

Lectures 16–17: Diffusion Models

October 19 and 21, 2026

Prof. Noah Golowich

In these notes we will explain how diffusion models generate data by reversing a noising process, and how consistency models learn to perform this sampling in one step. We generally take an informal approach, forgoing a completely rigorous treatment of stochastic calculus in preference for more intuitive explanations. For a more formal treatment, see [2].

1 Stochastic calculus and the evolution of densities

A *stochastic differential equation* (SDE) describes how a single random particle moves. Diffusion models use an SDE to interpolate between data and noise. In this section, we review some basics regarding SDEs.

1.1 Brownian motion and stochastic differential equations

We start by building a continuous-time source of noise. Write I_d for the $d \times d$ identity matrix and $\mathcal{N}(m, \Sigma)$ for the Gaussian distribution with mean m and covariance Σ . Fix a step size $h > 0$, draw independent $Z_0, Z_1, \dots \sim \mathcal{N}(0, I_d)$, and define a random walk on the grid $\{0, h, 2h, \dots\}$ by

$$B_0^{(h)} = 0, \quad B_{(k+1)h}^{(h)} = B_{kh}^{(h)} + \sqrt{h} Z_k,$$

interpolating linearly between grid points. Up to time t the walk takes t/h steps, each with variance h in every coordinate, so it accumulates variance $(t/h) \cdot h = t$ no matter how small h is. (With steps hZ_k instead, the variance would be th , which vanishes as $h \rightarrow 0$.) As $h \rightarrow 0$, the random paths $B^{(h)}$ converge in distribution to Brownian motion, viewed as random continuous functions on any bounded time interval with the topology of uniform convergence.

Definition 1.1 (Brownian motion). A standard d -dimensional *Brownian motion* $(B_t)_{t \geq 0}$ is a random process with $B_0 = 0$, continuous sample paths, and independent Gaussian increments: for $0 \leq s < t$,

$$B_t - B_s \sim \mathcal{N}(0, (t - s)I_d),$$

and increments on disjoint time intervals are independent.

Some additional work (which we omit) is needed to show that there exists a limiting process with continuous sample paths.

Stochastic differential equations. An SDE combines a deterministic velocity field with Brownian noise. Fix a time horizon $T > 0$, a *drift* $f : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, and a deterministic scalar *noise coefficient* $g : [0, T] \rightarrow [0, \infty)$, and consider the SDE

$$dX_t = f(t, X_t) dt + g(t) dB_t. \tag{1.1}$$

The simplest way to understand (1.1) is through its *Euler–Maruyama discretization*. Start from an initial point X_0 drawn independently of the noise, and repeatedly update

$$X_{t+h}^{(h)} = X_t^{(h)} + hf(t, X_t^{(h)}) + g(t)\sqrt{h}Z_t, \quad Z_t \sim \mathcal{N}(0, I_d), \quad (1.2)$$

using a fresh Gaussian Z_t at each grid time. In one step, the drift moves the particle a distance of order h , while the noise moves it a distance of order $\sqrt{h}$, with covariance $hg(t)^2 I_d$. So over very short times the noise dominates, while over a time interval of fixed length both terms matter. Under standard conditions (for example, f, g continuous, and f Lipschitz in x uniformly in t), (1.1) has a unique solution, and the Euler–Maruyama approximations converge to it in distribution as $h \rightarrow 0$.

Formally, (1.1) is shorthand for the integral equation

$$X_t = X_0 + \int_0^t f(u, X_u) du + \int_0^t g(u) dB_u. \quad (1.3)$$

The first integral is an ordinary time integral. Since g is deterministic and continuous, we can define the second as the mean-square limit of the sums

$$\sum_k g(t_k)(B_{t_{k+1}} - B_{t_k})$$

over partitions $0 = t_0 < t_1 < \dots < t_n = t$, as the largest gap $\max_k(t_{k+1} - t_k)$ tends to zero. Each such sum is a linear combination of the independent increments $B_{t_{k+1}} - B_{t_k} \sim \mathcal{N}(0, (t_{k+1} - t_k)I_d)$, so it is Gaussian with mean zero and covariance $\sum_k g(t_k)^2(t_{k+1} - t_k)I_d$. Passing to the limit gives

$$\int_0^t g(u) dB_u \sim \mathcal{N}\left(0, \left(\int_0^t g(u)^2 du\right) I_d\right). \quad (1.4)$$

This integral depends only on the Brownian motion, so it is independent of X_0 . We will use (1.4) repeatedly to solve linear SDEs in closed form.

Remark (Continuous local martingales). Brownian motion is not the only process that can drive an SDE; indeed, one can define SDEs with respect to *continuous local martingales*; see [2].

1.2 The Fokker–Planck equation

We now pass from individual paths to densities. Suppose X_t has a smooth density $p_t(x)$, so that $\mathbb{P}(X_t \in A) = \int_A p_t(x) dx$. All derivatives below are with respect to the spatial variable x : for a scalar function u and a vector field $v = (v_1, \dots, v_d)$, the *gradient*, *divergence*, and *Laplacian* are

$$\nabla u = (\partial_{x_1} u, \dots, \partial_{x_d} u), \quad \nabla \cdot v = \sum_{i=1}^d \partial_{x_i} v_i, \quad \Delta u = \nabla \cdot \nabla u = \sum_{i=1}^d \partial_{x_i}^2 u.$$

The divergence of v measures the net outward flow of v per unit volume. We write $v \cdot w$ for the Euclidean inner product and $\|v\|$ for the Euclidean norm.

The *Fokker–Planck equation* for (1.1) is

$$\partial_t p_t(x) = -\nabla \cdot (f(t, x)p_t(x)) + \frac{g(t)^2}{2} \Delta p_t(x). \quad (1.5)$$

To derive it, we think of probability as a fluid. The *probability flux* $J_t(x) \in \mathbb{R}^d$ is the vector in the direction of the net flow at x at time t , whose magnitude is the total amount of probability flow at x

(i.e., the total amount of probability mass crossing a small surface perpendicular to the vector, divided by the area of the surface). The drift and the noise each contribute to the flux:

$$J_t^{\text{drift}} = f(t, \cdot) p_t, \quad J_t^{\text{noise}} = -\frac{g(t)^2}{2} \nabla p_t, \quad J_t = f(t, \cdot) p_t - \frac{g(t)^2}{2} \nabla p_t. \quad (1.6)$$

The drift term is density times velocity, as for any fluid moving with velocity field f . The noise term points from higher to lower density, even though an individual Brownian particle has no preferred direction.

