AN OPTIMIZATION FRAMEWORK FOR DENOISING GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Noise-unconditioned denoising learns a single score, so generation can be viewed as gradient ascent on a fixed log-density landscape. This optimization interpretation leaves two questions: do trajectories approach meaningful data, and do different initializations retain diverse outcomes? We establish both phenomena for empirical and population multi-scale Gaussian-mixture objectives. Under explicit posterior and landscape conditions, ascent converges almost surely within $O(m\sigma_{\min}^2/\tau)$ of an m-dimensional reference manifold with reach at least τ. We derive a source–sink identity for the same fixed field and use it to prove finite-resolution regional coverage. In an affine construction, we further identify a positive-bandwidth regime where even empirical-score optimization can avoid training-sample memorization while preserving diversity at the training resolution. The same fixed objective naturally supports penalty-guided controllable generation. Experiments support our predicted mechanisms, show improved robustness under large-step guidance, and demonstrate downstream gains consistent with finite-data generative generalization.

1 INTRODUCTION AND BACKGROUND

Diffusion models have become a central technology of generative AI. They produce high-quality data (Ho et al., 2020; Rombach et al., 2022), support scientific applications such as protein design (Watson et al., 2023), and provide reusable priors for inverse problems (Chung et al., 2022). Increasingly, the goal is not just to generate a meaningful object: It must also agree with measurements, obey physical laws, or improve downstream tasks.

The standard score-based description learns denoising directions indexed by noise level and combines them through a prescribed reverse-time process (Song et al., 2020). Its distributional analysis uses evolution equations; some formulations of control additionally use stochastic optimal control (Han et al., 2024). These tools are powerful, while having high theoretical thresholds for the common generation and application community. However, the demands mentioned above can be formulated elegantly under an optimization framework, without losing comprehensibility. This observation motivates a simpler question:

Can optimization on a fixed objective be defined as a useful denoising generator?

To answer the question, we first revisit the development of diffusion to find its reasons for success.

From score matching to multi-scale generation. Score matching (Hyvärinen, 2005) learns the gradient of a log density without its normalizing constant. Denoising score matching (DSM) (Vincent, 2011) turns this estimation problem into regression on corrupted data. Diffusion probabilistic models (Sohl-Dickstein et al., 2015) connect forward corruption to reverse generation; noise-conditional score networks (NCSNs) (Song & Ermon, 2019) and denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) made multi-scale denoising an effective generative methodology. Reverse-time SDEs and probability-flow ODEs subsequently supplied a common language (Song et al., 2020).

Two ingredients in this progression are especially relevant here: a tractable denoising objective, and information from many noise-perturbed distributions. Under the manifold hypothesis, data occupy a low-dimensional part of ambient space (Fefferman et al., 2016). The ideal clean distribution can even lack an ambient density, whereas generation usually starts from an ambient Gaussian. For $G \sim \mathcal{N}(0, I_d)$, its norm is typically about $\sqrt{d}$ in high dimensions; Gaussian initialization can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

therefore be far from a bounded data region. Learning directions only near clean data does not constrain the entire route back. NCSN addresses this low-density difficulty by perturbing data at multiple scales (Song & Ermon, 2019). Broad perturbations reach the initialization region; fine perturbations resolve local data structure.

Noise unconditioning: keep the scales, remove the noise input. The multi-scale information need not be presented to a network as an explicit noise label. Noise unconditioning retains corrupted examples from all scales but gives the network only the current state. Optimizing this noise-unconditional DSM is equivalent to optimizing the score of a unified Gaussian mixture p_{MN} with M training centers and N noise levels (Zhou et al., 2026). The contrast is

$$\begin{aligned} \text{noise conditioning:} \quad & \{\nabla \log \widehat{p}_{\sigma_i}\}_{i=1}^{N} \quad \longrightarrow \text{ reverse-time denoising,} \\ \text{noise unconditioning:} \quad & \nabla \log p_{MN} \quad \longrightarrow \text{ ascent on a single landscape.} \end{aligned} \tag{1}$$

In the second case, the current state reweights all center–scale contributions. The objective remains fixed, while the current sample determines the contribution of each noise scale. This gives an optimization interpretation of the *exact target field*. Recent work has studied denoising without explicit noise conditioning: Sun et al. (2025) analyze conditioning-removal error, while Kadkhodaie et al. (2026) show that blind denoisers can implicitly infer the relevant noise level. Related autonomous perspectives also include Equilibrium Matching (Wang & Du, 2025) and Beckmann Transport Models (Cheuk-Kit et al., 2026). Building on the unified-density formulation, we study a complementary question: when does ascent on this fixed landscape provide manifold fidelity, regional coverage, and non-memorizing endpoints?

What else is missing to bridge the optimization and generation? The key missing step is not merely to observe that noise-unconditioned sampling can be written as optimization on a fixed landscape, but to explain why optimizing that landscape constitutes a meaningful generative process. More importantly, **convergence is necessary but not sufficient for generation.** A trajectory may converge perfectly and still be useless as a generator if its limit is away from the data manifold or if nearly all initializations collapse to known modes. A complete optimization-based account therefore has to establish both *fidelity* and *diversity*: trajectories should approach the data manifold, and different initializations should continue to populate its different regions.

We prove complementary geometric and probability guarantees for empirical and population noise-unconditioned landscapes. Positive reach τ and localized posterior geometry bound the endpoint's distance to the manifold by $O(m\sigma_{\min}^2/\tau)$, where m is the dimension of the manifold. A source-sink identity links points on the same ascent paths to a minimum-noise smoothing of the reference population; geometric retention transfers this probability to the stopped output.

The same framework also reveals a regime in which empirical-score optimization can generalize beyond the training set. In affine geometry (large reach), normal accuracy no longer requires vanishing noise, leaving a positive bandwidth where score smoothing (Zhou et al., 2026) suppresses sample-level memorization while preserving manifold variation. We construct finite training sets whose infinite-time endpoints remain separated from training samples while retaining coverage at the training resolution. But whether such a window extends to general manifolds remains open. Consistent with finite-data generative generalization, augmentation from a noise-unconditioned generator trained on only 100 minority-class examples outperforms repetition and noise-conditioned diffusion on an imbalanced CIFAR-10 task.

Why the optimization view matters. A fixed prior objective turns *controllable* generation into a familiar constrained- or penalized-optimization problem. Measurements, inequalities, or downstream preferences can be represented by penalty functions and added directly to the generative objective, without introducing a separate projected sampling process. This approach is particularly useful when condition-output pairs are scarce but a scientific requirement can be formulated explicitly, as a constraint or penalty. Guidance and diffusion inverse-problem methods already support inference-time control (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Chung et al., 2022); plug-and-play methods likewise use learned denoisers in reconstruction (Venkatakrishnan et al., 2013; Zhu et al., 2023). Our formulation makes the prior objective explicit and uses the same gradient-based interface, with or without control.

Our optimization-based approach also suggests an advantage under aggressive updates or strong guidance. A large update in a noise-conditioned sampler can push the iterate outside the high-density

region associated with its prescribed noise level (marginal distribution), producing a state–noise mismatch and an inaccurate subsequent score estimate (Zhong et al., 2026). Noise unconditioning instead learns a single field from data perturbed across all noise scales, giving the training distribution much broader coverage of state space. At inference, the current state itself reweights the contributions of different scales, allowing the field to adapt to the locally relevant noise scale rather than relying on an externally scheduled label. This motivates our large-step guided-sampling experiments.

Overall, our main contributions are as follows.

- **A generative theory for noise-unconditioned optimization.** Under explicit geometric and posterior conditions, we prove that fixed-objective ascent converges within $O(m\sigma_{\min}^2/\tau)$ of the data manifold for both empirical and population objectives.
- **Diversity beyond convergence.** We derive a source–sink identity for the same field and use it to prove finite-resolution regional coverage for stopped outputs, showing how initialized trajectories retain probability across data regions.
- **Non-memorizing empirical-score generation.** For a constructive affine family, we identify a positive-bandwidth regime in which exact empirical-score optimization avoids training-sample memorization while preserving diversity at the training resolution.
- **Control and practical utility.** The fixed landscape provides a direct penalty-guided interface for controllable generation. Experiments support the theoretical mechanisms, demonstrate large-step robustness, and show downstream gains consistent with finite-data generative generalization.

2 A Fixed Landscape and Manifold Convergence

We use empirical notation to expose finite-data effects (corresponding to practice); the same framework applies to population objectives (Remark 3.3).

2.1 Learning the gradient, optimizing the sample

Fix training centers $\mathcal{S}_M = \{\mu_1, \dots, \mu_M\} \subset \mathbb{R}^d$ and scales $0 < \sigma_{\min} = \sigma_1 \le \cdots \le \sigma_N = \sigma_{\max}$. Let $\widehat{Q}_M = M^{-1}\sum_j \delta_{\mu_j}$ be the empirical training distribution. Uniform sampling over centers and scales gives the noise-unconditioned loss

$$\mathcal{L}_u(\theta) = \frac{1}{2MN}\sum_{j,i}\sigma_i^2 \mathbb{E}_{X\sim\mathcal{N}(\mu_j,\sigma_i^2 I)}\left\| s_\theta(X) - \frac{\mu_j - X}{\sigma_i^2}\right\|^2. \tag{2}$$

The network receives only X. Define the unified density

$$p_{MN}(x) = \frac{1}{Z_{MN}}\sum_{j,i}\sigma_i^2 \mathcal{N}(x;\mu_j,\sigma_i^2 I), \qquad Z_{MN} = M\sum_i \sigma_i^2. \tag{3}$$

Completing the square identifies its gradient as the exact learning target (Zhou et al., 2026):

$$\mathcal{L}_u(\theta) = \frac{Z_{MN}}{2MN}\mathbb{E}_{X\sim p_{MN}}\|s_\theta(X) - \nabla \log p_{MN}(X)\|^2 + C, \tag{4}$$

where C is independent of θ. Thus *training optimizes the parameters to learn a gradient; generation optimizes the sample along that gradient*:

$$\dot{x} = \nabla \log p_{MN}(x), \qquad x_{n+1} = x_n + \eta_n s_\theta(x_n). \tag{5}$$

Our theory analyzes the exact empirical field $\nabla \log p_{MN}$ and its population analogue; s_θ is the learned approximation used in practice. Appendix A.2 discusses the relation among them.

How one objective uses multiple scales. Let $w_{ij}(x)$ be the mixture responsibility proportional to $\sigma_i^2\mathcal{N}(x;\mu_j,\sigma_i^2 I)$, and define

$$\beta = \sum_{j,i}\frac{w_{ij}}{\sigma_i^2}, \qquad \widetilde{w}_{ij} = \frac{w_{ij}}{\sigma_i^2\beta}, \qquad \bar{\mu} = \sum_{j,i}\widetilde{w}_{ij}\mu_j. \tag{6}$$

All three depend on x. Differentiation yields

$$\nabla \log p_{MN}(x) = \beta(x)(\bar{\mu}(x) - x), \qquad \sigma_{\max}^{-2} \le \beta(x) \le \sigma_{\min}^{-2}. \tag{7}$$

The field moves toward a state-dependent average of training centers, with β setting its speed. Broad kernels contribute away from the data and finer kernels resolve nearby structure, within one unchanged objective. Every critical point is a convex combination of training centers. This already bounds where optimization can end, but a convex combination can lie off a curved manifold. The next argument controls precisely that geometric error.

2.2 Geometry controls normal convergence

Let $\mathcal{M}$ be a compact embedded C^2 reference manifold without boundary, containing the centers, with dimension $1 \le m < d$ and reach at least $\tau > 0$. Reach measures the size of a neighborhood with unique nearest-point projection. Set $c = d - m$, $r(x) = \operatorname{dist}(x, \mathcal{M})$ and $z_x = \Pi_{\mathcal{M}}(x)$.

For initialization supported on a nonempty compact set K, define $\Omega = \{x : p_{MN}(x) \ge \min_K p_{MN}\}$. Ascent increases $\log p_{MN}$, so the trajectory stays in Ω; Gaussian decay makes this set compact. Analyticity then gives convergence to a critical point (Absil et al., 2005). To show that the limit represents the data geometry, choose $0 < r_0 < \tau$, $r_0 \le \sigma_{\max}\sqrt{c}$, and require, for some dimensionless constant $C_{\mathrm{loc}} > 0$

$$\sum_{j,i} \widetilde{w}_{ij}(x)\|\mu_j - z_x\|^2 \le C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r(x)^2/c\}, \quad x \in \Omega, \quad 0 < r(x) < r_0. \tag{8}$$

This condition controls how widely the active training centers spread around the nearest manifold point. The scale $r/\sqrt{c}$ is motivated by a flat-patch calculation; the inequality controls the full posterior, not just its most likely scale (Appendix A.4).

Put

$$\varepsilon_0 = \frac{C_{\mathrm{loc}} m \sigma_{\min}^2}{2\tau}, \qquad r_{\min} = \sigma_{\min}\sqrt{c}, \qquad \kappa_0 = \frac{C_{\mathrm{loc}} m r_0}{2c\tau}, \tag{9}$$

and assume $\kappa_0 < 1$ and $\varepsilon_0 < \min\{r_{\min}, r_0\}$. Reach bounds the normal component of a chord by its squared length divided by 2τ. Averaging this bound with (8) gives

$$\dot{r} \le \begin{cases} -\beta(x)(r - \varepsilon_0), & 0 < r \le r_{\min}, \\ -\beta(x)(1 - \kappa_0) r, & r_{\min} < r < r_0. \end{cases} \tag{10}$$

The direct pull toward the manifold dominates the curvature bias of the average, excluding critical points from $\varepsilon_0 < r < r_0$. Large reach allows stronger smoothing before that bias becomes significant.

We also need to exclude attracting equilibria farther away. Write $a_{ij} = (\mu_j - x)/\sigma_i^2$ and $C_x = \mathrm{Cov}_{w(x)}(a_{ij})$. Since $\nabla^2 \log p_{MN} = C_x - \beta I$, the conditions $\operatorname{tr} C_x \ge \alpha_d d\beta$ and $\operatorname{tr} C_x / \lambda_{\max}(C_x) \le R_d < \alpha_d d$ at off-tube critical points imply

$$\lambda_{\max}(\nabla^2 \log p_{MN}(x)) \ge \left(\frac{\alpha_d d}{R_d} - 1\right)\beta(x) > 0, \qquad 0 < \alpha_d \le 1. \tag{11}$$

Posterior variation in sufficiently few directions therefore makes such points unstable for ascent. Local attraction and off-tube instability together locate the generated endpoints.

Theorem 2.1 (Manifold convergence)**.** *Under the preceding conditions, suppose each off-tube critical point in Ω has a nonsingular Hessian $\nabla^2 \log p_{MN}$, and ν is absolutely continuous with respect to ambient Lebesgue measure, with $\nu(K) = 1$. Every exact trajectory from K converges, and*

$$\operatorname{dist}(x_\infty, \mathcal{M}) \le \varepsilon_0 = \frac{C_{\mathrm{loc}} m}{2\tau}\sigma_{\min}^2 \quad \nu\text{-almost surely}. \tag{12}$$

Every local maximizer in Ω and every global maximizer satisfy the same bound, without the nonsingularity or initialization-density conditions.

Appendix A proves the result. It supplies the fidelity part of the optimization framework: optimization converges, and the limit is close to the reference data manifold.

Flat geometry permits smoothing without normal error. If the centers lie in an affine subspace $\mathcal{A} = a_0 + V$, so does $\bar{\mu}(x)$, giving the exact identity

$$r_{\mathcal{A}}(x(t)) \le r_{\mathcal{A}}(x(0)) e^{-t/\sigma_{\max}^2} \longrightarrow 0. \tag{13}$$

This holds for every fixed positive schedule (Proposition A.18). It separates the smoothing used to form the objective from its final normal accuracy. Section 3.2 uses this separation to preserve smoothing against memorization while retaining variation along the manifold.

3 DIVERSITY, NON-MEMORIZATION, AND CONTROL

Manifold convergence makes the optimizer geometrically meaningful. To make it a generator, we must also control how initialization probability is distributed among its outcomes. We first establish that coverage/diversity, then combine it with positive smoothing to obtain non-memorizing endpoints, and finally use the same fixed objective as a prior for penalty-guided controllable generation.

3.1 RETAINING PROBABILITY ACROSS DATA REGIONS

Assume $\mathcal{M}$ is connected and specify a reference data distribution $Q_{\mathcal{M}}$ on it. The manifold describes geometry; $Q_{\mathcal{M}}$ supplies the regional probabilities that generation should retain. We first state the result for the empirical $\nabla \log p_{MN}$. For a Borel region $U \subseteq \mathcal{M}$, define

$$U^{-\delta} = \{z \in U : d_{\mathcal{M}}(z, \mathcal{M} \setminus U) > \delta\}, \tag{14}$$

where $d_{\mathcal{M}}$ is intrinsic distance. This safe interior ensures that any point displaced by at most δ remains within the region.

Local refinement preserves incoming diversity. Choose $\varepsilon_0 < r_{\text{stop}} < r_{\min} < r_{\text{tr}} < r_0$. Let τ_{tr} and τ_{stop} be the first entrance times into these distance tubes, and write $X_{\text{stop}} = x(\tau_{\text{stop}})$, $Z_{\text{stop}} = \Pi_{\mathcal{M}}(X_{\text{stop}})$ and $Z_{\text{tr}} = \Pi_{\mathcal{M}}(x(\tau_{\text{tr}}))$. Comparing projected tangential speed with inward speed gives

$$\int_{\tau_{\text{tr}}}^{\tau_{\text{stop}}} \left\| \frac{d}{dt} \Pi_{\mathcal{M}}(x(t)) \right\| dt \le \delta_{\text{stop}} \quad \text{almost surely.} \tag{15}$$

Proposition B.2 gives an explicit bound of order $\sqrt{m}\,\sigma_{\min}$ for admissible noise-proportional radii and controlled geometric constants. Hence the refinement stage can only induce limited redistribution of probability after trajectories enter the manifold neighborhood. The remaining question is therefore how much probability reaches each region.

The same multi-scale objective supplies the probability argument. Let $p_\sigma^Q = Q_{\mathcal{M}} * \mathcal{N}(0, \sigma^2 I)$. On a finite geometric grid $\sigma_i = \sigma_{\min} e^{(i-1)h}$, set $A_N = h \sum_i \sigma_i^2$. Gaussian scale calculus gives

$$\nabla \cdot (A_N p_{MN} \nabla \log p_{MN}) = p^Q_{\sigma_{\max}} - p^Q_{\sigma_{\min}} + \mathcal{R}_N, \qquad \epsilon_{\text{Pois}} = \tfrac{1}{2}\|\mathcal{R}_N\|_1. \tag{16}$$

The broad smoothing supplies probability and the fine smoothing receives it. The residual accounts for the finite grid and the difference between training and reference distributions. The σ^2 weighting is important: with log-scale integration, $\partial_\sigma p_\sigma^Q = \sigma \Delta p_\sigma^Q$ makes the balance telescope to the two cutoffs.

This stationary balance concerns paths, not their distributions at one common time. Its occupation-density argument places an auxiliary point Y on each original trajectory such that

$$\text{TV}(P_Y, p^Q_{\sigma_{\min}}(x)\,dx) \le \epsilon_{\text{Pois}} + \epsilon_{\text{init}}, \qquad \epsilon_{\text{init}} = \text{TV}(\nu, p^Q_{\sigma_{\max}}(x)\,dx), \tag{17}$$

where TV is the Total Variation distance and P_Y denotes the distribution of Y. The sampler still returns X_{stop}; the mark connects that path to reference data probabilities in the proof (Appendix C).

For confidence parameter $\zeta \in (0, 1)$, let a_d, a_m be upper Gaussian norm thresholds and b_c a lower normal threshold, each with tail probability $\zeta/3$. Define

$$\delta_{\text{noise}} = \frac{\sigma_{\min} a_m}{1 - \sigma_{\min} a_d/\tau}, \qquad \delta_{\text{cov}} = \delta_{\text{noise}} + \delta_{\text{stop}}. \tag{18}$$

If $\sigma_{\min} a_d \le r_{\text{tr}}$ and $r_{\text{stop}} < \sigma_{\min} b_c - \delta^2_{\text{noise}}/(2\tau)$, a reference perturbation falls in the refinement segment with probability at least $1-\zeta$. Its projection is within δ_{noise} of its reference point. Combining this displacement with (15) transfers reference mass to the returned sample.

Theorem 3.1 (Regional coverage)**.** *Under Theorem 2.1 and the preceding radius, grid and Gaussian conditions, $r(X_{\text{stop}}) \le r_{\text{stop}}$ almost surely, and for every Borel $U \subseteq \mathcal{M}$,*

$$\mathbb{P}_\nu[Z_{\text{tr}} \in U,\ Z_{\text{stop}} \in U] \ge (1-\zeta) Q_{\mathcal{M}}(U^{-\delta_{\text{cov}}}) - \epsilon_{\text{Pois}} - \epsilon_{\text{init}}. \tag{19}$$

Together, these theorems establish manifold fidelity and finite-resolution regional coverage for stopped outputs through the same fixed-objective ascent process. Appendix B gives the proof and error bounds. The next result adds a distinct property: diverse, non-memorizing endpoints even with an exact empirical score.

3.2 NON-MEMORIZING ENDPOINTS AT THE TRAINING RESOLUTION

Beyond fidelity and diversity, a positive smoothing floor can also prevent training-set memorization. At one scale, $w_{ij}/w_{i\ell} = \exp((\|x-\mu_\ell\|^2 - \|x-\mu_j\|^2)/(2\sigma_i^2))$: small noise amplifies a training center's distance advantage (Zhou et al., 2026). Affine contraction allows us to smooth this contrast while keeping zero normal endpoint error. We now show that this can also preserve generalization.

In $\mathcal{A} = \mathbb{R}^m \times \{0\}$, with $c = d - m \geq 2$, choose spacing $\ell > 0$ and the interlaced training grid

$$\mathcal{S}_M = \{(\ell(\mathbf{k} + \boldsymbol{\epsilon}/6), 0) : \mathbf{k} \in \{-L, \dots, L\}^m,\ \boldsymbol{\epsilon} \in \{-1, 1\}^m\}, \quad M = 2^m(2L+1)^m. \tag{20}$$

Let $Q_{\mathcal{M}}$ be uniform on the connected box $D_L = [-(L+1/2)\ell, (L+1/2)\ell]^m \times \{0\}$. Its training covering radius is $h_{\mathrm{cov},M} = \sqrt{m}\,\ell/3$. Each grid cell carries training and reference mass $2^m/M$, and contains the output core $C_{\mathbf{k}} = \{(u, 0) : |u_j - \ell k_j| \leq \ell/12 \text{ for all } j\}$.

Corollary 3.2 (Non-memorization with sample-resolution diversity)**.** *Fix $m \geq 1$, $c \geq 2$ and $\epsilon \in (0, 1/4)$. There are finite constants $s_0 \geq 1/2, L_0$ such that, for every $L \geq L_0$, one finite relative geometric schedule works throughout $\sigma_{\min} \in [s_0\ell, 2s_0\ell]$. Every exact trajectory converges to $x_\infty \in \mathcal{A} \cap \mathrm{conv}(\mathcal{S}_M)$. Under initialization $\mathcal{N}(0, \sigma_{\max}^2 I_d)$, for every Borel $U \subseteq \mathcal{A}$,*

$$\mathbb{P}[x_\infty \in U,\ \mathrm{dist}(x_\infty, \mathcal{S}_M) \geq h_{\mathrm{cov},M}/4] \geq Q_{\mathcal{M}}(U^{-7h_{\mathrm{cov},M}/4}) - \epsilon. \tag{21}$$

For every collection $I \subseteq \{-L, \dots, L\}^m$,

$$\left| \mathbb{P}\left[x_\infty \in \bigcup_{\mathbf{k} \in I} C_{\mathbf{k}} \right] - \frac{|I|}{(2L+1)^m} \right| \leq \epsilon. \tag{22}$$

The same outputs carry regional coverage and a training-distance margin. At least $(1-2\epsilon)M/2^m$ disjoint cores receive probability at least $2^{m-1}/M$, so a connected patch retains many outcomes rather than one representative. Both the margin and the coverage resolution scale with the training covering radius. The cell-probability comparison also controls *any union* of cells, giving distributional rather than merely geometric coverage.

The proof verifies finite-mixture gradient signs in small capture cells, then follows the source–sink marks to their endpoints (Appendix D). It establishes a nonempty bandwidth window and a sequence whose endpoint distributions converge in Wasserstein-1 distance W_1 to a continuous reference on a fixed box. Its significance is that *even exact empirical-score optimization can avoid memorization while retaining diversity at the data's sampling resolution*.

The window balances sample-level smoothing against the variation needed to distinguish nearby regions. Curvature adds the normal-accuracy restriction $C_{\mathrm{loc}} m \sigma_{\min}^2/(2\tau) \leq \varepsilon$; larger reach relaxes it when localization remains controlled. The contrast with diffusion is specifically its exact empirical zero-noise endpoint, which recovers $\widehat{Q}_M$ (Appendix D.8).

Remark 3.3 (Population distributions). Theorems 2.1 and 3.1 also hold for $p^{\mathrm{pop}}(x) = \sum_i \sigma_i^2 p_{\sigma_i}^Q(x) / \sum_i \sigma_i^2$ under the corresponding population assumptions. Center sums become $Q_{\mathcal{M}}$-expectations, and ϵ_{Pois} contains only noise-grid error; ϵ_{init} remains. The same coupling gives $P_{X_{\mathrm{stop}}} \to Q_{\mathcal{M}}$ in W_1 along admissible model families on a fixed compact manifold as spatial-resolution and probability errors vanish (Appendix I). Corollary 3.2 complements this population approximation with diverse, non-memorizing empirical-score endpoints; its fixed-box construction also approaches a continuous reference distribution in W_1.

3.3 CONTROLLABLE GENERATION BY PENALTY-GUIDED OPTIMIZATION

The fixed objective also provides a simple interface for controllable generation. Let y denote a task, observation, or desired property, and let $E_y(x) \geq 0$ be a differentiable task penalty whose smaller values indicate better task satisfaction. Rather than changing the generative prior, we maximize

$$F_y(x) = \log p_{MN}(x) - \lambda E_y(x), \quad \text{for some } \lambda \geq 0, \tag{23}$$

by ordinary gradient ascent:

$$x_{n+1} = x_n + \eta_n \left[\nabla \log p_{MN}(x_n) - \lambda \nabla E_y(x_n) \right], \quad n = 0, 1, 2, \dots, \tag{24}$$

for a sequence of steplengths $\{\eta_n\}$. The parameter λ trades off the generative landscape against the task objective. (In practice, $\nabla \log p_{MN}$ can be replaced by the learned state-only score s_θ.) With the exact score, this is gradient ascent on the single objective F_y; under the analytic conditions detailed in Appendix H, the corresponding flow converges to a stationary point satisfying

$$\nabla \log p_{MN}(x_{\infty,y}) = \lambda \nabla E_y(x_{\infty,y}).$$

Compatibility with generative geometry. For tasks that primarily modify semantic or latent degrees of freedom along the data manifold, the guidance near the manifold is naturally expected to be predominantly tangential; its normal component measures how strongly the task pushes against the learned data geometry. Inside the manifold tube, let $u_x = (x - \Pi_{\mathcal{M}}(x))/r(x)$. Assume that the same posterior-localization bound holds on the guided tube region and that there exists $\delta_y \geq 0$ such that

$$\lambda[-\langle u_x, \nabla E_y(x)\rangle]_+ \leq \beta(x)\delta_y, \tag{25}$$

where $\beta(x)$ is defined in (7). Here δ_y quantifies the outward normal guidance relative to the prior's local contraction scale. Combining this bound with (10) preserves inward attraction outside the inflated radius

$$\varepsilon_y = \max\left\{\varepsilon_0 + \delta_y,\ \frac{\delta_y}{1-\kappa_0}\right\}. \tag{26}$$

Provided $\varepsilon_y < r_0$, the guided flow retains this inward attraction after entering the tube. Tangential or inward guidance ($\delta_y = 0$) leaves the local normal bound unchanged, while a small outward component enlarges it continuously. Analogously, multiple task-compatible regions remain locally available when guidance does not overwhelm their separating/inward margins. Appendix H gives the convergence analysis and sufficient conditions for preserving local manifold attraction and multiple task-compatible regions under penalty guidance.

4 Experiments

We first test the main theoretical mechanisms using exact score fields, then evaluate practical noise unconditioning, large-step control, and downstream augmentation.

4.1 Numerical verification of the theoretical mechanisms

We test four mechanisms using exact population or empirical score fields. The normalized Euler updates in E1/E3 increase the exact log density for $0 < \eta < 2$ (Lemma F.1). Full settings and numerical limitations are in Appendix F.

E1: coverage and retention. Across five nonuniform targets on $S^2(1) \subset \mathbb{R}^{12}$, with $M = 10^4$ centers and 10^5 trajectories per target, the partition-level coverage diagnostic is $\widehat{\alpha}_{\mathcal{P}} = 0.789$–$0.987$, and the 99.9th-percentile projected refinement displacement is at most 0.112 radians. These results support regional probability allocation and local retention beyond population symmetry (Appendix F.1).

E2: off-tube instability. Population radial scores are positive at every tested off-tube grid node. For the $M = 2000$ uniform empirical sphere mixture, the located off-tube candidate has three positive and nine negative log-density Hessian eigenvalues, matching the predicted ascent instability. No off-tube root is found for the tested vMF cloud; the search is not exhaustive (Appendix F.2).

E3: positive-bandwidth non-memorization and diversity. On the 82-center affine grid with a fixed relative schedule, $s = \sigma_{\min}/\ell \in \{0.18, 0.20, 0.22, 0.24\}$ simultaneously yields 96.79–97.34% training-set separation (threshold $h_{\mathrm{cov},M}/4$), 39/41 populated cores, and core-label TV 0.0888–0.1060 (Figure 1). No output meets the separation threshold at the tested $s \leq 0.12$. These full-sample statistics exhibit the joint separation–diversity mechanism of Corollary 3.2 in a small finite instance, while the theorem uses more conservative sufficient constants (Appendix F.3).

E4: geometric scaling and posterior localization. Population spheres exhibit the predicted $O(\sigma_{\min}^2/\tau)$ bias, with fitted exponents 2.0006–2.0091. At the smallest tested $\sigma_{\min}/R = 0.005$, the normalized bias differs by only 0.00306% from $C_{\mathrm{fs}} = 1.199652\ldots$, computed from the schedule rather than fitted. On a separate 10^4-center sphere, the median localization ratio has 1.17% mean relative discrepancy from the matched flat reference over 15 radii; the largest sampled ratio is 2.3701 (Figure 2). These results support the geometric scaling and sampled posterior-localization mechanism underlying Theorem 2.1 (Appendix F.4).

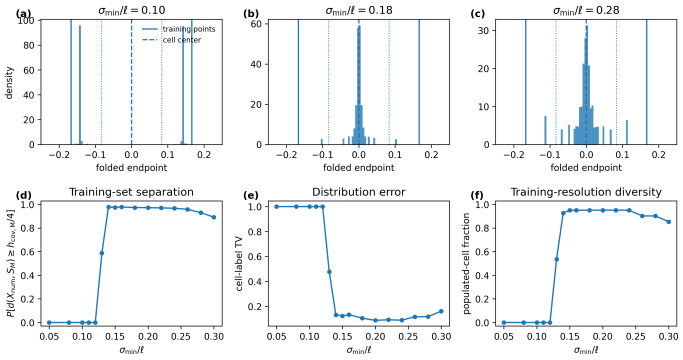

Figure 1: **Positive bandwidth separation and diversity (E3).** (a)-(c) Folded outputs at three bandwidths; solid, dashed, and dotted lines mark training locations, cell centers, and core boundaries. (d)-(f) Full-grid training-set separation, core-label TV, and populated-core fraction. Folding is used only for visualization.

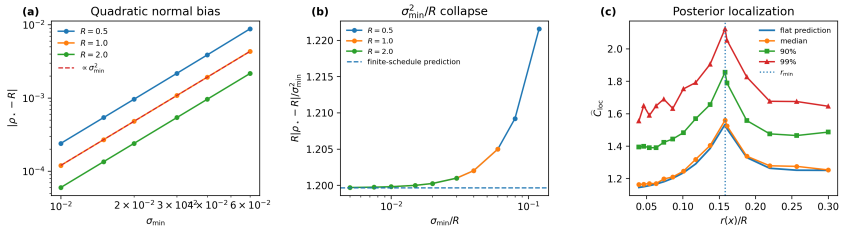

Figure 2: **Geometric scaling and localization (E4).** (a) Population bias. (b) Normalized bias versus the finite-schedule prediction. (c) Empirical localization ratios versus the matched flat reference.

4.2 Noise unconditioning across diffusion/flow models

Table 1: Comparison on CIFAR-10.

Model	FID ↓ w/ t	FID ↓ w/o t	Setting
DDPM/VP-SDE	3.42	5.77	remove t; SDE-1000
EDM [1]	1.99	3.36	remove t; Heun-35
uEDM [1]	2.04	2.23	uEDM precond.; Heun-35
FM / 1-RF [1]	3.01	**2.61**	remove t; Euler-100
VE-SDE [2]	6.43	7.47	remove emb.; PC-SDE-1K
Linear FM [3]	5.40	**5.31**	set $t = 0$; Euler-250

Table 2: Comparison on ImageNet 256×256.

Model	FID ↓ w/ t	FID ↓ w/o t	Setting
SiT-B/2 [1]	7.83	7.96	remove t; Euler-250 + CFG
DiT-B/2 [5]	19.83	38.54	remove t; VP-DDIM
DiT-B/2 [5]	19.83	**10.01**	new VP schedule; no t
EqM-B/2 [4]	38.21	42.74	remove t; const. target
SiT-XL/2→EqM [4]	2.42	**2.17**	EqM target + NAG-GD
FM-XL/2→EqF [6]	18.91	**15.15**	η-warp + NAG; no t

Unconditioning Sources: [1] Sun et al., 2025; [2] Zhou et al., 2026; [3] Helbling et al., 2026; [4] Wang & Du, 2025; [5] Li et al., 2026; [6] Lillemark et al., 2026. The table includes both direct removal ablations and method-level noise-unconditioned variants that additionally modify the schedule, target, preconditioning, or sampler.

Tables 1 and 2 compare models with and without explicit noise/time conditioning on CIFAR-10 and ImageNet 256×256, where we reproduced DDPM (Ho et al., 2020) and [2, 4, 5], and report the published results of [1, 3, 6]. They include both direct removal ablations and method-level variants that redesign the schedule, target, or solver. Removing explicit conditioning often causes only moderate Fréchet Inception Distance (FID) degradation and can even improve some flow-based models (Sun et al., 2025; Helbling et al., 2026); when naive removal is unfavorable, redesigned time-independent objectives or schedules can recover strong performance (Wang & Du, 2025; Li et al., 2026; Lillemark et al., 2026). These results support the practical viability of high-quality generation without explicit noise/time conditioning across a range of diffusion and flow-based models.

4.3 Large-step controllable generation

We test robustness to guided sampling on a controlled toy manifold and on CIFAR-10 using the same ODE sampler as (Zhou et al., 2026) with 1000 or 100 steps. The toy distribution is defined in a 100-dimensional latent space: its first two coordinates form four separated disks and the remaining 98 are Gaussian, followed by a canonical embedding into $\mathbb{R}^{3072}$. Requiring the first two latent coordinates to be positive selects the first-quadrant disk. On CIFAR-10, a pretrained classifier guides generation toward the *horse* class. Guidance decays linearly from 1 to 0; on CIFAR-10 its gradient is additionally normalized to the score norm. Full settings are in Appendix H.4.

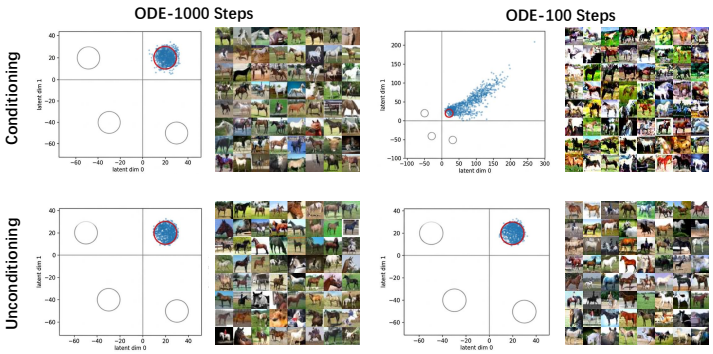

Figure 3: **Large-step controllable generation.** The top and bottom rows show noise-conditioned and noise-unconditioned sampling, respectively. Results with 1K steps are shown on the left, and 100-step results on the right. Each setting includes four-disk control and CIFAR-10 horse guidance.

