MAXIMUM LIKELIHOOD AS A PATH INTEGRAL OF SCORE MATCHING ALONG LANGEVIN DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the maximum-likelihood gradient of energy-based models with continuous observable variables. Writing the negative-phase sample as the endpoint of a Langevin diffusion started at the data point, and applying Itô's lemma, we show that the log-likelihood gradient is equivalent to a path integral of score matching along the diffusion. Three consequences follow. First, the integrand vanishes at the equilibrium of the diffusion Markov chain, providing a closed-form target for mixing diagnostics. Second, for free energies that consist of a Gaussian term plus a locally affine score of the interactions, the integrand reduces to the ϵ-prediction loss of denoising diffusion with likelihood weighting. Third, energy-based models can be trained by scheduled noise injection. All three claims are verified experimentally, furthermore, we show that an energy-based model can be trained on CIFAR-10 at roughly one tenth of the per-epoch cost of persistent contrastive divergence.

1 INTRODUCTION

Energy-based models (EBMs) define a density $p_\theta(\mathbf{x}) \propto \exp\{-\mathcal{F}_\theta(\mathbf{x})\}$ through a free energy $\mathcal{F}_\theta$ and leave the normalizing constant implicit (LeCun et al., 2006). The flexibility of modeling is gained at the expense of a heavier training and inference process, due to the sampling requirement to estimate the negative-phase gradient. The Maximum-likelihood (ML) gradient requires samples from p_θ, and contrastive divergence and its persistent variants (Hinton, 2002; Carreira-Perpiñán & Hinton, 2005; Tieleman, 2008) run a Markov chain to obtain them.

Score matching, as introduced by (Hyvärinen, 2005), avoids sampling altogether by minimizing the Fisher divergence between the data and the model, this however requires the computation of an expensive hessian term, or an approximation to bypass the intractability (Song et al., 2019). Furthermore, Koehler et al. (2023) showed that compared to maximum likelihood, score matching is efficient for well behaved distributions, and can be less efficient for poorly conditioned ones, hence appropriately designed or generalized score matching can recover much better statistical efficiency than vanilla score matching as demonstrated by (Qin & Risteski, 2024)

Denoising diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b), on the other hand, sidestep both issues by learning the score of noised data, the denoising objective has a form of score matching, and was originally justified as a Gedankenexperiment of sort (Vincent, 2011; Saremi & Hyvärinen, 2019). Recently, Song et al. (2021a); Kingma et al. (2021) showed that the denoising objective can be seen as a variational bound on the log-likelihood.

The three approaches of learning data distributions are often presented as alternatives to each other, yet Hyvärinen (2007) showed that score matching can be seen as the first order term of contrastive divergence with Langevin transitions, and that the two coincide in the limit of an infinitesimal step. We make that connection exact, by means of Itô's lemma, and our contributions are the following:

- *The ML-Path Integral of Score Matching Correspondence.* For an EBM with continuous observable variables, the log-likelihood gradient, equals the negative path integral along the Langevin diffusion started at the data point, of the gradient of the score matching objective evaluated on the marginals of the diffusion.
- *Stationarity of the Langevin Markov Chain.* The integrand has expectation zero at equilibrium. Monitoring it along a set of chains gives a diagnostic test for the burn-in of the

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

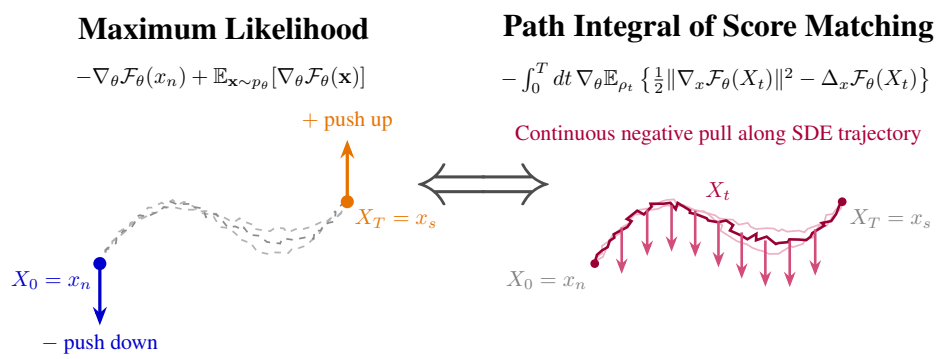

Figure 1: Two equivalent perspectives of the gradient update on a stochastic process. **(Left)** The boundary formulation forces optimization discretely at the terminals $X_0 = x_n$ and $X_T = x_s$. **(Right)** Our formulation distributes the gradient continuously along the stochastic paths $(X_t)_{t\geq 0}$, optimizing the full trajectory.

Langevin Markov chain. We demonstrate the test's properties experimentally, we show that it recovers the true theoretical gaussian relaxation rate, and that the test is a local equilibrium test, thus presenting its failure mode.

- *Denoising Diffusion as a special case.* For a Gaussian free energy plus an interaction term, whose derivative is locally affine, the marginals are Gaussian and the integrand is the ϵ-prediction loss of DDPM with likelihood weighting. This yields a sampler free training scheme for EBMs (Algorithm 1) that we test on CIFAR-10.

2 Background

Energy-based models. Let $\mathcal{F}_\theta : \mathbb{R}^K \to \mathbb{R}$ be $\mathcal{C}^2$ in $\mathbf{x}$ and in θ, and set $p_\theta(\mathbf{x}) = Z(\theta)^{-1}\exp\{-\mathcal{F}_\theta(\mathbf{x})\}$ with $Z(\theta) = \int \exp\{-\mathcal{F}_\theta(\mathbf{x})\}\,\mathrm{d}\mathbf{x} < \infty$. For an observation x_n,

$$\nabla_\theta \log p_\theta(x_n) = \mathbb{E}_{\mathbf{x}\sim p_\theta}\big[\nabla_\theta \mathcal{F}_\theta(\mathbf{x})\big] - \nabla_\theta \mathcal{F}_\theta(x_n). \tag{1}$$

The first term, the negative phase, is estimated with samples $x_s \sim p_\theta$, usually obtained by running a Markov chain (Hinton, 2002; Tieleman, 2008; Xie et al., 2016; Du & Mordatch, 2020). The sample is treated as a constant with respect to θ when the gradient is formed; we call this the *frozen-sampler* convention and use it throughout.

Score matching. The score-matching objective of Hyvärinen (2005) for a target density ρ is

$$J_{\mathrm{SM}}(\theta;\rho) = \mathbb{E}_\rho\big[\mathcal{J}_\theta(\mathbf{x})\big], \qquad \mathcal{J}_\theta(\mathbf{x}) := \tfrac{1}{2}\|\nabla_x \mathcal{F}_\theta(\mathbf{x})\|^2 - \Delta_x \mathcal{F}_\theta(\mathbf{x}), \tag{2}$$

which equals the Fisher divergence $\frac{1}{2}\mathbb{E}_\rho\|\nabla_x \log\rho + \nabla_x \mathcal{F}_\theta\|^2$ up to a θ-independent constant, under mild decay conditions. When ρ is a Gaussian smoothing of a data distribution, J_{SM} can be estimated without the Laplacian; this is denoising score matching (Vincent, 2011), the objective of score-based diffusion (Song & Ermon, 2019; Song et al., 2021b).