Why does unbiased noise produce a flux down the density gradient? Consider a random walk in one dimension on a grid with spacing ℓ . In each time interval of length h , every particle independently jumps one site to the left or to the right, with probability $1/2$ each. The probability masses at two adjacent sites x and $x + \ell$ are approximately $p_t(x)\ell$ and $p_t(x + \ell)\ell$. Half of the mass at x jumps right and half of the mass at $x + \ell$ jumps left, so the net rightward mass crossing the midpoint $x + \ell/2$ per unit time is

$$J_t^{\text{noise}}(x + \ell/2) = \frac{\ell}{2h} (p_t(x) - p_t(x + \ell)) = -\frac{\ell^2}{2h} \frac{p_t(x + \ell) - p_t(x)}{\ell}.$$

To match the SDE with g frozen at time t , the variance ℓ^2 of one jump should equal the variance $g(t)^2 h$ of the SDE increment over time h ; that is, we let $\ell, h \rightarrow 0$ with $\ell^2/(2h) \rightarrow g(t)^2/2$. In this limit the difference quotient becomes a derivative, and the flux tends to $-(g(t)^2/2) \partial_x p_t(x)$. In d dimensions, the same calculation along each coordinate, each with variance rate $g(t)^2$, gives the noise flux in (1.6).

Conservation of probability. The probability mass inside a region changes only by flowing across the region's boundary. Let $A \subset \mathbb{R}^d$ be a bounded region with smooth boundary ∂A and outward unit normal $n(x)$, and write dS for the surface-area element. The rate at which mass leaves A is the total outward flux through ∂A , so

$$\frac{d}{dt} \int_A p_t(x) dx = - \int_{\partial A} J_t(x) \cdot n(x) dS(x) = - \int_A \nabla \cdot J_t(x) dx,$$

where the last equality is the divergence theorem. Since this holds for every such region A and both integrands are continuous, the integrands agree, which gives the *continuity equation*

$$\partial_t p_t = -\nabla \cdot J_t. \quad (1.7)$$

Substituting the flux (1.6), and using that $g(t)$ does not depend on x so that $\nabla \cdot ((g(t)^2/2) \nabla p_t) = (g(t)^2/2) \Delta p_t$, gives (1.5).

Conservation in one dimension. In one dimension we can see this bookkeeping directly. Take the interval $[x - \Delta x, x]$, and let $J_t > 0$ mean flow to the right. Mass enters through the left endpoint at rate $J_t(x - \Delta x)$ and leaves through the right endpoint at rate $J_t(x)$, so the average density on the interval changes at rate

$$\frac{d}{dt} \left(\frac{1}{\Delta x} \int_{x-\Delta x}^x p_t(y) dy \right) = \frac{J_t(x - \Delta x) - J_t(x)}{\Delta x} \rightarrow -\partial_x J_t(x).$$

As $\Delta x \rightarrow 0$, the left side tends to $\partial_t p_t(x)$, which recovers (1.7) in one dimension. For example, if $J_t(y) = cy$ near x with $c > 0$, then outflow exceeds inflow by $c\Delta x$, and the density decreases at rate c . In Appendix A we derive (1.5) directly from the SDE more rigorously.

2 Diffusion models and time reversal

Let P_{data} be a distribution on $\mathbb{R}^d$ from which we have training samples. In these notes, a *diffusion process* is the solution of an SDE of the form (1.1). A *diffusion generative model* uses such a process to gradually turn data into noise, and then learns to run the process in reverse [5, 6, 7].

Concretely, we start the forward SDE at $X_0 \sim P_{\text{data}}$ and choose f and g so that the law p_T of X_T is close to a distribution p_{noise} that is easy to sample, typically a Gaussian. Running the forward process is easy, but it destroys the information in the data. Running it backward would do the opposite: the time-reversed process

$$Y_s = X_{T-s}, \quad 0 \leq s \leq T,$$

starts at $Y_0 \sim p_T \approx p_{\text{noise}}$ and ends at $Y_T \sim P_{\text{data}}$. Throughout, s denotes reverse time.

As we will see, the reverse process depends on the unknown data distribution only through the *score* of the forward marginal density,

$$\nabla_x \log p_t(x) = \frac{\nabla_x p_t(x)}{p_t(x)}. \quad (2.1)$$

At each point x , the score points in the direction in which the log density increases fastest.

Theorem 2.1 (Time reversal for state-independent noise). *Consider (1.1), suppose that f, g , and p_0 are smooth and bounded. Then for $0 \leq s \leq T$, the time reversal $Y_s = X_{T-s}$ satisfies*

$$dY_s = [-f(T-s, Y_s) + g(T-s)^2 \nabla \log p_{T-s}(Y_s)]ds + g(T-s) d\bar{B}_s, \quad Y_0 \sim p_T, \quad (2.2)$$

where $\bar{B}$ is a Brownian motion whose increments after reverse time s are independent of $(Y_r)_{r \leq s}$. In particular, $Y_s \sim p_{T-s}$.

We prove Theorem 2.1 in Appendix B; below we give an informal argument which explains why the marginals of the reverse process eq. (2.2) align with those of the forward process.

Guessing the reverse drift. The Fokker–Planck equation already suggests what the reverse drift should be. Write $q_s = p_{T-s}$ for the density of Y_s . Since $\partial_s q_s = -\partial_t p_t|_{t=T-s}$, flipping the sign of (1.5) gives

$$\partial_s q_s(x) = \nabla \cdot (f(T-s, x)q_s(x)) - \frac{g(T-s)^2}{2} \Delta q_s(x).$$

If the reversed process is again an SDE, with the same noise coefficient $g(T-s)$ and some drift $b_s(x)$, then its Fokker–Planck equation reads

$$\partial_s q_s(x) = -\nabla \cdot (b_s(x)q_s(x)) + \frac{g(T-s)^2}{2} \Delta q_s(x).$$

The choice $b_s(x) = -f(T-s, x) + g(T-s)^2 \nabla \log q_s(x)$ makes the two right-hand sides agree. Indeed, since $q_s(x) \nabla \log q_s(x) = \nabla q_s(x)$, the score part of b_s contributes $-\nabla \cdot (g(T-s)^2 \nabla q_s(x)) = -g(T-s)^2 \Delta q_s(x)$, and adding the diffusion term $\frac{g(T-s)^2}{2} \Delta q_s(x)$ leaves $-\frac{g(T-s)^2}{2} \Delta q_s(x)$, as required. Matching densities only shows that each Y_s has the correct distribution. Two processes can have the same marginal distributions at every time but different joint distributions across times, so the proof in Appendix B also checks the transition probabilities of the reversed process.

3 Examples of forward and backward diffusions

We now study three forward processes, each with a drift that is linear in x . For such processes the conditional distribution of X_t given a clean sample X_0 is an explicit Gaussian, so we can noise a training example to any time t in one step, and (as Section 4 shows) the training targets have a closed form. For each example below, we solve the forward SDE and then write down the reverse SDE from Theorem 2.1. The reversal is exact when $Y_0 \sim p_T$; in practice we start it from the Gaussian that p_T approaches when T is large.