At 1000 steps, both approaches satisfy the target with high-quality samples. At 100 steps, guided noise-conditioned sampling degrades sharply, whereas noise unconditioning remains stable on both tasks (Figure 3). The unguided 100-step conditioned samplers remain stable (Appendix H.4), identifying guidance as a key factor in the coarse-step degradation. These results demonstrate improved robustness in the tested settings and are consistent with the state–noise mismatch explanation.

4.4 Downstream benefits illustration

We next test whether finite-data generative generalization benefits a downstream task: if a generator trained on few data mainly reproduces its training samples (memorization), augmentation should resemble reweighting, whereas useful generalization should add within-class variation. On CIFAR-10, we use 100 cats and 1000 dogs; the minority class (cat) is either repeated or augmented with 900 samples from a diffusion or noise-unconditioned generator trained using *only those same 100 cats*. We train a binary ResNet-18 and evaluate on the cat–dog test set. A balanced set with 1000 real cats serves as a reference; full settings and pretrained-generator controls are in Appendix H.5.

Table 3: **Downstream classification under minority-class scarcity.** Test accuracy (%, mean±std).

Training condition	Cat	Dog	Bal. Acc.
Biased real	11.80±1.25	96.73±0.21	54.27±0.56
Repeated cats (1000)	13.63±0.76	97.13±0.46	55.38±0.41
Diffusion aug.	26.17±1.65	93.33±0.47	59.75±0.95
Unconditioning aug.	**38.80±3.45**	91.87±0.95	**65.33±2.03**
Balanced real	79.17±0.76	79.63±1.40	79.40±1.01

Repeating the minority samples barely improves balanced accuracy ($54.27\% \rightarrow 55.38\%$). With both generators trained on the same 100 cats, noise-unconditioned augmentation improves balanced accuracy by 5.58 percentage points over diffusion and 9.95 points over repetition; cat accuracy improves by 12.63 points over diffusion. These gains are consistent with the generalization picture, and show that the resulting variation can benefit downstream tasks. The same pattern persists in the ten-class extension (Appendix H.5).

5 Conclusion

Noise unconditioning offers a simple reinterpretation of multi-scale denoising: rather than following a prescribed sequence of noise levels, generation can be viewed as optimization on a single fixed landscape. From this perspective, we establish two basic generative properties—fidelity and diversity—through manifold convergence and finite-resolution regional coverage. We further show that, in a constructive affine setting, positive smoothing can prevent exact empirical-score optimization from collapsing onto the training set while preserving diversity at the training resolution. The same fixed objective also provides a natural interface for penalty-guided controllable generation.

Our experiments connect these mechanisms to practice: exact-score studies support the predicted geometric and diversity behavior, while learned models show robustness under large guided updates and downstream benefits. Overall, noise-unconditioned denoising provides an elegant framework for understanding the generation process from an optimization perspective.

AI USE STATEMENT

In this work, we used generative AI tools for polishing and drafting a few sections of the appendix. We also used it for theory proof brainstorming, checking, and correcting. We have **not** used generative AI tools for retrieval and discovery, research ideation or execution, and generating synthetic datasets does not apply to this work. Additionally, we used generative AI tools for partially coding E1-E4 and the downstream classification task. We have reviewed all AI-assisted work. For example, LLM-generated code, special theory case design, and mathematical calculations were verified and tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete assumptions and proofs for the theoretical results are provided in Appendices A–E and I. Appendix F reports the full numerical settings, procedures, diagnostics, and limitations for E1–E4. Appendices H.4–H.5 provide the implementation details for controllable generation and downstream classification, including sampling procedures, data splits, optimization settings, and random seeds. Anonymous reproduction code is partially provided in the supplementary material.

REFERENCES

Pierre-Antoine Absil, Robert Mahony, and Ben Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.

Lee Cheuk-Kit, Florentin Coeurdoux, Peter Potaptchik, Yilun Du, Michael Samuel Albergo, and Eric Vanden-Eijnden. Beckmann transport models: From autonomous flows to one-step maps. *arXiv preprint arXiv:2608.01692*, 2026.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Stochastic control for fine-tuning diffusion models: Optimality, regularity, and convergence. *arXiv preprint arXiv:2412.18164*, 2024.

Alec Helbling, Sebastian Gutierrez Hernandez, Benjamin Hoover, Duen Horng Chau, and Parikshit Ram. What time is it? how data geometry makes time conditioning optional for flow matching. *arXiv preprint arXiv:2605.08344*, 2026.

Morris W Hirsch, Charles Chapman Pugh, and Michael Shub. Invariant manifolds. *Bulletin of the American Mathematical Society*, 76(5):1015–1019, 1970.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695–709, 2005.

Zahra Kadkhodaie, Aram-Alexandre Pooladian, Sinho Chewi, and Eero Simoncelli. Blind denoising diffusion models and the blessings of dimensionality. *arXiv preprint arXiv:2602.09639*, 2026.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, Proceedings of Machine Learning Research. PMLR, 2023.

Gunther Leobacher and Alexander Steinicke. Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of global analysis and geometry*, 60(3):559–587, 2021.

Liuzhuozheng Li, Zhiyuan Zhan, Shuhong Liu, Dengyang Jiang, Zanyi Wang, Guang Dai, Jingdong Wang, and Mengmeng Wang. Exploring time conditioning in diffusion generative models from disjoint noisy data manifolds. *arXiv preprint arXiv:2604.25289*, 2026.

Hansen Jin Lillemark, Alex Rojas, Zachary Novack, Runqian Wang, Yilun Du, Yian Ma, Taylor Berg-Kirkpatrick, and Rose Yu. Equilibrium forcing: Adaptive video generation without noise conditioning. *arXiv preprint arXiv:2608.14706*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? *arXiv preprint arXiv:2502.13129*, 2025.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pp. 945–948. IEEE, 2013.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. *Advances in Neural Information Processing Systems*, 35:16782–16795, 2022.

Jincheng Zhong, Boyuan Jiang, Xin Tao, Pengfei Wan, Kun Gai, and Mingsheng Long. Mitigating noise shift in denoising generative models with noise awareness guidance. In *International Conference on Learning Representations*, volume 2026, pp. 34193–34211, 2026.

Xinyu Zhou, Jiawei Zhang, and Stephen J Wright. Smoothing the score function for generalization in diffusion models: An optimization-based explanation framework. *arXiv preprint arXiv:2601.19285*, 2026.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1219–1229. IEEE, 2023.

CONTENTS

1 Introduction and Background **1**

2 A Fixed Landscape and Manifold Convergence **3**

2.1 Learning the gradient, optimizing the sample 3

2.2 Geometry controls normal convergence 4

3 Diversity, Non-Memorization, and Control **5**

3.1 Retaining probability across data regions 5

3.2 Non-memorizing endpoints at the training resolution 6

3.3 Controllable generation by penalty-guided optimization 6

4 Experiments **7**

4.1 Numerical verification of the theoretical mechanisms 7

4.2 Noise unconditioning across diffusion/flow models 8

4.3 Large-step controllable generation 8

4.4 Downstream benefits illustration 9

5 Conclusion **9**

A Detailed Setup, Manifold Convergence, and Representation Geometry **15**

A.1 The empirical objective and posterior identities 15

A.2 Learning the gradient of the unified log density 15

A.3 Global existence and convergence 16

A.4 Local geometry and distance-matched scales 17

A.5 From posterior localization to normal attraction 18

A.6 Off-tube instability and saddle avoidance 19

A.7 Manifold convergence and optimizer localization 21

A.8 Affine representations and the meaning of larger reach 22

B Diversity: Retention and Regional Coverage **24**

B.1 Reference distribution and output time 24

B.2 Tangential motion during refinement 24

B.3 The source–sink identity and a point on the same path 25

B.4 From reference probability to output coverage 26

B.5 Noise, sampling error and coverage resolution 27

C Source–Sink Proofs and Error Bounds **29**

C.1 A marked trajectory and its occupation density 29

C.2 Projection of a Gaussian perturbation 30

C.3 Finite-grid, finite-sample and initialization errors 30
C.4 Why the scale weight matters 31

D Endpoint Non-Memorization at the Training-Set Resolution 32
D.1 Construction and probability coupling 32
D.2 Coupling a marked path to a resolved reference cell 32
D.3 Periodic comparison and the finite-cloud error 33
D.4 A fine-cell capture controls the full future trajectory 34
D.5 Explicit probability errors and a nonempty bandwidth window 35
D.6 Endpoint separation and regional coverage 37
D.7 Compatibility with a finite reach 38
D.8 Comparison with the zero-noise empirical diffusion endpoint 38

E Solvable Models 40
E.1 Flat model: posterior moments and finite schedules 40
E.2 Sphere model: curvature, localization, and exact regional coverage 42
E.3 Finite-schedule reference calculations for E4 49

F Numerical Verification: Full Settings and Results 51
F.1 E1: Regional probability beyond symmetry 52
F.2 E2: Off-tube critical structure 53
F.3 E3: Positive-bandwidth separation and sample-resolution diversity 57
F.4 E4: Geometric scaling and posterior-localization diagnostics 60

G Finite-Sample Refinements and Output Resolution 64
G.1 A posterior-moment residual 64
G.2 Training distance and the resolution of coverage 65
G.3 From a finite stop to an endpoint 65

H Penalty-Guided Generation: Convergence, Geometry, and Experimental Details 66
H.1 Penalty-guided ascent and stationary points 66
H.2 How guidance perturbs manifold attraction 67
H.3 Guidance strength and retained diversity 68
H.4 Large-step controllable-generation comparison 68
H.5 Downstream classification under data imbalance 69

I Population Objectives: Convergence and Regional Coverage 72
I.1 The population learning target and posterior identities 72
I.2 Manifold convergence for a population objective 72
I.3 Population coverage and the disappearance of sampling error 73

NOTATION AND ROUGH GUIDE

Appendix A proves convergence and geometric localization. Appendices B–C derive regional coverage for the same empirical trajectories. Appendix D proves the non-memorizing endpoint construction; Appendix H develops penalty-guided generation and its experimental instantiations; and Appendix I gives the population counterparts. All norms are Euclidean unless an intrinsic distance is specified. We write P_X for the distribution of a random variable X.

Notation	Meaning
Data, objective, and score	
d, M, N	Ambient dimension, number of training centers, and number of noise scales.
$\mathcal{S}_M = \{\mu_j\}_{j=1}^M$, $\widehat{Q}_M$	Training centers and empirical distribution $\widehat{Q}_M = M^{-1}\sum_j \delta_{\mu_j}$.
σ_i, $\sigma_{\min}$, $\sigma_{\max}$	Noise standard deviations and their lower and upper cutoffs.
$\widehat{p}_\sigma$	Gaussian smoothing of the empirical distribution, $\widehat{Q}_M * \mathcal{N}(0, \sigma^2 I)$.
p_{MN}, Z_{MN}	Unified empirical multi-scale density and its normalization constant.
s_θ	Learned state-only approximation to $\nabla \log p_{MN}$.
w_{ij}, $\widetilde{w}_{ij}$	Mixture responsibilities and precision-normalized responsibilities.
$a_{ij} = (\mu_j - x)/\sigma_i^2$	Component score vector.
β, $\bar{\mu}$	Average precision and posterior barycenter, with $\nabla \log p_{MN} = \beta(\bar{\mu} - x)$.
h, η_n	Log-noise grid spacing and optimization step size.
Geometry and convergence	
$\mathcal{M}$, $Q_{\mathcal{M}}$	Reference data manifold and its reference probability distribution.
m, $c = d - m$, τ	Manifold dimension, codimension, and lower bound on reach.
$r(x)$, $z_x = \Pi_{\mathcal{M}}(x)$	Distance to the manifold and nearest-point projection.
u_x	Unit normal $(x - z_x)/r(x)$ inside the reach tube.
K, ν, Ω	Initialization support, initialization distribution, and invariant prior superlevel set.
r_0, $r_{\min} = \sigma_{\min}\sqrt{c}$	Analysis-tube radius and lower-cutoff transition radius.
C_{loc}	Constant controlling the precision-normalized posterior second moment.
ε_0	Normal localization bound $C_{\mathrm{loc}} m \sigma_{\min}^2/(2\tau)$.
κ_0	Curvature/contraction ratio $C_{\mathrm{loc}} m r_0/(2c\tau)$.
C_x	Covariance of component scores, with $\nabla^2 \log p_{MN} = C_x - \beta I$.
$r_{\mathrm{eff}}(C)$	Effective rank $\mathrm{tr}(C)/\lambda_{\max}(C)$.
α_d, R_d	Trace and effective-rank constants in the off-tube instability condition.
Regional coverage	
r_{tr}, r_{stop}	Transition and stopping thresholds for normal distance.
τ_{tr}, τ_{stop}	Corresponding first entrance times.
X_{stop}, Z_{tr}, Z_{stop}	Stopped ambient output and manifold projections at transition/stop.
$d_{\mathcal{M}}$, $U^{-\delta}$	Intrinsic distance and δ-safe interior of a manifold region.
δ_{stop}	Bound on projected tangential motion during refinement.
p_σ^Q	Gaussian smoothing $Q_{\mathcal{M}} * \mathcal{N}(0, \sigma^2 I)$.
A_N	Positive flux normalization $h\sum_i \sigma_i^2$.
$\mathcal{R}_N$, ϵ_{Pois}	Source–sink residual and its half-L^1 norm.
ϵ_{init}	Initialization mismatch $\mathrm{TV}(\nu, p^Q_{\sigma_{\max}} dx)$.
Y, T_{mark}	Auxiliary marked point and marking time on the original trajectory.
ζ	Failure probability used in the Gaussian placement argument.
a_d, a_m, b_c	Gaussian norm thresholds used in the coverage proof.
δ_{noise}, δ_{cov}	Projection error from fine-scale noise and total coverage resolution.
$Q_k(a)$	a-quantile of χ_k^2, distinct from $Q_{\mathcal{M}}$.
W_1	Wasserstein-1 distance.
Finite-sample non-memorization	
$\mathcal{A} = a_0 + V$	Affine data subspace and its tangent linear subspace.
P_V, $P_{V^\perp}$	Orthogonal projections onto V and its complement.
ℓ, L	Training-grid spacing and half-width parameter of the affine construction.
D_L, $C_{\mathbf{k}}$	Reference box and sample-resolution output cores.
$h_{\mathrm{cov},M}$	Geometric covering radius of the finite training set.
$s = \sigma_{\min}/\ell$	Dimensionless lower bandwidth used in the affine construction and E3.
h_M	Additional posterior-moment residual for finite-sample refinements.
$\varepsilon_{\mathcal{M},h}$	Residual-aware normal localization radius.
Population and control	
p^{pop}	Unified population multi-scale density.
π	Distribution or finite measure over noise scales in the population objective.
E_y, λ	Differentiable task penalty and its guidance strength.
F_y	Penalized objective $F_y = \log p_{MN} - \lambda E_y$.
$\Omega_y(x_0)$, $\mathcal{R}_y$	Guided superlevel set and its relevant manifold-tube region.
δ_y	Outward normal guidance budget relative to the prior contraction scale.
ε_y	Resulting guided normal-attraction radius.

A Detailed Setup, Manifold Convergence, and Representation Geometry

This appendix gives the detailed version of Section 2. The formulas here use the exact empirical score $\nabla \log p_{MN}$; Appendix I proves their population counterparts. The manifold hypotheses locate its limiting points; they do not define or identify a learned manifold. The full convergence statement is Theorem A.14, and optimizer localization is Proposition A.15.

A.1 The empirical objective and posterior identities

We use the training centers and positive finite scale schedule of Section 2.1. All gradients are with respect to the sample x, and all vector norms are Euclidean. For reference,

$$\widehat{Q}_M = \frac{1}{M}\sum_j \delta_{\mu_j}, \qquad \widehat{p}_\sigma(x) = \frac{1}{M}\sum_j \mathcal{N}(x;\mu_j,\sigma^2 I), \tag{A.1}$$

$$p_{MN}(x) = \frac{\sum_{j,i}\sigma_i^2 \mathcal{N}(x;\mu_j,\sigma_i^2 I)}{Z_{MN}}, \qquad Z_{MN} = M\sum_i \sigma_i^2. \tag{A.2}$$

The analysis conditions on this finite training set; no independent-sampling assumption is needed. The reference manifold $\mathcal{M}$ describes its geometry, and the distribution $Q_{\mathcal{M}}$ introduced for coverage specifies the desired regional probabilities.

For $1 \le m < d$ and $c = d - m$, define

$$w_{ij}(x) := \frac{\sigma_i^2 \mathcal{N}(x;\mu_j,\sigma_i^2 I)}{\sum_{\ell,k}\sigma_k^2 \mathcal{N}(x;\mu_\ell,\sigma_k^2 I)}, \tag{A.3}$$

$$a_{ij}(x) := \frac{\mu_j - x}{\sigma_i^2}, \qquad \nabla \log p_{MN}(x) = \sum_{j,i} w_{ij}(x) a_{ij}(x), \tag{A.4}$$

$$\beta(x) := \sum_{j,i}\frac{w_{ij}(x)}{\sigma_i^2}, \qquad \widetilde{w}_{ij}(x) := \frac{w_{ij}(x)}{\sigma_i^2\beta(x)}, \qquad \bar{\mu}(x) := \sum_{j,i}\widetilde{w}_{ij}(x)\mu_j. \tag{A.5}$$

The precision-normalized weights equal the normalized Gaussian densities, whereas w_{ij} includes the additional scale weight. Direct substitution gives

$$\nabla \log p_{MN}(x) = \beta(x)(\bar{\mu}(x) - x), \qquad \sigma_{\max}^{-2} \le \beta(x) \le \sigma_{\min}^{-2}. \tag{A.6}$$

In particular every critical point satisfies

$$x = \bar{\mu}(x) \in \operatorname{conv}(\mathcal{S}_M). \tag{A.7}$$

We study

$$\dot{x} = \nabla \log p_{MN}(x). \tag{A.8}$$

The time change $u(t) = \int_0^t \beta(x(v))\,dv$ is increasing and unbounded, and gives $dx/du = \bar{\mu}(x) - x$. Thus this normalized continuous flow has the same paths and hitting locations; its Euler discretization can have different numerical errors.

A.2 Learning the gradient of the unified log density

The identity below retains the weights in the noise-unconditioned objective of Zhou et al. (2026):

$$\mathcal{L}_u(\theta) = \frac{1}{2MN}\sum_{j,i}\sigma_i^2 \mathbb{E}_{X\sim\mathcal{N}(\mu_j,\sigma_i^2 I)}\|s_\theta(X) - a_{ij}(X)\|^2, \tag{A.9}$$

$$= \frac{Z_{MN}}{2MN}\mathbb{E}_{X\sim p_{MN}}\|s_\theta(X) - \nabla \log p_{MN}(X)\|^2 + C. \tag{A.10}$$

Proof of (4). At fixed x, $w_{ij}(x)$ are the normalized loss coefficients and their component-score average is $\nabla \log p_{MN}(x)$. The variance decomposition gives

$$\sum_{j,i} w_{ij}\|s_\theta - a_{ij}\|^2 = \|s_\theta - \nabla \log p_{MN}\|^2 + \sum_{j,i} w_{ij}\|a_{ij} - \nabla \log p_{MN}\|^2.$$

Multiplication by $Z_{MN}p_{MN}(x)/(2MN)$ and integration proves the identity. The second term is independent of θ and integrable, being bounded by the component-score second moment. The unrestricted regression optimum is therefore $\nabla \log p_{MN}$, uniquely up to a set of p_{MN}-measure zero. □

Empirical, population, and learned scores. The three fields play distinct roles. For a fixed training set, (A.10) shows that the exact regression target of the empirical DSM objective is the empirical unified score $\nabla \log p_{MN}$. Replacing the empirical center average by the population distribution gives the population score $\nabla \log p^{\mathrm{pop}}$, defined in Appendix I.1. A learned network differs from the population field through the decomposition

$$s_\theta(x) - \nabla \log p^{\mathrm{pop}}(x) = \big[s_\theta(x) - \nabla \log p_{MN}(x)\big] + \big[\nabla \log p_{MN}(x) - \nabla \log p^{\mathrm{pop}}(x)\big].$$

The first term is the learned-to-empirical target error, while the second is the empirical-to-population target discrepancy. For the unrestricted empirical regression problem, the former is measured exactly by the excess DSM objective:

$$\mathcal{L}_u(\theta) - \inf_s \mathcal{L}_u(s) = \frac{Z_{MN}}{2MN}\mathbb{E}_{X\sim p_{MN}}\|s_\theta(X) - \nabla \log p_{MN}(X)\|^2.$$

Positive smoothing also gives useful intuition for the second term. At fixed positive cutoffs, both p_{MN} and p^{pop} are formed from averages of the same smooth Gaussian kernels, differing in whether the center distribution is empirical or population. Thus, when the training centers are sampled from the population, concentration of the smoothed densities and their spatial derivatives naturally suggests closeness of the corresponding score fields on regions where the smoothed density is not too small. This connection becomes more delicate toward vanishing bandwidth, where individual training centers can dominate the empirical field.

For the learned dynamics, a simple deterministic relation further explains why score approximation is relevant. Along a learned flow $\dot{x} = s_\theta(x)$,

$$\frac{d}{dt}\log p_{MN}(x(t)) \geq \frac{1}{2}\|\nabla \log p_{MN}(x(t))\|^2 - \frac{1}{2}\|s_\theta(x(t)) - \nabla \log p_{MN}(x(t))\|^2.$$

Hence sufficiently accurate local score approximation preserves the ascent mechanism. Turning the average DSM error above into a trajectory-wise guarantee, however, requires additional control of the learned field on the states visited during generation.

This separation between an exact target score and its learned neural approximation is also standard in score-based diffusion theory. For example, existing convergence analyses assume an L^2-accurate score estimator and derive sampling guarantees conditional on that error (Chen et al., 2023; Lee et al., 2023). Our geometric results therefore analyze the exact empirical and population target fields, while an end-to-end theory connecting finite-data neural training, score approximation, and the resulting autonomous trajectories is a complementary direction for future work.

A.3 Global existence and convergence

Let $K \subset \mathbb{R}^d$ be nonempty and compact, and let the initialization distribution ν be supported on K. Define

$$p_K := \min_{x\in K} p_{MN}(x) > 0, \qquad \Omega := \{x : p_{MN}(x) \geq p_K\}. \tag{A.11}$$

The set Ω is compact because the mixture is positive and continuous, and $p_{MN}(x) \to 0$ as $\|x\| \to \infty$.

Lemma A.1 (Global existence, finite length, and convergence)**.** *Every trajectory initialized in K exists for all $t \geq 0$, remains in Ω, has finite length, and converges to a critical point $x_\infty \in \Omega$ of $\log p_{MN}$.*

Proof. **A compact invariant region.** The positive finite Gaussian mixture and its logarithm are real analytic. Local existence and uniqueness follow from local Lipschitz continuity of the gradient. Along a solution,

$$\frac{d}{dt}\log p_{MN}(x(t)) = \|\nabla \log p_{MN}(x(t))\|^2 \geq 0. \tag{A.12}$$

Hence it stays in Ω. The vector field is bounded there, so a finite maximal existence time would give a limit in Ω from which the solution could be continued. The solution is therefore global.

Critical accumulation points. The objective increases to a finite value ℓ_∞, and

$$\int_0^\infty \|\nabla \log p_{MN}(x(t))\|^2 dt = \ell_\infty - \log p_{MN}(x(0)) < \infty. \tag{A.13}$$

The integrand has bounded derivative on Ω, since both the gradient and Hessian are bounded. Uniform continuity and integrability imply that it tends to zero. Thus every accumulation point is critical with value ℓ_∞.

Finite path length. The analytic gradient inequality (Absil et al., 2005), made uniform over the compact accumulation set by a finite cover, gives constants $C_L > 0$ and $0 < \vartheta < 1$ such that, for all sufficiently large t,

$$e(t)^\vartheta \le C_L \|\nabla \log p_{MN}(x(t))\|, \qquad e(t) = \ell_\infty - \log p_{MN}(x(t)). \tag{A.14}$$

The trajectory eventually lies in that cover; otherwise it would have an accumulation point outside it. If e reaches zero, the trajectory is stationary thereafter. Otherwise,

$$-\frac{d}{dt} e(t)^{1-\vartheta} \ge \frac{1-\vartheta}{C_L} \|\dot{x}(t)\|,$$

and hence

$$\int_T^\infty \|\dot{x}(t)\| dt \le \frac{C_L}{1-\vartheta} e(T)^{1-\vartheta} < \infty. \tag{A.15}$$

The initial segment also has finite length. The trajectory is Cauchy and converges to one of its critical accumulation points. □

Remark A.2 (Scope of the flow result). Applying the lemma with $K = \{x_0\}$ proves forward completeness and convergence for every $x_0 \in \mathbb{R}^d$. Denote the flow by $\phi_t(x_0) = x(t; x_0)$. No manifold assumption is needed for this fact. Geometric assumptions below locate the limiting point; they are not needed to establish the existence of a limit.

Gaussian initialization and confidence thresholds. For $G_k \sim \mathcal{N}(0, I_k)$, let

$$Q_k(a) = \inf\{q : \mathbb{P}(\|G_k\|^2 \le q) \ge a\}, \qquad 0 < a < 1, \tag{A.16}$$

be the a-quantile of χ^2_k. These scalar quantiles are distinct from the reference distribution $Q_{\mathcal{M}}$. For $0 < \delta_{\text{init}} < 1$, the shell

$$\Omega_{\text{init}} = \{x : \sigma_{\max}\sqrt{Q_d(\delta_{\text{init}}/2)} \le \|x\| \le \sigma_{\max}\sqrt{Q_d(1-\delta_{\text{init}}/2)}\} \tag{A.17}$$

has probability $1 - \delta_{\text{init}}$ under the centered Gaussian. The compact-support theorem applies to

$$\nu = \mathcal{N}(0, \sigma^2_{\max} I_d) \mid \Omega_{\text{init}}, \qquad K = \Omega_{\text{init}}. \tag{A.18}$$

A probability-one conclusion for this truncated initialization gives probability at least $1 - \delta_{\text{init}}$ for the untruncated Gaussian.

A.4 Local geometry and distance-matched scales

Definition A.3 (Reach)**.** For a closed set $A \subset \mathbb{R}^d$, its reach is the supremum of radii $r \ge 0$ such that every point at distance less than r from A has a unique nearest point in A.

Assumption A.4 (Local manifold geometry and available scales)**.** The manifold $\mathcal{M}$ is compact, embedded, C^2, and without boundary, with $\operatorname{reach}(\mathcal{M}) \ge \tau > 0$. Choose

$$0 < r_0 < \tau, \qquad r_0 \le \sigma_{\max}\sqrt{c}. \tag{A.19}$$

Write $r(x) = \operatorname{dist}(x, \mathcal{M})$. The unique nearest point $z_x = \Pi_{\mathcal{M}}(x)$ is defined on the entire reach tube $r(x) < \tau$; the smaller r_0-tube is where we impose posterior estimates. Write $T_z\mathcal{M}$ for the tangent space, $N_z\mathcal{M} = (T_z\mathcal{M})^\perp$ for the normal space, and $P_{T_z\mathcal{M}}, P_{N_z\mathcal{M}}$ for their orthogonal projectors. For $0 < r(x) < \tau$, let $u_x = (x - z_x)/r(x)$: this is the unit normal pointing from the nearest manifold point toward x, not a globally chosen orientation.

Reach controls both bending and the approach of distinct branches. The r_0-tube is the region $r(x) < r_0$. The smaller region $\varepsilon_{\mathcal{M}} < r(x) < r_0$, after its inner radius is defined below, will be called the intermediate contraction region. Both are restricted to the invariant region Ω whenever a posterior estimate is used. Within the reach tube the nearest-point projection is C^1, and $\nabla r(x) = u_x$ away from $\mathcal{M}$ (Leobacher & Steinicke, 2021, Theorem 2). These facts allow us to separate motion toward the manifold from motion along it.

The distance-matched scale. In a locally flat population patch, write $\mu = z_x + Ut$ with $U^\top U = I_m$ and $U^\top(x - z_x) = 0$. Integrating the tangent Gaussian gives

$$\int_{\mathcal{M}} \mathcal{N}(x; \mu, \sigma^2 I) q_{\mathcal{M}}(\mu) d\,\mathrm{vol}_{\mathcal{M}}(\mu) \approx q_{\mathcal{M}}(z_x)(2\pi\sigma^2)^{-c/2} e^{-r^2/(2\sigma^2)}. \tag{A.20}$$

The raw mixture and precision-normalized posterior therefore have profiles

$$H_r(\sigma) = \sigma^{2-c} e^{-r^2/(2\sigma^2)}, \qquad G_r(\sigma) = \sigma^{-c} e^{-r^2/(2\sigma^2)}. \tag{A.21}$$

Their logarithmic derivatives are respectively $[r^2 - (c-2)\sigma^2]/\sigma^3$ and $[r^2 - c\sigma^2]/\sigma^3$. Thus the precision profile peaks at

$$\sigma_{\mathrm{match}}(r) = r/\sqrt{c}, \qquad r_{\min} = \sigma_{\min}\sqrt{c}, \tag{A.22}$$

with lower-cutoff saturation when $r \le r_{\min}$. A mode of this scale profile motivates, but does not prove, the finite-posterior moment bound below. Appendix E.1 computes both the mode and the moment for flat reference models.

Assumption A.5 (Scale-dependent posterior localization)**.** There is a dimensionless constant $C_{\mathrm{loc}} > 0$ such that

$$\sum_{j,i} \widetilde{w}_{ij}(x) \|\mu_j - z_x\|^2 \le C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r(x)^2/c\}, \qquad x \in \Omega, \quad 0 < r(x) < r_0. \tag{A.23}$$

The substantive requirement is a controlled constant throughout the relevant tube and scale range. For fixed finite data and $\sigma_{\min} > 0$, compactness always allows the uninformative value $\mathrm{diam}(\mathcal{M})^2/(m\sigma_{\min}^2)$. The flat model in Appendix E.1 instead gives an order-one constant in a specified high-codimension regime.

A.5 From posterior localization to normal attraction

Lemma A.6 (Positive-reach normal inequality)**.** *Under Assumption A.4, for $z, \mu \in \mathcal{M}$ and every unit normal $u \in N_z\mathcal{M}$,*

$$\langle u, \mu - z \rangle \le \frac{\|\mu - z\|^2}{2\tau}. \tag{A.24}$$

Proof. For $0 < a < \tau$, the normal-tube property gives $\Pi_{\mathcal{M}}(z + au) = z$ (Leobacher & Steinicke, 2021). Hence

$$a^2 \le \|z + au - \mu\|^2 = a^2 - 2a\langle u, \mu - z\rangle + \|\mu - z\|^2.$$

Rearrange and let $a \uparrow \tau$. □

Lemma A.7 (Posterior alignment)**.** *Under Assumptions A.4 and A.5,*

$$\langle u_x, \bar{\mu}(x) - z_x \rangle \le \frac{C_{\mathrm{loc}} m}{2\tau} \max\{\sigma_{\min}^2, r(x)^2/c\}, \qquad x \in \Omega, \quad 0 < r(x) < r_0. \tag{A.25}$$

Proof. Average (A.24) over the nonnegative precision-normalized weights, then apply (A.23):

$$\langle u_x, \bar{\mu}(x) - z_x \rangle \le \frac{1}{2\tau} \sum_{j,i} \widetilde{w}_{ij}(x) \|\mu_j - z_x\|^2 \le \frac{C_{\mathrm{loc}} m}{2\tau} \max\{\sigma_{\min}^2, r(x)^2/c\}.$$

□

Define the geometric constants:

$$\varepsilon_0 := \frac{C_{\mathrm{loc}} m}{2\tau}\sigma_{\min}^2, \qquad \kappa_0 := \frac{C_{\mathrm{loc}} m r_0}{2c\tau}, \\ \varepsilon_{\mathcal{M}} := \min\{\varepsilon_0, r_{\min}\}. \tag{A.26}$$

Assumption A.8 (Contraction gap and consistent tube)**.** We require

$$\kappa_0 < 1, \tag{A.27}$$

$$\varepsilon_{\mathcal{M}} < r_0. \tag{A.28}$$

The gap asks that the curvature-induced normal bias be weaker than the direct inward displacement. It is favored by large reach and small m/c, provided C_{loc} is controlled.

The standing conditions also imply $\varepsilon_{\mathcal{M}} = \varepsilon_0 < r_{\min}$. Indeed,

$$\varepsilon_0 = \frac{C_{\mathrm{loc}} m}{2\tau}\sigma_{\min}^2 = \kappa_0 \frac{r_{\min}^2}{r_0}.$$

If $\varepsilon_0 \geq r_{\min}$, division by $r_{\min} > 0$ would give $r_{\min} \geq r_0/\kappa_0 > r_0$, and then $\varepsilon_{\mathcal{M}} = r_{\min} > r_0$, a contradiction. The minimum notation records the two radial regimes, but creates no extra admissible regime. The quantity ε_0 is an upper bound on endpoint distance, not a lower bound on the actual error or a rate of convergence in time.

Lemma A.9 (Strict inward score)**.** *Under the three local assumptions, along the flow in Ω whenever $0 < r(t) < r_0$,*

$$\dot{r}(t) \leq \begin{cases} -\beta(x(t))(r(t) - \varepsilon_0), & 0 < r(t) \leq r_{\min}, \\ -\beta(x(t))(1 - \kappa_0) r(t), & r_{\min} < r(t) < r_0. \end{cases} \tag{A.29}$$

Here and below $r(t) := r(x(t))$. In particular,

$$\langle x - z_x, \nabla \log p_{MN}(x) \rangle < 0 \quad \textit{for } x \in \Omega, \quad \varepsilon_{\mathcal{M}} < r(x) < r_0. \tag{A.30}$$

There is no critical point in this intermediate contraction region. For every $\rho \in (\varepsilon_{\mathcal{M}}, r_0)$, a trajectory in Ω that enters $\{r \leq \rho\}$ cannot subsequently leave it.

Proof. The distance gradient is u_x in the reach tube, so

$$\dot{r} = \beta(x)\big(\langle u_x, \bar{\mu}(x) - z_x \rangle - r\big). \tag{A.31}$$

For $r \leq r_{\min}$, posterior alignment bounds the first term by ε_0. For $r > r_{\min}$, it bounds it by $C_{\mathrm{loc}} m r^2/(2c\tau) = \kappa_0 r^2/r_0 < \kappa_0 r$. These are the two radial inequalities. As $\varepsilon_{\mathcal{M}} = \varepsilon_0 < r_{\min}$, both give $\dot{r} < 0$ for $\varepsilon_{\mathcal{M}} < r < r_0$, excluding critical points there. On $r = \rho$ the derivative points strictly inward. At a hypothetical first exit from $r \leq \rho$, continuity would instead require a nonnegative outward derivative. This contradiction proves invariance. □

A.6 Off-tube instability and saddle avoidance

Local contraction leaves one global question: could ascent stop at an attracting critical point outside the tube? Define

$$C_x := \mathrm{Cov}_{w(x)}(a_{ij}(x)). \tag{A.32}$$

Explicitly, since $\sum_{j,i} w_{ij}(x) a_{ij}(x) = \nabla \log p_{MN}(x)$,

$$C_x = \sum_{j,i} w_{ij}(x)(a_{ij}(x) - \nabla \log p_{MN}(x))(a_{ij}(x) - \nabla \log p_{MN}(x))^\top \succeq 0.$$

For a nonzero positive semidefinite matrix, use $r_{\mathrm{eff}}(C) = \mathrm{tr}(C)/\lambda_{\max}(C)$.

