CONVERGENCE THEORY OF DECENTRALIZED DIFFUSION MODELS VIA JOINT-SPACE GAUSSIAN CANCELLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion probabilistic models (DPMs) have demonstrated success in generative modeling, motivating a growing theory of their convergence. Recent decentralized DPMs combine experts trained on separate data partitions through a learned router, raising an additional question about how routing affects sampling accuracy. We develop *Joint-Space Gaussian Cancellation* (JGC), a flexible convergence analysis pipeline for standard and decentralized DPMs with DDPM sampling. Our key insight is to analyze the Gaussian reverse bridge jointly with its noisy input, retaining posterior cancellations before bounding the error. An exact Gaussian-channel interpolation expresses the complete finite-step defect through a signed divergence; a posterior-covariance potential then controls its accumulation without losing an extra dimension factor. This yields an explicit $O(\{d \log^2 N\}/N)$ total-variation discretization bound on equal log-SNR grids with polynomially moving SNR endpoints, merely under a finite second data moment. Here d is the data dimension and N the number of denoising steps. Extending the analysis classwise preserves this convergence order, while Gaussian overlap quantifies the additional routing error through expert separation without requiring a probability floor. These two results connect linear-dimensional convergence theory with the additional approximation cost of decentralized routing.

1 INTRODUCTION

Diffusion probabilistic models (DPMs) generate data by learning to reverse a process that adds Gaussian noise (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b). Their success in generative modeling has motivated extensive study of their convergence theory: how accurately does a finite sequence of learned reverse transitions approximate the target distribution? This question is especially relevant to decentralized DPMs, which combine experts trained on different data partitions through a learned router (McAllister et al., 2025; Dong et al., 2024; Chen et al., 2025).

The forward diffusion corrupts data into noise, while the reverse process uses a learned score or denoiser to recover the data distribution. Sampling introduces initialization, network, and discretization errors. The discretization order determines how the last term decreases with the number of steps. Establishing this order without strong data regularity is a central challenge. Figure 1 summarizes these error budgets and the additional routing term in decentralized models.

Significant progress has been made toward this goal. Under finite-moment assumptions, smoothness-free analyses give total variation (TV) discretization bounds of order $\widetilde{O}(d/\sqrt{N})$ or $\widetilde{O}(\sqrt{d/N})$ (Chen et al., 2023; Benton et al., 2024), where d is the data dimension and N the number of steps. More recent analyses establish first-order step dependence: Li & Yan (2025b) prove $\widetilde{O}(d/N)$ for one DDPM variance choice and $\widetilde{O}(k/N)$ for posterior variance under intrinsic covering assumptions, with covering dimension k; Liang et al. (2026) obtain a related result under bounded-support and covering conditions. Table 1 compares these rates together with their sampler choices and assumptions.

Decentralized models add a distinct question: *How does error in a router's probabilities affect the generated distribution?* For per-step routing, each reverse transition is a mixture of expert kernels. Its error depends on the separation of the experts as well as the router's accuracy: redistributing

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.

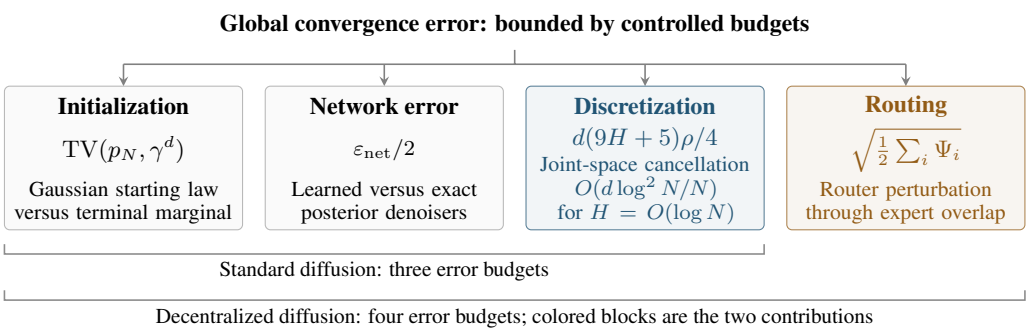


Figure 1: Global error budgets: three for standard diffusion and four with routing. Blue highlights $O(d\log^2 N/N)$ discretization; amber highlights routing. The rate uses $H = O(\log N)$ and $\rho = e^{H/N} - 1 \le 1$. Equation (20) gives additive upper bounds; Theorem 3.2 keeps the tighter unsplit network-and-routing term.

probability among nearly coincident predictions has little effect. Existing analysis of routed ODE steps gives Wasserstein guarantees under uniform approximation assumptions (Tang et al., 2026). Here we study Gaussian reverse transitions in TV, with learning errors measured under the true noisy marginals.

In this paper, we develop *Joint-Space Gaussian Cancellation* (JGC) to obtain $O(d\log^2 N/N)$ discretization for standard and decentralized DPMs using the posterior-variance DDPM sampler. Conditioning the reverse transition on the clean datum exposes a Gaussian bridge; averaging its centered posterior residual preserves cancellation before taking absolute values. We then interpolate through exact Gaussian observation channels and control the complete finite-step error by a posterior-covariance potential. The potential controls boundary terms that telescope across the log-SNR grid, yielding linear dimension dependence without a Taylor remainder (Figure 4). This argument requires only a finite second data moment, with no score smoothness, bounded support, or covering condition. A complementary two-time cancellation estimate applies on irregular grids.

Our contributions.

- **Linear-dimension, first-order discretization through joint-space cancellation.** Theorem 3.1 gives a finite-grid TV guarantee with constant, dimension, and grid dependence separated. On an equal log-SNR grid with $H = O(\log N)$, JGC establishes $O(d\log^2 N/N)$ discretization under a finite second moment (Theorem D.1). Its explicit bound holds whenever the relative SNR increment is at most one, with no dimension-dependent step restriction.
- **An explicit routing-error analysis for decentralized diffusion.** Theorem 3.2 transfers the accumulated estimate classwise, retaining $O(d\log^2 N/N)$ discretization while isolating routing error. Gaussian overlap quantifies its dependence on expert separation without a probability floor. Under bounded expert diameter, the routing coefficient stays bounded as a fixed-endpoint grid is refined (Corollary 3.3).

The rate matches the linear dimension and first-order step dependence of recent sharp TV theory, with sampler choices and assumptions distinguished in Table 1. Figure 2 maps the proof dependencies: Sections 3, 4, and 5 respectively state the guarantees, develop JGC, and analyze routing.

2 Preliminaries

We use the x-prediction parameterization: $\bar{y}$ denotes an exact posterior denoiser and y_θ its learned approximation. The corresponding routing probabilities are a and a_ϕ. Table 2 collects the notation, including the kernel and error symbols used in the proofs.

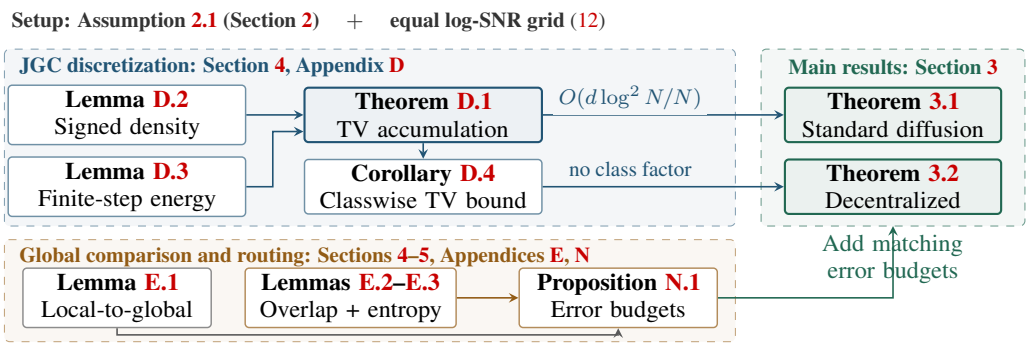


Figure 2: **Proof dependencies and section map.** Incoming arrows identify jointly used ingredients; the shared budget branch feeds both main theorems. The displayed rate assumes $H = O(\log N)$.

2.1 Diffusion models and reverse kernels

Assumption 2.1 (Data and grid)**.** Let $d, N \geq 1$ be integers and p_d a Borel probability measure on $\mathbb{R}^d$ with $M = \int \|y\|^2 p_d(dy) < \infty$. Fix a sampling grid $\mathcal{D} = \{0 < t_{\min} = t_0 < \cdots < t_N = t_{\max} < 1\}$, with $1 > \alpha_0 > \cdots > \alpha_N > 0$ and $\sigma_i^2 = 1 - \alpha_i^2$.

The data law may be discrete. On a common probability space, take $Y \sim p_d$ and independent standard Gaussian innovations $Z_0, \ldots, Z_N$, independent of the data and any class label introduced below. Define

$$X_{t_0} = \alpha_0 Y + \sigma_0 Z_0, \qquad X_{t_i} = \frac{\alpha_i}{\alpha_{i-1}} X_{t_{i-1}} + \sqrt{1 - \frac{\alpha_i^2}{\alpha_{i-1}^2}}\, Z_i. \tag{1}$$

Each time marginal therefore has the representation

$$X_t \overset{\text{law}}{=} \alpha_t Y + \sigma_t Z, \qquad Z \sim \mathcal{N}(0, I_d), \qquad \alpha_t^2 + \sigma_t^2 = 1. \tag{2}$$

Here Z is independent of Y; equation (2) specifies one marginal, while (1) specifies the joint process. Write $p_t = \mathcal{L}(X_t)$, $p_i = p_{t_i}$, $\alpha_i = \alpha_{t_i}$, and $\sigma_i = \sigma_{t_i}$; $\alpha_{t|s} = \alpha_t/\alpha_s$ and $\sigma_{t|s}^2 = 1 - \alpha_{t|s}^2$. In particular, p_0 denotes the *early-stopped marginal*, whereas p_d denotes the clean data law. We use $\mathrm{TV}(P, Q) = \sup_B |P(B) - Q(B)| = \frac{1}{2}|P - Q|(\mathbb{R}^d)$ and natural logarithms; B ranges over Borel sets. For a density p and a kernel K, pK denotes the output marginal $\int p(x)K(x, \cdot)\, dx$.

The sampler studied here. For a step $s = t_{i-1} < t = t_i$, define

$$\begin{aligned} r_i &:= 1 - \alpha_i^2/\alpha_{i-1}^2, \qquad A_i := \frac{\alpha_i \sigma_{i-1}^2}{\alpha_{i-1}\sigma_i^2}, \\ B_i &:= \frac{\alpha_{i-1} r_i}{\sigma_i^2}, \qquad v_i := \frac{r_i \sigma_{i-1}^2}{\sigma_i^2}. \end{aligned} \tag{3}$$

The exact Gaussian bridge, conditional on Y, is

$$X_{t_{i-1}} \mid (X_{t_i} = x, Y = y) \sim \mathcal{N}(A_i x + B_i y, v_i I_d). \tag{4}$$

Consequently, the exact reverse kernel $K_i(x, \cdot)$ is the mixture of these Gaussians over the posterior law of Y given $X_{t_i} = x$, and $p_i K_i = p_{i-1}$. DDPM replaces y by a denoising prediction (Ho et al., 2020). Its oracle and learned kernels in this paper are

$$\begin{aligned} Q_i^{\mathrm{std}}(x, \cdot) &= \mathcal{N}(A_i x + B_i \bar{y}(x, t_i), v_i I_d), \\ P_i^{\mathrm{std}}(x, \cdot) &= \mathcal{N}(A_i x + B_i y_\theta(x, t_i), v_i I_d), \end{aligned} \tag{5}$$

where $\bar{y}(x, t) = \mathbb{E}[Y \mid X_t = x]$. Canonical Gaussian Bayes versions of all posteriors are specified in Appendix B.1. We fix the *posterior bridge variance* v_i, also denoted $\sigma_{s|t}^2$; conclusions for other DDPM variance choices require a separate comparison. Denoiser and score parameterizations satisfy $\nabla \log p_t(x) = (\alpha_t \bar{y}(x, t) - x)/\sigma_t^2$. Network errors in the denoiser and score parameterizations differ by the factor α_t/σ_t^2.

Table 1: **Linear-dimension, first-order discretization with learned routing.** JGC gives $O(d \log^2 N/N) = \widetilde{O}(d/N)$ TV discretization for posterior-variance DDPM under a finite second moment, including expert routing. Initialization and learning terms are separate.

Analysis / sampler	TV rate	Data assumption	Learning control
Chen et al. (2023) Exponential integrator	$\widetilde{O}(d/\sqrt{N})$	Finite second moment	L^2 score errors
Benton et al. (2024) Exponential integrator	$\widetilde{O}(\sqrt{d/N})$	Second moment $O(d)$	L^2 score errors
Li & Yan (2025b, Thm. 1) DDPM; other variance*	$\widetilde{O}(d/N)$	Finite first moment†	L^2 score errors
Li & Yan (2025b, Thm. 3) Posterior-variance DDPM	$\widetilde{O}(k/N)$	First moment†; intrinsic covering	L^2 score errors
Liang et al. (2026, Thm. 4) Posterior-variance DDPM	$\widetilde{O}(k/N)$	Bounded support; intrinsic covering	L^2 score errors
Jain & Zhang (2025) ODE step plus forward noising	$\widetilde{O}(d/N)$	Finite second moment	L^2 score errors
JGC: standard DDPM Posterior variance	$\widetilde{\boldsymbol{O}}(\boldsymbol{d/N})$	**Finite second moment**	L^2 network errors
JGC: routed DDPM Posterior variance	$\widetilde{\boldsymbol{O}}(\boldsymbol{d/N})$	**Finite second moment**	L^2 network and routing errors; **no probability floor**‡

Both JGC rows use equal log-SNR grids of span H with $e^{H/N} - 1 \le 1$ (Theorem D.1), without support or covering conditions. Here $\widetilde{O}$ hides logs and k is covering dimension. For JGC, $H = O(\log N)$ gives $O(d \log^2 N/N)$; fixed H gives $O(d/N)$. Li–Yan and Liang et al. move the endpoints with N, and their k/N results can improve on ambient dimension. All rows target noisy marginals and use true-marginal L^2 errors, with the cited time weights. The first two rates and Jain–Zhang's rate follow by Pinsker. *Li–Yan's Theorem 1 variance differs from posterior variance; Appendix M gives the exact coefficients. †The first moment grows at most polynomially in N. ‡The L^2 router statement uses bounded expert diameter (Corollary 3.3).

2.2 Decentralized diffusion models

Let $L \ge 1$ be an integer and $J \in \{1, \dots, L\}$ a cluster label with positive priors π_l summing to one and conditional data laws p_d^l; $p_d = \sum_l \pi_l p_d^l$. The forward noise is independent of (Y, J). The router uses precisely this label set. No geometric separation or uniform lower bound on π_l is imposed. Define the normalized class marginal p_t^l, the posterior weight $a^l(x,t) = \mathbb{P}(J = l \mid X_t = x)$, and the class denoiser $\bar{y}^l(x,t) = \mathbb{E}[Y \mid X_t = x, J = l]$. Write $p_i^l = p_{t_i}^l$ and $a = (a^1, \dots, a^L)$. Bayes' rule gives the identity central to the extension:

$$p_t(x) a^l(x,t) = \pi_l p_t^l(x). \tag{6}$$

With $a_\phi(x,t) \in \Delta_L = \{w \in [0,1]^L : \sum_l w_l = 1\}$ a learned probability vector, the oracle and learned mixture kernels are

$$\begin{aligned} Q_i(x,\cdot) &= \sum_{l=1}^{L} a^l(x,t_i) \mathcal{N}(A_i x + B_i \bar{y}^l(x,t_i), v_i I_d), \\ P_i(x,\cdot) &= \sum_{l=1}^{L} a_\phi^l(x,t_i) \mathcal{N}(A_i x + B_i y_\theta^l(x,t_i), v_i I_d). \end{aligned} \tag{7}$$

The implemented step draws $J_i \sim a_\phi(x,t_i)$, then samples its Gaussian component, redrawing the label at every step (Figure 3). This is categorical expert selection (McAllister et al., 2025, Section 4.2); Appendix G distinguishes it from averaged-prediction and fixed-class sampling. All samplers start from $\nu = \gamma^d = \mathcal{N}(0, I_d)$ at t_N. Write $q_0^{\mathrm{std}} = \nu P_N^{\mathrm{std}} \cdots P_1^{\mathrm{std}}$ and $q_0 = \nu P_N \cdots P_1$ for their output laws.

Noise increments and error budgets. Network error concerns denoiser/expert predictions; routing error concerns router probabilities. *Learning* combines the two. Two signal-to-noise ratio (SNR)

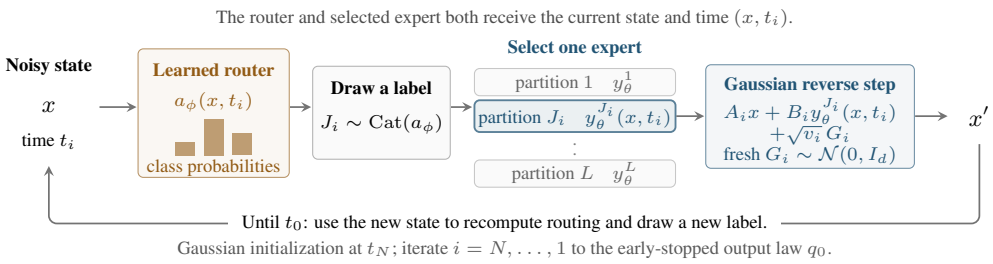


Figure 3: Per-step sampling in a decentralized diffusion model. Experts are trained on separate data partitions. At (x, t_i), draw J_i from the router's probabilities and use that expert's prediction in the Gaussian update, with posterior bridge variance v_i and fresh independent noise G_i. Recomputing routing and redrawing the label at each new state implements the mixture kernel (7).

increments control different errors:

$$\eta_i = \frac{\alpha_i^2}{\sigma_i^2}, \quad \rho_i = \frac{\eta_{i-1}}{\eta_i} - 1, \quad \Delta_i = \eta_{i-1} - \eta_i = \frac{B_i^2}{v_i}, \quad \Gamma = \sum_{i=1}^{N} \Delta_i = \eta_0 - \eta_N. \tag{8}$$

Relative increments ρ_i determine discretization, while absolute increments $\Delta_i = B_i^2/v_i$ measure squared network errors relative to the injected noise. For the mixture sampler, define

$$\begin{aligned} E_i^2 &:= \mathbb{E}_{X\sim p_i} \sum_l a^l(X, t_i) \|y_\theta^l(X, t_i) - \bar{y}^l(X, t_i)\|^2, \\ W_i^2 &:= \mathbb{E}_{X\sim p_i} \|a_\phi(X, t_i) - a(X, t_i)\|_1^2. \end{aligned} \tag{9}$$

Both expectations use the *true* noisy data marginal, as in population training objectives. For the standard sampler, set $e_i^2 = \mathbb{E}_{p_i}\|y_\theta - \bar{y}\|^2$. All denoisers are Borel functions into $\mathbb{R}^d$ (finite at every input), and the router is Borel into Δ_L; e_i, E_i are assumed finite where used. Routing stability depends on differences between learned experts. Define

$$\begin{aligned} D_i(x) &:= \max_{l,j} \|y_\theta^l(x, t_i) - y_\theta^j(x, t_i)\|, \\ \Psi_i &:= \mathbb{E}_{p_i} \log\left[1 + (e^{\Delta_i D_i(X)^2} - 1)\|a_\phi(X, t_i) - a(X, t_i)\|_1^2\right]. \end{aligned} \tag{10}$$

The main mixture bound uses this logarithmic routing energy; bounded predictions or predictor derivatives are not required. Exact routing gives $\Psi_i = 0$ even for unbounded diameters. We also use the sufficient upper bound

$$\Omega_i = \mathbb{E}_{p_i}\left[(e^{\Delta_i D_i(X)^2} - 1)\|a_\phi(X, t_i) - a(X, t_i)\|_1^2\right] \geq \Psi_i. \tag{11}$$

3 Discretization Rates and the Additional Routing Error

Write $\varepsilon_{\mathrm{net,std}}^2 = \sum_i \Delta_i e_i^2$ and $\varepsilon_{\mathrm{net,dec}}^2 = \sum_i \Delta_i E_i^2$ for the true-marginal network error energies. The two theorems use the explicit grid

$$H = \log(\eta_0/\eta_N), \qquad \eta_i = \eta_0 e^{-iH/N}, \qquad \rho = e^{H/N} - 1 \leq 1. \tag{12}$$

Thus $N \geq H/\log 2$ suffices for the mesh condition, independently of d.

Theorem 3.1 (Standard DDPM: explicit discretization rate)**.** *Under Assumption 2.1 and grid (12), use the standard posterior-variance DDPM kernels with finite e_i. Then*

$$\mathrm{TV}(p_0, q_0^{\mathrm{std}}) \leq \mathrm{TV}(p_N, \gamma^d) + \underbrace{\frac{d(9H+5)}{4}(e^{H/N} - 1)}_{\textit{discretization}} + \frac{1}{2}\varepsilon_{\mathrm{net,std}}. \tag{13}$$

Theorem 3.2 (Decentralized DDPM: explicit routing cost)**.** *Under Assumption 2.1 and grid (12), use the per-step mixture kernels and positive-prior class representation of Section 2. If the learning terms below are finite, then*

$$\mathrm{TV}(p_0, q_0) \le \mathrm{TV}(p_N, \gamma^d) + \underbrace{\frac{d(9H+5)}{4}(e^{H/N}-1)}_{\text{same linear-dimensional rate}} + \left[\frac{\varepsilon^2_{\mathrm{net,dec}}}{2} + \underbrace{\frac{1}{2}\sum_{i=1}^{N}\mathbb{E}_{p_i}\log\Big(1+(e^{\Delta_i D_i(X)^2}-1)\|a_\phi(X,t_i)-a(X,t_i)\|_1^2\Big)}_{\text{additional router energy}}\right]^{1/2}. \tag{14}$$

Explicit convergence order: $O(d\log^2 N/N)$**.** The discretization term separates dimension, steps, and noise interval:

$$\frac{d(9H+5)}{4}(e^{H/N}-1) \le \underbrace{\frac{e^{H/N}H(9H+5)}{4}}_{\text{explicit interval factor}}\ \underbrace{\frac{d}{N}}_{\text{dimension and steps}}. \tag{15}$$

Since $e^{H/N} \le 2$, polynomially moving SNR endpoints with $H = O(\log N)$ give $O(d\log^2 N/N) = \widetilde{O}(d/N)$ discretization in both theorems. Fixed H gives $O(d/N)$. The other error budgets remain separate; Appendix I maps the assumptions.

For subsequent proofs, we abbreviate these explicit terms as

$$\mathcal{D}_N := \frac{d(9H+5)}{4}(e^{H/N}-1), \qquad \mathcal{I}_N := \mathrm{TV}(p_N, \gamma^d). \tag{16}$$

The router energy in Theorem 3.2 is $\frac{1}{2}\sum_i \Psi_i$. Bounded expert diameter gives a simpler dependence on the router's L^2 error:

Corollary 3.3 (Bounded expert diameter)**.** *In the setting of Theorem 3.2, if $D_i(x) \le D < \infty$ for p_i-almost every x and all i, then*

$$\mathrm{TV}(p_0, q_0) \le \mathcal{I}_N + \mathcal{D}_N + \left[\frac{1}{2}\sum_i \left\{\Delta_i E_i^2 + (e^{D^2\Delta_i}-1)W_i^2\right\}\right]^{1/2}. \tag{17}$$

If $E_i \le E$, $W_i \le W$, and $\Delta_* = \max_i \Delta_i$, its learning term is at most

$$\sqrt{\Gamma/2}\left(E + De^{D^2\Delta_*/2}W\right). \tag{18}$$

A uniform output bound $\|y_\theta^l\| \le R$ is one sufficient condition, with $D = 2R$; a shared unbounded component of the experts cancels from D_i.

Appendix D.6 sharpens the interval factor using M; Appendix L retains within-class uncertainty, giving zero oracle discretization for one-atom classes. For fixed D and endpoints, the learning coefficients stay bounded under refinement; nonzero network or routing errors can still prevent convergence. Appendix H relates these errors to population risks with training-time masses Δ_i/Γ on the sampling grid.

Endpoint choice and learning accuracy. Taking $\eta_0 = N^2$ and $\eta_N = N^{-2}$ gives $\sigma_0 \asymp N^{-1}$, $H = 4\log N$, and initialization $O(\sqrt{M}/N + \sqrt{d}/N^2)$. For fixed d, p_d, the oracle clean-data error is $O(\log^2 N/N)$ in bounded-Lipschitz distance (Corollary F.2); clean-data TV is unavailable in general for singular laws. A vanishing learning term gives consistency; a term $O(\log^2 N/N)$ preserves this rate. These are population-accuracy conditions, not training guarantees (Appendix H). Even exact Gaussian sampling has $\Theta(\log N/N)$ TV error on this grid (Proposition C.6), ruling out an endpoint-independent $O(1/N)$ bound here, but not establishing necessity of the second logarithm. Appendix M.1 compares prior early-stopping analyses.

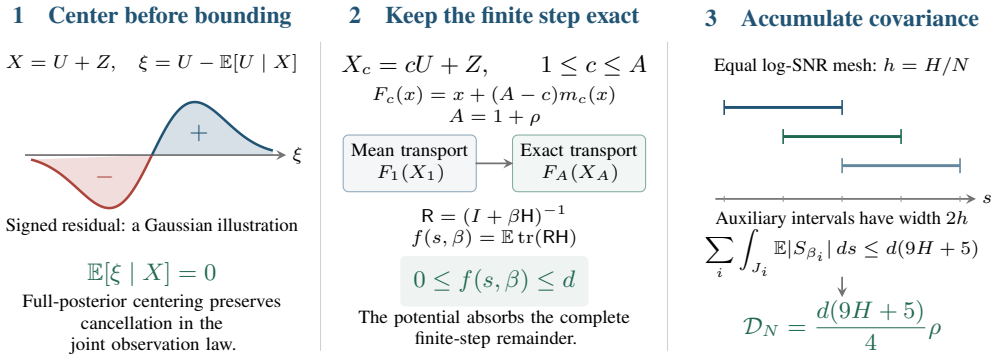


Figure 4: JGC: centering, exact interpolation, and covariance accumulation. Left: $\xi p(\xi \mid X = 0)$ for independent $U, Z \sim \mathcal{N}(0, 1)$, so $\xi \mid X = 0 \sim \mathcal{N}(0, 1/2)$; full-posterior centering holds for general data. Middle: exact channels retain the complete finite step. Right: covariance boundary terms telescope over overlapping log-SNR intervals, giving $O(d \log^2 N/N)$ for $H = O(\log N)$ (Theorem D.1).

4 Error Decomposition and Joint-Space Analysis

Figure 2 maps the main proof dependencies.

4.1 From Local Errors to a Global Guarantee

A dominated comparison (Li & Yan, 2025b, Sections 4.2–4.4) transfers true-marginal L^2 errors to the generated law through an absorbing-state construction. For oracle kernels Q_i and learned kernels P_i with transition densities, and reference densities p_i, the construction proves

$$\begin{aligned}\mathrm{TV}(p_0, q_0) \le \mathrm{TV}(p_N, \nu) + \sum_i \mathrm{TV}(p_{i-1}, p_i Q_i) \\ + \sqrt{\frac{1}{2}\sum_i \mathbb{E}_{p_i}\mathrm{KL}(Q_i(X, \cdot)\|P_i(X, \cdot))}.\end{aligned} \tag{19}$$

Lemma E.1 constructs auxiliary densities dominated by each p_i: discretization bounds lost mass, while path KL controls learning. For standard DDPM, equal Gaussian covariances give $\mathrm{KL}(Q_i^{\mathrm{std}}(x)\|P_i^{\mathrm{std}}(x)) = \frac{1}{2}\Delta_i\|\bar{y}(x, t_i) - y_\theta(x, t_i)\|^2$, which, together with the accumulated estimate below, proves Theorem 3.1.

With the budgets in (16), Figure 1 displays the additive bounds

$$\begin{aligned}\mathrm{TV}(p_0, q_0^{\mathrm{std}}) \le \mathcal{I}_N + \tfrac{1}{2}\varepsilon_{\mathrm{net,std}} + \mathcal{D}_N, \\ \mathrm{TV}(p_0, q_0) \le \mathcal{I}_N + \tfrac{1}{\sqrt{2}}\varepsilon_{\mathrm{net,dec}} + \mathcal{D}_N + \sqrt{\tfrac{1}{2}\sum_i \Psi_i}.\end{aligned} \tag{20}$$

Proposition N.1 proves these upper bounds. Theorem 3.2 retains the tighter unsplit square root; exact routing recovers coefficient $1/2$ (Corollary 5.1).

4.2 Joint-Space Cancellation through Exact Gaussian Channels

The exact bridge (4) allows us to remove its common final Gaussian noise by TV contraction. After rescaling, the problem becomes a Gaussian observation model $X = U + Z$, with $Z \sim \mathcal{N}(0, I_d)$ independent of U. Writing $m_U(x) = \mathbb{E}[U \mid X = x]$ and $\varepsilon = \rho_i$, we compare

$$\mathcal{L}(X + \varepsilon U) \quad \text{and} \quad \mathcal{L}(X + \varepsilon m_U(X)). \tag{21}$$

The residual $\xi = U - m_U(X)$ is centered under the full posterior. In particular, with $F_0 = X + \varepsilon m_U(X)$ and any smooth compactly supported test function f,

$$\mathbb{E}[\xi \mid X] = 0, \qquad \mathbb{E}\big[\nabla f(F_0)^T \xi\big] = 0. \tag{22}$$

Taking absolute values before this averaging loses an order even for Gaussian data: the conditional transition error is $\Theta(\rho_i)$ while its marginal error is $\Theta(\rho_i^2)$ (Proposition C.4).

To retain the cancellation for a finite step, set $A = 1 + \varepsilon$ and interpolate through the *exact* channels

$$X_c = cU + Z, \quad m_c(x) = \mathbb{E}[U \mid X_c = x], \quad F_c(x) = x + (A - c)m_c(x), \qquad 1 \le c \le A. \tag{23}$$

The laws of $F_1(X_1)$ and $F_A(X_A)$ are exactly the two laws in (21). At channel c, introduce the normalized covariance $\mathsf{H} = \mathrm{Cov}(cU \mid X_c)$, $\beta = (A-c)/c$, and the inverse Jacobian $\mathsf{R} = (I+\beta\mathsf{H})^{-1}$ of F_c. Posterior centering and Gaussian integration by parts identify the signed density

$$S_\beta = \frac{1}{p}\,\mathrm{div}\{\mathsf{R}\,\mathrm{div}(p\mathsf{H})\}. \tag{24}$$

The divergences act on the noisy observation, with β fixed; p is its density. For $s = \log(\eta_i c^2)$, the exact interpolation gives

$$\mathrm{TV}(p_{i-1}, p_i Q_i^{\mathrm{std}}) \le \frac{1}{4}\int_{\log\eta_i}^{\log\eta_i + 2\log(1+\rho_i)} \beta(s)\,\mathbb{E}|S_\beta(s)|\,ds. \tag{25}$$

Pointwise radial-moment estimates incur a quadratic dimension factor; we preserve covariance contractions until after accumulation.

4.3 Covariance accumulation gives linear dimension

Along $X_\eta = \sqrt{\eta}Y + Z$, write $r(s) = \mathbb{E}\,\mathrm{tr}\,\mathsf{H}$ and $a(s) = \mathbb{E}\,\mathrm{tr}(\mathsf{H}^2)$, where $s = \log\eta$. Gaussian observation identities give

$$0 \le r(s) \le d, \qquad r'(s) = r(s) - a(s). \tag{26}$$

The covariance can be large at individual noise levels; the integral of a is controlled by (26). Let ξ be the centered normalized residual at the current channel, $\mathsf{T} = \mathbb{E}[\xi^{\otimes 3} \mid X]$, and $\mathcal{B}_\beta = \mathbb{E}\|\mathsf{R}^{1/2}\mathsf{T}[\mathsf{R}\xi]\mathsf{R}^{1/2}\|_F^2$. This nonnegative contraction measures the finite-step contribution. The signed Gaussian divergence identity yields

$$\mathbb{E}|S_\beta| \le \sqrt{(r+a)(r+a+\beta^2\mathcal{B}_\beta)} + r + a. \tag{27}$$

The decisive estimate uses the bounded covariance potential $f(s,b) = \mathbb{E}\,\mathrm{tr}\{\mathsf{H}(I+b\mathsf{H})^{-1}\}$. Its Itô drift and the symmetry of the third posterior moment give

$$0 \le f(s,\beta) \le r \le d, \qquad \beta^2\mathcal{B}_\beta \le r - \frac{d}{ds}f(s,\beta(s)) \quad \text{when } \beta' = -(1+\beta)/2,\ \beta \le 1. \tag{28}$$

The total derivative includes the changing β; this absorbs the entire finite-step term without a Taylor truncation.

On grid (12), let $h = H/N$, $s_i = \log\eta_i$, and $J_i = [s_i, s_i + 2h]$. These intervals have twice the grid width because the exact transport multiplies the signal by $1 + \rho$. Their endpoint contributions telescope:

$$\sum_{i=1}^{N}\{r(s_i) - r(s_i + 2h)\} = r(s_N) + r(s_{N-1}) - r(s_0) - r(s_0 + h) \le 2d. \tag{29}$$

Consequently the indexed integrals of $r + a$ and $\beta_i^2\mathcal{B}_{\beta_i}$ are at most $d(4H + 2)$ and $d(2H + 2)$, respectively. Cauchy–Schwarz in (27), followed by (25) and $\beta_i \le \rho$, gives $\sum_i \mathrm{TV}(p_{i-1}, p_i Q_i^{\mathrm{std}}) \le d(9H+5)\rho/4$. Theorem D.1 proves every identity first for bounded data, then passes to finite second moments by posterior convergence and lower semicontinuity of TV.

The control $r \le d$ is in expectation, so large pointwise covariances are allowed. The equal grid aligns the boundary terms of r and f: they telescope before applying dimension bounds, avoiding a factor of N without uniform local regularity.

For the mixture kernel, identity (6) preserves the same argument within each class:

$$\sum_i \mathrm{TV}(p_{i-1}, p_i Q_i) \le \sum_l \pi_l \sum_i \mathrm{TV}(p_{i-1}^l, p_i^l Q_i^l) \le \mathcal{D}_N. \tag{30}$$

The identity $p_i a^l = \pi_l p_i^l$ preserves the state-dependent weights inside the integrals. Uniform classwise constants remove any class-count factor.

This cumulative guarantee is not a uniform $O(d\rho_i^2)$ local bound. Appendix C gives complementary irregular-grid estimates and Gaussian examples; Proposition D.5 refines the interval dependence using the data moment.

5 Routing stability through expert geometry

For Gaussians $g_l = \mathcal{N}(\mu_l, vI_d)$, probability vectors a, b, and $D_\mu = \max_{l,j} \|\mu_l - \mu_j\|$, Appendix E proves

$$\chi^2\left(\sum_l a_l g_l \middle\| \sum_l b_l g_l\right) \le (e^{D_\mu^2/v} - 1)\|a - b\|_1^2. \tag{31}$$

The simplex identity $\sum_l (a_l - b_l) = 0$ removes the common Gaussian component, and equal-covariance Gaussian overlap controls the remainder, even with zero weights. Here $D_\mu = B_i D_i(x)$, so $D_\mu^2/v_i = \Delta_i D_i(x)^2$: expert separation is measured relative to the step's noise.

Retaining the logarithmic routing energy. At a fixed input, let R_i use the true weights a^l and learned means y_θ^l. Gaussian KL and data processing give $\mathrm{KL}(Q_i\|R_i) \le \frac{1}{2}\Delta_i \sum_l a^l \|\bar{y}^l - y_\theta^l\|^2$. The entropy inequality

$$\mathrm{KL}(Q\|P) \le 2\mathrm{KL}(Q\|R) + \log\big(1 + \chi^2(R\|P)\big) \tag{32}$$

and (31) yield $k_i := \mathbb{E}_{p_i}\mathrm{KL}(Q_i\|P_i) \le \Delta_i E_i^2 + \Psi_i$, proving Theorem 3.2 through (19). Retaining the logarithm gives a weaker integrability requirement than using $\log(1+u) \le u$:

$$\Psi_i \le \Omega_i, \qquad \Psi_i \le 4\Delta_i \mathbb{E}_{p_i} D_i(X)^2. \tag{33}$$

Thus linear expert differences under Gaussian inputs may have finite Ψ_i even when $\Omega_i = \infty$ (Proposition E.6). Bounded diameter further gives the linear dependence on W in Corollary 3.3.

Combining overlap and coupling. A maximal coupling of labels gives $k_i \le \frac{1}{2}\Delta_i(E_i + T_i)^2$, where

$$T_i^2 = \mathbb{E}_{p_i}\left[D_i(X)^2 \|a - a_\phi\|_1/2\right]. \tag{34}$$

Proposition E.5 proves this quadratic-moment bound. The two estimates can be combined before summing.

Corollary 5.1 (Combined routing stability)**.** *Under Assumption 2.1, grid (12), and the mixture setup, suppose $E_i < \infty$ and $\sum_i \kappa_i < \infty$, where*

$$\begin{aligned} \kappa_i &= \min\left\{\Delta_i E_i^2 + \Psi_i, \frac{\Delta_i}{2}(E_i + T_i)^2\right\}, \\ \mathrm{TV}(p_0, q_0) &\le \mathcal{I}_N + \mathcal{D}_N + \sqrt{\frac{1}{2}\sum_i \kappa_i}. \end{aligned} \tag{35}$$

The bounds are interpreted in the extended sense; the entropy branch may be infinite. Exact routing gives $\Psi_i = T_i = 0$ and the standard network error coefficient $1/2$.

