Geometric Asymptotics of Score Mixing and Guidance in Diffusion Models

Kang Liu* Enrique Zuazua†

Abstract

Diffusion models are routinely guided in practice by combining multiple score fields, yet the mathematical structure of score mixing is still poorly understood. We study the small-time generation dynamics driven by mixed scores

$$s = \lambda \nabla \log u_1 + (1-\lambda)\nabla \log u_2, \qquad \lambda \geq 0,$$

in the heat-flow framework, where u_1, u_2 are heat evolutions of two compactly supported probability measures. This single formulation covers both the mixture-of-experts regime ($0 \leq \lambda \leq 1$) and the classifier-free guidance regime ($\lambda > 1$). Exploiting a Laplace–Varadhan principle under a similarity-time rescaling, we show that the small-time generation dynamics is governed by the explicit geometric potential

$$\Phi_\lambda = \lambda d_1^2 + (1-\lambda) d_2^2,$$

which depends only on the supports of the initial measures and on the mixing parameter. This gives a rigorous reduction from a singular, non-autonomous score-driven dynamics to autonomous Clarke-type subgradient inclusions. In the empirical setting of finite Dirac mixtures, the limiting potential is piecewise quadratic with a Voronoi-type structure; this rigidity yields convergence of all autonomous limiting trajectories to critical points and a conditional convergence criterion for the original generation flow toward local minimizers of the potential, with rate $\mathcal{O}(\sqrt{t})$ in the smooth stable case.

Contents

*Université Bourgogne Europe, CNRS, Institut de Mathématiques de Bourgogne, 21000 Dijon, France. kang.liu@u-bourgogne.fr

†[1] Friedrich–Alexander–Universität Erlangen–Nürnberg, Department of Mathematics, Chair for Dynamics, Control, Machine Learning, and Numerics (Alexander von Humboldt Professorship), 91058 Erlangen, Germany;
[2] Universidad Autónoma de Madrid, Departamento de Matemáticas, 28049 Madrid, Spain;
[3] Chair of Computational Mathematics, Fundación Deusto, 48007 Bilbao, Basque Country, Spain. enrique.zuazua@fau.de

1 Introduction

1.1 Background and motivation

Over the last few years, diffusion models have reshaped the landscape of generative AI, achieving state-of-the-art performance in image, audio, and multimodal generation [22, 49, 50] and complementing alternative paradigms such as GANs and VAEs [21, 28, 9]. At a high level, these models learn to transform simple Gaussian noise into structured samples that approximate a complex and

unknown data distribution. Algorithmically, they are often formulated through stochastic differential equations (SDEs) driven by an estimated *score function*, i.e., the gradient of the log-density of suitably noised data [49, 50]. The estimation of such scores is closely related to score matching [24, 27, 52].

From an analytical viewpoint, this framework is naturally connected to classical partial differential equations (PDEs), in particular to the heat equation and the associated Fokker–Planck dynamics [1, 35]. This PDE perspective provides a natural setting in which questions of well-posedness, stability, and asymptotic behavior can be rigorously addressed.

In the companion work [35], we developed a PDE-based framework for score-based diffusion models driven by the heat equation. There, the heat flow associated with the data distribution is taken as a minimal forward model, and its score field, the gradient of the log of the heat flow, is viewed as the idealized driving vector field for reverse-time generation. This leads to an interpretation of diffusion models as a backward Fokker–Planck evolution or, in the deterministic limit, as a non-autonomous gradient flow driven by a singular potential.

Within this setting, tools from geometric analysis and entropy methods become available. In particular, Li–Yau-type differential inequalities [31] control the divergence of the score and yield robust well-posedness and L^p-stability estimates for the backward dynamics. In the spirit of hypocoercivity and functional inequalities [51, 29], entropy-based arguments in terms of the Kullback–Leibler divergence [13, 10, 30, 8] show that reverse-time trajectories concentrate near the data manifold as the terminal time is approached.

In many modern applications, diffusion models are not used as isolated score fields, but are instead guided, composed, or modified at sampling time by combining several score-like vector fields [23, 17]. This naturally leads to the mechanism of **score mixing**:

$$s = \lambda s_1 + (1-\lambda)s_2, \qquad \lambda \geq 0.$$

Despite its widespread empirical success, the mathematical structure of score mixing and its impact on the associated generation dynamics remain largely unexplored. The central question addressed in this work is:

What is the intrinsic geometric and dynamical effect of combining score fields?

Working in the heat-flow setting allows us to isolate the mechanism of score mixing in a transparent mathematical framework. The main contribution of this paper is a geometric reduction principle for guided score-based generation: after a suitable similarity-time rescaling, the singular non-autonomous dynamics driven by mixed heat-flow scores is asymptotically governed by an autonomous nonsmooth dynamics associated with a geometric distance potential. This reduction shows that, in the small-time regime, the generation dynamics asymptotically forgets the full analytic structure of the heat flow and retains only the geometry of the supports of the initial measures, together with the mixing parameter λ.

More precisely, the main results of the paper are as follows:

- We identify the geometric small-time limit of score-mixing dynamics. Under a uniform lower small-ball mass condition on the initial measures, the rescaled mixed heat-flow potential converges, as $t \to 0^+$, to an explicit distance potential Φ_λ, determined only by the supports of the data and by the mixing parameter λ. The key mechanism behind this reduction is the Laplace–Varadhan principle.
- We derive the corresponding autonomous nonsmooth limiting dynamics. Every time-shift limit of the rescaled generation flow is a Carathéodory solution of a Clarke-type subgradient

inclusion: the genuine Clarke subgradient flow in the MoE regime and an outer Clarke inclusion in the CFG regime.

- In the empirical setting of finite Dirac mixtures, we exploit the piecewise quadratic, Voronoi-type structure of Φ_λ to prove convergence of every global solution of the autonomous limiting inclusion to a critical point. For the original non-autonomous generation flow, we prove a conditional convergence criterion: if non-minimizing critical points are excluded from the ω-limit set, then the flow converges to a local minimizer of Φ_λ. In the smooth stable case, we further obtain the convergence rate $\mathcal{O}(\sqrt{t})$.
- We complement the deterministic analysis with PDE and stochastic perspectives. On the PDE side, we relate Li–Yau-type Hessian bounds to the semiconcavity and Hamilton–Jacobi structure of the rescaled logarithmic potential, and derive L^p-energy estimates for the backward Fokker–Planck equation, revealing a polynomial-versus-exponential stability dichotomy between the MoE and CFG regimes. On the stochastic side, we interpret the noisy rescaled dynamics as a vanishing-viscosity perturbation of the limiting geometric dynamics, as a guide for future work rather than as a theorem of the present paper.

The analysis is carried out for exact heat-flow scores, which provide a tractable mathematical proxy for practical score-based models. We do not claim to analyze learned neural scores or general diffusion SDEs in full generality. Rather, our goal is to isolate a robust geometric mechanism underlying score mixing and guidance.

1.2 Related work

Classifier guidance was introduced in diffusion models as a way to produce "low-temperature" samples by correcting the model score with the gradient of a label-conditional density, thereby steering the reverse dynamics toward label-consistent regions [17]. This strategy, however, requires an additional classifier or conditional model, and may become less effective when the conditioning variable is high-dimensional.

Classifier-free guidance (CFG) avoids this external classifier by combining conditional and unconditional scores through a linear extrapolation [23]. In the notation of the present paper, this corresponds to the score-mixing regime $\lambda > 1$. CFG has since become a standard component of modern text-to-image diffusion pipelines. A recent fine-grained analysis of guidance in simple mixture models was carried out in [11]. There, the authors show that guidance does not merely sample from a naively tilted distribution, but can induce a geometric bias toward boundary or archetypal regions of the target component, and may lead to off-support behavior when the guidance strength is large and the score estimate is imperfect. This viewpoint is closely related in spirit to the geometric mechanism studied here, where the CFG regime is governed at small time by the extrapolative distance potential.

The averaging regime $0 \le \lambda \le 1$, which we refer to as the mixture-of-experts (MoE) regime, has also been explored; for instance, in [42], mixed scores are used to generate synthetic data that improves downstream recognition. Conceptually, this averaging viewpoint is also reminiscent of aggregation principles in federated learning, such as FedAvg [40], where several sources are combined through weighted averaging.

On the theoretical side, non-mixed diffusion models have been studied from PDE, stochastic, probabilistic, and numerical perspectives; see, for instance, [35, 13, 10, 30, 8]. In the heat-flow framework of [35], the generation process driven by a single heat-flow score was shown, under mild assumptions on the initial datum u_0, to return almost surely to $\mathrm{supp}(u_0)$. The proof relies on entropy contraction estimates for the backward Fokker–Planck equation, in the spirit of hypocoercivity and data-processing inequalities [51, 29]. Such an entropy-based approach is no longer directly

available for mixed scores, since the natural product candidate $u_1^\lambda u_2^{1-\lambda}$ is not in general a normalized probability density; see Remark 4.3. This obstruction motivates the geometric approach developed here.

Our analysis is based on the asymptotically autonomous structure of the generation dynamics. After a similarity-time rescaling [54], the non-autonomous reverse dynamics admits time-shift limits described by autonomous nonsmooth differential inclusions. This point of view is in the spirit of the classical theory initiated by Markus [39]; see also Galaktionov–Vázquez [19]. The limiting vector fields are governed by geometric potentials obtained from a Laplace–Varadhan principle for Gaussian convolutions; see, for instance, [16, Sec. 4.3] and [7, Sec. 6.4]. Since these potentials are generally nonsmooth, the limiting dynamics is formulated using Clarke and outer-Clarke differential inclusions [12]. Thus, the present work connects score-based diffusion models with geometric asymptotics and nonsmooth dynamical systems.

Relation to the companion paper [35]. The present paper builds on the heat-flow formulation introduced in [35], but addresses a different problem and requires different tools. The companion paper treats the case of a single heat-flow score, where entropy contraction, Li–Yau-type inequalities, and backward Fokker–Planck estimates provide the main analytical mechanism. By contrast, the present work studies mixed scores, including both the MoE regime $0 \le \lambda \le 1$ and the CFG regime $\lambda > 1$. In this setting, the entropy method is no longer the natural organizing principle, because the product $u_1^\lambda u_2^{1-\lambda}$, although it generates the mixed score, is not in general a normalized probability density and does not yield the same entropy-contraction mechanism; see Remark 4.3.

This is why the present paper develops a different approach, based on Laplace–Varadhan asymptotics, geometric distance potentials, and nonsmooth autonomous limiting dynamics. The L^p-energy estimates also differ from the single-score case: while the MoE estimate follows the strategy of [35, Thm. 3.1], the CFG regime involves an additional exponential amplification factor; see Theorem 4.1.

1.3 Organization of the paper

Section 2 introduces the heat-flow framework, the mixed-score construction, and the associated deterministic and stochastic generation dynamics. Section 3 states the main results, including the similarity-time rescaling, the geometric limiting potential, the limiting autonomous inclusions, and the convergence results in the empirical setting. Section 4 complements the dynamical analysis with PDE, Hamilton–Jacobi, and stochastic viewpoints, through Li–Yau/semiconcavity structures, energy estimates, and a stochastic-approximation interpretation of the noisy dynamics. Section 5 presents numerical experiments illustrating the asymptotic behavior of both deterministic and stochastic generation flows. Section 6 develops the Laplace–Varadhan asymptotics and gradient structure underlying the geometric reduction, and proves the time-shift convergence theorem for the rescaled generation dynamics. Section 7 studies the Clarke structure of the limiting geometric potential and the convergence of the corresponding autonomous inclusions. Section 8 proves the empirical convergence criterion and the rate estimates near smooth stable minimizers. Finally, Section 9 contains concluding remarks and perspectives.

2 Problem setting: score mixing and guided generation

In this section, we introduce the mathematical framework for score mixing and define the associated generation dynamics. This formulation highlights the interaction between multiple datasets through their score fields and prepares the ground for the small-time asymptotic analysis developed in Section 3, where the dynamics is shown to be governed by an explicit geometric potential.

2.1 Mixture of scores in diffusion models

To make the discussion concrete, let $u_i = u_i(x,t)$, $i = 1,2$, be two solutions of the forward heat equation on

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

where $T > 0$ is a fixed finite time horizon:

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

Here $\mathcal{P}(\mathbb{R}^d)$ denotes the space of Borel probability measures on $\mathbb{R}^d$, and the initial measures $u_{0,1}$ and $u_{0,2}$ represent the two datasets under consideration.

The associated score fields are

$$s_i(x,t) := \nabla \log u_i(x,t), \qquad (x,t) \in Q, \quad i = 1,2. \tag{2.2}$$

Rather than using each score separately, we consider the mixed score

$$s^{(\lambda)} := \lambda s_1 + (1-\lambda) s_2, \qquad \lambda \geq 0, \tag{2.3}$$

where λ is a mixing parameter.

Two regimes will play a central role in the sequel:

1. **Mixture-of-experts (MoE) regime $0 \leq \lambda \leq 1$.** In this case, $s^{(\lambda)}$ is a convex combination of the two scores.
2. **Classifier-free guidance (CFG) regime $\lambda > 1$.** In this case, $s^{(\lambda)}$ is an extrapolation. Writing $\alpha = \lambda - 1 > 0$, one has

$$s^{(\lambda)} = s_1 + \alpha(s_1 - s_2), \tag{2.4}$$

which matches the standard parametrization of classifier-free guidance in the diffusion-model literature [23, 47].

Equivalently, the mixed score is the logarithmic gradient of a product-of-experts-type density:

$$s^{(\lambda)} = \nabla \log u^{(\lambda)}, \qquad u^{(\lambda)} := u_1^{\lambda} u_2^{1-\lambda}. \tag{2.5}$$

For $0 \leq \lambda \leq 1$, this object can be interpreted as a geometric interpolation between the two heat flows. For $\lambda > 1$, it is an extrapolated product and is not, in general, a normalized probability density.

This distinction is important analytically. In the convex regime, the Li–Yau inequality yields one-sided divergence bounds for the mixed score, whereas such bounds may fail in the extrapolating CFG regime. Moreover, the entropy-based Fokker–Planck estimates used in the single-score setting of [35] do not directly extend to $u^{(\lambda)}$, precisely because $u^{(\lambda)}$ is not generally normalized. This motivates the strategy adopted in this paper: we analyze the associated characteristic ODEs and their geometric small-time limits directly.

2.2 Generation dynamics driven by the mixed score

The central object of this work is the generation dynamics driven by the mixed score $s^{(\lambda)}$. Starting from a prescribed terminal law $X_T \sim \nu_T$, typically chosen to be Gaussian, we consider the reverse-time stochastic differential equation

$$dX_t = -(1+\varepsilon)\, s^{(\lambda)}(X_t, t)\, dt + \sqrt{2\varepsilon}\, dW_t, \qquad t \in (0,T), \tag{2.6}$$

to be integrated from $t = T$ down to $t = 0$. Here $\varepsilon \geq 0$ and $(W_t)_{t\geq 0}$ is a standard Brownian motion. On every interval $[t_0, T]$, with $t_0 > 0$, existence and uniqueness follow from the smoothness and strict positivity of the heat flows $u_i(\cdot, t)$, hence from the standard Lipschitz theory for SDEs; see, for instance, [25, Ch. 5]. The difficulty is concentrated in the singular regime $t \to 0^+$, where the scores develop the small-time structures analyzed in this paper.

A generated sample is obtained by integrating (2.6) backward in time, for instance by Euler–Maruyama, and evaluating the trajectory at $t = 0$, or at a small positive stopping time in numerical implementations.

The density $\rho_\varepsilon(t, \cdot)$ associated with (2.6) satisfies the backward Fokker–Planck equation

$$\partial_t \rho_\varepsilon + \varepsilon\, \Delta \rho_\varepsilon - (1 + \varepsilon)\, \operatorname{div}\big(\rho_\varepsilon\, \nabla V_\lambda\big) = 0, \tag{2.7}$$

where

$$V_\lambda(x, t) := \log u^{(\lambda)}(x, t) = \lambda \log u_1(x, t) + (1 - \lambda) \log u_2(x, t). \tag{2.8}$$

In the deterministic case $\varepsilon = 0$, the generation process reduces to the characteristic ODE

$$\dot{X}_t = -\nabla V_\lambda(X_t, t), \tag{2.9}$$

whose associated transport equation is

$$\begin{cases} \partial_t \rho - \operatorname{div}\big(\rho\, \nabla V_\lambda(\cdot, t)\big) = 0, \quad (x, t) \in \mathbb{R}^d \times (0, T), \\ \rho(T, \cdot) = v_T. \end{cases} \tag{2.10}$$

This deterministic regime is also relevant computationally, as fast samplers such as DPM-Solver [37] are based on ODE-type reverse dynamics.

The potential V_λ is time-dependent and becomes singular as $t \to 0^+$. Thus, even in the deterministic setting, the behavior of (2.9) near the terminal generation time is not described by a standard autonomous gradient-flow picture. The central question is therefore:

What is the asymptotic behavior of the generation trajectories X_t as $t \to 0^+$?

The main results in the next section answer this question through a similarity-time rescaling: the singular non-autonomous dynamics is shown to converge, in an appropriate time-shift sense, toward an autonomous nonsmooth dynamics driven by a geometric potential depending only on the supports of the initial measures and on the mixing parameter λ.

3 Main results: geometric potential and limiting dynamics

3.1 Similarity-time rescaling and geometric limiting objects

To analyze the small-time regime of the generation dynamics (2.9), it is convenient to introduce the logarithmic, or similarity, time variable

$$\tau = \log\Big(\frac{T}{t}\Big), \qquad Y_\tau := X_{Te^{-\tau}}. \tag{3.1}$$

Thus the terminal time $t = T$ corresponds to $\tau = 0$, while the small-time regime $t \to 0^+$ corresponds to $\tau \to +\infty$.

Since $dt/d\tau = -Te^{-\tau}$, the chain rule gives

$$\dot{Y}_\tau = -Te^{-\tau}\, \dot{X}_{Te^{-\tau}}.$$

Using the generation ODE (2.9), we obtain the forward-time rescaled dynamics

$$\begin{cases} \dot{Y}_\tau = Te^{-\tau}\,\nabla V_\lambda\big(Y_\tau, Te^{-\tau}\big) = -\dfrac{1}{4}\,\nabla F_\lambda\big(Y_\tau, Te^{-\tau}\big), \quad \tau > 0, \\ Y_0 = x_T, \end{cases} \tag{3.2}$$

where the rescaled potential is defined by

$$F_\lambda(x,t) \coloneqq -4t\,V_\lambda(x,t). \tag{3.3}$$

Consequently, the small-time behavior of X_t as $t \to 0^+$ is equivalent to the long-time behavior of Y_τ as $\tau \to +\infty$. In particular, the two parametrizations have the same set of accumulation points:

$$\omega_t(X) = \omega_\tau(Y). \tag{3.4}$$

Definition 3.1 (ω-limit set)**.** Fix $T > 0$. Let $X = (X_t)_{t\in(0,T]} \in C((0,T];\mathbb{R}^d)$ and $Y = (Y_\tau)_{\tau\geq 0} \in C([0,\infty);\mathbb{R}^d)$. We define

$$\omega_t(X) \coloneqq \Big\{x^* \in \mathbb{R}^d :\ \exists\, t_n \to 0^+ \text{ such that } X_{t_n} \to x^*\Big\},$$

and

$$\omega_\tau(Y) \coloneqq \Big\{y^* \in \mathbb{R}^d :\ \exists\, \tau_n \to \infty \text{ such that } Y_{\tau_n} \to y^*\Big\}.$$

Remark 3.2 (Justification of (3.4))**.** In view of the definitions above, (3.4) follows directly from the bijectivity of the similarity-time change of variables. Indeed, if $x^* \in \omega_t(X)$, then there exists $t_n \to 0^+$ such that $X_{t_n} \to x^*$. Defining $\tau_n \coloneqq \log(T/t_n)$, we have $\tau_n \to \infty$ and

$$Y_{\tau_n} = X_{Te^{-\tau_n}} = X_{t_n} \to x^*,$$

hence $x^* \in \omega_\tau(Y)$. The converse inclusion is obtained identically by setting $t_n \coloneqq Te^{-\tau_n}$.

Geometric distance potential. The rescaled potential F_λ defined in (3.3) is the central object in the small-time analysis. As developed in Section 6, Laplace–Varadhan asymptotics for Gaussian convolutions show that, as $t \to 0^+$, $F_\lambda(\cdot,t)$ is governed at leading order by a purely geometric object depending only on the supports of the initial measures and on the parameter λ. This motivates the introduction of the limiting *geometric distance potential* Φ_λ, defined as the λ-weighted combination of the squared distances to the two data supports.

Definition 3.3 (Geometric distance potential)**.** Let $A_i \coloneqq \operatorname{supp}(u_{0,i})$ for $i = 1,2$, and define

$$d_i(x) \coloneqq \operatorname{dist}(x, A_i), \qquad i = 1,2.$$

For $\lambda \geq 0$, we introduce the *geometric distance potential*

$$\Phi_\lambda(x) \coloneqq \lambda\, d_1(x)^2 + (1-\lambda)\, d_2(x)^2. \tag{3.5}$$

The convergence

$$F_\lambda(\cdot,t) = -4t\,V_\lambda(\cdot,t) \to \Phi_\lambda \qquad \text{as } t \to 0^+, \tag{3.6}$$

is a manifestation of the Laplace–Varadhan principle; see [7, Sec. 6.4] and [16, Sec. 4.3]. Its rigorous presentation is stated in Section 6; see also Figure 6 for a numerical illustration.

The convergence in (3.6) reveals a fundamental structural simplification and reduction principle: in the small-time regime, the original non-autonomous dynamics is effectively governed by a *time-independent geometric energy landscape*. In other words, the generation mechanism asymptotically forgets the full analytical structure of the heat flow and retains only the geometry of the supports through Φ_λ. This provides a rigorous reduction from a high-dimensional, time-dependent score-driven system to a low-complexity autonomous dynamical system.

Remark 3.4. We emphasize that this nontrivial reduction hinges on the compact-support assumptions on the sets A_i. If, instead, both initial measures are sufficiently smooth and have full support on the whole space, the singular behavior of V_λ is no longer present. In that regime, consistently with the Li–Yau inequality [31, 35], $V_\lambda(\cdot, t)$ remains bounded as $t \to 0^+$, so that $F_\lambda(\cdot, t) = -4tV_\lambda(\cdot, t) \to 0$ as $t \to 0^+$. Thus the asymptotic reduction degenerates to the trivial zero limit.

Remark 3.5 (Semiconcavity of the geometric potential)**.** For any compact set $A \subset \mathbb{R}^d$, the squared-distance function d_A^2 is semiconcave. More precisely, $d_A^2 - \|\cdot\|^2$ is concave. Indeed, this follows from the identity

$$d_A(x)^2 = \|x\|^2 - 2\sup_{a\in A}\left(x \cdot a - \frac{1}{2}\|a\|^2\right),$$

since the supremum of affine functions is convex.

Consequently, in the MoE regime $0 \le \lambda \le 1$, the potential $\Phi_\lambda = \lambda d_1^2 + (1-\lambda)d_2^2$ is semiconcave with the same constant. This is consistent with the semiconcavity estimate for the rescaled logarithmic potentials $F_\lambda(\cdot, t)$, which is the matrix form behind the Li–Yau inequality.

By contrast, in the CFG regime $\lambda > 1$, the coefficient $1-\lambda$ is negative. Hence Φ_λ is a difference of squared-distance functions, and semiconcavity is not preserved in general. This loss of semiconcavity is one of the structural reasons why the MoE regime leads to the genuine Clarke subdifferential, whereas the CFG regime requires the outer Clarke structure introduced below.

The semiconcavity discussed above is closely connected to the Hamilton–Jacobi interpretation of the rescaled logarithmic density and of the limiting geometric potential. This connection is discussed in Section 4.1.

Clarke subdifferential. The limiting potential Φ_λ is, in general, only piecewise smooth. Its singularities arise from the non-uniqueness of nearest points in the supports A_1 and A_2, and are therefore located on Voronoi-type interfaces in the empirical case. Consequently, the limiting dynamics is not a classical gradient flow, but a generalized gradient flow. We describe it using Clarke's nonsmooth calculus [12, Thm. 2.5.1].

Definition 3.6 (Clarke subdifferential and nonsmooth interfaces)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz, and let $\mathcal{D}_f$ denote the set of points where f is differentiable. The *Clarke subdifferential* of f at $x \in \mathbb{R}^d$ is defined by

$$\partial^C f(x) := \operatorname{conv}\Big\{\lim_{k\to\infty} \nabla f(x_k) : \ x_k \in \mathcal{D}_f, \ x_k \to x\Big\}. \tag{3.7}$$

We denote by

$$\mathrm{ND}(f) := \big\{x \in \mathbb{R}^d : \partial^C f(x) \text{ is not a singleton}\big\} \tag{3.8}$$

the nonsmooth, or non-differentiability, set of f.

