

From Energy-Based Models to Diffusion Models

Score Matching, Denoising, and Noise-Conditional Networks

Dayeon Yoon

Uncertainty Quantification Lab
Seoul National University

July 20, 2026

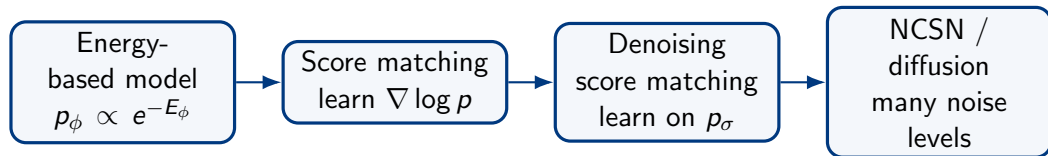
Table of contents

- 1 Overview
- 2 Energy-Based Models
- 3 Score Matching
- 4 Denoising Score Matching
- 5 Noise-Conditional Score Networks
- 6 Connection with Diffusion Models

Contents

- 1 Overview
- 2 Energy-Based Models
- 3 Score Matching
- 4 Denoising Score Matching
- 5 Noise-Conditional Score Networks
- 6 Connection with Diffusion Models

Overview



Central message

Diffusion models can be read as a sequence of statistical tricks that turn an intractable density-learning problem into tractable conditional regressions.

Contents

- 1 Overview
- 2 Energy-Based Models**
- 3 Score Matching
- 4 Denoising Score Matching
- 5 Noise-Conditional Score Networks
- 6 Connection with Diffusion Models

Energy-based model

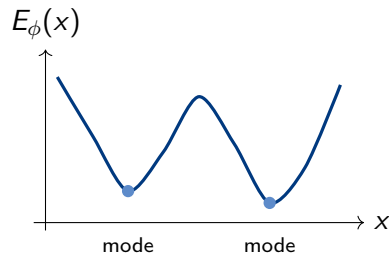
Energy-based model

For $x \in \mathbb{R}^D$ and energy $E_\phi : \mathbb{R}^D \rightarrow \mathbb{R}$,

$$p_\phi(x) = \frac{e^{-E_\phi(x)}}{Z_\phi}, \quad Z_\phi = \int_{\mathbb{R}^D} e^{-E_\phi(u)} du.$$

Low energy means high probability; adding a constant to E_ϕ leaves p_ϕ unchanged.

- The model specifies *relative* plausibility locally.
- Z_ϕ couples the whole state space globally.



Challenges of MLE training in EBM

$$\begin{aligned}\mathcal{L}_{\text{MLE}}(\phi) &= -\mathbb{E}_{x \sim p_{\text{data}}} \log p_{\phi}(x) \\ &= \mathbb{E}_{p_{\text{data}}} E_{\phi}(x) + \log Z_{\phi}, \\ \nabla_{\phi} \mathcal{L}_{\text{MLE}}(\phi) &= \underbrace{\mathbb{E}_{p_{\text{data}}} \nabla_{\phi} E_{\phi}(x)}_{\text{positive phase}} - \underbrace{\mathbb{E}_{p_{\phi}} \nabla_{\phi} E_{\phi}(x)}_{\text{negative phase}}.\end{aligned}$$

The bottleneck

The second expectation requires samples from the current model, but accurate sampling is itself a hard high-dimensional MCMC problem.

- Approximate route: contrastive divergence or persistent MCMC.
- Route pursued here: replace density matching by *score matching*.

Contents

- 1 Overview
- 2 Energy-Based Models
- 3 Score Matching**
- 4 Denoising Score Matching
- 5 Noise-Conditional Score Networks
- 6 Connection with Diffusion Models

Score Function

Score function

For a density $p(\mathbf{x})$ on $\mathbb{R}^D$, the score is

$$\mathbf{s}(\mathbf{x}) := \nabla_{\mathbf{x}} \log p(\mathbf{x}).$$

It forms a vector field pointing toward regions of higher probability.

Why model scores instead of densities?

1. No normalization constant:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}) = \nabla_{\mathbf{x}} \log \tilde{p}(\mathbf{x}) - \nabla_{\mathbf{x}} \log Z = \nabla_{\mathbf{x}} \log \tilde{p}(\mathbf{x}).$$

So the partition function Z is bypassed.