3.1 Ornstein–Uhlenbeck (OU) process

Fix a mean $m \in \mathbb{R}^d$, a mean-reversion rate $\kappa > 0$, and a noise level $\sigma > 0$. The OU process is

$$dX_t = \kappa(m - X_t) dt + \sigma dB_t. \quad (3.1)$$

The drift pulls X_t toward m at rate κ , while the noise spreads it out. To solve (3.1), multiply $X_t - m$ by the integrating factor $e^{\kappa t}$. Because $e^{\kappa t}$ is deterministic and $e^{\kappa t}(X_t - m)$ is linear in X_t , the ordinary product rule applies. (This is Itô's formula (A.2) for a function that is linear in x , so that the second-derivative term vanishes.) We obtain

$$d(e^{\kappa t}(X_t - m)) = \kappa e^{\kappa t}(X_t - m) dt + e^{\kappa t} dX_t = \sigma e^{\kappa t} dB_t.$$

Integrating from 0 to t and multiplying by $e^{-\kappa t}$ gives

$$X_t = m + e^{-\kappa t}(X_0 - m) + \sigma \int_0^t e^{-\kappa(t-u)} dB_u.$$

By (1.4), the last term is a Gaussian independent of X_0 , with mean zero and covariance

$$\sigma^2 \int_0^t e^{-2\kappa(t-u)} du I_d = \frac{\sigma^2}{2\kappa} (1 - e^{-2\kappa t}) I_d.$$

Consequently,

$$X_t \mid X_0 = x_0 \sim \mathcal{N}\left(m + e^{-\kappa t}(x_0 - m), \frac{\sigma^2}{2\kappa} (1 - e^{-2\kappa t}) I_d\right). \quad (3.2)$$

As $t \rightarrow \infty$, the mean forgets x_0 exponentially fast and the covariance increases to $\frac{\sigma^2}{2\kappa} I_d$, so X_t converges in distribution to $\mathcal{N}(m, \sigma^2 I_d / (2\kappa))$ from any initial distribution. This limit is the *stationary distribution* of the OU process. Since $-f(t, y) = \kappa(y - m)$ and $g \equiv \sigma$, Theorem 2.1 gives the reverse SDE

$$dY_s = [\kappa(Y_s - m) + \sigma^2 \nabla \log p_{T-s}(Y_s)] ds + \sigma d\bar{B}_s, \quad (3.3)$$

which we start from the stationary distribution $\mathcal{N}(m, \sigma^2 I_d / (2\kappa))$. A common normalization is $m = 0$, $\kappa = 1$, and $\sigma = \sqrt{2}$, for which the stationary distribution is $\mathcal{N}(0, I_d)$.

3.2 Variance exploding (VE) process

The variance exploding (VE) process keeps the original signal and adds Gaussian noise of increasing size [7]. It has zero drift, $dX_t = g(t) dB_t$, so by (1.4) the accumulated noise $X_t - X_0$ is Gaussian with covariance $\sigma(t)^2 I_d$, where $\sigma(t)^2 = \int_0^t g(u)^2 du$. Equivalently, we can specify the desired noise scale $\sigma(t)$, with $\sigma(0) = 0$, and set $g(t)^2 = \frac{d}{dt} \sigma(t)^2$. We take $\sigma(t) = t$, so that $g(t) = \sqrt{2t}$ and

$$dX_t = \sqrt{2t} dB_t, \quad X_t = X_0 + \int_0^t \sqrt{2u} dB_u. \quad (3.4)$$

The integral has covariance $(\int_0^t 2u du)I_d = t^2 I_d$, so

$$X_t \mid X_0 = x_0 \sim \mathcal{N}(x_0, t^2 I_d). \quad (3.5)$$

In particular, for each fixed t we can sample X_t as $X_0 + tZ$, where $Z \sim \mathcal{N}(0, I_d)$ is independent of X_0 ; that is, $X_t \stackrel{d}{=} X_0 + tZ$. (Note that the process $t \mapsto X_0 + tZ$ with a single Z moves along a straight line and is not the VE process.)

Unlike the OU process, the VE process has no stationary distribution. Instead, when T is large compared with the scale of the data, $X_T/T \stackrel{d}{=} X_0/T + Z$ is close to $\mathcal{N}(0, I_d)$, so we take $p_{\text{noise}} = \mathcal{N}(0, T^2 I_d)$. The reverse SDE is

$$dY_s = 2(T - s)\nabla \log p_{T-s}(Y_s) ds + \sqrt{2(T - s)} d\bar{B}_s. \quad (3.6)$$

It is easy to confuse two quantities here. The noise coefficient $g(t) = \sqrt{2t}$ is the rate at which noise is added at time t , while the accumulated noise level $\sigma(t) = t$ is the standard deviation of $X_t - X_0$.

3.3 Variance preserving (VP) process

The variance preserving (VP) process, a generalization of the OU process, shrinks the signal while it adds noise [7]. Choose a continuous, positive *noise schedule* $\beta(t)$ and let

$$dX_t = -\frac{1}{2}\beta(t)X_t dt + \sqrt{\beta(t)} dB_t. \quad (3.7)$$

This SDE is the continuous-time limit of the discrete noising step

$$X_{t+h} = \sqrt{1 - \beta(t)h} X_t + \sqrt{\beta(t)h} Z, \quad Z \sim \mathcal{N}(0, I_d),$$

which is the noising step of DDPM [6] with per-step noise variance $\beta(t)h$. Indeed, $\sqrt{1 - \beta(t)h} = 1 - \frac{1}{2}\beta(t)h + O(h^2)$, so up to $O(h^2)$ this step is the Euler–Maruyama step (1.2) for (3.7).

To solve (3.7), let $A(t) = \int_0^t \beta(u) du$ and $\alpha(t) = e^{-A(t)/2}$, and use the integrating factor $e^{A(t)/2}$. As for the OU process, the function $x \mapsto e^{A(t)/2}x$ is linear, so Itô's formula has no second-derivative term and the product rule gives

$$d(e^{A(t)/2}X_t) = \frac{1}{2}\beta(t)e^{A(t)/2}X_t dt + e^{A(t)/2}dX_t = e^{A(t)/2}\sqrt{\beta(t)}dB_t.$$

Integrating from 0 to t and multiplying by $e^{-A(t)/2} = \alpha(t)$ yields

$$X_t = \alpha(t)X_0 + e^{-A(t)/2} \int_0^t e^{A(u)/2} \sqrt{\beta(u)} dB_u.$$

By (1.4), the noise term is Gaussian and independent of X_0 , with covariance

$$e^{-A(t)} \int_0^t e^{A(u)} \beta(u) du I_d = e^{-A(t)} (e^{A(t)} - 1) I_d = (1 - \alpha(t)^2) I_d,$$

where we used $\frac{d}{du} e^{A(u)} = \beta(u) e^{A(u)}$. Therefore

$$X_t \mid X_0 = x_0 \sim \mathcal{N}(\alpha(t)x_0, (1 - \alpha(t)^2)I_d). \quad (3.8)$$

If X_0 has finite covariance, then

$$\text{Cov}(X_t) = \alpha(t)^2 \text{Cov}(X_0) + (1 - \alpha(t)^2)I_d.$$

So if the data are normalized to have $\text{Cov}(X_0) = I_d$, the covariance stays equal to I_d for all t , which is where the name comes from.

