WHY DIFFUSION SAMPLING CAN IGNORE AMBIENT DIMENSION: RATE–DISTORTION BOUNDS IN KL AND TV

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models sample high-dimensional data in few steps. Yet existing discretization bounds scale with an intrinsic dimension of the data or, for Gaussian mixtures, with the squared logarithm of the number of components. We show that what matters instead is the information the data carry at the target resolution. For the posterior-mean exponential integrator with K steps, the time-discretization error in total variation is $O(I_\Gamma L/K)$. Here I_Γ is the mutual information between the data and their Gaussian observation at the terminal signal-to-noise ratio, and L is the log-SNR range. The bound is linear in L/K, whereas Pinsker's inequality applied to information-theoretic KL bounds gives only $\sqrt{L/K}$. We bound I_Γ by the rate–distortion function of the data. Up to a constant, it is at most the metric entropy of the support, and for Gaussian mixtures it is at most the label entropy, so our bound improves on both. The implication: *as far as discretization error is concerned, diffusion in pixel space does as well as diffusion in latent space*, whatever the number of pixels.

1 INTRODUCTION

Diffusion models generate images, audio, video, and molecular structures by reversing a progressive noising process (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022). Their practical success is striking in view of classical nonasymptotic analyses: many bounds for the number of reverse steps depend explicitly on the dimension of the Euclidean representation (Chen et al., 2023b;a; Benton et al., 2024). This dependence can be pessimistic when a high-dimensional sample is generated from a compact latent description.

Modern image tokenizers make this mismatch concrete. Vector-quantized models encode an image by a sequence of discrete codebook indices before decoding it to pixels (van den Oord et al., 2017; Esser et al., 2021; Yu et al., 2024). The ambient pixel dimension can therefore be very large even when the uncertainty in the latent code is modest. Unless reconstruction is perfect, however, the decoded image only approximates the original. A useful bound must therefore account for both the uncertainty of the code and what the code misses. We ask whether these two quantities can replace the ambient pixel dimension in the sampling bound, in total variation as well as in KL divergence.

We answer this question through the Gaussian channel associated with the heat flow. Let $V \in \mathbb{R}^d$ be a latent signal, and let

$$X_t = V + \sqrt{t + \sigma^2}\, G, \qquad G \sim \mathcal{N}(0, I_d).$$

The case $\sigma = 0$ is the usual Brownian noising of an arbitrary target. The case $\sigma > 0$ includes isotropic Gaussian mixtures, where V is the mixture centre. When the sampler stops at time δ, the largest signal-to-noise ratio it reaches is $\Gamma = (\delta + \sigma^2)^{-1}$. The central quantity of this paper is the mutual information at that resolution,

$$I_\Gamma := I\big(V; \sqrt{\Gamma}\, V + G\big).$$

Main message. Suppose a diffusion model is run directly on pixels, with $\sigma = 0$ and $V = X_0$ the image, and no autoencoder is used anywhere. We show that the number of steps needed for accurate time discretization is governed by the *latent* structure of the images, not by the number of pixels.

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.

The relevant latent quantities are the entropy of the best code for the images and the part of that code's reconstruction error that is still resolved at the terminal resolution. The sampler never sees the code, and the guarantee holds simultaneously for every encoder–decoder pair, so it adapts to the most compressive one. In this precise sense, pixel-space diffusion already pays only for the latent, as far as discretization is concerned (Corollary 5 and Remark 1). The information complexity of pixel-space diffusion exceeds that of diffusion on the code itself by at most half a nat. The statement concerns time discretization; the difficulty of *learning* a pixel-space score is a separate question, isolated in the score-error term.

Write $R = \mathbb{E}\,\|V\|^2$, let h be the largest relative SNR step of the grid, let P_δ be the law of X_δ, and let $\widehat{P}_\delta$ be the sampler output. Our contributions are as follows.

- **KL (Theorem 1).** This part follows the information-theoretic approach of Reeves & Pfister (2025), who bound the KL divergence of discrete-time diffusion samplers through the I–MMSE relation. We state the bound in terms of I_Γ, because this is the form our rate–distortion results control. An exact MMSE-area identity together with summation by parts gives, for every grid,
$$\mathrm{KL}(P_\delta \| \widehat{P}_\delta) \le \frac{R}{2(T+\sigma^2)} + h\Big(I_\Gamma + \frac{R}{2(T+\sigma^2)}\Big) + \frac{1}{2} e_{\mathrm{apx}}.$$
- **Direct total variation (Theorem 2), our main new result.** Pinsker's inequality applied to a KL bound of order hI_Γ gives only $\sqrt{hI_\Gamma}$. We prove a bound that is linear in h. The exact reverse kernel is a Gaussian mixture over the posterior of V, and the sampler's kernel is its mean-matched collapse. A one-sided Jensen bound on the one-step density ratio holds at every point, and a one-step MMSE-drop inequality converts it into information. The result is
$$\mathrm{TV}(P_\delta, \widehat{P}_\delta) \le \sqrt{\frac{R}{4(T+\sigma^2)} + (3h+h^2)\, I_\Gamma + O\Big(\frac{hR}{T+\sigma^2}\Big)} + \frac{1}{2}\sqrt{e_{\mathrm{apx}}}.$$
It needs no density, typical set, truncation, or clipping.
- **Rate–distortion control (Theorem 3 and Corollary 2).** For every code Z, $I_\Gamma \le I(V;Z) + \frac{1}{2}\mathbb{E}\log\det\big(I + \Gamma\mathrm{Cov}(V \mid Z)\big)$. Hence $I_\Gamma \le \mathcal{R}_V(1/\Gamma) + \frac{1}{2}$, where $\mathcal{R}_V$ is the mean-square rate–distortion function. For a discrete code with reconstruction error D, the reconstruction cost is $\frac{d}{2}\log(1+\Gamma D/d)$, which is logarithmic in the terminal resolution rather than linear. For arbitrary data ($\sigma = 0$, $V = X_0$, with no density, support, or mixture assumption), sampling $X_0 + \varepsilon G$ therefore costs $O\big((\mathcal{R}_{X_0}(\varepsilon^2)+1)\log(T/\varepsilon^2)/K\big)$ in both KL and total variation (Corollary 2). This complexity is never larger than the metric entropy of the support at scale $\Gamma^{-1/2}$, plus half a nat. It is much smaller whenever the data spread their mass unevenly, so it improves on intrinsic-dimension and covering-number complexities (Corollary 3).
- **Autoencoders, tokenizers, and pixel versus latent diffusion (Corollaries 4 and 5).** For any encoder–decoder pair, and in particular for a vector-quantized tokenizer, the pixel-space discretization error is bounded by the cross-entropy of any code prior plus the spectrum of the reconstruction residual. At the code's own resolution, this is the code entropy plus at most $\frac{1}{2}$ nat. Hence diffusion on pixels and diffusion on the code have the same discretization complexity, up to half a nat.
- **Consequences (Section 6).** We obtain TV and KL guarantees for the unsmoothed Gaussian mixture in terms of $I(J;X_0) \le H(J)$, and TV to exact discrete targets after rounding. The framework is more general than approximate-mixture analyses (Li et al., 2026): it needs only a Wasserstein or rate–distortion approximation by a discrete law, and it covers singular data (Example 1).
- **Exact experiments (Section 7).** The exact TV increases with I for every geometry and number of components tested, until it saturates, with a common slope at small I. For a fixed code embedded in dimension up to 1024, it follows I_Γ rather than d: the TV at $d = 1024$ ranges over a factor of 30 depending only on how the distortion occupies the dimensions. The linear dependence on I_Γ is not tight.

No coordinate count appears in the bounds. The information I_Γ still grows with d for incompressible data; our results identify the structural quantity that replaces d.

2 Related work

General convergence analyses decompose diffusion sampling error into initialization, time discretization, and score approximation (Chen et al., 2023b;a; Benton et al., 2024). Direct total-variation analyses that avoid Girsanov's theorem were developed by Li et al. (2024) and Li & Yan (2025). Several works replace the ambient dimension by a manifold or intrinsic dimension (Liang et al., 2025a; Potaptchik et al., 2025; Huang et al., 2026). These quantities depend on the support of the data. Ours depends on the law, and it is never larger than the metric entropy of the support at the terminal scale, up to half a nat (Corollary 3). For Gaussian mixtures, Li et al. (2026) prove dimension-free total-variation guarantees for approximate finite Gaussian mixtures, with a polylogarithmic dependence on the number of components. Gaussian mixtures also arise in the study of when the DDPM training objective permits efficient score estimation (Shah et al., 2023).

Closest to our work is Reeves & Pfister (2025). They analyze discrete-time diffusion sampling by comparison with an idealized Gaussian-convolution process, and they bound the KL divergence to the target through the I–MMSE relation, using MMSE-based step sizes. Our KL result (Theorem 1) is of the same information-theoretic type. Our contributions go beyond KL in two directions. First, we give a direct total-variation bound that is linear in the step size, which does not follow from a KL bound through Pinsker's inequality. Second, we control I_Γ by the rate–distortion function. This connects the discretization error to autoencoders, vector-quantized tokenizers, latent versus pixel diffusion, Gaussian mixtures, and singular data. Our proofs use the I–MMSE identity of Guo et al. (2005) and the localization viewpoint of Eldan (2013) and El Alaoui & Montanari (2022), in which the sampler below is the Euler scheme of the localization process Also, Aghapour et al. (2026) derive dimension-free KL guarantees involving squared Rényi entropy, or squared Shannon entropy under a concentration assumption. Here the dependence is linear in a mutual information, it holds in total variation, and it extends from discrete targets to arbitrary data through the rate–distortion function.

3 The sampler and the Gaussian channel

Let V be an $\mathbb{R}^d$-valued random variable with $R := \mathbb{E}\,\|V\|^2 < \infty$, let $G_0 \sim \mathcal{N}(0, I_d)$ be independent of V, and define the target $X_0 = V + \sigma G_0$ with $\sigma \geq 0$. The forward heat flow is $X_t = X_0 + W_t$, so that $X_t \overset{d}{=} V + \sqrt{t+\sigma^2}\, G$. Let p_t denote the law of X_t, identified with its density, and set $M_t(x) := \mathbb{E}[V \mid X_t = x]$. Tweedie's formula gives

$$M_t(x) = x + (t+\sigma^2)\nabla \log p_t(x), \tag{1}$$

so M_t is recovered from the ordinary score, and no latent-variable oracle is required.

