Exercises MITP2025 – ML and QFT – Gert Aarts

Diffusion models and stochastic dynamics

In this exercise we will consider a system with one degree of freedom, $x(t)$, to work out the basics of stochastic differential equations (SDEs) as encountered in diffusion models. This exercise relies heavily on Refs. [1, 2] and I am indebted to Diaa Habibi, as well as Lingxiao Wang and Kai Zhou, for the joint work used in the preparation of this exercise.

1. Derivation of the Fokker-Plank equation (FPE).

   Consider the SDE

   $$\dot{x}(t) = \frac{1}{2}K[x(t),t] + g(t)\eta(t), \qquad \langle \eta(t)\eta(t')\rangle = \delta(t-t'). \tag{1}$$

   Here K is the (possibly time-dependent) drift and $g^2(t)$ is the (possibly time-dependent) diffusion coefficient. We want to derive the associated Fokker-Planck equation

   $$\partial_t p(x,t) = \frac{1}{2}\partial_x \left[g^2(t)\partial_x - K(x,t)\right] p(x,t), \tag{2}$$

   for the distribution $p(x,t)$, defined via

   $$\langle O[x(t)]\rangle_\eta = \int dx\, p(x,t)O(x), \tag{3}$$

   with $O(x)$ a generic observable. Here the subscript η denotes noise averaging and will be dropped from now on.

   To achieve this we consider the discretised process, with time $t = \epsilon n$,[1]

   $$x_{n+1} = x_n + \frac{1}{2}\epsilon K_n + \sqrt{\epsilon} g_n \eta_n, \qquad \langle \eta_n \eta_{n'}\rangle = \delta_{nn'}. \tag{4}$$

   Show that

   $$\begin{aligned}\langle O(x_{n+1})\rangle - \langle O(x_n)\rangle &= \langle O'(x_n)\delta_n + \frac{1}{2}O''(x_n)\delta_n^2 + \ldots\rangle \\ &= \frac{\epsilon}{2}\langle O'(x_n)K_n + g_n^2 O''(x_n)\rangle + \mathcal{O}(\epsilon^{3/2}),\end{aligned} \tag{5}$$

   where $\delta_n \equiv x_{n+1} - x_n$.

   In the $\epsilon \to 0$ limit, this gives

   $$\partial_t \langle O(x)\rangle = \frac{1}{2}\langle O'(x)K(x) + g^2(t)O''(x)\rangle. \tag{6}$$

   Use Eq. (3) to demonstrate that this yields the FPE (2) for $p(x,t)$.

[1] Sometimes, e.g. in stochastic quantisation, a convention is chosen where $\epsilon \to 2\epsilon$, which puts a $\sqrt{2}$ in the noise term and removes the factor of $1/2$ in front of the drift. This corresponds to a rescaling of time.

2. Stationary solution

   Consider the case of constant drift, derived from a potential $V(x)$, and constant diffusion, i.e.,

$$K[x(t),t] = K(x) = -V'(x), \qquad\qquad g^2(t) = T. \tag{7}$$

   Here T is not the final time in the diffusion model, but will be identified with the temperature. The FPE reads

$$\partial_t p(x,t) = \frac{1}{2}\partial_x \left[T\partial_x + V'(x)\right] p(x,t). \tag{8}$$

   Show that a stationary solution is given by

$$p(x) \sim e^{-V(x)/T}. \tag{9}$$

   One can show that this solution is (typically) reached exponentially fast [3]. This observation underpins stochastic quantisation, in which the drift is given by the derivative of the Euclidean action, $K(x) = -S'(x)$, and the stationary solution is the Euclidean path integral weight, $p(x) \sim e^{-S(x)/\hbar}$. The noise represents the "quantum fluctuations", with $T \to \hbar$.

3. Linear drift in the forward process of the diffusion model.

   For the forward process we consider a linear drift,

$$\dot{x}(t) = -\frac{1}{2}k(t)x(t) + g(t)\eta(t), \qquad\qquad \langle\eta(t)\eta(s)\rangle = \delta(t-s). \tag{10}$$

   Show that this equation is solved by

$$x(t) = x_0 f(t,0) + \int_0^t ds\, f(t,s)g(s)\eta(s), \qquad\qquad f(t,s) = e^{-\frac{1}{2}\int_s^t ds'\, k(s')}, \tag{11}$$

   where the initial condition $x(0) = x_0$ is determined by the target distribution $p_0(x_0)$.

