Probability flow ODEs in score-based and reflected diffusion models

Rama CONT
Mathematical Institute, University of Oxford

Abstract

Probability-flow ordinary differential equations (PF-ODEs) are widely used as deterministic samplers for score-based diffusion models. Their usual justification is that the Fokker–Planck equation of a diffusion can be rewritten as a continuity equation driven by the score function of the forward diffusion. This identity does not, however, guarantee that the resulting velocity field generates a well-posed flow. We provide theoretical insights into the design of such deterministic samplers for generative models based on diffusions and reflected diffusions. We identify sufficient conditions for a regular Lagrangian PF-ODE flow to exist; reverse sampling and invertibility require two-sided divergence control. For learned scores, sampler stability is controlled by an unweighted velocity error, exposing a mismatch with density-weighted score matching and motivating architectural control of Jacobians, divergence, growth, and compression. Under the manifold hypothesis, positive-time regularization justifies an early-stopped PF-ODE while constants deteriorate near the data endpoint; an explicit sphere example shows that the exact deterministic flow becomes singular as the noise level vanishes even though the diffusion marginals remain well defined.

These theoretical insights translate into concrete design principles for stable, invertible, and constraint-preserving diffusion samplers. We illustrate the practical relevance of these design principles using controlled numerical experiments.

1 Probability-flow ODEs in diffusion models

Score-based diffusion models [19] generate data by reversing a noising process and may be sampled either with a reverse-time SDE [3, 5] or with the deterministic probability-flow ODE (PF-ODE) introduced in Song et al. [18]. The PF-ODE is attractive because it removes stochastic simulation, supports deterministic encoding/decoding, and permits likelihood calculations [18, 19]. In constrained generative modeling, reflected diffusion models [12, 15] keep the forward process inside a prescribed domain and can likewise be paired with a deterministic PF-ODE.

The formal derivation of the PF-ODE in these references is based on the observation that, for the diffusion

$$\mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t, \tag{1}$$

the marginal density p_t satisfies the Fokker–Planck equation

$$\partial_t p_t + \nabla\cdot(f p_t) - \tfrac{1}{2}\sigma(t)^2 \Delta p_t = 0. \tag{2}$$

Introducing the score function $s_t = \nabla \log p_t$, (2) can be written as

$$\partial_t p_t + \nabla\cdot(p_t v) = 0 \quad \text{where} \quad v(x,t) = f(x,t) - \tfrac{1}{2}\sigma(t)^2 s_t(x) \tag{3}$$

which is the *continuity equation* for the ordinary differential equation (ODE) $\dot{Z}_t = v(Z_t, t)$ and concludes that Z_t has density p_t. It is then tempting to conclude that the ODE $\dot{Z}_t = v(Z_t, t)$ automatically transports p_0 to p_t. This conclusion is not automatic [7]: the PDE (2) determines an

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: AI for Stochastic Dynamics.

Eulerian density evolution, whereas a deterministic sampler requires a *Lagrangian* flow of individual samples. A unique and smooth density flow can exist even when the associated PF-ODE fails to generate a regular Lagrangian flow from the data time [7]. This distinction is especially relevant under the manifold hypothesis [10] i.e. when data law is supported on a lower-dimensional manifold.

Contributions. We use theoretical insights from [7, 8] to derive implications for the design and training of score-based and reflected diffusion models, and exhibit explicit examples and sampler diagnostics. First, when data is supported on a lower dimensional manifold, positive-time diffusion regularizes the singular data law but the PF-ODE may become singular as $t \downarrow 0$. In Section 3 we give an explicit example of this type, which leads to collapse of the normal directions at the data endpoint. This complements convergence results for diffusion models under the manifold hypothesis [4, 10, 17]. Second, for learned scores, deterministic sampler stability depends not only on score-matching error but also on regularity, divergence, growth, and compression of the induced velocity field. Third, for reflected models, the PF-ODE contains no explicit reflection term: confinement follows from tangency of the probability-flow velocity, rather than from a zero-normal-score condition.

These observations lead to a common design principle: the PF-ODE should be treated as a learned dynamical system with its own well-posedness and conditioning requirements, rather than as an automatic consequence of score estimation. In particular, early stopping, velocity-field regularization, divergence control, low-density validation, and appropriate boundary constraints should be viewed as structural components of deterministic diffusion sampling. We illustrate these principles with numerical experiments (Section 6 and Appendix B.1), with a focus on endpoint conditioning, the DSM–velocity-error gap, and Jacobian regularization.

2 From diffusion marginals to a deterministic sampler

2.1 Marginal matching requires a Lagrangian flow

The identity between (2) and the continuity equation establishes that p_t is an Eulerian solution for the velocity (3). To obtain a sampler one needs a flow map $Z(t,\cdot)$ whose trajectories solve $\dot{Z} = v(Z,t)$ which transports p_0 to p_t.

$$Z(t,\cdot)_{\#}(p_0\,\mathrm{d}x) = p_t\,\mathrm{d}x. \tag{4}$$

For non-smooth velocity fields the relevant object is a regular Lagrangian flow (RLF), which permits Sobolev or bounded-variation (BV) spatial regularity while imposing a quantitative compressibility bound. Under local BV regularity of the drift and score, linear growth, and a one-sided bound on the negative divergence of v, such a unique RLF exists and satisfies (4) [1, 7, 11]. This is substantially weaker than assuming a globally Lipschitz score.

The distinction matters for model validation. A well-posed SDE and a unique Fokker–Planck density do not certify the PF-ODE. The diffusion benefits from stochastic regularization, whereas the deterministic dynamics directly inherit score singularities. For a Brownian forward process initialized by the uniform law on an interval, for example, p_t is smooth and strictly positive for every $t > 0$, yet the score outside the initial support behaves as $O(1/t)$ as $t \downarrow 0$; the PF-ODE has no regular Lagrangian flow from time zero [7]. Thus "the score is smooth for every positive time" is weaker than the time-integrated regularity needed by the sampler.