When $A(T)$ is large, $\alpha(T)$ is small, so $X_T \stackrel{d}{=} \alpha(T)X_0 + \sqrt{1 - \alpha(T)^2} Z$ is close to $Z \sim \mathcal{N}(0, I_d)$, and we take $p_{\text{noise}} = \mathcal{N}(0, I_d)$. A common choice is the linear schedule $\beta(t) = \beta_{\min} + (t/T)(\beta_{\max} - \beta_{\min})$ with $0 < \beta_{\min} < \beta_{\max}$. The reverse SDE is

$$dY_s = \left[\frac{1}{2}\beta(T-s)Y_s + \beta(T-s)\nabla \log p_{T-s}(Y_s) \right] ds + \sqrt{\beta(T-s)} d\bar{B}_s. \quad (3.9)$$

When β is constant, (3.7) is the OU process (3.1) with $m = 0$, $\kappa = \beta/2$, and $\sigma = \sqrt{\beta}$; the choice $\beta = 2$ gives the normalization from the OU example.

4 Score matching

In the reverse SDE (2.2) we choose f and g ourselves, and the only unknown is the score $\nabla \log p_t(x)$, which depends on the data distribution. We approximate it by a function $s_\theta(t, x) \in \mathbb{R}^d$, typically a neural network with parameters θ that takes both the time and the noisy point as input. Substituting $s_\theta(t, x)$ for $\nabla \log p_t(x)$ and p_{noise} for p_T gives the learned reverse model

$$dY_s = [-f(T-s, Y_s) + g(T-s)^2 s_\theta(T-s, Y_s)] ds + g(T-s) d\bar{B}_s, \quad Y_0 \sim p_{\text{noise}}. \quad (4.1)$$

The question is how to train s_θ using only samples from P_{data} .

4.1 An unavailable regression target

The most direct approach is least-squares regression onto the true score.

Definition 4.1 (Explicit score matching). At a fixed time $t > 0$, the population least-squares objective for estimating the score is

$$J_{\text{ESM},t}(\theta) := \mathbb{E}_{X_t \sim p_t} \|s_\theta(t, X_t) - \nabla \log p_t(X_t)\|^2. \quad (4.2)$$

To learn the score at all noise levels, typically we choose a positive weight $\lambda(t)$ and minimize

$$\tilde{J}_{\text{ESM}}(\theta) := \mathbb{E}_{t \sim \text{Unif}(0,T)} [\lambda(t) J_{\text{ESM},t}(\theta)] = \frac{1}{T} \int_0^T \lambda(t) J_{\text{ESM},t}(\theta) dt. \quad (4.3)$$

The inputs to this regression are easy to sample: draw a training example X_0 and noise it to time t . Unfortunately, we cannot evaluate (4.2), because the label $\nabla \log p_t(X_t)$ is unknown; samples from a distribution do not tell us its density or the gradient of its density. In our examples each conditional law $X_t | X_0$ is Gaussian, but the marginal p_t is a mixture of these Gaussians over the unknown data distribution. We could replace P_{data} by the empirical distribution of the training set and compute the score of the resulting Gaussian mixture exactly. However, each label would then require a sum over the entire training set, and we would be fitting the score of the noised empirical distribution instead of the population score.

4.2 Denoising score matching

Although the marginal score is unknown, the *conditional* score is available, because we chose the noising process. Write $r_t(x | x_0)$ for the density of X_t given $X_0 = x_0$. The conditional score $\nabla_x \log r_t(x | x_0)$ depends on the clean example x_0 as well as on the noisy point x . When we plug in random points,

$\nabla_x \log r_t(X_t \mid X_0)$ denotes the gradient of $x \mapsto \log r_t(x \mid X_0)$ evaluated at $x = X_t$. The marginal density is a mixture of the conditional densities:

$$p_t(x) = \int r_t(x \mid x_0) P_{\text{data}}(dx_0). \quad (4.4)$$

The next lemma shows that averaging the conditional score over the clean examples consistent with a noisy point recovers the marginal score.

Lemma 4.2 (Conditional-score identity). *Fix $t > 0$. Suppose r_t is differentiable in x , and $p_t(x) > 0$. Then*

$$\mathbb{E}[\nabla_x \log r_t(X_t \mid X_0) \mid X_t = x] = \nabla \log p_t(x). \quad (4.5)$$

Proof. By Bayes' rule, the conditional distribution of X_0 given $X_t = x$ is $r_t(x \mid x_0)P_{\text{data}}(dx_0)/p_t(x)$. Thus

$$\begin{aligned} \mathbb{E}[\nabla_x \log r_t(X_t \mid X_0) \mid X_t = x] &= \frac{1}{p_t(x)} \int r_t(x \mid x_0) \nabla_x \log r_t(x \mid x_0) P_{\text{data}}(dx_0) \\ &= \frac{1}{p_t(x)} \int \nabla_x r_t(x \mid x_0) P_{\text{data}}(dx_0) = \frac{\nabla p_t(x)}{p_t(x)}. \end{aligned}$$

The second equality uses $r \nabla \log r = \nabla r$, and the third passes the gradient outside the integral and applies (4.4). $\square$

Lemma 4.2 says that $\nabla_x \log r_t(X_t \mid X_0)$ is a noisy label for $\nabla \log p_t(X_t)$ whose conditional mean is exactly the true label. This suggests regressing s_θ onto the conditional score in place of the unknown marginal score, which gives the *denoising score matching* objective

$$\tilde{J}_{\text{DSM}}(\theta) := \mathbb{E}_{t \sim \text{Unif}(0,T), X_0 \sim P_{\text{data}}, X_t \sim r_t(\cdot \mid X_0)} \left[\lambda(t) \|s_\theta(t, X_t) - \nabla_x \log r_t(X_t \mid X_0)\|^2 \right]. \quad (4.6)$$

Here t and X_0 are drawn independently, and then X_t is drawn from $r_t(\cdot \mid X_0)$. Every quantity in (4.6) can be sampled or computed.

Theorem 4.3 (Denoising score matching [10]). *Assume the conditions of Lemma 4.2 for almost every t , and suppose the λ -weighted second moments of $\nabla_x \log r_t(X_t \mid X_0)$ and $s_\theta(t, X_t)$ are finite. Then*

$$\tilde{J}_{\text{DSM}}(\theta) = \tilde{J}_{\text{ESM}}(\theta) + C, \quad C := \mathbb{E} \left[\lambda(t) \|\nabla_x \log r_t(X_t \mid X_0) - \nabla \log p_t(X_t)\|^2 \right] \geq 0, \quad (4.7)$$

where C does not depend on θ . Consequently, the two objectives have the same minimizers (when minimizers exist), even when θ ranges over a restricted parameterized class of functions.