Grid and sampler. Fix $T > \delta \geq 0$, with $\delta > 0$ when $\sigma = 0$, and a reverse grid $T = t_0 > t_1 > \cdots > t_K = \delta$. Define

$$\tau_k = t_k + \sigma^2, \quad \eta_k = \tau_k^{-1}, \quad \Delta_k = \eta_k - \eta_{k-1}, \quad h_k = \frac{\Delta_k}{\eta_{k-1}}, \quad h = \max_k h_k.$$

Thus $\eta_0 = (T+\sigma^2)^{-1}$ and $\eta_K = \Gamma = (\delta+\sigma^2)^{-1}$, and h_k is the relative increase of the SNR in step k. Given estimates $\widehat{M}_t$, obtained for instance from a learned score through (1), the posterior-mean exponential integrator is

$$\widehat{Y}_k = a_k \widehat{Y}_{k-1} + (1-a_k)\widehat{M}_{t_{k-1}}(\widehat{Y}_{k-1}) + \sqrt{\tau_k(1-a_k)}\,\xi_k, \qquad a_k = \frac{\tau_k}{\tau_{k-1}}, \tag{2}$$

where $\xi_k \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and $\widehat{Y}_0 \sim q_T = \mathcal{N}(0, (T+\sigma^2)I_d)$. We write $\widehat{P}_\delta$ for the law of $\widehat{Y}_K$. The variance $\tau_k(1-a_k)$ is that of the exact Gaussian bridge, so (2) replaces only the unknown latent by its posterior mean. The prediction error is

$$\mathcal{E}_{\text{apx}} := \sum_{k=1}^{K} \Delta_k\, \mathbb{E}\left\| M_{t_{k-1}}(X_{t_{k-1}}) - \widehat{M}_{t_{k-1}}(X_{t_{k-1}}) \right\|^2. \tag{3}$$

Section D shows that $\mathcal{E}_{\text{apx}}$ is bounded by the standard x_0-prediction loss. On the geometric grid $\eta_k = \eta_0 e^{kL/K}$, with $L = \log \frac{T+\sigma^2}{\delta+\sigma^2}$, all steps are equal and $h = e^{L/K} - 1 \approx L/K$.

The exact reverse step. For the Gaussian channel $\sqrt{\eta}\,V + G$, let

$$m(\eta) = \mathbb{E}\,\|V - \mathbb{E}[V \mid \sqrt{\eta}V + G]\|^2\,, \qquad I(\eta) = I(V; \sqrt{\eta}V + G).$$

The I–MMSE identity states that $\int_0^\eta m = 2I(\eta)$ (Guo et al., 2005). Let $C_t(x) = \mathrm{Cov}(V \mid X_t = x)$ denote the posterior covariance. Since X_t and $\sqrt{\eta}V + G$ carry the same information when $\eta = (t+\sigma^2)^{-1}$, we have $\mathbb{E}\,\mathrm{tr}\,C_t(X_t) = m(\eta)$. Given V and $X_{t_{k-1}}$, the forward process is a Brownian bridge:

$$X_{t_k} \;\big|\; (V, X_{t_{k-1}}) \;\sim\; \mathcal{N}\big(a_k X_{t_{k-1}} + (1-a_k)V,\; \tau_k(1-a_k)I_d\big). \tag{4}$$

Update (2) with the exact M is this step with the unknown V replaced by its posterior mean $M_{t_{k-1}}(X_{t_{k-1}})$. In the language of stochastic localization, it is the Euler scheme of the localization process. All our bounds measure the cost of this replacement through the posterior covariance, and hence through m and I.

4 Sampling guarantees from the information at resolution Γ

Lemma 1 (Exact MMSE-area identity)**.** *The discretization energy*

$$\mathcal{E}_{\mathrm{disc}} := \sum_{k=1}^{K} \int_{t_k}^{t_{k-1}} \mathbb{E}\,\big\|M_t(X_t) - M_{t_{k-1}}(X_{t_{k-1}})\big\|^2\,\frac{dt}{(t+\sigma^2)^2}$$

measures the change of the posterior mean within each step. It equals $\sum_{k=1}^{K}\int_{\eta_{k-1}}^{\eta_k}\big(m(\eta_{k-1}) - m(\eta)\big)\,d\eta$.

Theorem 1 (KL guarantee)**.** *If* $R < \infty$ *and* $\mathcal{E}_{\mathrm{apx}} \le e_{\mathrm{apx}}$*, then for every grid*

$$\mathrm{KL}(P_\delta \,\|\, \widehat{P}_\delta) \le \frac{R}{2(T+\sigma^2)} + h\Big(I_\Gamma + \frac{R}{2(T+\sigma^2)}\Big) + \frac{1}{2}e_{\mathrm{apx}}. \tag{5}$$

KL bounds of this information-theoretic type are due to Reeves & Pfister (2025). Theorem 1 states one in terms of I_Γ, for use with Section 5. The proof (Section B) uses a pathwise decomposition in which the discretization and prediction errors are orthogonal martingale increments. Lemma 1 then bounds the discretization part by $\sum_k \Delta_k(m(\eta_{k-1}) - m(\eta_k))$, and summation by parts together with I–MMSE turns this into $h\,(2I_\Gamma + \eta_0 m(\eta_0))$. The KL divergence is half of this energy plus half the prediction error, which gives (5).

For total variation, we compare one sampler step with one exact step, both started from the exact law of $X_{t_{k-1}}$. Let D_k be the total-variation distance between the law of X_{t_k} and the law obtained by applying one step of (2) with the exact M to $X_{t_{k-1}}$. Write $C = C_{t_{k-1}}(X_{t_{k-1}})$ for the posterior covariance at the start of the step.

Lemma 2 (One-step error)**.** $D_k \le \frac{1}{2}\,h_k\Delta_k\,\mathbb{E}\,\mathrm{tr}\,C + \Delta_k^2\,\mathbb{E}\,\mathrm{tr}\,\big(C^2(2I + \Delta_k C)^{-1}\big)$.

Lemma 3 (One-step MMSE drop)**.** *For* $\eta, \Delta > 0$ *and* $C = \mathrm{Cov}(V \mid \sqrt{\eta}V + G)$*,* $m(\eta) - m(\eta+\Delta) \ge \Delta\,\mathbb{E}\,\mathrm{tr}\,\big(C^2(I + \Delta C)^{-1}\big)$.

In Lemma 2, $\mathbb{E}\,\mathrm{tr}\,C = m(\eta_{k-1})$. Applying Lemma 3 with $\eta = \eta_{k-1}$ and $\Delta = \Delta_k/2$ bounds the second term by $\Delta_k\big(m(\eta_{k-1}) - m(\eta_k)\big)$. Hence

$$D_k \le \tfrac{1}{2}\,h_k\Delta_k\,m(\eta_{k-1}) + \Delta_k\big(m(\eta_{k-1}) - m(\eta_k)\big). \tag{6}$$

Summing over k, as for Theorem 1, yields the following result.

Theorem 2 (Direct total-variation guarantee)**.** *If* $R < \infty$ *and* $\mathcal{E}_{\mathrm{apx}} \le e_{\mathrm{apx}}$*, then for every grid*

$$\mathrm{TV}(P_\delta, \widehat{P}_\delta) \le \sqrt{\frac{R}{4(T+\sigma^2)}} + h\Big(3I_\Gamma + \frac{R}{T+\sigma^2}\Big) + h^2\Big(I_\Gamma + \frac{R}{2(T+\sigma^2)}\Big) + \frac{1}{2}\sqrt{e_{\mathrm{apx}}}. \tag{7}$$

Proof idea. By (4), the exact next state is a mixture of Gaussians, one for each value of V under the posterior. The sampler's next state is the single Gaussian obtained by plugging in the posterior mean. The two laws have the same mean, so the first-order error cancels. Comparing the laws reached after the step (Lemma 2) requires only a lower bound on a density ratio, and Jensen's inequality provides one at every point. The remaining second-order terms involve the posterior covariance. Its trace averages to m, and its square is controlled by the decrease of m over the step (Lemma 3). Summation by parts and I–MMSE then turn both terms into I_Γ, and a killed chain assembles the steps (Section C). Only a lower bound on the ratio is needed, so no typical set is required; such sets cost a union bound over components in Li et al. (2026).

Corollary 1 (Geometric grid)**.** *On the geometric grid with $K \geq 2L$, we have $h \leq 1.65L/K$. Consequently,*

$$\mathrm{KL}(P_\delta \| \widehat{P}_\delta) \leq \tfrac{R}{2(T+\sigma^2)} + 1.65\,\tfrac{L}{K}\Big(I_\Gamma + \tfrac{R}{2(T+\sigma^2)}\Big) + \tfrac{1}{2}e_{\mathrm{apx}},$$
$$\mathrm{TV}(P_\delta, \widehat{P}_\delta) \leq \sqrt{\tfrac{R}{4(T+\sigma^2)}} + \tfrac{L}{K}\Big(6.1\,I_\Gamma + 2.2\,\tfrac{R}{T+\sigma^2}\Big) + \tfrac{1}{2}\sqrt{e_{\mathrm{apx}}}.$$

5 The information at the terminal resolution

Theorems 1 and 2 reduce the sampling problem to bounding a single mutual information. The next result bounds it through any representation of the data.

Theorem 3 (Rate–distortion control of I_Γ)**.** *Let Z be any random element of a standard Borel space, jointly distributed with V. Then*

$$I_\Gamma \leq I(V;Z) + \tfrac{1}{2}\,\mathbb{E}\log\det\big(I_d + \Gamma\mathrm{Cov}(V \mid Z)\big). \tag{8}$$

Consequently:

(a) If V is discrete, then $I_\Gamma \leq H(V)$.

(b) If Z is discrete and $\mathbb{E}\,\|V - g(Z)\|^2 \leq D$ for some decoder g, then

$$I_\Gamma \leq H(Z) + \tfrac{d}{2}\log\big(1 + \tfrac{\Gamma D}{d}\big) \leq H(Z) + \tfrac{\Gamma D}{2}.$$

If moreover $\mathrm{rank}\,\mathrm{Cov}(V \mid Z) \leq s$ almost surely, then $I_\Gamma \leq H(Z) + \tfrac{s}{2}\log(1 + \tfrac{\Gamma D}{s})$.

(c) With the mean-square rate–distortion function $\mathcal{R}_V(D) = \inf\{I(V;Z) : \mathbb{E}\,\|V - g(Z)\|^2 \leq D\}$,

$$I_\Gamma \leq \inf_{D>0}\big\{\mathcal{R}_V(D) + \tfrac{\Gamma D}{2}\big\} \leq \mathcal{R}_V(1/\Gamma) + \tfrac{1}{2}.$$

(d) If Z is discrete and $N_z(r)$ is the r-covering number of the support of $\mathcal{L}(V \mid Z = z)$, then $I_\Gamma \leq H(Z) + \inf_{r>0}\{\mathbb{E}\log N_Z(r) + \Gamma r^2/2\}$.

Bound (8) is sharp. Its second term is exact when each conditional law $\mathcal{L}(V \mid Z = z)$ is Gaussian, and the whole bound becomes an equality in the limit of well-separated Gaussian conditional laws. Hence no bound that depends only on the rate and the conditional covariances can improve on it. Part (b) replaces the linear cost $\Gamma D/2$ of a pure mean-square argument by a logarithm once $\Gamma D \gtrsim d$, and by $\tfrac{s}{2}\log(\Gamma D/s)$ for rank-s reconstruction errors. Part (c) needs no discrete code and no density: with $\sigma = 0$ and $\Gamma = \varepsilon^{-2}$, sampling $X_0 + \varepsilon G$ costs about $(\mathcal{R}_{X_0}(\varepsilon^2) + \tfrac{1}{2})\log(T/\varepsilon^2)/K$. For data near an s-dimensional set, $\mathcal{R}_{X_0}(\varepsilon^2) \approx \tfrac{s}{2}\log(\mathrm{diam}^2/\varepsilon^2)$, which recovers adaptation to the intrinsic dimension without any manifold assumption. Corollary 3 shows that the bound is never worse than the metric entropy of the support, and can be much better. Table 1 collects these cases.