Langevin diffusion. The overdamped Langevin diffusion associated with $\mathcal{F}_\theta$ is

$$\mathrm{d}X_t = -\nabla_x \mathcal{F}_\theta(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad X_0 = x_n, \tag{3}$$

with $(B_t)_{t\geq 0}$ a K-dimensional Brownian motion. Its generator is $\mathcal{A}_\theta = \Delta_x - \nabla_x \mathcal{F}_\theta^\top \nabla_x$ (Pavliotis, 2014), and p_θ is its invariant law; under a dissipativity condition on $\nabla_x \mathcal{F}_\theta$ the convergence is exponential (Roberts & Tweedie, 1996; Mattingly et al., 2002). We write ρ_t for the law of X_t.

3 THE MAXIMUM LIKELIHOOD - PATH INTEGRAL OF SCORE MATCHING CORRESPONDANCE

Assumption 1. *(i) $\mathcal{F}_\theta$ is $\mathcal{C}^1$ in θ, and $\nabla_\theta \mathcal{F}_\theta$ is $\mathcal{C}^2$ in $\mathbf{x}$. (ii) $\nabla_x \mathcal{F}_\theta$ is locally Lipschitz and dissipative: $\mathbf{x}^\top \nabla_x \mathcal{F}_\theta(\mathbf{x}) \geq m\|\mathbf{x}\|^2 - b$ for some $m > 0$, $b \geq 0$. (iii) For every $T < \infty$, $\mathbb{E}\int_0^T \|\nabla_x \nabla_\theta \mathcal{F}_\theta(X_t)\|^2 \,\mathrm{d}t < \infty$.*

Condition (ii) guarantees a non-explosive, geometrically ergodic solution of (3) (Mattingly et al., 2002); condition (iii) makes the stochastic integral below a true martingale.

Theorem 1 (ML gradient as a path integral of score matching)**.** *Under Assumption 1, for any $T < \infty$ and any $x_n \in \mathbb{R}^K$,*

$$\mathbb{E}\big[\nabla_\theta \mathcal{F}_\theta(X_T)\big] - \nabla_\theta \mathcal{F}_\theta(x_n) \;=\; -\int_0^T \mathrm{d}t\; \nabla_\theta \mathbb{E}_{\rho_t}\big[\mathcal{J}_\theta(X_t)\big] \;=\; -\int_0^T \mathrm{d}t\; \nabla_\theta J_{\mathrm{SM}}(\theta;\rho_t), \tag{4}$$

where ∇_θ on the right acts on the explicit θ-dependence of $\mathcal{J}_\theta$ only, with ρ_t held fixed (the frozen-sampler convention). In particular, if $X_T \sim p_\theta$, the left-hand side is the log-likelihood gradient (1)*.*

Proof. Fix a coordinate θ_i and set $G(\mathbf{x}) := \frac{\partial}{\partial \theta_i}\mathcal{F}_\theta(\mathbf{x})$, a $\mathcal{C}^2$ function of $\mathbf{x}$. By Itô's formula along (3) (Øksendal, 2003),

$$\mathrm{d}G(X_t) = (\mathcal{A}_\theta G)(X_t)\,\mathrm{d}t + \sqrt{2}\,\nabla_x G(X_t)^\top \mathrm{d}B_t, \qquad \mathcal{A}_\theta G = \Delta_x G - \nabla_x \mathcal{F}_\theta^\top \nabla_x G. \tag{5}$$

By Assumption 1(iii) the stochastic integral has zero mean, so integrating over $[0,T]$ and taking expectations gives Dynkin's formula, $\mathbb{E}[G(X_T)] - G(x_n) = \int_0^T \mathbb{E}_{\rho_t}[\mathcal{A}_\theta G(X_t)]\,\mathrm{d}t$. It remains to identify the integrand. Since $\frac{\partial}{\partial \theta_i}$ commutes with ∇_x and Δ_x,

$$\mathcal{A}_\theta \frac{\partial}{\partial \theta_i}\mathcal{F}_\theta = \frac{\partial}{\partial \theta_i}\Delta_x \mathcal{F}_\theta - \nabla_x \mathcal{F}_\theta^\top \nabla_x \frac{\partial}{\partial \theta_i}\mathcal{F}_\theta = \frac{\partial}{\partial \theta_i}\Big[\Delta_x \mathcal{F}_\theta - \tfrac{1}{2}\|\nabla_x \mathcal{F}_\theta\|^2\Big] = -\frac{\partial}{\partial \theta_i}\mathcal{J}_\theta, \tag{6}$$

where $\frac{\partial}{\partial \theta_i}$ acts on $\mathcal{F}_\theta$ at a fixed point $\mathbf{x}$. Substituting and collecting the coordinates gives (4). □

The frozen sampler condition. The left-hand side of equation (4) is the standard ML estimator, in which the negative sample is not differentiated. The right-hand side inherits exactly that convention: the marginals ρ_t are those of the chain run at the current θ, and ∇_θ sees only the θ inside $\mathcal{J}_\theta$. No approximation is made; the identity is exact for every T. Differentiating the Langevin drift inside the integrand instead, i.e. treating $\mathrm{d}X_t$ as θ-dependent, produces a different integrand, $\nabla_\theta[\Delta_x \mathcal{F}_\theta - \|\nabla_x \mathcal{F}_\theta\|^2]$, whose expectation under p_θ does *not* vanish; the frozen convention is essential. Appendix B gives an elementary derivation of (4) by telescoping and Itô's multiplication rules that makes this explicit.

The Single-path estimator. Dropping the expectation in (5) and summing over an Euler discretisation with step Δt gives the unbiased estimator $\nabla_\theta \mathcal{F}_\theta(X_T) - \nabla_\theta \mathcal{F}_\theta(x_n) = -\sum_k \nabla_\theta \mathcal{J}_\theta(X_{t_k})\,\Delta t + \sqrt{2}\sum_k \nabla_x \nabla_\theta \mathcal{F}_\theta(X_{t_k})^\top \Delta B_k + o(1)$, whose second term is a mean-zero martingale. Section 5.1 checks the first term against the ML gradient as the step $\Delta t \to 0$.

Geometry. Equation (1) acts at the two endpoints of the path: it lowers $\mathcal{F}_\theta$ at x_n and raises it at x_s. Equation (4) distributes the same update along the path: at every time t it takes a step *down* the Fisher divergence between the model and the current marginal ρ_t (Figure 1). Ascending the likelihood is descending the time-integrated score-matching loss. Hyvärinen (2007) obtained the $T \to 0$ limit of this statement; (4) is the finite-T version.

