From Denoising to Scores: DAEs to Noise-Conditional Score Networks

Ilan Aliouchouche   Clément Marie   Karim Rochd

ENS Paris-Saclay — Master MVA, 2025

Abstract

Contents

1 Introduction

Generative modeling aims to learn the underlying distribution p_{data} of a dataset in order to generate new, realistic samples. One powerful approach is Energy-Based Modeling (EBM), which defines the probability density as $p_\theta(x) = e^{-E_\theta(x)}/Z_\theta$. However, training EBMs via maximum likelihood is notoriously difficult because the partition function Z_θ is generally intractable[cite: 17].

Score matching offers an alternative by modeling the gradient of the log-density, $\nabla_x \log p(x)$, effectively bypassing the normalizing constant Z_θ[cite: 18]. While theoretically elegant, standard score matching struggles in practice with high-dimensional data and low-density regions[cite: 300].

In this report, we trace the evolution of score-based generative modeling through two foundational papers. First, we examine [**?**], which established a surprising equivalence between Denoising Autoencoders (DAEs) and score matching, showing that DAEs implicitly learn the score of a Gaussian-smoothed data distribution[cite: 164]. Second, we explore [**?**], which scales this insight by training a single Noise-Conditional Score Network (NCSN) across multiple noise levels. This multi-scale approach overcomes the limitations of the manifold hypothesis [cite: 446], enabling high-quality synthesis via Annealed Langevin Dynamics[cite: 308].

2 Background

2.1 Energy-Based Models and Score Matching

Energy-based models define a density

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta},$$

where the normalizing constant Z_θ is generally intractable, making maximum-likelihood training difficult in high dimensions.

To avoid computing Z_θ, Hyvärinen (2005) introduced *score matching*, which trains EBMs by aligning the model score $\nabla_x \log p_\theta(x)$ with the data score $\nabla_x \log q(x)$. This leads to the Explicit Score Matching (ESM) objective

$$J_{\text{ESM}}(\theta) = \tfrac{1}{2}\, \mathbb{E}_{q(x)} \left[\|\nabla_x \log p_\theta(x) - \nabla_x \log q(x)\|^2 \right],$$

which can be interpreted as a least-squares fit between score fields.

However, the data score $\nabla_x \log q(x)$ is not accessible for general datasets, making ESM unusable in practice. Hyvärinen showed that minimizing ESM is equivalent to minimizing the tractable Implicit Score Matching (ISM) loss

$$J_{\text{ISM}}(\theta) = \mathbb{E}_{q(x)} \left[\operatorname{tr}\left(\nabla_x^2 E_\theta(x) \right) + \tfrac{1}{2} \|\nabla_x E_\theta(x)\|^2 \right].$$

Since J_{ISM} depends only on $E_\theta(x)$ and its derivatives, it removes all dependence on the unknown score $\nabla_x \log q(x)$ while remaining fully equivalent to the original ESM criterion.

2.2 Denoising Autoencoders

Denoising autoencoders (DAEs) modify the standard autoencoder objective by training the model to reconstruct a clean input from a corrupted version. Instead of reproducing the input itself, as a vanilla autoencoder does, the DAE receives a perturbed sample $\tilde{x}$ generated through a corru ption process $q(\tilde{x} \mid x)$, typically Gaussian noise $\tilde{x} = x + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$ and a chosen noise level σ.

Denoising Autoencoder Training Objective

Let $h = f_\theta(\tilde{x})$ denote the encoder output and $p_\theta(x \mid h)$ the decoder distribution. The denoising autoencoder is trained by minimizing

$$L_{\mathrm{DAE}}(\theta) = -\mathbb{E}_{q(x)}\, \mathbb{E}_{q(\tilde{x}|x)} \left[\log p_\theta(x \mid h = f_\theta(\tilde{x}))\right],$$

that is, the negative log-likelihood of reconstructing the clean input x from its corrupted version $\tilde{x}$.

This objective encourages the model to learn directions that map corrupted inputs back toward the data manifold rather than simply copying the input.