Proof. Fix t . Lemma 4.2 gives $\mathbb{E}[\nabla_x \log r_t(X_t \mid X_0) - \nabla \log p_t(X_t) \mid X_t] = 0$. Since $s_\theta(t, X_t) - \nabla \log p_t(X_t)$ is a function of X_t ,

$$\begin{aligned} &\mathbb{E}[(s_\theta(t, X_t) - \nabla \log p_t(X_t)) \cdot (\nabla_x \log r_t(X_t \mid X_0) - \nabla \log p_t(X_t))] \\ &= \mathbb{E}[(s_\theta(t, X_t) - \nabla \log p_t(X_t)) \cdot \mathbb{E}[\nabla_x \log r_t(X_t \mid X_0) - \nabla \log p_t(X_t) \mid X_t]] = 0. \end{aligned}$$

Now write

$$s_\theta(t, X_t) - \nabla_x \log r_t(X_t \mid X_0) = (s_\theta(t, X_t) - \nabla \log p_t(X_t)) - (\nabla_x \log r_t(X_t \mid X_0) - \nabla \log p_t(X_t))$$

and expand the squared norm. The cross term has mean zero by the above computation, so

$$\begin{aligned} \mathbb{E}\|s_\theta(t, X_t) - \nabla_x \log r_t(X_t \mid X_0)\|^2 &= \mathbb{E}\|s_\theta(t, X_t) - \nabla \log p_t(X_t)\|^2 \\ &\quad + \mathbb{E}\|\nabla_x \log r_t(X_t \mid X_0) - \nabla \log p_t(X_t)\|^2. \end{aligned} \quad (4.8)$$

Multiplying by $\lambda(t)$ and averaging over t turns the left side into $\tilde{J}_{\text{DSM}}(\theta)$, the first term on the right into $\tilde{J}_{\text{ESM}}(\theta)$, and the second into C , which does not involve θ . This proves (4.7). $\square$

In words, regressing onto a noisy label with the correct conditional mean has the same effect as regressing onto the conditional mean itself.

4.3 Gaussian targets and Tweedie's formula

In all three examples from Section 3, $X_t | X_0 \sim \mathcal{N}(\mu_t(X_0), \sigma_t^2 I_d)$ for $t > 0$, with the following conditional mean and variance:

Process	$\mu_t(x_0)$	σ_t^2
OU	$m + e^{-\kappa t}(x_0 - m)$	$\frac{\sigma^2}{2\kappa}(1 - e^{-2\kappa t})$
VE	x_0	t^2
VP	$\alpha(t)x_0$	$1 - \alpha(t)^2$

(For OU, the constant σ in the table is the coefficient in (3.1), which is different from σ_t .) Taking the gradient in x of the log of the Gaussian density (Appendix C) gives

$$\nabla_x \log r_t(X_t | X_0) = \frac{\mu_t(X_0) - X_t}{\sigma_t^2}. \quad (4.9)$$

Substituting this into (4.6) gives the computable training objective

$$\tilde{\mathcal{J}}_{\text{DSM}}(\theta) = \mathbb{E}_{t, X_0, X_t} \left[\lambda(t) \left\| s_\theta(t, X_t) - \frac{\mu_t(X_0) - X_t}{\sigma_t^2} \right\|^2 \right]. \quad (4.10)$$

The label points from the noisy sample X_t toward the conditional mean $\mu_t(X_0)$. For VE this is the clean example itself; for OU and VP it is a version of the clean example shrunk toward m and toward 0, respectively.

Remark. We remark that taking the conditional expectation of (4.9) given $X_t = x$ and applying Lemma 4.2 gives the following identity, often known as *Tweedie's formula* [11]:

$$\mathbb{E}[\mu_t(X_0) | X_t = x] = x + \sigma_t^2 \nabla \log p_t(x). \quad (4.11)$$

Tweedie's formula says that the score tells us how to denoise: a step of $\sigma_t^2 \nabla \log p_t(x)$ from x lands at the posterior mean of $\mu_t(X_0)$. Correspondingly, a score model s_θ defines a denoiser $D_\theta(t, x) := x + \sigma_t^2 s_\theta(t, x)$, which predicts $\mu_t(X_0)$ from $X_t = x$. Since $s_\theta(t, X_t) + (X_t - \mu_t(X_0))/\sigma_t^2 = (D_\theta(t, X_t) - \mu_t(X_0))/\sigma_t^2$, the loss (4.10) can be rewritten as

$$\mathbb{E} \left[\frac{\lambda(t)}{\sigma_t^4} \|D_\theta(t, X_t) - \mu_t(X_0)\|^2 \right].$$

So score matching is the same as training a denoiser by least squares, and the optimal denoiser is the conditional mean $\mathbb{E}[\mu_t(X_0) | X_t]$ (Appendix C). In all three examples, $\mu_t(x_0) = a_t x_0 + b_t$ with $a_t \neq 0$, so $(D_\theta(t, x) - b_t)/a_t$ predicts the clean example X_0 itself. For VE, $a_t = 1$ and $b_t = 0$, and D_θ predicts X_0 directly.

4.4 The weighted noise-prediction objective

The labels in (4.10) have very different sizes at different times. In training, we sample $X_t = \mu_t(X_0) + \sigma_t Z$ with $Z \sim \mathcal{N}(0, I_d)$ independent of X_0 , so by (4.9) the label is $\nabla_x \log r_t(X_t | X_0) = -Z/\sigma_t$, whose

expected squared norm $\mathbb{E}\|Z\|^2/\sigma_t^2 = d/\sigma_t^2$ blows up as $t \rightarrow 0$. With a constant weight, small times, where the labels are largest, would dominate the loss. A natural fix is to weight each time by the inverse of the expected squared norm of its label:

$$\lambda(t) = \sigma_t^2 \quad (\text{proportional to } 1/(d/\sigma_t^2) = \sigma_t^2/d). \quad (4.12)$$

We drop the factor $1/d$ (as it is independent of t). With this choice and $X_t = \mu_t(X_0) + \sigma_t Z$, (4.10) becomes

$$\begin{aligned} \tilde{J}_{\text{DSM}}(\theta) &= \mathbb{E}_{t, X_0, X_t} \left[\sigma_t^2 \left\| s_\theta(t, X_t) + \frac{X_t - \mu_t(X_0)}{\sigma_t} \right\|^2 \right] \\ &= \mathbb{E}_{\substack{t \sim \text{Unif}(0, T), X_0 \sim P_{\text{data}} \\ Z \sim \mathcal{N}(0, I_d)}} \left[\|\sigma_t s_\theta(t, \mu_t(X_0) + \sigma_t Z) + Z\|^2 \right], \end{aligned} \quad (4.13)$$

where t , X_0 , and Z are independent in the last expectation. Finally, if we parameterize the score as $s_\theta(t, x) = -\epsilon_\theta(t, x)/\sigma_t$, the loss becomes

$$\mathbb{E}\|\epsilon_\theta(t, \mu_t(X_0) + \sigma_t Z) - Z\|^2.$$

In words, we are training the network ϵ_θ to predict the noise Z that was added to the clean example.

Training. Each step of stochastic gradient descent on (4.13) does the following (in practice, averaged over a minibatch):

1. Sample a time $t \sim \text{Unif}(0, T)$, a training example X_0 , and noise $Z \sim \mathcal{N}(0, I_d)$.
2. Form the noisy example $X_t = \mu_t(X_0) + \sigma_t Z$.
3. Take a gradient step in θ on the loss $\|\epsilon_\theta(t, X_t) - Z\|^2$.

The network sees only t and X_t , not Z , so it cannot recover Z exactly. At each time t , the best it can do in mean square is the conditional mean $\mathbb{E}[Z \mid X_t = x]$. Since $Z = (X_t - \mu_t(X_0))/\sigma_t$, Tweedie's formula (4.11) gives $\mathbb{E}[Z \mid X_t = x] = -\sigma_t \nabla \log p_t(x)$, so the optimal noise predictor corresponds exactly to the true score.

