THE LOSS FLOOR OF DENOISING SCORE MATCHING: FISHER GEOMETRY FROM SCHRÖDINGER BRIDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Denoising score matching regresses onto a conditional score, while the generative dynamics require the marginal score. The two objectives share a population minimizer, but at fixed noisy state the conditional target is still random, and this inflates the training loss by a term no model can reduce. We show that this excess is exactly the trace of the Fisher–Rao metric of the conditional endpoint family, integrated along the diffusion, for any corruption kernel meeting mild regularity conditions. Deriving it from the Schrödinger bridge variational principle makes its origin precise: the first variation of the bridge functional gives the generative drift, the second gives the metric, so the information geometry observed in diffusion latent spaces is an intrinsic part of the training loss rather than a structure imposed on the model. For corruption diffusions the floor factorizes into a data-dependent information flow and a schedule-dependent weight, and for Gaussian corruption it evaluates to the mutual information the noisy state loses about the data. It is also measurable: one extra backward pass recovers it from a trained network, and on CIFAR-10 it is 99.86% of the reported loss on the noise range in standard use, large enough that two checkpoints of a single run are ranked backwards by a factor of 5.9 — an inversion that subtracting the floor repairs. The same decomposition applies to masked diffusion, where the floor is the entropy of the data and is likewise schedule-independent.

1 INTRODUCTION

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) are trained by denoising score matching (DSM) (Hyvärinen, 2005; Vincent, 2011; Song & Ermon, 2019), and in every formulation of them, including time-reversed SDEs (Anderson, 1982), variational bounds (Kingma et al., 2021), and flow matching and stochastic interpolants (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023), the same substitution is made. The generative dynamics require the *marginal* score $\mathbf{s}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log P_t(\mathbf{x})$ of the noisy data distribution, which is intractable, so training regresses instead onto the *conditional* score $\nabla_{\mathbf{x}} \log q(\mathbf{x}, t \mid \mathbf{y})$ of the corruption kernel for a clean sample $\mathbf{y}$. The conditional score is unbiased for the marginal score, so the two objectives share a population minimizer. At fixed $\mathbf{x}$, however, the conditional target is still a random variable, and sharing a minimizer does not imply sharing a loss value. The substitution inflates the objective by a term that no model can reduce. What that irreducible excess is, and what it depends on, is the subject of this work, in three contributions.

- We prove that the excess of the denoising objective over the ideal marginal-score objective is exactly the integrated trace of the Fisher–Rao metric of the conditional endpoint family, for any kernel meeting mild regularity conditions, with both terms derived from the Schrödinger bridge variational principle.
- We show that for corruption diffusions the floor factorizes into a schedule-dependent weight and a data-dependent information flow, evaluating for Gaussian corruption to a difference of mutual informations.
- We measure the floor on CIFAR-10 with a one-backward-pass estimator. It is 99.86% of the reported loss on the standard range, and subtracting it repairs a $5.9\times$ ranking inversion.

We show that the excess admits an exact geometric characterization. The fluctuation of the conditional score about its posterior mean is the score of the conditional endpoint distribution $p(\mathbf{y} \mid \mathbf{x}, t) = P_{\mathrm{data}}(\mathbf{y}) q(\mathbf{x}, t \mid \mathbf{y}) / P_t(\mathbf{x})$ with respect to the latent coordinate $\mathbf{x}$, so its second moment is by definition the Fisher information of that family (Barndorff-Nielsen, 1978; Amari & Nagaoka, 2000;

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.

Nielsen, 2020), giving an exact orthogonal decomposition of the denoising objective,

$$\Delta\mathcal{S}_{\mathrm{DSM}} = \Delta\mathcal{S}_{\mathrm{ideal}} + \frac{1}{2}\int_0^T \gamma(t)\,\mathbb{E}_{\mathbf{x}\sim P_t}[\operatorname{tr} g(\mathbf{x},t)]\,dt, \tag{1.1}$$

where $\Delta\mathcal{S}_{\mathrm{ideal}}$ is the ideal marginal-score objective and g the Fisher–Rao metric of the conditional endpoint family. The second term depends only on the corruption and the data, never on the model. It is an irreducible loss floor intrinsic to conditional denoising score matching.

The Schrödinger bridge formulation (Schrödinger, 1931; Schrödinger, 1932; Föllmer, 1988; Csiszár, 1975; Léonard, 2014) supplies the common origin of both terms. A bridge is the relative-entropy projection of a reference path measure onto endpoint constraints, with optimal dynamics given by a Doob h-transform (Doob, 1957), and bridges are by now used extensively in generative modeling (De Bortoli et al., 2021; Wang et al., 2021; Vargas et al., 2021; Chen et al., 2022; Tong et al., 2023; Shi et al., 2023). Here the same structure plays a different role. The *first* variation of the bridge functional produces the generative drift and hence $\Delta\mathcal{S}_{\mathrm{ideal}}$, while the *second* produces the quadratic form whose pullback to latent space is the Fisher metric in the floor. For affine Gaussian corruption this metric is $g_{ij} = \sigma_t^{-2}\delta_{ij} + \partial_i\partial_j \log P_t(\mathbf{x})$, precisely the Hessian/Fisher geometry recently identified in diffusion latent spaces (Karczewski et al., 2026; Lobashev et al., 2025). Equation (1.1) gives that geometry a variational origin, as the variance of a regression target already in the loss rather than a structure imposed on the model. The floor also admits an information-theoretic evaluation. For a corruption diffusion $d\mathbf{x}_t = f\,dt + \sqrt{D(t)}\,d\mathbf{W}_t$, $\mathbb{E}_{P_t}[\operatorname{tr} g] = -\frac{2}{D(t)}\frac{d}{dt}I(\mathbf{y};\mathbf{x}_t)$, so the floor separates into a data-dependent information flow and a schedule weight. When the weighting equals the diffusion coefficient the integral telescopes to a difference of mutual informations (Guo et al., 2005).

The practical consequences are large. Because the floor varies with the SNR range, schedule, and loss weighting, raw training losses are not comparable across training configurations. The factorization also separates what schedule design can and cannot affect, complementing work on information-based scheduling and noise allocation (Kingma et al., 2021; Stancevic et al., 2025; Kong et al., 2026; Zhang, 2025; Sabour et al., 2024; Williams et al., 2024; Ambrogioni et al., 2026).

studies an exact decomposition already implicit in the standard one.

2 From Path-Space Entropy to Denoising Score Matching

The training objective of diffusion models follows from the Schrödinger bridge variational principle in three steps, which we summarize here and derive in Appendix A. Reverse diffusion was studied from an action-principle perspective in Premkumar (2023), whose notation we borrow.

Let $\mathcal{G}$ be a Markovian reference path measure, discretized as $\mathcal{G}(\mathbf{x}_{0:N}) = P_0(\mathbf{x}_0)\prod_k G(\mathbf{x}_{k+1}\mid\mathbf{x}_k)$ on $N \gg 1$ segments of length $\Delta t = T/N$. $\mathcal{G}$ need be neither a probability measure nor a map from $P_0 = P_{\mathrm{prior}}$ to $P_T = P_{\mathrm{data}}$, since the endpoint constraints are imposed variationally. By Sanov's theorem (Dembo & Zeitouni, 1998; Csiszár, 1984) the empirical path measure of many independent realizations of $\mathcal{G}$ carries large-deviation weight $\exp[-nD_{\mathrm{KL}}^{\mathrm{path}}(\mathcal{H}\|\mathcal{G})]$, so conditioning on the endpoint marginals selects $\mathcal{H}^* = \operatorname{argmin}\{D_{KL}^{\mathrm{path}}(\mathcal{H}\|\mathcal{G}) : H_0 = P_0, H_T = P_T\}$ (Schrödinger, 1931; Schrödinger, 1932; Léonard, 2014). Lagrange multipliers give the factorized optimum $\mathcal{H}^*(\mathbf{x}_{0:N}) = f(\mathbf{x}_0)\,\mathcal{G}(\mathbf{x}_{0:N})\,g(\mathbf{x}_N)$, with f, g depending only on the endpoints, precisely the structure of a Doob h-transform (Doob, 1957), obtained here as a consequence of the variational principle rather than as a postulate. Propagating the terminal constraint backward through a field $\psi(\mathbf{x},t)$ tilts the reference kernel and shifts its drift to

$$\mathbf{v}_*(\mathbf{x},t) = \mathbf{v}(\mathbf{x},t) + \gamma(t)\,\nabla\ln\psi(\mathbf{x},t), \tag{2.1}$$

$\mathbf{v}$ being the reference drift and $\gamma(t)$ the diffusion coefficient. In the sector relevant to score-based generative modeling the forward field is constant and $P_t = \psi_t$, so Eq. (2.1) becomes the score-matching drift $\mathbf{v}_* = \mathbf{v} + \gamma\nabla\ln P_t$, and the learned score is not an arbitrary regression target but the optimal control field required to satisfy the terminal data constraint.

Since $\mathcal{H}^*$ is the I-projection of $\mathcal{G}$ onto the constraint set (Csiszár, 1975; Csiszár, 1984), relative entropy obeys the Pythagorean identity $D_{KL}^{\mathrm{path}}(\mathcal{H}\|\mathcal{G}) = D_{KL}^{\mathrm{path}}(\mathcal{H}^*\|\mathcal{G}) + D_{KL}^{\mathrm{path}}(\mathcal{H}\|\mathcal{H}^*)$, whose first term is independent of $\mathcal{H}$. Learning the bridge is therefore minimizing $\Delta\mathcal{S} = D_{KL}^{\mathrm{path}}(\mathcal{H}\|\mathcal{H}^*) \geq 0$, which for Markovian measures decomposes into per-step kernel divergences, each quadratic in the drift difference to leading order in Δt (Appendix A.5). The resulting on-policy weighting would

require simulating the generative process during training. But the integrand is point-wise minimized by $\mathbf{v}_H = \mathbf{v}_*$, so for unrestricted drift fields the minimizer does not depend on the positive weighting measure and we may use the reference noising marginal P_t instead (Tong et al., 2023). For a restricted parametric model this does affect the projection, a point we return to in Sec. 6. Writing the candidate drift as a control deformation $\mathbf{v}_H = \mathbf{v} + \gamma\,\mathbf{s}_\theta$ cancels the reference drift, the step by which the reference dynamics disappears from the objective, leaving the ideal score-matching loss

$$\Delta\mathcal{S}_{\text{ideal}} \;=\; \frac{1}{2}\int_0^T dt\,\gamma(t)\,\mathbb{E}_{\mathbf{x}\sim P_t}\left[\left\|\mathbf{s}_\theta(\mathbf{x},t) - \nabla_{\mathbf{x}}\log P_t(\mathbf{x})\right\|^2\right], \tag{2.2}$$

ideal in that its regression target, the marginal score, is not available in closed form.

Denoising score matching as conditional regression. Because ψ obeys a *linear* equation, superposition defines a conditional field $\psi^{(\mathbf{y})}$ for a bridge pinned to a Dirac terminal constraint at $\mathbf{y}$. The global field is their superposition over the data, $\psi(\mathbf{x},t) = \mathbb{E}_{\mathbf{y}\sim P_{\text{data}}}[\psi^{(\mathbf{y})}(\mathbf{x},t)] = P(\mathbf{x},t)$, with $\psi^{(\mathbf{y})}(\mathbf{x},t) = q(\mathbf{x},t \mid \mathbf{y})$ the corruption kernel itself. Replacing the data expectation by its Monte Carlo estimate and applying Jensen's inequality to the convex squared norm gives the tractable upper bound

$$\Delta\mathcal{S}_{\text{DSM}} := \frac{1}{2}\int_0^T dt\,\gamma(t)\,\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{x}\sim q(\cdot,t|\mathbf{y})}\left[\left\|\mathbf{s}_\theta(\mathbf{x},t) - \nabla_{\mathbf{x}}\ln q(\mathbf{x},t\mid\mathbf{y})\right\|^2\right] \;\geq\; \Delta\mathcal{S}_{\text{ideal}}, \tag{2.3}$$

the inequality by Jensen applied to $\nabla_{\mathbf{x}}\ln\mathbb{E}_{\mathbf{y}}[\psi^{(\mathbf{y})}]$. Every term is computable. Training samples $\mathbf{y}$, corrupts it, and regresses $\mathbf{s}_\theta$ onto $\nabla_{\mathbf{x}}\ln q(\mathbf{x},t\mid\mathbf{y})$, which for affine Gaussian corruption is linear in the noise and recovers the standard denoising score-matching loss (Vincent, 2011; Song & Ermon, 2019; Ho et al., 2020). The Jensen gap $\Delta\mathcal{S}_{\text{DSM}} - \Delta\mathcal{S}_{\text{ideal}}$ is the price of the substitution. The conditional score being unbiased, the gap does not shift the minimizer, but it shifts the objective's value by an additive, model-independent term. The remainder of the paper studies that term.

3 The Fisher Geometry of the Denoising Loss

Section 2 obtained the generative dynamics from the *first* variation of a path-space relative entropy. The *second* variation determines the geometry of the states the diffusion visits, and that geometry is exactly what inflates the denoising objective above its ideal value. It defines a canonical quadratic form on the bridge family measurable with respect to the endpoints alone. A regular, computable corruption selects a finite-dimensional reduction of it, and the Jensen gap of Sec. 2 is its integrated trace.

3.1 From the second variation to the latent metric

Every solution of the bridge problem is an endpoint tilt of the reference, $\ln\mathcal{H}^{(u,v)} = u(\mathbf{x}_0) + v(\mathbf{x}_N) + \ln\mathcal{G} - \Lambda[u,v]$, with Λ the log-partition functional. Its second variation is the endpoint variance $\delta^2\Lambda = \mathrm{Var}_\rho[\delta u(\mathbf{x}_0) + \delta v(\mathbf{x}_N)]$ (Lemma B.1), whose expansion carries a cross term, so the past and future sectors are *not* orthogonal in the ambient geometry. The first-order variation $\delta\ell = \delta u(\mathbf{x}_0) + \delta v(\mathbf{x}_N) - \mathbb{E}[\cdot]$ shows more: every infinitesimal likelihood ratio within the bridge family is measurable with respect to the endpoint σ-algebra. The endpoint pair is thus a sufficient statistic for the local statistical experiment, and the ambient form is its second moment. This is the classical setting in which Fisher information is monotone under Markov kernels and preserved under sufficient reductions (Chentsov, 1982; Ay et al., 2017; Pistone & Sempi, 1995; Cena & Pistone, 2007).