Assumption A.10 (Off-tube shell-rank geometry)**.** There are $\alpha_d \in (0, 1]$ and $R_d > 0$, with $R_d < \alpha_d d$, such that every critical point $x \in \Omega$ with $r(x) \geq r_0$ satisfies

$$\mathrm{tr}(C_x) \geq \alpha_d d\, \beta(x), \qquad r_{\mathrm{eff}}(C_x) \leq R_d. \tag{A.33}$$

The trace bound ensures $C_x \neq 0$. It supplies enough score variation; the rank bound concentrates that variation in sufficiently few directions to create an ascent-unstable direction.

Lemma A.11 (Hessian decomposition of the unified log density)**.** *For every x,*

$$\nabla^2 \log p_{MN}(x) = C_x - \beta(x) I. \tag{A.34}$$

Proof. Differentiating the normalized component weights gives $\nabla w_{ij} = w_{ij}(a_{ij} - \nabla \log p_{MN})$, while $D a_{ij} = -\sigma_i^{-2} I$. Therefore

$$\nabla^2 \log p_{MN} = \sum_{j,i} w_{ij} a_{ij} a_{ij}^\top - (\nabla \log p_{MN})(\nabla \log p_{MN})^\top - \beta I = C_x - \beta I.$$

□

Lemma A.12 (Instability of off-tube critical points)**.** *Under Assumption A.10, every off-tube critical point in Ω satisfies*

$$\lambda_{\max}(\nabla^2 \log p_{MN}(x)) \geq \left(\frac{\alpha_d d}{R_d} - 1 \right) \beta(x) > 0. \tag{A.35}$$

It is therefore not an ambient local maximum of $\log p_{MN}$.

Proof. The assumed trace is positive. By the effective-rank bound,

$$\lambda_{\max}(C_x) = \frac{\operatorname{tr} C_x}{r_{\text{eff}}(C_x)} \geq \frac{\alpha_d d}{R_d} \beta(x).$$

Subtracting βI yields (A.35). A positive Hessian eigenvalue at a critical point gives both an exponentially growing linearized direction for ascent and a direction of positive second variation. The point cannot be a local maximum. □

A sufficient trace mechanism. At a critical point, $\operatorname{tr} C_x = \sum_{j,i} w_{ij} \|x - \mu_j\|^2 / \sigma_i^4$. If an active set carries at least $(1 - \zeta_d)\beta(x)$ precision mass and satisfies $\|x - \mu_j\|^2 \geq (1 - \delta_d) d \sigma_i^2$, then $\operatorname{tr} C_x \geq (1 - \zeta_d)(1 - \delta_d) d \beta(x)$. Effective rank is a separate requirement. The sphere model verifies both ingredients at its off-tube center.

Why unstable endpoints have zero probability. A positive eigenvalue of $\nabla^2 \log p_{MN}$ does not exclude initialization on a stable set. We therefore use concrete sufficient conditions for the exact flow: nondegenerate off-tube critical points and an absolutely continuous initialization. The following lemma proves the required avoidance, rather than assuming the behavior of the limiting point. The limit map is Borel measurable because $x_\infty(x_0) = \lim_{n \to \infty} \phi_n(x_0)$ and each ϕ_n is continuous, so all endpoint events below are measurable.

Lemma A.13 (Almost-sure avoidance of off-tube critical points)**.** *Suppose Assumption A.10 holds, every point in*

$$\mathcal{C}_{\text{off}} := \{ x \in \Omega : \nabla \log p_{MN}(x) = 0,\ r(x) \geq r_0 \} \tag{A.36}$$

is hyperbolic, and ν is absolutely continuous with respect to ambient Lebesgue measure Leb_d, written $\nu \ll \text{Leb}_d$: every zero-volume set has ν-probability zero. Then

$$\mathbb{P}_{X_0 \sim \nu}[r(x_\infty) \geq r_0] = 0. \tag{A.37}$$

For this gradient flow, hyperbolicity means that $\nabla^2 \log p_{MN}$ at the critical point has no zero eigenvalue. The condition is imposed only at off-tube critical points, not at the limiting points inside the tube.

Proof. **Finitely many unstable equilibria.** The off-tube critical set is closed in compact Ω. A nonsingular Hessian makes each point isolated by the inverse function theorem. A compact set of such isolated critical points is finite. Each has a positive Hessian eigenvalue by Lemma A.12, so its local stable manifold has dimension less than d and zero ambient volume (Hirsch et al., 1970).

Pulling stable sets back to initialization. The derivative of the finite-time flow satisfies $\dot{J} = \nabla^2 \log p_{MN}(\phi_t(x_0))J$, $J(0) = I$, and hence

$$\det D\phi_t(x_0) = \exp\left(\int_0^t \Delta \log p_{MN}(\phi_u(x_0))du\right) > 0.$$

Uniqueness makes ϕ_t injective, and it is a diffeomorphism onto its image with locally Lipschitz inverse. A trajectory converging to an equilibrium eventually stays in its local stable neighborhood. Thus

$$W^s(x_*) \subseteq \bigcup_{n\geq 0} \phi_n^{-1}(W^s_{\mathrm{loc}}(x_*)). \tag{A.38}$$

Each preimage is null, using a countable cover by neighborhoods on which the inverse is Lipschitz. The countable union over n and finite union over equilibria remain null. Absolute continuity of ν gives the conclusion. □

Here, hyperbolicity provides a convenient sufficient condition for almost-sure avoidance of off-tube equilibria.

A.7 Manifold convergence and optimizer localization

Theorem A.14 (Almost-sure manifold-tube convergence)**.** *Let K, ν, Ω be as in* (A.11)*. Suppose Assumptions A.4, A.5, A.8, and A.10 hold. In addition, suppose all off-tube critical points in Ω are hyperbolic and $\nu \ll \mathrm{Leb}_d$. Then every trajectory from K converges to a critical point,*

$$x(t) \longrightarrow x_\infty \in \Omega, \qquad \nabla \log p_{MN}(x_\infty) = 0,$$

and its limit satisfies

$$\mathrm{dist}(x_\infty, \mathcal{M}) \leq \varepsilon_{\mathcal{M}} = \min\left\{\frac{C_{\mathrm{loc}} m}{2\tau}\sigma^2_{\min}, \sigma_{\min}\sqrt{c}\right\} \qquad \nu\text{-almost surely.} \tag{A.39}$$

Under the standing consistency conditions, $\varepsilon_{\mathcal{M}} = \varepsilon_0$, so the distance bound is $C_{\mathrm{loc}} m \sigma^2_{\min}/(2\tau)$. The conclusion is relative to this chosen reference manifold; it does not identify it uniquely or assert convergence to a global maximizer.

Proof. Every trajectory converges to a critical point by Lemma A.1. Lemma A.9 excludes a critical limit with $\varepsilon_{\mathcal{M}} < r < r_0$, and Lemma A.13 excludes $r \geq r_0$ almost surely. Thus $r(x_\infty) \leq \varepsilon_{\mathcal{M}}$. The consistency calculation above gives $\varepsilon_{\mathcal{M}} = \varepsilon_0$. □

Proposition A.15 (Location of local and global maximizers)**.** *Under the local and off-tube geometry assumptions of the theorem, without hyperbolicity or an absolute-continuity requirement on initialization, every ambient local maximizer $x^\star \in \Omega$ of $\log p_{MN}$ satisfies $r(x^\star) \leq \varepsilon_{\mathcal{M}}$. Moreover,*

$$\varnothing \neq \operatorname*{argmax}_{x\in\mathbb{R}^d} \log p_{MN}(x) \subseteq \{x : r(x) \leq \varepsilon_{\mathcal{M}}\}. \tag{A.40}$$

Proof. A local maximum has zero gradient and a negative semidefinite Hessian. The inward-score inequality rules out $\varepsilon_{\mathcal{M}} < r < r_0$; off-tube instability rules out $r \geq r_0$. A global maximum exists because p_{MN} is positive and tends to zero at infinity. Its value is at least p_K, so it belongs to Ω and satisfies the same bound. □

This proposition localizes the maximizers of the fixed objective; trajectory-level convergence is addressed separately by Theorem A.14.

Remark A.16 (Eventual entry, Gaussian starts, and sampling resolution). For every $\rho \in (\varepsilon_{\mathcal{M}}, r_0)$, almost every trajectory reaches $\{r \leq \rho\}$ in finite time and then remains there. This follows from convergence to a strictly smaller radius and the inward boundary condition. If the theorem holds under the conditional Gaussian initialization (A.18), the untruncated Gaussian satisfies

$$\mathbb{P}[r(x_\infty) \leq \varepsilon_{\mathcal{M}}] \geq 1 - \delta_{\mathrm{init}}. \tag{A.41}$$

Indeed, let $E = \{X_0 \in \Omega_{\text{init}}\}$ and $A = \{r(x_\infty) \le \varepsilon_{\mathcal{M}}\}$. The unconditional flow has a measurable limit for every initialization, and $\mathbb{P}(E) = 1 - \delta_{\text{init}}$. Applying the theorem to $P_{X_0|E}$ gives $\mathbb{P}(A \mid E) = 1$. Thus

$$\mathbb{P}(A) \ge \mathbb{P}(A \cap E) = \mathbb{P}(A \mid E)\mathbb{P}(E) = 1 - \delta_{\text{init}}.$$

For fixed finite data, a controlled posterior-localization constant need not persist as $\sigma_{\min} \to 0$. Sample-associated modes may develop under separation and smallest-scale dominance. Such modes can still be close to $\mathcal{M}$: memorization concerns allocation along the manifold, not a contradiction to normal localization.

Corollary A.17 (Vanishing tube radius)**.** *For a family of finite mixtures with centers on a fixed $\mathcal{M}$, with fixed m, c, τ, r_0, suppose the theorem's assumptions hold for each member, on its corresponding Ω and under its initialization distribution. Then, for each member,*

$$r(x_\infty^{(\sigma_{\min})}) \le \frac{C_{\text{loc}}(\sigma_{\min})m}{2\tau}\sigma_{\min}^2 < \frac{c}{r_0}\sigma_{\min}^2 \quad \textit{almost surely}. \tag{A.42}$$

Proof. For each model the contraction gap gives $C_{\text{loc}}(\sigma_{\min})m/(2\tau) < c/r_0$. Multiply by $\sigma_{\min}^2$ and use Theorem A.14. The probability-one conclusion is for each model separately. □

Finite empirical data can require a nonzero posterior-moment residual. Appendix G.1 propagates that residual through both the normal and tangential bounds. The zero-residual convergence result above does not assert uniform validity at arbitrarily small noise for a fixed finite cloud.

A.8 Affine representations and the meaning of larger reach

Proposition A.18 (Exact affine normal convergence)**.** *Suppose the finite set of training centers lies in a closed affine subspace $\mathcal{A} = a_0 + V$. Write $z_x = \Pi_{\mathcal{A}}(x)$, $r_{\mathcal{A}}(x) = \|x - z_x\|$, and let $P_{V^\perp}$ be the orthogonal projector onto $V^\perp$. Then*

$$P_{V^\perp}\nabla \log p_{MN}(x) = -\beta(x)(x - z_x), \qquad r_{\mathcal{A}}(x(t)) \le r_{\mathcal{A}}(x(0))e^{-t/\sigma_{\max}^2}. \tag{A.43}$$

Every critical point, and hence every limiting point, belongs to $\mathcal{A}$.

Proof. Since every training center lies in $\mathcal{A}$, so does $\bar{\mu}(x)$. With $n(t) = P_{V^\perp}(x(t) - a_0) = x(t) - \Pi_{\mathcal{A}}(x(t))$, the barycenter identity gives $\dot{n} = -\beta(x(t))n$. Consequently

$$\|n(t)\| = \|n(0)\|e^{-\int_0^t \beta(x(u))du} \le \|n(0)\|e^{-t/\sigma_{\max}^2}.$$

This identity also holds for $n(0) = 0$. At any critical point it forces $n = 0$. Global finite-mixture convergence then puts every endpoint in $\mathcal{A}$. □

The endpoint is also in $\text{conv}(\mathcal{S}_M)$, by (A.7). This normal result alone says nothing about allocation within the affine manifold. After proving regional coverage, Corollary 3.2 returns to this geometry and proves training-set separation together with coverage at the training resolution for a verified finite-cloud family.

Lemma A.19 (Reach controls the local Gaussian geometry)**.** *For $x = z_x + ru_x$ with $0 < r < \tau$ and every $\mu \in \mathcal{M}$,*

$$\begin{aligned} r^2 + (1 - r/\tau)\|\mu - z_x\|^2 &\le \|x - \mu\|^2 \\ &\le r^2 + (1 + r/\tau)\|\mu - z_x\|^2. \end{aligned} \tag{A.44}$$

Proof. Applying the reach inequality to both u_x and $-u_x$ gives $|\langle u_x, \mu - z_x\rangle| \le \|\mu - z_x\|^2/(2\tau)$. Insert this into $\|x - \mu\|^2 = r^2 + \|\mu - z_x\|^2 - 2r\langle u_x, \mu - z_x\rangle$. □

Remark A.20 (Geometry and representation). The relevant quantities are C_{loc}/τ, r_0/τ and the operating noise scales relative to reach. The distance sandwich connects curvature to Gaussian weights, but sample coverage and the scale schedule also affect their normalized moment. In the sphere model a valid comparison is $C_{\text{loc}} = \tau C_{\text{flat}}/(\tau - r_0)$, giving $\varepsilon_0 = C_{\text{flat}} m\sigma_{\min}^2/[2(\tau -$

$r_0)]$. These bounds favor locally near-affine, well-covered representations with separated branches. Uniform rescaling changes all lengths equally; nonlinear encoders additionally change the noise metric. Representation comparisons therefore require consistent normalization and a decoder-aware data interpretation.

B DIVERSITY: RETENTION AND REGIONAL COVERAGE

The convergence theorem locates individual endpoints. To obtain a generative guarantee, we must also relate the initialized trajectories to probabilities across the data manifold. We combine a geometric bound on refinement with a source–sink identity for the same empirical field.

B.1 REFERENCE DISTRIBUTION AND OUTPUT TIME

Keep the connected reference manifold $\mathcal{M}$ fixed and specify a probability distribution $Q_{\mathcal{M}}$ on it. If it has a density relative to manifold volume, write

$$Q_{\mathcal{M}}(U) = \int_U q_{\mathcal{M}}(z)\, d\,\mathrm{vol}_{\mathcal{M}}(z). \tag{B.1}$$

A density is not required by the coverage proof. Intrinsic distance $d_{\mathcal{M}}$ is the infimum of lengths of piecewise smooth curves on $\mathcal{M}$; it dominates Euclidean distance. For a Borel region U and $\delta \geq 0$, define

$$U^{-\delta} = \{z \in U : d_{\mathcal{M}}(z, \mathcal{M} \setminus U) > \delta\}. \tag{B.2}$$

Distance to the empty set is infinite. This safe interior protects region membership against a displacement of at most δ.

Choose the normal-distance thresholds

$$\varepsilon_{\mathcal{M}} = \varepsilon_0 < r_{\mathrm{stop}} < r_{\min} < r_{\mathrm{tr}} < r_0 < \tau, \qquad r_{\min} = \sigma_{\min}\sqrt{c}. \tag{B.3}$$

The transition radius starts the refinement analysis; the stopping radius specifies the returned sample. They are not radii of regions along the manifold. Define

$$\tau_{\mathrm{tr}} = \inf\{t \geq 0 : r(x(t)) \leq r_{\mathrm{tr}}\}, \qquad Z_{\mathrm{tr}} = \Pi_{\mathcal{M}}(x(\tau_{\mathrm{tr}})), \tag{B.4}$$

$$\tau_{\mathrm{stop}} = \inf\{t \geq 0 : r(x(t)) \leq r_{\mathrm{stop}}\}, \qquad Z_{\mathrm{stop}} = \Pi_{\mathcal{M}}(x(\tau_{\mathrm{stop}})), \tag{B.5}$$

$$X_{\mathrm{stop}} = x(\tau_{\mathrm{stop}}). \tag{B.6}$$

Theorem A.14 makes these times finite almost surely. On exceptional infinite-time events, assign a fixed point of $\mathcal{M}$ to the corresponding outputs. Strict inward crossing makes entrance into $r \leq \rho$ equal to the infimum of times with $r < \rho$. For $a > 0$, $\{\tau_\rho < a\} = \bigcup_{q \in \mathbb{Q} \cap [0,a)} \{r(\phi_q(X_0)) < \rho\}$, which proves measurability. Implementing this theoretical stop requires the reference distance or a justified surrogate.

B.2 TANGENTIAL MOTION DURING REFINEMENT

Lemma B.1 (Projected tangential velocity)**.** *Under the local geometry and posterior-localization assumptions, for* $0 < r(x(t)) \leq r_{\mathrm{tr}} < r_0$ *in* Ω*,*

$$\left\| \frac{d}{dt} \Pi_{\mathcal{M}}(x(t)) \right\| \leq L_\Pi \beta(x(t)) \sqrt{C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r(x(t))^2/c\}}, \qquad L_\Pi = \frac{\tau}{\tau - r_{\mathrm{tr}}}. \tag{B.7}$$

Proof. **Projection distortion.** The reach inequality at the projections z_x, z_y of two points in the tube gives

$$\langle x - y, z_x - z_y \rangle \geq \left(1 - \frac{r(x) + r(y)}{2\tau}\right) \|z_x - z_y\|^2.$$

Apply Cauchy–Schwarz, then take $y = x + hv$ and let $h \to 0$. The projection is C^1 on the reach tube (Leobacher & Steinicke, 2021), so

$$\|D\Pi_{\mathcal{M}}(x)\| \leq (1 - r(x)/\tau)^{-1} \leq L_\Pi. \tag{B.8}$$

Removing normal motion. Projection is constant along a normal fiber, hence $D\Pi_{\mathcal{M}}(x)(x - z_x) = 0$. The barycenter identity gives $\dot{z} = \beta D\Pi_{\mathcal{M}}(x)(\bar{\mu} - z)$. Jensen's inequality bounds $\|\bar{\mu} - z\|^2$ by the posterior second moment in (A.23). Combining the two estimates proves the result. □

Proposition B.2 (Explicit retention under early stopping)**.** *Under Theorem A.14 and* (B.3)*, almost surely,*

$$r(X_{\mathrm{stop}}) \le r_{\mathrm{stop}}, \qquad d_{\mathcal{M}}(Z_{\mathrm{tr}}, Z_{\mathrm{stop}}) \le \delta_{\mathrm{stop}}, \tag{B.9}$$

where

$$\begin{aligned}\delta_{\mathrm{stop}} = L_\Pi \frac{\sqrt{C_{\mathrm{loc}} m/c}}{1-\kappa_0}(r_{\mathrm{tr}} - r_{\mathrm{min}}) \\ + L_\Pi \sqrt{C_{\mathrm{loc}} m}\, \sigma_{\mathrm{min}} \log \frac{r_{\mathrm{min}} - \varepsilon_0}{r_{\mathrm{stop}} - \varepsilon_0}.\end{aligned} \tag{B.10}$$

The same bound controls the entire projected refinement path:

$$\int_{\tau_{\mathrm{tr}}}^{\tau_{\mathrm{stop}}} \left\| \frac{d}{dt} \Pi_{\mathcal{M}}(x(t)) \right\| dt \le \delta_{\mathrm{stop}}. \tag{B.11}$$

Proof. Convergence to radius at most $\varepsilon_0 < r_{\mathrm{stop}}$ implies finite entrance. If the entrance times coincide, the path length is zero. Otherwise inward invariance and first entrance imply $r_{\mathrm{stop}} < r(t) \le r_{\mathrm{tr}}$ during refinement, with strictly decreasing radius. Dividing the projected-speed bound by the radial contraction bound gives

$$\frac{\|\dot{z}\|}{-\dot{r}} \le \begin{cases} L_\Pi \sqrt{C_{\mathrm{loc}} m}\, \sigma_{\mathrm{min}}/(r-\varepsilon_0), & r \le r_{\mathrm{min}}, \\ L_\Pi \sqrt{C_{\mathrm{loc}} m/c}/(1-\kappa_0), & r > r_{\mathrm{min}}. \end{cases}$$

Integrate over the decreasing radius, extending the entrance radius up to r_{tr} if necessary. The constant integral on $[r_{\mathrm{min}}, r_{\mathrm{tr}}]$ and logarithmic integral on $[r_{\mathrm{stop}}, r_{\mathrm{min}}]$ give (B.10). Intrinsic endpoint distance is bounded by curve length. □

B.3 The source–sink identity and a point on the same path

Let

$$p_\sigma^Q(x) = \int_{\mathcal{M}} \mathcal{N}(x; z, \sigma^2 I) Q_{\mathcal{M}}(dz). \tag{B.12}$$

This is the density of $Z + \sigma G$ for independent $Z \sim Q_{\mathcal{M}}$ and $G \sim \mathcal{N}(0, I_d)$. It is a reference smoothing, rather than the distribution of the optimizer at a fixed time. Take $N \ge 2$, $\sigma_i = \sigma_{\mathrm{min}} e^{(i-1)h}$, $h > 0$, and define

$$A_N = h \sum_i \sigma_i^2, \qquad A_N p_{MN} = h \sum_i \sigma_i^2 \widehat{p}_{\sigma_i}. \tag{B.13}$$

Proposition B.3 (Multi-scale source–sink identity)**.** *For positive finite cutoffs,*

$$\Delta \int_{\sigma_{\mathrm{min}}}^{\sigma_{\mathrm{max}}} \sigma p_\sigma^Q \, d\sigma = p_{\sigma_{\mathrm{max}}}^Q - p_{\sigma_{\mathrm{min}}}^Q. \tag{B.14}$$

For the actual finite mixture, let

$$\mathcal{R}_N = A_N \Delta p_{MN} - (p_{\sigma_{\mathrm{max}}}^Q - p_{\sigma_{\mathrm{min}}}^Q), \qquad \epsilon_{\mathrm{Pois}} = \tfrac{1}{2} \|\mathcal{R}_N\|_1. \tag{B.15}$$

Then $\mathcal{R}_N$ is integrable with zero integral and

$$\nabla \cdot (A_N p_{MN} \nabla \log p_{MN}) = A_N \Delta p_{MN} = p_{\sigma_{\mathrm{max}}}^Q - p_{\sigma_{\mathrm{min}}}^Q + \mathcal{R}_N. \tag{B.16}$$

Proof. Differentiating a Gaussian gives $\partial_\sigma p_\sigma^Q = \sigma \Delta p_\sigma^Q$. Integration in σ proves the first identity. Compact reference support and positive finite cutoffs justify differentiation under the integral. The second follows from $p_{MN} \nabla \log p_{MN} = \nabla p_{MN}$. Gaussian Laplacians are integrable with zero integral, and both reference smoothings integrate to one. These facts also prove the residual assertions. □

Here Δp_{MN} is the density Laplacian; in particular,

$$\Delta p_{MN} = p_{MN} \big(\Delta \log p_{MN} + \|\nabla \log p_{MN}\|^2 \big). \tag{B.17}$$

The identity describes stationary flux along a fixed field. It does not prescribe time marginals. Related divergence-based constructions include Poisson flow and Beckmann transport (Xu et al., 2022; Cheuk-Kit et al., 2026).

For distributions P, Q, use

$$\mathrm{TV}(P,Q) = \sup_U |P(U) - Q(U)|, \tag{B.18}$$

which equals half the L^1 distance when densities exist. For measurable $0 \le f \le 1$,

$$\left| \int f\, dP - \int f\, dQ \right| \le \mathrm{TV}(P,Q). \tag{B.19}$$

Consequently applying the same conditional sampling kernel to two input distributions cannot increase TV, even when the input is retained in the output. Define the initialization mismatch by

$$\epsilon_{\mathrm{init}} = \mathrm{TV}(\nu, p^Q_{\sigma_{\max}}(x)dx). \tag{B.20}$$

Lemma B.4 (A fine-scale reference point on the score path)**.** *Let $\mathcal{R}^{\pm}_N = \max\{\pm\mathcal{R}_N, 0\}$ and set $\lambda_{\mathrm{mark}} = (p^Q_{\sigma_{\min}} + \mathcal{R}^-_N)/(A_N p_{MN})$. For an independent $E \sim \mathrm{Exp}(1)$, define on the original flow*

$$T_{\mathrm{mark}} = \inf\left\{ t \ge 0 : \int_0^t \lambda_{\mathrm{mark}}(\phi_u(X_0))du \ge E \right\}, \qquad Y = \phi_{T_{\mathrm{mark}}}(X_0). \tag{B.21}$$

For every start, T_{mark} is finite almost surely, and under $X_0 \sim \nu$,

$$\mathrm{TV}(P_Y, p^Q_{\sigma_{\min}}(x)dx) \le \epsilon_{\mathrm{Pois}} + \epsilon_{\mathrm{init}}. \tag{B.22}$$

The proof in Appendix C.1 identifies $A_N p_{MN}$ as the occupation density for a corrected nonnegative source and sink. It requires bounded reference support, so it also applies to the affine construction. The auxiliary mark leaves the trajectory and returned output unchanged. The next argument places it in the refinement segment where geometry controls its relation to the output.

B.4 From reference probability to output coverage

Choose $\zeta \in (0,1)$ and the Gaussian norm thresholds

$$a_d = \sqrt{Q_d(1-\zeta/3)}, \qquad a_m = \sqrt{Q_m(1-\zeta/3)}, \qquad b_c = \sqrt{Q_c(\zeta/3)}. \tag{B.23}$$

When $\sigma_{\min} a_d < \tau$, put

$$\delta_{\mathrm{noise}} = \frac{\sigma_{\min} a_m}{1 - \sigma_{\min} a_d/\tau}, \qquad \delta_{\mathrm{cov}} = \delta_{\mathrm{noise}} + \delta_{\mathrm{stop}}. \tag{B.24}$$

Lemma B.5 (Gaussian projection and placement in the refinement segment)**.** *For fixed $z \in \mathcal{M}$ and $y = z + \sigma_{\min} G$, consider $\mathcal{G}_z = \{\|G\| \le a_d,\ \|P_{T_z\mathcal{M}} G\| \le a_m,\ \|P_{N_z\mathcal{M}} G\| \ge b_c\}$. Its probability is at least $1-\zeta$, and on this event*

$$d_{\mathcal{M}}(z, \Pi_{\mathcal{M}}(y)) \le \delta_{\mathrm{noise}}, \qquad \sigma_{\min} b_c - \frac{\delta^2_{\mathrm{noise}}}{2\tau} \le r(y) \le \sigma_{\min} a_d. \tag{B.25}$$

Thus the conditions

$$\sigma_{\min} a_d \le r_{\mathrm{tr}}, \qquad r_{\mathrm{stop}} < \sigma_{\min} b_c - \frac{\delta^2_{\mathrm{noise}}}{2\tau} \tag{B.26}$$

place the perturbation between the two stopping surfaces.

Appendix C.2 proves the bounds by adding normal noise first, then projecting the segment that adds tangent noise. The parameter ζ affects both the probability and the admissible radii; it cannot be reduced arbitrarily with these radii fixed.

Theorem B.6 (Convergence and regional coverage)**.** *Under Proposition B.2, the geometric schedule and* (B.26)*, the exact empirical flow converges with $r(x_\infty) \le \varepsilon_{\mathcal{M}}$ almost surely, and returns X_{stop} with $r(X_{\mathrm{stop}}) \le r_{\mathrm{stop}}$. For every Borel $U \subseteq \mathcal{M}$,*

$$\mathbb{P}_\nu[Z_{\mathrm{tr}} \in U, Z_{\mathrm{stop}} \in U] \ge (1-\zeta) Q_{\mathcal{M}}(U^{-\delta_{\mathrm{cov}}}) - \epsilon_{\mathrm{Pois}} - \epsilon_{\mathrm{init}}. \tag{B.27}$$

For $z \in \mathrm{supp}(Q_{\mathcal{M}})$ and $\rho > \delta_{\mathrm{cov}}$,

$$\mathbb{P}_\nu[Z_{\mathrm{stop}} \in B_{\mathcal{M}}(z,\rho)] \ge (1-\zeta) Q_{\mathcal{M}}(B_{\mathcal{M}}(z, \rho - \delta_{\mathrm{cov}})) - \epsilon_{\mathrm{Pois}} - \epsilon_{\mathrm{init}}. \tag{B.28}$$

The latter also bounds $\mathbb{P}_\nu[\|X_{\mathrm{stop}} - z\| < r_{\mathrm{stop}} + \rho]$.

Proof. **A reference label for the actual path.** Construct Y and attach a label Z with the kernel

$$K_Q(y, dz) = \frac{\mathcal{N}(y; z, \sigma_{\min}^2 I)}{p^Q_{\sigma_{\min}}(y)} Q_{\mathcal{M}}(dz). \tag{B.29}$$

Call the augmented distribution $\mathbb{P}_{\mathrm{act}}$. Compare it with $\mathbb{P}_{\mathrm{id}}(dz, dy) = Q_{\mathcal{M}}(dz)\mathcal{N}(y; z, \sigma_{\min}^2 I)dy$. Both use the same kernel given Y, so the mark lemma and TV contraction give, for every joint event D,

$$|\mathbb{P}_{\mathrm{act}}((Z, Y) \in D) - \mathbb{P}_{\mathrm{id}}(D)| \le \epsilon_{\mathrm{Pois}} + \epsilon_{\mathrm{init}}. \tag{B.30}$$

Only under $\mathbb{P}_{\mathrm{id}}$ is the reference label exactly distributed as $Q_{\mathcal{M}}$.

A reference event with controlled probability. Let $\mathcal{G}$ be the Gaussian inequalities in Lemma B.5, expressed using $(Y - Z)/\sigma_{\min}$. Under the ideal distribution its conditional probability is at least $1 - \zeta$ for every $Z = z$. Thus

$$\mathbb{P}_{\mathrm{act}}[Z \in U^{-\delta_{\mathrm{cov}}}, \mathcal{G}] \ge (1 - \zeta) Q_{\mathcal{M}}(U^{-\delta_{\mathrm{cov}}}) - \epsilon_{\mathrm{Pois}} - \epsilon_{\mathrm{init}}. \tag{B.31}$$

All geometric implications of $\mathcal{G}$ are deterministic and therefore also hold for the actual labeled trajectory.

Following the same mark to the output. On this event, $r_{\mathrm{stop}} < r(Y) \le r_{\mathrm{tr}}$ and $d_{\mathcal{M}}(Z, \Pi_{\mathcal{M}}(Y)) \le \delta_{\mathrm{noise}}$. First entrance and inward invariance imply $\tau_{\mathrm{tr}} \le T_{\mathrm{mark}} < \tau_{\mathrm{stop}}$. The full projected path-length bound controls both subsegments adjacent to the mark, so each of Z_{tr} and Z_{stop} is within δ_{cov} of Z. Safe-interior membership puts both in U. Taking probabilities proves (B.27). Auxiliary labeling has not changed either output distribution.

For the ball version, a label in $B_{\mathcal{M}}(z, \rho - \delta_{\mathrm{cov}})$ gives $d_{\mathcal{M}}(z, Z_{\mathrm{stop}}) < \rho$. Finally, $\|X_{\mathrm{stop}} - z\| \le r_{\mathrm{stop}} + d_{\mathcal{M}}(Z_{\mathrm{stop}}, z)$. This yields both remaining assertions. □

B.5 Noise, sampling error and coverage resolution

For distributions with finite first moments, define

$$W_1(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} \int \|u - v\| \, \gamma(du, dv), \tag{B.32}$$

where $\Gamma(P, Q)$ denotes joint distributions with marginals P, Q. Let $B = \sup_{z \in \mathrm{supp}(Q_{\mathcal{M}})} \|z\|$.

Proposition B.7 (Probability-error decomposition and generic input bounds)**.** *Define the finite-grid error*

$$\epsilon_{\mathrm{grid}} := \frac{1}{2} \left\| A_N \Delta p_{MN} - \left(\widehat{p}_{\sigma_{\max}} - \widehat{p}_{\sigma_{\min}} \right) \right\|_1 \tag{B.33}$$

and the cutoff mismatch

$$\epsilon_{\mathrm{cut}} := \mathrm{TV}\left(\widehat{p}_{\sigma_{\max}} dx, p^Q_{\sigma_{\max}} dx \right) + \mathrm{TV}\left(\widehat{p}_{\sigma_{\min}} dx, p^Q_{\sigma_{\min}} dx \right). \tag{B.34}$$

Then

$$\epsilon_{\mathrm{Pois}} \le \epsilon_{\mathrm{grid}} + \epsilon_{\mathrm{cut}}. \tag{B.35}$$

For the finite geometric schedule, these terms admit the generic bounds

$$\epsilon_{\mathrm{grid}} \le dh \log \frac{\sigma_{\max}}{\sigma_{\min}} + \frac{h}{2} \sqrt{2d}, \tag{B.36}$$

$$\epsilon_{\mathrm{cut}} \le \frac{W_1(\widehat{Q}_M, Q_{\mathcal{M}})}{\sqrt{2\pi}} \left(\sigma_{\min}^{-1} + \sigma_{\max}^{-1} \right). \tag{B.37}$$

Consequently,

$$\begin{aligned} \epsilon_{\mathrm{Pois}} \le dh \log \frac{\sigma_{\max}}{\sigma_{\min}} + \frac{h}{2} \sqrt{2d} \\ + \frac{W_1(\widehat{Q}_M, Q_{\mathcal{M}})}{\sqrt{2\pi}} \left(\sigma_{\min}^{-1} + \sigma_{\max}^{-1} \right). \end{aligned} \tag{B.38}$$

For the truncated Gaussian (A.18)*,*

$$\epsilon_{\mathrm{init}} \le \delta_{\mathrm{init}} + \frac{\int \|z\| \, Q_{\mathcal{M}}(dz)}{\sqrt{2\pi}\sigma_{\max}} \le \delta_{\mathrm{init}} + \frac{B}{\sqrt{2\pi}\sigma_{\max}}. \tag{B.39}$$

Appendix C.3 proves this decomposition by first comparing the finite scale sum with its continuous log-scale integral and then comparing the empirical and reference smoothings only at the two cutoffs. Thus finite-sample mismatch enters through smoothed endpoint distributions rather than throughout the entire scale path.

The explicit bounds in (B.36)–(B.37) are distribution-free sufficient estimates. In particular, the $O(dh\log(\sigma_{\max}/\sigma_{\min}))$ grid term comes from uniform L^1 control of Gaussian scale derivatives and can be conservative in high ambient dimension. The coverage theorem itself depends on the actual errors ϵ_{Pois} and ϵ_{init}, for which structured models may admit substantially sharper control. These generic bounds are useful for making the scale tradeoffs explicit: decreasing $\sigma_{\min}$ improves geometric resolution but can increase finite-data cutoff mismatch, whereas increasing $\sigma_{\max}$ improves initialization overlap while requiring finer scale discretization.

Stopping policies. For $a > 1$, $0 < b < 1$, set $r_{\mathrm{tr}} = a\sigma_{\min}\sqrt{c} < r_0$ and $r_{\mathrm{stop}} = b\sigma_{\min}\sqrt{c}$, with $\varepsilon_0 \le (b/2)\sigma_{\min}\sqrt{c}$. Substitution into the retention bound gives

$$\delta_{\mathrm{stop}} \le \frac{\sqrt{C_{\mathrm{loc}}m}\sigma_{\min}}{1 - a\sigma_{\min}\sqrt{c}/\tau}\left[\frac{a-1}{1-\kappa_0} + \log\frac{2}{b}\right]. \tag{B.40}$$

Thus the retention scale is $O(\sqrt{m}\,\sigma_{\min})$ when the constants and geometric margins remain controlled. The Gaussian band conditions can hold by choosing $a > a_d/\sqrt{c}$, $b < b_c/\sqrt{c}$, and a sufficiently small relative noise scale. For $r_{\mathrm{stop}} = 2\varepsilon_0 < r_{\min}$ instead,

$$r(X_{\mathrm{stop}}) \le \frac{C_{\mathrm{loc}}m}{\tau}\sigma_{\min}^2, \tag{B.41}$$

$$\delta_{\mathrm{stop}} \le \frac{\sqrt{C_{\mathrm{loc}}m}\sigma_{\min}}{1 - a\sigma_{\min}\sqrt{c}/\tau}\left[\frac{a-1}{1-\kappa_0} + \log\frac{2\tau\sqrt{c}}{C_{\mathrm{loc}}m\sigma_{\min}}\right]. \tag{B.42}$$

This follows from $r_{\min} - \varepsilon_0 \le r_{\min}$ in the logarithm. It gives quadratic normal accuracy at a logarithmically larger retention scale. In either policy, coverage additionally requires the stated Gaussian band and a probability-error budget smaller than the reference region mass. The general endpoint needs further tangential control (Appendix G.3); the affine construction in Appendix D supplies it at the training resolution.