For the squared-distance functions

$$d_i(x) = \operatorname{dist}(x, A_i), \qquad i = 1, 2,$$

we define the combined nonsmooth interface by

$$\mathrm{ND}(A_1, A_2) := \mathrm{ND}(d_1^2) \cup \mathrm{ND}(d_2^2). \tag{3.9}$$

The set $\mathrm{ND}(A_1, A_2)$ records the points where at least one of the two squared-distance landscapes is nonsmooth. When A_1 and A_2 are finite, it is the union of the Voronoi interfaces associated with the two supports; see Figure 1.

Definition 3.7 (Outer Clarke subdifferential and critical sets)**.** For the geometric distance potential

$$\Phi_\lambda(x) = \lambda d_1(x)^2 + (1-\lambda) d_2(x)^2,$$

we define its *outer Clarke subdifferential* by

$$\widehat{\partial}\,\Phi_\lambda(x) := \lambda\,\partial^C(d_1^2)(x) + (1-\lambda)\,\partial^C(d_2^2)(x). \tag{3.10}$$

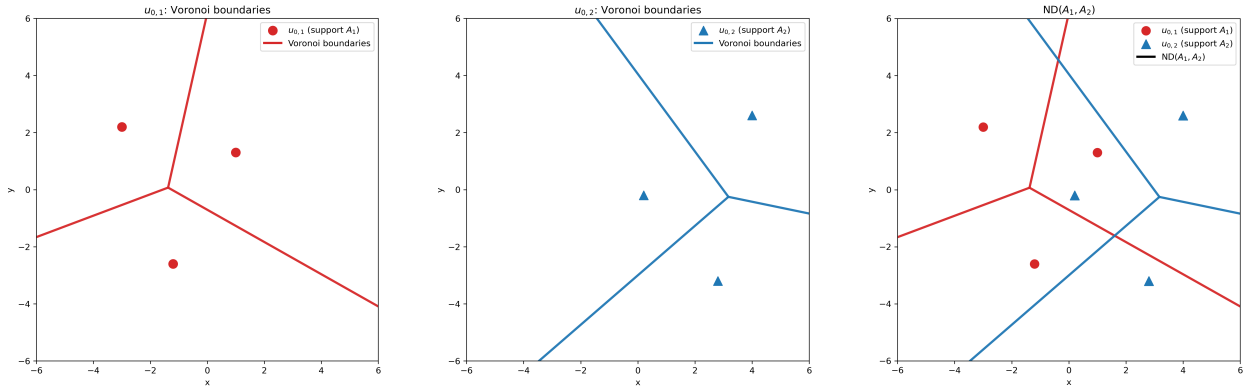
We also introduce the corresponding critical sets

$$\mathrm{Crit}(\Phi_\lambda) := \{x \in \mathbb{R}^d : \ 0 \in \partial^C \Phi_\lambda(x)\}, \tag{3.11}$$

and

$$\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda) := \{x \in \mathbb{R}^d : \ 0 \in \widehat{\partial}\,\Phi_\lambda(x)\}. \tag{3.12}$$

The terminology *outer Clarke subdifferential* is specific to the present paper. It emphasizes that $\widehat{\partial}\,\Phi_\lambda$ is formed by taking the Clarke subdifferentials of the two squared-distance functions separately and then taking their weighted Minkowski sum. In general, $\partial^C \Phi_\lambda(x) \subseteq \widehat{\partial}\,\Phi_\lambda(x)$, and the inclusion may be strict at points where d_1^2 and d_2^2 are nonsmooth, in the extrapolating regime $\lambda > 1$. This distinction is made precise in Lemma 3.8 below.



(a) $\mathrm{ND}(d_1^2)$ (b) $\mathrm{ND}(d_2^2)$ (c) $\mathrm{ND}(A_1, A_2)$

Figure 1: Non-differentiability sets associated with the empirical supports $A_1 = \mathrm{supp}(u_{0,1})$ and $A_2 = \mathrm{supp}(u_{0,2})$.

Lemma 3.8 (Clarke versus outer Clarke)**.** *Assume that A_1 and A_2 are compact subsets of $\mathbb{R}^d$. Then:*

1. *If $0 \le \lambda \le 1$, then*

$$\partial^C \Phi_\lambda(x) = \widehat{\partial}\,\Phi_\lambda(x) \qquad \forall x \in \mathbb{R}^d. \tag{3.13}$$

2. *If $\lambda > 1$, then the general Clarke sum rule gives*

$$\partial^C \Phi_\lambda(x) \subseteq \widehat{\partial}\Phi_\lambda(x) \qquad \forall x \in \mathbb{R}^d. \tag{3.14}$$

Moreover, if A_1 and A_2 are finite, then this inclusion is an equality away from the simultaneous interface:

$$\partial^C \Phi_\lambda(x) \begin{cases} = \widehat{\partial}\Phi_\lambda(x), & \forall x \in \mathbb{R}^d \setminus \big(\mathrm{ND}(d_1^2) \cap \mathrm{ND}(d_2^2)\big), \\ \subseteq \widehat{\partial}\Phi_\lambda(x), & \forall x \in \mathrm{ND}(d_1^2) \cap \mathrm{ND}(d_2^2). \end{cases} \tag{3.15}$$

Proof. The proof is given in Section 7.1. □

Remark 3.9. Lemma 3.8 highlights the structural difference between the two regimes. In the MoE case, the limiting dynamics is governed everywhere by the genuine Clarke subdifferential. In the CFG case, the natural limiting object is the larger outer Clarke subdifferential.

3.2 Time-shift limits

We now specialize to the case where the initial measures are finite Dirac mixtures:

$$A_1 = \{x_1, \dots, x_{n_1}\}, \qquad A_2 = \{y_1, \dots, y_{n_2}\},$$
$$u_{0,1} = \sum_{k=1}^{n_1} w_{1,k}\, \delta_{x_k}, \qquad u_{0,2} = \sum_{\ell=1}^{n_2} w_{2,\ell}\, \delta_{y_\ell},$$

with $w_{1,k} > 0$, $\sum_k w_{1,k} = 1$, and $w_{2,\ell} > 0$, $\sum_\ell w_{2,\ell} = 1$.

Our first main statement identifies the asymptotic autonomous limit of the non-autonomous rescaled dynamics Y.

Theorem 3.10 (Time-shift limit in the empirical case)**.** *Assume that $u_{0,1}$ and $u_{0,2}$ are finite Dirac mixtures. Fix $T > 0$. For every initial datum $x_T \in \mathbb{R}^d$, let $X = (X_t)_{t\in(0,T]}$ be the solution of (2.9) with $X_T = x_T$, and let $Y = (Y_\tau)_{\tau\geq 0}$ be the rescaled trajectory defined by (3.1).*

For every sequence $\tau_j \to \infty$, the family of time-shifted trajectories

$$Y_\tau^j := Y_{\tau+\tau_j}, \qquad \tau \geq 0,$$

is relatively compact in $C_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^d)$. Moreover, every subsequential limit

$$Y^j \to Z \qquad \text{in } C_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^d)$$

is a global Carathéodory solution (see Definition 3.15) of an autonomous limiting differential inclusion as follows:

1. ***MoE regime $0 \leq \lambda \leq 1$.*** *The limit curve Z satisfies the exact Clarke gradient inclusion*

$$\dot{Z}_\tau \in -\frac{1}{4}\, \partial^C \Phi_\lambda(Z_\tau) \qquad \text{for a.e. } \tau \geq 0. \tag{3.16}$$

2. ***CFG regime $\lambda > 1$.*** *The limit curve Z satisfies the outer Clarke gradient inclusion*

$$\dot{Z}_\tau \in -\frac{1}{4}\, \widehat{\partial}\, \Phi_\lambda(Z_\tau) \qquad \text{for a.e. } \tau \geq 0. \tag{3.17}$$

Proof. The proof of Theorem 3.10 is deferred to Section 6.4. It relies on the Laplace–Varadhan asymptotics for F_λ, on the mean-shift representation of the score, and on a compactness argument for time shifts in similarity time. □

Remark 3.11 (Existence of solutions to the limiting inclusion)**.** Since A_1 and A_2 are compact, both multifunctions

$$x \longmapsto -\frac{1}{4}\,\partial^C \Phi_\lambda(x) \qquad \text{and} \qquad x \longmapsto -\frac{1}{4}\,\widehat{\partial}\,\Phi_\lambda(x)$$

are upper semicontinuous, with nonempty, convex, compact values and at most linear growth. Hence, by the Viability Theorem [3, Thm. 10.1.6], each associated differential inclusion admits at least one global Carathéodory solution from every initial point. A more rigorous version is found in Lemma 7.1.

Remark 3.12 (Convergence implies criticality)**.** If the original trajectory X_t converges as $t \to 0^+$ to some point x^*, then every time-shift limit is necessarily the constant curve $Z_\tau \equiv x^*$. Therefore, by Theorem 3.10,

$$0 \in \partial^C \Phi_\lambda(x^*) \qquad \text{if } 0 \le \lambda \le 1,$$

whereas

$$0 \in \widehat{\partial}\,\Phi_\lambda(x^*) \qquad \text{if } \lambda > 1.$$

In particular, any convergent generation trajectory must converge to a critical point of the corresponding limiting nonsmooth dynamics.

Although the convergence assumption is strong, it is consistent with the numerical behavior observed in Figures 6, 7, and 8.

Remark 3.13 (Piecewise-affine structure of the limiting field)**.** Figure 2a shows the vector field appearing in (3.16) and (3.17) outside the interface set $\mathrm{ND}(A_1, A_2)$. The black stars mark the local minimizers of Φ_λ.

On this smooth region, the nearest points of x in A_1 and A_2 are uniquely determined; we denote them by a_1 and a_2, respectively. The limiting inclusion then reduces to the classical gradient-flow field

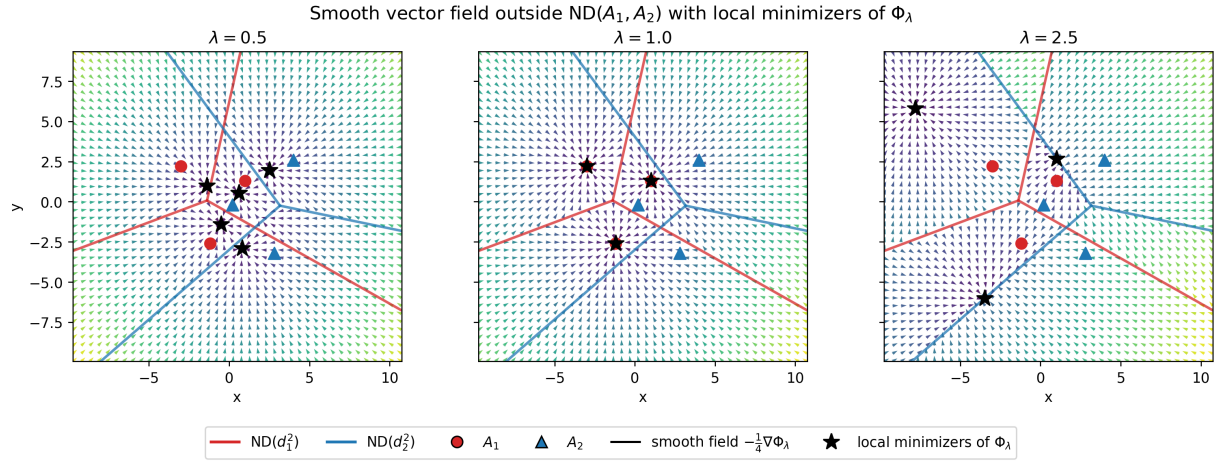
$$-\frac{1}{4}\nabla \Phi_\lambda(x) = -\frac{1}{2}\big(x - \lambda a_1 - (1-\lambda)a_2\big).$$

Since a_1 and a_2 remain constant on each smooth cell, this vector field is affine on each cell. Consequently, the limiting dynamics is piecewise affine on $\mathbb{R}^d \setminus \mathrm{ND}(A_1, A_2)$.

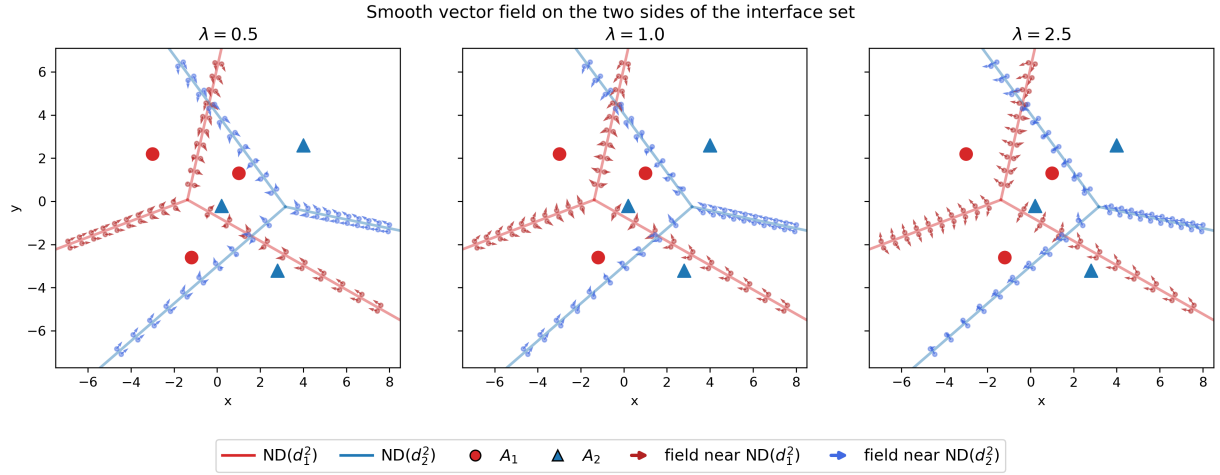
Figure 2b complements this picture near the interface set by plotting the same smooth field on the two sides of $\mathrm{ND}(A_1, A_2)$. In the MoE regime $0 \le \lambda \le 1$, the behavior is relatively simple: on the two sides of an interface, the vector fields either point away from the interface in opposite directions, or cross it with similar directions. In particular, no sliding phenomenon is expected. By contrast, in the CFG regime $\lambda > 1$, some interfaces have vector fields on both sides pointing toward the interface. This creates a sliding phenomenon, which makes the analysis more delicate; see [14, Fig. 8].

Finally, Figure 2c displays the three-dimensional landscape of the geometric potential Φ_λ for the same values of λ. The MoE landscapes retain the semiconcave squared-distance structure, whereas the CFG landscape exhibits stronger extrapolative and interface effects. This helps visualize the qualitative difference between the genuine Clarke dynamics in the MoE regime and the outer-Clarke dynamics in the CFG regime.

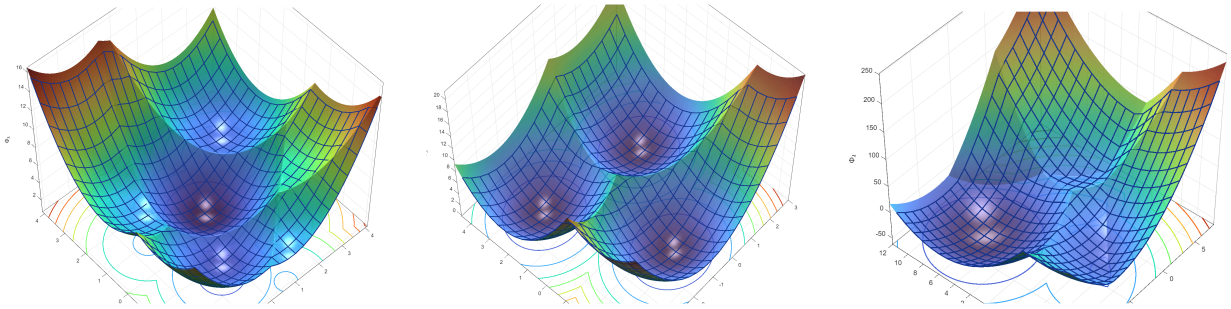
Remark 3.14 (Time-shift limits beyond the empirical setting)**.** Theorem 3.10 shows that every time-shift limit of the rescaled flow Y satisfies the corresponding autonomous limiting inclusion: (3.16) in the MoE regime and (3.17) in the CFG regime. Passing to time-shift limits is a classical tool in the asymptotic analysis of non-autonomous systems; see [19, Thm. 3] and [20].



(a) Smooth vector field $-\frac{1}{4}\nabla\Phi_\lambda$ outside the interface set $\mathrm{ND}(A_1, A_2)$. The black stars mark the local minimizers of Φ_λ.



(b) Smooth vector field $-\frac{1}{4}\nabla\Phi_\lambda$ plotted on the two sides of the interface set $\mathrm{ND}(A_1, A_2)$.



$\lambda = 0.5$ $\lambda = 1$ $\lambda = 2.5$

(c) 3D visualization of the geometric distance potential Φ_λ for three values of λ.

Figure 2: Vector fields of the limiting dynamics and the associated potential landscape.

Although the theorem is stated in the empirical setting, its proof only uses the quantitative Laplace–Varadhan estimates derived under Assumption A, introduced in Section 6. Consequently, the same time-shift conclusion remains valid for compactly supported probability measures satisfying this uniform lower small-ball mass condition. This class includes, in particular, finite Dirac mixtures ($\alpha = 0$); absolutely continuous measures on full-dimensional compact supports whose densities are bounded from below by a positive constant on their supports ($\alpha = d$); and measures supported on lower-dimensional manifolds, provided they have a uniformly positive density with respect to the corresponding intrinsic volume measure.

3.3 Convergence of the autonomous limiting system

Theorem 3.10 shows that every time-shift limit of the rescaled trajectory Y solves an autonomous limiting differential inclusion driven by the geometric potential Φ_λ. We now describe the qualitative dynamics of this limiting system in the empirical setting.

We begin by recalling the notion of solution for the differential inclusions.

Definition 3.15 (Global Carathéodory solution)**.** Let $\mathcal{F} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map and let $z_0 \in \mathbb{R}^d$. A curve

$$Z : [0, \infty) \to \mathbb{R}^d$$

is called a global *Carathéodory solution* of

$$\dot{Z}_\tau \in \mathcal{F}(Z_\tau) \qquad \text{for a.e. } \tau \geq 0, \qquad Z_0 = z_0,$$

if Z is absolutely continuous and there exists a measurable selection

$$\xi(\tau) \in \mathcal{F}(Z_\tau) \qquad \text{for a.e. } \tau \geq 0$$

such that

$$Z_\tau = z_0 + \int_0^\tau \xi(s)\, ds \qquad \forall \tau \geq 0.$$

In the empirical case, the geometric potential Φ_λ is piecewise quadratic, and this rigid structure is strong enough to force every autonomous limiting trajectory to converge to a single critical point.

Theorem 3.16 (Convergence of the autonomous limiting system in the empirical setting)**.** *Assume that $u_{0,1}$ and $u_{0,2}$ are finite Dirac mixtures, and fix $z_0 \in \mathbb{R}^d$. Then there exists at least one global Carathéodory solution starting from z_0 of the corresponding autonomous limiting inclusion, namely (3.16) in the MoE regime $0 \leq \lambda \leq 1$, and (3.17) in the CFG regime $\lambda > 1$.*

Moreover, every such global solution Z satisfies the following properties:

1. ***Lyapunov monotonicity.*** *For every $0 \leq a < b < \infty$,*

$$\Phi_\lambda(Z_b) - \Phi_\lambda(Z_a) = -4 \int_a^b \|\dot{Z}_\tau\|^2 \, d\tau.$$

In particular, the map $\tau \mapsto \Phi_\lambda(Z_\tau)$ is nonincreasing on $[0, \infty)$.

2. ***Convergence.*** *There exists $x^* \in \mathbb{R}^d$ such that*

$$Z_\tau \to x^* \qquad \text{as } \tau \to \infty.$$

Moreover, with the notation $\mathrm{Crit}(\Phi_\lambda)$ and $\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$ introduced in (3.11),

$$\begin{aligned} &x^* \in \mathrm{Crit}(\Phi_\lambda), && \text{if } 0 \leq \lambda \leq 1; \\ &x^* \in \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda), && \text{if } \lambda > 1. \end{aligned}$$

Proof. The proof is deferred to Section 7.4. It relies on the piecewise quadratic structure of Φ_λ in the empirical setting, on the finiteness of the critical set, and on the strict Lyapunov identity along Carathéodory solutions. □

Remark 3.17 (Role of the autonomous limiting system)**.** Theorem 3.16 shows that, in the empirical setting, every global solution of the autonomous limiting inclusion converges to a single critical point of the geometric potential Φ_λ. Thus the critical geometry of the time-independent landscape Φ_λ determines the possible asymptotic states of all time-shift limits of the rescaled generation flow.

This result explains the central role of Φ_λ. It does not, by itself, imply convergence of the original non-autonomous trajectory, since different time shifts could in principle select different limiting critical points. It does, however, reduce the remaining question to a selection problem: which critical points of Φ_λ can be reached by the original rescaled flow?

3.4 Convergence criterion and rates in the empirical setting

We now pass from the autonomous limiting inclusions back to the original non-autonomous generation flow in the empirical setting. The previous theorem shows that every time-shift limit of the rescaled trajectory Y converges to a critical point of Φ_λ, in the Clarke sense in the MoE regime and in the outer Clarke sense in the CFG regime. This asymptotic information is not yet enough to force convergence of Y itself: a priori, distinct time shifts could converge to distinct critical points.

The convergence criterion below rules out this possibility under a natural selection assumption. If the ω-limit set of Y contains no non-minimizing critical point, then the only possible limiting states are local minimizers of Φ_λ. The local trap property of such minimizers, established in Lemma 8.2, then prevents the trajectory from moving between different minimizing basins and yields convergence to a single local minimizer.

We denote by

$$\mathrm{Min}(\Phi_\lambda) := \{x \in \mathbb{R}^d : \ x \text{ is a local minimizer of } \Phi_\lambda\}.$$

Theorem 3.18 (Convergence criterion in the empirical setting)**.** *Assume that $u_{0,1}$ and $u_{0,2}$ are finite Dirac mixtures, and fix $T > 0$. For every initial datum $x_T \in \mathbb{R}^d$, let $X = (X_t)_{t\in(0,T]}$ be the solution of (2.9) with $X_T = x_T$, and let Y be the rescaled trajectory defined by (3.1).*

Assume that the ω-limit set of Y contains no non-minimizing critical point of the corresponding limiting notion. More precisely:

- *in the MoE regime $0 \le \lambda \le 1$,*

$$\omega_\tau(Y) \cap \big(\mathrm{Crit}(\Phi_\lambda) \setminus \mathrm{Min}(\Phi_\lambda)\big) = \varnothing;$$

- *in the CFG regime $\lambda > 1$,*

$$\omega_\tau(Y) \cap \big(\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda) \setminus \mathrm{Min}(\Phi_\lambda)\big) = \varnothing.$$

Then there exists $x^* \in \mathrm{Min}(\Phi_\lambda)$ such that

$$Y_\tau \to x^* \qquad \text{as } \tau \to \infty,$$

equivalently,

$$X_t \to x^* \qquad \text{as } t \to 0^+.$$

Remark 3.19 (On the conditional convergence criterion)**.** Theorem 3.18 is a conditional convergence result. Its assumption excludes the presence of non-minimizing critical points in the ω-limit

set of the particular rescaled trajectory Y. In the MoE regime, criticality is understood with respect to $\mathrm{Crit}(\Phi_\lambda)$, while in the CFG regime it is understood with respect to $\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$. Thus the hypothesis is a condition on the trajectory, not a property proved here.

The only possible obstruction comes from the nonsmooth interface. Indeed, on each smooth cell of the empirical decomposition, Φ_λ is quadratic with Hessian $2I$, so every smooth critical point is a strict local minimizer. Therefore any non-minimizing critical point must lie in $\mathrm{ND}(A_1, A_2)$. Proving that such points are avoided, for instance for Lebesgue-almost-every terminal datum x_T, would yield a generic unconditional convergence result. This remains open. The numerical experiments in Section 5 are consistent with the conditional hypothesis being satisfied along the simulated trajectories.

Corollary 3.20 (Rate near a smooth local minimizer in the empirical setting)**.** *Assume that the hypotheses of Theorem 3.18 hold, and let $x^* \in \mathrm{Min}(\Phi_\lambda)$ be the limit point of the rescaled trajectory Y.*

1. *In the MoE regime $0 \le \lambda \le 1$, the point x^* is automatically a smooth minimizer of Φ_λ. Then there exists a constant $C > 0$ such that*

$$\|Y_\tau - x^*\| \le Ce^{-\tau/2} \qquad \forall \tau \ge 0.$$

   Equivalently,

$$\|X_t - x^*\| \le C\sqrt{t} \qquad \forall t \in (0, T].$$

2. *In the CFG regime $\lambda > 1$, assume in addition that $x^* \notin \mathrm{ND}(A_1, A_2)$. Then there exists a constant $C > 0$ such that*

$$\|Y_\tau - x^*\| \le Ce^{-\tau/2} \qquad \forall \tau \ge 0.$$

   Equivalently,

$$\|X_t - x^*\| \le C\sqrt{t} \qquad \forall t \in (0, T].$$

Proof. The proofs are deferred to Section 8. They rely on the compactness approach for asymptotically autonomous systems described there: first, local minimizers are shown to be local traps for the autonomous limiting inclusions; next, this trap structure is combined with time-shift compactness and the convergence of limiting trajectories to prove Theorem 3.18; finally, Corollary 3.20 follows from a local perturbative argument near a smooth stable minimizer. □

Remark 3.21 (Endpoint regimes and pure imitation)**.** If $\lambda = 0$ (resp. $\lambda = 1$), then the mixed score reduces to the second (resp. first) expert. In this case, every local minimizer of Φ_λ belongs to A_2 (resp. A_1), corresponding to pure imitation of the selected dataset. Moreover, the rate in Corollary 3.20 is consistent with the one obtained in [35, Thm. 3.11].