Corollary 2 (Arbitrary data, $\sigma = 0$)**.** *Let $\sigma = 0$ and $V = X_0$, where the law of X_0 is arbitrary with $R = \mathbb{E}\,\|X_0\|^2 < \infty$. No density, bounded support, or mixture structure is assumed. Fix a resolution $\varepsilon > 0$, let $\delta = \varepsilon^2$ and $P_\varepsilon = \mathcal{L}(X_0 + \varepsilon G)$, and use the geometric grid with $L = \log(T/\varepsilon^2)$ and $K \geq 2L$. Then*

$$\mathrm{KL}(P_\varepsilon \| \widehat{P}_\varepsilon) \leq \frac{R}{2T} + 1.65\,\frac{L}{K}\Big(\mathcal{R}_{X_0}(\varepsilon^2) + \frac{1}{2} + \frac{R}{2T}\Big) + \frac{1}{2}e_{\mathrm{apx}},$$
$$\mathrm{TV}(P_\varepsilon, \widehat{P}_\varepsilon) \leq \sqrt{\frac{R}{4T}} + \frac{L}{K}\Big(6.1\big(\mathcal{R}_{X_0}(\varepsilon^2) + \tfrac{1}{2}\big) + 2.2\,\frac{R}{T}\Big) + \frac{1}{2}\sqrt{e_{\mathrm{apx}}}.$$

Corollary 2 is the form of our results for plain Gaussian diffusion on raw data. The number of steps needed to sample at resolution ε is set by the number of nats needed to describe X_0 to within mean-square error ε^2. The ambient dimension enters only through R, which appears in the initialization term and inside the logarithm L.

Corollary 3 (Never worse than metric entropy)**.** *Let $\mathcal{S} \supseteq \operatorname{supp}(V)$ be totally bounded, with r-covering number $N(r)$. Let Z be the index of the nearest centre of a minimal r-cover, with $r = \Gamma^{-1/2}$. Then*

$$I_\Gamma \le H(Z) + \tfrac{1}{2} = \log N(\Gamma^{-1/2}) - \mathrm{KL}\big(\mathcal{L}(Z) \,\big\|\, \mathrm{Unif}\big) + \tfrac{1}{2} \le \log N(\Gamma^{-1/2}) + \tfrac{1}{2}.$$

Comparison with intrinsic dimension and metric entropy. Analyses that adapt to low-dimensional structure (Liang et al., 2025a; Potaptchik et al., 2025; Huang et al., 2026) measure the data through its support. They use an intrinsic dimension k, or covering numbers whose logarithm at scale ε is about $k \log(\mathrm{diam}/\varepsilon)$. By Corollary 3, our complexity never exceeds this metric entropy at the terminal scale by more than half a nat. It is smaller by exactly $\mathrm{KL}(\mathcal{L}(Z)\|\mathrm{Unif})$, which measures how unevenly the data spread their mass over the cover. Support-based quantities cannot see this. Two examples show how large the gap can be.

- *Unbalanced weights.* Take m atoms with weights $1-p$ and $p/(m-1)$. The metric entropy is $\log m$ for every p, whereas $I_\Gamma \le H = h(p) + p\log(m-1) \to 0$ as $p \to 0$.
- *A rare full-dimensional part.* With probability $1-p$ the data lie on an s-dimensional set of diameter Δ, and with probability p they fill a ball of radius ϱ in $\mathbb{R}^d$. The support is d-dimensional, so support-based bounds pay about $d\log(\varrho\sqrt{\Gamma})$. Coding the part first and then the cell gives $I_\Gamma \lesssim \log 2 + \frac{1}{2} + (1-p)\, s \log(\Delta\sqrt{\Gamma}) + p\, d \log(3\varrho\sqrt{\Gamma})$. The full dimension is paid only with weight p.

In Figure 1(a), the support of the centres, and hence every support-based complexity of the latent, is fixed within each family. Yet the exact TV changes by a factor of up to 35 as the weights vary, and it follows I.

Corollary 4 (Vector-quantized autoencoders)**.** *Let $\sigma = 0$ and $V = X_0$. Let E be an encoder into a countable code set, for example n tokens from a codebook of size S, and let g be a decoder. Set $Z = E(X_0)$, $r = X_0 - g(Z)$, and $\Sigma_{\mathrm{res}} = \mathbb{E}[rr^\top]$. Then, for every probability mass function q on codes,*

$$I_\Gamma \;\le\; \mathbb{E}[-\log q(Z)] + \tfrac{1}{2}\textstyle\sum_{i=1}^{d} \log\big(1 + \Gamma\lambda_i(\Sigma_{\mathrm{res}})\big).$$

In particular, with q uniform on n tokens, $I_\Gamma \le n\log S + \frac{\Gamma}{2}\operatorname{tr}\Sigma_{\mathrm{res}}$.

All terms in Corollary 4 can be estimated from a trained tokenizer. The first term is the cross-entropy of a code prior, such as the transformer trained on the codes. The second is a principal-component analysis of the reconstruction residuals. The diffusion sampler runs in data space and need not know the autoencoder: the guarantee adapts to the best autoencoder that exists. The next corollary makes this explicit.

Corollary 5 (Pixel-space diffusion needs only the latent)**.** *Let $\sigma = 0$ and $V = X_0$, the pixels. Let (E, g) be any encoder–decoder pair, with code $Z = E(X_0)$ in a countable set, residual covariance Σ_{res}, and total reconstruction error $D = \operatorname{tr}\Sigma_{\mathrm{res}}$. Run the pixel-space sampler (2) to the code's resolution $\delta = D$ on the geometric grid, with $L = \log(T/D)$ and $K \ge 2L$. Then*

$$\mathrm{TV}(P_D, \widehat{P}_D) \le \sqrt{\frac{R}{4T}} + \frac{L}{K}\Big(6.1\big(H(Z) + \tfrac{1}{2}\big) + 2.2\,\frac{R}{T}\Big) + \frac{1}{2}\sqrt{e_{\mathrm{apx}}},$$

and the same KL bound as in Corollary 2 holds with $H(Z)$ in place of $\mathcal{R}_{X_0}(\varepsilon^2)$. Here $H(Z)$ may be replaced by the cross-entropy $\mathbb{E}[-\log q(Z)]$ of any code prior q, and the bound holds for all pairs (E, g) simultaneously. At a finer resolution $\delta < D$, the term $H(Z) + \frac{1}{2}$ becomes $H(Z) + \frac{1}{2}\sum_i \log\big(1 + \lambda_i(\Sigma_{\mathrm{res}})/\delta\big)$.

Remark 1 (Pixel space versus latent space)**.** Corollary 5 is the main conceptual message of the paper. A diffusion model trained and sampled directly on pixels, with no autoencoder, has a discretization guarantee no larger than the information of the best latent code of the data, plus half a nat at that code's resolution. Several features of this statement deserve emphasis.

- *The sampler is latent-agnostic.* It never sees (E, g). Since the bound holds for every encoder–decoder pair, pixel-space diffusion automatically inherits the most compressive representation that exists, including ones that no practitioner has built.
- *The pixel count does not enter.* The dimension d appears only through R, in the initialization term and in $L = \log(T/D)$. Both R and D are totals over pixels. With $T \propto R$ and $\delta = D$, the ratio T/D depends only on per-pixel quantities, so L is dimension-free as well.
- *The guarantee essentially matches that of latent diffusion.* Diffusion on the code itself, embedded in Euclidean space and followed by rounding (Corollary 7), has discretization term $6.1\, H(Z)\, L'/K$ by Theorems 2 and 3(a). The information complexities of the two approaches, $H(Z) + \frac{1}{2}$ and $H(Z)$, therefore differ by at most half a nat. Whatever practical advantage latent diffusion has must come from elsewhere: easier score learning, cheaper network evaluations, or a perceptually better choice of what to resolve.
- *Finer detail has a transparent price.* Resolving detail beyond the code's resolution costs $\frac{1}{2}\sum_i \log(1 + \lambda_i/\delta)$, one term per principal direction of the reconstruction residual. This is an effective dimension of what the code misses, not of the image.

Three caveats apply. The statement concerns discretization only; learning a pixel-space score is contained in e_{apx}. The target is the data blurred at the chosen resolution, which is unavoidable for singular data. Both sides of the comparison are upper bounds.

Table 1: The information at resolution Γ for several data models. Each entry, multiplied by a constant times L/K, bounds both the KL and the total-variation discretization errors (Corollary 1).

Structure of the data ($\sigma = 0, V = X_0$)	Bound on I_Γ	Source
Discrete, entropy H	H	Theorem 3(a)
Discrete code Z, reconstruction MSE D	$H(Z) + \frac{d}{2}\log(1 + \Gamma D/d)$	Theorem 3(b)
Same, with rank-s reconstruction error	$H(Z) + \frac{s}{2}\log(1 + \Gamma D/s)$	Theorem 3(b)
Arbitrary law	$\mathcal{R}_{X_0}(1/\Gamma) + \frac{1}{2}$	Theorem 3(c)
Near an s-dimensional set	$\approx \frac{s}{2}\log(\Gamma\,\mathrm{diam}^2)$	Theorem 3(c)
Support with covering number $N(r)$	$\log N(\Gamma^{-1/2}) - \mathrm{KL}(\mathcal{L}(Z)\|\mathrm{Unif}) + \frac{1}{2}$	Corollary 3
Full-dimensional density	$\approx \frac{d}{2}\log\Gamma$ (unavoidable)	tight for Gaussians
VQ tokenizer, residual spectrum (λ_i)	code cross-entropy $+\frac{1}{2}\sum_i \log(1 + \Gamma\lambda_i)$	Corollary 4
Gaussian mixture, $\sigma > 0, V = \mu_J$	$I(J; X_0) \le H(J)$	Corollary 6

6 Consequences

Gaussian mixtures. Let $X_0 = \mu_J + \sigma G_0$ with $\sigma > 0$ and a possibly countable label J. Take $V = \mu_J$. Since $t + \sigma^2 \ge \sigma^2$, the sampler may run to $\delta = 0$, and then P_0 is the unsmoothed mixture.

Corollary 6 (Entropy-adaptive mixture sampler)**.** *Let $R = \mathbb{E}\,\|\mu_J\|^2 < \infty$ and $\delta = 0$, so that $\Gamma = \sigma^{-2}$, and let $I_J := I(J; X_0) \le H(J)$. Then $I_\Gamma \le I_J$, and the bounds of Theorems 1 and 2 and Corollary 1 hold for the unsmoothed target P_0 with I_J in place of I_Γ and $L = \log(1 + T/\sigma^2)$. In particular, on the geometric grid with $K \ge 2L$,*

$$\mathrm{TV}(P_0, \widehat{P}_0) \le \sqrt{\tfrac{R}{4(T+\sigma^2)}} + \tfrac{L}{K}\left(6.1\, I_J + 2.2\, \tfrac{R}{T+\sigma^2}\right) + \tfrac{1}{2}\sqrt{e_{\mathrm{apx}}}.$$

If $H(J) = 0$, the discretization term vanishes.

The corollary needs no separation, no lower bound on the weights, and no bounded number of components. The mutual information I_J can be much smaller than $H(J)$ when components overlap, and $H(J)$ can be much smaller than $\log K_{\mathrm{mix}}$ when the weights are unbalanced.

Exact discrete targets. When $\sigma = 0$ and X_0 is discrete, every continuous sampler is at total variation 1 from X_0. A final rounding step removes this obstruction.