Overlap gives quadratic small-error sensitivity where coupling can be linear (Proposition E.7), without a larger finiteness domain (Proposition E.6). Appendix I gives applicability conditions; Proposition E.8 rules out conditional-KL control from unweighted routing error alone.

6 Conclusion

Joint-Space Gaussian Cancellation gives $O(\{d\log^2 N\}/N)$ TV discretization under a finite second moment on equal log-SNR grids with $H = O(\log N)$. For decentralized diffusion models, classwise analysis preserves this rate without an explicit factor in the number of experts. Gaussian overlap separates routing error from network error and quantifies its dependence on expert separation, without a router probability floor.

Ethics statement

The analysis concerns sampling accuracy in decentralized diffusion models. Keeping training data at separate sites does not by itself provide a formal privacy guarantee, and none is claimed here. As with other image generators, deployment requires attention to data rights and potential misuse of generated content.

AI use statement

During the paper revision, Codex assisted with figure polishing and proof checking, manuscript editing, and LaTeX preparation. The experimental values are retained from pure human experiments.

Reproducibility statement

The mathematical setup, noise grid, and error conventions are explicit in Sections 2–5. Appendices provide the notation table, complete proofs, and population-risk identities. Supplementary algorithms and numerical illustrations, including their limitations, appear in Appendix J.

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. *arXiv preprint arXiv:2206.07309*, 2022.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

Hansheng Chen, Kai Zhang, Hao Tan, Zexiang Xu, Fujun Luan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Gaussian mixture flow matching models, 2025. URL https://arxiv.org/abs/2504.05304.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023.

Yue-Jiang Dong, Hubery Yin, Fangyikang Wang, Yizhe Zhao, Chao Zhang, Chen Li, and Song-Hai Zhang. DC-DPM: A divide-and-conquer approach for diffusion reverse process, 2024. URL https://openreview.net/forum?id=VbAxCwV2e3.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Dongning Guo, Yihong Wu, Shlomo Shamai, and Sergio Verdú. Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Transactions on Information Theory*, 57(4): 2371–2385, 2011. doi: 10.1109/TIT.2011.2111010. URL https://arxiv.org/abs/1004.3332.

Hanzhong Guo, Cheng Lu, Fan Bao, Tianyu Pang, Shuicheng Yan, Chao Du, and Chongxuan Li. Gaussian mixture solvers for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/51373b6499708b6fcc38f1e8f8f5b376-Abstract-Conference.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Nishant Jain and Tong Zhang. A sharp KL-convergence analysis for diffusion models under minimal assumptions, 2025. URL https://arxiv.org/abs/2508.16306v1. Version 1, August 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Gen Li and Yuling Yan. O(d/t) convergence theory for diffusion probabilistic models under minimal assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=4EjdYiNRzE.

Gen Li and Yuling Yan. O(d/T) convergence theory for diffusion probabilistic models under minimal assumptions. *Journal of Machine Learning Research*, 26(292):1–55, 2025b. URL https://jmlr.org/papers/v26/25-0272.html.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation, 2026. URL https://arxiv.org/abs/2501.12982v3. Version 3, revised July 2026; first posted January 2025.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

David McAllister, Matthew Tancik, Jiaming Song, and Angjoo Kanazawa. Decentralized diffusion models. *arXiv preprint arXiv:2501.05450*, 2025.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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*, 2020b.

Chencheng Tang, Xuanyu Xue, Fangyikang Wang, Chao Zhang, and Hubery Yin. Wasserstein convergence of ODE-based samplers in decentralized diffusion model via velocity field decomposition, 2026. URL https://arxiv.org/abs/2606.15835.

Appendix

A Notation and reading guide

Table 2 collects the main symbols. We use the denoiser notation $\bar{y}, y_\theta$ and the router notation a, a_ϕ throughout. The clean data law p_d is distinguished from the positive-noise endpoint p_0. The kernels K_i, Q_i, and P_i distinguish exact, oracle, and learned reverse transitions. After the probabilistic foundations, Appendix C gives the Gaussian bridge reduction and a complementary local estimate. Appendix D proves the main linear-dimensional accumulated rate; Appendix E supplies routing stability and the global comparison. Endpoint limits, sampler distinctions, training identities, and applicability examples follow. Appendix L studies the partition dependence of discretization; Appendix M compares prior results. Symbols used only inside a proof are defined locally.

B Probabilistic foundations

B.1 Forward construction and canonical posteriors

We first specify versions of the conditional distributions used throughout the proofs. Abbreviate X_{t_i} by X_i in this subsection, and write $\varphi_v(z) = (2\pi v)^{-d/2}\exp\big(-\|z\|^2/(2v)\big)$ for the density of $\mathcal{N}(0, vI_d)$, where $v > 0$.

Lemma B.1 (Canonical forward and reverse laws)**.** *Under Assumption 2.1, the independent-innovation forward construction has, conditional on $Y = y$, a jointly Gaussian law with*

$$\begin{aligned} \mathbb{E}[X_i \mid Y = y] &= \alpha_i y, \\ \mathrm{Cov}(X_j, X_i \mid Y = y) &= \frac{\alpha_i}{\alpha_j}\sigma_j^2 I_d, \qquad 0 \le j \le i \le N. \end{aligned} \tag{36}$$

For every i and every $x \in \mathbb{R}^d$, define

$$\begin{aligned} p_i(x) &= \int \varphi_{\sigma_i^2}(x - \alpha_i y)\, p_d(dy), \\ \Pi_i(x, dy) &= \frac{\varphi_{\sigma_i^2}(x - \alpha_i y)}{p_i(x)}\, p_d(dy), \\ \bar{y}_i(x) &= \int y\, \Pi_i(x, dy). \end{aligned} \tag{37}$$

Then p_i is a strictly positive smooth density, Π_i is a posterior probability kernel for Y given X_i, and $\bar{y}_i$ is a finite smooth version of $\bar{y}(x, t_i)$. All posterior polynomial moments are finite and smooth at every x, and

$$\nabla \log p_i(x) = \frac{\alpha_i \bar{y}_i(x) - x}{\sigma_i^2}. \tag{38}$$

For every class with $\pi_l > 0$, the same definitions with p_d replaced by p_d^l give p_i^l, Π_i^l, and $\bar{y}_i^l = \bar{y}^l(\cdot, t_i)$. The canonical class probabilities are

$$\begin{aligned} a^l(x, t_i) &= \frac{\pi_l p_i^l(x)}{p_i(x)}, \qquad p_i(x)a^l(x, t_i) = \pi_l p_i^l(x), \\ \Pi_i(x, \cdot) &= \sum_l a^l(x, t_i)\Pi_i^l(x, \cdot). \end{aligned} \tag{39}$$

These identities hold for every x.

For $i \ge 1$, a canonical exact reverse kernel is

$$K_i(x, dz) = \left\{ \int \varphi_{v_i}(z - A_i x - B_i y)\, \Pi_i(x, dy) \right\} dz. \tag{40}$$

It satisfies $p_i K_i = p_{i-1}$. Similarly, replacing Π_i by Π_i^l defines the class reverse kernel K_i^l, and $p_i^l K_i^l = p_{i-1}^l$ and $K_i = \sum_l a^l K_i^l$.

Table 2: Main symbols and their meanings. The class laws p_d^l and p_t^l are normalized probability laws; their prior weights π_l are written explicitly. All learning-error expectations are under the true noisy data marginals.

Symbol	Meaning
Data and forward noising	
d, N, L	Data dimension, number of reverse steps, and number of classes.
$p_d,\ M$	Clean data law and its second moment $M = \mathbb{E}\|Y\|^2$.
$J,\ \pi_l,\ p_d^l$	Class label, class prior, and conditional data law: $p_d = \sum_l \pi_l p_d^l$.
$\gamma^d,\ \nu$	Standard Gaussian law and sampler initialization; here $\nu = \gamma^d$.
$\mathcal{D},\ t_{\min},\ t_{\max}$	Grid $\mathcal{D} = \{0 < t_{\min} = t_0 < \cdots < t_N = t_{\max} < 1\}$.
$\alpha_t,\ \sigma_t$	Signal and noise coefficients: $X_t \overset{\text{law}}{=} \alpha_t Y + \sigma_t Z$, $\alpha_t^2 + \sigma_t^2 = 1$.
$p_i,\ p_t^l$	True law of X_{t_i} and conditional law of X_t given $J = l$.
p_0	Early-stopped noisy law, generally different from p_d.
Denoisers, routing, and reverse kernels	
$\bar{y}(x,t),\ \bar{y}^l(x,t)$	Posterior means of Y given $X_t = x$, and given $(X_t = x, J = l)$.
$y_\theta,\ y_\theta^l$	Learned standard denoiser and learned class-l denoiser.
$a^l(x,t),\ a_\phi^l(x,t)$	True posterior class probability and learned routing probability. The vectors a, a_ϕ lie in Δ_L.
A_i, B_i, v_i	Exact bridge coefficients: $X_{t_{i-1}} \mid (X_{t_i} = x, Y = y) \sim \mathcal{N}(A_i x + B_i y, v_i I_d)$.
K_i	Exact reverse kernel, with density $p(x_{t_{i-1}} \mid x_{t_i})$; $p_i K_i = p_{i-1}$.
$Q_i^{\text{std}},\ P_i^{\text{std}}$	Standard Gaussian kernels with oracle and learned denoisers; their densities are $\widetilde{p}$ and p_θ.
$Q_i,\ P_i$	Oracle and learned mixture kernels, with densities $\widehat{p}$ and $p_{\theta,\phi}$.
$q_0^{\text{std}},\ q_0$	Output laws of the learned standard and mixture samplers.
Noise increments and error budgets	
$H,\ h,\ \rho$	Log-SNR span $H = \log(\eta_0/\eta_N)$, spacing $h = H/N$, and common relative increment $\rho = e^h - 1 \le 1$ on the main grid.
$\varepsilon_{\text{net,std}},\ \varepsilon_{\text{net,dec}}$	Weighted network error energies: $\varepsilon^2_{\text{net,std}} = \sum_i \Delta_i e_i^2$, $\varepsilon^2_{\text{net,dec}} = \sum_i \Delta_i E_i^2$.
$\eta_i,\ \rho_i$	SNR $\eta_i = \alpha_i^2/\sigma_i^2$ and relative increment $\rho_i = \eta_{i-1}/\eta_i - 1$.
$\Delta_i,\ \Gamma$	Absolute SNR increment $\Delta_i = \eta_{i-1} - \eta_i = B_i^2/v_i$ and sum $\Gamma = \eta_0 - \eta_N$.
$e_i,\ E_i,\ W_i$	Root mean-square network errors (standard and class-weighted), and ℓ^1 routing error; see (9).
$D_i(x),\ D,\ D_{\max,N}$	Pointwise diameter, fixed bound, and grid-dependent bound for the learned expert predictions.
$\Psi_i,\ \Omega_i,\ T_i$	Logarithmic routing energy, its exponential upper bound, and quadratic coupling moment; see (10), (11), and (34).
κ_i	Smaller conditional KL upper bound at step i; see (35).
$\mathcal{I}_N,\ \mathcal{D}_N$	Initialization error and accumulated discretization budget $\mathcal{D}_N = d(9H + 5)(e^{H/N} - 1)/4$; see (16).
$\text{TV},\ W_2,\ d_{\text{BL}}$	Total variation, quadratic Wasserstein, and bounded-Lipschitz distances.

Proof. We derive the posterior and reverse laws from the independent forward innovations.

A common probability space. In the class model, take the joint data law $\mathbb{P}(Y \in dy, J = l) = \pi_l p_d^l(dy)$ and its product with $N+1$ independent standard Gaussian laws. Without labels, use the product with p_d directly. Let $Z_0, \dots, Z_N$ be the Gaussian coordinate variables, independent of (Y, J), and set

$$\begin{aligned} X_0 &= \alpha_0 Y + \sigma_0 Z_0, \\ X_i &= c_i X_{i-1} + \sqrt{r_i} Z_i, \qquad c_i = \frac{\alpha_i}{\alpha_{i-1}}. \end{aligned} \tag{41}$$

The centered conditional variables $\zeta_i = X_i - \alpha_i Y$ satisfy $\zeta_0 = \sigma_0 Z_0$ and $\zeta_i = c_i \zeta_{i-1} + \sqrt{r_i} Z_i$. They form a jointly Gaussian vector independent of (Y, J). The recursion and $c_i^2 \sigma_{i-1}^2 + r_i = \sigma_i^2$ give $\mathrm{Var}(\zeta_i) = \sigma_i^2 I_d$; iterating the recursion gives (36). In particular, each single-time representation $X_i = \alpha_i Y + \sigma_i \widehat{Z}_i$ uses a standard Gaussian $\widehat{Z}_i = \zeta_i / \sigma_i$ independent of (Y, J), with the dependence between different $\widehat{Z}_i$ determined by this recursion.

Posterior versions and smoothness. The conditional density of X_i given $Y = y$ is $\varphi_{\sigma_i^2}(x - \alpha_i y)$, which proves the first identity in (37). Its integrand is positive and bounded, so $0 < p_i(x) < \infty$ for every x. For Borel sets $D, C \subseteq \mathbb{R}^d$, Tonelli's theorem gives

$$\begin{aligned} \int_D \Pi_i(x, C) p_i(x)\, dx &= \int_C \int_D \varphi_{\sigma_i^2}(x - \alpha_i y)\, dx\, p_d(dy) \\ &= \mathbb{P}(X_i \in D, Y \in C). \end{aligned}$$

The ratio in (37) is Borel in (x, y), so this defines a posterior probability kernel on all of $\mathbb{R}^d$.

For each compact set of x, every polynomial in y multiplied by any x-derivative of $\varphi_{\sigma_i^2}(x - \alpha_i y)$ is bounded uniformly in y. This follows from $\alpha_i, \sigma_i > 0$ and Gaussian decay. Differentiation under the prior integral is therefore justified by dominated convergence, without higher prior moments. The positive smooth denominator is bounded away from zero on each compact set. Consequently all posterior polynomial moments, including $\bar{y}_i$, are finite and smooth everywhere. Since $\mathbb{E}\|Y\| < \infty$, this posterior integral is a version of the usual conditional expectation. Differentiating p_i once gives (38).

Each class has finite second moment because $\pi_l \mathbb{E}[\|Y\|^2 \mid J = l] \le M$ and $\pi_l > 0$. The preceding argument applies to p_d^l. Applying Bayes' formula to the joint law of (J, Y, X_i) proves (39), with strictly positive $p_i^l(x)$ and denominator $p_i(x)$ for every x. Integrating y also gives $\bar{y}_i(x) = \sum_l a^l(x, t_i) \bar{y}_i^l(x)$.

Gaussian conditioning and the reverse kernel. For adjacent times, the conditional cross covariance is $c_i \sigma_{i-1}^2 I_d$. The Gaussian conditioning formula and (36) give

$$\begin{aligned} X_{i-1} \mid (X_i = x, Y = y) &\sim \mathcal{N}(A_i x + B_i y, v_i I_d), \\ A_i &= c_i \sigma_{i-1}^2 / \sigma_i^2, \\ B_i &= \alpha_{i-1} - A_i \alpha_i = \alpha_{i-1} r_i / \sigma_i^2, \\ v_i &= \sigma_{i-1}^2 - A_i^2 \sigma_i^2 = r_i \sigma_{i-1}^2 / \sigma_i^2 > 0. \end{aligned} \tag{42}$$

More explicitly, $X_{i-1} - A_i X_i - B_i Y = \zeta_{i-1} - A_i \zeta_i$ is Gaussian with covariance $v_i I_d$ and is uncorrelated with ζ_i. It is therefore independent of ζ_i, and the entire noise vector is independent of (Y, J). The residual is thus independent of (Y, J, X_i). This justifies the bridge representation with a fresh Gaussian independent of (Y, X_i).

Conditioning first on (Y, X_i) and then on X_i yields (40). In particular, for every Borel set C,

$$\int p_i(x) K_i(x, C)\, dx = \mathbb{E}[\mathbb{P}(X_{i-1} \in C \mid X_i)] = p_{i-1}(C).$$

The identical calculation conditional on $J = l$ gives $p_i^l K_i^l = p_{i-1}^l$; the posterior decomposition in (39) gives $K_i = \sum_l a^l K_i^l$. □

Measurable kernels and null-set versions. The density in (40) is jointly Borel in (x, z), strictly positive, and at most $(2\pi v_i)^{-d/2}$; its integral in z is one by Tonelli's theorem. The same properties hold for the oracle and learned Gaussian kernels in (5) and (7): their means are finite Borel functions and their weights are Borel probability vectors. Thus they define Markov kernels on every input state. Starting from $\nu = \mathcal{N}(0, I_d)$, every input and output marginal of these finite chains is absolutely continuous with respect to Lebesgue measure. Changing any predictor or posterior version on a Borel Lebesgue-null set, while retaining finite Borel means and probability-vector weights, leaves all output laws unchanged: at each step the current input assigns that set zero probability, and induction gives the same next marginal. Since $p_i(x) > 0$ everywhere, a p_i-null set is also Lebesgue-null. The canonical versions above additionally make all displayed posterior identities and derivatives valid at every x.

C A COMPLEMENTARY LOCAL MARGINAL ERROR BOUND

The main $O(d \log^2 N/N)$ rate for $H = O(\log N)$ follows from the accumulated estimate in Appendix D. This appendix gives a complementary estimate for each individual interval, including intervals of an irregular grid.

Lemma C.1 (Local marginal error)**.** *Under Assumption 2.1, the oracle kernel (5) satisfies, for every step i with $\rho_i \le 1/(d+6)$,*

$$\mathrm{TV}(p_{i-1}, p_i Q_i^{\mathrm{std}}) \le 261 d(d+2)\rho_i^2. \tag{43}$$

The constant is independent of the data second moment and the noise endpoints.

The argument keeps the input distributed according to the true forward marginal, so that conditional mean matching cancels the first-order error. We first remove the independent Gaussian noise shared by the exact and approximate reverse transitions. We then use two integrations by parts in the remaining Gaussian observation model. This avoids inverse powers of the vanishing reverse-step variance.

C.1 REDUCTION TO A GAUSSIAN OBSERVATION MODEL

Fix adjacent times $s < t$ with $0 < \alpha_t < \alpha_s < 1$, and let $Y \sim p_d$ have finite second moment. Write

$$r_{st} = 1 - \frac{\alpha_t^2}{\alpha_s^2}, \qquad A_{st} = \frac{\alpha_t \sigma_s^2}{\alpha_s \sigma_t^2}, \qquad B_{st} = \frac{\alpha_s r_{st}}{\sigma_t^2}, \qquad v_{st} = \frac{r_{st}\sigma_s^2}{\sigma_t^2}.$$

Here v_{st} is the covariance scalar, rather than its square root. Conditional on Y, the Gaussian bridge from X_t to X_s has the representation

$$X_s = A_{st} X_t + B_{st} Y + \sqrt{v_{st}}\, Z', \qquad Z' \sim \mathcal{N}(0, I_d), \tag{44}$$

where Z' can be taken independent of (X_t, Y). The coefficients follow by conditioning the jointly Gaussian forward process given Y. Its marginal can also be verified directly using

$$A_{st}\alpha_t + B_{st} = \alpha_s, \qquad A_{st}^2 \sigma_t^2 + v_{st} = \sigma_s^2. \tag{45}$$

Let $\bar{y}_t(x) = \mathbb{E}[Y \mid X_t = x]$. Applying the oracle DDPM kernel $\widetilde{K}_{s|t}(x, \cdot) = \mathcal{N}(A_{st}x + B_{st}\bar{y}_t(x), v_{st} I_d)$ to the true marginal therefore gives

$$\begin{aligned} p_s &= \mathcal{L}(A_{st} X_t + B_{st} Y + \sqrt{v_{st}}\, Z'), \\ p_t \widetilde{K}_{s|t} &= \mathcal{L}(A_{st} X_t + B_{st} \bar{y}_t(X_t) + \sqrt{v_{st}}\, Z'). \end{aligned} \tag{46}$$

We use $p_t \widetilde{K}_{s|t}$ to denote the output marginal after integrating against p_t.

Adding a common independent Gaussian contracts total variation. Total variation is also invariant under a common invertible rescaling. Thus, upon setting

$$U = \frac{\alpha_t}{\sigma_t} Y, \qquad X = \frac{X_t}{\sigma_t} = U + Z, \qquad m(x) = \mathbb{E}[U \mid X = x], \tag{47}$$

with $Z \sim \mathcal{N}(0, I_d)$ independent of U, we obtain the normalized posterior $m(x) = (\alpha_t/\sigma_t)\bar{y}(\sigma_t x, t)$ and

$$\mathrm{TV}(p_s, p_t \widetilde{K}_{s|t}) \le \mathrm{TV}\big(\mathcal{L}(X + \varepsilon U), \mathcal{L}(X + \varepsilon m(X))\big), \tag{48}$$

where

$$\varepsilon = \frac{B_{st}}{A_{st}\alpha_t} = \frac{\alpha_s^2 - \alpha_t^2}{\alpha_t^2 \sigma_s^2} = \frac{\eta_s}{\eta_t} - 1, \qquad \eta_t = \frac{\alpha_t^2}{\sigma_t^2}. \tag{49}$$

The reduction preserves the Gaussian observation noise in X, whose covariance is the identity. The following lemma controls the right-hand side.

C.2 A POSTERIOR-MEAN TRANSPORT ESTIMATE

Lemma C.2 (Posterior-mean transport). *Let U be an $\mathbb{R}^d$-valued random variable with finite second moment, let $Z \sim \mathcal{N}(0, I_d)$ be independent of U, and define $X = U + Z$ and $m(x) = \mathbb{E}[U \mid X = x]$. For every $\varepsilon \geq 0$,*

$$\begin{aligned} &\mathrm{TV}\big(\mathcal{L}(X + \varepsilon U), \mathcal{L}(X + \varepsilon m(X))\big) \\ &\quad \leq \varepsilon^2 \left\{ \frac{d + 124d(d+2)}{4} + \frac{208}{3}\varepsilon d(d+2)(d+4) + 160\varepsilon^2 d(d+2)(d+4)(d+6) \right\}. \end{aligned} \tag{50}$$

In particular, if $\varepsilon \leq 1/(d+6)$, then

$$\mathrm{TV}\big(\mathcal{L}(X + \varepsilon U), \mathcal{L}(X + \varepsilon m(X))\big) \leq 261 d(d+2)\varepsilon^2. \tag{51}$$

Proof. We use the convention $\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|_1$. All conditional expectations in this proof refer to the joint law of (U, X). The case $\varepsilon = 0$ is immediate, so assume $\varepsilon > 0$. For a vector field b we use $(Db)_{ij} = \partial_j b_i$; for a matrix field M, set $(\operatorname{div} M)_j = \sum_i \partial_i M_{ij}$, so that $\operatorname{div}(Mb) = (\operatorname{div} M)^T b + \operatorname{tr}(MDb)$ when M is symmetric. Directional derivatives are denoted by $D[\cdot][w]$.

Posterior regularity and derivative identities. Let μ denote the law of U. The posterior measure at $X = x$ is

$$\mu_x(du) = \frac{\exp\big(-\|x - u\|^2/2\big)\, \mu(du)}{\int \exp(-\|x - z\|^2/2)\, \mu(dz)}. \tag{52}$$

Its denominator is strictly positive. For any compact set K, polynomial h, and multi-index β, the derivative $\partial_x^\beta \{h(u) e^{-\|x-u\|^2/2}\}$ is a polynomial in (x, u) times the same Gaussian factor, and is bounded uniformly over $x \in K$ and $u \in \mathbb{R}^d$. These bounds are integrable against the probability measure μ. Dominated convergence therefore makes the numerator and denominator smooth; continuity and positivity bound the denominator away from zero on K. Thus all posterior moments are finite at each x and posterior polynomial moments are locally differentiable to every order. We only use finiteness, not differentiability, of the absolute centered moments r_k defined below. No higher moments of μ are needed. Moreover,

$$m(x) = \nabla \psi(x), \qquad \psi(x) = \log \int \exp\big(x^T u - \|u\|^2/2\big)\, \mu(du).$$

Define the residual and conditional moments by

$$\xi = U - m(X), \quad R = \|\xi\|, \quad G = \|Z\|, \qquad r_k(x) = \mathbb{E}[\|U - m(x)\|^k \mid X = x]. \tag{53}$$

For a polynomial h independent of x, differentiating (52) gives $\partial_j \mathbb{E}[h(U) \mid X = x] = \mathbb{E}[h(U)(U_j - m_j(x)) \mid X = x]$. For an x-dependent polynomial $h(x, u)$ with smooth coefficients, the product rule adds $\mathbb{E}[\partial_j h(x, U) \mid X = x]$ to this identity. In particular, $\partial_j(U_i - m_i(x)) = -H_{ij}$: differentiating the second centered moment gives no centering contribution because $\mathbb{E}[\xi \mid X = x] = 0$, whereas differentiating the third gives three covariance products. With $T_{ijk} = \partial_k H_{ij}$ and $K_{ijkl} = \partial_l T_{ijk}$, this yields

$$\begin{aligned} H := Dm &= \mathbb{E}[\xi\xi^T \mid X = x] \succeq 0, \\ T := DH &= \mathbb{E}[\xi^{\otimes 3} \mid X = x], \\ K_{ijkl} := (DT)_{ijkl} &= \mathbb{E}[\xi_i\xi_j\xi_k\xi_l \mid X = x] - H_{ij}H_{kl} - H_{ik}H_{jl} - H_{il}H_{jk}. \end{aligned} \tag{54}$$

All these tensors are symmetric. For a vector w, $T[w]$ denotes the symmetric matrix with entries $\sum_k T_{ijk} w_k$. In particular,

$$\|H\|_{\mathrm{op}} \le \operatorname{tr} H = r_2, \qquad \|T[w]\|_{\mathrm{op}} \le r_3 \|w\|, \qquad r_j \le r_k^{j/k} \quad (1 \le j \le k). \tag{55}$$

The integrated residual moments are controlled by the observation noise. Indeed, $\xi = \mathbb{E}[Z \mid X] - Z$, so conditional Jensen gives

$$\mathbb{E}R^k \le 2^k \mathbb{E}G^k \quad (k \ge 1), \qquad \mathbb{E}R^2 = d - \mathbb{E}\|\mathbb{E}[Z \mid X]\|^2 \le d. \tag{56}$$

We will use the Gaussian moments

$$\mu_4 := \mathbb{E}G^4 = d(d+2), \quad \mu_6 := \mathbb{E}G^6 = d(d+2)(d+4), \quad \mu_8 := \mathbb{E}G^8 = d(d+2)(d+4)(d+6). \tag{57}$$

These identities follow by differentiating $\mathbb{E}\exp\bigl(t\|Z\|^2\bigr) = (1-2t)^{-d/2}$ at zero.

Interpolation and cancellation. For $\lambda \in [0,1]$, set

$$F_\lambda(x,u) = x + \varepsilon\{(1-\lambda)m(x) + \lambda u\}, \qquad \beta_\lambda = \varepsilon(1-\lambda), \tag{58}$$

and define

$$\mathsf{A} = (I + \beta_\lambda H)^{-1}, \qquad \mathsf{C} = \mathsf{A}H, \qquad \boldsymbol{v} = \mathsf{A}\xi. \tag{59}$$

Here β_λ is an interpolation parameter local to this proof. In particular, $\|\mathsf{A}\|_{\mathrm{op}} \le 1$. The matrices H, A, and C commute and are positive semidefinite, since A and C are functions of H. No commutation with $T[w]$ is assumed. For each fixed u, the derivative of F_λ with respect to x is $I + \beta_\lambda H \succeq I$, and $F_\lambda(\cdot, u)$ is a global smooth diffeomorphism. To see surjectivity explicitly, the function $\|x\|^2/2 + \beta_\lambda \psi(x) - z^T x$ is strongly convex and coercive for every z: convexity of ψ bounds it below by its affine tangent at zero. Its unique minimizer solves $x + \beta_\lambda m(x) = z$. Injectivity follows from strong monotonicity, and the nonsingular Jacobian gives a smooth inverse. Adding $\varepsilon\lambda u$ does not change these conclusions. Strong monotonicity also gives $\|F_\lambda(x,u)\| \ge \|x\| - \|\beta_\lambda m(0) + \varepsilon\lambda u\|$. Hence, for each fixed u, compact supports pull back to compact sets, uniformly over $\lambda \in [0,1]$.

Take $f \in C_c^\infty(\mathbb{R}^d)$. Taylor's formula in λ and $\mathbb{E}[\xi \mid X] = 0$ imply

$$\begin{aligned}
\mathbb{E}f(F_1(X,U)) - \mathbb{E}f(F_0(X,U)) &= \varepsilon \mathbb{E}[\nabla f(F_0(X,U))^T \xi] \\
&\quad + \varepsilon^2 \int_0^1 (1-\lambda) \mathbb{E}[\xi^T \nabla^2 f(F_\lambda(X,U)) \xi]\, d\lambda \\
&= \varepsilon^2 \int_0^1 (1-\lambda) \mathbb{E}[\xi^T \nabla^2 f(F_\lambda(X,U)) \xi]\, d\lambda.
\end{aligned} \tag{60}$$

The first-order term vanishes because $F_0(X,U)$ depends only on X. Taylor's formula is first applied pointwise. The bounds $|\partial_\lambda f(F_\lambda)| \le \varepsilon \|\nabla f\|_\infty R$ and $|\partial_\lambda^2 f(F_\lambda)| \le \varepsilon^2 \|\nabla^2 f\|_\infty R^2$, together with $\mathbb{E}R^2 \le d$, justify taking expectations and integrating in λ.

Two integrations by parts. In the following calculation, u and λ are fixed and all derivatives are with respect to x. Write

$$p(x \mid u) = (2\pi)^{-d/2} e^{-\|x-u\|^2/2}, \qquad g = \nabla_x \log p(x \mid u) = u - x = -Z, \qquad Dg = -I,$$

and let $q = \operatorname{div} \boldsymbol{v} + g^T \boldsymbol{v}$. We claim that

$$\mathbb{E}[\xi^T \nabla^2 f(F_\lambda) \xi] = \mathbb{E}[f(F_\lambda) W_\lambda], \qquad W_\lambda = q^2 + \boldsymbol{v}^T \nabla q - \operatorname{div}(\mathsf{C}\boldsymbol{v}) - (\mathsf{C}\boldsymbol{v})^T g. \tag{61}$$

Here and below F_λ without arguments means $F_\lambda(X,U)$. To verify the claim, use summed coordinate indices and $D_x F_\lambda = \mathsf{A}^{-1}$ to obtain

$$\begin{aligned}
\int p\, \xi_i \xi_j f_{ij}(F_\lambda)\, dx &= \int p\, \xi_j v_k \partial_k [f_j(F_\lambda)]\, dx \\
&= -\int f_j(F_\lambda) \partial_k (p v_k \xi_j)\, dx.
\end{aligned}$$

Since $D_x\xi = -H$, $\partial_k(pv_k\xi_j) = \xi_j \operatorname{div}(p\boldsymbol{v}) - p(H\boldsymbol{v})_j$. The vector coefficient after this first integration by parts is $\xi q - H\boldsymbol{v}$. Using (64), established below, it obeys

$$\|\xi q - H\boldsymbol{v}\| \le 2r_2 R + \beta_\lambda r_3 R^2 + GR^2. \tag{62}$$

This coefficient is integrable under the joint law. Conditional Lyapunov and Hölder inequalities give

$$\mathbb{E}[r_2 R] \le \mathbb{E}R^3, \qquad \mathbb{E}[r_3 R^2] \le \mathbb{E}R^5,$$
$$\mathbb{E}[GR^2] \le (\mathbb{E}G^3)^{1/3}(\mathbb{E}R^3)^{2/3}.$$

All are finite by (56). Since ∇f is bounded, averaging the first integration-by-parts identity over U is therefore justified. Converting the remaining gradient of f to an x-gradient and integrating once more gives a second minus sign. Explicitly, since $f_j(F_\lambda) = \mathsf{A}_{kj}\partial_k[f(F_\lambda)]$ and $\mathsf{A}(\xi q - H\boldsymbol{v}) = \boldsymbol{v}q - \mathsf{C}\boldsymbol{v}$, the first integral after integration by parts is $-\int p(\boldsymbol{v}q - \mathsf{C}\boldsymbol{v})^T \nabla_x[f(F_\lambda)]\,dx$. The second is therefore

$$\int f(F_\lambda) \operatorname{div}\{\boldsymbol{v} \operatorname{div}(p\boldsymbol{v}) - p\mathsf{A}H\boldsymbol{v}\}\,dx$$
$$= \int p f(F_\lambda)\{q^2 + \boldsymbol{v}^T\nabla q - \operatorname{div}(\mathsf{C}\boldsymbol{v}) - (\mathsf{C}\boldsymbol{v})^T g\}\,dx.$$

There are no boundary terms: for each fixed u, the inverse image under $F_\lambda(\cdot, u)$ of the compact support of f is compact. The original Hessian integral is absolutely integrable by $\mathbb{E}R^2 \le d$, and (62) handles the intermediate expression. Finally, (72) below establishes integrability of $f(F_\lambda)W_\lambda$, uniformly for $\lambda \in [0,1]$ at fixed ε. Thus averaging the second identity over U and the subsequent integration in λ are both justified.

Bounds for the differential weight. The derivative identities

$$D\mathsf{A}[w] = -\beta_\lambda \mathsf{A}T[w]\mathsf{A}, \qquad D\mathsf{C}[w] = \mathsf{A}T[w]\mathsf{A}, \qquad D\boldsymbol{v}[w] = -\mathsf{C}w - \beta_\lambda \mathsf{A}T[w]\boldsymbol{v} \tag{63}$$

follow by differentiating (59). In the second identity the matrix order is retained: $D\mathsf{C}[w] = \mathsf{A}T[w] - \beta_\lambda \mathsf{A}T[w]\mathsf{A}H = \mathsf{A}T[w]\mathsf{A}$, using $I - \beta_\lambda \mathsf{A}H = \mathsf{A}$. Define the contraction vector

$$\tau_i = \sum_{jk} T_{ijk}\mathsf{A}_{jk} = \mathbb{E}[\xi_i(\xi^T\mathsf{A}\xi) \mid X = x], \qquad \|\tau\| \le r_3.$$

Full symmetry of T gives $T[w]\boldsymbol{v} = T[\boldsymbol{v}]w$ and hence $D\boldsymbol{v} = -\mathsf{C} - \beta_\lambda \mathsf{A}T[\boldsymbol{v}]$. Taking its trace uses $\operatorname{tr}(\mathsf{A}T[\boldsymbol{v}]) = \tau^T\boldsymbol{v}$. Then

$$\operatorname{div}\boldsymbol{v} = -\operatorname{tr}\mathsf{C} - \beta_\lambda \tau^T\boldsymbol{v},$$
$$|\operatorname{div}\boldsymbol{v}| \le r_2 + \beta_\lambda r_3 R, \qquad \|D\boldsymbol{v}\|_{\mathrm{op}} \le r_2 + \beta_\lambda r_3 R. \tag{64}$$

The gradient of $\operatorname{tr}\mathsf{C}$ is $t_{\mathsf{A}^2} := \mathbb{E}[\xi(\xi^T\mathsf{A}^2\xi) \mid X = x]$, the contraction of T with A^2, so its norm is at most r_3. We also have

$$\|D\tau\|_{\mathrm{op}} \le 4r_4 + \beta_\lambda r_3^2. \tag{65}$$

Indeed, the two terms in $(D\tau[w])_i = \sum_{jkl} K_{ijkl}\mathsf{A}_{jk}w_l + \sum_{jk} T_{ijk}(D\mathsf{A}[w])_{jk}$ expand to

$$D\tau[w] = \mathbb{E}[\xi(\xi^T w)(\xi^T\mathsf{A}\xi) \mid X = x] - 2H\mathsf{A}Hw - \operatorname{tr}(\mathsf{A}H)Hw$$
$$+ \mathbb{E}[\xi\{\xi^T D\mathsf{A}[w]\xi\} \mid X = x]. \tag{66}$$

The first term is bounded by $r_4\|w\|$, and each of the three covariance-product terms by $r_2^2\|w\|$ because $\|\mathsf{A}\|_{\mathrm{op}} \le 1$ and $\operatorname{tr}(\mathsf{A}H) \le r_2$. The last term is bounded by $r_3\|D\mathsf{A}[w]\|_{\mathrm{op}} \le \beta_\lambda r_3^2\|w\|$. Finally, $r_2^2 \le r_4$ proves (65). Differentiating the first identity in (64) gives

$$\nabla \operatorname{div}\boldsymbol{v} = -t_{\mathsf{A}^2} - \beta_\lambda\{(D\tau)^T\boldsymbol{v} + (D\boldsymbol{v})^T\tau\}.$$

Its norm is at most $r_3 + \beta_\lambda(4r_4 + \beta_\lambda r_3^2)R + \beta_\lambda(r_2 + \beta_\lambda r_3 R)r_3$, namely

$$\|\nabla \operatorname{div}\boldsymbol{v}\| \le r_3 + \beta_\lambda r_2 r_3 + 4\beta_\lambda r_4 R + 2\beta_\lambda^2 r_3^2 R. \tag{67}$$

Expanding the matrix divergence in coordinates gives

$$(\operatorname{div}\mathsf{C})_j = \sum_{iab} \mathsf{A}_{ia}T_{abi}\mathsf{A}_{bj} = \sum_b \tau_b \mathsf{A}_{bj} = (\mathsf{A}\tau)_j.$$

Consequently,

$$\operatorname{div}(\mathsf{C}\boldsymbol{v}) = \boldsymbol{v}^T \mathsf{A}\tau - \operatorname{tr}(\mathsf{C}^2) - \beta_\lambda \operatorname{tr}(\mathsf{C}\mathsf{A}T[\boldsymbol{v}]). \tag{68}$$

Using $\operatorname{tr}(\mathsf{C}^2) \le (\operatorname{tr}\mathsf{C})^2 \le r_2^2$ gives

$$|\operatorname{div}(\mathsf{C}\boldsymbol{v})| \le r_3 R + r_2^2 + \beta_\lambda r_2 r_3 R. \tag{69}$$

For the last term, the matrix CA is positive semidefinite and $\operatorname{tr}(\mathsf{CA}) \le \operatorname{tr} H = r_2$; hence $|\operatorname{tr}(\mathsf{C}\mathsf{A}T[\boldsymbol{v}])| \le r_2 r_3 R$. These contraction estimates introduce no additional dimension factor.

