Score-Based Generative Models: Probability Flow ODE

Nathan Cantafio

April 23, 2026

The SDE framework for diffusion models admits a remarkable deterministic counterpart: for every Itô diffusion there exists an ordinary differential equation whose trajectories share the same time-marginals. This note reviews the *probability flow ODE* of Song et al. [1], deriving it from the Fokker–Planck equation and showcasing it on the VE, VP, and sub-VP families. We then follow Chen et al. [2] in applying the instantaneous change-of-variables formula to show how this deterministic flow enables exact likelihood computation: the log-likelihood decomposes into a tractable prior term and a time-integrated divergence of the ODE vector field, the latter approximable via Hutchinson's estimator without materializing the full Jacobian.

1. Introduction

The previous note developed denoising score matching as a principled way to learn the score function $\nabla_x \log p_t(x)$ across a continuum of noise scales. Given such a model, Song et al. [1] propose two routes to sampling. The first is the *reverse-time SDE*: run the time-reversed diffusion, which requires adding a score-dependent drift and integrating a stochastic equation backward. The second, and the subject of this note, is the *probability flow ODE*: an ordinary differential equation, derived from the same score, whose trajectories trace out the same distribution as the SDE at every time t.

The distinction matters for two reasons. First, deterministic ODE solvers are often faster and more controllable than SDE integrators: step sizes can be adapted, and there is no variance from injected noise. Second, and more fundamentally, the probability flow ODE is *invertible*: a data point x_0 maps to a unique "noised" x_T under the forward ODE, and the reverse ODE recovers x_0 exactly.

We proceed as follows. Section 2 recalls the SDE framework and the three standard noise schedules. Section 3 derives the probability flow ODE by identifying a family of ODEs that are consistent with the Fokker–Planck equation of the SDE. Section 4 applies the instantaneous change-of-variables formula to obtain a tractable likelihood estimate.

2. The SDE framework

2.1. The forward SDE

Song et al. [1] model diffusion as a solution to an Itô SDE on the time interval $[0, T]$:

$$dx = f(x,t)\,dt + g(t)\,dW, \tag{1}$$

where $f : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ is the *drift* coefficient, $g : [0,T] \to \mathbb{R}$ is the scalar *diffusion* coefficient, and W is a standard Brownian motion. $x_0 \sim p_0$ is the data distribution and we denote the "density of the data over time" by $p_t(x)$. The drift f and diffusion g are chosen so that p_T is approximately a standard Gaussian. The reverse-time SDE takes the form [3]:

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

where $\bar{W}$ is a Brownian motion running backward in time.

2.2. Standard noise schedules

Three families of SDEs are considered in Song et al. [1].

Variance Exploding (VE). Uses zero drift and a time-varying diffusion coefficient,

$$f(x,t) = 0, \qquad g(t) = \sqrt{\frac{d[\sigma(t)^2]}{dt}},$$

so the SDE reduces to $dx = \sqrt{d[\sigma(t)^2]/dt}\; dW$. The forward process is therefore $x_t = x_0 + \sigma(t)\, z$ with $z \sim \mathcal{N}(0, I)$, and the marginal variance grows without bound as $\sigma(t) \to \infty$, hence the name. The score of the marginal is $\nabla_x \log p_t(x) = -\mathbb{E}[(x - x_0)/\sigma(t)^2 \mid x_t = x]$.

Variance Preserving (VP). The continuous-time limit of the DDPM noise schedule [4],

$$f(x,t) = -\tfrac{1}{2}\beta(t)\, x, \qquad g(t) = \sqrt{\beta(t)},$$

for a positive noise schedule $\beta(t)$. The affine drift contracts the signal while the diffusion injects noise, keeping the variance bounded. The conditional distribution $p_{t|0}(x_t \mid x_0)$ is Gaussian with mean $e^{-\frac{1}{2}\int_0^t \beta(s)\, ds}\, x_0$ and variance $1 - e^{-\int_0^t \beta(s)\, ds}$, so the marginal variance is preserved at one when initialized from $p_0 = \mathcal{N}(0, I)$.