Remark 3.22 (Rates at interface minimizers in the CFG regime)**.** In the CFG regime $\lambda > 1$, Corollary 3.20 establishes the rate $\mathcal{O}(\sqrt{t})$ only when the limiting minimizer x^* is a smooth point of Φ_λ, namely when $x^* \notin \mathrm{ND}(A_1, A_2)$. No such rate is proved here for minimizers lying on the interface set $\mathrm{ND}(A_1, A_2)$. This restriction is genuine: as illustrated in Figure 4, for sufficiently large λ, CFG local minimizers may occur precisely on this interface. In that case, the present analysis yields qualitative convergence, under the hypothesis of Theorem 3.18, but not a quantitative rate. Establishing sharp convergence rates at interface minimizers, possibly under suitable transversality assumptions on the Voronoi geometry, remains an open problem.

3.5 Geometric interpretation in the MoE and CFG regimes

We conclude this section with a geometric interpretation of the two guidance regimes. Recall that A_1 and A_2 denote the supports of the initial measures $u_{0,1}$ and $u_{0,2}$, and that

$$\Phi_\lambda(x) = \lambda d_1(x)^2 + (1-\lambda) d_2(x)^2$$

depends only on A_1, A_2, and the mixing parameter λ. The preceding results show that, in the small-time regime, the generation dynamics is organized by this geometric landscape: time-shift limits are governed by the critical structure of Φ_λ, while local minimizers represent the stable states selected by the dynamics.

MoE regime $0 \le \lambda \le 1$. In this regime,

$$\Phi_\lambda(x) = \min_{y_1 \in A_1,\ y_2 \in A_2} \Big(\lambda \|x - y_1\|^2 + (1-\lambda)\|x - y_2\|^2 \Big).$$

Thus Φ_λ is a cooperative combination of the squared distances to the two supports. Its local minimizers may be interpreted as geometric barycenters balancing proximity to both datasets.

The MoE regime therefore has a comparatively rigid geometric structure. The nonsmoothness of Φ_λ arises only from projection-switching interfaces, while the potential retains the semiconcavity of squared-distance functions. Consequently, the limiting autonomous dynamics is governed by the genuine Clarke subgradient flow of Φ_λ. Hence, under the conditional criterion of Theorem 3.18, the original generation flow converges to a local minimizer of Φ_λ.

If $A_1 \cap A_2 \neq \varnothing$, then every point in the overlap is a global minimizer of Φ_λ. If the supports are disjoint, the minimizers are typically located between the two supports rather than on either one. In this case, the generated states interpolate geometrically between the two datasets. Figure 3 illustrates this behavior in a two-dimensional empirical example.

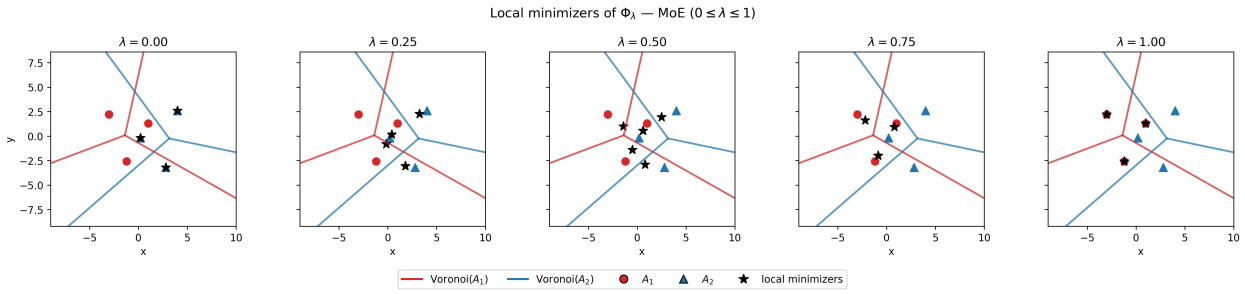


Figure 3: Local minimizers of Φ_λ in the MoE regime. As λ increases from 0 to 1, the minimizers trace a geometric interpolation from A_2 toward A_1.

CFG regime $\lambda > 1$. In the CFG regime,

$$\Phi_\lambda(x) = \lambda d_1(x)^2 - (\lambda - 1) d_2(x)^2 = \min_{y_1 \in A_1} \max_{y_2 \in A_2} \Big(\lambda \|x - y_1\|^2 - (\lambda - 1)\|x - y_2\|^2 \Big).$$

The potential is therefore no longer a cooperative average. It has an extrapolative, competitive structure: the dynamics is attracted toward the target support A_1 while being pushed away from the nearest regions of A_2. This provides a geometric interpretation of classifier-free guidance, where

A_1 represents the target conditional structure and A_2 plays the role of a broader background or reference distribution.

This extrapolative structure also makes the CFG regime more singular than the MoE regime. The negative coefficient in front of d_2^2 destroys the semiconcavity mechanism in general, and the nonsmooth interfaces may become active components of the limiting dynamics. This is why the asymptotic description naturally involves the larger outer Clarke structure rather than only the genuine Clarke subdifferential.

In particular, if $A_1 \subset A_2$, then the set of minimizers of Φ_λ is exactly A_1. Thus, the dynamics is naturally biased toward points in A_1. This provides a geometric interpretation of guidance: the flow favors configurations that remain compatible with the target class A_1 while being repelled from features that are only characteristic of the broader class A_2. This is consistent with the intuition behind CFG in diffusion models, as introduced in [23]: one starts from a more general generative model and guides it toward a target conditional class in order to improve the fidelity and visual quality of the generated samples.

Figure 4 illustrates the local minimizers of Φ_λ in the CFG regime for the same finite Dirac-mixture example. As λ increases, minimizers may approach the interface $\mathrm{ND}(A_1, A_2)$ and then evolve along it, reflecting the stronger role of nonsmooth geometry in the extrapolative regime.

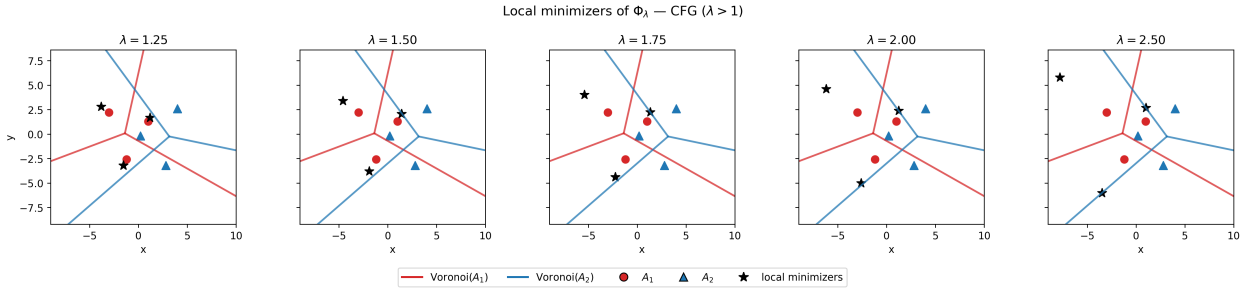


Figure 4: Local minimizers of Φ_λ in the CFG regime. As λ increases, the minimizers may touch the interface $\mathrm{ND}(A_1, A_2)$ and then evolve along it.

In summary, the MoE regime has a cooperative geometric structure, leading to barycentric interpolation between the two supports and to a relatively rigid Clarke gradient dynamics. The CFG regime has a competitive extrapolative structure, combining attraction toward A_1 with repulsion from the nearest regions of A_2. This destroys semiconcavity in general and amplifies interface effects, which is reflected in the appearance of the larger outer Clarke structure in the limiting dynamics.

4 PDE and SDE viewpoints: Hamilton–Jacobi structure, energy estimates, and stochastic approximation

The previous sections focused on the deterministic generation dynamics and its asymptotic reduction, via similarity-time rescaling, to autonomous nonsmooth systems driven by the geometric potential Φ_λ. In particular, we showed that time-shift limits are governed by the Clarke differential inclusion in the MoE regime and by the outer Clarke inclusion in the CFG regime. In the empirical setting, the associated autonomous limiting systems converge to critical points of Φ_λ. Under the additional assumption that the ω-limit set contains no non-minimizing critical point, this yields convergence of the original generation flow toward a local minimizer, together with explicit rates in the smooth stable case.

The goal of the present section is to complement this deterministic geometric analysis with PDE, Hamilton–Jacobi, and stochastic viewpoints. We first explain how the heat-flow Hessian bounds imply Li–Yau estimates and, after logarithmic rescaling, uniform semiconcavity of the potential $F = -4t \log u$. The same rescaled potential satisfies a non-autonomous viscous Hamilton–Jacobi equation, whose small-time limit is consistent with the squared-distance eikonal structure of the geometric potential. We then use the Li–Yau bounds to derive L^p-energy estimates for the backward Fokker–Planck equation with mixed score drift, revealing a sharp stability difference between the MoE and CFG regimes. Finally, we rewrite the noisy generation process in similarity-time variables and interpret it as a vanishing-viscosity perturbation of the limiting geometric dynamics.

4.1 Li–Yau, semiconcavity, and Hamilton–Jacobi structure

We first describe the Hamilton–Jacobi structure underlying the rescaled logarithmic potential. For clarity, we begin with a single heat flow.

Let u solve the heat equation with compactly supported initial measure u_0, and set

$$A = \operatorname{supp}(u_0).$$

For $t > 0$, define the rescaled logarithmic potential

$$F(x,t) := -4t \log u(x,t).$$

By the matrix lower bound established in [35, Lem. 5.1], one has

$$\operatorname{Hess}(\log u(x,t)) \succeq -\frac{1}{2t} I_d.$$

Taking traces yields the classical Li–Yau inequality [31]

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

Equivalently, in terms of $F = -4t \log u$, the same estimate becomes

$$\operatorname{Hess}(F(x,t)) \preceq 2I_d, \qquad \Delta F(x,t) \leq 2d.$$

Thus $F(\cdot,t)$ is semiconcave with constant 2, uniformly for $t > 0$. In this sense, the Li–Yau inequality is the trace form of a stronger semiconcavity estimate for the rescaled logarithmic potential.

The same potential also satisfies a viscous Hamilton–Jacobi equation. Indeed, introducing the logarithmic variable

$$V = \log u,$$

namely the Hopf–Cole transform of the positive heat solution u, the heat equation becomes

$$\partial_t V = \Delta V + |\nabla V|^2, \qquad t > 0.$$

Since $F = -4tV$, a direct computation gives

$$\partial_t F = \frac{F}{t} + \Delta F - \frac{1}{4t}|\nabla F|^2, \qquad t > 0.$$

Therefore F is a classical solution of a non-autonomous viscous Hamilton–Jacobi equation.

On the other hand, the Laplace–Varadhan principle yields the local uniform convergence

$$F(\cdot,t) \to d_A^2 \qquad \text{as } t \to 0^+,$$

where $d_A(x) = \operatorname{dist}(x, A)$; see Lemma 6.2. The limiting profile d_A^2 retains the same semiconcavity constant 2; see Remark 3.5.

The limiting profile is also related to the eikonal equation. More precisely, d_A is the viscosity solution of

$$|\nabla d_A| = 1$$

away from the target set A, with $d_A = 0$ on A, in the standard target-set sense; see [4, Ch. II] and [15]. Consequently, the selected squared-distance profile

$$\Phi = d_A^2$$

satisfies

$$|\nabla \Phi|^2 = 4\Phi$$

in the viscosity sense. This stationary Hamilton–Jacobi relation can also be viewed as the formal small-time limit of the equation satisfied by F. Indeed, rewriting the viscous Hamilton–Jacobi equation as

$$|\nabla F|^2 = 4F - 4t\,\partial_t F + 4t\,\Delta F,$$

and formally letting $t \to 0^+$ at the level of the equation, one obtains again

$$|\nabla \Phi|^2 = 4\Phi.$$

Altogether, the discussion is summarized in Figure 5.

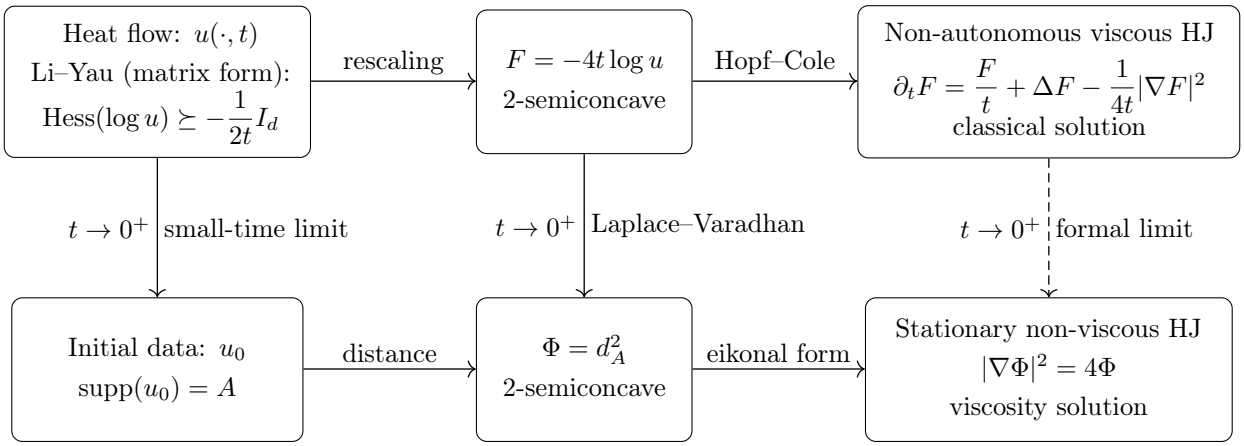


Figure 5: Li–Yau, semiconcavity, and Hamilton–Jacobi structure for the rescaled logarithmic potential.

4.2 Energy estimate in the mixed case

We now return to the mixed-score setting. Recall the backward Fokker–Planck equation with noise (2.7):

$$\begin{cases} \partial_t \rho_\varepsilon + \varepsilon\, \Delta \rho_\varepsilon - (1+\varepsilon)\, \mathrm{div}\big(\rho_\varepsilon\, \nabla V_\lambda\big) = 0, & (x,t) \in \mathbb{R}^d \times (0,T), \\ \rho_\varepsilon(\cdot, T) = v_T, & x \in \mathbb{R}^d, \end{cases} \tag{4.1}$$

where

$$\nabla V_\lambda = \lambda \nabla \log u_1 + (1-\lambda) \nabla \log u_2.$$

By the Li–Yau bound recalled above, in the MoE regime $0 \le \lambda \le 1$, the convexity of the coefficients gives

$$\Delta V_\lambda(x,t) \ge -\frac{d}{2t}, \qquad \forall (x,t) \in \mathbb{R}^d \times (0,\infty). \tag{4.2}$$

In the CFG regime $\lambda > 1$, the coefficient $1-\lambda$ is negative. Thus the lower Li–Yau bound for $\log u_2$ is no longer sufficient, and one also needs an upper bound. Assume that $u_{0,2}$ is compactly supported, and let

$$R = \sup\{\|x\| : \ x \in \mathrm{supp}(u_{0,2})\}.$$

By the matrix upper estimate in [35, Lem. 5.1],

$$\mathrm{Hess}(\log u_2(x,t)) \preceq \left(-\frac{1}{2t} + \frac{R^2}{4t^2}\right) I_d.$$

Taking traces gives

$$\Delta \log u_2(x,t) \le -\frac{d}{2t} + \frac{dR^2}{4t^2}.$$

Combining this estimate with the Li–Yau lower bound for u_1, we obtain

$$\Delta V_\lambda(x,t) \ge -\frac{d}{2t} - (\lambda-1)\frac{dR^2}{4t^2}, \tag{4.3}$$

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

Based on these lower bounds on ΔV_λ, we obtain the following L^p-estimate.

Theorem 4.1 (Energy estimate)**.** *Let u_1 and u_2 be the solutions of the heat equations (2.1) with initial data $u_{0,1}$ and $u_{0,2}$, both probability measures. Fix $p \in [1,\infty)$, and let $v_T \in L^p(\mathbb{R}^d)$.*

- ***MoE regime $0 \le \lambda \le 1$.*** *Equation (4.1) admits a unique solution*

$$\rho_\varepsilon \in C\big((0,T]; L^p(\mathbb{R}^d)\big),$$

and

$$\|\rho_\varepsilon(t)\|_{L^p} \le \left(\frac{T}{t}\right)^{\frac{d(1+\varepsilon)(p-1)}{2p}} \|v_T\|_{L^p}, \qquad \forall t \in (0,T]. \tag{4.4}$$

- ***CFG regime $\lambda > 1$.*** *Assume that $u_{0,2}$ has compact support, and let R be its radius. Then equation (4.1) admits a unique solution*

$$\rho_\varepsilon \in C\big((0,T]; L^p(\mathbb{R}^d)\big),$$

and

$$\|\rho_\varepsilon(t)\|_{L^p} \le \left(\frac{T}{t}\right)^{\frac{d(1+\varepsilon)(p-1)}{2p}} \exp\left(\frac{(1+\varepsilon)(p-1)(\lambda-1)dR^2}{4p}\left(\frac{1}{t}-\frac{1}{T}\right)\right) \|v_T\|_{L^p}, \qquad \forall t \in (0,T]. \tag{4.5}$$

Proof. The MoE estimate (4.4) follows by the energy method of [35, Thm. 3.1], where the Li–Yau lower bound $\Delta V_\lambda \geq -d/(2t)$ in (4.2) is used to control the divergence term in the time derivative of $\|\rho_\varepsilon(t)\|_{L^p}^p$.

The CFG case is genuinely different: the lower bound (4.3) contains an additional $1/t^2$-singular term, and the corresponding energy inequality must be integrated explicitly. Arguing as in the proof of [35, Thm. 3.1], but using (4.3), we obtain

$$\frac{d}{dt}\|\rho_\varepsilon(t)\|_{L^p}^p \geq -\left(\frac{A}{t}+\frac{B}{t^2}\right)\|\rho_\varepsilon(t)\|_{L^p}^p,$$

where

$$A=\frac{d(1+\varepsilon)(p-1)}{2}, \qquad B=\frac{(1+\varepsilon)(p-1)(\lambda-1)dR^2}{4}.$$

Dividing by $\|\rho_\varepsilon(t)\|_{L^p}^p$, we get

$$\frac{d}{dt}\log\|\rho_\varepsilon(t)\|_{L^p}^p \geq -\frac{A}{t}-\frac{B}{t^2}.$$

Integrating from t to T, we infer

$$\log\|v_T\|_{L^p}^p - \log\|\rho_\varepsilon(t)\|_{L^p}^p \geq -A\log\left(\frac{T}{t}\right) - B\left(\frac{1}{t}-\frac{1}{T}\right).$$

Exponentiating yields

$$\|\rho_\varepsilon(t)\|_{L^p}^p \leq \left(\frac{T}{t}\right)^A \exp\left(B\left(\frac{1}{t}-\frac{1}{T}\right)\right)\|v_T\|_{L^p}^p.$$

Taking the p-th root gives (4.5). □

Remark 4.2 (MoE versus CFG: a stability perspective)**.** The previous estimates reveal a qualitative difference between the MoE and CFG regimes. In the MoE case $0 \leq \lambda \leq 1$, the L^p-bound grows at most polynomially as $t \to 0^+$. In the CFG case $\lambda > 1$, the estimate contains the additional singular factor

$$\exp\left(C\left(\frac{1}{t}-\frac{1}{T}\right)\right),$$

which reflects a substantially stronger instability near $t = 0$.

This is consistent with the asymptotic dynamical picture obtained in the previous section. In the deterministic case $\varepsilon = 0$, the backward characteristic dynamics in the MoE regime is asymptotically governed by the genuine Clarke gradient flow associated with Φ_λ. By contrast, in the CFG regime, the asymptotic description requires the larger outer Clarke differential inclusion, which allows more possible limiting vector fields, especially near nonsmooth interfaces. Thus, both at the PDE level through the energy estimates and at the dynamical level through the limiting inclusions, the CFG regime exhibits a less rigid and less stable behavior than the MoE regime.

Remark 4.3 (On the lack of an entropy estimate in the mixing setting)**.** In the non-mixing case, one can derive an entropy estimate that is stronger than the L^p-energy estimate, since it is independent of the space dimension; see [35, Sec. 3.2]. More precisely, one proves that the Kullback–Leibler divergence between the solution of the backward generative Fokker–Planck equation and the corresponding solution of the original heat equation is nonincreasing backward in time. This contraction property plays a key role in showing that the generative distribution concentrates toward the support of the initial data.

Such an argument is no longer directly available in the present mixing framework. Indeed, in general there is no natural reference probability density associated with the mixed score field. A natural candidate would be $u_1^\lambda u_2^{1-\lambda}$, but this function is not, in general, normalized. In the MoE regime $0 \le \lambda \le 1$, its total mass is bounded by 1 by Hölder's inequality, while in the CFG regime $\lambda > 1$, it may even have mass larger than 1. Moreover, this product satisfies a reaction–diffusion equation and therefore does not evolve by a mass-preserving dynamics. Consequently, the relative entropy with respect to this candidate no longer has a direct probabilistic meaning, and the contraction argument used in the non-mixing case does not apply.

This is one of the reasons why, in the present work, we adopt a geometric approach to the convergence of generative flows. We focus on the deterministic generation dynamics, for which the asymptotic analysis is cleaner and more transparent. Extending this geometric picture to the stochastic setting remains a natural direction for future work. We now provide a first step in this direction by rewriting the noisy generation process in similarity time, where it appears as a vanishing-viscosity perturbation of the limiting geometric dynamics.

4.3 Diffusive generation and stochastic approximation

As in the deterministic case, we perform the similarity-time change of variables for (4.1),

$$\tau = \log(T/t), \qquad t = Te^{-\tau}, \qquad \tilde\rho_\varepsilon(y,\tau) := \rho_\varepsilon\big(y, Te^{-\tau}\big).$$

A direct computation gives

$$\partial_\tau \tilde\rho_\varepsilon(y,\tau) = -t\,\partial_t \rho_\varepsilon(y,t)\big|_{t=Te^{-\tau}},$$

so that $\tilde\rho_\varepsilon$ solves

$$\begin{cases} \partial_\tau \tilde\rho_\varepsilon - \varepsilon\, Te^{-\tau}\Delta\tilde\rho_\varepsilon - \dfrac{1+\varepsilon}{4}\operatorname{div}\Big(\tilde\rho_\varepsilon\, \nabla F_\lambda(\cdot, Te^{-\tau})\Big) = 0, \quad (y,\tau)\in\mathbb{R}^d\times(0,\infty), \\ \tilde\rho_\varepsilon(\cdot,0) = v_T, \end{cases} \tag{4.6}$$

where

$$F_\lambda(x,t) = -4t\, V_\lambda(x,t)$$

is the rescaled potential.

The corresponding stochastic generation flow is

$$\begin{cases} dY_{\varepsilon,\tau} = -\dfrac{1+\varepsilon}{4}\,\nabla F_\lambda(Y_{\varepsilon,\tau}, Te^{-\tau})\,d\tau + \sqrt{2\varepsilon T}\,e^{-\tau/2}\,dW_\tau, \qquad \tau>0, \\ Y_{\varepsilon,0}\sim v_T. \end{cases} \tag{4.7}$$

This is the similarity-time formulation of the original stochastic generation process (2.6). Here, $(W_\tau)_{\tau\ge0}$ denotes a standard Brownian motion in similarity time, obtained from the original Brownian motion by the usual deterministic time change; the two formulations are equivalent in distribution.

The key feature of (4.6) and (4.7) is the exponentially decaying diffusion strength

$$\varepsilon T e^{-\tau} \to 0 \qquad \text{as } \tau\to\infty.$$

Thus the similarity-time scaling turns the noisy generation process into a vanishing-viscosity perturbation of the deterministic rescaled dynamics. At the same time, the drift becomes asymptotically

geometric: by Lemma 6.3, the field $\nabla F_\lambda(\cdot, t)$ converges locally toward the outer Clarke subdifferential $\widehat{\partial}\Phi_\lambda$ as $t \to 0^+$; moreover, by Lemma 6.4, it converges locally uniformly to $\nabla\Phi_\lambda$ away from the nondifferentiability set $\mathrm{ND}(A_1, A_2)$.