This sufficiency is what makes a latent geometry possible. Conditioning the path measure directly on an interior state $X_t = \mathbf{x}$ does not give a regular family. For $\mathbf{x} \neq \mathbf{x}'$ the conditioned measures live on disjoint path sets and are mutually singular, so no local Kullback–Leibler expansion exists from which a metric could be extracted (Kallenberg, 2002, Sec. 2.2). Retaining the endpoint experiment and conditioning *it* on the latent state, $\iota_t : \mathbf{x} \mapsto p(\mathbf{x}_0, \mathbf{x}_N \mid X_t = \mathbf{x})$, gives a chain of reductions $\mathcal{P}_{\text{path}} \to \mathcal{P}_{\text{end}} \to \mathcal{P}_{\text{end|lat}}$ that discards precisely the path information on which the ambient form does not depend. The reduction need not be unique. It is minimal given the tangent structure, and it is the one the denoising objective uses.

The reduction ι_t is still general. The construction used in practice follows from what a trainable model requires: (i) The latent state lives in the data space $\mathbb{R}^d$, not a learned code space. (ii) The states

form an ordered Markov hierarchy from the data to a fixed, data-independent prior left invariant by the transition. (iii) The corruption is fixed before training, hence state-independent, isotropic and continuous. Requirement (ii) already constrains the information flow, since information about the data is only discarded, never created (Cover & Thomas, 2006). Requirement (iii) forces an additive diffusion with Gaussian kernel $q(\mathbf{x}, t \mid \mathbf{x}_N) = \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{x}_N, \sigma_t^2 \mathbf{I})$. This is not a uniqueness theorem but the natural continuous realization of the requirements, and the standard family in use (Ho et al., 2020; Song et al., 2021; Kingma et al., 2021). As a function of the clean endpoint this is an exponential family with $(d{+}1)$-component sufficient statistic $\mathbf{s}(\mathbf{x}_N) = (\mathbf{x}_N, -\frac{1}{2}\|\mathbf{x}_N\|^2)$ and natural parameter

$$A(\mathbf{x}, t) \;=\; \Big(\frac{\alpha_t}{\sigma_t^2}\mathbf{x},\; \frac{\alpha_t^2}{\sigma_t^2}\Big) \;=\; \mathrm{SNR}_t \cdot \big(\tilde{\mathbf{x}},\, 1\big), \qquad \tilde{\mathbf{x}} := \mathbf{x}/\alpha_t. \tag{3.1}$$

Latent spacetime is therefore a cone in natural-parameter space. The direction of A is the rescaled spatial coordinate, its magnitude is the signal-to-noise ratio, and the extra dimension is the second-moment statistic of the corruption, which isotropy keeps one-dimensional. An anisotropic Gaussian instead gives the family codimension, making it *curved* (Efron, 1975; Amari, 1982). A monotone reparametrization of time slides along the ray without changing it, which is the geometric content of schedule reparametrization invariance (Kingma et al., 2021). Pulling the ambient form back along A factorizes the metric as $g_{\mu\nu} = \partial_\mu A^a \partial_\nu A^b \,\mathrm{Cov}_{p(\mathbf{x}_N \mid \mathbf{x}, t)}[\mathbf{s}_a, \mathbf{s}_b]$, the first factor fixed by the schedule and the second carrying the data.

Proposition 3.1 (Endpoint split and the loss metric)**.** *The two endpoints are conditionally independent given the latent state, so $g_{\mu\nu} = g^{\mathrm{past}}_{\mu\nu} + g^{\mathrm{future}}_{\mu\nu}$, and for the Gaussian corruption above,*

$$g^{\mathrm{future}}_{ij}(\mathbf{x}, t) = \frac{\alpha_t^2}{\sigma_t^4}\mathrm{Cov}\big[x_N^i, x_N^j \mid X_t = \mathbf{x}\big] = \frac{1}{\sigma_t^2}\delta_{ij} + \partial_i \partial_j \ln P_t(\mathbf{x}). \tag{3.2}$$

The ambient cross term vanishes only after conditioning on the intermediate state, by the Markov property. The past sector is the Fisher information of the seed posterior $p(\mathbf{x}_0 \mid \mathbf{x}, t)$, generically nonzero yet invisible to denoising score matching, whose target involves the clean endpoint alone. The full latent metric therefore contains strictly more structure than the loss probes, and the geometry is not reverse-engineered from the objective. Equation (3.2), whose trace appears in the theorem below, has been studied descriptively (Karczewski et al., 2026; Lobashev et al., 2025). The derivation here shows that it follows from the second variation of the bridge objective. The cone also makes the information hierarchy quantitative. The information the latent retains about the endpoint decays at a rate set by the radial component of the metric, the point-wise counterpart of the integrated identity of Sec. 4 (Appendix B).

3.2 The irreducible loss floor

It remains to identify the geometry of Sec. 3.1 with the Jensen gap of Sec. 2. Write $q(\mathbf{y} \mid \mathbf{x}, t) = P_{\mathrm{data}}(\mathbf{y}) q(\mathbf{x}, t \mid \mathbf{y}) / P(\mathbf{x}, t)$ for the posterior over clean data. Expanding the squared norm in Eq. (2.3) around the ideal target gives a quadratic term and a cross term, and the cross term vanishes for *every* θ, not merely at convergence, because $\mathbb{E}_{\mathbf{y}|\mathbf{x}}[\nabla_{\mathbf{x}} \ln q(\mathbf{y} \mid \mathbf{x}, t)] = \nabla_{\mathbf{x}} \int d\mathbf{y}\, q = 0$ point-wise in $\mathbf{x}$. Only the model-independent term survives.

$$\begin{aligned} \Delta\mathcal{S}_{\mathrm{DSM}} - \Delta\mathcal{S}_{\mathrm{ideal}} &= \frac{1}{2}\int_0^T dt\, \gamma(t)\, \mathbb{E}_{\mathbf{x}\sim P_t} \mathbb{E}_{\mathbf{y}|\mathbf{x}} \left\| \nabla_{\mathbf{x}} \ln q(\mathbf{y} \mid \mathbf{x}, t) \right\|^2 \\ &= \frac{1}{2}\int_0^T dt\, \gamma(t)\, \mathbb{E}_{\mathbf{x}\sim P_t} \big[\, \mathrm{tr}\, \mathrm{Cov}_{\mathbf{y}|\mathbf{x}}[\nabla_{\mathbf{x}} \ln q(\mathbf{x}, t \mid \mathbf{y})] \,\big], \end{aligned} \tag{3.3}$$

the second equality by Bayes' rule, $\nabla_{\mathbf{x}} \ln q(\mathbf{y} \mid \mathbf{x}, t) = \nabla_{\mathbf{x}} \ln q(\mathbf{x}, t \mid \mathbf{y}) - \nabla_{\mathbf{x}} \ln P(\mathbf{x}, t)$, which makes the posterior score the centered conditional score.

Theorem 3.2 (The loss floor is a Fisher information)**.** *Let $q(\mathbf{x}, t \mid \mathbf{y})$ be a corruption kernel that is positive on the support of interest, with $\mathbf{x} \mapsto \ln q(\mathbf{x}, t \mid \mathbf{y})$ differentiable and $\nabla_{\mathbf{x}} q(\mathbf{x}, t \mid \mathbf{y})$ locally dominated by an integrable function of $\mathbf{y}$, so that differentiation under the integral sign is permitted. Let P_t be the induced marginal, let $p(\mathbf{y} \mid \mathbf{x}, t) = P_{\mathrm{data}}(\mathbf{y}) q(\mathbf{x}, t \mid \mathbf{y}) / P_t(\mathbf{x})$ be the* conditional endpoint family *— the laws over clean endpoints $\mathbf{y}$ indexed by the latent coordinate $\mathbf{x}$, which plays the role of the parameter — and let $g_{ij}(\mathbf{x}, t) := \mathbb{E}_{\mathbf{y}|\mathbf{x}}[\partial_i \ln p\, \partial_j \ln p]$ be its Fisher–Rao metric. Then*

$$\Delta\mathcal{S}_{\mathrm{DSM}} - \Delta\mathcal{S}_{\mathrm{ideal}} \;=\; \frac{1}{2}\int_0^T dt\, \gamma(t)\, \mathbb{E}_{\mathbf{x}\sim P_t}\big[\, \mathrm{tr}\, g(\mathbf{x}, t)\big]. \tag{3.4}$$

The proof is immediate. Since $P_{\mathrm{data}}(\mathbf{y})$ does not depend on $\mathbf{x}$, differentiating $\ln p(\mathbf{y} \mid \mathbf{x}, t)$ gives

$$\nabla_{\mathbf{x}} \ln p(\mathbf{y} \mid \mathbf{x}, t) \;=\; \nabla_{\mathbf{x}} \ln q(\mathbf{x}, t \mid \mathbf{y}) \;-\; \nabla_{\mathbf{x}} \ln P_t(\mathbf{x}), \tag{3.5}$$

whose posterior expectation is $\nabla_{\mathbf{x}} \int d\mathbf{y}\, p = 0$, the unbiasedness identity of denoising score matching. The posterior score is therefore centered, its second moment equals its covariance $g(\mathbf{x}, t)$, and this is simultaneously the Fisher–Rao metric of the theorem and the integrand of Eq. (3.3). Taking the trace gives Eq. (3.4). Beyond the stated regularity the theorem assumes nothing about the kernel, no Gaussianity, affine structure, or exponential-family form, and its content is a conditional-variance identity. The ideal objective compares the model against $\nabla \ln P_t$, which by Eq. (3.5) is the posterior *mean* of the conditional score. The implementable objective cannot evaluate that mean and substitutes a single draw $\mathbf{y} \sim p(\cdot \mid \mathbf{x}, t)$, unbiased but noisy. The point-wise decomposition reads

$$\mathbb{E}\big\|\mathbf{s}_\theta - \mathbf{s}_{\mathrm{cond}}\big\|^2 \;=\; \big\|\mathbf{s}_\theta - \nabla \ln P_t\big\|^2 \;+\; \mathrm{Var}\big(\mathbf{s}_{\mathrm{cond}} \mid \mathbf{x}\big), \tag{3.6}$$

that is, denoising objective = estimation error + irreducible conditional variance, in which the second term is the Fisher information of the reduction ι_t of Sec. 3.1. This is a bias–variance decomposition, though the first term is a squared estimation error rather than a bias, since $\mathbf{s}_\theta$ is deterministic given $\mathbf{x}$. Because the conditional variance of a regression target does not involve the regressor, the floor cannot depend on θ. *The latent Fisher metric is the covariance of the random target introduced by denoising score matching*, and the loss floor is its integrated trace along the flow.

For the affine Gaussian kernel the conditional score is $(\alpha_t \mathbf{y} - \mathbf{x})/\sigma_t^2$, so Eq. (3.5) gives $\nabla_{\mathbf{x}} \ln p(\mathbf{y} \mid \mathbf{x}, t) = (\alpha_t/\sigma_t^2)(\mathbf{y} - \mathbb{E}[\mathbf{y} \mid \mathbf{x}])$ and the metric reduces to Eq. (3.2).

4 The Loss Floor as Information Flow

Theorem 3.2 identified the irreducible part of the denoising objective with a Fisher–Rao metric of the endpoint posterior, given explicitly for affine kernels by Eq. (3.2). What the metric measures when accumulated along the trajectory has a simple information-theoretic answer. It is the information about the clean endpoint that is lost under corruption. Proofs are in Appendix C and numerical details in Appendix D.

4.1 The mutual-information identity

The Fisher–Rao representation is local in the corruption parameter. Its integrand involves only the corruption kernel, not affine Gaussian structure. Writing $J(p) := \mathbb{E}_p \|\nabla \ln p\|^2$ for the translation Fisher information, the trace of the posterior metric averaged over the noisy marginal is exactly the Fisher information lost when the endpoint-conditioned kernels are mixed over the data, and along a diffusion the same gap is the rate of forgetting.

Let the corruption be generated by a diffusion $d\mathbf{x}_t = f(\mathbf{x}_t, t)\, dt + \sqrt{D(t)}\, d\mathbf{W}_t$ started from the data distribution, with f and D independent of $\mathbf{y}$ and t increasing in the corruption direction. We write $D(t)$ rather than $\gamma(t)$ to distinguish it from the bridge dynamics of Sec. 2.

Theorem 4.1 (Information flow and general factorization of the floor)**.** *For any corruption kernel,* $\mathbb{E}_{P_t}[\mathrm{tr}\, g(\cdot, t)] = \mathbb{E}_{\mathbf{y}}[J(q_t(\cdot \mid \mathbf{y}))] - J(P_t)$*, and along a corruption diffusion as above,*

$$\frac{d}{dt} I(\mathbf{y}; \mathbf{x}_t) = -\frac{D(t)}{2} \Big(\mathbb{E}_{\mathbf{y}} J(q_t(\cdot \mid \mathbf{y})) - J(P_t) \Big) = -\frac{D(t)}{2}\, \mathbb{E}_{P_t}\big[\, \mathrm{tr}\, g(\cdot, t)\big] \le 0. \tag{4.1}$$

Consequently, for an arbitrary loss weighting $w(t)$*,*

$$\Delta\mathcal{S}_{\mathrm{DSM}} - \Delta\mathcal{S}_{\mathrm{ideal}} \;=\; \frac{1}{2} \int w(t)\, \mathbb{E}_{P_t}\big[\, \mathrm{tr}\, g\big]\, dt \;=\; -\int \underbrace{\frac{w(t)}{D(t)}}_{\textit{schedule}}\, \underbrace{\frac{dI(\mathbf{y}; \mathbf{x}_t)}{dt}}_{\textit{data}}\, dt. \tag{4.2}$$

In particular, if $w = D$ *then* $\Delta\mathcal{S}_{\mathrm{DSM}} - \Delta\mathcal{S}_{\mathrm{ideal}} = I(\mathbf{y}; \mathbf{x}_{t_0}) - I(\mathbf{y}; \mathbf{x}_{t_1})$.

The instantaneous floor is the gap between the Fisher information available when the clean endpoint is known and that remaining after it is marginalized out. Its non-negativity is the monotonicity of Fisher information under mixing (Cover & Thomas, 2006), and up to the local diffusion scale it is the rate at which information about the endpoint is erased, so integrating turns the local Fisher geometry of the loss into a global information-flow law (Guo et al., 2005). No Gaussian, affine-interpolant, or SNR assumption enters. The data appear only through $-dI/dt$, the schedule only through w/D. Equation (4.1) is verified numerically for a nonlinear corruption drift by direct Fokker–Planck solution (Appendix D).

4.2 GAUSSIAN CORRUPTION, SCHEDULE AND WEIGHTING

The affine case is the specialization in which the information flow has a closed form. Equation (3.2) makes $\operatorname{tr} g$ proportional to the minimum mean squared error of endpoint estimation. Throughout, $\mathrm{snr}_t := \alpha_t^2/\sigma_t^2$ is the signal-to-noise ratio and $\lambda_t := \ln \mathrm{snr}_t$ the log-SNR used in Sec. 5. The bridge construction further singles out one combination of schedule and weighting, and it is the one that makes the floor an information.