4 Implications and a Notable Special Case

4.1 Stationarity and a burn-in diagnostic

Proposition 1 (Stationarity). *Under Assumption 1 and if* $\|\nabla_x g(\mathbf{x})\, p(\mathbf{x})\| = o\left(\frac{1}{\|\mathbf{x}\|^{K-1}}\right)$ *as* $\|\mathbf{x}\| \to \infty$, *with* $g = \mathcal{F}_\theta$ *or* $\nabla_\theta \mathcal{F}_\theta$ *then*

$$\mathbb{E}_{p_\theta}\big[\nabla_\theta \mathcal{J}_\theta\big] = 0, \qquad \mathbb{E}_{p_\theta}\big[\mathcal{J}_\theta\big] = -\tfrac{1}{2}\,\mathcal{I}(p_\theta), \quad \mathcal{I}(p_\theta) := \mathbb{E}_{p_\theta}\|\nabla_x \log p_\theta\|^2. \tag{7}$$

The proof (Appendix A) is an integration by parts. The first identity says that θ_0 is a stationary point of $\theta \mapsto J_{\mathrm{SM}}(\theta; p_{\theta_0})$; the second is Hyvärinen's value of the score-matching objective at the optimum.

Two consequences follow. First, in (4) the integrand vanishes once $\rho_t = p_\theta$: after the chain has mixed, the path carries no further information about the gradient. If the target is multimodal and the chain equilibrates within a single mode at some local mixing time $t^\star$, the integral over $[t^\star, T]$ vanishes for the *restriction* of p_θ to that mode, and the gradient is that of a local likelihood. Second, (7) is a stationarity test with a known target. Run C chains and monitor the windowed average

$$\hat{G}(t) \;:=\; \frac{1}{C}\sum_{c=1}^{C} \frac{1}{W}\int_t^{t+W} \nabla_\theta \mathcal{J}_\theta\big(X_s^{(c)}\big)\,\mathrm{d}s, \tag{8}$$

and declare burn-in complete once $\|\hat{G}(t)\|$ falls below its across-chain standard error. Unlike the Gelman–Rubin statistic (Gelman & Rubin, 1992), which estimates within-chain variance from a correlated segment, (8) compares against the exact value 0. Like every single-run diagnostic (Cowles & Carlin, 1996), it certifies equilibration only within the region the chains have visited; Section 5.2 quantifies this.

4.2 Denoising diffusion as a special case

Integration by parts turns (4) into a form that is useful whenever ρ_t is known: under the decay conditions of Proposition 1, $\mathbb{E}_{\rho_t}[\Delta_x \mathcal{F}_\theta] = -\mathbb{E}_{\rho_t}[\nabla_x \mathcal{F}_\theta^\top \nabla_x \log \rho_t]$, hence

$$\mathbb{E}[\nabla_\theta \mathcal{F}_\theta(X_T)] - \nabla_\theta \mathcal{F}_\theta(x_n) = -\frac{1}{2}\int_0^T \mathrm{d}t\; \nabla_\theta \mathbb{E}_{\rho_t}\Big[\big\|\nabla_x \mathcal{F}_\theta(X_t) + \nabla_x \log \rho_t(X_t)\big\|^2\Big], \tag{9}$$

the time integral of the gradient of the Fisher divergence between ρ_t and p_θ.

Now let $\mathcal{F}_\theta(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 + \mathcal{I}_\theta(\mathbf{x})$, a quadratic confinement plus an interaction with score $s_\theta := \nabla_x \mathcal{I}_\theta$, and replace s_θ by its affine expansion around the datum, $s_\theta(\mathbf{x}) \approx s_\theta(x_n) + \nabla_x^2 \mathcal{I}_\theta(x_n)(\mathbf{x} - x_n)$. The diffusion (3) becomes an Ornstein–Uhlenbeck process with Gaussian marginals $\rho_t = \mathcal{N}(\mu_t, \Sigma_t)$ (closed-forms are detailed in Appendix C). In the simplest case where the interaction is flat at the datum, $\nabla_x^2 \mathcal{I}_\theta(x_n) \approx 0$,

$$\mu_t = e^{-t}x_n - (1 - e^{-t})\, s_\theta(x_n), \qquad \Sigma_t = \sigma_t^2\, \mathbf{I}_K, \quad \sigma_t^2 = 1 - e^{-2t}, \tag{10}$$

so $X_t = \mu_t + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I}_K)$ and $\nabla_x \log \rho_t(X_t) = -\epsilon/\sigma_t$. Substituting in (9),

$$\mathbb{E}[\nabla_\theta \mathcal{F}_\theta(X_T)] - \nabla_\theta \mathcal{F}_\theta(x_n) = -\int_0^T \mathrm{d}t\; \lambda_t\, \nabla_\theta \mathbb{E}_\epsilon\Big[\big\|\epsilon_\theta(X_t, t) - \epsilon\big\|^2\Big], \tag{11}$$

with $\epsilon_\theta(\mathbf{x}, t) := \sigma_t\, \nabla_x \mathcal{F}_\theta(\mathbf{x})$ and $\lambda_t = 1/(2\sigma_t^2)$. This is the ϵ-prediction loss of Ho et al. (2020). With $\bar{\alpha}_t := e^{-2t}$ the forward kernel $X_t = \sqrt{\bar{\alpha}_t}\, x_n + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$ is the variance-preserving kernel of Song et al. (2021b), plus the correction $-(1 - e^{-t})s_\theta(x_n)$, and the weight $\lambda_t \propto 1/\sigma_t^2$ is the likelihood weighting of Song et al. (2021a) . Two features distinguish this from a standard diffusion model: the network parametrises a scalar $\mathcal{F}_\theta$ whose gradient is the score, so ϵ_θ is a conservative field by construction (Song & Kingma, 2021).

Algorithm 1 implements (11) with one time sample per datum. It needs no Markov chain and one Hessian-free backward pass through $\mathcal{F}_\theta$, against the T sequential Langevin steps of persistent contrastive divergence (PCD) or the Laplacian of exact score matching.

