The Mathematical Builidng Blocks of Diffusion Generative Models

Tonghe Zhang[1]

[1] Carnegie Mellon University
tonghez@andrew.cmu.edu

October, 2025

Content

Abstract

We provide a proof of the necessary condition of the Fokker-Planck (FP) equation and derive some important conclusions, including the continuity equation and the reverse-time diffusion process. Then we study a typical variant of diffusion process, namely, flow matching processes with linear interpolation paths, and study the relationship between its score and velocity. Finally, we study how to adopt the FP equation to derive an ODE-SDE conversion formula that links flow ODEs with diffusion SDEs.

1 Preliminaries

To fully understand the proof of FP equations, we need to recall several results from probability, analysis and linear algebra.

1.1 Wiener Process

Recall that a Wiener process W_t in $\mathbb{R}^d$ possesses the following property:

$$\forall t, h \in \mathbb{R}_+ : \ W_{t+h} - W_t \sim \mathcal{N}(0, h\mathbb{I}_{d\times d}) \tag{1}$$

† Corresponding author.

This naturally implies that, for any matrix $A \in \mathbb{R}^{d\times d}$ independent of W_t,

$$\begin{aligned}&\mathbb{E}\left[W_{t+h} - W_t\right] = 0\\ &\mathbb{E}\left[(W_{t+h} - W_t)^\top A(W_{t+h} - W_t)\right] = \mathbb{E}_{\epsilon\sim}\left[\epsilon^\top A\epsilon\right],\ \epsilon \sim \mathcal{N}(0, h\mathbb{I}_{d\times d})\end{aligned} \tag{2}$$

Here, $\epsilon \in \mathbb{R}^d$, which has independent coordinates. The expectations are taken with respect to the randomness in W_t.

1.2 Hutchinson's Trace Estimator

For $\epsilon \sim \mathcal{N}(0, h\mathbb{I}_{d\times d})$ and matrix $A \in \mathbb{R}^{d\times d}$, the trace of A is the expected quadratic form of ϵ and A.

$$\mathbb{E}_\epsilon\left[\epsilon^\top A\epsilon\right] = \mathbb{E}_{\epsilon_1,\epsilon_2,\ldots,\epsilon_d}\left[\sum_{i,j}\epsilon_i A_{i,j}\epsilon_j\right] = \sum_{i,j} A_{i,j}\ \mathbb{E}_{\epsilon_1,\epsilon_2,\ldots,\epsilon_d}\left[\epsilon_i\epsilon_j\right] = \sum_{i,j} A_{i,j}\ h\delta_{i,j} = h\cdot \mathrm{tr}A \tag{3}$$

Consequently, $\widehat{A}_\epsilon = \frac{1}{h}\epsilon^\top A\epsilon$ is an unbiased estimator of $\mathrm{tr}\, A$.

1.3 Hessian Matrix and Laplacian

For a scalar functional $u : \mathbb{R}^d \to \mathbb{R}$ (which usually takes the physical meaning of a potential field over $\mathbb{R}^d$), the Hessian matrix is defined by packing the second-order derivatives into a matrix:

$$H_u = \nabla^2 u = \left(\frac{\partial^2}{\partial x_i x_j}u(\vec{x})\right)_{i,j\in[d]^2} \in \mathbb{R}^d$$

The Laplacian of f is defined as the sum of second order derivatives at the same coordindates:

$$\Delta u = \nabla\cdot\nabla u = \sum_{i=1}^{d}\ \frac{\partial}{\partial x_i}\ u(\vec{x}) \in \mathbb{R}$$

By definition, the trace of the Hessian matrix is the Laplacian.

$$\mathrm{tr}\,\nabla^2 u = \Delta u \tag{4}$$

Combining Eqs. (2), (3), and (4), we immediately obtain the following.

