

A PDE PERSPECTIVE ON GENERATIVE DIFFUSION MODELS

KANG LIU¹ AND ENRIQUE ZUAZUA^{2,3,4}

ABSTRACT. Score-based diffusion models have emerged as a powerful class of generative methods, with successful applications across diverse domains. Despite their empirical success, their mathematical foundations remain only partially understood, particularly regarding the stability and consistency of the stochastic and partial differential equations underlying their dynamics.

In this work, we develop a partial differential equation (PDE) framework for score-based diffusion processes associated with the heat flow of a compactly supported data measure. After establishing weak well-posedness on positive-time intervals, we derive a sharp one-sided divergence bound for the score and use it to obtain sharp uniform L^p -stability estimates for the generative Fokker–Planck equation. This bound is the Euclidean form of the Li–Yau differential inequality. Through entropy stability methods, we prove that the generated laws converge in the 2-Wasserstein distance at rate $\sqrt{t}$ as $t \rightarrow 0^+$ to a probability measure supported on the data support. This holds for both deterministic and stochastic generation, for positive terminal densities with finite absolute entropy and second moment. In the deterministic case, almost every trajectory converges to a point of the data support; for empirical data, we further obtain a trajectory-wise square-root rate.

These results clarify how exact-score dynamics concentrate on the data support without necessarily recovering the data distribution. An explicitly solvable case identifies the limiting weights and their dependence on Gaussian initialization, while a relative-entropy estimate quantifies the effect of score approximation. Together, these findings connect the stability and concentration mechanisms with score learning and early stopping, providing a mathematical perspective on the balance between imitation fidelity and generative diversity.

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

¹UNIVERSITÉ BOURGOGNE EUROPE, CNRS, INSTITUT DE MATHÉMATIQUES DE BOURGOGNE, 21000 DIJON, FRANCE.

²FRIEDRICH - ALEXANDER - UNIVERSITÄT ERLANGEN - NÜRNBERG, DEPARTMENT OF MATHEMATICS, CHAIR FOR DYNAMICS, CONTROL, MACHINE LEARNING, AND NUMERICS (ALEXANDER VON HUMBOLDT PROFESSORSHIP), 91058 ERLANGEN, GERMANY.

³UNIVERSIDAD AUTÓNOMA DE MADRID, DEPARTAMENTO DE MATEMÁTICAS, 28049 MADRID, SPAIN.

⁴CHAIR OF COMPUTATIONAL MATHEMATICS, FUNDACIÓN DEUSTO, 48007 BILBAO, BASQUE COUNTRY, SPAIN.

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

2020 *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 partially supported by the European Research Council (ERC) under the European Union’s Horizon Europe research and innovation programme (grant agreement No. 101096251-CoDeFeL); by the Alexander von Humboldt Professorship program; the European Union’s Horizon Europe MSCA project ModConFlex (HORIZON-MSCA-2021-DN-01, project 101073558); the Transregio 154 Project “Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks” of the DFG; the AFOSR 24IOE027 project; the SURE-AI Norwegian Centre for Sustainable, Risk-Averse, and Ethical AI grant 357482, Research Council of Norway; by the Grant PID2023-146872OB-I00-DyCMaMod of MICIU (Spain) and by the COST Actions CA24122 – Multiscale Stochastics, Patterns, and Analysis of Combinatorial Environments and CA24136 – Interactions between Control Theory and Machine Learning.

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 (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 the dimension may be much higher for multimodal data 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* [26], face severe challenges in high-dimensional settings due to the *curse of dimensionality* [32, Ch. 4], which limits their use for modern data domains.

To overcome these limitations, deep generative models have been developed over the past decade. Among them, *diffusion models* [15, 29] have emerged as a particularly powerful class. These models seek to sample from u_0 by learning the *score function* of progressively noised data and modeling the reverse-time dynamics of an underlying diffusion process [2]. 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.

The forward noising process is governed by a *Fokker–Planck* (FP) or heat equation. Generation starts from a simple terminal law, usually Gaussian, and evolves backward using the score as a prescribed drift in a *stochastic differential equation* (SDE), or in the corresponding deterministic dynamics.

The present work concerns the stability and small-time behavior of this reverse evolution. We distinguish concentration on the data support from recovery of the probability weights on that support. For empirical data, these questions concern memorization of the training locations and the frequencies with which they are generated.

1.1. Formulation of diffusion models. Diffusion models combine a forward noising process with a backward generative process. We describe these two processes from the PDE perspective adopted in this work.

Forward noising process. Given a time horizon $0 < T < \infty$ and the ambient space $\mathbb{R}^d$, we first introduce the *forward heat equation*:

$$(1.1) \quad \begin{cases} \partial_t u(x, t) - \Delta u(x, t) = 0, & (x, t) \in Q, \\ u(\cdot, 0) = u_0 \in \mathcal{P}(\mathbb{R}^d). \end{cases}$$

Here, Q denotes the space-time cylinder

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

and $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$. The initial datum u_0 represents the unknown data distribution and need not admit a Lebesgue density. We use $\langle \cdot, \cdot \rangle$ and $|\cdot|$ for the Euclidean scalar product and norm, respectively, and $\text{dist}(x, A) := \inf_{y \in A} |x - y|$ for the distance from a point to a set.

The solution of (1.1) is given by convolution with the heat kernel:

$$(1.2) \quad u(x, t) = (G_t * u_0)(x), \quad (x, t) \in Q,$$

where

$$G_t(x) = (4\pi t)^{-d/2} \exp\left(-\frac{|x|^2}{4t}\right), \quad x \in \mathbb{R}^d, \quad t > 0.$$

This convolution progressively smooths the data distribution. Since the heat kernel is smooth and strictly positive, u is smooth and strictly positive at every positive time, even when u_0 is singular. We may therefore introduce the associated *score function*

$$(1.3) \quad s(x, t) = \nabla \log u(x, t) = \frac{\nabla u(x, t)}{u(x, t)}, \quad (x, t) \in Q.$$

The score provides the connection between forward noising and backward generation. Indeed, since $su = \nabla u$, the same heat-flow solution satisfies, for every $\epsilon \geq 0$,

$$(1.4) \quad \partial_t u + \epsilon \Delta u - (1 + \epsilon) \operatorname{div}(su) = 0.$$

Once the reference heat flow u is fixed, its score s is regarded as a prescribed vector field. With this convention, (1.4) is parabolic in the backward time direction when $\epsilon > 0$ and reduces to a transport equation when $\epsilon = 0$. This rewriting motivates a score-driven backward evolution, distinct from directly inverting the heat equation: the information about the data distribution is encoded in the prescribed score.

Backward generative process. Motivated by (1.4), we define the generative dynamics by the backward SDE

$$(1.5) \quad \begin{cases} dX_t = -(1 + \epsilon)s(X_t, t) dt + \sqrt{2\epsilon} dW_t, & t \in (0, T], \\ X_T \sim v_T, \end{cases}$$

where v_T is a probability distribution from which samples can be drawn easily, such as a Gaussian. The equation is integrated from T towards zero. Here, W_t denotes a Brownian motion in backward time, independent of X_T . When $\epsilon = 0$, (1.5) reduces to the deterministic ODE $\dot{X}_t = -s(X_t, t)$; when $\epsilon > 0$, it describes stochastic generation. Unless otherwise stated, our theoretical analysis uses the exact score defined by (1.3); its approximation by neural networks is discussed in Section 4.

In practice, the backward dynamics are discretized using a numerical method, such as the Euler–Maruyama scheme for stochastic generation or an ODE solver in the deterministic case. The idealized output distribution is the limiting law as $t \rightarrow 0^+$, whose existence and relation to u_0 are central questions of our analysis.

At the level of probability laws, the associated backward FP equation for $v(t) = \operatorname{law}(X_t)$ takes the form

$$(1.6) \quad \begin{cases} \partial_t v(x, t) + \epsilon \Delta v(x, t) - (1 + \epsilon) \operatorname{div}(s(x, t)v(x, t)) = 0, & (x, t) \in \mathring{Q}, \\ v(\cdot, T) = v_T. \end{cases}$$

Two features make the small-time behavior of this evolution nontrivial: the score may become singular as $t \rightarrow 0^+$, and the terminal initialization v_T need not coincide with the forward law $u(T)$.

1.2. Main results. For a compactly supported data measure u_0 , possibly singular, we study the stability of (1.6) and the concentration of the generated law near the singular endpoint $t = 0$. We establish three main results.

- **Stability of the generated densities (Theorem 2.2).** The Li–Yau inequality [21] provides the one-sided bound

$$\operatorname{div} s(x, t) \geq -\frac{d}{2t}$$

on the divergence of the score function. Combining this estimate with an L^p energy argument for the backward FP equation, we obtain, for $1 \leq p < \infty$, every $\epsilon \geq 0$, and terminal probability densities $v_T \in L^p(\mathbb{R}^d)$,

$$\|v(t)\|_{L^p} \leq \left(\frac{T}{t}\right)^{\frac{d(1+\epsilon)(p-1)}{2p}} \|v_T\|_{L^p}, \quad 0 < t \leq T.$$

The amplification factor is sharp in general as a uniform bound over terminal data; see Remark 2.4.

- **Convergence of the generated laws (Theorem 3.1).** We prove that, as $t \rightarrow 0^+$, the generated law converges to a limiting distribution v_0^* supported on the data support. This holds for both deterministic and stochastic generation, provided that the terminal law v_T has a positive density, finite absolute entropy, and a finite second moment. More precisely, the generated laws converge in the 2-Wasserstein distance with the rate

$$W_2(v(t), v_0^*) \leq C_T \sqrt{t}, \quad 0 < t \leq T.$$

Under the additional condition on v_T stated in the theorem, which is satisfied by every nondegenerate Gaussian initialization, the limit has exactly the same support as the data:

$$\text{supp}(v_0^*) = \text{supp}(u_0).$$

The limiting law may nevertheless differ from u_0 , since its weights depend on the terminal initialization.

- **Convergence of individual trajectories (Theorem 3.7).** For deterministic generation, we prove that for Lebesgue-almost every terminal point, the trajectory converges to a point of the data support as $t \rightarrow 0^+$, for an arbitrary compactly supported data measure. Moreover, in the empirical case $\text{supp}(u_0) = \{y_i\}_{i=1}^N$, the limiting point is one of the training samples and the convergence is quantitative:

$$|X_t - y_i| \leq C\sqrt{t}$$

for some $i \in \{1, \dots, N\}$.

Two complementary results identify the limiting weights in the explicit case $\epsilon = 1$ (Section 3.3) and quantify the error caused by score approximation (Proposition 4.1).

1.3. Related work and our contribution. We discuss the relation between our results and existing analyses of density stability, endpoint convergence, and memorization.

Stability of the generative equation. Positive-time well-posedness follows from the classical theory of FP and transport equations [7, 19]. Theorem 2.2 provides quantitative control as the score becomes singular near $t = 0$: the one-sided Li–Yau bound on its divergence yields an explicit L^p amplification factor that is sharp in general as a uniform bound over terminal data. Positivity of the evolution then gives the corresponding stability estimate for differences of probability solutions in Remark 2.3.

For approximate scores, [28, Thm. 1] bounds the data-to-generated Kullback–Leibler (KL) divergence by the terminal mismatch and a likelihood-weighted score-matching loss. In the heat-flow normalization, its coefficient agrees with Proposition 4.1 when $\epsilon = 1$. Our proposition places this established connection within the present PDE framework, treating every $\epsilon > 0$ through a direct entropy argument on positive-time intervals.

Convergence of the generated laws. For exact reverse diffusions, [25, Thm. 1] establishes endpoint existence under its time-reversal hypotheses and shows that absolute continuity of the terminal initialization with respect to the forward terminal law is inherited at the data endpoint. Equivalent terminal laws give equivalent endpoint laws, and the corresponding f -divergences contract. Thus support recovery under a change of initialization is already established in this stochastic setting.

Theorem 3.1 complements these endpoint results with the quantitative estimate $W_2(v(t), v_0^*) \leq C_T \sqrt{t}$. A common entropy and continuity-equation argument covers every $\epsilon \geq 0$, including deterministic generation, under the stated terminal-density, entropy, and moment assumptions. The data measure may have an arbitrary compact support, without manifold or other geometric regularity.

Gaussian approximation of the forward terminal law by moment matching is considered in [25, Lem. 2]. Our explicit analysis at $\epsilon = 1$ instead quantifies its effect on the generated endpoint. The expansion of the weight a_{v_T} in $dv_0^* = a_{v_T} du_0$, through order T^{-2} , identifies how Gaussian initialization changes the sampling weights and why matching its center to the data barycenter cancels the first-order correction.

Ornstein–Uhlenbeck noising and sampling estimates. Several sampling analyses use Ornstein–Uhlenbeck (OU) noising rather than the heat flow [5, 9, 10, 20]. After a normalization of time, the forward equation is

$$(1.7) \quad \partial_\tau w - \Delta w - \operatorname{div}(xw) = 0, \quad w(\cdot, 0) = u_0,$$

whose stationary probability density is $G_{1/2}$. The change to self-similar variables [4, 34] gives the relation

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

Thus heat and OU noising are related by a deterministic change of time and space, and the singular small-noise endpoint occurs in both formulations.

The cited sampling results control initialization, score approximation, and discretization errors in total variation or relative entropy under their respective assumptions. For compactly supported, possibly singular targets, [11, Thm. 1] gives a W_1 sampling bound under weighted score-error and step-size conditions, including an early-stopping error. Our estimates address the continuous exact-score dynamics at a fixed horizon and quantify its approach to its own endpoint. They are therefore complementary to numerical sampling guarantees, rather than bounds on the same error under weaker assumptions.

Convergence of individual trajectories. For empirical-score dynamics, [3, Thm. 4.3] uses Voronoi geometry to characterize possible limit points. If a limit point lies in the interior of a Voronoi cell, the whole trajectory converges to the corresponding training sample at the scale of the forward noise standard deviation. The theorem also allows limit points on Voronoi boundaries. In [3, Sect. 4.2], the authors conjecture that Gaussian initialization almost surely excludes these boundary limit points.

Theorem 3.7 establishes this conjectured conclusion in the heat-flow setting. For empirical data, Lebesgue-almost every terminal point gives a trajectory converging to a training sample, with the $\mathcal{O}(\sqrt{t})$ rate corresponding to the noise scale. The conclusion therefore holds almost surely for any absolutely continuous terminal law, including every nondegenerate Gaussian. More generally, part (i) proves convergence to a point of the data support for an arbitrary compactly supported measure. The finite-mean-length argument in Section 5.4 requires neither manifold regularity nor a lower-mass condition; Voronoi geometry is used only to obtain the empirical rate.

A complementary direction is pursued in [24, Thm. 2.5], where an inertia update replaces the final stage of empirical diffusion to mitigate memorization. Under a compact C^2 manifold assumption and regularity and positivity conditions on the data density, the modified method admits a high-probability W_1 approximation bound with a sample-size exponent governed by the intrinsic dimension, for the terminal initializations and horizon choices specified there. Our theorem concerns the unmodified dynamics and describes the endpoint concentration that motivates such modifications. Further aspects of empirical-score accuracy, large-dimensional dynamics, and training-data copying are studied in [6, 8, 22, 27].

1.4. Organization. Section 2 establishes positive-time well-posedness and L^p stability of the generative Fokker–Planck equation. Section 3 studies the small-time asymptotics of the probability laws and deterministic trajectories, followed by the explicit case $\epsilon = 1$. Section 4 relates score approximation to training and regularization. Section 5 begins with the heat-flow estimates and the classical well-posedness argument, and then presents the energy, endpoint, trajectory, and entropy

calculations. Section 6 concludes the paper. Appendices A and B provide the localization and approximation arguments for the stability estimates; Appendix C contains the explicit-case proofs.

2. WELL-POSEDNESS AND STABILITY OF THE GENERATIVE FOKKER–PLANCK EQUATION

We study the backward FP equation (1.6) with the prescribed score

$$s = \nabla \log u, \quad u(t) = G_t * u_0,$$

associated with the reference heat flow. We first establish well-posedness on intervals bounded away from $t = 0$. We then use the Li–Yau inequality to derive L^p estimates and stability with respect to the terminal law.