Design implication: early stopping is structural. Under the manifold hypothesis linear VP/VE forward diffusions regularize the law at every $\delta > 0$. On $[\delta, T]$ the score has explicit Hessian, divergence, and linear-growth bounds sufficient for PF-ODE well-posedness, with constants that deteriorate as $\delta \downarrow 0$ [7]. The principled deterministic pipeline is therefore

$$\mu_0 \xrightarrow{\text{forward diffusion}} p_\delta \xrightarrow{\text{PF-ODE on } [\delta,T]} p_T, \qquad p_T \xrightarrow{\text{reverse PF-ODE}} p_\delta, \tag{5}$$

and the final passage $p_\delta \to \mu_0$ requires a limiting argument or a separate terminal denoising step. Reducing δ improves endpoint fidelity but worsens conditioning. This gives an analytic interpretation of the empirical difficulty of the final denoising steps and suggests treating the truncation time as a model-design parameter, not merely an ODE-solver tolerance.

2.2 Reverse sampling is stronger than forward marginal matching

The RLF criterion for forward transport uses a one-sided compression bound, schematically

$$[\nabla \cdot v]^- \in L_t^1 L_x^\infty. \tag{6}$$

Reversing time sends v to $-v$ and exchanges the positive and negative parts of the divergence. Consequently, a reverse deterministic sampler additionally needs

$$[\nabla \cdot v]^+ \in L_t^1 L_x^\infty. \tag{7}$$

Under both bounds the flow is essentially invertible, the reverse RLF transports p_T to p_0 (or to p_δ in the early-stopped setting), and encode–decode inversion and change-of-variables likelihoods become justified [7]. This asymmetry matters for diffusion-model objectives. Forward marginal matching, reverse deterministic generation, and exact likelihood evaluation are not equivalent specifications. A model intended only for one-way generation may tolerate weaker divergence control than a model intended as an invertible continuous normalizing flow. Architectural and regularization choices should therefore be aligned with the required downstream operation.

3 Singularity at the data endpoint under the manifold hypothesis

By *data endpoint* we mean the zero-noise time $t = 0$, not the boundary of a spatial domain. The manifold hypothesis posits that high-dimensional data concentrate on, or near, a set whose intrinsic dimension is much smaller than the ambient dimension. We focus here on the idealized but important limiting case in which the data distribution is *exactly* supported on a lower-dimensional manifold and is therefore singular with respect to ambient Lebesgue measure. The conclusions below need not hold in the same form for distributions that merely concentrate near a manifold while retaining a positive ambient density.

Diffusion models can nevertheless be well behaved in this singular setting. Convergence to targets supported on lower-dimensional manifolds or empirical measures has been established without assuming an ambient target density [10]; more recent results show that score-learning and sampling rates can depend primarily on intrinsic rather than ambient dimension [4, 17]. The question addressed here is different: does such favorable distributional behavior imply that the deterministic probability-flow ODE remains a regular and invertible ambient-space flow all the way to $t = 0$? The following example shows that it does not.

The construction is a higher-dimensional version of the time-zero vacuum mechanism in [7]: diffusion immediately fills directions in which the initial law has no ambient density, while the corresponding deterministic probability-flow velocity becomes singular as this regularization is removed.

Proposition 1 (Sphere data: singularity at the data endpoint)**.** *Let μ_0 be the uniform distribution on*

$$\mathcal{M} = S_R^m \times \{0\} \subset \mathbb{R}^{m+1} \times \mathbb{R}^{d-m-1} = \mathbb{R}^d, \qquad m < d,$$

and consider the variance-exploding forward diffusion

$$X_t = X_0 + W_t, \qquad X_0 \sim \mu_0.$$

For every $t > 0$, X_t has a smooth strictly positive density p_t.

If $d \geq m + 2$, write $x = (u, z) \in \mathbb{R}^{m+1} \times \mathbb{R}^{d-m-1}$. Then the velocity $v = -\frac{1}{2}\nabla \log p_t$ satisfies

$$v_z(u, z, t) = \frac{z}{2t}. \tag{8}$$

Consequently, no absolutely continuous PF-ODE trajectory on $[0, T]$ can start from a point with $z_0 \neq 0$. Since $\{z \neq 0\}$ has full ambient Lebesgue measure, there is no regular Lagrangian flow from $t = 0$ with respect to ambient Lebesgue measure.

In reverse time, however, a trajectory initialized at $T > 0$ satisfies

$$z_t = \sqrt{\frac{t}{T}}\, z_T, \tag{9}$$

and hence converges continuously to the lower-dimensional subspace as $t \downarrow 0$. Thus the obstruction at the data endpoint concerns ambient-flow regularity, compression, and invertibility; it does not imply nonexistence of individual reverse trajectories.

Derivation. Because the last $d-m-1$ coordinates of X_0 vanish and Brownian noise is isotropic,

$$p_t(u,z) = q_t(u)(2\pi t)^{-(d-m-1)/2} \exp\left(-\frac{|z|^2}{2t}\right), \tag{10}$$

where q_t is the Gaussian smoothing of the uniform law on $S_R^m \subset \mathbb{R}^{m+1}$. Therefore

$$\nabla_z \log p_t(u,z) = -\frac{z}{t},$$

which proves (8). Every solution of $\dot{z} = z/(2t)$ on $(0,T]$ has the form $z_t = C\sqrt{t}$. Continuity at $t=0$ therefore forces $z_0 = 0$, so an ambient trajectory cannot be launched at time zero from Lebesgue-a.e. point. Conversely, fixing z_T yields (9). In the $c = d-m-1$ displayed normal directions, the reverse map contracts volume by

$$\left(\frac{t}{T}\right)^{c/2},$$

which tends to zero at the data endpoint.

The conclusion concerns an *ambient* regular Lagrangian flow, whose reference measure is d-dimensional Lebesgue measure. Since μ_0 itself is concentrated on a Lebesgue-null set, Proposition 1 does not rule out a deterministic transport defined only μ_0-a.s. on $\mathcal{M}$. Rather, it shows that such a transport cannot remain a uniformly regular, invertible ambient-space flow as the noise level vanishes.