2. Complete representation: Since the score is the gradient of the log-density, it characterizes the distribution up to a constant.

The generative score is not the statistician's score

	Generative / data score	Statistician's score
Definition	$s_{\theta}(x) = \nabla_x \log p_{\theta}(x)$	$U_{\theta}(x) = \nabla_{\theta} \log p_{\theta}(x)$
Differentiates	the observation x	the parameter θ
Geometry	vector field on sample space	tangent direction in parameter space
Main use	sampling and denoising	estimation, tests, Fisher information
Zero mean	requires decay at the boundary	$\mathbb{E}_{\theta} U_{\theta}(X) = 0$ under regularity

Throughout this talk, “score” means $\nabla_x \log p(x)$ unless stated otherwise.

Score Matching

Instead of fitting normalized probabilities, score matching trains EBMs by aligning the model score with the data score.

Population objective

$$\mathcal{L}_{\text{SM}}(\phi) := \frac{1}{2} \mathbb{E}_{X \sim p} \|s_{\phi}(X) - \nabla_x \log p(X)\|^2.$$

The target is unknown because only samples $X_1, \dots, X_n \sim p$ are observed.

Removing the unknown data score

Proposition 3.2.1 (tractable score matching)

Assume p and s_ϕ are differentiable and $p(x)s_{\phi,i}(x) \rightarrow 0$ at the boundary for every coordinate i . Then

$$\mathcal{L}_{\text{SM}}(\phi) = \tilde{\mathcal{L}}_{\text{SM}}(\phi) + C_p, \quad \tilde{\mathcal{L}}_{\text{SM}}(\phi) = \mathbb{E}_p \left[\operatorname{div} s_\phi(X) + \frac{1}{2} \|s_\phi(X)\|^2 \right],$$

where C_p does not depend on ϕ .

Therefore any sufficiently expressive minimizer satisfies

$$s_{\phi^*}(x) = \nabla_x \log p(x) \quad \text{for } p\text{-a.e. } x.$$

Proof, part I

Coordinatewise integration by parts

For differentiable u, v with vanishing boundary term,

$$\int u(x) \partial_i v(x) dx = - \int v(x) \partial_i u(x) dx.$$

Take $u = s_{\phi,i}$ and $v = p$. Since $p \partial_i \log p = \partial_i p$,

$$\begin{aligned} \mathbb{E}_p \langle s_{\phi}(X), \nabla \log p(X) \rangle &= \sum_{i=1}^D \int s_{\phi,i}(x) \partial_i p(x) dx \\ &= - \sum_{i=1}^D \int p(x) \partial_i s_{\phi,i}(x) dx \\ &= -\mathbb{E}_p[\text{div } s_{\phi}(X)]. \end{aligned}$$

Proof, part II

$$\begin{aligned}\mathcal{L}_{\text{SM}}(\phi) &= \frac{1}{2} \mathbb{E}_p \|s_\phi\|^2 - \mathbb{E}_p \langle s_\phi, \nabla \log p \rangle + \frac{1}{2} \mathbb{E}_p \|\nabla \log p\|^2 \\ &= \underbrace{\mathbb{E}_p \left[\frac{1}{2} \|s_\phi\|^2 + \text{div } s_\phi \right]}_{\tilde{\mathcal{L}}_{\text{SM}}(\phi)} + \underbrace{\frac{1}{2} \mathbb{E}_p \|\nabla \log p\|^2}_{C_p}.\end{aligned}$$

Conclusion

C_p is independent of ϕ , so the tractable and original objectives have identical minimizers.

New bottleneck

Exact divergence is a trace over D input coordinates; for neural networks, its derivative also introduces expensive second-order automatic differentiation.