Finally, $|q| \le r_2 + \beta_\lambda r_3 R + GR$, and $Dg = -I$ gives

$$\boldsymbol{v}^T \nabla q = \boldsymbol{v}^T \nabla \operatorname{div} \boldsymbol{v} + (D\boldsymbol{v}\,\boldsymbol{v})^T g - \|\boldsymbol{v}\|^2.$$

Combining the preceding bounds with $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ in (61), we obtain

$$\begin{aligned} |W_\lambda| \le\ & 4r_2^2 + 5\beta_\lambda^2 r_3^2 R^2 + 3G^2R^2 + 2r_3R + 2\beta_\lambda r_2 r_3 R \\ & + 4\beta_\lambda r_4 R^2 + 2r_2 RG + \beta_\lambda r_3 R^2 G + R^2. \end{aligned} \tag{70}$$

Integrability and dimension dependence. Conditional Lyapunov inequalities give

$$\begin{aligned} \mathbb{E}r_2^2 &\le \mathbb{E}R^4, & \mathbb{E}[r_3^2R^2] &= \mathbb{E}[r_3^2 r_2] \le \mathbb{E}R^8, \\ \mathbb{E}[r_3 R] &= \mathbb{E}[r_3 r_1] \le \mathbb{E}R^4, & \mathbb{E}[r_2 r_3 R] &= \mathbb{E}[r_2 r_3 r_1] \le \mathbb{E}R^6, \\ \mathbb{E}[r_4 R^2] &= \mathbb{E}[r_4 r_2] \le \mathbb{E}R^6. \end{aligned}$$

For terms involving Z, Hölder's inequality and (56) imply

$$\begin{aligned} \mathbb{E}[G^2R^2] &\le \mu_4^{1/2}(\mathbb{E}R^4)^{1/2} \le 4\mu_4, \\ \mathbb{E}[r_2 RG] &\le (\mathbb{E}R^4)^{3/4}\mu_4^{1/4} \le 8\mu_4, \\ \mathbb{E}[r_3 R^2 G] &\le (\mathbb{E}R^6)^{5/6}\mu_6^{1/6} \le 32\mu_6. \end{aligned} \tag{71}$$

For example, the second inequality uses Hölder exponents $2, 4, 4$ on r_2, R, G and $\mathbb{E}r_2^2 \le \mathbb{E}R^4$; the third uses exponents $2, 3, 6$ and $\mathbb{E}r_3^2 \le \mathbb{E}R^6$.

Inserting $\mathbb{E}R^4 \le 16\mu_4$, $\mathbb{E}R^6 \le 64\mu_6$, $\mathbb{E}R^8 \le 256\mu_8$, and (71) into (70) gives

$$\mathbb{E}|W_\lambda| \le d + 124\mu_4 + 416\beta_\lambda\mu_6 + 1280\beta_\lambda^2\mu_8. \tag{72}$$

The μ_4 coefficient is $64 + 12 + 32 + 16 = 124$; the $\beta_\lambda\mu_6$ coefficient is $128 + 256 + 32 = 416$; and the $\beta_\lambda^2\mu_8$ coefficient is $5 \cdot 256 = 1280$. Thus all terms in the integrations by parts are absolutely integrable, including after averaging over the prior. More explicitly, replace each β_λ on the right-hand side of (70) by ε and call the resulting random variable Z_ε. Then $|W_\lambda| \le Z_\varepsilon$ for every $\lambda \in [0,1]$, and $\mathbb{E}Z_\varepsilon \le d + 124\mu_4 + 416\varepsilon\mu_6 + 1280\varepsilon^2\mu_8 < \infty$. This supplies one integrable dominating variable for the whole interpolation.

Combining (60), (61), and (72), with $\beta_\lambda = \varepsilon(1-\lambda)$ and $\int_0^1 (1-\lambda)^k\, d\lambda = 1/(k+1)$ for $k = 1, 2, 3$, yields

$$\begin{aligned} &|\mathbb{E}f(F_1) - \mathbb{E}f(F_0)| \\ &\quad \le \|f\|_\infty \varepsilon^2 \left\{ \frac{d + 124\mu_4}{2} + \frac{416}{3}\varepsilon\mu_6 + 320\varepsilon^2\mu_8 \right\}. \end{aligned} \tag{73}$$

The supremum over $f \in C_c^\infty$ with $\|f\|_\infty \le 1$ is the variation norm of the signed difference of the two laws. Here is a direct justification. For a finite signed Borel measure $\nu = \nu^+ - \nu^-$ and $\delta > 0$, inner regularity of its mutually singular Jordan parts gives disjoint compact sets K_+, K_- with $\nu^\pm(\mathbb{R}^d \setminus K_\pm) < \delta$. Choose smooth functions $0 \le \chi_\pm \le 1$ equal to one on $K_\pm$, with disjoint compact supports, and put $f = \chi_+ - \chi_-$. Then $\|f\|_\infty \le 1$ and $\int f\, d\nu \ge |\nu|(\mathbb{R}^d) - 4\delta$; the opposite bound $|\int f\, d\nu| \le \|f\|_\infty |\nu|(\mathbb{R}^d)$ always holds. Letting $\delta \downarrow 0$ proves the characterization, without requiring densities. Apply it to the difference of the two laws and divide (73) by two to obtain (50).

If $\varepsilon \le 1/(d+6)$, then $\varepsilon\mu_6 \le \mu_4$ and $\varepsilon^2\mu_8 \le \mu_4$. Consequently, the expression inside braces in (50) is bounded by $d/4 + (781/3)\mu_4 < 261\mu_4$. This proves (51). □

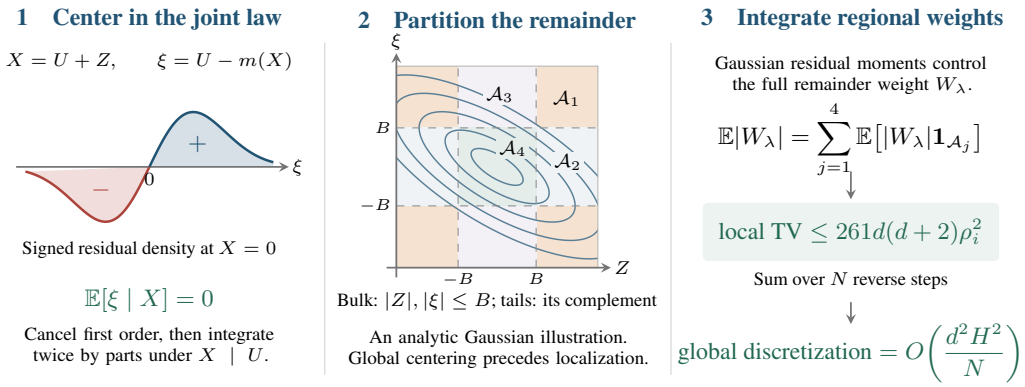


Figure 5: Complementary local view of JGC: full-posterior centering precedes localization of the signed remainder. The Gaussian illustration concerns the local bound of Lemma C.1; the sharper accumulated $O(d \log^2 N/N)$ bound for $H = O(\log N)$ uses the covariance potential in Appendix D.

C.3 A JOINT-SPACE GAUSSIAN PARTITION OF THE SIGNED REMAINDER

The preceding proof admits a precise bulk–tail decomposition in the joint observation space. The order of operations is essential: first cancel the conditional first-order term in (60), then perform the Gaussian integrations by parts, and only then partition the resulting signed remainder. This makes the localization compatible with the unrestricted posterior moments that enter the derivative identities.

Retain the notation of Lemma C.2. For $B > 0$, partition the joint (X, U) space according to the observation noise $Z = X - U$ and the posterior residual $\xi = U - m(X)$:

$$\begin{aligned} \mathcal{A}_1 &= \{\|Z\| > B,\ \|\xi\| > B\}, \quad \mathcal{A}_2 = \{\|Z\| > B,\ \|\xi\| \le B\}, \\ \mathcal{A}_3 &= \{\|Z\| \le B,\ \|\xi\| > B\}, \quad \mathcal{A}_4 = \{\|Z\| \le B,\ \|\xi\| \le B\}. \end{aligned} \tag{74}$$

The last cell is the bulk region. The residual ξ is not generally Gaussian; its tail and moment bounds below are inherited from the Gaussian observation.

Lemma C.3 (Joint-space localization after cancellation)**.** *Let $\mathbb{P}$ denote the joint law of (X, U), and let W_λ be the integrable differential weight in (61). Define finite signed measures on the output space by*

$$\mathsf{R}_j = \varepsilon^2 \int_0^1 (1-\lambda)(F_\lambda)_{\#}\big(\mathbf{1}_{\mathcal{A}_j} W_\lambda \mathbb{P}\big)\, d\lambda, \qquad j = 1, \dots, 4, \tag{75}$$

where $(F_\lambda)_{\#}$ denotes pushforward and the integral is evaluated on Borel sets. Then

$$\mathcal{L}(F_1) - \mathcal{L}(F_0) = \sum_{j=1}^{4} \mathsf{R}_j, \tag{76}$$

and, writing $\|\cdot\|_{\mathrm{var}}$ for the variation norm,

$$\begin{aligned} \mathrm{TV}\big(\mathcal{L}(F_1), \mathcal{L}(F_0)\big) &\le \frac{1}{2} \sum_{j=1}^{4} \|\mathsf{R}_j\|_{\mathrm{var}} \\ &\le \frac{\varepsilon^2}{2} \int_0^1 (1-\lambda) \mathbb{E}|W_\lambda|\, d\lambda. \end{aligned} \tag{77}$$

Thus the same constant $261 d(d+2)$ applies when $\varepsilon \le 1/(d+6)$. Moreover, set $\mu_{2k} = \mathbb{E}\|Z\|^{2k} = \prod_{h=0}^{k-1}(d+2h)$ and

$$\mathsf{L}_{d,\varepsilon} = 4\sqrt{\mu_4} + 124\sqrt{\mu_8} + 416\varepsilon\sqrt{\mu_{12}} + 1280\varepsilon^2\sqrt{\mu_{16}}. \tag{78}$$

The three tail cells satisfy

$$\frac{1}{2} \sum_{j=1}^{3} \|\mathsf{R}_j\|_{\mathrm{var}} \le \frac{\varepsilon^2 \mathsf{L}_{d,\varepsilon}}{4} \min\{1, 2^{1/2+d/4} e^{-B^2/32}\}. \tag{79}$$

Proof. The common integrable envelope following (72) makes the signed measures well-defined. Equations (60) and (61) prove (76) against every smooth compactly supported test function; these tests determine finite signed Borel measures. Variation contracts under pushforward and is subadditive. The cells form a disjoint partition, so summing their absolute-weight integrals gives exactly $\mathbb{E}|W_\lambda|$. This proves (77) and preserves the existing dimension constant.

For completeness, the tail estimate controls the remainder weights as well as the probability of the tail region. Write $\bar{Z} = \mathbb{E}[Z \mid X]$. Since $\xi = \bar{Z} - Z$, conditional Jensen and Cauchy–Schwarz give

$$\mathbb{E}e^{\|\xi\|^2/16} \le \mathbb{E}e^{(\|\bar{Z}\|^2+\|Z\|^2)/8} \le \left(\mathbb{E}e^{\|\bar{Z}\|^2/4}\mathbb{E}e^{\|Z\|^2/4}\right)^{1/2} \le 2^{d/2}.$$

Markov's inequality, the Gaussian exponential moment, and a union bound imply

$$\mathbb{P}(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3) \le \min\{1, 2^{1+d/2}e^{-B^2/16}\}.$$

To control W_λ, apply Hölder and conditional Jensen to the envelope in (70). In detail, for a monomial $G^c R^b \prod_k r_k^{a_k}$ of total degree $m = c + b + \sum_k k a_k > 0$, the residual moment bound gives

$$\Big\|G^c R^b \prod_k r_k^{a_k}\Big\|_{L^2(\mathbb{P})} \le 2^{m-c}\sqrt{\mu_{2m}}.$$

Here $\|r_k\|_{2m/k} \le (\mathbb{E}R^{2m})^{k/(2m)}$ and $\mathbb{E}R^{2m} \le 2^{2m}\mu_{2m}$ justify every factor. Minkowski's inequality applied to the nine terms of the envelope, with $\beta_\lambda \le \varepsilon$, yields $\|W_\lambda\|_2 \le \mathsf{L}_{d,\varepsilon}$: the degree-four coefficients sum to $64 + 12 + 32 + 16 = 124$, the degree-six coefficients to $128 + 256 + 32 = 416$, and the degree-eight coefficient is 1280. Cauchy–Schwarz on the tail event, followed by $\int_0^1 (1-\lambda)\, d\lambda = 1/2$, proves (79). □

For $0 < \varepsilon \le 1/(d+6)$, choosing $B^2 = 16\{(1 + d/2)\log 2 + 2\log(1/\varepsilon)\}$ makes the tail contribution at most $\mathsf{L}_{d,\varepsilon}\varepsilon^3/4$. The Gaussian moment products show $\mathsf{L}_{d,\varepsilon} = O(d^2)$ uniformly under this step restriction. Consequently the tail cells are of higher order, while the complete remainder retains the $O(d^2\varepsilon^2)$ bound without a cutoff-dependent logarithmic loss. The Gaussian scale is $B = O(\sqrt{d + \log(1/\varepsilon)})$.

This localization does not assert a separate first-order cancellation on each cell. Indeed $\mathbb{E}[\xi \mathbf{1}_{\mathcal{A}_j} \mid X]$ need not vanish. Nor is R_j the raw kernel difference restricted to $\mathcal{A}_j$: inserting cell indicators before integration by parts would introduce boundary terms. The second-order gain comes from global posterior centering together with Gaussian integration by parts; the joint-space partition separates the bulk and tail contributions of that already centered remainder.

C.4 Conclusion for the reverse kernel

Apply Lemma C.2 to (48) with the value of ε in (49). Since $U = (\alpha_t/\sigma_t)Y$ has finite second moment whenever Y does, all its assumptions hold. This proves

$$\mathrm{TV}(p_s, p_t\widetilde{K}_{s|t}) \le 261 d(d+2)\left(\frac{\eta_s}{\eta_t} - 1\right)^2 \qquad \text{if } \frac{\eta_s}{\eta_t} - 1 \le \frac{1}{d+6}, \tag{80}$$

and hence Lemma C.1. For a grid $t_0 < \cdots < t_N$, the relative increment for the step $t_i \to t_{i-1}$ is $\rho_i = \eta_{i-1}/\eta_i - 1$; it is distinct from the local interpolation parameter β_λ used above.

The coefficient in (80) is independent of the value of the data second moment. In particular, the same estimate applies to every positive-probability class of a finite data partition, including a discrete data measure. The resulting dimension dependence is $d(d+2)$; no linear-in-d bound is asserted by this proof.

C.5 An explicit separation between conditional and marginal errors

Proposition C.4 (First-order conditional error, second-order marginal error)**.** *Take $d = 1$ and $p_d = \mathcal{N}(0,1)$. Fix $u = \alpha_t^2 \in (0,1)$ and choose $s < t$ with $\eta_s = (1+\rho)\eta_t$. For the exact reverse kernel $K_{s|t}$ and the oracle posterior-variance kernel $Q_{s|t}$, as $\rho \downarrow 0$ with u fixed,*

$$\mathbb{E}_{p_t}\,\mathrm{TV}(K_{s|t}(X,\cdot), Q_{s|t}(X,\cdot)) = \varphi(1)u\rho + o(\rho), \tag{81}$$

$$\mathrm{TV}(p_s, p_t Q_{s|t}) = \varphi(1)u(1-u)\rho^2 + o(\rho^2), \tag{82}$$

where φ is the standard normal density.

Proof. Every forward marginal is standard normal. The VP relation gives

$$\alpha_s^2 = \frac{u(1+\rho)}{1+u\rho}, \qquad c^2 := \frac{\alpha_t^2}{\alpha_s^2} = \frac{1+u\rho}{1+\rho}, \qquad r := 1-c^2 = \frac{(1-u)\rho}{1+\rho}, \qquad v = \frac{r\sigma_s^2}{\sigma_t^2} = \frac{r}{1+u\rho}.$$

Gaussian conditioning yields $K_{s|t}(x) = \mathcal{N}(cx, r)$. Since $\bar{y}(x,t) = \alpha_t x$ and $A_{st} + B_{st}\alpha_t = c$, the oracle kernel is $Q_{s|t}(x) = \mathcal{N}(cx, v)$. Thus, for every x, translation and scale invariance give

$$\mathrm{TV}(K_{s|t}(x), Q_{s|t}(x)) = \mathrm{TV}(\mathcal{N}(0, 1+u\rho), \mathcal{N}(0,1)).$$

In contrast, $p_t Q_{s|t} = \mathcal{N}(0, c^2+v) = \mathcal{N}(0, 1-\delta_\rho)$, where

$$\delta_\rho = \frac{u(1-u)\rho^2}{(1+\rho)(1+u\rho)} \in (0,1).$$

For completeness, the density g_h of $\mathcal{N}(0,1+h)$ satisfies $\partial_h g_0(z) = \frac{1}{2}(z^2-1)\varphi(z)$. For $|h| \le 1/2$, direct differentiation gives the integrable envelope $|\partial_h g_h(z)| \le C(1+z^2)e^{-z^2/3}$. The fundamental theorem of calculus and dominated convergence therefore imply $(g_h - g_0)/h \to \partial_h g_0$ in L^1. Moreover, $\mathbb{E}|Z^2-1| = 4\int_1^\infty (z^2-1)\varphi(z)\,dz = 4\varphi(1)$, using symmetry, $\mathbb{E}(Z^2-1) = 0$, and $(z\varphi(z))' = (1-z^2)\varphi(z)$. Consequently,

$$\mathrm{TV}(\mathcal{N}(0,1+h), \mathcal{N}(0,1)) = \varphi(1)|h| + o(|h|).$$

Taking $h = u\rho$ and $h = -\delta_\rho$ proves both statements. □

The positive leading constants show that the two local orders differ even for Gaussian data. These local asymptotics hold for fixed u, without a uniform positive leading constant as u approaches 0 or 1. The following calculation gives a separate global example; neither result asserts dimension-optimality.

C.6 A FIXED-INTERVAL GLOBAL FIRST-ORDER EXAMPLE

Proposition C.5 (Global Gaussian discretization order)**.** *Let $d = 1$ and $p_d = \mathcal{N}(0,1)$. Fix SNR endpoints $\eta_+ > \eta_- > 0$, independently of N, and use $\eta_i = \eta_+ e^{-iH/N}$, where $H = \log(\eta_+/\eta_-)$. Set $u_i = \alpha_i^2 = \eta_i/(1+\eta_i)$ and $u_\pm = \eta_\pm/(1+\eta_\pm)$, so $u_0 = u_+$ and $u_N = u_-$. The standard oracle sampler initialized from γ^1 has output $q_{0,N}^{\mathrm{oracle}} = \mathcal{N}(0, 1-V_{0,N})$, with*

$$V_{0,N} = \frac{1}{u_+}\sum_{i=1}^{N} \frac{(u_{i-1}-u_i)^2}{1-u_i} \in (0,1). \tag{83}$$

Moreover,

$$\lim_{N\to\infty} N\,\mathrm{TV}(p_0, q_{0,N}^{\mathrm{oracle}}) = \varphi(1)\frac{H(u_+^2-u_-^2)}{2u_+} > 0. \tag{84}$$

Thus the global TV discretization error is $\Theta(N^{-1})$ for this fixed-interval sampler family.

Proof. All forward marginals are $\mathcal{N}(0,1)$, so initialization is exact and $\bar{y}(x,t_i) = \alpha_i x$. Substituting into (3) gives the oracle update

$$\widehat{X}_{i-1} = c_i \widehat{X}_i + \sqrt{v_i}\,\widehat{Z}_i, \qquad c_i = \frac{\alpha_i}{\alpha_{i-1}}, \qquad v_i = (1-c_i^2)\frac{1-u_{i-1}}{1-u_i},$$

where each $\widehat{Z}_i$ is a fresh standard Gaussian independent of the current state, and $\widehat{X}_N \sim \mathcal{N}(0,1)$. In particular, every generated marginal is a centered Gaussian. If its variance is w_i, then $w_N = 1$ and $w_{i-1} = c_i^2 w_i + v_i$. Defining $V_i = 1 - w_i$ gives

$$V_N = 0, \qquad V_{i-1} = c_i^2 V_i + \epsilon_i, \qquad \epsilon_i = 1 - c_i^2 - v_i = \frac{(u_{i-1}-u_i)^2}{u_{i-1}(1-u_i)} > 0. \tag{85}$$

Because $v_i > 0$ and $0 < c_i^2 + v_i < 1$, the variance recursion implies $0 < w_i < 1$ for $i < N$, while (85) implies $V_i > 0$. Unrolling the latter recursion and using $\prod_{j=1}^{i-1} c_j^2 = u_{i-1}/u_0$, with the empty product equal to one, proves (83), where $V_{0,N} = V_0$.

Let $h = H/N$ and $u(s) = \eta_+ e^{-s}/(1+\eta_+ e^{-s})$ for $0 \le s \le H$. This smooth function stays in the fixed compact interval $[u_-, u_+] \subset (0,1)$ and satisfies $u'(s) = -u(s)(1-u(s))$. Its first two derivatives are bounded on $[0,H]$. Taylor expansion at $s_i = ih$, uniformly in i and N, therefore gives

$$u_{i-1} - u_i = hu_i(1-u_i) + O(h^2), \qquad \frac{(u_{i-1}-u_i)^2}{1-u_i} = h^2 u_i^2 (1-u_i) + O(h^3).$$

The second remainder is uniform because $1 - u_i \ge 1 - u_+ > 0$. There are N terms; hence the remainder after summation and multiplication by N is $O(N^2 h^3) = O(N^{-1})$. The Riemann-sum limit yields

$$\begin{aligned} \lim_{N\to\infty} NV_{0,N} &= \frac{H}{u_+}\int_0^H u(s)^2(1-u(s))\,ds \\ &= \frac{H}{u_+}\int_{u_-}^{u_+} u\,du = \frac{H(u_+^2 - u_-^2)}{2u_+}. \end{aligned} \tag{86}$$

In particular, $V_{0,N} \to 0$. The L^1 differentiability of Gaussian densities proved in Proposition C.4 gives

$$\mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(0, 1-v)) = \varphi(1)v + o(v) \qquad (v \downarrow 0).$$

Combining this expansion with (86) proves (84). □

The recursion holds for every $N \ge 1$, and the main theorem's step condition holds for all sufficiently large N. Here $p_0 = p_d = \gamma^1$, so the error contains no initialization or smoothing discrepancy. The example establishes first-order global behavior for these fixed endpoints and this prescribed log-SNR grid. It is not a lower bound over other samplers or grids, and does not establish optimality of the dimension dependence or of the moving-endpoint clean-data rate.

C.7 Growing intervals: a Gaussian obstruction to an endpoint-uniform rate

The preceding fixed-endpoint asymptotic does not by itself justify substituting diverging endpoints. The next result treats that limit directly, with a remainder bound uniform over the entire log-SNR axis.

Proposition C.6 (Gaussian error on a growing log-SNR interval)**.** *Let $d = 1$, $p_d = \mathcal{N}(0,1)$, and use the standard oracle posterior-variance sampler initialized from γ^1. Suppose $\eta_{0,N} \to \infty$, $\eta_{N,N} \to \eta_- $ with $0 \le \eta_- < \infty$, and the equal log-SNR spacing $h_N = H_N/N \to 0$, where $H_N = \log(\eta_{0,N}/\eta_{N,N})$. Write $u_- = \eta_-/(1+\eta_-)$. Then*

$$\mathrm{TV}(p_{0,N}, q_{0,N}^{\mathrm{oracle}}) = \left(\frac{\varphi(1)}{2}(1-u_-^2) + o(1)\right)\frac{H_N}{N}, \qquad \varphi(1) = \frac{e^{-1/2}}{\sqrt{2\pi}}. \tag{87}$$

In particular, the polynomial endpoints $\eta_{0,N} = N^a$ and $\eta_{N,N} = N^{-b}$, with fixed $a, b > 0$, give $\Theta(\log N/N)$ TV error. A bound C/N with C independent of the endpoints is therefore false for this grid family, even with this fixed Gaussian data law and exact predictions.

Proof. Write $h = h_N$, $\rho = e^h - 1$, $s_i = \log \eta_{i,N}$, and $u_i = \eta_{i,N}/(1+\eta_{i,N})$. The exact finite-grid variance recursion (85) does not require fixed endpoints. It gives $q_{0,N}^{\mathrm{oracle}} = \mathcal{N}(0, 1 - V_{0,N})$ and

$$V_{0,N} = \frac{\rho^2}{u_0}\sum_{i=1}^N g_h(s_i), \qquad g_h(s) = \frac{e^{2s}}{(1+e^{s+h})^2(1+e^s)}. \tag{88}$$

Let $g(s) = e^{2s}/(1+e^s)^3$. Since $1 + e^s \le 1 + e^{s+h} \le e^h(1+e^s)$, $e^{-2h}g(s) \le g_h(s) \le g(s)$ for all s. Also, g increases to $4/27$ at $s = \log 2$, then decreases to zero, and hence

$$\int_{\mathbb{R}} |g'(s)|\,\mathrm{d}s = \frac{8}{27}, \qquad \int_{\mathbb{R}} g(s)\,\mathrm{d}s = \int_0^\infty \frac{u}{(1+u)^3}\,\mathrm{d}u = \frac{1}{2}.$$

On each interval $[s_i, s_i + h]$, the fundamental theorem of calculus bounds the error of its left-endpoint rectangle by $h\int_{s_i}^{s_i+h} |g'(s)|\,\mathrm{d}s$. Summing over these disjoint grid intervals proves the uniform estimate

$$\left| h\sum_{i=1}^N g(s_i) - \int_{s_N}^{s_0} g(s)\,\mathrm{d}s \right| \le \frac{8}{27}h. \tag{89}$$

As $\eta_{N,N} \to \eta_-$, $s_0 \to \infty$, and $h \to 0$, the integral tends to $(1-u_-^2)/2$; when $\eta_- = 0$, the lower integration limit tends to $-\infty$. The comparison with g_h therefore gives $h\sum_i g_h(s_i) \to (1-u_-^2)/2$. Since $\rho/h \to 1$ and $u_0 \to 1$, equation (88) implies $V_{0,N}/h \to (1-u_-^2)/2$. Finally, $p_{0,N} = \gamma^1$ and the L^1 Gaussian-density expansion in Proposition C.4 gives $\mathrm{TV}(\gamma^1, \mathcal{N}(0, 1-V_{0,N})) = \varphi(1)V_{0,N} + o(V_{0,N})$. This proves (87). The polynomial-endpoint specialization has $H_N = (a+b)\log N$, so the mesh condition holds eventually and the claimed obstruction follows. □

The positive coefficient also applies when the terminal SNR is fixed, so the obstruction persists when only the stopping noise shrinks. There is no initialization, network, or smoothing discrepancy in this example, since every forward marginal equals γ^1. It proves the necessity of some interval dependence for the specified grid, not the optimality of H_N^2/N, nor a lower bound over all samplers or grids. For example, on the same Gaussian law, an equally spaced noise-amplitude grid $\sigma_i = \sigma_0 + i(\sigma_N - \sigma_0)/N$ gives a different conclusion. Put $u_i = 1 - \sigma_i^2$ and $\delta_\sigma = (\sigma_N - \sigma_0)/N$. The variance recursion, valid on any strictly increasing positive-noise grid, yields

$$V_{0,N} = \frac{1}{u_0}\sum_i \delta_\sigma^2 \frac{(\sigma_i + \sigma_{i-1})^2}{\sigma_i^2} \le \frac{4(\sigma_N - \sigma_0)^2}{u_0 N}.$$

Thus, if u_0 is bounded away from zero, this Gaussian example has $O(1/N)$ error uniformly in the endpoints on that alternative grid. This calculation is not a general-data convergence theorem: extending the JGC accumulation to such nonuniform log-SNR grids requires new control of the covariance boundary terms.

D Linear-dimensional accumulated discretization error

The Gaussian-channel reduction in (48) isolates the posterior-mean transport defect of the original sampler. We now control these defects jointly across an equal log-SNR grid. The essential additional step is to interpolate through *exact* Gaussian channels and retain the covariance resolvent throughout the calculation. Its finite-step contribution is absorbed into a bounded covariance potential whose boundary terms telescope across the grid. This gives a linear dependence on dimension without replacing any finite-step term by an asymptotic remainder.

Theorem D.1 (Accumulated discretization on an equal log-SNR grid)**.** *Let $Y \sim p_d$ have finite second moment. Let $N \ge 1$, $H > 0$, and $\eta_i = \eta_0 e^{-iH/N} > 0$ for $i = 0, \dots, N$. For the oracle posterior-variance kernels Q_i^{std} of (5), if $\rho = e^{H/N} - 1 \le 1$, then*

$$\begin{aligned}\sum_{i=1}^N \mathrm{TV}(p_{i-1}, p_i Q_i^{\mathrm{std}}) &\le \frac{d(9H+5)}{4}\left(e^{H/N} - 1\right) \\ &\le \frac{e^{H/N} H(9H+5)}{4}\,\frac{d}{N}.\end{aligned} \tag{90}$$

The step condition is $N \ge H/\log 2$, with no dimension dependence. The result bounds the *sum* of marginal defects. It does not assert a uniform $Cd\rho^2$ estimate for every individual step, or a linear-dimensional rate for an arbitrary nonuniform grid. Lemma C.1 remains a complementary local estimate. We prove the theorem first for bounded data and then pass to finite second moments by approximation in Section D.5. Throughout this appendix, H denotes the log-SNR span, whereas H denotes a posterior covariance matrix.

D.1 Posterior centering and a resolvent-weighted Gaussian identity

Let V be bounded, let $Z \sim \mathcal{N}(0, I_d)$ be independent of V, and let $X = V + Z$ have density p. Define

$$\begin{aligned} &m(x) = \mathbb{E}[V \mid X = x], && \xi = V - m(X), && h(x) = \mathbb{E}[Z \mid X = x], \\ &\mathsf{H}(x) = \mathbb{E}[\xi\xi^\top \mid X = x], && \mathsf{T}(x) = \mathbb{E}[\xi^{\otimes 3} \mid X = x], && t(x) = \mathbb{E}[\xi\|\xi\|^2 \mid X = x]. \end{aligned}$$

For a scalar $\beta \ge 0$, let

$$\mathsf{R} = (I + \beta\mathsf{H})^{-1}, \qquad v = \mathsf{R}\xi, \qquad \mathfrak{q}_0 = -\operatorname{tr}\mathsf{H} - Z^\top\xi, \qquad \mathfrak{q}_\beta = \operatorname{div}_x v - Z^\top v.$$

All derivatives of joint fields hold V fixed. We use $(Dv)_{ij} = \partial_j v_i$ and $(\operatorname{div} M)_i = \sum_j \partial_j M_{ij}$ for a matrix field. The contraction $\mathsf{T}[v]$ has entries $\sum_k \mathsf{T}_{ijk} v_k$. Differentiating the Gaussian posterior gives

$$D_x \xi = -\mathsf{H}, \qquad D_x \mathsf{H} = \mathsf{T}, \qquad D_x v = -\mathsf{R}\mathsf{H} - \beta \mathsf{R}\mathsf{T}[v]. \tag{91}$$

The posterior is centered under its full conditional law; no restriction to a localization region is made in these identities. Since $Z = h - \xi$ and $\nabla \log p = -h$, we have

$$w := p^{-1} \operatorname{div}(p\mathsf{H}) = t - \mathsf{H}h = \mathbb{E}[\mathfrak{q}_0 \xi \mid X]. \tag{92}$$

Lemma D.2 (Resolvent-weighted signed density)**.** *With the notation above, put*

$$r = \mathbb{E} \operatorname{tr} \mathsf{H}, \qquad a = \mathbb{E} \operatorname{tr}(\mathsf{H}^2), \qquad \mathcal{B}_\beta = \mathbb{E} \operatorname{tr}\{\mathsf{R}\mathsf{T}[v]\mathsf{R}\mathsf{T}[v]\}.$$

Then the signed density $S_\beta = p^{-1} \operatorname{div}(p\mathsf{R}w)$ satisfies

$$S_\beta = \mathbb{E}[\mathfrak{q}_0 \mathfrak{q}_\beta \mid X] - \operatorname{tr}(\mathsf{R}\mathsf{H}) - \operatorname{tr}(\mathsf{R}\mathsf{H}^2), \tag{93}$$

$$\mathbb{E}|S_\beta| \le \sqrt{(r+a)(r+a+\beta^2 \mathcal{B}_\beta)} + r + a. \tag{94}$$

Proof. Differentiating the joint Gaussian integral gives, for a smooth joint vector field $F(x, V)$ whose value and first derivatives have at most polynomial growth in x, uniformly over the bounded support of V,

$$p^{-1} \operatorname{div}\{p\mathbb{E}[F \mid X]\} = \mathbb{E}[\operatorname{div}_x F - Z^\top F \mid X].$$

Apply this identity to $F = \mathfrak{q}_0 v$. Direct differentiation gives $\nabla \mathfrak{q}_0 = -t - \xi + \mathsf{H}Z$. The identities $\mathbb{E}[v \mid X] = 0$ and $\mathbb{E}[\xi Z^\top \mid X] = -\mathsf{H}$ therefore imply

$$\mathbb{E}[v^\top \nabla \mathfrak{q}_0 \mid X] = -\operatorname{tr}(\mathsf{R}\mathsf{H}) - \operatorname{tr}(\mathsf{R}\mathsf{H}^2),$$

which proves (93).

For completeness, the Gaussian divergence isometry for a possibly nonsymmetric Jacobian is

$$\mathbb{E}(\operatorname{div} v - Z^\top v)^2 = \mathbb{E}\|v\|^2 + \mathbb{E} \operatorname{tr}((Dv)^2). \tag{95}$$

This is a matrix square, not a squared Frobenius norm. To verify it, condition on V and write $\delta_G v = \operatorname{div} v - Z^\top v$. Using $\partial_i \delta_G v = \delta_G(\partial_i v) - v_i$, two Gaussian integrations by parts give

$$\mathbb{E}(\delta_G v)^2 = -\mathbb{E} \sum_i v_i \partial_i \delta_G v = \mathbb{E}\|v\|^2 + \mathbb{E} \sum_{ij} (\partial_j v_i)(\partial_i v_j).$$

Equation (91) now yields

$$\begin{aligned} \mathbb{E}\mathfrak{q}_0^2 &= r + a, \\ \mathbb{E}\mathfrak{q}_\beta^2 &= \mathbb{E} \operatorname{tr}(\mathsf{R}^2\mathsf{H}) + \mathbb{E} \operatorname{tr}((\mathsf{R}\mathsf{H})^2) + \beta^2 \mathcal{B}_\beta \le r + a + \beta^2 \mathcal{B}_\beta. \end{aligned} \tag{96}$$

The cross term containing $\mathsf{T}[v]$ vanishes after conditioning on X, because this contraction is linear in v and $\mathbb{E}[v \mid X] = 0$. Also $\operatorname{tr}(\mathsf{R}\mathsf{T}[v]\mathsf{R}\mathsf{T}[v]) = \|\mathsf{R}^{1/2}\mathsf{T}[v]\mathsf{R}^{1/2}\|_F^2 \ge 0$. Taking absolute values in (93), then applying Cauchy–Schwarz after full expectation and using $0 \preceq \mathsf{R} \preceq I$, proves (94). No conditional sign is assumed for $\mathbb{E}[\mathfrak{q}_0 \mathfrak{q}_\beta \mid X]$. For bounded V, posterior moments and the derivatives used here are bounded. Smooth cutoffs justify both integrations by parts, since all remaining factors grow at most polynomially in Z. □

D.2 Exact-channel interpolation controls the full finite step

Fix bounded U, let $X = U + Z$, and let $\varepsilon \in [0, 1]$. Write $A = 1 + \varepsilon$. For $c \in [1, A]$, define

$$\begin{aligned} X_c &= cU + Z, & m_c(x) &= \mathbb{E}[U \mid X_c = x], & \delta &= A - c, \\ F_c(x) &= x + \delta m_c(x), & \mu_c &= (F_c)_\# \mathcal{L}(X_c). \end{aligned}$$

These laws connect exactly the two outputs in the channel reduction:

$$\mu_1 = \mathcal{L}(X + \varepsilon m_1(X)), \qquad \mu_A = \mathcal{L}(X + \varepsilon U).$$

Use the preceding subsection's notation with current prior $V = cU$. If $\Sigma_c = \mathrm{Cov}(U \mid X_c)$, then $\mathsf{H} = c^2\Sigma_c$, $\beta = \delta/c$, and $DF_c = I + \beta\mathsf{H}$ has inverse R. The map F_c is a smooth global diffeomorphism: its Jacobian is at least I, so it is strongly monotone; boundedness of m_c makes it proper, and its nonsingular Jacobian gives a smooth inverse.