Codimension one. The case $d = m+1$, where there are no additional flat normal coordinates, exhibits the same mechanism in the radial normal direction. Writing $r = |x|$ and $\nu = (m-1)/2$, the radial score of the Gaussian smoothing of the uniform law on S_R^m can be written as

$$\partial_r \log p_t(r) = -\frac{r}{t} + \frac{R}{t}\frac{I_{\nu+1}(Rr/t)}{I_\nu(Rr/t)}, \tag{11}$$

where I_ν denotes the modified Bessel function. Since

$$\frac{I_{\nu+1}(a)}{I_\nu(a)} \longrightarrow 1, \qquad a \to \infty,$$

we obtain, uniformly on compact subsets of $\{x \in \mathbb{R}^{m+1} : 0 < |x| \neq R\}$,

$$\partial_r \log p_t(r) = -\frac{r-R}{t} + O(1), \qquad t \downarrow 0. \tag{12}$$

Thus, at every fixed ambient point separated from the sphere and from the origin, the normal component of the probability-flow velocity has a $1/t$ blow-up. If an absolutely continuous trajectory were continuous at $t=0$ with $r_0 \neq R$, then for sufficiently small t its distance from the sphere would remain bounded away from zero, making the time integral of the leading velocity term divergent. Hence such a trajectory cannot start at that point. This again excludes an ambient regular Lagrangian flow from the data endpoint.

Notice that this statement does not conflict with reverse-time collapse onto the sphere. Along a reverse trajectory the normal distance itself decreases with t, so the velocity experienced along that trajectory need not scale as $1/t$. The $1/t$ statement concerns the velocity away from the data manifold.

Relation to convergence results under the manifold hypothesis. De Bortoli [10] showed that denoising diffusion models can converge to targets supported on lower-dimensional manifolds or empirical measures and obtained quantitative W_1 guarantees without assuming an ambient target density. Azangulov et al. [4] obtain score-learning and reverse-SDE sampling rates largely independent of ambient dimension, while Potaptchik et al. [17] showed that the number of reverse-SDE discretization steps required for KL convergence can scale linearly, up to logarithmic factors, with intrinsic dimension. These are distributional convergence results for stochastic diffusion samplers.

Proposition 1 addresses a different question: whether the associated deterministic probability flow remains a regular and invertible ambient-space flow as the noise level vanishes. The sphere example shows that, even with the *exact* score, approaching a lower-dimensional target forces collapse of normal directions. Thus favorable intrinsic-dimensional sampling rates do not imply Lagrangian regularity, bounded compression or invertibility of the PF-ODE at $t=0$.

Design implication. Under the manifold hypothesis, when the target distribution is idealized as being exactly supported on a lower-dimensional manifold, the zero-noise limit should be viewed as a *singular collapse limit*, rather than as an ordinary well-conditioned ODE time. If stable deterministic transport, invertible encoding/decoding, or change-of-variables likelihood evaluation is required, a natural strategy is to stop the PF-ODE at some $\delta > 0$ and treat the passage from p_δ to the data distribution separately, for example through a terminal denoising or manifold-valued correction step.

If reverse trajectories are integrated closer to zero, one should expect normal Jacobian factors and ambient volume elements to collapse, while the conditioning constants of the deterministic flow deteriorate even for a perfect score. This is an intrinsic geometric consequence of removing diffusion regularization, not merely an effect of score-estimation error.

4 Learned scores: training error versus sampler stability

Let s_θ be a learned approximation of $s = \nabla \log p$ and

$$v_\theta = f - \tfrac{1}{2}\sigma^2 s_\theta. \tag{13}$$

On an early-stopped interval $[\delta, T]$, the RLF stability theory gives trajectory convergence under local Sobolev regularity, a uniform compression bound, and convergence of the velocity in an unweighted L^1 norm. Writing

$$E_v := \|v - v_\theta\|_{L^1([\delta,T]\times B_R)}, \tag{14}$$

the measure of initial points whose exact and learned trajectories differ by more than η admits a logarithmic modulus in E_v for Sobolev velocities [7, 9]. In the globally Lipschitz regime, a standard Grönwall argument yields a linear dependence on a uniform score error, but with an amplification factor that can grow rapidly as $\delta \downarrow 0$. The important point is that E_v is not the standard denoising score-matching objective

$$E_\theta^2 := \int_\delta^T \int p_t(x)\,|s_\theta(x,t) - s_t(x)|^2 \,\mathrm{d}x\,\mathrm{d}t. \tag{15}$$

On a ball B_R, Cauchy–Schwarz gives the conversion

$$E_v \le \frac{1}{2}\|\sigma\|_\infty^2 E_\theta \sqrt{\kappa_{\delta,R}}, \qquad \kappa_{\delta,R} := \int_\delta^T \int_{B_R} \frac{1}{p_t(x)}\,\mathrm{d}x\,\mathrm{d}t. \tag{16}$$

Thus low-density regions receive little weight in (15) but are amplified in the conversion to the velocity norm controlling deterministic trajectories. For standard linear VP/VE diffusions, $\kappa_{\delta,R}$ is finite for every $\delta > 0$ and finite R, but can grow sharply as $\delta \downarrow 0$ or R increases [7]. This is a concrete reason why a small score loss need not imply a stable PF-ODE sampler.

Design implication: architecture should enforce regularity. The learned flow must control more than pointwise error. Sufficient conditions include local $W^{1,p}$ regularity ($p > 1$), a one-sided divergence bound for v_θ, and linear growth. These quantities are not controlled by density-weighted score matching alone. Practical mechanisms suggested by the theory include spectral normalization, Jacobian penalties, explicit divergence regularization, and uniform control of network growth. For a sequence of increasingly accurate score models, compression and growth constants must remain uniformly bounded; otherwise $E_\theta \to 0$ does not imply convergence of the samplers. This implication complements statistical score-estimation guarantees [13] and convergence analyses that assume smooth PF-ODE dynamics [14] and also connects to schedule design [6, 16]. A schedule optimized only for score-estimation difficulty or terminal SNR may be poorly conditioned for deterministic transport near the data endpoint.