C SOURCE–SINK PROOFS AND ERROR BOUNDS

C.1 A MARKED TRAJECTORY AND ITS OCCUPATION DENSITY

Proof of Lemma B.4. **Complete trajectories and finite marking.** With $B_M = \max_j \|\mu_j\|$, the barycenter identity gives $\|\nabla \log p_{MN}(x)\| \le \sigma_{\min}^{-2}(B_M + \|x\|)$. The score and its negative have at most linear growth, so Gronwall's inequality gives a complete flow ϕ_t for all real t. It is a smooth diffeomorphism with inverse ϕ_{-t}. Each forward path stays in its compact positive superlevel set. The continuous marking rate is strictly positive on that set, hence its integral along the path diverges and $T_{\mathrm{mark}} < \infty$ almost surely. Joint continuity of the integrated rate ensures measurability.

Nonnegative source and sink. For this proof put $a = p^Q_{\sigma_{\max}} + \mathcal{R}_N^+$ and $b = p^Q_{\sigma_{\min}} + \mathcal{R}_N^-$. Both have integral $1 + \epsilon_{\mathrm{Pois}}$, since the positive and negative parts of the residual have the same integral ϵ_{Pois}. The flux identity becomes

$$\nabla \cdot (A_N p_{MN} \nabla \log p_{MN}) = a - b, \qquad \lambda_{\mathrm{mark}} A_N p_{MN} = b. \tag{C.1}$$

These are unnormalized source and sink densities of equal mass.

Occupation density. Let $\mathsf{T}_t f$ transport an integrable nonnegative density f along the score flow while removing each trajectory at its mark. The inverse-flow Jacobian and conditional survival probability give

$$(\mathsf{T}_t f)(y) = f(\phi_{-t}(y)) \exp\left[- \int_0^t (\Delta \log p_{MN} + \lambda_{\mathrm{mark}})(\phi_{-u}(y)) du \right]. \tag{C.2}$$

Consequently,

$$\|\mathsf{T}_t f\|_1 = \int f(x_0) e^{-\int_0^t \lambda_{\mathrm{mark}}(\phi_u(x_0)) du} dx_0 \le \|f\|_1. \tag{C.3}$$

Differentiate the characteristic formula for $f = A_N p_{MN}$. The product rule and (C.1) give $\frac{d}{dt}\mathsf{T}_t(A_N p_{MN}) = -\mathsf{T}_t a$. The rate is continuous, which suffices to differentiate the time integral; derivatives of $\mathcal{R}_N^-$ are not needed. Integrating gives

$$A_N p_{MN} - \mathsf{T}_t(A_N p_{MN}) = \int_0^t \mathsf{T}_u a \, du. \tag{C.4}$$

The right-hand side increases pointwise. As each mark is finite, dominated convergence in (C.3) gives $\|\mathsf{T}_t(A_N p_{MN})\|_1 \to 0$. Monotone convergence therefore implies

$$\int_0^\infty \mathsf{T}_t a \, dt = A_N p_{MN} \quad \text{almost everywhere.} \tag{C.5}$$

Distribution of the marked point. The mark kernel is

$$\mathsf{K}(x_0, U) = \int_0^\infty \mathbf{1}_U(\phi_t(x_0)) \lambda_{\mathrm{mark}}(\phi_t(x_0)) e^{-\int_0^t \lambda_{\mathrm{mark}}(\phi_u(x_0)) du} dt.$$

It has total mass one. Tonelli's theorem, a flow change of variables and the occupation identity give

$$\int a(x_0) \mathsf{K}(x_0, U) dx_0 = \int_U \lambda_{\mathrm{mark}}(y) \int_0^\infty (\mathsf{T}_t a)(y) dt \, dy = \int_U b(y) dy.$$

In measure notation this is $(a\,dx)\mathsf{K} = b\,dx$. Subtract the residual terms:

$$(p^Q_{\sigma_{\max}} dx)\mathsf{K} - p^Q_{\sigma_{\min}} dx = \mathcal{R}_N^- dx - (\mathcal{R}_N^+ dx)\mathsf{K}.$$

Both measures on the right have mass ϵ_{Pois}, so their difference on every event has magnitude at most that value. Finally, TV contracts under the same kernel, yielding

$$\mathrm{TV}(P_Y, p^Q_{\sigma_{\min}} dx) \le \mathrm{TV}(\nu, p^Q_{\sigma_{\max}} dx) + \epsilon_{\mathrm{Pois}}.$$

This proves the claim for the unchanged initialized trajectories. □

C.2 Projection of a Gaussian perturbation

Proof of Lemma B.5. Write $G_T = P_{T_z\mathcal{M}}G$ and $G_N = P_{N_z\mathcal{M}}G$. The squared norms of G, G_T, G_N have the indicated chi-squared distributions. Each excluded tail has probability $\zeta/3$, so their union has probability at most ζ.

Tangential projection error. On the good event the segment $\gamma(v) = z + \sigma_{\min}G_N + v\sigma_{\min}G_T$, $0 \le v \le 1$, stays in the reach tube because $\|\gamma(v) - z\| \le \sigma_{\min}a_d < \tau$. Its initial projection is z, and its final projection is $q = \Pi_{\mathcal{M}}(y)$. By the projection derivative bound,

$$d_{\mathcal{M}}(z, q) \le \int_0^1 \|D\Pi_{\mathcal{M}}(\gamma(v))\sigma_{\min}G_T\|dv \le \frac{\sigma_{\min}a_m}{1 - \sigma_{\min}a_d/\tau} = \delta_{\text{noise}}.$$

Normal placement. The reach inequality for all unit normals at z gives $\|P_{N_z\mathcal{M}}(q - z)\| \le \|q - z\|^2/(2\tau)$. Hence

$$r(y) \ge \|\sigma_{\min}G_N - P_{N_z\mathcal{M}}(q - z)\| \ge \sigma_{\min}b_c - \delta_{\text{noise}}^2/(2\tau).$$

The upper bound is $r(y) \le \|y - z\| \le \sigma_{\min}a_d$. The band conditions place y in the refinement segment. □

C.3 Finite-grid, finite-sample and initialization errors

Proof of Proposition B.7. **Log-scale quadrature.** Write $\varphi_\sigma = \mathcal{N}(\cdot; 0, \sigma^2 I)$ and $k_\sigma = \sigma^2\Delta\varphi_\sigma = (\|x\|^2/\sigma^2 - d)\varphi_\sigma$. For $t = \log(\sigma/\sigma_{\min})$ and $u = \|x\|^2/\sigma^2$,

$$\|k_\sigma\|_1 \le \sqrt{2d}, \qquad \partial_t k_\sigma = [(u - d)^2 - 2u]\varphi_\sigma, \qquad \|\partial_t k_\sigma\|_1 \le 4d.$$

The first bound uses the variance $2d$ of $\|G\|^2$; the last uses its centered second moment and mean. Convolution with $\widehat{Q}_M$ does not increase L^1 norms. Thus $f(t) = \sigma^2\Delta\widehat{p}_\sigma$ is $4d$-Lipschitz in L^1. Set $b = \log(\sigma_{\max}/\sigma_{\min}) = (N - 1)h$. The scale identity gives $\int_0^b f(t)dt = \widehat{p}_{\sigma_{\max}} - \widehat{p}_{\sigma_{\min}}$. Comparing each integral cell with its left endpoint costs at most $2dh^2$ in L^1. The extra upper endpoint of the original mixture contributes at most $h\sqrt{2d}$. Therefore

$$\tfrac{1}{2}\|A_N\Delta p_{MN} - (\widehat{p}_{\sigma_{\max}} - \widehat{p}_{\sigma_{\min}})\|_1 \le dhb + \tfrac{h}{2}\sqrt{2d}. \tag{C.6}$$

Empirical versus reference smoothings. For any centers u, v, integration along their joining segment gives

$$\|\varphi_\sigma(\cdot - u) - \varphi_\sigma(\cdot - v)\|_1 \le \|u - v\| \sup_{\|e\|=1} \|\partial_e\varphi_\sigma\|_1 = \sqrt{2/\pi}\,\|u - v\|/\sigma.$$

Average over a coupling and take the infimum:

$$\mathrm{TV}(\widehat{p}_\sigma dx, p_\sigma^Q dx) \le \frac{W_1(\widehat{Q}_M, Q_{\mathcal{M}})}{\sqrt{2\pi}\sigma}. \tag{C.7}$$

Decompose the residual as

$$\mathcal{R}_N = \Big[A_N\Delta p_{MN} - (\widehat{p}_{\sigma_{\max}} - \widehat{p}_{\sigma_{\min}})\Big] + (\widehat{p}_{\sigma_{\max}} - p^Q_{\sigma_{\max}}) - (\widehat{p}_{\sigma_{\min}} - p^Q_{\sigma_{\min}}).$$

Using $\mathrm{TV}(P, Q) = \frac{1}{2}\|p - q\|_1$ for densities and the triangle inequality gives

$$\epsilon_{\text{Pois}} \le \epsilon_{\text{grid}} + \epsilon_{\text{cut}}.$$

Equation (C.6) gives (B.36), while applying (C.7) at the two cutoffs gives (B.37). Combining the two proves (B.38).

Gaussian initialization. Conditioning a distribution on an event of mass $1 - \delta_{\text{init}}$ changes TV by exactly δ_{init}. The same translation estimate gives

$$\mathrm{TV}(\mathcal{N}(0, \sigma_{\max}^2 I), p^Q_{\sigma_{\max}}dx) \le \frac{\int \|z\|Q_{\mathcal{M}}(dz)}{\sqrt{2\pi}\sigma_{\max}} \le \frac{B}{\sqrt{2\pi}\sigma_{\max}}.$$

Add the truncation cost. For an untruncated Gaussian, this cost is zero. □

Continuous-scale specializations. For a bounded log-continuous mixture, A_N is replaced by $\int_{\sigma_{\min}}^{\sigma_{\max}} \sigma \, d\sigma$. The quadrature error vanishes; when its center distribution is also $Q_{\mathcal{M}}$, the entire residual vanishes. Compact center support and positive finite cutoffs ensure the same analyticity, Gaussian decay and complete-flow properties. The local geometry must still hold for that instance.

C.4 Why the scale weight matters

For a positive C^1 weight $w(\sigma)$ aggregated against $d\sigma/\sigma$, integration by parts gives

$$\Delta \int_{\sigma_{\min}}^{\sigma_{\max}} w(\sigma) p_\sigma^Q \frac{d\sigma}{\sigma} = \left[\frac{w(\sigma)}{\sigma^2} p_\sigma^Q \right]_{\sigma_{\min}}^{\sigma_{\max}} - \int_{\sigma_{\min}}^{\sigma_{\max}} \left(\frac{w(\sigma)}{\sigma^2} \right)' p_\sigma^Q \, d\sigma.$$

The choice $w(\sigma) \propto \sigma^2$ removes intermediate-scale terms and yields the endpoint balance used in the proof. Other weights still define fixed objectives, but this identity generally has additional source or sink terms. When integrating against $d\sigma$ instead, the corresponding weight is proportional to σ.

D ENDPOINT NON-MEMORIZATION AT THE TRAINING-SET RESOLUTION

This appendix proves Corollary 3.2 using finite-mixture convergence, exact affine normal contraction, and the source–sink marked-distribution lemma. Its main ingredients are a finite-grid sign estimate at the training resolution and a coupling that preserves regional probability. The reference box has compact support, as required by the marked-distribution argument; the containing affine space may be unbounded.

D.1 CONSTRUCTION AND PROBABILITY COUPLING

Use the finite grid, reference box D_L, and output cores $C_{\mathbf{k}}$ of Section 3.2. Directly from the largest one-dimensional gap and the outer boundary distance,

$$h_{\mathrm{cov},M} = \sup_{z \in D_L} \operatorname{dist}(z, \mathcal{S}_M) = \sqrt{m}\,\ell/3, \tag{D.1}$$

and the cores are

$$C_{\mathbf{k}} = \{(u,0) : |u_j - \ell k_j| \le \ell/12 \text{ for all } j\}. \tag{D.2}$$

The proof verifies one finite relative schedule on the full interval

$$\sigma_{\min} \in [s_0\ell, 2s_0\ell], \qquad \sigma_i = \sigma_{\min} e^{(i-1)h}, \quad N \ge 2. \tag{D.3}$$

Every trajectory converges, with

$$x_\infty \in \mathcal{A} \cap \operatorname{conv}(\mathcal{S}_M), \tag{D.4}$$

by the finite-mixture and affine results. We prove that the covered endpoints additionally satisfy

$$\mathbb{P}[\operatorname{dist}(x_\infty, \mathcal{S}_M) \ge h_{\mathrm{cov},M}/4] \ge 1 - \epsilon. \tag{D.5}$$

Additional consequences of the main corollary. At every covered output cell, the center marginal obeys

$$\max_j \sum_i \widetilde{w}_{ij}(x_\infty) \le \left(1 + e^{-1/(36 s_0^2)}\right)^{-m} < 0.529^m \quad \text{on } \{x_\infty \in \bigcup_{\mathbf{k}} C_{\mathbf{k}}\}. \tag{D.6}$$

The proof constructs a coupling with $Z \sim Q_{\mathcal{M}}$ such that

$$\mathbb{P}[\|x_\infty - Z\| \le 7h_{\mathrm{cov},M}/4,\ \operatorname{dist}(x_\infty, \mathcal{S}_M) \ge h_{\mathrm{cov},M}/4] \ge 1 - \epsilon. \tag{D.7}$$

It also preserves the reference grid-cell label with that probability. Consequently,

$$W_1(P_{x_\infty}, Q_{\mathcal{M}}) \le 7h_{\mathrm{cov},M}/4 + \epsilon \operatorname{diam}(D_L). \tag{D.8}$$

Thus the coupling controls endpoint discrepancy at the training-resolution scale, with an explicit probability error.

D.2 COUPLING A MARKED PATH TO A RESOLVED REFERENCE CELL

We first record the probability step used below. Work on $\mathcal{A} = \mathbb{R}^m \times \{0\}$, and write $Q_{\mathcal{M}} * \mathcal{N}(0, \sigma^2 I_m)$ for convolution in its *intrinsic* coordinates. The ambient smoothing then factors as this distribution times $\mathcal{N}(0, \sigma^2 I_c)$ in the normal coordinates. This factorization concerns a smoothing, not the time marginal of the actual optimization flow.

Suppose disjoint capture sets $B_{\mathbf{k}} \subseteq \mathcal{A}$ and output sets $C_{\mathbf{k}} \subseteq \mathcal{A}$ have the following verified property: a path through

$$\{x : \Pi_{\mathcal{A}}(x) \in B_{\mathbf{k}},\ r_{\mathcal{A}}(x) \le r_{\mathrm{tr}}\}$$

converges to $C_{\mathbf{k}}$. Let $Z \sim Q_{\mathcal{M}}$, and let its reference cell label be $\mathbf{k}(Z)$. Assume that membership in $B_{\mathbf{k}}$ agrees with this label. Define, only for this comparison,

$$e_{\mathrm{blur}} := \mathrm{TV}\big(Q_{\mathcal{M}} * \mathcal{N}(0, \sigma_{\min}^2 I_m), Q_{\mathcal{M}}\big),$$

$$e_{\mathrm{cell}} := Q_{\mathcal{M}}\left(\mathcal{A} \setminus \bigcup_{\mathbf{k}} B_{\mathbf{k}}\right),$$

$$e_{\mathrm{normal}} := \mathbb{P}[\sigma_{\min}\|G_c\| > r_{\mathrm{tr}}].$$

Lemma B.4 gives a mark Y on the actual path with total-variation error at most $\epsilon_{\mathrm{Pois}} + \epsilon_{\mathrm{init}}$ from the ambient reference smoothing. Couple it maximally to $Y^Q \sim p^Q_{\sigma_{\min}}(x)\,dx$. Then couple $\Pi_{\mathcal{A}}(Y^Q)$ maximally to $Z \sim Q_{\mathcal{M}}$. These constructions can be attached conditionally to the original path; neither changes that path or its endpoint. The event

$$Y = Y^Q, \qquad Z = \Pi_{\mathcal{A}}(Y^Q), \qquad Z \in \bigcup_{\mathbf{k}} B_{\mathbf{k}}, \qquad r_{\mathcal{A}}(Y^Q) \le r_{\mathrm{tr}} \tag{D.9}$$

has probability at least

$$1 - \epsilon_{\mathrm{Pois}} - \epsilon_{\mathrm{init}} - e_{\mathrm{blur}} - e_{\mathrm{cell}} - e_{\mathrm{normal}}. \tag{D.10}$$

On this event $x_\infty \in C_{\mathbf{k}(Z)}$.

A maximal coupling places the common density $\min\{dP/d(P+Q), dQ/d(P+Q)\}$ on the diagonal, giving equality probability $1 - \mathrm{TV}(P, Q)$. Conditional kernels attach these couplings to the existing trajectory. A union bound gives (D.10).

This differs from the Gaussian-posterior reference label in (B.29). There the label comparison pays a spatial Gaussian displacement. Here the reference has an intrinsic density, and its small total-variation change under smoothing permits an equality coupling except on a controlled probability error. Such a comparison is *not* available for an arbitrary reference distribution.

D.3 Periodic comparison and the finite-cloud error

Rescale lengths by ℓ, so the grid spacing is one. Write $s = \sigma_{\min}/\ell$; in these coordinates we again use σ_i for the rescaled noise levels. Scores are multiplied by a positive constant under this rescaling, and the time variable is rescaled by ℓ^2; trajectories and endpoint statements are otherwise unchanged. The following quantities are local to this proof.

Let φ_σ denote the one-dimensional normalized Gaussian and consider the infinite paired grid

$$g_\sigma(u) := \frac{1}{2} \sum_{k \in \mathbb{Z}} \left[\varphi_\sigma(u - k - 1/6) + \varphi_\sigma(u - k + 1/6)\right]. \tag{D.11}$$

This is a periodic comparison function, not a replacement density for the finite sampler. Integrating against $e^{-2\pi i n u}$ over one period, shifting each Gaussian, and summing the intervals gives its Fourier series

$$g_\sigma(u) = 1 + 2 \sum_{n \ge 1} \cos(n\pi/3) e^{-2\pi^2 n^2 \sigma^2} \cos(2\pi n u). \tag{D.12}$$

The Gaussian Fourier integral follows, for example, by differentiating its characteristic function and integrating by parts. Gaussian decay makes the displayed series and its spatial derivatives absolutely and uniformly convergent, so all differentiations below are legitimate.

For $\sigma \ge 1/2$, put $q = e^{-2\pi^2\sigma^2} < 1/100$. Then

$$|g_\sigma - 1| \le \frac{2q}{1-q}, \qquad g_\sigma \ge \tfrac{1}{2}, \qquad \sum_{n \ge 2} n^2 q^{n^2} \le \frac{4q^4}{1-4q} \le \frac{q}{4}.$$

To verify the sum bound, use $n^2 \le 4^{n-1}$ and $n^2 \ge n + 2$ for integers $n \ge 2$. For $0 < t < \pi$, $|\sin(nt)| \le n \sin t$; induction follows from the angle-addition formula. Since $\cos(\pi/3) = 1/2$, the first harmonic therefore controls the derivative:

$$g'_\sigma(u) \le -\pi q \sin(2\pi u) < 0 \quad (0 < u < 1/2), \tag{D.13}$$

with the opposite sign on $(-1/2, 0)$, repeated every period. The maxima of this comparison lie between, not on, the two training positions in each cell. Importantly, every fixed scale has the same coordinate signs.

At a state $x = (u, n)$, the infinite-array fixed-scale sum in $d = m + c$ dimensions is

$$\mathscr{G}_\sigma(u, n) = (2\pi\sigma^2)^{-c/2} e^{-\|n\|^2/(2\sigma^2)} \prod_{j=1}^{m} g_\sigma(u_j).$$

The scale mixture does not generally factor across coordinates; we use factorization only *at a fixed scale*, then sum same-sign contributions. Choose $0 < \theta \le 1/12$ and use the Gaussian threshold a_d already defined in (B.23). For $\|n\| \le sa_d$, $1/12 \le u_j - k_j \le 1/2 - \theta$, and arbitrary other coordinates, the derivative of $h\sum_i \sigma_i^2 \mathscr{G}_{\sigma_i}$ is negative with magnitude at least

$$A(s,\theta) := \frac{\pi \log 2}{2^{m+c+1}(2\pi)^{c/2}} s^{2-c} e^{-a_d^2/2 - 4\pi^2 s^2} \sin(2\pi\theta) > 0, \tag{D.14}$$

provided $\sigma_{\max}/s \ge 2$ and $h \le (\log 2)/4$. Indeed, all scales have the required sign. Restrict the sum to $s \le \sigma_i \le \sqrt{2}s$; its total h-weight is at least $(\log 2)/4$. On that band use $\sigma_i^2 \ge s^2$, $\sigma_i^{-c} \ge (2s)^{-c}$, the normal lower bound $e^{-a_d^2/2}$, $q \ge e^{-4\pi^2 s^2}$, and $\prod_{\ell \ne j} g_{\sigma_i} \ge 2^{-(m-1)}$. Also $\sin(2\pi(u_j - k_j)) \ge \sin(2\pi\theta)$. Reflection gives the positive margin for negative coordinate offsets. For $c \ge 2$, $A(s,\theta)$ decreases with s.

We next control the error caused by truncating the training array. For $d > 1$, a geometric grid with $h \le 1$ satisfies

$$h\sum_i \sigma_i^{-d} e^{-r^2/(2\sigma_i^2)} \le K_d r^{-d}, \qquad K_d := 2^{d/2-1}\Gamma(d/2) + 2d^{d/2}e^{-d/2}. \tag{D.15}$$

To see this, extend the positive sum to the whole geometric grid. The log-scale integrand is unimodal. Its Riemann sum is at most its integral plus $2h$ times its maximum, by summing the variation over grid intervals. The substitution $v = r^2/(2\sigma^2)$ gives the integral $2^{d/2-1}\Gamma(d/2)r^{-d}$; the maximum is $d^{d/2}e^{-d/2}r^{-d}$.

Let $J \ge \max\{1, \sqrt{m}/3\}$ be an integer padding width. Suppose $|k_j| \le L - J$, $|u_j - k_j| \le 1/2 - \theta$, and $\|n\| \le sa_d$. Every omitted training position has tangential distance at least J from u. The error in any coordinate derivative of the scaled actual density $(2L+1)^m A_N p_{MN}$, compared with $h\sum_i \sigma_i^2 \mathscr{G}_{\sigma_i}$, is at most

$$C_{m,d} J^{1-c}, \qquad C_{m,d} := (2\pi)^{-d/2} K_d (3/2)^{m+d-1} \frac{2^{c-1}|S^{m-1}|}{c-1}. \tag{D.16}$$

Here $|S^{m-1}| = 2\pi^{m/2}/\Gamma(m/2)$; in particular $|S^0| = 2$. This bound is uniform in the upper cutoff and in the number of scales. It is the reason for the sufficient condition $c \ge 2$ in this construction, not a claim of impossibility for codimension one.

For the details, differentiating one weighted Gaussian gives $\partial_{u_j}(\sigma^2 \mathcal{N}(x;\mu,\sigma^2 I)) = (\mu_j - u_j)\mathcal{N}(x;\mu,\sigma^2 I)$. Equation (D.15) bounds its scale sum in absolute value by $(2\pi)^{-d/2}K_d\|x-\mu\|^{1-d}$, and hence by the same constant times $\|u - \mu_\|\|^{1-d}$. The infinite paired grid has disjoint cubes of side $1/3$ around its positions, each weighted by 2^{-m}. On such a cube, $\|u - y\| \le (3/2)\|u - \mu_\|\|$ when the latter is at least J. All these cubes are outside the ball of radius $J/2$. Replacing the sum by the integral over those disjoint cubes gives

$$\begin{aligned} 2^{-m} \sum_{\text{omitted } \mu} \|u - \mu_\|\|^{1-d} &\le (3/2)^{m+d-1} \int_{\|y-u\| \ge J/2} \|y-u\|^{1-d}\, dy \\ &= (3/2)^{m+d-1} \frac{2^{c-1}|S^{m-1}|}{c-1} J^{1-c}. \end{aligned}$$

This proves the finite-cloud derivative bound. All infinite sums are absolute in the estimates used here. The actual mixture remains finite and normalized throughout.

Choose J so large that

$$C_{m,d} J^{1-c} < \tfrac{1}{2} A(2s_0, \theta). \tag{D.17}$$

Then the actual score has the stated inward coordinate signs, uniformly for $s \in [s_0, 2s_0]$, on every retained cell's slabs $1/12 \le |u_j - k_j| \le 1/2 - \theta$.

D.4 A FINE-CELL CAPTURE CONTROLS THE FULL FUTURE TRAJECTORY

For retained indices $\mathbf{k} \in \{-L + J, \dots, L - J\}^m$, let

$$B_{\mathbf{k}} = \{(u,0) : |u_j - k_j| \le 1/2 - \theta \text{ for all } j\}.$$

The closed cylinder over this cell with normal radius $r_{\mathrm{tr}} = sa_d$ is forward invariant. On every tangential face, the corresponding coordinate derivative of $\log p_{MN}$ points strictly inward by (D.17). On the normal boundary the exact affine identity gives $\dot{n} = -\beta(x)n$, so normal distance decreases. A first-exit argument, also at intersections of faces, proves invariance.

Every path in this compact cylinder converges by the finite-mixture lemma. Its limit has zero normal component and is critical. The same coordinate signs exclude any critical point with an offset of magnitude at least $1/12$. Thus the endpoint lies in $C_{\mathbf{k}}$. Every coordinate of such an endpoint is at distance at least $1/12$ from every training coordinate in that dimension. Therefore

$$\operatorname{dist}(x_\infty, \mathcal{S}_M) \geq \sqrt{m}/12, \qquad \|x_\infty - z\| \leq 7\sqrt{m}/12 \quad (z \text{ in the full unit reference cell } \mathbf{k}). \tag{D.18}$$

This controls the endpoint by a *sample-resolution* cell, not by the center of an arbitrarily large data region. Uniqueness of the endpoint inside each cell is neither needed nor asserted. Maximizing the log density on each compact capture cylinder also gives an interior local maximum: an inward derivative rules out a boundary maximizer. Hence distinct retained cells contain distinct off-training local maxima of the actual finite density, not only candidate reference points.

The posterior bound is also explicit. At a covered endpoint, the nearest training center belongs to its own cell. For one coordinate, the squared-distance difference between the two positions $k_j \pm 1/6$ is at most $1/18$. At any fixed scale $\sigma_i \geq s_0$, the larger of their two normalized weights is at most $(1 + e^{-1/(36s_0^2)})^{-1}$. Taking the product over the m coordinates bounds the largest conditional weight among its own 2^m centers. Including other cells only increases the denominator. Finally, the precision-normalized center marginal is a convex combination of the fixed-scale center posteriors. This proves (D.6). The center maximizing that marginal is still a nearest center, since each radial Gaussian kernel decreases with distance. The same bound holds throughout every output cell, including normal offsets.

D.5 Explicit probability errors and a nonempty bandwidth window

We verify that the geometry and the probability requirements can hold simultaneously for one finite cloud and a whole positive-bandwidth interval. Continue with unit spacing. Fix $\epsilon \in (0, 1/4)$, set $\zeta = \epsilon/8$, and use the associated a_d. Choose

$$s_0 \geq \max\left\{\frac{1}{2}, \sqrt{\frac{\log(128m/\epsilon)}{2\pi^2}}\right\}, \qquad \theta = \frac{\epsilon}{32m}. \tag{D.19}$$

Choose a finite integer J satisfying (D.17), and then $L > J$ so large that

$$2L + 1 \geq \max\left\{\frac{32mJ}{\epsilon}, \frac{256ms_0}{\epsilon}\right\}. \tag{D.20}$$

These are finite choices: $A(2s_0, \theta) > 0$, and $J^{1-c} \to 0$ for $c \geq 2$. This explicit order of choices establishes nonemptiness; no incompatible lower and upper bandwidth requirements are hidden in the statement.

Let the fixed relative cutoff ratio $\Lambda > 1$ satisfy

$$\Lambda \geq \max\left\{2, \frac{8\sqrt{m}(L + 1/2)}{\sqrt{2\pi}\,\epsilon s_0}\right\}. \tag{D.21}$$

Choose finite N and $h = \log\Lambda/(N-1)$ with

$$h \leq \min\left\{1, \frac{\log 2}{4}, \frac{\epsilon}{8(d\log\Lambda + \sqrt{2d}/2)}\right\}. \tag{D.22}$$

Use this same Λ, h, N for all $s \in [s_0, 2s_0]$, with $\sigma_i = se^{(i-1)h}$. Thus $\sigma_{\max} = \Lambda s$ may be much larger than the data box. Its local influence has already been included in the finite-array sign estimate, rather than excluded by an artificial small upper cutoff.

Reference cells and intrinsic smoothing. For the uniform distribution on a box of side $2L + 1$, translation by $v \in \mathbb{R}^m$ changes total variation by

$$1 - \prod_{j=1}^{m}(1 - |v_j|/(2L+1))_+ \leq \sum_j |v_j|/(2L+1).$$

Averaging over $v = sG_m$ gives

$$e_{\text{blur}} \le \frac{ms\sqrt{2/\pi}}{2L+1} \le \epsilon/8. \tag{D.23}$$

The probability excluded by the omitted boundary cells and the thin strips between capture cells satisfies

$$e_{\text{cell}} \le \frac{2mJ}{2L+1} + 2m\theta \le \epsilon/8. \tag{D.24}$$

Indeed the exact retained fraction is $[((2(L-J)+1)/(2L+1))(1-2\theta)]^m$, and $1-\prod_j(1-b_j) \le \sum_j b_j$. Also $e_{\text{normal}} = \mathbb{P}[\|G_c\| > a_d] \le \zeta$, since the c-dimensional norm is dominated by the d-dimensional one.

The empirical fine smoothing is close to the reference smoothing. An elementary cellwise coupling sends each coordinate of a uniform reference point to $k - 1/6$ on the left half-cell and to $k + 1/6$ on the right. It has the correct empirical marginal and coordinate error at most $1/3$. Consequently

$$W_1(\widehat{Q}_M, Q_{\mathcal{M}}) \le \sqrt{m}/3. \tag{D.25}$$

For a sharper estimate at the lower cutoff, truncate the infinite periodic product density to D_L and divide it by $(2L+1)^m$. This is a probability density, because each coordinate interval contains an integer number of periods and $\int_0^1 g_s = 1$. On D_L the finite-array smoothed density is no larger than this periodic density. Their total-variation distance is therefore exactly the mass of the finite smoothing outside D_L.

The preceding coordinatewise coupling bounds that mass by $2m(s\sqrt{2/\pi} + 1/3)/(2L+1)$: a point can leave the box only if the coupled uniform coordinate is within $|sG_1| + 1/3$ of a boundary. The periodic density differs from the uniform density by at most

$$\rho_m(s) := \frac{1}{2}\left[\left(1 + \frac{2q}{1-q}\right)^m - 1\right], \qquad q = e^{-2\pi^2 s^2}.$$

Our choice of s_0 gives $q \le \epsilon/(128m)$ and $\rho_m(s) \le 4mq \le \epsilon/32$. For example, use $2q/(1-q) \le 3q$ and $e^v - 1 \le 2v$ for $0 \le v \le 1$. Adding (D.23), and using $s \ge 1/2$, yields

$$\mathrm{TV}(\widehat{p}_s\, dx, p_s^Q\, dx) \le \rho_m(s) + \frac{4ms}{2L+1} \le \epsilon/16. \tag{D.26}$$

The ambient normal Gaussian is common to both measures, so their ambient TV equals their intrinsic TV. This calculation improves the generic $W_1/\sigma_{\min}$ estimate for this particular resolved grid; it does not assert that improvement for arbitrary training clouds.

The source–sink and initialization errors. The residual decomposition used in Proposition B.7 gives

$$\epsilon_{\text{Pois}} \le dh \log \Lambda + \frac{h}{2}\sqrt{2d} + \mathrm{TV}(\widehat{p}_s\, dx, p_s^Q\, dx) + \mathrm{TV}(\widehat{p}_{\Lambda s}\, dx, p_{\Lambda s}^Q\, dx).$$

The grid term is at most $\epsilon/8$. At the upper cutoff, (C.7) and (D.25) give TV at most $\sqrt{m}/(3\sqrt{2\pi}\Lambda s) \le \epsilon/16$. For the *untruncated* initialization $\nu = \mathcal{N}(0, (\Lambda s)^2 I_d)$, the same Gaussian translation calculation gives

$$\epsilon_{\text{init}} \le \frac{\sqrt{m}(L+1/2)}{\sqrt{2\pi}\Lambda s} \le \epsilon/8.$$

No compact-initialization condition is needed here: finite-mixture convergence holds for every start, and the marked-distribution lemma is global. Together with (D.26), these bounds imply

$$\epsilon_{\text{Pois}} + \epsilon_{\text{init}} \le 3\epsilon/8. \tag{D.27}$$

Thus the total error in (D.10) is at most $3\epsilon/4 < \epsilon$, uniformly on the entire window.

D.6 Endpoint separation and regional coverage

Proof of Corollary 3.2. Take the finite constants and schedule just constructed. By finite-mixture convergence and the affine normal identity, every trajectory has a critical limit in $\mathcal{A} \cap \mathrm{conv}(\mathcal{S}_M)$. Apply the coupling in Appendix D.2 with the capture cells verified above. Its good event has probability at least $1 - \epsilon$, and on this event the actual endpoint lies in $C_{\mathbf{k}(Z)}$. The reference cell label is uniform on $\{-L, \dots, L\}^m$, since $Z \sim Q_{\mathcal{M}}$. For any index collection I, the two indicators

$$\mathbf{1}_{\{x_\infty \in \bigcup_{\mathbf{k} \in I} C_{\mathbf{k}}\}}, \qquad \mathbf{1}_{\{\mathbf{k}(Z) \in I\}}$$

agree on the good event. Their expectation difference has magnitude at most ϵ, proving (22). Equivalently, the output-cell label distribution, with one extra label for points outside all output cells, is within TV ϵ of the uniform reference label distribution. The reference assigns that extra label zero mass.

Restore ℓ in (D.18). On the same good event,

$$\mathrm{dist}(x_\infty, \mathcal{S}_M) \geq \sqrt{m}\ell/12, \qquad \|x_\infty - Z\| \leq 7\sqrt{m}\ell/12.$$

Each training coordinate has covering radius $\ell/3$ on its interval: the largest adjacent gap is $2\ell/3$, and the outer boundary has the same $\ell/3$ distance. Coordinatewise nearest points give $h_{\mathrm{cov},M} = \sqrt{m}\ell/3$, with equality at a grid-cell corner. This proves the stated separation and coupling constants. A reference point in $U^{-7h_{\mathrm{cov},M}/4}$ cannot leave U under that distance bound. Subtracting the failure probability proves the *joint* regional inequality (21).

For the count of populated cells, let I contain those with probability less than $1/[2(2L+1)^m]$. Their total deficit exceeds $|I|/[2(2L+1)^m]$, which (22) bounds by ϵ. Thus at least $(1-2\epsilon)(2L+1)^m$ cells have the claimed mass. Each cell contains endpoints with positive training distance. In addition, every retained capture cylinder contains an open set of positive Gaussian initialization probability, so it supplies an endpoint in its own output cell independently of the quantitative count.