Among marginal-preserving diffusions realizing a given affine interpolant, the bridge singles out the memoryless one, the unique state-independent coefficient for which $\mathbf{x}_t/\alpha_t$ has independent increments, namely $\gamma(t) = 2\sigma_t^2 \frac{d}{dt} \ln(\alpha_t/\sigma_t)$, and for it $\gamma(t)\alpha_t^2/\sigma_t^4 = d\mathrm{snr}_t/dt$ for every affine interpolant (Lemma C.1, Appendix C). This is verified by direct computation for each interpolant family (Appendix D), and it says that *the memoryless schedule is precisely the one whose loss weighting is the SNR measure*, so the bridge member of the marginal-preserving family is the canonical object singled out by the objective rather than one convenient choice among many. Changing variables from t to λ for a general weight $w(t)$ factorizes the floor,

$$\Delta\mathcal{S}_{\mathrm{DSM}} - \Delta\mathcal{S}_{\mathrm{ideal}} \;=\; \frac{1}{2}\int \omega(\mathrm{snr})\,\mathrm{MMSE}(\mathrm{snr})\,d\mathrm{snr}, \qquad \omega := \frac{w}{\gamma}, \tag{4.3}$$

with $\mathrm{MMSE}(\mathrm{snr}) := \mathbb{E}\,\|\mathbf{y} - \mathbb{E}[\mathbf{y} \mid \mathbf{x}_{\mathrm{snr}}]\|^2$ and $\mathbf{x}_{\mathrm{snr}} = \sqrt{\mathrm{snr}}\,\mathbf{y} + \mathbf{z}$, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. The data enter only through MMSE, a scalar function of the signal-to-noise ratio determined by the distribution alone. The schedule enters only through ω and the endpoints, with the bridge weighting corresponding to $\omega \equiv 1$. At large snr the quantity $\mathrm{snr}\,\mathrm{MMSE}(\mathrm{snr})$ saturates at the information dimension of the data, which is what makes the floor diverge at the data end and connects it to intrinsic-dimension estimation. With that weighting the floor becomes an entropy outright.

Theorem 4.2 (Loss floor = information gained)**.** *With the bridge schedule of Lemma C.1,*

$$\Delta\mathcal{S}_{\mathrm{DSM}} - \Delta\mathcal{S}_{\mathrm{ideal}} \;=\; \frac{1}{2}\int \mathrm{MMSE}(\mathrm{snr})\,d\mathrm{snr} \;=\; I(\mathbf{y};\mathbf{x}_{\mathrm{snr}_1}) - I(\mathbf{y};\mathbf{x}_{\mathrm{snr}_0}) \;=\; h(\mathbf{x}_{\mathrm{snr}_1}) - h(\mathbf{x}_{\mathrm{snr}_0}), \tag{4.4}$$

where I is mutual information and h differential entropy.

The proof combines Theorem 3.2 with Lemma C.1 and the I–MMSE relation of Guo, Shamai and Verdú (Guo et al., 2005) (Appendix C). On a Gaussian mixture, for which $h(\mathbf{x}_{\mathrm{snr}})$ is computable by quadrature, the two sides of Eq. (4.4) agree to a relative error of 2×10^{-5} (Fig. 2, Appendix D).

Because the right-hand side of Eq. (4.4) is a difference of endpoint values, the floor depends on the schedule only through the endpoints of the SNR range, not the path between them. This is the invariance theorem of variational diffusion models (Kingma et al., 2021), obtained here as an integral rather than by direct computation. A schedule is a *parametrization* of a fixed curve in the information manifold, and the floor a parametrization-invariant line integral along it, pinned by its endpoints. Invariance is thus no algebraic accident of the ELBO but the statement that a geometric quantity does not depend on how one traverses the curve. A schedule cannot change the floor, only redistribute estimation error along the path (Zhang, 2025; Kong et al., 2026).

5 CONSEQUENCES FOR TRAINING: HOW LARGE IS THE FLOOR?

For the bridge weighting the decomposition reads $\Delta\mathcal{S}_{\mathrm{DSM}} = \Delta\mathcal{S}_{\mathrm{ideal}} + \Delta h$, so the floor depends on the schedule through the endpoints of its SNR range and on the data through their entropy, and raw training losses are not directly comparable across runs using different ranges, schedules, or weightings. Under fixed training conditions the floor is common to all runs and the conventional loss remains a valid comparator. It is cross-configuration comparison that is at risk. The size of that risk is an empirical question, and this section answers it on CIFAR-10.

5.1 AN ESTIMATOR FOR THE FLOOR

Theorem 3.2 expresses the floor through a posterior covariance that is not directly available. In the log-SNR coordinate λ the evidence-bound weighting makes $\omega \equiv 1$, so the floor is $\frac{1}{2}\int M(\lambda)\,d\lambda$ with integrand $M(\lambda) := \sigma_\lambda^2\,\mathbb{E}_{P_\lambda}[\operatorname{tr} g] = e^\lambda\,\mathbb{E}_{P_\lambda}[\operatorname{tr}\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}]]$, the second equality by Eq. (3.2). Differentiating Tweedie's identity gives $\partial_{\mathbf{x}}\mathbb{E}[\mathbf{y} \mid \mathbf{x}] = (\alpha_\lambda/\sigma_\lambda^2)\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}]$, and substituting yields, for an ϵ-prediction network,

$$M(\lambda) \;=\; d \;-\; \sigma_\lambda\,\mathbb{E}_{P_\lambda}\big[\operatorname{tr}\partial_{\mathbf{x}}\epsilon_\theta\big]. \tag{5.1}$$

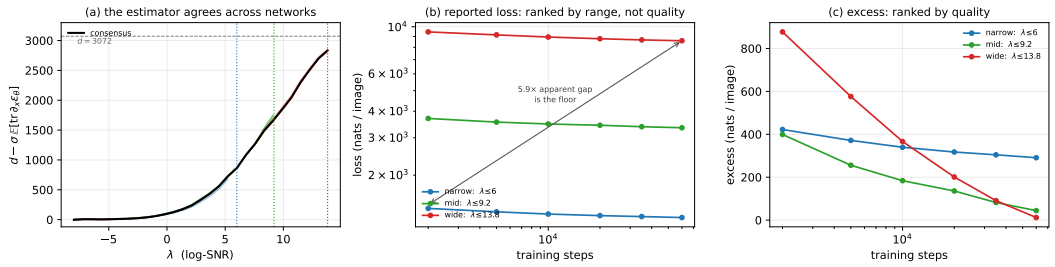

Figure 1: The loss floor on CIFAR-10. **(a)** Floor integrand from each converged checkpoint over its own training range, consensus in black. The curves are indistinguishable, so the estimator returns the same floor from independently trained networks. Dashed line: the saturation value $d = 3072$. **(b)** Reported loss against training steps. The three series are separated by their floors, not by model quality, and a 2k-step model on the narrow range reports a loss $5.9\times$ smaller than a 60k-step model on the wide range. **(c)** After floor subtraction the excess is ordered by training budget on every range.

Table 1: Reported loss, estimated floor, and the model-dependent remainder for the 60k-step model on each log-SNR range, in nats per image. Widening λ_1 from 6.0 to 13.8 multiplies the reported loss by $6.8\times$ and divides the model-dependent part by $24\times$.

range	λ_1	reported	floor	floor share	model-dependent
narrow	6.0	1263.7	972.9	76.98%	290.9
mid	9.2	3345.0	3300.7	98.68%	44.3
wide	13.8	8589.9	8577.8	99.86%	12.0

The trace is estimated with Rademacher probes, $\mathrm{tr}\, J \approx \mathbb{E}_v[v^\top J v]$, in which $v^\top J$ is a single vector–Jacobian product. One probe per sample suffices, since a batch mean is all that is required, so the correction costs one backward pass and needs nothing but the trained denoiser. On a Gaussian mixture, where the posterior covariance is closed-form, Eq. (5.1) reproduces $M(\lambda)$ to 4×10^{-11} and the integrated floor to 7×10^{-4} (Appendix D.3). The identity holds at the optimum. For an undertrained network the Jacobian is too small and the floor is *over*estimated, approaching d as ϵ_θ becomes uninformative, so estimates must not be taken outside a network's training range.

5.2 The floor on CIFAR-10

We train a 12M-parameter UNet on CIFAR-10 with the objective of Sec. 2 on three log-SNR ranges, $\lambda \in [-8, \lambda_1]$ with $\lambda_1 \in \{6.0, 9.2, 13.8\}$, the latter two corresponding to $\sigma_{\min} \approx 10^{-2}$ and 10^{-3}, both in common use. The quality axis comes from checkpoints rather than separate runs. One run per range, saved at six step counts, yields a ladder of model qualities at no extra compute. Losses are evaluated on held-out data with common random numbers, every model scored on an identical stream of $(\mathbf{y}, \lambda, \epsilon)$ triples. This matters because the quality differences being resolved are far smaller than the floor. A checkpoint contributes to the floor at λ only if λ lies inside its own training range (Appendix E).

Table 1 compares each range's 60k-step model against its floor. On the wide range, the regime of standard practice, about one part in 700 of the reported number carries information about the model. The floor is not a correction to a reported diffusion loss but the bulk of it.

That the floor does not depend on the model is not a prediction to be tested, since a conditional variance cannot involve the regressor. What is testable is Eq. (5.1), whose accuracy licenses correcting one model's loss with a floor obtained from others. Integrated floors from independently trained networks agree to 4.92% on the narrow range (nine checkpoints across all three configurations), 1.31% on the mid and 0.63% on the wide, with absolute scatter ≈ 50 nats irrespective of range, the estimator's noise level, not anything configuration-dependent.

Subtracting the floor gives an excess that decreases monotonically with training budget on all three ranges (Fig. 1c), from 877 to 12 nats on the wide range, 422 to 291 on the narrow. The raw loss does not order models across ranges. Taking two checkpoints of the *same* run, so no specialization confound arises, the 60k-step model on the wide range reports 8589.9 while the 2k-step model on

the narrow range reports 1450.0. The $30\times$ less-trained model appears $5.9\times$ better. The raw gap is 7140 nats and the floor difference alone is 7605 ± 72, accounting for all of it, while the floor-subtracted excesses, 12.0 and 477.1, order the two correctly. Recomputing each excess from a floor built without that model changes it negligibly ($44.3 \to 45.6$, $90.7 \to 87.3$ nats), so the correction is not self-fulfilling. The exception is the best wide-range model, whose excess moves from 12.0 to -13.0 and is thus indistinguishable from zero, so its excess should be read as an upper bound. Floor subtraction is necessary but not sufficient, and we recommend pairing it with a common-range evaluation. Appendix E explains why.

6 Discussion

This paper took the standard training objective seriously as an object of study in its own right, not as a surrogate to be justified and set aside. When read through the bridge variational principle, the objective already contained the geometry of the latent space, the information flow of the corruption, and the reason its value cannot be compared across configurations. None of this structure was inserted by hand. It follows from one modeling choice common to every diffusion pipeline, and it therefore holds for every model trained with denoising score matching. A geometric quantity, an information-theoretic quantity, and a loss value turned out to be the same number, and each viewpoint licenses consequences the others do not make obvious.

Scope and relation to prior work. The primary results are Theorem 3.2, its factorization in Theorem 4.1, and the measurement of Sec. 5. The ingredients of Sec. 3 are classical (Chentsov, 1982; Efron, 1975; Amari, 1982). Novelty is claimed only for their assembly into a derivation of the latent metric from the bridge principle. Nearby work addresses optimization problems easily conflated with ours, namely sampler discretization (Sabour et al., 2024; Williams et al., 2024; Stancevic et al., 2025), noise density (Ambrogioni et al., 2026), loss weighting (Kingma et al., 2021), and time reparametrization (Zhang, 2025; Kong et al., 2026). Ours is not a new member of these families. It is Eq. (4.3), which delimits what any of them can affect. Two boundaries are worth marking. The objective as an entropy-production functional is developed independently and in more detail by Ding et al. (2026), whereas we identify the irreducible term with a Fisher–Rao metric and an entropy change. Intrinsic-dimension estimation from a trained model is already established (Pope et al., 2021; Stanczuk et al., 2024; Kamkari et al., 2024), and we do not claim it here.

6.1 The discrete case of masked diffusion

The decomposition also has a consequence beyond the continuous setting, one of the most useful observations of this work. Discrete diffusion (Austin et al., 2021; Lou et al., 2024; Nie et al., 2025) admits the same treatment. The masked-diffusion objective is a cross-entropy, hence the Bregman divergence generated by the negative entropy, so the algebra behind Eq. (3.6) applies verbatim. The expected divergence to a random target splits into a model-dependent term at the posterior mean and an irreducible remainder, the Bregman information of the target (Savage, 1971; Banerjee et al., 2005b;a), which for cross-entropy is a conditional entropy. Writing m_t for the masking probability, decreasing from 1 to 0, and U for the set of unmasked tokens, a still-masked token is revealed over an infinitesimal step with probability $-\dot{m}_t/m_t$ at a cost $H(y_i \mid y_U)$. With a perfect model the residual loss is therefore

$$\text{floor} \;=\; \int_0^1 \frac{dm}{m}\, \mathbb{E}_U\Big[\sum_{i \notin U} H\big(y_i \mid y_U\big)\Big], \tag{6.1}$$

where each token is unmasked independently with probability $1 - m$. Two properties follow immediately. Substituting $u = m_t$ in the time integral removes the schedule, so the floor depends only on the endpoint masking rates and not on the shape of m_t. Evaluating Eq. (6.1) then gives the entropy $H(y_1, \ldots, y_n)$ of the data. We have verified both numerically for correlated tokens (Appendix D).

The same algebra settles the other direction of generality, namely the choice of regression target within the continuous setting. Every objective built on an affine interpolant, including $\mathbf{x}_0$- and v-prediction, flow matching, and stochastic interpolants, regresses onto a target $T = a_t\mathbf{y} + b_t\mathbf{z}$, and eliminating $\mathbf{z}$ in favour of $\mathbf{x}_t$ leaves $\mathrm{Cov}[T \mid \mathbf{x}_t] = c_t^2\,\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}_t]$ with $c_t = a_t - b_t\alpha_t/\sigma_t$. All members of the family therefore share one geometric object and differ only by the closed-form weight c_t^2. No affine parametrization removes the floor, and the conversion between a reported flow-matching

loss and a reported diffusion loss is explicit. Appendix F gives the derivation, the weight for each standard target, and numerical verification.

This is the discrete counterpart of the continuous statement. There the floor is an accumulated mutual information fixed by the endpoints of the SNR range, here it is the data entropy, fixed by the endpoint masking rates, and in both the schedule redistributes the cost without changing the total. It also accounts for the known schedule invariance of the masked-diffusion evidence bound (Sahoo et al., 2024; Shi et al., 2024), whose irreducible part telescopes along the filtration. We leave open a discrete analogue of the cone Eq. (3.1) and of the latent metric itself, which would require the α-geometry of the simplex.