Sub-VP. A variant of the VP SDE with a tighter variance bound,

$$f(x,t) = -\tfrac{1}{2}\beta(t)\, x, \qquad g(t) = \sqrt{\beta(t)\big(1 - e^{-2\int_0^t \beta(s)\, ds}\big)}.$$

The sub-VP process stays closer to the data distribution early in the diffusion, which can improve score estimation at low noise levels.

3. The probability flow ODE

3.1. The Fokker–Planck equation

Associated to the SDE (1) is the *Fokker–Planck equation* (also called the forward Kolmogorov equation), which describes how the marginal density $p_t(x)$ evolves over time. For the scalar diffusion coefficient $g(t)$ in (1), the noise injected at each step is isotropic: it is proportional to the identity matrix $g(t)^2 I$. Under this assumption the Fokker–Planck equation takes the form

$$\frac{\partial p_t}{\partial t} = -\nabla_x \cdot \big[f(x,t)\, p_t\big] + \frac{g(t)^2}{2}\, \Delta_x p_t, \tag{3}$$

where $\nabla_x \cdot$ denotes the divergence and $\Delta_x = \nabla_x \cdot \nabla_x = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ is the Laplacian. The Laplacian arises here precisely because the diffusion is isotropic: the second-order term is $\frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}[D_{ij}\, p_t]$, where $D = g(t)^2 I$ is the diffusion tensor. Since D is diagonal and spatially constant, the off-diagonal mixed partials vanish identically and the sum collapses to $\frac{g(t)^2}{2}\Delta_x p_t$. In the non-isotropic case when the diffusion coefficient is a general matrix $G(x,t)$ and $D = GG^\top$ has off-diagonal entries or spatial dependence, the mixed partial terms do not vanish, and the Fokker–Planck equation must retain the full second-order operator.

The first term in (3) describes transport of probability mass by the drift f; the second describes its spreading by the injected noise. Any process, stochastic or deterministic, whose marginal density satisfies (3) will have the same time-marginals $\{p_t\}$ as the original SDE.

3.2. Deriving the probability flow ODE

Our goal is to find an ODE $dx = h(x,t)\,dt$ whose marginal density evolves identically to that of the SDE (1). The *continuity equation* governs how a density evolves under a deterministic flow with vector field h:

$$\frac{\partial p_t}{\partial t} = -\nabla_x \cdot \big[h(x,t)\, p_t\big]. \tag{4}$$

This is a first-order equation in p_t, whereas the Fokker–Planck equation (3) contains a second-order term $\frac{g(t)^2}{2}\Delta_x p_t$. If we can rewrite this Laplacian as a divergence of something proportional to p_t, we can fold it into the drift and recover the form of (4). The key step is:

$$\Delta_x p_t = \nabla_x \cdot (\nabla_x p_t) = \nabla_x \cdot \big[p_t\, \nabla_x \log p_t\big],$$

where the second equality uses $\nabla_x p_t = p_t\, \nabla_x \log p_t$. Substituting into (3) and combining both terms under a single divergence:

$$\begin{aligned}\frac{\partial p_t}{\partial t} &= -\nabla_x \cdot \big[f(x,t)\, p_t\big] + \frac{g(t)^2}{2}\, \nabla_x \cdot \big[p_t\, \nabla_x \log p_t\big] \\ &= -\nabla_x \cdot \left[\Big(f(x,t) - \frac{g(t)^2}{2}\, \nabla_x \log p_t(x)\Big)\, p_t\right].\end{aligned}$$

This is now exactly in the form of the continuity equation (4), with vector field:

$$dx = \left[f(x,t) - \frac{g(t)^2}{2}\, \nabla_x \log p_t(x)\right] dt. \tag{5}$$

This is the *probability flow ODE*. Any ODE with this vector field will induce the same evolution of p_t as the SDE (1), and therefore shares the same time-marginals for all $t \in [0,T]$.