Sampling. To generate a new sample, draw $Y_0 \sim p_{\text{noise}}$ and simulate the learned reverse SDE (4.1) with the Euler–Maruyama scheme. At reverse time s , a step of size h is

$$Y_{s+h} = Y_s + h[-f(T-s, Y_s) + g(T-s)^2 s_\theta(T-s, Y_s)] + g(T-s)\sqrt{h} Z_s,$$

where $Z_s \sim \mathcal{N}(0, I_d)$ is drawn fresh at each step. Each step evaluates the score at the forward time $T-s > 0$. Because the score can be singular at forward time 0 (for instance, when P_{data} has no density), one usually stops at $s = T - \varepsilon$ and outputs the denoiser's prediction of the clean example from $Y_{T-\varepsilon}$. The output distribution differs from P_{data} because of three approximations: $s_\theta(t, x)$ in place of $\nabla \log p_t(x)$, p_{noise} in place of p_T , and the time discretization. With the exact score, exact initialization, and exact continuous-time simulation, Theorem 2.1 shows that Y_T has distribution P_{data} .

5 Probability flow ODEs and consistency models

The reverse SDE generates a sample by making many updates, each using a score evaluation and fresh noise. We now show that a deterministic ODE can produce the same marginal distributions.

5.1 The probability flow ODE

Recall from (1.6) that the forward SDE has probability flux $f(t, \cdot)p_t - \frac{1}{2}g(t)^2\nabla p_t$. An ODE with velocity $v(t, x)$ has flux $v(t, x)p_t(x)$. Dividing the SDE's flux by its density therefore suggests the velocity

$$v(t, x) := f(t, x) - \frac{g(t)^2}{2}\nabla \log p_t(x). \quad (5.1)$$

The resulting *probability flow ODE* reproduces the density evolution of the SDE [7].

Theorem 5.1 (Probability flow ODE for state-independent noise). *Assume the conditions of Theorem 2.1 hold with the same regularity down to time 0, and let $(X_t)_{0 \leq t \leq T}$ solve (1.1). Let $\tilde{X}_t$ solve*

$$\frac{d\tilde{X}_t}{dt} = v(t, \tilde{X}_t) = f(t, \tilde{X}_t) - \frac{g(t)^2}{2}\nabla \log p_t(\tilde{X}_t), \quad \tilde{X}_0 \sim P_{\text{data}}. \quad (5.2)$$

Then $\tilde{X}_t$ and X_t have the same distribution for every $t \in [0, T]$.

Proof. We first check that p_t satisfies the continuity equation¹ for velocity v . Since $p_t \nabla \log p_t = \nabla p_t$ and $g(t)$ does not depend on x , the Fokker–Planck equation (1.5) gives

$$\partial_t p_t = -\nabla \cdot (f(t, \cdot)p_t) + \frac{g(t)^2}{2}\Delta p_t = -\nabla \cdot \left[\left(f(t, \cdot) - \frac{g(t)^2}{2}\nabla \log p_t \right) p_t \right] = -\nabla \cdot (v(t, \cdot)p_t). \quad (5.3)$$

Let $\tilde{p}_t$ denote the density of $\tilde{X}_t$. Applying the Fokker–Planck derivation in Appendix A with drift v and noise coefficient zero gives

$$\partial_t \tilde{p}_t = -\nabla \cdot (v(t, \cdot)\tilde{p}_t), \quad \tilde{p}_0 = p_0.$$

Thus p_t and $\tilde{p}_t$ solve the same continuity equation with the same initial density. The assumptions of Theorem 2.1 make v smooth, locally Lipschitz in x , and of at most linear growth on $[0, T]$: indeed, v is half the difference between the forward drift and the reverse drift evaluated at reverse time $T - t$. For such a velocity field, uniqueness for the continuity equation gives $\tilde{p}_t = p_t$ on $[0, T]$. $\square$

The conclusion concerns distributions at individual times. Once $\tilde{X}_0$ is fixed, its entire ODE trajectory is deterministic, whereas the SDE continues to introduce noise. Consequently, the two processes generally have different joint distributions across times. If P_{data} has no smooth density, we may apply the theorem on $[\varepsilon, T]$ for $\varepsilon > 0$.

5.2 Sampling with the reverse ODE

As before, use reverse time $s = T - t$. Writing $\tilde{Y}_s = \tilde{X}_{T-s}$ and changing the sign of the velocity gives

$$\frac{d\tilde{Y}_s}{ds} = -f(T - s, \tilde{Y}_s) + \frac{g(T - s)^2}{2}\nabla \log p_{T-s}(\tilde{Y}_s), \quad \tilde{Y}_0 \sim p_T. \quad (5.4)$$

The ODE flow is invertible, so Theorem 5.1 implies $\tilde{Y}_s \sim p_{T-s}$.

Compare (5.4) with the reverse SDE (2.2): the ODE removes the Brownian term and halves the coefficient of the score. Merely deleting the Brownian term from the reverse SDE would give the wrong

¹The *continuity equation* describes the flow of mass for an ODE, i.e., the Fokker–Planck equation with no Brownian motion term.

density evolution. In practice, we replace the score by s_θ and initialize with p_{noise} . An Euler step is then

$$\tilde{Y}_{s+h} = \tilde{Y}_s + h \left[-f(T-s, \tilde{Y}_s) + \frac{g(T-s)^2}{2} s_\theta(T-s, \tilde{Y}_s) \right]. \quad (5.5)$$

For the VE process $f = 0$, $g(t) = \sqrt{2t}$, the velocity in this update is $(T-s)s_\theta(T-s, \tilde{Y}_s)$. For VP it is $\frac{1}{2}\beta(T-s)(\tilde{Y}_s + s_\theta(T-s, \tilde{Y}_s))$. Sampling is deterministic conditional on the initial noise, but still requires evaluating the score at each numerical step.

A A derivation of the Fokker–Planck equation

This appendix derives the Fokker–Planck equation (1.5) directly from the SDE.