Limitations. Evaluating the floor in closed form as ΔI assumes affine Gaussian corruption, and Theorem 4.1 covers general diffusions only by leaving the information flow implicit. The high-SNR rate is set by the information dimension, which need not exist for singular data. Replacing the on-policy weighting by the simulation-free one leaves the functional minimizer unchanged but not the parametric projection, so our statements concern the objective's value rather than that projection. The CIFAR-10 experiment measures the floor and the resulting inversion, and makes no claim about sample quality. Finally, the floor is only as estimable as the posterior covariance, hardest to estimate at the low-noise end where it diverges.

One caution on reading the metric. The sharp peak that $\mathbb{E}[\mathrm{tr}\, g]$ develops in a low-dimensional mixture does not survive on CIFAR-10, because natural images carry structure on many scales at once, and the transitions at the individual scales are spread across the corruption range rather than coinciding, so the phase-transition picture should not be expected to transfer to real data.

Concluding remarks. The denoising objective splits exactly into a model-dependent estimation error and a model-independent floor, the integrated trace of the Fisher–Rao metric of the conditional endpoint family, and for affine Gaussian corruption the accumulated mutual information between data and noisy state. The consequence we emphasize is the least glamorous. A reported diffusion loss mixes model fit with a constant that on CIFAR-10 is 99.86% of it, and the two should be separated before runs are compared.

Reproducibility Statement

Every numerical claim in this paper is reproducible from the code released with the submission. The code is available at `https://anonymous.4open.science/r/loss-floor-dsm-BE06`. The analytic identity checks of Appendix D are self-contained and run in seconds on a CPU. The CIFAR-10 measurement of Sec. 5 requires two training runs of 60k steps, described in Appendix E, and the released code contains the saved evaluation output, so Table 1 and Figure 1 can be regenerated without retraining.

Use of Large Language Models

We used large language models in the following ways. For *writing*, they assisted with drafting and editing prose, improving readability, and formatting references. For *literature search*, they were used to find and summarize related work, and every reference so identified was read and verified by the authors before citation. For *code*, they assisted in writing and debugging the verification scripts and the experimental code released with the paper. Every line of code was verified independently by the authors. We take responsibility for the final content, including anything produced with generative AI, and any errors that remain are ours.

References

Michael S. Albergo, Nicholas Matthew Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *J. Mach. Learn. Res.*, 26:209:1–209:80, 2023. URL `https://api.semanticscholar.org/CorpusID:257532329`.

S. Amari and H. Nagaoka. *Methods of Information Geometry*. Fields Institute Communications. American Mathematical Society, 2000. ISBN 9780821805312. URL `https://books.google.co.jp/books?id=qGzCwAEACAAJ`.

Shun-ichi Amari. Differential geometry of curved exponential families—curvatures and information loss. *The Annals of Statistics*, 10(2):357–385, 1982.

Luca Ambrogioni et al. From atoms to entropy: Optimal noise allocation for diffusion training in the convex regime, 2026.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 2017.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005a.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005b.

Ole Barndorff-Nielsen. *Information and exponential families in statistical theory*. John Wiley & Sons Ltd, 1978. ISBN 0-471-99545-2.

Lawrence D. Brown. *Fundamentals of statistical exponential families : with applications in statistical decision theory*. Lecture notes-monograph series Fundamentals of statistical exponential families. Institute of Mathematical Statistics, 1986. ISBN 0-940600-10-2.

Alberto Cena and Giovanni Pistone. Exponential statistical manifold. *Annals of the Institute of Statistical Mathematics*, 59(1):27–56, 2007.

Tianrong Chen, Guan-Horng Liu, and Evangelos A. Theodorou. Likelihood training of Schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022.

N. N. Chentsov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, 1982.

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

Imre Csiszár. Information geometry and alternating minimization procedures. *Statistics and Decisions, Dedewicz*, 1:205–237, 1984.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 13(1):146–158, 1975. URL https://cir.nii.ac.jp/crid/1573950400419085440.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/940392f5f32a7ade1cc201767cf83e31-Paper.pdf.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2 edition, 1998.

Xuehao Ding, H. T. Quan, and Yuhai Tu. Stochastic thermodynamics of score matching in diffusion models, 2026.

Joseph L. Doob. Conditional brownian motion and the boundary limits of harmonic functions. *Bulletin de la Société Mathématique de France*, 85:431–458, 1957.

Bradley Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242, 1975.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106:1602 – 1614, 2011. URL https://api.semanticscholar.org/CorpusID:23284154.

Hans Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Mathematics*, pp. 101–203. Springer, 1988. doi: 10.1007/BFb0086180.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL https://arxiv.org/abs/2006.11239.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, dec 2005. ISSN 1532-4435.

Olav Kallenberg. *Processes, Distributions, and Independence*, pp. 45–61. Springer New York, New York, NY, 2002. ISBN 978-1-4757-4015-8. doi: 10.1007/978-1-4757-4015-8_3. URL https://doi.org/10.1007/978-1-4757-4015-8_3.

Hamidreza Kamkari, Brendan Leigh Ross, Rasa Hosseinzadeh, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. The spacetime of diffusion models: An information geometry perspective, 2026. URL https://arxiv.org/abs/2505.17517.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Seo Taek Kong, Weina Wang, and R. Srikant. Noise schedule design for diffusion models: An optimal control perspective, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ArXiv*, abs/2210.02747, 2022. URL https://api.semanticscholar.org/CorpusID:252734897.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ArXiv*, abs/2209.03003, 2022. URL https://api.semanticscholar.org/CorpusID:252111177.

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models, 2025. URL https://arxiv.org/abs/2506.10632.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *ICML*, 2024.

Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014. ISSN 1078-0947. doi: 10.3934/dcds.2014.34.1533. URL https://www.aimsciences.org/article/id/d5bcf817-901d-4104-b7da-eade7847c53e.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020. doi: 10.3390/e22101100.

Giovanni Pistone and Carlo Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 23(5):1543–1561, 1995. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2242533.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Akhil Premkumar. Generative diffusion from an action principle. *ArXiv*, abs/2310.04490, 2023. URL https://api.semanticscholar.org/CorpusID:263831231.

Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. Springer, 2 edition, 1996.

Herbert E. Robbins. An empirical bayes approach to statistics. In Samuel Kotz and Norman L. Johnson (eds.), *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 388–394. Springer New York, New York, NY, 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_26. URL https://doi.org/10.1007/978-1-4612-0919-5_26.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

E. Schrödinger. *Über die Umkehrung der Naturgesetze*. Sitzungsberichte der Preussischen Akademie der Wissenschaften. Physikalisch-mathematische Klasse. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u. Company, 1931.

E. Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Annales de l'institut Henri Poincaré*, 2(4):269–310, 1932. URL http://eudml.org/doc/78968.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, 2024.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion Schrödinger bridge matching. 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 Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Ognjen Stancevic, Julius Handke, and Luca Ambrogioni. Entropic time schedulers for generative diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Tsolaki, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Alexander Tong, Esmeralda S. Whitammer, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL https://api.semanticscholar.org/CorpusID:259375451.

Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9), 2021. ISSN 1099-4300. doi: 10.3390/e23091134. URL https://www.mdpi.com/1099-4300/23/9/1134.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via schrödinger bridge. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10794–10804. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21l.html.

Christopher Williams, Andrew Campbell, Arnaud Doucet, and Saifuddin Syed. Score-optimal diffusion schedules. In *Advances in Neural Information Processing Systems*, 2024.

Leo Zhang. The cosine schedule is Fisher-Rao-optimal for masked discrete diffusion models, 2025.

APPENDIX

A VARIATIONAL DERIVATION OF THE ENDPOINT-TILTED PATH MEASURE

This appendix collects the derivations underlying the variational formulation of the Schrödinger bridge in Sec. 2. We first solve the constrained minimization that yields the factorized optimal path measure, and then derive the recursion relations for the auxiliary fields together with their continuum limit. The remaining subsections expand the tilted kernel at short times, construct the conditional fields that connect the bridge to the score-matching objective, and compute the short-time KL divergence between diffusion kernels that underlies the excess action.

For orientation, the main text is organized as follows. Section 2 derives the generative drift from the bridge functional. Section 3 takes its second variation to obtain the floor as a Fisher trace. Section 4 specializes to Gaussian corruption and identifies the floor with an information flow. Section 5 gives the estimator and the CIFAR-10 measurement. Section 6 treats the discrete case and closes with limitations.

A.1 SOLUTION OF THE BRIDGE PROBLEM

Consider the time interval $[0, T]$ discretized into N equal segments of length $\Delta t = T/N$, and let $\mathbf{x}_k$ denote the state at time $t_k = k\Delta t$. The joint path measure under the reference process is

$$\mathcal{G}(\mathbf{x}_{0:N}) = P_0(\mathbf{x}_0) \prod_{k=0}^{N-1} G(\mathbf{x}_{k+1} \mid \mathbf{x}_k), \tag{A.1}$$

and under the candidate process,

$$\mathcal{H}(\mathbf{x}_{0:N}) = P_0(\mathbf{x}_0) H(\mathbf{x}_{1:N} \mid \mathbf{x}_0) = P_0(\mathbf{x}_0) \prod_{k=0}^{N-1} H(\mathbf{x}_{k+1} \mid \mathbf{x}_k). \tag{A.2}$$

The Schrödinger bridge is formulated directly as the minimization of relative entropy on path space,

$$\begin{aligned} \mathcal{H}^* &= \operatorname*{argmin}_{\mathcal{H}:\, \mathcal{H}_0 = P_0,\, \mathcal{H}_N = P_T} D_{KL}^{\text{path}}(\mathcal{H} \| \mathcal{G}), \\ D_{KL}^{\text{path}}(\mathcal{H} \| \mathcal{G}) &= \int d\mathbf{x}_{0:N}\, \mathcal{H}(\mathbf{x}_{0:N}) \ln \frac{\mathcal{H}(\mathbf{x}_{0:N})}{\mathcal{G}(\mathbf{x}_{0:N})}, \end{aligned} \tag{A.3}$$

with the endpoint marginals imposed as constraints. No equivalence between a marginal-propagator divergence and the path-space divergence is required, and we do not assert one. Equation (A.3) is the definition of the problem in the sense of (Föllmer, 1988; Léonard, 2014), and the derivation below is a variation of that functional alone.

To enforce the marginal constraints at the initial and final times, we introduce Lagrange multiplier functions $\lambda_0(\mathbf{x}_0)$ and $\lambda_N(\mathbf{x}_N)$ and form the augmented functional

$$\begin{aligned} \mathcal{F}[\mathcal{H}; \lambda_0, \lambda_N] = D_{KL}^{\text{path}}(\mathcal{H} \| \mathcal{G}) &+ \int d\mathbf{y}\, \lambda_0(\mathbf{y}) \left[\int d\mathbf{x}_{1:N}\, \mathcal{H}(\mathbf{y}, \mathbf{x}_{1:N}) - P_0(\mathbf{y}) \right] \\ &+ \int d\mathbf{y}\, \lambda_N(\mathbf{y}) \left[\int d\mathbf{x}_{0:N-1}\, \mathcal{H}(\mathbf{x}_{0:N-1}, \mathbf{y}) - P_T(\mathbf{y}) \right]. \end{aligned} \tag{A.4}$$

Treating the variation of $\mathcal{F}$ with respect to $\mathcal{H}$ as a functional derivative in the space of probability measures, and setting it to zero, gives

$$\ln \frac{\mathcal{H}^*}{\mathcal{G}} + 1 + \lambda_0(\mathbf{x}_0) + \lambda_N(\mathbf{x}_N) = 0, \tag{A.5}$$

which rearranges to $\mathcal{H}^*(\mathbf{x}_{0:N}) = f(\mathbf{x}_0)\, \mathcal{G}(\mathbf{x}_{0:N})\, g(\mathbf{x}_N)$, where the endpoint functions are identified as

$$f(\mathbf{x}_0) = e^{-\frac{1}{2} - c - \lambda_0(\mathbf{x}_0)}, \tag{A.6a}$$

$$g(\mathbf{x}_N) = e^{-\frac{1}{2} + c - \lambda_N(\mathbf{x}_N)}. \tag{A.6b}$$

The constant c reflects a global symmetry, since the transformation $f \to e^{\alpha} f$, $g \to e^{-\alpha} g$ leaves the product fg invariant, and it is fixed once the endpoint marginals are enforced below.

To fix the multipliers explicitly, we enforce the endpoint marginals. The initial marginal of $\mathcal{H}^*$ is

$$P_0(\mathbf{x}_0) = \int d\mathbf{x}_{1:N}\, \mathcal{H}^*(\mathbf{x}_{0:N}) = f(\mathbf{x}_0) P_0(\mathbf{x}_0) \int d\mathbf{x}_{1:N}\, \mathcal{G}(\mathbf{x}_{1:N} \mid \mathbf{x}_0)\, g(\mathbf{x}_N), \tag{A.7}$$

which implies

$$f(\mathbf{x}_0) = \frac{1}{\psi_0(\mathbf{x}_0)}, \qquad \psi_0(\mathbf{x}_0) := \int d\mathbf{x}_{1:N}\, \mathcal{G}(\mathbf{x}_{1:N} \mid \mathbf{x}_0)\, g(\mathbf{x}_N). \tag{A.8}$$

Similarly, enforcing the final marginal gives

$$P_N(\mathbf{x}_N) = g(\mathbf{x}_N) \int d\mathbf{x}_{0:N-1}\, \mathcal{H}^*(\mathbf{x}_{0:N-1}, \mathbf{x}_N) = g(\mathbf{x}_N)\, \bar{\psi}_N(\mathbf{x}_N), \tag{A.9}$$

where the integral over the preceding slices reproduces the forward field $\bar{\psi}_N$ defined below. These relations close the system, confirm the consistency of the forward and backward factorization, and recover the product identity $P_k(\mathbf{x}_k) = \psi_k(\mathbf{x}_k)\, \bar{\psi}_k(\mathbf{x}_k)$ at the endpoints.