Posterior differentiation gives

$$\partial_c m_c = \Sigma_c x - c\,\mathrm{Cov}(U, \|U\|^2 \mid X_c), \qquad D_x m_c = c\Sigma_c.$$

Expanding around m_c and using (92) gives the acceleration identity

$$\begin{aligned} b_c := \partial_c m_c + (D_x m_c) m_c &= \Sigma_c h - c\mathbb{E}[(U - m_c)\|U - m_c\|^2 \mid X_c] \\ &= -c^{-2} w. \end{aligned} \tag{97}$$

For $g \in C_c^\infty(\mathbb{R}^d)$, differentiate along the coupling $X_c = cU + Z$. Conditional averaging cancels $U - m_c$, and therefore

$$\begin{aligned} \frac{d}{dc}\mathbb{E}g(F_c(X_c)) &= \delta\mathbb{E}[\nabla g(F_c)^\top b_c] \\ &= -\frac{\delta}{c^2}\mathbb{E}[\nabla_x\{g(F_c)\}^\top \mathsf{R} w] = \frac{\delta}{c^2}\mathbb{E}[g(F_c) S_\beta]. \end{aligned} \tag{98}$$

All coefficients are integrable uniformly over this compact c interval. Integrating (98) thus gives an identity of finite signed measures, first against smooth compactly supported tests and then by uniqueness of Radon measures. With the convention $\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|_{\mathrm{var}}$, it follows that

$$\mathrm{TV}(\mu_A, \mu_1) \le \frac{1}{2}\int_1^A \frac{\delta}{c^2}\mathbb{E}|S_\beta|\, dc. \tag{99}$$

This controls the complete finite step, including all orders in ε.

For step i, set $U = \sqrt{\eta_i}Y$, $s_i = \log\eta_i$, and $s = \log(\eta_i c^2)$. Then $dc = (c/2)ds$ and

$$\begin{aligned} J_i &= [s_i, s_i + 2\log(1+\varepsilon)], \qquad \beta_i(s) = e^{(s_i + 2\log(1+\varepsilon) - s)/2} - 1, \\ \mathrm{TV}(\mu_A, \mu_1) &\le \frac{1}{4}\int_{J_i} \beta_i(s)\mathbb{E}|S_{\beta_i(s)}|\, ds. \end{aligned} \tag{100}$$

The interpolation interval has twice the log-SNR length of an algorithmic cell. This follows from the squared signal scale $\eta_i c^2$ and must be retained in the accumulation below.

D.3 A COVARIANCE POTENTIAL ABSORBS THE FINITE-STEP TENSOR ENERGY

We next couple the exact Gaussian posteriors across SNR. The underlying Gaussian observation and MMSE identities are standard (see Guo et al., 2011 for background on Gaussian MMSE); the derivation below makes explicit the normalization and matrix contractions needed by the finite-step argument.

For bounded Y, consider the canonical observation process $O_\tau = \tau Y + W_\tau$, with Brownian motion W independent of Y. The likelihood of the full observation history is proportional to $\exp\{y^\top O_\tau - \tau\|y\|^2/2\}$, so its endpoint is sufficient. Its posterior is consequently the same as that for $\sqrt{\tau}Y + Z$. Let $\mathcal{F}_\tau$ be the observation filtration, $m_\tau = \mathbb{E}[Y \mid \mathcal{F}_\tau]$, and $I_\tau = O_\tau - \int_0^\tau m_u\, du$. This innovation is a continuous martingale in the observation filtration with quadratic variation τI_d, hence Brownian. Itô differentiation of the Gaussian likelihood quotient gives, for a bounded scalar function g,

$$d\mathbb{E}[g(Y) \mid \mathcal{F}_\tau] = \mathrm{Cov}(g(Y), Y \mid \mathcal{F}_\tau)^\top dI_\tau.$$

Apply this to coordinates and coordinate products. With $\Sigma_\tau = \mathrm{Cov}(Y \mid \mathcal{F}_\tau)$ and third central posterior tensor $\mathsf{T}_\tau^{\mathrm{raw}}$, the result is

$$dm_\tau = \Sigma_\tau\, dI_\tau, \qquad d\Sigma_\tau = \mathsf{T}_\tau^{\mathrm{raw}}\, dI_\tau - \Sigma_\tau^2 d\tau.$$

The negative covariance drift is the Itô product $-dm_\tau\, dm_\tau^\top$. Put $s = \log\tau$, $\mathsf{H} = \tau\Sigma_\tau$, and $\mathsf{T} = \tau^{3/2}\mathsf{T}_\tau^{\mathrm{raw}}$. On any finite log-SNR interval, Brownian time change then yields

$$d\mathsf{H} = (\mathsf{H} - \mathsf{H}^2)ds + \mathsf{T}\, dB_s, \qquad r'(s) = r(s) - a(s). \tag{101}$$

The one-time posterior laws in this coupling coincide with those in the exact-channel interpolation. Here $(\mathsf{T}\,dB_s)_{ij} = \sum_k \mathsf{T}_{ijk} dB_{k,s}$, $r(s) = \mathbb{E}\,\mathrm{tr}\,\mathsf{H}$, and $a(s) = \mathbb{E}\,\mathrm{tr}(\mathsf{H}^2)$. All coefficients are bounded on compact positive SNR intervals. Also

$$0 \le r(s) \le d, \tag{102}$$

because $\mathsf{H} = \mathrm{Cov}(Z \mid \sqrt{\tau}Y + Z)$ and the law of total covariance bounds its expected trace by $\mathbb{E}\|Z\|^2 = d$.

Define the two-variable covariance potential

$$f(s,b) = \mathbb{E}\,\mathrm{tr}\{\mathsf{H}(s)(I + b\mathsf{H}(s))^{-1}\}.$$

Let $b = \beta(s) \ge 0$ be differentiable and $\mathsf{R} = (I + \beta(s)\mathsf{H})^{-1}$. Matrix differentiation, without commuting a perturbation K past H, gives

$$\begin{aligned} D_{\mathsf{H}}\,\mathrm{tr}(\mathsf{RH})[K] &= \mathrm{tr}(\mathsf{R}^2 K), \\ D^2_{\mathsf{H}}\,\mathrm{tr}(\mathsf{RH})[K,K] &= -2\beta\,\mathrm{tr}(\mathsf{R}K\mathsf{R}K\mathsf{R}), \\ \partial_\beta\,\mathrm{tr}(\mathsf{RH}) &= -\,\mathrm{tr}(\mathsf{R}^2\mathsf{H}^2). \end{aligned}$$

Itô's formula applied to (101) therefore gives

$$\frac{d}{ds} f(s,\beta(s)) = \mathbb{E}\,\mathrm{tr}(\mathsf{R}^2\mathsf{H}) - (1+\beta')\mathbb{E}\,\mathrm{tr}(\mathsf{R}^2\mathsf{H}^2) - \beta\mathcal{K}_\beta, \tag{103}$$

where $\mathcal{K}_\beta = \mathbb{E}\sum_k \mathrm{tr}(\mathsf{R}\mathsf{T}_k\mathsf{R}\mathsf{T}_k\mathsf{R})$ and $(\mathsf{T}_k)_{ij} = \mathsf{T}_{ijk}$. The derivative on the left includes the changing resolvent parameter; it is not a partial derivative in s.

Lemma D.3 (Finite-step tensor energy)**.** *For every $\beta \ge 0$, $\beta\mathcal{B}_\beta \le \mathcal{K}_\beta$. Along the exact-channel interpolation of step i, if $0 \le \beta_i \le 1$, then*

$$\beta_i^2 \mathcal{B}_{\beta_i} \le \beta_i \mathcal{K}_{\beta_i} \le r(s) - \frac{d}{ds} f(s,\beta_i(s)). \tag{104}$$

Proof. Diagonalize H pointwise, with eigenvalues $\lambda_j \ge 0$, and write $R_j = (1+\beta\lambda_j)^{-1}$. Since $\mathbb{E}[vv^\top \mid X] = \mathsf{RHR}$,

$$\mathcal{B}_\beta = \mathbb{E}\sum_{ijk} R_i R_j R_k^2 \lambda_k \mathsf{T}^2_{ijk}, \qquad \mathcal{K}_\beta = \mathbb{E}\sum_{ijk} R_i^2 R_j \mathsf{T}^2_{ijk}.$$

Using $\beta R_k \lambda_k \le 1$ and the full symmetry of T gives

$$\beta\mathcal{B}_\beta \le \mathbb{E}\sum_{ijk} R_i R_j R_k \mathsf{T}^2_{ijk} \le \mathcal{K}_\beta. \tag{105}$$

For the second inequality, apply $3R_iR_jR_k \le R_i^2R_j + R_j^2R_k + R_k^2R_i$ and relabel indices; each of the three tensor sums equals $\mathcal{K}_\beta$. The eigenbasis is used only for this algebraic inequality and is never differentiated.

Along (100), $\beta_i' = -(1+\beta_i)/2$, so $1 + \beta_i' = (1-\beta_i)/2 \ge 0$. Use this sign in (103), together with $\mathbb{E}\,\mathrm{tr}(\mathsf{R}^2\mathsf{H}) \le r$, to obtain (104). □

D.4 Exact accumulation on the discrete grid

We prove Theorem D.1 for bounded Y. Put $\ell = H/N$ and $s_i = s_0 - i\ell$. The Gaussian bridge reduction (48) uses $\varepsilon = \eta_{i-1}/\eta_i - 1 = \rho = e^\ell - 1$ at every step. Thus the auxiliary exact-channel intervals are $J_i = [s_i, s_i + 2\ell]$. Introduce $s_{-1} = s_0 + \ell$ solely for this interpolation; the sampler's output endpoint remains s_0. Their boundary terms telescope as

$$\begin{aligned} \Delta_r &:= \sum_{i=1}^{N} \{r(s_i) - r(s_i + 2\ell)\} \\ &= r(s_{N-1}) + r(s_N) - r(s_{-1}) - r(s_0) \le 2d. \end{aligned} \tag{106}$$

This also holds for $N = 1$: the two occurrences of $r(s_0)$ cancel. Counting interval overlaps, $\sum_i |J_i| = 2H$. Consequently (101) and (102) give

$$\mathfrak{A} := \sum_i \int_{J_i} (r + a)\, ds = 2 \sum_i \int_{J_i} r\, ds + \Delta_r \le 4dH + 2d. \tag{107}$$

Since $\beta_i(s_i) = \rho$, $\beta_i(s_i + 2\ell) = 0$, $f(s, 0) = r(s)$, and $f(s, \rho) \le r(s)$, (104) also yields

$$\begin{aligned} \mathfrak{B} &:= \sum_i \int_{J_i} \beta_i^2 \mathcal{B}_{\beta_i}\, ds \\ &\le \sum_i \int_{J_i} r\, ds + \sum_i \{f(s_i, \rho) - r(s_i + 2\ell)\} \le 2dH + 2d. \end{aligned} \tag{108}$$

Neither step replaces a discrete sum by an uncontrolled Riemann approximation, and no monotonicity of r is required.

Apply Cauchy–Schwarz to (94) on the disjoint union of the *indexed* intervals J_i, so that overlaps retain their multiplicity. Then

$$\begin{aligned} \sum_i \int_{J_i} \mathbb{E}|S_{\beta_i}|\, ds &\le \sqrt{\mathfrak{A}(\mathfrak{A} + \mathfrak{B})} + \mathfrak{A} \\ &\le 2\mathfrak{A} + \tfrac{1}{2}\mathfrak{B} \le d(9H + 5). \end{aligned} \tag{109}$$

Since $0 \le \beta_i \le \rho$, equations (48) and (100) imply

$$\sum_i \mathrm{TV}(p_{i-1}, p_i Q_i^{\mathrm{std}}) \le \frac{\rho}{4} \sum_i \int_{J_i} \mathbb{E}|S_{\beta_i}|\, ds \le \frac{d\rho(9H + 5)}{4}.$$

Finally, $e^x - 1 \le x e^x$ proves the second inequality in (90). One may retain the slightly smaller bound

$$\frac{d\rho}{2} \left\{ 2H + 1 + \sqrt{(2H + 1)(3H + 2)} \right\}$$

before the arithmetic relaxation in (109). The simpler constant in the theorem makes the interval dependence explicit.

D.5 Finite-second-moment data and classwise extension

Removing boundedness. Let μ be the law of Y and, for sufficiently large M, put $\mu_M = \mu(\cdot \mid \|Y\| \le M)$. Then $\mu_M \to \mu$ in total variation. At a fixed positive channel scale c, posterior numerators and denominators can be written, with the conditioning normalization canceled, as

$$N_M(x) = \int_{\|y\| \le M} y \varphi(x - cy)\, \mu(dy), \qquad D_M(x) = \int_{\|y\| \le M} \varphi(x - cy)\, \mu(dy),$$

where φ is the standard Gaussian density. Both converge uniformly on every compact observation set: Gaussian weighting uniformly bounds each polynomial in y there, and the discarded prior mass tends to zero. The limiting denominator is continuous and positive, hence bounded away from zero on that compact set. Thus the posterior means N_M / D_M converge locally uniformly.

Channel laws converge in total variation by contraction under their common Gaussian kernel and are therefore tight. Locally uniform convergence of the posterior means implies weak convergence of the posterior-mean transport outputs. Indeed, restrict a bounded continuous test function to a compact input set carrying all but an arbitrarily small amount of channel mass; the transport maps converge uniformly there and their images lie in a common compact set. The exact transport outputs converge even in total variation by their common Gaussian kernel. Total variation is jointly lower semicontinuous under weak convergence on $\mathbb{R}^d$, by its dual over bounded continuous functions. Hence, for each fixed step,

$$\mathrm{TV}(P, Q) \le \liminf_{M \to \infty} \mathrm{TV}(P_M, Q_M)$$

for the two normalized transport output laws. The same passage applies to their finite sum, since a finite sum of nonnegative lower limits is at most the lower limit of the sums. The bound already proved is independent of M and therefore passes to μ. Applying the Gaussian bridge contraction once more proves Theorem D.1 as stated. This argument does not require high data moments or convergence of learned denoisers; it concerns only the oracle marginal defects.

Corollary D.4 (Classwise accumulated discretization)**.** *Under the grid and data assumptions of Theorem D.1, let $p_d = \sum_{l=1}^{L} \pi_l p_d^l$ be the fixed finite class decomposition with $\pi_l > 0$. Let Q_i be the oracle mixture kernel of Section 2, using the true posterior class weights and classwise posterior means. Then*

$$\sum_{i=1}^{N} \mathrm{TV}(p_{i-1}, p_i Q_i) \le \frac{d(9H+5)}{4}(e^{H/N} - 1). \tag{110}$$

There is no multiplicative dependence on the number of classes.

Proof. Every positive-probability class has finite conditional second moment. Let p_i^l be its forward marginal and Q_i^l its oracle Gaussian kernel. The class-joint identity $p_i(x)a^l(x,t_i) = \pi_l p_i^l(x)$ implies

$$p_i Q_i = \sum_l \pi_l (p_i^l Q_i^l), \qquad p_{i-1} = \sum_l \pi_l p_{i-1}^l.$$

Convexity of total variation and Theorem D.1, applied separately to each fixed class law on the same grid, give

$$\sum_i \mathrm{TV}(p_{i-1}, p_i Q_i) \le \sum_l \pi_l \sum_i \mathrm{TV}(p_{i-1}^l, p_i^l Q_i^l) \le \sum_l \pi_l \frac{d(9H+5)\rho}{4} = \frac{d(9H+5)\rho}{4}.$$

□

The preceding proof sharpens the discretization budget for the *same* oracle kernels used in the global comparison lemma. Initialization, network, and routing errors are added by that comparison; no conditional reverse-kernel approximation is substituted for the marginal defect. The full-posterior centering and joint Gaussian cancellation are shared with the local proof, while the resolvent potential retains the finite-step cancellation needed for the linear-dimensional accumulated rate.

A Gaussian illustration. For $Y \sim \mathcal{N}(0, I_d)$, the normalized posterior covariance at SNR $\eta = e^s$ is deterministic. Writing $\lambda = \eta/(1+\eta)$ gives

$$\mathsf{H} = \lambda I_d, \qquad r = d\lambda, \qquad a = d\lambda^2, \qquad \mathsf{T} = 0, \qquad f(s,b) = \frac{d\lambda}{1+b\lambda}. \tag{111}$$

Thus covariance contractions carry one dimension factor, whereas $\mathbb{E}\|Z\|^4 = d(d+2)$ has two. For non-Gaussian data, H can be random and T nonzero. The potential controls this additional tensor contribution while retaining integrated linear dimension dependence. The example illustrates the benefit of matrix structure, without proving dimension optimality.

D.6 A MOMENT-SENSITIVE REFINEMENT OF THE INTERVAL FACTOR

The factor H in the prefactor of (90) results from replacing the posterior-covariance trace by d throughout the interval. Retaining its small-SNR behavior gives a sharper bound without changing the sampler or adding a data assumption.

Proposition D.5 (Moment-sensitive interval bound)**.** *Under the assumptions of Theorem D.1, put $h = H/N$, $m = M/d$, and*

$$\begin{aligned} L_M &= \min\left\{H, \log\frac{1+e^h\eta_0 m}{1+\eta_N m}\right\}, \\ B_M &= \frac{\eta_N m}{1+\eta_N m} + \frac{e^h \eta_N m}{1+e^h\eta_N m}. \end{aligned} \tag{112}$$

Then the standard oracle marginal defects satisfy

$$\sum_i \mathrm{TV}(p_{i-1}, p_i Q_i^{\mathrm{std}}) \le \mathcal{D}_N^{\mathrm{mom}} := \frac{d\rho}{4}\left(9L_M + \frac{5}{2}B_M\right) \le \mathcal{D}_N. \tag{113}$$

The same bound holds for the oracle mixture kernels Q_i, with M the second moment of the full data law. Consequently $\mathcal{D}_N^{\mathrm{mom}}$ may replace $\mathcal{D}_N$ in both main theorems.

Proof. First suppose Y is bounded and retain $r(s)$, $a(s)$, and $f(s,b)$ from the preceding proof. For $\eta = e^s$, the posterior mean minimizes the mean-square prediction error of Y from $X = \sqrt{\eta}Y + Z$. Comparing with the linear predictor cX, and taking $c = \sqrt{\eta}M/(d+\eta M)$, gives

$$\mathbb{E}\|Y - \mathbb{E}[Y \mid X]\|^2 \le (1 - c\sqrt{\eta})^2 M + c^2 d = \frac{dM}{d+\eta M}.$$

No centering of Y is required, since Z is independent and has mean zero. Thus

$$r(s) = \eta\,\mathbb{E}\|Y - \mathbb{E}[Y \mid X]\|^2 \le d\frac{\eta m}{1+\eta m}. \tag{114}$$

For $J_i = [s_i, s_i + 2h]$, set $G_N = \sum_i \int_{J_i} r(s)\,ds$ and retain the signed boundary sum Δ_r in (106). The indexed intervals have total length $2H$ and overlap at most twice, within $[s_N, s_0 + h]$. Therefore

$$G_N \le 2d \min\left\{H, \int_{\eta_N}^{e^h\eta_0} \frac{m}{1+\eta m}\,\mathrm{d}\eta\right\} = 2dL_M, \qquad \Delta_r \le r(s_N) + r(s_{N-1}) \le dB_M. \tag{115}$$

The second inequality drops the nonnegative upper-endpoint traces; it is valid also for $N = 1$.

The exact energy identities used above give $\mathfrak{A} = 2G_N + \Delta_r$ and $\mathfrak{B} \le G_N + \Delta_r$. Hence (109), before the constant trace bound, yields

$$\sum_i \int_{J_i} \mathbb{E}|S_{\beta_i}|\,ds \le 2\mathfrak{A} + \tfrac{1}{2}\mathfrak{B} \le \tfrac{9}{2}G_N + \tfrac{5}{2}\Delta_r \le 9dL_M + \tfrac{5}{2}dB_M.$$

Multiplication by $\rho/4$ proves the first bound in (113). The second follows from $L_M \le H$ and $B_M \le 2$. For general finite-second-moment data, the truncated laws in Section D.5 have moments $M_R \to M$. The same lower-semicontinuity argument passes the finite sum of defects to the limit; the right side of (113) is continuous in $M \ge 0$. This proves the standard bound, including $M = 0$.

For the classwise statement, let $M_l = \mathbb{E}[\|Y\|^2 \mid J = l]$, so $\sum_l \pi_l M_l = M$. At fixed endpoints, the function $M \mapsto \log[(1+AM)/(1+BM)]$ is concave for $A > B \ge 0$, since its derivative is $(A-B)/[(1+AM)(1+BM)]$. Its minimum with a constant remains concave, as does $M \mapsto CM/(1+CM)$ for $C \ge 0$. Thus $M \mapsto \mathcal{D}_N^{\mathrm{mom}}$ is concave. Applying the standard bound to each class and using the classwise TV comparison and Jensen's inequality gives $\sum_l \pi_l \mathcal{D}_N^{\mathrm{mom}}(M_l) \le \mathcal{D}_N^{\mathrm{mom}}(M)$. □

For fixed d, M, η_0 and $\eta_N = N^{-b}$ with $b > 0$, the new factor L_M stays bounded, $B_M = O(N^{-b})$, and $\rho = O(\log N/N)$. The discretization bound therefore improves to $O(\log N/N)$ with the dimension and moment dependence displayed in (113). This is convergence to a fixed positive-noise target, not to clean data. If also $\eta_0 = N^a$ with $a > 0$, then $L_M = O(\log N)$ in general; the main $O(d\log^2 N/N)$ order remains valid. The refinement removes an unnecessarily large small-SNR contribution, without asserting independence of the shrinking stopping noise.

E Network error and routing stability

We first prove a comparison lemma that combines a local marginal total variation error with a conditional KL error evaluated under the true forward marginals. The proof uses the clipping construction of Li & Yan (2025a, Sections 4.2–4.4). The auxiliary processes serve only to compare probability measures; the sampling algorithm remains unchanged.

E.1 A comparison lemma under the true marginals

For a density p and a Markov kernel K admitting a Lebesgue transition density, write $(pK)(z) = \int p(x)K(x,z)\,\mathrm{d}x$. Throughout this appendix, transition i runs from t_i to t_{i-1}, for $i = 1, \dots, N$. Use finite jointly Borel versions of the transition densities; identities and inequalities between densities are understood Lebesgue-almost everywhere.

Lemma E.1 (Marginal-TV and conditional-KL comparison)**.** *Let $p_0, \dots, p_N$ and ν be probability densities on $\mathbb{R}^d$, and let Q_i, P_i be Markov kernels admitting Lebesgue transition densities. Let $q_N = \nu$ and $q_{i-1} = q_i P_i$. Define*

$$\delta_i = \mathrm{TV}(p_{i-1}, p_i Q_i), \tag{116}$$

$$k_i = \int p_i(x)\, \mathrm{KL}\big(Q_i(x,\cdot) \| P_i(x,\cdot)\big)\, \mathrm{d}x. \tag{117}$$

If $\sum_{i=1}^N k_i < \infty$, then

$$\mathrm{TV}(p_0, q_0) \le \mathrm{TV}(p_N, \nu) + \sum_{i=1}^N \delta_i + \left(\frac{1}{2}\sum_{i=1}^N k_i\right)^{1/2}. \tag{118}$$

Proof. Choose finite Borel versions of $p_0, \dots, p_N, \nu$ and adjoin an absorbing state $\dagger$ to $\mathbb{R}^d$. Define subprobability densities recursively by

$$\begin{aligned} r_N &= \min\{p_N, \nu\}, \qquad u_{i-1} = r_i Q_i, \\ c_{i-1}(z) &= \min\{1, p_{i-1}(z)/u_{i-1}(z)\}, \\ r_{i-1}(z) &= c_{i-1}(z) u_{i-1}(z) = \min\{p_{i-1}(z), u_{i-1}(z)\}. \end{aligned} \tag{119}$$

At each step, Tonelli gives $\int u_{i-1} = \int r_i \le 1$, so the raw integral defining u_{i-1} is finite outside a Lebesgue-null set. Replace it by zero on that set before forming the ratio, and set $c_{i-1} = 1$ wherever $u_{i-1} = 0$. These choices leave the represented measures unchanged. Every transition kernel assigns zero mass to each such output null set at every input, so they also leave all retention probabilities unchanged. These densities are the nonabsorbed parts of a Markov process: initialize with density r_N on $\mathbb{R}^d$ and the remaining mass at $\dagger$; from a state in $\mathbb{R}^d$, apply Q_i, then retain the proposed output z with probability $c_{i-1}(z)$, sending it to $\dagger$ otherwise. The state $\dagger$ is absorbing.

By construction, $0 \le r_i \le p_i$. Let $\ell_i = 1 - \int r_i$. Since both p_N and ν integrate to one, $\ell_N = \mathrm{TV}(p_N, \nu)$. Positivity of Q_i gives $u_{i-1} \le p_i Q_i$, and hence

$$\begin{aligned} \ell_{i-1} - \ell_i &= \int (u_{i-1} - p_{i-1})_+ \\ &\le \int (p_i Q_i - p_{i-1})_+ = \delta_i. \end{aligned} \tag{120}$$

It follows that $\ell_0 \le \mathrm{TV}(p_N, \nu) + \sum_i \delta_i$.

Construct a second absorbing process with the same initial distribution and the same fixed retention functions c_{i-1}, replacing each Q_i by P_i. Let s_i be its nonabsorbed density. Induction gives

$$s_N = r_N \le q_N, \qquad s_{i-1} = c_{i-1}(s_i P_i) \le q_i P_i = q_{i-1}. \tag{121}$$

More explicitly, write $\mathcal{X} = \mathbb{R}^d \cup \{\dagger\}$, with $\dagger$ an isolated point. For $K_i = Q_i$ or P_i, the extended transition from $x \in \mathbb{R}^d$ is

$$\overline{K}_i(x, dz) = c_{i-1}(z) K_i(x,z)\, dz + \left[\int_{\mathbb{R}^d} (1 - c_{i-1}(y)) K_i(x,y)\, dy\right] \delta_\dagger(dz), \tag{122}$$

where the first term is supported on $\mathbb{R}^d$, and $\overline{K}_i(\dagger, \cdot) = \delta_\dagger$. Let R and S denote the resulting probability laws on $\mathcal{X}^{N+1}$, using $\overline{Q}_i$ and $\overline{P}_i$, respectively. Both have initial law $r_N(x)\, dx + \ell_N \delta_\dagger$.

Finiteness of k_i implies $Q_i(x,\cdot) \ll P_i(x,\cdot)$ for p_i-almost every x, hence also for r_i-almost every x. At such an input, the two extended transitions apply the same retention channel to Q_i and P_i. They therefore preserve this absolute continuity, and data processing bounds their conditional KL by $\mathrm{KL}(Q_i(x,\cdot)\|P_i(x,\cdot))$. At $\dagger$, both transitions equal $\delta_\dagger$. Induction along the finite path gives $\mathsf{R} \ll \mathsf{S}$. We detail the integrability needed for the path-space chain rule. Relative to the dominating measure $dz + \delta_\dagger$ on $\mathcal{X}$, take the ratio $\theta_i(x,z)$ of the two extended transition densities at inputs where $\overline{Q}_i(x,\cdot) \ll \overline{P}_i(x,\cdot)$, with arbitrary measurable values on irrelevant sets and $\theta_i(\dagger,\dagger) = 1$. This ratio is finite and strictly positive $\overline{Q}_i(x,\cdot)$-almost everywhere at r_i-almost every input. The common initial law and the finite product of transition densities give

$$\frac{d\mathsf{R}}{d\mathsf{S}}(X_N, \dots, X_0) = \prod_{i=1}^N \theta_i(X_i, X_{i-1}) \qquad \mathsf{R}\text{-almost surely.}$$

For each such input,

$$\int (-\log \theta_i(x,z))_+ \, \overline{Q}_i(x,dz) = \int_{\{0<\theta_i<1\}} -\theta_i \log \theta_i \, \overline{P}_i(x,dz) \le 1/e.$$

Consequently, data processing and $r_i \le p_i$ imply $\mathbb{E}_{\mathsf{R}}|\log \theta_i(X_i, X_{i-1})| \le k_i + 2/e < \infty$. We may therefore integrate the logarithm of the finite product term by term:

$$\begin{aligned} \mathrm{KL}(\mathsf{R}\|\mathsf{S}) &= \sum_{i=1}^{N} \int r_i(x) \, \mathrm{KL}(\overline{Q}_i(x,\cdot)\|\overline{P}_i(x,\cdot)) \, dx \\ &\le \sum_{i=1}^{N} \int r_i(x) \, \mathrm{KL}(Q_i(x,\cdot)\|P_i(x,\cdot)) \, \mathrm{d}x \\ &\le \sum_{i=1}^{N} k_i, \end{aligned} \tag{123}$$

where the last step uses $r_i \le p_i$ and nonnegativity of conditional KL.

Let $\mathsf{R}_0, \mathsf{S}_0$ be the extended endpoint distributions. For any Borel set $E \subseteq \mathbb{R}^d$,

$$\begin{aligned} p_0(E) - q_0(E) &\le (p_0 - r_0)(E) + (r_0 - s_0)(E) \\ &\le \ell_0 + \mathrm{TV}(\mathsf{R}_0, \mathsf{S}_0). \end{aligned}$$

Taking the supremum over E, then using Pinsker's inequality and data processing from the path to its endpoint, proves (118). □

The defect in (116) is measured after integrating the input state against p_i. Thus this lemma preserves any cancellation used to bound a local marginal error. It does not require a bound on $\int p_i(x) \, \mathrm{TV}(K_i(x,\cdot), Q_i(x,\cdot)) \, \mathrm{d}x$ for an exact reverse kernel K_i.

E.2 Two inequalities for Gaussian mixtures

For probability densities f, g, use the convention $\chi^2(f\|g) = \int (f-g)^2/g$, with $0/0 = 0$ and a positive numerator over zero contributing $+\infty$. Thus finite $\chi^2(f\|g)$ implies $f \ll g$. Whenever $f \ll g$, expansion gives $1 + \chi^2(f\|g) = \int_{\{g>0\}} f^2/g$, also in the extended sense. For KL, likewise set $0 \log(0/g) = 0$ and $\mathrm{KL}(f\|g) = +\infty$ if $f \not\ll g$.

Lemma E.2 (Perturbing the weights of nearby Gaussian components)**.** *Let g_l be the density of $\mathcal{N}(\mu_l, vI)$, with $v > 0$, and let a, b belong to the probability simplex in $\mathbb{R}^L$. Set $f = \sum_l a_l g_l$, $g = \sum_l b_l g_l$, and $D_\mu = \max_{l,j} \|\mu_l - \mu_j\|$. Then*

$$\chi^2(f\|g) \le \left(e^{D_\mu^2/v} - 1\right) \|a - b\|_1^2. \tag{124}$$

No positive lower bound on the weights is required.

Proof. Convexity of the reciprocal function gives the pointwise bound $1/g \le \sum_j b_j/g_j$. Write $z_l = a_l - b_l$. Since $\sum_l z_l = 0$, $f - g = \sum_l z_l (g_l - g_j)$ for every j. The triangle inequality in $L^2(g_j^{-1} \, \mathrm{d}x)$ implies

$$\begin{aligned} \chi^2(f\|g) &\le \sum_j b_j \int \frac{[\sum_l z_l (g_l - g_j)]^2}{g_j} \\ &\le \sum_j b_j \left[\sum_l |z_l| \sqrt{\chi^2(g_l\|g_j)} \right]^2. \end{aligned} \tag{125}$$

For equal-covariance Gaussian densities, completing the square gives

$$\frac{g_l(z)^2}{g_j(z)} = \exp\left(\frac{\|\mu_l - \mu_j\|^2}{v} \right) (2\pi v)^{-d/2} \exp\left(-\frac{\|z - (2\mu_l - \mu_j)\|^2}{2v} \right).$$

Integrating this identity gives

$$\int \frac{g_l^2}{g_j} = \exp\left(\frac{\|\mu_l - \mu_j\|^2}{v}\right), \qquad \chi^2(g_l\|g_j) = \exp\left(\frac{\|\mu_l - \mu_j\|^2}{v}\right) - 1. \tag{126}$$

Substitute this into (125) and use $\sum_j b_j = 1$ to obtain (124). □

Lemma E.3 (An entropy comparison through an intermediate density)**.** *For probability densities f, h, g, whenever the right-hand side is finite,*

$$\mathrm{KL}(f\|g) \le 2\,\mathrm{KL}(f\|h) + \log\big(1 + \chi^2(h\|g)\big). \tag{127}$$

Proof. The finite-right-hand-side assumption implies $f \ll h \ll g$ and $Z := \int_{\{g>0\}} h^2/g = 1 + \chi^2(h\|g) < \infty$. For every bounded measurable φ, the entropy variational inequality applied to $\varphi/2$ under h gives

$$\int f\varphi \le 2\,\mathrm{KL}(f\|h) + 2\log\int h e^{\varphi/2}.$$

Indeed, for $\psi = \varphi/2$, let $h_\psi = he^\psi / \int he^\psi$. Boundedness of ψ and finiteness of $\mathrm{KL}(f\|h)$ justify

$$0 \le \mathrm{KL}(f\|h_\psi) = \mathrm{KL}(f\|h) - \int f\psi + \log\int he^\psi.$$

Cauchy–Schwarz, using $h \ll g$, implies

$$\int he^{\varphi/2} = \int_{\{g>0\}} \frac{h}{\sqrt{g}}\sqrt{ge^\varphi} \le Z^{1/2}\left(\int ge^\varphi\right)^{1/2}.$$

Consequently,

$$\int f\varphi - \log\int ge^\varphi \le 2\,\mathrm{KL}(f\|h) + \log Z.$$

Taking the supremum over bounded measurable φ gives (127) by the variational formula for KL, including when the left-hand side is initially allowed to be infinite. To recall why bounded tests suffice, take $\varphi_n = \max\{-n, \min\{n, \log(f/g)\}\}$ on $\{g > 0\}$, with $\log(0/g) = -\infty$ and $\varphi_n = 0$ on $\{g = 0\}$. Then $ge^{\varphi_n} \le g + f$ and $\int ge^{\varphi_n} \to 1$ by dominated convergence. The negative part of $f\log(f/g)$ has integral at most $1/e$; dominated convergence for that part and monotone convergence for the positive part give $\int f\varphi_n \to \mathrm{KL}(f\|g)$ in the extended sense. Thus no integrability assumption on $\log(h/g)$ is needed. □

E.3 The SNR coefficient and the standard sampler

For the applications below, abbreviate the positive-time forward density p_{t_i} by p_i. Recall that A_i, B_i are the coefficients in the Gaussian mean and v_i is the actual transition variance. For $s = t_{i-1}$ and $t = t_i$, they are

$$A_i = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}, \qquad B_i = \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}, \qquad v_i = \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2}. \tag{128}$$

Write $\eta_i = \alpha_{t_i}^2/\sigma_{t_i}^2$ and $\Delta_i = \eta_{i-1} - \eta_i$. The identity $\alpha_s^2\sigma_{t|s}^2 = \alpha_s^2 - \alpha_t^2$ yields

$$\frac{B_i^2}{v_i} = \frac{\alpha_s^2 - \alpha_t^2}{\sigma_s^2\sigma_t^2} = \sigma_s^{-2} - \sigma_t^{-2} = \Delta_i, \tag{129}$$

$$\sum_{i=1}^N \Delta_i = \eta_0 - \eta_N =: \Gamma. \tag{130}$$

For the standard sampler, take

$$Q_i(x,\cdot) = \mathcal{N}(A_i x + B_i\bar{y}(x,t_i), v_i I),$$
$$P_i(x,\cdot) = \mathcal{N}(A_i x + B_i y_\theta(x,t_i), v_i I),$$

where $\bar{y}(x,t) = \mathbb{E}[Y \mid X_t = x]$. The Gaussian KL formula gives

$$k_i = \frac{\Delta_i}{2}\mathbb{E}_{X_{t_i}\sim p_i}\|\bar{y}(X_{t_i},t_i) - y_\theta(X_{t_i},t_i)\|^2. \tag{131}$$

Thus a uniform true-marginal squared error bound ε_y^2 contributes at most $\frac{1}{2}\sqrt{\Gamma}\,\varepsilon_y$ to (118). This step requires neither a bounded true posterior mean nor control of the error under a generated marginal.