5 Reflected diffusion models: deterministic confinement without reflection

Constrained diffusions [12, 15] replace (1) by a normally reflected diffusion in a bounded domain:

$$\mathrm{d}X_t = f(X_t,t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t - \nu(X_t)\,\mathrm{d}L_t, \qquad X_t \in \overline{\Omega} \subset \mathbb{R}^d, \tag{17}$$

where ν is the outward unit normal and L_t is boundary local time. Reflected diffusion models use such forward dynamics to prevent mass from leaving the feasible set [12, 15]. Their marginal density satisfies a Fokker–Planck equation with co-normal no-flux condition

$$J \cdot \nu = 0, \qquad J := fp - \tfrac{1}{2}\sigma^2 \nabla p. \tag{18}$$

Define again $v = J/p = f - \frac{1}{2}\sigma^2 \nabla \log p$. If $p > 0$ is differentiable up to the boundary (18) implies

$$v \cdot \nu = 0 \quad \text{on } \partial\Omega. \tag{19}$$

The local-time term has disappeared: its effect on the marginal flow is encoded by the boundary behavior of the score. Under positive-time regularization and drift–score regularity, the ODE

$$\dot{Z}_t = f(Z_t, t) - \tfrac{1}{2}\sigma(t)^2 \nabla \log p_t(Z_t) \tag{20}$$

generates a confined regular Lagrangian flow and transports the same marginals as the reflected diffusion [8]. In particular, for the reflected VE ($f = 0$) and reflected VP models ($f(x,t) = -\beta(t)x$, $\sigma^2 = 2\beta$), this conclusion holds on every $[\delta, T]$ for arbitrary initial laws in a smooth bounded domain.

5.1 The boundary condition acts on the full velocity

Equation (19) is the correct design target. It does *not* generally mean that the normal score is zero. Dividing the no-flux condition by $p > 0$ gives

$$\partial_\nu \log p = \frac{2f \cdot \nu}{\sigma^2} \quad \text{on } \partial\Omega. \tag{21}$$

For reflected VE dynamics with $f = 0$, this reduces to the Neumann condition. For reflected VP dynamics, however, $f \cdot \nu$ is generally nonzero, so imposing $\partial_\nu \log p = 0$ would be inconsistent with the forward no-flux PDE. Boundary-aware score parameterizations should enforce or approximate the co-normal relation (21), or penalize the normal component of the *learned velocity*

$$v_\theta \cdot \nu = \big(f - \tfrac{1}{2}\sigma^2 s_\theta\big) \cdot \nu, \tag{22}$$

rather than zeroing the score normal component independently of the drift. This has a useful computational consequence: in the regular regime, one does not need to solve a reflected ODE or add projection/reflection events to the PF-ODE. Confinement follows from tangency of the velocity and the regular Lagrangian-flow structure. This separates the numerical treatment of the reverse SDE, which retains reflection, from the deterministic PF-ODE sampler, which does not.

5.2 Endpoint and boundary failure modes

Two failure mechanisms clarify what a constrained sampler must control.

Data-endpoint singularity. Reflected Brownian motion started from a compactly supported density inside an interval has a unique smooth positive density for every $t > 0$ and satisfies tangency at the walls, yet the score in the initial vacuum still behaves like $1/t$ [8]. The PF-ODE has no regular Lagrangian flow from time zero, while it is well posed on every $[\delta, T]$ [8]. Thus reflection does not cure the data-endpoint pathology; early stopping remains necessary for compactly supported data.

Boundary-current compression. For more general no-flux density–flux pairs, if the density degenerates at the boundary then $p\, v \cdot \nu = 0$ no longer implies $v \cdot \nu = 0$. A boundary current can coexist with exact marginal transport and confined integral curves while producing unbounded compression and no regular Lagrangian flow [8]. Uniformly elliptic reflected VE/VP models after positive-time regularization avoid this mechanism because their densities are positive up to the boundary, but the example is a warning for learned or generalized constrained flows: preserving mass and keeping sample paths inside the domain are not sufficient for the stability of the deterministic sampler.

Geometry also enters through score regularity. Smooth convex domains are favorable for Neumann/co-normal regularity, whereas nonconvex or nonsmooth boundaries can degrade second-derivative estimates near the wall [8]. Domain choice and boundary representation are therefore part of the conditioning of a reflected score model, not merely semantic constraints on the output space.

6 Implications: Design principles for diffusion samplers

These results suggest treating the PF-ODE as a learned dynamical system, not just a by-product of score estimation. The central object is the learned probability-flow velocity, not the score in isolation.

6.1 Regularize the quantity that is actually integrated

The training objective should distinguish score accuracy from regularity of the velocity. One can use

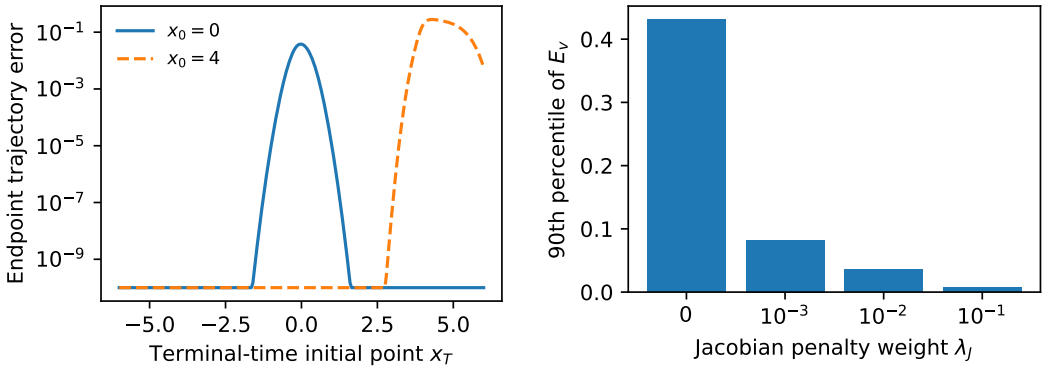
$$\mathcal{L}_{\mathrm{PF}} = \mathcal{L}_{\mathrm{DSM}} + \lambda_J \mathcal{R}_J(v_\theta) + \lambda_{\mathrm{div}} \mathcal{R}_{\mathrm{div}}(v_\theta) + \lambda_g \mathcal{R}_{\mathrm{growth}}(v_\theta) + \lambda_\partial \mathcal{R}_\partial(v_\theta), \tag{23}$$