Hutchinson's Estimator

Let $u \in \mathbb{R}^D$ satisfy $\mathbb{E}[u] = 0$ and $\mathbb{E}[uu^\top] = I$ (Gaussian, Rademacher). For any square matrix A ,

$$\text{Tr}(A) = \mathbb{E}_u[u^\top A u].$$

Hence, with $J_{s_\phi}(x) = \nabla_x s_\phi(x)$,

$$\text{div } s_\phi(x) = \text{Tr } J_{s_\phi}(x) = \mathbb{E}_u[u^\top J_{s_\phi}(x) u].$$

Computational gain

Each probe uses a JVP/VJP rather than enumerating all D diagonal entries.

$$\text{JVP: } \frac{f(x + \varepsilon v) - f(x)}{\varepsilon} \approx \frac{\partial f}{\partial x} \cdot v$$

Statistical price

K probes are unbiased, but have Monte Carlo variance of order $1/K$.

Sliced Score Matching

Combining Hutchinson's identity with $\mathbb{E}_u[(u^\top s)^2] = \|s\|^2$ gives

$$\tilde{\mathcal{L}}_{\text{SSM}}(\phi) = \mathbb{E}_{X,u} \left[u^\top J_{s_\phi}(X)u + \frac{1}{2}(u^\top s_\phi(X))^2 \right].$$

Interpretation

Instead of fitting every coordinate of the vector field simultaneously, each probe checks one random one-dimensional slice; averaging restores the full loss.

Sampling with Langevin Dynamics

The Langevin SDE with invariant density p is

$$dX_t = \nabla \log p(X_t) dt + \sqrt{2} dW_t.$$

Euler–Maruyama with learned score $s_{\hat{\phi}}$ gives

$$X_{k+1} = X_k + \eta s_{\hat{\phi}}(X_k) + \sqrt{2\eta} Z_k, \quad Z_k \sim \mathcal{N}(0, I).$$

Drift

$s_{\hat{\phi}}$ pushes particles toward increasing log-density.

Diffusion

Gaussian noise prevents deterministic collapse into a local mode.

Mixing can still be extremely slow when modes are separated by low-density regions.

Q. SDE-based Sampling

- **Overdamped Langevin:**

$$dx = -\nabla E(x)dt + \sqrt{2} dw$$

Position-only diffusion in an energy landscape.

- **Underdamped / Kinetic Langevin:** adds velocity or momentum, so sampling can move faster.
- **OU / Score-SDE:** uses time-dependent noising and reverse-time generation.
- **Probability-Flow ODE:** removes noise and transports data deterministically.

Contents

- 1 Overview
- 2 Energy-Based Models
- 3 Score Matching
- 4 Denoising Score Matching**
- 5 Noise-Conditional Score Networks
- 6 Connection with Diffusion Models

Limitations of SM/SSM

Vanilla or sliced score matching faces three related problems:

- 1 **Manifold hypothesis:** image data may concentrate near a lower-dimensional set, so the ambient density score is undefined or unstable.

Manifold

- A manifold is a low-dimensional structure embedded in a high-dimensional space.
- Score matching learns $s(x) = \nabla_x \log p_{\text{data}}(x)$.
- If the data distribution is supported only on a lower-dimensional manifold, it does not have a regular density in the full ambient space.
- Therefore, the score can be undefined away from the manifold and unstable near it.

Limitations of SM/SSM

- 2 **Probe variance:** Hutchinson directions add Monte Carlo noise.
- 3 **Derivative cost:** JVP/VJP evaluations remain expensive.

Denoising Score Matching

Corrupt $X \sim p$ through a known kernel $q_\sigma(y | x)$ and learn the score of the smoothed marginal

$$p_\sigma(y) = \int q_\sigma(y | x)p(x) dx.$$

For Gaussian noise, p_σ has full support and a tractable conditional score.

Denoising Score Matching

The desired noisy-score regression is

$$\mathcal{L}_{\text{SM}}(\phi; \sigma) = \frac{1}{2} \mathbb{E}_{Y \sim p_\sigma} \|s_\phi(Y; \sigma) - \nabla_Y \log p_\sigma(Y)\|^2.$$

We can sample Y but generally cannot evaluate $\nabla_Y \log p_\sigma(Y)$.

Condition on what generated the noise

The conditional score $\nabla_Y \log q_\sigma(Y | X)$ is known, suggesting

$$\mathcal{L}_{\text{DSM}}(\phi; \sigma) = \frac{1}{2} \mathbb{E}_{X, Y|X} \|s_\phi(Y; \sigma) - \nabla_Y \log q_\sigma(Y | X)\|^2.$$

Two elementary score identities explain why this substitution is exact.