E.4 JOINT DENOISER AND CLASSIFIER PERTURBATIONS

Let $\pi_l > 0$ be the class priors and p_i^l the normalized class-conditional forward densities, so that

$$p_i = \sum_{l=1}^{L} \pi_l p_i^l, \qquad a^l(x,t_i)p_i(x) = \pi_l p_i^l(x). \tag{132}$$

Let $\bar{y}^l(x,t)$ be the true class-conditional denoiser, and let $y_\theta^l(x,t)$ be its learned estimate. We use finite-valued measurable versions of these means. Write $a_\phi(x,t)$ for the measurable learned classifier, assumed to take values in the probability simplex. Define the three kernels

$$\begin{aligned} Q_i(x,\cdot) &= \sum_l a^l(x,t_i)\mathcal{N}(A_i x + B_i \bar{y}^l(x,t_i), v_i I), \\ R_i(x,\cdot) &= \sum_l a^l(x,t_i)\mathcal{N}(A_i x + B_i y_\theta^l(x,t_i), v_i I), \\ P_i(x,\cdot) &= \sum_l a_\phi^l(x,t_i)\mathcal{N}(A_i x + B_i y_\theta^l(x,t_i), v_i I). \end{aligned} \tag{133}$$

The kernel R_i is used only as an intermediate conditional distribution; no assumption is made about the marginals of a chain driven by R_i.

Define the state-dependent diameter of the learned denoisers by

$$D_i(x) = \max_{l,j} \|y_\theta^l(x,t_i) - y_\theta^j(x,t_i)\|. \tag{134}$$

This is a finite measurable function because the number of classes is finite. The error budgets are

$$E_i^2 = \mathbb{E}_{p_i}\left[\sum_l a^l(X_{t_i},t_i)\|\bar{y}^l(X_{t_i},t_i) - y_\theta^l(X_{t_i},t_i)\|^2\right], \tag{135}$$

$$W_i^2 = \mathbb{E}_{p_i}\|a(X_{t_i},t_i) - a_\phi(X_{t_i},t_i)\|_1^2, \tag{136}$$

$$\Psi_i = \mathbb{E}_{p_i}\log\left[1 + \left(e^{\Delta_i D_i(X_{t_i})^2} - 1\right)\|a(X_{t_i},t_i) - a_\phi(X_{t_i},t_i)\|_1^2\right], \tag{137}$$

$$\Omega_i = \mathbb{E}_{p_i}\left[\left(e^{\Delta_i D_i(X_{t_i})^2} - 1\right)\|a(X_{t_i},t_i) - a_\phi(X_{t_i},t_i)\|_1^2\right]. \tag{138}$$

All these quantities are nonnegative extended-real expectations under the true forward marginal p_i. The quantity W_i^2 is always at most 4, but its finiteness alone does not imply that Ψ_i is finite. For the global entropy bound below, assume $E_i^2 < \infty$ and $\Psi_i < \infty$ for every i; no finite Ω_i is required.

At a fixed input x, augment Q_i and R_i with their common latent class label. The joint KL is the class-weighted sum of the component KLs, and projection can only decrease KL. Hence

$$\mathrm{KL}(Q_i(x,\cdot)\|R_i(x,\cdot)) \le \frac{\Delta_i}{2}\sum_l a^l(x,t_i)\|\bar{y}^l(x,t_i) - y_\theta^l(x,t_i)\|^2. \tag{139}$$

The diameter of the component means in R_i and P_i is at most $|B_i|D_i(x)$: their common $A_i x$ term cancels. Therefore Lemma E.2 and (129) give

$$\chi^2(R_i(x,\cdot)\|P_i(x,\cdot)) \le \left(e^{\Delta_i D_i(x)^2} - 1\right)\|a(x,t_i) - a_\phi(x,t_i)\|_1^2. \tag{140}$$

Combining Lemma E.3 with (139)–(140) gives the pointwise bound

$$\begin{aligned} \mathrm{KL}(Q_i(x,\cdot)\|P_i(x,\cdot)) \le \Delta_i \sum_l a^l(x,t_i)\|\bar{y}^l(x,t_i) - y_\theta^l(x,t_i)\|^2 \\ + \log\left[1 + \left(e^{\Delta_i D_i(x)^2} - 1\right)\|a(x,t_i) - a_\phi(x,t_i)\|_1^2\right]. \end{aligned} \tag{141}$$

The entropy lemma applies at every input: the finitely many component means and their diameter are finite, so both terms on its right-hand side are finite pointwise. Integrating the nonnegative bound against p_i therefore yields, also when an expectation is infinite,

$$k_i \le \Delta_i E_i^2 + \Psi_i. \tag{142}$$

When the right-hand sides are summable, Lemma E.1 implies

$$\mathrm{TV}(p_0, q_0) \le \mathrm{TV}(p_N, \nu) + \sum_i \delta_i + \left\{ \frac{1}{2} \sum_i \left(\Delta_i E_i^2 + \Psi_i \right) \right\}^{1/2}. \tag{143}$$

A uniform bound on the learned diameter. Suppose $D_i(x) \le D < \infty$ for p_i-almost every x and every i. Then

$$\Psi_i \le \Omega_i \le \left(e^{D^2 \Delta_i} - 1 \right) W_i^2. \tag{144}$$

In particular, the required routing energies are finite. Let $E = \max_i E_i$, $W = \max_i W_i$, and $\Delta_{\max} = \max_i \Delta_i$. Since

$$\sum_i (e^{D^2 \Delta_i} - 1) \le D^2 e^{D^2 \Delta_{\max}} \sum_i \Delta_i = D^2 e^{D^2 \Delta_{\max}} \Gamma,$$

the square-root term in (143) is bounded by

$$\sqrt{\Gamma/2} \left(E + D e^{D^2 \Delta_{\max}/2} W \right). \tag{145}$$

For fixed endpoints and a fixed diameter bound D, this coefficient is bounded independently of the number of steps. A uniform output bound $\|y_\theta^l\| \le R$ is sufficient with $D = 2R$, but a shared unbounded shift of all learned means does not affect their diameter. The factor $e^{\Delta_i D_i(x)^2} - 1$ in (140) records that, for fixed denoiser differences, the Gaussian components approach one another relative to their transition noise as the step shrinks. For general state-dependent diameters, finiteness of each Ψ_i alone does not provide a bound uniform over refinements: such a conclusion also requires uniform control of their sum, for example through (144).

Remark E.4 (Per-class and aggregate error conventions). If each expert has squared network error at most ε_y^2 under the global density p_i, then $E_i^2 \le L\varepsilon_y^2$. If each classifier coordinate has squared error at most ε_a^2 under p_i, Cauchy–Schwarz gives $W_i^2 \le L^2 \varepsilon_a^2$. These bounds give $E \le \sqrt{L}\varepsilon_y$ and $W \le L\varepsilon_a$ in (145). Alternatively, a per-class network error bound under the normalized class-conditional density p_i^l gives $E_i^2 \le \varepsilon_y^2$ by (132). The latter is a different error convention and must be stated explicitly.

E.5 A COUPLING BOUND WITHOUT EXPONENTIAL ROUTING ENERGIES

The following alternative uses a weighted second moment of the learned diameter. For bounded geometry, its dependence on a small routing error can be weaker than the entropy bound, but it avoids an exponential coefficient and recovers the standard network error coefficient under exact routing. Proposition E.6 below compares its integrability condition with that of Ψ_i.

Proposition E.5 (Stability by coupling the Gaussian components)**.** *Under the DDPM kernel setup of Section 2, use the equal log-SNR grid of Theorem 3.2 with $\rho = e^{H/N} - 1 \le 1$. Let all prediction functions be finite-valued and measurable, with $a_\phi(x, t_i)$ in the probability simplex. Define*

$$T_i^2 = \mathbb{E}_{p_i} \left[D_i(X_{t_i})^2 \frac{\|a(X_{t_i}, t_i) - a_\phi(X_{t_i}, t_i)\|_1}{2} \right]. \tag{146}$$

If $E_i^2 < \infty$ and $T_i^2 < \infty$ for every i, then

$$\mathrm{TV}(p_0, q_0) \le \mathcal{I}_N + \mathcal{D}_N + \frac{1}{2} \left[\sum_{i=1}^N \Delta_i (E_i + T_i)^2 \right]^{1/2}. \tag{147}$$

This statement requires neither finite Ψ_i nor a uniform bound on the learned outputs or their diameter.

Proof. Fix a step i and input x, and abbreviate the probability vectors by a, a_ϕ, the true means by $\bar{y}^l = \bar{y}^l(x, t_i)$, and the learned means by $h_l = y_\theta^l(x, t_i)$. Set

$$\zeta = \frac{1}{2} \|a - a_\phi\|_1 = \sum_l (a^l - a_\phi^l)_+ = \sum_j (a_\phi^j - a^j)_+.$$

When $\zeta = 0$, let $c_{lj} = a^l \mathbf{1}_{\{l=j\}}$. Otherwise, define

$$c_{lj} = \min(a^l, a^l_\phi)\mathbf{1}_{\{l=j\}} + \frac{(a^l - a^l_\phi)_+ (a^j_\phi - a^j)_+}{\zeta}. \tag{148}$$

The row sums of c are a, its column sums are a_ϕ, and $\sum_{l \neq j} c_{lj} = \zeta$: the two residual supports are disjoint. The formula, including its separate zero case, is measurable in a, a_ϕ. No positive lower bound on their coordinates is used.

Consider two joint distributions on (l, j, z) with respective densities

$$c_{lj}\mathcal{N}(A_i x + B_i \bar{y}^l, v_i I)(z), \qquad c_{lj}\mathcal{N}(A_i x + B_i h_j, v_i I)(z).$$

Their z marginals are $Q_i(x, \cdot)$ and $P_i(x, \cdot)$. They have the same label-pair weights, so projection and the Gaussian KL formula give

$$\mathrm{KL}(Q_i(x,\cdot) \| P_i(x,\cdot)) \leq \frac{\Delta_i}{2} \sum_{l,j} c_{lj} \|\bar{y}^l - h_j\|^2. \tag{149}$$

Terms with $c_{lj} = 0$ contribute zero. All component means are finite, so the displayed conditional KL bound is finite at every input.

The triangle inequality in the Hilbert space with weights c_{lj} gives

$$\begin{aligned}\left(\sum_{l,j} c_{lj}\|\bar{y}^l - h_j\|^2\right)^{1/2} &\leq \left(\sum_{l,j} c_{lj}\|\bar{y}^l - h_l\|^2\right)^{1/2} + \left(\sum_{l,j} c_{lj}\|h_l - h_j\|^2\right)^{1/2} \\ &\leq \left(\sum_{l} a^l \|\bar{y}^l - h_l\|^2\right)^{1/2} + D_i(x)\sqrt{\zeta}.\end{aligned}$$

The second line uses the row sums and the total off-diagonal mass. Applying the triangle inequality once more, now in $L^2(p_i)$, yields

$$k_i \leq \frac{\Delta_i}{2}(E_i + T_i)^2 < \infty. \tag{150}$$

The class-joint identity and Theorem D.1, applied to each normalized class law, give $\sum_i \delta_i \leq \mathcal{D}_N$. Substituting this accumulated bound and the conditional-KL estimates into Lemma E.1, with $\nu = \gamma^d$, proves (147). □

If $D_i \leq D$ almost everywhere, Cauchy–Schwarz implies

$$T_i^2 \leq \frac{D^2}{2}\mathbb{E}_{p_i}\|a - a_\phi\|_1 \leq \frac{D^2}{2} W_i.$$

Consequently, $E_i \leq E$ and $W_i \leq W$ give the simpler learning term

$$\frac{\sqrt{\Gamma}}{2}\left(E + D\sqrt{W/2}\right). \tag{151}$$

This removes the exponential coefficient at the cost of replacing the linear dependence on W in (145) by a square-root dependence. If the router is exact, $T_i = 0$ and the network error coefficient in (147) is the same as for standard DDPM.

The diameter is a relaxation of pairwise geometry. The joint-density argument in (149) holds for every coupling of a and a_ϕ. Its right-hand side can therefore be minimized over that finite transport polytope at each input. The diameter bound instead controls every mismatched pair by $D_i(x)$; it can be loose when a remote expert is unrelated to the routing error. We retain it because it separates network accuracy from an explicit routing moment. Neither this relaxation nor the minimum of the two integrated estimates characterizes optimal mixture sensitivity.

Proof of Corollary 5.1. Equation (142) holds in the extended sense. The same is true of $k_i \le \Delta_i(E_i + T_i)^2/2$: when $T_i < \infty$, the proof of (150) applies, and when $T_i = \infty$ the inequality is automatic. Thus, with each $E_i < \infty$,

$$k_i \le \kappa_i := \min\left\{\Delta_i E_i^2 + \Psi_i,\ \frac{\Delta_i}{2}(E_i + T_i)^2\right\}.$$

The assumption $\sum_i \kappa_i < \infty$ therefore supplies the finite conditional KL required by Lemma E.1, even if one of the two candidate bounds is infinite at a step. Using $\sum_i \delta_i \le \mathcal{D}_N$ from the classwise application of Theorem D.1 and $\nu = \gamma^d$ gives (35). The minimum is between deterministic integrated quantities at finitely many steps; no input-dependent selection is involved. Under exact routing, $\Psi_i = T_i = 0$, so $\kappa_i = \Delta_i E_i^2/2$ and the learning term becomes $\frac{1}{2}(\sum_i \Delta_i E_i^2)^{1/2}$. □

E.6 Moment conditions for the logarithmic routing energy

Proposition E.6 (Comparing routing energies)**.** *For any step in the setup of Section 2, let the prediction functions be finite-valued and measurable, and let a_ϕ take values in the probability simplex. With expectations allowed to be infinite,*

$$\Psi_i \le \Omega_i, \qquad \Psi_i \le 4\Delta_i \mathbb{E}_{p_i} D_i(X)^2, \qquad \Delta_i T_i^2 \le \Psi_i + 2/e. \tag{152}$$

Consequently a finite quadratic moment of the learned diameter suffices for finite Ψ_i, and finite Ψ_i implies finite T_i at any fixed positive step. There are finite-variance data and learned predictors with $E_i, T_i, \Psi_i < \infty$ but $\Omega_i = \infty$.

Proof. At a fixed input, put $s = \Delta_i D_i(x)^2 \ge 0$ and $\zeta = \|a(x,t_i) - a_\phi(x,t_i)\|_1/2 \in [0,1]$. The logarithmic integrand is $f = \log\big(1 + 4\zeta^2(e^s - 1)\big) \ge 0$. The inequality $\log(1+u) \le u$ gives the first assertion. Moreover,

$$1 + 4\zeta^2(e^s - 1) \le 1 + 4(e^s - 1) \le e^{4s},$$

where the last inequality is the tangent-line bound $z^4 \ge 1 + 4(z-1)$ at $z = 1$, with $z = e^s$. Thus $f \le 4s$. For $\zeta > 0$,

$$1 + 4\zeta^2(e^s - 1) - \zeta^2 e^s = 1 + \zeta^2(3e^s - 4) \ge 1 - \zeta^2 \ge 0.$$

It follows that $f \ge s + 2\log\zeta$, and hence

$$s\zeta \le \zeta f + 2\zeta\log(1/\zeta) \le f + 2/e.$$

The same inequality holds at $\zeta = 0$ by direct substitution. Integrating these nonnegative pointwise inequalities proves (152).

For the example, choose any admissible positive one-dimensional step, and write its input coefficients as $\alpha, \sigma > 0$ and its absolute SNR increment as $\Delta > 0$. Let $Y = \pm 1$ with equal probabilities, with sign as the class label, and let $X = \alpha Y + \sigma Z$ for an independent $Z \sim \mathcal{N}(0,1)$. The oracle denoisers and router are

$$\bar{y}^{\pm}(x) = \pm 1, \qquad a_+(x) = (1 + e^{-2\alpha x/\sigma^2})^{-1}.$$

Choose learned means $h_+(x) = cx$, $h_-(x) = 0$ and router $b = (1/2, 1/2)$, where $c > 0$ satisfies $\Delta c^2 > 1/(2\sigma^2)$. Then $D(x) = c|x|$ and $\|a(x) - b\|_1 = 2|a_+(x) - 1/2| \to 1$ as $|x| \to \infty$. The class-joint identity expresses E^2 as

$$E^2 = \tfrac{1}{2}\mathbb{E}[(1 - cX)^2 \mid Y = 1] + \tfrac{1}{2} < \infty.$$

Also $T^2 \le c^2 \mathbb{E}X^2 < \infty$ and $\Psi \le 4\Delta c^2 \mathbb{E}X^2 < \infty$. For all sufficiently large positive x, the squared routing discrepancy is bounded below by $1/4$. Since the marginal density is at least half the $\mathcal{N}(\alpha, \sigma^2)$ density, the integral defining Ω dominates, up to a positive constant, the tail integral of

$$\exp\left[\left(\Delta c^2 - \frac{1}{2\sigma^2}\right)x^2 + \frac{\alpha x}{\sigma^2}\right].$$

This integral diverges. The choice of c is made after the step is fixed, so the example also respects any prescribed relative-SNR step restriction. □

Retaining the logarithm therefore weakens the sufficient condition based on Ω_i. It does not enlarge the finiteness domain of the coupling bound: finite Ψ_i already implies finite T_i at each fixed step. The bounds can nevertheless differ in their numerical values and their behavior along a sequence of grids. In particular, the additive constant in $\Delta_i T_i^2 \le \Psi_i + 2/e$ cannot transfer a vanishing sum of logarithmic energies to a vanishing sum of coupling energies. Neither finite data variance nor finite class-weighted E_i^2 alone implies $\mathbb{E}_{p_i} D_i^2 < \infty$, as the tail construction below demonstrates.

E.7 A Bayes-router perturbation with quadratic KL sensitivity

Proposition E.7 (Overlap and coupling under a small routing error)**.** *Fix an admissible one-dimensional step and $m > 0$. Let $Y = \pm m$ with equal probabilities, with one atom per class, and use the exact expert predictions. Write $X = \alpha Y + \sigma Z$ for the true input, where $Z \sim \mathcal{N}(0,1)$ is independent, and set*

$$a(x) = \Pr(Y = m \mid X = x) = (1 + e^{-2\alpha m x/\sigma^2})^{-1}, \qquad b_\epsilon(x) = a(x) + \epsilon a(x)(1 - a(x)).$$

Use learned weights $(b_\epsilon, 1 - b_\epsilon)$, with $|\epsilon| \le 1$. Let $k(\epsilon)$ be the conditional KL averaged under the true input law. With the step and m fixed, as $\epsilon \to 0$,

$$k(\epsilon) = \Theta(\epsilon^2), \qquad \Psi(\epsilon) = \Theta(\epsilon^2), \qquad \frac{\Delta}{2} T(\epsilon)^2 = \Theta(|\epsilon|). \tag{153}$$

Here $E = 0$, $D = 2m$, and all constants are finite and strictly positive.

Proof. Put $u(x) = a(x)(1 - a(x))$. Since $a(x)^2 \le b_\epsilon(x) \le 2a(x) - a(x)^2$, the learned weights are valid everywhere. For $g_\pm^x(z) = \mathcal{N}(Ax \pm Bm, v)(z)$, write

$$q_x = a g_+^x + (1 - a) g_-^x, \qquad w_x = u(g_+^x - g_-^x), \qquad p_{\epsilon,x} = q_x + \epsilon w_x.$$

Then $\int w_x \, dz = 0$ and $|w_x| \le q_x$, since $u \le \min(a, 1 - a)$. For $|\epsilon| \le 1/2$, the Taylor remainder in $-\log(1 + \epsilon w_x/q_x)$ is bounded by a universal constant times $|\epsilon|^3 |w_x/q_x|^3$. Integrating against q_x removes the linear term and bounds the remainder uniformly in x by a constant times $|\epsilon|^3$. Thus

$$k(\epsilon) = \frac{\epsilon^2}{2} \mathcal{J} + O(|\epsilon|^3), \qquad \mathcal{J} = \mathbb{E}_X \int \frac{w_X(z)^2}{q_X(z)} \, dz \in (0, 1].$$

The upper bound follows from $|w_x| \le q_x$; strict positivity follows from $0 < a(x) < 1$ and $Bm > 0$, so the two Gaussian densities differ at every fixed input. This also justifies integration of the expansion over the true input law.

The ℓ_1 routing discrepancy is $2|\epsilon| u(X)$. Writing $s = 4\Delta m^2 > 0$ and using $0 < u \le 1/4$ gives

$$\Psi(\epsilon) = 4(e^s - 1)\epsilon^2 \mathbb{E} u(X)^2 + O(\epsilon^4),$$
$$\frac{\Delta}{2} T(\epsilon)^2 = \frac{s}{2} |\epsilon| \mathbb{E} u(X).$$

Both expectations are finite and positive, proving (153). □

The overlap estimate captures the quadratic order of the actual KL, whereas label coupling pays for the first-order mass of mismatched labels. For two distinct expert values, optimizing the component coupling still transports this mass over distance $2m$, so it does not remove the difference in orders. On a fixed finite grid, the resulting learning terms have orders $O(|\epsilon|)$ and $O(\sqrt{|\epsilon|})$; these are upper bounds, not matching output-TV lower bounds. In this example $\Omega = 4(e^s - 1)\epsilon^2 \mathbb{E} u(X)^2$ has the same quadratic order as Ψ. The small-error advantage comes from Gaussian overlap; retaining the logarithm has the separate integrability benefit in Proposition E.6.

E.8 Completion of the two main convergence proofs

Use the equal log-SNR grid $\eta_i = \eta_0 e^{-iH/N}$, where $H = \log(\eta_0/\eta_N)$, and suppose $\rho = e^{H/N} - 1 \le 1$. For the standard oracle kernel, Theorem D.1 gives

$$\sum_{i=1}^{N} \delta_i \le \mathcal{D}_N = \frac{d(9H + 5)}{4} (e^{H/N} - 1). \tag{154}$$

Combining this accumulated bound, (131), and Lemma E.1, with $e_i^2 = \mathbb{E}_{p_i}\|\bar{y}(X_{t_i}, t_i) - y_\theta(X_{t_i}, t_i)\|^2$, gives

$$\mathrm{TV}(p_0, q_0^{\mathrm{std}}) \le \mathrm{TV}(p_N, \nu) + \mathcal{D}_N + \frac{1}{2}\left(\sum_i \Delta_i e_i^2\right)^{1/2}. \tag{155}$$

Taking $\nu = \gamma^d$ proves Theorem 3.1. If $e_i \le \varepsilon_y$ for every i, telescoping $\sum_i \Delta_i = \Gamma$ also gives the uniform-error specialization $\sqrt{\Gamma}\,\varepsilon_y/2$.

For the decentralized oracle, let Q_i^l denote its lth Gaussian component. Identity (132) implies

$$p_i Q_i = \sum_l \pi_l (p_i^l Q_i^l), \qquad p_{i-1} = \sum_l \pi_l p_{i-1}^l. \tag{156}$$

Consequently, convexity of total variation, followed by Theorem D.1 applied to each normalized class distribution on the common grid, gives

$$\begin{aligned}\sum_{i=1}^N \delta_i &\le \sum_l \pi_l \sum_{i=1}^N \mathrm{TV}(p_{i-1}^l, p_i^l Q_i^l) \\ &\le \sum_l \pi_l \mathcal{D}_N = \mathcal{D}_N.\end{aligned} \tag{157}$$

Every positive-prior class has finite second moment because $\mathbb{E}[\|Y\|^2 \mid J = l] \le M/\pi_l < \infty$. The accumulated theorem is independent of this second-moment value. The integration against each p_i^l occurs before taking the absolute value, and the sum over steps is bounded before averaging over classes. No pointwise estimate $\delta_i \le \mathcal{D}_N/N$ is used.

Substitute (157) into (143) and set $\nu = \gamma^d$. This proves Theorem 3.2. Under a uniform learned-diameter bound, (144) and (145) also prove Corollary 3.3. In particular, under this diameter bound the uniform per-class global-marginal error convention gives

$$\begin{aligned}\mathrm{TV}(p_0, q_0) \le\ & \mathrm{TV}(p_N, \nu) + \mathcal{D}_N \\ &+ \sqrt{\Gamma/2}\left(\sqrt{L}\,\varepsilon_y + D e^{D^2\Delta_{\max}/2} L \varepsilon_a\right).\end{aligned} \tag{158}$$

Here L is the number of classes; H is the log-SNR interval length.

Finally, $e^x - 1 \le x e^x$ for $x \ge 0$ gives

$$\mathcal{D}_N \le \frac{e^{H/N} H(9H+5)}{4}\,\frac{d}{N}. \tag{159}$$

Thus the explicit discretization term has order d/N for fixed endpoints, under a mesh restriction independent of dimension. With a fixed learned-diameter bound and uniform E, W, the learning terms in (145) and (151) do not acquire a factor of N.

E.9 Why unweighted routing error does not control conditional KL

Proposition E.8 (A limitation of unweighted error budgets)**.** *For any fixed one-dimensional step satisfying the setup, there are a two-class data law with finite second moment and a family of finite Borel learned predictors for which $E_i, W_i \to 0$, while $\mathbb{E}_{p_i}\mathrm{KL}(Q_i(X,\cdot)\|P_i(X,\cdot)) = +\infty$ throughout the family. Thus E_i, W_i alone do not ensure the finite conditional KL used by the comparison proof.*

Proof. Suppress the step index and write $A, B, v > 0$, $\Delta = B^2/v$, $\alpha = \alpha_i \in (0,1)$, and $\sigma^2 = 1 - \alpha^2$. Let $Y = \pm 1$ with equal probabilities and let its sign be the class label. The noisy class densities are $p_\pm(x) = \varphi_{\sigma^2}(x \mp \alpha)$, where φ_{σ^2} is the centered normal density of variance σ^2; $p = (p_+ + p_-)/2$. The oracle experts are $\bar{y}^\pm = \pm 1$, and $a_+(x) = p_+(x)/(p_+(x) + p_-(x)) = (1 + e^{-2\alpha x/\sigma^2})^{-1}$. For $R \ge 0$, set

$$g_R(x) = \mathbf{1}_{\{x < -R\}} e^{x^2/(4\sigma^2)}, \qquad h_-(x) = -1, \qquad h_+(x) = 1 + g_R(x).$$

Use learned experts $h_\pm$ and learned routing weights $b_+ = 1$, $b_- = 0$ on $x < -R$, with $b = a$ elsewhere. Then

$$E^2 = \frac{1}{2}\int_{-\infty}^{-R} p_+(x)e^{x^2/(2\sigma^2)}\,dx = \frac{\sigma e^{-\alpha^2/(2\sigma^2)}}{2\alpha\sqrt{2\pi}}e^{-\alpha R/\sigma^2} \longrightarrow 0,$$
$$W^2 = 4\int_{x<-R} p(x)(1-a_+(x))^2\,dx \le 4\mathbb{P}_p(X < -R) \longrightarrow 0.$$

On this tail the conditional kernels are

$$Q_x = a_+(x)\mathcal{N}(Ax+B,v) + (1-a_+(x))\mathcal{N}(Ax-B,v), \qquad P_x = \mathcal{N}(Ax+Bh_+(x),v).$$

Write $m_x = Ax + B(2a_+(x)-1)$, $\mu_x = Ax + Bh_+(x)$, and $G_x = \mathcal{N}(m_x, v)$. For each fixed x, comparison after adding the class label bounds both $\mathrm{KL}(Q_x\|G_x)$ and $\mathrm{KL}(Q_x\|P_x)$ by finite weighted Gaussian KLs. The log density ratio $\log(dG_x/dP_x)$ is affine and hence Q_x-integrable. Splitting this log ratio therefore gives

$$\mathrm{KL}(Q_x\|P_x) = \mathrm{KL}(Q_x\|G_x) + \frac{(m_x-\mu_x)^2}{2v} \ge \frac{\Delta}{2}g_R(x)^2,$$

where $2a_+ - 1 \le 1$ was used in the last step. Averaging and retaining only the negative class yields

$$\begin{aligned}\mathbb{E}_p\,\mathrm{KL}(Q_X\|P_X) &\ge \frac{\Delta}{4}\int_{-\infty}^{-R} p_-(x)e^{x^2/(2\sigma^2)}\,dx \\ &= \frac{\Delta e^{-\alpha^2/(2\sigma^2)}}{4\sqrt{2\pi}\sigma}\int_{-\infty}^{-R} e^{-\alpha x/\sigma^2}\,dx = +\infty.\end{aligned}$$

□

The theorem's routing conditions exclude this example: on $x < -R$, $a_+ \le 1/2$, $D(x) = 2 + g_R(x)$, and $\|a-b\|_1 = 2(1-a_+) \ge 1$. On that tail,

$$\log\left[1 + (e^{\Delta D(x)^2} - 1)\|a-b\|_1^2\right] \ge \Delta D(x)^2.$$

The same divergent integral therefore gives $\Psi_i = +\infty$ and $T_i^2 = +\infty$; moreover $\Omega_i \ge \Psi_i = +\infty$. This is a limitation of controlling conditional KL by E_i, W_i alone, not an impossibility result for all TV bounds. Indeed, $Q_x = P_x$ off the tail, so $\mathrm{TV}(pQ, pP) \le \mathbb{P}_p(X < -R) \to 0$ in this example.

F Endpoints and schedules

This appendix proves the endpoint statements and Corollaries F.1 and F.2. It separates oracle endpoint rates from the learning conditions needed when the SNR interval grows. Throughout, $M = \mathbb{E}\|Y\|^2 < \infty$, and p_0 denotes the noisy law at the positive time t_0, rather than the clean-data distribution p_d.

F.1 Gaussian initialization

The standard initialization estimate is

$$\mathcal{I}_N \le \tfrac{1}{2}\sqrt{\alpha_N^2 M + d(\sigma_N^2 - 1 - \log\sigma_N^2)}. \tag{160}$$

Let $\varphi_{a,\Sigma}$ denote the density of $\mathcal{N}(a,\Sigma)$, and define joint probability measures

$$\mathsf{P}(dy,dx) = p_d(dy)\varphi_{\alpha_N y,\sigma_N^2 I_d}(x)\,dx, \qquad \mathsf{G}(dy,dx) = p_d(dy)\varphi_{0,I_d}(x)\,dx.$$

Their x-marginals are p_N and γ^d, respectively. Since $\sigma_N > 0$, both conditional densities are strictly positive, so $\mathsf{P} \ll \mathsf{G}$ even when p_d is singular. Their density ratio is the ratio of the two displayed Gaussian densities. Its logarithm is integrable under P because $M < \infty$, and direct Gaussian integration gives

$$\begin{aligned}\mathrm{KL}(\mathsf{P}\|\mathsf{G}) &= \mathbb{E}_Y\mathrm{KL}\big(\mathcal{N}(\alpha_N Y, \sigma_N^2 I_d)\|\mathcal{N}(0,I_d)\big) \\ &= \frac{1}{2}\left[\alpha_N^2 M + d\sigma_N^2 - d - d\log\sigma_N^2\right].\end{aligned}$$

Here the conditional quadratic terms have expectations $\mathbb{E}\|X_{t_N} - \alpha_N Y\|^2 = d\sigma_N^2$ and $\mathbb{E}\|X_{t_N}\|^2 = \alpha_N^2 M + d\sigma_N^2$. Data processing under $(y, x) \mapsto x$ therefore gives

$$\begin{aligned} \mathrm{KL}(p_N \| \gamma^d) &\le \mathrm{KL}(\mathsf{P} \| \mathsf{G}) \\ &= \frac{1}{2}\left[\alpha_N^2 M + d(\sigma_N^2 - 1 - \log \sigma_N^2)\right]. \end{aligned} \tag{161}$$

Pinsker's inequality, with $\mathrm{TV}(P, Q) \le \sqrt{\mathrm{KL}(P\|Q)/2}$, yields (160). In particular, the argument does not require $\mathrm{KL}(p_d\|\gamma^d)$ to be finite.

Writing $\eta = \eta_N$ gives $\alpha_N^2 = \eta/(1+\eta)$ and $\sigma_N^2 = 1/(1+\eta)$. Since

$$\log(1+\eta) - \frac{\eta}{1+\eta} = \int_0^\eta \frac{u}{(1+u)^2}\,\mathrm{d}u \le \frac{\eta^2}{2},$$

we obtain the explicit estimate

$$\mathcal{I}_N \le \frac{1}{2}\sqrt{\eta_N M + \tfrac{1}{2} d \eta_N^2}. \tag{162}$$

For $\eta_N = N^{-2}$, using $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for $u, v \ge 0$ gives the explicit rate

$$\mathcal{I}_N \le \frac{\sqrt{M}}{2N} + \frac{\sqrt{d}}{2\sqrt{2}\,N^2},$$

which gives the initialization rate discussed after the main theorems. A fixed nonzero terminal SNR instead leaves a fixed initialization term; refining the discretization alone does not force that term to zero.

F.2 The clean-data endpoint

The smoothing error obeys

$$W_2^2(p_d, p_0) \le (1-\alpha_0)^2 M + d\sigma_0^2. \tag{163}$$

Couple $Y \sim p_d$ and $X_{t_0} = \alpha_0 Y + \sigma_0 Z$, with $Z \sim \mathcal{N}(0, I_d)$ independent of Y. Both variables have finite second moments. The cross term is integrable by Cauchy–Schwarz and vanishes because independence and $\mathbb{E}Z = 0$ give $\mathbb{E}\langle Y, Z\rangle = 0$; no centering assumption on Y is needed. Hence

$$\mathbb{E}\|Y - X_{t_0}\|^2 = (1-\alpha_0)^2 \mathbb{E}\|Y\|^2 + \sigma_0^2 \mathbb{E}\|Z\|^2 = (1-\alpha_0)^2 M + d\sigma_0^2.$$

Taking the infimum over couplings gives (163). If p_d is supported on a countable set A, then $p_d(A) = 1$ while $p_0(A) = 0$ because $\sigma_0 > 0$ gives p_0 a density. Hence $\mathrm{TV}(p_d, p_0) = 1$. More generally, the same conclusion holds whenever p_d is supported on a Lebesgue-null set. Consequently, the main TV theorem targets the early-stopped distribution; it does not assert TV convergence to arbitrary clean data. The bounded-Lipschitz argument below gives a common distance between the generated law and clean data. It is compatible with this obstruction to TV convergence: every finite-grid output also has a density, because the last transition has strictly positive Gaussian variance.

F.3 Equal log-SNR spacing

Corollary F.1 (Equal spacing in log SNR). *Let $H = \log(\eta_0/\eta_N)$ and $\eta_i = \eta_0 e^{-iH/N}$. Under the remaining assumptions of the corresponding main theorem, if $e^{H/N} - 1 \le 1$, then*

$$\begin{aligned} \mathcal{D}_N &= \frac{d(9H+5)}{4}(e^{H/N} - 1) \le \frac{e^{H/N} H (9H+5)}{4}\frac{d}{N}, \\ \Delta_* &= \eta_0(1 - e^{-H/N}). \end{aligned} \tag{164}$$

For fixed endpoints, discretization is $O(d/N)$ and, for fixed D, the learning coefficients in (18) are bounded independently of N.

For $\eta_i = \eta_0 e^{-iH/N}$ and $H = \log(\eta_0/\eta_N) > 0$,

$$\rho_i = e^{H/N} - 1, \qquad \Delta_i = \eta_0 e^{-(i-1)H/N}(1 - e^{-H/N}).$$

Thus Δ_i is largest at $i = 1$, its sum is $\Gamma = \eta_0 - \eta_N$, and the accumulated discretization theorem requires exactly

$$e^{H/N} - 1 \le 1 \quad \Longleftrightarrow \quad \frac{H}{N} \le \log 2.$$

For fixed SNR endpoints, this holds whenever $N \ge H/\log 2$, independently of dimension. Theorem D.1 bounds the sum of marginal defects by $\mathcal{D}_N$; this is not a bound of $\mathcal{D}_N/N$ on each individual defect. Finally, $e^x - 1 \le xe^x$ for $x \ge 0$ gives the second inequality in (164). This proves Corollary F.1, including its step-size condition. For $0 \le u \le \Delta_*$,

$$e^{D^2 u} - 1 \le D^2 u e^{D^2 \Delta_*}.$$

Consequently, if the learned-denoiser diameters satisfy $D_i(x) \le D$ for p_i-almost every x and every i, then $\Psi_i \le \Omega_i \le (e^{D^2\Delta_i} - 1)W_i^2$. When also $E_i \le E$ and $W_i \le W$, the expression inside the square root in Theorem 3.2 is at most

$$\frac{\Gamma}{2}\left(E^2 + D^2 e^{D^2\Delta_*} W^2\right).$$

Taking square roots and using $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for $u, v \ge 0$ proves (18). For fixed SNR endpoints, Γ is fixed and $0 < \Delta_* \le \Gamma$, so these learning coefficients are bounded independently of N for fixed D. Under the output bound $\|y_\theta^l\| \le R$, one may take $D = 2R$; an absolute output bound is not needed for the diameter estimate itself.

For completeness, write $h_{i,l}(x) = \|y_\theta^l(x, t_i) - \bar{y}^l(x, t_i)\|^2$. Identity (6) gives

$$E_i^2 = \sum_l \pi_l \int h_{i,l}(x) p_i^l(x)\, dx.$$

Thus $\int h_{i,l} p_i^l \le \varepsilon_y^2$ for every l implies $E_i^2 \le \varepsilon_y^2 \sum_l \pi_l = \varepsilon_y^2$. If instead $\int h_{i,l} p_i \le \varepsilon_y^2$, then $0 \le a^l \le 1$ yields

$$E_i^2 = \sum_l \int a^l(x, t_i) h_{i,l}(x) p_i(x)\, dx \le \sum_l \int h_{i,l}(x) p_i(x)\, dx \le L\varepsilon_y^2.$$

Finally, Cauchy–Schwarz in the class index gives $\|a_\phi - a\|_1^2 \le L \sum_l |a_\phi^l - a^l|^2$. Therefore $\mathbb{E}_{p_i}|a_\phi^l - a^l|^2 \le \varepsilon_a^2$ for every class implies $W_i^2 \le L^2\varepsilon_a^2$. These calculations use the posterior weights inside each integral throughout.