Corollary 1.1. *For a scalar functional $u : \mathbb{R}^d \to \mathbb{R}$ and white Gaussian vector $\epsilon \sim \mathcal{N}(0, h\mathbb{I}_{d\times d})$, the expected value of the quadratic form obtained from the difference of Wiener process W_t and Hessian matrix $\nabla^2 u$ is equivalent to the Laplacian of the potential field u.*

$$\mathbb{E}\left[(W_{t+h} - W_t)^\top \nabla^2 u(W_{t+h} - W_t)\right] = h\Delta u \tag{5}$$

1.4 The Laplacian of probability density; score function

The Laplacian of a probability density function p often appears in the literature of diffusion models. By definition, we have

$$\nabla^2 p = \nabla\cdot(\nabla p)$$

Furthermore, we relate the density gradient to the score function, $s(x) = \nabla \log p(x) = \frac{\nabla p(x)}{p(x)}$, which implies

$$\nabla^2 p = \nabla\cdot(p\nabla\log p)$$

So we know that

Theorem 1.2. *The Laplacian of a probability density is the divergence of the product between density and its score function:*

$$\nabla^2 p = \nabla\cdot(p(x)s(x)) \tag{6}$$

1.5 Point-wise Equivalence, Test Functions, and Integration by Parts

To prove that two real functionals are point-wise equal, we can resort to evaluating the inner products of these functions with the same test function. If the test results (i.e. the inner products) are always the same for any test function used, then the functions are the same. Rigorously speaking,

Theorem 1.3. *For arbitrary integrable functions $g_1, g_2 : \mathbb{R}^d \to \mathbb{R}$ it holds that*

$$g_1(x) = g_2(x) \text{ for all } x \in \mathbb{R}^d \quad \Leftrightarrow \quad \int f(x)g_1(x)\mathrm{d}x = \int f(x)g_2(x)\mathrm{d}x \text{ for all test functions } f$$

Recall that a test function is an infinitely differentiable function with a compact support, and it includes dirac delta functions. So, from RHS to LHS is simple, just pick the dirac delta. The transition from LHS to RHS is also straightforward.

More properties of test functions

- First, the test functions f are zero on the border of the compact support $\partial\Omega$. Because by definition, f is differentiable in the support first, so it is continuous from the outside to the boundary. Second, the support is compact, so for any point on the boundary, there is a sequence of points outside the support (whose values are zero) that converges to it. Since the function is continuous on this trajectory, $f\Big|_{\partial\Omega}$ must also be zero.
- For arbitrary test functions u, v, we can use integration by parts to derive

$$\begin{aligned}\int_D f_1(x)\frac{\partial}{\partial x_i}f_2(x)\mathrm{d}x =& f_1(x)f_2(x)\Big|_{\partial D} - \int f_2(x)\frac{\partial}{\partial x_i}f_1(x)\mathrm{d}x = 0 - \int f_2(x)\frac{\partial}{\partial x_i}f_1(x)\mathrm{d}x \\ =& -\int f_2(x)\frac{\partial}{\partial x_i}f_1(x)\mathrm{d}x\end{aligned} \tag{7}$$

  Using this together with the definition of the divergence and Laplacian, we get the identities:

$$\begin{aligned}\int \nabla^T u(x)\vec{v}(x)\mathrm{d}x &= -\int u(x)\,\mathrm{div}\,(\vec{v})\,(x)\mathrm{d}x \quad \left(u : \mathbb{R}^d \to \mathbb{R}, \vec{v} : \mathbb{R}^d \to \mathbb{R}^d\right) \\ \int u(x)\Delta v(x)\mathrm{d}x &= \int v(x)\Delta u(x)\mathrm{d}x \quad \left(u : \mathbb{R}^d \to \mathbb{R}, v : \mathbb{R}^d \to \mathbb{R}\right)\end{aligned} \tag{8}$$

  A even more symbolic expression is

$$\begin{aligned}\langle \nabla u, \vec{v}\rangle =& -\langle \nabla\cdot\vec{v}, u\rangle \\ \langle u, \Delta v\rangle =& \langle v, \Delta u\rangle = -\langle \nabla v, \nabla u\rangle\end{aligned} \tag{9}$$

  And writing like this immediately tells us that the Laplacian is self-adjoint when applying on test functions. For completeness, we also provide a quick proof below.

