Score-Based Generative Models: An Intuitive Path from Langevin Dynamics to Diffusion

Nathan Cantafio

April 9, 2026

This note traces the development of score-based generative models through three landmark papers. We begin with the foundational insight of Song & Ermon (2019): that learning the score function $\nabla_x \log p(x)$ sidesteps the intractable normalizing constant, and that annealing across multiple noise levels enables effective Langevin-based sampling. We then examine Ho et al. (2020), who derived a closely related procedure through the lens of variational inference over a Gaussian Markov chain, introducing the noise reparametrization that reframes learning as noise prediction. Finally, we present the unifying perspective of Song et al. (2023), which shows that both frameworks are discretizations of a continuous-time stochastic differential equation (SDE), and that the reverse of any such SDE depends only on the time-dependent score. Together, these works establish a framework for generative modelling that has since become the dominant paradigm for high-quality image, audio, and video synthesis.

1. Introduction

Generative modelling asks the question: given samples from some unknown distribution $p(x)$, can we learn a model that generates new samples from that same distribution? For data in high-dimensional spaces such as images, audio, or text, this is a highly challenging problem. The dimensionality d can reach into the millions, the distribution $p(x)$ can be multimodal and is likely to be concentrated on a low-dimensional manifold, Moreover, we observe only samples from $p(x)$, never the density itself. In the last decade or so *score-based generative models*, also known as *diffusion models*, have emerged as the dominant paradigm. We follow the development through three papers. Song & Ermon (2019) introduced the idea of learning the score function via denoising score matching and sampling with annealed Langevin dynamics [1]. Ho et al. (2020) derived a similar procedure from a Markov chain perspective via the ELBO, introducing the noise reparametrization that became standard [2]. Song et al. (2023) showed that both frameworks are discretizations of a single continuous-time SDE, yielding a unifying theoretical picture and new algorithmic possibilities [3].

2. Background: Score Functions and Langevin Dynamics

2.1. Score functions

For a differentiable probability density $p(x)$ over $\mathbb{R}^d$, the *score function* is defined as

$$s(x) := \nabla_x \log p(x).$$

The score is a vector field on $\mathbb{R}^d$ that points in the direction of increasing log-density. It does not require the normalizing constant of p: if $p(x) = \tilde{p}(x)/Z$, then $\nabla_x \log p(x) = \nabla_x \log \tilde{p}(x)$. This observation is the key motivation for score-based methods: learning the score bypasses the intractability that afflicts direct density estimation with general neural networks, which would require computing $Z = \int \tilde{p}(x)\, dx$ for some unnormalized density $\tilde{p}$.

2.2. Langevin dynamics

One approach to sampling from a non-trivial distribution $p(x)$ over $\mathbb{R}^d$ is *Langevin dynamics*: starting from any x_0, iterate

$$x_{t+1} = x_t + \frac{\epsilon}{2}\nabla_x \log p(x_t) + \sqrt{\epsilon}\, z_t, \qquad z_t \sim \mathcal{N}(0, I). \tag{1}$$

As $\epsilon \to 0$ and $T \to \infty$, x_T converges in distribution to $p(x)$ [4]. Thus, access to the score $\nabla_x \log p(x)$ is sufficient for sampling. One need not evaluate the density itself.

3. Song & Ermon (2019): Noise-Conditioned Score Networks

3.1. Score matching

The goal is to train a neural network $s_\theta(x)$ to approximate $\nabla_x \log p(x)$ without first estimating p. The ideal objective would be $\frac{1}{2}\mathbb{E}_p[\|s_\theta(x) - \nabla_x \log p(x)\|^2]$, but this requires knowledge of the very score we wish to learn. Hyvärinen (2005) [5] showed that under mild regularity conditions, an equivalent objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_p\left[\operatorname{tr}\big(\nabla_x s_\theta(x)\big) + \tfrac{1}{2}\|s_\theta(x)\|^2\right]. \tag{2}$$