Lemma 1

Let $p(x, y) = p(x)q(y | x)$, with marginal $p_Y(y) > 0$. Then

Lemma 1 (score decomposition)

$$\nabla_y \log q(y | x) = \nabla_y \log p_Y(y) + \nabla_y \log p(x | y).$$

Proof

Bayes' rule gives

$$\log p(x | y) = \log q(y | x) + \log p(x) - \log p_Y(y).$$

Differentiate with respect to y ; $\log p(x)$ is constant in y , and rearrange. □

Lemma 2

Lemma 2 (posterior-score centering)

For every regular y ,

$$\mathbb{E}[\nabla_y \log p(X | y) | Y = y] = 0.$$

Proof

$$\begin{aligned}\mathbb{E}[\nabla_y \log p(X | y) | y] &= \int p(x | y) \nabla_y \log p(x | y) dx \\ &= \int \nabla_y p(x | y) dx \\ &= \nabla_y \int p(x | y) dx = \nabla_y 1 = 0.\end{aligned}$$

Interchanging derivative and integral is the usual regularity condition. □

Equivalence of SM and DSM

Theorem 3.3.1 (equivalence of SM and DSM)

For fixed $\sigma > 0$ and suitable regularity,

$$\mathcal{L}_{\text{DSM}}(\phi; \sigma) = \mathcal{L}_{\text{SM}}(\phi; \sigma) + C_\sigma,$$

where

$$C_\sigma = \frac{1}{2} \mathbb{E} \|\nabla_y \log p(X | Y)\|^2$$

does not depend on ϕ . Both losses therefore have the minimizer

$$s_{\phi^*}(y; \sigma) = \nabla_y \log p_\sigma(y) \quad \text{for } p_\sigma\text{-a.e. } y.$$

Therefore, intractable marginal target of SM can be replaced by tractable conditional target of DSM.

Proof

Write

$$a(Y) = s_\phi(Y; \sigma) - \nabla_Y \log p_\sigma(Y), \quad r(X, Y) = \nabla_Y \log p(X | Y).$$

By Lemma 1, $s_\phi - \nabla \log q_\sigma = a - r$. Therefore

$$\begin{aligned} 2\mathcal{L}_{\text{DSM}} &= \mathbb{E} \|a(Y) - r(X, Y)\|^2 \\ &= \mathbb{E} \|a(Y)\|^2 + \mathbb{E} \|r(X, Y)\|^2 - 2\mathbb{E} \langle a(Y), r(X, Y) \rangle. \end{aligned}$$

The cross-term is

$$\mathbb{E} \langle a(Y), r(X, Y) \rangle = \mathbb{E}_Y \langle a(Y), \mathbb{E}[r(X, Y) | Y] \rangle = 0$$

by Lemma 2. Hence

$$\mathcal{L}_{\text{DSM}} = \mathcal{L}_{\text{SM}} + \frac{1}{2} \mathbb{E} \|r(X, Y)\|^2.$$

The last term is independent of ϕ . □

DSM with Gaussian kernel

Let

$$Y = X + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \quad q_\sigma(y | x) = \mathcal{N}(y; x, \sigma^2 I).$$

Then

$$\nabla_y \log q_\sigma(y | x) = \frac{x - y}{\sigma^2} = -\frac{\varepsilon}{\sigma}.$$

Thus

$$\mathcal{L}_{\text{DSM}}(\phi; \sigma) = \frac{1}{2} \mathbb{E}_{X, \varepsilon} \left\| s_\phi(X + \sigma\varepsilon; \sigma) + \frac{\varepsilon}{\sigma} \right\|^2.$$

Three equivalent parameterizations

Learn the score s_ϕ , the noise $\varepsilon_\phi = -\sigma s_\phi$, or the denoised mean $D_\phi(y) = y + \sigma^2 s_\phi(y)$.