2.1. Existence and uniqueness. Assume that $u_0 \in \mathcal{P}(\mathbb{R}^d)$ has compact support contained in $B_R(0)$ for some $R > 0$. By Lemma 5.1, the score is smooth at positive times and satisfies

$$|s(x, t)| \leq \frac{|x| + R}{2t}, \quad (x, t) \in Q.$$

Moreover, the Hessian estimate in Lemma 5.1 gives, for every $0 < \tau < T$,

$$\sup_{t \in [\tau, T]} \|\nabla_x s(\cdot, t)\|_{L^\infty(\mathbb{R}^d)} < \infty.$$

Hence, on each interval $[\tau, T]$ bounded away from zero, the drift is globally Lipschitz in space and has at most linear growth. These properties place (1.6) within the classical positive-time well-posedness theory.

We equip $\mathcal{P}(\mathbb{R}^d)$ with the topology of narrow convergence, that is, convergence against bounded continuous functions. For probability measures, $\rightarrow$ denotes convergence in this topology. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures with finite second moment, endowed with the 2-Wasserstein distance W_2 ; see [30, Ch. 6].

A narrowly continuous curve $v : (0, T] \rightarrow \mathcal{P}(\mathbb{R}^d)$ is called a weak solution of (1.6) with terminal law v_T if $v(T) = v_T$ and

$$(2.1) \quad \int_0^T \int_{\mathbb{R}^d} (-\partial_t \varphi + \epsilon \Delta \varphi + (1 + \epsilon) \langle s, \nabla \varphi \rangle) dv(t) dt = 0, \quad \varphi \in C_c^\infty(\dot{Q}).$$

Proposition 2.1 (Well-posedness). *Let $T > 0$, $\epsilon \geq 0$, and let $u_0 \in \mathcal{P}(\mathbb{R}^d)$ have compact support.*

- (i) *For every $v_T \in \mathcal{P}(\mathbb{R}^d)$, there exists a unique weak solution of (1.6) in the sense of (2.1) such that*

$$v \in C((0, T]; \mathcal{P}(\mathbb{R}^d)), \quad v(T) = v_T.$$

If $\epsilon > 0$, the solution admits a smooth, strictly positive density on $\dot{Q}$.

- (ii) *If, in addition, $v_T \in \mathcal{P}_2(\mathbb{R}^d)$, then*

$$v \in C((0, T]; \mathcal{P}_2(\mathbb{R}^d), W_2).$$

Proof. The proof is given in Section 5.1. □

For $\epsilon > 0$, the classical correspondence between diffusion processes and FP equations identifies this weak solution with the time marginals of the generative SDE (1.5); see [7, Sect. 1.3].

2.2. The Li–Yau inequality and L^p stability. The preceding growth and regularity estimates ensure well-posedness on every interval bounded away from zero. To control the evolution as $t \rightarrow 0^+$, where the score may become singular, we use a one-sided estimate on its divergence.

The classical Li–Yau inequality [21] for positive solutions of the heat equation gives

$$(2.2) \quad \Delta \log u(x, t) \geq -\frac{d}{2t}, \quad (x, t) \in Q.$$

Since $s = \nabla \log u$, this is equivalent to

$$(2.3) \quad \operatorname{div} s(x, t) \geq -\frac{d}{2t}, \quad (x, t) \in Q.$$

In the present Euclidean setting, the same estimate also follows directly from the Gaussian covariance identity established in the proof of Lemma 5.1.

In contrast with the spatial Lipschitz estimate used for well-posedness, (2.3) controls only the negative part of the divergence. This is precisely the quantity entering the L^p energy argument and leads to the following estimate.

Theorem 2.2 (Energy estimate for the score-based FP equation). *Let $u_0 \in \mathcal{P}(\mathbb{R}^d)$ have compact support, let $T > 0$, $\epsilon \geq 0$, and $1 \leq p < \infty$. Suppose that $v_T \in \mathcal{P}(\mathbb{R}^d)$ admits a density in $L^p(\mathbb{R}^d)$. Then the unique weak solution of Proposition 2.1(i) admits a density, still denoted by v , such that*

$$v \in C((0, T]; L^p(\mathbb{R}^d)), \quad v(T) = v_T,$$

and

$$(2.4) \quad \|v(t)\|_{L^p} \leq \left(\frac{T}{t}\right)^{\frac{d(1+\epsilon)(p-1)}{2p}} \|v_T\|_{L^p}, \quad 0 < t \leq T.$$

Proof. The key argument, showing how the Li–Yau inequality enters the L^p energy estimate, is presented in Section 5.2 under auxiliary regularity assumptions. The rigorous proof under the hypotheses above, based on spatial cutoffs and approximation, is given in Appendix A. $\square$

Remark 2.3 (Stability with respect to the terminal law). Under the assumptions on u_0, T, ϵ , and p in Theorem 2.2, let $v_{1,T}$ and $v_{2,T}$ be probability densities in $L^p(\mathbb{R}^d)$. The corresponding weak solutions of (1.6), with the same score field, satisfy

$$(2.5) \quad \|v_1(t) - v_2(t)\|_{L^p} \leq \left(\frac{T}{t}\right)^{\frac{d(1+\epsilon)(p-1)}{2p}} \|v_{1,T} - v_{2,T}\|_{L^p}, \quad 0 < t \leq T.$$

In particular, the evolution is contractive in L^1 . The estimate follows from positivity and linearity of the solution operator together with Theorem 2.2; see Section 5.2.

Remark 2.4 (Sharpness of the estimate). The factor in (2.4) is optimal in general as a uniform bound over terminal probability densities, for every $\epsilon \geq 0$. Indeed, consider the Gaussian example

$$u_0 = \delta_0, \quad s(x, t) = -\frac{x}{2t}, \quad v_T^M = G_{M/2}, \quad M > 0.$$

The corresponding solution is

$$v^M(t) = G_{B_M(t)/2}, \quad B_M(t) = M \left(\frac{t}{T}\right)^{1+\epsilon} + 2t \left(1 - \left(\frac{t}{T}\right)^\epsilon\right),$$

where B_M solves

$$B'_M(t) = \frac{1+\epsilon}{t} B_M(t) - 2\epsilon, \quad B_M(T) = M.$$

Setting $\alpha_p = d(p-1)/(2p)$, for every fixed $0 < t < T$ we obtain

$$\lim_{M \rightarrow \infty} \frac{\|v^M(t)\|_{L^p}}{\|v_T^M\|_{L^p}} = \lim_{M \rightarrow \infty} \left(\frac{M}{B_M(t)} \right)^{\alpha_p} = \left(\frac{T}{t} \right)^{(1+\epsilon)\alpha_p},$$

which is exactly the factor in (2.4). Hence this factor cannot be improved uniformly over the terminal data.

This sharpness statement is uniform in the terminal datum and does not mean that every fixed datum has the same small-time behavior. For instance, if M is fixed and $\epsilon > 0$, then $B_M(t) \sim 2t$ as $t \rightarrow 0^+$. For $p = 1$, the estimate reduces to conservation of mass.

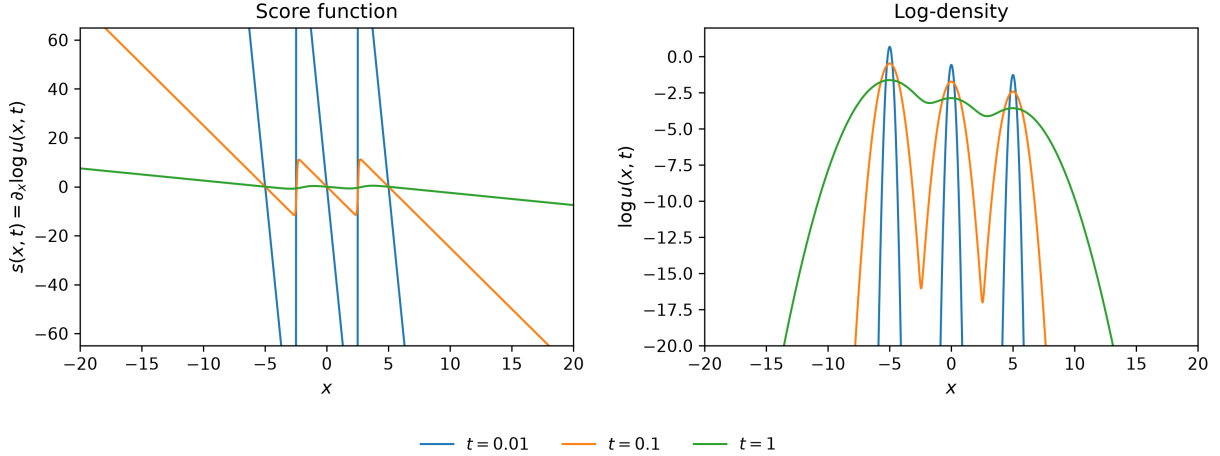


FIGURE 1. Score function (left) and log-density (right) of the heat flow originating from $u_0 = 0.7\delta_{-5} + 0.2\delta_0 + 0.1\delta_5$, shown at $t = 0.01, 0.1, 1$. The steep negative score slopes near the atoms illustrate the $1/t$ scale in the Li-Yau estimate. The log-density and score are evaluated using log-sum-exp and normalized Gaussian weights, respectively.

3. SMALL-TIME ASYMPTOTICS OF THE GENERATIVE PROCESS

We now study the behavior of the generative dynamics as $t \rightarrow 0^+$. We first establish concentration and convergence of the generated laws, and then analyze individual deterministic trajectories. The explicit case $\epsilon = 1$ further describes the dependence on terminal initialization.

3.1. Endpoint convergence and concentration on the data support. To analyze the endpoint behavior of the diffusion model, we study whether its solution concentrates on $\text{supp}(u_0)$ as $t \rightarrow 0^+$. We address this question by an entropy stability analysis of the FP equation.

Recall that the KL divergence (or relative entropy) between two probability measures $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$(3.1) \quad \text{KL}(\mu_1 \parallel \mu_2) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{d\mu_1}{d\mu_2}\right) d\mu_1, & \text{if } \mu_1 \ll \mu_2, \\ +\infty, & \text{otherwise,} \end{cases}$$

where $\mu_1 \ll \mu_2$ means that μ_1 is absolutely continuous with respect to μ_2 .

As we shall see in Lemma 5.3, the KL divergence between v (the solution of (1.6)) and u (the reference solution of the heat equation defining the score function) decays when viewed backward in time:

$$(3.2) \quad \text{KL}(v(t_1) \| u(t_1)) \leq \text{KL}(v(t_2) \| u(t_2)), \quad \forall 0 < t_1 \leq t_2 \leq T.$$

The entropy controls the generated-law score estimate used in the proof of the following theorem. In addition to positive-time well-posedness, the endpoint argument requires the following density and entropy assumptions.

Theorem 3.1 (Concentration of support and endpoint convergence). *Let $T > 0$, $\epsilon \geq 0$, and let $u_0 \in \mathcal{P}(\mathbb{R}^d)$ have compact support S . Assume that $v_T \in \mathcal{P}_2(\mathbb{R}^d)$ has a density satisfying*

$$(3.3) \quad v_T > 0 \quad \mathcal{L}^d\text{-almost everywhere}, \quad \int_{\mathbb{R}^d} v_T(x) |\log v_T(x)| dx < \infty.$$

Let v be the solution of (1.6), and set

$$K_T := \text{KL}(v_T \| u(T)), \quad C_T := 2\sqrt{2K_T + d \log 2} + \sqrt{\epsilon K_T}.$$

Then $K_T < \infty$, and the following assertions hold.

(i) *There exists $v_0^* \in \mathcal{P}_2(\mathbb{R}^d)$ such that*

$$(3.4) \quad W_2(v(t), v_0^*) \leq C_T \sqrt{t}, \quad 0 < t \leq T.$$

(ii) *One has $v_0^* \ll u_0$ and $\text{supp}(v_0^*) \subset S$. If there exist $C, q > 0$ such that $|\log v_T(x)| \leq C(1 + |x|^q)$ for $\mathcal{L}^d$ -almost every x , then $u_0 \ll v_0^*$ and*

$$\text{supp}(v_0^*) = \text{supp}(u_0) = S.$$

Proof. The proof is presented in Section 5.3. □

Remark 3.2 (Deterministic generation). For $\epsilon = 0$, the generative dynamics follow the gradient of the time-dependent potential $\log u$ in the backward time direction. Although local maximizers of $\log u(\cdot, t)$ at positive times need not belong to $\text{supp}(u_0)$, Theorem 3.1 shows that the generated law concentrates on the data support as $t \rightarrow 0^+$. At the trajectory level, Theorem 3.7 further shows that Lebesgue-almost every deterministic trajectory converges to a point of $\text{supp}(u_0)$; for empirical data, the convergence occurs toward a training sample with rate $\mathcal{O}(\sqrt{t})$.

Remark 3.3 (Concentration and dependence on initialization). Under the assumptions of Theorem 3.1, since $\text{supp}(v_0^*) \subset S$, we have

$$(3.5) \quad \left(\int_{\mathbb{R}^d} \text{dist}(x, S)^2 dv(t)(x) \right)^{1/2} \leq W_2(v(t), v_0^*) \leq C_T \sqrt{t}.$$

Consequently, for every $r > 0$,

$$(3.6) \quad v(t)\{x : \text{dist}(x, S) \geq r\} \leq \min \left\{ 1, \frac{C_T^2 t}{r^2} \right\}.$$

In particular, $v(t)(U) \rightarrow 1$ for every open neighborhood U of S .

The L^1 stability estimate also controls the dependence of the endpoint on the terminal initialization. Indeed, since $v(t) \rightarrow v_0^*$ and $u(t) \rightarrow u_0$ narrowly as $t \rightarrow 0^+$, lower semicontinuity of the total variation norm and Remark 2.3 with $p = 1$ give

$$(3.7) \quad \|v_0^* - u_0\|_{\text{TV}} \leq \|v_T - u(T)\|_{L^1}.$$

Here $\|\sigma\|_{\text{TV}} := |\sigma|(\mathbb{R}^d)$, so that $\|f dx\|_{\text{TV}} = \|f\|_{L^1}$. More generally, the same argument compares the limiting laws associated with any two admissible terminal densities.

Remark 3.4 (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 (1.5) 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 $t = 0$. The displayed positive-time samples lie away from the support; no monotonicity in the stopping time is asserted.

Panels (C)-(D) depict analogous trajectories when the generation flow (1.5) 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. This illustrates the discrepancy caused by stopping before the singular endpoint.

The two experiments illustrate concentration toward different reference supports: a curve for the population measure and finitely many points for the empirical measure. This comparison does not by itself quantify statistical generalization.

Remark 3.5 (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

$$(3.8) \quad s(x, t) = \frac{1}{2t} \left(\frac{\sum_{k=1}^N e^{-\frac{|x-y_k|^2}{4t}} (y_k - x)}{\sum_{k=1}^N e^{-\frac{|x-y_k|^2}{4t}}} \right) = \frac{1}{2t} \left(\frac{\sum_{k=1}^N e^{-\frac{|x-y_k|^2}{4t}} y_k}{\sum_{k=1}^N e^{-\frac{|x-y_k|^2}{4t}}} - x \right),$$

for all $(x, t) \in Q$.

Theorem 3.7 shows that almost every deterministic trajectory converges to one of these training points. This does not assert that their sampling weights are recovered.

Remark 3.6 (Examples of terminal-law dependence). Nondegenerate Gaussian densities satisfy the hypotheses of Theorem 3.1. The entropy and moment conditions ensure $\text{KL}(v_T \| u(T)) < \infty$; they are sufficient conditions, not claimed to be necessary. Strict positivity is not needed for support inclusion, but is relevant to recovering the entire support. The additional global bound $|\log v_T(x)| \leq C(1 + |x|^q)$ is a convenient sufficient condition for finite reverse entropy. The following examples illustrate possible failures for singular initializations or for laws confined to one basin.