Finally, both the endpoint and Z lie in D_L, by the convex-hull identity for the former. Their coupled expected distance is at most $7h_{\mathrm{cov},M}/4 + \epsilon\,\mathrm{diam}(D_L)$, proving (D.8). The posterior bound was proved in Appendix D.4. All events are measurable because the limit map is the pointwise limit of continuous integer-time flow maps. □

A positive training-distance margin for every start. For each fixed certified cloud and relative schedule, a smaller closed bandwidth interval $I \subset (s_0, 2s_0)$ of positive length admits a uniform $\eta > 0$ with $\mathrm{dist}(x_\infty, \mathcal{S}_M) \geq \eta$ for every start. To see this in unit-spacing coordinates, put $t_i = e^{(i-1)h}$ and denote the density at scales st_i by $p^{(s)}_{MN}$. Each function $f_j(s) = \langle \mu_j, \nabla \log p^{(s)}_{MN}(\mu_j) \rangle$ is analytic on $(0, \infty)$, and the finite Gaussian expansion gives

$$s^2 \nabla \log p^{(s)}_{MN}(\mu_j) \longrightarrow -\frac{\sum_i t_i^{-d}}{\sum_i t_i^{2-d}} \mu_j \quad (s \to \infty).$$

The centered grid contains no zero training center, so no f_j is identically zero. Their finitely many zeros on $[s_0, 2s_0]$ leave such an interval I. The training-center gradient norms have a strictly positive minimum over j and $s \in I$. Uniform continuity extends this nonvanishing property to common neighborhoods of the centers, excluding critical points there. Every endpoint is critical. Restoring ℓ rescales the margin; this every-start margin need not be comparable to $h_{\mathrm{cov},M}$, unlike the explicit high-probability margin.

A continuous population limit. Choose $\epsilon_n \downarrow 0$ and a certified $s_{0,n}, J_n, L_n$ for each n, with $L_n \to \infty$. Setting $\ell_n = (2L_n + 1)^{-1}$ keeps the reference box fixed at $[-1/2, 1/2]^m \times \{0\}$, and (D.8) gives convergence to its continuous reference distribution in W_1. Each finite model uses a positive bandwidth fixed throughout generation and retains a positive training-distance threshold $\sqrt{m}/[12(2L_n + 1)]$, proportional to its training covering radius.

Small perturbations of a fixed certified cloud. The sign inequalities have a strict positive margin on a finite union of compact slabs. For fixed finite L, N and positive cutoffs, the finite Gaussian density and its derivatives depend continuously on the training positions, uniformly on those compact

sets. Therefore sufficiently small perturbations *within the same affine space* preserve the capture and core-exclusion signs. Moving every center by less than $h_{\mathrm{cov},M}/8$ preserves a separation of at least $h_{\mathrm{cov},M}/8$, where this displayed radius denotes the unperturbed grid value. The residual is also continuous in L^1 under these perturbations, by Gaussian translation continuity and finiteness of the mixture. Hence the spare error margin preserves the cell-mass and coverage bounds under sufficiently small within-subspace perturbations, with either a slightly enlarged failure budget or the original budget after starting from a smaller one. Thus the guarantees persist on an open neighborhood of each certified finite cloud.

Finite-resolution interpretation. For a finite Gaussian mixture, an open m-dimensional critical plateau on $\mathcal{A}$ is impossible: the restricted density is analytic; vanishing derivatives on an open set would make it constant on the connected affine space, contradicting Gaussian decay at infinity. If its critical points are all isolated, compactness of their convex-hull location makes their number finite. In that common nondegenerate case, the endpoint distribution is discrete. Neither fact prevents the many-cell, finite-resolution coverage and continuous *limit of distributions* proved above. No continuum of endpoints for one fixed empirical mixture is assumed.

Interpretation of the bandwidth window. The sufficient lower floor controls sample-scale contrast and the fine-smoothing approximation. A larger floor weakens the useful periodic modulation and therefore requires a smaller finite-array perturbation in (D.17). For a fixed finite array this creates an upper restriction on the certified floor. Choosing J and then L as above leaves a nonempty interval between these requirements. The constants are conservative, and weak modulation may lead to slow final tangential convergence; nevertheless, the result establishes an exact positive-bandwidth window for the finite-cloud family.

D.7 Compatibility with a finite reach

For a curved instance satisfying Theorem A.14, the additional normal-accuracy requirement is

$$\varepsilon_{\mathcal{M}} = \frac{C_{\mathrm{loc}} m}{2\tau}\sigma_{\min}^2 \le \varepsilon \quad \Longleftarrow \quad \sigma_{\min} \le \sqrt{\frac{2\tau\varepsilon}{C_{\mathrm{loc}} m}}. \tag{D.28}$$

Here ε is a prescribed normal tolerance. Increasing reach relaxes this sufficient upper bound only when the actual localization constant and the other hypotheses remain controlled. For example, an independently established smoothing window $[\sigma_0, 2\sigma_0]$ meets this tolerance throughout if $C_{\mathrm{loc}} \le C_*$ and $\tau \ge 2C_* m\sigma_0^2/\varepsilon$, along with the tube and landscape conditions. This is a compatibility calculation, not a claim that arbitrary curved data inherit the affine retention condition. In the affine case the normal error is exactly zero, so only the sample-wise and regional requirements remain. Large reach does not by itself supply separation between distinct data regions.

D.8 Comparison with the zero-noise empirical diffusion endpoint

We compare with exact empirical reverse diffusion initialized from its correct terminal noisy marginal. At noise level σ, its prescribed marginal is $\widehat{p}_\sigma$ (Song et al., 2020). For a uniform training index J independent of $G \sim \mathcal{N}(0, I_d)$,

$$Y_\sigma \overset{\mathrm{d}}{=} \mu_J + \sigma G.$$

Consequently, for every fixed $\eta > 0$,

$$\mathbb{P}[\mathrm{dist}(Y_\sigma, \mathcal{S}_M) > \eta] \le \mathbb{P}[\sigma\|G\| > \eta] \longrightarrow 0, \qquad P_{Y_\sigma} \Longrightarrow \widehat{Q}_M \quad (\sigma \to 0).$$

The distance bound follows from $\mathrm{dist}(\mu_J + \sigma G, \mathcal{S}_M) \le \sigma\|G\|$, while weak convergence follows by dominated convergence against bounded continuous test functions. Thus the exact empirical zero-noise diffusion marginal recovers the empirical center distribution $\widehat{Q}_M$.

Corollary 3.2 describes a different positive-bandwidth regime. The smoothing floor remains fixed throughout optimization, yet the same infinite-time endpoints satisfy a training-distance margin proportional to the training resolution together with regional and cell-mass coverage. In the constructive

affine family, positive smoothing can therefore retain sample-resolution diversity without forcing the empirical-score endpoints back onto the training centers.

A positive-cutoff posterior-mean readout provides a related geometric comparison. Direct differentiation gives

$$y + \sigma^2 \nabla \log \widehat{p}_\sigma(y) = \frac{\sum_j \mu_j \mathcal{N}(y; \mu_j, \sigma^2 I)}{\sum_j \mathcal{N}(y; \mu_j, \sigma^2 I)} \in \mathrm{conv}(\mathcal{S}_M).$$

Hence, when the training centers lie in the affine space $\mathcal{A}$, this readout also lies in $\mathcal{A}$. The unified-score construction studied here additionally controls the infinite-time endpoint distribution itself: for the certified finite-cloud family, the endpoints remain separated from the training set while retaining regional probability at the training resolution. Along the fixed-box sequence constructed above, these endpoint distributions further converge in W_1 to a continuous reference distribution.

E SOLVABLE MODELS

The flat calculation identifies the posterior moment used in normal contraction. The sphere exhibits its reach dependence, unstable off-tube geometry, and exact coverage under symmetry. The proper sphere mixtures are population instances of our framework; the translation-invariant flat model is a local reference, not a probability distribution.

Proper population mixtures. For compactly supported $Q_{\mathcal{M}}$ and a finite nonzero scale measure π on $[\sigma_{\min}, \sigma_{\max}] \subset (0, \infty)$, define

$$p^{\mathrm{pop}}(x) = \frac{\int \sigma^2 \int \mathcal{N}(x; \mu, \sigma^2 I) Q_{\mathcal{M}}(d\mu) \pi(d\sigma)}{\int \sigma^2 \pi(d\sigma)}. \tag{E.1}$$

Tonelli's theorem gives a positive normalized density. Bounded centers and positive cutoffs make the Gaussian power series locally uniformly convergent over the mixing variables, so the density and its logarithm are analytic. With $B = \sup \|\mu\|$, its tail is bounded by $(2\pi\sigma_{\min}^2)^{-d/2} \exp[-(\|x\| - B)^2/(2\sigma_{\max}^2)]$ for $\|x\| \geq B$. Thus positive superlevels are compact, and the analytic convergence proof applies. Posterior sums become integrals in the geometric arguments; Appendix I gives the full population convergence and coverage statements.

The unbounded log-scale measure on $(0, \infty)$ has infinite mixture normalizer and is used only for reference kernels and posteriors away from the data. The flat model also has an improper translation-invariant center measure. Neither is substituted for a proper sampling density.

E.1 FLAT MODEL: POSTERIOR MOMENTS AND FINITE SCHEDULES

Take $\mathcal{M}_\flat = \mathbb{R}^m \times \{0\}$ with Lebesgue base measure. Write $x = (x_\parallel, x_\perp)$, $r = \|x_\perp\| > 0$, and $z_x = (x_\parallel, 0)$. Conditional on any fixed σ, the precision-weighted center posterior is $\mathcal{N}(x_\parallel, \sigma^2 I_m)$. Its conditional second moment around z_x is $m\sigma^2$.

Proposition E.1 (Exact unbounded-reference localization constant)**.** *With scale measure $d\sigma/\sigma$ on $(0, \infty)$, the precision-weighted posterior is proper for every $r > 0$, and*

$$u := \frac{r^2}{2\sigma^2} \sim \mathrm{Gamma}(c/2, 1). \tag{E.2}$$

For $c > 2$,

$$\mathbb{E}_{\widetilde{w}} \|\mu - z_x\|^2 = m \mathbb{E}_{\widetilde{w}} \sigma^2 = \frac{m r^2}{c - 2}. \tag{E.3}$$

Thus the distance-matched moment bound has the sharp constant

$$C_{\mathrm{loc}} = \frac{c}{c - 2}. \tag{E.4}$$

Proof. At fixed σ, normalizing the tangent Gaussian gives $\mu_\parallel \mid x, \sigma \sim \mathcal{N}(x_\parallel, \sigma^2 I_m)$, with conditional squared displacement $m\sigma^2$. Integrating the tangent coordinates leaves the scale weight $\sigma^{-c} e^{-r^2/(2\sigma^2)} d\sigma/\sigma$. Under $u = r^2/(2\sigma^2)$ this is proportional to $u^{c/2-1} e^{-u} du$, with finite Gamma normalizer for $r > 0, c > 0$. For $c > 2$,

$$\mathbb{E}\sigma^2 = \frac{r^2}{2} \mathbb{E}[u^{-1}] = \frac{r^2}{2} \frac{\Gamma(c/2 - 1)}{\Gamma(c/2)} = \frac{r^2}{c - 2}.$$

Multiplication by m gives the result, and comparison with mr^2/c gives the exact constant. □

The mode squared is r^2/c, whereas the mean squared scale is $r^2/(c - 2)$. Their ratio explains the moment constant without identifying a mode with a mean.

Proposition E.2 (Cutoff dependence in codimensions one and two)**.** *For $c = 1, 2$, with a truncated log-continuous schedule satisfying $\sigma_{\min} \leq r/\sqrt{c}$ and $\sigma_{\max} \geq 2r$,*

$$\mathbb{E}_{\widetilde{w}} \|\mu - z_x\|^2 \asymp m \begin{cases} r\sigma_{\max}, & c = 1, \\ r^2 \log(\sigma_{\max}/r), & c = 2, \end{cases} \tag{E.5}$$

with absolute comparison constants. Thus no upper-cutoff-independent constant controls the distance-matched second moment as $\sigma_{\max}/r \to \infty$.

Proof. The same substitution truncates the Gamma integrals to $u_- = r^2/(2\sigma_{\max}^2) \le 1/8$ and $u_+ = r^2/(2\sigma_{\min}^2) \ge c/2$. Thus

$$\mathbb{E}\sigma^2 = \frac{r^2}{2}\frac{\int_{u_-}^{u_+} u^{c/2-2}e^{-u}du}{\int_{u_-}^{u_+} u^{c/2-1}e^{-u}du}. \tag{E.6}$$

For $c = 1, 2$, the denominator lies between the positive integral on $[1/8, 1/2]$ and $\Gamma(c/2)$. For $c = 1$, bounding e^{-u} from below on $[u_-, 1/2]$ and from above by one gives a numerator comparable to $u_-^{-1/2}$. For $c = 2$, integration on $[u_-, 1]$ and a bounded exponential tail give a numerator comparable to $\log(1/u_-)$. Substitution yields respectively $r\sigma_{\max}$ and $r^2\log(\sigma_{\max}/r)$. Multiply by m. □

This is a threshold for a cutoff-independent moment in this model, not an impossibility theorem for finite schedules in low codimension. Nor does it say that the contraction gap must fail whenever $c \le 2$.

Lemma E.3 (Discrete geometric schedules)**.** *Let $c \ge 3$, $\sigma_i = \sigma_{\min}\rho^{i-1}$, $\sigma_N = \sigma_{\max}$, $\rho > 1$, and*

$$h := \log\rho \le \frac{1}{27\sqrt{c}}, \qquad \sigma_{\max}^2 \ge \frac{8c}{c-2}\sigma_{\min}^2.$$

Then the flat-model posterior satisfies, for $0 < r \le \sigma_{\max}\sqrt{(c-2)/2}$,

$$\mathbb{E}_{\widetilde{w}}\|\mu - z_x\|^2 \le \big(1 + 30(h\sqrt{c} + e^{-c/2})\big)\frac{m\max\{r^2, 4c\sigma_{\min}^2\}}{c-2}. \tag{E.7}$$

In particular, the choice

$$C_{\mathrm{loc}} := (1 + 30(h\sqrt{c} + e^{-c/2}))\frac{4c}{c-2}$$

is valid on this range.

Proof. It is enough to bound the scale mean, since the conditional center moment is $m\sigma^2$.

Reduction to radii above the lower cutoff. Differentiating with respect to $v = r^2/2$ gives $\frac{d}{dv}\mathbb{E}\sigma^2 = -\operatorname{Cov}(\sigma^2, \sigma^{-2}) = \mathbb{E}\sigma^2\mathbb{E}\sigma^{-2} - 1 \ge 0$. Set $r_\dagger = 2\sqrt{c}\,\sigma_{\min}$, which is in the admissible range by the scale-range assumption. It suffices to treat $r \ge r_\dagger$ and then use monotonicity.

Riemann sums relative to their integrals. For $k = c, c-2$, let $f_k(t) = \exp(-kt - r^2e^{-2t}/2)$, $I_k = \int_{\log\sigma_{\min}}^{\log\sigma_{\max}} f_k(t)dt$, and $S_k = h\sum_i f_k(\log\sigma_i)$. Unimodality gives $|S_k - I_k| \le 3h\sup f_k$: two peak heights bound total variation and one accounts for the extra endpoint. Both modes $t_k^* = \log(r/\sqrt{k})$ are at least $(\log 2)/2$ from the integration boundaries. For $|v| \le 1/2$,

$$\log f_k(t_k^*) - \log f_k(t_k^* + v) = k[v + (e^{-2v} - 1)/2] \le ekv^2.$$

Integrating on $|v| \le \min\{1/(2\sqrt{k}), 1/3\}$ gives $I_k \ge \sup f_k/(3\sqrt{k})$. Therefore, for $b = h\sqrt{c} \le 1/27$,

$$\frac{S_{c-2}}{S_c} \le \frac{1+9b}{1-9b}\frac{I_{c-2}}{I_c} \le (1+27b)\,\mathbb{E}_{\mathrm{cont,trunc}}\sigma^2.$$

Truncation of the scale posterior. For $u \sim \mathrm{Gamma}(c/2, 1)$, the upper cutoff in u is $u_+ = r^2/(2\sigma_{\min}^2) \ge 2c$. Removing the lower-$u$ part only decreases the conditional mean of u^{-1}. Moreover, $\mathbb{P}(u > u_+) \le e^{-c}\mathbb{E}e^{u/2} = e^{-c}2^{c/2} \le e^{-c/2}$. Hence

$$\mathbb{E}_{\mathrm{cont,trunc}}\sigma^2 \le \frac{r^2}{c-2}\frac{1}{1-e^{-c/2}}.$$

For $b \le 1/27$ and $a = e^{-c/2} < 1/4$, $(1+27b)/(1-a) \le 1 + 30(b+a)$, as is seen by multiplying by $1-a$. This proves the bound for $r \ge r_\dagger$. Monotonicity replaces r^2 by $r_\dagger^2$ below that threshold. Finally, $\max\{r^2, 4c\sigma_{\min}^2\} \le 4c\max\{r^2/c, \sigma_{\min}^2\}$, giving the stated localization constant. □

The grid condition is sufficient, not necessary. E1 uses 24 scales on two decades, with $h\sqrt{c} \approx 0.6$, outside this sufficient range. Its downstream diagnostics do not certify this particular uniform moment bound.

Remark E.4 (Limits of the flat idealization). For finitely many centers in an affine subspace, normal contraction is exact by Proposition A.18. The translation-invariant flat calculation instead describes a local posterior and is not a normalizable data distribution. In particular, its unbounded-scale singular field should not be substituted into the finite-mixture global-flow theorem. Curvature and off-tube instability are addressed by the sphere model next.

E.2 Sphere model: curvature, localization, and exact regional coverage

Let $\mathcal{M}_\circ = S^m(R) \subset V := \mathbb{R}^{m+1} \times \{0\} \subset \mathbb{R}^d$, with uniform reference distribution, reach R, and codimension $c = d - m$. Use the proper population density p^{pop} from (E.1) for this uniform distribution and the stated bounded scale schedule; reference-limit kernels are distinguished explicitly. Write P_V for the orthogonal projector onto V, not for the nonlinear sphere projection $\Pi_{\mathcal{M}}$. The scalar $\varrho = \|P_V x\|$ denotes subspace radius in the sphere calculations.

Proposition E.5 (Sharpness of the reach inequality)**.** *For $z, \mu \in S^m(R)$, set $e = z/R$. Then*

$$\langle e, \mu - z\rangle = -\frac{\|\mu - z\|^2}{2R}.$$

Consequently the positive-reach upper bound is attained for the inward normal $u = -e$, including after averaging over any posterior.

Proof. For $\|\mu\| = \|z\| = R$, $\|\mu - z\|^2 = 2R^2 - 2\langle \mu, z\rangle$. Dividing by $2R$ gives the identity for $e = z/R$ and its positive version for $-e$. Averaging preserves the equality. □

This shows sharpness of the normal chord coefficient $1/(2\tau)$, not sharpness of every subsequent bound on a trajectory.

Proposition E.6 (Sphere localization by an exact flat-model comparison)**.** *Suppose $m \geq 2$. Use the same finite or truncated-continuous scale schedule in the sphere and flat models. For any x with $0 < r(x) < R$, write $r = r(x)$, set $\varrho = \|P_V x\| > 0$, and $z_x = RP_V x/\varrho$. Then*

$$\mathbb{E}_{\widetilde{w},\mathrm{sph}}\|\mu - z_x\|^2 \leq \frac{R}{\varrho}\, m\, \mathbb{E}_{\widetilde{w},\mathrm{flat},r(x)}\sigma^2. \tag{E.8}$$

Hence, if the flat-model bound is valid with constant C_{flat} on $0 < r < r_0 < R$, the sphere moment bound is valid with the choice

$$C_{\mathrm{loc}} := \frac{R}{R - r_0} C_{\mathrm{flat}}. \tag{E.9}$$

For the unbounded reference posterior and $c > 2$,

$$\mathbb{E}_{\widetilde{w},\mathrm{sph}}\|\mu - z_x\|^2 \leq \frac{R}{\varrho}\frac{mr^2}{c-2}, \qquad \mathbb{E}_{\widetilde{w},\mathrm{sph}}\|\mu - z_x\|^2 = \frac{R}{\varrho}\frac{mr^2}{c-2}(1+o(1)) \quad (r/R \to 0,\ m, c \text{ fixed}). \tag{E.10}$$

The local-limit equality does not assert the same leading constant uniformly for arbitrary finite cutoffs.

Proof. We compare the sphere to a flat posterior at the same normal distance and with exactly the same scale measure. All auxiliary quantities in this proof are local to this comparison: set

$$\varrho := \|P_V x\|, \quad k := \varrho/R, \quad a := m/2, \quad z_x := RP_V x/\varrho, \quad v := \|\mu - z_x\|^2.$$

Because $r(x) < R$, $\varrho > 0$: if $P_V x = 0$, its distance to any sphere point is $\sqrt{R^2 + \|(I - P_V)x\|^2} \geq R$.

Exact chord coordinates. Writing θ for the angle between μ and z_x,

$$v = 2R^2(1 - \cos\theta), \qquad r^2 = (\varrho - R)^2 + \|(I - P_V)x\|^2.$$

Therefore

$$\begin{aligned}\|x-\mu\|^2 &= \varrho^2 + R^2 - 2\varrho R\cos\theta + \|(I-P_V)x\|^2 \\ &= r^2 + 2\varrho R(1-\cos\theta) = r^2 + kv.\end{aligned}$$

This is why the argument covers nonradial ambient points as well. Uniform spherical measure has angular factor $\sin^{m-1}\theta\, d\theta$. Since

$$dv = 2R^2\sin\theta\, d\theta, \qquad \sin^2\theta = \frac{v}{R^2}\left(1-\frac{v}{4R^2}\right),$$

its density in v, up to a positive constant independent of v and σ, is

$$v^{a-1}\left(1-\frac{v}{4R^2}\right)^{a-1} dv, \qquad 0<v<4R^2.$$

Define

$$\chi(v) := \begin{cases}(1-v/(4R^2))^{a-1}, & 0\le v<4R^2,\\ 0, & v\ge 4R^2.\end{cases}$$

Because $m\ge 2$, we have $a\ge 1$, so $0\le\chi\le 1$ and χ is nonincreasing. This sign is the only reason for the restriction $m\ge 2$.

The fixed-scale chord moment. At fixed σ, the sphere posterior in v is proportional to $v^{a-1}e^{-kv/(2\sigma^2)}\chi(v)\,dv$. Without $\chi(v)$, this is a Gamma distribution of shape a and scale $2\sigma^2/k$, whose mean is $2a\sigma^2/k = m\sigma^2/k$. For any integrable random variable Y and an independent copy Y', a bounded nonincreasing weight satisfies

$$2\,\mathrm{Cov}(Y,\chi(Y)) = \mathbb{E}[(Y-Y')(\chi(Y)-\chi(Y'))]\le 0.$$

Every integrand is nonpositive. Thus, when $\mathbb{E}\chi(Y)>0$,

$$\frac{\mathbb{E}[Y\chi(Y)]}{\mathbb{E}\chi(Y)} = \mathbb{E}Y + \frac{\mathrm{Cov}(Y,\chi(Y))}{\mathbb{E}\chi(Y)} \le \mathbb{E}Y.$$

Apply this identity to the fixed-scale Gamma distribution. Its tilt has a positive normalizer, and we obtain

$$\mathbb{E}_{\mathrm{sph}}[v\mid\sigma] \le \frac{m}{k}\sigma^2. \tag{E.11}$$

Comparing the scale posteriors. For the common scale measure $\pi(d\sigma)$, integration in v gives the sphere scale weight

$$\sigma^{-d}e^{-r^2/(2\sigma^2)}\left[\int_0^\infty v^{a-1}e^{-kv/(2\sigma^2)}\chi(v)\,dv\right]\pi(d\sigma).$$

Set $v=2\sigma^2u/k$ inside this integral and define

$$\omega_{\mathrm{sph}}(\sigma) := \int_0^\infty u^{a-1}e^{-u}\chi(2\sigma^2u/k)\,du.$$

The bracket becomes $(2\sigma^2/k)^a\omega_{\mathrm{sph}}(\sigma)$. Since $2a=m$, the full scale weight is, up to a scale-independent factor,

$$\sigma^{-c}e^{-r^2/(2\sigma^2)}\omega_{\mathrm{sph}}(\sigma)\,\pi(d\sigma).$$

The flat scale weight is exactly the same expression without $\omega_{\mathrm{sph}}(\sigma)$. Moreover, ω_{sph} is nonincreasing: increasing σ decreases $\chi(2\sigma^2u/k)$ at every u. The covariance inequality above, now with the increasing variable σ^2 and decreasing weight ω_{sph}, gives

$$\begin{aligned}\mathbb{E}_{\mathrm{sph}}\sigma^2 &= \frac{\mathbb{E}_{\mathrm{flat},r}[\sigma^2\omega_{\mathrm{sph}}(\sigma)]}{\mathbb{E}_{\mathrm{flat},r}\omega_{\mathrm{sph}}(\sigma)}\\ &\le \mathbb{E}_{\mathrm{flat},r}\sigma^2.\end{aligned}$$

Taking expectations in (E.11) therefore yields

$$\mathbb{E}_{\mathrm{sph}}\|\mu - z_x\|^2 = \mathbb{E}_{\mathrm{sph}} v \le \frac{m}{k}\mathbb{E}_{\mathrm{sph}}\sigma^2 \le \frac{R}{\varrho}\, m\mathbb{E}_{\mathrm{flat},r}\sigma^2,$$

which is (E.8). If $r < r_0$, then $|\varrho - R| \le r < r_0$, so $\varrho > R - r_0$. A valid flat constant consequently gives

$$\begin{aligned}\mathbb{E}_{\mathrm{sph}}\|\mu - z_x\|^2 &\le \frac{R}{\varrho} C_{\mathrm{flat}} m \max\{\sigma_{\min}^2, r^2/c\} \\ &\le \frac{R}{R - r_0} C_{\mathrm{flat}} m \max\{\sigma_{\min}^2, r^2/c\}.\end{aligned}$$

This proves the stated choice of C_{loc}.

The unbounded reference and its local limit. For $c > 2$, the flat reference scale moment is finite, so the same comparison applies and gives $\mathbb{E}_{\mathrm{sph}} v \le (R/\varrho) m r^2/(c-2)$. For the asymptotic equality, integrating the precision weight over $\sigma \in (0, \infty)$ gives the chord density proportional to

$$v^{a-1}\chi(v)(r^2 + kv)^{-d/2}\, dv.$$

This follows by the same Gamma substitution as in the flat proof, with $r^2 + kv$ in place of r^2. Substituting $v = r^2 u/k$ in its numerator and denominator yields

$$\mathbb{E}_{\mathrm{sph}} v = \frac{r^2}{k} \frac{\displaystyle\int_0^\infty u^a (1+u)^{-d/2} \chi(r^2 u/k)\, du}{\displaystyle\int_0^\infty u^{a-1}(1+u)^{-d/2}\chi(r^2 u/k)\, du}.$$

As $r/R \to 0$, $|\varrho - R| \le r$ implies $k \to 1$, and $\chi(r^2 u/k) \to 1$ for each fixed u. Both integrands are dominated by their versions with $\chi = 1$. The denominator is integrable for $c > 0$, and the numerator for $c > 2$; these are precisely the required hypotheses for dominated convergence. Using the beta integral, their limiting ratio is

$$\begin{aligned}\frac{B(a+1, c/2-1)}{B(a, c/2)} &= \frac{\Gamma(a+1)\Gamma(c/2-1)}{\Gamma(a)\Gamma(c/2)} \\ &= \frac{a}{c/2-1} = \frac{m}{c-2}.\end{aligned}$$

Here $B(p,q) = \int_0^\infty u^{p-1}(1+u)^{-p-q}\, du = \Gamma(p)\Gamma(q)/\Gamma(p+q)$ for positive p, q. Since $1/k = R/\varrho$, this proves the local-limit equality in (E.10). It does not replace the finite-range moment by the unbounded moment without a truncation bound.

□

Remark E.7 (Usable finite-range constants). For $m \ge 2, c \ge 3$, assume the schedule hypotheses of Lemma E.3 and $r_0 \le \sigma_{\max}\sqrt{(c-2)/2}$. One valid flat constant is

$$C_{\mathrm{flat}} := \big(1 + 30(h\sqrt{c} + e^{-c/2})\big)\frac{4c}{c-2}.$$

The sphere comparison then gives the valid uniform choice

$$C_{\mathrm{loc}} := \frac{R}{R - r_0} C_{\mathrm{flat}}.$$

For this choice, substituting $\tau = R$ and the definition of ε_0 yields

$$\begin{aligned}\varepsilon_0 &= \frac{C_{\mathrm{loc}} m}{2R}\sigma_{\min}^2 = \frac{C_{\mathrm{flat}} m}{2(R - r_0)}\sigma_{\min}^2, \\ \kappa_0 &= \frac{C_{\mathrm{loc}} m r_0}{2cR} = \frac{C_{\mathrm{flat}} m r_0}{2c(R - r_0)}.\end{aligned}$$

The contraction gap can therefore be checked without an implicit substitution:

$$\kappa_0 < 1 \quad \Longleftrightarrow \quad C_{\mathrm{flat}} m r_0 < 2c(R - r_0) \quad \Longleftrightarrow \quad \frac{r_0}{R} < \frac{2c}{C_{\mathrm{flat}} m + 2c}.$$

These are valid finite-schedule constants at every point of the tube, including points outside the sphere's containing subspace. The value $c/(c-2)$ remains the flat unbounded-reference benchmark; it must not replace a uniform finite-schedule constant without a separate justification. The $m = 1$ experiments concern radial scores and spectra, not an application of this $m \geq 2$ comparison.

Proposition E.8 (Sphere-center spectrum: finite schedule and reference limit)**.** *At $x = 0$, the center posterior is uniform and independent of scale, so $\nabla \log p^{\mathrm{pop}}(0) = 0$. For any proper bounded scale schedule,*

$$\beta(0) = \mathbb{E}_w \sigma^{-2}, \qquad C_0 = \frac{R^2}{m+1} \mathbb{E}_w \sigma^{-4} \, P_V. \tag{E.12}$$

The Hessian eigenvalues are $\lambda_{\|} = R^2 \mathbb{E}_w \sigma^{-4}/(m+1) - \beta(0)$ on V and $\lambda_{\perp} = -\beta(0)$ on $V^{\perp}$.

For the unbounded reference and $d > 2$, $u = R^2/(2\sigma^2) \sim \mathrm{Gamma}((d-2)/2, 1)$ under w, and exactly

$$\beta(0) = \frac{d-2}{R^2}, \quad C_0 = \frac{(d-2)d}{(m+1)R^2} P_V, \quad \frac{\mathrm{tr}(C_0)}{\beta(0)} = d, \quad r_{\mathrm{eff}}(C_0) = m + 1. \tag{E.13}$$

For $c \geq 2$, the reference eigenvalues are

$$\lambda_+ = \frac{(d-2)(c-1)}{(m+1)R^2} \quad (m+1 \textit{ directions}), \qquad \lambda_- = -\frac{d-2}{R^2} \quad (c-1 \textit{ directions}). \tag{E.14}$$

Thus the reference shell-rank constants are $\alpha_d = 1$ and $R_d = m + 1$, with a strict gap when $c \geq 2$.

Proof. **Finite schedule.** At the origin all centers are at distance R. The posterior center distribution is therefore uniform and independent of scale, with $\mathbb{E}\mu = 0$ and $\mathbb{E}\mu\mu^{\top} = R^2 P_V/(m+1)$. Thus the gradient vanishes and $C_0 = \mathbb{E}_w[\sigma^{-4}]\mathbb{E}[\mu\mu^{\top}]$. Subtracting $\beta(0)I$ gives the two eigenspaces and eigenvalues.

Unbounded reference. Here the raw scale weight is $\sigma^{2-d} e^{-R^2/(2\sigma^2)} d\sigma/\sigma$. Under $u = R^2/(2\sigma^2)$ it becomes Gamma with shape $(d-2)/2$. Its first two moments give

$$\mathbb{E}_w \sigma^{-2} = (d-2)/R^2, \qquad \mathbb{E}_w \sigma^{-4} = (d-2)d/R^4.$$

Substitution gives (E.13) and (E.14). The positive eigenspace has dimension $m+1$, and the negative one dimension $c-1$; both signs are strict for $c \geq 2$. □

Remark E.9 (Codimension one is schedule dependent). For $c = 1, d > 2$, the reference eigenvalue on V is zero because $2\mathbb{E}[u^2]/d = \mathbb{E}[u]$. A finite schedule need not preserve this sign: a single scale with $\sigma^2 > R^2/(m+1)$ makes $\lambda_{\|} < 0$. For $(m, c) = (1, 1)$, the formal Gamma shape is zero, so even that unbounded posterior is not proper. The zero reference curve in that numerical case is only a formal benchmark. None of these observations rules out useful finite-scale models in codimension one.

Lemma E.10 (Reduction of the off-tube critical set)**.** *For a proper uniform-sphere mixture, every critical point lies in $V \cap \overline{B}_R(0)$. In this subspace, write $g(\varrho) = \langle \nabla \log p^{\mathrm{pop}}(\varrho e), e \rangle$, where $e \in V$ is any unit vector. This is independent of e. If $g(\varrho) > 0$ on $(0, R - r_0]$, the only off-tube critical point is 0.*

Proof. The barycenter lies in V, so $(I - P_V)\nabla \log p^{\mathrm{pop}} = -\beta(I - P_V)x$. Criticality forces $x \in V$, and the barycenter identity gives $\|x\| \leq R$. Rotational symmetry makes the gradient radial on V. For $x \in V$ of radius $\varrho \leq R$, its distance to the sphere is $R - \varrho$. Thus positivity of g on $(0, R - r_0]$ excludes all nonzero off-tube critical points; the origin remains critical by symmetry. □

Proposition E.11 (Reference radial positivity and finite-range persistence)**.** *For $c \geq 2$, the radial score of the unbounded reference kernel is strictly positive on $(0, R)$. For each fixed $0 < r_0 < R$, this positivity on $(0, R - r_0]$, and hyperbolicity at the origin, persist for sufficiently small lower cutoff, sufficiently large upper cutoff, and then a sufficiently fine geometric discretization. No numerical cutoff threshold is asserted here.*

Proof. We first obtain a positive margin for the reference radial score, then show that finite cutoffs and a fine grid preserve that margin. Positivity at grid points alone would not give this conclusion near the origin.

Integrating the reference scale. Because $m \geq 1$ and $c \geq 2$, we have $d > 2$. For $a > 0$, the substitution $u = a^2/(2\sigma^2)$ gives

$$\begin{aligned}\int_0^\infty \sigma^{1-d} e^{-a^2/(2\sigma^2)}\, d\sigma &= \frac{1}{2}\left(\frac{a^2}{2}\right)^{1-d/2} \int_0^\infty u^{d/2-2} e^{-u}\, du \\ &= 2^{d/2-2}\Gamma(d/2-1) a^{2-d}.\end{aligned}$$

After including the Gaussian factor $(2\pi)^{-d/2}$, the multiplier is a positive constant depending only on d. Since constant multipliers do not change a score, we may use the reference kernel

$$\mathscr{P}_\infty(y) := \int_{S^m(R)} \|y - \mu\|^{2-d}\, Q_{\mathcal{M}}(d\mu).$$

It is positive and smooth away from the sphere, and radial when restricted to V.

Positivity from the subspace Laplacian. The Laplacian here is taken in the $(m+1)$-dimensional subspace V, not in $\mathbb{R}^d$. In an n-dimensional Euclidean space, differentiation gives

$$\nabla\|y\|^p = p\|y\|^{p-2} y, \qquad \Delta\|y\|^p = p(p+n-2)\|y\|^{p-2}.$$

Take $p = 2 - d$, $n = m + 1$, and replace y by $y - \mu$. The coefficient becomes

$$(2-d)(2-d+m+1-2) = (2-d)(m+1-d) = (d-2)(c-1) > 0.$$

For $y \in V$ with $\|y\| < R$, differentiation under the compact center integral is valid locally because $\|y - \mu\|$ is bounded away from zero. Thus

$$\Delta_V \mathscr{P}_\infty(y) = (d-2)(c-1) \int \|y - \mu\|^{-d}\, Q_{\mathcal{M}}(d\mu) > 0. \tag{E.15}$$

Write $P_\infty(\varrho) = \mathscr{P}_\infty(\varrho e)$ for a unit vector $e \in V$. Smooth radial symmetry gives $P'_\infty(0) = 0$, and the radial Laplacian formula reads

$$P''_\infty(\varrho) + \frac{m}{\varrho} P'_\infty(\varrho) = \Delta_V \mathscr{P}_\infty(\varrho e), \quad \text{or} \quad (\varrho^m P'_\infty(\varrho))' = \varrho^m \Delta_V \mathscr{P}_\infty(\varrho e).$$

Integrating from zero to $\varrho > 0$, using $\lim_{u \downarrow 0} u^m P'_\infty(u) = 0$, gives

$$P'_\infty(\varrho) = \varrho^{-m} \int_0^\varrho u^m \Delta_V \mathscr{P}_\infty(ue)\, du > 0.$$

Since $P_\infty(\varrho) > 0$, the reference score $g_\infty(\varrho) = P'_\infty(\varrho)/P_\infty(\varrho)$ is positive for every $0 < \varrho < R$.