A.2 Recursion relations and the continuum limit

We now derive the recursion relations for the auxiliary fields ψ_k and $\bar{\psi}_k$ that characterize the bridge at intermediate times. The backward field is the propagation of the terminal tilt,

$$\psi_k(\mathbf{x}_k) = \int d\mathbf{x}_{k+1:N}\, g(\mathbf{x}_N) \prod_{s=k}^{N-1} G(\mathbf{x}_{s+1} \mid \mathbf{x}_s), \tag{A.10}$$

and by the Chapman–Kolmogorov equation it satisfies the backward recursion

$$\psi_{k-1}(\mathbf{x}_{k-1}) = \int d\mathbf{x}_k\, \psi_k(\mathbf{x}_k)\, G(\mathbf{x}_k \mid \mathbf{x}_{k-1}), \tag{A.11}$$

with terminal condition $\psi_N(\mathbf{x}_N) = g(\mathbf{x}_N)$. Similarly, the forward field is

$$\bar{\psi}_k(\mathbf{x}_k) = \int d\mathbf{x}_{0:k-1}\, f(\mathbf{x}_0) P_0(\mathbf{x}_0) \prod_{s=0}^{k-1} G(\mathbf{x}_{s+1} \mid \mathbf{x}_s), \tag{A.12}$$

and satisfies the forward recursion

$$\bar{\psi}_{k+1}(\mathbf{x}_{k+1}) = \int d\mathbf{x}_k\, G(\mathbf{x}_{k+1} \mid \mathbf{x}_k)\, \bar{\psi}_k(\mathbf{x}_k), \tag{A.13}$$

with initial condition $\bar{\psi}_0(\mathbf{x}_0) = f(\mathbf{x}_0) P_0(\mathbf{x}_0)$. The product identity $P_k(\mathbf{x}_k) = \psi_k(\mathbf{x}_k)\, \bar{\psi}_k(\mathbf{x}_k)$ follows directly from the factorized joint measure by integrating out all other time slices.

To obtain the continuum limit of these recursions, consider the backward recursion Eq. (A.11). As $\Delta t \to 0$, the transition kernel $G(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ is the short-time propagator of a diffusion with drift $\mathbf{v}$ and diffusion coefficient $\gamma(t)$, and the standard Kramers–Moyal expansion (Risken, 1996) yields the adjoint (backward) Kolmogorov equation

$$\partial_t \psi + \mathbf{v} \cdot \nabla \psi + \frac{\gamma(t)}{2} \nabla^2 \psi - V_G \psi = 0, \tag{A.14}$$

with terminal condition $\psi(\mathbf{x}, T) = g(\mathbf{x})$. The killing term V_G appears if the reference process includes a potential allowing for killing, so that probability is not conserved. Similarly, the forward recursion Eq. (A.13) yields the forward Fokker–Planck equation

$$\partial_t \bar{\psi} + \nabla \cdot (\mathbf{v} \bar{\psi}) - \frac{\gamma(t)}{2} \nabla^2 \bar{\psi} + V_G \bar{\psi} = 0, \tag{A.15}$$

with initial condition $\bar{\psi}(\mathbf{x}, 0) = f(\mathbf{x}) P_0(\mathbf{x})$. These two equations are dual under time reversal and form the continuous-time backbone of the Schrödinger bridge.

A.3 SHORT-TIME EXPANSION OF THE TILTED KERNEL

The optimal transition kernel is

$$H^*(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \frac{\psi_{k+1}(\mathbf{x}_{k+1})}{\psi_k(\mathbf{x}_k)}\, G(\mathbf{x}_{k+1} \mid \mathbf{x}_k). \tag{A.16}$$

To extract its short-time form, begin with the short-time reference kernel

$$G(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \frac{1}{(2\pi\gamma_k\Delta t)^{d/2}} \exp\left[-\frac{\|\Delta\mathbf{x} - \mathbf{v}(\mathbf{x}_k)\Delta t\|^2}{2\gamma_k\Delta t} - V_G(\mathbf{x}_k)\Delta t\right], \tag{A.17}$$

where $\Delta\mathbf{x} = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\gamma_k = \gamma(t_k)$. Under the reference measure the increment scales as $\Delta\mathbf{x} \sim \sqrt{\Delta t}$, so the log-ratio of the ψ fields must be expanded to first order in Δt and second order in $\Delta\mathbf{x}$,

$$\begin{aligned}\ln \frac{\psi_{k+1}(\mathbf{x}_{k+1})}{\psi_k(\mathbf{x}_k)} = \Delta\mathbf{x} \cdot \nabla \ln \psi_k(\mathbf{x}_k) + \frac{1}{2}\Delta\mathbf{x}^a \Delta\mathbf{x}^b\, \partial_a\partial_b \ln \psi_k(\mathbf{x}_k) \\ + \Delta t\, \partial_t \ln \psi_k(\mathbf{x}_k) + \mathcal{O}(\Delta t^{3/2}).\end{aligned} \tag{A.18}$$

Substituting this expansion into Eq. (A.16) and combining with the Gaussian exponent of Eq. (A.17), the linear term in $\Delta\mathbf{x}$ completes the square against the reference exponent and produces the effective drift $\mathbf{v}_* = \mathbf{v} + \gamma_k\nabla \ln \psi_k$. The remaining terms of order Δt are exactly those needed to keep the tilted kernel normalized. The tilted kernel is therefore, to leading order, a Gaussian diffusion with the same noise coefficient γ_k and the effective drift

$$\mathbf{v}_*(\mathbf{x}, t) = \mathbf{v}(\mathbf{x}, t) + \gamma(t)\, \nabla \ln \psi(\mathbf{x}, t) \tag{A.19}$$

in the continuum limit. Equation (A.19) is the central dynamical result of the Schrödinger bridge, since the optimal process is obtained by adding the gradient of the log-backward field to the reference drift.

A.4 CONDITIONAL FIELD CONSTRUCTION AND THE POSTERIOR MIXTURE

In the score-matching sector of Sec. 2, we set $V_G = -\nabla \cdot \mathbf{v}$, for which the forward equation admits the constant solution $\bar{\psi} \equiv 1$. The marginal is therefore $P_t = \psi_t$ and the effective drift is $\mathbf{v}_* = \mathbf{v} + \gamma\nabla \ln P_t$. It remains to express $\nabla \ln P_t$ in terms of the training data.

Because the equation governing ψ is linear, its solution can be obtained by superposition. For a single data point $\mathbf{y}$, define the conditional field

$$\psi^{(\mathbf{y})}(\mathbf{x}, t) = G(\mathbf{y}, T \mid \mathbf{x}, t), \tag{A.20}$$

the transition kernel from $(\mathbf{x}, t)$ to the terminal point $\mathbf{y}$ at time T. This field satisfies the same backward Kolmogorov equation as ψ, with terminal condition $\psi^{(\mathbf{y})}(\mathbf{x}, T) = \delta^d(\mathbf{x} - \mathbf{y})$, and by linearity the full field is the average over the data distribution,

$$\psi(\mathbf{x}, t) = \int d\mathbf{y}\, P_{\text{data}}(\mathbf{y})\, \psi^{(\mathbf{y})}(\mathbf{x}, t). \tag{A.21}$$

For the affine Gaussian diffusions used in practice, namely variance-preserving or variance-exploding schedules, the reference kernel is Gaussian, and reversing time via $\tau = T - t$ to pass from the generative orientation to the noising one gives

$$\psi^{(\mathbf{y})}(\mathbf{x}, t) = q_\tau(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\big(\mathbf{x}; \alpha_\tau\mathbf{y},\ \sigma_\tau^2\mathbf{I}\big)\,, \tag{A.22}$$

where α_τ and σ_τ are the signal and noise schedules of the noising process. Suppressing the orientation change in the notation from here on, the marginal is consequently

$$P_t(\mathbf{x}) = \int d\mathbf{y}\, P_{\text{data}}(\mathbf{y})\, q_t(\mathbf{x} \mid \mathbf{y}), \tag{A.23}$$

which is precisely the forward noising marginal.

The conditional bridge kernel for a fixed data point $\mathbf{y}$ follows from the tilted kernel formula,

$$H^{(\mathbf{y})}(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \frac{\psi^{(\mathbf{y})}_{k+1}(\mathbf{x}_{k+1})}{\psi^{(\mathbf{y})}_k(\mathbf{x}_k)}\, G(\mathbf{x}_{k+1} \mid \mathbf{x}_k), \tag{A.24}$$

and by the Markov property of the noising chain this simplifies to the reverse transition of the chain conditioned on the clean sample,

$$H^{(\mathbf{y})}(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = q(\mathbf{x}_{k+1} \mid \mathbf{x}_k, \mathbf{y}) = \frac{q(\mathbf{x}_k \mid \mathbf{x}_{k+1})\, q(\mathbf{x}_{k+1} \mid \mathbf{y})}{q(\mathbf{x}_k \mid \mathbf{y})}. \tag{A.25}$$

Averaging over the data distribution with the posterior weight

$$p(\mathbf{y} \mid \mathbf{x}_k) = \frac{P_{\mathrm{data}}(\mathbf{y})\, q(\mathbf{x}_k \mid \mathbf{y})}{P_k(\mathbf{x}_k)} \tag{A.26}$$

yields the full optimal reverse kernel,

$$H^*(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \int d\mathbf{y}\, p(\mathbf{y} \mid \mathbf{x}_k)\, q(\mathbf{x}_{k+1} \mid \mathbf{x}_k, \mathbf{y}). \tag{A.27}$$

This is the posterior-weighted mixture representation of the optimal bridge. The generative process averages over the posterior of the clean data point given the current noisy observation and then steps according to the conditional noising posterior.

Finally, the same construction expresses the marginal score as a posterior expectation,

$$\nabla \ln P_t(\mathbf{x}) = \frac{\int d\mathbf{y}\, P_{\mathrm{data}}(\mathbf{y})\, q_t(\mathbf{x} \mid \mathbf{y})\, \nabla \ln q_t(\mathbf{x} \mid \mathbf{y})}{P_t(\mathbf{x})} = \mathbb{E}_{p(\mathbf{y}|\mathbf{x})}\left[\nabla \ln q_t(\mathbf{x} \mid \mathbf{y})\right]. \tag{A.28}$$

This is the identity underlying the simulation-free score-matching objective. The unknown marginal score is replaced by a tractable posterior expectation, estimated in practice by sampling a clean data point and adding noise.

A.5 Short-time KL divergence between diffusion kernels

We derive the per-step KL divergence between two diffusion processes that share the same diffusion coefficient but differ in drift, the computation underlying the excess action of Sec. 2. Consider two transition kernels over a small interval Δt with drifts $\mathbf{v}_1(\mathbf{x}_k)$, $\mathbf{v}_2(\mathbf{x}_k)$ and common diffusion coefficient γ_k,

$$H_i(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \frac{1}{(2\pi\gamma_k \Delta t)^{d/2}} \exp\left[-\frac{\|\Delta\mathbf{x} - \mathbf{v}_i(\mathbf{x}_k)\Delta t\|^2}{2\gamma_k \Delta t}\right], \quad i = 1, 2, \tag{A.29}$$

where $\Delta\mathbf{x} = \mathbf{x}_{k+1} - \mathbf{x}_k$. The drifts are evaluated at the initial point $\mathbf{x}_k$, and higher-order terms in the drift do not affect the leading-order divergence. The per-step KL divergence is

$$\begin{aligned} D_{KL}\big(H^{(1)}(\cdot \mid \mathbf{x}_k) \,\|\, H^{(2)}(\cdot \mid \mathbf{x}_k)\big) = \int d\mathbf{x}_k\, d\mathbf{x}_{k+1}\, P_k^{(1)}(\mathbf{x}_k)\, H^{(1)}(\mathbf{x}_{k+1} \mid \mathbf{x}_k) \\ \times \ln \frac{H^{(1)}(\mathbf{x}_{k+1} \mid \mathbf{x}_k)}{H^{(2)}(\mathbf{x}_{k+1} \mid \mathbf{x}_k)}, \end{aligned} \tag{A.30}$$

where $P_k^{(1)}$ is the marginal under the first process, or any measure with the same support, since the leading-order result is independent of the weighting. The log-ratio simplifies because the normalization factors cancel,

$$\ln \frac{H_1}{H_2} = \frac{1}{\gamma_k} \Delta\mathbf{x} \cdot (\mathbf{v}_1 - \mathbf{v}_2) + \frac{\Delta t}{2\gamma_k}\left(\|\mathbf{v}_2\|^2 - \|\mathbf{v}_1\|^2\right). \tag{A.31}$$

Under H_1 the increment is distributed as

$$\Delta\mathbf{x} = \mathbf{v}_1(\mathbf{x}_k)\Delta t + \sqrt{\gamma_k \Delta t}\, \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}), \tag{A.32}$$

and substituting into the log-ratio and averaging over $\boldsymbol{\xi}$ gives

$$\begin{aligned} \mathbb{E}_{H_1}\left[\ln \frac{H_1}{H_2}\right] &= \frac{\Delta t}{\gamma_k}\, \mathbf{v}_1 \cdot (\mathbf{v}_1 - \mathbf{v}_2) + \frac{\Delta t}{2\gamma_k}\left(\|\mathbf{v}_2\|^2 - \|\mathbf{v}_1\|^2\right) \\ &= \frac{\Delta t}{2\gamma_k}\|\mathbf{v}_1 - \mathbf{v}_2\|^2, \end{aligned} \tag{A.33}$$

where the term proportional to $\mathbb{E}[\boldsymbol{\xi}]$ vanishes. This is exact for the expectation of the log-ratio under H_1, because the log-ratio is linear in $\Delta\mathbf{x}$ plus a deterministic constant. The only higher-order correction arises from the time-dependence of the drifts over the interval, which contributes $\mathcal{O}(\Delta t^{3/2})$ or higher. Therefore

$$D_{KL}\big(H_1(\cdot \mid \mathbf{x}_k) \,\|\, H_2(\cdot \mid \mathbf{x}_k)\big) = \frac{\Delta t}{2\gamma_k}\|\mathbf{v}_1(\mathbf{x}_k) - \mathbf{v}_2(\mathbf{x}_k)\|^2 + \mathcal{O}(\Delta t^{3/2}). \tag{A.34}$$

Integrating over the initial state $\mathbf{x}_k$ with weight $P_k^{(1)}(\mathbf{x}_k)$ and taking the continuum limit yields the integral form used in the main text,

$$\Delta\mathcal{S} = \frac{1}{2}\int_0^T \frac{dt}{\gamma(t)}\,\mathbb{E}_{\mathbf{x}\sim P_t}\left[\|\mathbf{v}_H(\mathbf{x},t) - \mathbf{v}_*(\mathbf{x},t)\|^2\right] + \mathcal{O}(\Delta t), \tag{A.35}$$

where the error term vanishes as the discretization is refined. This establishes the connection between the path-space KL divergence and the squared drift difference underlying the score-matching objective.

B Details of the Latent-Geometry Construction

This appendix supplies the derivations for Sec. 3. We compute the ambient second variation and establish the endpoint measurability of the bridge tangent space. We then show that exact path conditioning is singular, work out the exponential-family form of the Gaussian corruption, and carry out the pullback with its endpoint split. The final subsections give the Gaussian evaluation and the radial information identity.

B.1 The ambient second variation

Let $\mathcal{G}$ be a σ-finite reference path measure and

$$\Lambda[u,v] = \ln \int d\mathbf{x}_{0:N}\, \mathcal{G}(\mathbf{x}_{0:N})\, e^{u(\mathbf{x}_0)+v(\mathbf{x}_N)}, \qquad \mathcal{D} = \{(u,v) : \Lambda[u,v] < \infty\}. \tag{B.1}$$

Write $\rho^{(u,v)}$ for the joint law of $(\mathbf{x}_0, \mathbf{x}_N)$ under the tilt family $\mathcal{H}^{(u,v)}$ of Sec. 3.1, and let $S = \delta u(\mathbf{x}_0) + \delta v(\mathbf{x}_N)$ for an admissible direction $(\delta u, \delta v)$.