   Show that the second moment $\mu_2(t)$ is given by

$$\mu_2(t) \equiv \langle x^2(t)\rangle = \mu_2(0)f^2(t,0) + \Xi(t), \qquad\qquad \Xi(t) = \int_0^t ds\, f^2(t,s)g^2(s). \tag{12}$$

   Here $\mu_2(0)$ is the second moment of the target distribution and the brackets denote an average over both this distribution and the noise. Time runs between 0 and T, $0 \leq t \leq T$.

   Note:

   - a typical choice in the variance-expanding scheme is $k(t) = 0$ and $g(t) = \sigma^{t/T}$. In this case the second moment (or variance) grows exponentially in t.
   - a typical choice in the variance-preserving scheme is $k(t) = g^2(t)$. In this case the variance goes to 1 exponentially fast, independent of the choice of $g(t)$.

4. Transition amplitude.

   The linear process above has the nice property that any update starting from $x(0) = x_0$, for arbitrary $0 \leq t \leq T$, can be expressed by a Gaussian transition kernel, i.e.,

   $$p(x,t|x_0) = \mathcal{N}\left(x; x_0 f(t,0), \Xi(t)\right), \tag{13}$$

   such that

   $$p(x,t) = \int dx_0\, p(x,t|x_0)p_0(x_0) \tag{14}$$

   satisfies the FPE. As always, the Gaussian or normal distribution is defined as

   $$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{15}$$

   To show this, demonstrate that $x(t) - x_0 f(t,0)$ is a random variable with mean zero and variance $\Xi(t)$, i.e,

   $$x(t) - x_0 f(t,0) = \sqrt{\Xi(t)}\,\eta(t), \qquad\qquad \eta(t) \sim \mathcal{N}(0,1), \tag{16}$$

   see previous question. This immediately gives Eq. (13). You may also show that Eq. (14) indeed satisfies the FPE (2) with a linear drift, but this is somewhat tedious.

5. Score.

   The score is given by $s(x,t) = \nabla \log p(x,t)$. Using Jensen's inequality,

   $$\varphi\left(\int_{-\infty}^{\infty} dx\, g(x)p(x)\right) \leq \int_{-\infty}^{\infty} dx\, \varphi(g(x))p(x), \tag{17}$$

   where $p(x)$ is a normalised distribution, $g(x)$ is some function and $\phi(.)$ is a convex function, show that

   $$\int dx_0\, \log\left[p(x,t|x_0)\right] p_0(x_0) \leq \log p(x,t), \tag{18}$$

   (the log is concave). To determine the score, the gradient is taken under the integral and it is understood that an average over the target distribution is carried out. The inequality is usually ignored. Show that the score is then given by

   $$s(x,t) = \nabla \log p(x,t) \to \nabla \log p(x,t|x_0) = -\frac{x - x_0 f(t,0)}{\Xi(t)} = -\frac{\eta(t)}{\sqrt{\Xi(t)}}. \tag{19}$$

   The final expression is computable, i.e., it does not require knowledge of $p_0(x_0)$ or $p(x,t)$. It can hence be used in the loss function.

6. Loss function.

   We want to approximate the "true" score $s(x,t)$ with a learned score $s_\theta(x,t)$, represented by some neural network with trainable parameters collectively indicated as θ. For this one minimises a loss function,

   $$\mathcal{L}(\theta) = \frac{1}{2}\int_0^T dt\, \mathbb{E}_{p(x,t)}\left[\lambda(t)\big|\big|s_\theta(x,t) - \nabla \log p(x,t)\big|\big|_2^2\right], \tag{20}$$

where the weight $\lambda(t)$ is usually chosen to be equal to the variance $\Xi(t)$ of the noise at time t. The expectation value is taken over the time-dependent distribution $p(x,t)$, including the target distribution $p_0(x_0)$ at $t = 0$.