  Proof. Notice that in each equation, x is understood as $(x_1, x_2, \ldots, x_d)$ and

$$\int_\Omega f(x)\mathrm{d}x = \int_{\Omega_1\times\ldots\times\Omega_d} f(x_1, x_2, \ldots, x_d)\ \mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_d$$

  So we can do integration by parts recurssively on each coordinate and peel of the vector inner products. Specifically, For the first statement,

$$\begin{aligned}\int_\Omega \nabla^\top u\ \vec{v}\ \mathrm{d}x =& \int_\Omega \sum_{i=1}^d \frac{\partial u}{\partial x_i}v\mathrm{d}x = \sum_{i=1}^d \int_{\prod_{j\neq i}\Omega_j}\prod_{j\neq i}\mathrm{d}x_j\ \int_{\Omega_i}\frac{\partial u}{\partial x_i}v(x_1,\ldots,x_{i-1},\mathbf{x_i},x_{i+1},\ldots,x_d)\mathrm{d}\mathbf{x_i} \\ =& \sum_{i=1}^d \int_{\prod_{j\neq i}\Omega_j}\prod_{j\neq i}\mathrm{d}x_j\ \int_{\Omega_i}\left(uv\big|_{\partial\Omega_i} - u\frac{\partial v}{\partial x_i}\right) \\ =& -\sum_{i=1}^d\int_\Omega \mathrm{d}x\ u\frac{\partial v}{\partial x_i} = -\int_\Omega \mathrm{d}x\ u\nabla\cdot v\end{aligned} \tag{10}$$

As for the second statement,

$$\begin{aligned}
\int_\Omega v\Delta u \mathrm{d}x &= \int_\Omega v \sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right)\mathrm{d}x \\
&= \sum_{i=1}^d \int_{\prod_{j\neq i}\Omega_j} \prod_{j\neq i}\mathrm{d}x_j \int_{\Omega_i} \mathrm{d}x_i\ v(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_d)\frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right)\mathrm{d}x_i \\
&= \sum_{i=1}^d \int_{\prod_{j\neq i}\Omega_j} \prod_{j\neq i}\mathrm{d}x_j \left\{v(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_d)\frac{\partial u}{\partial x_i}\bigg|_{\partial\Omega_i} - \int_{\Omega_i}\mathrm{d}x_i\ \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\right\} \\
&= \sum_{i=1}^d \int_{\prod_{j\neq i}\Omega_j} \prod_{j\neq i}\mathrm{d}x_j \left\{0 - \int_{\Omega_i}\mathrm{d}x_i\ \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\right\} = -\int_\Omega \sum_{i=1}^d \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\mathrm{d}x_i
\end{aligned} \tag{11}$$

From here you can see that this result is symmetric w.r.t. u and v. So we can replace u with v and vice versa while the value will not change, hence the second statement also holds. □

2 Fokker-Planck Equations

For a stochastic process determined by a stochastic differential equation (SDE),

$$\mathrm{d}X_t = \text{some function of } X_t,\ t, \text{ and random noise}$$

the Fokker-Planck (FP) equation tells us how the marginal density of this process $X_t \sim p_t(\cdot)$ changes when the time evolves, and it expresses the marginal density in terms of the coefficients of the underlying SDE. The FP equations precisely characterize the *dynamics* of this stochastic process, connecting the *differential equations* of evolution with the *statistical* properties of this system.

In the following sections, we will detail the proof of the Fokker-Planck equation in its most general form. Afterwards, we will use the FP equation to study two special types of stochastic processes, the first one is deterministic, and the second is stochastic, but the noise is irrelevant to the current state. These two cases are of particular interest to machine learning studies, as the former leads to flow-matching process, and the latter leads to diffusion models.

2.1 How an SDE Drives the Evolution of Marginal Density

First, we provide a statement of the Fokker-Planck equation.