From random increments to the generator

The *generator* of the SDE at time t is the operator L_t that describes how the expected value of a function of X_t changes over a short time:

$$L_t \phi(x) = \lim_{h \downarrow 0} \frac{\mathbb{E}[\phi(X_{t+h}) \mid X_t = x] - \phi(x)}{h}.$$

We begin with an informal computation of $L_t \phi(x)$. Let us condition on $X_t = x$ and freeze the coefficients at (t, x) for a step of length h . The Euler–Maruyama increment is

$$\delta X = hf(t, x) + g(t)\sqrt{h}Z, \quad Z \sim \mathcal{N}(0, I_d),$$

with mean $\mathbb{E}[\delta X] = hf(t, x)$ and second moments

$$\mathbb{E}[\delta X_i \delta X_j] = hg(t)^2 \mathbf{1}_{\{i=j\}} + h^2 f_i(t, x) f_j(t, x).$$

For a smooth test function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ with bounded third derivatives, a second-order Taylor expansion gives

$$\phi(x + \delta X) - \phi(x) = \nabla \phi(x) \cdot \delta X + \frac{1}{2} \sum_{i,j} \partial_{x_i x_j} \phi(x) \delta X_i \delta X_j + R,$$

where $|R| \leq C \|\delta X\|^3$ for a constant C . We take expectations term by term. The first-order term contributes $hf(t, x) \cdot \nabla \phi(x)$. In the second-order term, the part $hg(t)^2 \mathbf{1}_{\{i=j\}}$ of $\mathbb{E}[\delta X_i \delta X_j]$ contributes $\frac{hg(t)^2}{2} \Delta \phi(x)$, and the part $h^2 f_i f_j$ contributes $O(h^2)$. Finally, $\mathbb{E}\|\delta X\|^3 = O(h^{3/2})$, because $\|\delta X\| \leq h\|f(t, x)\| + g(t)\sqrt{h}\|Z\|$ and $\mathbb{E}\|Z\|^3 < \infty$. Dividing by h and letting $h \downarrow 0$ gives

$$L_t \phi(x) = f(t, x) \cdot \nabla \phi(x) + \frac{g(t)^2}{2} \Delta \phi(x). \quad (\text{A.1})$$

Given eq. (A.1), we have that $\partial_t \int \phi(x) p_t(x) dx = \int L_t \phi(x) p_t(x) dx$ for all smooth test functions ϕ . Integrating by parts moves the derivatives from ϕ onto p_t , and since this holds for all ϕ , we conclude that p_t satisfies the Fokker–Planck equation (1.5).

Itô's formula and the weak density equation

Now we proceed with a formal derivation, using Itô's formula. For a function $F(t, x)$ that is continuously differentiable in t and twice continuously differentiable in x , Itô's formula states that $F(t, X_t)$ satisfies the SDE

$$dF(t, X_t) = \left(\partial_t F + f \cdot \nabla F + \frac{g(t)^2}{2} \Delta F \right) (t, X_t) dt + g(t) \nabla F(t, X_t) \cdot dB_t. \quad (\text{A.2})$$

Compared with the ordinary chain rule, the extra term $\frac{g(t)^2}{2} \Delta F$ comes from the squared Brownian increments in the Taylor expansion above.

Now let $\phi \in C_c^\infty(\mathbb{R}^d)$ be a time-independent test function, i.e., a smooth function that vanishes outside a bounded set. Apply (A.2) with $F(t, x) = \phi(x)$, integrate from 0 to t , and take expectations. The stochastic integral has mean zero, so $\mathbb{E}[\phi(X_t)] - \mathbb{E}[\phi(X_0)] = \int_0^t \mathbb{E}[\phi(X_t)] = \mathbb{E} \left[f(t, X_t) \cdot \nabla \phi(X_t) + \frac{g(t)^2}{2} \Delta \phi(X_t) \right] du$, and differentiating in t gives

$$\frac{d}{dt} \mathbb{E}[\phi(X_t)] = \mathbb{E} \left[f(t, X_t) \cdot \nabla \phi(X_t) + \frac{g(t)^2}{2} \Delta \phi(X_t) \right].$$

Writing both expectations as integrals against p_t , we obtain

$$\int \phi \partial_t p_t dx = \int f p_t \cdot \nabla \phi dx + \frac{g(t)^2}{2} \int p_t \Delta \phi dx.$$

Integrating the first term on the right by parts once and the second term twice (and using that ϕ vanishes outside a bounded set, so there are no boundary terms) gives:

$$\int f p_t \cdot \nabla \phi dx = - \int \phi \nabla \cdot (f p_t) dx, \quad \int p_t \Delta \phi dx = \int \phi \Delta p_t dx.$$

Consequently,

$$\int \phi(x) \left[\partial_t p_t(x) + \nabla \cdot (f(t, x) p_t(x)) - \frac{g(t)^2}{2} \Delta p_t(x) \right] dx = 0.$$

This holds for every $\phi \in C_c^\infty(\mathbb{R}^d)$, and the bracket is continuous under our smoothness assumptions, so the bracket vanishes everywhere. (If it were nonzero at some point, it would have a constant sign on a small ball around that point, and a nonnegative ϕ supported in that ball and not identically zero would make the integral nonzero.) This proves (1.5). If X_t has no smooth density, the identity for $\frac{d}{dt} \mathbb{E}[\phi(X_t)]$ still holds; it is called the *weak form* of the Fokker–Planck equation.

B Proof of the time reversal theorem

This appendix proves Theorem 2.1, in the notation of Section 2: X solves the forward SDE (1.1), p_t is the density of X_t , and $Y_s = X_{T-s}$ is its time reversal. We compute the transition densities of Y by Bayes' rule and show that they solve the Fokker–Planck equation of the reverse SDE (2.2). The uniqueness assumptions then identify Y with the solution of (2.2).

Proof of Theorem 2.1. The reverse transition density. For $u < v$, let $k_{u,v}(y, x)$ be the transition density of the forward process, i.e., the conditional density of X_v at x given $X_u = y$. Fix a reverse time s and a state x , and let $v = T - s$. For a later reverse time $r > s$, let $u = T - r < v$. Since $Y_s = X_v$ and $Y_r = X_u$, Bayes' rule gives the conditional density of Y_r at y given $Y_s = x$:

$$q_{s,r}(x, y) = \frac{p_u(y) k_{u,v}(y, x)}{p_v(x)}. \quad (\text{B.1})$$

The reversed process is also Markov, because the Markov property is symmetric in time: conditional on the present state X_v , the history before time v is independent of the future after time v . Hence the transition densities (B.1), together with the initial law p_T , determine the law of the reversed process.

Forward and backward equations. For smooth functions ϕ and ρ on $\mathbb{R}^d$, we consider the *generator* L_u of the forward process and its adjoint L_u^* as follows:

$$L_u \phi = f(u, \cdot) \cdot \nabla \phi + \frac{g(u)^2}{2} \Delta \phi, \quad L_u^* \rho = -\nabla \cdot (f(u, \cdot) \rho) + \frac{g(u)^2}{2} \Delta \rho.$$

The generator describes how the conditional expectation of ϕ changes over a short time,

$$L_u \phi(y) = \lim_{h \downarrow 0} \frac{\mathbb{E}[\phi(X_{u+h}) \mid X_u = y] - \phi(y)}{h},$$

as we showed in Appendix A. Integration by parts gives $\int \rho L_u \phi dx = \int \phi L_u^* \rho dx$ for compactly supported ϕ , and the Fokker–Planck equation (1.5) says that $\partial_u p_u = L_u^* p_u$. The transition density satisfies a second equation, in its initial variables: with the endpoint (v, x) fixed, it satisfies the *backward equation*

$$\partial_u k_{u,v}(y, x) = -L_u k_{u,v}(y, x),$$

where L_u acts on the variable y . To see this, condition on what happens in a short time $h > 0$ after u . By the Markov property,

$$k_{u,v}(y, x) = \mathbb{E}[k_{u+h,v}(X_{u+h}, x) \mid X_u = y].$$

Applying the definition of the generator to $\phi = k_{u+h,v}(\cdot, x)$, the right side equals $k_{u+h,v}(y, x) + h L_u k_{u+h,v}(y, x) + o(h)$. Expanding $k_{u+h,v}(y, x) = k_{u,v}(y, x) + h \partial_u k_{u,v}(y, x) + o(h)$, this is $k_{u,v}(y, x) + h(\partial_u + L_u)k_{u,v}(y, x) + o(h)$. Comparing with the left side shows that $(\partial_u + L_u)k_{u,v}(y, x) = 0$.