Remark (non-isotropic diffusion). The argument extends to the general case where the scalar $g(t)$ is replaced by a matrix-valued diffusion coefficient $G(x,t) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. The SDE becomes $dx = f(x,t)\,dt + G(x,t)\,dW$, and the Fokker–Planck equation now involves the full diffusion tensor $D(x,t) = G(x,t)G(x,t)^\top$:

$$\frac{\partial p_t}{\partial t} = -\nabla_x \cdot \big[f(x,t)\, p_t\big] + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i\, \partial x_j}\big[D_{ij}(x,t)\, p_t\big].$$

The mixed partial terms no longer vanish in general. Applying the product rule and the identity $\nabla_x p_t = p_t \nabla_x \log p_t$ to each entry of the second-order operator, and collecting all terms under a single divergence, yields the corresponding probability flow ODE:

$$dx = \left[f(x,t) - \frac{1}{2}\big(\nabla_x \cdot D(x,t)\big) - \frac{1}{2} D(x,t)\, \nabla_x \log p_t(x)\right] dt,$$

where $[\nabla_x \cdot D]_i = \sum_j \partial D_{ij}/\partial x_j$ captures the spatial variation of G. When G depends only on t and is proportional to the identity (the isotropic scalar case), both $D = g(t)^2 I$ and its divergence vanish, recovering (5).

3.3. Instantiations for standard schedules

Substituting the drift and diffusion coefficients for each schedule into (5):

VE. With $f = 0$ and $g(t)^2 = d[\sigma(t)^2]/dt$:

$$dx = -\frac{1}{2}\frac{d[\sigma(t)^2]}{dt}\, \nabla_x \log p_t(x)\, dt.$$

VP. With $f(x,t) = -\frac{1}{2}\beta(t)\,x$ and $g(t)^2 = \beta(t)$:

$$dx = \left[-\frac{1}{2}\beta(t)\,x - \frac{1}{2}\beta(t)\,\nabla_x \log p_t(x)\right] dt.$$

Sub-VP. Let $\bar{\beta}(t) = \int_0^t \beta(s)\,ds$. With $g(t)^2 = \beta(t)(1 - e^{-2\bar{\beta}(t)})$:

$$dx = \left[-\frac{1}{2}\beta(t)\,x - \frac{1}{2}\beta(t)\big(1 - e^{-2\bar{\beta}(t)}\big)\,\nabla_x \log p_t(x)\right] dt.$$

In practice, the score $\nabla_x \log p_t(x)$ is replaced by the learned model $s_\theta(x,t)$ at inference time.

3.4. The ODE as a generative model

To generate a sample, one initializes $x_T \sim p_T \approx \mathcal{N}(0, I)$ and integrates the probability flow ODE backward from $t = T$ to $t = 0$. Because the ODE is deterministic, the only source of randomness is the initialization. Samples drawn this way have the same distribution at every time t as samples from the reverse SDE; in particular, x_0 is distributed approximately as p_0, the data distribution. Moreover, because the ODE flow is a diffeomorphism, it defines a bijective map between p_T and p_0, which we exploit in the next section for exact likelihood computation.

4. Likelihood computation

4.1. The instantaneous change of variables

We derive how $\log p_t(x_t)$ evolves along a trajectory of the ODE $\dot{x} = h(x,t)$. Denote by T_ε the flow map that advances the state by ε time, so $x(t+\varepsilon) = T_\varepsilon(x(t))$. Applying the standard finite-dimensional change-of-variables formula to the diffeomorphism T_ε:

$$\log p_{t+\varepsilon}(x(t+\varepsilon)) = \log p_t(x(t)) - \log \left|\det \frac{\partial T_\varepsilon}{\partial x}(x(t))\right|.$$

Rearranging and dividing by ε, then taking the limit $\varepsilon \to 0^+$ via L'Hôpital's rule (the numerator and denominator both tend to zero):

$$\frac{d}{dt}\log p_t(x_t) = -\lim_{\varepsilon \to 0^+} \frac{\partial}{\partial \varepsilon} \log \left|\det \frac{\partial T_\varepsilon}{\partial x}(x(t))\right|.$$

Since $\frac{\partial T_\varepsilon}{\partial x} \to I$ as $\varepsilon \to 0$, we have $\left|\det \frac{\partial T_\varepsilon}{\partial x}\right| \to 1$, and so

$$\frac{d}{d\varepsilon}\log|\det A| = \frac{1}{|\det A|}\frac{d}{d\varepsilon}|\det A| \to \frac{d}{d\varepsilon}|\det A|$$

at $\varepsilon = 0$. Applying Jacobi's formula $\frac{\partial}{\partial \varepsilon}\det A = \det A \cdot \operatorname{tr}\left(A^{-1}\frac{\partial A}{\partial \varepsilon}\right)$:

$$\frac{d}{dt}\log p_t(x_t) = -\operatorname{tr}\left(\lim_{\varepsilon \to 0^+} \frac{\partial}{\partial \varepsilon}\frac{\partial T_\varepsilon}{\partial x}(x(t))\right).$$