Q. DSM with non-Gaussian kernel: heavy-tailed DSM

Use a generalized normal kernel

$$Y = X + \sigma \varepsilon, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \text{GN}(0, 1, \beta),$$

where

$$q_{\sigma, \beta}(y | x) \propto \exp \left\{ - \sum_i \left| \frac{y_i - x_i}{\sigma} \right|^\beta \right\}.$$

Here, $\beta = 2$ gives Gaussian-type noise, while $\beta = 1$ gives Laplace noise; smaller β means heavier tails. In this case,

$$\mathcal{L}_{\text{HTDSM}}(\phi; \sigma, \beta) = \frac{1}{2} \mathbb{E}_{X, \varepsilon} \left\| s_\phi(X + \sigma \varepsilon; \sigma, \beta) + \frac{\beta}{\sigma} \text{sign}(\varepsilon) \odot |\varepsilon|^{\beta-1} \right\|^2.$$

Why heavy-tailed DSM?

Heavy-tailed noise spreads samples farther in high dimensions, helping score learning in low-density regions and reducing the thin-annulus problem of Gaussian DSM.

Smoothing makes Langevin sampling better posed

At a fixed noise level, sample from p_σ via

$$Y_{k+1} = Y_k + \eta s_{\hat{\phi}}(Y_k; \sigma) + \sqrt{2\eta} Z_k.$$

Noise injection helps in two ways:

- 1 **Well-defined gradients:** Gaussian convolution gives full support and a smooth score in ambient space.
- 2 **Improved coverage:** smoothing fills low-density gaps and lowers barriers between modes, improving training signal and MCMC movement.

But one σ cannot do everything

Small σ preserves detail but mixes poorly; large σ mixes well but erases fine structure. This tension motivates multiple noise levels.

DSM is Denoising: Tweedie's Formula

Let $X \sim p$, $Y | X = x \sim q(\cdot | x)$, and assume differentiation under the integral is valid.

Generalized Tweedie / Fisher identity

$$\nabla_y \log p_Y(y) = \mathbb{E}[\nabla_y \log q(y | X) | Y = y].$$

For additive noise $Y = X + \xi$ with density r , this becomes

$$\nabla_y \log p_Y(y) = \mathbb{E}[\nabla \log r(y - X) | Y = y].$$

Beyond Gaussian noise

The identity always targets the posterior mean of the *noise score*. A posterior mean of X appears when that noise score is affine in X .

Tweedie's formula

Proof

From Lemma 1,

$$\nabla_y \log q(y | X) = \nabla_y \log p_Y(y) + \nabla_y \log p(X | y).$$

Take the conditional expectation given $Y = y$ and apply Lemma 2. □

If $Y = X + \sigma\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$, then

$$\nabla_y \log q_\sigma(y | X) = \frac{X - y}{\sigma^2}.$$

Therefore

Tweedie's formula

$$\mathbb{E}[X | Y = y] = y + \sigma^2 \nabla_y \log p_\sigma(y).$$

SURE estimates

Let $Y = x + \sigma Z$, $Z \sim \mathcal{N}(0, I_D)$, with fixed unknown x , and let $D : \mathbb{R}^D \rightarrow \mathbb{R}^D$ be weakly differentiable.

SURE theorem

$$\text{SURE}(D; Y) := \|D(Y) - Y\|^2 + 2\sigma^2 \text{div } D(Y) - D\sigma^2$$

is unbiased for conditional MSE:

$$\mathbb{E}[\text{SURE}(D; Y) \mid x] = \mathbb{E}[\|D(Y) - x\|^2 \mid x].$$

Hence a denoiser can be tuned using noisy observations alone.

Stein's identity

Stein identity

If $Z \sim \mathcal{N}(0, I_D)$ and $g : \mathbb{R}^D \rightarrow \mathbb{R}^D$ is regular enough,

$$\mathbb{E}[Z^\top g(Z)] = \mathbb{E}[\text{div } g(Z)].$$

Proof sketch with all essential steps

For each coordinate, the standard normal density φ satisfies $\partial_i \varphi(z) = -z_i \varphi(z)$. Integration by parts gives

$$\int z_i g_i(z) \varphi(z) \, dz = - \int g_i(z) \partial_i \varphi(z) \, dz = \int \partial_i g_i(z) \varphi(z) \, dz,$$

where the boundary term vanishes. Sum over i . □

With $Y = x + \sigma Z$, the chain rule yields $\mathbb{E}[(Y - x)^\top a(Y) \mid x] = \sigma^2 \mathbb{E}[\text{div } a(Y) \mid x]$.