[>=stealth,thick]
[circle, draw, minimum size=0.9cm] (x) at (0,0) x; [circle, draw, minimum size=0.9cm] (xt) at (2,1) $\tilde{x}$; [circle, draw, minimum size=0.9cm] (h) at (4,0) h; [circle, draw, minimum size=0.9cm] (L) at (2,-1.2) L;
[->] (x) – node[midway, above, xshift=-10pt] corruption (xt); [->] (xt) – node[midway, above, xshift=9pt] encoder (h);
[->] (h) – node[midway, right, xshift=-3pt, yshift=-5pt] decoder (L);
[->] (x) – (L);

Figure 1: Computation graph of a DAE. The clean input x is corrupted into $\tilde{x}$, encoded into h, and decoded to produce a reconstruction that is compared to x through the loss $\mathcal{L}$.

3 Vincent (2011): Denoising Autoencoders and Score Matching

3.1 Denoising score matching and its connection to denoising autoencoders

For simplicity, we denote the score model of our energy-based density by

$$\psi(x;\theta) \;\approx\; \nabla_x \log p_\theta(x),$$

so that $\psi(\cdot;\theta)$ is a vector field defined over the data space. Denoising Score Matching (DSM) introduces a corrupted observation $\tilde{x}$ obtained from a clean sample $x \sim q(x)$ through a corruption process $q_\sigma(\tilde{x} \mid x)$. The corresponding joint distribution factorizes as

$$q(x, \tilde{x}) = q(x)\, q_\sigma(\tilde{x} \mid x).$$

Denoising Score Matching Objective

The Denoising Score Matching (DSM) objective is defined as

$$J_{\mathrm{DSM}}(\theta) = \mathbb{E}_{q(x,\tilde{x})}\left[\frac{1}{2}\left\|\psi(\tilde{x};\theta) - \nabla_{\tilde{x}} \log q(\tilde{x} \mid x)\right\|^2\right],$$

that is, a mean-squared regression objective where the model predicts the posterior score $\nabla_{\tilde{x}} \log q(\tilde{x} \mid x)$ from the corrupted input $\tilde{x}$.

This has exactly the structure of a mean squared error regression problem where the input is $\tilde{x}$, the prediction is $\psi(\tilde{x};\theta)$, and the regression target is the posterior score $\nabla_{\tilde{x}} \log q(\tilde{x} \mid x)$.

Consider now the Gaussian corruption process used in denoising autoencoders:

$$q(\tilde{x} \mid x) = \mathcal{N}(\tilde{x} \mid x, \sigma^2 I).$$

Since DSM is a least-squares objective in $\psi(\tilde{x};\theta)$, the optimal vector field is the conditional expectation

$$\psi^\star(\tilde{x}) = \mathbb{E}_{q(x|\tilde{x})}[\nabla_{\tilde{x}} \log q(\tilde{x} \mid x)] .$$

For Gaussian corruption, the posterior score satisfies

$$\nabla_{\tilde{x}} \log q(\tilde{x} \mid x) = \frac{x - \tilde{x}}{\sigma^2}.$$

Thus,

$$\psi^\star(\tilde{x}) = \mathbb{E}_{q(x|\tilde{x})}\left[\frac{x - \tilde{x}}{\sigma^2}\right].$$

This expression makes the denoising interpretation explicit: the optimal vector field $\psi^\star(\tilde{x})$ points, in expectation, from a noisy sample $\tilde{x}$ toward its clean counterpart x, scaled by σ^{-2}.

The term $(x - \tilde{x})$ is precisely the reconstruction direction learned by denoising autoencoders, whose objective is to recover x from $\tilde{x}$.

A visual illustration is shown in Figure **??**, where a one-dimensional data manifold (in red) is corrupted with Gaussian noise. The Gaussian corruption kernels around manifold points are shown in green, the noisy samples in grey, and the vector field learned by a denoising autoencoder is represented in blue. The resulting field points in the direction of decreasing corruption, back toward the true manifold.

