(\int_{\mathbb{R}^d} e^{\|z\|}\, G_t(z)\, dz\right) \le e^R \int_{\mathbb{R}^d} e^{\|z\|}\, G_t(z)\, dz,$$

since $\operatorname{supp}(u_0) \subset B_R(0)$.

Finally, because G_t is a Gaussian measure on $\mathbb{R}^d$, the *Fernique* theorem (see [11]) implies that the exponential moment $\int_{\mathbb{R}^d} e^{\|z\|}\, G_t(z)\, dz$ is finite and, moreover, uniformly bounded for $t \in (0, T]$ by a constant depending only on d and T. This proves the desired estimate (5.5). □

Proof of Lemma 5.2. Since u_0 has compact support, the function $u(T)$ is smooth, strictly positive, and absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^d$. In particular, we have $v_T \ll u(T)$.

By the definition of the KL divergence,

$$\mathrm{KL}\big(v_T \,\|\, u(T)\big) = \int_{\mathbb{R}^d} v_T(x) \log \frac{v_T(x)}{u(x,T)}\, dx = \underbrace{\int_{\mathbb{R}^d} v_T(x) \log v_T(x)\, dx}_{\text{negative Shannon entropy}} - \underbrace{\int_{\mathbb{R}^d} v_T(x) \log u(x,T)\, dx}_{\text{negative cross-entropy}}.$$

By Assumption A, the first term (negative Shannon entropy) is finite. To control the second term, using (5.3), there exists $C > 0$ such that

$$\left|\int_{\mathbb{R}^d} v_T(x) \log u(x,T)\, dx\right| \le \int_{\mathbb{R}^d} v_T(x)\, \big(C(1 + \|x\|^2)\big)\, dx, \quad \forall x \in \mathbb{R}^d.$$

By Assumption A, v_T has a finite second moment; hence, the integral on the right-hand side is finite. This proves that $\mathrm{KL}(v_T \,\|\, u(T)) < \infty$.

For the second part, consider $\mathrm{KL}(u(T) \,\|\, v_T)$. The negative Shannon entropy of $u(T)$ is finite due to the Gaussian-type decay of $u(T)$. The negative cross-entropy term involves

$$\int_{\mathbb{R}^d} u(x,T) \log v_T(x)\, dx.$$

If $\log v_T(x)$ is uniformly bounded by a polynomial function of $\|x\|$, then it is integrable against $u(x,T)$. Consequently, $\mathrm{KL}(u(T) \,\|\, v_T) < \infty$. □

Proof of Lemma 5.3. The proof consists of two steps.

Step 1 (Data-processing inequality). By (5.4), the score function

$$s(x,t) = \nabla \log u(x,t)$$

is Lipschitz continuous with respect to x for every $t > 0$. Moreover, the Lipschitz constant can be chosen uniformly on any interval $[t, T]$ with $t > 0$.

Therefore, by the data-processing inequality for the Markov semigroup (see [5, Theorem 9.8.25 and Remark 9.8.27] for the PDE version), we obtain

$$\mathrm{KL}\big(v(t) \,\|\, u(t)\big) \;\le\; \mathrm{KL}\big(v_T \,\|\, u(T)\big).$$

By Lemma 5.2, the right-hand side is finite.

Furthermore, if $|\log v_T(x)|$ is bounded above by a polynomial function of $\|x\|$, then Lemma 5.2 implies that

$$\mathrm{KL}\big(u(T) \,\|\, v_T\big) < \infty.$$

Applying the data-processing inequality once more yields

$$\mathrm{KL}\big(u(t) \,\|\, v(t)\big) \;\le\; \mathrm{KL}\big(u(T) \,\|\, v_T\big) \;<\; \infty.$$

Step 2 (Strong data-processing inequality). We prove the identity (5.6); the proof of (5.7) follows by the same argument upon exchanging u and v. For any $t \in (0,T]$, by step 1, $\mathrm{KL}\left(v(t) \,\|\, u(t)\right) < \infty$. Set

$$w(t) \coloneqq \mathrm{KL}\left(v(t) \,\|\, u(t)\right) = \int_{\mathbb{R}^d} v(x,t)\, L(x,t)\, dx, \qquad L(x,t) \coloneqq \log \frac{v(x,t)}{u(x,t)}.$$