Corollary 7 (Rounding). *Let X_0 take values in a finite set $\mathcal{A}$ with minimal separation $\Delta_{\min}$. Let $f(y)$ be the point of $\mathcal{A}$ nearest to $\widehat{M}_\delta(y)$, and let $\varepsilon_\delta^2 = \mathbb{E}\left\|\widehat{M}_\delta(X_\delta) - M_\delta(X_\delta)\right\|^2$. Then*

$$\mathrm{TV}\big(\mathcal{L}(X_0), \mathcal{L}(f(\widehat{Y}_K))\big) \le \mathrm{TV}(P_\delta, \widehat{P}_\delta) + \frac{8}{\Delta_{\min}^2}\Big(4\delta\, H\big(X_0 \mid (2\delta)^{-1/2}X_0 + G\big) + \varepsilon_\delta^2\Big).$$

The first term is bounded by Theorem 2 with $I_\Gamma \le H(X_0)$.

The conditional entropy decays exponentially in $\Delta_{\min}^2/\delta$ for separated atoms, so δ can be taken polynomially small at only a logarithmic cost in L. This covers Gaussian diffusion on embedded codes and language tokens (Rombach et al., 2022; Li et al., 2022), where $H(X_0) \le n \log S$ for n tokens over a vocabulary of size S. Discrete-state guarantees, by contrast, scale polynomially in S (Liang et al., 2025b).

Comparison with approximate-mixture analyses. Li et al. (2026) prove dimension-free TV guarantees when the data are close *in total variation* to an isotropic Gaussian mixture with unit variance. Their discretization term is $\log^2(K_{\mathrm{mix}}N)\log^2 N/N$ for K_{mix} components and N steps. Our results are more general in two ways. For exact mixtures, Corollary 6 depends on $I(J; X_0) \le H(J)$ instead of $\log^2 K_{\mathrm{mix}}$, and it allows countably many components. More importantly, Theorem 3 needs no mixture structure: a Wasserstein, or rate–distortion, approximation by a discrete law suffices. If $W_2(\mathcal{L}(V), \nu) \le \varepsilon$ for a discrete law ν, then taking $Z \sim \nu$ optimally coupled with V in Theorem 3(b) gives $I_\Gamma \le H(\nu) + \frac{d}{2}\log(1 + \Gamma\varepsilon^2/d)$. Such an approximation always exists, whereas total-variation approximation by Gaussian mixtures is impossible for singular data.

Example 1 (Singular data approximated in Wasserstein distance). Let X_0 be uniform on segment $S_j = \{\mu_j + su_j : |s| \le \varepsilon/2\} \subset \mathbb{R}^d$, chosen with probability π_j, $j \le m$, with arbitrary centres μ_j and unit directions u_j. The law is singular, so its TV distance to every Gaussian mixture is 1. With $Z = \mu_J$, we have $W_2^2(\mathcal{L}(X_0), \mathcal{L}(Z)) \le \mathbb{E}\,\|X_0 - Z\|^2 = \varepsilon^2/12$, and $\mathrm{Cov}(X_0 \mid Z)$ has rank one. Hence (8) gives, in every dimension, $I_\Gamma \le H(J) + \frac{1}{2}\log(1 + \Gamma\varepsilon^2/12)$. With $\sigma = 0$, Corollary 1 then gives a d-free TV bound with $6.1\frac{L}{K}[H(J) + \frac{1}{2}\log(1 + \frac{\varepsilon^2}{12\delta})]$ as discretization term.

7 Experiments: Does the Total Variation Follow the Latent Information?

We test the two predictions behind the main message. First, the error is driven by the latent information, whatever the geometry or the number of components. Second, it does not grow with the ambient dimension when the latent information does not. All runs use the exact posterior mean and the exact p_T, so the error is pure time discretization. TV is computed exactly from the output density, with no density fitting (Section G).

Setup. In (a), $X_0 = \mu_J + 0.5\,G_0 \in \mathbb{R}$ is run with the mixture sampler ($\sigma = 0.5$, $\delta = 0$, $T = 50$, $K = 32$). There are five families of centres, with $K_{\mathrm{mix}} \in \{2, 3, 8, 32\}$ and separated, overlapping, or irregular placement, and only the weights vary within a family. In (b), $X_0 = (U, E_2, \ldots, E_d) \in \mathbb{R}^d$, where U is uniform on the eight atoms $\{3j - 10.5\}_{j=0}^{7}$ and the $E_i \sim \mathcal{N}(0, v_i)$ are independent. We use $\sigma = 0$, $\delta = 10^{-3}$, $T = 50$ and $K = 73$. There are three distortion families: low-rank ($v_2 = v_3 = 0.005$ and $v_i = 0$ for $i \ge 4$), fixed total ($v_i = 0.01/(d-1)$), and per-pixel ($v_i = 0.005$). Further details are in Section G.

TV grows with the latent information, whatever the geometry. In Figure 1(a), the exact TV increases with $I(J; X_0)$ in every family until it saturates, and near saturation it varies by less than 3%. For small I the curves nearly coincide: TV/I lies between 0.036 and 0.041 in all five families, although K_{mix} ranges from 2 to 32. For larger I the curves saturate, whereas the bound of Theorem 2 is linear in I.

Dimension is misleading; the latent is not. In Figure 1(b), all three families share the same code and the same dimension.

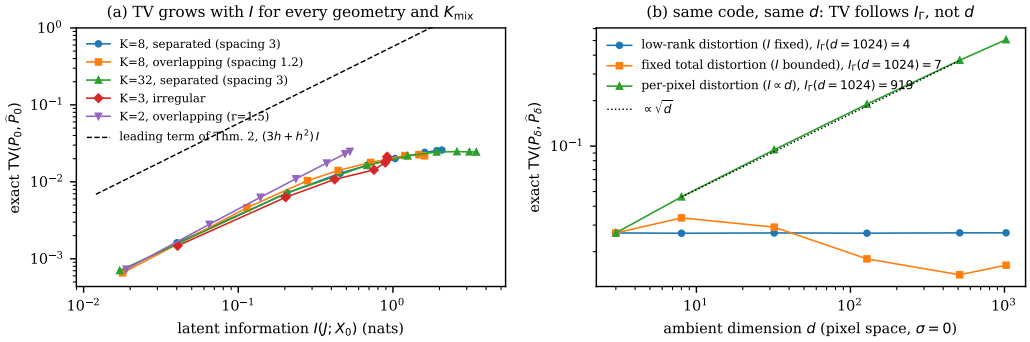


Figure 1: Exact TV of the posterior-mean exponential integrator. (a) Five mixture families, in which only the weights vary. (b) Pixel-space diffusion in $\mathbb{R}^d$ of one eight-component code, with three kinds of distortion.

- With low-rank distortion, $I_\Gamma = 3.9$ and $\mathrm{TV} = 0.027$ at every d from 3 to 1024.
- With fixed total distortion, $I_\Gamma \le 7.1$ and $\mathrm{TV} \le 0.034$.
- With per-pixel distortion, $I_\Gamma \propto d$ and $\mathrm{TV} \propto \sqrt{d}$.

At $d = 1024$ the TVs are 0.016, 0.027 and 0.51. The two full-dimensional families even share the same support, and hence the same intrinsic dimension $d - 1$, yet their TVs differ by a factor of 30. The error is predicted neither by the dimension nor by the support, but by the information at resolution Γ.

8 Limitations and Conclusion

The bounds are informative only when the data are compressible at the chosen resolution. For full-dimensional smooth data, $I_\Gamma \approx \frac{d}{2} \log \Gamma$. The target is the smoothed law P_δ, unless $\sigma > 0$ or rounding applies. The sampler is the posterior-mean exponential integrator (2), not DDPM with variance $1 - \alpha_t$. The bounds are probably not tight. They are linear in I_Γ, whereas our experiments suggest that the true dependence is sublinear: the exact TV saturates for large I, and it grows more slowly than I_Γ when the information is spread over independent directions. The experiments have product structure and use exact scores.

In summary, the reverse process needs to resolve only the information that the data carry at the terminal resolution. That information is bounded by the rate–distortion function, exceeds the metric entropy of the support by at most half a nat, and is bounded through any code together with its reconstruction error. A diffusion model run on pixels therefore pays only for the latent structure of the data (Remark 1).

AI use statement

Generative AI tools were used to assist with mathematical exposition, the derivation and checking of arguments, LaTeX restructuring, and numerical sanity checks. The authors reviewed the resulting arguments, calculations, citations, and code, and take responsibility for the complete contents of the paper.

References

Ahmad Aghapour, Erhan Bayraktar, and Ziqing Zhang. Entropy-based dimension-free convergence and loss-adaptive schedules for diffusion models. *arXiv preprint arXiv:2601.21943*, 2026.

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.

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, 2023a.

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

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

Ahmed El Alaoui and Andrea Montanari. An information-theoretic view of stochastic localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426, 2022.

Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23:532–569, 2013.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Robert M. Gray. *Entropy and Information Theory*. Springer, 2nd edition, 2011.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. doi: 10.1109/TIT.2005.844072.

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *Mathematics of Operations Research*, 2026.

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*, 2025.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4VGEeER6W9.

Gen Li, Changxiao Cai, and Yuting Wei. Dimension-free convergence of diffusion models for approximate gaussian mixtures. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=HMu24dTKkJ.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*, 2025a.

Yuchen Liang, Yingbin Liang, Lifeng Lai, and Ness Shroff. Discrete diffusion models: Novel analysis and new sampler guarantees. *arXiv preprint arXiv:2509.16756*, 2025b.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4668–4685. PMLR, 30 Jun–04 Jul 2025.

Galen Reeves and Henry D. Pfister. Information-theoretic proofs for diffusion sampling. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6. IEEE, 2025. doi: 10.1109/ISIT63088.2025.11195404.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

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

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

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024.

A Preliminaries

Throughout, $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$, φ_s denotes the density of $\mathcal{N}(0, sI_d)$, and $\succeq$ is the Loewner order. The only standing assumption is $R = \mathbb{E}\,\|V\|^2 < \infty$.

Lemma 4 (Posterior facts)**.** *Fix t with $\tau = t + \sigma^2 > 0$. The density $p_t(x) = \mathbb{E}\,\varphi_\tau(x - V)$ is smooth and positive. The posterior of V given $X_t = x$ is $\pi_x(dv) \propto \varphi_\tau(x - v)\,\mu_V(dv)$, and it has moments of all orders. Moreover, M_t is smooth, with*

$$\nabla M_t(x) = C_t(x)/\tau,$$

and $\mathbb{E}\,\mathrm{tr}\, C_t(X_t) = m(1/\tau)$.

Proof. Write $\varphi_\tau(x - v) = \varphi_\tau(x)\, e^{\langle x, v\rangle/\tau - \|v\|^2/(2\tau)}$. The Gaussian factor $e^{-\|v\|^2/(2\tau)}$ dominates every polynomial in v, locally uniformly in x, so all x-derivatives may be taken under the expectation. This gives smoothness and the moment bounds. The posterior is an exponential family in the parameter x/τ, so differentiating its mean in x gives its covariance divided by τ. Finally, $X_t/\sqrt{\tau} \overset{d}{=} \sqrt{1/\tau}\, V + G$ jointly with V, so conditioning on X_t is the same as conditioning on the channel output at SNR $1/\tau$. □

Lemma 5 (Brownian bridge)**.** *For $s < t$, put $a = \tau_s/\tau_t$ with $\tau_s = s + \sigma^2$ and $\tau_t = t + \sigma^2$. Conditionally on (V, X_t),*

$$X_s \sim \mathcal{N}\big(aX_t + (1-a)V,\ \tau_s(1-a)I_d\big).$$

The centred noise $X_s - aX_t - (1-a)V$ is independent of (V, X_t).