where the last term is used only for constrained models. The role of each term follows directly from the flow conditions. $\mathcal{R}_J$ controls a Sobolev/Jacobian norm, $\mathcal{R}_{\text{div}}$ controls the divergence component that drives compression, and $\mathcal{R}_{\text{growth}}$ prevents rapid escape to spatial regions in which neither the score nor its error is controlled. For generation in one time direction, the divergence penalty may be one-sided; for invertible generation/encoding it should control both signs. Spectral normalization and Jacobian-trace penalties are concrete mechanisms for implementing this [7]. For reflected models, the natural boundary regularizer is not $\|s_\theta \cdot \nu\|^2$ in general, but

$$\mathcal{R}_\partial(v_\theta) = \mathbb{E}_{(x,t)\sim q_\partial} \left| \left(f(x,t) - \tfrac{1}{2}\sigma(t)^2 s_\theta(x,t) \right) \cdot \nu(x) \right|^2, \tag{24}$$

for a suitable distribution q_∂ on boundary/time pairs. This enforces the co-normal condition and reduces to a zero-normal-score condition only when $f \cdot \nu = 0$. It is particularly relevant when a score network is trained in ambient coordinates and has no built-in knowledge of the domain geometry.

Numerical illustration To test whether such regularization can suppress an instability that is weakly identified by score matching, we consider the one-dimensional VE model $p_t = N(0, 1+t)$ and a learned score $s_a = s + ah$, where $h(x) = \exp(-(x-4)^2/(2w^2))$ is localized in a low-density region. With $N = 1024$ noisy score observations, the coefficient a is poorly identified by the density-weighted loss. Over 5000 repetitions, adding a spatially uniform Jacobian penalty reduces the 90th percentile of the unweighted velocity error E_v from 0.431 at $\lambda_J = 0$ to $0.0825, 0.0351,$ and 0.00678 for $\lambda_J = 10^{-3}, 10^{-2}, 10^{-1}$, respectively (see Figure 1b). This example illustrates that velocity regularization can suppress modes that receive little weight under the training distribution and substantially reduce the stability-relevant velocity error. Details are given in Appendix B.1.



(a) Matched-DSM trajectory errors

(b) Tail-mode regularization

Figure 1: Low-density diagnostics and a controlled regularization test. (a) Endpoint trajectory errors for the score perturbations with $E_\theta^2 = 10^{-3}$; (b) 90th percentile of unweighted velocity error over 5000 noisy-score regressions decreases with Jacobian penalty (matched DSM error not imposed).

6.2 Use density-stratified validation

The inverse-density factor in (16) suggests that global denoising-score-matching is a poor diagnostic for ODE sampling. A more informative report stratifies validation points into bins according to $\log p_t(x)$ or another proxy for density and evaluates score and velocity/Jacobian error in each bin. The same principle applies at the boundary of a constrained domain: evaluate tangency and compression as a function of distance to $\partial\Omega$. Such diagnostics target regions that density weighting tends to hide.

Numerical illustration In the same Gaussian model, we perturb the score by a localized bump of width $w = 0.2$, centered either at $x_0 = 0$ or at the low-density location $x_0 = 4$, and choose the amplitudes so that both perturbations have the same integrated DSM error $E_\theta^2 = 10^{-3}$. Despite identical DSM loss, the tail perturbation produces a $12.3\times$ larger unweighted velocity error, a $152\times$ larger excess Jacobian penalty, and a $15.9\times$ larger mean endpoint trajectory error on a uniform validation grid (Figure 1a). Thus a global density-weighted score metric can make two models appear equivalent while density-stratified or trajectory-level diagnostics distinguish their PF-ODE behavior.

Trajectory-level tests are also important. If Z_θ is used for deterministic decoding or likelihood evaluation, one should test forward–reverse round trips, sensitivity to score perturbations, and volume

Table 1: Theory-to-design translation for deterministic diffusion samplers.

Requirement/ failure mode	Analytic diagnostic	Design consequence
PF-ODE existence	Sobolev/BV regularity, growth, $[\nabla \cdot v_\theta]^-$ control	Regularize Jacobian/divergence and network growth; do not rely on score loss alone.
Reverse/invertible flow	Both $[\nabla \cdot v_\theta]^-$ and $[\nabla \cdot v_\theta]^+$	Stronger regularization for deterministic decoding, round trips, and likelihood evaluation.
Under the manifold hypothesis	Normal score/compression constants diverge as $t \downarrow 0$	Train/sample on $[\delta, T]$ and use a terminal denoising or limiting step; tune δ jointly with schedule.
Low-density regions	E_v depends on an unweighted norm; conversion from score loss uses $\int 1/p_t$	Add localization or low-density diagnostics/regularization; validate trajectories, not only DSM loss.
Reflected boundary	$v_\theta \cdot \nu \approx 0$; generally $s_\theta \cdot \nu \neq 0$	Enforce co-normal relation for the drift–score field; avoid generic zero-normal-score constraints.
Boundary degeneracy	Loss of positivity can permit large compression near $\partial\Omega$	Maintain positive-time regularization; monitor boundary density, normal trace, and compression.

distortion along trajectories. These are aligned with properties controlled by regular Lagrangian-flow theory and can distinguish a low-loss score estimator from a stable ODE vector field.