Let

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

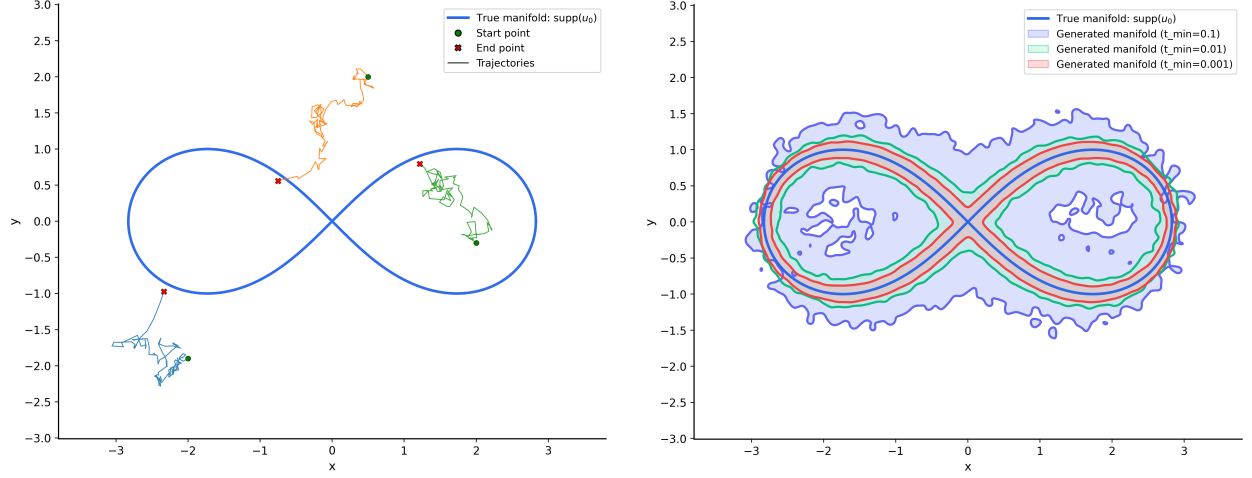
and consider the deterministic flow case ($\epsilon = 0$):

- (1) *Case 1.* Suppose that $\text{supp}(v_T) \subset \{(0, y) : y \in \mathbb{R}\}$. Then v_T does not possess a density in $\mathbb{R}^2$. From (3.8), one verifies that

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

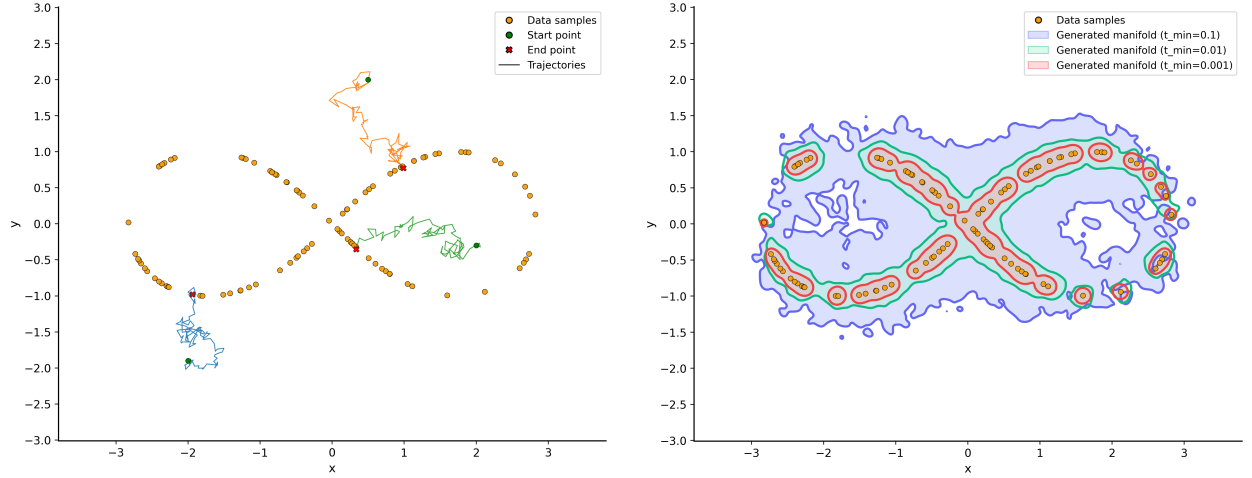
Consequently, the generative flow collapses every point in $\text{supp}(v_T)$ to the single point $(0, 0)$, which lies outside $\text{supp}(u_0)$, see the vertical-axis trajectories in Figure 3.

- (2) *Case 2.* Assume that $v_T \ll \mathcal{L}^2$ and $v_T((0, \infty) \times \mathbb{R}) = 1$, so the terminal density is not strictly positive on $\mathbb{R}^2$. By symmetry and uniqueness, the right half-plane is invariant under the deterministic flow. The almost-everywhere convergence in Theorem 3.7(i) therefore implies that v_T -almost every trajectory converges to $(1, 0)$. Consequently, the limiting law is $\delta_{(1,0)}$, rather than a law supported on both data points; see the right-hand trajectories in Figure 3.



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

(B) Early-stopped samples with the true score function.



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

(D) Early-stopped samples with the empirical score function.

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 (3.8). In all experiments, we set the backward diffusion parameter to be $\epsilon = 0.2$. The displayed sample region encompasses 95% of the 10,000 reverse-time samples, reducing the effect of extreme stochastic trajectories.

3.2. Trajectory convergence rate in the ODE setting. Theorem 3.1 gives concentration and a W_2 rate for the generated laws. Here, we study the *quantitative* concentration rate for the

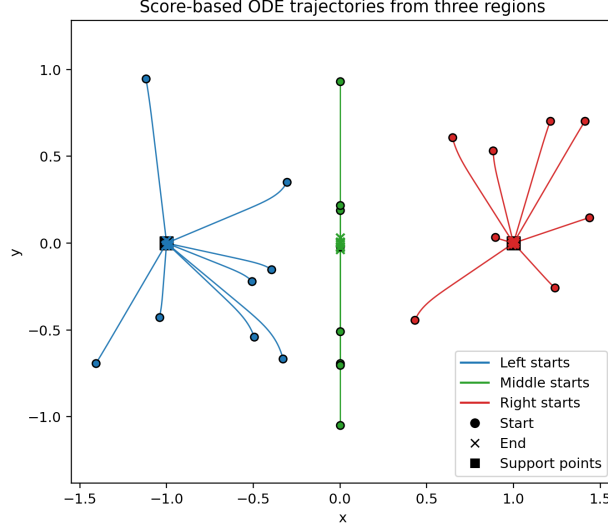


FIGURE 3. Trajectories of the score-based ODE ($\epsilon = 0$) with $u_0 = \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)})$. Initial points with $x < 0$ converge to $(-1, 0)$, those with $x > 0$ converge to $(1, 0)$, and points with $x = 0$ converge to the barycenter $(0, 0)$.

deterministic case of the generative flow (1.5), i.e., $\epsilon = 0$. The resulting ODE is

$$(3.9) \quad \begin{cases} \frac{dX_t}{dt} = -s(X_t, t), & t \in (0, T), \\ X_T = x_T, \end{cases}$$

where s is the score function (1.3) and x_T is any point in $\mathbb{R}^d$.

Theorem 3.7 (Trajectory convergence). *Let $u_0 \in \mathcal{P}(\mathbb{R}^d)$ have compact support $S = \text{supp}(u_0)$, and let X_t solve (3.9) with terminal condition $X_T = x_T \in \mathbb{R}^d$. Then the following assertions hold:*

- (i) *For Lebesgue-almost every $x_T \in \mathbb{R}^d$, the trajectory X_t has a limit in S as $t \rightarrow 0^+$. In particular,*

$$(3.10) \quad \lim_{t \rightarrow 0^+} \text{dist}(X_t, S) = 0.$$

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

$$(3.11) \quad \text{dist}(X_t, K) \leq \text{dist}(x_T, K) T^{-1/2} \sqrt{t}, \quad t \in (0, T].$$

- (iii) *If $S = \{y_1, \dots, y_N\}$ is finite, then for Lebesgue-almost every $x_T \in \mathbb{R}^d$, there exist $i \in \{1, \dots, N\}$ and $C = C(u_0, x_T, T) > 0$ such that*

$$(3.12) \quad |X_t - y_i| \leq C\sqrt{t}, \quad 0 < t \leq T.$$

Proof. The proof is presented in Section 5.4. □

Remark 3.8. Assertion (i) concerns Lebesgue-almost every terminal point, whereas (ii) holds for every terminal point but only controls the distance to the convex hull of the data support. The symmetry example in Figure 3 shows that exceptional initializations may converge outside S .

A geometric description of such exceptional limits is developed in [23]. For

$$f_S(x) = \text{dist}(x, S)^2, \quad \Pi_S(x) = \underset{y \in S}{\text{argmin}} |x - y|,$$

one has

$$\partial^C f_S(x) = 2(x - \text{conv } \Pi_S(x)),$$

where ∂^C denotes the Clarke generalized gradient. The corresponding equilibrium set is therefore

$$\mathcal{E}_S = \{x \in \mathbb{R}^d : x \in \text{conv } \Pi_S(x)\}.$$

The companion work shows that, for compact supports, every accumulation point of a deterministic trajectory belongs to $\mathcal{E}_S$.

The set $\mathcal{E}_S$ may strictly contain S . For instance, if S is a sphere centered at a , then

$$\mathcal{E}_S = S \cup \{a\}.$$

For the uniform measure on the sphere, symmetry gives $s(a, t) = 0$, so the trajectory initialized at a remains there. Thus the center provides a simple example of an exceptional limit outside the data support.

When S is finite, assertion (iii) strengthens (i) by providing the quantitative rate $\mathcal{O}(\sqrt{t})$ toward the limiting data point. The constant depends on u_0 , the terminal point x_T , and the time horizon T ; no estimate uniform over all data configurations or terminal points is asserted. For general compact supports, Theorem 3.1 provides the corresponding square-root rate at the level of probability laws.

3.3. The explicit case $\epsilon = 1$. Theorem 3.1 describes the support and convergence of the limiting law, but does not in general identify its probability weights. When $\epsilon = 1$, the quotient v/u satisfies a backward heat equation, which yields an explicit representation of the generated law and makes the effect of the terminal initialization transparent.

Proposition 3.9. *Assume the hypotheses of Theorem 3.1 and let $\epsilon = 1$. Then, for $0 < t < T$,*

$$(3.13) \quad v(x, t) = u(x, t) \int_{\mathbb{R}^d} \frac{G_{T-t}(x-y)}{u(y, T)} dv_T(y).$$

Moreover, the limiting law v_0^* of Theorem 3.1 is given by

$$(3.14) \quad dv_0^*(x) = a_{v_T}(x) du_0(x), \quad a_{v_T}(x) := \int_{\mathbb{R}^d} \frac{G_T(x-y)}{u(y, T)} dv_T(y).$$

One has $0 < a_{v_T} < \infty$ u_0 -almost everywhere and

$$\int_{\mathbb{R}^d} a_{v_T} du_0 = 1.$$

Hence v_0^* and u_0 are equivalent.

Proof. The proof is given in Appendix C. □

The explicit formula also allows us to quantify the effect of the terminal horizon for Gaussian initialization. Let

$$(3.15) \quad m_0 := \int_{\mathbb{R}^d} x du_0(x), \quad \Sigma_0 := \int_{\mathbb{R}^d} (x - m_0)(x - m_0)^\top du_0(x).$$

Proposition 3.10. *Let $u_0 \in \mathcal{P}(\mathbb{R}^d)$ have compact support S . For $v_T = G_T$, write $a_T = a_{v_T}$. Then, uniformly for $x \in S$,*

$$(3.16) \quad a_T(x) = 1 + \frac{\langle m_0, m_0 - x \rangle}{2T} + \frac{B_0(x)}{8T^2} + \mathcal{O}(T^{-3}), \quad T \rightarrow \infty,$$

where

$$(3.17) \quad B_0(x) = \langle m_0, x - m_0 \rangle^2 + \text{tr}(\Sigma_0^2) - \langle \Sigma_0(x - 2m_0), x - 2m_0 \rangle.$$

If instead $v_T = G_T(\cdot - m_0)$ and $a_T^c = a_{v_T}$, then, uniformly for $x \in S$,

$$(3.18) \quad a_T^c(x) = 1 + \frac{\text{tr}(\Sigma_0^2) - \langle \Sigma_0(x - m_0), x - m_0 \rangle}{8T^2} + \mathcal{O}(T^{-3}), \quad T \rightarrow \infty.$$

Proof. The proof is given in Appendix C. □

Remark 3.11 (Terminal initialization and limiting weights). For $\epsilon = 1$, under the preceding assumptions, the exact score fixes the endpoint support, while the terminal law may change the probability weights through $dv_0^* = a_{v_T} du_0$. Proposition 3.10 shows that for both Gaussian initializations considered above,

$$a_{v_T}(x) \longrightarrow 1 \quad \text{uniformly for } x \in S, \quad T \rightarrow \infty.$$

Thus the original data weights are recovered in the large-horizon limit. Centering the Gaussian at the data barycenter removes the first-order correction, while the second-order term retains information on the covariance of u_0 .

4. SCORE APPROXIMATION, LEARNING, AND REGULARIZATION

The preceding results concern the exact score. We now turn to its approximation by neural networks, the stability of the resulting generative equation, and the role of regularization.

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 $s(x, t) = \nabla_x \log u(x, t)$. We write $s_\theta(x, t)$ for the neural network parameterized by $\theta \in \mathbb{R}^P$. In this explanatory subsection, the time-integrated losses on $(0, T)$ are written formally; when they diverge at zero, a positive lower cutoff is needed, as discussed below. This leads to the standard score-matching formulation [31]:

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

A motivation for this score-matching problem is provided by Proposition 4.1 below, which quantifies the relative-entropy error caused by replacing s with s_θ .

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.

The Gaussian convolution gives

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

where G_t is the Gaussian heat kernel. At each positive time, this identity gives the following formal reformulation of (4.1), up to a θ -independent term:

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

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, 31]:

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

Moreover, by performing the change of variables $z = (x - y_k)/(2t)$, one obtains a score-parameterized form of the DDPM loss [15]:

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

Here, each clean data sample y_k is perturbed to $x = 2tz + y_k$, where z has density $G_{(4t)}^{-1}$. The target $-z$ is the negative of the scaled noise $(x - y_k)/(2t)$. Equivalently, setting $\zeta = \sqrt{2t}z$ gives standard Gaussian noise with density $G_{1/2}$; the usual noise-prediction parameterization then carries the time weight $1/(2t)$. Thus changing that weight changes the objective.

Another classical reformulation is the implicit score-matching 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, this loss takes the form, up to a parameter-independent constant,

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

At each positive time, Gaussian decay justifies these integrations by parts for differentiable scores of at most linear growth with bounded first derivatives.

At each positive time, if there were no constraints on s_θ , the empirical score function (3.8) would minimize the corresponding denoising objective by satisfying its first-order optimality condition; see also [3, 14, 33]. By Theorem 3.7, for Lebesgue-almost every terminal point its deterministic trajectory converges to a training point. Restricting the score class can change this behavior, but a small empirical fitting error alone does not control the population error in Proposition 4.1.

4.2. Stability under score approximation. The score-matching error has a direct interpretation in the generative equation. The next result separates the terminal mismatch from the accumulated score error in continuous time.