Theorem 2.1 (Fokker-Planck Equation)**.** *Let p_t be a probability path and consider Itô SDE*

$$\vec{X}_0 \sim p_0, \quad \mathrm{d}\vec{X}_t = \vec{\mu}\left(\vec{X}_t, t\right)\mathrm{d}t + \vec{\sigma}\left(\vec{X}_t, t\right)\ \mathrm{d}W_t$$

where $\vec{X}_t \in \mathbb{R}^d$. Then $\vec{X}_t$ has distribution p_t for all $0 \le t \le 1$ if and only if the Fokker-Planck equation holds:

$$\partial_t p_t(x) = -\vec{\nabla}\cdot\left(p_t\vec{\mu}_t\right)(x) + \frac{1}{2}\Delta(\sigma_t^2 p_t)(x) \quad \text{for all } x \in \mathbb{R}^d, 0 \le t \le 1$$

where we use the abbreviations $\vec{\mu}_t(x) = \vec{\mu}(x_t, t)$ and $\sigma_t^2(x) = \langle\vec{\sigma}(x_t,t), \vec{\sigma}(x_t,t)\rangle$.

Proof. We only show the necessary condition, which is nontrivial.

This equation shows point-wise equivalence between two real functionals. To show that, we pick an arbitrary test function $f : \mathbb{R}^d \to \mathbb{R}$, and show that the left and the right are equivalent after applying f to form inner products. We want to show that

$$\int \partial_t p_t(x) f(x)\mathrm{d}x = \int \left[-\nabla\cdot\left(p_t\vec{\mu}_t\right)(x) + \frac{1}{2}\Delta(\sigma_t^2 p_t)(x)\right] f(x)\mathrm{d}x$$

holds for all test functions. And then for any point x, we pick $f(z) = \delta(z-x)$ and finish proving that LHS=RHS at any point.

Let us start from LHS. We will use the property that $p_t(x)$ is in fact a probability density, and the test function is time-irrelevant. Consequently, LHS can be written into an expectation:

$$LHS = \partial_t \mathbb{E}_{x\sim p_t}\left[f(x)\right] = \lim_{h\to 0}\frac{1}{h}\mathbb{E}\left[f(\vec{X}_{t+h}) - f(\vec{X}_t)\right] = \lim_{h\to 0}\frac{1}{h}\mathbb{E}\left[\mathbb{E}\left[f(\vec{X}_{t+h}) - f(\vec{X}_t)\middle|\vec{X}_t\right]\right]$$

We will use the second-order Taylor expansion to evaluate the enumerator, since this limit eliminates any remaining terms of order $o(h)$. Then we will call the helper functions derived in the preliminary section and take the limit to finish the proof. In what follows, we neglect the higher order terms for better readability.

