CMPTG 5 - Statistical Physics of Neural Networks

Sidharth Kannan

April 2025

1 Lecture 05 - Fokker-Planck Equations

Concepts to review prior to this lecture: Dirac Delta functions, Hessian matrices.

In the previous lecture, we saw how we could sample from a diffusion model by reversing the stochastic differential equation that represents our diffusion model. Now, because we are solving a stochastic equation, with a numerical solver, we have to take rather small time steps. This comes down to the fact that since we are modelling a stochastic process, the function is not smooth, nor is it even approximately smooth.

If there was a way to convert this stochastic differential equation to an ordinary differential equation, or even a partial differential equation, we could bring the considerable machinery of the applicable numerical methods to bear, and solve our differential equation much more efficiently. We can also then use the resulting ODE to accurately compute the actually probabilities (likelihoods) of particular samples.

Lucky for us, there is a way to perform precisely this conversion, and it is known as the Fokker-Planck equation. The Fokker-Planck equation was introduced by Max Planck and Adriaan Fokker in the early 1900s, to assist in the modeling of Brownian motion.

For an Itô SDE of the form

$$dX_t = f(X_t, t)dt + g(X_t, t)dW_t \tag{1}$$

with an initial condition $p_0(x)$, the Fokker-Planck equation describes the *probability flow.* That is, it tells us how the probability density changes over time.

$$\begin{aligned}\frac{d}{dt}p_t(x) &= -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left[f_i(x,t)p_t(x)\right] + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial x_i \partial x_j}\left[[g(x,t)g(x,t)^T]_{ij}p_t(x)\right] \\ &= -\nabla\cdot[fp_t] + \frac{1}{2}[\nabla^2\cdot[gg^Tp_t]]\end{aligned}$$

Which we can rewrite as

$$\frac{d}{dt}p_t(x) = -\nabla\cdot[fp_t + \frac{1}{2}\nabla^T[gg^Tp_t]] \tag{2}$$

Which yields

$$\frac{d}{dt}p_t(x) + \nabla\cdot[fp_t + \frac{1}{2}\nabla^T[gg^Tp_t]] = 0 \tag{3}$$

This is a gnarly mess, so let's take a moment to paint an intuitive picture, before we go into a formal derivation. Equations of this form are called *continuity equations*, and they appear everywhere in physics, from studies of fluid flow, to electric currents, to quantum mechanics. What this equation is essentially saying is that the time rate of change of the probability density is equation to the amount of probability flowing into that region of space. To see how, the first concept to understand here is the $\nabla\cdot$ operator, called the *divergence*.

The divergence is a sum over partial derivatives. Each derivative tells us how much our quantity is changing in each direction, so, intuitively, we can understand the divergence as representing the net outflow of the quantity from a particular region of space. Armed with this understanding, we can try and understand the Fokker-Planck equation. Using the form in Eq. 2, we have that the quantity whose divergence we care about is

$$fp_t + \frac{1}{2}\nabla^T[gg^T p_t] \tag{4}$$

The first term is easier to tease apart. The function f is our deterministic drift velocity, so naturally, the first term can be interpreted as a kind of momentum density, whose divergence then gives the total amount of outflow due to the drift. The second term is harder, so let's break it apart.

gg^T is what is commonly called the *diffusion matrix.* As we learned before, g represents random drift, and so this matrix tells us how much the probability density is spreading in each direction. Now, diffusion on its own does not change the probability density, because all points are diffusing at equal rates. Thus, we care more about the (spatial) rate of change of the diffusion. I.e. are the points near me diffusing more or less than I am.

The actual derivation of the Fokker-Planck equation is rather hairy, and involves a bit of mathematical machinery known as Itô's Lemma.

1.1 Itô's Lemma

In order to derive the Fokker-Planck equation, we must first derive a sort of "chain rule" for the stochastic calculus. This chain rule is called Itô's formula, and it tells us how to calculate the differential of a *function* of a stochastic process. That is,

Given an SDE to describe the stochastic process, X_t, is the stochastic process $Y_t \equiv \eta(X_t)$ also described by an SDE? If so, what are the drift coefficients of that SDE? The SDE can be written

$$dY_t = \bar{f}(X_t, t)dt + \bar{g}(X_t, t)dt$$

Itô's lemma gives us a way to compute these coefficients:

$$\begin{aligned}\bar{f}(x,t) &= \nabla\eta(x)^T f(x,t) + \mathrm{Tr}(g(x,t)^T\nabla^2\eta(x)g(x,t))\\ \bar{g}(x,t) &= \sqrt{\nabla\eta(x)^T g(x,t)g(x,t)\nabla\eta(x)}\end{aligned}$$