Proposition 4.1. *Let $T > 0$, $\epsilon > 0$, and let $u_0 \in \mathcal{P}(\mathbb{R}^d)$ have compact support. Suppose that s_θ is smooth and globally Lipschitz in space, with at most linear growth, uniformly on every $[\tau, T]$, $0 < \tau < T$. Under these assumptions, the positive-time well-posedness argument of Proposition 2.1 applies with s replaced by s_θ . Let v_θ denote the resulting unique narrowly continuous weak probability solution of (1.6), with score s_θ and terminal probability density $v_{\theta, T}$ satisfying $\operatorname{KL}(u(T) \| v_{\theta, T}) < \infty$. Then, for $0 < t < T$,*

$$(4.6) \quad \operatorname{KL}(u(t) \| v_\theta(t)) \leq \operatorname{KL}(u(T) \| v_{\theta, T}) + \frac{(1 + \epsilon)^2}{4\epsilon} \int_t^T \int_{\mathbb{R}^d} |s - s_\theta|^2 u dx dr.$$

Proof. The proof is presented in Section 5.6. $\square$

Remark 4.2. For $\epsilon = 1$, (4.6) has the heat-flow form of the likelihood-weighted estimate in [28, Thm. 1]. Here we obtain it within the present PDE framework by entropy truncation. The reference law in (4.6) is always the heat flow $u(t)$. Thus, if $v_{\theta, T} = u(T)$, the terminal mismatch vanishes and the estimate measures solely the accumulated score error. For a different terminal initialization, however, $u(t)$ is generally not the exact-score generative solution issued from $v_{\theta, T}$.

The direction of the relative entropy is also important. Even if $v_\theta(t)$ converges narrowly to $v_{\theta, 0}$ as $t \rightarrow 0^+$ and the right-hand side remains bounded, the estimate controls $\operatorname{KL}(u_0 \| v_{\theta, 0})$. In particular, finite endpoint relative entropy implies $u_0 \ll v_{\theta, 0}$ and does not exclude additional generated mass outside $\operatorname{supp}(u_0)$. In particular, (4.6) does not imply preservation of the data support under score approximation.

4.3. Other regularization techniques. The Li–Yau bound $\operatorname{div} s \geq -d/(2t)$ becomes singular as $t \rightarrow 0^+$. The following strategies are motivated by the preceding analysis.

- (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 a time $0 < t_{\min} < T$. On $[t_{\min}, T]$, the score divergence is bounded below by $-d/(2t_{\min})$. This approach has been mentioned in [5, 22, 29]; see also [11].

Theorem 3.1 bounds the W_2 distance between the stopped law and its exact-score endpoint by a constant times $\sqrt{t_{\min}}$. Theorem 3.7 gives a corresponding empirical trajectory-wise

bound. These estimates quantify a stopping discrepancy, not an optimal stopping rule or a diversity guarantee.

- (2) *Regularization in the loss function.* On the same positive-time interval, 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:

$$(4.7) \quad \inf_{\theta \in \mathbb{R}^P} \frac{1}{N} \sum_{k=1}^N \int_{t_{\min}}^T \int_{\mathbb{R}^d} \left(\left| s_{\theta}(x, t) - \frac{y_k - x}{2t} \right|^2 + \lambda (\operatorname{div}_x s_{\theta}(x, t))^2 \right) G_t(x - y_k) dx dt,$$

where $\lambda \geq 0$ is a hyperparameter controlling the strength of the regularization.

Related regularization ideas appear in [3, 17], although (4.7) is not intended to reproduce those schemes. Our L^p analysis instead highlights the negative part of the score divergence as the quantity directly affecting stability. This provides a motivation for divergence-based regularization, while leaving questions of sample quality and optimal penalty design beyond the scope of the present work.

5. PROOFS OF MAIN RESULTS

Throughout this section, we denote by u the solution of the heat equation (1.1), and by v the solution of the backward FP equation (1.6).

5.1. Heat-flow estimates and positive-time well-posedness. We collect the estimates for the reference heat flow that are used throughout the proofs. They also verify the hypotheses of the classical well-posedness results applied below.

Lemma 5.1 (A priori estimates for the heat equation). *Let $u_0 \in \mathcal{P}(\mathbb{R}^d)$, $u(t) = G_t * u_0$, and $s = \nabla \log u$. Then u is strictly positive, u and s are smooth at positive times, and*

$$(5.1) \quad \operatorname{div} s(x, t) \geq -\frac{d}{2t}, \quad (x, t) \in Q.$$

If, in addition, $\operatorname{supp}(u_0) \subset B_R(0)$ for some $R > 0$, the following estimates hold, where $a_+ = \max\{a, 0\}$.

(i) For every $(x, t) \in Q$,

$$(5.2) \quad (4\pi t)^{-d/2} \exp\left(-\frac{(|x| + R)^2}{4t}\right) \leq u(x, t) \leq (4\pi t)^{-d/2} \exp\left(-\frac{(|x| - R)_+^2}{4t}\right).$$

(ii) For every $(x, t) \in Q$,

$$(5.3) \quad |s(x, t)| \leq \frac{|x| + R}{2t},$$

and

$$(5.4) \quad -\frac{1}{2t} I_d \preceq \nabla_x s(x, t) \preceq \left(-\frac{1}{2t} + \frac{R^2}{4t^2}\right) I_d.$$

Proof. Positivity and smoothness follow from the Gaussian convolution. For $(x, t) \in Q$, define the probability measure and its mean

$$(5.5) \quad d\pi_{x,t}(y) := \frac{G_t(x - y)}{u(x, t)} du_0(y), \quad m(x, t) := \int_{\mathbb{R}^d} y d\pi_{x,t}(y).$$

Differentiating the convolution gives

$$(5.6) \quad s(x, t) = \frac{m(x, t) - x}{2t}.$$

A second differentiation yields

$$\nabla_x s(x, t) = -\frac{I_d}{2t} + \frac{\Sigma(x, t)}{4t^2}, \quad \Sigma(x, t) = \int_{\mathbb{R}^d} (y - m(x, t))(y - m(x, t))^\top d\pi_{x,t}(y).$$

These differentiated integrals are finite even without moments of u_0 , since a polynomial times a Gaussian is bounded at every fixed positive time, locally uniformly in (x, t) . The covariance matrix is nonnegative. Taking the trace proves (5.1).

Assume now that $\text{supp}(u_0) \subset B_R(0)$.

Step 1 (Proof of Gaussian bounds). Since $\text{supp}(u_0) \subset B_R(0)$, 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

$$(|x| - R)_+ \leq |x - y| \leq |x| + R,$$

Therefore,

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

By the definition of G_t and the fact that $\int_{B_R(0)} du_0(y) = 1$, the estimate (5.2) follows.

Step 2 (Score and Hessian bounds).

The measure $\pi_{x,t}$ is supported in $B_R(0)$, so $|m(x, t)| \leq R$. Formula (5.6) gives (5.3). For any unit vector $e \in \mathbb{R}^d$, the covariance satisfies

$$0 \leq \langle \Sigma(x, t)e, e \rangle = \int_{\mathbb{R}^d} \langle e, y \rangle^2 d\pi_{x,t}(y) - \langle e, m(x, t) \rangle^2 \leq R^2.$$

Thus $0 \preceq \Sigma(x, t) \preceq R^2 I_d$. Substitution into the covariance formula above yields (5.4). $\square$

Proof of Proposition 2.1. Fix $0 < \tau < T$. By Lemma 5.1, the score $s = \nabla \log u$ is smooth and satisfies

$$(5.7) \quad |s(x, t)| \leq C_\tau(1 + |x|), \quad \|\nabla_x s(\cdot, t)\|_{L^\infty(\mathbb{R}^d)} \leq C_\tau, \quad \tau \leq t \leq T.$$

Thus the coefficients of (1.6) are regular on every positive-time strip, although the corresponding bounds may diverge as $\tau \rightarrow 0^+$.

For the application of the classical Cauchy theory, introduce the increasing time variable

$$r = T - t, \quad \rho(r) = v(T - r), \quad b(x, r) = (1 + \epsilon)s(x, T - r).$$

Then (1.6) becomes

$$(5.8) \quad \partial_r \rho = \epsilon \Delta \rho - \text{div}(b\rho), \quad \rho(0) = v_T, \quad 0 < r < T - \tau.$$

Step 1 (Existence and uniqueness). Suppose first that $\epsilon > 0$. The diffusion matrix $A = \epsilon I_d$ is constant and positive definite, while b is smooth and has at most linear growth. Hence [7, Thm. 6.6.4] gives a nonnegative weak solution of (5.8). Moreover, for $V(x) = 1 + |x|^2$ and

$$\mathcal{L}_r \varphi = \epsilon \Delta \varphi + \langle b(\cdot, r), \nabla \varphi \rangle,$$

we have

$$\mathcal{L}_r V(x) = 2\epsilon d + 2\langle b(x, r), x \rangle \leq C_\tau V(x).$$

By [7, Cor. 6.6.6], the solution conserves mass. We take its probability-valued representative; the continuity of this representative is specified below.

Moreover,

$$\frac{\|A\|_{\text{op}}}{1 + |x|^2} + \frac{|b(x, r)|}{1 + |x|} \leq C_\tau.$$

Here $\|A\|_{\text{op}}$ denotes the operator norm. The local coefficient conditions and the coefficient-integrability assumption of [7, Thm. 9.4.3] are therefore satisfied, and uniqueness follows in the weak probability class. The weak formulation gives continuity against compactly supported smooth tests. Conservation of mass then upgrades this representative to narrow continuity, including its prescribed initial trace. Returning to the original time variable gives

$$v \in C([\tau, T]; \mathcal{P}(\mathbb{R}^d)), \quad v(T) = v_T.$$

Since the diffusion matrix is nondegenerate and the coefficients are smooth on the strip, standard interior parabolic regularity and positivity imply that v has a smooth, strictly positive density for $\tau < t < T$; see [7, Sects. 6.3–6.4 and 8.1].

When $\epsilon = 0$, (5.8) reduces to a transport equation. By (5.7), b is globally Lipschitz in space. The characteristic representation therefore gives a unique narrowly continuous weak probability solution; see [7, Sect. 9.8(iv)]. This proves part (i) on $[\tau, T]$.

Step 2 (Wasserstein continuity). Assume now that $v_T \in \mathcal{P}_2(\mathbb{R}^d)$. The same Lyapunov estimate with $V(x) = 1 + |x|^2$ gives propagation of the second moment on every positive-time strip; see [7, Sect. 7.1]. In particular,

$$(5.9) \quad \sup_{t \in [\tau, T]} \int_{\mathbb{R}^d} |x|^2 dv(t)(x) \leq C_\tau \left(1 + \int_{\mathbb{R}^d} |x|^2 dv_T(x) \right).$$

Testing with $\varphi_L(x) = |x|^2 \chi(x/L)$, where $\chi \in C_c^\infty(\mathbb{R}^d)$ equals one near the origin, and removing the cutoff, using linear growth and the preceding moment bound, gives, for $\tau \leq t_1 < t_2 \leq T$,

$$\int_{\mathbb{R}^d} |x|^2 dv(t_2)(x) - \int_{\mathbb{R}^d} |x|^2 dv(t_1)(x) = \int_{t_1}^{t_2} \int_{\mathbb{R}^d} (-2\epsilon d - 2(1 + \epsilon) \langle s(x, r), x \rangle) dv(r)(x) dr.$$

Indeed, $|\nabla \varphi_L| \leq C(1 + |x|)$ and $|\Delta \varphi_L| \leq C$, uniformly in $L \geq 1$. The integrands are therefore dominated by $C_\tau(1 + |x|^2)$; (5.9) justifies the space–time passage to the limit. The identity shows that the second moment is absolutely continuous on $[\tau, T]$. Together with narrow continuity, the characterization of Wasserstein convergence [30, Thm. 6.9] yields

$$v \in C([\tau, T]; \mathcal{P}_2(\mathbb{R}^d), W_2).$$

Finally, uniqueness implies that the solutions obtained on different positive-time strips coincide on their overlaps. Since $\tau > 0$ is arbitrary, the preceding conclusions hold on $(0, T]$, proving both assertions of the proposition. $\square$

5.2. The Li–Yau bound and the energy estimate. We explain the role of the Li–Yau inequality in the energy estimate. For this calculation, let $1 < p < \infty$ and assume that v is a smooth, strictly positive solution of (1.6), with sufficient decay at infinity to justify the differentiations and integrations by parts below. We also assume that $v(t) \rightarrow v_T$ in L^p as $t \rightarrow T^-$. These additional assumptions are used only in this subsection. Appendix A proves Theorem 2.2 under the stated hypotheses.

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

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

By the convolution representation of u , the score function s is well-defined and belongs to $\mathcal{C}^\infty(Q)$. The one-sided divergence bound is (5.1), established in Lemma 5.1; it is the Euclidean Li–Yau inequality recalled in Section 2.2.

Step 2 (Energy identity). Multiply equation (1.6) by v^{p-1} and integrate over $\mathbb{R}^d$:

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

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:

$$\begin{aligned} \int_{\mathbb{R}^d} v^{p-1} \operatorname{div}(s v) \, dx &= - \int_{\mathbb{R}^d} \langle \nabla(v^{p-1}), v s \rangle \, dx \\ &= -(p-1) \int_{\mathbb{R}^d} v^{p-2} \langle \nabla v, v s \rangle \, dx \\ &= -(p-1) \int_{\mathbb{R}^d} v^{p-1} \langle \nabla v, s \rangle \, dx \\ &= -\frac{p-1}{p} \int_{\mathbb{R}^d} \langle \nabla(v^p), s \rangle \, dx = \frac{p-1}{p} \int_{\mathbb{R}^d} v^p \operatorname{div}(s) \, dx. \end{aligned}$$

Putting all terms together yields the energy identity:

$$(5.10) \quad \frac{1}{p} \frac{d}{dt} \|v\|_{L^p}^p - \epsilon(p-1) \int_{\mathbb{R}^d} v^{p-2} |\nabla v|^2 \, dx - (1 + \epsilon) \frac{p-1}{p} \int_{\mathbb{R}^d} v^p \operatorname{div}(s) \, dx = 0.$$

Step 3 (Energy estimate). Now, using the lower bound on $\operatorname{div}(s)$ from (5.1), we deduce from (5.10) the following differential inequality:

$$\begin{cases} \frac{d}{dt} \|v(t)\|_{L^p}^p \geq -\frac{d(1+\epsilon)(p-1)}{2t} \|v(t)\|_{L^p}^p, & \text{for a.e. } t \in (0, T), \\ \|v(T)\|_{L^p}^p = \|v_T\|_{L^p}^p. \end{cases}$$

Since $v > 0$, dividing by $\|v(t)\|_{L^p}^p$ yields

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

Consequently,

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

Integrating this differential inequality from t to T gives

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

Exponentiating both sides, we obtain

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

This gives (2.4) under the additional assumptions of this subsection. For a nonnegative probability solution, the case $p = 1$ is simply conservation of mass. A rigorous proof under the original hypotheses

of Theorem 2.2, without the auxiliary smoothness, strict positivity, and decay assumptions used here, is given in Appendix A.

Proof of Remark 2.3. By scaling Proposition 2.1, finite nonnegative terminal data define nonnegative weak solutions with conserved mass, unique in this class. Set

$$w_T = v_{1,T} - v_{2,T}, \quad M = \|w_T\|_{L^1}.$$

If $M = 0$, the conclusion follows from uniqueness. Otherwise, let q^+ and q^- be the nonnegative solutions with terminal data $(w_T)_+$ and $(w_T)_-$, respectively, and set $q = q^+ + q^-$.

The nonnegative solutions $v_1 + q^-$ and $v_2 + q^+$ have the same terminal datum, since

$$v_{1,T} + (w_T)_- = v_{2,T} + (w_T)_+ = \max\{v_{1,T}, v_{2,T}\}.$$

After normalization by their common mass, uniqueness in Proposition 2.1 gives $v_1 + q^- = v_2 + q^+$. Consequently, for every $0 < t \leq T$,

$$|v_1(t) - v_2(t)| = |q^+(t) - q^-(t)| \leq q(t) \quad \text{a.e. in } \mathbb{R}^d.$$

The function q/M is the probability solution with terminal density $|w_T|/M \in L^p(\mathbb{R}^d)$. Applying Theorem 2.2 therefore yields

$$\|v_1(t) - v_2(t)\|_{L^p} \leq \|q(t)\|_{L^p} \leq \left(\frac{T}{t}\right)^{\frac{d(1+\epsilon)(p-1)}{2p}} \|w_T\|_{L^p}.$$

This proves (2.5). $\square$

5.3. Proof of Theorem 3.1. We use the heat-flow estimates in Lemma 5.1 and the following two entropy lemmas. Their proofs are in Section 5.5; Appendix B provides the localization and terminal-passage arguments.