$$\begin{aligned}
&f\left(\vec{X}_{t+h}\right) - f\left(\vec{X}_t\right)\\
=&f\left(\vec{X}_t + h\vec{\mu}_t\left(\vec{X}_t\right) + \sigma_t\left(W_{t+h} - W_t\right)\right) - f\left(\vec{X}_t\right) \qquad //\mathrm{d}\vec{X}_t = \vec{\mu}_t(\vec{X}_t)\mathrm{d}t + \sigma_t(\vec{X}_t)\mathrm{dt}\\
\overset{(i)}{=}& \nabla f\left(\vec{X}_t\right)^T\left(h\vec{\mu}_t\left(\vec{X}_t\right) + \sigma_t\left(W_{t+h} - W_t\right)\right)\Bigg)\\
&+\frac{1}{2}\left(h\vec{\mu}_t\left(\vec{X}_t\right) + \sigma_t\left(W_{t+h} - W_t\right)\right)\Bigg)^T\nabla^2 f\left(\vec{X}_t\right)\left(h\vec{\mu}_t\left(\vec{X}_t\right) + \sigma_t\left(W_{t+h} - W_t\right)\right)\Bigg) + o(h^2) \qquad //\text{2nd order Taylor}\\
\overset{(ii)}{=}& h\nabla f\left(\vec{X}_t\right)^T\vec{\mu}_t\left(\vec{X}_t\right) + \sigma_t\nabla f\left(\vec{X}_t\right)^T\left(W_{t+h} - W_t\right)\\
&+\frac{1}{2}h^2\vec{\mu}_t\left(\vec{X}_t\right)^T\nabla^2 f\left(\vec{X}_t\right)\vec{\mu}_t\left(\vec{X}_t\right) + h\sigma_t\vec{\mu}_t\left(\vec{X}_t\right)^T\nabla^2 f\left(\vec{X}_t\right)\left(W_{t+h} - W_t\right)\\
&+\frac{1}{2}\sigma_t^2\left(W_{t+h} - W_t\right)^T\nabla^2 f\left(\vec{X}_t\right)\left(W_{t+h} - W_t\right) + o(h^2) \quad //\text{Brute-force expansion}\\
=& h\nabla f\left(\vec{X}_t\right)^T\vec{\mu}_t\left(\vec{X}_t\right) + \sigma_t\nabla f\left(\vec{X}_t\right)^T\left(W_{t+h} - W_t\right) + \frac{1}{2}\sigma_t^2\left(W_{t+h} - W_t\right)^T\nabla^2 f\left(\vec{X}_t\right)\left(W_{t+h} - W_t\right) + o(h^2)
\end{aligned} \tag{12}$$

Now take the conditonal expectation w.r.t. $\vec{X}_t$ to both sides. We will nullify the second term using the fact that the Wiener process is independent with $\vec{X}_t$ and has zero mean. The third term on RHS will be simplified by Eq. (5), we suggests that

$$\mathbb{E}\left[(W_{t+h} - W_t)^\top\nabla^2 u(W_{t+h} - W_t)\right] = h\Delta u$$

With these observations, we arrive at

$$\mathbb{E}\left[f\left(\vec{X}_{t+h}\right) - f\left(\vec{X}_t\right) \mid \vec{X}_t\right] = h\nabla f\left(\vec{X}_t\right)^T\vec{\mu}_t\left(\vec{X}_t\right) + \frac{h}{2}\sigma_t^2(\vec{X}_t)\Delta\left(f(\vec{X}_t)\right) + o(h^2) \tag{13}$$

Finnaly, marginalize the conditional expectation by computing integrals, and call the integration by parts formula for vector fields, we have

$$\begin{aligned}
&\partial_t\mathbb{E}\left[f\left(\vec{X}_t\right)\right]\\
=&\int f(x)\left(\partial_t p_t(x)\right)\mathrm{d}x\\
=&\lim_{h\to 0}\frac{1}{h}\mathbb{E}\left[f\left(\vec{X}_{t+h}\right) - f\left(\vec{X}_t\right)\right]\\
=&\lim_{h\to 0}\frac{1}{h}\mathbb{E}\left[\mathbb{E}\left[f\left(\vec{X}_{t+h}\right) - f\left(\vec{X}_t\right) \mid \vec{X}_t\right]\right.\\
=&\mathbb{E}\left[\lim_{h\to 0}\frac{1}{h}\left(h\nabla f\left(\vec{X}_t\right)^T\vec{\mu}_t\left(\vec{X}_t\right) + \frac{h}{2}\sigma_t^2(\vec{X}_t)\Delta f\left(\vec{X}_t\right) + o(h^2)\right)\right]\\
=&\mathbb{E}\left[\nabla f\left(\vec{X}_t\right)^T\vec{\mu}_t\left(\vec{X}_t\right) + \frac{1}{2}\sigma_t^2(\vec{X}_t)\Delta f\left(\vec{X}_t\right)\right]\\
=&\int\nabla f(x)^T\vec{\mu}_t(x)p_t(x)\mathrm{d}x + \frac{1}{2}\int\Delta f(x)\cdot\sigma_t^2(x)p_t(x)\mathrm{d}x\\
=&-\int f(x)\,\mathrm{div}\left(\vec{\mu}_t p_t\right)(x)\mathrm{d}x + \frac{1}{2}\int f(x)\Delta\left(\sigma_t^2(x)p_t(x)\right)\mathrm{d}x\\
=&\int f(x)\left(-\,\mathrm{div}\left(\vec{\mu}_t p_t\right)(x) + \frac{1}{2}\Delta\left(\sigma_t^2 p_t\right)(x)\right)\mathrm{d}x \qquad //\text{Integration by parts}
\end{aligned}$$