For fixed η_0 and $\eta_N = N^{-2}$, $H = \log(\eta_0 N^2)$ and $H/N \to 0$. The step-size condition therefore holds for all sufficiently large N, independently of d. The explicit bound is

$$\mathcal{D}_N \le \frac{e^{H/N} H(9H+5)}{4} \frac{d}{N},$$

so the stated $O(d \log^2 N/N)$ discretization bound follows, with constants allowed to depend on the fixed η_0. If d varies with N, this displayed dimension factor must be retained when assessing convergence. If the early-stopping SNR η_0 also grows, its effects on H, Γ, and Δ_* must all be retained. In particular, the bounded learning coefficient on a fixed SNR interval is not a uniform bound as the early-stopping noise tends to zero.

Logarithmic factors and the bounded exponential. The exact finite-grid budget has no implicit N-dependence beyond H and the displayed exponential. For $0 \le h = H/N \le \log 2$, Taylor's integral remainder gives

$$0 \le e^h - 1 - h = \int_0^h (h-u)e^u\, \mathrm{d}u \le \tfrac{1}{2} e^h h^2 \le h^2, \qquad e^h \le 2.$$

Thus $\mathcal{D}_N \le dH(9H+5)/(2N)$. For endpoints $\eta_0 = c_0 N^a$ and $\eta_N = c_1 N^{-b}$, with fixed $c_0, c_1 > 0$, $a, b \ge 0$, and $a + b > 0$, one has $H = (a+b)\log N + \log(c_0/c_1)$ and the mesh condition holds for all sufficiently large N. Consequently, $e^{H/N} = 1 + O(\log N/N)$ and $\mathcal{D}_N = O(d \log^2 N/N) = \widetilde{O}(d/N)$, the convention used in Table 1. The logarithms arise from the growing interval factor $H(9H+5)$; $e^{H/N}$ stays bounded. Fixed H gives the special case $O(d/N)$ with $e^{H/N} = 1 + O(N^{-1})$. These are discretization upper bounds, not lower bounds or rates for the entire learned error without controlling the remaining terms. The routing coefficient $e^{D^2\Delta_*/2}$ is different: $\Delta_* = \eta_0(1 - e^{-H/N})$ need not stay bounded on moving endpoints, even when $e^{H/N} \le 2$.

When the bound is informative. The mesh condition $H/N \le \log 2$ ensures validity, but alone does not force a small error. A sufficient condition for the displayed discretization budget to vanish is $dH(H+1)/N \to 0$. Thus $\eta_0 \to \infty$ does not by itself make the theorem vacuous: for example, $\eta_0 = N^2$ and $\eta_N = N^{-2}$ give $H = 4\log N$ and vanishing discretization for fixed d. If the numerical upper bound exceeds one on a coarse grid, it can only certify the trivial TV bound there. Choosing a zero-noise endpoint at finite N is outside the theorem. Moreover, for general singular data, clean-data convergence must be interpreted in the weaker metric below. Proposition D.5 supplies a smaller finite-grid bound, and Proposition C.6 separates unavoidable grid dependence from potentially loose upper-bound factors.

F.4 A COMMON DISTANCE AND CONVERGENCE TO CLEAN DATA

We use the bounded-Lipschitz metric

$$d_{\mathrm{BL}}(P,Q) = \sup_{\|f\|_\infty \le 1,\, \mathrm{Lip}(f)\le 1} |\mathbb{E}_P f - \mathbb{E}_Q f|, \tag{165}$$

Corollary F.2 (Convergence to clean data)**.** *Under Assumption 2.1, for either sampler output q,*

$$d_{\mathrm{BL}}(p_d, q) \le \sqrt{(1-\alpha_0)^2 M + d\sigma_0^2} + 2\mathrm{TV}(p_0, q). \tag{166}$$

Fix d and p_d. For $N > 1$, take $\eta_{0,N} = N^2$, $\eta_{N,N} = N^{-2}$ and equal log-SNR spacing. If $4\log N/N \le \log 2$, exact denoisers and, for mixtures, exact routing give

$$\begin{aligned} d_{\mathrm{BL}}(p_d, q) \le{}& \frac{\sqrt{d}+\sqrt{M}}{N} + \frac{\sqrt{M}+\sqrt{d/2}}{N^2} \\ &+ 2de^{4\log N/N}\frac{\log N(36\log N + 5)}{N}. \end{aligned} \tag{167}$$

Thus the oracle rate is $O(\log^2 N/N)$ for fixed d, p_d, with no diameter assumption. Learned outputs also converge weakly if $\sum_i \Delta_i e_i^2 \to 0$ in the standard case, or $\sum_i \kappa_i \to 0$ in Corollary 5.1; $\sum_i(\Delta_i E_i^2 + \Psi_i) \to 0$ is a sufficient mixture condition. All quantities use the N-dependent grid.

Finite-grid bounded-Lipschitz bound. Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy $\|f\|_\infty \le 1$ and $\mathrm{Lip}(f) \le 1$, with the Euclidean norm defining the Lipschitz constant. The coupling used for (163) gives

$$|\mathbb{E}f(Y) - \mathbb{E}f(X_{t_0})| \le \mathbb{E}\|Y - X_{t_0}\| \le \sqrt{(1-\alpha_0)^2 M + d\sigma_0^2}.$$

For any probability measure q_0, the convention $\mathrm{TV}(P,Q) = \frac{1}{2}\|P-Q\|_1$ gives

$$\left|\int f\, dp_0 - \int f\, dq_0\right| \le 2\mathrm{TV}(p_0, q_0).$$

The triangle inequality and the supremum over these test functions prove (166). Applying either main theorem bounds the TV term by its initialization, discretization, and learning contributions. This argument does not require q_0 to have a finite second moment.

Vanishing error budgets with moving endpoints. Fix d and the clean-data law p_d. For each N, allow a different N-step grid and different learned predictors, subject to the corresponding theorem's assumptions. Write $p_{i,N}$, $q_{0,N}$, $\eta_{i,N}$, and similarly for the remaining grid-dependent quantities. For the standard sampler set

$$\mathcal{L}_N = \frac{1}{2}\left(\sum_{i=1}^N \Delta_{i,N} e_{i,N}^2\right)^{1/2},$$

and for the mixture sampler one may take

$$\mathcal{L}_N = \left[\frac{1}{2}\sum_{i=1}^N \{\Delta_{i,N} E_{i,N}^2 + \Psi_{i,N}\}\right]^{1/2}.$$

The sum of these logarithmic energies can be replaced by the sum of the stepwise minima $\kappa_{i,N}$ in Corollary 5.1. The upper bound $\Psi_{i,N} \le \Omega_{i,N}$ also makes the former exponential-energy condition sufficient, but it is not required by the logarithmic bound. Suppose the step-size condition holds for every sufficiently large N, $\sigma_{0,N} \to 0$, and

$$\mathcal{I}_N + \mathcal{D}_N + \mathcal{L}_N \longrightarrow 0.$$

Variance preservation and positive $\alpha_{0,N}$ imply $\alpha_{0,N} \to 1$, so (166) gives

$$d_{\mathrm{BL}}(p_d, q_{0,N}) \le \sqrt{(1-\alpha_{0,N})^2 M + d\sigma_{0,N}^2} + 2\mathcal{I}_N + 2\mathcal{D}_N + 2\mathcal{L}_N \longrightarrow 0. \tag{168}$$

The metric argument at the end of this appendix proves $q_{0,N} \Rightarrow p_d$. These are sufficient conditions for convergence; no necessity is asserted.

A family of moving-endpoint schedules. For fixed $a, b > 0$ and integer $N > 1$, choose

$$\eta_{0,N} = N^a, \qquad \eta_{N,N} = N^{-b}, \qquad H_N = (a+b)\log N, \qquad \eta_{i,N} = N^a e^{-iH_N/N}. \tag{169}$$

These coefficients are realized by the variance-preserving schedule $\alpha_t^2 = 1 - t$, $\sigma_t^2 = t$ and grid $t_{i,N} = 1/(1+\eta_{i,N})$. Every analyzed time lies strictly between zero and one. The step-size condition is exactly

$$\frac{(a+b)\log N}{N} \le \log 2,$$

which holds for all sufficiently large N, independently of d. The learning coefficients satisfy

$$\begin{aligned} \Gamma_N &= N^a - N^{-b} \sim N^a, \\ \Delta_{*,N} &= N^a\big(1 - e^{-(a+b)\log N/N}\big) \sim (a+b)N^{a-1}\log N. \end{aligned} \tag{170}$$

The asymptotics use $\log N/N \to 0$ and $(1-e^{-x})/x \to 1$ as $x \downarrow 0$.

The initialization bound above and $\sigma_{0,N}^2 = (1+N^a)^{-1}$ give

$$\begin{aligned} \mathcal{I}_N &\le \frac{\sqrt{M}}{2} N^{-b/2} + \frac{\sqrt{d}}{2\sqrt{2}} N^{-b}, \\ 1 - \alpha_{0,N} &= \frac{\sigma_{0,N}^2}{1+\alpha_{0,N}} \le N^{-a}, \\ \sqrt{(1-\alpha_{0,N})^2 M + d\sigma_{0,N}^2} &\le \sqrt{M}\, N^{-a} + \sqrt{d}\, N^{-a/2}, \\ \mathcal{D}_N &\le \frac{d e^{H_N/N} H_N (9H_N + 5)}{4N}. \end{aligned}$$

For exact denoisers and exact routing, every learning energy vanishes. No bound on oracle outputs or their diameters is needed: the finitely many canonical posterior means are finite at every input, and the routing discrepancy is pointwise zero. Thus there is no undefined product of infinity and zero in either routing energy. For the standard oracle, likewise $e_{i,N} = 0$. Applying (168) yields

$$\begin{aligned} d_{\mathrm{BL}}(p_d, q_{0,N}^{\mathrm{oracle}}) \le{}& \sqrt{d}\, N^{-a/2} + \sqrt{M}\, N^{-a} + \sqrt{M}\, N^{-b/2} + \sqrt{d/2}\, N^{-b} \\ &+ \frac{d e^{H_N/N} H_N (9H_N + 5)}{2N}. \end{aligned} \tag{171}$$

Here the oracle is the chain of oracle Gaussian or Gaussian-mixture kernels initialized from γ^d; it is not assumed to equal the exact reverse chain.

The faster oracle rate in Corollary F.2. Set $a = b = 2$ in (169). The exact step-size requirement becomes $4\log N/N \le \log 2$, and

$$\mathcal{I}_N \le \frac{\sqrt{M}}{2N} + \frac{\sqrt{d}}{2\sqrt{2}\, N^2}, \qquad \mathcal{D}_N \le d e^{4\log N/N} \frac{\log N (36\log N + 5)}{N}.$$

The smoothing term is at most $\sqrt{M}/N^2 + \sqrt{d}/N$. Consequently,

$$d_{\mathrm{BL}}(p_d, q_{0,N}^{\mathrm{oracle}}) \le \frac{\sqrt{d}+\sqrt{M}}{N} + \frac{\sqrt{M}+\sqrt{d/2}}{N^2} + 2de^{4\log N/N}\frac{\log N(36\log N+5)}{N}. \tag{172}$$

Since $e^{4\log N/N}$ is bounded for $N > 1$, this proves (167) and oracle weak convergence for every fixed finite-second-moment data law, including singular laws. The final term includes the factor of two converting TV to the bounded-Lipschitz convention in (165). The claim concerns this sufficient oracle rate; no optimality over schedules is asserted.

Learning conditions depend on the growing SNR interval. On any schedule (169), it suffices that $\sum_i\{\Delta_{i,N}E_{i,N}^2 + \Psi_{i,N}\} \to 0$ for the mixture sampler, or that $\sum_i \kappa_{i,N} \to 0$. These conditions retain the dependence between expert separation and routing error. The standard condition is $\sum_i \Delta_{i,N}e_{i,N}^2 \to 0$. For a more explicit sufficient condition, suppose $D_{i,N}(x) \le D_{\max,N} < \infty$ for $p_{i,N}$-almost every x and every i, and define $E_{\max,N} = \max_i E_{i,N}$ and $W_{\max,N} = \max_i W_{i,N}$. Corollary 3.3 gives

$$\mathcal{L}_N \le \sqrt{\Gamma_N/2}\left(E_{\max,N} + D_{\max,N}e^{D_{\max,N}^2\Delta_{*,N}/2}W_{\max,N}\right). \tag{173}$$

For fixed $D_{\max,N} = D$ and $0 < a < 1$, $\Delta_{*,N} \to 0$, so $E_{\max,N} + DW_{\max,N} = o(N^{-a/2})$ suffices. At $a = 1$, Taylor's formula gives $\Delta_{*,N} = (1+b)\log N + O(\log^2 N/N)$, so for $D > 0$ the exponential coefficient is asymptotic to $N^{D^2(1+b)/2}$. At $a > 1$ its logarithm grows as $D^2(a+b)N^{a-1}\log N/2$. Thus the linear-in-$W_{\max,N}$ conditions for $0 < a < 1$ cannot simply be transferred to the faster oracle choice $a = 2$.

The coupling bound in Proposition E.5 supplies a different sufficient condition for every fixed $a > 0$. If $D_{i,N} \le D$, then Cauchy–Schwarz gives

$$T_{i,N}^2 \le \frac{D^2}{2}\mathbb{E}_{p_{i,N}}\|a - a_\phi\|_1 \le \frac{D^2W_{\max,N}}{2}.$$

Since $\sum_i \Delta_{i,N} = \Gamma_N$, its learning contribution is bounded by

$$\mathcal{L}_N^{\mathrm{coup}} \le \frac{\sqrt{\Gamma_N}}{2}\left(E_{\max,N} + D\sqrt{W_{\max,N}/2}\right). \tag{174}$$

Hence $E_{\max,N} = o(N^{-a/2})$ and $W_{\max,N} = o(N^{-a})$ suffice for fixed D. In particular, on the $a = b = 2$ oracle schedule, $E_{\max,N} = o(N^{-1})$ and $W_{\max,N} = o(N^{-2})$ imply learned weak convergence. This bound has square-root dependence on $W_{\max,N}$ and imposes no exponential routing moment. For the standard sampler on any such schedule, $\max_i e_{i,N} = o(N^{-a/2})$ suffices. All these conditions concern population errors, not finite-sample or optimization guarantees.

Matching the oracle rate is stronger than consistency. On the $a = b = 2$ grid, write $r_N = \log^2 N/N$. The mixture learning term in Corollary 5.1 is $\mathcal{L}_N = (\frac{1}{2}\sum_i \kappa_{i,N})^{1/2}$. Thus $\sum_i \kappa_{i,N} = O(r_N^2)$ makes $\mathcal{L}_N = O(r_N)$ and preserves the oracle order in (168). Under a fixed diameter bound, (174) and $\sqrt{\Gamma_N} \asymp N$ show that

$$E_{\max,N} = O(\log^2 N/N^2), \qquad W_{\max,N} = O(\log^4 N/N^4) \tag{175}$$

are sufficient. Indeed, both $NE_{\max,N}$ and $ND\sqrt{W_{\max,N}/2}$ are $O(r_N)$. For standard DDPM, $\sum_i \Delta_{i,N}e_{i,N}^2 = O(r_N^2)$ is sufficient, in particular $\max_i e_{i,N} = O(\log^2 N/N^2)$. In contrast, the consistency conditions above only make these learning terms tend to zero: for example, with exact routing, $E_{\max,N} = 1/(N\log N)$ yields a sufficient upper bound of order $1/\log N$, which is slower than r_N. These conditions preserve the rate of the displayed bound; they are not necessary conditions on the actual output error, or a claim that training attains the stipulated population accuracy.

A milder schedule for learned experts with growing diameter. The choice $a = 1/2$, $b = 2$ provides a useful trade-off for learned samplers. It has

$$H_N = \tfrac{5}{2}\log N, \qquad \Gamma_N = \sqrt{N} - N^{-2} \le \sqrt{N}, \qquad \Delta_{*,N} \le \frac{5\log N}{2\sqrt{N}}.$$

Its oracle bounded-Lipschitz rate is $O(\sqrt{d}\,N^{-1/4} + \sqrt{M}\,N^{-1/2} + d\log^2 N/N)$, by (171). Its advantage for learned errors is the vanishing absolute SNR step. Specifically,

$$D_{\max,N}^2 \frac{\log N}{\sqrt{N}} = O(1), \qquad N^{1/4}(E_{\max,N} + D_{\max,N} W_{\max,N}) \longrightarrow 0 \tag{176}$$

imply $\mathcal{L}_N \to 0$ by (173). The symbol $D_{\max,N}$ denotes a diameter bound, while $\mathcal{D}_N$ denotes the discretization budget. Appendix I gives a projection construction with $D_{\max,N} = O(\sqrt{\log N})$ satisfying these conditions under an explicitly stronger tail assumption.

If the dimension or data law varies. The finite-grid comparison also applies to a clean law $p_d^{(N)}$ on $\mathbb{R}^{d_N}$ with second moment M_N. In that case retain

$$S_N^2 = (1-\alpha_{0,N})^2 M_N + d_N \sigma_{0,N}^2$$

and require $S_N \to 0$, vanishing initialization, discretization, and learning budgets, and the same dimension-independent step-size condition. On the milder schedule $a = 1/2$, $b = 2$, sufficient endpoint and dimension conditions are

$$\frac{M_N}{N} \to 0, \qquad \frac{d_N}{\sqrt{N}} \to 0.$$

Indeed, $S_N^2 \le M_N/N + d_N/\sqrt{N}$, and these conditions also make the initialization bound vanish. Moreover, $d_N \log^2 N/N = (d_N/\sqrt{N})(\log^2 N/\sqrt{N}) \to 0$, so the new discretization budget vanishes. The mesh condition $H_N/N \le \log 2$ holds for all sufficiently large N regardless of d_N. Together with a vanishing learning budget, the conclusion is a vanishing bounded-Lipschitz discrepancy between the two laws on each $\mathbb{R}^{d_N}$. Ordinary weak convergence to a fixed law is the fixed-dimension statement above; no common ambient space for changing dimensions is presumed.

F.5 The bounded-Lipschitz metric and weak convergence

Why vanishing bounded-Lipschitz distance implies weak convergence. We give a direct argument on the fixed space $\mathbb{R}^d$. Let μ_n, μ be probability measures with $\beta_n = d_{\mathrm{BL}}(\mu_n, \mu) \to 0$. For $r > 0$, the function $g_r(x) = \min\{1, (\|x\| - r)_+\}$ belongs to the defining test class. Hence

$$\mu_n(\{\|x\| > r+1\}) \le \int g_r \, d\mu_n \le \mu(\{\|x\| > r\}) + \beta_n.$$

Because closed balls increase to $\mathbb{R}^d$, their complements have arbitrarily small μ-mass. The display therefore controls the tails of all sufficiently large n. Enlarging the ball to handle the finitely many remaining μ_n shows that $\{\mu, \mu_1, \mu_2, \ldots\}$ is uniformly tight.

Fix a bounded continuous f and write $K = \|f\|_\infty$. The case $K = 0$ is immediate. Given $\epsilon, \delta > 0$, uniform tightness supplies a closed ball B whose complement has mass at most ϵ under every measure in this family. Uniform continuity of f on B supplies $h > 0$ such that $|f(x) - f(y)| \le \delta$ for $x, y \in B$ with $\|x - y\| < h$. Choose $A > 0$ with $Ah \ge 2K$, and define on all of $\mathbb{R}^d$

$$v(x) = \inf_{y \in B}\{f(y) + A\|x - y\|\}, \qquad \widetilde{f}(x) = \max\{-K, \min\{K, v(x)\}\}.$$

The triangle inequality shows that v, and hence $\widetilde{f}$, is A-Lipschitz. For $x \in B$, choosing $y = x$ gives $v(x) \le f(x)$. If $\|x - y\| < h$, the expression defining v is at least $f(x) - \delta$; if $\|x - y\| \ge h$, it is at least $-K + Ah \ge K \ge f(x)$. Thus $|\widetilde{f} - f| \le \delta$ on B and $\|\widetilde{f}\|_\infty \le K$ globally. Scaling $\widetilde{f}$ by $\max\{K, A\}$ gives

$$\left| \int f \, d\mu_n - \int f \, d\mu \right| \le 2\delta + 4K\epsilon + \max\{K, A\}\beta_n.$$

First let $n \to \infty$, with B, h, A fixed, and then let $\epsilon, \delta \downarrow 0$. This proves convergence against every bounded continuous f, namely $\mu_n \Rightarrow \mu$.

The converse and the metric property. Suppose now that $\mu_n \Rightarrow \mu$. For each fixed r, weak convergence gives $\int g_r \, d\mu_n \to \int g_r \, d\mu$. The same cutoff argument above, followed by enlarging the ball for finitely many initial indices, shows uniform tightness. Given $\epsilon, \delta > 0$, choose a closed ball B with complement of mass at most ϵ under all these measures, and cover B by finitely many balls of radius δ centered at points $x_1, \dots, x_m \in B$.

Let $\mathcal{F} = \{f : \|f\|_\infty \le 1, \mathrm{Lip}(f) \le 1\}$. Partition $[-1, 1]$ into finitely many intervals of length at most δ. Group functions in $\mathcal{F}$ by the intervals containing their values at $x_1, \dots, x_m$, and retain one representative $f^{(j)}$ from each nonempty group. This gives finitely many globally defined functions $f^{(1)}, \dots, f^{(K)} \in \mathcal{F}$. For any $f \in \mathcal{F}$, its representative satisfies $|f - f^{(j)}| \le 3\delta$ on B: at a point within distance δ of x_k, the two Lipschitz errors each contribute at most δ, and the values at x_k differ by at most δ. Outside B, the difference is at most 2. Therefore

$$d_{\mathrm{BL}}(\mu_n, \mu) \le \max_{1 \le j \le K} \left| \int f^{(j)} \, d\mu_n - \int f^{(j)} \, d\mu \right| + 6\delta + 4\epsilon.$$

The maximum tends to zero because it is over finitely many bounded continuous functions. Letting $\epsilon, \delta \downarrow 0$ proves the converse.

Nonnegativity, symmetry, finiteness, and the triangle inequality follow directly from the defining supremum. To verify separation, suppose $d_{\mathrm{BL}}(\mu, \nu) = 0$. For every nonempty closed set F and integer $k \ge 1$, the function $h_k(x) = (1 - k \, \mathrm{dist}(x, F))_+$ satisfies $h_k / k \in \mathcal{F}$. Thus $\int h_k \, d\mu = \int h_k \, d\nu$. Since $h_k \downarrow \mathbf{1}_F$, dominated convergence gives $\mu(F) = \nu(F)$, and agreement on closed sets implies agreement on all Borel sets. Hence $\mu = \nu$, completing the assertion that d_{BL} is a metric that metrizes weak convergence on $\mathbb{R}^d$.

G WHICH EXPERT-COMBINATION SAMPLER IS ANALYZED?

Several expert-combination rules use the same predictions and probabilities but define different transition laws. We distinguish them here to specify the scope of Theorem 3.2 and its relation to a simpler classwise analysis. Throughout this appendix the variance is the posterior bridge variance $v_i I_d$ from (3).

Drawing an expert at every step versus averaging predictions. At a fixed input x and time t_i, abbreviate $b_l = a^l_\phi(x, t_i)$, $h_l = y^l_\theta(x, t_i)$, and $\bar{h} = \sum_l b_l h_l$. The kernel analyzed in the mixture theorem and the kernel that averages predictions are, respectively,

$$\begin{aligned} P_i(x, \cdot) &= \sum_l b_l \mathcal{N}(A_i x + B_i h_l, v_i I_d), \\ P_i^{\mathrm{avg}}(x, \cdot) &= \mathcal{N}(A_i x + B_i \bar{h}, v_i I_d). \end{aligned} \tag{177}$$

Both have conditional mean $A_i x + B_i \bar{h}$. The conditional variance identity gives

$$\begin{aligned} \mathrm{Cov}_{P_i(x,\cdot)} &= v_i I_d + B_i^2 \sum_l b_l (h_l - \bar{h})(h_l - \bar{h})^T, \\ \mathrm{Cov}_{P_i^{\mathrm{avg}}(x,\cdot)} &= v_i I_d. \end{aligned} \tag{178}$$

Indeed, conditional on the selected label the covariance is $v_i I_d$, and the covariance of the conditional mean is exactly the second term. This term is positive semidefinite. Since $B_i > 0$, it vanishes exactly when all positive-weight predictions agree, in which case the two kernels coincide; otherwise their covariances, and hence their laws, differ. This identity provides no ranking of their sampling errors or empirical quality. With oracle predictions and true weights, the averaged prediction is $\sum_l a^l \bar{y}^l = \bar{y}$, so the averaged kernel is precisely Q_i^{std}. Learned averaging is likewise a standard DDPM with the effective denoiser $\sum_l a^l_\phi y^l_\theta$, whose error can be bounded using the same true-marginal quantities as the routed sampler.

Proposition G.1 (Stability of averaged predictions). *Under Assumption 2.1, grid (12), and the mixture setup, suppose $E_i < \infty$ and $D_i(X) \le D < \infty$ p_i-almost surely for each i. For $q_0^{\mathrm{avg}} = \gamma^d P_N^{\mathrm{avg}} \cdots P_1^{\mathrm{avg}}$,*

$$\mathrm{TV}(p_0, q_0^{\mathrm{avg}}) \le \mathcal{I}_N + \mathcal{D}_N + \frac{1}{2} \left[\sum_i \Delta_i (E_i + D W_i / 2)^2 \right]^{1/2}. \tag{179}$$

In particular, $E_i \le E$ and $W_i \le W$ give the learning term $\sqrt{\Gamma}(E + DW/2)/2$.

Proof. Fix an input and step, writing $m_l = \bar{y}^l$, $h_l = y^l_\theta$, $a_l = a^l$, and $b_l = a^l_\phi$. The tower identity gives

$$\sum_l b_l h_l - \bar{y} = \sum_l a_l(h_l - m_l) + \sum_l (b_l - a_l)h_l.$$

Jensen bounds the squared norm of the first sum by $\sum_l a_l \|h_l - m_l\|^2$. Set $z_l = b_l - a_l$ and $\zeta = \|b-a\|_1/2$. Since $\sum_l z_l = 0$, when $\zeta > 0$ the second sum is $\zeta(h_+ - h_-)$, where $h_+ = \sum_l (z_l)_+ h_l/\zeta$ and $h_- = \sum_l (-z_l)_+ h_l/\zeta$ are convex combinations of the predictions. Their distance is at most $D_i(x)$: express their difference as the average of all pairwise differences with the product of the two convex weights. When $\zeta = 0$ the sum vanishes. Thus

$$\left\| \sum_l (b_l - a_l)h_l \right\| \le D_i(x)\|b - a\|_1/2.$$

Minkowski in $L^2(p_i)$ now gives the effective network error $e_{\mathrm{avg},i} \le E_i + DW_i/2 < \infty$; here $W_i \le 2$. The effective denoiser is finite and Borel. Theorem 3.1 therefore proves (179). □

The same proof permits state-dependent diameters whenever $S_i^2 := \frac{1}{4}\mathbb{E}_{p_i}[D_i(X)^2 \|a - a_\phi\|_1^2] < \infty$, by replacing $DW_i/2$ with S_i. This uses no probability floor and no exponential coefficient. A smaller available upper bound does not rank the samplers' actual errors: averaging and categorical routing have different oracle kernels. In particular, the latter retains between-class posterior uncertainty, as quantified in Appendix L.

If the router is evaluated independently and experts require separate calls, categorical selection permits one router call and one selected expert call per step, whereas full averaging requires every positive-weight expert prediction. Shared computation, batching, and parallel execution can change the cost; no wall-clock advantage is asserted by these bounds.

Keeping one expert for the entire trajectory. Define

$$P_i^l(x, \cdot) = \mathcal{N}(A_i x + B_i y^l_\theta(x, t_i), v_i I_d), \qquad q_0^l = \gamma^d P_N^l \cdots P_1^l.$$

A sampler that first draws $J \sim (\pi_1, \dots, \pi_L)$, independently of its Gaussian initial state, and then retains this label has output $q_0^{\mathrm{fixed}} = \sum_l \pi_l q_0^l$. It admits the following direct bound.

Proposition G.2 (Classwise bound for a fixed expert)**.** *Under Assumption 2.1 and grid (12), let the finite class representation have $\pi_l > 0$ and assume $E_i < \infty$. Then*

$$\mathrm{TV}(p_0, q_0^{\mathrm{fixed}}) \le \sum_l \pi_l \,\mathrm{TV}(p_N^l, \gamma^d) + \mathcal{D}_N + \frac{1}{2}\left(\sum_i \Delta_i E_i^2\right)^{1/2}. \tag{180}$$

Proof. Each class has finite second moment because $\pi_l \mathbb{E}[\|Y\|^2 \mid J = l] \le M$ and $\pi_l > 0$. Set $e_{i,l}^2 = \mathbb{E}_{p_i^l}\|y^l_\theta - \bar{y}^l\|^2$. The class-joint identity (6) and Tonelli's theorem give $E_i^2 = \sum_l \pi_l e_{i,l}^2$; thus all $e_{i,l}$ are finite. Theorem 3.1 applied to p_d^l gives

$$\mathrm{TV}(p_0^l, q_0^l) \le \mathrm{TV}(p_N^l, \gamma^d) + \mathcal{D}_N + \frac{1}{2}\left(\sum_i \Delta_i e_{i,l}^2\right)^{1/2}.$$

The discretization budget is the same for every class because its constant depends on the dimension and grid, not the data law's second-moment value. Convexity of TV and $p_0 = \sum_l \pi_l p_0^l$ imply $\mathrm{TV}(p_0, q_0^{\mathrm{fixed}}) \le \sum_l \pi_l \,\mathrm{TV}(p_0^l, q_0^l)$. Finally, Cauchy–Schwarz for the probability weights gives

$$\sum_l \pi_l \left(\sum_i \Delta_i e_{i,l}^2\right)^{1/2} \le \left(\sum_l \pi_l \sum_i \Delta_i e_{i,l}^2\right)^{1/2} = \left(\sum_i \Delta_i E_i^2\right)^{1/2},$$

which proves (180). □

This classwise argument requires no routing network. Its initialization term is an average of classwise discrepancies, which can exceed $\mathrm{TV}(p_N, \gamma^d)$. Its generated law is generally different from $\gamma^d P_N \cdots P_1$: selecting a label once and retaining it does not equal redrawing a state-dependent label at each transition. The learned-routing theorem controls the latter chain and its true-marginal expert and routing errors; it does not assert that classwise convergence analysis is unavailable.

One-hot routing and other reverse dynamics. A deterministic rule $l_*(x, t_i)$, with a measurable convention for ties, is represented by $a^l_\phi(x, t_i) = \mathbf{1}_{\{l=l_*(x,t_i)\}}$. It is a special case of (7) when it uses the same Gaussian mean and posterior bridge variance. However, using the maximizer of the true posterior is generally not exact routing: for $L > 1$, canonical Gaussian Bayes probabilities satisfy $0 < a^l(x, t_i) < 1$ for every x, and

$$\|a(x, t_i) - a_\phi(x, t_i)\|_1 = 2\{1 - a^{l_*(x,t_i)}(x, t_i)\} > 0.$$

Thus exact knowledge of the maximizing label does not set the routing-error budget to zero. Coincident expert predictions can still remove the kernel's sensitivity to that error through $D_i(x) = 0$. Changing the variance, using a probability-flow ODE, or choosing a different numerical update changes the transition law. Such samplers require their own comparison; this appendix makes no convergence or quality claim for them.

H Population objectives and sampling errors

This appendix identifies the population denoising and routing targets and connects excess risks to the sampling-error budgets. The results concern the joint probability model in Section 2; they require no empirical performance claim.

H.1 Routing and denoising objectives

For an empirical data law $p_d = \sum_{j=1}^n \beta_j \delta_{y_j}$, where $\beta_j > 0$ and $\sum_j \beta_j = 1$, let L_j be the one-hot label of y_j. At a positive noise level, Bayes' rule gives

$$w_j(x, t) = \frac{\beta_j \exp\big(-\|x - \alpha_t y_j\|^2/(2\sigma_t^2)\big)}{\sum_{k=1}^n \beta_k \exp(-\|x - \alpha_t y_k\|^2/(2\sigma_t^2))}, \qquad a(x, t) = \sum_{j=1}^n w_j(x, t) L_j. \tag{181}$$

Uniform empirical weights correspond to $\beta_j = 1/n$. More generally, for the joint data-label law (Y, J) in Section 2, write $U = e_J$ for the one-hot class vector. Then $a(X_t, t) = \mathbb{E}[U \mid X_t, t]$. For the deterministic partitions used in the algorithms, $U = L(Y) = e_{c(Y)}$.

Proposition H.1 (Population routing objectives)**.** *Let $\gamma_d = \mathcal{N}(0, I_d)$ and draw (t, Y, J, Z) from the product law*

$$\mathbb{P}_{\mathrm{train}}(dt, dy, \{l\}, dz) = \tau(dt)\, \pi_l p_d^l(dy)\, \gamma_d(dz). \tag{182}$$

Thus the time, the data-label pair, and the Gaussian noise are mutually independent. Suppose α_t, σ_t are positive measurable functions on the support of τ, with $\alpha_t^2 + \sigma_t^2 = 1$, and set $X_t = \alpha_t Y + \sigma_t Z$. For measurable probability-vector predictions $b(x, t)$, define

$$\begin{aligned} \mathcal{R}_2(b) &= \mathbb{E}\|U - b(X_t, t)\|^2, \\ \mathcal{R}_{\mathrm{CE}}(b) &= \mathbb{E}\left[-\sum_{l=1}^L U_l \log b_l(X_t, t)\right]. \end{aligned} \tag{183}$$

Let $\Lambda(dt, dx) = \tau(dt) p_t(x)\, dx$ be the law of (t, X_t). The exact excess-risk identities are

$$\begin{aligned} \mathcal{R}_2(b) - \mathcal{R}_2(a) &= \mathbb{E}_\Lambda \|b - a\|_2^2, \\ \mathcal{R}_{\mathrm{CE}}(b) - \mathcal{R}_{\mathrm{CE}}(a) &= \mathbb{E}_\Lambda\, \mathrm{KL}(a\|b). \end{aligned} \tag{184}$$

Hence, over all such measurable predictions, each objective is minimized exactly by $b = a$ up to Λ-null sets. The second equality allows the value $+\infty$.

Proof. The product law (182) and Gaussian convolution give

$$\mathbb{P}(t \in dt, X_t \in dx, J = l) = \tau(dt)\, \pi_l p_t^l(x)\, dx.$$

Summing over l proves that Λ is the stated joint law; Bayes' rule identifies $\mathbb{E}[U \mid X_t = x, t]$ with the canonical router $a^l(x,t) = \pi_l p_t^l(x)/p_t(x)$ from Section 2. Condition on $(X_t, t) = (x, t)$ and write $a = \mathbb{E}[U \mid x, t]$. Expanding the square gives

$$\mathbb{E}[\|U - b\|^2 \mid x, t] = \mathbb{E}[\|U - a\|^2 \mid x, t] + \|a - b\|^2,$$

because the cross term contains $\mathbb{E}[U - a \mid x, t] = 0$. Conditional cross-entropy is

$$-\sum_l a^l \log b_l = H(a) + \mathrm{KL}(a\|b), \qquad H(a) = -\sum_l a^l \log a^l.$$

Both conditional losses are uniquely minimized by $b = a$. We use $0 \log 0 = 0$ and $0 \log(0/b_l) = 0$, including $b_l = 0$; a positive a^l with $b_l = 0$ gives infinite KL and cross-entropy. Averaging proves (184): the optimal risks are finite because $\mathcal{R}_2(a) \le 2$ and $0 \le H(a) \le \log L$, while the excess terms are nonnegative. Zero excess implies equality of the predictors Λ-almost everywhere, rather than at unobserved time-state pairs. □

Network error and the theorem's error budgets. For the learned router, we apply Proposition H.1 with $b = a_\phi$. For a collection $h = (h_l)_{l=1}^L$ of measurable denoisers, define $\mathcal{R}_y(h) = \mathbb{E}\|Y - h_J(X_t, t)\|^2$ and let $\bar{y}^l(x,t) = \mathbb{E}[Y \mid X_t = x, t, J = l]$. Write $\bar{\boldsymbol{y}} = (\bar{y}^l)_{l=1}^L$ and $\boldsymbol{y}_\theta = (y_\theta^l)_{l=1}^L$ for the oracle and learned collections. Finite second moment gives $\mathcal{R}_y(\bar{\boldsymbol{y}}) \le M$. For any h of finite risk, both $h_J(X_t, t)$ and $\bar{y}^J(X_t, t)$ are square-integrable. Conditional orthogonality therefore gives

$$\mathcal{R}_y(h) - \mathcal{R}_y(\bar{\boldsymbol{y}}) = \mathbb{E}\|h_J(X_t,t) - \bar{y}^J(X_t,t)\|^2 = \int \sum_l \pi_l \mathbb{E}_{p_t^l}\|h_l - \bar{y}^l\|^2\, \tau(dt). \tag{185}$$

This identifies the population target over unrestricted measurable denoisers; it is not a finite-sample or optimization guarantee for a neural network.