When the drift of an SDE is autonomous and the noise intensity vanishes, the problem falls within the scope of stochastic approximation theory; see, for instance, the classical convergence results for stochastic gradient algorithms in [44] and the asymptotic pseudotrajectory framework in [5]. In particular, the exponential decay rate $e^{-\tau}$ is much faster than the regimes considered in [5, Prop. 4.1]. Therefore, if the drift were already autonomous, one would expect the stochastic dynamics to share the same late-time behavior as its deterministic counterpart.

In the present setting, however, the drift itself varies simultaneously and converges only toward a nonsmooth limiting regime. This coupled evolution of the drift and the noise makes the rigorous stochastic analysis substantially more delicate. We do not pursue such a theory at the level of theorems in this paper. Rather, the discussion above suggests a natural future strategy: introduce vanishing-noise perturbations of the autonomous differential inclusions (3.16) and (3.17), and use them as intermediate stochastic approximations of the limiting systems. The relevant technical framework is the stochastic approximation theory for differential inclusions developed by Benaïm–Hofbauer–Sorin [6]. A precise convergence theorem for the noisy rescaled dynamics (4.7) toward a Carathéodory solution of the limiting inclusion remains a natural direction for future work.

5 Numerical simulations

This section provides numerical illustrations of the main results. We focus on three aspects. First, we visualize the geometric potential Φ_λ and its connection with mixed heat-flow scores through the Laplace–Varadhan principle, presented in Section 6. Second, we examine the convergence of the deterministic backward generation flow (2.9), in connection with the main results of Section 3, and also illustrate its stochastic counterpart (2.6), discussed in Section 4.3. Finally, we investigate the effect of guidance on a real image dataset, CIFAR–10.

5.1 The finite Dirac mixtures setting

We first focus on the empirical setting and present one- and two-dimensional experiments in order to make the underlying phenomena more transparent.

In this framework, when the initial distribution takes the form

$$u_0 = \sum_{k=1}^{n} w_k\, \delta_{x_k}, \qquad w_k > 0, \quad \sum_{k=1}^{n} w_k = 1,$$

the corresponding solution of the heat equation with initial data u_0 is the Gaussian mixture

$$u(x,t) = \sum_{k=1}^{n} w_k\, G_t(x - x_k), \qquad G_t(z) = (4\pi t)^{-d/2} \exp\Big(-\frac{\|z\|^2}{4t}\Big). \tag{5.1}$$

Using the identity

$$\nabla G_t(z) = -\frac{z}{2t}\, G_t(z),$$

the exact score function can be written explicitly as

$$s(x,t) = \nabla \log u(x,t) = \frac{1}{2t}\left(\sum_{k=1}^{n} p_k(x,t)\, x_k - x\right), \qquad p_k(x,t) = \frac{w_k\, G_t(x - x_k)}{\sum_{j=1}^{n} w_j\, G_t(x - x_j)}. \tag{5.2}$$

Applying (5.2) to the two initial measures $u_{0,1}$ and $u_{0,2}$, we obtain the corresponding exact score fields s_1 and s_2. The mixed score used for guidance is then given by

$$s^{(\lambda)}(x,t) = \lambda\, s_1(x,t) + (1-\lambda)\, s_2(x,t), \qquad \lambda \geq 0.$$

Numerical setup. The generation processes are simulated in similarity time $\tau = \log(T/t)$. Unless otherwise specified, we set $T = 1$ and $t_{\min} = 10^{-4}$, hence $\tau_{\max} = \log(T/t_{\min}) \simeq 9.2$. The deterministic rescaled ODE (3.2), equivalent to (2.9), is integrated with an explicit Euler scheme using the uniform step $\Delta\tau = 10^{-2}$. The stochastic rescaled SDE (4.7), corresponding to (2.6), is integrated by the Euler–Maruyama scheme with the same step size. For stochastic experiments, the random seed is fixed across runs to ensure reproducibility of the displayed trajectories. In this finite-mixture setting, all score fields are computed exactly from (5.2); no neural-network approximation or score-matching estimation is involved.

1D. In the one-dimensional setting, we consider

$$u_{0,1} = \frac{1}{3}\big(\delta_{-1} + \delta_1 + \delta_2\big), \qquad u_{0,2} = \frac{1}{3}\big(\delta_0 + \delta_{1.5} + \delta_5\big).$$

Figure 6 compares the geometric potential Φ_λ with the corresponding dynamical behavior in the one-dimensional empirical setting, for $\lambda \in \{0.5, 1, 2\}$. In the first row, we compare the log-product potential $V_\lambda(\cdot,t)$ at $t = 10^{-4}$ with the geometric potential Φ_λ. We observe a clear agreement between the two profiles: by Lemma 6.2, when t is small, the rescaled potential $-4t\, V_\lambda(\cdot,t)$ is close to Φ_λ. Moreover, the local maximizers of $V_\lambda(\cdot,t)$ match well with the local minimizers of Φ_λ, in accordance with Lemmas 6.3 and 6.4. In the second row, we numerically integrate the deterministic generation dynamics (2.9) in rescaled time. In all three cases, the trajectories converge toward local minimizers of Φ_λ, consistently with Theorem 3.18. In the third row, we simulate the stochastic rescaled generation dynamics with noise level $\varepsilon = 0.2$. The trajectories exhibit the same overall attraction toward the local minimizers of Φ_λ, up to visible stochastic fluctuations. This provides numerical evidence that the deterministic geometric picture remains relevant in the stochastic setting, as discussed in Section 4.3.

2D. In the two-dimensional case, we consider

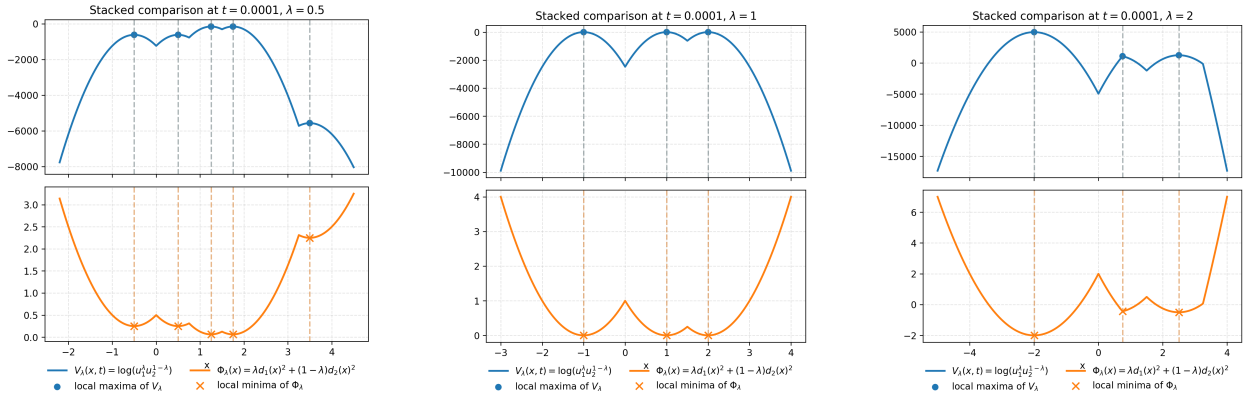
$$u_{0,1} = \frac{1}{3}\Big(\delta_{(-3.0,\,2.2)} + \delta_{(-1.2,\,-2.6)} + \delta_{(1.0,\,1.3)}\Big), \qquad u_{0,2} = \frac{1}{3}\Big(\delta_{(2.8,\,-3.2)} + \delta_{(4.0,\,2.6)} + \delta_{(0.2,\,-0.2)}\Big).$$

The corresponding Voronoi interfaces are displayed in Figure 1. The associated landscapes of Φ_λ and gradient fields for $\lambda \in \{0.5, 1, 2.5\}$ were presented earlier in Figure 2.

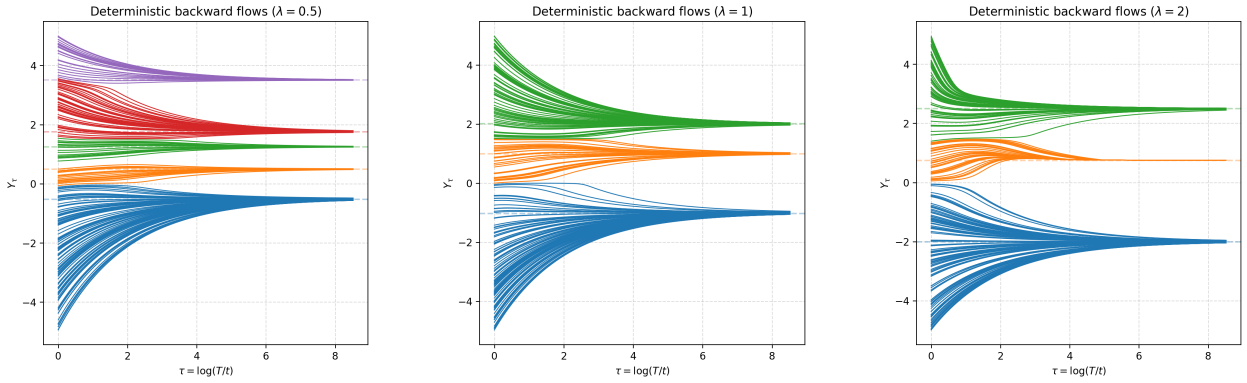
Figure 7 compares the deterministic generation dynamics (2.9) with the stochastic generation process (2.6) for $\varepsilon = 0.2$, with $T = 1$ and $t_{\min} = 10^{-4}$ fixed. In the deterministic setting, the numerically observed behavior is consistent with Theorem 3.18. In the stochastic case, although diffusion induces visible fluctuations, the trajectories still exhibit a similar concentration near the local minimizers of Φ_λ. This agrees with the discussion in Subsection 4.3 and further supports the relevance of the geometric picture in the stochastic regime.

5.2 The continuous data setting

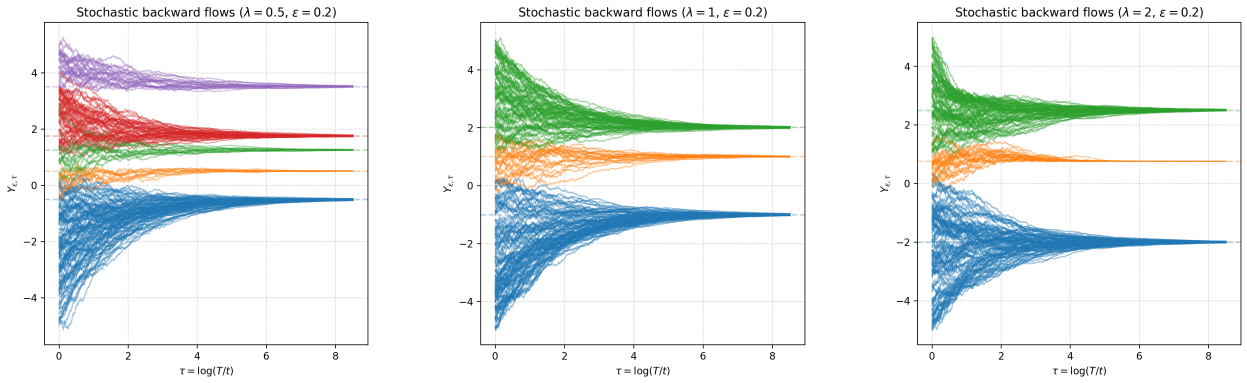
We next consider a two-dimensional example with continuous data distributions. The first distribution is supported on horizontal segments, while the second is supported on vertical segments. More



(a) $\lambda = 0.5$ (b) $\lambda = 1$ (c) $\lambda = 2$

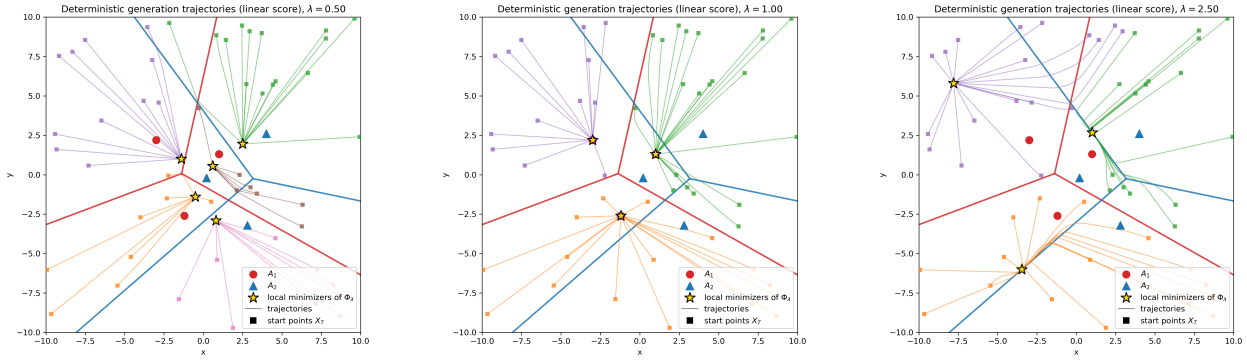


(d) $\lambda = 0.5$, $\varepsilon = 0$ (e) $\lambda = 1$, $\varepsilon = 0$ (f) $\lambda = 2$, $\varepsilon = 0$

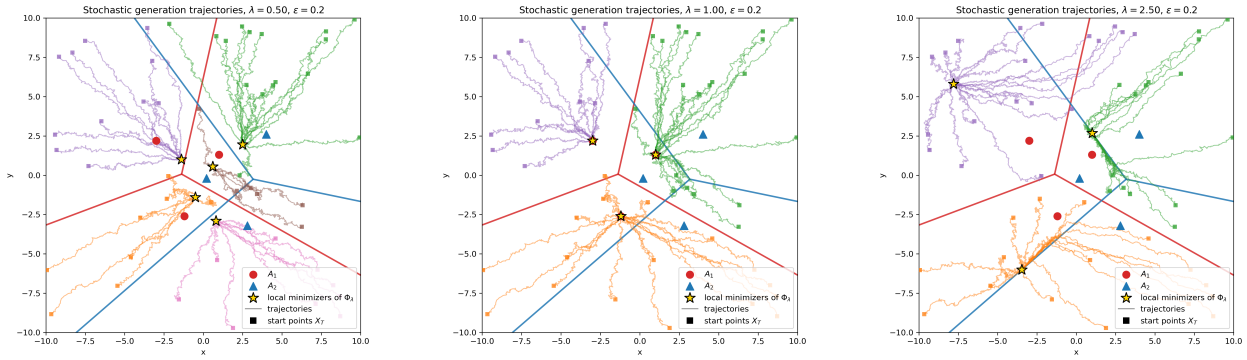


(g) $\lambda = 0.5$, $\varepsilon = 0.2$ (h) $\lambda = 1$, $\varepsilon = 0.2$ (i) $\lambda = 2$, $\varepsilon = 0.2$

Figure 6: 1D case. Top row: stacked visualization of the log-product potential $V_\lambda(\cdot, t)$ at $t = 10^{-4}$ and the geometric potential Φ_λ, with dashed lines indicating local maximizers of $V_\lambda(\cdot, t)$ and local minimizers of Φ_λ. For $\lambda = 2$, the middle local minimizer lies in the non-differentiability set $\mathrm{ND}(A_1, A_2)$: it is the Voronoi interface point 0.75, midway between 0 and 1.5 in A_2. Middle row: numerical integration of the deterministic rescaled gradient-flow dynamics. Bottom row: numerical integration of the stochastic rescaled dynamics with noise level $\varepsilon = 0.2$. In all cases, the trajectories are attracted toward the local minimizers of Φ_λ.



(a) $\lambda = 0.5$, $\varepsilon = 0$ (b) $\lambda = 1$, $\varepsilon = 0$ (c) $\lambda = 2.5$, $\varepsilon = 0$



(d) $\lambda = 0.5$, $\varepsilon = 0.2$ (e) $\lambda = 1$, $\varepsilon = 0.2$ (f) $\lambda = 2.5$, $\varepsilon = 0.2$

Figure 7: 2D case. Backward generation trajectories in $\mathbb{R}^2$ driven by the mixed exact score associated with the two empirical measures $u_{0,1}$ and $u_{0,2}$. Top row: deterministic trajectories solving (2.9). Bottom row: stochastic trajectories solving (2.6) with $\varepsilon = 0.2$. Left (MoE, $\lambda = 0.5$): the local minimizers of Φ_λ lie in the interior of strict Voronoi cells, and the trajectories are attracted toward them. Middle (pure imitation, $\lambda = 1$): Φ_λ reduces to the squared distance to A_1, so the trajectories concentrate near points of A_1. Right (CFG, $\lambda = 2.5$): one local minimizer lies in the interior of a Voronoi cell, while two others lie on Voronoi interfaces; both deterministic and stochastic trajectories concentrate near these minimizers.

precisely, let

$$I = [-1.5, -1.2] \cup [-0.8, 0.8] \cup [1.2, 1.5].$$

We define $u_{0,1}$ and $u_{0,2}$ as the uniform probability measures supported on

$$A_1 = I \times \{-1, 1\}, \qquad A_2 = \{-1, 1\} \times I.$$

Thus, A_1 consists of two segmented horizontal lines, while A_2 consists of two segmented vertical lines. The first two panels of Figure 8 illustrate these supports and the corresponding generation trajectories driven by each dataset. Since the data are uniformly distributed along the segments, the generated points spread along them accordingly.

The third panel of Figure 8 displays the generation flow driven by the MoE score ($\lambda = 0.5$). Since the two support sets do not intersect, the geometric potential Φ_λ favors points balancing proximity to both A_1 and A_2. This is clearly visible in the simulation: the trajectories are attracted toward the central junction region between the two supports, even though $A_1 \cap A_2 = \varnothing$.

The last panel illustrates the CFG regime ($\lambda = 2$). In this case, the trajectories are drawn toward points that remain close to A_1 while being repelled from the vertical structure A_2. This is consistent with the geometric interpretation of CFG as favoring the first dataset while suppressing features associated with the second one.

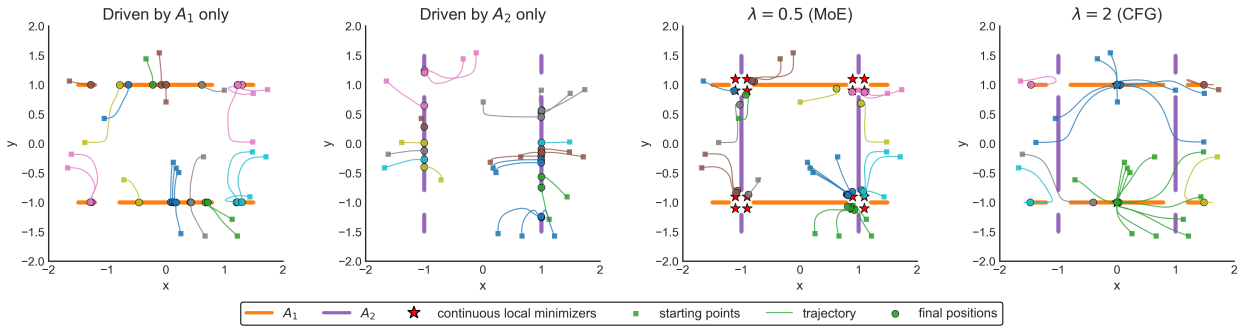


Figure 8: Backward deterministic generation trajectories driven by pure or mixed scores associated with continuous datasets supported on horizontal and vertical segmented lines.

Remark 5.1 (Relation to practical guided generation)**.** Although highly idealized, the previous continuous-data example captures a mechanism that is closely related to practical guided generation. In real applications, high-dimensional data such as images, audio, or text embeddings are often concentrated near structured subsets of the ambient space rather than being spread uniformly. From this perspective, the two supports A_1 and A_2 may be interpreted as simplified proxies for two distinct semantic sources.

In the MoE regime ($0 \leq \lambda \leq 1$), the mixed score combines the two sources in a cooperative way. Geometrically, the generation dynamics is driven toward minimizers of the potential Φ_λ, which balance proximity to both supports. This corresponds to generating samples that remain compatible with two families of constraints at the same time.

By contrast, in the CFG regime ($\lambda > 1$), the mixing becomes extrapolative: the dynamics favors the first support while being repelled from the second one. This is analogous to practical guidance mechanisms in diffusion models, where one seeks samples that strongly express a target attribute while suppressing an undesired background component. The real-data experiment below provides a concrete illustration of this phenomenon in a high-dimensional image dataset.

5.3 CIFAR-10 Guidance via Score Mixing

We conclude the numerical section with a qualitative real-data illustration of the geometric mechanism on CIFAR-10. The aim is not to propose a competitive image-generation method, but to test whether the phenomena predicted by the theory remain visible in a high-dimensional empirical setting: cooperative attraction in the MoE regime, discriminative guidance in the CFG regime, and loss of fidelity under excessive amplification. Accordingly, the outputs below should be interpreted as backward-flow samples generated by exact empirical heat-flow scores in pixel space, rather than as samples from a trained state-of-the-art diffusion model.

Each CIFAR-10 image is represented directly in pixel space as a vector in $\mathbb{R}^{3072}$, corresponding to the RGB values of a 32×32 image. For each class c, we consider the empirical measure

$$u_{0,c} = \frac{1}{N_c} \sum_{k=1}^{N_c} \delta_{x_k^{(c)}},$$

where $x_k^{(c)} \in \mathbb{R}^{3072}$ denotes an image of class c. We use the standard CIFAR-10 training split, with $N_c = 5000$ images per class. In the experiment below, we fix

$$A_1 = \text{airplane class}, \qquad A_2 = \text{union of all non-airplane classes},$$

so that $N_{\text{airplane}} = 5000$ and $N_{\text{others}} = 45000$. The heat-flow scores are evaluated exactly using the Gaussian-mixture formula (5.2), and the deterministic rescaled backward dynamics is integrated with the same parameters as in the finite-mixture experiments above, namely $\Delta\tau = 10^{-2}$ and $\tau_{\max} \simeq 9.2$.

The mixed score is therefore

$$s^{(\lambda)} = \lambda s_{\text{airplane}} + (1-\lambda) s_{\text{others}},$$

which provides a transparent class-versus-complement guidance setting. The first score attracts the dynamics toward the target class, while the second score represents the empirical geometry of all remaining classes. Figure 9 shows the generated samples for

$$\lambda \in \{0, 0.5, 1, 2, 5\}.$$

For $\lambda = 0$, the dynamics is driven entirely by the complementary dataset A_2, and the resulting samples display generic non-airplane structures. As λ increases through the MoE and moderate CFG regimes, airplane-like features become progressively more pronounced: elongated bodies, wing-like silhouettes, and sky-like backgrounds appear more frequently. This indicates that the mixed score effectively transfers the dynamics toward the target support.

The large-guidance regime exhibits a different behavior. For $\lambda = 2$, and more clearly for $\lambda = 5$, the samples become more strongly biased toward the airplane class, but their visual fidelity deteriorates: colors become less natural, contrast is amplified, and several images show oversaturated or distorted structures. This reproduces, in the present exact-score geometric setting, the familiar over-guidance trade-off observed in classifier-free guidance: increasing the guidance strength improves semantic alignment up to an intermediate regime, while excessive guidance degrades realism; see, for instance, [17, 23, 26, 53, 11].

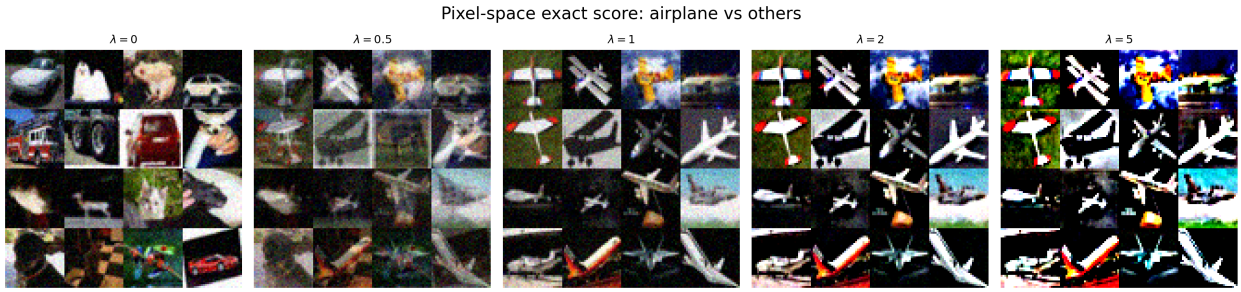


Figure 9: Real-data experiment on CIFAR-10 in pixel space. Here A_1 is the airplane class and A_2 is the union of all non-airplane classes. The mixed score is $s^{(\lambda)} = \lambda s_{\text{airplane}} + (1-\lambda) s_{\text{others}}$. As λ increases, the generated samples become more clearly aligned with the airplane class, but large values of λ also induce a visible loss of fidelity, illustrating the over-guidance effect.

This behavior has a direct geometric interpretation. Since the two empirical supports are separated at the class level, the score difference

$$s_{\text{airplane}} - s_{\text{others}}$$

acts as a discriminative direction between the target class and its complement. Moderate guidance amplifies this direction and sharpens the target-class features. When λ is too large, however, the same discriminative direction is over-amplified: the dynamics is pushed away from the balanced high-density region of the empirical distribution, producing samples that are more class-specific but less faithful.