A uniform margin including the origin. The reference center spectrum gives

$$g_\infty(0) = 0, \qquad g'_\infty(0) = \lambda_+ = \frac{(d-2)(c-1)}{(m+1)R^2} > 0.$$

Thus $g_\infty(\varrho)/\varrho$ extends continuously to $\varrho = 0$, with value λ_+. For fixed $r_0 \in (0, R)$, define

$$\gamma := \min_{0 \leq \varrho \leq R - r_0} \frac{g_\infty(\varrho)}{\varrho} > 0,$$

using this continuous extension at zero. Compactness and strict positivity justify the positive minimum.

Removing tails and discretizing in C^2. On the full ambient closed ball $\|x\| \le R - r_0$, every center is at distance at least r_0, and the distances are also uniformly bounded above. After omission of the mixture normalization, the scale integrand is $\sigma^2 \mathcal{N}(x; \mu, \sigma^2 I) d\sigma/\sigma$. The Gaussian derivative identities

$$\nabla_x \mathcal{N} = -\frac{x-\mu}{\sigma^2}\mathcal{N}, \qquad \nabla_x^2 \mathcal{N} = \left(\frac{(x-\mu)(x-\mu)^\top}{\sigma^4} - \frac{I}{\sigma^2}\right)\mathcal{N}$$

give the following bounds for the scale-integrand coefficients relative to $d\sigma$, including their first two spatial derivatives (the unit cutoff is only for expressing integrability):

$$C\sigma^{-d-3} e^{-r_0^2/(2\sigma^2)} \quad (0 < \sigma \le 1), \qquad C\sigma^{1-d} \quad (\sigma \ge 1),$$

where C is finite and independent of x, μ, σ on the specified compact ball. The first bound is integrable at zero because of Gaussian suppression; the second is integrable at infinity because $d > 2$. Their tail integrals tend to zero. Hence the unnormalized bounded-scale kernels, after multiplication by the same harmless positive constant obtained by scale integration, converge to $\mathscr{P}_\infty$ uniformly with their first two derivatives as the lower cutoff tends to zero and the upper cutoff tends to infinity.

The reference kernel has a positive minimum on this compact ball. The identities

$$\nabla \log P = \frac{\nabla P}{P}, \qquad \nabla^2 \log P = \frac{\nabla^2 P}{P} - \frac{\nabla P \nabla P^\top}{P^2}$$

therefore imply convergence of the log-kernel derivatives as well. For fixed positive finite cutoffs, set $t = \log \sigma$. The scale integrand and its first two spatial derivatives are uniformly continuous on the compact (x, t)-domain. Their geometric-grid sums multiplied by h are Riemann sums, so they also converge uniformly with these derivatives as $h \to 0$. Neither multiplication by h nor normalization of the density changes its score or log Hessian.

Transferring the sign to the full interval. Let g_{fin} be a score obtained after the chosen cutoffs and discretization. Symmetry ensures $g_{\mathrm{fin}}(0) = g_\infty(0) = 0$. For $\varrho > 0$, the fundamental theorem of calculus gives

$$\begin{aligned}\left|\frac{g_{\mathrm{fin}}(\varrho) - g_\infty(\varrho)}{\varrho}\right| &= \left|\frac{1}{\varrho}\int_0^\varrho (g'_{\mathrm{fin}}(u) - g'_\infty(u))\, du\right| \\ &\le \sup_{0 \le u \le R - r_0} |g'_{\mathrm{fin}}(u) - g'_\infty(u)|.\end{aligned}$$

Choose the log-density Hessian approximation error below $\gamma/2$. Then

$$\frac{g_{\mathrm{fin}}(\varrho)}{\varrho} \ge \frac{g_\infty(\varrho)}{\varrho} - \frac{\gamma}{2} \ge \frac{\gamma}{2} > 0 \quad (0 < \varrho \le R - r_0).$$

This also handles arbitrarily small positive ϱ, where ordinary uniform convergence of scores alone would not suffice. Finally, the reference center eigenvalues $\lambda_+ > 0$ and $\lambda_- < 0$ are separated from zero. Taking the same Hessian error below $\frac{1}{2}\min\{\lambda_+, -\lambda_-\}$ preserves their signs, and hence hyperbolicity at the origin. All cutoffs and grid sizes in this argument are qualitative existence choices, not an explicit certificate for the schedules used in E2.

□

This argument supplies an analytic reference result and a qualitative finite-range existence result. It does not certify the particular coarse schedules in E2; those remain numerical tests.

Corollary E.12 (Saddle avoidance in a proper sphere mixture)**.** *For a bounded-scale uniform-sphere mixture with $c \ge 2$, assume $g > 0$ on $(0, R - r_0]$ and $\lambda_\parallel > 0$ at the origin. Then the only possible off-tube endpoint is the hyperbolic center, and an absolutely continuous initialization avoids it almost surely.*

Proof. The preceding reduction leaves only the origin off the tube. Its eigenvalues are $\lambda_{\|} > 0$ on V and $-\beta(0) < 0$ on $V^{\perp}$, so it is hyperbolic and its stable manifold has dimension $c - 1 < d$. The proof of Lemma A.13 applies to this proper population mixture, whose analytic and flow properties were established above. □

Compatibility in one proper model. For $m \geq 2, c \geq 3$, let $\overline{C}_c = 4c[1+30(1/27+e^{-c/2})]/(c-2)$. It bounds C_{flat} whenever $h\sqrt{c} \leq 1/27$. Choose r_0 so that $\overline{C}_c m r_0/[2c(R-r_0)] < 1$, then choose $\sigma_{\min}$ small enough that $\overline{C}_c m \sigma_{\min}^2/[2(R-r_0)] < \min\{r_0, \sigma_{\min}\sqrt{c}\}$. The finite-range persistence proposition permits a sufficiently large upper cutoff and sufficiently fine grid giving radial positivity and center hyperbolicity, while the preceding uniform moment bound remains valid. At the only off-tube critical point, C_0 is proportional to a rank-$(m+1)$ projector and $\lambda_{\|} > 0$ implies $\operatorname{tr} C_0/\beta(0) > m+1$. Thus $R_d = m+1$ and $\alpha_d = \min\{1, \operatorname{tr} C_0/[d\beta(0)]\}$ satisfy the actual shell-rank gap. This verifies compatibility of the geometric assumptions in a proper model; it does not certify the particular coarse numerical schedules.

Proposition E.13 (Exact regional probabilities and zero projected drift on a uniform sphere)**.** *Consider a proper bounded-scale uniform-sphere mixture and an initialization distribution invariant under $O(m+1) \times O(c-1)$, with $\mathbb{P}[P_V X_0 = 0] = 0$. Suppose the transition and stopping tubes are reached almost surely. Then*

$$Z_{\mathrm{stop}} = Z_{\mathrm{tr}} = R\frac{P_V X_0}{\|P_V X_0\|} \quad \textit{almost surely}, \qquad Z_{\mathrm{tr}}, Z_{\mathrm{stop}} \sim \mathrm{Unif}(S^m(R)). \tag{E.16}$$

The actual tangential retention radius is zero.

Proof. Rotations in V fixing $P_V x$ also fix its posterior barycenter, which lies in V. Thus $\dot{y} = b(t)y$ for $y = P_V x$ and a scalar $b(t)$. The subspace $y = 0$ is invariant; uniqueness prevents a path starting outside it from reaching it at finite time. Therefore $y(t)/\|y(t)\|$ is constant. The unique nearest point on the sphere is $RP_V x/\|P_V x\|$, so the two projected outputs equal the initial direction scaled by R. Rotation-invariant initialization makes that direction uniform. Finite entrance is assumed in the proposition, completing the probability statement. □

Corollary E.14 (Concrete sphere coverage)**.** *Under Proposition E.13, for every $z \in S^m(R)$ and $\rho > 0$,*

$$\mathbb{P}[Z_{\mathrm{stop}} \in B_{\mathcal{M}}(z, \rho)] = \frac{\displaystyle\int_0^{\min\{\rho/R, \pi\}} \sin^{m-1}\theta \, d\theta}{\displaystyle\int_0^{\pi} \sin^{m-1}\theta \, d\theta}. \tag{E.17}$$

Proof. The preceding proposition makes Z_{stop} uniform. In colatitude θ around the ball center, the surface element is proportional to $\sin^{m-1}\theta \, d\theta$ and intrinsic distance is $R\theta$. The ratio of the cap integral to the full integral is the stated probability. At $\rho = \pi R$ the missing antipode has zero mass. □

When the general posterior and radius assumptions also hold, their reach–noise retention bound is valid but conservative on the uniform sphere: symmetry gives zero projected drift directly. The zero-drift result itself needs no localization constant once finite entrance is known. Nonuniform densities or finite center clouds need not preserve this cancellation, which is why E1 tests genuine projected motion.

Remark E.15 (Population-to-empirical scope). The exact sphere statements concern a uniform population, not a fixed random center cloud. A uniform empirical transfer requires separate control of posterior moments and landscape perturbations. The h_M^2 extension is a conditional way to express a moment residual; E1's nearest-center quantile is a sampling-resolution diagnostic, not a proof of that residual bound.

E.3 FINITE-SCHEDULE REFERENCE CALCULATIONS FOR E4

The E4 curves use the same precision-normalized weights as the preceding analysis. We record their finite-schedule forms to distinguish these references from both the unbounded-scale limit and fitted empirical curves.

The flat finite-schedule moment. For any fixed positive finite schedule with equal scale sampling, integrating the tangent Gaussian in the flat model gives

$$\pi_i^{\mathrm{flat}}(r) = \frac{\sigma_i^{-c}\exp[-r^2/(2\sigma_i^2)]}{\sum_{a=1}^N \sigma_a^{-c}\exp[-r^2/(2\sigma_a^2)]}, \qquad \mathbb{E}_{\widetilde{w},\mathrm{flat},r}\|\mu - z_x\|^2 = m\sum_{i=1}^N \pi_i^{\mathrm{flat}}(r)\sigma_i^2. \tag{E.18}$$

Thus the radius-dependent reference plotted in E4 is

$$\mathcal{C}_{\mathrm{flat},N}(r) := \frac{\sum_{i=1}^N \pi_i^{\mathrm{flat}}(r)\sigma_i^2}{\max\{\sigma_{\min}^2, r^2/c\}}. \tag{E.19}$$

This is an exact ratio for the translation-invariant flat posterior, not a normalizable flat data distribution. It is also a function of r, distinct from a uniform constant C_{flat} used in Proposition E.6. The sphere comparison supplies the factor R/ϱ for a population moment; it does not make (E.19) an empirical upper bound or predict an exact finite-cloud median.

Small-noise radial bias on the uniform two-sphere. For the proper population $S^2(R)$ model used in E4, fix a finite relative schedule $1 = q_1 \le \cdots \le q_N < \infty$ and write $\sigma_i = sq_i$, with $s = \sigma_{\min}$. At $x = \varrho e \in V$, let $a_i = \varrho R/\sigma_i^2$. The variable $t = \langle e, \mu/R\rangle$ is uniform on $[-1,1]$ before conditioning. Direct integration gives

$$\begin{aligned} I_i(\varrho) &:= \int_{S^2(R)} \mathcal{N}(\varrho e;\mu,\sigma_i^2 I) Q_{\mathcal{M}}(d\mu) \\ &= (2\pi\sigma_i^2)^{-d/2} e^{-(\varrho-R)^2/(2\sigma_i^2)} \frac{\sigma_i^2}{2\varrho R}\left(1 - e^{-2a_i}\right). \end{aligned} \tag{E.20}$$

The mean radial coordinate conditional on scale is $R(\coth a_i - a_i^{-1})$. Consequently the normalized radial field is

$$F_s(\varrho) = R\sum_{i=1}^N \pi_i^{\mathrm{sph}}(\varrho)(\coth a_i - a_i^{-1}) - \varrho, \qquad \pi_i^{\mathrm{sph}}(\varrho) = \frac{I_i(\varrho)}{\sum_{a=1}^N I_a(\varrho)}. \tag{E.21}$$

These are precision-normalized scale probabilities: the extra σ_i^2 in the unified density has canceled against the score precision. Since $\beta > 0$, zeros of F_s are zeros of the original radial score as well.

To examine the near-sphere branch, put $\varrho = R - bs^2/R$ with bounded b. For the fixed finite relative schedule, $s/R \to 0$ implies $a_i \to \infty$ at every scale. Equations (E.20)–(E.21), together with $\coth a_i = 1 + O(e^{-2a_i})$, give

$$\pi_i^{\mathrm{sph}}(R - bs^2/R) = \frac{q_i^{-c}}{\sum_{a=1}^N q_a^{-c}} + O((s/R)^2), \qquad c = d-2, \tag{E.22}$$

and hence, uniformly for b in a fixed bounded neighborhood of the limiting root,

$$\frac{R}{s^2}F_s(R - bs^2/R) = b - \frac{\sum_i q_i^{2-c}}{\sum_i q_i^{-c}} + O((s/R)^2). \tag{E.23}$$

The expansion also holds after differentiation in b: it follows directly from the finite sums above, whose omitted exponential terms and their derivatives vanish faster than any power. The leading root is positive and simple, so there is a local stationary branch with

$$R - \varrho_\star = \frac{s^2}{R}\left[\frac{\sum_i q_i^{2-c}}{\sum_i q_i^{-c}} + O((s/R)^2)\right]. \tag{E.24}$$

For $m = 2$, the leading constant can therefore be written as

$$\frac{R|\varrho_\star - R|}{\sigma_{\min}^2} \longrightarrow C_{\mathrm{fs}} = \frac{m}{2} \frac{\sum_{i=1}^{N} q_i^{2-c}}{\sum_{i=1}^{N} q_i^{-c}} = 1.199652317\ldots, \tag{E.25}$$

where the numerical value substitutes $c = 10$, $N = 64$, and $q_i = 16^{(i-1)/63}$. This calculation concerns a proper population sphere with a fixed finite relative schedule, all of whose scales shrink in the limit. It is not an assertion that the general finite-cloud localization bound is attained, nor does it identify a sampled value of C_{loc} with C_{fs}.

F Numerical Verification: Full Settings and Results

The closed-form calculations above determine the flat-model localization constant and the population log-density Hessian spectrum at the sphere center exactly. Combined with the general results of Appendix A, they establish the predicted quadratic localization in $\sigma_{\min}$ whenever the standing landscape assumptions hold. The experiments below address the complementary questions that are not settled by those calculations: finite samples, coarse scale schedules, non-uniform target densities, and the global off-tube critical structure. Experiment E1 tests finite-resolution regional probability allocation and local tangential retention, while Experiment E2 tests population radial-score positivity and the instability geometry of the off-tube critical points located in finite empirical mixtures. Experiment E3 examines positive-bandwidth separation and sample-resolution core coverage, while Experiment E4 compares spherical bias and sampled empirical posterior moments with finite-schedule references.

For the empirical mixtures, the score (A.4) is evaluated by the exact finite log-sum-exp over all scale–center pairs, up to floating-point error. The remaining numerical approximations are the Euler time discretization and event-location error in E1, and quadrature and nonlinear root-finding error in E2. E3 additionally records finite-tolerance termination diagnostics, while E4 separates population quadrature from sampled finite-cloud moment calculations. Trajectories in E1 and E3 are integrated by

$$x \leftarrow x + \eta\big(\bar{\mu}(x) - x\big),$$

an Euler discretization of the time-reparameterized flow $\dot{x} = \bar{\mu}(x) - x$. Its continuous-time trajectories and hitting locations coincide with those of (A.8); sensitivity to the step size η is reported explicitly below.

Lemma F.1 (Ascent of the normalized Euler update)**.** *For the exact finite-mixture field, let*

$$x^+ = x + \eta\big(\bar{\mu}(x) - x\big). \tag{F.1}$$

Then, in exact arithmetic,

$$\log p_{MN}(x^+) - \log p_{MN}(x) \geq \beta(x)\eta\left(1 - \frac{\eta}{2}\right)\|\bar{\mu}(x) - x\|^2. \tag{F.2}$$

In particular, $0 < \eta < 2$ gives strict ascent unless x is stationary. If the iteration uses steps $0 < \underline{\eta} \leq \eta_n \leq \overline{\eta} < 2$, then, from every starting point, it has finite total length and converges to a stationary point of $\log p_{MN}$.

Proof. For any displacement Δ, the Gaussian density ratio and Jensen's inequality give

$$\begin{aligned}\log \frac{p_{MN}(x+\Delta)}{p_{MN}(x)} &= \log \sum_{j,i} w_{ij}(x) \exp\left(\frac{\langle \mu_j - x, \Delta\rangle}{\sigma_i^2} - \frac{\|\Delta\|^2}{2\sigma_i^2}\right) \\ &\geq \langle \nabla \log p_{MN}(x), \Delta\rangle - \frac{\beta(x)}{2}\|\Delta\|^2.\end{aligned}$$

Substituting $\Delta = \eta(\bar{\mu}(x) - x)$ and using $\nabla \log p_{MN}(x) = \beta(x)(\bar{\mu}(x) - x)$ proves (F.2).

For the convergence statement, let $\Delta_n = x_{n+1} - x_n$. The step bounds and $\sigma_{\max}^{-2} \leq \beta(x) \leq \sigma_{\min}^{-2}$ imply

$$\log p_{MN}(x_{n+1}) - \log p_{MN}(x_n) \geq \sigma_{\max}^{-2}\left(\overline{\eta}^{-1} - \frac{1}{2}\right)\|\Delta_n\|^2, \qquad \|\nabla \log p_{MN}(x_n)\| \leq \frac{\sigma_{\min}^{-2}}{\underline{\eta}}\|\Delta_n\|.$$

All iterates remain in the compact initial superlevel set. The two estimates imply vanishing steps and critical accumulation points. Applying the analytic-gradient inequality on the compact accumulation set to the same estimates gives $\sum_n \|\Delta_n\| < \infty$, and therefore convergence to a stationary point. □

The step sizes used in E1 and E3 lie in this ascent regime. This result justifies the normalized update as an ascent method for the exact objective; it does not identify its finite-step hitting locations or output probabilities with those of the continuous flow. The step-size and termination diagnostics below address these remaining numerical effects.

F.1 E1: Regional probability beyond symmetry

Setup. We use the sphere $S^2(1) \subset \mathbb{R}^{12}$, so $m = 2$ and $c = 10$, with $M = 10^4$ centers drawn from either (i) a von Mises–Fisher density with concentration $\kappa \in \{1, 4, 16\}$, or (ii) a two-cluster mixture of antipodal $\mathrm{vMF}(8)$ components with masses $(p, 1-p)$, $p \in \{0.5, 0.9\}$. The geometric schedule contains 24 scales on $[0.02, 2]$. We take

$$r_0 = 0.3, \qquad r_{\mathrm{tr}} = 0.2, \qquad r_{\mathrm{stop}} = 0.03.$$

Using the population benchmark constant specified below, the reference values associated with (B.3) are

$$r_{\min} = \sigma_{\min}\sqrt{c} \approx 0.063, \qquad \varepsilon_0 \approx 5 \times 10^{-4}.$$

For the unbounded local-limit benchmark $C_{\mathrm{loc}} = c/(c-2) = 1.25$ from Proposition E.6, the constants are

$$C_{\mathrm{loc}}/2 = 0.625, \qquad L_\Pi = 1.25, \qquad \kappa_0 = 0.0375, \qquad \varepsilon_0 = 0.0005.$$

Consequently (B.10) evaluates to

$$\begin{aligned}\delta_{\mathrm{stop}}^{\mathrm{pop}} &= 1.25\left[\frac{0.5}{0.9625}(0.2 - 0.02\sqrt{10}) + 0.02\sqrt{2.5}\log\frac{0.02\sqrt{10} - 0.0005}{0.03 - 0.0005}\right] \\ &= 0.1186336\ldots \approx 0.119.\end{aligned}$$

Replacing the benchmark constant by $1.25\,R/(R - r_{\mathrm{tr}}) = 1.5625$ and recomputing the same formula gives the curvature-inflated reference value 0.134. These numbers are theory-guided benchmarks for the finite empirical experiment, rather than uniform certificates for its realized center cloud and coarse schedule.

We run 10^5 trajectories from the truncated Gaussian initialization (A.18) with $\delta_{\mathrm{init}} = 10^{-3}$ and default step size $\eta = 0.15$. We record Z_{tr} at the first crossing of $r(x) \le r_{\mathrm{tr}}$ and Z_{stop} at the first subsequent crossing of $r(x) \le r_{\mathrm{stop}}$. Since every center lies in $\mathbb{R}^3 \times \{0\}$, the Euler update closes over the pair $(P_V x, \|(I - P_V)x\|)$, which we exploit exactly.

For the population vMF model, rotational symmetry about the vMF axis makes the regional probability ratio one-dimensional in the colatitude. A fixed finite center cloud has only approximate rotational symmetry, so the same statistic should be interpreted as an axial, finite-resolution probability diagnostic. Let $\mathcal{P}$ denote the prescribed partition of the sphere. For each cell $B \in \mathcal{P}$, let $\underline{p}_B$ be the lower endpoint of the two-sided 95% Wilson interval for the transition probability estimated from the sampled trajectories. We report

$$\widehat{\alpha}_{\mathcal{P}} := \min_{B \in \mathcal{P}} \frac{\underline{p}_B}{Q_{\mathcal{M}}(B)}. \tag{F.3}$$

For the vMF targets, $\mathcal{P}$ consists of twelve equal- $Q_{\mathcal{M}}$-mass colatitude bins, constructed using the closed-form inverse CDF of the $t = \cos\theta$ marginal. For the cluster mixtures, $\mathcal{P}$ consists of the two hemispheres centered at the component poles. If $\widehat{\mathbb{P}}$ denotes the empirical trajectory distribution, then $\underline{p}_B \le \widehat{\mathbb{P}}[Z_{\mathrm{tr}} \in B]$ for every cell and therefore

$$\widehat{\mathbb{P}}[Z_{\mathrm{tr}} \in A] \ge \widehat{\alpha}_{\mathcal{P}} Q_{\mathcal{M}}(A) \qquad \text{for every union } A \text{ of cells of } \mathcal{P}. \tag{F.4}$$

Thus $\widehat{\alpha}_{\mathcal{P}}$ directly quantifies finite-resolution regional routing on the partition.

For retention, we report the 99% and 99.9% quantiles of

$$d_{\mathcal{M}}\big(Z_{\mathrm{stop}}, Z_{\mathrm{tr}}\big).$$

These quantiles test the high-probability analogue of the deterministic retention bound in Proposition B.2. We also report the sampling-resolution statistic $\varepsilon_M^{\mathrm{loc}}$, defined as the $Q_{\mathcal{M}}$-weighted 99% quantile of the ambient Euclidean distance to the nearest training center. This is a quantile-based local coverage radius, not a certificate of the posterior-moment residual h_M in Appendix G.1. An ablation at $\kappa = 4$ with $M \in \{10^3, 5 \cdot 10^3, 10^4\}$, using three independent center seeds at each sample size, tests whether the observed finite-sample resolution is consistent with the expected $M^{-1/2}$ behavior on a two-dimensional manifold.

A high-probability displacement bound gives a coverage bound by subtracting the probability of its failure event.

A population 99.9% retention guarantee would lose at most 10^{-3} in the coverage lower bound. For an exact empirical statement, let $\widehat{\delta}_{0.999}$ be the upper empirical quantile of the unrounded displacements, that is, the smallest observed threshold for which at least 99.9% of trajectories have displacement at most that threshold. This order-statistic definition, rather than a rounded table entry or an arbitrary interpolation convention, guarantees

$$\widehat{\mathbb{P}}[d_{\mathcal{M}}(Z_{\mathrm{stop}}, Z_{\mathrm{tr}}) \le \widehat{\delta}_{0.999}] \ge 0.999.$$

For any Borel region U, retain only partition cells entirely inside its safe interior. The event-inclusion proof and (F.4) give

$$\widehat{\mathbb{P}}[Z_{\mathrm{stop}} \in U] \ge \widehat{\alpha}_{\mathcal{P}}\, Q_{\mathcal{M}}\left(\bigcup_{\substack{B \in \mathcal{P}: \\ B \subseteq U^{-\widehat{\delta}_{0.999}}}} B \right) - 10^{-3}. \tag{F.5}$$

When the safe interior is itself a union of cells, the union equals $U^{-\widehat{\delta}_{0.999}}$. This is a finite-resolution statement about the empirical trajectory distribution, not a population confidence guarantee. The printed p99 and p99.9 values below remain the original numerical summaries; rounding does not certify an exact exceedance count at a printed threshold.

Results. Table 4 summarizes the five target distributions. At the tested partition resolution, the regional probability lower bound remains uniformly bounded away from zero: $\widehat{\alpha}_{\mathcal{P}} \ge 0.789$ across all vMF concentrations and $\widehat{\alpha}_{\mathcal{P}} \ge 0.937$ for the cluster mixtures. The binding cell lies at or near the high-density pole region for the vMF targets and in one of the two component hemispheres for the cluster targets; for $p = 0.9$, the measured mass split is 0.846. The ratio curves are consistent with mild broad-scale flattening: the transition distribution undersamples the highest- density cells relative to $Q_{\mathcal{M}}$, while lower-density cells are oversampled. The high-density cells determine the reported lower bound; oversampling the other cells is not penalized by the one-sided criterion. The strictly positive values of $\widehat{\alpha}_{\mathcal{P}}$ provide direct finite-resolution evidence for regional probability allocation beyond the exact symmetry setting of Proposition E.13.

The retention quantiles are also consistent with the predicted scale. Across all configurations, the p99.9 displacement is at most 0.112, below both the population benchmark $\delta_{\mathrm{stop}}^{\mathrm{pop}} = 0.119$ and the curvature-inflated reference value 0.134. Together with (F.5), this shows that the observed regional allocation is largely retained through refinement at the tested resolution.

The M-ablation in Figure 4 examines the observed dependence on sampling resolution. The measured $\varepsilon_M^{\mathrm{loc}}$ contracts by a factor 3.27 from $M = 10^3$ to $M = 10^4$, consistent with the $M^{-1/2}$ reference rate $\sqrt{10} \approx 3.16$. Across the three tested sample sizes, the descriptive affine fit is

$$\mathrm{p99} \approx 0.037 + 0.31\, \varepsilon_M^{\mathrm{loc}}.$$

The intercept is an extrapolation toward vanishing sampling-resolution radius, while the second term records the observed finite-sample dependence; the fit is descriptive rather than a decomposition asserted by the theory. At $M = 10^3$, the empirical quantile 0.108 ± 0.003 over three center seeds nearly reaches the 0.119 benchmark, whereas at $M = 10^4$ the seed-to-seed spread of both $\widehat{\alpha}_{\mathcal{P}}$ and the retention quantile is below 2%. These observations are consistent with the sampling- resolution limitation discussed in Appendix G.1.

Reducing the step size by a factor of three, $\eta = 0.15 \to 0.05$, with identical centers and initializations, changes $\widehat{\alpha}_{\mathcal{P}}$ by 0.04 and the retention quantile by 6%. This sensitivity is not negligible at the scale of the narrowest numerical margins, but it leaves the two qualitative conclusions unchanged: partition-level coverage remains bounded away from zero, and the upper retention quantiles remain on the scale predicted by (B.10).

F.2 E2: Off-tube critical structure

Setup. We probe the off-tube geometry of the multi-scale score field in two complementary ways, with $R = 1$ throughout.

Table 4: E1: finite-resolution probability and retention across target densities ($M = 10^4$, 10^5 trajectories; every trajectory reached both stopping surfaces; $\delta_{\mathrm{stop}}^{\mathrm{pop}} = 0.119$). $\widehat{\alpha}_{\mathcal{P}}$ is the minimum, over twelve equal-$Q_{\mathcal{M}}$-mass colatitude bins, of the marginal two-sided 95% Wilson lower endpoint for the empirical bin mass divided by its target mass; the binding bin is reported in $t = \cos\theta$. For the cluster mixtures, the partition consists of the two component hemispheres, and the measured mass split is also reported. The Wilson intervals are conditional on the realized center cloud and are not a simultaneous confidence band.

Density	$\widehat{\alpha}_{\mathcal{P}}$	binding bin	split	p99	p99.9
vMF, $\kappa = 1$	0.789	$[0.93, 1.00]$	—	0.039	0.052
vMF, $\kappa = 4$	0.799	$[0.98, 1.00]$	—	0.058	0.106
vMF, $\kappa = 16$	0.852	$[0.989, 0.995]$	—	0.046	0.085
clusters, $p = 0.5$	0.987	binding hemisphere	0.503	0.056	0.109
clusters, $p = 0.9$	0.937	heavier-component hemisphere	0.846	0.059	0.112

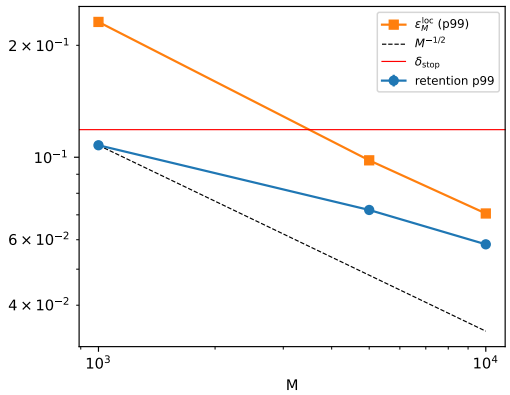

Figure 4: E1 ablation ($\kappa = 4$, three center seeds per M): retention quantile p99 and the measured 99% nearest-center coverage radius $\varepsilon_M^{\mathrm{loc}}$ versus M, with an $M^{-1/2}$ guide and the idealized population benchmark $\delta_{\mathrm{stop}}^{\mathrm{pop}}$. The displayed affine relation p99 $\approx 0.037 + 0.31\,\varepsilon_M^{\mathrm{loc}}$ is a descriptive fit over the three tested sample sizes.

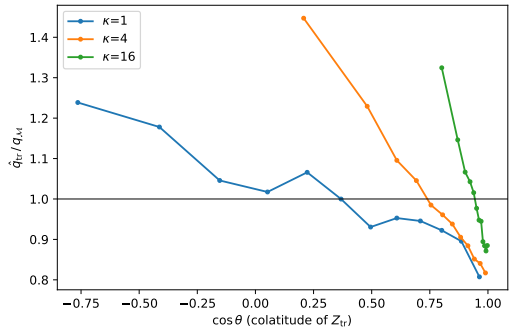

Figure 5: E1: binned density ratio $\widehat{q}_{\mathrm{tr}}/q_{\mathcal{M}}$ versus $\cos\theta$ on equal-$Q_{\mathcal{M}}$-mass bins, each plotted at the $\cos\theta$ value splitting its target mass in half, for vMF targets with $\kappa \in \{1, 4, 16\}$. The ratio is smallest in or near the high-density pole region and exceeds one in the tails, a pattern consistent with mild flattening by the broad scales.

First, for the population uniform measure on $S^m(1) \subset \mathbb{R}^{m+c}$, we evaluate the radial score $g(\varrho)$ of Lemma E.10 by summing over a finite geometric scale schedule and applying Gauss–Legendre quadrature in colatitude. The baseline uses 200 scales on $[0.01, 2]$ and 400 quadrature nodes; we

also test a coarser 40-scale schedule on $[0.01, 2]$ and a wider 200-scale schedule on $[0.005, 4]$. We evaluate positivity on $(0, R - r_0]$ with $r_0 = 0.3$.

Second, for empirical mixtures with $M = 2000$ centers on $S^2(1) \subset \mathbb{R}^{12}$, sampled either uniformly or from $\mathrm{vMF}(\kappa = 4)$, we search for zeros of the exact score using a damped Gauss–Newton (Levenberg–Marquardt) iteration. At each $\sigma_{\min} \in \{0.2, 0.1, 0.05, 0.02\}$, we use 24 geometrically spaced scales on $[\sigma_{\min}, 2]$, keeping the center cloud fixed throughout the coarse-to-fine sweep.

At every level, we draw 2048 fresh off-tube initializations: the direction is uniform on the ambient unit sphere, the ambient radius is sampled uniformly from $[0.02, 2R]$, and only points with $r(x) > r_0$ are retained. Deduplicated roots from the preceding noise level are additionally used as warm starts. Writing $s(x) = \nabla \log p_{MN}(x)$ and $J(x) = \nabla^2 \log p_{MN}(x)$, each LM step solves

$$\left[J^\top J + \lambda \bar{d}_J I\right] p = J^\top s, \qquad \bar{d}_J = \frac{1}{d} \operatorname{tr}(J^\top J), \tag{F.6}$$

and proposes $x^+ = x - p$. The proposal is accepted when the normalized residual decreases. The damping starts at 10^{-6}, is multiplied by 0.3 after acceptance and by 10 after rejection, and the search is run for at most 150 iterations. A numerical damping floor of 10^{-14} is used, and runs with $\|x\| > 3\sigma_{\max}$ or $\lambda > 10^{10}$ are marked as diverged or stalled. Convergence is declared when

$$\frac{\|\nabla \log p_{MN}(x)\|}{\beta(x)} = \|\bar{\mu}(x) - x\| \le 10^{-10}. \tag{F.7}$$

Converged candidates are greedily deduplicated in ambient Euclidean distance with tolerance 10^{-5}; we additionally sweep tolerances from 10^{-7} to 10^{-2}.

At each located critical point we evaluate the analytic log-density Hessian

$$\nabla^2 \log p_{MN}(x) = C_x - \beta(x) I \tag{F.8}$$

and record its spectrum. We also measure the precision-normalized posterior spread

$$\Sigma_x := \sum_{j=1}^{M} \sum_{i=1}^{N} \widetilde{w}_{ij}(x)(\mu_j - x)(\mu_j - x)^\top. \tag{F.9}$$

This quantity captures the geometry of the active centers but is distinct from the component-score covariance C_x in Assumption A.10. At a critical point,

$$C_x = \beta(x) \sum_{j,i} \widetilde{w}_{ij}(x) \frac{(\mu_j - x)(\mu_j - x)^\top}{\sigma_i^2}. \tag{F.10}$$

At the population center of the uniform sphere, center and scale posteriors factorize, so $C_x \propto \Sigma_x$ and their effective ranks coincide.

For a complementary center-utilization diagnostic, define the center-marginal responsibility

$$q_j(x) = \frac{\sum_i \sigma_i^2 \mathcal{N}(x; \mu_j, \sigma_i^2 I)}{\sum_{\ell,i} \sigma_i^2 \mathcal{N}(x; \mu_\ell, \sigma_i^2 I)}, \qquad \mathrm{ESS}(x) = \frac{1}{\sum_j q_j(x)^2}. \tag{F.11}$$

Finally, when all centers lie on $S^m(R)$,

$$\operatorname{tr}(\Sigma_x) = R^2 - \|x\|^2 - 2\left\langle x, \frac{\nabla \log p_{MN}(x)}{\beta(x)} \right\rangle, \tag{F.12}$$

$$\operatorname{tr}(\Sigma_x) = R^2 - \|x\|^2 \qquad \text{at a critical point.} \tag{F.13}$$

We use this identity as a numerical consistency check.

(i) Population radial-score positivity. For every tested pair $(m, c) \in \{1, 2, 4\} \times \{1, 2, 3, 4, 8, 16\}$ and all three scale schedules, $g(\varrho) > 0$ at every node of the 120-point grid on $[10^{-3}, R - r_0]$ (Figure 6). The minimum always occurs at the smallest tested radius, consistent with linear vanishing near the origin for $c \ge 2$.