6.3 Choose schedules together with the endpoint cutoff

For compactly supported data, early-stopping estimates depend on accumulated noise variance ς_t^2. In the VE case $\varsigma_t^2 \asymp t$, representative score-Hessian bounds scale like $\varsigma_t^{-2} + R^2\varsigma_t^{-4}$, so integrated conditioning deteriorates rapidly as the sampler is extended toward $t = 0$. The relevant design variable is therefore the pair *(noise schedule, endpoint cutoff)* rather than either one alone. A schedule that injects enough noise to regularize the score quickly can make a given positive cutoff better conditioned, while an aggressive reduction of δ can erase this benefit. This perspective is complementary to recent schedule-design work based on Lipschitz conditioning [6].

Numerical illustration (Appendix B.1) The cutoff tradeoff is explicit for the VE model $X_0 = (Y, 0)$, $Y \sim N(0, 1)$, in $\mathbb{R}^{1+c}$. In the c normal directions the reverse PF-ODE from $T = 1$ to δ is $z_\delta = \sqrt{\delta}\, z_T$. Hence the residual normal scale decreases as $\sqrt{\delta}$, while inverse sensitivity grows as $\delta^{-1/2}$ and normal volume contracts by $\delta^{c/2}$. For example, at $\delta = 10^{-2}$ the residual normal scale is 0.1, but inverse sensitivity is already 10; for codimension $c = 8$, the normal volume factor is 10^{-8}. Thus improved endpoint fidelity is accompanied by deteriorating invertibility even at positive cutoff.

Here the normal contribution to the Wasserstein-2 distance from the singular target is exactly $\sqrt{c\delta}$; at $\delta = 10^{-2}$ this equals 0.283 for c=8, while the inverse normal sensitivity is 10. Thus finite-cutoff distributional error can be small while the ambient deterministic map is already strongly compressed. This is arguably one of the central messages: distributional approximation can improve toward the manifold while ambient deterministic invertibility deteriorates.

For reflected models the same principle interacts with geometry. Parabolic regularization gives a smooth positive density up to a smooth boundary, but constants depend on boundary regularity and can deteriorate at corners or nonconvex singularities. If the constraint set is a modeling choice, a smooth geometric representation can therefore improve score regularity and deterministic sampling. These design principles, summarized in Table 1, suggest that validation protocols for ODE-based diffusion models should, in addition to score loss and sample-quality metrics, evaluate: (i) the time profile of Jacobian/divergence norms of v_θ; (ii) empirical volume compression along trajectories; (iii) sensitivity of the sampler to decreasing δ; (iv) score/velocity errors stratified by density level; and, for constrained models, (v) the boundary residual $|v_\theta \cdot \nu|$. These diagnostics target quantities that affect the flow and can reveal failures invisible to density-weighted training loss.

The results also separate two modeling choices that are often conflated. The reverse SDE is a stochastic process whose score error is naturally averaged under its evolving law, while the PF-ODE is a deterministic vector field whose global trajectories can be affected by low-density regions. A score architecture and schedule optimized for SDE sampling need not be optimal for ODE sampling. If deterministic decoding, fast ODE generation, or exact likelihoods are central goals, PF-ODE conditioning should be included explicitly in training and schedule selection.

References

[1] Luigi Ambrosio. Transport equation and cauchy problem for BV vector fields. *Inventiones Mathematicae*, 158:227–260, 2004.

[2] Luigi Ambrosio. Transport equation and cauchy problem for non-smooth vector fields. In *Calculus of Variations and Non-Linear Partial Differential Equations*, volume 1927 of *Lecture Notes in Mathematics*, pages 1–41. Springer, 2008.

[3] Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

[4] Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2025. URL `https://arxiv.org/abs/2409.18804`.

[5] Patrick Cattiaux. Time reversal of diffusion processes with a boundary condition. *Stochastic processes and their Applications*, 28(2):275–292, 1988.

[6] Sitan Chen, Eric Vanden-Eijnden, and Yunan Xu. Lipschitz-guided design of interpolation schedules in generative models. *arXiv preprint arXiv:2509.01629*, 2025.

[7] Rama Cont. Mimicking diffusion processes with differential equations. *arXiv preprint arXiv:2607.25685*, 2026.

[8] Rama Cont. Reflected diffusions, no-flux continuity equations and confined lagrangian flows in bounded domains. *arXiv preprint arXiv:2607.28344*, 2026.

[9] Gianluca Crippa and Camillo De Lellis. Estimates and regularity results for the DiPerna–Lions flow. *Journal für die reine und angewandte Mathematik*, 616:15–46, 2008.

[10] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. URL `https://arxiv.org/abs/2208.05314`.

[11] Ronald J. DiPerna and Pierre-Louis Lions. Ordinary differential equations, transport theory and Sobolev spaces. *Inventiones Mathematicae*, 98:511–547, 1989.

[12] Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael J. Hutchinson. Diffusion models for constrained domains. *Transactions on Machine Learning Research*, 2023.

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

[14] Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ODE for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

[15] Aaron Lou and Stefano Ermon. Reflected diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[16] Connor Mooney, Zhongjian Wang, Jack Xin, and Yifeng Yu. Global well-posedness and convergence analysis of score-based generative models via sharp lipschitz estimates. In *International Conference on Learning Representations*, 2025.

[17] Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4668–4685. PMLR, 2025. URL `https://arxiv.org/abs/2410.09046`.

[18] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[19] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

A Some mathematical results on ODE samplers for diffusion models

This appendix summarizes results from [7, 8]. For mathematical definitions we refer to Ambrosio [2].

A.1 PF-ODE in the whole space

Consider (1) with scalar uniformly nondegenerate $\sigma(t) \geq \sigma_{\min} > 0$, and let p_t be its marginal density. Suppose the drift is locally bounded variation (BV) in space with integrable growth, its distributional divergence is absolutely continuous with an integrable one-sided L^∞ bound, and the score $\nabla \log p_t$ has corresponding Sobolev/BV regularity, linear growth, and an integrable upper bound on $\Delta \log p_t$. Then the velocity (3) generates a unique regular Lagrangian flow and transports p_0 to p_t. If the opposite sign of $\nabla \cdot v$ is also integrably bounded in L^∞, the time-reversed regular Lagrangian flow exists and is the a.e. inverse of the forward flow.