Proof of SURE

Write $a(Y) = D(Y) - Y$. Then

$$\begin{aligned}\|D(Y) - x\|^2 &= \|a(Y) + (Y - x)\|^2 \\ &= \|a(Y)\|^2 + 2a(Y)^\top(Y - x) + \|Y - x\|^2.\end{aligned}$$

Conditional on x , Stein's identity gives

$$\mathbb{E}[a(Y)^\top(Y - x) \mid x] = \sigma^2 \mathbb{E}[\operatorname{div} a(Y) \mid x] = \sigma^2 \mathbb{E}[\operatorname{div} D(Y) - D \mid x],$$

and $\mathbb{E}\|Y - x\|^2 = D\sigma^2$. Substitution yields

$$\mathbb{E}[\|D(Y) - x\|^2 \mid x] = \mathbb{E}[\|D(Y) - Y\|^2 + 2\sigma^2 \operatorname{div} D(Y) - D\sigma^2 \mid x].$$

This is exactly the SURE identity. □

Parameterizing the denoiser reveals score matching inside SURE

Set

$$D_\phi(y) = y + \sigma^2 s_\phi(y; \sigma).$$

Then $D_\phi(y) - y = \sigma^2 s_\phi(y; \sigma)$ and $\text{div } D_\phi(y) = D + \sigma^2 \text{div } s_\phi(y; \sigma)$. Hence

$$\frac{1}{2\sigma^4} \text{SURE}(D_\phi; y) = \frac{1}{2} \|s_\phi(y; \sigma)\|^2 + \text{div } s_\phi(y; \sigma) + \text{const}(\sigma).$$

Loss-level equivalence

Expected SURE is Hyvärinen's tractable score-matching loss on p_σ , up to scale and an additive constant.

Tweedie supplies the same population minimizer: $D^*(y) = \mathbb{E}[X | y] = y + \sigma^2 \nabla \log p_\sigma(y)$.

Q. weakly differentiable, integrable divergence term assumption

In the canonical DSM pipeline, they hold automatically.

- **1. Gaussian smoothing** makes the density smooth and positive:

$$p_{\sigma} = p_{\text{data}} * \mathcal{N}(0, \sigma^2 I).$$

So the score and denoiser are well-defined, and Gaussian tails remove boundary terms.

- **2. Network parameterization** gives weak differentiability. Smooth activations are directly differentiable; ReLU networks are piecewise-linear and continuous, so they are weakly differentiable and have an a.e. divergence.

Thus, with Gaussian noise and standard networks, DSM regularity assumptions usually hold automatically.

Q. Pre-determining the structure of the score

Approach	Structure imposed / what is learned
Conservative	$\nabla \log p_\phi = -\nabla_{\mathbf{x}} E_\phi$ (curl-free); learn a scalar energy E_ϕ
EDM precondition.	$D_\theta = c_{\text{skip}} \mathbf{x} + c_{\text{out}} F_\theta(\cdot)$; learn only the residual F_θ
Symmetry-constrained	$\nabla \log p_\phi(U\mathbf{x}) = U \nabla \log p_\phi(\mathbf{x}) \quad \forall U \in G$; learn a G -equivariant field ($E(3)$, permutations)
FP-consistent	penalize the Fokker–Planck residual of $\{\nabla \log p_\phi(\cdot, t)\}$

All keep the DSM/NCSN loss; they only shrink the hypothesis class.

Contents

- 1 Overview
- 2 Energy-Based Models
- 3 Score Matching
- 4 Denoising Score Matching
- 5 Noise-Conditional Score Networks**
- 6 Connection with Diffusion Models

Noise-Conditional Score Networks

Choose

$$0 < \sigma_1 < \sigma_2 < \dots < \sigma_L.$$



NCSN idea

A single network $s_\phi(y, \sigma)$ learns the score of every smoothed marginal, using the noise level as an input.