Lemma 5.2 (Terminal bound). *Under the hypotheses of Theorem 3.1, one has:*

$$\text{KL}(v_T \| u(T)) < \infty.$$

Moreover, if $|\log v_T(x)| \leq C(1 + |x|^q)$ for almost every $x \in \mathbb{R}^d$, for some $C, q > 0$, then

$$\text{KL}(u(T) \| v_T) < \infty.$$

Proof. The proof is presented in Section 5.5. $\square$

Lemma 5.3 (Relative-entropy contraction and dissipation). *Let v solve (1.6) under the hypotheses of Theorem 3.1. Write $H(t) = \text{KL}(v(t) \| u(t))$. Then, for $0 < t_1 \leq t_2 \leq T$,*

$$H(t_1) \leq H(t_2) \leq H(T) < \infty.$$

If $\epsilon > 0$, for every $0 < t_1 < t_2 \leq T$,

$$(5.11) \quad H(t_2) - H(t_1) = \epsilon \int_{t_1}^{t_2} \int_{\mathbb{R}^d} v \left| \nabla \log \frac{v}{u} \right|^2 dx dt.$$

In particular, H is absolutely continuous on every $[\tau, T]$ and

$$(5.12) \quad \frac{d}{dt} \text{KL}(v(t) \| u(t)) = \epsilon \int_{\mathbb{R}^d} v \left| \nabla \log \frac{v}{u} \right|^2 dx \quad \text{for a.e. } t \in (0, T).$$

If also $\text{KL}(u(T) \| v_T) < \infty$, the same statements hold with u and v interchanged; in particular,

$$(5.13) \quad \frac{d}{dt} \text{KL}(u(t) \| v(t)) = \epsilon \int_{\mathbb{R}^d} u \left| \nabla \log \frac{u}{v} \right|^2 dx \quad \text{for a.e. } t \in (0, T).$$

For $\epsilon = 0$, each finite relative entropy above is constant in time.

Proof. The proof is presented in Section 5.5; its localization and terminal-passage arguments are given in Appendix B. $\square$

Proof of Theorem 3.1. The solution is supplied by Proposition 2.1(ii). Recall that

$$K_T = \text{KL}(v_T \| u(T)), \quad C_T = 2\sqrt{2K_T + d \log 2} + \sqrt{\epsilon K_T}.$$

By Lemma 5.3,

$$(5.14) \quad \text{KL}(v(t) \| u(t)) \leq K_T, \quad 0 < t \leq T.$$

Step 1 (Estimate of the score along the generated law). Recall the probability measure $\pi_{x,t}$ defined in (5.5):

$$d\pi_{x,t}(y) = \frac{G_t(x-y)}{u(x,t)} du_0(y).$$

By the mean-shift identity (5.6),

$$s(x,t) = \int_{\mathbb{R}^d} \frac{y-x}{2t} d\pi_{x,t}(y).$$

For each fixed $t > 0$, the function $z \mapsto e^{t|z|^2/2}$ is convex. Hence, by Jensen's inequality,

$$\begin{aligned} e^{t|s(x,t)|^2/2} &= \exp\left(\frac{t}{2} \left| \int_{\mathbb{R}^d} \frac{y-x}{2t} d\pi_{x,t}(y) \right|^2\right) \\ &\leq \int_{\mathbb{R}^d} e^{|x-y|^2/(8t)} d\pi_{x,t}(y) = \int_{\mathbb{R}^d} e^{|x-y|^2/(8t)} \frac{G_t(x-y)}{u(x,t)} du_0(y). \end{aligned}$$

Multiplying by $u(x,t)$, integrating in x , and using Tonelli's theorem, we obtain

$$\begin{aligned} \int_{\mathbb{R}^d} e^{t|s(x,t)|^2/2} u(x,t) dx &\leq \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{|x-y|^2/(8t)} G_t(x-y) dx du_0(y) \\ &= (4\pi t)^{-d/2} \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{-|x-y|^2/(8t)} dx du_0(y) \\ &= 2^{d/2}. \end{aligned}$$

We transfer this estimate from $u(t)$ to $v(t)$ using the Donsker–Varadhan variational inequality [12]. For

$$\phi_n(x) := \min \left\{ \frac{t}{2} |s(x,t)|^2, n \right\},$$

we have

$$\int_{\mathbb{R}^d} \phi_n(x) v(x,t) dx \leq \text{KL}(v(t) \| u(t)) + \log \int_{\mathbb{R}^d} e^{\phi_n(x)} u(x,t) dx.$$

Using (5.14) and the preceding exponential estimate,

$$\int_{\mathbb{R}^d} \phi_n(x) v(x,t) dx \leq K_T + \frac{d}{2} \log 2.$$

Letting $n \rightarrow \infty$ by monotone convergence gives

$$(5.15) \quad \int_{\mathbb{R}^d} |s(x,t)|^2 v(x,t) dx \leq \frac{2K_T + d \log 2}{t} =: \frac{C_0^2}{t}, \quad C_0 := \sqrt{2K_T + d \log 2}.$$

Step 2 (W_2 convergence and rate). We rewrite the backward FP equation as a continuity equation. Suppose first that $\epsilon > 0$. Since v is smooth and strictly positive on $\dot{Q}$,

$$\nabla v = v \nabla \log v = v \left(s + \nabla \log \frac{v}{u} \right).$$

Therefore

$$\begin{aligned}\epsilon \nabla v - (1 + \epsilon)sv &= \epsilon v \left(s + \nabla \log \frac{v}{u} \right) - (1 + \epsilon)sv \\ &= v \left(-s + \epsilon \nabla \log \frac{v}{u} \right).\end{aligned}$$

When $\epsilon = 0$, the flux is simply $-sv$. Thus, in both cases,

$$(5.16) \quad \partial_t v + \operatorname{div}(b_v v) = 0, \quad b_v(\cdot, t) = \begin{cases} -s + \epsilon \nabla \log(v/u), & \epsilon > 0, \\ -s, & \epsilon = 0, \end{cases}$$

in the sense of distributions on $\mathring{Q}$.

Fix $0 < t_1 < t_2 \leq T$. If $\epsilon > 0$, set

$$I(t) := \int_{\mathbb{R}^d} v(x, t) \left| \nabla \log \frac{v(x, t)}{u(x, t)} \right|^2 dx.$$

By the entropy dissipation identity (5.11),

$$\epsilon \int_{t_1}^{t_2} I(t) dt = \operatorname{KL}(v(t_2) \| u(t_2)) - \operatorname{KL}(v(t_1) \| u(t_1)) \leq K_T.$$

By (5.15) and the entropy dissipation estimate, the velocity in (5.16) satisfies

$$r \mapsto \|b_v(\cdot, r)\|_{L^2(v(r))} \in L^1((t_1, t_2)).$$

Indeed, for $\epsilon > 0$,

$$\begin{aligned}\int_{t_1}^{t_2} \|b_v(\cdot, r)\|_{L^2(v(r))} dr &\leq \int_{t_1}^{t_2} \|s(\cdot, r)\|_{L^2(v(r))} dr + \epsilon \int_{t_1}^{t_2} \sqrt{I(r)} dr \\ &\leq 2C_0(\sqrt{t_2} - \sqrt{t_1}) + \sqrt{\epsilon K_T(t_2 - t_1)} < \infty.\end{aligned}$$

When $\epsilon = 0$, only the first term is present.

Since v is W_2 -continuous on $[t_1, t_2]$ by Proposition 2.1(ii), the converse implication in [1, Thm. 8.3.1], applied to the continuity equation (5.16), yields metric absolute continuity and

$$(5.17) \quad W_2(v(t_1), v(t_2)) \leq \int_{t_1}^{t_2} \|b_v(\cdot, r)\|_{L^2(v(r))} dr \leq 2C_0(\sqrt{t_2} - \sqrt{t_1}) + \sqrt{\epsilon K_T(t_2 - t_1)}.$$

The right-hand side tends to zero as $t_1, t_2 \rightarrow 0^+$. Hence $v(t)$ is Cauchy in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. By completeness, there exists a unique $v_0^* \in \mathcal{P}_2(\mathbb{R}^d)$ such that

$$v(t) \longrightarrow v_0^* \quad \text{in } W_2 \quad \text{as } t \rightarrow 0^+.$$

Fixing $0 < t \leq T$, setting $t_2 = t$, and letting $t_1 \rightarrow 0^+$ in (5.17), we obtain

$$W_2(v(t), v_0^*) \leq (2C_0 + \sqrt{\epsilon K_T}) \sqrt{t} = C_T \sqrt{t}.$$

This proves (3.4) and gives a W_2 -continuous extension to $[0, T]$ by setting $v(0) = v_0^*$.

Step 3 (Identification of the endpoint). The convolution representation gives the coupling bound

$$W_2(u(t), u_0)^2 \leq \iint_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 G_t(x - y) dx du_0(y) = 2dt.$$

Thus $u(t) \rightarrow u_0$ in W_2 as $t \rightarrow 0^+$. Together with Step 2, this implies the narrow convergences

$$v(t) \rightharpoonup v_0^*, \quad u(t) \rightharpoonup u_0.$$

By the joint lower semicontinuity of relative entropy,

$$\operatorname{KL}(v_0^* \| u_0) \leq \liminf_{t \rightarrow 0^+} \operatorname{KL}(v(t) \| u(t)) \leq K_T < \infty.$$

Therefore

$$v_0^* \ll u_0, \quad \text{supp}(v_0^*) \subset \text{supp}(u_0) = S.$$

Assume now, in addition, that

$$|\log v_T(x)| \leq C(1 + |x|^q) \quad \text{for almost every } x \in \mathbb{R}^d.$$

By Lemma 5.2, $\text{KL}(u(T)\|v_T) < \infty$, and Lemma 5.3 yields

$$\text{KL}(u(t)\|v(t)) \leq \text{KL}(u(T)\|v_T), \quad 0 < t \leq T.$$

Applying joint lower semicontinuity once more gives

$$\text{KL}(u_0\|v_0^*) \leq \liminf_{t \rightarrow 0^+} \text{KL}(u(t)\|v(t)) < \infty.$$

Hence $u_0 \ll v_0^*$. Together with the previous inclusion,

$$\text{supp}(v_0^*) = \text{supp}(u_0) = S.$$

Step 4 (Concentration on the data support). Since $\text{supp}(v_0^*) \subset S$, every coupling π of $v(t)$ and v_0^* satisfies $\pi(\mathbb{R}^d \times S) = 1$. Therefore

$$\text{dist}(x, S) \leq |x - y| \quad \pi\text{-almost everywhere,}$$

and hence

$$\int_{\mathbb{R}^d} \text{dist}(x, S)^2 dv(t)(x) \leq \iint_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\pi(x, y).$$

Taking the infimum over all couplings yields

$$\left(\int_{\mathbb{R}^d} \text{dist}(x, S)^2 dv(t)(x) \right)^{1/2} \leq W_2(v(t), v_0^*) \leq C_T \sqrt{t},$$

which proves (3.5). By Markov's inequality, for every $r > 0$,

$$v(t)\{x : \text{dist}(x, S) \geq r\} \leq \min \left\{ 1, \frac{C_T^2 t}{r^2} \right\},$$

which is (3.6).

Finally, let U be an open neighborhood of S . Since S is compact, there exists $r > 0$ such that

$$\{x : \text{dist}(x, S) < r\} \subset U.$$

Consequently,

$$v(t)(U^c) \leq \frac{C_T^2 t}{r^2} \longrightarrow 0, \quad t \rightarrow 0^+,$$

and therefore

$$v(t)(U) \longrightarrow 1.$$

For $\epsilon > 0$, the FP–SDE correspondence gives $v(t) = \text{law}(X_t)$; for $\epsilon = 0$ the same identity follows from the characteristic representation. The conclusion follows. $\square$

5.4. **Proof of Theorem 3.7.** For convenience, we recall the generation ODE (3.9):

$$(5.18) \quad \begin{cases} \frac{dX_t}{dt} = -s(X_t, t), & t \in (0, T), \\ X_T = x_T \in \mathbb{R}^d. \end{cases}$$

The proof combines a finite-mean-length argument with quantitative estimates from the Gaussian mean-shift formula. Its three steps correspond to assertions (i)–(iii) of the theorem.

Step 1 (Almost-everywhere convergence). We use the reference terminal density $u(T)$ to control the total length of the deterministic trajectories. The heat flow itself solves (1.6); uniqueness and the characteristic representation therefore give

$$u(t) = (\Phi_{t,T})_{\#} u(T),$$

where $\Phi_{t,T}(x)$ is the trajectory at time t starting from x at time T .

The mean-shift formula (5.6) expresses $s(x, t)$ as the average of $(y - x)/(2t)$ under $\pi_{x,t}$. Jensen's inequality, followed by integration in x , gives

$$\begin{aligned} \int_{\mathbb{R}^d} |s(x, t)|^2 u(x, t) dx &\leq \frac{1}{4t^2} \iint_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 G_t(x - y) dx du_0(y) \\ &= \frac{d}{2t}. \end{aligned}$$

Using the push-forward identity and Cauchy–Schwarz, we find

$$\begin{aligned} \int_{\mathbb{R}^d} \int_0^T |s(\Phi_{t,T}(x), t)| dt u(x, T) dx &= \int_0^T \int_{\mathbb{R}^d} |s(y, t)| u(y, t) dy dt \\ &\leq \int_0^T \sqrt{\frac{d}{2t}} dt = \sqrt{2dT}. \end{aligned}$$

Thus for $u(T)$ -almost every x the trajectory has finite length:

$$\int_0^T |s(\Phi_{t,T}(x), t)| dt < \infty.$$

For such an x , the ODE implies, for $0 < a < b \leq T$,

$$|\Phi_{b,T}(x) - \Phi_{a,T}(x)| \leq \int_a^b |s(\Phi_{t,T}(x), t)| dt.$$

The right-hand side tends to zero as $a, b \rightarrow 0^+$, so the trajectory has a finite limit, denoted by $\Phi_{0,T}(x)$. The finite-length set is Borel. Define $\Phi_{0,T}$ to be a fixed point of S on its complement. The resulting map is Borel measurable, as follows by taking the pointwise limit along any fixed sequence $t_k \rightarrow 0^+$ on the finite-length set.

To locate that limit, let φ be bounded and continuous. Dominated convergence and the heat-flow limit give

$$\begin{aligned} \int_{\mathbb{R}^d} \varphi(\Phi_{0,T}(x)) u(x, T) dx &= \lim_{t \rightarrow 0^+} \int_{\mathbb{R}^d} \varphi(\Phi_{t,T}(x)) u(x, T) dx \\ &= \lim_{t \rightarrow 0^+} \int_{\mathbb{R}^d} \varphi(y) u(y, t) dy = \int_{\mathbb{R}^d} \varphi du_0. \end{aligned}$$

Therefore $(\Phi_{0,T})_{\#} u(T) = u_0$, and $\Phi_{0,T}(x) \in S$ for $u(T)$ -almost every x . Since $u(x, T) > 0$ everywhere, $u(T)$ and Lebesgue measure have the same null sets. This proves assertion (i).

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

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

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

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

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

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

Along any trajectory of (5.18),

$$(5.19) \quad \frac{d}{dt} V(X_t) = \langle \nabla V(X_t), \dot{X}_t \rangle = \langle X_t - p(X_t), -\nabla \log u(X_t, t) \rangle.$$

By the Gaussian mean-shift formula (5.6),

$$s(x, t) = \frac{m(x, t) - x}{2t},$$

where $m(x, t)$ is the barycenter in (5.5). Thus, using (5.19), we have

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

Since $m(x, t)$ is a barycenter with positive weights supported on $\text{supp}(u_0)$, 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:

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

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

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

Therefore,

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

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

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

The desired estimate (3.11) follows immediately from the definition of V .