Proof. Given V, we have $X_s = V + \sqrt{\tau_s}\, G$ and $X_t = X_s + \sqrt{t-s}\, G'$, with G and G' independent. So (X_s, X_t) is jointly Gaussian given V. The Gaussian conditioning formulas give mean $V + \frac{\tau_s}{\tau_t}(X_t - V)$ and covariance $\tau_s - \tau_s^2/\tau_t = \tau_s(1-a)$. Since this conditional law depends on (V, X_t) only through its mean, the centred noise is independent of (V, X_t). □

Lemma 6 (I–MMSE)**.** *For every $\eta > 0$, $\int_0^\eta m(u)\,du = 2I(\eta)$ (Guo et al., 2005), and $I(\eta) \le \frac{1}{2}\log\det(I + \eta \mathrm{Cov} V) < \infty$. Moreover, m is nonincreasing.*

Proof. The bound on $I(\eta)$ is (8) with Z constant, proved in Section E. Monotonicity follows from Lemma 3. □

A.1 Proof of Lemma 3

Raising the SNR from η to $\eta + \Delta$ is the same as receiving one extra observation of V at SNR Δ, with noise independent of everything else. The MMSE drops by the energy of the correction that this observation brings to the posterior mean. Already the best *linear* use of the extra observation achieves the claimed amount.

Proof of Lemma 3. Step 1: two nested observations. Let $\tau = 1/\eta$ and $\tau' = 1/(\eta + \Delta)$, so that $\tau' < \tau$. Let G and G' be standard Gaussian vectors, independent of each other and of V, and put

$$X' = V + \sqrt{\tau'}\, G, \qquad X = X' + \sqrt{\tau - \tau'}\, G'.$$

Then $X = V + \sqrt{\tau}\,\tilde{G}$ with $\tilde{G}$ standard Gaussian and independent of V. So $X/\sqrt{\tau}$ and $X'/\sqrt{\tau'}$ are channel outputs at SNR η and $\eta + \Delta$. With

$$\theta := \mathbb{E}[V \mid X], \qquad \theta' := \mathbb{E}[V \mid X'],$$

we therefore have

$$m(\eta) = \mathbb{E}\,\|V - \theta\|^2, \qquad m(\eta + \Delta) = \mathbb{E}\,\|V - \theta'\|^2, \qquad C = \mathrm{Cov}(V \mid X).$$

Moreover X is X' plus noise that is independent of (V, X'). So once X' is known, X carries no further information on V, and $\mathbb{E}[V \mid X, X'] = \theta'$.

Step 2: the MMSE drop is the energy of the update. By the tower property, $\theta = \mathbb{E}[\theta' \mid X]$. Write $V - \theta = (V - \theta') + (\theta' - \theta)$. The first term is orthogonal in L^2 to every square-integrable function of (X, X'), and in particular to $\theta' - \theta$. Hence

$$m(\eta) = m(\eta + \Delta) + \mathbb{E}\,\|\theta' - \theta\|^2.$$

Since $\mathbb{E}[\theta' \mid X] = \theta$, this gives

$$m(\eta) - m(\eta + \Delta) = \mathbb{E}\,\|\theta' - \theta\|^2 = \mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(\theta' \mid X). \tag{9}$$

It remains to show that $\mathrm{Cov}(\theta' \mid X) \succeq \Delta\, C^2 (I + \Delta C)^{-1}$.

Step 3: given X, the finer observation is a fresh observation at SNR Δ. The Gaussian conditioning in the proof of Lemma 5 uses only $\tau' < \tau$, so it applies to (X', X) with $a = \tau'/\tau$. Conditionally on (V, X),

$$X' \sim \mathcal{N}\big(aX + (1 - a)V,\ \tau'(1 - a) I_d\big),$$

and the centred noise is independent of (V, X). Define $Y := (X' - aX)/(1 - a)$. Then $Y = V + \zeta$, where ζ is centred Gaussian, independent of (V, X), with covariance $\frac{\tau'}{1-a} I_d$. Moreover

$$\frac{\tau'}{1 - a} = \frac{\tau\tau'}{\tau - \tau'} = \Big(\frac{1}{\tau'} - \frac{1}{\tau}\Big)^{-1} = \frac{1}{\Delta}.$$

Since (X, X') and (X, Y) determine each other, $\theta' = \mathbb{E}[V \mid X, Y]$. Fix x, and let $\mathbb{E}_x$ and Cov_x denote conditional expectation and covariance given $X = x$. Under $\mathbb{E}_x$, the vector V has mean $\mu := \theta(x)$ and covariance $C = C(x)$. The observation is $Y = V + \Delta^{-1/2}\xi$, with ξ standard Gaussian and independent of V. So θ' is the Bayes estimator of V from Y.

Step 4: comparison with a linear estimator. For every $u \in \mathbb{R}^d$, the variable $u^\top \theta'$ is the best square-integrable predictor of $u^\top V$ from Y. Hence the Bayes error matrix is smallest in the Loewner order:

$$\mathbb{E}_x\big[(V - \theta')(V - \theta')^\top\big] \preceq \mathbb{E}_x\big[(V - \widehat{V})(V - \widehat{V})^\top\big]$$

for every estimator $\widehat{V} = f(Y)$. Take the linear estimator

$$\widehat{V} = \mu + K(Y - \mu), \qquad K = C\big(C + \Delta^{-1} I\big)^{-1}.$$

Then $V - \widehat{V} = (I - K)(V - \mu) - \Delta^{-1/2} K\xi$. The two terms are independent, with covariances C and I for $V - \mu$ and ξ. So, with $S := C + \Delta^{-1} I$,

$$\mathbb{E}_x\big[(V - \widehat{V})(V - \widehat{V})^\top\big] = (I - K) C (I - K)^\top + \Delta^{-1} K K^\top = C - C S^{-1} C = C(I + \Delta C)^{-1}.$$

All matrices here are functions of C, so they commute. The second equality expands the product and uses $KSK^\top = CS^{-1}C$. The third uses $C - CS^{-1}C = CS^{-1}(S - C) = \Delta^{-1}CS^{-1} = C(I + \Delta C)^{-1}$.

Step 5: conclusion. Since $V - \theta'$ is orthogonal to $\theta' - \mu$, the law of total covariance reads $C = \mathbb{E}_x[(V - \theta')(V - \theta')^\top] + \mathrm{Cov}_x(\theta')$. With Step 4,

$$\mathrm{Cov}_x(\theta') \succeq C - C(I + \Delta C)^{-1} = \Delta\, C^2 (I + \Delta C)^{-1}.$$

Take traces, then the expectation over X, and use (9). □

B Proof of the KL guarantee

B.1 Pathwise KL and orthogonal error decomposition

On $s \in (s_{k-1}, s_k]$, where $s_k := T - t_k$, the continuous-time interpolation of (2) is the exponential integrator

$$d\widetilde{Y}_s = \frac{\widehat{M}_{t_{k-1}}(\widetilde{Y}_{s_{k-1}}) - \widetilde{Y}_s}{T - s + \sigma^2}\, ds + dB_s. \tag{10}$$

The exact reverse process is

$$dY_s = \frac{M_{T-s}(Y_s) - Y_s}{T - s + \sigma^2}\, ds + dB_s, \qquad 0 \le s \le T - \delta, \tag{11}$$

started from p_T. Let P and $\widehat{\mathsf{P}}$ be the path laws of (11) and of (10), the latter started from q_T.

Since $p_T = \int \mathcal{N}(v, (T + \sigma^2) I_d)\, \mu_V(dv)$, convexity of relative entropy gives

$$\mathrm{KL}(p_T \| q_T) \le \frac{R}{2(T + \sigma^2)}. \tag{12}$$

For $s \in (s_{k-1}, s_k]$ define

$$A_s := \frac{M_{T-s}(Y_s) - M_{t_{k-1}}(Y_{s_{k-1}})}{T - s + \sigma^2}, \qquad B_s := \frac{M_{t_{k-1}}(Y_{s_{k-1}}) - \widehat{M}_{t_{k-1}}(Y_{s_{k-1}})}{T - s + \sigma^2}.$$

Their sum is the drift difference evaluated along the exact path.

To see that they are orthogonal, let $\mathcal{G}_s := \sigma(X_u : T - s \le u \le T)$. The future Brownian increments are conditionally independent of V given X_{T-s}. Hence $M_{T-s}(X_{T-s}) = \mathbb{E}[V \mid \mathcal{G}_s]$ is a square-integrable $(\mathcal{G}_s)$-martingale. Therefore $\mathbb{E}[A_s \mid \mathcal{G}_{s_{k-1}}] = 0$, while B_s is $\mathcal{G}_{s_{k-1}}$-measurable, and so $\mathbb{E}\langle A_s, B_s\rangle = 0$. Integrating gives

$$\mathbb{E}_{\mathsf{P}} \int_0^{T-\delta} \|A_s + B_s\|^2\, ds = \mathcal{E}_{\mathrm{disc}} + \mathcal{E}_{\mathrm{apx}}, \tag{13}$$

where

$$\mathcal{E}_{\mathrm{disc}} = \sum_k \int_{s_{k-1}}^{s_k} \frac{\mathbb{E}\left\|M_{T-s}(Y_s) - M_{t_{k-1}}(Y_{s_{k-1}})\right\|^2}{(T - s + \sigma^2)^2}\, ds.$$

For the second term, since $Y_{s_{k-1}} \sim X_{t_{k-1}}$ and $\int_{s_{k-1}}^{s_k} (T - s + \sigma^2)^{-2} ds = \Delta_k$, the prediction part equals (3).

The entropy identity for diffusions with identical diffusion coefficient (Girsanov's theorem with localization, see Chen et al., 2023b; Benton et al., 2024) gives $\mathrm{KL}(\mathsf{P} \| \widehat{\mathsf{P}}) \le \mathrm{KL}(p_T \| q_T) + \frac{1}{2}\mathbb{E}_{\mathsf{P}} \int \|A_s + B_s\|^2$. Data processing from paths to endpoints then yields

$$\mathrm{KL}(P_\delta \| \widehat{P}_\delta) \le \frac{R}{2(T + \sigma^2)} + \frac{1}{2}\mathcal{E}_{\mathrm{disc}} + \frac{1}{2}\mathcal{E}_{\mathrm{apx}}. \tag{14}$$

B.2 Proof of the MMSE-area identity

Let $(W_\eta)_{\eta\geq 0}$ be a Brownian motion independent of V. Set $Z_\eta = \eta V + W_\eta$ and $\mathcal{F}_\eta = \sigma(Z_u : u \leq \eta)$. The likelihood of the path depends on v only through $\exp\{\langle v, Z_\eta\rangle - \eta \|v\|^2/2\}$. Hence Z_η is sufficient, and $N_\eta := \mathbb{E}[V \mid \mathcal{F}_\eta] = \mathbb{E}[V \mid Z_\eta]$ is a square-integrable martingale. For $\eta \geq \eta_0$, the projection identity gives

$$\mathbb{E}\,\|N_\eta - N_{\eta_0}\|^2 = m(\eta_0) - m(\eta).$$

The forward process is linked to (Z_η) by time inversion. Write $X_t = V + B_{t+\sigma^2}$, where B is a Brownian motion independent of V with $B_{\sigma^2} = \sigma G_0$. Then $(\eta B_{1/\eta})_\eta$ is again a Brownian motion, so the process $\eta X_{1/\eta-\sigma^2}$, $0 < \eta \leq \sigma^{-2}$, has the same joint law with V as (Z_η). Hence $M_t(X_t) = N_\eta$ for $\eta = (t+\sigma^2)^{-1}$, jointly in t. With $\eta = (t+\sigma^2)^{-1}$, so that $d\eta = dt/(t+\sigma^2)^2$ up to orientation, each summand of $\mathcal{E}_{\text{disc}}$ becomes $\int_{\eta_{k-1}}^{\eta_k} \mathbb{E}\left\|N_\eta - N_{\eta_{k-1}}\right\|^2 d\eta = \int_{\eta_{k-1}}^{\eta_k} (m(\eta_{k-1}) - m(\eta))\, d\eta$. This proves Lemma 1.