NCSN training

For $Y_i = X + \sigma_i \varepsilon$,

$$\mathcal{L}_{\text{NCSN}}(\phi) = \sum_{i=1}^L \lambda(\sigma_i) \frac{1}{2} \mathbb{E}_{X, \varepsilon} \left\| s_{\phi}(X + \sigma_i \varepsilon, \sigma_i) + \frac{\varepsilon}{\sigma_i} \right\|^2.$$

The population optimum is

$$s_{\phi^*}(y, \sigma_i) = \nabla_y \log p_{\sigma_i}(y) \quad i = 1, \dots, L.$$

Weighting

Since the Gaussian target has scale $1/\sigma_i$, a common choice $\lambda(\sigma_i) \propto \sigma_i^2$ balances contributions across levels.

Sampling runs Langevin dynamics from coarse to fine

- 1: Choose $\sigma_L > \dots > \sigma_1$ and initialize $x \sim \mathcal{N}(0, \sigma_L^2 I)$
- 2: **for** $i = L, L-1, \dots, 1$ **do**
- 3: Set a step size η_i (typically proportional to σ_i^2)
- 4: **for** $k = 1, \dots, K_i$ **do**
- 5: Draw $z \sim \mathcal{N}(0, I)$
- 6: $x \leftarrow x + \eta_i s_{\hat{\phi}}(x, \sigma_i) + \sqrt{2\eta_i} z$
- 7: **end for**
- 8: Use the current x to initialize the next lower noise level
- 9: **end for**
- 10: **return** x

Why annealing helps

High noise connects modes and finds global structure; lower noise successively restores detail while starting from an already plausible sample.

NCSN sampling is accurate but expensive

If there are L noise levels and K Langevin steps per level, generation needs roughly

$O(LK)$ sequential score-network evaluations.

Two forces make LK large:

- 1 **Local accuracy and stability:** large steps amplify score error and discretization bias.
- 2 **Slow high-dimensional mixing:** local MCMC moves need many steps to traverse multimodal targets.

Contents

- 1 Overview
- 2 Energy-Based Models
- 3 Score Matching
- 4 Denoising Score Matching
- 5 Noise-Conditional Score Networks
- 6 Connection with Diffusion Models**

NCSN loss: Relationship with DDPM loss

For the variance-exploding corruption $Y = X + \sigma\varepsilon$,

$$s^*(y, \sigma) = \nabla_y \log p_\sigma(y) = -\frac{1}{\sigma} \mathbb{E}[\varepsilon \mid Y = y].$$

Thus

$$\varepsilon^*(y, \sigma) = \mathbb{E}[\varepsilon \mid y] = -\sigma s^*(y, \sigma).$$

For DDPM notation

$$X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon,$$

Tweedie's identity gives

$$\nabla_{x_t} \log p_t(x_t) = -\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \mathbb{E}[\varepsilon \mid X_t = x_t].$$

The two networks learn the same conditional information in different units.

NCSN and DDPM losses differ by scale and weighting

With $Y = X + \sigma\varepsilon$ and $\varepsilon_\phi = -\sigma s_\phi$,

$$\|s_\phi(Y, \sigma) + \varepsilon/\sigma\|^2 = \frac{1}{\sigma^2} \|\varepsilon_\phi(Y, \sigma) - \varepsilon\|^2.$$

Therefore a DSM weight $\lambda(\sigma) = \sigma^2$ produces the familiar unweighted noise-prediction MSE.

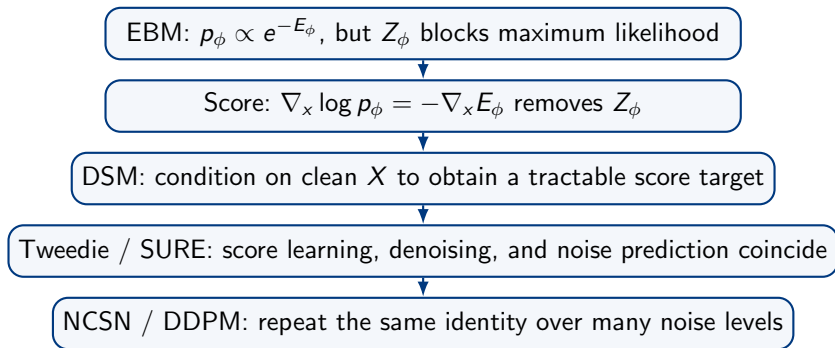
What is genuinely different?