At first glance, the expectation in (2) might appear a problem but this can be cheaply computed via a Monte Carlo estimate over the training data. The equivalence between (2) and the ideal objective follows by expanding the squared error $\mathbb{E}_p[\|s_\theta(x) - \nabla_x \log p(x)\|^2]$ and integrating. The cross-term $\mathbb{E}_p[s_\theta(x)^\top \nabla_x \log p(x)]$ equals $\int s_\theta(x)^\top \nabla_x p(x)\, dx$; integrating by parts (and assuming $p(x) \to 0$ at the boundary) converts this to $-\mathbb{E}_p[\nabla_x \cdot s_\theta(x)]$, which is precisely the trace term in (2). The remaining $\mathbb{E}_p[\|\nabla_x \log p(x)\|^2]$ term is constant in θ, so minimizing (2) is equivalent to minimizing the ideal objective. The minimizer therefore satisfies $s_{\theta^\star}(x) = \nabla_x \log p(x)$ almost surely. Computing the trace term is expensive in high dimensions, but denoising score matching (below) circumvents this.

3.2. The manifold problem and its fix

There is another issue: score matching is only guaranteed to work when p satisfies regularity conditions including full support over $\mathbb{R}^d$. In practice, real data lives on a low-dimensional manifold embedded in a high-dimensional ambient space. Where $p(x) \approx 0$, there are no training samples, so the score network receives no gradient signal. The full-support condition is violated and the score simply cannot be learned off the data manifold.

The fix is to perturb the data with (Gaussian) noise. Adding noise with variance σ^2 yields the smoothed density

$$\tilde{p}_\sigma(\tilde{x}) = \int p(x)\, \mathcal{N}(\tilde{x};\, x,\, \sigma^2 I)\, dx,$$

which has full support over $\mathbb{R}^d$. Score matching now works and the score network receives well-defined gradient signal everywhere.

3.3. Denoising score matching

A key insight of Vincent (2011) is that fitting the score of the noised density $\tilde{p}_\sigma$ via the explicit score matching objective is *exactly equal*, up to a constant in θ, to minimising the denoising score matching (DSM) objective [6]:

$$\mathcal{J}_{\mathrm{DSM}}(\theta; \sigma) = \mathbb{E}_{p(x)}\, \mathbb{E}_{\tilde{x}\sim\mathcal{N}(x,\sigma^2 I)}\left[\left\| s_\theta(\tilde{x}) + \frac{\tilde{x} - x}{\sigma^2}\right\|^2\right]. \tag{3}$$

The derivation proceeds in three steps. Expanding the ideal objective $\mathbb{E}_{\tilde{p}_\sigma}[\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log \tilde{p}_\sigma(\tilde{x})\|^2]$ and discarding the term $\mathbb{E}_{\tilde{p}_\sigma}[\|\nabla_{\tilde{x}} \log \tilde{p}_\sigma\|^2]$, which is constant in θ, it suffices to rewrite the

cross-term. First, differentiate under the integral sign using $\tilde{p}_\sigma(\tilde{x}) = \int p(x)\,\mathcal{N}(\tilde{x}; x, \sigma^2 I)\,dx$:

$$\nabla_{\tilde{x}} \log \tilde{p}_\sigma(\tilde{x}) \;=\; \frac{\nabla_{\tilde{x}} \tilde{p}_\sigma(\tilde{x})}{\tilde{p}_\sigma(\tilde{x})} \;=\; \frac{\int p(x)\,\nabla_{\tilde{x}} \mathcal{N}(\tilde{x}; x, \sigma^2 I)\,dx}{\tilde{p}_\sigma(\tilde{x})}.$$