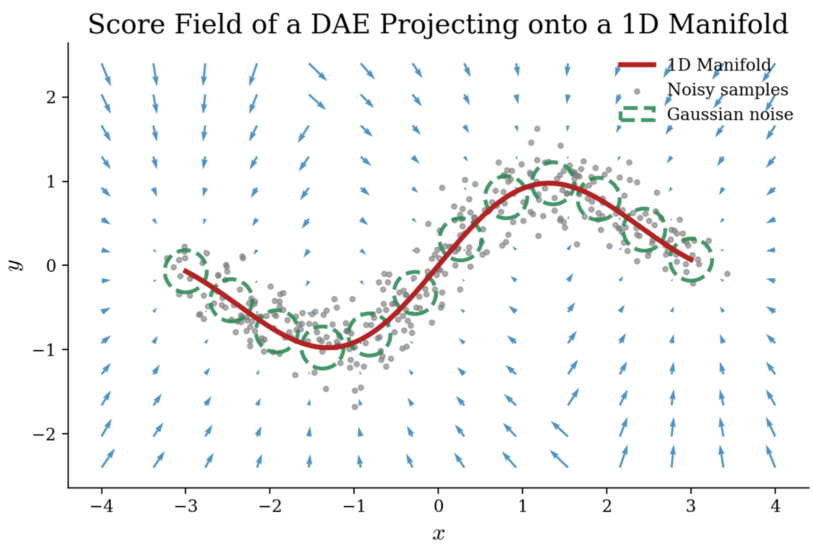

Figure 2: One-dimensional data manifold (red), Gaussian corruption kernels (green), noisy samples (grey), and the denoising vector field learned by a DAE (blue).

Vincent (2011) proved that the DSM objective is equivalent to the ESM and the proof does not depend on the particular form of $q(\tilde{x} \mid x)$ or $q(x)$.

3.2 An EBM Yielding the DAE Objective

Connection between DSM and the DAE objective

Consider the denoising autoencoder reconstruction loss

$$L_{\mathrm{DAE}}(\theta) = \mathbb{E}_{q_\sigma(x,\tilde{x})}\left[\left\|W^\top \sigma(W\tilde{x} + b) + c - x\right\|^2\right],$$

obtained under Gaussian corruption $q(\tilde{x} \mid x) = \mathcal{N}(\tilde{x} \mid x, \sigma^2 I)$.

For the energy-based model

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\big(-E(x;\theta)\big), \qquad E(x;\theta) = -\langle c, x\rangle - \tfrac{1}{2}\|x\|^2 + \sum_{j=1}^{d_h} \mathrm{softplus}(\langle W_j, x\rangle + b_j),$$

the associated score field satisfies

$$\psi(x;\theta) = \frac{1}{\sigma^2}\Big(W^\top \sigma(Wx + b) + c - x\Big).$$

Substituting this expression into the denoising score matching loss gives

$$J_{\mathrm{DSM}}(\theta) = \frac{1}{2\sigma^4} L_{\mathrm{DAE}}(\theta).$$

In other words, minimizing the DSM objective for this EBM is exactly equivalent to minimizing the denoising autoencoder reconstruction loss.

4 Song & Ermon (2019): Noise-Conditional Score Networks

4.1 Motivation: the manifold and low-density problems

Although Vincent (2011) showed that denoising autoencoders learn the score of a smoothed data distribution, using a single fixed noise level σ remains insufficient for high-dimensional generative modeling. Real-world data, such as natural images, lie near a low-dimensional manifold in $\mathbb{R}^d$, where the score $\nabla_x \log p_{\mathrm{data}}(x)$ is well defined but becomes unstable in low-density regions. When sampling with Langevin dynamics, the estimated gradients outside the data support are unreliable, causing poor mixing or divergence.

To address this, Song and Ermon (2019) proposed to add Gaussian noise to the data and train the score network across multiple noise levels. This smoothing makes the data distribution well behaved over all of $\mathbb{R}^d$, allowing the model to learn stable coarse-scale gradients for large σ and progressively refine fine details as σ decreases.

4.2 Noise-conditional network formulation

The key idea is to make the score model explicitly depend on the noise level σ so that one shared network can learn multiple smoothed versions of the data distribution.