B.3 Summation and proof of the KL theorem

Lemma 7 (Summation by parts)**.** *With $m_k = m(\eta_k)$,*

$$\sum_{k=1}^{K} \eta_{k-1}(m_{k-1} - m_k) \leq \eta_0 m_0 + 2I_\Gamma \qquad \textit{and} \qquad \sum_{k=1}^{K} \Delta_k m_{k-1} \leq 2I_\Gamma + h\,(\eta_0 m_0 + 2I_\Gamma).$$

Proof. For the first bound, rearranging the sum gives

$$\sum_{k} \eta_{k-1}(m_{k-1} - m_k) = \eta_0 m_0 - \eta_{K-1} m_K + \sum_{j=1}^{K-1} \Delta_j m_j.$$

Since m is nonincreasing, $m \geq m_j$ on $[\eta_{j-1}, \eta_j]$, so $\Delta_j m_j \leq \int_{\eta_{j-1}}^{\eta_j} m$. By Lemma 6, $\int_{\eta_0}^{\eta_{K-1}} m \leq 2I_\Gamma$.

For the second bound, $m \leq m_{k-1}$ on $[\eta_{k-1}, \eta_k]$, so $\Delta_k m_{k-1} \leq \int_{\eta_{k-1}}^{\eta_k} m + \Delta_k(m_{k-1} - m_k)$. Moreover $\Delta_k = h_k \eta_{k-1} \leq h\,\eta_{k-1}$, and the first bound applies to the remaining sum. □

Proof of Theorem 1. Since $m(\eta) \geq m_k$ on $[\eta_{k-1}, \eta_k]$, Lemma 1 gives $\mathcal{E}_{\text{disc}} \leq \sum_k \Delta_k(m_{k-1} - m_k) \leq h \sum_k \eta_{k-1}(m_{k-1} - m_k)$. By Lemma 7 this is at most $h(\eta_0 m_0 + 2I_\Gamma)$. Moreover $\eta_0 m_0 \leq R/(T+\sigma^2)$. Insert both facts into (14). □

C Proof of the total-variation guarantee

Let $\mathsf{K}_k^\star$ and $\widehat{\mathsf{K}}_k$ be the transition kernels of (2) with M and with $\widehat{M}$ respectively:

$$\mathsf{K}_k^\star(x, \cdot) = \mathcal{N}\big(a_k x + (1-a_k) M_{t_{k-1}}(x),\ \tau_k(1-a_k) I_d\big).$$

For a sub-probability density ν, write $(\nu\mathsf{K})(y) = \int \nu(x)\mathsf{K}(x,y)\,dx$ and $|\nu| = \int \nu$.

C.1 Killed chains

Proposition 1. *With $D_k := \mathrm{TV}(p_{t_{k-1}}\mathsf{K}_k^\star,\, p_{t_k})$,*

$$\mathrm{TV}(P_\delta, \widehat{P}_\delta) \leq \mathrm{TV}(p_T, q_T) + \sum_{k=1}^{K} D_k + \tfrac{1}{2}\sqrt{\mathcal{E}_{\text{apx}}}, \qquad \mathrm{TV}(p_T, q_T) \leq \sqrt{\tfrac{R}{4(T+\sigma^2)}}.$$

Proof. Set $\nu_0 = p_T \wedge q_T$ and $\nu_k = p_{t_k} \wedge (\nu_{k-1}\mathsf{K}_k^\star)$. Define the thinning functions $\theta_0 = (p_T/q_T) \wedge 1$ and $\theta_k = \big(p_{t_k}/(\nu_{k-1}\mathsf{K}_k^\star)\big) \wedge 1$, with the convention $0/0 = 0$. All densities p_{t_k} are positive and continuous.

Two killed chains. On $\mathbb{R}^d \cup \{\partial\}$, where ∂ is a cemetery state, run two chains from a common start: draw a point from q_T and keep it with probability θ_0, otherwise send it to ∂. At step k, a chain at ∂ stays there. A chain alive at x moves by $\mathsf{K}_k^\star(x,\cdot)$ in the first chain, or by $\widehat{\mathsf{K}}_k(x,\cdot)$ in the second. The new point x' is then kept with probability $\theta_k(x')$ and otherwise sent to ∂.

Alive densities. By induction, the alive sub-density of the first chain is $\nu_k \le p_{t_k}$. That of the second chain, $\widehat{\mu}_k$, satisfies $\widehat{\mu}_k \le \widehat{\mu}_{k-1}\widehat{\mathsf{K}}_k \le \mathcal{L}(\widehat{Y}_k)$. Hence

$$\mathrm{TV}(P_\delta, \widehat{P}_\delta) \le \int (p_{t_K} - \widehat{\mu}_K)_+ \le (1 - |\nu_K|) + \int (\nu_K - \widehat{\mu}_K)_+.$$

The deficit $1 - |\nu_K|$. Write $p_{t_k} - \nu_{k-1}\mathsf{K}_k^\star = (p_{t_k} - p_{t_{k-1}}\mathsf{K}_k^\star) + (p_{t_{k-1}} - \nu_{k-1})\mathsf{K}_k^\star$, where the second term is nonnegative. Integrating the positive part gives $1 - |\nu_k| \le D_k + 1 - |\nu_{k-1}|$. Moreover $1 - |\nu_0| = \mathrm{TV}(p_T, q_T)$.

The score term. The quantity $\int (\nu_K - \widehat{\mu}_K)_+$ is at most the total variation between the two path laws, which Pinsker's inequality bounds by $\sqrt{\mathrm{KL}/2}$. The two chains share their initial law and thinning kernels, and their move kernels coincide at ∂. The chain rule for relative entropy (Cover & Thomas, 2006; Gray, 2011) therefore gives

$$\mathrm{KL} \le \sum_k \int \nu_{k-1} \frac{(1-a_k)\left\|M - \widehat{M}\right\|^2}{2\tau_k} \le \tfrac{1}{2}\mathcal{E}_{\mathrm{apx}},$$

where we used $\nu_{k-1} \le p_{t_{k-1}}$ and $(1-a_k)/\tau_k = \Delta_k$.

Initialization. The bound on $\mathrm{TV}(p_T, q_T)$ follows from (12) and Pinsker's inequality. □

C.2 Proof of Lemma 2

The proof compares the two laws reached after one step, both started from the exact law $p_{t_{k-1}}$. The Gaussian noise common to both is removed first. What remains is a deterministic map applied to $X_{t_{k-1}}$, compared with a Gaussian perturbation of V. The total variation between these two laws is the expectation of $(1-\rho)_+$ for an explicit density ratio ρ. A lower bound on $\log \rho$, valid at every point, gives the lemma.

Notation. Fix k. Write $X = X_{t_{k-1}}$, with density $p = p_{t_{k-1}}$. Also write $M = M_{t_{k-1}}$, $C(x) = C_{t_{k-1}}(x)$, $\tau = \tau_{k-1}$, $a = a_k$, $h = h_k$ and $\Delta = \Delta_k$. The step size enters only through the identities

$$a = \frac{\tau_k}{\tau_{k-1}} = \frac{\eta_{k-1}}{\eta_k} = \frac{1}{1+h}, \qquad \frac{1-a}{a} = h, \qquad \frac{h}{\tau} = h\,\eta_{k-1} = \Delta. \tag{15}$$

Proof of Lemma 2. Step 1: remove the common noise. By (4), X_{t_k} has the law of

$$aX + (1-a)V + \sqrt{\tau_k(1-a)}\,\xi,$$

where $\xi \sim \mathcal{N}(0, I_d)$ is independent of (V, X). One step of (2) with the exact M, started from X, gives

$$aX + (1-a)M(X) + \sqrt{\tau_k(1-a)}\,\xi,$$

with ξ of the same law. Both outputs have the form $U + N$, where N is the same independent Gaussian noise. Adding independent noise is a Markov kernel, and total variation cannot increase under a Markov kernel. Therefore

$$D_k \le \mathrm{TV}\big(\mathcal{L}(c(X)),\, \mathcal{L}(U)\big), \qquad c(x) := ax + (1-a)M(x), \qquad U := aX + (1-a)V. \tag{16}$$

Step 2: c is a smooth bijection. By Tweedie's formula (1), $M(x) = x + \tau\nabla \log p(x)$, so

$$c(x) = x + (1-a)\tau\nabla \log p(x) = \nabla\phi(x), \qquad \phi(x) := \tfrac{1}{2}\|x\|^2 + (1-a)\tau \log p(x).$$

By Lemma 4, $\nabla M = C/\tau$, and hence

$$\nabla^2\phi(x) = \nabla c(x) = aI + (1-a)\,C(x)/\tau \succeq aI.$$

So ϕ is a-strongly convex. For every $u \in \mathbb{R}^d$, the function $x \mapsto \phi(x) - \langle u, x\rangle$ is strongly convex, so it has a unique minimizer. That minimizer is the unique solution of $c(x) = u$. Thus c is a bijection of $\mathbb{R}^d$, and its inverse is smooth by the inverse function theorem. By the change-of-variables formula, $c(X)$ has density

$$p_c(u) = \frac{p(c^{-1}(u))}{J(c^{-1}(u))}, \qquad J(x) := \det\big(aI + (1-a)C(x)/\tau\big).$$

Step 3: the density of U. Since $X = V + \sqrt{\tau}\, G$ with G independent of V, we have $U = aX + (1-a)V = V + a\sqrt{\tau}\, G$. Hence U has the positive density $q(u) = \mathbb{E}\,\varphi_{a^2\tau}(u - V)$.

Step 4: total variation as an expectation of a ratio. Since $p_c > 0$,

$$\mathrm{TV}\big(\mathcal{L}(c(X)), \mathcal{L}(U)\big) = \int \big(p_c(u) - q(u)\big)_+ du = \int \Big(1 - \frac{q(u)}{p_c(u)}\Big)_+ p_c(u)\, du.$$

Substitute $u = c(x)$, so that $du = J(x)\, dx$ and $p_c(c(x))J(x) = p(x)$. This gives

$$\mathrm{TV}\big(\mathcal{L}(c(X)), \mathcal{L}(U)\big) = \mathbb{E}\big[(1 - \rho(X))_+\big], \qquad \rho(x) := \frac{q(c(x))\, J(x)}{p(x)}. \tag{17}$$

It therefore suffices to bound ρ from below.