Step 3 (Convergence rate for finite support). Let

$$u_0 = \sum_{k=1}^N w_k \delta_{y_k}, \quad w_k = u_0(\{y_k\}) > 0, \quad \sum_{k=1}^N w_k = 1.$$

When $N = 1$, the score formula gives $X_t - y_1 = \sqrt{t/T}(x_T - y_1)$, so the result is immediate. Assume $N \geq 2$ and fix a terminal point for which Step 1 applies. Since the sites are distinct and the support is finite, the full limit is a site y_i . We use this convergence only to enter the corresponding Voronoi core. Define

$$(5.20) \quad \gamma := \frac{1}{2} \min_{j \neq i} |y_i - y_j|^2 > 0,$$

and introduce the corresponding γ -Voronoi core of y_i , illustrated in Figure 4:

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

We use the following two claims, proved below:

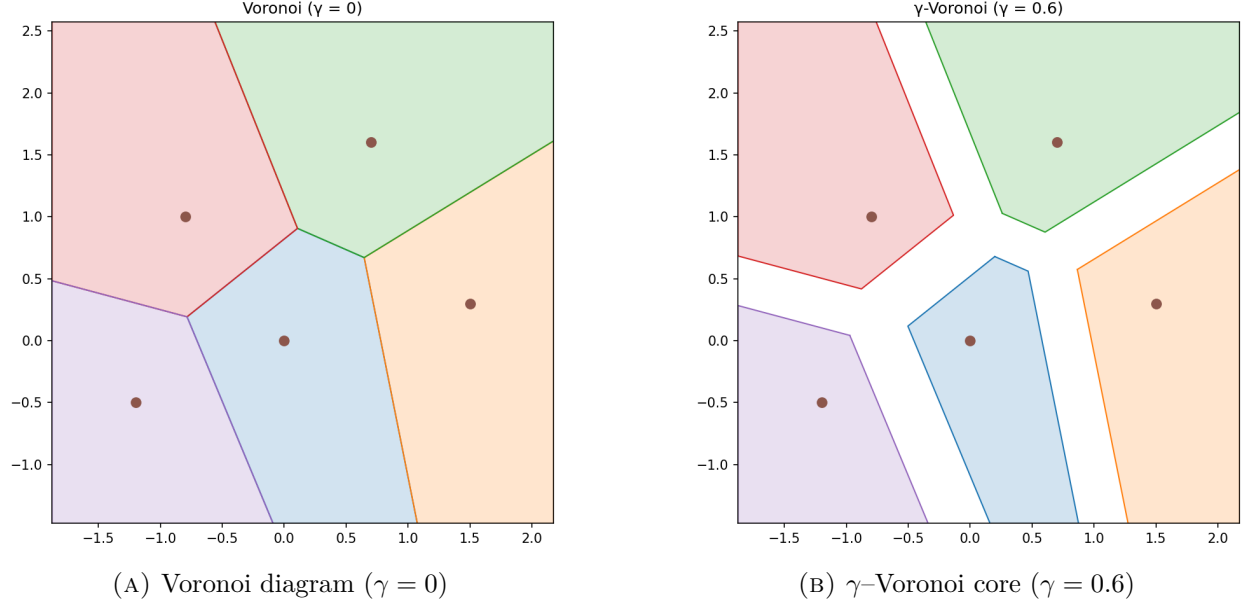


FIGURE 4. Voronoi diagram and γ -Voronoi cores for five points in $\mathbb{R}^2$. When $\gamma = 0$, we recover the classical Voronoi diagram.

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

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

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

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

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

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

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

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

Introduce the similarity time variable $\tau := -\log t$ (so that $d\tau/dt = -1/t$) and define $\rho(\tau) := |Z_{t(\tau)}|$. The function ρ is locally Lipschitz. The inequality below holds almost everywhere, also at its zeros (where its derivative is zero almost everywhere). Using the chain rule,

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

Substituting the inequality above, we obtain

$$\frac{d\rho}{d\tau} \leq -\frac{1}{2}\rho(\tau) + \frac{C_i}{2} e^{-\frac{\gamma}{4}e^\tau}, \quad \tau \geq \tau_* := -\log t_*.$$

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

$$\frac{d}{d\tau} \left(e^{\frac{\tau}{2}} \rho(\tau) \right) \leq \frac{C_i}{2} e^{\frac{\tau}{2}} e^{-\frac{\gamma}{4}e^\tau}.$$

Integrating this differential inequality from τ_* to $\tau \geq \tau_*$ yields

$$e^{\frac{\tau}{2}} \rho(\tau) - e^{\frac{\tau_*}{2}} \rho(\tau_*) \leq \frac{C_i}{2} \int_{\tau_*}^{\tau} e^{\frac{\sigma}{2}} e^{-\frac{\gamma}{4}e^\sigma} d\sigma.$$

Thus,

$$\rho(\tau) \leq e^{-\frac{\tau}{2}} e^{\frac{\tau_*}{2}} \rho(\tau_*) + \frac{C_i}{2} e^{-\frac{\tau}{2}} \int_{\tau_*}^{\tau} e^{\frac{\sigma}{2}} e^{-\frac{\gamma}{4}e^\sigma} d\sigma.$$

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

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

The integral term is bounded by

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

where the last equality follows by a change of variables and the Gamma integral. Therefore,

$$|Z_t| \leq \left((t_*)^{-1/2} |Z_{t_*}| + C_i \sqrt{\frac{\pi}{\gamma}} \right) t^{1/2}, \quad \forall t \in (0, t_*).$$

This is the required square-root rate near zero. Here, the constants t_* , C_i , and γ also depend on the data sites and weights. Since Z_t is continuous, the desired convergence rate (3.12) holds over the entire interval $(0, T]$ upon choosing a sufficiently large constant C .

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

$$(5.21) \quad \psi_j(x) := |x - y_j|^2 - |x - y_i|^2, \quad \forall j \neq i.$$

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 at least w_i , we obtain

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

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 $t \in (0, \tau_0)$,

$$(5.22) \quad C_i e^{-\frac{\gamma}{4t}} \leq \frac{1}{8} d_{ij} \quad \forall j \neq i.$$

By Step 1, $X_t \rightarrow y_i$, and $y_i \in \text{int } V_i(\gamma)$. We may therefore choose $0 < t_* < \tau_0$ such that $X_{t_*} \in \text{int } V_i(\gamma)$.

Suppose that X_t leaves the core when time decreases from t_* . Let $\tau_1 \in (0, t_*)$ be the lower endpoint of the connected component of $\{t : X_t \in \text{int } V_i(\gamma)\}$ containing t_* . By continuity, $X_{\tau_1} \in \partial V_i(\gamma)$. For an active index $j \neq i$, we have

$$(5.23) \quad \psi_j(X_{\tau_1}) = \gamma, \quad \left. \frac{d}{dt} \psi_j(X_t) \right|_{t=\tau_1} \geq 0,$$

since $\psi_j(X_t) > \gamma$ immediately to the right of τ_1 .

Differentiating along the trajectory gives

$$\left. \frac{d}{dt} \psi_j(X_t) \right|_{t=\tau_1} = \frac{1}{\tau_1} \langle y_i - y_j, X_{\tau_1} - m(X_{\tau_1}, \tau_1) \rangle.$$

Decompose the scalar product:

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

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

$$(5.24) \quad |\langle y_i - y_j, y_i - m(X_{\tau_1}, \tau_1) \rangle| \leq d_{ij} C_i e^{-\frac{\gamma}{4\tau_1}} \leq \frac{1}{8} d_{ij}^2,$$

where the last inequality uses (5.22) and $\tau_1 < t_* < \tau_0$.

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

$$(5.25) \quad \frac{1}{2}(\psi_j(X_{\tau_1}) - d_{ij}^2) = \frac{1}{2}(\gamma - d_{ij}^2) \leq -\frac{1}{4} d_{ij}^2,$$

where the final bound follows from (5.20).

Combining (5.24) and (5.25) gives

$$\left. \frac{d}{dt} \psi_j(X_t) \right|_{t=\tau_1} \leq \frac{1}{\tau_1} \left(\frac{1}{8} d_{ij}^2 - \frac{1}{4} d_{ij}^2 \right) = -\frac{1}{8\tau_1} d_{ij}^2 < 0.$$

This contradicts the nonnegative derivative in (5.23). Hence, $X_t \in V_i(\gamma)$ for all $t < t_*$. $\square$

5.5. Proofs of the entropy lemmas. We first establish the terminal entropy bounds. We then prove entropy contraction and present the dissipation calculation under additional integrability and decay assumptions. The justification of this calculation under the stated hypotheses, including the terminal passage, is given in Appendix B.

Proof of Lemma 5.2. The Gaussian bounds in Lemma 5.1 imply

$$|\log u(x, T)| \leq C(1 + |x|^2), \quad x \in \mathbb{R}^d.$$

Since v_T has finite absolute entropy and second moment,

$$\int_{\mathbb{R}^d} v_T \left| \log \frac{v_T}{u(T)} \right| dx \leq \int_{\mathbb{R}^d} v_T |\log v_T| dx + C \int_{\mathbb{R}^d} (1 + |x|^2) v_T dx < \infty.$$

Thus $\text{KL}(v_T \| u(T)) < \infty$.

For the reverse divergence, the Gaussian upper bound shows that $u(T)$ has moments of every order. Under the additional assumption $|\log v_T(x)| \leq C(1 + |x|^q)$, we therefore have

$$\int_{\mathbb{R}^d} u(x, T) \left| \log \frac{u(x, T)}{v_T(x)} \right| dx \leq C' \int_{\mathbb{R}^d} (1 + |x|^2 + |x|^q) u(x, T) dx < \infty.$$

This proves $\text{KL}(u(T) \| v_T) < \infty$. $\square$

Proof of Lemma 5.3. Write $H(t) = \text{KL}(v(t) \| u(t))$. We first establish contraction under the stated hypotheses, and then explain the dissipation identity.

Step 1 (Entropy contraction).

On every interval $[t_1, t_2] \subset (0, T]$, the two laws evolve under the same mass-preserving Markov transition operator, supplied by the well-posedness theory of Section 5.1. The output density ratio is a conditional average of the input ratio. Jensen's inequality applied to $z \mapsto z \log z$ therefore gives

$$H(t_1) \leq H(t_2), \quad 0 < t_1 \leq t_2 \leq T.$$

In particular,

$$\text{KL}(v(t) \| u(t)) \leq \text{KL}(v_T \| u(T)) < \infty,$$

where finiteness follows from Lemma 5.2.

If $\text{KL}(u(T) \| v_T) < \infty$, the same argument with the two laws interchanged gives

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

The polynomial bound in Lemma 5.2 is a sufficient condition for this terminal finiteness.

When $\epsilon = 0$, the transition is induced by an invertible characteristic map. The Jacobian factors cancel in the density ratio, so a change of variables gives equality:

$$H(t_1) = H(t_2).$$

The same conclusion holds for the reverse relative entropy whenever it is finite. This proves the deterministic case.

Step 2 (Entropy dissipation identity).

Assume now that $\epsilon > 0$. Interior smoothness and positivity follow from Proposition 2.1. For the calculation below, we additionally assume sufficient integrability and decay of the displayed terms to justify differentiation under the integral and integrations by parts on compact time intervals contained in $(0, T)$. These auxiliary assumptions are used only to exhibit the diffusion–transport cancellation; they are not imposed in Lemma 5.3.

The calculation is the FP analogue of the heat-flow identity in [18, Prop. 1.6]. We derive it directly for the present equation.

Set

$$L(x, t) := \log \frac{v(x, t)}{u(x, t)}, \quad H(t) = \int_{\mathbb{R}^d} v(x, t) L(x, t) dx.$$

Differentiating in time and using $\partial_t L = (\partial_t v)/v - (\partial_t u)/u$, we obtain

$$H'(t) = \int_{\mathbb{R}^d} \left((L + 1) \partial_t v - \frac{v}{u} \partial_t u \right) dx.$$

Both densities satisfy the same backward FP equation:

$$\partial_t v = -\epsilon \Delta v + (1 + \epsilon) \text{div}(sv), \quad \partial_t u = -\epsilon \Delta u + (1 + \epsilon) \text{div}(su).$$

Consequently,

$$\begin{aligned} H'(t) &= -\epsilon \int_{\mathbb{R}^d} (L+1) \Delta v \, dx + \epsilon \int_{\mathbb{R}^d} \frac{v}{u} \Delta u \, dx \\ &\quad + (1+\epsilon) \int_{\mathbb{R}^d} (L+1) \operatorname{div}(sv) \, dx - (1+\epsilon) \int_{\mathbb{R}^d} \frac{v}{u} \operatorname{div}(su) \, dx. \end{aligned}$$

We treat the diffusion and transport contributions separately.

Diffusion contribution. Integration by parts gives

$$\begin{aligned} \mathcal{D} &:= -\epsilon \int_{\mathbb{R}^d} (L+1) \Delta v \, dx + \epsilon \int_{\mathbb{R}^d} \frac{v}{u} \Delta u \, dx \\ &= \epsilon \int_{\mathbb{R}^d} \langle \nabla L, \nabla v \rangle \, dx - \epsilon \int_{\mathbb{R}^d} \left\langle \nabla \left(\frac{v}{u} \right), \nabla u \right\rangle \, dx \\ &= \epsilon \int_{\mathbb{R}^d} \left\langle \nabla v - \frac{v}{u} \nabla u, \nabla L \right\rangle \, dx \\ &= \epsilon \int_{\mathbb{R}^d} v |\nabla L|^2 \, dx, \end{aligned}$$

where we used $\nabla(v/u) = (v/u)\nabla L$ and $\nabla v - (v/u)\nabla u = v\nabla L$.

Transport contribution. The same integrations by parts yield

$$\begin{aligned} \mathcal{T} &:= (1+\epsilon) \int_{\mathbb{R}^d} (L+1) \operatorname{div}(sv) \, dx - (1+\epsilon) \int_{\mathbb{R}^d} \frac{v}{u} \operatorname{div}(su) \, dx \\ &= -(1+\epsilon) \int_{\mathbb{R}^d} v \langle s, \nabla L \rangle \, dx + (1+\epsilon) \int_{\mathbb{R}^d} u \left\langle s, \nabla \left(\frac{v}{u} \right) \right\rangle \, dx \\ &= 0. \end{aligned}$$

Thus the common transport field does not contribute to the entropy derivative.

Combining the two contributions, we conclude that

$$H'(t) = \epsilon \int_{\mathbb{R}^d} v(x, t) \left| \nabla \log \frac{v(x, t)}{u(x, t)} \right|^2 \, dx.$$

Integrating on an interior time interval gives

$$H(t_2) - H(t_1) = \epsilon \int_{t_1}^{t_2} \int_{\mathbb{R}^d} v \left| \nabla \log \frac{v}{u} \right|^2 \, dx \, dt, \quad 0 < t_1 < t_2 < T.$$

Interchanging u and v gives the corresponding reverse-entropy identity.

This establishes the dissipation identities under the auxiliary assumptions of Step 2. Appendix B proves the integrated identity under the original hypotheses, using convex entropy truncations and spatial cutoffs. The L^1 terminal traces permit the passage to $t_2 = T$ for each truncated entropy, after which monotone convergence removes the truncation. This gives (5.11), absolute continuity of H on every $[\tau, T]$, and the almost-everywhere identities (5.12)–(5.13). $\square$

5.6. Proof of Proposition 4.1.

Proof of Proposition 4.1. We first derive the estimate under the same auxiliary integrability and decay assumptions used in the entropy calculation above. The justification under the hypotheses of the proposition is given in Appendix B.