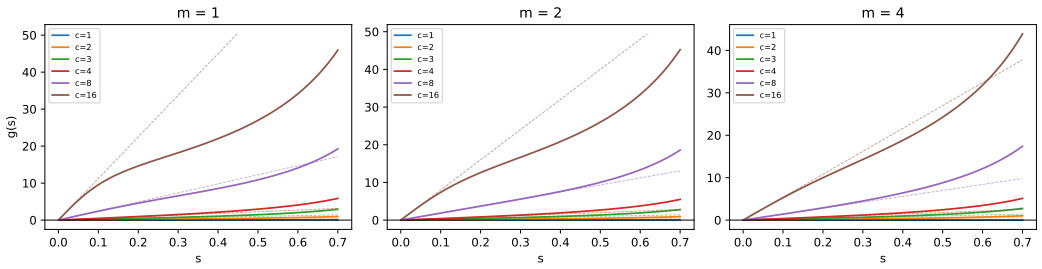

Figure 6: **Population off-tube radial score.** Radial score $g(\varrho)$ on $[10^{-3}, R-r_0]$ for $m \in \{1,2,4\}$ and $c \in \{1,2,3,4,8,16\}$. The finite scale sum is combined with Gauss–Legendre quadrature in colatitude. Dashed lines show the post-hoc small-ϱ reference $k(m,c)\varrho$, where $k = (c-1)(c+m-2)/(m+1)$. The score is positive at every tested grid node.

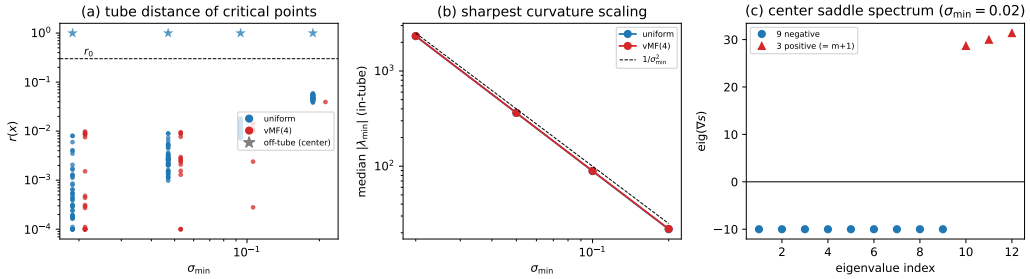

Figure 7: **Critical structure of empirical sphere mixtures.** $M = 2000$ centers lie on $S^2(1) \subset \mathbb{R}^{12}$. (a) Tube distance of all located deduplicated critical points across the $\sigma_{\min}$ sweep; the only located off-tube point is the near-center saddle of the uniform cloud. (b) Median magnitude of the most negative Hessian eigenvalue for in-tube critical points, compared with the $1/\sigma_{\min}^2$ scale. (c) Hessian spectrum of the near-center uniform-cloud saddle, with three positive and nine negative eigenvalues.

The post-hoc small-ϱ slope, estimated from the first six grid nodes, approaches

$$k(m,c) = \frac{(c-1)(c+m-2)}{m+1}, \tag{F.14}$$

which equals the unstable eigenvalue λ_+ in Proposition E.8 at $R = 1$. The agreement is within 0.15% for $c \geq 8$; for example, the measured and predicted values are 112.38 and 112.5 at $(m,c) = (1,16)$ and 14.0006 and 14 at $(m,c) = (4,8)$. At moderate codimension, the finite-schedule excess decreases as the scale range widens: for $(m,c) = (1,2)$, the ratio of measured to predicted slope decreases from 1.70 to 1.25 when the schedule is extended from $[0.01, 2]$ to $[0.005, 4]$.

The $c = 1$ cases show the expected schedule sensitivity. Their measured slope remains positive for all three tested schedules, but decreases toward the continuum knife edge as the range widens; at $m = 4$, for example, it decreases from 4.3×10^{-3} to 1.4×10^{-4}. For $c \geq 2$, these results support the radial positivity mechanism used to exclude noncentral off-tube critical points in the sphere model.

(ii) Empirical critical points. For the uniform cloud, the search locates one deduplicated off-tube critical point at every $\sigma_{\min}$ and no second off-tube root (Figure 7a). It is a finite-sample perturbation of the population center, with $\|x\| \approx 3 \times 10^{-3}$ and hence $r(x) \approx R$. Across the sweep, this root receives 685–742 converged solver starts and remains a single root throughout the deduplication-tolerance sweep.

Its local geometry closely matches the population prediction. The log-density Hessian has $m+1 = 3$ positive and nine negative eigenvalues (Figure 7c), giving the predicted three-dimensional ascent-unstable subspace. The posterior-spread diagnostics give

$$\mathrm{tr}(\Sigma_x) = 0.99999 \approx R^2, \qquad r_{\mathrm{eff}}(\Sigma_x) = 2.90 \approx m+1, \tag{F.15}$$

with $r_{\text{eff}}(\Sigma_x)/(m+1) = 0.97$. The center-marginal ESS is approximately 1999 out of 2000, showing that the near-center saddle is supported by essentially the entire uniform cloud rather than a single training center or a small local cluster.

Every other located critical point lies well inside the tube. For the uniform cloud, $r(x) \le 0.059$; for the vMF cloud, $r(x) \le 0.040$ at $\sigma_{\min} = 0.2$ and $r(x) \le 0.015$ at all sharper scales. Their center ESS ranges from approximately 1 for sample-dominated modes to approximately 176. The median most negative Hessian eigenvalues across the four lower cutoffs are

$$-21.8,\ -88.7,\ -362,\ -2.33 \times 10^3, \tag{F.16}$$

close to the finest-scale reference $-1/\sigma_{\min}^2 = -25, -100, -400, -2500$.

For the $\mathrm{vMF}(4)$ cloud, no off-tube critical point is located at any tested $\sigma_{\min}$. The score at the origin instead points toward the density pole. At $\sigma_{\min} = 0.2$, the located critical structure consists only of the density mode near the pole, at polar angle 3.1°. At finer scales, additional in-tube finite-sample structures appear predominantly in lower-density regions; at $\sigma_{\min} = 0.02$, 16 of the 20 located points lie beyond 50° from the pole.

Summary and scope. The two parts give complementary evidence for the global landscape picture. Population radial scores remain positive throughout the tested off-tube grid, while the off-tube critical point located in the finite uniform mixture has precisely the predicted ascent-unstable geometry. Breaking uniform symmetry with a vMF cloud removes the center critical point, and the search locates no replacement outside the tube.

The empirical root search is not exhaustive, particularly at the finest vMF setting where cold-start convergence becomes sparse; coarse-to-fine continuation substantially improves the root yield. We therefore use E2 as numerical support for the off-tube instability mechanism rather than as a census of all finite-mixture critical points. Together with E1, these results support the geometric ingredients of the convergence and coverage analysis in finite, nonuniform settings.

F.3 E3: Positive-bandwidth separation and sample-resolution diversity

Setup and exact field evaluation. We take $\mathcal{A} = \mathbb{R} \times \{0\} \subset \mathbb{R}^3$ and the interlaced grid (20) with $L = 20$ and $\ell = 1$. Thus $M = 82$, $Q_{\mathcal{M}}$ is uniform on $[-20.5, 20.5] \times \{0\}$, and there are 41 reference cells with output cores $C_k = \{(u, 0) : |u - k| \le 1/12\}$. The training covering radius and the separation threshold are, respectively, $h_{\mathrm{cov},M} = 1/3$ and $h_{\mathrm{cov},M}/4 = 1/12$. For each $s = \sigma_{\min}/\ell$ listed in Table 5, we use

$$\sigma_i = s\, 128^{(i-1)/23}, \quad i = 1, \dots, 24, \qquad X_0 \sim \mathcal{N}(0, (128s)^2 I_3). \tag{F.17}$$

The relative schedule is fixed across the sweep. The absolute upper cutoff and the Gaussian initialization variance therefore change with s as well; this is not a fixed-initialization ablation of the lower cutoff alone.

At each setting we simulate $n_{\mathrm{traj}} = 8192$ initializations using the update $x_{n+1} = x_n + \eta(\bar{\mu}(x_n) - x_n)$, with $\eta = 0.2$ in the main sweep. Gaussian sums are evaluated in PyTorch single precision using stabilized log weights. At a fixed scale, the center sum on the Cartesian-product grid factorizes into one-dimensional sums; the scale contributions are then combined with their precision-normalized responsibilities. This computes the original finite-mixture barycenter, up to floating-point error, rather than replacing it by the infinite periodic comparison function. No neural network is trained.

Termination and numerical output. The implementation checks $\max_a \|\bar{\mu}(x_n^{(a)}) - x_n^{(a)}\|$ every 50 iterations, with threshold 10^{-5} and a cap of 5000 updates. The residual is evaluated before the checked update, and the returned state includes that update. We denote the returned numerical output by X_{num}, distinguishing it from the exact limit x_∞. Every reported run reaches the threshold before the cap. In the main sweep the update counts range from 101 to 2451, and the largest recorded maximum residual is 9.72×10^{-6}.

Metrics at the training resolution. Let $X^{(a)}_{\mathrm{num}}$ be the returned ambient outputs and $U^{(a)} = \Pi_{\mathcal{A}}(X^{(a)}_{\mathrm{num}})$ their affine projections. The separation statistic and output-core frequencies are

$$\widehat{p}_{\mathrm{sep}} = \frac{1}{n_{\mathrm{traj}}} \sum_{a=1}^{n_{\mathrm{traj}}} \mathbf{1}\Big\{ \mathrm{dist}(X^{(a)}_{\mathrm{num}}, \mathcal{S}_M) \geq 1/12 \Big\}, \tag{F.18}$$

$$\widehat{p}_k = \frac{1}{n_{\mathrm{traj}}} \sum_{a=1}^{n_{\mathrm{traj}}} \mathbf{1}\{U^{(a)} \in C_k\}, \qquad \widehat{p}_{\mathrm{out}} = 1 - \sum_{k=-20}^{20} \widehat{p}_k. \tag{F.19}$$

The categorical reference assigns mass $1/41$ to each reference-cell label and zero to the outside-core label. As in (22), reference labels describe the full cells of $Q_{\mathcal{M}}$, whereas output labels describe the smaller cores; the comparison is not with $Q_{\mathcal{M}}(C_k)$. We report

$$\widehat{\epsilon}_{\mathrm{cell}} = \frac{1}{2} \left[\sum_{k=-20}^{20} |\widehat{p}_k - 1/41| + \widehat{p}_{\mathrm{out}} \right], \qquad \widehat{n}_{\mathrm{pop}} = \#\{k : \widehat{p}_k \geq 1/82\}. \tag{F.20}$$

The populated-core fraction is $\widehat{n}_{\mathrm{pop}}/41$. All 8192 outputs enter the denominators; neither the separation statistic nor the core-label distribution is conditioned on the other. We also report empirical quantiles of the ambient distance to $\mathcal{S}_M$.

Membership of a central core implies the training-distance margin, but the reverse implication does not hold for arbitrary locations between cells. We therefore retain separation and core mass as distinct diagnostics. For example, at $s = 0.26$ their measured values are 0.9583 and 0.9319. Likewise, zero populated *cores* at small bandwidth does not mean that outputs occupy only one full reference cell: sample-associated modes can remain distributed over the grid while missing the prescribed off-training cores.

Bandwidth sweep. Table 5 reports the main sweep. For the tested $s \leq 0.12$, no output meets the separation threshold or lies in a central core. At $s = 0.13$, the separated fraction is 0.5875; it reaches 0.9780 at $s = 0.14$. The sampled bandwidths from 0.15 through 0.24 retain 39 populated cores. At $s \in \{0.18, 0.20, 0.22, 0.24\}$, separation is 0.9679–0.9734, core-label TV is 0.0888–0.1060, and the fifth-percentile distance is 0.0992–0.1228, above $1/12$. Increasing s to 0.30 gives TV 0.1623, 35 populated cores, and separation 0.8916. Thus the strongest joint separation–coverage behavior occurs over an intermediate bandwidth range rather than varying monotonically with smoothing.

Table 5: **E3 bandwidth sweep.** Each row uses 8192 numerical outputs with $\eta = 0.2$. The separation threshold is $1/12$; core mass is $1 - \widehat{p}_{\mathrm{out}}$; $\widehat{n}_{\mathrm{pop}}$ counts the 41 cores whose estimated mass is at least $1/82$; $q_{0.05}$ is the empirical fifth-percentile distance to the training set. All rows meet the recorded residual threshold 10^{-5}. Statistics are estimates for the simulated outputs, not confidence bounds for the exact flow.

s	$\widehat{p}_{\mathrm{sep}}$	core mass	$\widehat{\epsilon}_{\mathrm{cell}}$	$\widehat{n}_{\mathrm{pop}}$	$q_{0.05}$	updates
0.05	0.0000	0.0000	1.0000	0	0.00250	101
0.08	0.0000	0.0000	1.0000	0	0.01096	151
0.10	0.0000	0.0000	1.0000	0	0.02388	151
0.11	0.0000	0.0000	1.0000	0	0.03499	251
0.12	0.0000	0.0000	1.0000	0	0.05004	301
0.13	0.5875	0.5875	0.4779	22	0.06941	751
0.14	0.9780	0.9780	0.1325	38	0.09861	2451
0.15	0.9753	0.9753	0.1237	39	0.10467	451
0.16	0.9780	0.9780	0.1347	39	0.12682	301
0.18	0.9734	0.9734	0.1060	39	0.12276	251
0.20	0.9729	0.9729	0.0888	39	0.11978	201
0.22	0.9701	0.9701	0.0934	39	0.11185	251
0.24	0.9679	0.9679	0.0901	39	0.09919	301
0.26	0.9583	0.9319	0.1173	37	0.09669	701
0.28	0.9305	0.9305	0.1187	37	0.05256	301
0.30	0.8916	0.8916	0.1623	35	0.00687	351

Folded visualization. For a returned tangential coordinate u, Figure 8 displays

$$v = \frac{u}{\ell} - \left\lfloor \frac{u}{\ell} + \frac{1}{2} \right\rfloor \in [-1/2, 1/2). \tag{F.21}$$

The training locations fold to $\pm 1/6$, the cell center to zero, and core boundaries to $\pm 1/12$. The displayed histograms pool all returned outputs across the full finite grid, without an interior-cell filter. At $s = 0.10$ they retain two sample-associated peaks; at $s = 0.18$ and 0.24 most mass is near the central part of the cores. Finite-array effects and variation between cells remain visible after folding. Folding is used only for visualization; all separation and diversity metrics above are computed on the original finite grid.

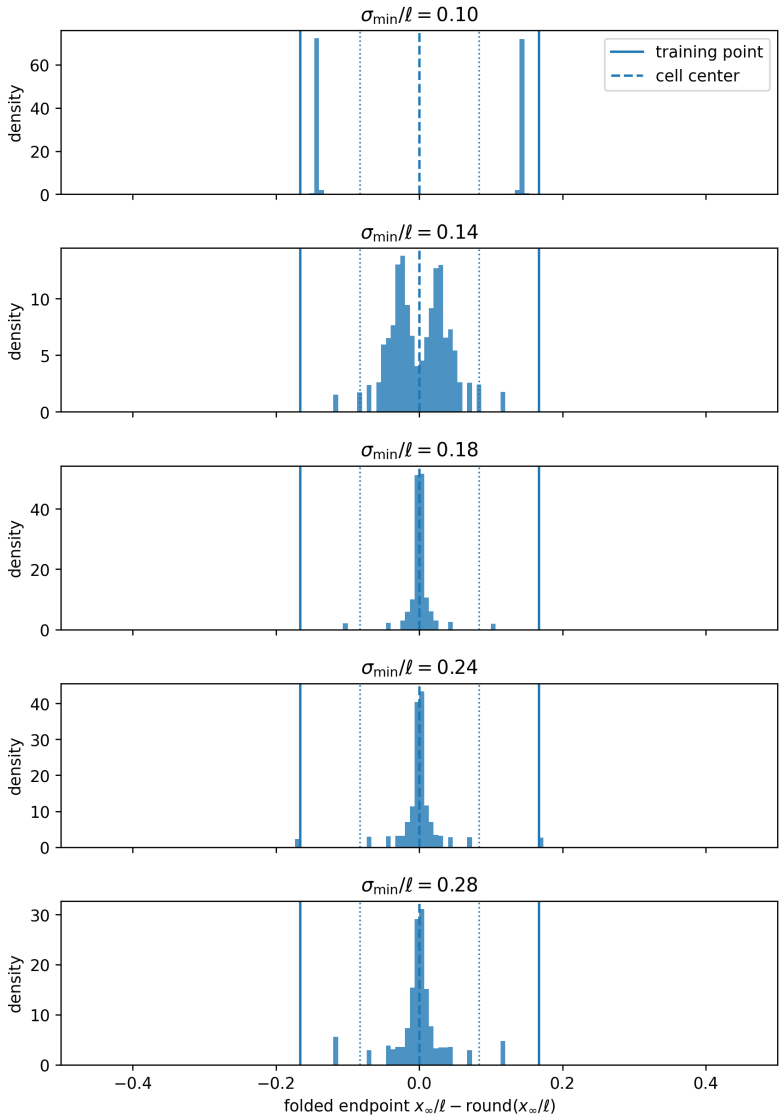

Figure 8: **Folded numerical outputs across the bandwidth sweep (E3).** The 150 equal-width histogram bins cover $[-1/2, 1/2)$; training locations, cell centers, and output-core boundaries are marked. All full-grid outputs are pooled. This plot visualizes within-cell displacement only; diversity across cells is quantified by (F.20). The finite-array histograms need not consist of a single exact central atom.

Step-size sensitivity. We repeat $s \in \{0.14, 0.18, 0.24\}$ with $\eta = 0.1$, the same 8192 initializations at each bandwidth, and the same residual threshold. The seed is $\mathrm{round}(10^4 s)$ for both step sizes;

different bandwidths use different seeds. Table 6 gives the paired results. The largest absolute changes are 0.00428 in separation and 0.00574 in core-label TV. At $s = 0.18$ and 0.24, the fifth-percentile training distances change by less than 4.4×10^{-5}, and the number of populated cores stays at 39. At $s = 0.14$, however, the fifth percentile changes from 0.09861 to 0.08411, close to the $1/12$ threshold, and the populated-core count changes from 38 to 39. This setting requires 2451 and 4901 updates at the two step sizes. The qualitative separation–coverage pattern persists under step refinement, with the largest sensitivity at $s = 0.14$, where the lower-tail distance lies closest to the separation threshold.

Table 6: **E3 paired step-size comparison.** The initializations are matched within each bandwidth. The last column is the maximum normalized residual at the final recorded check. The $s = 0.14$ lower-tail margin remains sensitive despite similar residuals, whereas the fifth percentiles at $s = 0.18$ and 0.24 change by less than 4.4×10^{-5}.

s	η	$\widehat{p}_{\mathrm{sep}}$	$\widehat{\epsilon}_{\mathrm{cell}}$	$q_{0.05}$	updates	max. residual
0.14	0.2	0.9780	0.1325	0.09861	2451	9.72×10^{-6}
0.14	0.1	0.9764	0.1268	0.08411	4901	9.78×10^{-6}
0.18	0.2	0.9734	0.1060	0.12276	251	3.81×10^{-6}
0.18	0.1	0.9700	0.1039	0.12278	451	9.54×10^{-6}
0.24	0.2	0.9679	0.0901	0.09919	301	3.81×10^{-6}
0.24	0.1	0.9636	0.0844	0.09914	501	9.54×10^{-6}

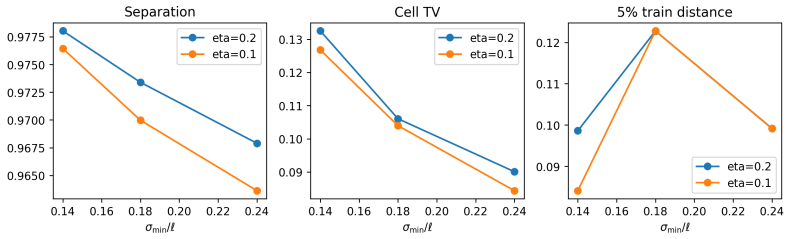

Figure 9: **Step-size sensitivity of the E3 numerical outputs.** The two step sizes use matched initializations at each bandwidth and the same stopping diagnostic. Separation and core-label TV change modestly at all three settings; the lower-tail distance at $s = 0.14$ is more sensitive than at $s = 0.18$ and 0.24. This is a finite-step comparison, not an extrapolation to zero step size.

Relation to the constructive theorem. E3 uses the same finite-grid geometry and theorem-aligned statistics as Corollary 3.2, but its parameters are chosen for a small numerical instance rather than the theorem's conservative sufficient regime; in particular, all tested $s < 1/2$, whereas the proof uses $s_0 \geq 1/2$. It therefore provides a numerical realization of simultaneous separation and sample-resolution diversity.

F.4 E4: Geometric scaling and posterior-localization diagnostics

E4 separates two complementary questions: whether the near-manifold bias follows the finite-schedule population prediction, and whether a finite empirical cloud exhibits controlled posterior moments at sampled tube locations. The two diagnostics use their own matched scale schedules.

(i) Population bias: setup and computation. We use the proper population density (E.1) with $Q_{\mathcal{M}} = \mathrm{Unif}(S^2(R))$, $m = 2$, $c = 10$, and equal sampling of the 64 scales

$$\sigma_i = \sigma_{\min} 16^{(i-1)/63}, \quad i = 1, \ldots, 64. \tag{F.22}$$

The radii are $R \in \{0.5, 1, 2\}$ and the lower cutoffs are $\sigma_{\min} \in \{0.01, 0.015, 0.02, 0.03, 0.04, 0.06\}$. Thus the absolute schedule is held fixed when comparing different R at one lower cutoff, and the relative schedule is held fixed throughout.

For $x = \varrho e \in V$, we evaluate the normalized radial field as $R\mathbb{E}_{\widetilde{w}}[\cos\theta] - \varrho$. The angular integral uses 1024 Gauss–Legendre nodes in $t = \cos\theta \in [-1, 1]$, whose reference density is constant on S^2; the 64 noise scales are summed exactly. The calculation uses double precision and stabilized log weights. We search 600 equally spaced radii in $[0.7R, 1.3R]$, select the sign-change bracket closest to R, and bisect for at most 120 iterations, stopping when the absolute field residual is below 10^{-14} or the bracket width is below 10^{-13}. The resulting radius is denoted $\widehat{\varrho}_\star$, with reported bias $\widehat{b}_\star = |\widehat{\varrho}_\star - R|$. This directly estimates the near-sphere stationary branch of the population field.

Population bias: results and reference. For each R, ordinary least squares on the six pairs $(\log\sigma_{\min}, \log\widehat{b}_\star)$ gives the slopes in Table 7. Figure 10 displays both the unnormalized biases and $R\widehat{b}_\star/\sigma_{\min}^2$. At the smallest relative cutoff $\sigma_{\min}/R = 0.005$, the latter is 1.1996890, compared with the prescribed-schedule prediction $C_{\mathrm{fs}} = 1.1996523$ from Appendix E.3; their relative discrepancy is 0.00306%. Across the full sweep the discrepancy ranges up to 1.828% at $\sigma_{\min}/R = 0.12$, consistent with the local rather than exact finite-noise nature of the prediction. The nearly quadratic slopes, rescaled curves, and limiting constant together closely track the predicted finite-schedule noise–curvature dependence.

Table 7: **E4 population bias scaling.** Slopes use all six tested lower cutoffs. Here $B(s) = R\widehat{b}_\star/s^2$ at fixed R and $s = \sigma_{\min}$. The last column reports the relative discrepancy between $B(0.01)$ and $C_{\mathrm{fs}} = 1.199652317\ldots$.

R	fitted slope	$B(0.01)$	$B(0.06)$	error at 0.01 (%)
0.5	2.0091	1.2002404	1.2215817	0.04902
1.0	2.0022	1.1997992	1.2049867	0.01225
2.0	2.0006	1.1996890	1.2009770	0.00306

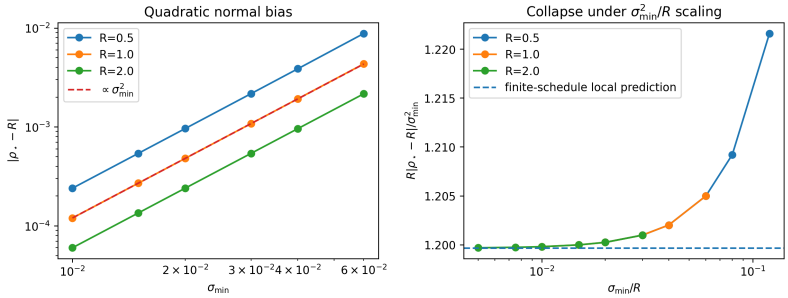

Figure 10: **Population spherical bias with a fixed relative noise schedule (E4).** Left: the located near-sphere bias for three sphere radii, with a quadratic guide. Right: the normalized bias approaches $C_{\mathrm{fs}} = 1.199652317\ldots$, computed from the 64-scale schedule rather than fitted to the numerical outputs. The quadratic guide on the left is anchored to one numerical point and illustrates slope only; the horizontal line on the right is the analytic finite-schedule local reference. All reported biases come from population quadrature and radial root finding.

(ii) Empirical posterior localization: setup. We draw a single cloud of $M = 10^4$ uniform centers on $S^2(1) \subset \mathbb{R}^{12}$ and hold it fixed for the diagnostic. Here $R = 1$, $\sigma_{\min} = 0.05$, and the 24 geometric scales span $[0.05, 0.8]$. At each radius r we sample 512 independent base points $z \sim \mathrm{Unif}(S^2(1))$ and uniformly distributed unit directions $n_z \in N_z\mathcal{M}$, and evaluate $x = z + rn_z$. There are 14 logarithmically spaced radii between $r_{\min}/4$ and 0.30, together with the additional radius $r = r_{\min} = 0.05\sqrt{10} \approx 0.158114$. Thus 15 radii and 7680 query locations are tested, all with $r < R$; the nearest-point projection is z. The normal space has one sphere-radial direction and nine directions orthogonal to the containing three-dimensional subspace. Sampling a normalized ten-dimensional Gaussian in this basis includes both inward and outward radial components, rather than only radial points in V.

The PyTorch implementation uses seed 999, one fixed empirical cloud, new queries at each radius, and single-precision log-sum-exp evaluation of the finite center–scale weights. It measures

$$\widehat{C}_{\mathrm{loc}}(x) = \frac{\sum_{j,i} \widetilde{w}_{ij}(x)\|\mu_j - z\|^2}{m \max\{\sigma_{\min}^2, r^2/c\}}, \qquad z = \Pi_{\mathcal{M}}(x), \quad m = 2,\ c = 10. \tag{F.23}$$

For each radius we record the median, 90th and 99th percentiles, and sampled maximum. These statistics summarize 7680 sampled tube locations for one fixed empirical cloud over the tested radius range.

Empirical localization: comparison and results. The reference curve is $\mathcal{C}_{\mathrm{flat},N}(r)$ from (E.19), evaluated with the actual 24-scale schedule of this diagnostic, not the 64-scale schedule of the bias experiment. For each radius, let $\widehat{C}_{0.5}(r)$ denote the empirical median and define

$$e_{\mathrm{med}}(r) = \frac{|\widehat{C}_{0.5}(r) - \mathcal{C}_{\mathrm{flat},N}(r)|}{\mathcal{C}_{\mathrm{flat},N}(r)}. \tag{F.24}$$

The arithmetic mean of e_{med} over the 15 radii is 1.174% and its maximum is 2.180%. The largest observed per-radius 99th percentile is 2.1233, and the maximum over all queried ratios is 2.3701. At $r = r_{\min}$, the flat reference is 1.533454 and the empirical median is 1.558910 (Table 8). Figure 11 shows that the median follows the finite-schedule reference while the tails remain larger. The peak near $r_{\min}$ coincides with the switch in the normalization $\max\{\sigma_{\min}^2, r^2/c\}$.

Table 8: **E4 localization at selected radii.** All entries use $R = 1$, one fixed cloud of 10^4 centers, and 512 queries per radius. The row $r = 0.158114$ corresponds to $r_{\min}$ before rounding. Aggregate discrepancies reported in the text use all 15 radii, not only these five rows.

r	$\mathcal{C}_{\mathrm{flat},N}(r)$	median	99th percentile	sampled maximum
0.039528	1.144095	1.162255	1.5547	1.7547
0.117725	1.289768	1.317887	1.7915	2.1698
0.158114	1.533454	1.558910	2.1233	2.3701
0.219636	1.262671	1.278057	1.6759	1.8251
0.300000	1.250007	1.252572	1.6455	1.9018

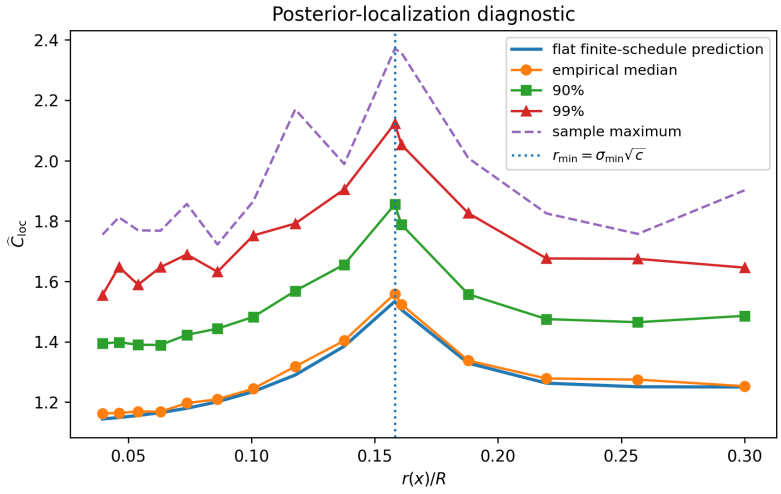

Figure 11: **Sampled posterior-localization diagnostic (E4).** At each of 15 radii, 512 tube queries are evaluated against one fixed cloud of 10^4 uniform spherical centers. The median closely follows the radius-dependent flat reference for the same 24-scale schedule, while the higher quantiles and sampled maximum show the variation across query locations.

Combined numerical evidence. E1–E4 provide complementary numerical support for regional retention, off-tube instability, positive-bandwidth separation with sample-resolution diversity, and the predicted geometric/localization scaling. E4 evaluates posterior localization on sampled tube queries, while the source–sink probability errors remain analytic quantities. These exact-field studies therefore test the mechanisms underlying the theory; learned-model behavior is evaluated separately below.

G Finite-Sample Refinements and Output Resolution

The exact empirical model can have gaps between its training centers. This appendix records how an additional posterior-moment residual changes the normal and tangential bounds, and why coverage must be stated at an explicit resolution.

G.1 A posterior-moment residual

Replace (A.23) by

$$\sum_{j,i} \widetilde{w}_{ij}(x)\|\mu_j - z_x\|^2 \le C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r(x)^2/c\} + h_M^2, \quad x \in \Omega, \quad 0 < r(x) < r_0. \tag{G.1}$$

Here $h_M \ge 0$ has units of length and describes an assumed posterior residual, not automatically a fill distance or a sampling rate. Set $b_M = h_M^2/(2\tau)$. Averaging the reach inequality gives

$$\dot r \le \begin{cases} -\beta(x)(r - \varepsilon_0 - b_M), & r \le r_{\min}, \\ -\beta(x)((1-\kappa_0)r - b_M), & r > r_{\min}. \end{cases}$$

With $\kappa_0 < 1$, define

$$\varepsilon_{\mathcal{M},h} = \max\left\{\min\{r_{\min}, \varepsilon_0 + b_M\}, \frac{b_M}{1-\kappa_0}\right\}, \tag{G.2}$$

and assume $\varepsilon_{\mathcal{M},h} < r_0$. Above this radius each applicable radial bound is strictly inward. In the first regime either $r > \varepsilon_0 + b_M$ or that regime is empty; in the second, $(1-\kappa_0)r > b_M$. Keeping the off-tube hypotheses of Theorem A.14 therefore gives $r(x_\infty) \le \varepsilon_{\mathcal{M},h}$ almost surely and the same localization for maximizers. Also,

$$\varepsilon_{\mathcal{M},h} \le \varepsilon_{\mathcal{M}} + \frac{h_M^2}{2\tau(1-\kappa_0)},$$

by $\min\{r, a+b\} \le \min\{r, a\} + b$.

Proposition G.1 (Residual-aware tangential retention)**.** *Under these residual-aware convergence conditions, choose $\varepsilon_{\mathcal{M},h} < r_{\mathrm{stop}} < r_{\mathrm{tr}} < r_0$ and set*

$$D_h(r) = \begin{cases} r - \varepsilon_0 - b_M, & r \le r_{\min}, \\ (1-\kappa_0)r - b_M, & r > r_{\min}. \end{cases} \tag{G.3}$$

Then $r(X_{\mathrm{stop}}) \le r_{\mathrm{stop}}$ almost surely and the projected refinement length is at most

$$\delta_{\mathrm{stop},h} = L_\Pi \int_{r_{\mathrm{stop}}}^{r_{\mathrm{tr}}} \frac{\sqrt{C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r^2/c\} + h_M^2}}{D_h(r)}\, dr, \qquad L_\Pi = \frac{\tau}{\tau - r_{\mathrm{tr}}}. \tag{G.4}$$

Under the corresponding Gaussian band conditions, the source–sink coverage theorem holds with $\delta_{\mathrm{cov}} = \delta_{\mathrm{noise}} + \delta_{\mathrm{stop},h}$ and the same probability errors.

Proof. The residual-aware normal bound gives finite entrance and $-\dot r \ge \beta D_h(r) > 0$ between the two surfaces. Projection and Jensen's inequality now give $\|\dot z\| \le L_\Pi \beta \sqrt{C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r^2/c\} + h_M^2}$. Divide these inequalities and integrate over decreasing r, splitting at $r_{\min}$ if necessary. Each denominator is bounded away from zero on the relevant closed interval. The source–sink proof uses this full path-length bound and is otherwise unchanged. With $h_M = 0$ and the original radius ordering, the two integrals evaluate to (B.10). □

A fixed positive h_M prevents these sufficient bounds from supporting $r_{\mathrm{stop}} \to 0$ at fixed reach. Thus both the normal error and the retention calculation must account for a nonvanishing posterior residual.

G.2 Training distance and the resolution of coverage

For a compactly supported reference distribution define

$$h_{\mathrm{cov},M} = \sup_{z \in \mathrm{supp}(Q_{\mathcal{M}})} \mathrm{dist}(z, \mathcal{S}_M). \tag{G.5}$$

This is the geometric covering radius, distinct from the posterior residual h_M. The affine grid has the explicit value in (D.1).

Corollary G.2 (Training distance for the stopped output)**.** *Under Theorem B.6, for every $\eta \geq 0$,*

$$\begin{aligned}\mathbb{P}[\mathrm{dist}(X_{\mathrm{stop}}, \mathcal{S}_M) > \eta] \geq (1-\zeta) Q_{\mathcal{M}}\big(\{z : \mathrm{dist}(z, \mathcal{S}_M) > \eta + r_{\mathrm{stop}} + \delta_{\mathrm{cov}}\}\big) \\ - \epsilon_{\mathrm{Pois}} - \epsilon_{\mathrm{init}}.\end{aligned} \tag{G.6}$$

Proof. Distance to the training set is one-Lipschitz. A reference label at the displayed distance, within δ_{cov} of the stopped projection, gives a projection at distance greater than $\eta + r_{\mathrm{stop}}$. The ambient output differs from that projection by at most r_{stop}. Apply the same safe-interior event comparison as in the coverage theorem. □

Conversely, if a manifold-valued output has coverage $\mathbb{P}[Z_{\mathrm{out}} \in U] \geq \alpha Q_{\mathcal{M}}(U^{-\delta}) - \epsilon_{\mathrm{cov}}$ and training-set concentration $\mathbb{P}[\mathrm{dist}(Z_{\mathrm{out}}, \mathcal{S}_M) \leq \eta] \geq 1 - \epsilon_{\mathrm{mem}}$, then

$$\alpha Q_{\mathcal{M}}(\{z : \mathrm{dist}(z, \mathcal{S}_M) > \eta + \delta\}) \leq \epsilon_{\mathrm{mem}} + \epsilon_{\mathrm{cov}}. \tag{G.7}$$

Indeed, apply coverage to $U = \{z : \mathrm{dist}(z, \mathcal{S}_M) > \eta\}$ and use the one-Lipschitz property to include the displayed reference set in $U^{-\delta}$. For $\alpha > 0$ and zero errors this implies $h_{\mathrm{cov},M} \leq \eta + \delta$ by continuity and the definition of support. This distinguishes coarse coverage from novelty: assigning each point of a uniformly populated circle to its nearest training center preserves coverage at half the training spacing, but exactly memorizes that finite set.