Now since this result holds for any test functions, at any point x_0 we can pick $f(x) = \delta(x - x_0)$ and obtain

$$\partial_t p_t(x)\Big|_{x=x_0} = -\operatorname{div}\left(\vec{\mu}_t(x)p_t(x)\right) + \frac{1}{2}\Delta\left(\sigma_t^2(x)p_t\left(x\right)\right)\Big|_{x=x_0}$$

thus these two functionals are equivalent point-wise.

□

2.2 Special Case I: ODE, Continuity Equation, and Flow-matching

With the Fokker-Planck equation, we naturally obtain an ODE counterpart by setting the noise magnitude $\sigma_t(X_t)$ to zero. This is known as the continuity equation.

Corollary 2.2 (Continuity Equation)**.** *Let p_t be a probability path and consider the ODE*

$$\vec{X}_0 \sim p_0, \quad \mathrm{d}\vec{X}_t = \vec{\mu}\left(\vec{X}_t, t\right)\mathrm{d}t$$

where $\vec{X}_t \in \mathbb{R}^d$. Then $\vec{X}_t$ has distribution p_t for all $0 \le t \le 1$ if and only if the continuity equation holds:

$$\partial_t p_t(x) = -\vec{\nabla}\cdot\left(p_t\vec{\mu}_t\right)(x) \quad \text{for all } x \in \mathbb{R}^d, 0 \le t \le 1$$

where we use the abbreviations $\vec{\mu}_t(x) = \vec{\mu}(x_t, t)$.

This relationship is very useful in the derivation of flow-matching process's log probabilities.

2.3 Special Case II: Forward and Reverse-time Diffusion Process

Definition 2.3 (Diffusion Process)**.** A **diffusion process** is a continuous-time stochastic process $\{X_t\}_{t\ge 0}$ governed by the Itô stochastic differential equation (SDE):

$$\mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t \tag{14}$$

where:

- $f(X_t, t) : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$ is the drift coefficient,
- $g(t) : \mathbb{R}_+ \to \mathbb{R}_+$ is the state-independent diffusion coefficient,
- W_t is a standard d-dimensional Brownian motion (Wiener Process)

The key characteristic of a diffusion process is that the diffusion coefficient $g(t)$ depends only on time t and not on the current state X_t, distinguishing it from the general Itô SDE where $\sigma(X_t, t)$ may depend on the state.

By the Fokker-Planck equation, the marginal probability density of a forward diffusion process is given by

$$\frac{\partial \mathbf{p}_t(\mathbf{x})}{\partial t} = -\nabla\cdot\left(\mathbf{f}(\mathbf{x}, t)\mathbf{p}_t(\mathbf{x})\right) + \frac{1}{2}g^2(t)\nabla^2\mathbf{p}_t(\mathbf{x}) \tag{15}$$

It turns out that the forward diffusion process is actually invertible mathematically. In detail, there exists a reverse-time diffusion process with density $\tilde{\mathbf{p}}_s(\mathbf{x})$, such that $\tilde{\mathbf{p}}_s(\mathbf{x}) = \mathbf{p}_{T-s}(\mathbf{x})$. When the reverse-time process evolves from $s = 0$ to $s = T$, the marginal probability is exactly the same as the forward process evolved backward. Next, we find the SDE for the time-reversed process.