Differentiating in time and using $\partial_t L = \frac{\partial_t v}{v} - \frac{\partial_t u}{u}$ yields

$$\frac{d}{dt} w(t) = \int_{\mathbb{R}^d} \left((L+1)\, \partial_t v - \tfrac{v}{u}\, \partial_t u \right) dx.$$

Here, the interchange of time differentiation and integration follows from the argument in [19, Prop. 1.6, Appx. A], which requires the finiteness of the fourth-order moment of $u(T)$. Since $u(T)$ is the heat flow of compactly supported initial data, this condition holds. For the case of $\mathrm{KL}\big(u(t) \,\|\, v(t)\big)$, the same reasoning applies provided that v_T has a finite fourth-order moment.

Now use the FP equations satisfied by v and u:

$$\partial_t v = -\epsilon \Delta v + (1+\epsilon)\, \mathrm{div}(vs), \qquad \partial_t u = -\epsilon \Delta u + (1+\epsilon)\, \mathrm{div}(us).$$

Substituting these into the above gives

$$\begin{aligned} \frac{d}{dt} w(t) = &-\epsilon \int_{\mathbb{R}^d} (L+1)\, \Delta v\, dx + (1+\epsilon) \int_{\mathbb{R}^d} (L+1)\, \mathrm{div}(vs)\, dx \\ &+ \epsilon \int_{\mathbb{R}^d} \frac{v}{u}\, \Delta u\, dx - (1+\epsilon) \int_{\mathbb{R}^d} \frac{v}{u}\, \mathrm{div}(us)\, dx. \end{aligned}$$

We treat the diffusion and transport contributions separately.

- Diffusion part: Integrating by parts, we have

$$\int_{\mathbb{R}^d} (L+1)\, \Delta v\, dx = -\int_{\mathbb{R}^d} \langle \nabla L, \nabla v \rangle\, dx,$$

and

$$\int_{\mathbb{R}^d} \frac{v}{u}\, \Delta u\, dx = -\int_{\mathbb{R}^d} \left\langle \nabla\left(\frac{v}{u}\right), \nabla u \right\rangle dx = -\int_{\mathbb{R}^d} \left\langle \frac{\nabla v}{u} - \frac{v \nabla u}{u^2}, \nabla u \right\rangle dx.$$

Combining these terms yields

$$\begin{aligned} \mathcal{D} &= \epsilon \int_{\mathbb{R}^d} \left(\langle \nabla L, \nabla v \rangle - \frac{\langle \nabla v, \nabla u \rangle}{u} + \frac{v}{u^2} \|\nabla u\|^2 \right) dx \\ &= \epsilon \int_{\mathbb{R}^d} v \left(\frac{\|\nabla v\|^2}{v^2} + \frac{\|\nabla u\|^2}{u^2} - 2\, \frac{\langle \nabla v, \nabla u \rangle}{v\, u} \right) dx \\ &= \epsilon \int_{\mathbb{R}^d} v\, \|\nabla L\|^2\, dx, \end{aligned}$$

since $\nabla L = \frac{\nabla v}{v} - \frac{\nabla u}{u}$.

- Transport part: Integration by parts gives

$$\int_{\mathbb{R}^d} (L+1)\, \mathrm{div}(vs)\, dx = -\int_{\mathbb{R}^d} v\, \langle s, \nabla L \rangle\, dx,$$

and

$$-\int_{\mathbb{R}^d} \frac{v}{u}\, \mathrm{div}(us)\, dx = \int_{\mathbb{R}^d} v\, \langle s, \nabla L \rangle\, dx.$$

These two terms cancel exactly, so the total transport contribution satisfies $\mathcal{T} = 0$.

