[AI2613 Lecture 22] Kolmogorov Forward / Backward Equations

May 31, 2024

Today we develop differential equations governing the evolution of a continuous-time Markov chain. In particular, for a continuous Markov process like Langevin Dynamics

$$\mathrm{d}X_t = \nabla f(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

In this lecture, we will show that its stationary distribution is $p(x) \sim e^{-f(x)}$.

1 Kolmogorov Backward Equation

Recall in the last lecture we introduced the notion of the Markov semigroup $(Q_t)_{t\geq 0}$ and its generator $\mathcal{L}$.

The following equation captures the evolution of the function $f_t := Q_t f_0$.

Theorem 1 (Kolmogorov Backward Equation, KBE)**.**

$$\frac{\mathrm{d}}{\mathrm{d}t} Q_t f = \mathcal{L} Q_t f,$$

or equivalently

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t = \mathcal{L} f_t.$$

Proof.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} Q_t f &= \lim_{s\to 0} \frac{Q_{t+s} f - Q_t f}{s} \\ &= Q_t \lim_{s\to 0} \frac{Q_s f - f}{s} = Q_t \mathcal{L} f.\end{aligned}$$

□

You might be curious that why this is called a *backward* equation. What is moving backward? Later, we will introduce the *Kolmogorov forward equation*, and it will be clear from its statement that the forward equation describes how the density of X_t evolves when time is moving forward.

I will explain this in the next lecture, where a form of KBE involving certain backward probabilities will be used.

2 Kolmogorov Forward Equation

We now come back to our original view of the transition matrix for a discrete Markov chain: if we multiple P^t by an initial distribution π_0, we can get $\pi_t = (P^\top)^t \pi_0$, which is the distribution at the t-th step. We can ask similar questions for a continuous-time Markov:

(1) What corresponds to $P^\top$?

(2) How does p_t evolve when starting with distribution p_0?

Note that the conjugate matrix P^* of P is defined as follows: for any vectors x, y, $\langle Px, y\rangle = \langle x, P^* y\rangle$. If P is a real matrix, then $P^* = P^\top$. Similarly, when an inner product $\langle , \rangle$ is present, one can define the adjoint operator Q^* of Q as the one satisfying $\langle Qx, y\rangle = \langle x, Q^* y\rangle$. Its existence is a consequence of the Riesz representation theorem.

Our answer to the first question above is obtained by calculating $\mathbf{E}[f(X_t)]$ in two ways: Note that we have the inner product $\langle f, g\rangle = \int_{\mathbb{R}} f(x)g(x)\mathrm{d}x$ for $f, g \in L^2(\mathbb{R})$. On the one hand,

$$\mathbf{E}[f(X_t)] = \int_{\mathbb{R}} f(x)p_t(x)\mathrm{d}x = \langle f, p_t\rangle.$$

On the other hand,

$$\begin{aligned}
\mathbf{E}[f(X_t)] &= \mathbf{E}[\mathbf{E}[f(X_t)|X_0]] \\
&= \int_{\mathbb{R}} \mathbf{E}[f(X_t)|X_0 = x]\, p_0(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} Q_t f(x) p_0(x)\mathrm{d}x \\
&= \langle Q_t f, p_0\rangle = \langle f, Q_t^* p_0\rangle.
\end{aligned}$$

Thus $p_t = Q_t^* p_0$. We have an illuminating analogue of $(P^\top)^t$, the operator Q_t^*.

Now we answer the second question by calculate $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[f(X_t)]$ in two ways. On one hand,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[f(X_t)] &= \lim_{s\to 0}\frac{1}{s}\mathbf{E}[f(X_{t+s}) - f(X_t)] \\
&= \lim_{s\to 0}\frac{1}{s}\mathbf{E}[\mathbf{E}[f(X_{t+s} - f(X_t))|X_t]] \\
&= \lim_{s\to 0}\frac{1}{s}\int_{\mathbb{R}} \mathbf{E}[\mathbf{E}[f(X_{t+s} - f(X_t))|X_t = x]]\, p_t(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} \mathcal{L}f(x)p_t(x)\mathrm{d}x = \langle \mathcal{L}f, p_t\rangle = \langle f, \mathcal{L}^* p_t\rangle.
\end{aligned}$$

On the other hand,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[f(X_t)] = \frac{\mathrm{d}}{\mathrm{d}t}\langle f, p_t\rangle = \langle f, \frac{\mathrm{d}}{\mathrm{d}t}p_t\rangle.$$

Hence $\frac{\mathrm{d}}{\mathrm{d}t}p_t = \mathcal{L}^* p_t$. This identity is called *Kolmogorov forward equation* by probabilists or *Fokker-Planck equation* by physicists.