Step 5: an explicit formula for ρ. By Lemma 4, the posterior of V given $X = x$ is $\pi_x(dv) = \varphi_\tau(x - v)\mu_V(dv)/p(x)$. Let $\mathbb{E}_x$ denote expectation under π_x. Then

$$\frac{q(c(x))}{p(x)} = \int \frac{\varphi_{a^2\tau}(c(x) - v)}{p(x)}\, \mu_V(dv) = \mathbb{E}_x\Big[\frac{\varphi_{a^2\tau}(c(x) - V)}{\varphi_\tau(x - V)}\Big].$$

Introduce the posterior error $e := V - M(x)$ and the deterministic vector $w := x - M(x)$. They satisfy

$$\mathbb{E}_x[e] = 0, \qquad \mathbb{E}_x \|e\|^2 = \operatorname{tr} C(x), \qquad x - V = w - e, \qquad c(x) - V = aw - e.$$

The last identity holds because $c(x) - V = a(x - M(x)) + (M(x) - V)$. The ratio of the two Gaussian densities is

$$\frac{\varphi_{a^2\tau}(aw - e)}{\varphi_\tau(w - e)} = a^{-d} \exp\Big(-\frac{\|aw - e\|^2}{2a^2\tau} + \frac{\|w - e\|^2}{2\tau}\Big).$$

Expand $\|aw - e\|^2 / a^2 = \|w\|^2 - \frac{2}{a}\langle w, e\rangle + \frac{1}{a^2}\|e\|^2$ and $\|w - e\|^2 = \|w\|^2 - 2\langle w, e\rangle + \|e\|^2$. The terms in $\|w\|^2$ cancel, and the exponent equals

$$\frac{1/a - 1}{\tau}\langle w, e\rangle - \frac{1/a^2 - 1}{2\tau}\|e\|^2 = \Delta\langle w, e\rangle - \Delta\big(1 + \tfrac{h}{2}\big)\|e\|^2.$$

Here we used (15), since $1/a - 1 = h$ and $(1/a^2 - 1)/2 = h + h^2/2$. Similarly, $(1-a)/(a\tau) = h/\tau = \Delta$ gives

$$J(x) = a^d \det\big(I + \tfrac{1-a}{a\tau}\, C(x)\big) = a^d \det\big(I + \Delta C(x)\big).$$

The factors a^{-d} and a^d cancel, and we obtain

$$\rho(x) = \det\big(I + \Delta C(x)\big)\, \mathbb{E}_x\Big[\exp\Big(\Delta\langle w, e\rangle - \Delta\big(1 + \tfrac{h}{2}\big)\|e\|^2\Big)\Big]. \tag{18}$$

The determinant comes from the Jacobian of c. The expectation compares the exact Gaussian mixture with its mean-matched collapse.

Step 6: a lower bound on $\log\rho$ valid at every point. We bound the two factors of (18) separately.

- *The expectation.* By Jensen's inequality, $\log \mathbb{E}_x[e^Y] \geq \mathbb{E}_x[Y]$. Since w is fixed and $\mathbb{E}_x[e] = 0$, the linear term averages to zero. This is where the mean matching of the sampler enters. Hence

$$\log \mathbb{E}_x\Big[\exp\Big(\Delta\langle w, e\rangle - \Delta\big(1 + \tfrac{h}{2}\big)\|e\|^2\Big)\Big] \geq -\Delta\big(1 + \tfrac{h}{2}\big)\operatorname{tr} C(x).$$

- *The determinant.* For $z \geq 0$, $\log(1+z) \geq z - z^2/(2+z)$. Indeed, both sides vanish at $z = 0$, and the derivative of their difference is $z^2/\big((1+z)(2+z)^2\big) \geq 0$. Apply this to the eigenvalues $\lambda_i \geq 0$ of $\Delta C(x)$:

$$\log\det\big(I + \Delta C\big) = \sum_i \log(1+\lambda_i) \geq \sum_i \lambda_i - \sum_i \frac{\lambda_i^2}{2+\lambda_i} \\ = \Delta \operatorname{tr} C - \Delta^2 \operatorname{tr}\big(C^2(2I + \Delta C)^{-1}\big).$$

Add the two bounds. The first-order terms $\Delta \operatorname{tr} C$ cancel, leaving $-\frac{h}{2}\Delta \operatorname{tr} C$, so for every x

$$\log \rho(x) \geq -\tfrac{1}{2}h\Delta \operatorname{tr} C(x) - \Delta^2 \operatorname{tr}\big(C(x)^2(2I + \Delta C(x))^{-1}\big). \tag{19}$$

Only second-order terms in the step remain. This is the pointwise form of the first-order cancellation described after Theorem 2.

Step 7: conclusion. Since $1 - r \leq -\log r$ for every $r > 0$, and the right-hand side of (19) is nonpositive,

$$(1-\rho(x))_+ \leq (-\log\rho(x))_+ \leq \tfrac{1}{2}h\Delta \operatorname{tr} C(x) + \Delta^2 \operatorname{tr}\big(C(x)^2(2I + \Delta C(x))^{-1}\big).$$

Take the expectation over X, and combine with (16) and (17):

$$D_k \leq \mathbb{E}\big[(1-\rho(X))_+\big] \leq \tfrac{1}{2}h_k\Delta_k \,\mathbb{E}\operatorname{tr} C + \Delta_k^2 \,\mathbb{E}\operatorname{tr}\big(C^2(2I + \Delta_k C)^{-1}\big).$$

This is Lemma 2. □

C.3 Proof of the total-variation theorem

Proof of Theorem 2. By (6) and Lemma 7,

$$\sum_k D_k \leq \tfrac{h}{2}\big(2I_\Gamma + h(\eta_0 m_0 + 2I_\Gamma)\big) + h\,(\eta_0 m_0 + 2I_\Gamma) = h(3I_\Gamma + \eta_0 m_0) + h^2\big(I_\Gamma + \tfrac{1}{2}\eta_0 m_0\big).$$

Use $\eta_0 m_0 \leq R/(T+\sigma^2)$ and Proposition 1. □

Proof of Corollary 1. On the geometric grid, $h = e^{L/K} - 1$. For $x \in [0, \frac{1}{2}]$, $e^x - 1 \leq xe^x \leq 1.65x$. Hence, if $K \geq 2L$, then $h \leq 1.65L/K$ and $h \leq 0.65$. The KL bound follows directly from Theorem 1. For total variation,

$$h(3I + r) + h^2(I + r/2) \leq h(3.65\,I + 1.33\,r) \leq \frac{L}{K}(6.1\,I + 2.2\,r), \qquad r = \frac{R}{T+\sigma^2}.$$ □

D The prediction error and the x_0-prediction loss

For $t > 0$, set $D_t(x) = \mathbb{E}[X_0 \mid X_t = x] = x + t\nabla \log p_t(x)$ and $\widehat{D}_t(x) = x + t\widehat{s}_t(x)$. With $\widehat{M}_t = x + (t+\sigma^2)\widehat{s}_t$, we have

$$M_t - \widehat{M}_t = \frac{t+\sigma^2}{t}(D_t - \widehat{D}_t).$$

Since $t_k \leq t_{k-1}$,

$$\Big(\frac{t_{k-1}+\sigma^2}{t_{k-1}}\Big)^2\Big(\frac{1}{t_k+\sigma^2} - \frac{1}{t_{k-1}+\sigma^2}\Big) = \frac{(t_{k-1}-t_k)(t_{k-1}+\sigma^2)}{t_{k-1}^2(t_k+\sigma^2)} \leq \frac{1}{t_k} - \frac{1}{t_{k-1}},$$

where the last inequality is equivalent to $t_k\sigma^2 \leq t_{k-1}\sigma^2$. Therefore, when $\delta > 0$,

$$\mathcal{E}_{\mathrm{apx}} \leq \sum_k \Big(\frac{1}{t_k} - \frac{1}{t_{k-1}}\Big)\mathbb{E}\left\|D_{t_{k-1}}(X_{t_{k-1}}) - \widehat{D}_{t_{k-1}}(X_{t_{k-1}})\right\|^2,$$

which is the usual weighted x_0-prediction loss.

E PROOFS OF THE INFORMATION BOUNDS

Proof of (8). Let $Y = \sqrt{\Gamma}V + G$ with G independent of (V, Z). Since $Z \to V \to Y$ is a Markov chain, the chain rule on standard Borel spaces (Gray, 2011) gives $I(V;Y) = I(Z;Y) + I(V;Y \mid Z)$. By data processing, $I(Z;Y) \le I(Z;V)$.

For z outside a null set, $C_z = \mathrm{Cov}(V \mid Z = z)$ is finite, and Y given $Z = z$ has a density with covariance $I + \Gamma C_z$. Since adding independent noise does not decrease differential entropy, and by the Gaussian maximum-entropy inequality,

$$h(G) \le h(Y \mid Z = z) \le \tfrac{1}{2} \log \big((2\pi e)^d \det(I + \Gamma C_z)\big).$$

Hence $I(V;Y \mid Z = z) = h(Y \mid Z = z) - h(G) \le \frac{1}{2} \log\det(I + \Gamma C_z)$. Integrating over z proves (8). □

Proof of Theorem 3(a)–(c). (a) Take $Z = V$. Then $\mathrm{Cov}(V \mid Z) = 0$ and $I(V;Z) = H(V)$.

(b) For $C \succeq 0$ of rank at most r, the arithmetic–geometric mean inequality on the nonzero eigenvalues gives $\log\det(I + \Gamma C) \le r \log(1 + \Gamma \operatorname{tr} C / r)$. This function of $\operatorname{tr} C$ is concave, so Jensen's inequality applies over Z. Moreover $\mathbb{E} \operatorname{tr} \mathrm{Cov}(V \mid Z) \le \mathbb{E} \|V - g(Z)\|^2 \le D$, because the conditional mean is the best predictor. Finally $I(V;Z) \le H(Z)$. Take $r = d$ in general and $r = s$ under the rank condition. The last inequality in (b) follows from $\log(1 + x) \le x$.

(c) By $\log(1 + x) \le x$, $\frac{1}{2}\mathbb{E} \log\det(I + \Gamma C_Z) \le \frac{\Gamma}{2}\mathbb{E} \operatorname{tr} C_Z \le \frac{\Gamma D}{2}$. Take the infimum over admissible (Z, g) and then over D. Choosing $D = 1/\Gamma$ gives the last bound. □

Proof of Theorem 3(d). Fix $r > 0$. For each z, assign every point of $S_z = \operatorname{supp} \mathcal{L}(V \mid Z = z)$ measurably to the centre of an r-net, and let Q_r denote the resulting cell index. Given $Z = z$, Q_r takes at most $N_z(r)$ values. The chain rule gives

$$I(V;Y \mid Z = z) \le \log N_z(r) + I(V;Y \mid Z = z, Q_r).$$

On the cell with centre c, $\|V - c\| \le r$. Using $\mathcal{N}(\sqrt{\Gamma}c, I)$ as a reference output law bounds the last term by $\Gamma r^2/2$. Average over (Z, Q_r), use $I(Z;Y) \le H(Z)$, and minimize over r. □

Proof of Corollary 2. With $\sigma = 0$ and $\Gamma = \varepsilon^{-2}$, Theorem 3(c) gives $I_\Gamma \le \mathcal{R}_{X_0}(\varepsilon^2) + \frac{1}{2}$. Insert this bound into Corollary 1, with $T + \sigma^2 = T$ and $\delta = \varepsilon^2$. □

Proof of Corollary 3. Let $c_1, \dots, c_N$ be the centres of a minimal r-cover of $\mathcal{S}$, and let Z be the index of the centre nearest to V, with ties broken measurably. Then $\|V - c_Z\| \le r$ almost surely, so $\mathbb{E} \|V - c_Z\|^2 \le r^2 = 1/\Gamma$. Theorem 3(b) with $D = 1/\Gamma$ gives $I_\Gamma \le H(Z) + \Gamma D/2 = H(Z) + \frac{1}{2}$. Finally, $H(Z) = \log N - \mathrm{KL}(\mathcal{L}(Z) \| \mathrm{Unif}\{1, \dots, N\})$. □