Combining the diffusion and transport contributions, we conclude that

$$\frac{d}{dt}w(t) = \epsilon \int_{\mathbb{R}^d} v(x,t)\,\|\nabla L(x,t)\|^2\,dx.$$

Since $L = \log \frac{v}{u}$, this is precisely the identity (5.6). □

6. Conclusions and perspectives

6.1. **Conclusions.** In this work, we developed a rigorous PDE framework for the analysis of score-based diffusion models. By reformulating the classical heat equation in terms of its associated score field, we demonstrated that the fundamental behavior of diffusion models can be characterized through the dynamics of a backward Fokker–Planck equation. A central component of our analysis is the Li–Yau inequality, which yields sharp divergence bounds for the score function and ensures the well-posedness of the backward flow, even in the presence of singularities in the empirical score near $t = 0$.

Building on this foundation, we derived sharp L^p estimates and established an entropy stability framework based on the Kullback–Leibler divergence. These results allowed us to rigorously characterize the behavior of the reverse-time dynamics and, in particular, to prove a concentration phenomenon: the solutions of the backward flow converge to the data manifold with probability one as $t \to 0$. In the deterministic setting, we further quantified this concentration rate, showing that it scales as $\sqrt{t}$ when the initial measure is a finite sum of Dirac masses.

This concentration result provides a fresh perspective on the imitative capacity of diffusion models, unifying various phenomena such as the implicit regularization induced by neural network training, the use of early-stopping strategies, and the introduction of Li–Yau-type regularizers.

The dual *imitation-generation* viewpoint proposed in this work to analyze diffusion models paves the way for a deeper theoretical understanding of generative modeling and for the development of new principles in algorithm design.

Note also that an instructive analogy can be drawn when comparing the classical adjoint methodology in optimal control with the forward-backward structure of diffusion models. In control theory, the adjoint equation arises from the variational analysis of an optimal control problem, coupling a forward state dynamics with a backward adjoint system that propagates sensitivity information and defines the optimal control. Similarly, diffusion models involve a forward diffusion process and a backward generative flow, whose interaction encodes the transformation between data and noise distributions. However, unlike in the adjoint framework, where the backward variable depends functionally on the forward trajectory through a well-defined costate or adjoint equation, in diffusion models, the backward evolution is not an adjoint in the variational sense but rather a probabilistic reversal of the forward stochastic dynamics, guided by the score field. This distinction highlights a shared mathematical symmetry between the two paradigms, while underlining their fundamentally different interpretative and analytical natures.

6.2. **Perspectives.** Future research directions can be broadly divided into two complementary lines: imitation and generation.

On the imitation side, several extensions of the present work are particularly promising:

- *Other PDE settings.* The theoretical framework developed in this article can be extended to more general dynamics beyond the linear heat equation or its FP variants. For instance, one may consider forward processes with nonlinear diffusion or transport operators, such as porous media or the p-Laplacian. Similar results to those obtained here via the Li–Yau estimate can be established using advanced nonlinear analogues, such as Aronson–Bénilan-type inequalities for porous-medium equations, the semiconcavity of viscous Hamilton–Jacobi equations, etc. Beyond the intrinsic interest of adapting our analysis to those settings, it would be worthwhile to explore the potential advantages for generative AI.

- *Stochastic generation flow and large deviation theory.* The convergence rate established in Theorem 3.11 concerns the deterministic generation flow (3.8) associated with the empirical score function. A natural next step is to extend this analysis to the stochastic generation flow (2.7). Since the diffusion term becomes negligible compared to the drift term of order $\mathcal{O}(1/t)$ near $t = 0$, one may expect to employ the theory of large deviations [12], to ensure the concentration phenomena for (2.7) with the empirical score, with high probability. This topic requires substantial further investigation.
- *Concentration rates for continuous data distributions.* In the finite-sample setting, the key step in proving the convergence rate is a Varadhan-type concentration result for Gaussian mean-shift distributions (see Claim 1 in the proof). In the continuous-data case, however, establishing such an inequality becomes significantly more challenging and strongly depends on the geometric properties of the boundary of the data manifold. Investigating this continuous setting constitutes an important direction for future work.