Algorithm 1 Path-integral Score Matching (PISM) for EBMs

Require: data q_{data}, horizon T, network $\mathcal{F}_\theta$, weight $\lambda_t = 1/(2\sigma_t^2)$ (or uniform)
repeat
Sample $x_0 \sim q_{\text{data}}$, $t \sim \text{Uniform}(0, T]$, $\epsilon \sim \mathcal{N}(0, \mathbf{I}_K)$; set $\sigma_t = \sqrt{1 - e^{-2t}}$.
$x_t \leftarrow e^{-t} x_0 + \sigma_t \epsilon$ (optionally $-(1 - e^{-t})\,\text{stopgrad}[\nabla_x \mathcal{F}_\theta(x_0)]$)
$g \leftarrow \nabla_x \mathcal{F}_\theta(x_t)$ by automatic differentiation.
Take a gradient step on $\lambda_t \left\| \sigma_t\, g - \epsilon \right\|^2$.
until convergence

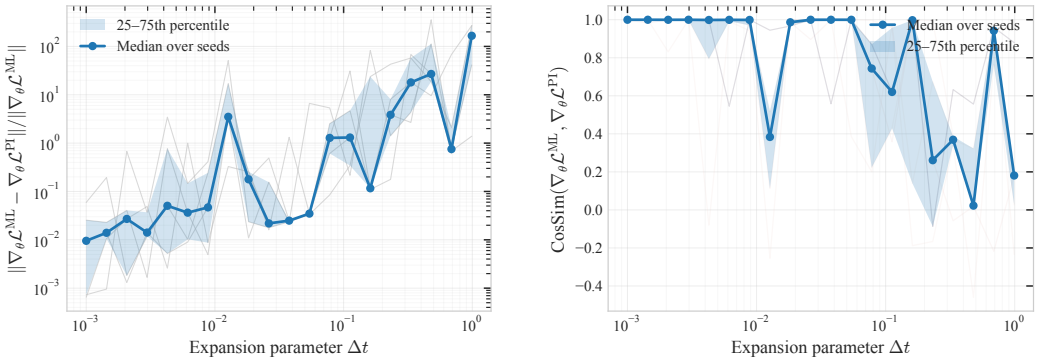

Figure 2: Numerical check of Theorem 1 on a GRBM on MNIST, as a function of the Euler step Δt (median and inter-quartile range over five seeds). **Left:** relative error between the ML gradient and the discretized path integral. **Right:** cosine similarity between the two. The gradients agree to within 10^{-2} for $\Delta t \lesssim 10^{-2}$; the discretization error grows with the step.

5 Experiments

Two models with closed forms are used throughout. The *diagonal Gaussian* $\mathcal{F}_\theta(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \text{diag}(a)\mathbf{x}$ has an exact Ornstein–Uhlenbeck kernel. The *Gaussian restricted Boltzmann machine* (GRBM) with one binary hidden unit $h \in \{-1, +1\}$ (Welling et al., 2005),

$$\mathcal{F}_\theta(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|^2 - \log\cosh(b + \mathbf{x}^\top w), \qquad \theta = \{b, w\}, \tag{12}$$

has the exactly known law $\sigma(2b)\,\mathcal{N}(w, \mathbf{I}) + \sigma(-2b)\,\mathcal{N}(-w, \mathbf{I})$, where σ represents the logistic function, so exact samples are available and the barrier between the modes is controlled by $\|w\|$. Its score and Hessian are given in Appendix D.

5.1 The gradient identity

Gradient agreement. For each MNIST digit x_n we run one Langevin path (3) from x_n under the GRBM (12), compute the ML gradient (1) with the endpoint as the negative sample, and the discretised single-path integral of Theorem 1, for 20 logarithmically spaced steps $\Delta t \in [10^{-3}, 0.99]$ and five seeds. Figure 2 reports relative error and cosine similarity: the two gradients agree up to discretisation error, which decays with Δt as predicted.

Optimisation trajectories. We draw $N = 1000$ samples from a two-dimensional GRBM with $b = 0$, $w^\star = (-2, -1)$ by Langevin dynamics ($T = 1000$, $\Delta t = 10^{-3}$), and train two identically initialised GRBMs by SGD ($\eta = 10^{-3}$), one with the ML gradient and one with the path-integral gradient. The free energy is symmetric under $w \mapsto -w$, so $\pm w^\star$ are both optima. Figure 3 shows the two trajectories on the loss landscape: they follow the same path and reach the same optimum.

5.2 The burn-in diagnostic

Setup. For the diagonal Gaussian with $K = 8$ and rates a_i geometrically spaced in $[\kappa^{-1}, 1]$, the slowest relaxation rate is $2a_{\min} = 2/\kappa$ and, since $\frac{\partial \mathcal{J}_\theta}{\partial a_i} = a_i x_i^2 - 1$, the diagnostic has the exact

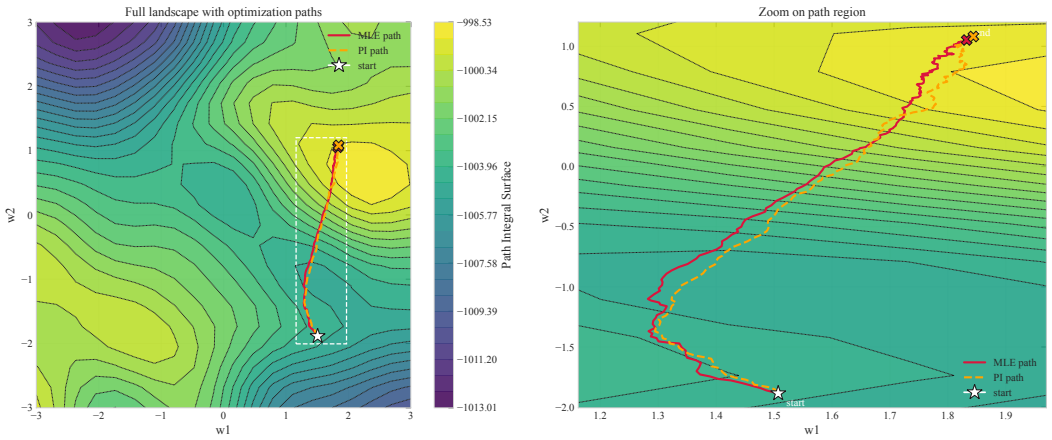

Figure 3: Parameter trajectories of a 2-D GRBM trained by SGD with the ML gradient (solid) and the path-integral gradient (dashed), from the same initialisation (star). **Left:** full landscape. **Right:** zoom on the path region.

Table 1: Stationarity check from 4×10^5 exact samples. "fd. check" is the finite-difference error on $\nabla_\theta \mathcal{J}_\theta$; $z = \max_i |\mathbb{E}[\frac{\partial \mathcal{J}_\theta}{\partial \theta_i}]|/\mathrm{SE}$, a value of order one being consistent with (7); the last two columns compare the scalar integrand to its predicted value.

model	K	fd. check	z	$\mathbb{E}_{p_\theta}[\mathcal{J}_\theta]$	$-\frac{1}{2}\mathcal{I}(p_\theta)$
Gaussian	8	5.8×10^{-8}	1.57	-4.9641 ± 0.0048	-4.9660
GRBM	6	1.8×10^{-7}	1.96	-2.9409 ± 0.0027	-2.9367

expectation

$$\mathbb{E}_{\rho_t}\left[\frac{\partial \mathcal{J}_\theta}{\partial a_i}\right] = e^{-2a_i t}\big(a_i(x_{0,i}^2 + v_0) - 1\big) \tag{13}$$

for chains started at $\mathcal{N}(x_0, v_0\mathbf{I})$ (Appendix E). At large t the norm is carried by the slowest mode, so $\|\hat{G}\|$ and $D_{\mathrm{KL}}(\rho_t\|p_\theta)$ should both decay at rate $2a_{\min}$. That is the claim we test, with 512 chains started three model standard deviations from equilibrium and exact Ornstein–Uhlenbeck steps. For the GRBM we use 4×10^5 exact samples for the stationarity check and 4096 chains started inside the dominant well for the metastability sweep.