Lemma B.1 (Ambient Fisher form)**.** *For (u,v) interior to $\mathcal{D}$, $\delta^2\Lambda[(\delta u,\delta v)] = \mathrm{Var}_{\rho^{(u,v)}}[\delta u(\mathbf{x}_0) + \delta v(\mathbf{x}_N)]$.*

Consider $\tau \mapsto \Lambda[u + \tau\,\delta u,\ v + \tau\,\delta v]$, a scalar tilt parameter along the ray $(\delta u, \delta v)$ and unrelated to the log-SNR λ of Sec. 5. By construction

$$\Lambda[u + \tau\delta u, v + \tau\delta v] = \Lambda[u,v] + \ln \mathbb{E}_{\rho^{(u,v)}}\left[e^{\tau S}\right], \tag{B.2}$$

since tilting by τS reweights the endpoint law by exactly $e^{\tau S}$. The second term is the cumulant-generating function of S under $\rho^{(u,v)}$, and differentiating twice at $\tau = 0$ gives

$$\frac{d}{d\tau}\Big|_0 = \mathbb{E}_\rho[S], \qquad \frac{d^2}{d\tau^2}\Big|_0 = \mathrm{Var}_\rho[S], \tag{B.3}$$

which is Lemma B.1. Differentiation under the integral is justified for (u,v) interior to $\mathcal{D}$ by the standard exponential-integrability argument for exponential families (Brown, 1986; Barndorff-Nielsen, 1978). In the interior of $\mathcal{D}$, the expectation $\mathbb{E}_{\mathcal{G}}[e^{(u+\tau\delta u)(\mathbf{x}_0)+(v+\tau\delta v)(\mathbf{x}_N)}]$ is finite on a neighbourhood of $\tau = 0$ and is also analytic there.

Nothing beyond σ-finiteness and exponential integrability was used. In particular, $\mathcal{G}$ need not be a probability measure, need not be Markov, and no finite-dimensional parameterization is involved.

The construction of this appendix, namely the ambient second variation, the finite-parameter immersion, and the pullback, is verified numerically on three reference kernels whose coupling maps $A(\mathbf{x},t)$ differ in structure. These are a Gaussian kernel, a Gaussian kernel with nonlinear drift, and a cubic exponential family in which the latent immersion has codimension one. The identification of the loss floor with the trace of the pullback holds in all three to a relative error below 4×10^{-14}, confirming that it requires neither Gaussianity nor flatness of the latent immersion. The third case also separates the two notions. A nonlinear drift leaves the immersion e-affine, since it only reparametrizes the ray, whereas codimension genuinely curves it. The radial identity $dK/dr = r\, g_{rr}$ is reproduced to a relative error below 10^{-8}.

B.2 Endpoint measurability of the tangent space

The log-likelihood ratio between two nearby members of the tilt family follows directly from its definition in Sec. 3.1,

$$\ln \frac{\mathcal{H}^{(u+\delta u,\, v+\delta v)}}{\mathcal{H}^{(u,v)}}(\mathbf{x}_{0:N}) = \delta u(\mathbf{x}_0) + \delta v(\mathbf{x}_N) - \delta\Lambda, \tag{B.4}$$

so the centered score is

$$\delta\ell = \delta u(\mathbf{x}_0) + \delta v(\mathbf{x}_N) - \mathbb{E}_\rho\big[\delta u(\mathbf{x}_0) + \delta v(\mathbf{x}_N)\big], \tag{B.5}$$

a function of the endpoints alone. Every tangent vector of the bridge family is therefore represented by an endpoint-measurable score. The whole local statistical experiment, namely the collection of likelihood ratios distinguishing infinitesimally separated bridges, is measurable with respect to $\sigma(\mathbf{x}_0, \mathbf{x}_N)$. Lemma B.1 is the second moment of Eq. (B.5).

This is the precise sense in which passing from the path to the endpoint pair loses nothing relevant. By the monotonicity of Fisher information under Markov kernels, any reduction of the path variable can only decrease the Fisher form, with equality for a sufficient reduction (Chentsov, 1982; Ay et al., 2017). Equation (B.5) exhibits the endpoint pair as sufficient for the tangent experiment, so the reduction is lossless at this order.

B.3 Singularity of exact path conditioning

Let $\mathcal{G}$ be a nondegenerate diffusion reference and consider $\mathcal{H}(\cdot \mid X_t = \mathbf{x})$. The event $\{X_t = \mathbf{x}\}$ has probability zero, and the conditional measures for distinct $\mathbf{x}$ are carried by the disjoint path sets $\{\omega : \omega(t) = \mathbf{x}\}$. Hence for $\mathbf{x} \neq \mathbf{x}'$ the two measures are mutually singular and

$$D_{KL}\big(\mathcal{H}(\cdot \mid X_t = \mathbf{x}) \,\big\|\, \mathcal{H}(\cdot \mid X_t = \mathbf{x}')\big) = +\infty. \tag{B.6}$$

A Fisher metric requires $D_{KL}(P_{\theta+d\theta} \| P_\theta) = \frac{1}{2} g_{ij} d\theta^i d\theta^j + O(|d\theta|^3)$, but no such expansion exists here at any order. The obstruction is not specific to Schrödinger bridges but is generic for exact point conditioning of continuous-path measures, and it is the reason the reduction ι_t of Sec. 3.1 conditions the endpoint experiment rather than the path.

B.4 Exponential-family form of the Gaussian corruption

With $q(\mathbf{x}, t \mid \mathbf{x}_N) = \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{x}_N, \sigma_t^2 \mathbf{I})$,

$$\ln q = -\frac{\|\mathbf{x} - \alpha_t \mathbf{x}_N\|^2}{2\sigma_t^2} + \text{const} = \frac{\alpha_t}{\sigma_t^2}\, \mathbf{x} \cdot \mathbf{x}_N \;-\; \frac{\alpha_t^2}{2\sigma_t^2} \|\mathbf{x}_N\|^2 \;-\; \frac{\|\mathbf{x}\|^2}{2\sigma_t^2} + \text{const}, \tag{B.7}$$

which is the exponential-family form of Sec. 3.1 with

$$\mathbf{s}(\mathbf{x}_N) = \Big(\mathbf{x}_N, -\tfrac{1}{2}\|\mathbf{x}_N\|^2\Big), \qquad A(\mathbf{x}, t) = \Big(\frac{\alpha_t}{\sigma_t^2}\mathbf{x},\; \frac{\alpha_t^2}{\sigma_t^2}\Big), \tag{B.8}$$

and $B(\mathbf{x}, t)$ collecting the remaining $\mathbf{x}$-dependent terms. No $\mathbf{x}_N$-dependent term remains outside the statistic, so the carrier of the endpoint family is P_{data} itself. The coefficient of the second-moment statistic is the signal-to-noise ratio $\text{SNR}_t = \alpha_t^2/\sigma_t^2$, so with $\tilde{\mathbf{x}} = \mathbf{x}/\alpha_t$,

$$A(\mathbf{x}, t) = \text{SNR}_t \cdot (\tilde{\mathbf{x}}, 1), \tag{B.9}$$

which is Eq. (3.1). At fixed $\tilde{\mathbf{x}}$ the time derivative is $\partial_t A = (\partial_t \text{SNR}_t)\,(\tilde{\mathbf{x}}, 1) \parallel A$, so the temporal direction of latent spacetime is radial in natural-parameter space and the spatial directions are angular. For an *anisotropic* Gaussian corruption with covariance Σ_t, the term $-\frac{1}{2}\mathbf{x}_N^\top \Sigma_t^{-1} \mathbf{x}_N$ does not reduce to a multiple of $\|\mathbf{x}_N\|^2$, and the sufficient statistic must carry the full symmetric tensor $-\frac{1}{2}\mathbf{x}_N \otimes \mathbf{x}_N$. The natural-parameter space then has $m = d + d(d+1)/2$ dimensions while the latent manifold still has $d + 1$. The image of ι_{lat} therefore acquires codimension, the immersion is no longer onto an open set, and the latent family becomes a curved exponential family with nonvanishing embedding curvature $\Pi^\perp[\partial_\mu \partial_\nu A]$. Isotropy is exactly the condition $m = d + 1$.

B.5 PULLBACK AND ENDPOINT SPLIT

The conditional endpoint law of Sec. 3.1 is an exponential family with carrier $\nu = P_{\mathrm{data}}$ and natural parameter $A(\mathbf{x}, t)$, so Lemma B.1 applies to it verbatim and its Fisher form in natural coordinates is $\mathrm{Cov}_p[\mathbf{s}_a, \mathbf{s}_b]$. Composing with ι_{lat} and using the chain rule gives the factorization of Sec. 3.1,

$$g_{\mu\nu} = \partial_\mu A^a \, \partial_\nu A^b \, \mathrm{Cov}_p[\mathbf{s}_a, \mathbf{s}_b]. \tag{B.10}$$

For the split, the bridge is Markov, so conditioning on $X_t = \mathbf{x}$ makes past and future independent,

$$p(\mathbf{x}_0, \mathbf{x}_N \mid X_t = \mathbf{x}) = p(\mathbf{x}_0 \mid X_t = \mathbf{x}) \, p(\mathbf{x}_N \mid X_t = \mathbf{x}). \tag{B.11}$$

The score of the joint conditional with respect to $\mathbf{x}$ is then the sum of the two individual scores, and because the two factors are independent under the conditional measure the cross-covariance vanishes, giving Proposition 3.1 with

$$g_{ij}^{\mathrm{future}} = \mathrm{Cov}_{p(\mathbf{x}_N \mid \mathbf{x}, t)} \big[\partial_i \ln q(\mathbf{x}, t \mid \mathbf{x}_N), \; \partial_j \ln q(\mathbf{x}, t \mid \mathbf{x}_N) \big], \tag{B.12}$$

$$g_{ij}^{\mathrm{past}} = \mathrm{Cov}_{p(\mathbf{x}_0 \mid \mathbf{x}, t)} \big[\partial_i \ln G(\mathbf{x}, t \mid \mathbf{x}_0), \; \partial_j \ln G(\mathbf{x}, t \mid \mathbf{x}_0) \big]. \tag{B.13}$$

No such cancellation occurs in the three-term expansion of Sec. 3.1, where the endpoints are not conditioned and their covariance is generically nonzero.

B.6 GAUSSIAN EVALUATION

From the Gaussian corruption kernel of Sec. 3.1,

$$\partial_i \ln q(\mathbf{x}, t \mid \mathbf{x}_N) = -\frac{x_i - \alpha_t x_{N,i}}{\sigma_t^2}, \tag{B.14}$$

whose only $\mathbf{x}_N$-dependence is through $\alpha_t x_{N,i}/\sigma_t^2$. Hence

$$g_{ij}^{\mathrm{future}}(\mathbf{x}, t) = \frac{\alpha_t^2}{\sigma_t^4} \, \mathrm{Cov} \big[x_N^i, x_N^j \mid X_t = \mathbf{x} \big]. \tag{B.15}$$

To convert this into Hessian form, write the marginal as an integral over the data distribution, $P_t(\mathbf{x}) = \int d\mathbf{x}_N \, P_{\mathrm{data}}(\mathbf{x}_N) \, q(\mathbf{x}, t \mid \mathbf{x}_N)$. Differentiating once,

$$\partial_i \ln P_t(\mathbf{x}) = \mathbb{E} \big[\partial_i \ln q \mid \mathbf{x} \big] = \frac{\alpha_t \mathbb{E}[x_N^i \mid \mathbf{x}] - x_i}{\sigma_t^2}, \tag{B.16}$$

which is the Tweedie relation (Robbins, 1992; Efron, 2011). Differentiating a second time and using $\partial_j \mathbb{E}[x_N^i \mid \mathbf{x}] = (\alpha_t/\sigma_t^2) \mathrm{Cov}[x_N^i, x_N^j \mid \mathbf{x}]$,

$$\partial_i \partial_j \ln P_t(\mathbf{x}) = -\frac{\delta_{ij}}{\sigma_t^2} + \frac{\alpha_t^2}{\sigma_t^4} \mathrm{Cov} \big[x_N^i, x_N^j \mid \mathbf{x} \big]. \tag{B.17}$$

Combining with Eq. (B.15) gives Eq. (3.2),

$$g_{ij}^{\mathrm{future}}(\mathbf{x}, t) = \frac{1}{\sigma_t^2} \delta_{ij} + \partial_i \partial_j \ln P_t(\mathbf{x}). \tag{B.18}$$

Comparison with Theorem 3.2 is immediate. The theorem's metric is $\mathbb{E}[\partial_i \ln p(\mathbf{x}_N \mid \mathbf{x}, t) \, \partial_j \ln p(\mathbf{x}_N \mid \mathbf{x}, t)]$, and by Eq. (3.5) the posterior score is $\partial_i \ln q - \partial_i \ln P_t$. Its second moment under the posterior is the covariance of $\partial_i \ln q$, which is exactly g^{future}.

B.7 THE RADIAL INFORMATION IDENTITY

Let $p_\theta(\mathbf{x}_N) \propto \nu(\mathbf{x}_N) e^{\theta \cdot \mathbf{s}}$ and $K(\theta) = D_{KL}(p_\theta \| \nu)$. Using $\Lambda(\theta) = \ln \int \nu \, e^{\theta \cdot \mathbf{s}}$ and $\eta = \nabla \Lambda = \mathbb{E}_{p_\theta}[\mathbf{s}]$,

$$K(\theta) = \mathbb{E}_{p_\theta} \big[\theta \cdot \mathbf{s} - \Lambda(\theta) \big] = \theta \cdot \eta(\theta) - \Lambda(\theta), \tag{B.19}$$

so that

$$\nabla_\theta K = \eta + \theta^\top \nabla^2 \Lambda - \nabla \Lambda = g(\theta) \, \theta, \qquad g = \nabla^2 \Lambda = \mathrm{Cov}_{p_\theta}[\mathbf{s}]. \tag{B.20}$$

Restricting to the ray $\theta = r \, n$ with n fixed and normalized,

$$\frac{dK}{dr} = n \cdot \nabla_\theta K = r \, n^\top g \, n = r \, g_{rr}, \tag{B.21}$$

which is the radial identity quoted in Sec. 3.1. Since by Eq. (3.1) the radial coordinate is SNR_t and the angular coordinate is the rescaled latent, the corruption moves along the ray and K decreases monotonically as SNR_t decreases, at a rate set by the radial component of the Fisher metric. At $r = 0$ the conditional endpoint law coincides with the carrier and the latent retains nothing about the data.