Write

$$g = v_\theta, \quad e = s_\theta - s, \quad L = \log \frac{u}{g}, \quad H_\theta(t) = \operatorname{KL}(u(t) \| g(t)).$$

The equation for g takes the form

$$\partial_t g = -\epsilon \Delta g + (1+\epsilon) \operatorname{div}(sg) + (1+\epsilon) \operatorname{div}(eg).$$

The diffusion calculation in the proof of Lemma 5.3 is unchanged, and the transport terms involving the common field s cancel. Only the additional drift e remains:

$$\begin{aligned} H'_\theta(t) &= \epsilon \int_{\mathbb{R}^d} u |\nabla L|^2 dx - (1 + \epsilon) \int_{\mathbb{R}^d} \frac{u}{g} \operatorname{div}(eg) dx \\ &= \epsilon \int_{\mathbb{R}^d} u |\nabla L|^2 dx + (1 + \epsilon) \int_{\mathbb{R}^d} g \left\langle e, \nabla \left(\frac{u}{g} \right) \right\rangle dx \\ &= \epsilon \int_{\mathbb{R}^d} u |\nabla L|^2 dx + (1 + \epsilon) \int_{\mathbb{R}^d} u \langle e, \nabla L \rangle dx. \end{aligned}$$

Since $\epsilon > 0$, completing the square gives

$$\epsilon |\nabla L|^2 + (1 + \epsilon) \langle e, \nabla L \rangle = \epsilon \left| \nabla L + \frac{1 + \epsilon}{2\epsilon} e \right|^2 - \frac{(1 + \epsilon)^2}{4\epsilon} |e|^2.$$

Therefore,

$$H'_\theta(t) \geq -\frac{(1 + \epsilon)^2}{4\epsilon} \int_{\mathbb{R}^d} u(x, t) |e(x, t)|^2 dx.$$

Integrating from t to $t_2 < T$, we obtain

$$H_\theta(t) \leq H_\theta(t_2) + \frac{(1 + \epsilon)^2}{4\epsilon} \int_t^{t_2} \int_{\mathbb{R}^d} u(x, r) |e(x, r)|^2 dx dr.$$

Appendix B establishes this inequality under the original assumptions and permits the passage $t_2 \rightarrow T^-$. Substituting $e = s_\theta - s$ then gives (4.6). It uses the same entropy truncations and spatial cutoffs as for Lemma 5.3, with an additional flux arising from the score error. That flux is controlled by the linear growth of e and the masses of u and g . $\square$

6. CONCLUSIONS AND PERSPECTIVES

In this paper, we have studied the stability and the singular endpoint of exact-score generative dynamics. The L^p estimate controls individual probability densities and differences of probability solutions, while relative-entropy dissipation gives a quantitative W_2 limit for both deterministic and stochastic generation. At the trajectory level, finite mean length yields almost-everywhere convergence to the data support. For empirical data, the Voronoi estimate provides the corresponding square-root rate.

The explicit case $\epsilon = 1$ further separates the two roles of the reference score and the terminal law: the former fixes the support, and the latter changes the endpoint weights. The expansion through order T^{-2} describes this change for Gaussian initialization and explains the cancellation obtained by matching the barycenter. The score-error estimate then relates stability to the loss used in training, without requiring the perturbed endpoint to have the original support.

Several questions remain. For non-empirical supports, trajectory-wise rates require information beyond the law-level estimate proved here. The structure of exceptional deterministic initializations and the small-time fluctuations of stochastic trajectories also deserve further study. On the approximation side, a quantitative comparison with the unknown population law must account for the learned score and the numerical solver. The present estimates provide complementary controls on initialization, score perturbation, and positive-time stopping that may serve as ingredients in such an analysis.

APPENDIX A. LOCALIZATION AND APPROXIMATION FOR THE ENERGY ESTIMATE

We justify the energy calculation in Section 5.2 under the hypotheses of Theorem 2.2. We first approximate the terminal density by compactly supported Lipschitz probability densities and establish the energy estimate for the corresponding solutions using spatial cutoffs. Passing to the limit then gives the estimate and the strong L^p terminal trace for the original datum.

Proof of Theorem 2.2. Let v_T be a probability density in $L^p(\mathbb{R}^d)$ and fix $0 < \tau < T$. By Lemma 5.1, the score s is smooth, globally Lipschitz in space, and of at most linear growth, uniformly on $[\tau, T]$. Moreover,

$$\operatorname{div} s(x, t) \geq -\frac{d}{2t}.$$

Step 1 (Approximation of the terminal density). Let $\operatorname{Lip}_c(\mathbb{R}^d)$ denote the space of compactly supported Lipschitz functions. Choose nonnegative $v_T^n \in \operatorname{Lip}_c(\mathbb{R}^d)$ of mass one such that

$$v_T^n \longrightarrow v_T \quad \text{in } L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d).$$

We first establish the estimate for nonnegative terminal data in $\operatorname{Lip}_c(\mathbb{R}^d)$, a class containing both v_T^n and $|v_T^n - v_T^m|$.

Step 2 (Spatial localization and terminal traces). Let $w_T \in \operatorname{Lip}_c(\mathbb{R}^d)$ be nonnegative. The zero datum is immediate. For $\epsilon > 0$, the construction in [19, Prop. 5], applied after the time change in (5.8), gives a nonnegative solution

$$w \in L^\infty([\tau, T]; L^1(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)).$$

The coefficient hypotheses follow from the constant diffusion matrix and the preceding bounds on s and its spatial derivative. For $\epsilon = 0$, the characteristic formula gives the same bounded-integrable solution class.

Choose a nonnegative radial nonincreasing $\chi \in C_c^\infty(\mathbb{R}^d)$, equal to one on B_1 and zero outside B_2 , and set

$$\chi_L(x) = \chi(x/L), \quad A_L = \{x \in \mathbb{R}^d : L \leq |x| \leq 2L\}, \quad L \geq 1.$$

Then $\chi_L \uparrow 1$, $|\nabla \chi_L| \leq C/L$, and $|\Delta \chi_L| \leq C/L^2$, with both derivatives supported in A_L . Testing the equation with χ_L and using its weak terminal trace gives

$$\left| \int_{\mathbb{R}^d} \chi_L w(t) dx - \int_{\mathbb{R}^d} \chi_L w_T dx \right| \leq C_{\tau, \epsilon} \int_t^T \int_{A_L} w(x, r) dx dr.$$

Linear growth of s makes the constant independent of L . The right-hand side tends to zero as $L \rightarrow \infty$, proving conservation of mass.

Local continuity, including at T , follows from parabolic regularity for the continuous terminal datum when $\epsilon > 0$, and from characteristics when $\epsilon = 0$. Nonnegativity and conservation of mass then give strong L^1 continuity by Scheffé's lemma. Interpolation yields

$$w \in C([\tau, T]; L^q(\mathbb{R}^d)), \quad 1 \leq q < \infty.$$

By uniqueness after normalization, $w(t) = \mathcal{S}_{t,T} w_T$, where $\mathcal{S}_{t,T}$ is the probability solution operator of Proposition 2.1, extended to finite nonnegative data by scaling, with $\mathcal{S}_{t,T} 0 = 0$. This operator is positive and additive.

Step 3 (Energy estimate). Assume first that $p > 1$ and define

$$\beta_\delta(z) = \frac{(z^2 + \delta^2)^{p/2} - \delta^p}{p}, \quad A_\delta(z) = z\beta'_\delta(z) - \beta_\delta(z), \quad \delta > 0.$$

Then $\beta''_\delta \geq 0$ and $A_\delta \geq 0$. Set $E_{L,\delta}(r) = \int_{\mathbb{R}^d} \chi_L \beta_\delta(w(r)) dx$. Multiplying the equation by $\chi_L \beta'_\delta(w)$, integrating by parts in space, and integrating from t to $t_2 < T$ gives

$$\begin{aligned} (A.1) \quad & E_{L,\delta}(t) + \epsilon \int_t^{t_2} \int_{\mathbb{R}^d} \chi_L \beta''_\delta(w) |\nabla w|^2 dx dr \\ & = E_{L,\delta}(t_2) - (1 + \epsilon) \int_t^{t_2} \int_{\mathbb{R}^d} \chi_L A_\delta(w) \operatorname{div} s dx dr + \mathcal{R}_{L,\delta}(t, t_2), \end{aligned}$$

where

$$\mathcal{R}_{L,\delta}(t, t_2) = \int_t^{t_2} \int_{\mathbb{R}^d} \beta_\delta(w) (\epsilon \Delta \chi_L + (1 + \epsilon) \langle s, \nabla \chi_L \rangle) dx dr.$$

For $\epsilon = 0$, the same identity follows from the chain rule along characteristics. This is the time-integrated, localized form of (5.10).

Discard the nonnegative diffusion term and use $\operatorname{div} s(x, r) \geq -d/(2r)$. For $z \geq 0$ and $0 < \delta \leq 1$,

$$0 \leq \beta_\delta(z) + A_\delta(z) = z^2(z^2 + \delta^2)^{(p-2)/2} \leq C_p(z + z^p),$$

while $\beta_\delta(z) \rightarrow z^p/p$ and $A_\delta(z) \rightarrow (p-1)z^p/p$ as $\delta \rightarrow 0^+$. The bounds from Step 2 therefore justify dominated convergence as $\delta \rightarrow 0^+$ in both endpoint terms $E_{L,\delta}(t)$ and $E_{L,\delta}(t_2)$, and in the remaining space-time integrals.

Multiplying the resulting inequality by p and then letting $t_2 \rightarrow T^-$, using the strong L^p terminal trace, yields

$$(A.2) \quad F_L(t) \leq F_L(T) + c_p \int_t^T \frac{F_L(r)}{r} dr + \mathcal{R}_L(t),$$

where

$$F_L(t) = \int_{\mathbb{R}^d} \chi_L w(t)^p dx, \quad c_p = \frac{d(1+\epsilon)(p-1)}{2},$$

and

$$|\mathcal{R}_L(t)| \leq C_{\tau,\epsilon} \int_t^T \int_{A_L} w^p dx dr \longrightarrow 0 \quad \text{as } L \rightarrow \infty.$$

Here the cutoff derivatives compensate for the linear growth of s , and the preliminary $L^1 \cap L^\infty$ bound makes w^p integrable in space and time.

Monotone convergence in (A.2) gives

$$F(t) \leq F(T) + c_p \int_t^T \frac{F(r)}{r} dr, \quad F(t) = \|w(t)\|_{L^p}^p.$$

The backward Gronwall argument of Section 5.2 therefore yields

$$(A.3) \quad \|w(t)\|_{L^p} \leq \left(\frac{T}{t}\right)^{c_p/p} \|w_T\|_{L^p}, \quad \tau \leq t \leq T.$$

For $p = 1$, nonnegativity and mass conservation give $\|w(t)\|_{L^1} = \|w_T\|_{L^1}$, so the same estimate holds with $c_1 = 0$.

Step 4 (Passage to the limit). Set $v^n(t) = \mathcal{S}_{t,T} v_T^n$. Decomposing the terminal densities into their common part $\min\{v_T^n, v_T^m\}$ and their nonnegative remainders, positivity and additivity give

$$(A.4) \quad |v^n(t) - v^m(t)| \leq \mathcal{S}_{t,T} |v_T^n - v_T^m| \quad \text{a.e. in } \mathbb{R}^d.$$

Since $|v_T^n - v_T^m| \in \operatorname{Lip}_c(\mathbb{R}^d)$, Step 3 implies

$$\begin{aligned} \sup_{t \in [\tau, T]} \|v^n(t) - v^m(t)\|_{L^p} &\leq \left(\frac{T}{\tau}\right)^{c_p/p} \|v_T^n - v_T^m\|_{L^p}, \\ \sup_{t \in [\tau, T]} \|v^n(t) - v^m(t)\|_{L^1} &\leq \|v_T^n - v_T^m\|_{L^1}. \end{aligned}$$

Thus v^n converges in $C([\tau, T]; L^1) \cap C([\tau, T]; L^p)$ to a nonnegative unit-mass limit with terminal density v_T . Passing to the limit against compactly supported test functions shows that it solves (1.6). By Proposition 2.1, this limit is the unique weak probability solution v .

Passing to the limit in (A.3) proves (2.4). The convergence in $C([\tau, T]; L^p)$ gives the strong terminal trace. Since $\tau > 0$ is arbitrary, the conclusion holds on $(0, T]$. $\square$

APPENDIX B. ENTROPY TRUNCATION AND TERMINAL PASSAGES

We justify the entropy calculations of Sections 5.5 and 5.6 under the hypotheses of Lemma 5.3 and Proposition 4.1. For $\epsilon > 0$, interior smoothness and positivity follow from the positive-time theory. What remains is to justify spatial integration and the terminal passage without the additional integrability and decay assumptions used in the main text.

We first replace the entropy by convex truncations of at most linear growth and localize in space using the cutoffs χ_L and annuli A_L from Appendix A, Step 2. The truncated entropy densities are bounded by a constant times the sum of the two probability densities, so their integrability is known before any entropy estimate is used. We then remove the spatial cutoff, pass to the terminal time using the L^1 traces, and finally remove the entropy truncation by monotone convergence. This proves the integrated identities and the score-error estimate, including their terminal values.

Proof of Lemma 5.3. We prove the integrated identity directly from the equation. The entropy is truncated before integration, so its finiteness at intermediate times is not assumed.

Step 1 (The deterministic case). If $\epsilon = 0$, both laws are transported by the same invertible characteristic map on every interval $[t_1, t_2] \subset (0, T]$. The Jacobian factors cancel in the density ratio. A change of variables therefore gives

$$\text{KL}(v(t_1) \| u(t_1)) = \text{KL}(v(t_2) \| u(t_2)).$$

The same argument applies with u and v interchanged.

Step 2 (A truncated entropy identity). Assume $\epsilon > 0$ and fix $0 < t_1 < t_2 < T$. Both densities are smooth and positive on this time interval. Set $h = v/u$. Subtracting the equation for u , multiplied by h , from that for v gives

$$(B.1) \quad \partial_t h + \epsilon \Delta h + (\epsilon - 1) \langle s, \nabla h \rangle = 0.$$

Let $\theta_n \in C_c^\infty((0, \infty))$ satisfy $0 \leq \theta_n \uparrow 1$, and define

$$(B.2) \quad \Psi_n''(z) = \frac{\theta_n(z)}{z}, \quad \Psi_n(1) = \Psi_n'(1) = 0.$$

Then $0 \leq \Psi_n \uparrow \Psi$, where $\Psi(z) = z \log z - z + 1$, and $\Psi_n(z) \leq C_n(1 + z)$. In addition, Ψ_n' and $\Psi_n - z\Psi_n'$ are bounded. Put $\mathcal{H}_n = u\Psi_n(h)$ and $H_n(t) = \int_{\mathbb{R}^d} \mathcal{H}_n(x, t) dx$. The product rule and (B.1) yield

$$(B.3) \quad \partial_t \mathcal{H}_n + \epsilon \Delta \mathcal{H}_n - (1 + \epsilon) \text{div}(s \mathcal{H}_n) = \epsilon u \Psi_n''(h) |\nabla h|^2.$$

Notice that $0 \leq \mathcal{H}_n \leq C_n(u + v)$, so this density is integrable in space and time on $[t_1, t_2]$.

Multiply (B.3) by the spatial cutoff χ_L from Appendix A, Step 2, and integrate. We obtain

$$\begin{aligned} \int_{\mathbb{R}^d} \chi_L \mathcal{H}_n(t_2) dx - \int_{\mathbb{R}^d} \chi_L \mathcal{H}_n(t_1) dx &= \epsilon \int_{t_1}^{t_2} \int_{\mathbb{R}^d} \chi_L u \Psi_n''(h) |\nabla h|^2 dx dt + \mathcal{E}_{n,L}, \\ |\mathcal{E}_{n,L}| &\leq C_{t_1, \epsilon} \int_{t_1}^{t_2} \int_{A_L} \mathcal{H}_n dx dt \longrightarrow 0 \quad (L \rightarrow \infty). \end{aligned}$$

The estimate uses $|s(x, t)| \leq C_{t_1}(1 + |x|)$ and $|\nabla \chi_L| \leq C/L$ on A_L . Since $\chi_L \uparrow 1$, monotone convergence on the nonnegative right-hand side gives

$$(B.4) \quad H_n(t_2) - H_n(t_1) = \epsilon \int_{t_1}^{t_2} \int_{\mathbb{R}^d} u \Psi_n''(h) |\nabla h|^2 dx dt.$$

Step 3 (Terminal trace and removal of truncation). For fixed n , the perspective function $(f, g) \mapsto g\Psi_n(f/g)$ extends continuously to the nonnegative quadrant and is Lipschitz there: its partial derivatives in the interior are $\Psi_n'(f/g)$ and $\Psi_n(f/g) - (f/g)\Psi_n'(f/g)$, which are bounded. The L^1 terminal traces of u and v therefore imply $H_n(t_2) \rightarrow H_n(T)$ as $t_2 \rightarrow T^-$. Thus (B.4) holds also for $t_2 = T$.