2.3.1 Time Reversal and Density Evolution

We denote by s the increasing time index of the reverse process, and let t be the increasing time index of the forward process. With a time endpoint T, these two time indices are related with $s + t = T$. Since $\mathrm{d}s = -\mathrm{d}t$, we obtain that the evolution of the reversed density with respect to the forward time s is:

$$\frac{\partial \tilde{\mathbf{p}}_s(\mathbf{x})}{\partial s} = \frac{\partial \mathbf{p}_{T-s}(\mathbf{x})}{\partial s} = -\frac{\partial \mathbf{p}_t(\mathbf{x})}{\partial t}\Big|_{t=T-s} \tag{16}$$

Substituting the Forward Fokker-Planck Equation (15) into (16), we can express the probability density of the reverse process in terms of the forward diffusion coefficients:

$$\begin{aligned}\frac{\partial \tilde{\mathbf{p}}_s(\mathbf{x})}{\partial s} &= -\left[-\nabla\cdot\left(\mathbf{f}(\mathbf{x}, t)\mathbf{p}_t(\mathbf{x})\right) + \frac{1}{2}g^2(t)\nabla^2\mathbf{p}_t(\mathbf{x})\right]\Big|_{t=T-s} \\ &= \nabla\cdot\left(\mathbf{f}(\mathbf{x}, T-s)\tilde{\mathbf{p}}_s(\mathbf{x})\right) - \frac{1}{2}g^2(T-s)\nabla^2\tilde{\mathbf{p}}_s(\mathbf{x})\end{aligned} \tag{17}$$

2.3.2 Diffusion, Laplacian, and Score

Since the magnitude of the noise of a diffusion process is irrelevant to the state, the Laplacian in the Fokker-Planck equation will be applied directly to the probability density. By Eq. (6), this naturally results in score functions; hence in diffusion processes, the Laplacian are very closely related to the score function. This is why the score appears so many times in diffusion model learning. Concretely speaking, the Laplacian in the diffusion term can be simplified as

$$\frac{1}{2}g^2(T-s)\nabla^2\tilde{\mathbf{p}}_s(\mathbf{x}) = \frac{1}{2}g^2(T-s)\nabla\cdot(\tilde{\mathbf{p}}_s(\mathbf{x})\nabla\log\tilde{\mathbf{p}}_s(\mathbf{x})) \tag{18}$$

Substituting this into the reversed density evolution in Eq. (17) and factoring out the divergence operator and the reverse density, we have

$$\frac{\partial\tilde{\mathbf{p}}_s(\mathbf{x})}{\partial s} = \nabla\cdot\left(\tilde{\mathbf{p}}_s(\mathbf{x})\left[\mathbf{f}(\mathbf{x},T-s) - \frac{1}{2}g^2(T-s)\nabla\log\tilde{\mathbf{p}}_s(\mathbf{x})\right]\right) \tag{19}$$

2.3.3 The Reverse-Time SDE

If the reverse process exists and is also a diffusion process, it must also be governed by an FP equation with state-irrelevant noise. Let the density $\tilde{\mathbf{p}}_s(\mathbf{x})$ of the hypothesized reverse-time SDE be determined by

$$d\tilde{\mathbf{X}}_s = \mathbf{h}(\tilde{\mathbf{X}}_s, s)ds + \tilde{g}(s)d\mathbf{W}_s \tag{20}$$

since it is the reverse process of $\mathbf{p}$, it must also satisfy its own Fokker-Planck equation:

$$\frac{\partial\tilde{\mathbf{p}}_s(\mathbf{x})}{\partial s} = -\nabla\cdot(\mathbf{h}(\mathbf{x},s)\tilde{\mathbf{p}}_s(\mathbf{x})) + \frac{1}{2}\tilde{g}^2(s)\nabla^2\tilde{\mathbf{p}}_s(\mathbf{x}) \tag{21}$$

If we assume that the diffusion coefficient remains the same, $\tilde{g}(s) = g(T-s)$, we can [1] equate (19) with (21) to determine the reverse drift $\mathbf{h}(\mathbf{x},s)$.