Second, apply the identity $\nabla_{\tilde{x}} \mathcal{N}(\tilde{x}; x, \sigma^2 I) = \mathcal{N}(\tilde{x}; x, \sigma^2 I) \cdot \nabla_{\tilde{x}} \log \mathcal{N}(\tilde{x}; x, \sigma^2 I)$ and the explicit Gaussian score $\nabla_{\tilde{x}} \log \mathcal{N}(\tilde{x}; x, \sigma^2 I) = -(\tilde{x} - x)/\sigma^2$ to obtain

$$\nabla_{\tilde{x}} \log \tilde{p}_\sigma(\tilde{x}) \;=\; \frac{\int p(x)\,\mathcal{N}(\tilde{x}; x, \sigma^2 I)\,\left(-\frac{\tilde{x}-x}{\sigma^2}\right) dx}{\tilde{p}_\sigma(\tilde{x})} \;=\; \mathbb{E}_{p(x|\tilde{x})}\left[-\frac{\tilde{x} - x}{\sigma^2}\right],$$

where the last equality uses Bayes' theorem to recognise the ratio as the posterior $p(x \mid \tilde{x}) \propto p(x)\,\mathcal{N}(\tilde{x}; x, \sigma^2 I)$. Third, substitute into the cross-term and apply the tower property to collapse the nested expectation into a joint expectation over $p(x)\,\mathcal{N}(\tilde{x}; x, \sigma^2 I)$. Reassembling and completing the square yields exactly $\mathcal{J}_{\mathrm{DSM}}(\theta; \sigma)$ plus a constant, establishing the equality.

Writing the perturbation as $\tilde{x} = x + \sigma z$ with $z \sim \mathcal{N}(0, I)$, the target inside (3) becomes $-z/\sigma$: the score network learns the direction back toward the clean data, scaled by $1/\sigma^2$. In other words, it learns to denoise. This equivalence between score estimation and denoising is central to everything that follows.

Notably, the argument above does not fundamentally depend on the Gaussian form of the noise kernel. For any kernel $q(\tilde{x} \mid x)$, the same three steps go through with $\nabla_{\tilde{x}} \log q(\tilde{x} \mid x)$ in place of $-(\tilde{x} - x)/\sigma^2$, yielding a DSM objective whose target is the posterior mean of the kernel score $\mathbb{E}_{p(x|\tilde{x})}[\nabla_{\tilde{x}} \log q(\tilde{x} \mid x)]$. The Gaussian case is simply a nice setting in which the target takes a particularly clean closed form.

3.4. Annealed Langevin dynamics and joint training

Using a single noise level σ introduces a tension. A large σ spreads the distribution so that Langevin dynamics mixes well globally, but the gradients are for a heavily perturbed target. A small σ keeps the target close to $p(x)$, but Langevin dynamics mixes poorly from a random initialization far from the data manifold. Song & Ermon resolve this with annealed Langevin dynamics: choose a geometric sequence of noise levels $\sigma_1 > \sigma_2 > \cdots > \sigma_L$, and run Langevin dynamics sequentially at each level, using the output of level i as the initialization for level $i + 1$. Large noise provides global mixing; small noise provides local refinement.

To avoid training L separate networks, a single noise-conditioned network $s_\theta(x, \sigma)$ is trained jointly on the unified objective

$$\mathcal{L}(\theta) = \frac{1}{L} \sum_{i=1}^{L} \lambda(\sigma_i)\,\ell(\theta; \sigma_i), \tag{4}$$

where $\ell(\theta; \sigma_i)$ is the per-level DSM loss. Since empirically $\|s_\theta(x, \sigma)\|_2 \propto 1/\sigma$, the per-level losses scale as $\ell(\theta; \sigma) \propto 1/\sigma^2$. The choice $\lambda(\sigma) = \sigma^2$ exactly cancels this, balancing contributions across all noise scales.