To prove it, we can start by writing down an expression for Y_{t+h}, in the limit when the time step h is small.

$$\begin{aligned}Y_{t+h} =& \eta(X_{t+h})\\ \approx& \eta(X_t + hf(X_t,t) + \sqrt{h}g(X_t,t)\epsilon)\end{aligned}$$

Then, we Taylor expand η, keeping the terms up to linear order in h.

$$\begin{aligned}&= \eta(X_t) + \nabla\eta(X_t)^T \left[hf(X_t,t) + \sqrt{h}g(X_t,t)\epsilon\right]\\ &\quad + \frac{1}{2}\left[hf(X_t,t) + \sqrt{h}g(X_t,t)\epsilon\right]^T \nabla^2\eta(X_t)^T \left[hf(X_t,t) + \sqrt{h}g(X_t,t)\epsilon\right] + \mathcal{O}(h^2) \qquad (5)\end{aligned}$$

Here, ∇^2 is the Hessian matrix, not the Laplacian. To make progress, let's look at that last term:

$$\frac{1}{2}\left[hf(X_t,t) + \sqrt{h}g(X_t,t)\epsilon\right]^T \nabla^2\eta(X_t)^T \left[hf(X_t,t) + \sqrt{h}g(X_t,t)\epsilon\right] \qquad (6)$$

We see it has the form

$$\frac{1}{2}(A+B)^T M(A+B) \qquad (7)$$

where $A = hf(X_t,t)$, $B = \sqrt{h}g(X_t,t)\epsilon$, and M is the symmetric linear operator $\nabla^2\eta(X_t)$. We recognize this as a quadratic form, and expand it as

$$\frac{1}{2}A^T MA + A^T MB + \frac{1}{2}B^T MB \qquad (8)$$

We substitute our definitions of A, B, M back in to get.

$$\begin{aligned}&= Y_t + \nabla\eta(X_t)^T \left[hf(X_t,t) + \sqrt{h}g(X_t,t)\epsilon\right] + \frac{1}{2}h^2 f(X_t,t)^T \nabla^2\eta(X_t) f(X_t,t)\\ &\quad + \sqrt{h}h[g(X_t,t)\epsilon]^T \nabla^2\eta(X_t) f(X_t,t) + h\epsilon^T g(X_t,t)^T \nabla^2\eta(X_t) g(X_t,t)\epsilon + \mathcal{O}(h^2) \qquad (9)\end{aligned}$$

The reason for this factorization is to clearly separate out what terms are of order h, $h^{\frac{3}{2}}$, and h^2. We know that

$$\bar{f}(x,t) = \lim_{h\to 0}\frac{1}{h}\left(\mathbb{E}[Y_{t+h} - Y_t|X_t]\right)$$

and so we can substitute in our expression for Y_{t+h} to get,

$$\begin{aligned}&= \lim_{h\to 0}\mathbb{E}\Bigg[\nabla\eta(X_t)^T \left[f(X_t,t) + \frac{1}{\sqrt{h}}g(X_t,t)\epsilon\right] + \frac{1}{2}hf(X_t,t)^T \nabla^2\eta(X_t) f(X_t,t)\\ &\quad + \sqrt{h}[g(X_t,t)\epsilon]^T \nabla^2\eta(X_t) f(X_t,t) + \epsilon^T g^T(X_t,t) \nabla^2\eta(X_t) g(X_t,t)\epsilon|X_t\Bigg] \qquad (10)\end{aligned}$$

Using the fact that ϵ is independent of X_t, and that $\mathbb{E}(\epsilon^T A\epsilon) = \mathrm{Tr}(\epsilon)$, we have that

$$= \nabla\eta(X_t)^T f(X_t,t) + \mathrm{Tr}\left(g(X_t,t)^T \nabla^2\eta(X_t) g(X_t,t)\right) \qquad (11)$$