By factoring the reverse FP equation (21) using the identity $\frac{1}{2}\tilde{g}^2\nabla^2\tilde{\mathbf{p}}_s = -\nabla\cdot\left(-\frac{1}{2}\tilde{g}^2\tilde{\mathbf{p}}_s\nabla\log\tilde{\mathbf{p}}_s\right)$:

$$\frac{\partial\tilde{\mathbf{p}}_s(\mathbf{x})}{\partial s} = -\nabla\cdot\left(\tilde{\mathbf{p}}_s(\mathbf{x})\left[\mathbf{h}(\mathbf{x},s) - \frac{1}{2}g^2(T-s)\nabla\log\tilde{\mathbf{p}}_s(\mathbf{x})\right]\right) \tag{22}$$

Equating the drift terms from this and equation (19):

$$-\left[\mathbf{h}(\mathbf{x},s) - \frac{1}{2}g^2(T-s)\nabla\log\tilde{\mathbf{p}}_s(\mathbf{x})\right] = \mathbf{f}(\mathbf{x},T-s) - \frac{1}{2}g^2(T-s)\nabla\log\tilde{\mathbf{p}}_s(\mathbf{x})$$

Solving for the reverse drift $\mathbf{h}(\mathbf{x},s)$:

$$\mathbf{h}(\mathbf{x},s) = -\mathbf{f}(\mathbf{x},T-s) + g^2(T-s)\nabla\log\tilde{\mathbf{p}}_s(\mathbf{x})$$

Substituting $\mathbf{h}(\tilde{\mathbf{X}}_s,s)$ back into (20) and replacing $\tilde{\mathbf{p}}_s(\tilde{\mathbf{X}}_s)$ with $\mathbf{p}_{T-s}(\tilde{\mathbf{X}}_s)$, we get the Reverse-Time Diffusion Process Equation:

$$\text{Reverse diffusion SDE}: \quad d\tilde{\mathbf{X}}_s = \left[-\mathbf{f}(\tilde{\mathbf{X}}_s,T-s) + g^2(T-s)\nabla\log\mathbf{p}_{T-s}(\tilde{\mathbf{X}}_s)\right]ds + g(T-s)d\mathbf{W}_s \;\; s\in[0,T] \tag{23}$$

The convention of reversed time index

The equation derived above is valid for a process $\tilde{\mathbf{X}}_s$ running in a forward-time index $s\in[0,T]$. However, literature, particularly in generative modeling, often maintains the original time index $t\in[0,T]$ but interprets the process as running backward from T to 0.

Let us use the original time index t, but define the differential as a negative change $d(-t)$. Let $\mathbf{x}(t)$ be the reverse process, where t now runs from T (start) to 0 (end). Eq. (23) states can be written as

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t,t) - g^2(t)\nabla\log\mathbf{p}_t(\mathbf{x}_t)\right](-\mathrm{d}t) + g(t)\mathrm{d}\bar{\mathbf{W}}_t$$

[1] This is the only assumption in solving the two terms in reverse-process, apparently an additional assumption is required because we have one equation for two unknowns.

and we remind the reader that now t decreases from T to 0. Here, $(-\mathrm{d}t)$ indicates the backward direction and $\mathrm{d}\bar{\mathbf{W}}_t$ is a backward Wiener process, which means that when t decreases, it is statistically equivalent to $\mathrm{d}\mathbf{W}_s$ with s increasing.

To fully express the backward time, we abuse notations and write $-\mathrm{d}t$ as $\mathrm{d}t$, though it violates the convention that infinitesimal time should be positive. Then with a negative '$\mathrm{d}t$', we arrive at what is seen in the diffusion model literature:

$$\text{Reverse diffusion SDE (ML convention)}: \ \mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t) \underbrace{\nabla \log \mathbf{p}_t(\mathbf{x}_t)}_{\text{Score function}}\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{W}}_t, \ t \in [T, 0]$$

and we remind the readers again that this $\mathrm{d}t$ is negative in machine learning's convention.

2.4 Discussion: Connecting Flows with Diffusion

References