For linear VP/VE diffusions with compactly supported initial law, no density or score at time zero is required: for each $\delta > 0$, Gaussian regularization yields the required score bounds on $[\delta, T]$. In the VE scaling $\varsigma_t^2 \asymp t$, worst-case bounds grow polynomially in δ^{-1}, which explains the deterioration toward the data endpoint.

A.2 Learned-score stability

Let Z^δ and Z_θ^δ be the exact and learned regular Lagrangian flows, both initialized as the identity at time δ. Under local $W^{1,p}$ regularity ($p > 1$), growth, and compression bounds, for x in a ball B_r one obtains an estimate of the form

$$\mathcal{L}^d\big(\{x \in B_r : \big|Z^\delta(T,x) - Z_\theta^\delta(T,x)\big| > \eta\}\big) \leq \frac{C_p + L_\theta E_v/\lambda}{\log(1+\eta/\lambda)}, \tag{25}$$

where E_v is defined in (14), L_θ is the compression constant of the learned flow, and C_p depends on the Sobolev norm of the exact velocity on a confinement ball. Taking $\lambda = \sqrt{E_v}$ yields convergence in measure as $E_v \to 0$ with fixed regularity/compression constants. Equation (16) relates E_v to the denoising score-matching loss.

A.3 PF-ODE in a bounded domain

Let $\Omega \subset \mathbb{R}^d$ be a bounded smooth domain and consider the normally reflected diffusion (17). Its probability flux is $J = fp - \frac{1}{2}\sigma^2 \nabla p$. If, on $[\delta, T]$, p is positive up to the boundary and the velocity $v = J/p$ has the interior/collar BV regularity [9] and one-sided divergence bounds required for a regular Lagrangian flow [2], then the no-flux condition gives a vanishing normal trace of v. Extending v by zero outside Ω introduces no singular boundary divergence term, and the resulting RLF stays in $\overline{\Omega}$ and transports p_δ to p_t.

For $C^{2,\alpha}$ domains and sufficiently regular drift, parabolic boundary regularity verifies these conditions for every $\delta > 0$ and arbitrary initial law. The reflected VE and VP diffusions used in constrained score-based modeling satisfy the resulting early-stopped transport statement. The constants degenerate as $\delta \downarrow 0$.

B Practical checklist for a PF-ODE implementation

For an implementation intended to use deterministic PF-ODE sampling, the theory motivates the following checks in addition to ordinary score-model evaluation:

1. report the data-endpoint truncation δ and the terminal denoising procedure;
2. measure Jacobian/divergence norms of v_θ over time and across density quantiles;

3. test sensitivity of sample trajectories and likelihood estimates to the ODE solver and to perturbations of s_θ;
4. if invertibility is used, test both forward and reverse compression rather than forward generation alone;
5. for reflected models, evaluate $v_\theta \cdot \nu$ on boundary samples and use the drift-dependent co-normal target;
6. stratify score error by density level so that tail or boundary errors are not hidden by density weighting.

B.1 Numerical experiments

We report three numerical experiments that isolate the mechanisms behind the design principles. Exact reference scores allow us to distinguish score approximation, flow conditioning, and numerical integration error.

Jacobian regularization of a poorly identified tail mode. We test a simplified instance of the regularization principle in (23). Retain $h = h_4$ and the model $s_a = s + ah$ above, whose true coefficient is $a = 0$. For each of 5000 independent training sets, draw $N = 1024$ pairs with t_i uniform on $[0.05, 1]$ and $x_i \sim p_{t_i}$, and use targets $s(x_i, t_i)+\varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 0.2^2)$ independently. Penalizing the excess velocity Jacobian under a spatially uniform reference distribution gives

$$\mathcal{R}_{J,\mathrm{err}}(a) = K_J a^2, \qquad K_J = \frac{1}{2R}\int_{-R}^{R} \left|\tfrac{1}{2}h'(x)\right|^2 \,\mathrm{d}x, \qquad R = 6. \tag{26}$$

Minimizing the empirical squared regression loss plus $\lambda_J \mathcal{R}_{J,\mathrm{err}}(a)$ gives

$$\widehat{a}_{\lambda_J} = \frac{N^{-1}\sum_{i=1}^{N} h(x_i)\varepsilon_i}{N^{-1}\sum_{i=1}^{N} h(x_i)^2 + \lambda_J K_J}. \tag{27}$$

The same training sets are used for all penalty weights. Since the tail basis is rarely sampled, its coefficient can be poorly identified without regularization. Table 2 and Figure 1(b) show that the penalty suppresses this residual mode: the 90th percentile of E_v decreases from 0.4308 without regularization to 0.03510 at $\lambda_J = 10^{-2}$ and 0.006779 at $\lambda_J = 10^{-1}$. The population DSM error also decreases.

Table 2: Tail-mode regularization across 5000 independent training sets. Each column reports a separate empirical 90th percentile. The penalty is spatially normalized as in (26), whereas E_v and E_θ^2 retain the integrated definitions in (14) and (15).

λ_J	$\lvert\widehat{a}_{\lambda_J}\rvert$	E_θ^2	E_v
0	1.8091	2.361×10^{-3}	0.4308
10^{-3}	0.34648	8.659×10^{-5}	0.08251
10^{-2}	0.14741	1.567×10^{-5}	0.03510
10^{-1}	0.02847	5.846×10^{-7}	0.006779

Equal DSM error, different velocity and trajectory errors. To isolate (16), consider $p_t = \mathcal{N}(0, 1+t)$ on $[\delta, T] = [0.05, 1]$, with $s(x,t) = -x/(1+t)$ and $v = -s/2$. Define

$$s_\theta(x,t) = s(x,t) + Ah_{x_0}(x), \qquad h_{x_0}(x) = \exp\left(-\frac{(x-x_0)^2}{2w^2}\right), \qquad w = 0.2. \tag{28}$$

We compare $x_0 = 0$ and $x_0 = 4$, choosing A by quadrature so that both perturbations have $E_\theta^2 = 10^{-3}$ in (15). Here "DSM error" denotes that population score error, not a noisy empirical denoising loss. Besides E_v on $B_R = [-6, 6]$, we evaluate the excess Jacobian error $J_{\mathrm{err}} := \int_\delta^T \int_{-6}^{6} |\partial_x(v_\theta - v)|^2 \,\mathrm{d}x\,\mathrm{d}t$. Reverse trajectories are initialized at 241 equally spaced points $x_T \in [-6, 6]$ and compared with the exact endpoint $x_T\sqrt{(1+\delta)/(1+T)}$.