Let $\{\sigma_1 > \sigma_2 > \cdots > \sigma_L\}$ denote a geometric sequence of noise scales. For each σ_i, define the corresponding Gaussian-smoothed data distribution

$$q_{\sigma_i}(x) = \int p_{\mathrm{data}}(x')\, \mathcal{N}(x \mid x', \sigma_i^2 I)\, dx'.$$

Each q_{σ_i} has full support in $\mathbb{R}^d$, making its score well defined everywhere. Instead of training separate networks for each noise level, the authors introduced a single model $s_\theta(x, \sigma)$ that receives both the noisy input x and the noise level σ as arguments and is trained to approximate

$$s_\theta(x, \sigma) \approx \nabla_x \log q_\sigma(x).$$

Conditioning on σ lets the same network learn score fields at different noise levels: smoother ones for large σ and more detailed ones for small σ.

Noise-Conditional Score Network (NCSN)

Given a ladder of noise scales $\{\sigma_i\}_{i=1}^L$, a single neural network $s_\theta(x, \sigma)$ is trained to estimate the score of the Gaussian-smoothed data distribution q_σ:

$$s_\theta(x, \sigma) \;\approx\; \nabla_x \log q_\sigma(x).$$

This formulation extends Vincent's denoising autoencoder interpretation. In Vincent (2011), training a DAE with Gaussian corruption of variance σ^2 makes its reconstruction field $(r_\theta(x) - x)/\sigma^2$ approximate the score $\nabla_x \log q_\sigma(x)$ of the data distribution smoothed by Gaussian noise. Song and Ermon (2019) generalize this idea by learning a single network $s_\theta(x, \sigma)$ that predicts these scores jointly for multiple noise levels σ, thereby capturing both coarse and fine structures of the data manifold.

4.3 Training objective: denoising score matching (DSM)

Once the model structure is defined, the next step is to specify how it is trained. Song and Ermon use a denoising score-matching objective that teaches the network to recover the exact gradient of the log-density for each noise level.

For each fixed noise level σ, the network is trained using the *denoising score-matching* objective

$$\ell(\theta; \sigma) = \tfrac{1}{2}\, \mathbb{E}_{x \sim p_{\text{data}}} \mathbb{E}_{\tilde{x} \sim \mathcal{N}(x, \sigma^2 I)} \Big[\big\| s_\theta(\tilde{x}, \sigma) + \tfrac{\tilde{x} - x}{\sigma^2} \big\|_2^2 \Big], \tag{1}$$

where $\tilde{x}$ denotes a corrupted version of the clean data x. The vector $(\tilde{x} - x)/\sigma^2$ corresponds to the exact score of the Gaussian corruption kernel, so the network learns to reproduce this field from noisy samples.

Training across all noise scales amounts to minimizing the averaged loss

$$\mathcal{L}(\theta) = \frac{1}{L} \sum_{i=1}^{L} \lambda(\sigma_i)\, \ell(\theta; \sigma_i), \qquad \lambda(\sigma_i) = \sigma_i^2, \tag{2}$$

where the weighting $\lambda(\sigma) = \sigma^2$ balances gradient magnitudes at different noise levels and prevents small-σ terms from dominating. This objective is simple, stable, and free of adversarial or likelihood components; it can be applied to any differentiable network architecture.

Key insight

By training $s_\theta(x, \sigma)$ to predict the gradient of the log-density of Gaussian-smoothed data for multiple noise levels, NCSNs learn a single multi-scale score field that spans from coarse global structure to fine details.

Sampling from these learned scores is performed using the *annealed Langevin dynamics* procedure described in Section 5.

5 Sampling with Annealed Langevin Dynamics

5.1 Motivation

Once the network has learned score functions $s_\theta(x, \sigma)$ for multiple noise levels, it can be used to generate new samples. The goal is to draw samples from the original data distribution $p_{\text{data}}(x)$ using only its estimated score field. A natural approach is *Langevin dynamics*, a stochastic process that iteratively perturbs a particle in the direction of increasing log-density while adding Gaussian noise to maintain exploration.