In particular, $H_n(t) \leq H_n(T) \leq H(T)$, where $H(T) = \text{KL}(v_T \| u(T)) < \infty$. Letting $n \rightarrow \infty$ shows that $H(t) = \int_{\mathbb{R}^d} u \Psi(v/u) dx = \text{KL}(v(t) \| u(t))$ is finite and bounded by $H(T)$. Since

$$u \Psi_n''(h) |\nabla h|^2 \uparrow \frac{u}{h} |\nabla h|^2 = v \left| \nabla \log \frac{v}{u} \right|^2,$$

monotone convergence in (B.4) proves (5.11) for $0 < t_1 < t_2 \leq T$. This identity gives the absolute continuity of H on every $[\tau, T]$ and (5.12). The local balance (B.3) depends only on the fact that the two densities solve the same FP equation. Repeating it with $v \Psi_n(u/v)$, followed by the same terminal passage, proves (5.13) whenever $\text{KL}(u(T) \| v_T) < \infty$. $\square$

Proof of Proposition 4.1. The coefficient assumptions on s_θ give the same positive-time probability construction as in Section 5.1. The mass-conservation and L^1 approximation arguments of Appendix A, Steps 2 and 4, give the L^1 terminal trace; these arguments do not use the Li–Yau bound. Set

$$f = u, \quad g = v_\theta, \quad h = f/g, \quad e = s_\theta - s,$$

and use the same truncations Ψ_n as above. Both densities are smooth and positive for $0 < t < T$, and

$$\partial_t g + \epsilon \Delta g - (1 + \epsilon) \text{div}(sg) = (1 + \epsilon) \text{div}(eg),$$

whereas the left-hand operator annihilates f . For $\mathcal{H}_n = g \Psi_n(h)$, the product and chain rules give

$$\begin{aligned} (B.5) \quad & \partial_t \mathcal{H}_n + \epsilon \Delta \mathcal{H}_n - (1 + \epsilon) \text{div}(s \mathcal{H}_n) \\ &= \epsilon g \Psi_n''(h) |\nabla h|^2 + (1 + \epsilon) g h \Psi_n''(h) \langle e, \nabla h \rangle + \text{div} J_n, \\ & J_n = -(1 + \epsilon) e g (h \Psi_n'(h) - \Psi_n(h)). \end{aligned}$$

Completing the square and using $h \Psi_n''(h) = \theta_n(h) \leq 1$ yield

$$\epsilon g \Psi_n''(h) |\nabla h|^2 + (1 + \epsilon) g h \Psi_n''(h) \langle e, \nabla h \rangle \geq -\frac{(1 + \epsilon)^2}{4\epsilon} f |e|^2.$$

The additional cutoff term comes from J_n . Since

$$|h \Psi_n'(h) - \Psi_n(h)| = \left| \int_1^h \theta_n(z) dz \right| \leq |h - 1|,$$

linear growth of e gives $|J_n|/(1 + |x|) \leq C(f + g)$ on every positive-time strip. Together with $0 \leq \mathcal{H}_n \leq C_n(f + g)$, this justifies removing the spatial cutoff as in the preceding proof. Writing $H_n(t) = \int_{\mathbb{R}^d} \mathcal{H}_n(x, t) dx$, we obtain

$$H_n(t) \leq H_n(t_2) + \frac{(1 + \epsilon)^2}{4\epsilon} \int_t^{t_2} \int_{\mathbb{R}^d} u(x, r) |e(x, r)|^2 dx dr, \quad 0 < t < t_2 < T.$$

The forcing is integrable on $[t, T]$ because e has at most linear growth and u has uniformly bounded second moments there. For fixed n , the Lipschitz property of the truncated perspective and the L^1 terminal traces allow $t_2 \rightarrow T^-$. Letting then $n \rightarrow \infty$ by monotone convergence gives (4.6). $\square$

APPENDIX C. PROOFS FOR THE EXPLICIT CASE

Proof of Proposition 3.9. The special structure at $\epsilon = 1$ is seen by writing $v = uh$. Since $\partial_t u = \Delta u$ and $su = \nabla u$,

$$\partial_t(uh) + \Delta(uh) - 2 \text{div}(suh) = u(\partial_t h + \Delta h).$$

Hence the quotient satisfies

$$(C.1) \quad \partial_t h + \Delta h = 0.$$

This suggests

$$h(x, t) = \int_{\mathbb{R}^d} G_{T-t}(x - y) \frac{1}{u(y, T)} dv_T(y).$$

Define $\tilde{v} = uh$ by this formula. On each compact subset of $\mathring{Q}$, the Gaussian bounds of Lemma 5.1 bound the integrand and its derivatives by $C(1 + |y|)^k e^{-c|y|^2 + C|y|}$ for suitable $c > 0$, $C > 0$, and k . This is bounded in y and integrable against v_T . Differentiation under the integral is therefore justified, and $\tilde{v}$ solves (1.6) on $\mathring{Q}$.

It remains to identify the two endpoint traces. Introduce the probability measure

$$(C.2) \quad d\Gamma_T(z, y) := \frac{G_T(y - z)}{u(y, T)} du_0(z) dv_T(y).$$

Indeed, integration in z gives

$$\int_{\mathbb{R}^d} \frac{G_T(y - z)}{u(y, T)} du_0(z) = 1.$$

Its y -marginal is therefore v_T , while its z -marginal is $a_{v_T} u_0$.

Set

$$c_t(z, y) = \left(1 - \frac{t}{T}\right) z + \frac{t}{T} y, \quad \kappa_t = t \left(1 - \frac{t}{T}\right).$$

Completing the square gives the Gaussian bridge identity

$$(C.3) \quad G_t(x - z) G_{T-t}(y - x) = G_T(y - z) G_{\kappa_t}(x - c_t(z, y)).$$

Using $u(t) = G_t * u_0$ in the definition of $\tilde{v}$, we obtain

$$(C.4) \quad \tilde{v}(x, t) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} G_{\kappa_t}(x - c_t(z, y)) d\Gamma_T(z, y).$$

This representation defines a narrowly continuous curve of probability measures on $(0, T)$. Let $\varphi \in C_b(\mathbb{R}^d)$. Since the inner Gaussian in (C.4) is a probability density, dominated convergence gives

$$\lim_{t \rightarrow T^-} \int_{\mathbb{R}^d} \varphi d\tilde{v}(t) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(y) d\Gamma_T(z, y) = \int_{\mathbb{R}^d} \varphi dv_T,$$

because $\kappa_t \rightarrow 0$ and $c_t(z, y) \rightarrow y$. Thus the candidate $\tilde{v}$ has terminal law v_T . Uniqueness in Proposition 2.1 gives $\tilde{v} = v$, proving (3.13).

Similarly, as $t \rightarrow 0^+$, $\kappa_t \rightarrow 0$ and $c_t(z, y) \rightarrow z$, so

$$\lim_{t \rightarrow 0^+} \int_{\mathbb{R}^d} \varphi dv(t) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(z) d\Gamma_T(z, y) = \int_{\mathbb{R}^d} \varphi(z) a_{v_T}(z) du_0(z).$$

Theorem 3.1 already gives convergence of $v(t)$ in W_2 to the unique limiting law v_0^* . Hence the preceding narrow limit identifies

$$dv_0^* = a_{v_T} du_0.$$

Finally, Tonelli's theorem gives

$$\int_{\mathbb{R}^d} a_{v_T}(z) du_0(z) = \int_{\mathbb{R}^d} \frac{\int_{\mathbb{R}^d} G_T(y - z) du_0(z)}{u(y, T)} dv_T(y) = 1.$$

Since the kernel in the definition of a_{v_T} is strictly positive, we have $0 < a_{v_T} < \infty$ u_0 -almost everywhere. Thus v_0^* and u_0 are equivalent. $\square$

Proof of Proposition 3.10. The proof follows the moment-expansion approach of [13], applied to the rescaled Gaussian convolution. Centering removes the first-moment term, while Gaussian averaging determines the corrections that survive in the endpoint weights. Since the explicit formula involves the reciprocal of the heat density, we also control the normalized denominator and the integrated reciprocal remainder.

Step 1 (Centering and rescaling). Consider first $v_T = G_T$. Set

$$\nu = (x \mapsto x - m_0)_{\#} u_0, \quad \delta = (2T)^{-1/2}, \quad \varphi = G_{1/2}.$$

Thus $\int_{\mathbb{R}^d} z \, d\nu(z) = 0$ and $\int_{\mathbb{R}^d} z z^\top \, d\nu(z) = \Sigma_0$. Define

$$D_\delta(\xi) = \int_{\mathbb{R}^d} e^{\delta \langle \xi, z \rangle - \delta^2 |z|^2 / 2} \, d\nu(z).$$

The substitution $y = m_0 + \xi / \delta$ in (3.14), followed by completion of the square, gives

$$(C.5) \quad a_T(x) = e^{-\delta^2 c} \int_{\mathbb{R}^d} \frac{\varphi(\xi)}{D_\delta(\xi + \delta b)} \, d\xi, \quad b = x - 2m_0, \quad c = \langle m_0, x - m_0 \rangle.$$

Since S is compact, b and c remain bounded for $x \in S$. All remainder estimates below are uniform on S .

Step 2 (Gaussian moments of the denominator). Put $k_\delta(\xi) = D_\delta(\xi + \delta b) - 1$ and

$$H(z, w) = \langle z, w \rangle + \langle b, z + w \rangle.$$

The Gaussian identity

$$\int_{\mathbb{R}^d} e^{\delta \langle \xi, z \rangle} \varphi(\xi) \, d\xi = e^{\delta^2 |z|^2 / 2}$$

allows us to integrate in ξ before expanding. Since ν is centered,

$$\iint_{\mathbb{R}^d \times \mathbb{R}^d} H \, d\nu(z) \, d\nu(w) = 0, \quad \iint_{\mathbb{R}^d \times \mathbb{R}^d} H^2 \, d\nu(z) \, d\nu(w) = \text{tr}(\Sigma_0^2) + 2\langle \Sigma_0 b, b \rangle.$$

Consequently, Taylor's formula gives

$$(C.6) \quad \begin{aligned} \int_{\mathbb{R}^d} k_\delta \varphi \, d\xi &= \int_{\mathbb{R}^d} (e^{\delta^2 \langle b, z \rangle} - 1) \, d\nu(z) \\ &= \frac{\delta^4}{2} \langle \Sigma_0 b, b \rangle + \mathcal{O}(\delta^6), \\ \int_{\mathbb{R}^d} k_\delta^2 \varphi \, d\xi &= \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{\delta^2 H(z, w)} \, d\nu(z) \, d\nu(w) - 2 \int_{\mathbb{R}^d} e^{\delta^2 \langle b, z \rangle} \, d\nu(z) + 1 \\ (C.7) \quad &= \frac{\delta^4}{2} \text{tr}(\Sigma_0^2) + \mathcal{O}(\delta^6). \end{aligned}$$

The remainders are uniform because z, w , and b remain in bounded sets.

Step 3 (Reciprocal expansion and remainder). Jensen's inequality and the centering of ν give

$$(C.8) \quad D_\delta(\xi + \delta b) \geq e^{-\delta^2 \text{tr}(\Sigma_0)/2}.$$

Furthermore, Taylor's formula, with the linear term in δ canceled by $\int_{\mathbb{R}^d} z \, d\nu(z) = 0$, yields

$$|k_\delta(\xi)| \leq C \delta^2 (1 + |\xi|^2) e^{C\delta|\xi| + C\delta^2}.$$

Gaussian integrability therefore implies

$$(C.9) \quad \int_{\mathbb{R}^d} |k_\delta(\xi)|^3 \varphi(\xi) \, d\xi = \mathcal{O}(\delta^6).$$

Using the exact identity

$$\frac{1}{1+k} = 1 - k + k^2 - \frac{k^3}{1+k},$$

the lower bound (C.8) controls the integrated remainder by a constant times (C.9). Substituting (C.6)–(C.7) into (C.5) thus gives

$$\begin{aligned} a_T(x) &= e^{-\delta^2 c} \left(1 + \frac{\delta^4}{2} (\text{tr}(\Sigma_0^2) - \langle \Sigma_0 b, b \rangle) + \mathcal{O}(\delta^6) \right) \\ &= 1 - \delta^2 c + \frac{\delta^4}{2} (c^2 + \text{tr}(\Sigma_0^2) - \langle \Sigma_0 b, b \rangle) + \mathcal{O}(\delta^6). \end{aligned}$$

Since $\delta^2 = (2T)^{-1}$, this proves (3.16)–(3.17).

For $v_T = G_T(\cdot - m_0)$, translate the data and the terminal Gaussian by $-m_0$. The translated data have zero mean and covariance Σ_0 . Applying the preceding expansion at $x - m_0$, with the mean-mismatch factor absent, gives (3.18). $\square$

REFERENCES

- [1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich, 2nd ed., Birkhäuser, Basel, 2008.
- [2] B. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- [3] R. Baptista, A. Dasgupta, N. B. Kovachki, A. Oberai, and A. M. Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785v2*, 2025.
- [4] G. I. Barenblatt. *Scaling, Self-similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics*. Cambridge University Press, 1996.
- [5] 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.
- [6] G. Biroli, T. Bonnaire, V. De Bortoli, and M. Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15:9957, 2024.
- [7] V. I. Bogachev, N. V. Krylov, M. Röckner, and S. V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*. American Mathematical Society, Mathematical Surveys and Monographs, Vol. 207, 2015.
- [8] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwal, F. Tramèr, 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.
- [9] 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.
- [10] 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.
- [11] V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.
- [12] 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.
- [13] J. Duoandikoetxea and E. Zuazua. Moments, masses de Dirac et décomposition de fonctions. *C. R. Acad. Sci. Paris Sér. I Math.*, 315(6):693–698, 1992.
- [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. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.
- [17] D. P. Kingma and Y. LeCun. Regularized estimation of image statistics by score matching. *Advances in Neural Information Processing Systems*, 23:1126–1134, 2010.
- [18] B. Klartag and O. Ordentlich. The strong data processing inequality under the heat flow. *IEEE Transactions on Information Theory*, 71(5):3317–3333, 2025.
- [19] C. Le Bris and P.-L. Lions. Existence and uniqueness of solutions to Fokker–Planck type equations with irregular coefficients. *Communications in Partial Differential Equations*, 33(7):1272–1317, 2008.
- [20] 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.

- [21] P. Li and S.-T. Yau. On the parabolic kernel of the Schrödinger operator. *Acta Mathematica*, 156:153–201, 1986.
- [22] S. Li, S. Chen, and Q. Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.
- [23] K. Liu and E. Zuazua. Convergence of guided generative dynamics in diffusion models: score mixing and classifier-free guidance. Manuscript in preparation, 2026.
- [24] Y. Lyu, T. M. Nguyen, Y. Qian, and X. T. Tong. Resolving memorization in empirical diffusion model for manifold data in high-dimensional spaces. *arXiv preprint arXiv:2505.02508v3*, 2025.
- [25] J. Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.
- [26] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.
- [27] 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.
- [28] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021.
- [29] 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.
- [30] C. Villani. *Optimal Transport: Old and New*. Springer, Grundlehren der Mathematischen Wissenschaften, Vol. 338, 2009.
- [31] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.
- [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.