4. Ho et al. (2020): Denoising Diffusion Probabilistic Models

4.1. The forward and reverse processes

Ho et al. approached the same problem from a different direction: rather than score matching, they modelled the noising process explicitly through a Markov chain. The forward process is a Gaussian Markov chain that gradually corrupts data $x_0 \sim p_{\mathrm{data}}$ into pure noise:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t; \sqrt{1 - \beta_t}\, x_{t-1},\, \beta_t I\Big), \tag{5}$$

where $\{\beta_t\}_{t=1}^T$ is a fixed noise schedule chosen so that $x_T \approx \mathcal{N}(0, I)$. A crucial property is that any intermediate x_t can be sampled directly from x_0 without simulating the chain step by step. Setting $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{6}$$

The reverse process is parametrized as a learned Gaussian Markov chain backward in time:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1};\, \mu_\theta(x_t, t),\, \sigma_t^2 I),$$

where the variance σ_t^2 is fixed and only the mean μ_θ is learned. This ansatz is justified by the classical result of Feller (1949) that the time-reversal of a Gaussian diffusion with small step size is also Gaussian [7]. Note that the variance is treated as an unlearned parameter since this did better empirically.

4.2. Variational training and the noise reparametrisation

Because directly minimizing the KL divergence between the forward and reverse chains is intractable, Ho et al. instead maximize the Evidence Lower Bound (ELBO), as is standard in variational inference. After expanding and simplifying, the ELBO reduces to a sum of denoising terms, one per time-step, each comparing the learned reverse step to the forward posterior:

$$\sum_{t=2}^{T} \mathbb{E}_q \left[D_{\mathrm{KL}}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)) \right]. \tag{7}$$

Because the forward process is Gaussian, the posterior $q(x_{t-1} \mid x_t, x_0)$ is tractable in closed form as a Gaussian with computable mean $\tilde{\mu}_t(x_t, x_0)$ and variance $\tilde{\beta}_t$. Each KL term is therefore a tractable KL between two Gaussians. Two boundary terms in the full ELBO are handled as follows. The prior matching term at $t = T$, namely $D_{\mathrm{KL}}(q(x_T \mid x_0) \,\|\, \mathcal{N}(0, I))$, is driven to zero by design of the noise schedule and is treated as a constant. The reconstruction term at $t = 1$, $\mathbb{E}_q[\log p_\theta(x_0 \mid x_1)]$, is retained in the full objective but dropped in the simplified version without meaningful loss in practice. The remaining sum (7) is the operative training objective. Crucially, the one-step sampling property of equation (6) means each term of (7) can be estimated independently from a randomly sampled t. Which would enable, for example, efficient stochastic gradient descent.

To make the mean $\tilde{\mu}_t(x_t, x_0)$ a function of x_t alone (as required at inference time when x_0 is unavailable) one substitutes (6) to eliminate x_0 in favour of the noise ϵ. This gives

$$\tilde{\mu}_t(x_t, \epsilon) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\, \epsilon \right),$$

motivating noise reparametrization: instead of learning μ_θ directly, one parametrizes the network to predict the noise,

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\, \epsilon_\theta(x_t, t) \right),$$

where $\alpha_t = 1 - \beta_t$. Each denoising KL term then reduces to a weighted squared error in noise space. Ho et al. found empirically that dropping the t-dependent weightings and uniformly sampling t yields better sample quality, giving the simplified objective:

$$\mathcal{L}_{\mathrm{simple}}(\theta) = \mathbb{E}_{t, x_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon,\ t \right) \right\|^2 \right]. \tag{8}$$

This is precisely denoising score matching over noise scales indexed by t, establishing a connection with the Song & Ermon framework.

5. Song et al. (2023): A Unifying SDE Perspective

5.1. The general forward SDE

Both frameworks involve a discrete noising process indexed by a finite number of steps. Song et al. (2023) observed that as the number of steps $T \to \infty$, these discrete chains converge to continuous-time stochastic differential equations. The general forward process takes the form of an Itô SDE:

$$dx = f(x,t)\,dt + g(t)\,dW, \tag{9}$$

where W is a standard Wiener process and the drift f and diffusion coefficient g are chosen so that $x(0) \sim p_{\text{data}}$ is gradually transformed into $x(T) \sim \mathcal{N}(0, I)$. The two prior models emerge as special cases:

$$\begin{aligned} &\text{VE-SDE (NCSN):} \quad dx = \sqrt{\tfrac{d\sigma^2(t)}{dt}}\,dW, \\ &\text{VP-SDE (DDPM):} \quad dx = -\tfrac{1}{2}\beta(t)x\,dt + \sqrt{\beta(t)}\,dW. \end{aligned}$$

VE refers to variance exploding (the variance of $x(t)$ grows without bound) and VP to variance preserving (the variance stays approximately constant). They are different parametrizations of the same family.