On the generation side, building on the discussions in Section 4, several research directions naturally emerge:

- *Network architecture based on gradient flow.* As discussed earlier, the generative process can be formulated as a gradient flow, where the underlying vector field corresponds to the gradient of the log-density. Consequently, instead of directly learning the score function, a natural and simpler approach is to learn the log-density ($\log u$), which is a scalar potential function, and use its gradient in the learned generative dynamics. The number of neurons required to approximate scalar functions is significantly smaller than that needed for approximating high-dimensional vector fields, see [24]. Combining this gradient-based architecture with the proposed loss functions yields a coherent and well-structured neural network learning framework. Moreover, during both training and generation, the required gradient can be computed efficiently via backpropagation without introducing discretization error. This design thus reduces model complexity while maintaining expressive power. Future work may compare this gradient-based neural architecture with alternative structures, both numerically and theoretically, particularly regarding the stability of the learned dynamics.
- *Trade-off between imitation and generation.* This trade-off is central to the early-stopping strategy. A larger choice of $t_{\min}$ promotes greater generative diversity but reduces fidelity to the original data distribution. Since Theorem 3.11 provides an upper bound on the distance to the true data manifold, it is meaningful to investigate principled strategies for selecting an optimal $t_{\min}$, potentially leveraging a priori geometric information about the underlying data manifold.
- *Li–Yau type regularization.* Incorporating a penalty term involving the divergence of the neural network score function, as in (4.6), is motivated by the Li–Yau stability analysis. Nevertheless, the most significant contribution from the Li–Yau framework arises from the negative part of the divergence of the score function, as demonstrated in the energy estimate. It is therefore of particular interest to compare two regularization strategies: penalizing only the negative part of $\operatorname{div}(s)$ or its full modulus, as in (4.6). A deeper theoretical and numerical investigation is required to understand how such regularization influences the learned score function and how the choice of its hyperparameter affects the trade-off between data imitation and sample generation.

References

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

[2] G. I. Barenblatt. *Scaling, Self-similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics*. Cambridge University Press, 1996.
[3] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
[5] V. I. Bogachev, N. V. Krylov, M. RÃ¶ckner, and S. V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*. American Mathematical Society, Mathematical Surveys and Monographs, Vol. 207, 2022.
[6] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.
[7] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
[8] G. Conforti, A. Durmus, and M. Gentiloni Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025.
[9] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.
[10] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Mathematics and Its Applications, Vol. 375, 1996.
[11] X. Fernique. Intégrabilité des vecteurs gaussiens. *Comptes Rendus de l'Académie des Sciences de Paris, Série A–B*, 270:A1698–A1699, 1970.
[12] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, Grundlehren der Mathematischen Wissenschaften, Vol. 260, 3rd ed., 2012.
[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.
[14] X. Gu, C. Du, T. Pang, C. Li, M. Lin, and Y. Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.
[15] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
[16] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.
[17] D. P. Kingma and Y. LeCun. Regularized estimation of image statistics by score matching. *Advances in Neural Information Processing Systems*, 23, 2010.
[18] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
[19] B. Klartag and O. Ordentlich. The strong data processing inequality under the heat flow. *IEEE Transactions on Information Theory*, 2025.
[20] A. Klenke. *Probability Theory: A Comprehensive Course*. Springer, 2008.
[21] H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.
[22] P. Li and S.-T. Yau. On the parabolic kernel of the Schrödinger operator. *Acta Mathematica*, 156:153–201, 1986.
[23] S. Li, S. Chen, and Q. Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.
[24] Z. Li, K. Liu, L. Liverani, and E. Zuazua. Universal approximation of dynamical systems by semi-autonomous neural ODEs and applications. *arXiv preprint arXiv:2407.17092*, 2024.
[25] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.
[26] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.
[27] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.
[28] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
[29] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.
[30] C. Villani. *Optimal Transport: Old and New*. Springer, Grundlehren der Mathematischen Wissenschaften, Vol. 338, 2008.
[31] C. Villani. *Hypocoercivity*. American Mathematical Society, 2009.

[32] M. P. Wand and M. C. Jones. *Kernel Smoothing*. CRC Press, 1994.
[33] M. Yi, J. Sun, and Z. Li. On the generalization of diffusion model. *arXiv preprint arXiv:2305.14712*, 2023.
[34] E. Zuazua. Asymptotic behavior of scalar convection–diffusion equations. *arXiv preprint arXiv:2003.11834*, 2020.