Thus, even in this simplified exact-score experiment, the CIFAR-10 simulations support the central geometric message of the paper. Score mixing acts as a guidance mechanism determined by the relative position of the underlying supports. In the disjoint class-versus-complement setting, increasing λ enhances target-class discrimination, but beyond an intermediate range the extrapolative nature of CFG produces a visible loss of fidelity.

6 Laplace–Varadhan principle and gradient structure

This section develops the small-time asymptotic structure of the rescaled mixed potential $F_\lambda(\cdot, t)$ and proves Theorem 3.10.

The analysis relies on Laplace–Varadhan asymptotics for Gaussian convolutions, which identify the limiting distance potential; see, for instance, Dembo–Zeitouni [16, Sec. 4.3], Bender–Orszag [7, Sec. 6.4], and Fukunaga–Hostetler [18]. We first illustrate this mechanism in the elementary case of a single empirical source, where the role of Voronoi interfaces is already transparent. We then return to the general mixed setting and establish quantitative value and gradient estimates for F_λ. These estimates are subsequently used to prove Theorem 3.10, by compactness of time shifts and identification of weak-$*$ limits of the rescaled drifts.

6.1 Laplace–Varadhan principle for a single empirical source

We start with the elementary case of a single empirical source. Although simple, it already contains the essential geometric mechanism of the small-time limit: after heat regularization, the leading-order behavior is determined only by the distance to the support of the initial measure, while the weights contribute only at lower order.

Let

$$u_0 = \sum_{k=1}^{n} w_k \, \delta_{x_k}, \qquad w_k > 0, \qquad \sum_{k=1}^{n} w_k = 1,$$

and denote its support by

$$A = \operatorname{supp}(u_0) = \{x_1, \dots, x_n\}.$$

We write

$$d(x) := \operatorname{dist}(x, A) = \min_{1 \le k \le n} \|x - x_k\|, \qquad I(x) := \operatorname*{argmin}_{1 \le k \le n} \|x - x_k\|.$$

Using the explicit Gaussian-mixture formula (5.1), we factor out the dominant exponential scale:

$$u(x,t) = (4\pi t)^{-d/2} \exp\Big(-\frac{d(x)^2}{4t}\Big) \sum_{k=1}^{n} w_k \exp\Big(-\frac{\|x - x_k\|^2 - d(x)^2}{4t}\Big).$$

The renormalized sum satisfies

$$0 < \sum_{k \in I(x)} w_k \le \sum_{k=1}^{n} w_k \exp\Big(-\frac{\|x - x_k\|^2 - d(x)^2}{4t}\Big) \le \sum_{k=1}^{n} w_k = 1,$$

and therefore its logarithm is $O(1)$ as $t \to 0^+$. It follows that

$$-4t \log u(x,t) \longrightarrow d(x)^2 \qquad \text{as } t \to 0^+. \tag{6.1}$$

Thus, to leading order, the heat flow forgets the weights and retains only the squared distance to the support. This is the basic Laplace–Varadhan principle in the present setting; see [7, Sec. 6.4] and [16, Sec. 4.3].

To understand the limiting gradient, we use the explicit score formula (5.2):

$$\nabla\big(-4t \log u(x,t)\big) = 2\left(x - \sum_{k=1}^{n} p_k(x,t)\, x_k\right), \qquad p_k(x,t) := \frac{w_k G_t(x - x_k)}{\sum_{j=1}^{n} w_j G_t(x - x_j)}.$$

For each fixed $x \in \mathbb{R}^d$, the coefficients $p_k(x,t)$ satisfy

$$p_k(x,t) \longrightarrow \begin{cases} \dfrac{w_k}{\sum_{j \in I(x)} w_j}, & k \in I(x), \\ 0, & k \notin I(x), \end{cases} \qquad \text{as } t \to 0^+.$$

Indeed, after factoring out the common dominant exponential $e^{-d(x)^2/(4t)}$, the terms corresponding to $k \in I(x)$ remain of order one, whereas the others decay exponentially fast.

This leads to two regimes.

- If $I(x) = \{k^*\}$ is a singleton, then

  $$\nabla\big(-4t \log u(x,t)\big) \to 2(x - x_{k^*}) = \nabla(d(x)^2), \qquad \text{as } t \to 0^+.$$

  Thus, away from the Voronoi interfaces, both the rescaled potential and its gradient converge to the classical squared-distance landscape.
- If $|I(x)| \geq 2$, then several support points are tied at the same minimal distance, and

  $$\nabla\big(-4t \log u(x,t)\big) \to 2 \sum_{k \in I(x)} \frac{w_k}{\sum_{j \in I(x)} w_j} (x - x_k) \in \partial^C(d^2)(x).$$

  Thus, on the interface, the limiting gradient is no longer unique; it belongs to the Clarke subdifferential of the squared-distance function.

This single-source picture already contains the essential geometry of the limit: the rescaled heat potential converges to a squared-distance function, and Voronoi interfaces are the unique source of nonsmoothness.

6.2 Laplace–Varadhan principle for the mixed setting

We now return to the mixed setting and allow general compactly supported initial measures. For $i \in \{1,2\}$, let $u_{0,i}$ be a probability measure on $\mathbb{R}^d$ with compact support

$$A_i = \operatorname{supp}(u_{0,i}), \qquad d_i(x) = \operatorname{dist}(x, A_i), \qquad \Pi_i(x) := \operatorname*{argmin}_{a \in A_i} \|x - a\|$$

Throughout this subsection we assume the following quantitative lower-mass condition.

Assumption A (Uniform lower mass bound (power law))**.** There exist constants $c > 0$, $\alpha \geq 0$, and $r_0 > 0$ such that, for each $i \in \{1,2\}$,

$$\inf_{x \in A_i} u_{0,i}\big(B(x,r)\big) \geq c\, r^{\alpha} \qquad \forall r \in (0, r_0]. \tag{6.2}$$

Remark 6.1. Assumption A is the key quantitative input allowing one to extend the Laplace–Varadhan argument beyond the empirical setting. Its role is to provide a uniform lower bound on the mass of small balls centered on the support, which is precisely what is needed to prove the convergence of the rescaled potential F_λ toward the geometric potential Φ_λ.

Geometrically, this assumption means that the measure is supported on a set of effective dimension α and has a uniformly positive density at that scale. This includes, in particular, finite Dirac measures ($\alpha = 0$), absolutely continuous measures on full-dimensional supports ($\alpha = d$), and more generally measures concentrated on a finite union of lower-dimensional manifolds carrying a uniform intrinsic density bound. The last example is especially natural in applications to imaging and generative modeling, where one often assumes that data are concentrated near a low-dimensional latent manifold embedded in a high-dimensional ambient space.

Gaussian mean-shift formula. We recall that

$$u_i(\cdot, t) = G_t * u_{0,i}, \qquad i = 1, 2,$$

where

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

is the heat kernel. Differentiating under the integral sign gives

$$\nabla u_i(x,t) = \int_{A_i} \nabla_x G_t(x-y)\, du_{0,i}(y) = -\frac{1}{2t}\int_{A_i} (x-y)\, G_t(x-y)\, du_{0,i}(y),$$

and therefore

$$\nabla \log u_i(x,t) = \frac{\nabla u_i(x,t)}{u_i(x,t)} = \frac{1}{2t}\big(m_i(x,t) - x\big), \tag{6.3}$$

where

$$m_i(x,t) := \int_{A_i} y\, d\nu^x_{t,i}(y), \qquad \nu^x_{t,i}(dy) := \frac{G_t(x-y)}{u_i(x,t)}\, du_{0,i}(y). \tag{6.4}$$

Since $\nu^x_{t,i}$ is a probability measure supported on A_i, one has

$$m_i(x,t) \in \operatorname{conv}(A_i) \qquad \forall (x,t) \in \mathbb{R}^d \times (0,T].$$

Substituting (6.3) into the definition of the rescaled potential

$$F_\lambda(x,t) = -4t\,\big(\lambda \log u_1(x,t) + (1-\lambda)\log u_2(x,t)\big),$$

we obtain

$$\nabla F_\lambda(x,t) = 2\Big(x - \lambda m_1(x,t) - (1-\lambda) m_2(x,t)\Big). \tag{6.5}$$

This is exactly the drift field driving the similarity-time dynamics (3.2).

Laplace–Varadhan-type estimates. Formula (6.5) shows that the asymptotic behavior of ∇F_λ is encoded in the concentration of the barycenters $m_i(x,t)$ toward nearest points of the supports A_i. We now state three lemmas quantifying this principle; their proofs are given in the next subsection.

Lemma 6.2 (Quantitative value convergence)**.** *Let Assumption A hold. Then, for each $i \in \{1,2\}$ and all (x,t) with $0 < t \le r_0^2$, one has*

$$u_i(x,t) \le (4\pi t)^{-d/2} \exp\Big(-\frac{d_i(x)^2}{4t}\Big), \tag{6.6}$$

$$u_i(x,t) \ge c\, t^{\alpha/2} (4\pi t)^{-d/2} \exp\Big(-\frac{(d_i(x)+\sqrt{t})^2}{4t}\Big). \tag{6.7}$$

Consequently, for all $x \in \mathbb{R}^d$ and $0 < t \le r_0^2$,

$$\big|\Phi_\lambda(x) - F_\lambda(x,t)\big| \le C_1(x)\sqrt{t} + C_2\, t(1 + |\log t|), \tag{6.8}$$

where

$$C_1(x) := 2|\lambda|\, d_1(x) + 2|1-\lambda|\, d_2(x),$$

and

$$C_2 = \max\big\{1 + 4|\log c| + 2d\log(4\pi),\ 2\alpha + 2d\big\}\,(|\lambda| + |1-\lambda|).$$

Lemma 6.3 (Qualitative gradient convergence toward the outer Clarke)**.** *Let Assumption A hold. Fix $x^* \in \mathbb{R}^d$. Then, for every $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\|x - x^*\| + t \le \delta \qquad \Longrightarrow \qquad \operatorname{dist}\big(\nabla F_\lambda(x,t),\, \widehat{\partial}\,\Phi_\lambda(x^*)\big) \le \varepsilon,$$

where $\widehat{\partial}\,\Phi_\lambda$ is the outer Clarke subdifferential defined in (3.10).

Lemma 6.4 (Local uniform quantitative gradient convergence outside the active interface)**.** *Assume that*

$$u_{0,1} = \sum_{k=1}^{n_1} w_{1,k}\,\delta_{x_k}, \qquad u_{0,2} = \sum_{\ell=1}^{n_2} w_{2,\ell}\,\delta_{y_\ell}, \qquad A_1 = \{x_1,\dots,x_{n_1}\}, \qquad A_2 = \{y_1,\dots,y_{n_2}\},$$

with $w_{i,k} > 0$ and $\sum_k w_{i,k} = 1$. Assume that

$$x_0 \notin \mathrm{ND}(A_1,A_2) = \mathrm{ND}(d_1^2) \cup \mathrm{ND}(d_2^2).$$

Then there exist a neighborhood U of x_0 and constants $C, \eta > 0$ such that

$$\|\nabla F_\lambda(x,t) - \nabla \Phi_\lambda(x)\| \le C e^{-\eta/t} \qquad \forall x \in U,\ \forall t > 0. \tag{6.9}$$

In the endpoint cases, the assumption may be weakened to the active interface: if $\lambda = 0$, it is enough to assume $x_0 \notin \mathrm{ND}(d_2^2)$; if $\lambda = 1$, it is enough to assume $x_0 \notin \mathrm{ND}(d_1^2)$.

6.3 Proofs of the technical lemmas

Proof of Lemma 6.2. We prove the estimates for a fixed $i \in \{1,2\}$.

Step 1: Upper bound. Since $u_i(x,t) = \int_{A_i} G_t(x-y)\, du_{0,i}(y)$ and $u_{0,i}$ is a probability measure,

$$u_i(x,t) \le \sup_{y \in A_i} G_t(x-y).$$

Because $\min_{y\in A_i}\|x-y\| = d_i(x)$ and $r \mapsto e^{-r^2/(4t)}$ is decreasing,

$$\sup_{y\in A_i} G_t(x-y) = (4\pi t)^{-d/2} \exp\Big(-\frac{d_i(x)^2}{4t}\Big),$$

which gives (6.6).

Step 2: Lower bound. Fix $x \in \mathbb{R}^d$, and choose $y_x \in A_i$ such that $\|x - y_x\| = d_i(x)$, which is possible since A_i is compact. For every $y \in B(y_x, \sqrt{t})$,

$$\|x - y\| \le \|x - y_x\| + \|y - y_x\| \le d_i(x) + \sqrt{t},$$

and therefore

$$G_t(x - y) \ge (4\pi t)^{-d/2} \exp\Big(-\frac{(d_i(x) + \sqrt{t})^2}{4t}\Big).$$

Hence

$$u_i(x,t) \ge \int_{B(y_x,\sqrt{t})} G_t(x-y)\, du_{0,i}(y) \ge (4\pi t)^{-d/2} \exp\Big(-\frac{(d_i(x) + \sqrt{t})^2}{4t}\Big)\, u_{0,i}\big(B(y_x, \sqrt{t})\big).$$

Since $t \le r_0^2$, we have $\sqrt{t} \le r_0$, so Assumption A yields

$$u_{0,i}\big(B(y_x, \sqrt{t})\big) \ge c\,(\sqrt{t})^\alpha = c\, t^{\alpha/2}.$$

This proves (6.7).

Step 3: Estimate for $F_\lambda - \Phi_\lambda$. From (6.6),

$$\log u_i(x,t) \le -\frac{d}{2}\log(4\pi t) - \frac{d_i(x)^2}{4t},$$

hence

$$4t \log u_i(x,t) + d_i(x)^2 \le -2dt \log(4\pi t).$$

From (6.7),

$$\log u_i(x,t) \ge \log c + \frac{\alpha}{2}\log t - \frac{d}{2}\log(4\pi t) - \frac{(d_i(x) + \sqrt{t})^2}{4t},$$

so

$$4t \log u_i(x,t) + d_i(x)^2 \ge -\big((d_i(x) + \sqrt{t})^2 - d_i(x)^2\big) + 4t \log c + 2\alpha t \log t - 2dt \log(4\pi t).$$

Since

$$(d_i(x) + \sqrt{t})^2 - d_i(x)^2 = 2d_i(x)\sqrt{t} + t,$$

we obtain

$$\big|4t \log u_i(x,t) + d_i(x)^2\big| \le 2d_i(x)\sqrt{t} + C\, t(1 + |\log t|),$$

for a constant C depending only on c, α, d. Finally, using

$$\Phi_\lambda(x) - F_\lambda(x,t) = \lambda\big(d_1(x)^2 + 4t \log u_1(x,t)\big) + (1-\lambda)\big(d_2(x)^2 + 4t \log u_2(x,t)\big),$$

we deduce (6.8). □

Proof of Lemma 6.3. We divide the proof into three steps.

Step 1: Reduction to barycenter concentration. Fix $x^* \in \mathbb{R}^d$. For $i = 1, 2$, recall that

$$\partial^C(d_i^2)(x^*) = 2\big(x^* - \mathrm{conv}(\Pi_i(x^*))\big).$$

Recall mean-shift formula (6.3)-(6.5),

$$\nabla F_\lambda(x,t) = 2\big(x - \lambda m_1(x,t) - (1-\lambda)m_2(x,t)\big).$$

Hence

$$\begin{aligned}\operatorname{dist}\big(\nabla F_\lambda(x,t), \widehat{\partial}\Phi_\lambda(x^*)\big) \le\ & 2\|x-x^*\| \\ & + 2|\lambda|\ \operatorname{dist}\big(m_1(x,t), \operatorname{conv}(\Pi_1(x^*))\big) \\ & + 2|1-\lambda|\ \operatorname{dist}\big(m_2(x,t), \operatorname{conv}(\Pi_2(x^*))\big).\end{aligned}$$

Thus it suffices to prove that, for each $i = 1, 2$,

$$\forall \varepsilon > 0\ \exists \delta > 0: \quad \|x - x^*\| + t \le \delta \Longrightarrow \operatorname{dist}\big(m_i(x,t), \operatorname{conv}(\Pi_i(x^*))\big) \le \varepsilon. \tag{6.10}$$

Step 2: Compactness and contradiction. Fix $i \in \{1,2\}$, and suppose that (6.10) fails. Then there exist $\varepsilon_0 > 0$, $x_n \to x^*$, and $t_n \to 0^+$ such that

$$\operatorname{dist}\big(m_i(x_n,t_n), \operatorname{conv}(\Pi_i(x^*))\big) > \varepsilon_0 \qquad \forall n.$$

Recall that

$$\nu_{t_n,i}^{x_n}(dy) = \frac{G_{t_n}(x_n - y)}{u_i(x_n,t_n)}\, du_{0,i}(y)$$

is a probability measure supported on A_i, and that

$$m_i(x_n,t_n) = \int_{A_i} y\, d\nu_{t_n,i}^{x_n}(y).$$

Since $m_i(x_n,t_n) \in \operatorname{conv}(A_i)$ and A_i is compact, up to a subsequence,

$$m_i(x_n,t_n) \to \bar{m}_i \qquad \text{for some } \bar{m}_i \in \operatorname{conv}(A_i).$$

Up to a further subsequence, the probability measures $\nu_{t_n,i}^{x_n}$ converge weakly to some probability measure ν supported on A_i. We shall prove that

$$\operatorname{supp}(\nu) \subset \Pi_i(x^*).$$

This will imply

$$\bar{m}_i = \lim_{n\to\infty} \int_{A_i} y\, d\nu_{t_n,i}^{x_n}(y) = \int_{A_i} y\, d\nu(y) \in \operatorname{conv}(\Pi_i(x^*)),$$

contradicting the strict distance bound above.

Step 3: Identification of the limiting support. If $\Pi_i(x^*) = A_i$, there is nothing to prove. Otherwise, let O be an open neighborhood of $\Pi_i(x^*)$ in A_i. Since $A_i \setminus O$ is compact and disjoint from $\Pi_i(x^*)$, there exists $\eta > 0$ such that

$$\|x^* - y\|^2 \ge d_i(x^*)^2 + \eta \qquad \forall y \in A_i \setminus O.$$

Moreover,

$$\|x_n - y\|^2 - d_i(x_n)^2 \longrightarrow \|x^* - y\|^2 - d_i(x^*)^2 \qquad \text{uniformly on } A_i.$$

Therefore, for n large enough,

$$\|x_n - y\|^2 \ge d_i(x_n)^2 + \frac{\eta}{2} \qquad \forall y \in A_i \setminus O.$$

It follows that

$$\nu_{t_n,i}^{x_n}(A_i \setminus O) \le \frac{(4\pi t_n)^{-d/2} \exp\big(-(d_i(x_n)^2 + \eta/2)/(4t_n)\big)}{u_i(x_n, t_n)}.$$

Using the lower bound (6.7), we obtain

$$\nu_{t_n,i}^{x_n}(A_i \setminus O) \le C\, t_n^{-\alpha/2} \exp\Big(\frac{d_i(x_n)}{2\sqrt{t_n}} - \frac{\eta}{8t_n}\Big) \longrightarrow 0.$$

Hence the mass of $\nu_{t_n,i}^{x_n}$ outside every neighborhood of $\Pi_i(x^*)$ vanishes as $n \to \infty$.

Now take a decreasing sequence of open neighborhoods O_m of $\Pi_i(x^*)$ in A_i such that

$$\bigcap_{m\ge 1} O_m = \Pi_i(x^*), \qquad \overline{O_{m+1}} \subset O_m.$$

The estimate above gives

$$\nu_{t_n,i}^{x_n}(A_i \setminus O_{m+1}) \to 0 \qquad \forall m.$$

Choose a continuous cutoff $\psi_m : A_i \to [0,1]$ such that

$$\psi_m = 0 \text{ on } \overline{O_{m+1}}, \qquad \psi_m = 1 \text{ on } A_i \setminus O_m.$$

Then

$$\int_{A_i} \psi_m \, d\nu_{t_n,i}^{x_n} \le \nu_{t_n,i}^{x_n}(A_i \setminus O_{m+1}) \to 0.$$

Passing to the weak limit gives

$$\int_{A_i} \psi_m \, d\nu = 0,$$

and therefore

$$\nu(A_i \setminus O_m) = 0.$$

Since this holds for every m, we obtain

$$\nu(A_i \setminus \Pi_i(x^*)) = 0.$$

and hence $\operatorname{supp}(\nu) \subset \Pi_i(x^*)$. The conclusion follows by step 2. □

Proof of Lemma 6.4. Since

$$x_0 \notin \mathrm{ND}(A_1, A_2) = \mathrm{ND}(d_1^2) \cup \mathrm{ND}(d_2^2),$$

both d_1^2 and d_2^2 are differentiable at x_0. Hence the nearest points of x_0 in A_1 and A_2 are unique, denoted respectively by x_{k^*} and y_{ℓ^*}. In particular,

$$\delta_1(x_0) := \min_{k \ne k^*} \big(\|x_0 - x_k\|^2 - \|x_0 - x_{k^*}\|^2\big) > 0,$$

and

$$\delta_2(x_0) := \min_{\ell \ne \ell^*} \big(\|x_0 - y_\ell\|^2 - \|x_0 - y_{\ell^*}\|^2\big) > 0.$$

By continuity of the functions

$$x \longmapsto \|x - x_k\|^2 - \|x - x_{k^*}\|^2, \qquad x \longmapsto \|x - y_\ell\|^2 - \|x - y_{\ell^*}\|^2,$$

there exist a neighborhood U of x_0 and constants $\eta_1, \eta_2 > 0$ such that, for every $x \in U$,

$$\|x - x_k\|^2 - \|x - x_{k^*}\|^2 \geq \eta_1 \qquad \forall k \neq k^*,$$

and

$$\|x - y_\ell\|^2 - \|x - y_{\ell^*}\|^2 \geq \eta_2 \qquad \forall \ell \neq \ell^*.$$

In particular, for every $x \in U$, the nearest points of x in A_1 and A_2 remain uniquely given by x_{k^*} and y_{ℓ^*}. Therefore

$$\nabla(d_1^2)(x) = 2(x - x_{k^*}), \qquad \nabla(d_2^2)(x) = 2(x - y_{\ell^*}),$$

and hence

$$\nabla\Phi_\lambda(x) = \lambda\nabla(d_1^2)(x) + (1-\lambda)\nabla(d_2^2)(x) = 2\big(x - \lambda x_{k^*} - (1-\lambda)y_{\ell^*}\big) \qquad \forall x \in U.$$

We now estimate the discrepancy between $m_i(x,t)$ and the corresponding nearest point, uniformly for $x \in U$. For $i = 1$,

$$m_1(x,t) = \frac{\sum_{k=1}^{n_1} w_{1,k} e^{-\|x-x_k\|^2/(4t)} x_k}{\sum_{k=1}^{n_1} w_{1,k} e^{-\|x-x_k\|^2/(4t)}}.$$

Factoring out the dominant exponential $e^{-\|x-x_{k^*}\|^2/(4t)}$, we obtain

$$m_1(x,t) - x_{k^*} = \frac{\sum_{k\neq k^*} w_{1,k} e^{-(\|x-x_k\|^2 - \|x-x_{k^*}\|^2)/(4t)}(x_k - x_{k^*})}{w_{1,k^*} + \sum_{k\neq k^*} w_{1,k} e^{-(\|x-x_k\|^2 - \|x-x_{k^*}\|^2)/(4t)}}.$$

Since, for every $x \in U$,

$$\|x - x_k\|^2 - \|x - x_{k^*}\|^2 \geq \eta_1 \qquad \forall k \neq k^*,$$

and

$$\|x_k - x_{k^*}\| \leq \operatorname{diam}(A_1), \qquad w_{1,k^*} \geq \min_k w_{1,k} > 0,$$

it follows that there exists a constant $C_1 > 0$, independent of $x \in U$ and $t > 0$, such that

$$\|m_1(x,t) - x_{k^*}\| \leq C_1 e^{-\eta_1/(4t)} \qquad \forall x \in U,\ \forall t > 0.$$

Exactly the same argument yields a constant $C_2 > 0$ such that

$$\|m_2(x,t) - y_{\ell^*}\| \leq C_2 e^{-\eta_2/(4t)} \qquad \forall x \in U,\ \forall t > 0.$$

Finally, using the mean-shift formula

$$\nabla F_\lambda(x,t) = 2\big(x - \lambda m_1(x,t) - (1-\lambda)m_2(x,t)\big),$$

we obtain, for every $x \in U$ and $t > 0$,

$$\begin{aligned}\|\nabla F_\lambda(x,t) - \nabla\Phi_\lambda(x)\| &\leq 2|\lambda|\,\|m_1(x,t) - x_{k^*}\| + 2|1-\lambda|\,\|m_2(x,t) - y_{\ell^*}\| \\ &\leq 2|\lambda|\,C_1 e^{-\eta_1/(4t)} + 2|1-\lambda|\,C_2 e^{-\eta_2/(4t)}.\end{aligned}$$

Setting

$$\eta \coloneqq \frac{1}{4}\min\{\eta_1, \eta_2\},$$

and enlarging the constant if necessary, we conclude that there exists $C > 0$ such that

$$\|\nabla F_\lambda(x,t) - \nabla \Phi_\lambda(x)\| \le Ce^{-\eta/t} \qquad \forall x \in U,\ \forall t > 0.$$

This proves (6.9).