G.3 From a finite stop to an endpoint

Finite ambient path length and the projection derivative bound imply

$$\int_{\tau_{\mathrm{stop}}}^{\infty} \left\| \frac{d}{dt} \Pi_{\mathcal{M}}(x(t)) \right\| dt \leq \frac{\tau}{\tau - r_{\mathrm{stop}}} \int_{\tau_{\mathrm{stop}}}^{\infty} \|\dot{x}(t)\| dt < \infty. \tag{G.8}$$

If this projected tail has a uniform bound δ_{tail}, the same reference coupling gives endpoint coverage with resolution $\delta_{\mathrm{cov}} + \delta_{\mathrm{tail}}$. Pathwise finiteness does not supply a small uniform bound. The sample-resolution result instead proves endpoint retention from finite-mixture gradient signs in the capture cells, so its endpoint guarantee does not require this additional tail assumption.

H Penalty-Guided Generation: Convergence, Geometry, and Experimental Details

This appendix expands Section 3.3. We first analyze the generic penalty-guided objective, then explain how the task gradient perturbs manifold attraction and resolved diversity. We finally record the concrete penalties used in the controllable-generation experiments. The theoretical statements use the exact unified score $\nabla \log p_{MN}$; replacing it by the learned state-only score s_θ gives an inexact-gradient implementation.

H.1 Penalty-guided ascent and stationary points

Let $E_y : \mathbb{R}^d \to \mathbb{R}_+$ be a differentiable task penalty, with smaller values indicating better task agreement. In general, we do *not* require

$$E_y(x) = 0 \quad \Longleftrightarrow \quad x \text{ satisfies the task.}$$

For example, a squared measurement residual may vanish at exact feasibility, whereas smooth inequality penalties and classifier losses generally need not. For a finite penalty weight $\lambda \geq 0$, define

$$F_y(x) = \log p_{MN}(x) - \lambda E_y(x), \qquad \dot{x} = \nabla F_y(x) = \nabla \log p_{MN}(x) - \lambda \nabla E_y(x). \tag{H.1}$$

The task is imposed through a soft penalty relaxation rather than an exact projection.

Assume that E_y is nonnegative and real analytic. Along the exact flow,

$$\frac{d}{dt} F_y(x(t)) = \|\nabla F_y(x(t))\|^2 \geq 0. \tag{H.2}$$

Since $F_y \leq \log p_{MN}$ and $\log p_{MN}(x) \to -\infty$ as $\|x\| \to \infty$, the superlevel set

$$\Omega_y(x_0) = \{x : F_y(x) \geq F_y(x_0)\} \tag{H.3}$$

is compact and forward invariant. The same analytic-gradient argument as in Lemma A.1 therefore gives global forward existence, finite path length, and convergence to a stationary point $x_{\infty,y}$, which satisfies

$$\nabla \log p_{MN}(x_{\infty,y}) = \lambda \nabla E_y(x_{\infty,y}). \tag{H.4}$$

This is stationarity of the penalized objective, not a claim of global optimality or exact satisfaction of a hard constraint.

For the explicit exact-score update

$$x_{n+1} = x_n + \eta_n \left[\nabla \log p_{MN}(x_n) - \lambda \nabla E_y(x_n)\right], \tag{H.5}$$

choose $\overline{\eta} > 0$ and $L_y > 0$ such that ∇F_y is L_y-Lipschitz on a compact convex set containing all trial segments

$$\left\{x + t\nabla F_y(x) : x \in \Omega_y(x_0),\ 0 \leq t \leq \overline{\eta}\right\},$$

and require

$$0 < \underline{\eta} \leq \eta_n \leq \overline{\eta} < 2/L_y \qquad \text{for every } n. \tag{H.6}$$

Such choices are possible under the analytic assumption: first fix a finite trial-step cap, bound the Hessian on a compact convex set containing its trial segments, and then reduce $\overline{\eta}$ if necessary. This smoothness condition covers the update segments, not merely the possibly nonconvex superlevel set.

With $\Delta_n := x_{n+1} - x_n$, the ascent estimate gives

$$\begin{aligned} F_y(x_{n+1}) - F_y(x_n) &\geq \left(\eta_n^{-1} - \frac{L_y}{2}\right) \|\Delta_n\|^2 \\ &\geq \left(\overline{\eta}^{-1} - \frac{L_y}{2}\right) \|\Delta_n\|^2. \end{aligned} \tag{H.7}$$

Thus, by induction, all iterates stay in $\Omega_y(x_0)$ and the objective values increase to a finite limit. Moreover,

$$\|\nabla F_y(x_n)\| = \frac{\|\Delta_n\|}{\eta_n} \leq \underline{\eta}^{-1} \|\Delta_n\|.$$

The sufficient-ascent estimate and this gradient–step bound, together with the analytic-gradient inequality on the compact accumulation set, yield $\sum_n \|\Delta_n\| < \infty$. Hence the iterates converge to a stationary point satisfying (H.4). The uniform lower step bound and strict upper margin in (H.6) are part of this discrete convergence guarantee.

H.2 How guidance perturbs manifold attraction

Inside the reach tube, write $z_x = \Pi_{\mathcal{M}}(x)$, $r(x) = \|x - z_x\|$, and $u_x = (x - z_x)/r(x)$ for $r(x) > 0$. For the guided flow initialized at x_0, define the relevant tube region by

$$\mathcal{R}_y := \Omega_y(x_0) \cap \{x : r(x) < r_0\}. \tag{H.8}$$

Retain the local geometry and contraction-gap conditions used in Lemma A.9, and require the posterior-localization estimate on this guided region:

$$\sum_{j,i} \widetilde{w}_{ij}(x)\|\mu_j - z_x\|^2 \le C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r(x)^2/c\}, \qquad x \in \mathcal{R}_y, \quad r(x) > 0. \tag{H.9}$$

Here $\varepsilon_0, r_{\min}, \kappa_0$ are defined as before using this C_{loc}, with $\kappa_0 < 1$ and $\varepsilon_0 < \min\{r_{\min}, r_0\}$. The domain in (H.9) is essential: guidance preserves a superlevel set of F_y, but need not preserve the unguided superlevel set of $\log p_{MN}$. Localization on that unguided set alone therefore does not justify the guided radial estimate.

The guided radial velocity is

$$\dot r = \langle u_x, \nabla \log p_{MN}(x)\rangle - \lambda \langle u_x, \nabla E_y(x)\rangle. \tag{H.10}$$

Only the outward component of $-\lambda \nabla E_y$ opposes the prior's normal attraction. Define

$$\delta_y := \sup_{\substack{x \in \mathcal{R}_y \\ 0 < r(x) < r_0}} \frac{\lambda[-\langle u_x, \nabla E_y(x)\rangle]_+}{\beta(x)}, \tag{H.11}$$

with $\delta_y = 0$ if the indexed set is empty. This quantity is finite under the assumptions above, since ∇E_y is bounded on the compact set $\Omega_y(x_0)$ and $\beta(x) \ge \sigma_{\max}^{-2}$. Thus

$$\lambda[-\langle u_x, \nabla E_y(x)\rangle]_+ \le \beta(x)\delta_y \qquad (x \in \mathcal{R}_y,\ r(x) > 0). \tag{H.12}$$

The quantity δ_y measures harmful normal guidance relative to the local contraction scale of the prior.

Applying the proof of Lemma A.9 on $\mathcal{R}_y$ and combining it with (H.12) gives

$$\dot r \le \begin{cases} -\beta(x)(r - \varepsilon_0 - \delta_y), & 0 < r \le r_{\min}, \\ -\beta(x)((1-\kappa_0)r - \delta_y), & r_{\min} < r < r_0. \end{cases} \tag{H.13}$$

Set

$$\varepsilon_y = \max\left\{\varepsilon_0 + \delta_y,\ \frac{\delta_y}{1-\kappa_0}\right\}, \qquad \text{and assume } \varepsilon_y < r_0. \tag{H.14}$$

If $r(x(t_0)) < r_0$ at some time t_0, a first-exit argument keeps the trajectory in the tube thereafter. The two radial inequalities imply

$$[r(x(t)) - \varepsilon_y]_+ \le [r(x(t_0)) - \varepsilon_y]_+ \exp\left(-\frac{1-\kappa_0}{\sigma_{\max}^2}(t - t_0)\right), \qquad t \ge t_0. \tag{H.15}$$

Consequently, the stationary limit from the preceding subsection obeys $\mathrm{dist}(x_{\infty,y}, \mathcal{M}) \le \varepsilon_y$. In particular, tangential or inward guidance has $\delta_y = 0$ and leaves the local normal bound unchanged, provided the localization estimate holds on $\mathcal{R}_y$. A small outward component enlarges the bound continuously. If $\varepsilon_y \ge r_0$, this sufficient contraction guarantee no longer applies. This is an after-entry statement: global entry from arbitrary initialization requires the corresponding off-tube behavior of the guided field, which does not follow automatically from the unguided assumptions.

The same local radial calculation covers norm-clipped guidance. If a clipped penalty gradient g_y is used in the field $\dot x = \nabla \log p_{MN}(x) - \lambda g_y(x)$ and $\|g_y(x)\| \le c_{\mathrm{clip}}$, then

$$\frac{\lambda[-\langle u_x, g_y(x)\rangle]_+}{\beta(x)} \le \lambda c_{\mathrm{clip}} \sigma_{\max}^2. \tag{H.16}$$

This is a coarse worst-case budget; predominantly tangential directions can have a much smaller effective δ_y. The posterior-localization estimate must hold on the region visited by this clipped flow. Since the clipped field need not be a gradient of F_y, the normal estimate alone does not imply the stationary-point convergence proved above.

Exact affine compatibility. If the data lie in an affine space $\mathcal{A} = a_0 + V$ and $P_{V^\perp} \nabla E_y(x) = 0$, then the penalty changes only motion within V. The normal component remains

$$P_{V^\perp} \dot{x} = -\beta(x) P_{V^\perp}(x - a_0), \qquad r_{\mathcal{A}}(x(t)) \le r_{\mathcal{A}}(x(0)) e^{-t/\sigma_{\max}^2}. \tag{H.17}$$

Hence a tangential task penalty can alter which point is selected inside the data subspace without changing exact normal convergence.

H.3 Guidance strength and retained diversity

The normal calculation controls fidelity but not probability allocation along the manifold. A simple sufficient condition follows from the capture regions in Appendix D.4. Suppose an unguided region has a strict inward margin $\gamma > 0$ on a boundary face with outward tangential normal v:

$$\langle v, \nabla \log p_{MN}(x) \rangle \le -\gamma.$$

If, on the same face,

$$\lambda |\langle v, \nabla E_y(x) \rangle| < \gamma, \tag{H.18}$$

then the guided field still points inward on that face. If this condition holds on every tangential face and the guided field also points inward on the normal boundary, the compact capture region remains forward invariant. Under the analytic assumptions above, trajectories entering it converge to stationary points within that region. Multiple disjoint task-compatible regions can therefore retain distinct outcomes.

The source–sink balance itself changes under guidance:

$$\begin{aligned} \nabla \cdot (A_N p_{MN} [\nabla \log p_{MN} - \lambda \nabla E_y]) = p^Q_{\sigma_{\max}} - p^Q_{\sigma_{\min}} + \mathcal{R}_N \\ - \lambda A_N \nabla \cdot (p_{MN} \nabla E_y). \end{aligned} \tag{H.19}$$

Thus penalty guidance modifies regional probability through the additional flux term. Structurally, manifold fidelity is controlled by the normal component of guidance, while retained diversity depends on whether its tangential component preserves multiple task-compatible regions.

H.4 Large-step controllable-generation comparison

The practical comparison uses a time-varying guidance schedule and, on CIFAR-10, score-norm matching. It therefore evaluates robustness of learned fields to guided displacements, complementing the fixed-λ exact-gradient analysis above.

We compare noise-conditioned and noise-unconditioned learned score models with the same ODE sampling framework (Zhou et al., 2026). For each task we use two step budgets, $K \in \{1000, 100\}$. The conditioned model is evaluated with the externally scheduled noise level, whereas the noise-unconditioned model receives only the current state. Unless stated otherwise, the sampling setup is matched across the two models.

For both tasks, the guidance strength is annealed over the sampling trajectory:

$$\lambda_k = 1 - \frac{k}{K-1}, \qquad k = 0, \dots, K-1. \tag{H.20}$$

Synthetic four-disk constraint. We use the same 100-dimensional latent construction as in the toy studies: the first two coordinates form four separated disks and the remaining 98 coordinates are Gaussian, followed by the canonical embedding into $\mathbb{R}^{3072}$. Both learned score models use the same standard U-Net backbone implemented in PyTorch, differing in whether the noise level is provided as an input.

The task requires the first two latent coordinates to be positive, selecting the first-quadrant disk. Writing $y = W^\top x$, where W denotes the canonical projection onto the 100-dimensional latent coordinates, we use

$$E_{\text{toy}}(x) = \text{softplus}(-\alpha y_1) + \text{softplus}(-\alpha y_2), \tag{H.21}$$

and replace the prior score s_k by

$$d_k = s_k - \lambda_k \nabla E_{\text{toy}}(x_k), \qquad s_k = \begin{cases} s_\theta(x_k, \sigma_k), & \text{noise conditioned,} \\ s_\theta(x_k), & \text{noise unconditioned.} \end{cases} \tag{H.22}$$

The same penalty and linear schedule are used for both models.

CIFAR-10 classifier guidance. For CIFAR-10 we use the NCSN++ architecture of Song et al. (2020) for both models, again differing only in whether the scheduled noise level is supplied. A fixed classifier pretrained on CIFAR-10 provides guidance toward the *horse* class, following the classifier-guidance principle of Dhariwal & Nichol (2021). Define

$$E_{\mathrm{cls}}(x) = -\log p_\phi(y_{\mathrm{horse}} \mid \mathrm{prep}(x)), \tag{H.23}$$

where prep denotes the classifier preprocessing. At every sampling step, PyTorch autograd computes the classifier direction $g_k = -\nabla E_{\mathrm{cls}}(x_k)$ together with the current generative score s_k. We normalize the guidance *per sample* toward the current score magnitude,

$$\rho_k = \frac{\|s_k\|_2}{\|g_k\|_2 + \epsilon_{\mathrm{stab}}}, \qquad d_k = s_k + \lambda_k \rho_k g_k, \tag{H.24}$$

with the norms taken over all image coordinates and $\epsilon_{\mathrm{stab}} > 0$ preventing division by zero. The classifier direction is therefore rescaled toward the current score magnitude before multiplication by the common factor λ_k. This removes guidance-scale differences between the two score parameterizations and makes the 1000-versus-100-step comparison focus on robustness to the guided displacement.

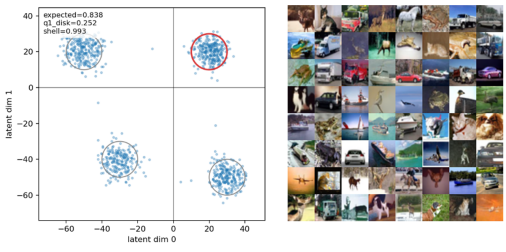

Figure 12: **Unguided 100-step conditioning baseline.** This ablation isolates the failure in Figure 3 to the interaction between coarse steps and guidance rather than the reduced step count alone.

Unguided 100-step control. To distinguish a coarse-solver failure from a guidance-induced failure, we also run the noise-conditioned models with the same 100-step sampler but with guidance removed as in the Fig. 12. These unguided runs still generate well-formed samples, showing that the degradation in Figure 3 is associated with the interaction between large guided updates and noise conditioning rather than the 100-step discretization by itself.

Interpretation. In a noise-conditioned sampler, the scheduled σ_k determines which marginal the network is expected to score. A sufficiently large guided move can place x_{k+1} outside the high-density region of the next scheduled marginal, so $s_\theta(x_{k+1}, \sigma_{k+1})$ is evaluated at a mismatched state–noise pair. Noise unconditioning aggregates perturbations from all scales into one state-only field. The current state therefore determines the effective contribution of different scales, making the learned field substantially more robust to such guidance-induced displacements. The 100-step results in Figure 3, together with the unguided control above, provide empirical support for this mechanism.

H.5 Downstream classification under data imbalance

Task and fixed real-data split. We construct a binary CIFAR-10 cat–dog task from the official training split. The low-data training set contains 100 cats and 1000 dogs, and the same selected images are reused in every comparison. The balanced-real reference keeps these 100 cats and adds 900 distinct real cats. The repetition baseline repeats the same 100 cats to obtain 1000 cat instances, increasing minority-class weight without adding new source images. Evaluation uses all cat and dog images from the official CIFAR-10 test split, and balanced accuracy is the mean of the two per-class accuracies.

Generated-data conditions. For the main low-data comparison, the noise-conditioned diffusion model and noise-unconditioned model are each trained using *only the same 100 real cats* available to the downstream task. Each generates 900 additional cats, which are combined with the 100 real cats

and 1000 real dogs. As a separate control, we also evaluate 900 cats generated by models pretrained on the full CIFAR-10 training distribution; these serve as broad-data augmentation controls. For the low-data comparison, both generators follow the training and sampling configuration of Zhou et al. (2026), with the training subset restricted to the same 100 cats.

Why this probes finite-data generalization. Sample-level memorization and generalization of noise-unconditioned scores are studied directly in Zhou et al. (2026); here we ask whether the difference is useful downstream. In the memorization limit, generated augmentation contributes little new information beyond repeatedly presenting the same finite minority set, so its effect approaches reweighting. A generator that generalizes beyond those examples can instead provide useful within-class variation, effectively broadening the minority-class support available to the classifier. We therefore use held-out classification accuracy as a functional measure of whether this additional variation is useful downstream.

Downstream classifier. Every condition trains the same CIFAR-style ResNet-18 binary classifier from scratch. We replace the ImageNet stem by a 3×3, stride-one convolution and remove max pooling. Training uses random 32×32 crops with padding four, horizontal flips, CIFAR-10 normalization, SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, Nesterov momentum, and a 100-epoch cosine schedule. The batch size is 128. To match the optimization budget, every method draws 2000 training examples per epoch; for the 1100-image biased baseline this sampling is with replacement and therefore preserves the original 10:1 imbalance in expectation. Within each classifier seed, all methods use the same initial ResNet-18 weights. Results report mean and standard deviation over three classifier seeds (0, 1, 2).

Table 9: **Full downstream augmentation results.** Test accuracy (%, mean±std). The two 100-cat generator rows are the data-scarce comparison emphasized in Section 4.4.

Training condition	Cat	Dog	Bal. Acc.
Biased real	11.80±1.25	96.73±0.21	54.27±0.56
Balanced real reference	79.17±0.76	79.63±1.40	79.40±1.01
Repeated cats (1000)	13.63±0.76	97.13±0.46	55.38±0.41
Pretrained diffusion aug.	77.53±1.34	82.07±1.36	79.80±1.35
Pretrained unconditioning aug.	77.00±2.62	82.33±2.01	79.67±2.28
100-cat diffusion aug.	26.17±1.65	93.33±0.47	59.75±0.95
100-cat unconditioning aug.	38.80±3.45	91.87±0.95	65.33±2.03

The pretrained-generator controls reach 79.80% and 79.67% balanced accuracy, close to the 79.40% balanced-real reference, showing that both model classes can provide effective augmentation when trained on broad CIFAR-10 data. The data-scarce comparison is more informative: with both generators trained on only the same 100 cats, noise unconditioning improves balanced accuracy by 5.58 percentage points over diffusion and by 9.95 points over repetition. The corresponding cat-class gains are 12.63 and 25.17 points.

The gain over repetition cannot be explained by class-count balancing alone, while the comparison with diffusion holds the scarce real training set fixed. The results are therefore consistent with finite-data generative generalization providing additional minority-class variation that is useful for classification of held-out real data.

Ten-class CIFAR-10 extension. To test whether the downstream effect persists beyond binary classification, we repeat the experiment on the full CIFAR-10 ten-class task. The cat class remains data-scarce, with only 100 fixed real training images, while each of the other nine classes uses 1000 fixed real images. The seven conditions mirror the binary experiment: biased real data, a balanced-real reference, repeated minority examples, and augmentation from pretrained or 100-cat-trained diffusion and noise-unconditioned generators. Each augmentation condition adds 900 generated cats, producing 1000 cat entries in total.

We use the same CIFAR-style ResNet-18 training protocol, now with ten output classes. Each method uses 10,000 sampled training examples per epoch; within each classifier seed, all seven conditions share the same network initialization. Results are averaged over the same three classifier seeds.

Evaluation uses the complete CIFAR-10 test set. Since the test set is balanced across classes, overall and macro accuracy coincide; we additionally report cat accuracy and the mean accuracy over the other nine classes.

Table 10: **Ten-class CIFAR-10 classification with a scarce cat class.** Accuracy (%, mean±std over three classifier seeds). "Other-9" is the mean per-class accuracy over the nine non-cat classes.

Training condition	Overall	Cat	Other-9
Biased real	81.49±1.02	22.40±2.25	88.06±0.95
Balanced real reference	86.00±1.01	71.30±2.31	87.63±0.90
Repeated cats (1000)	79.90±1.43	14.73±1.04	87.14±1.53
Pretrained diffusion aug.	86.17±0.93	69.40±2.26	88.03±0.83
Pretrained unconditioning aug.	84.77±1.05	66.87±1.72	86.76±1.27
100-cat diffusion aug.	81.59±0.29	29.50±0.53	87.37±0.31
100-cat unconditioning aug.	83.50±0.73	42.97±3.30	88.00±0.48

The ten-class results preserve the low-data pattern observed in the binary task. With both generators trained on the same 100 cats, noise-unconditioned augmentation improves overall accuracy by 1.91 percentage points over diffusion (83.50% versus 81.59%) and cat accuracy by 13.47 points (42.97% versus 29.50%). The advantage appears in all three classifier seeds. At the same time, mean accuracy over the other nine classes is essentially preserved, and is slightly higher for noise unconditioning (88.00% versus 87.37%). Relative to simply repeating the same 100 cats, noise-unconditioned augmentation improves overall accuracy by 3.60 points and cat accuracy by 28.23 points. Thus the downstream gain is not specific to binary cat–dog discrimination and remains visible in the full ten-class recognition problem.

I Population Objectives: Convergence and Regional Coverage

We prove Remark 3.3 directly for a *smoothed population mixture*. No training-size limit is needed, and no score of the possibly singular clean distribution $Q_{\mathcal{M}}$ is assumed.

I.1 The population learning target and posterior identities

Let $Q_{\mathcal{M}}$ be a probability distribution with compact support in $\mathcal{M}$, and let π be a probability measure on a fixed interval $[\sigma_{\min}, \sigma_{\max}] \subset (0, \infty)$. Normalizing a finite nonzero scale measure does not change the density in (E.1), which we reuse here:

$$p^{\mathrm{pop}}(x) = \frac{\int \sigma^2 p_\sigma^Q(x)\, \pi(d\sigma)}{\int \sigma^2\, \pi(d\sigma)}, \qquad p_\sigma^Q(x) = \int \mathcal{N}(x; \mu, \sigma^2 I) Q_{\mathcal{M}}(d\mu). \tag{I.1}$$

The distribution $Q_{\mathcal{M}}$ need not have a density. Taking $\pi = N^{-1} \sum_i \delta_{\sigma_i}$ gives the finite-schedule population objective in the main-text remark.

For independent $\mu \sim Q_{\mathcal{M}}$ and $\sigma \sim \pi$, draw $X \mid \mu, \sigma \sim \mathcal{N}(\mu, \sigma^2 I)$. The population state-only loss obeys

$$\begin{aligned} \mathcal{L}_u^{\mathrm{pop}}(\theta) &= \frac{1}{2}\mathbb{E}\left[\sigma^2 \left\| s_\theta(X) - \frac{\mu - X}{\sigma^2} \right\|^2\right] \\ &= \frac{\mathbb{E}_\pi \sigma^2}{2} \mathbb{E}_{X \sim p^{\mathrm{pop}}} \| s_\theta(X) - \nabla \log p^{\mathrm{pop}}(X) \|^2 + C, \end{aligned} \tag{I.2}$$

where C is independent of θ. This follows by the same conditional variance decomposition as (A.10). The weighted component-score second moment is $\mathbb{E}[\sigma^2 \|(\mu - X)/\sigma^2\|^2] = d$, so the score-independent remainder is finite. Thus population noise unconditioning also learns the gradient of a single fixed log density.

For this population objective, use the existing posterior notation with integrals in place of center–scale sums:

$$\begin{aligned} w_x(d\mu, d\sigma) &= \frac{\sigma^2 \mathcal{N}(x; \mu, \sigma^2 I) Q_{\mathcal{M}}(d\mu) \pi(d\sigma)}{p^{\mathrm{pop}}(x) \int \sigma^2 \pi(d\sigma)}, \\ \beta(x) &= \int \sigma^{-2} w_x(d\mu, d\sigma), \qquad \bar{\mu}(x) = \int \mu\, \widetilde{w}_x(d\mu, d\sigma), \\ \widetilde{w}_x(d\mu, d\sigma) &= \frac{\sigma^{-2}}{\beta(x)} w_x(d\mu, d\sigma). \end{aligned} \tag{I.3}$$

These symbols now refer to the population posterior. Differentiation under the integral gives

$$\begin{aligned} \nabla \log p^{\mathrm{pop}}(x) &= \beta(x)(\bar{\mu}(x) - x), \qquad \sigma_{\max}^{-2} \le \beta(x) \le \sigma_{\min}^{-2}, \\ \nabla^2 \log p^{\mathrm{pop}}(x) &= C_x - \beta(x) I, \qquad C_x = \mathrm{Cov}_{w_x}\left(\frac{\mu - x}{\sigma^2}\right). \end{aligned} \tag{I.4}$$

In particular, every critical point belongs to $\mathrm{conv}(\mathrm{supp}\, Q_{\mathcal{M}})$.

I.2 Manifold convergence for a population objective

The same proof applies with posterior and landscape conditions imposed on the population objective.

Analyticity, bounded trajectories, and complete flow. By Tonelli's theorem, p^{pop} is a positive probability density. Compact center support and positive bounded scales allow the Gaussian kernels to be integrated locally uniformly in a complex neighborhood of each real x. The resulting density is real analytic, as is its logarithm since the density is strictly positive. With $B = \sup_{\mu \in \mathrm{supp}\, Q_{\mathcal{M}}} \|\mu\|$, for $\|x\| \ge B$,

$$p^{\mathrm{pop}}(x) \le (2\pi\sigma_{\min}^2)^{-d/2} \exp\left[-\frac{(\|x\| - B)^2}{2\sigma_{\max}^2}\right]. \tag{I.5}$$

Every positive-density superlevel is therefore compact. For nonempty compact K, define $\Omega = \{x : p^{\mathrm{pop}}(x) \ge \min_K p^{\mathrm{pop}}\}$. The objective increases along $\dot{x} = \nabla \log p^{\mathrm{pop}}(x)$, and the proof

of Lemma A.1 gives global forward existence, finite path length, and convergence from every point of K. Applying this with $K = \{x_0\}$ gives convergence for every initialization. Moreover, $\|\nabla \log p^{\mathrm{pop}}(x)\| \le \sigma_{\min}^{-2}(B + \|x\|)$. Both the field and its negative have at most linear growth, so the two-sided complete-flow and Jacobian arguments used for source–sink marking also apply.

Normal attraction and off-tube instability. Impose the population version of posterior localization on the corresponding Ω:

$$\int \|\mu - z_x\|^2 \, \widetilde{w}_x(d\mu, d\sigma) \le C_{\mathrm{loc}} m \max\{\sigma_{\min}^2, r(x)^2/c\}, \qquad x \in \Omega, \quad 0 < r(x) < r_0. \tag{I.6}$$

The reach inequality holds for every support point μ. Integrating it against $\widetilde{w}_x$ gives the same posterior-alignment and radial bounds (A.25) and (A.29). Jensen's inequality gives the same projected-speed and retention bounds. The covariance identity (I.4) likewise gives the same positive log-density Hessian eigenvalue under the population shell-rank condition.

Proposition I.1 (Population manifold convergence)**.** *Let $\mathcal{M}$ have the compact C^2 geometry of Theorem A.14, and suppose its local geometry, posterior-localization, contraction-gap, and off-tube shell-rank assumptions hold for* (I.1) *on its own Ω. If all off-tube critical points have nonsingular $\nabla^2 \log p^{\mathrm{pop}}$ and $\nu \ll \mathrm{Leb}_d$ with $\nu(K) = 1$, then every trajectory from K converges and*

$$\operatorname{dist}(x_\infty, \mathcal{M}) \le \varepsilon_0 = \frac{C_{\mathrm{loc}} m}{2\tau} \sigma_{\min}^2 \quad \nu\text{-almost surely}. \tag{I.7}$$

Every local maximizer in Ω and every global maximizer obey the same bound without the hyperbolicity or initialization-density conditions. For a compactly supported population in an affine space $\mathcal{A} = a_0 + V$, the exact normal identity and convergence in (A.43) *hold for every start, without the curved manifold assumptions.*

Proof. The analytic argument gives a critical limit. The population radial bound excludes the intermediate contraction region. Off-tube hyperbolicity makes the compact off-tube critical set finite; its ascent-unstable equilibria have zero-volume stable sets by the same stable-manifold and inverse-flow argument as Lemma A.13. This proves the almost-sure bound. The Hessian test and Gaussian decay prove the optimizer statements as in Proposition A.15. In the affine case the posterior barycenter remains in $\mathcal{A}$, giving $\dot{n} = -\beta(x)n$ for $n = P_{V^\perp}(x - a_0)$. The analytic limit therefore lies in $\mathcal{A}$ and in the convex hull of the population support. □

I.3 Population coverage and the disappearance of sampling error

For the finite geometric schedule $\sigma_i = \sigma_{\min} e^{(i-1)h}$, $N \ge 2$ and $h > 0$, take $\pi = N^{-1} \sum_i \delta_{\sigma_i}$. With the same $A_N = h \sum_i \sigma_i^2$ as before,

$$A_N p^{\mathrm{pop}} = h \sum_i \sigma_i^2 p^Q_{\sigma_i}, \qquad \mathcal{R}_N = A_N \Delta p^{\mathrm{pop}} - (p^Q_{\sigma_{\max}} - p^Q_{\sigma_{\min}}). \tag{I.8}$$

For the population objective, the empirical–reference cutoff discrepancy vanishes exactly, so ϵ_{Pois} consists only of finite-grid quadrature error. The same L^1 quadrature argument used in the proof of Proposition B.7 applies with $Q_{\mathcal{M}}$ in place of $\widehat{Q}_M$, giving

$$\epsilon_{\mathrm{Pois}} \le dh \log \frac{\sigma_{\max}}{\sigma_{\min}} + \frac{h}{2}\sqrt{2d}. \tag{I.9}$$

This estimate is sufficient to show that the error vanishes as the log-scale grid is refined, but need not be quantitatively sharp in high ambient dimension.

The residual is integrable with zero integral. The preceding complete-flow and compact-superlevel properties allow the occupation-density proof of Lemma B.4 to be applied with p^{pop} in place of p_{MN}. Its mark lies on the *population ascent trajectory* and satisfies the same total-variation bound, with $\epsilon_{\mathrm{init}} = \mathrm{TV}(\nu, p^Q_{\sigma_{\max}}(x)\,dx)$. The reference-label coupling and Gaussian projection argument are unchanged.

Corollary I.2 (Population regional coverage)**.** *Under Proposition I.1, suppose $\mathcal{M}$ is connected and the radius ordering and Gaussian band conditions of Theorem B.6 hold for the population flow and the finite geometric schedule. Then $r(X_{\mathrm{stop}}) \le r_{\mathrm{stop}}$ almost surely, and for every Borel $U \subseteq \mathcal{M}$,*

$$\mathbb{P}_\nu[Z_{\mathrm{tr}} \in U,\, Z_{\mathrm{stop}} \in U] \ge (1 - \zeta) Q_{\mathcal{M}}(U^{-\delta_{\mathrm{cov}}}) - \epsilon_{\mathrm{Pois}} - \epsilon_{\mathrm{init}}, \tag{I.10}$$

where

$$\epsilon_{\mathrm{Pois}} = \frac{1}{2} \left\| A_N \Delta p^{\mathrm{pop}} - (p^Q_{\sigma_{\max}} - p^Q_{\sigma_{\min}}) \right\|_1 .$$

Here the stopping times, projections, and coverage resolution are defined as before, using the population field and its posterior-localization constant.

Proof. The population radial and projected-speed inequalities give the same full refinement-length bound. Apply the marked-point comparison above, then follow its reference label through the proof of Theorem B.6. □

Continuous log-scale mixing. For $\pi(d\sigma) = d\sigma/[\sigma \log(\sigma_{\max}/\sigma_{\min})]$ on $0 < \sigma_{\min} < \sigma_{\max} < \infty$, the same normalized objective becomes

$$p^{\mathrm{pop}}(x) = \frac{\int_{\sigma_{\min}}^{\sigma_{\max}} \sigma p^Q_\sigma(x)\, d\sigma}{\int_{\sigma_{\min}}^{\sigma_{\max}} \sigma\, d\sigma}. \tag{I.11}$$

Using $\partial_\sigma p^Q_\sigma = \sigma \Delta p^Q_\sigma$ gives the exact stationary balance

$$\nabla \cdot \left[\left(\int_{\sigma_{\min}}^{\sigma_{\max}} \sigma\, d\sigma \right) p^{\mathrm{pop}} \nabla \log p^{\mathrm{pop}} \right] = p^Q_{\sigma_{\max}} - p^Q_{\sigma_{\min}}. \tag{I.12}$$

Thus $\epsilon_{\mathrm{Pois}} = 0$. Under the corresponding population geometric and initialization hypotheses, coverage holds with only the Gaussian event loss and ϵ_{init}. Initializing exactly from $p^Q_{\sigma_{\max}}$ also removes ϵ_{init}, provided the geometric assumptions hold on the superlevels visited by that unbounded initialization. For the compactly supported initialization of the main theorem, or a Gaussian approximation to the broad smoothing, its mismatch must still be retained. A general scale measure π gives the convergence result above, but this exact two-cutoff balance specifically uses log-scale mixing with the stated σ^2 weight.

Distributional meaning of the coverage bound. The proof also yields an ambient Wasserstein bound for the stopped *population* output:

$$W_1(P_{X_{\mathrm{stop}}}, Q_{\mathcal{M}}) \le r_{\mathrm{stop}} + \delta_{\mathrm{cov}} + \mathrm{diam}(\mathcal{M}) \min\{1, \zeta + \epsilon_{\mathrm{Pois}} + \epsilon_{\mathrm{init}}\}. \tag{I.13}$$

Indeed, couple the actual mark to $Y^Q \sim p^Q_{\sigma_{\min}}$ with mismatch probability at most $\epsilon_{\mathrm{Pois}} + \epsilon_{\mathrm{init}}$, and, given Y^Q, attach Z through the reference posterior kernel in (B.29). Then $Z \sim Q_{\mathcal{M}}$. On mark agreement and the Gaussian good event, $d_{\mathcal{M}}(Z, Z_{\mathrm{stop}}) \le \delta_{\mathrm{cov}}$; on their complement the Euclidean distance between these two manifold points is at most $\mathrm{diam}(\mathcal{M})$. Adding $\|X_{\mathrm{stop}} - Z_{\mathrm{stop}}\| \le r_{\mathrm{stop}}$ proves the bound. In a family satisfying the hypotheses with the displayed resolution and probability errors tending to zero on a fixed compact manifold, the stopped output distributions therefore converge to $Q_{\mathcal{M}}$ in W_1. This is a conditional population-approximation conclusion, not exact distribution recovery at fixed positive cutoffs.

Scope and relation to empirical non-memorization. The population results require the corresponding posterior and landscape conditions on p^{pop}. General regional coverage concerns the stopped output, while endpoint diversity additionally requires the tail-retention control of Appendix G.3. The proper sphere mixtures provide concrete population examples, and the translation-invariant flat model serves as a local analytic reference. Penalty guidance extends under the guided-region conditions of Appendix H.

Corollary 3.2 addresses the complementary empirical question of diverse endpoints separated from a finite training set. Exact reverse diffusion initialized from the appropriate noisy distribution instead approaches its center distribution in the zero-noise limit (Song et al., 2020): $Q_{\mathcal{M}}$ for a population score and $\widehat{Q}_M$ for an empirical score (Appendix D.8).