Proof of the two examples after Corollary 3. For the first, $Z = V$ and $H(V) = h(p) + p \log(m - 1)$. For the second, let B indicate the part and let $Z = (B, Z_B)$, where Z_0 and Z_1 are nearest-centre indices of minimal $\Gamma^{-1/2}$-covers of the two parts. Then $H(Z) \le \log 2 + (1 - p) \log N_0 + p \log N_1$, where $\log N_1 \le d \log(3\varrho\sqrt{\Gamma})$ for $\Gamma^{-1/2} \le \varrho$, and $\log N_0 \lesssim s \log(\Delta\sqrt{\Gamma})$ for an s-dimensional set. Since $\mathbb{E} \|V - c_Z\|^2 \le 1/\Gamma$, Theorem 3(b) concludes. □

Proof of Corollary 4. Take $Z = E(X_0)$. Because the encoder is deterministic, $I(X_0; Z) = H(Z)$, and Gibbs' inequality gives $H(Z) \le \mathbb{E}[-\log q(Z)]$. Next, $\log\det$ is concave, so $\mathbb{E} \log\det(I + \Gamma C_Z) \le \log\det(I + \Gamma\, \mathbb{E} C_Z)$. The conditional mean minimizes the mean-square error matrix, so $\mathbb{E} C_Z \preceq \mathbb{E}[(X_0 - g(Z))(X_0 - g(Z))^\top] = \Sigma_{\mathrm{res}}$. Since $\log\det$ is monotone in the Loewner order, $\log\det(I + \Gamma\, \mathbb{E} C_Z) \le \sum_i \log(1 + \Gamma \lambda_i(\Sigma_{\mathrm{res}}))$. For the uniform q, $\mathbb{E}[-\log q(Z)] = n \log S$; the final bound also uses $\log(1 + x) \le x$. □

Proof of Corollary 5. By Corollary 4 with $\Gamma = 1/\delta$,

$$I_\Gamma \le H(Z) + \tfrac{1}{2} \sum_i \log\big(1 + \lambda_i(\Sigma_{\mathrm{res}})/\delta\big).$$

The same bound holds with $H(Z)$ replaced by $\mathbb{E}[-\log q(Z)]$. For $\delta = D$, $\log(1+x) \le x$ gives $\frac{1}{2}\sum_i \log(1+\lambda_i/D) \le \frac{1}{2}\sum_i \lambda_i/D = \frac{1}{2}$. Insert the resulting bound on I_Γ into Corollary 1 with $\sigma = 0$. Since the argument applies to each pair (E, g) separately, the bound holds for all pairs simultaneously. □

F PROOFS OF THE CONSEQUENCES

Proof of Corollary 6. With $V = \mu_J$ and $\delta = 0$, we have $\Gamma = \sigma^{-2}$ and $\sqrt{\Gamma}V + G = \sigma^{-1}(\mu_J + \sigma G) \overset{d}{=} \sigma^{-1}X_0$ jointly with J. Since μ_J is a function of J, data processing gives $I_\Gamma \le I(J; X_0) \le H(J)$. Theorems 1 and 2 and Corollary 1 apply because $\tau_K = \sigma^2 > 0$. If $H(J) = 0$, then V is deterministic, so $m \equiv 0$, and every D_k and $\mathcal{E}_{\text{disc}}$ vanishes. □

Proof of the claims in Example 1. Given $J = j$, we have $X_0 = \mu_j + s u_j$ with s uniform on $[-\varepsilon/2, \varepsilon/2]$. Hence $\mathbb{E}[\|X_0 - \mu_J\|^2 \mid J] = \mathbb{E}s^2 = \varepsilon^2/12$ and $\mathrm{Cov}(X_0 \mid J) = \frac{\varepsilon^2}{12} u_J u_J^\top$. Take $Z = J$ (or $Z = \mu_J$) in (8). Then $I(X_0; Z) \le H(J)$, and $\frac{1}{2}\log\det(I + \Gamma\frac{\varepsilon^2}{12} u_J u_J^\top) = \frac{1}{2}\log(1 + \Gamma\varepsilon^2/12)$. The law of X_0 is supported on a finite union of segments, a Lebesgue-null set when $d \ge 2$. Every Gaussian mixture is absolutely continuous, so the total-variation distance between the two is 1. □

Proof of Corollary 7. Apply the same map f to X_δ and to $\widehat{Y}_K$. By data processing, $\mathrm{TV}(\mathcal{L}f(X_\delta), \mathcal{L}f(\widehat{Y}_K)) \le \mathrm{TV}(P_\delta, \widehat{P}_\delta)$. Moreover, $\mathrm{TV}(\mathcal{L}(X_0), \mathcal{L}f(X_\delta)) \le \mathbb{P}(f(X_\delta) \ne X_0)$.

Rounding is correct whenever $\left\|\widehat{M}_\delta(X_\delta) - X_0\right\| < \Delta_{\min}/2$. Markov's inequality therefore gives $\mathbb{P}(f(X_\delta) \ne X_0) \le 8\big(m(\Gamma) + \varepsilon_\delta^2\big)/\Delta_{\min}^2$, with $\Gamma = 1/\delta$. Since m is nonincreasing,

$$\frac{\Gamma}{2} m(\Gamma) \le \int_{\Gamma/2}^{\Gamma} m = 2(I_\Gamma - I_{\Gamma/2}) \le 2\, H\big(X_0 \mid \sqrt{\Gamma/2}\, X_0 + G\big),$$

where we used $I_\Gamma \le H(X_0)$. Hence $m(\Gamma) \le 4\delta\, H\big(X_0 \mid (2\delta)^{-1/2} X_0 + G\big)$, which proves the claim. □

G EXPERIMENTAL DETAILS

How the TV is computed. Let q_k be the density of $\widehat{Y}_k$. For a one-dimensional target, update (2) maps q_{k-1} to

$$q_k = (c_k)_{\#} q_{k-1} * \varphi_{\tau_k(1-a_k)}, \qquad c_k(x) = a_k x + (1 - a_k) M_{t_{k-1}}(x), \tag{20}$$

where φ_s is the $\mathcal{N}(0, s)$ density. The push-forward has density $q_{k-1}(c_k^{-1}(y))/c_k'(c_k^{-1}(y))$, with $c_k' = a_k + (1 - a_k) C_{t_{k-1}}/\tau_{k-1} > 0$. We evaluate (20) on a grid of spacing 0.002, computing the convolution by FFT and starting from $q_0 = p_T$. We then report $\mathrm{TV} = \frac{1}{2}\int |q_K - p_\delta|$.

In (b) the coordinates of X_0 are independent. Then $p_t(x) = \prod_i p_t^{(i)}(x_i)$ and M_t acts coordinate-wise, so update (2) runs independently in each coordinate, and P_δ and $\widehat{P}_\delta$ are both products of one-dimensional laws. Hence

$$\mathrm{TV}(P_\delta, \widehat{P}_\delta) = \mathbb{E}_{X \sim P_\delta}\Big[\Big(1 - \prod_{i=1}^{d} r_i(X_i)\Big)_+\Big], \qquad r_i = q_K^{(i)}/p_\delta^{(i)}, \tag{21}$$

where each one-dimensional ratio r_i comes from (20). We estimate the expectation from 2×10^5 exact samples of P_δ, with standard error below 10^{-3}. Mutual informations are exact: from posterior entropies in (a), and as $I_\Gamma = \sum_i \big[h(p_\delta^{(i)}) - \frac{1}{2}\log(2\pi e\delta)\big]$ in (b).

Numerical accuracy. We evaluate the propagation (20) with linear interpolation for c_k^{-1} and an exact Gaussian multiplier in Fourier space. Halving the grid spacing of 0.002 changes the mixture-sampler TVs by less than 10^{-5} in relative terms, and the pixel-space TVs by at most 1.5%, including the products (21) at $d = 1024$. Widening the domain from $[-40, 40]$ to $[-60, 60]$ changes no reported digit. In (a), $I(J; X_0) = H(J) - \mathbb{E}H(J \mid X_0)$ is computed from the exact label posterior. In (b), each coordinate's information is computed from the differential entropy of $p_\delta^{(i)}$.

Panel (a). We use $T = 50$, $K = 32$ geometric steps in the SNR, component standard deviation 0.5, the mixture sampler ($\sigma = 0.5$) and $\delta = 0$. The families are:

- eight atoms with spacing 3;
- eight atoms with spacing 1.2;
- 32 atoms with spacing 3;
- three atoms at $\{-4, 0, 1.2\}$;
- two atoms at ± 0.75.

In the first four families the weights are proportional to $e^{-\beta k}$, with $\beta \in \{0, .25, .5, 1, 1.5, 2, 3, 5\}$ for the two eight-atom families, $\beta \in \{0, .1, .2, .4, .8, 1.5, 3, 6\}$ for 32 atoms, and $\beta \in \{0, .25, .5, 1, 2, 3, 5\}$ for three atoms. In the last family they are $(p, 1-p)$ with $p \in \{.5, .35, .2, .1, .05, .02, .005\}$.

Panel (b). We use $\sigma = 0$, $T = 50$ and $\delta = 10^{-3}$, with $K = 73$ steps, so that $h = 0.16$. Coordinate 1 carries eight atoms with spacing 3 and uniform weights. In the other coordinates the distortion is Gaussian:

- low rank: variance 0.005 in two coordinates, and exactly zero in the rest;
- fixed total energy: variance $0.01/(d-1)$ in each of the other coordinates;
- fixed per-pixel size: variance 0.005 in each of the other coordinates.

The TV is computed by (21). Coordinates with zero distortion have $r_i \equiv 1$ exactly, since for a deterministic coordinate $M_t \equiv 0$ and update (2) reproduces the exact Gaussian law of X_{t_k}.

Tables 2 and 3 list the exact values behind Figure 1. Table 2 also reports $\sum_k D_k$, the sum of the exact one-step errors in Proposition 1.

Table 2: **Exact values for the first family of Figure 1(a).** Eight separated atoms (spacing 3) with weights proportional to $e^{-\beta k}$, run with the mixture sampler ($\sigma = 0.5$, $\delta = 0$, $T = 50$, $K = 32$). The last column is the sum of the exact one-step errors D_k of Proposition 1. All values are computed exactly.

run	$I(J; X_0)$	exact TV	$\sum_k D_k$
$\beta = 0$	2.07	0.0259	0.112
$\beta = 1$	1.03	0.0201	0.083
$\beta = 2$	0.46	0.0129	0.046
$\beta = 5$	0.040	0.0016	0.0045

Table 3: **Exact values behind Figure 1(b).** Pixel-space diffusion ($\sigma = 0$, $\delta = 10^{-3}$, $T = 50$, $K = 73$) of an eight-atom code with three kinds of distortion. Each entry gives I_Γ and the exact TV. Standard errors of the TV are at most 8×10^{-4}.

	$d = 3$	$d = 8$	$d = 32$	$d = 128$	$d = 512$	$d = 1024$
low rank: I_Γ / TV	3.9 / .027	3.9 / .027	3.9 / .027	3.9 / .027	3.9 / .027	3.9 / .027
fixed total: I_Γ / TV	3.9 / .027	5.2 / .034	6.4 / .029	6.9 / .018	7.0 / .014	7.1 / .016
per pixel: I_Γ / TV	3.9 / .027	8.4 / .046	30 / .095	116 / .19	460 / .37	919 / .51