The identity holds. With exact i.i.d. samples (Table 1), the largest coordinate of $\mathbb{E}_{p_\theta}[\nabla_\theta \mathcal{J}_\theta]$ is 1.6 and 2.0 standard errors from zero for the two models, and $\mathbb{E}_{p_\theta}[\mathcal{J}_\theta]$ matches $-\frac{1}{2}\mathcal{I}(p_\theta)$ to three digits, as Proposition 1 predicts.

The diagnostic recovers the relaxation rate. Table 2 and Figure 4(a,d) compare the fitted decay rate of $\|\hat{G}\|$ with $2a_{\min}$. Rates are fitted only where the signal exceeds three across-chain standard errors, so the noise floor cannot bias the fit. Agreement is within 1.2% at $\kappa = 100$ and 5.0% at $\kappa = 30$, degrading to 14% at $\kappa = 3$. The residual at small κ is a property of the fitting window, not of the estimator: fitting the *exact* curve (13) the same way gives at most 4.9%, and fitting the exact D_{KL} gives 13.2% in the same direction. Burn-in times read off the diagnostic exceed those read off D_{KL} by a factor 1.2–2.3, which is conservative in the right direction for a stopping rule.

Baselines. Figure 4(b,c) places two common heuristics on the same run. The energy trace $\mathbb{E}[\mathcal{F}_\theta(X_t)]$ flattens at roughly a quarter of its initial value while the chain is still far from equilibrium. The Gelman–Rubin statistic $\hat{R}$ does decrease, but more slowly than $\|\hat{G}\|$, because it must estimate within-chain variance from a correlated segment whereas (8) exploits an exactly known equilibrium value.

Table 2: Decay-rate recovery on the diagonal Gaussian, 512 chains, exact OU steps. The three fitted columns should equal $2a_{\min}$. The last two columns give the time at which each quantity has fallen to 1% of its initial value.

		fitted decay rate			t at 1%	
κ	$2a_{\min}$	$\|\hat{G}\|$	$\|\hat{G}\|$ exact	D_{KL}	$\|\hat{G}\|$	D_{KL}
3	0.6667	0.7599	0.6993	0.7543	5.55	4.76
10	0.2000	0.2292	0.2042	0.2171	24.61	13.62
30	0.0667	0.0633	0.0673	0.0704	72.44	37.74
100	0.0200	0.0202	0.0201	0.0207	272.36	118.49

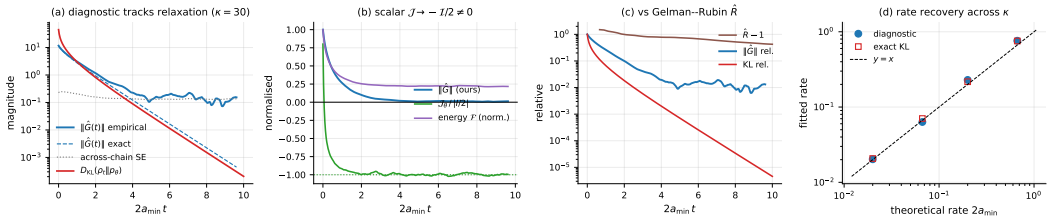

Figure 4: Diagonal Gaussian, 512 chains, exact Ornstein–Uhlenbeck steps; time in units of the slowest relaxation time. **(a)** $\|\hat{G}(t)\|$ follows the exact prediction (13) and decays at the rate of $D_{\mathrm{KL}}(\rho_t\|p_\theta)$ until it meets the across-chain noise floor. **(b)** The scalar integrand converges to $-\frac{1}{2}\mathcal{I}(p_\theta)$, and the energy trace flattens well before equilibrium. **(c)** Gelman–Rubin $\hat{R}$, computed by discarding the first half of the run, certifies convergence later. **(d)** Fitted rates against $2a_{\min}$ across condition numbers.

Table 3: Metastability sweep, GRBM with $K = 6$, $b = 0.6$, 4096 chains started in the dominant well, $T = 20$, $\Delta t = 5 \times 10^{-3}$. "Occ." columns are the exact mode occupancy under the true law and the occupancy the chains reach. "Signal" is the exact bias under the single-well law, which the diagnostic must resolve against its noise floor.

$\|w\|$	occ. exact	occ. reached	mixed?	signal	$\|\hat{G}\|$ / SE	verdict
0.5	0.602	0.597	yes	0.0835 ± 0.0035	0.94	true positive
1.0	0.684	0.680	yes	0.1534 ± 0.0044	0.94	true positive
1.5	0.734	0.740	yes	0.1456 ± 0.0045	0.93	true positive
2.0	0.757	0.765	yes	0.0848 ± 0.0044	0.89	true positive
3.0	0.767	0.930	no	0.0116 ± 0.0040	1.03	**false positive**
4.0	0.769	0.996	no	0.0026 ± 0.0039	0.95	**false positive**

Failure mode: local, not global, equilibration. We sweep the mode separation $\|w\|$ of the GRBM at fixed $b = 0.6$ and compare the mode occupancy reached by the chains with the exact occupancy under the true law (Table 3, Figure 5). For $\|w\| \leq 2$ the chains reach the correct occupancy to within 0.01 and the diagnostic correctly reports equilibration. For $\|w\| = 3$ and 4 the occupancy is wrong by 0.16 and 0.23—the chains never leave the dominant well—and the diagnostic still reports equilibration: two false positives.

The reason is structural, not statistical. As $\|w\|$ grows, $\tanh$ saturates inside each well, $\mathcal{F}_\theta(\mathbf{x}) \approx \frac{1}{2}\|\mathbf{x}\|^2 - (b + \mathbf{x}^\top w)$ there, and the single-well restriction becomes a near-exact stationary law of the local dynamics. The bias the diagnostic would have to detect, $\|\mathbb{E}_{\text{one well}}[\nabla_\theta \mathcal{J}_\theta]\|$, collapses from 0.153 at $\|w\| = 1$ to 0.0026 ± 0.0039 at $\|w\| = 4$, indistinguishable from zero at 4×10^5 samples. The estimand vanishes, so more chains do not help. This is the local-likelihood phenomenon of Section 4.1: the diagnostic certifies equilibration within the basin the chains occupy and says nothing about basins they have not found. It should be used alongside, not instead of, a multi-start check.