- the forward noise schedule (variance exploding versus variance preserving),
- the generative sampler (annealed Langevin versus learned reverse transitions),
- time/noise weighting and discretization.

NCSN and DDPM

	NCSN	DDPM
$\mathbf{x}_{i+1} \mid \mathbf{x}_i$	Derive as $\mathbf{x}_{i+1} = \mathbf{x}_i + \sqrt{\sigma_{i+1}^2 - \sigma_i^2} \epsilon$	Define as $\mathbf{x}_{i+1} = \sqrt{1 - \beta_i} \mathbf{x}_i + \sqrt{\beta_i} \epsilon$
$\mathbf{x}_i \mid \mathbf{x}$	Define as $\mathbf{x}_i = \mathbf{x} + \sigma_i \epsilon$	Derive as $\mathbf{x}_i = \bar{\alpha}_i \mathbf{x} + \sqrt{1 - \bar{\alpha}_i^2} \epsilon$
p_{prior}	$\mathcal{N}(\mathbf{0}, \sigma_L^2 \mathbf{I})$	$\mathcal{N}(\mathbf{0}, \mathbf{I})$
Loss	$\mathbb{E}_i \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\ s_\phi(\mathbf{x}_i, \sigma_i) + \frac{\epsilon}{\sigma_i} \right\ _2^2 \right]$	$\mathbb{E}_i \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\ \epsilon_\phi(\mathbf{x}_i, i) - \epsilon \right\ _2^2 \right]$
Sampling	Apply Langevin per layer; use output to initialize the next	Traversing the Markovian chain with $p_{\phi^\times}(\mathbf{x}_{i-1} \mid \mathbf{x}_i)$

Conclusion



Takeaway

A diffusion model makes score estimation identifiable, tractable, and usable for generation by organizing a sequence of conditional regressions.

References and source map I

- [1] *Score-Based Perspective: From EBMs to NCSN*, Chapter 3 and Supplement D.2, PDF supplied for this seminar.
- [2] A. Hyvärinen (2005). Estimation of non-normalized statistical models by score matching.
- [3] P. Vincent (2011). A connection between score matching and denoising autoencoders.
- [4] Y. Song et al. (2020). Sliced score matching: A scalable approach to density and score estimation.
- [5] Y. Song and S. Ermon (2019). Generative modeling by estimating gradients of the data distribution.
- [6] B. Efron (2011). Tweedie's formula and selection bias.
- [7] J. Ho, A. Jain, and P. Abbeel (2020). Denoising diffusion probabilistic models.
- [8] J. Deasy, N. Simidjievski, and P. Liò (2022). Heavy-tailed denoising score matching.
- [9] T. Salimans and J. Ho (2021). Should EBMs model the energy or the score?

References and source map II

- [10] C.-H. Chao, W.-F. Sun, B.-W. Cheng, and C.-Y. Lee (2023). On investigating the conservative property of score-based generative models.
- [11] T. Karras, M. Aittala, T. Aila, and S. Laine (2022). Elucidating the design space of diffusion-based generative models.
- [12] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling (2022). Equivariant diffusion for molecule generation in 3D.
- [13] C.-H. Lai, Y. Takida, N. Murata, T. Uesaka, Y. Mitsufuji, and S. Ermon (2023). FP-Diffusion: Improving score-based diffusion models by enforcing the underlying score Fokker–Planck equation.

References and source map III

Where the proofs appear in the supplied PDF

- Proposition 3.2.1 and proof: Sections 3.2.1 and D.2.1.
- Theorem 3.3.1 and original proof: Sections 3.3.2 and D.2.2.
- Tweedie's formula and proof: Sections 3.3.4 and D.2.3.
- SURE and Stein proof: Sections 3.3.5 and D.2.4.
- Generalized score matching: Section 3.3.6.
- NCSN and comparison with DDPM: Sections 3.4–3.5.