C Proofs of the Information-Flow Identities

Lemma C.1 (Bridge weighting is the SNR measure)**.** *Write* $\mathrm{snr}_t := \alpha_t^2/\sigma_t^2$ *for the signal-to-noise ratio. Among the marginal-preserving diffusions that realize a given affine interpolant, the bridge construction singles out the memoryless one, that is, the unique state-independent coefficient for which the rescaled process* $\mathbf{x}_t/\alpha_t$ *has independent increments, namely* $\gamma(t) = 2\sigma_t^2 \frac{d}{dt}\ln(\alpha_t/\sigma_t)$*, and for this coefficient, for any affine interpolant,* $\gamma(t)\,\alpha_t^2/\sigma_t^4 = d\mathrm{snr}_t/dt$*.*

Proof of Lemma C.1. Parametrize time in the generative orientation, so that α_t/σ_t is increasing. A diffusion $d\mathbf{x}_t = f(\mathbf{x}_t, t)\,dt + \sqrt{\gamma(t)}\,d\mathbf{W}_t$ with state-independent coefficients that preserves the marginals $q(\mathbf{x}, t \mid \mathbf{y}) = \mathcal{N}(\mathbf{x}; \alpha_t\mathbf{y}, \sigma_t^2\mathbf{I})$ must match the first two conditional moments, which gives $f(\mathbf{x}, t) = (\dot\alpha_t/\alpha_t)\,\mathbf{x}$ and $\dot\sigma_t^2 = 2(\dot\alpha_t/\alpha_t)\sigma_t^2 - \gamma(t)$. Hence $\gamma(t) = 2\sigma_t^2\frac{d}{dt}\ln(\alpha_t/\sigma_t)$, and the rescaled process $\mathbf{x}_t/\alpha_t$ is then a time-changed Brownian motion, so its increments are independent. A state-dependent coefficient would couple the increment variance to the state, so the memoryless coefficient is unique. Finally, $\gamma(t)\,\alpha_t^2/\sigma_t^4 = 2(\alpha_t^2/\sigma_t^2)\,\frac{d}{dt}\ln(\alpha_t/\sigma_t) = \mathrm{snr}_t\,\frac{d}{dt}\ln\mathrm{snr}_t = d\mathrm{snr}_t/dt$. □

Proof of Theorem 4.1, first equality (the Fisher-information gap). By Eq. (3.5) and the centering identity, for fixed $\mathbf{x}$

$$\begin{aligned}\mathbb{E}_{\mathbf{y}|\mathbf{x}}\big\|\nabla_{\mathbf{x}}\ln p(\mathbf{y}\mid\mathbf{x},t)\big\|^2 &= \mathbb{E}_{\mathbf{y}|\mathbf{x}}\big\|\nabla_{\mathbf{x}}\ln q(\mathbf{x},t\mid\mathbf{y})\big\|^2 - 2\,\nabla\ln P_t\cdot\mathbb{E}_{\mathbf{y}|\mathbf{x}}\big[\nabla_{\mathbf{x}}\ln q\big]\\ &\quad + \big\|\nabla\ln P_t\big\|^2\\ &= \mathbb{E}_{\mathbf{y}|\mathbf{x}}\big\|\nabla_{\mathbf{x}}\ln q(\mathbf{x},t\mid\mathbf{y})\big\|^2 - \big\|\nabla\ln P_t(\mathbf{x})\big\|^2.\end{aligned}$$

Averaging over $\mathbf{x}\sim P_t$ and using the tower property, $\mathbb{E}_{P_t}\mathbb{E}_{\mathbf{y}|\mathbf{x}}\|\nabla\ln q\|^2 = \mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{x}|\mathbf{y}}\|\nabla\ln q(\cdot\mid\mathbf{y})\|^2 = \mathbb{E}_{\mathbf{y}}J(q_t(\cdot\mid\mathbf{y}))$. □

Proof of Theorem 4.1, second equality (de Bruijn). Any solution of the Fokker–Planck equation $\partial_t p = -\nabla\cdot(fp) + \frac{D}{2}\nabla^2 p$ satisfies $\frac{d}{dt}h(p_t) = \mathbb{E}_{p_t}[\nabla\cdot f] + \frac{D}{2}J(p_t)$, by two integrations by parts. Apply this to P_t and to $q_t(\cdot\mid\mathbf{y})$ for each $\mathbf{y}$, and average the latter over $\mathbf{y}$. The drift terms then agree by the tower property, $\mathbb{E}_{P_t}[\nabla\cdot f] = \mathbb{E}_{\mathbf{y}}\,\mathbb{E}_{q_t(\cdot|\mathbf{y})}[\nabla\cdot f]$, and cancel in $I(\mathbf{y};\mathbf{x}_t) = h(\mathbf{x}_t) - h(\mathbf{x}_t\mid\mathbf{y})$, leaving the Fisher-information gap, which the first equality converts into the metric trace. □

Proof of Theorem 4.2. Theorem 3.2 with the Gaussian evaluation Eq. (3.2) gives the integrand $\gamma(t)(\alpha_t^2/\sigma_t^4)\,\mathrm{tr}\,\mathrm{Cov}[\mathbf{y}\mid\mathbf{x}]$. By Lemma C.1 the prefactor is $d\mathrm{snr}/dt$, so the t-integral becomes $\frac{1}{2}\int \mathrm{MMSE}\,d\mathrm{snr}$. The second equality is the I–MMSE relation of Guo, Shamai and Verdú (Guo et al., 2005), $dI(\mathbf{y};\mathbf{x}_{\mathrm{snr}})/d\mathrm{snr} = \frac{1}{2}\mathrm{MMSE}(\mathrm{snr})$. The third holds because $h(\mathbf{x}_{\mathrm{snr}}\mid\mathbf{y}) = h(\mathbf{z})$ is independent of snr, so $I = h(\mathbf{x}_{\mathrm{snr}}) - h(\mathbf{z})$ and the constant cancels in the difference. In the log-SNR coordinate $\lambda = \ln\mathrm{snr}$ the same statement reads $\frac{1}{2}\int M(\lambda)\,d\lambda$ with M of Sec. 5.1, since $d\mathrm{snr} = \mathrm{snr}\,d\lambda$ and $M = \mathrm{snr}\,\mathrm{MMSE}$. □

D Numerical and Experimental Details

This appendix collects the numerical verifications quoted in the main text. The identity checks in Secs. D.1–D.6 are analytic or low-dimensional by design, so that every quantity can be computed exactly or by quadrature. They establish identities rather than performance. Section E gives the experimental details of the CIFAR-10 measurement of Sec. 5.2, and Appendix F verifies the interpolant results of Proposition F.1.

D.1 de Bruijn identity for a nonlinear drift

For the nonlinear corruption drift $f(x) = -ax^3$, we solved the Fokker–Planck equation directly, by explicit finite-difference time-stepping on a uniform grid of 3001 points over $[-12, 12]$ ($dt = 2\times10^{-5}$, 3×10^4 steps), and evaluated both sides of Eq. (4.1). The two sides agree with a median relative error of 1.7×10^{-3} point-wise in time, and the integrated form agrees to 0.4%.

D.2 THE SCHEDULE IDENTITY

For the linear interpolant $\alpha_t = t$, $\sigma_t = 1 - t$, the bridge coefficient of Lemma C.1 is $\gamma(t) = 2\sigma_t^2 \frac{d}{dt} \ln(\alpha_t/\sigma_t) = 2(1-t)/t$, so the left-hand side of the schedule identity of Lemma C.1 is $2\frac{1-t}{t}\frac{t^2}{(1-t)^4} = \frac{2t}{(1-t)^3}$, which equals $\frac{d}{dt}\frac{t^2}{(1-t)^2} = \frac{d\mathrm{snr}_t}{dt}$, as Lemma C.1 claims. The trigonometric and polynomial interpolants follow by the same computation, and we have verified the schedule identity of Lemma C.1 numerically to a relative error of 3×10^{-6} for four distinct interpolants.

D.3 ESTIMATOR VALIDATION ON A GAUSSIAN MIXTURE

We validate Eq. (5.1) on an isotropic Gaussian mixture in $d = 12$ dimensions, where the posterior mean and covariance are closed-form and $M(\lambda)$ can be evaluated exactly. Across six log-SNR values spanning $\lambda \in [-4, 4]$, the identity $d - \sigma_\lambda \mathbb{E}[\mathrm{tr}\, \partial_{\mathbf{x}} \epsilon_\theta] = e^\lambda \mathbb{E}[\mathrm{tr}\, \mathrm{Cov}[\mathbf{y} \mid \mathbf{x}]]$ holds to a worst-case relative error of 4×10^{-11} (finite-difference Jacobian, exact posterior on both sides). Replacing the finite-difference Jacobian with a single-probe Hutchinson estimator and integrating over $\lambda \in [-6, 6]$ on a 49-point grid, the resulting floor agrees with the one computed directly from the exact posterior covariance to a relative difference of 7×10^{-4}, the residual expected from a single Rademacher probe per sample. Code: `floor_jacobian.py`.

D.4 THE FLOOR–ENTROPY IDENTITY

Figure 2 evaluates Theorem 4.2 on a two-mode Gaussian mixture for which the differential entropy $h(\mathbf{x}_{\mathrm{snr}})$ is computable by quadrature. The accumulated floor $\frac{1}{2}\int \mathrm{MMSE}\, d\mathrm{snr}$ and the entropy change Δh agree to a relative error of 2×10^{-5}.

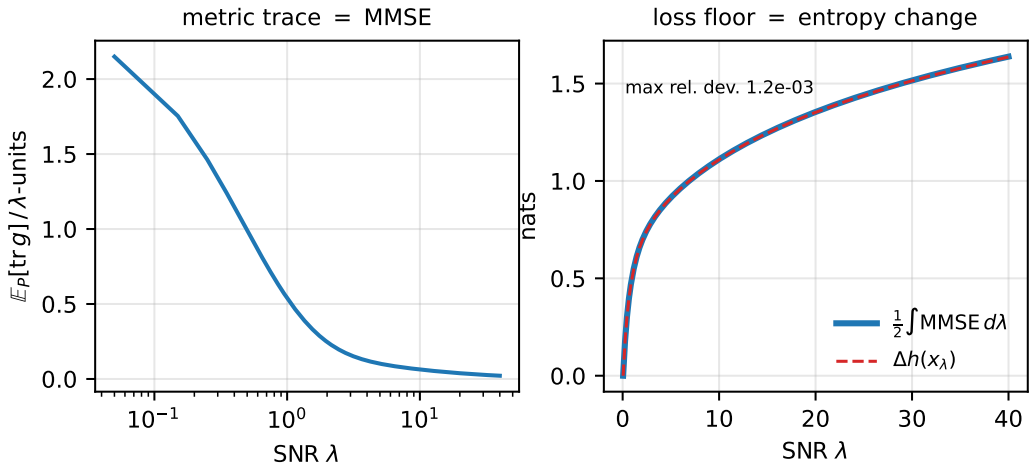

Figure 2: The loss floor is an entropy. **Left:** the trace of the latent Fisher–Rao metric, equivalently the MMSE of endpoint estimation, along the flow. **Right:** the accumulated floor $\frac{1}{2}\int \mathrm{MMSE}\, d\mathrm{snr}$ (solid) against the differential-entropy change $\Delta h(\mathbf{x}_{\mathrm{snr}})$ (dashed), verifying Theorem 4.2. Two-mode Gaussian mixture, entropies by quadrature.

D.5 RANKING INVERSION IN AN ANALYTIC MODEL

Before the CIFAR-10 measurement of Sec. 5.2 we verified that the inversion is not an artefact of estimation, by exhibiting it in a model where every quantity is exact. The example uses the analytic two-mode mixture. The two SNR ranges are A (narrow) and B (wide). The two models are exact scores scaled by $c = 0.97$ (better) and $c = 0.90$ (worse). Raw losses are 1.0490 and 1.1267 on A and 1.5455 and 1.8601 on B. The floors, computed from Theorem 3.2 as expectations of the posterior covariance, are 1.0414 and 1.5143. The floor-subtracted excesses are $0.0077 < 0.0854$ on A and $0.0311 < 0.3458$ on B. Read across the two ranges, the raw loss ranks the worse model on A above the better model on B ($1.1267 < 1.5455$), while the floor-subtracted excesses rank them correctly ($0.0311 < 0.0854$).

D.6 THE DISCRETE FLOOR

For $n = 3$ tokens over an alphabet of size 3 with total correlation 0.46 nats, Eq. (6.1) reproduces the data entropy $H(y_{1:n}) = 2.58757$ to a relative error of 8×10^{-8}. Integrating in t against the weight $-\dot{m}_t/m_t$ returns the same value for the masking schedules $m = 1 - t$, $\cos(\pi t/2)$, and $(1 - t)^3$. The remaining schedule we tried, $m = \sqrt{1 - t}$, returns 2.830 on a uniform grid, a 9.4% discrepancy. This is a quadrature artefact rather than a failure of the identity. The weight $-\dot{m}_t/m_t = 1/(2(1 - t))$ has an integrable singularity at $t = 1$ for this schedule, which a uniform grid resolves poorly. Refining the grid towards the endpoint moves the value monotonically back. Clustering as $t = 1 - s^2$ gives 2.601 (0.5%), and the substitution $u = m_t$, which is what the schedule-independence argument itself uses, removes the singularity and returns H to the same accuracy as the other three schedules.

E EXPERIMENTAL DETAILS FOR THE CIFAR-10 MEASUREMENT

Corruption and objective. All runs use the variance-preserving corruption with $\alpha_\lambda^2 = \mathrm{sigmoid}(\lambda)$ and $\sigma_\lambda^2 = 1 - \alpha_\lambda^2$, so that λ is the log-SNR exactly. Training samples λ uniformly on the configured range and minimizes the ϵ-prediction loss under the evidence-bound weighting $w = |\dot{\lambda}|$, for which $\omega \equiv 1$ in the notation of Eq. (4.3) and the reported loss is in nats per image. Data are the 50k CIFAR-10 training images scaled to $[-1, 1]$, with no augmentation. The floor is a property of the data distribution, so the distribution the models see must be fixed and identical across runs. The 10k test images are held out for every number reported.

Configurations. Three log-SNR ranges $\lambda \in [-8, \lambda_1]$ with $\lambda_1 \in \{6.0, 9.2, 13.8\}$, called *narrow*, *mid*, and *wide*. The upper endpoints correspond to $\sigma_{\min} \approx 5 \times 10^{-2}$, 10^{-2}, and 10^{-3}. The lower endpoint is common to all three, which is what makes the floor difference attributable to λ_1 alone.

Network and optimization. A standard UNet with 12.0M parameters: base width 64, channel multipliers $(1, 2, 2)$ over three resolutions, two residual blocks per resolution, self-attention at 16×16, dropout 0.1, and sinusoidal λ embedding. Adam at learning rate 2×10^{-4} with 1k linear warmup steps, batch size 128, gradient-norm clipping at 1.0, 60k steps per run, no exponential moving average (the quantity of interest is the training objective, not sample quality). Checkpoints are saved at 2k, 5k, 10k, 20k, 35k, and 60k steps, giving six model qualities per configuration from one run.