Substitute Eq. (19) into the loss (20) to show that

$$\begin{aligned}\mathcal{L}(\theta) &= \frac{1}{2}\sum_{t=0}^{T} \mathbb{E}_{p(x,t)} \left[\Xi(t)\left|\left|s_\theta(x,t) + \frac{\eta(t)}{\sqrt{\Xi(t)}}\right|\right|_2^2\right] \\ &= \frac{1}{2}\sum_{t=0}^{T} \mathbb{E}_{p(x,t)} \left[\left|\left|s_\theta(x,t)\sqrt{\Xi(t)} + \eta(t)\right|\right|_2^2\right], \qquad \eta(t) \sim \mathcal{N}(0,1). \qquad (21)\end{aligned}$$

Here we also discretised the time interval. The thing to note here is that the score $s_\theta(x,t)$ can now be learned by minimising this loss function. The average is obtained by sampling configurations at different times t, starting from configurations provided by the target distribution at $t = 0$, which are subjected to different levels of noise. This procedure is known as score matching.

7. Gaussian mixture.

As an example, consider the Gaussian mixture

$$p_0(x_0) = \frac{1}{2}\left[\mathcal{N}(x_0;\mu_0,\sigma_0^2) + \mathcal{N}(x_0;-\mu_0,\sigma_0^2)\right]. \qquad (22)$$

This is our target distribution. Plot this distribution. Show that the time-dependent distribution under the forward process is given by

$$p(x,t) = \int dx_0\, p(x,t|x_0)p_0(x_0) = \frac{1}{2}\left[\mathcal{N}(x;\mu(t),\sigma^2(t)) + \mathcal{N}(x;-\mu(t),\sigma^2(t))\right], \qquad (23)$$

by performing the Gaussian integrals. Here

$$\mu(t) = \mu_0 f(t,0), \qquad \sigma^2(t) = \sigma_0^2 f^2(t,0) + \Xi(t). \qquad (24)$$

Sketch the time-dependent distribution, starting from $t = 0$. You may use that

- in the variance-expanding scheme $f(t,0) = 1$ for all t, while $\Xi(t)$ grows exponentially.
- in the variance-preserving scheme $f(t,0) \to 0$ while $\Xi(t) \to 1$, both exponentially.

Explain how the information from the target distribution is erased ("noising").

The corresponding backward process reads

$$\dot{x}(t) = -\frac{1}{2}k(t)x(t) - g^2(t)\nabla \log p(x,t) + g(t)\eta(t), \qquad (25)$$

where now $t = T \to 0$, i.e. time runs backwards. The extra term in the drift is due to the score, which can in this case be given analytically. Show that it is given by

$$s(x,t) = \nabla \log p(x,t) = -\frac{x}{\sigma^2(t)} + \frac{\mu(t)}{\sigma^2(t)}\tanh\left(\frac{\mu(t)x}{\sigma^2(t)}\right). \qquad (26)$$

The dynamics is no longer linear and cannot be solved analytically. However, the solution of the backward FPE is given by $p(x,t)$ run backwards, as shown in the lecture.

To solve Eq. (25) numerically, it is useful to write it in terms of $\tau = T - t$, such that $0 \leq \tau \leq T$, i.e.,

$$x'(\tau) = \frac{1}{2}k(T-\tau)x(\tau) + g^2(T-\tau)\nabla \log p(x, T-\tau) + g(T-\tau)\eta(\tau), \tag{27}$$

and then discretise it, as in Eq. (4). See Ref. [1] for details. The target distribution should emerge ("denoising").

Of course, in general the target distribution is more complicated and the score is not known analytically. The main task is then to learn the score using ML methods and assess its reliability, see Ref. [4] for an example in two-dimensional $\lambda\phi^4$ theory.

References

[1] G. Aarts, D. E. Habibi, L. Wang and K. Zhou, "On learning higher-order cumulants in diffusion models," Mach. Learn. Sci. Tech. **6** (2025) no.2, 025004 [2410.21212 [hep-lat]].

[2] G. Aarts, "Introductory lectures on lattice QCD at nonzero baryon number," J. Phys. Conf. Ser. **706** (2016) no.2, 022004, App. C [1512.05145 [hep-lat]].

[3] P.H. Damgaard and H. Hüffel, "Stochastic quantization," Phys. Rept. **152** (1987) 227.

[4] L. Wang, G. Aarts and K. Zhou, "Diffusion models as stochastic quantization in lattice field theory," JHEP **05** (2024) 060 [2309.17082 [hep-lat]].