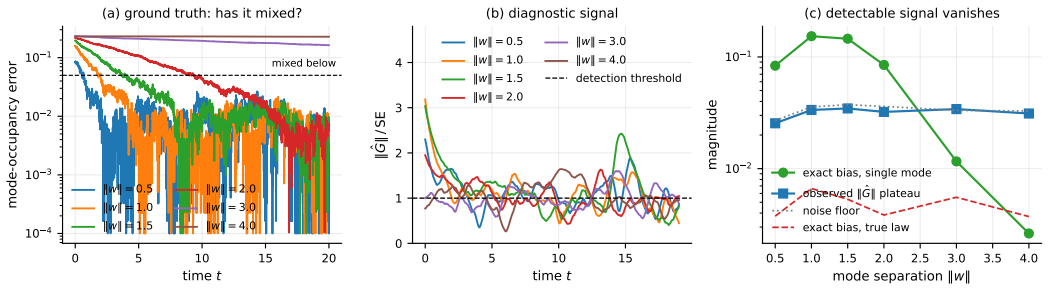

Figure 5: Metastability sweep. **(a)** Ground truth: distance of the chains' mode occupancy from its exact value; curves below the dashed line have mixed, $\|w\| = 3, 4$ have not. **(b)** The diagnostic signal, in units of its own standard error, converges to 1 for $\|w\| \leq 2$. **(c)** Why: the exact bias under the single-well law rises and then collapses below the noise floor as the barrier grows, while the true law gives zero throughout

Table 4: OOD detection AUROC (↑) using the raw energy, CIFAR-10 in-distribution, and wall-clock time per epoch on the same hardware. Ours: mean ± std over training seeds; the first three rows are as reported by Du & Mordatch (2020) and Grathwohl et al. (2020).

Method	SVHN	CIFAR-100	DTD	CelebA	Interp.	t/epoch (s)
Glow (likelihood)	0.05	0.51	–	0.57	0.51	–
IGEBM	0.63	0.50	0.48	0.70	0.70	–
JEM	0.67	0.67	–	0.75	0.65	–
SM	0.06±0.00	**0.52±0.00**	0.52±0.01	0.50±0.01	0.30±0.00	14.86±0.13
PCD	0.55±0.01	0.49±0.01	**0.60±0.02**	0.31±0.00	0.62±0.02	65.97±0.06
PISM	**0.97±0.00**	0.47 ±0.00	0.55±0.00	**0.51±0.00**	**0.78±0.00**	**5.79±0.26**

5.3 Training a Convolutional EBM without a sampler

We train a Convolutional EBM following the method of Du & Mordatch (2020) on CIFAR-10 with three objectives and otherwise identical architecture, optimizer and schedule: exact score matching (SM), PCD with a Langevin sampler capped at 60 steps in the spirit of short-run MCMC (Nijkamp et al., 2019), and PISM (Algorithm 1). Following Du & Mordatch (2020) and Grathwohl et al. (2020) we evaluate by out-of-distribution (OOD) detection with the raw energy as score, reporting AUROC against SVHN, CIFAR-100, Textures (DTD), CelebA and CIFAR-10 interpolations (Table 4). Likelihood-based models are known to fail on SVHN (Nalisnick et al., 2019).

PISM trains in 5.8 s per epoch against 66 s for PCD and 15 s for SM, because the sampling burden is replaced by a single noise draw and the Laplacian by a first-order gradient. On SVHN, CelebA, and Interpolations it is the strongest method in the table; on CIFAR-100, and DTD it is at chance, as is PCD under the same capped sampler. We do not claim state-of-the-art OOD detection. The point of the experiment is that an objective derived from the ML gradient trains a deep EBM at a fraction of the cost of the sampler-based estimator, with comparable or better energies on the benchmarks where the latter is not itself at chance.

6 Conclusion

We have shown that, for continuous energy-based models, the maximum-likelihood gradient can be expressed as the gradient of a path integral of the score-matching objective along Langevin dynamics. At equilibrium, this gradient vanishes in expectation, yielding a natural criterion for diagnosing the burn-in of Langevin Markov chains with general energy functions. Under quadratic confinement, corresponding to a Gaussian term, the DDPM ϵ-loss is recovered as a special case of this path-integral formulation.

AI USE STATEMENT

We used LLMs to help improve the planning and design of the experiments and to polish the final draft of the paper. LLMs were not used for the original proposal or derivation of the path-integral score matching method. All AI-assisted text and experimental suggestions were reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

All models in Sections 5.1 and 5.2 have closed-form free energies, scores, Hessians and, for the Burn-in experiments; all parameters are stated in the text and table captions, and the remaining details in Appendix F. The CIFAR-10 experiment protocol follows Du & Mordatch (2020) with the modifications listed in Section 5.3. Code to reproduce all figures and tables will be released as supplementary material.

REFERENCES

Miguel Á. Carreira-Perpiñán and Geoffrey E. Hinton. On contrastive divergence learning. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pp. 33–40. PMLR, 2005.

Mary Kathryn Cowles and Bradley P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models, 2020. URL https://arxiv.org/abs/1903.08689.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

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

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

Aapo Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531, 2007. doi: 10.1109/TNN.2007.895819.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Frederic Koehler, Andrew Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *International Conference on Learning Representations*, 2023.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

Jonathan C. Mattingly, Andrew M. Stuart, and Desmond J. Higham. Ergodicity for SDEs and approximations: Locally lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002. doi: 10.1016/S0304-4149(02)00150-3.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run MCMC toward energy-based model. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6 edition, 2003.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*. Springer, 2014.

Y. Qin and A. Risteski. Fit like you sample: Sample-efficient generalized score matching from fast mixing diffusions. In *Proceedings of the 37th Annual Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Saeed Saremi and Aapo Hyvärinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019. URL https://jmlr.org/papers/v20/19-216.html.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song and Diederik P. Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation, 2019. URL https://arxiv.org/abs/1905.07088.

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

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

Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1064–1071, 2008.

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

Max Welling, Michal Rosen-Zvi, and Geoffrey E. Hinton. Exponential family harmoniums with an application to information retrieval. *Advances in Neural Information Processing Systems*, 17, 2005.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. A theory of generative ConvNet. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2635–2644, 2016.