Table 3: Matched population DSM error for central and tail score perturbations. Trajectory errors are evaluated at $t = \delta$; the mean is an arithmetic mean over the uniform spatial grid, not an expectation under p_T. The Jacobian error is the unnormalized space–time integral J_{err}.

Diagnostic	Central defect ($x_0 = 0$)	Tail defect ($x_0 = 4$)
Score amplitude A	0.09560	1.17745
Population DSM error E_θ^2	0.001	0.001
Unweighted velocity error E_v	0.02276	0.28039
Excess Jacobian error J_{err}	0.009618	1.45902
Mean grid trajectory error	0.001854	0.02953
Maximum grid trajectory error	0.03788	0.28037

Despite equal DSM error, the tail perturbation has a 12.3-fold larger E_v, a 151.7-fold larger J_{err}, and a 15.9-fold larger mean grid trajectory error (Table 3). Figure 1(a) shows the spatial localization of the discrepancies. This supports spatially resolved and low-density validation of the deterministic field.

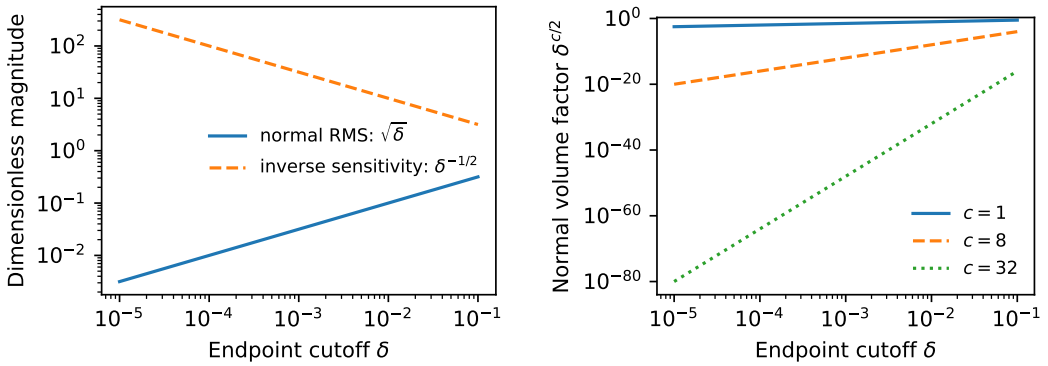
Endpoint fidelity versus conditioning. As a closed-form analogue of Proposition 1, let $X_0 = (Y, 0_c) \in \mathbb{R}^{1+c}$, with $Y \sim \mathcal{N}(0,1)$ and $c \geq 1$, and set $X_t = X_0 + W_t$. Then $p_t = \mathcal{N}(0, 1+t) \otimes \mathcal{N}(0, tI_c)$ and $v(y,z,t) = (y/[2(1+t)], z/[2t])$. The reverse PF-ODE map from T to δ is

$$\Phi_{\delta,T}(y,z) = \left(\sqrt{\frac{1+\delta}{1+T}}\, y, \sqrt{\frac{\delta}{T}}\, z \right). \tag{29}$$

For initialization from p_T, the normal component $\Phi^z_{\delta,T}$ satisfies

$$\left(\frac{\mathbb{E}|z_\delta|^2}{c} \right)^{1/2} = \sqrt{\delta}, \qquad \left\| \left(D_z \Phi^z_{\delta,T} \right)^{-1} \right\| = \sqrt{\frac{T}{\delta}}, \qquad \det D_z \Phi^z_{\delta,T} = \left(\frac{\delta}{T} \right)^{c/2}. \tag{30}$$

Figure 2 shows the resulting fidelity–conditioning tradeoff for $T = 1$. At $\delta = 10^{-2}$, the normal RMS per coordinate is 0.1 and the inverse normal sensitivity is 10, while the normal volume factors are 10^{-8} for $c = 8$ and 10^{-32} for $c = 32$. For the scalar equation $\dot{z} = z/(2t)$ with $z(1) = 1$, adaptive Runge-Kutta45 with relative tolerance 10^{-6} uses 98, 194, 290, and 470 function evaluations at cutoffs 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-5}, respectively. Across the 41 tested cutoffs, the relative endpoint error remains below 2.1×10^{-6}. This example exhibits increasing solver work and inverse sensitivity: the exact reverse map is contractive in the normal directions.



(a) Fidelity and inverse sensitivity

(b) Normal volume collapse

Figure 2: Exact-score endpoint diagnostics for the flat-manifold VE model, with $T = 1$ and $\delta \in [10^{-5}, 10^{-1}]$. Decreasing the normal residual improves fidelity to the data subspace while increasing the sensitivity of the inverse map. Normal volume contraction is amplified by codimension c; these curves follow the exact expressions in (30).

Implementation details DSM calibration uses trapezoidal quadrature with 501 time nodes and 16001 spatial nodes on $[-8, 8]$; E_v and J_{err} use 30001 spatial nodes on $[-6, 6]$, and K_J uses 100001 nodes. The matched-DSM trajectories use RK45 with relative tolerance 10^{-8}, absolute tolerance 10^{-10}, and maximum step 0.01. The regression experiment uses NumPy's default random generator with seed 42. The regularization test uses a known reference score and a single coefficient.