Taylor-expanding the flow map gives $T_\varepsilon(x) = x + \varepsilon\,h(x,t) + O(\varepsilon^2)$, so

$$\frac{\partial T_\varepsilon}{\partial x}(x(t)) = I + \varepsilon\,\frac{\partial h}{\partial x}(x(t),t) + O(\varepsilon^2).$$

Differentiating with respect to ε and taking $\varepsilon \to 0$ completes the proof:

$$\frac{d}{dt}\log p_t(x_t) = -\operatorname{tr}\left(\frac{\partial h}{\partial x}(x_t,t)\right) = -\nabla_x \cdot h(x_t,t). \tag{6}$$

4.2. Deriving the log-likelihood

Integrating (6) along the forward ODE trajectory from $t = 0$ to $t = T$:

$$\log p_T(x_T) - \log p_0(x_0) = -\int_0^T \nabla_x \cdot h(x_t, t)\, dt.$$

Rearranging:

$$\log p_0(x_0) = \log p_T(x_T) + \int_0^T \nabla_x \cdot h(x_t, t)\, dt. \tag{7}$$

In practice, both the trajectory $\{x_t\}$ and the integral of the divergence are computed simultaneously by augmenting the ODE state with the scalar $\ell_t = -\int_0^t \nabla_x \cdot h(x_s, s)\, ds$ and solving the joint system:

$$\frac{d}{dt}\begin{pmatrix} x_t \\ \ell_t \end{pmatrix} = \begin{pmatrix} h(x_t, t) \\ -\nabla_x \cdot h(x_t, t) \end{pmatrix}, \quad \begin{pmatrix} x_0 \\ \ell_0 \end{pmatrix} = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}. \tag{8}$$

At $t = T$, the log-likelihood follows from $\log p_0(x_0) = \log p_T(x_T) + \ell_T$.

4.3. Approximating the divergence

Computing $\nabla_x \cdot h(x_t, t) = \mathrm{tr}(J_h(x_t, t))$ exactly requires $\mathcal{O}(d)$ backward passes, which is intractable in high dimensions. The Hutchinson estimator resolves this: for any random vector ε with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon\varepsilon^\top] = I$,

$$\mathrm{tr}(J_h(x_t, t)) = \mathbb{E}_\varepsilon\left[\varepsilon^\top J_h(x_t, t)\, \varepsilon\right]. \tag{9}$$

Each evaluation requires only the Jacobian-vector product $J_h(x_t, t)\, \varepsilon$, which is a single forward-mode pass through h and costs $\mathcal{O}(1)$ per step. Averaging over a few draws of ε controls variance. The log-likelihood (7) is then computable at cost proportional to a standard ODE solve.

5. Discussion and Conclusion

The probability flow ODE is not merely a computational convenience. It reveals a deeper structure in the diffusion model framework: the stochastic reverse process and the deterministic ODE are two sides of the same coin, both encoding the same family of marginals $\{p_t\}$ through the score. The SDE samples stochastically by injecting noise at each step; the ODE samples deterministically by following the mean drift. Neither is universally better, but the ODE offers invertibility and exact likelihood computation that the SDE cannot provide.

Additionally, the connection to continuous normalizing flows [2] is worth making explicit. The probability flow ODE is exactly a continuous normalizing flow whose vector field is constrained to take the form $f - \frac{g^2}{2}\nabla_x \log p_t$. The diffusion model framework provides a principled way to specify and train this vector field via denoising score matching, sidestepping the difficulties of training normalizing flows directly.

References

[1] Y. Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. *International Conference on Learning Representations*. 2021.

[2] R. T. Q. Chen et al. “Neural Ordinary Differential Equations”. *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.

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

[4] J. Ho, A. Jain, and P. Abbeel. “Denoising Diffusion Probabilistic Models”. *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.