A Proof of Proposition 1

Let $p = p_\theta \propto e^{-\mathcal{F}_\theta}$ and let $g : \mathbb{R}^K \to \mathbb{R}$ be $\mathcal{C}^2$ with $\lim_{R\to\infty} \sup_{\|\mathbf{x}\|=R} \|\nabla_x g\, p\| \cdot R^{K-1} = 0$. Using integration by parts over a ball $\mathcal{B}_R = \{\mathbf{x};\, \|\mathbf{x}\| < R\}$,

$$\int_{\mathcal{B}_R} \mathrm{d}\mathbf{x} \nabla_x^T [\nabla_x g\, p] = \int_{\mathcal{B}_R} \mathrm{d}\mathbf{x} \Delta_x g\, p + \int_{\mathcal{B}_R} \mathrm{d}\mathbf{x} \nabla_x g^T\, \nabla_x \log p\, p \tag{14}$$

Using the divergence theorem, the left-hand integral is rewritten over the boundary sphere $\|\mathbf{x}\| = R$, and by the Cauchy-Schwarz inequality:

$$\left| \int_{\mathcal{B}_R} \mathrm{d}\mathbf{x} \nabla_x^T [\nabla_x g\, p] \right| = \left| \oint_{\|\mathbf{x}\|=R} \mathrm{d}\mathbf{s}\, n^\top \nabla_x g\, p \right| \tag{15}$$

$$\leq \oint_{\|\mathbf{x}\|=R} \mathrm{d}\mathbf{s}\, \|n\| \cdot \|\nabla_x g\, p\|, \quad \|n\| = 1 \tag{16}$$

$$\leq \left(\sup_{\|\mathbf{x}\|=R} \|\nabla_x g\, p\| \right) \oint_{\|\mathbf{x}\|=R} \mathrm{d}\mathbf{s} \tag{17}$$

$$= \left(\sup_{\|\mathbf{x}\|=R} \|\nabla_x g\, p\| \right) \cdot C_K R^{K-1} \tag{18}$$

where $C_K = \frac{2\pi^{K/2}}{\Gamma(K/2)}$ is the surface area of a unit K-dimensional sphere. Taking the limit as $R \to \infty$, the vanishing condition at infinity directly applies to the upper bound:

$$0 \leq \lim_{R\to\infty} \left| \int_{\mathcal{B}_R} \mathrm{d}\mathbf{x} \nabla_x^T [\nabla_x g\, p] \right| \leq C_K \cdot \lim_{R\to\infty} \left(\sup_{\|\mathbf{x}\|=R} \|\nabla_x g\, p\| \cdot R^{K-1} \right) = 0 \tag{19}$$

the limit of the contour integral is zero. It follows that,

$$\mathbb{E}_p[\Delta_x g] = -\int \nabla_x g^\top \nabla_x p\, \mathrm{d}\mathbf{x} = \mathbb{E}_p\big[\nabla_x \mathcal{F}_\theta^\top \nabla_x g\big], \tag{20}$$

since $\nabla_x p = -p\, \nabla_x \mathcal{F}_\theta$. Apply (20) with $g = \frac{\partial}{\partial \theta_i} \mathcal{F}_\theta$:

$$\mathbb{E}_p\big[\frac{\partial}{\partial \theta_i} \mathcal{J}_\theta\big] = \mathbb{E}_p\big[\nabla_x \mathcal{F}_\theta^\top \nabla_x \frac{\partial}{\partial \theta_i} \mathcal{F}_\theta\big] - \mathbb{E}_p\big[\Delta_x \frac{\partial}{\partial \theta_i} \mathcal{F}_\theta\big] = 0. \tag{21}$$

Apply (20) with $g = \mathcal{F}_\theta$: $\mathbb{E}_p[\Delta_x \mathcal{F}_\theta] = \mathbb{E}_p \|\nabla_x \mathcal{F}_\theta\|^2 = \mathcal{I}(p)$, hence $\mathbb{E}_p[\mathcal{J}_\theta] = \frac{1}{2}\mathcal{I}(p) - \mathcal{I}(p) = -\frac{1}{2}\mathcal{I}(p)$. □

The same computation gives $\mathbb{E}_p[\mathcal{A}_\theta \mathcal{F}_\theta] = \mathbb{E}_p[\Delta_x \mathcal{F}_\theta - \|\nabla_x \mathcal{F}_\theta\|^2] = 0$, the statement that p is invariant for the generator.

B An elementary derivation of Theorem 1

Dynkin's formula packages three elementary steps: a telescoping sum, a second-order Taylor expansion, and Itô's multiplication rules. We spell them out here; this is the derivation we first used, and it makes explicit where the frozen-sampler convention enters.

Telescoping. Discretise (3) by Euler–Maruyama with step dt and $N = T/dt$ steps, $t_k = k\, dt$,

$$X_{t_{k+1}} = X_{t_k} + dX_{t_k}, \qquad dX_{t_k} = -\nabla_x \mathcal{F}_\theta(X_{t_k})\, dt + \sqrt{2}\, dB_{t_k}, \qquad dB_{t_k} \sim \mathcal{N}(0, dt\, \mathbf{I}_K). \tag{22}$$

By linearity of ∇_θ the ML estimator is a telescoping sum over the path,

$$\nabla_\theta \mathcal{F}_\theta(X_T) - \nabla_\theta \mathcal{F}_\theta(x_n) = \sum_{k=0}^{N-1} \Big[\nabla_\theta \mathcal{F}_\theta(X_{t_{k+1}}) - \nabla_\theta \mathcal{F}_\theta(X_{t_k}) \Big], \tag{23}$$

in which, by the frozen-sampler convention, the points X_{t_k} are constants with respect to θ.

Second-order expansion. Fix one coordinate and write $G := \frac{\partial}{\partial \theta_i}\mathcal{F}_\theta$. Expanding one increment to second order in dX_{t_k},

$$G(X_{t_{k+1}}) - G(X_{t_k}) = \underbrace{\nabla_x G(X_{t_k})^\top dX_{t_k}}_{I_1} + \underbrace{\tfrac{1}{2}\, dX_{t_k}^\top \nabla_x^2 G(X_{t_k})\, dX_{t_k}}_{I_2} + o(dt). \tag{24}$$

Substituting dX_{t_k} into I_1 gives

$$I_1 = -\nabla_x G^\top \nabla_x \mathcal{F}_\theta\, dt + \sqrt{2}\, \nabla_x G^\top dB_{t_k} = -\frac{\partial}{\partial \theta_i}\Big[\tfrac{1}{2}\|\nabla_x \mathcal{F}_\theta\|^2\Big] dt + \sqrt{2}\, \nabla_x G^\top dB_{t_k}, \tag{25}$$

where the last equality uses $\nabla_x G = \nabla_x \frac{\partial}{\partial \theta_i}\mathcal{F}_\theta = \frac{\partial}{\partial \theta_i}\nabla_x \mathcal{F}_\theta$ and the chain rule at a fixed point $\mathbf{x}$.