The endpoint cases admit the announced weaker assumptions. Indeed, if $\lambda = 1$, then

$$\nabla F_1(x,t) = 2(x - m_1(x,t)), \qquad \nabla \Phi_1(x) = \nabla d_1^2(x),$$

so only the uniqueness of the nearest point in A_1 is needed. Similarly, if $\lambda = 0$, then

$$\nabla F_0(x,t) = 2(x - m_2(x,t)), \qquad \nabla \Phi_0(x) = \nabla d_2^2(x),$$

so only the uniqueness of the nearest point in A_2 is needed. The same argument above, with the inactive support omitted, proves the estimate under these weaker endpoint assumptions. □

6.4 Proof of Theorem 3.10

We will prove Theorem 3.10 under Assumption A, which covers the finite Dirac mixture case. Let $X = (X_t)_{t\in(0,T]}$ be the solution of the deterministic generation flow (2.9), and let $Y = (Y_\tau)_{\tau\ge 0}$ be its similarity-time rescaling defined in (3.1), equivalently the solution of (3.2).

Step 1: Lyapunov confinement and uniform drift bounds. Set

$$K := \operatorname{conv}\big(\lambda A_1 + (1-\lambda)A_2\big).$$

Recall from the Gaussian mean-shift formula (6.3)–(6.4) that

$$m_i(x,t) \in \operatorname{conv}(A_i), \qquad i = 1, 2.$$

Hence

$$\lambda m_1(x,t) + (1-\lambda)m_2(x,t) \in \lambda \operatorname{conv}(A_1) + (1-\lambda)\operatorname{conv}(A_2) = K.$$

Using (6.5), we obtain

$$\nabla F_\lambda(x,t) \in 2(x - K) \qquad \forall (x,t) \in \mathbb{R}^d \times (0,T]. \tag{6.11}$$

Define the Lyapunov function

$$L(x) := \frac{1}{2}\operatorname{dist}(x,K)^2.$$

Since K is nonempty, closed, and convex, $L \in C^1(\mathbb{R}^d)$ and

$$\nabla L(x) = x - \Pi_K(x),$$

where Π_K denotes the metric projection onto K.

By (6.11), for every $\tau \ge 0$, there exists $z_\tau \in K$ such that

$$\nabla F_\lambda\big(Y_\tau, Te^{-\tau}\big) = 2(Y_\tau - z_\tau).$$

Since Y solves (3.2), we have

$$\dot{Y}_\tau = -\frac{1}{4}\nabla F_\lambda\big(Y_\tau, Te^{-\tau}\big) = -\frac{1}{2}(Y_\tau - z_\tau).$$

Let $\pi_\tau := \Pi_K(Y_\tau)$. Then

$$\begin{aligned}\frac{d}{d\tau}L(Y_\tau) &= \langle \nabla L(Y_\tau), \dot{Y}_\tau\rangle \\ &= \big\langle Y_\tau - \pi_\tau, -\tfrac{1}{2}(Y_\tau - z_\tau)\big\rangle \\ &= -\frac{1}{2}\|Y_\tau - \pi_\tau\|^2 - \frac{1}{2}\langle Y_\tau - \pi_\tau, \pi_\tau - z_\tau\rangle. \end{aligned} \tag{6.12}$$

By the characterization of the metric projection onto a closed convex set,

$$\langle Y_\tau - \pi_\tau, z - \pi_\tau\rangle \le 0 \qquad \forall z \in K.$$

Taking $z = z_\tau \in K$, we get

$$\langle Y_\tau - \pi_\tau, \pi_\tau - z_\tau\rangle \ge 0.$$

Therefore (6.12) yields

$$\frac{d}{d\tau}L(Y_\tau) \le -\frac{1}{2}\|Y_\tau - \pi_\tau\|^2 = -L(Y_\tau).$$

Hence

$$L(Y_\tau) \le e^{-\tau}L(Y_0) = e^{-\tau}L(x_T) \qquad \forall \tau \ge 0. \tag{6.13}$$

Define the compact sublevel set

$$\Omega := \{x \in \mathbb{R}^d : \ L(x) \le L(x_T)\}.$$

Then (6.13) implies

$$Y_\tau \in \Omega \qquad \forall \tau \ge 0.$$

This proves that Y remains in a compact subset of $\mathbb{R}^d$.

Moreover, by (6.11),

$$\|\nabla F_\lambda(x,t)\| \le 2\,\operatorname{dist}(x,K) \le 2\Big(\|x\| + \sup_{z\in K}\|z\|\Big).$$

Since $Y_\tau \in \Omega$ for all $\tau \ge 0$, we deduce that there exists $C > 0$ such that

$$\|\dot{Y}_\tau\| = \frac{1}{4}\big\|\nabla F_\lambda(Y_\tau, Te^{-\tau})\big\| \le C \qquad \forall \tau \ge 0. \tag{6.14}$$

Step 2: Time shifts and compactness. Let $(\tau_j)_{j\ge1}$ be any sequence such that $\tau_j \to \infty$, and define

$$Y^j_\tau := Y_{\tau+\tau_j}, \qquad \tau \ge 0.$$

Since $Y_\tau \in \Omega$ for all $\tau \ge 0$, each Y^j takes values in the fixed compact set Ω. Moreover, by (6.14),

$$\Big\|\frac{d}{d\tau}Y^j_\tau\Big\| = \|\dot{Y}_{\tau+\tau_j}\| \le C \qquad \forall \tau \ge 0.$$

Thus the family $(Y^j)_j$ is uniformly bounded and equi-Lipschitz on $\mathbb{R}_+$. By Arzelà–Ascoli, it is relatively compact in $C_{\mathrm{loc}}(\mathbb{R}_+;\mathbb{R}^d)$.

Hence, up to extraction of a subsequence, there exists a Lipschitz curve

$$Z = (Z_\tau)_{\tau\ge0} \in C_{\mathrm{loc}}(\mathbb{R}_+;\mathbb{R}^d)$$

such that

$$Y^j \to Z \qquad \text{in } C_{\mathrm{loc}}(\mathbb{R}_+;\mathbb{R}^d). \tag{6.15}$$

Step 3: Integral formulation and weak-$*$ limit of the drifts. Fix $0 \le a < b < \infty$. For each j, integrating the equation for Y gives

$$Y_b^j - Y_a^j = -\frac{1}{4}\int_a^b p_j(\tau)\,d\tau, \qquad p_j(\tau) := \nabla F_\lambda\big(Y_\tau^j, Te^{-(\tau+\tau_j)}\big). \tag{6.16}$$

By the uniform bound on ∇F_λ along Y, the sequence $(p_j)_j$ is bounded in $L^\infty((a,b);\mathbb{R}^d)$. Thus, after extracting a further subsequence if necessary, there exists

$$p \in L^\infty((a,b);\mathbb{R}^d)$$

such that

$$p_j \rightharpoonup^* p \qquad \text{in } L^\infty((a,b);\mathbb{R}^d). \tag{6.17}$$

Passing to the limit in (6.16) using (6.15) and (6.17), we obtain

$$Z_b - Z_a = -\frac{1}{4}\int_a^b p(\tau)\,d\tau.$$

Hence Z is absolutely continuous and

$$\dot{Z}_\tau = -\frac{1}{4}\,p(\tau) \qquad \text{for a.e. } \tau \in (a,b). \tag{6.18}$$

To identify the limit $p(\tau)$, we use the following lemma, whose proof is given after the proof of Theorem 3.10.

Lemma 6.5 (L^∞ weak-$*$ limits preserve pointwise convex constraints)**.** *Let $I = (a,b)$ and let $p_j \in L^\infty(I;\mathbb{R}^d)$ satisfy*

$$p_j \rightharpoonup^* p \qquad \text{in } L^\infty(I;\mathbb{R}^d).$$

Assume that there exists a measurable set-valued map $C : I \rightrightarrows \mathbb{R}^d$ such that:

1. *$C(\tau)$ is nonempty, closed, convex, and uniformly bounded for a.e. $\tau \in I$;*
2. *$\operatorname{dist}\big(p_j(\tau), C(\tau)\big) \to 0$ for a.e. $\tau \in I$.*

Then,

$$p(\tau) \in C(\tau) \qquad \text{for a.e. } \tau \in I.$$

Step 4: Identification of the limiting inclusion. Fix $\tau \in (a,b)$. By (6.15) and $\tau_j \to +\infty$,

$$Y_\tau^j \to Z_\tau, \qquad Te^{-(\tau+\tau_j)} \to 0.$$

Hence Lemma 6.3 gives

$$\operatorname{dist}\Big(p_j(\tau), \widehat{\partial}\Phi_\lambda(Z_\tau)\Big) \to 0, \qquad p_j(\tau) = \nabla F_\lambda\big(Y_\tau^j, Te^{-(\tau+\tau_j)}\big).$$

We now apply Lemma 6.5 with

$$C(\tau) = \widehat{\partial}\,\Phi_\lambda(Z_\tau).$$

Since $Z([a,b])$ is compact and $\widehat{\partial}\,\Phi_\lambda$ is upper semicontinuous with nonempty compact convex values, C is measurable and uniformly bounded on (a,b). Therefore,

$$p(\tau) \in \widehat{\partial}\Phi_\lambda(Z_\tau) \qquad \text{for a.e. } \tau \in (a,b). \tag{6.19}$$

Together with (6.18), this yields

$$\dot{Z}_\tau \in -\frac{1}{4}\,\widehat{\partial}\Phi_\lambda(Z_\tau) \qquad \text{for a.e. } \tau \in (a,b).$$

Since a, b are arbitrary, the inclusion holds for a.e. $\tau \geq 0$.

In the MoE regime $0 \leq \lambda \leq 1$, Lemma 3.8 gives $\widehat{\partial}\Phi_\lambda = \partial^C \Phi_\lambda$, and therefore

$$\dot{Z}_\tau \in -\frac{1}{4}\,\partial^C \Phi_\lambda(Z_\tau) \qquad \text{for a.e. } \tau \geq 0.$$

Thus Z is a global Carathéodory solution of the corresponding limiting inclusion. This completes the proof of Theorem 3.10.

Proof of Lemma 6.5. Since $C(\tau)$ is uniformly bounded, there exists $M > 0$ such that

$$C(\tau) \subset B(0, M) \qquad \text{for a.e. } \tau \in I.$$

For $\xi \in \mathbb{R}^d$, define the support function

$$h(\tau, \xi) \coloneqq \sup_{q \in C(\tau)} \langle \xi, q \rangle.$$

Then $|h(\tau,\xi)| \leq M\|\xi\|$ for a.e. τ, so $h(\cdot, \xi) \in L^\infty(I)$.

Fix $\xi \in \mathbb{R}^d$. For each j and a.e. $\tau \in I$,

$$\langle \xi, p_j(\tau) \rangle \leq h(\tau, \xi) + \|\xi\| \operatorname{dist}\big(p_j(\tau), C(\tau)\big).$$

Indeed, this follows by taking $q = \Pi_{C(\tau)} p_j(\tau)$ and using the definition of $h(\tau, \xi)$.

Let $\varphi \in L^1(I)$ with $\varphi \geq 0$. Integrating, we get

$$\int_I \varphi(\tau)\, \langle \xi, p_j(\tau) \rangle\, d\tau \leq \int_I \varphi(\tau)\, h(\tau, \xi)\, d\tau + \|\xi\| \int_I \varphi(\tau)\, \operatorname{dist}\big(p_j(\tau), C(\tau)\big)\, d\tau.$$

Since $\operatorname{dist}\big(p_j(\tau), C(\tau)\big) \to 0$ for a.e. $\tau \in I$ and is uniformly bounded, by dominated convergence, the last term tends to 0. Passing to the limit and using $p_j \rightharpoonup^* p$, we obtain

$$\int_I \varphi(\tau)\, \langle \xi, p(\tau) \rangle\, d\tau \leq \int_I \varphi(\tau)\, h(\tau, \xi)\, d\tau \qquad \forall \varphi \in L^1(I),\ \varphi \geq 0.$$

Hence

$$\langle \xi, p(\tau) \rangle \leq h(\tau, \xi) \qquad \text{for a.e. } \tau \in I.$$

Applying this for every $\xi \in \mathbb{Q}^d$, and using that $\mathbb{Q}^d$ is countable, we obtain a full-measure set $I_0 \subset I$ such that for every $\tau \in I_0$,

$$\langle \xi, p(\tau) \rangle \leq h(\tau, \xi) \qquad \forall \xi \in \mathbb{Q}^d.$$

By continuity of $\xi \mapsto h(\tau, \xi)$, the inequality extends to all $\xi \in \mathbb{R}^d$. Therefore,

$$\langle \xi, p(\tau) \rangle \leq \sup_{q \in C(\tau)} \langle \xi, q \rangle \qquad \forall \xi \in \mathbb{R}^d,\ \forall \tau \in I_0.$$

If $p(\tau) \notin C(\tau)$ for some $\tau \in I_0$, the Hahn–Banach separation theorem yields

$$\langle \xi, p(\tau) \rangle > \sup_{q \in C(\tau)} \langle \xi, q \rangle$$

for some $\xi \in \mathbb{R}^d$, a contradiction. Hence $p(\tau) \in C(\tau)$ for every $\tau \in I_0$, that is, for a.e. $\tau \in I$. □

7 Clarke analysis on the geometric potential

The purpose of this section is to analyze the autonomous limiting inclusions arising in Theorem 3.10. Since the geometric potential Φ_λ is only locally Lipschitz in general, the natural framework is that of Clarke generalized gradients and differential inclusions; see Clarke [12] and Aubin–Frankowska [3]. We first clarify the Clarke structure of Φ_λ, then establish the existence of global Carathéodory solutions, and finally show that, in the empirical setting, the piecewise quadratic geometry of Φ_λ forces every such solution to converge to a critical point.

In this section, the support sets A_1 and A_2 are assumed to be compact. Recall that

$$d_i(x) = \operatorname{dist}(x, A_i), \qquad \Phi_\lambda(x) = \lambda d_1(x)^2 + (1-\lambda) d_2(x)^2.$$

Since A_i is compact, the projection set

$$\Pi_i(x) = \operatorname*{argmin}_{a \in A_i} \|x - a\|$$

is nonempty and compact for every $x \in \mathbb{R}^d$.

7.1 Clarke structure of the geometric potential

We begin by proving Lemma 3.8, which clarifies the relation between the Clarke and outer Clarke subdifferentials in the two guidance regimes.

The standard formula for the Clarke subdifferential of the squared distance to a compact set gives

$$\partial^C(d_i^2)(x) = \operatorname{conv}\{\, 2(x-a) : \ a \in \Pi_i(x) \,\}, \qquad i = 1, 2. \tag{7.1}$$

In particular, d_i^2 is differentiable at x if and only if $\Pi_i(x)$ is a singleton.

Proof of Lemma 3.8. We divide the proof into the two regimes.

Step 1: MoE case $(0 \le \lambda \le 1)$**.** In this regime,

$$\Phi_\lambda(x) = \lambda d_1(x)^2 + (1-\lambda) d_2(x)^2 = \min_{a_1 \in A_1} \min_{a_2 \in A_2} \Big(\lambda \|x - a_1\|^2 + (1-\lambda) \|x - a_2\|^2 \Big).$$

Thus Φ_λ is the pointwise minimum of the C^1 family

$$Q_{a_1,a_2}(x) := \lambda \|x - a_1\|^2 + (1-\lambda) \|x - a_2\|^2, \qquad (a_1, a_2) \in A_1 \times A_2.$$

A pair (a_1, a_2) is active at x if and only if

$$Q_{a_1,a_2}(x) = \Phi_\lambda(x).$$

Since both coefficients are nonnegative, this is equivalent to

$$a_1 \in \Pi_1(x), \qquad a_2 \in \Pi_2(x),$$

for $0 < \lambda < 1$; the endpoint cases $\lambda = 0$ and $\lambda = 1$ follow by the same argument after the obvious simplification.

Applying the Clarke–Danskin formula to

$$-\Phi_\lambda(x) = \max_{a_1 \in A_1,\ a_2 \in A_2} \big(-Q_{a_1,a_2}(x) \big),$$

and using $\partial^C(-f)(x) = -\partial^C f(x)$, we obtain

$$\partial^C \Phi_\lambda(x) = \overline{\mathrm{conv}}\Big\{\nabla Q_{a_1,a_2}(x) : (a_1,a_2) \in \Pi_1(x) \times \Pi_2(x)\Big\}.$$

Since

$$\nabla Q_{a_1,a_2}(x) = 2\lambda(x-a_1) + 2(1-\lambda)(x-a_2),$$

we obtain

$$\partial^C \Phi_\lambda(x) = \overline{\mathrm{conv}}\Big\{2\lambda(x-a_1) + 2(1-\lambda)(x-a_2) : a_1 \in \Pi_1(x),\ a_2 \in \Pi_2(x)\Big\}.$$

Because $\Pi_1(x)$ and $\Pi_2(x)$ are compact, the set of active gradients is compact, hence its convex hull is compact, and the closure may be removed. Therefore

$$\partial^C \Phi_\lambda(x) = \mathrm{conv}\Big\{2\lambda(x-a_1) + 2(1-\lambda)(x-a_2) : a_1 \in \Pi_1(x),\ a_2 \in \Pi_2(x)\Big\}. \tag{7.2}$$

Together with (7.1) and the definition of $\widehat{\partial}\,\Phi_\lambda$ in (3.10), this yields (3.13).

Step 2: CFG case ($\lambda > 1$)**.** For arbitrary compact supports A_1, A_2, the general Clarke sum rule gives

$$\partial^C \Phi_\lambda(x) = \partial^C\big(\lambda d_1^2 + (1-\lambda) d_2^2\big)(x) \subseteq \lambda\, \partial^C(d_1^2)(x) + (1-\lambda)\, \partial^C(d_2^2)(x) = \widehat{\partial}\Phi_\lambda(x).$$

Assume now that A_1 and A_2 are finite. If

$$x \notin \mathrm{ND}(d_1^2) \cap \mathrm{ND}(d_2^2),$$

then at least one of d_1^2 and d_2^2 is smooth in a neighborhood of x, since outside the Voronoi interface the nearest point is locally unique and locally constant. The exact Clarke sum rule with one smooth term therefore yields

$$\partial^C \Phi_\lambda(x) = \lambda\, \partial^C(d_1^2)(x) + (1-\lambda)\, \partial^C(d_2^2)(x) = \widehat{\partial}\Phi_\lambda(x).$$

On the simultaneous interface $\mathrm{ND}(d_1^2) \cap \mathrm{ND}(d_2^2)$, the general inclusion above is the only statement available in general. This proves the CFG claim. □

7.2 Existence of global Carathéodory solutions

Lemma 7.1. *Assume that A_1 and A_2 are compact. Then, for every $z_0 \in \mathbb{R}^d$, the Cauchy problems (3.16) and (3.17) with initial datum $Z_0 = z_0$ admit at least one global Carathéodory solution.*

Proof. By the Viability Theorem [3, Thm. 10.1.6], it suffices to check that the right-hand sides of (3.16) and (3.17) are Peano maps, namely upper semicontinuous set-valued maps with nonempty, convex, compact values and at most linear growth.

Since d_1^2 and d_2^2 are locally Lipschitz on $\mathbb{R}^d$, the Clarke subdifferentials

$$\partial^C(d_1^2), \qquad \partial^C(d_2^2)$$

are upper semicontinuous and take nonempty, convex, compact values. Hence the same is true for

$$z \longmapsto -\frac{1}{4}\,\partial^C \Phi_\lambda(z)$$

in the MoE case, and for

$$z \longmapsto -\frac{1}{4}\,\widehat{\partial}\,\Phi_\lambda(z) = -\frac{1}{4}\Big(\lambda\,\partial^C(d_1^2)(z) + (1-\lambda)\,\partial^C(d_2^2)(z)\Big)$$

in the CFG case.

It remains to check the linear growth. Since A_1 and A_2 are compact, there exists $R > 0$ such that

$$A_1 \cup A_2 \subset B(0,R).$$

Moreover, for every $z \in \mathbb{R}^d$ and $i \in \{1,2\}$,

$$\partial^C(d_i^2)(z) \subset 2\big(z - \mathrm{conv}(A_i)\big),$$

so every $p_i \in \partial^C(d_i^2)(z)$ satisfies

$$\|p_i\| \le 2(\|z\| + R).$$

It follows that every element of both multifunctions

$$-\frac{1}{4}\,\partial^C\Phi_\lambda(z) \qquad \text{and} \qquad -\frac{1}{4}\,\widehat{\partial}\,\Phi_\lambda(z)$$

has norm bounded by $C(1+\|z\|)$ for some constant $C > 0$ independent of z.

Therefore both right-hand sides are Peano maps. The Viability Theorem then yields the existence of at least one global Carathéodory solution to (3.16) and (3.17). □

7.3 Empirical rigidity of the geometric potential

We now specialize to the empirical setting

$$A_1 = \{x_1, \dots, x_{n_1}\}, \qquad A_2 = \{y_1, \dots, y_{n_2}\}.$$

Then the geometric potential Φ_λ is piecewise quadratic on a finite stratification of $\mathbb{R}^d$. This yields the rigidity properties needed for the convergence analysis of autonomous solutions.

For each nonempty pair of index sets

$$I \subset \{1, \dots, n_1\}, \qquad J \subset \{1, \dots, n_2\},$$

we define the corresponding stratum

$$\Sigma_{I,J} := \Big\{x \in \mathbb{R}^d :\ \Pi_1(x) = \{x_k :\ k \in I\},\ \Pi_2(x) = \{y_\ell :\ \ell \in J\}\Big\}.$$

Since A_1 and A_2 are finite, only finitely many such strata are nonempty, and

$$\mathbb{R}^d = \bigcup_{I,J} \Sigma_{I,J}.$$

For $(k,\ell) \in \{1, \dots, n_1\} \times \{1, \dots, n_2\}$, we set

$$Q^\lambda_{k\ell}(x) := \lambda\|x - x_k\|^2 + (1-\lambda)\|x - y_\ell\|^2.$$

If $x \in \Sigma_{I,J}$, then

$$d_1(x)^2 = \|x - x_k\|^2 \quad \forall k \in I, \qquad d_2(x)^2 = \|x - y_\ell\|^2 \quad \forall \ell \in J,$$

so all active branches $Q^\lambda_{k\ell}$, $(k,\ell) \in I \times J$, coincide on $\Sigma_{I,J}$.

Let $M_{I,J}$ denote the affine hull of $\Sigma_{I,J}$, and let $V_{I,J}$ be the vector subspace parallel to $M_{I,J}$. Since the differences $Q^\lambda_{k\ell} - Q^\lambda_{k'\ell'}$ are affine and vanish on $\Sigma_{I,J}$, they also vanish on $M_{I,J}$. Hence the restriction

$$q_{I,J} \coloneqq Q^\lambda_{k\ell}\Big|_{M_{I,J}}$$

is well defined, independently of the choice of $(k,\ell) \in I \times J$. Moreover, each $Q^\lambda_{k\ell}$ has Hessian $2I$, so $q_{I,J}$ is a strongly convex quadratic function on $M_{I,J}$.

Lemma 7.2 (Empirical rigidity of critical points and MoE minimizers)**.** *In the empirical setting, the following properties hold:*

1. *For every $\lambda \geq 0$, the outer Clarke critical set*

$$\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$$

is finite. Consequently, the Clarke critical set

$$\mathrm{Crit}(\Phi_\lambda)$$

is finite.

2. *Assume in addition that $0 \leq \lambda \leq 1$. If x^* is a local minimizer of Φ_λ, then there exists a neighborhood U of x^* such that*

$$\Phi_\lambda(x) = \Phi_\lambda(x^*) + \|x - x^*\|^2 \qquad \forall x \in U.$$

In particular, x^* is an isolated strict local minimizer, Φ_λ is smooth in U, and

$$\nabla\Phi_\lambda(x) = 2(x - x^*) \qquad \forall x \in U.$$

3. *If $0 < \lambda < 1$, then every local minimizer of Φ_λ lies outside the full interface set:*

$$x^* \notin \mathrm{ND}(A_1, A_2) = \mathrm{ND}(d_1^2) \cup \mathrm{ND}(d_2^2).$$

If $\lambda = 0$ (resp. $\lambda = 1$), then every local minimizer of Φ_λ lies outside the active interface:

$$x^* \notin \mathrm{ND}(d_2^2) \qquad (\text{resp. } x^* \notin \mathrm{ND}(d_1^2)).$$

Proof. We prove the three assertions separately.