Let $\mathcal{T}_{\mathrm{in}} = \{t_1, \dots, t_N\}$ be the input times of the N reverse updates; t_0 is their final output time. At these input times, use the same $\alpha_{t_i}, \sigma_{t_i}$ in training and sampling, so that the training conditional laws are the theorem's p_i and p_i^l. If τ assigns mass $\tau_i = \tau(\{t_i\}) > 0$ to each $t_i \in \mathcal{T}_{\mathrm{in}}$, nonnegativity of the integrands in (184)–(185) gives

$$E_i^2 \le \frac{\mathcal{R}_y(\boldsymbol{y}_\theta) - \mathcal{R}_y(\bar{\boldsymbol{y}})}{\tau_i},$$
$$W_i^2 \le \frac{L[\mathcal{R}_2(a_\phi) - \mathcal{R}_2(a)]}{\tau_i}, \qquad W_i^2 \le \frac{2[\mathcal{R}_{\mathrm{CE}}(a_\phi) - \mathcal{R}_{\mathrm{CE}}(a)]}{\tau_i}. \tag{186}$$

The two routing bounds use $\|a_\phi - a\|_1^2 \le L\|a_\phi - a\|_2^2$ and Pinsker's inequality, respectively. These are worst-case conversions with an explicit dependence on training-time mass, not uniform per-time error bounds obtained for free from an averaged loss. A continuous training-time law, such as $\mathrm{Uniform}(0,1)$, assigns zero mass to every finite grid. A population minimizer can then be modified at the sampling times without changing its loss. Such a loss alone does not imply the sampling-grid assumptions of the main theorems; those assumptions must be established separately.

H.2 From population risks to sampling energy

Corollary H.2 (Population risks control the full sampling energy)**.** *Use the mixture setup and grid (12), with $D_i \le D < \infty$ p_i-almost everywhere. Suppose the training coefficients agree with the sampling coefficients on $\mathcal{T}_{\mathrm{in}}$ and the training law gives mass $\tau_i = \tau(\{t_i\}) > 0$ to every input time $i = 1, \dots, N$. Define*

$$\mathcal{X}_y = \mathcal{R}_y(\boldsymbol{y}_\theta) - \mathcal{R}_y(\bar{\boldsymbol{y}}) < \infty, \quad \mathcal{X}_2 = \mathcal{R}_2(a_\phi) - \mathcal{R}_2(a), \quad \mathcal{X}_{\mathrm{CE}} = \mathcal{R}_{\mathrm{CE}}(a_\phi) - \mathcal{R}_{\mathrm{CE}}(a),$$

and $C_\tau = \max_i \Delta_i/\tau_i$, $\Delta_* = \max_i \Delta_i$. Then

$$\mathrm{TV}(p_0, q_0) \le \mathcal{I}_N + \mathcal{D}_N + \sqrt{\frac{C_\tau}{2}\left[\mathcal{X}_y + D^2 e^{D^2\Delta_*} \min\{L\mathcal{X}_2, 2\mathcal{X}_{\mathrm{CE}}\}\right]}. \tag{187}$$

The standard sampler has the corresponding bound $\mathrm{TV}(p_0, q_0^{\mathrm{std}}) \le \mathcal{I}_N + \mathcal{D}_N + \frac{1}{2}\sqrt{C_\tau \mathcal{X}_y^{\mathrm{std}}}$, where $\mathcal{X}_y^{\mathrm{std}} < \infty$ is the standard denoiser's population excess squared loss under the same training law.

Proof. The excess-risk identities and nonnegativity give

$$\sum_i \tau_i E_i^2 \le \mathcal{X}_y, \qquad \sum_i \tau_i W_i^2 \le \min\{L\mathcal{X}_2, 2\mathcal{X}_{\mathrm{CE}}\}.$$

Each inequality follows by restricting the training-time integral to its sampling-grid atoms; in the second inequality use the pointwise norm comparison or Pinsker before summing. Since $\Delta_i \le C_\tau \tau_i$ and $e^{D^2\Delta_i} - 1 \le D^2\Delta_i e^{D^2\Delta_*}$,

$$\sum_i \{\Delta_i E_i^2 + (e^{D^2\Delta_i} - 1)W_i^2\} \le C_\tau \left[\mathcal{X}_y + D^2 e^{D^2\Delta_*} \min\{L\mathcal{X}_2, 2\mathcal{X}_{\mathrm{CE}}\}\right].$$

Insert this into Corollary 3.3. The standard case follows from conditional orthogonality with $L = 1$ and Theorem 3.1. □

This conversion controls the sum of learning energies directly, using $\Psi_i \le \Omega_i$ in the logarithmic theorem. A concrete matching design samples input time t_i with probability Δ_i/Γ and uses its forward-noising coefficients; the risk identities then have no sampling-to-training penalty beyond Γ. Over all training laws with positive mass on $\mathcal{T}_{\mathrm{in}}$, $\Gamma = \sum_i \Delta_i \le C_\tau \sum_i \tau_i \le C_\tau$. Equality $C_\tau = \Gamma$ holds exactly when all training mass is on $\mathcal{T}_{\mathrm{in}}$ and $\tau_i = \Delta_i/\Gamma$: equality in the finite sum forces $\Delta_i = C_\tau \tau_i$ for every i, and equality in the last inequality forces $\sum_i \tau_i = 1$. On an equal-log-SNR grid, uniform training on the N input times gives $C_\tau = N\Delta_* \le \eta_0 H$, using $1 - e^{-H/N} \le H/N$. Uniform training on the full $N+1$ node grid $\mathcal{D} = \{t_0, \dots, t_N\}$ instead gives $C_\tau = (N+1)\Delta_* \le (1+1/N)\eta_0 H$. Both conversions remain bounded independently of N for fixed endpoints. These statements describe how a specified population objective controls the theorem's budgets; they do not change or certify the training law used for the reported experiments.

Interpreting classwise training errors. Equation (6) implies $E_i^2 = \sum_l \pi_l \mathbb{E}_{p_i^l} \|y_\theta^l - \bar{y}^l\|^2$. Uniform class-conditional error ε_y^2 therefore gives $E_i \le \varepsilon_y$. If instead each expert is controlled under the global marginal p_i, then the bound guaranteed by $\mathbb{E}_{p_i}\|y_\theta^l - \bar{y}^l\|^2 \le \varepsilon_y^2$ is $E_i \le \sqrt{L}\,\varepsilon_y$. Similarly, $\mathbb{E}_{p_i}|a_\phi^l - a^l|^2 \le \varepsilon_a^2$ for every class gives $W_i \le L\varepsilon_a$. These are explicit conversions between error conventions, not assumptions that the class weights are constant in x. Corollary H.2 bounds the learning term by population excess risks through $C_\tau = \max_i \Delta_i/\tau_i$. Training on the actual sampling input times with masses $\tau_i = \Delta_i/\Gamma$ gives $C_\tau = \Gamma$. This is a specified population design; an atomless time-averaged loss alone does not control a fixed sampling grid.

Network error when the stopping noise shrinks. The factor $\Gamma = \eta_0 - \eta_N$ in a bound based on a uniform x-prediction error can grow rapidly. The exact weighted network energy admits a more informative conversion. At a positive-noise node define the learned score and noise prediction by

$$s_\theta(x, t_i) = \frac{\alpha_i y_\theta(x, t_i) - x}{\sigma_i^2}, \qquad \epsilon_\theta(x, t_i) = \frac{x - \alpha_i y_\theta(x, t_i)}{\sigma_i}.$$

Let $e_{s,i}^2 = \mathbb{E}_{p_i}\|s_\theta - \nabla \log p_i\|^2$ and $e_{\epsilon,i}^2 = \mathbb{E}\|\epsilon_\theta(X_{t_i}, t_i) - \mathbb{E}[Z \mid X_{t_i}]\|^2$, where $X_{t_i} = \alpha_i Y + \sigma_i Z$ in this marginal representation. On the equal log-SNR grid, $\Delta_i = \rho\alpha_i^2/\sigma_i^2$, so the exact identities are

$$\Delta_i e_i^2 = \rho\sigma_i^2 e_{s,i}^2 = \rho e_{\epsilon,i}^2, \qquad \varepsilon_{\mathrm{net,std}}^2 = \rho \sum_i \sigma_i^2 e_{s,i}^2 = \rho \sum_i e_{\epsilon,i}^2. \tag{188}$$

They follow by subtracting the corresponding posterior-mean identities; no bound on a predictor derivative is used. The class-weighted versions hold by applying the same conversion to each expert and averaging with $p_i a^l = \pi_l p_i^l$.

For example, if $e_{s,i} \le e_s$, put $u_i = \alpha_i^2$ and $h = H/N$. The function $f(s) = 1/(1+e^s)$ is decreasing; its left-rectangle error on each log-SNR interval is at most the interval width times the decrease in f. These errors telescope, giving

$$h \sum_i \sigma_i^2 \le \int_{\log \eta_N}^{\log \eta_0} \frac{\mathrm{d}s}{1+e^s} + h = \log(u_0/u_N) + h.$$

Consequently,

$$\varepsilon^2_{\mathrm{net,std}} \le e_s^2 \frac{\rho}{h}\big(\log(u_0/u_N) + h\big), \qquad \frac{\rho}{h} \le e^h \le 2. \tag{189}$$

Since $u_0 \le 1$, this coefficient has no divergence from $\eta_0 \to \infty$ at fixed η_N under the mesh condition. For $\eta_N = N^{-b}$ it grows at most as $O(\log N)$. These are alternative sufficient population-error controls, not a claim that trained score or noise predictors attain them. A fixed nonzero error budget still does not establish learned convergence.

I When the routing and learning conditions hold

The finite-second-moment assumption controls the data and the canonical posterior means. A bound on the diameter of learned experts is a separate, sufficient condition for the simpler learning estimate. This appendix gives concrete cases that distinguish these requirements and a projection construction that is compatible with moving endpoints. None of the constructions assumes that a particular training algorithm achieves the stated population errors.

I.1 Which conditions belong to which result?

The main theorems separate assumptions on the data, choices of the sampler, and conditions on learned errors. Table 3 identifies their scope. Bounded support in intermediate proofs is removed by the truncation argument in Appendix D; it is not a standing assumption. Neither score smoothness nor a fourth data moment is required.

Table 3: Conditions and scope. The last three rows concern specializations, not additional assumptions of both main theorems.

Result / role	Requirement	Applicability and limitation
Main data setting	$M < \infty$; fixed finite class law with positive priors	Includes empirical, singular, bounded, and finite-variance Gaussian-mixture data; excludes infinite-variance laws. No class separation or prior floor is needed.
Main sampler	Gaussian VP Markov noising; posterior bridge variance; positive stopping noise; per-step categorical routing	A different variance, deterministic final update, averaged prediction, or fixed class changes the kernel; see Appendix G.
Main discretization rate	Equal log-SNR spacing, $e^{H/N} - 1 \le 1$	$O(d \log^2 N/N)$ for $H = O(\log N)$; fixed H gives $O(d/N)$. A generic time or noise schedule need not satisfy this grid condition.
Main learned bound	Finite true-marginal network errors and $\sum_i \Psi_i < \infty$	Linear-growth experts suffice for finiteness. Small population errors are substantive requirements, not guaranteed by empirical loss or finite parameters alone.
Simple routing bound	Uniform learned expert diameter $D_i \le D$	Optional for Corollary 3.3. Even Gaussian oracle experts can violate it. Finiteness of Ω_i is not required when Ψ_i is finite.
Training-risk conversion	Matching population class/noising law and positive training mass at each sampling node	Sufficient for Corollary H.2; neither empirical risk nor an atomless time average alone certifies nodewise errors.
Quantitative projection example	An exponential norm-square data moment and the stated improving errors	Gives an explicit projection-bias rate; this stronger tail condition does not follow from $M < \infty$.

Endpoints and learned convergence. The TV target is the positive-noise law p_0. If the clean law is supported on a Lebesgue-null set, a positive-variance sampler output has TV distance one from it. Clean-data convergence therefore uses bounded-Lipschitz distance, shrinking stopping noise, and vanishing initialization, discretization, and combined network/routing budgets (Corollary F.2).

Table 4: Examples of the stability conditions. Here D_i is the diameter of the learned experts, E_i is their class-weighted error, and the router is kept fixed when experts are projected. The quantitative Gaussian-tail condition in the final row is additional to Assumption 2.1.

Setting	Expert geometry	Applicable conclusion
$\|Y\| \le R$; learned experts projected onto the radius-R ball	$D_i \le 2R$; projection cannot increase E_i	Bounded-diameter bound with linear routing-error dependence
Unequal-variance Gaussian classes; oracle experts	Unbounded diameter, but $\mathbb{E}_{p_i} D_i^2 < \infty$	Finite logarithmic routing energy for every router; no uniform diameter bound
Exact routing; arbitrary finite learned experts	No diameter moment required; $E_i < \infty$ is sufficient	$\Psi_i = T_i = 0$; coupling gives the standard network error coefficient
$\mathbb{E}e^{\|Y\|^2/K^2} \le C$; projection radius $R_N = K\sqrt{q \log N}$	$D_{\max,N} = 2R_N$; projection bias is $O(N^{-q/2})$	Learned clean-data convergence under the errors specified below, for $q > 1/2$

Fixed endpoints or fixed nonzero network errors do not by themselves give this limit. The stipulated accuracy rates on moving grids are sufficient population conditions, not learning guarantees; their constants must also be tracked if the dimension, data law, or networks vary.

Training time is a separate issue. The main theorems impose errors at the actual sampling nodes, without specifying a training algorithm. Positive training mass is sufficient for the discrete conversion proved here, not necessary for every possible training-to-sampling argument. Even continuity in time alone gives no quantitative conversion from an averaged loss: a continuous bump of height one and width δ around a sampling node has time-integrated squared size $O(\delta)$ under a bounded time density, but unchanged value at that node. A separate uniform temporal regularity bound could permit another conversion. Class-balanced training must also be matched or reweighted to the class priors used in the theorem.

I.2 A SUFFICIENT INTEGRABILITY CONDITION FOR NETWORKS

Proposition I.1 (Linear-growth predictions)**.** *Under Assumption 2.1, suppose that at every sampling input time there is $c_i < \infty$ such that the finite Borel experts satisfy*

$$\|y_\theta^l(x, t_i)\| \le c_i(1 + \|x\|) \qquad \text{for every } x, l.$$

Then $E_i < \infty$ and $\mathbb{E}_{p_i} D_i^2 < \infty$. In particular, $\Psi_i < \infty$ for every Borel simplex-valued router. The same condition on the standard predictor gives $e_i < \infty$.

Proof. Put $m_i = \mathbb{E}\|X_{t_i}\|^2 = \alpha_i^2 M + d\sigma_i^2 < \infty$. Conditional Jensen gives $\sum_l \pi_l \mathbb{E}_{p_i^l} \|\bar{y}^l\|^2 \le M$. The growth bound and $(1+u)^2 \le 2(1+u^2)$ imply

$$E_i^2 \le 4c_i^2(1 + m_i) + 2M, \qquad \mathbb{E}_{p_i} D_i^2 \le 8c_i^2(1 + m_i).$$

The first inequality uses $p_i a^l = \pi_l p_i^l$ and $\|u - v\|^2 \le 2\|u\|^2 + 2\|v\|^2$; the second uses $D_i(x) \le 2c_i(1 + \|x\|)$. For $u \ge 0$ and $0 \le z \le 4$, $\log(1 + (e^u - 1)z) \le 4u$: indeed $1 + 4(e^u - 1) \le e^{4u}$, since both sides agree at zero and the derivative of the right side is no smaller. As $\|a_\phi - a\|_1^2 \le 4$, this yields $\Psi_i \le 4\Delta_i \mathbb{E}_{p_i} D_i^2 < \infty$. Taking a single class proves the standard claim. □

A globally Lipschitz predictor with finite value at zero has this growth property; finite compositions of affine maps and Lipschitz activations are examples. This is a sufficient condition, not an additional hypothesis of the main theorems or a statement about every architecture. It establishes finiteness on each fixed grid, not small error or uniform control as the endpoints and networks change. Conversely, finite-valued Borel predictions can grow too rapidly for their squared errors to be integrable under Gaussian inputs.

I.3 Unbounded Gaussian experts and exact routing

Consider one-dimensional classes $Y \mid J = l \sim \mathcal{N}(0, \tau_l^2)$ with $\tau_l > 0$, and a positive-noise time with $\alpha, \sigma > 0$. Joint Gaussian conditioning, using $\mathrm{Cov}(Y, X \mid J = l) = \alpha\tau_l^2$ and $\mathrm{Var}(X \mid J = l) = \alpha^2\tau_l^2 + \sigma^2$, gives

$$\bar{y}^l(x, t) = \frac{\alpha\tau_l^2}{\alpha^2\tau_l^2 + \sigma^2} x.$$

The coefficient is strictly increasing in the class variance, since

$$\frac{\mathrm{d}}{\mathrm{d}u} \frac{\alpha u}{\alpha^2 u + \sigma^2} = \frac{\alpha\sigma^2}{(\alpha^2 u + \sigma^2)^2} > 0.$$

Two unequal variances therefore produce an oracle diameter proportional to $|x|$, which is unbounded. Nevertheless, its second moment under the Gaussian-mixture input law is finite. For these oracle experts, the bound $\Psi_i \le 4\Delta_i \mathbb{E}_{p_i} D_i^2$ makes the logarithmic energy finite for every measurable probability-vector router. Thus a bounded diameter is not needed even in a simple nontrivial example with routing error.

More generally, exact routing makes $\|a - a_\phi\|_1 = 0$ pointwise, so $\Psi_i = T_i = 0$ for any finite learned expert values, regardless of their diameter moments. If $E_i < \infty$, the coupling bound then gives network error contribution $\frac{1}{2}\sqrt{\sum_i \Delta_i E_i^2}$. Conversely, finite M and finite class-weighted errors E_i do not imply $\mathbb{E}_{p_i} D_i^2 < \infty$: the tail construction in Proposition E.8 has finite network error and infinite diameter second moment. The diameter moment is a convenient sufficient condition for finite Ψ_i, not a consequence of the standing assumptions or a necessary condition for the theorem.

I.4 Projection and its population error

Let Π_R be Euclidean projection onto the closed ball $\{y : \|y\| \le R\}$, where $R \ge 0$. It is continuous and nonexpansive, so projecting Borel experts preserves measurability.

Proposition I.2 (Projection controls expert geometry)**.** *At each input time t_i, let h_i^l be raw finite Borel predictions with class-weighted error*

$$\widetilde{E}_i^2 = \sum_l \pi_l \mathbb{E}_{p_i^l} \|h_i^l(X) - \bar{y}^l(X, t_i)\|^2 < \infty.$$

Use the projected experts $y_\theta^l(x, t_i) = \Pi_R h_i^l(x)$ and retain the same router. Then $D_i(x) \le 2R$ for every x, and

$$E_i \le \widetilde{E}_i + B_{i,R}, \qquad B_{i,R}^2 := \sum_l \pi_l \mathbb{E}_{p_i^l} (\|\bar{y}^l(X, t_i)\| - R)_+^2 \le \mathbb{E}(\|Y\| - R)_+^2. \tag{190}$$

In particular, if $\|Y\| \le R$ almost surely, projection does not increase E_i. If, in addition to finite second moment, there are constants $K > 0$ and $C < \infty$ such that $\mathbb{E}\exp\big(\|Y\|^2/K^2\big) \le C$, then uniformly in the grid and its input times,

$$B_{i,R} \le K\sqrt{C/e}\, e^{-R^2/(2K^2)}, \qquad E_i \le \widetilde{E}_i + K\sqrt{C/e}\, e^{-R^2/(2K^2)}. \tag{191}$$

Proof. For $m = \bar{y}^l(x, t_i)$, nonexpansiveness gives

$$\|\Pi_R h_i^l(x) - m\| \le \|\Pi_R h_i^l(x) - \Pi_R m\| + \|\Pi_R m - m\| \le \|h_i^l(x) - m\| + (\|m\| - R)_+.$$

Taking the L^2 norm under the joint law of (J, X_{t_i}) and applying Minkowski's inequality proves the first inequality in (190). The function $y \mapsto (\|y\| - R)_+^2$ is convex, since it is a convex nondecreasing function of the norm. Conditional Jensen and $\bar{y}^J(X_{t_i}, t_i) = \mathbb{E}[Y \mid J, X_{t_i}]$ therefore give

$$B_{i,R}^2 = \mathbb{E}(\|\mathbb{E}[Y \mid J, X_{t_i}]\| - R)_+^2 \le \mathbb{E}(\|Y\| - R)_+^2.$$

The right-hand side is at most M, so the norms used in Minkowski's inequality are finite. Both projected expert outputs lie in the same radius-R ball, proving $D_i \le 2R$.

When $\|Y\| \le R$ almost surely, every canonical class posterior mean lies in this ball. Indeed, each positive-prior class law has this support, and its posterior at every input is a reweighting of that law

by a positive Gaussian density. Thus $\Pi_R m = m$ and the nonexpansive inequality proves pointwise nonincrease of the error, in addition to the aggregate conclusion $B_{i,R} = 0$.

For the quantitative bound, for $u > R$ put $z = u - R$. Since $u^2 - R^2 = z^2 + 2Rz \geq z^2$ and $\sup_{z \geq 0} z^2 e^{-z^2/K^2} = K^2/e$,

$$(u - R)_+^2 \leq \frac{K^2}{e} \exp\left(\frac{u^2 - R^2}{K^2}\right).$$

For $0 \leq u \leq R$ the same inequality is immediate. Substitute $u = \|Y\|$ into the last inequality in (190) and apply the assumed exponential moment. This gives $B_{i,R}^2 \leq (K^2 C/e) e^{-R^2/K^2}$ and proves (191). □

Under finite second moment alone, dominated convergence gives $\mathbb{E}(\|Y\| - R)_+^2 \to 0$ as $R \to \infty$ for each fixed data law. There is no rate uniform over all laws with second moment $M > 0$: put mass M/A^2 at a point of norm $A > \max\{R, \sqrt{M}\}$ and the remaining mass at zero. Its second moment is M, while its squared truncation residual is $M(1 - R/A)^2$, arbitrarily close to M as $A \to \infty$. The exponential-tail condition in Proposition I.2 is therefore an explicit additional assumption for the quantitative example below.

I.5 A GROWING PROJECTION RADIUS COMPATIBLE WITH LEARNED CONVERGENCE

Fix a data law satisfying the exponential moment in Proposition I.2, and use the milder endpoint schedule $\eta_{0,N} = N^{1/2}$, $\eta_{N,N} = N^{-2}$. Let

$$R_N = K\sqrt{q \log N}, \qquad q > 1/2, \qquad D_{\max,N} = 2R_N,$$

with q a fixed exponent. Write $\widetilde{E}_{\max,N} = \max_i \widetilde{E}_{i,N}$, $E_{\max,N} = \max_i E_{i,N}$ and $W_{\max,N} = \max_i W_{i,N}$. Projection gives

$$E_{\max,N} \leq \widetilde{E}_{\max,N} + K\sqrt{C/e}\, N^{-q/2}, \qquad D_{\max,N}^2 \frac{\log N}{\sqrt{N}} = 4K^2 q \frac{\log^2 N}{\sqrt{N}} \longrightarrow 0.$$

Consequently the population conditions

$$\widetilde{E}_{\max,N} = o(N^{-1/4}), \qquad W_{\max,N} = o\left(\frac{N^{-1/4}}{\sqrt{\log N}}\right)$$

imply $N^{1/4}(E_{\max,N} + D_{\max,N} W_{\max,N}) \to 0$, because $1/4 - q/2 < 0$. Equation (176) then gives a vanishing learning budget, and hence learned weak convergence for fixed d and p_d. A fixed finite mixture of Gaussian data laws satisfies the required exponential moment for sufficiently large K; the construction therefore includes unbounded data. It establishes compatibility of explicit sufficient assumptions, rather than the statistical attainability of the raw-error or routing-error rates.

Training-time alignment is a separate requirement. One population design places mass $\tau_i = \Delta_i/\Gamma$ at each actual sampling input time t_i, with matching noising coefficients and class priors. Corollary H.2 then has $C_\tau = \Gamma$. This connects the assumed errors to a specified training objective; an atomless time-averaged loss alone does not control values at finitely many sampling nodes. These examples make no claim about the training configuration of the retained empirical results.

J SUPPLEMENTARY NUMERICAL ILLUSTRATIONS

The following experiments illustrate the behavior of the mixture-of-Gaussians sampler, MoG-DPM, on synthetic and image datasets. We compare unconditional generation, a class sampled once at initialization, and a class redrawn at each reverse step. All reported values and generated images are retained from the existing experiments.

Scope and limitations. These results complement the theoretical analysis with numerical examples. They measure MMD or FID, not the total variation or bounded-Lipschitz distances in the theorems, and do not establish the predicted convergence rates. The implementation uses one shared conditional denoiser rather than training across separate institutions. The tables do not certify the global network error or routing error assumptions, isolate the classifier loss or expert-diameter effects, or establish statistical significance of differences. Relative performance depends on the schedule, base sampler, and step count. Results for ODE integrators and alternative DDPM variances are included as empirical comparisons; the posterior-variance DDPM theorems do not provide guarantees for those updates.

J.1 Training and sampling procedures

The shared conditional network $y_\theta(x_t, t, l) = y^l_\theta(x_t, t)$ represents the class denoisers, and $a_\phi(x_t, t)$ predicts the class probabilities. In the training algorithms, Y denotes clean data, $c(Y)$ its class index, and $L(Y)$ its one-hot label. Appendix H gives the population objectives and the conditions under which their excess risks control sampling errors.

Algorithms 1 and 2 describe a training design using a discrete time law τ_{train} (uniform on a chosen training grid). Applying the discrete-time risk conversion requires positive training mass at each sampling input time and matching noising coefficients, as in Corollary H.2. The retained records do not establish that alignment for the reported models. The available tables do not support a separate empirical comparison of cross-entropy and squared loss. Each iteration draws the data, time, and Gaussian noise mutually independently, using fresh randomness.

Algorithm 1 Training the conditional denoiser y_θ

1: **repeat**
2: Sample $Y \sim p_d$, $t \sim \tau_{\text{train}}$, and $\epsilon \sim \mathcal{N}(0, I)$.
3: Set $x_t = \alpha_t Y + \sigma_t \epsilon$.
4: Update θ using the gradient of $\|y_\theta(x_t, t, c(Y)) - Y\|^2$.
5: **until** the training stopping criterion is met

Algorithm 2 Training the class-probability network a_ϕ

1: **repeat**
2: Sample $Y \sim p_d$, $t \sim \tau_{\text{train}}$, and $\epsilon \sim \mathcal{N}(0, I)$.
3: Set $x_t = \alpha_t Y + \sigma_t \epsilon$.
4: Update ϕ using the gradient of either $\|a_\phi(x_t, t) - L(Y)\|^2$ or $\mathrm{CE}(L(Y), a_\phi(x_t, t))$.
5: **until** the training stopping criterion is met

J.2 Sampling

MoG-DPM samples a class at each reverse step, whereas the Fixed-class baseline samples a class once at initialization. The fixed-class sampler uses the class prior π_l from the mixture representation of p_d. Define the DDPM update mean

$$\mu_\theta(x_t; s, t, l) = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2} x_t + \frac{\alpha_s \sigma^2_{t|s}}{\sigma_t^2} y_\theta(x_t, t, l). \tag{192}$$

Algorithms 3 and 4 make the routing difference explicit. Every Gaussian draw is fresh and independent of earlier randomness; at each step it is also independent of the categorical draw conditional on the current state. In the fixed-class sampler, the initial Gaussian and class draws are independent. Thus the per-step sampler implements exactly the kernel in (7).

Algorithm 3 DDPM sampling with per-step class routing (MoG-DPM)

1: Sample $x_{t_N} \sim \mathcal{N}(0, I)$.
2: **for** $i = N, N-1, \ldots, 1$ **do**
3: Sample $l \sim \text{Categorical}(a_\phi(x_{t_i}, t_i))$ and $z \sim \mathcal{N}(0, I)$.
4: Set $x_{t_{i-1}} = \mu_\theta(x_{t_i}; t_{i-1}, t_i, l) + \sigma_{t_{i-1}|t_i} z$.
5: **end for**
6: Return x_{t_0}.

Algorithm 4 DDPM sampling with a fixed class (Fixed-class)

1: Sample $x_{t_N} \sim \mathcal{N}(0, I)$ and $l \sim \text{Categorical}(\pi_1, \ldots, \pi_L)$.
2: **for** $i = N, N-1, \ldots, 1$ **do**
3: Sample $z \sim \mathcal{N}(0, I)$.
4: Set $x_{t_{i-1}} = \mu_\theta(x_{t_i}; t_{i-1}, t_i, l) + \sigma_{t_{i-1}|t_i} z$.
5: **end for**
6: Return x_{t_0}.

We also combine these routing conventions with ODE-based integrators. In Algorithms 5 and 6, $\text{Step}(x, t, s, l)$ denotes the chosen integrator's update using the conditional denoiser for class l. The pseudocode specifies the routing convention; solver-specific updates are supplied by the base integrator. The retained records do not determine whether a label was held across all internal stages or how multistep prediction histories were handled; these pseudocodes do not resolve those implementation choices. Per-step random routing makes the overall MoG-DPM sampler stochastic even when Step is deterministic. The DDPM theory does not by itself imply a convergence guarantee for this composite sampler.

Algorithm 5 ODE-based integration with per-step class routing

1: Sample $x_{t_N} \sim \mathcal{N}(0, I)$.
2: **for** $i = N, N-1, \ldots, 1$ **do**
3: Sample $l \sim \text{Categorical}(a_\phi(x_{t_i}, t_i))$.
4: Set $x_{t_{i-1}} = \text{Step}(x_{t_i}, t_i, t_{i-1}, l)$.
5: **end for**
6: Return x_{t_0}.

Algorithm 6 ODE-based integration with a fixed class

1: Sample $x_{t_N} \sim \mathcal{N}(0, I)$ and $l \sim \text{Categorical}(\pi_1, \ldots, \pi_L)$.
2: **for** $i = N, N-1, \ldots, 1$ **do**
3: Set $x_{t_{i-1}} = \text{Step}(x_{t_i}, t_i, t_{i-1}, l)$.
4: **end for**
5: Return x_{t_0}.

J.3 Reported results

The comparisons below evaluate unconditional generation (Uncond), fixed-class routing (Fixed-class), and per-step routing (MoG-DPM). They measure sample quality through MMD or FID, rather than the distributional distances used in the convergence theorems. Relative performance depends on the base sampler, schedule, and step count.

J.3.1 Synthetic 2D datasets

The five synthetic datasets contain continuous points in $\mathbb{R}^2$ with labels determined by their natural clusters. Each experiment generates 4K samples, evaluated using the reported Maximum Mean

Table 5: Reported MMD ↓ on the five synthetic datasets. Values use the 10^{-5} scaling of the evaluation. Each entry averages eight computations using 4K generated samples and a Laplace kernel with bandwidth 0.1. Bold marks the lowest reported value within each dataset and step count.

Dataset (classes)	Steps	Uncond	Fixed-class	MoG-DPM
Circles (2)	500	2.788	8.017	**0.8717**
	100	7.685	5.066	**2.088**
	50	12.42	4.764	**2.594**
	30	20.82	**10.86**	15.55
	20	50.96	**18.90**	40.57
	10	121.0	**33.44**	71.94
Moons (2)	500	3.935	7.051	**2.590**
	100	5.724	4.111	**0.7585**
	50	13.21	5.457	**0.08196**
	30	21.38	8.021	**3.508**
	20	40.36	10.44	**10.12**
	10	149.1	**43.29**	44.30
Pinwheel (5)	500	3.672	6.658	**3.056**
	100	3.891	9.002	**3.789**
	50	14.05	14.79	**3.265**
	30	24.37	17.64	**8.619**
	20	52.51	20.48	**9.817**
	10	169.3	21.07	**12.32**
Checkerboard (8)	500	-1.301	6.561	**-2.912**
	100	0.1639	7.913	**-4.804**
	50	9.516	8.759	**0.8225**
	30	33.57	12.18	**7.559**
	20	115.5	19.88	**10.22**
	10	494.9	**54.94**	90.71
Gaussians (8)	500	6.339	7.945	**2.567**
	100	6.364	11.80	**5.768**
	50	18.93	13.66	**7.536**
	30	45.70	8.005	**1.572**
	20	90.04	6.098	**3.801**
	10	211.1	22.03	**16.62**

Discrepancy (MMD) statistic with a Laplace kernel of bandwidth 0.1 (Gretton et al., 2012). Table 5 reports the mean of eight computations; the retained records do not distinguish independent training runs from repeated generation or evaluation. MoG-DPM improves on the unconditional baseline in the reported configurations, while its comparison with Fixed-class varies across datasets and step counts. The negative entries are retained finite-sample estimates and are not values of the nonnegative population MMD. Such signs are compatible with an unbiased squared-MMD estimator, but the exact estimator and normalization could not be verified from the retained records.

J.3.2 CIFAR-10 WITH STOCHASTIC BASE SAMPLERS

Table 6 reports FID (Heusel et al., 2017) on CIFAR-10 using 50K generated samples for DDPM (Ho et al., 2020), SN-DDPM (Bao et al., 2022), and GMS (Guo et al., 2023). Both routing strategies improve on unconditional DDPM for the reported linear and cosine schedules. The gains are less uniform with the other base samplers. In particular, MoG-DPM has higher FID than unconditional GMS at several cosine-schedule step counts, and Fixed-class often has lower FID than MoG-DPM with SN-DDPM and GMS. These results do not establish that per-step routing is uniformly superior to fixed-class routing.

J.3.3 UNSUPERVISED LABELS ON CELEBA-HQ-256

The CelebA-HQ-256 experiment uses labels obtained without ground-truth classes. Images are encoded into VAE latent vectors (Kingma & Welling, 2013), processed by principal component

Table 6: FID ↓ on CIFAR-10 using 50K generated samples. Column headings give denoising steps. The two panels retain all reported linear- and cosine-schedule results. Bold marks the lowest FID within each base sampler and step count, including ties.

Linear schedule						
Method	10	25	50	100	200	1000
DDPM (Uncond)	43.14	21.63	15.21	10.94	8.23	5.11
+ Fixed-class	37.58	21.05	15.12	10.67	7.82	**4.50**
+ MoG-DPM	**37.50**	**20.72**	**13.81**	**9.38**	**6.66**	**4.50**
SN-DDPM (Uncond)	21.87	6.91	4.58	3.74	3.34	3.71
+ Fixed-class	11.90	**4.98**	**3.62**	**2.98**	**2.55**	**2.93**
+ MoG-DPM	**11.66**	6.39	4.29	3.40	2.97	3.30
GMS (Uncond)	17.43	5.96	4.16	3.26	3.01	**2.76**
+ Fixed-class	**10.40**	**4.84**	**3.61**	**3.00**	**3.00**	2.86
+ MoG-DPM	14.54	5.89	3.82	3.11	3.01	2.82
Cosine schedule						
Method	10	25	50	100	200	1000
DDPM (Uncond)	34.76	16.18	11.11	8.38	6.66	4.92
+ Fixed-class	26.90	14.97	9.94	**7.02**	**5.33**	**3.76**
+ MoG-DPM	**24.28**	**13.69**	**9.44**	**7.02**	5.82	4.69
SN-DDPM (Uncond)	16.33	6.05	4.19	3.83	3.72	4.08
+ Fixed-class	**9.92**	**4.95**	**3.35**	**2.67**	**2.53**	**2.74**
+ MoG-DPM	10.77	5.87	4.08	3.63	3.35	3.51
GMS (Uncond)	13.80	5.48	4.00	3.46	3.34	4.23
+ Fixed-class	**8.76**	**4.91**	**3.43**	**2.76**	**2.60**	**3.35**
+ MoG-DPM	10.80	6.22	4.53	3.64	3.34	4.35

analysis (PCA) (Abdi & Williams, 2010), and partitioned into ten clusters by K-means. Table 7 reports FID results. Both routing strategies improve on the unconditional baseline at the three reported step counts; MoG-DPM has lower FID at 10 and 25 steps, while Fixed-class has lower FID at 50 steps.

This experiment illustrates the use of an unsupervised partition. It does not establish that every partition yields the same network and routing errors or generation quality. The experiments use AdamW; no claim about the optimization rate as a function of partition geometry is inferred from these tables.

Table 7: FID ↓ on CelebA-HQ-256 with labels obtained by K-means clustering of VAE features after PCA. Both merging strategies improve on the reported unconditional baseline at these three step counts; their relative ordering varies with the step count.

Method	10 steps	25 steps	50 steps
Uncond	35.21	18.60	14.16
Fixed-class	30.67	15.73	**12.16**
MoG-DPM	**30.58**	**15.46**	12.25

Table 8: FID ↓ on CIFAR-10 with ODE-based integrators. MoG-DPM uses stochastic class routing even when the base integrator is deterministic. These results illustrate empirical compatibility with the listed integrators; the DDPM analysis does not establish their convergence rates. Bold marks the lowest FID within each base integrator and step count.