Standard Langevin dynamics updates a sample x_t as

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

where α is the step size. If the score $\nabla_x \log p(x)$ is exact and α is small enough, this procedure asymptotically produces samples from $p(x)$. However, when applied directly to high-dimensional data, the dynamics tend to mix very slowly and may get trapped within a single mode. Song and Ermon (2019) propose an *annealed* version that leverages the multi-scale scores learned by the Noise-Conditional Score Network.

5.2 Annealed Langevin dynamics algorithm

The central idea is to sample gradually from high noise to low noise. At large σ, the smoothed distribution $q_\sigma(x)$ is simple and well connected; as σ decreases, the process refines the sample toward the true data manifold. For each noise level σ_i, the network $s_\theta(x, \sigma_i)$ provides an estimate of the score $\nabla_x \log q_{\sigma_i}(x)$, and Langevin steps are taken accordingly.

Algorithm 1: Annealed Langevin Dynamics [?]

1. **Input:** trained score network $s_\theta(x, \sigma)$, noise levels $\sigma_1 > \cdots > \sigma_L$, number of iterations T per level, step constant ϵ.
2. Initialize $x_0 \sim \mathcal{N}(0, \sigma_1^2 I)$.
3. **for** $i = 1, \ldots, L$ **do**
   (a) Set $\alpha_i = \epsilon\, (\sigma_i/\sigma_L)^2$.
   (b) **for** $t = 1, \ldots, T$ **do**
      i. Sample $z_t \sim \mathcal{N}(0, I)$.
      ii. Update
$$x \leftarrow x + \frac{\alpha_i}{2}\, s_\theta(x, \sigma_i) + \sqrt{\alpha_i}\, z_t.$$
4. **Output:** final sample x.

Here the step size α_i scales proportionally to σ_i^2 to maintain a roughly constant signal-to-noise ratio across levels. Each stage uses the score field at its own noise level, producing a gradual denoising trajectory from random Gaussian noise to the data manifold. The total number of iterations $L \times T$ is a trade-off between sample quality and computational cost.

Key insight

Annealed Langevin dynamics generate samples by successively refining Gaussian noise using the learned multi-scale score fields. High-noise levels ensure global coverage, while low-noise levels restore fine structure consistent with the data distribution.

5.3 Practical considerations

Empirically, Song and Ermon (2019) report that many short updates per noise level yield better results than a few large steps. Choosing an adequate noise schedule is crucial: σ_1 must be large enough to cover the data manifold, and σ_L must be small but nonzero to avoid numerical instability. They also monitor the norm of $\sigma\, s_\theta(x, \sigma)$ as a diagnostic, which should remain roughly constant across noise scales if the model is properly trained.

The resulting samples on datasets such as CIFAR-10 and MNIST demonstrate that score-based generative models can produce realistic images without adversarial training or explicit likelihood estimation, confirming the effectiveness of the annealed Langevin framework.

6 A Unifying View: From DAEs to NCSNs

The transition from Vincent's Denoising Autoencoders to Song and Ermon's Noise-Conditional Score Networks represents a shift from representation learning to full generative modeling.

6.1 The Bridge

Vincent (2011) proved that minimizing the reconstruction error of a DAE with noise level σ is mathematically equivalent to performing score matching on the smoothed density $q_\sigma(x) = \int p_{\text{data}}(y)\mathcal{N}(x|y, \sigma^2 I)dy$[cite: 156, 164]. Effectively, the "denoising vector" learned by the autoencoder points towards the mode of the data distribution[cite: 129].

However, a single DAE trained with a fixed, small σ fails as a generative model. As noted by Song and Ermon (2019), real-world data resides on low-dimensional manifolds[cite: 363]. Ideally, we want $\sigma \to 0$, but in this limit, the score is undefined in the ambient space, and the training signal vanishes in low-density regions[cite: 300, 301]. Conversely, a large σ makes score estimation stable but destroys fine data details.

6.2 The Multi-Scale Solution

The key innovation of the NCSN framework is to accept this trade-off and utilize *all* noise levels. Instead of training separate DAEs, Song and Ermon train a single network $s_\theta(x, \sigma)$ conditioned on the noise level[cite: 459].