Itô's multiplication rules. Expanding I_2 and keeping terms up to order dt with $dt^2 = dt\, dB_t = 0$ and $dB_t\, dB_t^\top = dt\, \mathbf{I}_K$,

$$\begin{aligned} I_2 &= \tfrac{1}{2}\operatorname{Tr}\Big\{\nabla_x^2 G\,\big(\nabla_x \mathcal{F}_\theta \nabla_x \mathcal{F}_\theta^\top dt^2 + 2\, dB_{t_k} dB_{t_k}^\top\big)\Big\} - \sqrt{2}\, \nabla_x \mathcal{F}_\theta^\top \nabla_x^2 G\, dB_{t_k}\, dt \\ &= \Delta_x G\, dt = \frac{\partial}{\partial \theta_i}\big[\Delta_x \mathcal{F}_\theta\big] dt. \end{aligned} \tag{26}$$

Adding the two terms,

$$G(X_{t_{k+1}}) - G(X_{t_k}) = -\frac{\partial}{\partial \theta_i}\mathcal{J}_\theta(X_{t_k})\, dt + \sqrt{2}\, \nabla_x \frac{\partial}{\partial \theta_i}\mathcal{F}_\theta(X_{t_k})^\top dB_{t_k} + o(dt), \tag{27}$$

with $\mathcal{J}_\theta = \frac{1}{2}\|\nabla_x \mathcal{F}_\theta\|^2 - \Delta_x \mathcal{F}_\theta$. Summing (27) over k in (23) gives the single-path estimator of Remark 2. The noise term is a sum of increments, each independent of the past and multiplied by a function of the past, so it is a martingale with zero mean; taking expectations and letting $dt \to 0$ gives (4). Averaging S independent paths $X^{(s)}$ from the same x_n and letting $S \to \infty$ replaces the empirical average over paths by $\mathbb{E}_{\rho_t}$.

C Ornstein–Uhlenbeck marginals for the affine score

With $\mathcal{F}_\theta(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 + \mathcal{I}_\theta(\mathbf{x})$, $s_\theta = \nabla_x \mathcal{I}_\theta$ and the affine expansion $s_\theta(\mathbf{x}) \approx s_\theta(x_n) + H_n(\mathbf{x} - x_n)$, $H_n := \nabla_x^2 \mathcal{I}_\theta(x_n)$, the diffusion (3) reads

$$\mathrm{d}X_t = -\big[A_n X_t + b_n\big]\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad A_n = \mathbf{I}_K + H_n, \quad b_n = s_\theta(x_n) - H_n x_n. \tag{28}$$

For A_n symmetric positive definite its solution is

$$X_t = e^{-A_n t} x_n - A_n^{-1}\big(\mathbf{I}_K - e^{-A_n t}\big) b_n + \sqrt{2}\int_0^t e^{-A_n (t-s)}\, \mathrm{d}B_s, \tag{29}$$

so $\rho_t = \mathcal{N}(\mu_t, \Sigma_t)$ with

$$\mu_t = e^{-A_n t} x_n - A_n^{-1}\big(\mathbf{I}_K - e^{-A_n t}\big) b_n, \qquad \Sigma_t = A_n^{-1}\big(\mathbf{I}_K - e^{-2A_n t}\big). \tag{30}$$

Setting $H_n = 0$ gives (10). Substituting $\nabla_x \log \rho_t(X_t) = -\Sigma_t^{-1}(X_t - \mu_t)$ in (9) gives a Gaussian denoising loss with anisotropic noise; the isotropic case is (11). Under the frozen-sampler convention μ_t and Σ_t are constants with respect to θ.

D Closed forms for the Gaussian RBM

For (12), with $u := b + \mathbf{x}^\top w$,

$$\nabla_x \mathcal{F}_\theta(\mathbf{x}) = \mathbf{x} - w \tanh u, \tag{31}$$

$$\nabla_x^2 \mathcal{F}_\theta(\mathbf{x}) = \mathbf{I}_K - (1 - \tanh^2 u)\, w w^\top, \tag{32}$$

$$\mathcal{J}_\theta(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x} - w \tanh u\|^2 - K + (1 - \tanh^2 u)\|w\|^2. \tag{33}$$

Since $e^{-\mathcal{F}_\theta(\mathbf{x})} = \frac{1}{2} e^{\|w\|^2/2}\big[e^b e^{-\|\mathbf{x}-w\|^2/2} + e^{-b} e^{-\|\mathbf{x}+w\|^2/2}\big]$, the model law is the mixture $\sigma(2b)\,\mathcal{N}(w, \mathbf{I}_K) + \sigma(-2b)\,\mathcal{N}(-w, \mathbf{I}_K)$, from which exact samples and exact mode occupancies are computed.

E EXACT DIAGNOSTIC FOR THE DIAGONAL GAUSSIAN

For $\mathcal{F}_\theta(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathrm{diag}(a)\mathbf{x}$, $\mathcal{J}_\theta(\mathbf{x}) = \frac{1}{2}\sum_i a_i^2 x_i^2 - \sum_i a_i$ and $\frac{\partial \mathcal{J}_\theta}{\partial a_i} = a_i x_i^2 - 1$. Each coordinate of (3) is an independent Ornstein–Uhlenbeck process; started at $X_0 \sim \mathcal{N}(x_0, v_0\mathbf{I})$ its law at time t is Gaussian with mean $e^{-a_i t}x_{0,i}$ and variance $v_0 e^{-2a_i t} + (1 - e^{-2a_i t})/a_i$. Hence $\mathbb{E}_{\rho_t}[x_i^2] = e^{-2a_i t}(x_{0,i}^2 + v_0) + (1 - e^{-2a_i t})/a_i$ and (13) follows. The Kullback–Leibler divergence between Gaussians with these parameters is available in closed form and is the reference curve in Figure 4.

F EXPERIMENTAL DETAILS

Gradient identity (Section 5.1). GRBM (12) on MNIST ($K = 784$, we normalize the mnist samples using the mean and standard deviation of the dataset at each pixel). For each dt the path is simulated by Euler–Maruyama over horizon $T = 1000$; the ML gradient uses the endpoint as negative sample and the path-integral gradient sums $-\nabla_\theta \mathcal{J}_\theta(X_{t_k})dt$ along the same path. Relative error is $\|\nabla_\theta \mathcal{L}^{\mathrm{ML}} - \nabla_\theta \mathcal{L}^{\mathrm{PI}}\| / \|\nabla_\theta \mathcal{L}^{\mathrm{ML}}\|$.

Diagnostic (Section 5.2). Decay rates are fitted by least squares on $\log|\hat{G}(t)|$ over the window where $|\hat{G}| > 3\,\mathrm{SE}$. Window $W = 0.4$ in units of $\frac{1}{2a_{\min}}$. Gelman–Rubin $\hat{R}$ is computed componentwise on the state vector X, after discarding the first half of each run.

CIFAR-10 (Section 5.3). Architecture, augmentation, and optimiser follow Du & Mordatch (2020). For PCD, we use a replay buffer of size $10{,}000$ and perform 60 Langevin steps per update, with step size $\eta = 10$, noise scale 5×10^{-3}, and per-step gradient clipping at 0.01; 5% of chains are reinitialized from noise. Training uses the SGD optimizer with a batch size of 128, learning rate 10^{-4}, and 60 epochs. PISM: horizon $T = 1000$, likelihood weighting λ_t is used, a discrete noise schedule of $\beta_t \in [10^{-4}, 0.02]$, with $\alpha_t = 1 - \beta_t$ as in most implmentations of denoising diffusion models. The number of seeds is set to 3.