Step 1: Finiteness of outer Clarke critical points. Fix a nonempty stratum $\Sigma_{I,J}$, and let $x \in \Sigma_{I,J}$ satisfy

$$0 \in \widehat{\partial}\Phi_\lambda(x).$$

Since

$$\Pi_1(x) = \{x_k : \ k \in I\}, \qquad \Pi_2(x) = \{y_\ell : \ \ell \in J\},$$

formula (7.1) gives

$$\partial^C(d_1^2)(x) = \mathrm{conv}\{2(x - x_k) : \ k \in I\}, \qquad \partial^C(d_2^2)(x) = \mathrm{conv}\{2(x - y_\ell) : \ \ell \in J\}.$$

Therefore

$$\widehat{\partial}\Phi_\lambda(x) = \mathrm{conv}\Big\{\nabla Q^\lambda_{k\ell}(x) : \ (k,\ell) \in I \times J\Big\}.$$

Thus there exist coefficients $\theta_{k\ell} \geq 0$, with $\sum_{(k,\ell)\in I\times J} \theta_{k\ell} = 1$, such that

$$0 = \sum_{(k,\ell)\in I\times J} \theta_{k\ell} \nabla Q^{\lambda}_{k\ell}(x).$$

Let $M_{I,J}$ be the affine hull of $\Sigma_{I,J}$, and let $V_{I,J}$ be its direction space. Since all active branches coincide on $M_{I,J}$, their tangential gradients coincide:

$$P_{V_{I,J}} \nabla Q^{\lambda}_{k\ell}(x) = \nabla q_{I,J}(x) \qquad \forall (k,\ell) \in I \times J.$$

Projecting the convex-combination identity onto $V_{I,J}$, we obtain

$$0 = \nabla q_{I,J}(x).$$

Since $q_{I,J}$ is a strongly convex quadratic on $M_{I,J}$, it has at most one critical point. Hence each stratum contains at most one outer Clarke critical point. Since there are only finitely many nonempty strata, $\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$ is finite. The finiteness of $\mathrm{Crit}(\Phi_\lambda)$ follows from

$$\partial^C \Phi_\lambda(x) \subseteq \widehat{\partial} \Phi_\lambda(x) \qquad \forall x \in \mathbb{R}^d.$$

Step 2: Local minimizers in the MoE regime. Assume $0 \leq \lambda \leq 1$. Then

$$\Phi_\lambda(x) = \min_{1\leq k\leq n_1,\ 1\leq \ell\leq n_2} Q^{\lambda}_{k\ell}(x).$$

Let x^* be a local minimizer of Φ_λ, and denote by

$$\mathcal{A}(x^*) \coloneqq \big\{(k,\ell) :\ Q^{\lambda}_{k\ell}(x^*) = \Phi_\lambda(x^*)\big\}$$

the active branch set at x^*.

For each $(k,\ell) \in \mathcal{A}(x^*)$, Taylor's formula gives

$$Q^{\lambda}_{k\ell}(x^* + h) = Q^{\lambda}_{k\ell}(x^*) + \langle \nabla Q^{\lambda}_{k\ell}(x^*), h\rangle + \|h\|^2.$$

Since x^* is a local minimizer of $\Phi_\lambda = \min Q^{\lambda}_{k\ell}$, for all sufficiently small h,

$$0 \leq \Phi_\lambda(x^* + h) - \Phi_\lambda(x^*) \leq \min_{(k,\ell)\in\mathcal{A}(x^*)} \Big(\langle \nabla Q^{\lambda}_{k\ell}(x^*), h\rangle + \|h\|^2\Big).$$

Taking $h = r\xi$, dividing by r, and letting $r \to 0^+$, we obtain

$$\min_{(k,\ell)\in\mathcal{A}(x^*)} \langle \nabla Q^{\lambda}_{k\ell}(x^*), \xi\rangle \geq 0 \qquad \forall \xi \in \mathbb{S}^{d-1}.$$

Applying this also to $-\xi$, we conclude that

$$\nabla Q^{\lambda}_{k\ell}(x^*) = 0 \qquad \forall (k,\ell) \in \mathcal{A}(x^*).$$

Since

$$\nabla Q^{\lambda}_{k\ell}(x) = 2\big(x - \lambda x_k - (1-\lambda) y_\ell\big),$$

we get

$$x^* = \lambda x_k + (1-\lambda) y_\ell \qquad \forall (k,\ell) \in \mathcal{A}(x^*).$$

Hence all active branches have the same center x^*. Since they also coincide at x^*, they are identical:

$$Q^\lambda_{k\ell}(x) \equiv \Phi_\lambda(x^*) + \|x - x^*\|^2 \qquad \forall (k,\ell) \in \mathcal{A}(x^*).$$

All inactive branches are strictly above $\Phi_\lambda(x^*)$ at x^*, and therefore remain inactive in a neighborhood of x^*. Thus there exists a neighborhood U of x^* such that

$$\Phi_\lambda(x) = \Phi_\lambda(x^*) + \|x - x^*\|^2 \qquad \forall x \in U.$$

The remaining conclusions of (2) follow immediately.

Step 3: Location of MoE minimizers. Assume first that $0 < \lambda < 1$, and let x^* be a local minimizer of Φ_λ. Define

$$I^* = \{k : \ x_k \in \Pi_1(x^*)\}, \qquad J^* = \{\ell : \ y_\ell \in \Pi_2(x^*)\}.$$

Then the active branch set is

$$\mathcal{A}(x^*) = I^* \times J^*.$$

By Step 2, for every $(k,\ell) \in I^* \times J^*$,

$$\nabla Q^\lambda_{k\ell}(x^*) = 0,$$

or equivalently

$$x^* = \lambda x_k + (1-\lambda) y_\ell.$$

Since $0 < \lambda < 1$, both coefficients λ and $1-\lambda$ are strictly positive. Fixing $\ell \in J^*$ and comparing two indices $k, k' \in I^*$ gives $x_k = x_{k'}$. Similarly, fixing $k \in I^*$ and comparing $\ell, \ell' \in J^*$ gives $y_\ell = y_{\ell'}$. Hence the nearest support point is unique in both A_1 and A_2, and therefore

$$x^* \notin \mathrm{ND}(d_1^2) \cup \mathrm{ND}(d_2^2) = \mathrm{ND}(A_1, A_2).$$

If $\lambda = 0$, then $\Phi_\lambda = d_2^2$, and the same argument applied to the active branches of d_2^2 shows that $x^* \notin \mathrm{ND}(d_2^2)$. Similarly, if $\lambda = 1$, then $\Phi_\lambda = d_1^2$, and every local minimizer satisfies $x^* \notin \mathrm{ND}(d_1^2)$. This proves (3). □

Remark 7.3 (Local minimizers in the MoE and CFG regimes)**.** Lemma 7.2 shows that the MoE regime has a rigid local structure. If $0 \le \lambda \le 1$, every local minimizer x^* of Φ_λ is a strict quadratic minimizer: in a neighborhood of x^*,

$$\Phi_\lambda(x) = \Phi_\lambda(x^*) + \|x - x^*\|^2.$$

In particular, Φ_λ is smooth near x^*. Moreover, in the interior MoE regime $0 < \lambda < 1$, the nearest points in both supports are unique at x^*, and hence

$$x^* \notin \mathrm{ND}(A_1, A_2).$$

Consequently, near MoE local minimizers the dynamics falls into the smooth quadratic setting covered by the local estimate (6.9), with the endpoint convention stated in Lemma 6.4. Thus the non-autonomous drift converges locally uniformly and exponentially fast to the smooth limiting gradient field.

By contrast, in the CFG regime $\lambda > 1$, local minimizers may genuinely lie on the interface set $\mathrm{ND}(A_1, A_2)$. In that case the limiting potential can remain nonsmooth at the minimizer, and no analogous local quadratic rigidity or uniform exponential drift estimate is available in general. This is why the rate result in the CFG regime is stated under the additional assumption that the limiting minimizer lies outside $\mathrm{ND}(A_1, A_2)$.

7.4 Proof of Theorem 3.16

We prove the convergence theorem for the autonomous limiting inclusions in the empirical setting.

The existence statement follows from Lemma 7.1. It remains to prove the Lyapunov identity and the convergence of global solutions.

Step 1: Lyapunov identity. Let Z be a global Carathéodory solution of either (3.16) or (3.17). Thus there exists a measurable selection ξ_τ such that

$$Z_\tau = Z_0 + \int_0^\tau \xi_s \, ds, \qquad \dot{Z}_\tau = \xi_\tau \quad \text{for a.e. } \tau \geq 0,$$

and

$$\xi_\tau \in -\frac{1}{4}\mathcal{G}(Z_\tau) \quad \text{for a.e. } \tau \geq 0,$$

where

$$\mathcal{G}(z) = \partial^C \Phi_\lambda(z) \quad \text{in the MoE case,} \qquad \mathcal{G}(z) = \widehat{\partial}\, \Phi_\lambda(z) \quad \text{in the CFG case.}$$

Equivalently,

$$-4\xi_\tau \in \mathcal{G}(Z_\tau) \qquad \text{for a.e. } \tau \geq 0. \tag{7.3}$$

We claim that

$$\frac{d}{d\tau}\Phi_\lambda(Z_\tau) = -4\|\dot{Z}_\tau\|^2 \qquad \text{for a.e. } \tau \geq 0. \tag{7.4}$$

Since Φ_λ is locally Lipschitz and Z is locally absolutely continuous, the composition $\Phi_\lambda \circ Z$ is locally absolutely continuous. Hence it is differentiable for a.e. τ.

Fix $0 \leq a < b < \infty$. For each nonempty stratum $\Sigma_{I,J}$, set

$$E_{I,J} := \{\tau \in [a,b] : Z_\tau \in \Sigma_{I,J}\}.$$

The sets $E_{I,J}$ are measurable, and since the family of nonempty strata is finite and covers $\mathbb{R}^d$, the interval $[a,b]$ is covered by the sets $E_{I,J}$.

We now fix a nonempty stratum $\Sigma_{I,J}$ and prove (7.4) for a.e. $\tau \in E_{I,J}$. Let $\tau \in E_{I,J}$ be such that

1. τ is a Lebesgue density point of $E_{I,J}$;
2. Z is differentiable at τ, with $\dot{Z}_\tau = \xi_\tau$;
3. $\Phi_\lambda \circ Z$ is differentiable at τ;
4. the inclusion (7.3) holds at τ.

These properties hold for a.e. $\tau \in E_{I,J}$.

We first show that

$$\xi_\tau = \dot{Z}_\tau \in V_{I,J}, \tag{7.5}$$

where $V_{I,J}$ denotes the direction space of the affine hull $M_{I,J}$ of the stratum. Since τ is a density point of $E_{I,J}$, there exists a sequence $h_n \to 0$ such that $\tau + h_n \in E_{I,J}$. Therefore

$$Z_{\tau+h_n} \in \Sigma_{I,J} \subset M_{I,J}, \qquad Z_\tau \in \Sigma_{I,J} \subset M_{I,J}.$$

Hence

$$\frac{Z_{\tau+h_n} - Z_\tau}{h_n} \in V_{I,J}.$$

Passing to the limit and using the differentiability of Z at τ, we obtain (7.5).

Next, fix any active pair $(k,\ell) \in I \times J$. On the stratum $\Sigma_{I,J}$, the potential Φ_λ agrees with the active branch $Q^\lambda_{k\ell}$. Since τ is a density point of $E_{I,J}$, the two functions

$$s \mapsto \Phi_\lambda(Z_s), \qquad s \mapsto Q^\lambda_{k\ell}(Z_s)$$

coincide on a set of density one at τ. Both are differentiable at τ. Hence their derivatives agree, and

$$\frac{d}{d\tau}\Phi_\lambda(Z_\tau) = \frac{d}{d\tau}Q^\lambda_{k\ell}(Z_\tau) = \langle \nabla Q^\lambda_{k\ell}(Z_\tau), \xi_\tau \rangle.$$

Since $\xi_\tau \in V_{I,J}$, only the tangential component of $\nabla Q^\lambda_{k\ell}(Z_\tau)$ contributes. Therefore

$$\frac{d}{d\tau}\Phi_\lambda(Z_\tau) = \left\langle P_{V_{I,J}} \nabla Q^\lambda_{k\ell}(Z_\tau), \xi_\tau \right\rangle = \langle \nabla_{M_{I,J}} q_{I,J}(Z_\tau), \xi_\tau \rangle. \tag{7.6}$$

Here $q_{I,J}$ denotes the common restriction of the active branches to $M_{I,J}$, and $\nabla_{M_{I,J}} q_{I,J} \in V_{I,J}$ denotes its gradient on the affine space $M_{I,J}$.

We now use the differential inclusion. From (7.3) and from the subdifferential formula on the stratum, we have

$$-4\xi_\tau \in \operatorname{conv}\Big\{ \nabla Q^\lambda_{k\ell}(Z_\tau) : (k,\ell) \in I \times J \Big\}.$$

Indeed, this follows from (7.2) in the MoE case, and from the definition of $\widehat{\partial}\Phi_\lambda$ together with (7.1) in the CFG case. Thus there exist coefficients $\theta_{k\ell} \geq 0$, with

$$\sum_{(k,\ell)\in I\times J} \theta_{k\ell} = 1,$$

such that

$$-4\xi_\tau = \sum_{(k,\ell)\in I\times J} \theta_{k\ell} \nabla Q^\lambda_{k\ell}(Z_\tau). \tag{7.7}$$

All active branches have the same restriction $q_{I,J}$ on $M_{I,J}$. Therefore their tangential gradients coincide:

$$P_{V_{I,J}} \nabla Q^\lambda_{k\ell}(Z_\tau) = \nabla_{M_{I,J}} q_{I,J}(Z_\tau), \qquad (k,\ell) \in I \times J. \tag{7.8}$$

Projecting (7.7) onto $V_{I,J}$ and using (7.8), we obtain

$$P_{V_{I,J}}(-4\xi_\tau) = \nabla_{M_{I,J}} q_{I,J}(Z_\tau).$$

Since $\xi_\tau \in V_{I,J}$, this gives

$$\nabla_{M_{I,J}} q_{I,J}(Z_\tau) = -4\xi_\tau.$$

Combining this identity with (7.6), we conclude that

$$\frac{d}{d\tau}\Phi_\lambda(Z_\tau) = \langle -4\xi_\tau, \xi_\tau \rangle = -4\|\xi_\tau\|^2 = -4\|\dot{Z}_\tau\|^2.$$

This proves (7.4) for a.e. $\tau \in E_{I,J}$.

Since there are only finitely many nonempty strata, the identity holds for a.e. $\tau \in [a,b]$. Integrating over $[a,b]$, we obtain

$$\Phi_\lambda(Z_b) - \Phi_\lambda(Z_a) = -4\int_a^b \|\dot{Z}_\tau\|^2 \, d\tau. \tag{7.9}$$

In particular, $\tau \mapsto \Phi_\lambda(Z_\tau)$ is nonincreasing.

Step 2: Compactness of the trajectory. We next show that the trajectory is bounded. Since $A_1 \cup A_2$ is finite, there exists $R_A > 0$ such that

$$\|a\| \le R_A \qquad \forall a \in A_1 \cup A_2.$$

For every $x \in \mathbb{R}^d$, choose $a_i \in \Pi_i(x)$, $i = 1, 2$. Then

$$\Phi_\lambda(x) = \lambda\|x - a_1\|^2 + (1-\lambda)\|x - a_2\|^2.$$

Expanding the squares gives

$$\Phi_\lambda(x) = \|x\|^2 - 2\langle x, \lambda a_1 + (1-\lambda)a_2\rangle + \lambda\|a_1\|^2 + (1-\lambda)\|a_2\|^2.$$

The last two terms are bounded from below by $-C$, while the linear term is bounded by $C\|x\|$. Hence

$$\Phi_\lambda(x) \ge \|x\|^2 - C\|x\| - C,$$

and therefore

$$\Phi_\lambda(x) \to +\infty \qquad \text{as } \|x\| \to \infty.$$

Thus Φ_λ is coercive.

By the Lyapunov monotonicity,

$$\Phi_\lambda(Z_\tau) \le \Phi_\lambda(Z_0) \qquad \forall \tau \ge 0.$$

Therefore Z_τ remains in the compact sublevel set

$$K \coloneqq \{x \in \mathbb{R}^d : \Phi_\lambda(x) \le \Phi_\lambda(Z_0)\}.$$

Consequently, the ω-limit set $\omega_\tau(Z)$ is nonempty, compact, and connected.

Moreover, since $\Phi_\lambda(Z_\tau)$ is nonincreasing and bounded from below, there exists $\ell \in \mathbb{R}$ such that

$$\Phi_\lambda(Z_\tau) \to \ell \qquad \text{as } \tau \to \infty.$$

By continuity of Φ_λ, every $x \in \omega_\tau(Z)$ satisfies

$$\Phi_\lambda(x) = \ell.$$

Hence Φ_λ is constant on $\omega_\tau(Z)$.

Step 3: The ω-limit set is critical. We prove that

$$\omega_\tau(Z) \subset \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda). \tag{7.10}$$

We argue by contradiction. Let

$$x^* \in \omega_\tau(Z) \setminus \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda).$$

Since $0 \notin \widehat{\partial}\,\Phi_\lambda(x^*)$, and since $\widehat{\partial}\,\Phi_\lambda$ is upper semicontinuous with nonempty compact values, there exist $r > 0$ and $\eta > 0$ such that

$$\inf_{\zeta \in \widehat{\partial}\,\Phi_\lambda(x)} \|\zeta\| \ge \eta \qquad \forall x \in B(x^*, r). \tag{7.11}$$

If $\omega_\tau(Z) = \{x^*\}$, then $Z_\tau \to x^*$. For all large τ, $Z_\tau \in B(x^*, r)$. From the inclusion and (7.11), we get

$$\|\dot{Z}_\tau\| \geq \frac{\eta}{4} \qquad \text{for a.e. large } \tau.$$

The Lyapunov identity then gives

$$\frac{d}{d\tau}\Phi_\lambda(Z_\tau) = -4\|\dot{Z}_\tau\|^2 \leq -\frac{\eta^2}{4} \qquad \text{for a.e. large } \tau,$$

which contradicts the convergence of $\Phi_\lambda(Z_\tau)$ to ℓ. Thus the singleton case is impossible unless x^* is critical.

We now consider the case where $\omega_\tau(Z)$ contains another point. Choose

$$y^* \in \omega_\tau(Z), \qquad y^* \neq x^*.$$

Shrinking $r > 0$ if necessary, we may assume that

$$y^* \notin B(x^*, 2r).$$

Since the trajectory remains in the compact set K, and the multifunction $-\frac{1}{4}\mathcal{G}$ is bounded on K, there exists $M > 0$ such that

$$\|\dot{Z}_\tau\| \leq M \qquad \text{for a.e. } \tau \geq 0. \tag{7.12}$$

Let $\tau_j \to \infty$ be such that

$$Z_{\tau_j} \to x^*.$$

For j large enough,

$$Z_{\tau_j} \in B(x^*, r/2) \qquad \text{and} \qquad \Phi_\lambda(Z_{\tau_j}) < \ell + \frac{\eta^2 r}{32M}.$$

Since $y^* \in \omega_\tau(Z)$ and $y^* \notin B(x^*, 2r)$, the trajectory cannot remain forever in $B(x^*, r)$ after time τ_j. Thus it must exit $B(x^*, r)$ at some later time. Starting from $B(x^*, r/2)$, and using the velocity bound (7.12), the trajectory spends at least time $r/(2M)$ inside $B(x^*, r)$ before exiting.

During this time interval, (7.11) and the inclusion imply

$$\|\dot{Z}_\tau\| \geq \frac{\eta}{4} \qquad \text{for a.e. } \tau$$

as long as $Z_\tau \in B(x^*, r)$. Hence, by the Lyapunov identity, the value of Φ_λ decreases by at least

$$4\left(\frac{\eta}{4}\right)^2 \frac{r}{2M} = \frac{\eta^2 r}{8M}.$$

Consequently, after the trajectory exits $B(x^*, r)$, we have

$$\Phi_\lambda(Z_\tau) \leq \Phi_\lambda(Z_{\tau_j}) - \frac{\eta^2 r}{8M} < \ell - \frac{3\eta^2 r}{32M}.$$

Since $\Phi_\lambda(Z_\tau)$ is nonincreasing, this strict inequality persists for all later times, contradicting

$$\Phi_\lambda(Z_\tau) \to \ell.$$

This contradiction proves (7.10).

Step 4: Reduction to a single point. By Lemma 7.2, $\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$ is finite. Since $\omega_\tau(Z)$ is compact, connected, and satisfies

$$\omega_\tau(Z) \subset \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda),$$

it must be a singleton. Thus there exists $x^* \in \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$ such that

$$\omega_\tau(Z) = \{x^*\}, \qquad Z_\tau \to x^* \quad \text{as } \tau \to \infty.$$

In the MoE regime, Lemma 3.8 gives $\widehat{\partial}\Phi_\lambda = \partial^C \Phi_\lambda$, hence $x^* \in \mathrm{Crit}(\Phi_\lambda)$. This completes the proof.

8 Autonomous stability and empirical convergence

This section is devoted to the proof of Theorem 3.18 and Corollary 3.20. The strategy follows the classical compactness approach for asymptotically autonomous systems, in the spirit of Markus [39] and Galaktionov–Vázquez [19, Thm. 3], adapted here to our nonsmooth limiting differential inclusions.

The proof proceeds in three steps. We first identify the local trap structure of local minimizers for the autonomous limiting inclusions. We then use this autonomous stability property, together with time-shift compactness and the convergence of limiting trajectories established earlier, to prove the empirical convergence criterion. Finally, we establish the local convergence rate near a limiting local minimizer by a perturbative argument around the linearized limiting field.

8.1 Local trap properties of local minimizers

We begin by introducing the trapping notion associated with the autonomous limiting inclusions.

Definition 8.1 (Local trap for a differential inclusion)**.** Let $\mathcal{F} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map, and consider the differential inclusion

$$\dot{Z}_\tau \in \mathcal{F}(Z_\tau), \qquad \tau \geq 0.$$

A point $p \in \mathbb{R}^d$ is called a *local trap* for this differential inclusion if, for every $\varepsilon > 0$ sufficiently small, there exists $\delta \in (0, \varepsilon)$ such that every global Carathéodory solution Z with

$$Z_0 \in B(p, \delta)$$

satisfies

$$Z_\tau \in B(p, \varepsilon) \qquad \forall \tau \geq 0,$$

and

$$Z_\tau \to p \qquad \text{as } \tau \to \infty.$$

Lemma 8.2 (Local minimizers are local traps for the autonomous limiting system)**.** *Assume that $u_{0,1}$ and $u_{0,2}$ are finite Dirac mixtures. Then*

$$p \in \mathrm{Min}(\Phi_\lambda)$$

if and only if p is a local trap for (3.16) when $0 \leq \lambda \leq 1$, and for (3.17) when $\lambda > 1$.

Proof. We first prove that every local minimizer is a local trap.

Assume that

$$p \in \mathrm{Min}(\Phi_\lambda).$$

Since every local minimizer is in particular an outer Clarke critical point, one has

$$\mathrm{Min}(\Phi_\lambda) \subset \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda).$$

By Lemma 7.2, the set $\mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$ is finite. Hence p is isolated. Choose $\varepsilon_0 > 0$ such that

$$\overline{B}(p, \varepsilon_0) \cap \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda) = \{p\}.$$

Since p is a local minimizer, after possibly reducing ε_0 we may also assume that

$$\Phi_\lambda(x) \geq \Phi_\lambda(p) \qquad \forall x \in B(p, \varepsilon_0).$$

Since p is an isolated local minimizer, after reducing $\varepsilon_0 > 0$ if necessary, p is a strict minimizer on $\overline{B}(p, \varepsilon_0)$. Hence, for every $\varepsilon \in (0, \varepsilon_0]$,

$$m_\varepsilon := \min_{x \in \partial B(p,\varepsilon)} \Phi_\lambda(x) > \Phi_\lambda(p).$$

Again by continuity of Φ_λ at p, there exists $\delta \in (0, \varepsilon)$ such that

$$\Phi_\lambda(x) < m_\varepsilon \qquad \forall x \in B(p, \delta).$$

Let Z be a global Carathéodory solution of the corresponding autonomous limiting inclusion with

$$Z_0 \in B(p, \delta).$$

By Theorem 3.16,

$$\Phi_\lambda(Z_b) - \Phi_\lambda(Z_a) = -4 \int_a^b \|\dot{Z}_\tau\|^2 \, d\tau \qquad \forall 0 \le a \le b < \infty.$$

In particular, $\tau \mapsto \Phi_\lambda(Z_\tau)$ is nonincreasing, hence

$$\Phi_\lambda(Z_\tau) \le \Phi_\lambda(Z_0) < m_\varepsilon \qquad \forall \tau \ge 0.$$

Therefore Z_τ can never reach the sphere $\partial B(p, \varepsilon)$, since every point of that sphere has energy at least m_ε. Thus

$$Z_\tau \in B(p, \varepsilon) \qquad \forall \tau \ge 0.$$

By Theorem 3.16, there exists

$$x^* \in \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda)$$

such that

$$Z_\tau \to x^* \qquad \text{as } \tau \to \infty.$$

Since $Z_\tau \in B(p, \varepsilon)$ for all $\tau \ge 0$, one has

$$x^* \in \overline{B}(p, \varepsilon) \cap \mathrm{Crit}_{\mathrm{out}}(\Phi_\lambda) = \{p\}.$$

Hence $x^* = p$, and therefore

$$Z_\tau \to p \qquad \text{as } \tau \to \infty.$$

This proves that p is a local trap.