A similar argument for the variance gives

$$\begin{aligned}
\bar{g}(x,t)^2 &= \lim_{h\to 0}\frac{1}{h}\mathrm{Var}[Y_{t+h}-Y_t|X_t] \\
&= \lim_{h\to 0}\frac{1}{h}\mathrm{Var}\left[\nabla\eta(X_t)^T\sqrt{h}g(X_t,t)\epsilon+\sqrt{h}h[g(X_t,t)\epsilon]^T\nabla^2\eta(X_t)f(X_t,t)+h\epsilon^T g(X_t,t)^T\nabla^2\eta(X_t)g(X_t,t)\epsilon\Big|\, X_t\right] \\
&= \lim_{h\to 0}\mathrm{Var}\left[\nabla\eta(X_t)^T g(X_t,t)\epsilon+h[g(X_t,t)\epsilon]^T\nabla^2\eta(X_t)f(X_t,t)+\sqrt{h}\epsilon^T g(X_t,t)^T\nabla^2\eta(X_t)g(X_t,t)\epsilon\Big|\, X_t\right] \\
&=\mathrm{Var}\left[\nabla\eta(X_t)^T g(X_t,t)\epsilon|X_t\right] \\
&=\nabla\eta(X_t)^T g(X_t,t)[\nabla\eta(X_t)^T g(X_t,t)]^T \\
&=\nabla\eta(X_t)^T g(X_t,t)g(X_t,t)^T\nabla\eta(X_t)
\end{aligned}$$

In going from the first to the second line, we use the definition of Y_{t+h}, and in going from the third to the fourth line, we use the fact that $\mathrm{Var}(A\epsilon)$ for some standard normally distributed ϵ is the covariance matrix, AA^T.

This concludes our derivation of Itô's Lemma.

1.2 Deriving the Fokker-Planck Equation

Armed with Itô's Lemma, the next task becomes to derive the Fokker Planck equation. Let us fix a training sample, x_0. We will define $\eta_k(x)\sim\mathcal{N}(x_0,\frac{1}{k^2})$ to be a Gaussian kernel centered at x_0.

For some arbitrary, smooth function F, we have that

$$F(x_0)=\lim_{k\to\infty}\int\eta_k(x)F(x)dx \tag{12}$$

This is a direct consequence of the sifting property of the Dirac δ-function. As $k\to\infty$, the Gaussian limits to a Dirac δ, and so we may apply this sifting property. In particular, for $F=p_t$, we have

$$\begin{aligned}
p_t(x_0) &= \lim_{k\to\infty}\int\eta_k(x)p_t(x)dx \\
&= \lim_{k\to\infty}\mathbb{E}[\eta_k(X_t)]
\end{aligned}$$

By Itô's Lemma, we can come up with the stochastic differential equation that $Y_t^k=\eta_k(X_t)$ follows.

$$\frac{d}{dt}\mathbb{E}[\eta_k(X_t)] =\mathbb{E}\left[\nabla\eta_k(X_t)^T f(X_t,t)+\mathrm{Tr}(g(t)^T\nabla^2\eta_k(X_t)g(t))\right]$$

Note that we can ignore the stochastic term here, because we are taking an expectation, and the stochastic term is mean-zero.

All that remains is to compute this expectation:

$$
\begin{aligned}
&\frac{d}{dt}p_t(x_0)\\
=&\frac{d}{dt}\lim_{k\to\infty}\mathbb{E}[\eta_k(X_t)]\\
=&\lim_{k\to\infty}\frac{d}{dt}\mathbb{E}[\eta_k(X_t)]\\
=&\lim_{k\to\infty}\mathbb{E}\left[\nabla\eta_k(X_t)^T f(X_t,t)+\mathrm{Tr}(g(X_t,t)^T\nabla^2\eta_k(X_t)g(X_t,t))\right]\\
=&\lim_{k\to\infty}\int\left[\nabla\eta_k(x)^T f(x,t)+\mathrm{Tr}(g(X_t,t)^T\nabla^2\eta_k(x)g(X_t,t))\right]p_t(x)dx\\
=&\lim_{k\to\infty}\left[\int\nabla\eta_k(x)^T f(x,t)p_t(x)dx+\int\mathrm{Tr}(g(X_t,t)^T\nabla^2\eta_k(x)g(X_t,t))p_t(x)dx\right]\\
=&\lim_{k\to\infty}\left[-\int\eta_k(x)^T\nabla\cdot[f(x,t)p_t(x)]dx+\int\mathrm{Tr}(g(X_t,t)^T\nabla^2\eta_k(x)g(X_t,t))p_t(x)dx\right]\quad\text{(using partial integration)}\\
=&-\nabla\cdot[f(x_0,t)p_t(x_0)]+\lim_{k\to\infty}\int\mathrm{Tr}(g(X_t,t)^T\nabla^2\eta_k(x)g(X_t,t))p_t(x)dx
\end{aligned}
$$