Method	10 steps	25 steps	35 steps	50 steps
DDIM	21.31	10.70	9.12	7.74
+ Fixed-class	**16.54**	**9.15**	8.63	**6.60**
+ MoG-DPM	19.44	10.19	**8.44**	7.33
DPM Solver	7.95	6.54	6.17	3.37
+ Fixed-class	5.78	3.49	3.28	**2.90**
+ MoG-DPM	**5.45**	**3.27**	**3.19**	3.36
DPM Solver++	11.11	4.78	3.95	3.42
+ Fixed-class	10.94	4.29	3.80	**2.99**
+ MoG-DPM	**10.37**	**3.69**	**3.39**	3.34
Method	18 steps (reported NFE = 35)			
EDM	2.00			
+ Fixed-class	**1.78**			
+ MoG-DPM	1.95			

J.3.4 ODE-BASED INTEGRATORS

Table 8 reports results for DDIM (Song et al., 2020a), DPM Solver (Lu et al., 2022a), DPM Solver++ (Lu et al., 2022b), and EDM (Karras et al., 2022). The two routing conventions can be used with these base integrators, but their relative FID varies with the integrator and step count. Denoising steps and the reported EDM function-evaluation count are distinguished in the table; the table does not measure classifier overhead or wall-clock speed.

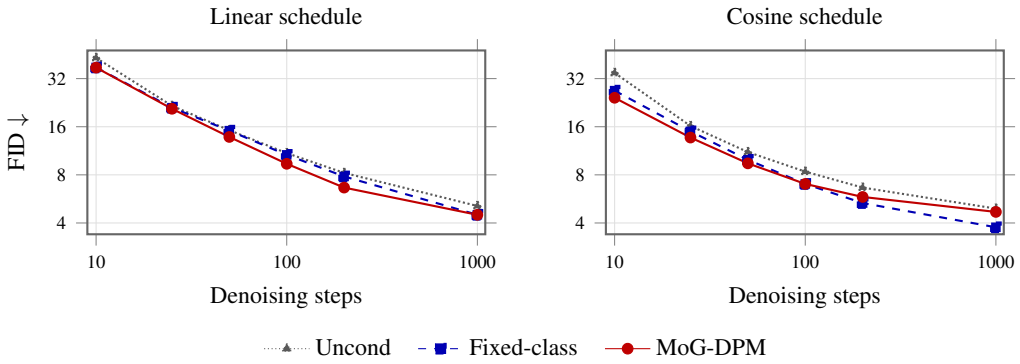


Figure 6: CIFAR-10 DDPM: FID ↓ versus denoising steps, with logarithmic axes. Each marker is an existing entry from Table 6; lines connect reported values.

J.4 A DDPM ILLUSTRATION

Figure 6 replots all DDPM entries from the existing CIFAR-10 evaluation, using 50K generated samples per setting. The plot uses the same shared conditional denoiser and routing conventions as the tables.

Both routing variants improve on unconditional DDPM in the reported settings. MoG-DPM matches or improves on Fixed-class routing for the linear schedule. With the cosine schedule, Fixed-class has lower FID at 200 and 1000 steps.

The theory controls TV under explicit network error and geometry-weighted routing error assumptions. FID is a different quantity: these results illustrate sampling behavior rather than verify those assumptions or measure the theoretical TV rate. The complete tables above also include base samplers with different variances or ODE updates; those settings remain outside the analyzed DDPM kernel.

J.5 QUALITATIVE SAMPLES

Figures 7–10 show CIFAR-10 sample grids from DDPM with MoG-DPM at 10, 25, 50, and 100 denoising steps. Figure 11 shows latent-diffusion examples with ODE-based sampling. These grids illustrate individual outputs and are not paired comparisons between methods.

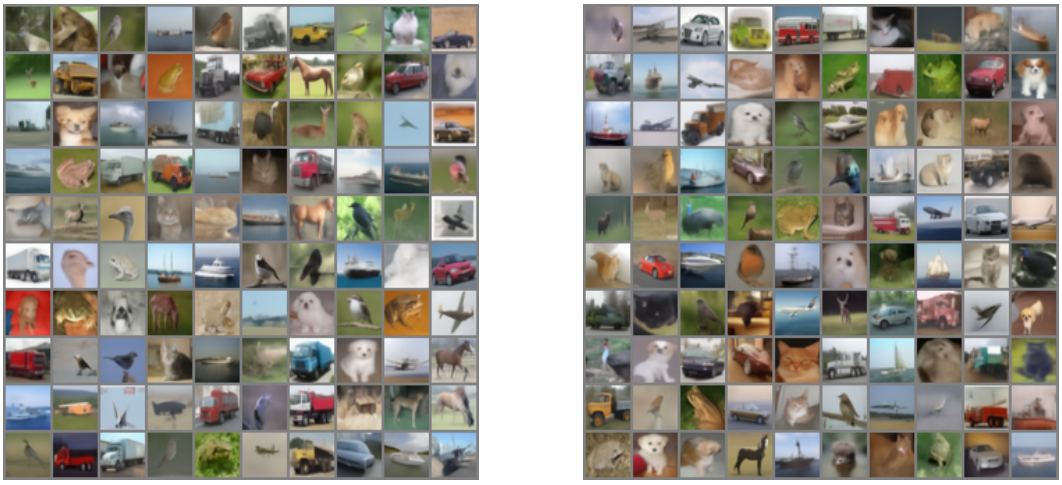


Figure 7: CIFAR-10 sample grids from DDPM with MoG-DPM at 10 denoising steps.

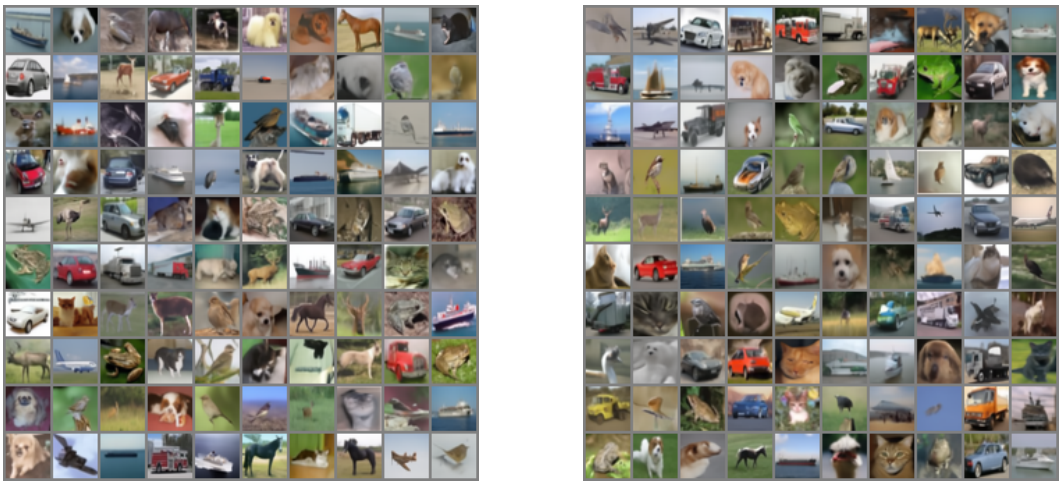


Figure 8: CIFAR-10 sample grids from DDPM with MoG-DPM at 25 denoising steps.

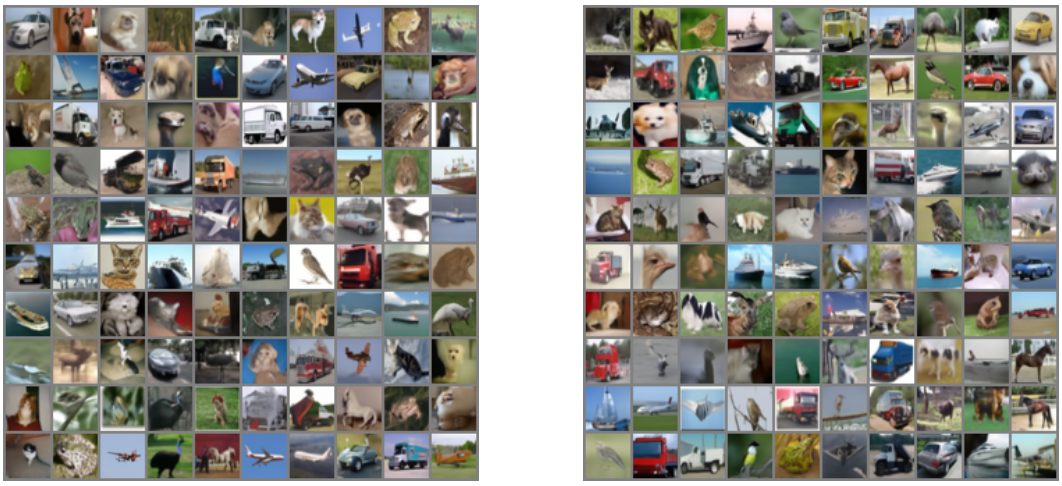


Figure 9: CIFAR-10 sample grids from DDPM with MoG-DPM at 50 denoising steps.

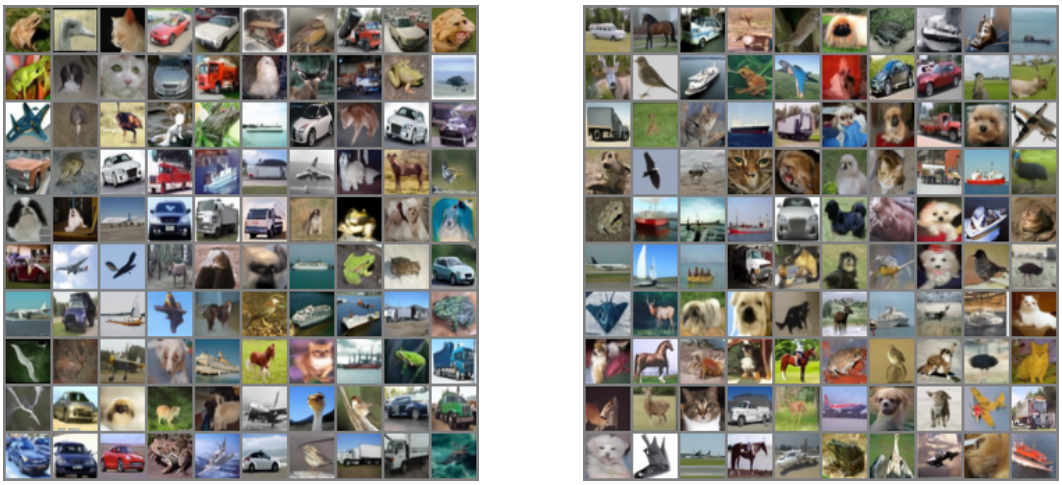


Figure 10: CIFAR-10 sample grids from DDPM with MoG-DPM at 100 denoising steps.



Figure 11: Retained latent-diffusion samples using an ODE-based integrator, at 10 steps (left) and 25 steps (right). The checkpoint and solver configuration are not specified in the retained records.

K SCOPE OF THE GUARANTEES

The analyzed transition. Theorems 3.1 and 3.2 concern the posterior-variance stochastic kernels in (5) and (7). A weighted average of expert predictions produces a single Gaussian transition and generally differs from drawing a class and then sampling its Gaussian component. Likewise, replacing the reverse step by an ODE solver or a different noise covariance changes the kernel. The corresponding empirical results in Appendix J are comparisons, rather than additional consequences of these theorems.

Learning assumptions. Squared denoising loss identifies the population conditional means. Proposition H.1 and its following discussion give the corresponding population characterizations for routing and denoising. This population characterization alone does not provide finite-sample or optimization guarantees for E_i and W_i. Per-time errors must also be distinguished from an average over training times: equation (186) gives the explicit conversions when those times have positive training probability. Theorem 3.2 uses logarithmic routing energy Ψ_i, which is no larger than the sufficient bound Ω_i. Corollary 3.3 gives a sufficient condition in terms of the pairwise diameter of learned expert predictions, allowing an unbounded common component. Projecting all predictions onto a ball of radius R enforces diameter at most $2R$, but its approximation error must be included in E_i. Proposition E.5 instead uses a quadratic diameter-weighted routing moment. Appendix I proves when projection controls both diameter and approximation bias. The reported experiments do not certify these global network conditions. Corollary H.2 converts population excess risks directly into a sampling-error bound when the sampling times have positive training mass; it does not supply a statistical generalization guarantee.

The clean-data limit. Corollary F.2 uses a common bounded-Lipschitz metric to combine smoothing and sampling errors. It proves weak convergence for fixed dimension and data law with shrinking early-stopping noise, sufficient discretization refinement, and vanishing learning energies. A fixed nonzero learning-error budget does not give this conclusion. Neither finite TV error nor bounded-Lipschitz convergence alone controls generated second moments, and no clean-data TV guarantee is asserted for singular data.

Constants, dimension, and the grid. The main discretization budget is $d(9H+5)(e^{H/N}-1)/4$ for an equal log-SNR grid of span H, with the dimension-free step condition $e^{H/N}-1 \le 1$. With $H = O(\log N)$, it is $O(d\log^2 N/N)$; fixed H gives the sharper special case $O(d/N)$. The factor $e^{H/N} \le 2$ is uniformly bounded; the logarithms arise from $H(9H+5)$. This is an accumulated guarantee, not a uniform $O(d\rho_i^2)$ bound at each time. For irregular grids, the complementary local bound $C_d\rho_i^2$, with $C_d = 261d(d+2)$ and $\rho_i \le 1/(d+6)$, remains available. Appendix L further exposes the dependence of this local budget on classwise posterior uncertainty. These upper bounds are not practical step-count predictions or certificates for the retained image experiments. Appendix M distinguishes prior sharp rates, kernels, and assumptions; intrinsic-dimension adaptation is not established here.

Routing and the data partition. The accumulated discretization bound is independent of the number of classes because it averages classwise estimates with the normalized priors π_l. Appendix L makes the discarded within-class posterior uncertainty explicit and shows how it can reduce the discretization budget. The difficulty of learning the experts and router, the expert diameter, and conversions from individual class-error budgets also depend on the partition. A reduction in the oracle budget need not improve the full learned bound.

L HOW THE PARTITION AFFECTS DISCRETIZATION

The main theorems use the class-uniform budget $\mathcal{D}_N$. This appendix examines a complementary quantity that this worst-case bound does not retain: the posterior uncertainty remaining within each class. Throughout, use Assumption 2.1, the finite positive-prior class representation, and the oracle

mixture kernels of (7). Write $X_i = X_{t_i}$ and define

$$C_i^l(x) = \mathrm{Cov}(Y \mid X_i = x, J = l),$$
$$V_i(J) = \mathbb{E}\|Y - \bar{y}^J(X_i, t_i)\|^2 = \mathbb{E}_{p_i} \sum_l a^l(X_i, t_i) \operatorname{tr} C_i^l(X_i) \le M. \tag{193}$$

Here $\bar{y}^J$ denotes the denoiser indexed by the random data label. The canonical Gaussian posteriors have finite moments at every input; conditional expectation as an L^2 projection gives the final bound.

Retained and discarded posterior uncertainty. At a fixed input, put $m_l = \bar{y}^l(x, t_i)$ and $m = \sum_l a^l(x, t_i) m_l = \bar{y}(x, t_i)$. The exact and routed oracle kernels both have conditional mean $A_i x + B_i m$. The bridge representation and conditional total covariance give

$$\begin{aligned} \mathrm{Cov}_{K_i(x)} &= v_i I_d + B_i^2 \,\mathrm{Cov}(Y \mid X_i = x), \\ \mathrm{Cov}_{Q_i(x)} &= v_i I_d + B_i^2 \sum_l a^l(x, t_i)(m_l - m)(m_l - m)^T, \\ \mathrm{Cov}_{K_i(x)} - \mathrm{Cov}_{Q_i(x)} &= B_i^2 \sum_l a^l(x, t_i) C_i^l(x) \succeq 0. \end{aligned} \tag{194}$$

Thus the oracle retains between-class posterior uncertainty and omits within-class posterior uncertainty. Its matching conditional mean also gives

$$\mathrm{Cov}(p_{i-1}) - \mathrm{Cov}(p_i Q_i) = B_i^2 \mathbb{E}_{p_i} \sum_l a^l(X_i, t_i) C_i^l(X_i). \tag{195}$$

Indeed, the covariance of the common conditional mean cancels when the input is integrated against p_i. All terms are integrable by (193) and conditional Jensen. Covariance ordering alone does not order total variation errors.

Lemma L.1 (A within-class bound)**.** *Let φ be the standard one-dimensional Gaussian density and $C_d = 261d(d+2)$. For every step, without a restriction on ρ_i,*

$$\mathrm{TV}(K_i(x), Q_i(x)) \le \varphi(1)\Delta_i \sum_l a^l(x, t_i) \operatorname{tr} C_i^l(x). \tag{196}$$

Consequently, if $\rho_i \le 1/(d+6)$, then

$$\mathrm{TV}(p_{i-1}, p_i Q_i) \le \min\{C_d \rho_i^2,\ \varphi(1)\Delta_i V_i(J)\}. \tag{197}$$

Proof. We first prove a Gaussian convolution estimate. Let R be a centered random vector with finite second moment, let $G \sim \mathcal{N}(0, vI_d)$ be independent, and let $b \in \mathbb{R}$. For the density g_v of G, Taylor's formula gives

$$g_v(z - br) - g_v(z) = -br^T \nabla g_v(z) + b^2 \int_0^1 (1-u) r^T \nabla^2 g_v(z - ubr) r \, du.$$

Averaging cancels the first term. Rotation and Gaussian scaling yield

$$\int |r^T \nabla^2 g_v(z) r| \, dz = \frac{\|r\|^2}{v} \mathbb{E}|Z_1^2 - 1| = \frac{4\varphi(1)}{v}\|r\|^2, \qquad Z_1 \sim \mathcal{N}(0,1).$$

The last equality follows by symmetry and $\int_1^\infty (z^2 - 1)\varphi(z)\, dz = \varphi(1)$. Taking half the density L^1 norm, using translation invariance and $\int_0^1 (1-u)\, du = 1/2$, gives

$$\mathrm{TV}(\mathcal{L}(G + bR), \mathcal{L}(G)) \le \varphi(1) \frac{b^2}{v} \mathbb{E}\|R\|^2. \tag{198}$$

These exchanges require only a second moment: the Gaussian Hessian is bounded, so its pointwise Taylor remainder is integrable in R, and the displayed Hessian norm makes its absolute integral over (z, R, u) finite. The canceled first derivative is integrable because $\mathbb{E}\|R\| < \infty$. Thus Fubini, or Tonelli for the absolute remainder, applies.

At fixed (x, l), apply (198) to the centered posterior $R = Y - \bar{y}^l(x, t_i)$, with $b = B_i$ and $v = v_i$. Translate by $A_i x + B_i \bar{y}^l(x, t_i)$ and use $B_i^2 / v_i = \Delta_i$. Convexity under the common posterior weights $a^l(x, t_i)$ proves (196). Integrating against p_i and using $p_i K_i = p_{i-1}$ bounds the local marginal defect by $\varphi(1)\Delta_i V_i(J)$. Under the stated step restriction, Lemma C.1, applied to each normalized class law, and the class-joint identity give its other upper bound $C_d \rho_i^2$. Taking the minimum proves (197). □

Corollary L.2 (Partition-dependent discretization budget)**.** *Under the assumptions of Corollary 5.1, define*

$$\mathcal{D}_N(J) = \min\left\{\mathcal{D}_N,\ \varphi(1)\sum_{i=1}^{N}\Delta_i V_i(J)\right\} \le \mathcal{D}_N. \tag{199}$$

If, in addition, $\rho \le 1/(d+6)$, one may replace this budget by the smaller quantity

$$\mathcal{D}_N^{\mathrm{fine}}(J) = \min\left\{\mathcal{D}_N,\ \sum_{i=1}^{N}\min\{C_d\rho^2,\ \varphi(1)\Delta_i V_i(J)\}\right\}. \tag{200}$$

With exactly the same learning quantities κ_i as in (35),

$$\mathrm{TV}(p_0, q_0) \le \mathcal{I}_N + \mathcal{D}_N(J) + \sqrt{\frac{1}{2}\sum_i \kappa_i}. \tag{201}$$

The same inequality holds with $\mathcal{D}_N^{\mathrm{fine}}(J)$ under its additional step restriction.

Proof. Integrating (196) bounds the sum of local marginal defects by $\varphi(1)\sum_i \Delta_i V_i(J)$. Independently, Theorem D.1, applied classwise and averaged with the prior probabilities, bounds that same sum by $\mathcal{D}_N$. The minimum of these two accumulated bounds proves (199). Under the finer step restriction, (197) gives the second budget in (200). The same entropy and coupling estimates as in Corollary 5.1 give $\mathbb{E}_{p_i}\mathrm{KL}(Q_i\|P_i) \le \kappa_i$. Their sum is finite by assumption, so Lemma E.1 proves the endpoint inequality. □

The outer minimum is essential: the sum of the coarse per-step bounds need not be smaller than the new accumulated budget $\mathcal{D}_N$. The variance branch alone also gives a valid budget $\varphi(1)\sum_i \Delta_i V_i(J)$ on arbitrary grids, without a small-step condition, provided the same finite learning-energy assumptions hold. It complements the accumulated cancellation estimate: for fixed nonzero within-class uncertainty it is generally first order in an individual absolute increment and does not improve the universal asymptotic grid order by itself.

Nested refinements. If $J = g(J')$ in the same joint data model, orthogonal projection gives

$$V_i(J) = V_i(J') + \mathbb{E}\|\mathbb{E}[Y \mid X_i, J'] - \mathbb{E}[Y \mid X_i, J]\|^2.$$

Hence neither (199) nor, when applicable, (200) can increase under this refinement. This statement concerns the discretization budget. It neither orders actual TV errors nor ensures that the full learned bound improves, since changing the class representation also changes its learning quantities.

Exactness for atomic classes. If $p_d^l = \delta_{y_l}$, then $\bar{y}^l(x, t_i) = y_l$ and $C_i^l(x) = 0$. The posterior distribution of Y places weights $a^l(x, t_i)$ at these atoms, so the bridge formula gives $Q_i(x) = K_i(x)$ for every input and every admissible grid. Thus

$$p_N Q_N \cdots Q_1 = p_0, \qquad \mathrm{TV}(p_0, \gamma^d Q_N \cdots Q_1) \le \mathcal{I}_N. \tag{202}$$

There is no discretization error, but Gaussian initialization can still produce an error; the target here is the positive-noise law p_0. For at least two distinct atoms with positive priors, the standard oracle instead has a strictly positive local marginal defect: its missing covariance is $B_i^2\mathbb{E}\,\mathrm{Cov}(Y \mid X_i) \ne 0$, since all canonical posterior weights are positive. Its output therefore cannot equal p_{i-1}. This is a local comparison under the true input marginal, not a ranking of complete learned samplers.

A nondegenerate Gaussian-class example. Suppose, only for this example, that $p_d^l = \mathcal{N}(\mu_l, \tau^2 I_d)$ with $\tau > 0$ common to all classes. Set $S_i = \alpha_i^2\tau^2 + \sigma_i^2$ and $c_i = \tau^2/(1+\eta_i\tau^2)$. Gaussian conditioning gives

$$\bar{y}^l(x, t_i) = \mu_l + \frac{\alpha_i\tau^2}{S_i}(x - \alpha_i\mu_l), \qquad C_i^l(x) = c_i I_d.$$

Each normalized input class is Gaussian and its oracle prediction is affine. The bridge has the same mean; its omitted conditional covariance is the constant $B_i^2 c_i I_d$. Consequently

$$p_{i-1}^l = \mathcal{N}(\alpha_{i-1}\mu_l, S_{i-1} I_d), \qquad p_i^l Q_i^l = \mathcal{N}(\alpha_{i-1}\mu_l, (S_{i-1} - B_i^2 c_i) I_d).$$

The latter variance is positive because the update adds $v_i I_d$. The relative variance deficit is therefore in $(0, 1)$ and equals

$$\begin{aligned}\delta_i^{\mathrm{G}} &= \frac{B_i^2 c_i}{S_{i-1}} = \frac{\tau^2 \Delta_i^2}{\eta_{i-1}(1+\eta_i\tau^2)(1+\eta_{i-1}\tau^2)} \\ &= \frac{u\rho_i^2}{(1+\rho_i)(1+u)(1+u(1+\rho_i))}, \qquad u = \eta_i\tau^2.\end{aligned} \tag{203}$$

Here substitution uses $B_i^2 = \Delta_i^2/[\eta_{i-1}(1+\eta_{i-1})]$ and $S_{i-1} = (1+\eta_{i-1}\tau^2)/(1+\eta_{i-1})$. Writing $H_d(\delta) = \mathrm{TV}(\mathcal{N}(0, I_d), \mathcal{N}(0, (1-\delta)I_d))$, classwise scaling and (6) give

$$\mathrm{TV}(p_{i-1}, p_i Q_i) \le \sum_l \pi_l H_d(\delta_i^{\mathrm{G}}) = H_d(\delta_i^{\mathrm{G}}). \tag{204}$$

Equality holds for each normalized class comparison, not necessarily for the mixture TV distance. Differentiating the centered Gaussian density in its variance gives $H_d(\delta) = c_d\delta + o(\delta)$, where $c_d = \frac{1}{4}\mathbb{E}|\chi_d^2 - d|$ and $c_1 = \varphi(1)$. Indeed, at variance one the density derivative is $\frac{1}{2}(\|z\|^2 - d)\varphi_d(z)$, with φ_d the standard d-dimensional Gaussian density; for variances in $[1/2, 1]$ its absolute value is bounded by an integrable Gaussian times a quadratic polynomial. Dominated convergence gives the derivative in L^1 and the claimed TV expansion. For fixed $\eta_i\tau^2 = u > 0$, the right side of (204) is therefore $c_d u(1+u)^{-2}\rho_i^2 + o(\rho_i^2)$. For a fixed step and $\tau \downarrow 0$, it is instead $c_d\Delta_i^2\tau^2/\eta_{i-1} + o(\tau^2)$. This example retains second-order marginal cancellation while making its coefficient sensitive to within-class spread. Small spread need not make learning the router easier: here $\bar{y}^l - \bar{y}^j = (\mu_l - \mu_j)/(1+\eta_i\tau^2)$, so expert separation can remain large in the atomic limit.

M Relation to prior local convergence estimates

Table 1 distinguishes the sampler, assumptions, and treatment of error accumulation as well as the final rate. Theorem D.1 proves $d(9H+5)(e^{H/N}-1)/4$ discretization for posterior-variance DDPM on equal log-SNR grids under a finite second moment. Exact-channel interpolation and a resolvent covariance potential preserve the cancellations needed for linear dimension dependence. This is an accumulated bound, with $e^{H/N} - 1 \le 1$; it is not a claim of a universal $O(d\rho_i^2)$ local bound. The complementary Lemma C.1 applies to every two-time interval with $\rho_i \le 1/(d+6)$, with constant $261d(d+2)$ independent of the endpoints and moment value, and can be summed on irregular grids.

Posterior variance and the Li–Yan results. Theorem numbers here refer to the journal extension (Li & Yan, 2025b) of Li & Yan (2025a). To compare coefficients, set $\lambda_i = \alpha_i^2/\alpha_{i-1}^2$ and $u_i = \alpha_i^2$. Our score-form conditional mean and variance are

$$\frac{x + (1-\lambda_i)\nabla \log p_i(x)}{\sqrt{\lambda_i}}, \qquad v_i = \frac{(1-\lambda_i)(\lambda_i - u_i)}{\lambda_i(1-u_i)}. \tag{205}$$

Li–Yan's Theorem 1 uses the same mean but conditional variance $(1-\lambda_i)/\lambda_i$ and gives $\widetilde{O}(d/N)$ under a finite first moment that grows at most polynomially in N. Their Theorem 3 uses exactly v_i and gives $\widetilde{O}(k/N)$ under an intrinsic covering condition. Its proof combines typical sets and posterior matrices (Lemmas 9, 14, and 15) with schedule-dependent accumulation (Lemmas 18 and 22). Our ambient-dimension rate matches the d/N exponents through a different argument for posterior variance; it does not subsume their weaker moment assumption or their intrinsic-dimension adaptation.

Sharp posterior-variance theory with support geometry. Theorem 4 of Liang et al. (2026), in the July 2026 revision, also uses exactly (205). It assumes bounded support and a covering-number condition, with weighted true-marginal L^2 score errors (Assumptions 1, 2, and 4). Appendix D, equations (93)–(105), bounds a one-step marginal density defect through posterior covariances; Lemma 18 includes differences of adjacent posterior uncertainties, accumulated using the schedule in equation (107). Thus posterior-uncertainty accumulation already plays an important role in sharp theory. Our exact-channel interpolation supplies an explicit finite-grid bound under the present moment assumption, while their result captures intrinsic dimension.

What transfers classwise. The Bayes averaging in (30) applies to *any* valid classwise local bound, including one obtained by another proof. Shared support and covering assumptions can likewise permit classwise use of the preceding estimates. We do not claim that such an extension is unavailable to prior methods. Our accumulated estimate applies to every positive-probability class law with the same dimension, grid, and constant. Summing over steps and averaging over the fixed class priors therefore preserves the $O(d \log^2 N/N)$ budget when $H = O(\log N)$ without a class-count factor. The learned-router terms then require the separate Gaussian-mixture analysis in Section 5.

Sharp KL analysis and other Gaussian mixtures. Jain & Zhang (2025, Theorem 3.1) obtain a $\widetilde{O}(d^2/N^2)$ KL discretization term under finite second moment and true-marginal L^2 score errors. Their update combines a probability-flow ODE step with forward noising and differs from (205). It demonstrates that KL analysis can yield first-order TV bounds by Pinsker; our cancellation discussion concerns the particular conditional-transition comparison in Section 4. Finally, Gaussian Mixture Solvers (Guo et al., 2023) estimate transition moments and fit mixture parameters during sampling. Our mixture instead uses class-conditioned denoisers and a posterior router. These constructions share Gaussian components but specify different approximation mechanisms.

M.1 How prior bounds handle a vanishing stopping time

Early stopping is coupled to the step count or kept as an explicit accuracy parameter in the smoothness-free results compared here; the associated logarithms are not removed by suppressing them with $\widetilde{O}$ notation.

Li & Yan (2025b, equation (3), Theorems 1 and 3) choose the first forward-noising variance $\beta_1 = N^{-c_0}$ and a step-count-dependent schedule. Their early-stopped target thus has SNR $(1-\beta_1)/\beta_1 = N^{c_0} - 1$, which diverges polynomially. The displayed TV discretization terms are $O(d \log^3 N/N)$ for their first variance choice and $O(k \log^3 N/N)$ for posterior variance with intrinsic covering. This handles the endpoint through a joint schedule limit, not an endpoint-uniform constant.

Benton et al. (2024, Theorem 1 and Corollary 1) keep a forward horizon T and a positive stopping time δ. Their mesh parameter satisfies $\kappa = \Theta((T + \log(1/\delta))/N)$, and the KL discretization terms are $\kappa^2 dN + \kappa dT$. The shrinking endpoint therefore enters logarithmically through the mesh. Here T denotes elapsed diffusion time, not the number of steps.

For the ODE-plus-noising sampler, Jain & Zhang (2025, Theorem 3.1) likewise use $c = \Theta((T + \log(1/\delta))/N)$ and obtain a KL discretization term $d^2 c^3 N$. Pinsker gives a TV contribution of order $d(T + \log(1/\delta))^{3/2}/N$ under that theorem's assumptions and step condition. Their different kernel cannot be substituted into our posterior-variance theorem.

These results motivate retaining endpoint parameters until an explicit schedule is chosen. In our setting, polynomial SNR endpoints keep $H = O(\log N)$ while the stopping noise vanishes. Proposition D.5 further avoids charging the entire low-SNR interval at the maximal covariance trace. Proposition C.6 identifies an actual H/N cost on our equal log-SNR grid, while leaving open whether the second logarithm can be removed for general finite-second-moment data. Improving the grid or preserving more signed cancellation could help; neither improvement follows by treating H as constant.

N Three- and four-part global error bounds

The decomposition in Figure 1 separates the quantities controlled at different stages of the proof. It follows from the marginal-TV and conditional-KL comparison in Lemma E.1, which uses the absorbing-state construction of Li & Yan (2025b, Sections 4.2–4.4). The comparison principle is a tool for assembling the bounds; the two contributions analyzed here are the linear-dimensional accumulated discretization bound and the additional learned-router term.

Write the actual local oracle defects as

$$\delta_i^{\mathrm{std}} = \mathrm{TV}(p_{i-1}, p_i Q_i^{\mathrm{std}}), \qquad \delta_i^{\mathrm{dec}} = \mathrm{TV}(p_{i-1}, p_i Q_i). \tag{206}$$

Recall that $\mathcal{I}_N = \mathrm{TV}(p_N, \gamma^d)$, $\varepsilon^2_{\mathrm{net,std}} = \sum_i \Delta_i e_i^2$, and $\varepsilon^2_{\mathrm{net,dec}} = \sum_i \Delta_i E_i^2$. The two local defects generally differ; the common upper bound $\mathcal{D}_N$ does not assert that they are equal.

Proposition N.1 (Global error budgets)**.** *Under Assumption 2.1, initialize the corresponding learned sampler from γ^d. If $\varepsilon_{\mathrm{net,std}} < \infty$, then*

$$\mathrm{TV}(p_0, q_0^{\mathrm{std}}) \le \underbrace{\mathcal{I}_N}_{\textit{initialization}} + \underbrace{\tfrac{1}{2}\varepsilon_{\mathrm{net,std}}}_{\textit{network error}} + \underbrace{\sum_i \delta_i^{\mathrm{std}}}_{\textit{discretization}} . \tag{207}$$

For the decentralized sampler, if $\varepsilon_{\mathrm{net,dec}} < \infty$ and $\sum_i \Psi_i < \infty$, then

$$\begin{aligned} \mathrm{TV}(p_0, q_0) &\le \mathcal{I}_N + \sum_i \delta_i^{\mathrm{dec}} + \sqrt{\tfrac{1}{2}\varepsilon_{\mathrm{net,dec}}^2 + \tfrac{1}{2}\sum_i \Psi_i} \\ &\le \underbrace{\mathcal{I}_N}_{\textit{initialization}} + \underbrace{\tfrac{1}{\sqrt{2}}\varepsilon_{\mathrm{net,dec}}}_{\textit{network error}} + \underbrace{\sum_i \delta_i^{\mathrm{dec}}}_{\textit{discretization}} + \underbrace{\sqrt{\tfrac{1}{2}\sum_i \Psi_i}}_{\textit{routing}} . \end{aligned} \tag{208}$$

No small-step restriction is needed for this proposition: the local defects have not yet been estimated. The final line separates four nonnegative *upper-bound budgets*; it is not an equality between four independently measurable portions of the actual endpoint error. In particular, Ψ_i depends on the learned expert geometry as well as the routing error.

Proof. For either sampler, set $k_i = \mathbb{E}_{p_i}\mathrm{KL}(Q_i(X,\cdot)\|P_i(X,\cdot))$ with its corresponding oracle and learned kernels. Lemma E.1 bounds the endpoint TV by $\mathcal{I}_N + \sum_i \delta_i + \sqrt{\frac{1}{2}\sum_i k_i}$. For the standard sampler, the equal-covariance Gaussian KL formula and $B_i^2/v_i = \Delta_i$ give $k_i = \Delta_i e_i^2/2$. Substitution proves (207), including the coefficient $1/2$.

For the decentralized sampler, introduce R_i with the true weights a^l and learned denoisers y_θ^l, as in (133). At a fixed input x, data processing from the common class label yields

$$\mathrm{KL}(Q_i(x)\|R_i(x)) \le \frac{\Delta_i}{2}\sum_l a^l(x,t_i)\|\bar{y}^l(x,t_i) - y_\theta^l(x,t_i)\|^2.$$

The entropy bridge in Lemma E.3 and the Gaussian-overlap estimate in Lemma E.2 then give

$$\mathrm{KL}(Q_i(x)\|P_i(x)) \le 2\mathrm{KL}(Q_i(x)\|R_i(x)) + \log\Big[1 + (e^{\Delta_i D_i(x)^2} - 1)\|a(x,t_i) - a_\phi(x,t_i)\|_1^2\Big] .$$

Integration against p_i gives $k_i \le \Delta_i E_i^2 + \Psi_i$; the stated finiteness assumptions therefore justify applying the comparison lemma. This proves the first line of (208). The second follows from $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for $u, v \ge 0$. □

Inserting the two contributions. On the equal log-SNR grid $\eta_i = \eta_0 e^{-iH/N}$, with $H = \log(\eta_0/\eta_N)$ and $\rho = e^{H/N} - 1 \le 1$, Theorem D.1 and its classwise application give

$$\sum_i \delta_i^{\mathrm{std}}, \sum_i \delta_i^{\mathrm{dec}} \le \mathcal{D}_N = \frac{d(9H+5)}{4}(e^{H/N} - 1) \le \frac{e^{H/N}H(9H+5)}{4}\frac{d}{N}. \tag{209}$$

The theorem controls the accumulated defects by cancellation of posterior-covariance boundary terms; it does not assert an equal share $\mathcal{D}_N/N$ for each step. Replacing the local-defect sums by $\mathcal{D}_N$ in (207)–(208) gives the three- and four-part bounds displayed in the main text. The initialization and network error contributions then sit alongside the discretization estimate; the decentralized bound adds the routing contribution controlled in Section 5.

Theorem 3.2 retains the smaller, unsplit square root. Moreover, under exact routing the direct comparison of mixtures with identical weights gives $k_i \le \Delta_i E_i^2/2$, recovering network error contribution $\varepsilon_{\mathrm{net,dec}}/2$; equivalently, use the coupling branch of Corollary 5.1 with $T_i = 0$. The coefficient $1/\sqrt{2}$ in the four-part presentation is a cost of the entropy bridge, not an intrinsic penalty for decentralization.