- **At high σ:** The noise "fills" the low-density regions, allowing the network to learn global structure and mix between distant modes[cite: 447].
- **At low σ:** The network refines the samples, recovering the sharp details of the data manifold[cite: 508].

This transforms the DAE from a feature extractor into a multi-scale gradient estimator capable of guiding a sampling process from pure noise to realistic data.

7 Experimental Results and Observations

We evaluate the efficacy of Noise-Conditional Score Networks (NCSN) combined with Annealed Langevin Dynamics. The experiments focus on two main aspects: the ability to generate high-fidelity samples on complex image datasets and the ability to mix between separated modes in a distribution.

7.1 Image Generation Performance

[cite$_{s}tart]Themodelwastrainedon MNIST, CelebA, andCIFAR-10datasets[cite:40].[cite_{s}tart]Thegenerated$
$basedmodelscanproducehigh-qualityimagescomparabletomodernlikelihood-basedmodelsandGANswithout$
$11, 41]$.

Quantitative evaluation on CIFAR-10 highlights the effectiveness of the method:

- **Inception Score:** The NCSN achieves a score of **8.87**, setting a new state-of-the-art for unconditional generative models at the time of publication[cite: 11, 42]. [cite$_{s}tart]$**FID Score:**$ThemodelachievesaFréch$
$42, 276]$.

7.2 Mode Mixing (The Toy Experiment)

A crucial observation from [**?**] is the failure of standard Langevin dynamics to mix properly when modes are separated by low-density regions.

- **Standard Langevin:** When applied to a mixture of two Gaussians with disjoint supports, standard sampling fails to recover the correct relative weights of the modes, often becoming trapped in a single mode[cite: 124, 133].
- **Annealed Langevin:** By starting with large noise (σ_1), the sampler bridges the low-density gap. [cite$_{s}tart]Asthenoiselevelisannealeddownto\sigma_L$, the method faithfully recovers the relative weights of the mixture distribution[cite: 171, 243].

7.3 Representation Learning

Beyond generation, the score network learns semantic representations of the data. [cite$_{s}tart]Thisisevidencedbyim$
$12, 278]$.

8 Practical Notes Limitations

8.1 Hyperparameters and Stability

Implementing NCSNs requires careful tuning of the noise "ladder" $\{\sigma_i\}_{i=1}^{L}$:

- **Geometric Sequence:** The noise levels are chosen as a geometric progression where $\frac{\sigma_1}{\sigma_2} = \cdots = \frac{\sigma_{L-1}}{\sigma_L} > 1$[cite: 457]. σ_1 must be large enough to cover the data variance, while σ_L should be small enough to be invisible to the human eye[cite: 458].
- **Step Size (α):** For the annealed Langevin dynamics, the step size α_i must scale with the noise variance. Song and Ermon propose $\alpha_i \propto \sigma_i^2$ to maintain a constant signal-to-noise ratio throughout the sampling process[cite: 509].

8.2 Limitations

While avoiding adversarial training is a major advantage, the framework has limitations:

- **Computational Cost:** The sampling process is iterative (e.g., $T = 100$ steps per noise level [cite: 522]), which is significantly slower than single-pass generators like GANs.
- **Bias:** As noted by Vincent (2011), the score estimator is biased when $\sigma > 0$. While beneficial for stability, it means we are theoretically sampling from a slightly smoothed distribution rather than the exact p_{data}[cite: 187].

9 Conclusion

We have presented a coherent narrative linking Denoising Autoencoders to modern score-based generative models. Vincent (2011) laid the theoretical groundwork by identifying the connection between denoising objectives and score matching[cite: 164]. Song and Ermon (2019) resolved the geometric instabilities of the single-noise approach by introducing noise-conditional networks and annealed sampling[cite: 281]. This lineage demonstrates that learning to "undo" noise at multiple scales is equivalent to learning the gradient of the data distribution, providing a stable and principled foundation for generative modeling without adversarial objectives.

A Derivation Snippets

B Glossary of Notation