5.2. The reverse-time SDE and its consequences

Anderson (1982) established that any forward diffusion SDE of the form (9) admits an exact reverse-time counterpart [8]:

$$dx = \left[f(x,t) - g(t)^2\,\nabla_x \log p_t(x)\right] dt + g(t)\,d\bar{W}, \tag{10}$$

where $\bar{W}$ is a reverse-time Wiener process and p_t denotes the marginal density of the forward process at time t. The reverse process depends only on the time-dependent score $\nabla_x \log p_t(x)$. This is precisely what is estimated by denoising score matching. Substituting a trained score network $s_\theta(x,t)$ (we don't have to necessarily use denoising score matching) and running any numerical SDE solver backwards yields a generative model.

This SDE lens delivers several benefits beyond unification. First, every SDE has an associated deterministic probability flow ODE with the same marginal densities p_t; this enables fast sampling via ODE solvers and exact log-likelihood computation without an ELBO. Second, the framework admits predictor–corrector samplers that interleave numerical SDE steps (predictor) with Langevin correction steps (corrector), combining the complementary strengths of DDPM and NCSN sampling. Third, conditional generation tasks such as inpainting and colourization can be handled by modifying the reverse SDE with a likelihood gradient $\nabla_x \log p(y \mid x)$, requiring no retraining. Finally, the SDE framework allows engineering new drift and diffusion functions to yield processes with special properties.

6. Discussion and Conclusion

The three papers reviewed here tell a coherent story, but one that was not obvious in advance. Score matching and variational inference are superficially quite different enterprises: one minimizes a functional of the score field, the other maximizes a likelihood lower bound over a latent Markov chain. That both procedures reduce, after algebraic simplification, to the same denoising regression problem is a non-trivial coincidence. The SDE perspective explains why; both are discretizations of the same continuous-time object, and the denoising target is simply the score of the marginal p_t, which is all the reverse SDE ever needs.

A deeper recurring theme is the role of noise as a computational device. In score matching, noise is introduced to repair the manifold problem: a defect of the data geometry. In DDPM,

noise is the latent variable of a hierarchical model: a modelling choice. The SDE perspective reveals these as two descriptions of the same object, and in doing so clarifies that the noise schedule is not merely a nuisance parameter but the central design choice that determines the geometry of the generative process.

This review has followed only one thread of a much larger story. The foundations laid by the three papers examined here have since supported an explosion of work on faster sampling schedules, improved architectures, and applications to image, audio, video, and scientific data. But understanding these foundations remains the essential first step: the subsequent developments are, in large part, refinements of the ideas traced here.

References

[1] Y. Song and S. Ermon. "Generative Modeling by Estimating Gradients of the Data Distribution". *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. NeurIPS'19. 2019, pp. 11918–11930.

[2] J. Ho, A. Jain, and P. Abbeel. "Denoising Diffusion Probabilistic Models". *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NeurIPS'20. 2020, pp. 6840–6851.

[3] Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NeurIPS'23. 2023, pp. 37799–37812.

[4] G. O. Roberts and R. L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". *Bernoulli* 2 (1996), pp. 341–363.

[5] A. Hyvärinen. "Estimation of Non-Normalized Statistical Models by Score Matching". *Journal of Machine Learning Research* 6 (2005), pp. 695–709.

[6] P. Vincent. "A Connection Between Score Matching and Denoising Autoencoders". *Neural Computation* 23.7 (2011), pp. 1661–1674.

[7] W. Feller. "On the Theory of Stochastic Processes, with Particular Reference to Applications". *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (1949).

[8] B. D. Anderson. "Reverse-time diffusion equation models". *Stochastic Processes and their Applications* 12.3 (May 1982), pp. 313–326.