Evaluation with common random numbers. Every checkpoint is scored on an identical stream of $(\mathbf{y}, \lambda, \epsilon)$ triples drawn once and reused. The reported and common-range losses each average 20 batches of 100 held-out images, and the floor curve is evaluated on a grid of 29 log-SNR nodes with 8 batches of 50 images (400 images) per node. Without the shared stream the paired differences that separate checkpoints, of order 10 nats, would be buried under a Monte-Carlo standard error of order 10^2 nats. The same stream serves the loss and the floor, so the excess is a paired difference throughout.

Floor estimation. The floor integrand is Eq. (5.1), $M(\lambda) = d - \sigma_\lambda \mathbb{E}[\mathrm{tr}\, \partial_{\mathbf{x}} \epsilon_\theta]$ with $d = 3072$, and the trace is taken with one Rademacher probe per evaluation sample through a single vector–Jacobian product. The model is run in evaluation mode. Integration in λ uses the trapezoid rule on the 29-node grid. Only converged checkpoints ($\geq$ 20k steps) contribute to the consensus floor, and such a checkpoint contributes at λ only if λ lies within its own training range. Outside that range the network is uninformative, the Jacobian is near zero, and Eq. (5.1) saturates at d, which would bias the consensus upward. The consensus floor at each node is the mean over contributing checkpoints, and the quoted uncertainties are the standard deviation across them.

Cost. Two runs of 60k steps at batch 128, roughly 1.5–4 hours each on a single modern GPU, plus about 20 minutes of evaluation. The third configuration is optional and strengthens the rank statistic without being needed for the inversion claim. The experiment is therefore reproducible on one GPU in under a day.

What floor subtraction does not fix. Excesses computed on different ranges remain integrals over different domains, so subtracting the floor removes the dominant model-independent confound without making the residuals exactly commensurate. Since all models on a range share one floor constant, floor uncertainty cancels in comparisons within a range and enters cross-range comparisons

only through the difference, which is determined here to 0.95%. The practice we recommend is to report the floor-subtracted excess together with an evaluation on one common range.

Caveat on the exact estimator. We also implemented the exact empirical-posterior floor, which computes $\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}]$ against the 50k training images directly. It is correct at low SNR and we use it there as a cross-check, but it cannot be used across the full range. The information available from an empirical distribution over N points is capped at $\ln N \approx 10.8$ nats, while the population floor over a useful range is in the thousands, and in $d = 3072$ the posterior collapses onto the nearest training image by about $\lambda = 0$. This is why the Jacobian estimator is used at scale.

F The Loss Floor for Flow Matching and Stochastic Interpolants

The decomposition of Theorem 3.2 was stated for the denoising score-matching objective, in which the regression target is the conditional score of the corruption kernel. Nothing in the argument depends on that particular target. This appendix shows that the entire family of affine interpolant objectives, namely ϵ-prediction, $\mathbf{x}_0$-prediction, v-prediction, flow matching, and stochastic interpolants, shares the same floor, up to a weight that is available in closed form. The geometric content of the floor is therefore a property of the corruption, not of the parametrization chosen to learn against it.

F.1 An affine target family

Let the interpolant be

$$\mathbf{x}_t \;=\; \alpha_t\,\mathbf{y} \;+\; \sigma_t\,\mathbf{z}, \qquad \mathbf{y} \sim P_{\mathrm{data}}, \quad \mathbf{z} \sim \mathcal{N}(0, I), \tag{F.1}$$

with $\alpha_t, \sigma_t > 0$ differentiable and σ_t bounded away from zero on the interior of the training range. Equation (F.1) covers the variance-preserving and variance-exploding diffusions, for which $\alpha_t^2 + \sigma_t^2 = 1$ or $\alpha_t \equiv 1$, and the linear interpolant $\alpha_t = t$, $\sigma_t = 1 - t$ of rectified flow and flow matching (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023).

Every objective in the family regresses a network $T_\theta(\mathbf{x}_t, t)$ onto a target that is affine in the two sources of randomness,

$$T(\mathbf{y}, \mathbf{z}, t) \;=\; a_t\,\mathbf{y} \;+\; b_t\,\mathbf{z}, \tag{F.2}$$

under a loss of the form $\mathcal{L} = \frac{1}{2}\int dt\, w(t)\, \mathbb{E}\,\|T_\theta(\mathbf{x}_t, t) - T(\mathbf{y}, \mathbf{z}, t)\|^2$. The choices in use are collected in Table 2.

F.2 The floor of an affine target

Proposition F.1 (Floor of an affine interpolant objective)**.** *Define*

$$c_t \;:=\; a_t - b_t\,\frac{\alpha_t}{\sigma_t}. \tag{F.3}$$

Then the conditional covariance of the target given the interpolant state is

$$\mathrm{Cov}\big[\,T \,\big|\, \mathbf{x}_t\,\big] \;=\; c_t^2\;\mathrm{Cov}\big[\,\mathbf{y} \,\big|\, \mathbf{x}_t\,\big], \tag{F.4}$$

and consequently the objective decomposes as $\mathcal{L} = \mathcal{L}_{\mathrm{ideal}} + \mathrm{floor}$ *with*

$$\mathrm{floor} \;=\; \frac{1}{2}\int dt\; w(t)\, c_t^2\; \mathbb{E}_{\mathbf{x}\sim P_t}\big[\,\mathrm{tr}\,\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}_t]\big]. \tag{F.5}$$

Proof. Solving Eq. (F.1) for the noise gives $\mathbf{z} = (\mathbf{x}_t - \alpha_t\mathbf{y})/\sigma_t$, so that on the event $\mathbf{x}_t = \mathbf{x}$,

$$T \;=\; a_t\mathbf{y} + \frac{b_t}{\sigma_t}\big(\mathbf{x} - \alpha_t\mathbf{y}\big) \;=\; \Big(a_t - \frac{b_t\alpha_t}{\sigma_t}\Big)\mathbf{y} \;+\; \frac{b_t}{\sigma_t}\,\mathbf{x}.$$

The second term is $\mathbf{x}$-measurable and therefore contributes nothing to the conditional covariance, which gives Eq. (F.4). Since T is an unbiased estimator of $\mathbb{E}[T \mid \mathbf{x}_t]$, the argument of Theorem 3.2 applies verbatim. The expected squared error splits into a term evaluated at the conditional mean and the conditional variance of the target, and the latter is independent of θ. Taking the trace and integrating against w yields Eq. (F.5). □

For the flow-matching velocity the coefficient takes a compact form. With $r_t := \alpha_t/\sigma_t$ and $(a_t, b_t) = (\dot\alpha_t, \dot\sigma_t)$,

$$c_t \;=\; \dot\alpha_t - \dot\sigma_t \frac{\alpha_t}{\sigma_t} \;=\; \frac{\dot\alpha_t\sigma_t - \alpha_t\dot\sigma_t}{\sigma_t} \;=\; \sigma_t\,\dot r_t, \tag{F.6}$$

so the flow-matching floor is governed by the rate at which the signal-to-noise amplitude ratio changes along the interpolant. Comparing with the v-prediction entry of Table 2 gives

$$\frac{c_t^{\mathrm{fm}}}{c_t^{v}} \;=\; \frac{W_t}{\alpha_t^2 + \sigma_t^2}, \qquad W_t := \alpha_t\dot\sigma_t - \dot\alpha_t\sigma_t, \tag{F.7}$$

with W_t the Wronskian of the interpolant, equal to 1 for the canonical parametrizations in use (the linear interpolant, and $\alpha_t = \cos t$, $\sigma_t = \sin t$). In this normalization the rectified-flow velocity is a v-parametrization in disguise. The two carry identical floors exactly when the interpolant is variance-preserving, $\alpha_t^2 + \sigma_t^2 = 1$. For the linear interpolant $\alpha_t = 1 - t$, $\sigma_t = t$ the identification fails by the factor $\alpha_t^2 + \sigma_t^2$, which ranges over $[\frac{1}{2}, 1]$. The two objectives then differ by a bounded but non-constant reweighting of the same floor rather than by nothing at all.

Table 2: The coefficient c_t of Eq. (F.3) for the standard targets, written with $r_t := \alpha_t/\sigma_t$, so that $r_t^2 = e^{\lambda_t}$ is the signal-to-noise ratio. Every entry shares the same conditional covariance $\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}_t]$; only the prefactor differs.

target	(a_t, b_t)	c_t	c_t^2
ϵ-prediction	$(0,\ 1)$	$-r_t$	e^{λ_t}
$\mathbf{x}_0$-prediction	$(1,\ 0)$	1	1
v-prediction	$(-\sigma_t,\ \alpha_t)$	$-(\sigma_t^2 + \alpha_t^2)/\sigma_t$	$(\alpha_t^2 + \sigma_t^2)^2/\sigma_t^2$
flow-matching velocity	$(\dot\alpha_t,\ \dot\sigma_t)$	$\sigma_t\,\dot r_t$	$\sigma_t^2\dot r_t^{\,2}$

F.3 Consequences

Equation (F.5) contains a single data-dependent object, the denoising posterior covariance $\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}_t]$, whose trace is the minimum mean squared error of estimating the clean sample from the interpolant state. Writing the floor in the log-SNR coordinate and using $\mathrm{tr}\,\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}] = e^{-\lambda}M(\lambda)$ with M the integrand of Eq. (5.1),

$$\text{floor} \;=\; \frac{1}{2}\int \omega(\lambda)\, M(\lambda)\, d\lambda, \qquad \omega(\lambda) \;:=\; \frac{w}{|\dot\lambda|}\, c^{\,2}\, e^{-\lambda}, \tag{F.8}$$

the integral running over the traversed range of λ in the direction of increasing corruption. For ϵ-prediction with the evidence-bound weighting, $c^2 = e^{\lambda}$ and $w = |\dot\lambda|$, so $\omega \equiv 1$ and Eq. (F.8) reduces to the expression used in the main text. Every other member of the family is obtained by substituting its entry from Table 2. The geometry is unchanged and only ω moves.

Two objectives from this family, evaluated on the same data and the same interpolant, report floors in the ratio $\omega_1(\lambda)/\omega_2(\lambda)$ point-wise in λ. Since that ratio is explicit, the conversion between a reported flow-matching loss and a reported diffusion loss is available in closed form. But the two numbers are not interchangeable as reported, for the same reason that two diffusion losses on different log-SNR ranges are not. The recommendation of Sec. 5 therefore extends across objective families, not only across schedules within one.

The floor vanishes only if $c_t \equiv 0$, that is, $a_t/b_t = \alpha_t/\sigma_t$, in which case by Eq. (F.4) the target is a deterministic multiple of $\mathbf{x}_t$ and the regression problem is empty. For every parametrization that carries information about $\mathbf{y}$, the floor is strictly positive wherever the denoising posterior has non-zero variance. Choosing ϵ- rather than $\mathbf{x}_0$-prediction, or flow matching rather than score matching, redistributes the floor across noise levels. It cannot remove it.

For interpolants that bridge two data distributions rather than data and noise, $\mathbf{x}_t = \alpha_t\mathbf{y}_1 + \beta_t\mathbf{y}_0 + \gamma_t\mathbf{z}$ with $\mathbf{y}_0 \sim P_0$, $\mathbf{y}_1 \sim P_1$ and $\mathbf{z}$ independent (Albergo et al., 2023), the velocity target is $\dot\alpha_t\mathbf{y}_1 + \dot\beta_t\mathbf{y}_0 + \dot\gamma_t\mathbf{z}$ and the same elimination applies after solving for any one of the three sources. The floor is then the conditional covariance of the remaining pair, contracted with the corresponding coefficient vector. It is again a property of the interpolant and the two marginals alone. The single-source case of Proposition F.1 is recovered when $\beta_t \equiv 0$. We do not pursue the general case here, since the objectives in practical use are affine in a single latent source.

F.4 NUMERICAL VERIFICATION

Every claim above is checked against a Gaussian mixture, for which the denoising posterior $p(\mathbf{y} \mid \mathbf{x}_t)$ is available in closed form, in $d = 1$ on the trigonometric variance-preserving interpolant and in $d = 8$ on the linear one. Table 3 collects the results.

Table 3: Numerical verification of Appendix F. Identities that are exact in the estimator are reported as relative error; identities involving Monte-Carlo averages are reported as the deviation of the paired difference from zero in units of its own standard error, which is the appropriate test when two unbiased estimators of the same quantity are compared on shared samples.

claim	tested quantity	$d = 1$, VP	$d = 8$, linear
Eq. (F.4)	$\operatorname{tr}\operatorname{Cov}[T \mid \mathbf{x}] / \operatorname{tr}\operatorname{Cov}[\mathbf{y} \mid \mathbf{x}]$ vs. c_t^2	4×10^{-16}	4×10^{-16}
posterior	bin-conditioned $\operatorname{Cov}[\mathbf{y} \mid \mathbf{x}]$ vs. closed form	$1.4\,\sigma$	—
Eq. (F.5)	$\mathcal{L} - \mathcal{L}_{\mathrm{ideal}}$ vs. floor	$0.5\,\sigma$	$0.9\,\sigma$
Eq. (F.8)	$\int dt$ form vs. $\int \omega M \, d\lambda$ form	2×10^{-16}	0
Eq. (5.1)	$d - \sigma \mathbb{E}[\operatorname{tr} \partial_{\mathbf{x}} \epsilon^\star]$ vs. $M(\lambda)$	3×10^{-15}	2×10^{-14}
Table 2	each c_t vs. differentiated schedule	6×10^{-11}	5×10^{-11}
Eq. (F.7)	$c_t^{\mathrm{fm}} / c_t^v$ vs. $(\alpha_t^2 + \sigma_t^2)^{-1}$	exact	exact

Three of these deserve comment. The first row is exact to machine precision rather than merely accurate, because Eq. (F.4) is an identity between the conditional laws and not an approximation. Once $\mathbf{z}$ is eliminated, the four targets are deterministic affine functions of the same random variable $\mathbf{y}$, so the measured ratio is c_t^2 for any sample whatsoever. The floor itself is not exact in this sense, which is why the decomposition is tested separately and statistically in the third row. There, $\mathcal{L} - \mathcal{L}_{\mathrm{ideal}}$ and the floor are estimated from the same draws, the first as a plain sample average and the second in Rao–Blackwellized form, so their difference has a standard error an order of magnitude below that of either estimate alone, and agreement at a fraction of a standard error is a sharper statement than the raw relative discrepancy of about 3×10^{-4} would suggest. The predictors used are the exact conditional mean and two deliberately spoiled versions of it, a shrunk one and a biased one, confirming that the floor does not move with θ.