The second term can be simplified as

$$
\begin{aligned}
&\int\mathrm{Tr}(g(x,t)^T\nabla^2\eta_k(x)g(x,t))p_t(x)dx\\
=&\sum_{l=1}^{d}\int\left[g(t)^T\nabla^2\eta_k(x)g(t)\right]_{ll}p_t(x)dx\\
=&\sum_{l=1}^{d}\int p_t(x)\sum_{j=1}^{d}g_{jl}(t)[\nabla^2\eta_k(x)g(t)]_{jl}dx\\
=&\sum_{l=1}^{d}\sum_{j=1}^{d}\sum_{i=1}^{d}\int\left[p_t(x)g_{jl}(t)g_{il}(t)\right]\frac{\partial^2}{\partial x_j\partial x_i}\eta_k(x)dx\\
=&\sum_{l=1}^{d}\sum_{j=1}^{d}\sum_{i=1}^{d}\int\eta_k(x)\frac{\partial^2}{\partial x_j\partial x_i}\left[p_t(x)g_{jl}(t)g_{il}(t)\right]dx\quad\text{(using partial integration twice)}\\
\to&\sum_{l=1}^{d}\sum_{j=1}^{d}\sum_{i=1}^{d}\frac{\partial^2}{\partial x_j\partial x_i}\left[p_t(x_0)g_{jl}(x_0,t)g_{il}(x_0,t)\right]\quad(\text{for } k\to\infty)\\
=&\sum_{j=1}^{d}\sum_{i=1}^{d}\frac{\partial^2}{\partial x_j\partial x_i}\left[p_t(x_0)[g(x_0,t)g^T(x_0,t)]_{ji}\right]
\end{aligned}
$$

This is the Fokker-Planck equation for the probability flow under our SDE. In essence, what we did was we took our initial distribution, and we applied a Dirac-delta to it, and then used Itô's lemma to derive an expression for the time evolution of our sample.

1.3 Applications to Diffusion

That was a lot of math. Let's take a step back and bring this back to diffusion models. Our goal is to come up with an ODE to sample from, instead of a stochastic differential equation. All that remains now is to then take the Fokker-Planck equation, and apply it to our time-reversed SDE for our diffusion model.

$$dX_t = f(X_t, t)dt + g(t)dW_t$$

If we start with a distribution, $X_0 \sim p_0$, and apply our SDE, we will get a stochastic process X_t, with some distribution. Now we ask, is there an *ordinary* differential equation of the form

$$dZ_t = F(Z_t, t)dt$$

such that if $Z_0 \sim X_0$, then $Z_t \sim X_t$. The answer is yes, as we will now derive. Applying the Fokker-Planck equation,

$$\begin{aligned}
\frac{d}{dt}p_t(x) &= -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left[f_i(x,t)p_t(x)\right] + \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}\left[[g(t)g(t)^T]_{ij} p_t(x)\right] \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left[f_i(x,t)p_t(x)\right] + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \sum_{j=1}^{n} \frac{1}{2}\left[[g(t)g(t)^T]_{ij}\right] \frac{\partial}{\partial x_j} p_t(x) \quad \text{(derivatives are linear)} \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left[f_i(x,t)p_t(x)\right] + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \sum_{j=1}^{n} \frac{1}{2}\left[[g(t)g(t)^T]_{ij}\right] \left[\frac{\partial}{\partial x_j} \log p_t(x)\right] p_t(x) \quad \text{(derivative of log)} \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left[\left[f_i(x,t) - \frac{1}{2}\sum_{j=1}^{n}\left[[g(t)g(t)^T]_{ij}\right]\left[\frac{\partial}{\partial x_j}\log p_t(x)\right]\right] p_t(x)\right] \quad \text{(derivatives are linear)} \\
&= -\nabla \cdot [F p_t](x)
\end{aligned}$$

For the function

$$F(x,t) = f(x,t) - \frac{1}{2}g(t)g(t)^T \nabla \log p_t(x) \quad \text{(SDE-2-ODE)}$$

So, we have constructed a formula for transforming our SDE into an ODE, which we can now use to sample from our diffusion model.

Recall that in diffusion, we sample by simulating the reverse SDE

$$d\bar{X}_t = [f(X_t, t) - g^2(t)\nabla \log p_t(\bar{X}_t)]dt + g(t)d\bar{W}_t$$

where f, g are user defined functions, and the score is learned during training. Using the SDE-2-ODE equation, we have that

$$d\bar{X}_t = [f(X_t, t) - \frac{1}{2}g^2(t)\nabla \log p_t(\bar{X}_t)]dt$$

Note that now there is no stochastic term, and so we have an ordinary differential equation. All we need access to is the score function, which is learned during training. Sampling in this manner yields equivalent results, but is significantly faster than sampling a diffusion model.

As a quick side note, this ODE formalism results in a close correspondence to what is known as a continuous flow model. These models are the continuous time limit of a normalizing flow.