The reverse transition density solves the reverse Fokker–Planck equation. We show that, as a function of (r, y) , the density $q_{s,r}(x, y)$ satisfies the Fokker–Planck equation of (2.2). Throughout this step, ∇ and Δ act on the variable y , and $u = T - r$. The drift of (2.2) at reverse time r is

$$b_r(y) = -f(T - r, y) + g(T - r)^2 \nabla \log p_{T-r}(y) = -f(u, y) + g(u)^2 \nabla \log p_u(y).$$

The denominator $p_v(x)$ in (B.1) does not depend on r . Since $u = T - r$, we have $\partial_r = -\partial_u$, so the Fokker–Planck equation $\partial_u p_u = L_u^* p_u$ and the backward equation give

$$\begin{aligned} \partial_r(p_u(y) k_{u,v}(y, x)) &= -k_{u,v}(y, x) \partial_u p_u(y) - p_u(y) \partial_u k_{u,v}(y, x) \\ &= -k_{u,v}(y, x) L_u^* p_u(y) + p_u(y) L_u k_{u,v}(y, x). \end{aligned}$$

On the other hand, applying the reverse density operator to $p_u(y) k_{u,v}(y, x)$ and expanding products gives

$$\begin{aligned} & -\nabla \cdot (b_r(y) p_u(y) k_{u,v}(y, x)) + \frac{g(u)^2}{2} \Delta(p_u(y) k_{u,v}(y, x)) \\ &= \nabla \cdot (f(u, y) p_u(y) k_{u,v}(y, x)) - g(u)^2 \nabla \cdot (k_{u,v}(y, x) \nabla p_u(y)) + \frac{g(u)^2}{2} \Delta(p_u(y) k_{u,v}(y, x)) \\ &= k_{u,v}(y, x) \nabla \cdot (f(u, y) p_u(y)) + p_u(y) f(u, y) \cdot \nabla k_{u,v}(y, x) \\ &\quad - \frac{g(u)^2}{2} k_{u,v}(y, x) \Delta p_u(y) + \frac{g(u)^2}{2} p_u(y) \Delta k_{u,v}(y, x) \\ &= -k_{u,v}(y, x) L_u^* p_u(y) + p_u(y) L_u k_{u,v}(y, x). \end{aligned}$$

The first equality uses $b_r(y)p_u(y) = -f(u, y)p_u(y) + g(u)^2\nabla p_u(y)$. The second uses the product rules

$$\begin{aligned}\nabla \cdot (f(u, y)p_u(y)k_{u,v}(y, x)) &= k_{u,v}(y, x) \nabla \cdot (f(u, y)p_u(y)) + p_u(y)f(u, y) \cdot \nabla k_{u,v}(y, x), \\ \nabla \cdot (k_{u,v}(y, x)\nabla p_u(y)) &= \nabla k_{u,v}(y, x) \cdot \nabla p_u(y) + k_{u,v}(y, x)\Delta p_u(y), \\ \Delta(p_u(y)k_{u,v}(y, x)) &= k_{u,v}(y, x)\Delta p_u(y) + 2\nabla p_u(y) \cdot \nabla k_{u,v}(y, x) + p_u(y)\Delta k_{u,v}(y, x),\end{aligned}$$

after which the two terms involving $\nabla p_u(y) \cdot \nabla k_{u,v}(y, x)$ cancel. The third is the definition of L_u and L_u^* . Combining the two displays and dividing by $p_v(x)$ shows that

$$\partial_r q_{s,r}(x, y) = -\nabla \cdot (b_r(y)q_{s,r}(x, y)) + \frac{g(T-r)^2}{2}\Delta q_{s,r}(x, y),$$

which is the Fokker–Planck equation with drift b_r and noise coefficient $g(T-r)$. As $r \downarrow s$, $q_{s,r}(x, \cdot)$ converges to the point mass at x , because the reversed paths are continuous.

Identifying the reversed process. By assumption (i), the SDE (2.2) has a unique solution from every starting point, and its transition probabilities solve the same Fokker–Planck equation. By assumption (ii), this equation with point-mass initial data has only one solution, so $q_{s,r}$ is the transition density of (2.2). The reversed process and the solution of (2.2) are both Markov, with the same initial law p_T and the same transition probabilities, so all of their finite-dimensional distributions agree. This proves the theorem. Since Y has the same law as a solution of (2.2), it can be written in the form (2.2) for a suitable Brownian motion $\bar{B}$. $\square$

C The Gaussian score and Tweedie’s formula

Fix $t > 0$ and suppose that $X_t \mid X_0 = x_0$ has the Gaussian density

$$r_t(x \mid x_0) = (2\pi\sigma_t^2)^{-d/2} \exp\left(-\frac{\|x - \mu_t(x_0)\|^2}{2\sigma_t^2}\right).$$

The normalizing constant does not depend on x , so the gradient of the logarithm is

$$\nabla_x \log r_t(x \mid x_0) = -\frac{x - \mu_t(x_0)}{\sigma_t^2},$$

which is (4.9).

To obtain the marginal score, differentiate the mixture density (4.4) under the integral:

$$\begin{aligned}\nabla p_t(x) &= \int \nabla_x r_t(x \mid x_0) P_{\text{data}}(dx_0) \\ &= \int \frac{\mu_t(x_0) - x}{\sigma_t^2} r_t(x \mid x_0) P_{\text{data}}(dx_0).\end{aligned}$$

Divide by $p_t(x) > 0$. By Bayes’ rule, $r_t(x \mid x_0)P_{\text{data}}(dx_0)/p_t(x)$ is the conditional distribution of X_0 given $X_t = x$, so

$$\nabla \log p_t(x) = \frac{\mathbb{E}[\mu_t(X_0) \mid X_t = x] - x}{\sigma_t^2}.$$

Rearranging gives Tweedie’s formula (4.11). Here we assume that the conditional expectations exist; finite second moments of $\mu_t(X_0)$ also justify the least-squares interpretation below.

References

- [1] Wenpin Tang and Hanyang Zhao. *Score-based diffusion models via stochastic differential equations: A technical tutorial*. arXiv:2402.07487, 2024. <https://arxiv.org/abs/2402.07487>.
- [2] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. 6th edition, Springer, 2003.
- [3] Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- [4] Ulrich G. Haussmann and Étienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986.
- [5] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ICML*, 2015. <https://arxiv.org/abs/1503.03585>.
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. <https://arxiv.org/abs/2006.11239>.
- [7] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021. <https://arxiv.org/abs/2011.13456>.
- [8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. <https://arxiv.org/abs/2010.02502>.
- [9] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *ICML*, 2023. <https://arxiv.org/abs/2303.01469>.
- [10] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.
- [11] Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.