Conversely, assume that p is a local trap for the corresponding autonomous limiting inclusion. We show that p is a local minimizer of Φ_λ.

Suppose by contradiction that p is not a local minimizer. Then for every $r > 0$ there exists $z_r \in B(p, r)$ such that

$$\Phi_\lambda(z_r) < \Phi_\lambda(p).$$

By the local trap property, for $r > 0$ small enough, every global Carathéodory solution Z with initial datum

$$Z_0 = z_r$$

satisfies

$$Z_\tau \to p \qquad \text{as } \tau \to \infty.$$

On the other hand, by Lyapunov monotonicity,

$$\Phi_\lambda(Z_\tau) \le \Phi_\lambda(Z_0) = \Phi_\lambda(z_r) < \Phi_\lambda(p) \qquad \forall \tau \ge 0.$$

Passing to the limit as $\tau \to \infty$, and using the continuity of Φ_λ, we obtain

$$\Phi_\lambda(p) = \lim_{\tau \to \infty} \Phi_\lambda(Z_\tau) \le \Phi_\lambda(z_r) < \Phi_\lambda(p),$$

a contradiction. Therefore $p \in \mathrm{Min}(\Phi_\lambda)$. □

8.2 Proof of Theorem 3.18

We first prove that

$$\omega_\tau(Y) \subset \mathrm{Min}(\Phi_\lambda).$$

It is enough to show that

$$\mathrm{dist}\big(Y_\tau, \mathrm{Min}(\Phi_\lambda)\big) \to 0 \qquad \text{as } \tau \to \infty.$$

Assume by contradiction that this is false. Then there exist $\varepsilon_0 > 0$ and a sequence $\tau_j \to \infty$ such that

$$\mathrm{dist}\big(Y_{\tau_j}, \mathrm{Min}(\Phi_\lambda)\big) > \varepsilon_0 \qquad \forall j.$$

For each $\varepsilon \in (0, \varepsilon_0]$, define

$$\mathcal{O}_\varepsilon := \Big\{\tau > 0 : \ \mathrm{dist}\big(Y_\tau, \mathrm{Min}(\Phi_\lambda)\big) > \varepsilon\Big\}.$$

Since the map

$$\tau \longmapsto \mathrm{dist}\big(Y_\tau, \mathrm{Min}(\Phi_\lambda)\big)$$

is continuous, the set $\mathcal{O}_\varepsilon$ is open. For each j, let I_j^ε denote the connected component of $\mathcal{O}_\varepsilon$ containing τ_j. In particular, I_j^ε is a nonempty open interval. Moreover, if $0 < \varepsilon_2 \le \varepsilon_1 \le \varepsilon_0$, then

$$\mathcal{O}_{\varepsilon_1} \subset \mathcal{O}_{\varepsilon_2},$$

and therefore

$$I_j^{\varepsilon_1} \subset I_j^{\varepsilon_2}.$$

We will need the following lemma, which provides a uniform bound on the bad intervals and is the key ingredient in the contradiction argument. This type of estimate already appears in [19, Lem. 3.2]. For completeness, we defer its proof until after the proof of Theorem 3.18.

Lemma 8.3. *For every $\varepsilon \in (0, \varepsilon_0]$, there exists a constant $C_\varepsilon > 0$ such that*

$$|I_j^\varepsilon| \le C_\varepsilon \qquad \forall j \ge 1.$$

We now return to the proof of the theorem. Fix

$$\rho \in (0, \varepsilon_0)$$

sufficiently small. By Lemma 8.2 and the finiteness of $\mathrm{Min}(\Phi_\lambda)$, there exists $\delta_\rho \in (0, \rho/2)$ such that, for every global Carathéodory solution Z of the corresponding autonomous limiting inclusion,

$$\mathrm{dist}\big(Z_0, \mathrm{Min}(\Phi_\lambda)\big) < \delta_\rho$$

implies

$$\mathrm{dist}\big(Z_\tau, \mathrm{Min}(\Phi_\lambda)\big) < \rho/2 \qquad \forall \tau \ge 0.$$

Now set

$$\varepsilon := \min\{\rho/2, \delta_\rho/2\}.$$

For each j, write

$$I_j^\varepsilon = (a_j^\varepsilon, b_j^\varepsilon).$$

Since $\tau_j \in I_j^\varepsilon$ and $\tau_j \to \infty$, it follows that $a_j^\varepsilon \to \infty$. In particular, $a_j^\varepsilon > 0$ for all large j. Hence, by continuity,

$$\mathrm{dist}\big(Y_{a_j^\varepsilon}, \mathrm{Min}(\Phi_\lambda)\big) = \varepsilon.$$

By Theorem 3.10, after extraction of a subsequence, the shifted trajectories

$$Y_{a_j^\varepsilon + \cdot}$$

converge locally uniformly on $[0,\infty)$ to a global Carathéodory solution Z of the corresponding autonomous limiting inclusion. Since

$$\operatorname{dist}\big(Y_{a_j^\varepsilon}, \operatorname{Min}(\Phi_\lambda)\big) = \varepsilon,$$

we obtain

$$\operatorname{dist}\big(Z_0, \operatorname{Min}(\Phi_\lambda)\big) = \varepsilon < \delta_\rho.$$

Hence

$$\operatorname{dist}\big(Z_\tau, \operatorname{Min}(\Phi_\lambda)\big) < \rho/2 \qquad \forall \tau \geq 0.$$

By Lemma 8.3, there exists $C_\varepsilon > 0$ such that

$$|I_j^\varepsilon| = b_j^\varepsilon - a_j^\varepsilon \leq C_\varepsilon \qquad \forall j.$$

Since $Y_{a_j^\varepsilon + \cdot} \to Z$ locally uniformly on $[0,\infty)$, the convergence is uniform on the compact interval $[0, C_\varepsilon]$. Therefore, after possibly enlarging j,

$$\operatorname{dist}\big(Y_{a_j^\varepsilon + \tau}, \operatorname{Min}(\Phi_\lambda)\big) \leq \rho \qquad \forall \tau \in [0, C_\varepsilon].$$

Now, if $\tau \in I_j^\varepsilon$, then

$$a_j^\varepsilon < \tau < b_j^\varepsilon \leq a_j^\varepsilon + C_\varepsilon,$$

hence

$$0 < \tau - a_j^\varepsilon < C_\varepsilon.$$

Therefore

$$\operatorname{dist}\big(Y_\tau, \operatorname{Min}(\Phi_\lambda)\big) \leq \rho \qquad \forall \tau \in I_j^\varepsilon$$

for all sufficiently large j in the extracted subsequence.

Since $\rho < \varepsilon_0$, this is impossible because

$$\tau_j \in I_j^{\varepsilon_0} \subset I_j^\varepsilon$$

and

$$\operatorname{dist}\big(Y_{\tau_j}, \operatorname{Min}(\Phi_\lambda)\big) > \varepsilon_0.$$

The contradiction proves that

$$\operatorname{dist}\big(Y_\tau, \operatorname{Min}(\Phi_\lambda)\big) \to 0 \qquad \text{as } \tau \to \infty.$$

Hence

$$\omega_\tau(Y) \subset \operatorname{Min}(\Phi_\lambda).$$

Since $\omega_\tau(Y)$ is nonempty, compact, and connected, while $\operatorname{Min}(\Phi_\lambda)$ is finite in the empirical setting, it follows that

$$\omega_\tau(Y) = \{x^*\}$$

for some $x^* \in \operatorname{Min}(\Phi_\lambda)$. Therefore

$$Y_\tau \to x^* \qquad \text{as } \tau \to \infty.$$

The equivalence with the convergence of X follows from (3.4). We conclude Theorem 3.18.

Proof of Lemma 8.3. Fix $\varepsilon \in (0, \varepsilon_0]$. Suppose by contradiction that no such constant exists. Then, after extraction of a subsequence, still indexed by j, we may assume that

$$|I_j^\varepsilon| \to \infty \qquad \text{as } j \to \infty.$$

Write

$$I_j^\varepsilon = (a_j^\varepsilon, b_j^\varepsilon),$$

with $0 \le a_j^\varepsilon < b_j^\varepsilon \le +\infty$. Since $|I_j^\varepsilon| \to \infty$, we may choose $s_j \in I_j^\varepsilon$ so that

$$s_j \to \infty, \qquad b_j^\varepsilon - s_j \to \infty.$$

For example, if $b_j^\varepsilon < \infty$, one may take

$$s_j := a_j^\varepsilon + \frac{1}{4}|I_j^\varepsilon|,$$

while if $b_j^\varepsilon = +\infty$, any choice $s_j \in I_j^\varepsilon$ with $s_j \to \infty$ is sufficient.

Then, for every fixed $T > 0$, one has

$$[s_j, s_j + T] \subset I_j^\varepsilon$$

for all sufficiently large j. Hence

$$\operatorname{dist}\big(Y_{s_j+\tau}, \operatorname{Min}(\Phi_\lambda)\big) > \varepsilon \qquad \forall \tau \in [0, T],$$

for all sufficiently large j.

By Theorem 3.10, after extraction of a subsequence, the shifted trajectories

$$Y_{s_j+\cdot}$$

converge locally uniformly on $[0, \infty)$ to a global Carathéodory solution Z of the corresponding autonomous limiting inclusion. Passing to the limit in the previous inequality yields

$$\operatorname{dist}\big(Z_\tau, \operatorname{Min}(\Phi_\lambda)\big) \ge \varepsilon \qquad \forall \tau \ge 0.$$

On the other hand, by Theorem 3.16, there exists a point x^* such that

$$Z_\tau \to x^* \qquad \text{as } \tau \to \infty,$$

where

$$x^* \in \operatorname{Crit}(\Phi_\lambda) \quad \text{if } 0 \le \lambda \le 1, \qquad x^* \in \operatorname{Crit}_{\text{out}}(\Phi_\lambda) \quad \text{if } \lambda > 1.$$

Moreover, for every fixed $\tau \ge 0$, since $s_j + \tau \to \infty$ and

$$Y_{s_j+\tau} \to Z_\tau,$$

we have

$$Z_\tau \in \omega_\tau(Y).$$

Since $\omega_\tau(Y)$ is closed, passing to the limit as $\tau \to \infty$ gives

$$x^* \in \omega_\tau(Y).$$

By the assumption of Theorem 3.18, no non-minimizing critical point of the relevant limiting notion belongs to $\omega_\tau(Y)$. Therefore

$$x^* \in \operatorname{Min}(\Phi_\lambda).$$

This contradicts the fact that

$$\operatorname{dist}\big(Z_\tau, \operatorname{Min}(\Phi_\lambda)\big) \ge \varepsilon \qquad \forall \tau \ge 0,$$

since $Z_\tau \to x^* \in \operatorname{Min}(\Phi_\lambda)$. The contradiction proves the lemma. □

8.3 Proof of Corollary 3.20

We now establish the convergence rate when the limiting point is a local minimizer.

Let $x^* \in \mathrm{Min}(\Phi_\lambda)$ be the limit point given by Theorem 3.18.

In both regimes, the proof reduces to the same local perturbation argument. We first record the local estimates needed below. There exist a neighborhood U of x^* and constants $C_0, \eta > 0$ such that

$$\nabla \Phi_\lambda(x) = 2(x - x^*) \qquad \forall x \in U, \tag{8.1}$$

and

$$\|\nabla F_\lambda(x,t) - \nabla \Phi_\lambda(x)\| \le C_0 e^{-\eta/t} \qquad \forall x \in U,\ \forall t > 0. \tag{8.2}$$

Indeed, in the MoE regime $0 \le \lambda \le 1$, this follows from Lemma 7.2(2)-(3) together with Lemma 6.4. In the CFG regime $\lambda > 1$, the additional assumption $x^* \notin \mathrm{ND}(A_1, A_2)$ implies that Φ_λ is locally a single quadratic branch. Since x^* is a local minimizer, this branch is centered at x^*, giving (8.1); and (8.2) follows from Lemma 6.4, after shrinking U if necessary.

Since $Y_\tau \to x^*$, there exists $\tau_0 \ge 0$ such that

$$Y_\tau \in U \qquad \forall \tau \ge \tau_0.$$

Set

$$E_\tau := Y_\tau - x^*.$$

Using (3.2), we obtain, for every $\tau \ge \tau_0$,

$$\dot{E}_\tau = -\frac{1}{4} \nabla F_\lambda(Y_\tau, Te^{-\tau}).$$

Adding and subtracting $\nabla \Phi_\lambda(Y_\tau)$, and using (8.1), we get

$$\dot{E}_\tau = -\frac{1}{4} \nabla \Phi_\lambda(Y_\tau) - \frac{1}{4}\Big(\nabla F_\lambda(Y_\tau, Te^{-\tau}) - \nabla \Phi_\lambda(Y_\tau)\Big) = -\frac{1}{2} E_\tau + R(\tau),$$

where

$$R(\tau) := -\frac{1}{4}\Big(\nabla F_\lambda(Y_\tau, Te^{-\tau}) - \nabla \Phi_\lambda(Y_\tau)\Big).$$

By (8.2),

$$\|R(\tau)\| \le \frac{C_0}{4}\, e^{-\eta e^\tau / T} \qquad \forall \tau \ge \tau_0.$$

After renaming the constants, we may simply write

$$\|R(\tau)\| \le C_1 e^{-\kappa e^\tau} \qquad \forall \tau \ge \tau_0$$

for some $C_1, \kappa > 0$.

Applying the variation-of-constants formula on $[\tau_0, \tau]$, we obtain

$$E_\tau = e^{-(\tau - \tau_0)/2} E_{\tau_0} + \int_{\tau_0}^{\tau} e^{-(\tau - s)/2} R(s)\, ds.$$

Therefore

$$\|E_\tau\| \le e^{-(\tau-\tau_0)/2} \|E_{\tau_0}\| + \int_{\tau_0}^{\tau} e^{-(\tau-s)/2} \|R(s)\|\, ds \le C e^{-\tau/2}$$

for a suitable constant $C > 0$. Hence

$$\|Y_\tau - x^*\| \le C e^{-\tau/2} \qquad \forall \tau \ge 0.$$

Since $t = Te^{-\tau}$, this is equivalent to

$$\|X_t - x^*\| = \|Y_{\log(T/t)} - x^*\| \le C\sqrt{t}, \qquad t \in (0, T].$$

This proves the result.

9 Conclusions and perspectives

9.1 Conclusions

In this article, we studied the small-time asymptotic behavior of diffusion-model generation dynamics driven by linear score mixing in the heat-flow setting. The central result is a geometric reduction principle: after a natural similarity-time rescaling, the singular non-autonomous score-driven dynamics is asymptotically governed by an autonomous nonsmooth dynamics associated with the distance potential Φ_λ. The Laplace–Varadhan principle is the mechanism that turns heat-flow scores into squared-distance geometry. In the general compact-support setting satisfying the stated lower mass condition, every time-shift limit of the rescaled dynamics satisfies the corresponding limiting inclusion: the genuine Clarke differential inclusion in the MoE regime and the outer Clarke inclusion in the CFG regime. In particular, any convergent generation trajectory must converge to a critical point of the limiting geometric potential.

In the finite-support empirical setting, we obtained a sharper description thanks to the piecewise quadratic structure of the limiting potential and its Voronoi-type geometry. We proved that every global solution of the autonomous limiting inclusion converges to a critical point. For the original non-autonomous generation flow, we proved a conditional convergence criterion: once non-minimizing critical points are excluded from its ω-limit set, the trajectory converges to a local minimizer of the geometric potential. In the smooth stable case, we further established the convergence rate $\mathcal{O}(\sqrt{t})$. The numerical experiments are consistent with this geometric picture and illustrate the role of nonsmooth local minimizers and interface effects, especially in the guided regime.

We also complemented the deterministic geometric analysis with PDE, Hamilton–Jacobi, and stochastic viewpoints. The Li–Yau-type Hessian bounds give a semiconcavity interpretation of the rescaled logarithmic potentials and connect the Laplace–Varadhan limit with a stationary eikonal/Hamilton–Jacobi structure. The L^p-energy estimates for the backward Fokker–Planck equation then reveal a clear polynomial-versus-exponential stability distinction between the MoE and CFG regimes. Finally, the noisy rescaled dynamics suggests a natural connection with stochastic approximation of differential inclusions. Altogether, these results provide a rigorous framework for understanding score mixing and guidance in diffusion models from the viewpoint of support geometry, semiconcavity, and nonsmooth asymptotic dynamics.

9.2 Learning the geometric potential with neural networks

Our analysis suggests a possible application-oriented use of the limiting geometric description. When the score functions are exact, the mixed generation dynamics is asymptotically organized by the geometric potential Φ_λ, and convergent trajectories are driven toward its local minimizers. This raises the following natural perspective: rather than learning or simulating the full time-dependent score field, one may try to learn the reduced potential Φ_λ itself and then generate samples through a gradient-type dynamics on this learned energy landscape.

Of course, this should be understood as a heuristic direction rather than a result of the present paper. In high dimension, the direct computation of $\nabla\Phi_\lambda$ may be expensive or unstable, especially if it is approximated by finite differences. A natural alternative is to approximate the scalar potential by a neural network

$$f_{\mathrm{NN}}(z;\theta) \simeq \Phi_\lambda(z),$$

and then recover the vector field $\nabla_z f_{\mathrm{NN}}(z;\theta)$ by automatic differentiation. This is in the same spirit as several neural-network approaches in scientific computing, including physics-informed neural

networks and operator-learning methods; see, for example, [43, 38, 32, 33, 34].

More concretely, given two datasets

$$A_1 = \{x_i\}_{i=1}^{n_1}, \qquad A_2 = \{y_i\}_{i=1}^{n_2},$$

one could sample points z_j in the ambient space, evaluate $\Phi_\lambda(z_j)$ through the distance formula, and train f_{NN} by minimizing a regression loss such as

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{j=1}^{N}\big(f_{\mathrm{NN}}(z_j;\theta) - \Phi_\lambda(z_j)\big)^2.$$

Once trained, the network could be used to define a reduced generative dynamics of the form

$$dZ_\tau = -\nabla_z f_{\mathrm{NN}}(Z_\tau;\theta^*)\,d\tau + \epsilon(\tau)\,dW_\tau,$$

where $\epsilon(\tau)$ is a prescribed noise schedule. This viewpoint provides a possible bridge between diffusion-based generation and gradient-based sampling on learned geometric energies.

9.3 Other future directions

We conclude by briefly summarizing several directions suggested by the present work.

1. **Generic convergence to local minimizers.** In the empirical setting, our convergence criterion shows that convergence to a local minimizer follows once non-minimizing critical points are excluded from the ω-limit set; see Theorem 3.18. It would be very desirable to prove that this condition is satisfied for almost every initial datum, which would imply generic convergence of trajectories toward local minimizers of the geometric potential. One natural avenue is via a Łojasiewicz-type inequality for piecewise-quadratic functions (cf. [36, 48, 2]), which would reduce the question to a topological argument on the basin of attraction of saddle critical points. This is precisely the behavior observed throughout our numerical simulations, including for nonsmooth minimizers lying on the non-differentiable interface set $\mathrm{ND}(A_1, A_2)$.
2. **Beyond the empirical setting.** The convergence results for the original generation flow rely on the finite Dirac structure, through the piecewise quadratic geometry of the limiting potential and the associated piecewise affine limiting dynamics. For general compactly supported continuous data, we currently obtain only the characterization of time-shift limits; see Remark 3.14. Extending the convergence theory to this setting would require additional assumptions on the geometry of the supports, such as suitable prox-regularity properties [45, Sec. 13. F], and deserves further investigation.
3. **Stochastic generative dynamics.** As discussed in Section 4.3, the similarity-time rescaling reveals the noisy generation process as a vanishing-viscosity perturbation of the limiting geometric dynamics. Developing a rigorous stochastic theory for this regime is a natural extension of the present analysis. Concretely, the open problem is to prove a precise pathwise (or in-probability) convergence theorem of the noisy rescaled dynamics (4.7) to a Carathéodory solution of the limiting Clarke (resp. outer Clarke) inclusion as $\tau \to \infty$. The Benaïm–Hofbauer–Sorin framework of stochastic approximation for differential inclusions [6] appears to provide the natural technical tools for such a result.
4. **Implications for training and sampling schedules.** The similarity-time rescaling also suggests discretizing uniformly in $\tau = \log(T/t)$ rather than in the physical time t, which better resolves the singular regime near $t = 0$. This simple idea could be tested directly on high-dimensional problems and realistic datasets, and may fit naturally with standard architectures used in diffusion models, such as U-Net [46, 22] and, more recently, transformer-based backbones [41].

Acknowledgments

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.

References

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

[2] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1–2):5–16, 2009.

[3] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*. Systems & Control: Foundations & Applications. Birkhäuser Boston, 1990.

[4] Martino Bardi and Italo Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Systems & Control: Foundations & Applications. Birkhäuser Boston, Boston, MA, 1997.

[5] M. Benaïm and M. W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

[6] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

[7] C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Springer, 2013.

[8] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, 2024.

[9] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[10] 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 *International Conference on Learning Representations*, 2023.

[11] M. Chidambaram, K. Gatmiry, S. Chen, H. Lee, and J. Lu. What does guidance do? a fine-grained analysis in a simple setting. In *Advances in Neural Information Processing Systems*, volume 37, pages 84968–85005, 2024.

[12] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley, 1989.

[13] 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.

[14] J. Cortés. Discontinuous dynamical systems: A tutorial on solutions, nonsmooth analysis, and stability. *IEEE Control Systems Magazine*, 28(3):36–73, 2008.

[15] Michael G. Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1):1–67, 1992.

[16] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Applications of Mathematics*. Springer, New York, 2 edition, 1998.

[17] P. Dhariwal and A. Q. Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[18] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

[19] V. A. Galaktionov and J. L. Vázquez. Asymptotic behaviour of nonlinear parabolic equations with critical exponents: A dynamical systems approach. *Journal of Functional Analysis*, 100(2):435–462, 1991.

[20] V. A. Galaktionov and J. L. Vázquez. *A Stability Technique for Evolution Partial Differential Equations: A Dynamical Systems Approach*, volume 56 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, 2004.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[23] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. OpenReview.net, 2021.

[24] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695–709, 2005.

[25] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer, 2 edition, 1991.

[26] T. Karras, M. Aittala, T. Kynkäänniemi, J. Lehtinen, T. Aila, and S. Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[27] D. P. Kingma and Y. LeCun. Regularized estimation of image statistics by score matching. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[28] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[29] B. Klartag and O. Ordentlich. The strong data processing inequality under the heat flow. *IEEE Transactions on Information Theory*, 2025.

[30] H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, volume 35, pages 22870–22882, 2022.

[31] P. Li and S.-T. Yau. On the parabolic kernel of the Schrödinger operator. *Acta Mathematica*, 156:153–201, 1986.

[32] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. M. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

[33] Z. Li, K. Liu, L. Liverani, and E. Zuazua. Universal approximation of dynamical systems by semiautonomous neural ODEs and applications. *SIAM Journal on Numerical Analysis*, 64(1):193–223, 2026.

[34] Z. Li, K. Liu, Y. Song, H. Yue, and E. Zuazua. Deep neural ODE operator networks for PDEs. *Mathematical Models and Methods in Applied Sciences*, 2026.

[35] K. Liu and E. Zuazua. A PDE perspective on generative diffusion models, 2025. Preprint at https://arxiv.org/abs/2511.05940.

[36] S. Łojasiewicz. Ensembles semi-analytiques, 1965. Lecture notes, Institut des Hautes Études Scientifiques.

[37] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pages 5775–5787, 2022.

[38] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[39] L. Markus. Asymptotically autonomous differential systems. In *Contributions to the Theory of Nonlinear Oscillations, Vol. III*, volume 36 of *Annals of Mathematics Studies*, pages 17–29. Princeton University Press, Princeton, NJ, 1956.

[40] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

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

[42] P. Rahimi and S. Marcel. ScoreMix: Improving face recognition via score composition in diffusion generators, 2025. Preprint at https://arxiv.org/abs/2506.10226.

[43] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[44] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[45] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 1998.

[46] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[47] S. Sadat, M. Kansy, O. Hilliges, and R. M. Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *International Conference on Learning Representations*. OpenReview.net, 2025.

[48] L. Simon. Asymptotics for a class of nonlinear evolution equations, with applications to geometric problems. *Annals of Mathematics*, 118(3):525–571, 1983.

[49] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[50] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[51] C. Villani. *Hypocoercivity*. American Mathematical Society, 2009.

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

[53] X. Wang, N. Dufour, N. Andreou, M.-P. Cani, V. Fernández Abrevaya, D. Picard, and V. Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research*, 2024.

[54] E. Zuazua. Asymptotic behavior of scalar convection–diffusion equations, 2020. Preprint at https://arxiv.org/abs/2003.11834.