Definition 2 (Kolmogorov forward equation, KFE)**.**

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t = \mathcal{L}^* p_t$$

Its counterpart for the discrete-time Markov chain is the following simple identity:

$$\pi_t - \pi_{t-1} = (P - I)^\top \pi_{t-1}.$$

3 Distribution Evolution of a Diffusion

Consider the time-independent Markov process:

$$\mathrm{d}X_t = \mu(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t.$$

We will study the evolution of the density $p_t(x)$ of X_t. We already know that p_t satisfies KFE and it remains to find the operator $\mathcal{L}^*$.

We already know $\mathcal{L}$ from the last lecture:

$$\begin{aligned}
\mathcal{L}f(x) &= \lim_{t\to 0} \frac{\mathbf{E}\left[f(X_t)|X_0 = x\right] - f(x)}{t} \\
&= \lim_{t\to 0} \frac{1}{t}\mathbf{E}\left[f(X_t) - f(X_0)|X_0 = x\right] \\
&= \mathbf{E}\left[\frac{f(X_{\mathrm{d}t}) - f(X_0)}{\mathrm{d}t}|X_0 = x\right] \quad \left(\text{Recall } \mathrm{d}f(X_t) = \left(\mu(X_t)f'X_t + \frac{\sigma^2(X_t)}{2}f''(X_t)\right)\mathrm{d}t + \sigma(X_t)f'\mathrm{d}B_t\right) \\
&= f'(x)\mu(x) + \frac{1}{2}f''(x)\sigma^2(x),
\end{aligned}$$

thus

$$\mathcal{L}(\cdot) = \mu(x)\frac{\partial}{\partial x}(\cdot) + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}(\cdot).$$

One can use some rules for adjoint operators such as $(A+B)^* = A^* + B^*$, $\left(\frac{\partial}{\partial x}\right)^* = -\frac{\partial}{\partial x}$ to obtain $\mathcal{L}^*$ from $\mathcal{L}$ directly. Nevertheless, we use the definition to calculate $\mathcal{L}^*$ here. Fix a twice-differentiable function f. On the one hand,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[f(X_t)\right] = \frac{\mathrm{d}}{\mathrm{d}t}\langle p_t, f\rangle = \langle \frac{\mathrm{d}}{\mathrm{d}t}p_t, f\rangle = \langle \mathcal{L}^* p_t, f\rangle,$$

where the last identity follows from KFE. On the other hand,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[f(X_t)\right] &= \lim_{s\to 0}\mathbf{E}\left[\frac{f(X_{t+s}) - f(X_t)}{s}\right] \\
&= \lim_{s\to 0}\mathbf{E}\left[\mathbf{E}\left[\frac{f(X_{t+s}) - f(X_t)}{s} \,\middle|\, X_t\right]\right] \\
&= \mathbf{E}\left[f'(X_t)\mu(X_t) + \frac{1}{2}f''(X_t)\sigma(X_t)\right],
\end{aligned}$$

where the last identity follows from Itô's formula. Using the formula of integration by parts, we obtain

$$\begin{aligned}
\mathbf{E}\left[f'(X_t)\mu(X_t)\right] &= \int f'(x)\mu(x)p_t(x)\mathrm{d}x \\
&= \mu(x)p_t(x)f(x)\Big|_{-\infty}^{\infty} - \int f(x)\frac{\partial}{\partial x}\left(\mu(x)p_t(x)\right)\mathrm{d}x \quad (f(x) \in C_0) \\
&= -\langle f, \frac{\partial}{\partial x}(\mu \cdot p_t)\rangle,
\end{aligned}$$

where the last inequality follows from the fact that $p_t(x)$ vanishes at infinity. Similarly, using the formula of integration by parts twice, we obtain

$$\begin{aligned}\mathbf{E}\left[f''(X_t)\sigma^2(X_t)\right] &= \int f''(x)\sigma(x)p_t(x)\mathrm{d}x \\ &= -\int f'(x)\frac{\partial}{\partial x}\left(\sigma^2(x)p_t(x)\right)\mathrm{d}x \\ &= \int \frac{\partial^2}{\partial x^2}\left(\sigma^2(x)p_t(x)\right)\mathrm{d}x \\ &= \langle f, \frac{\partial^2}{\partial x^2}\left(\sigma^2 p_t\right)\rangle,\end{aligned}$$

thus $\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{E}\left[f(X_t)\right] = \langle -\frac{\partial}{\partial x}\mu p_t + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2 p_t\right), f\rangle$. Hence

$$\mathcal{L}^*(g) = -\frac{\partial}{\partial x}\mu g + \frac{1}{2}\frac{\partial^2}{\partial x^2}\sigma^2 g.$$

Therefore the evolution of p_t for Diffusion is

$$\frac{\partial}{\partial t}p_t(x) = -\frac{\partial}{\partial x}\left(\mu(x)p_t(x)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2(x)p_t(x)\right), \tag{1}$$

which is usually referred to as the KFE for Itô diffusion.

4 *Computing Stationary Distribution*

If $\pi(x)$ is the stationary distribution of a diffusion, by plugging it into (1), we obtain

$$\frac{\partial}{\partial x}\left(\mu(x)\pi(x)\right) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2(x)\pi(x)\right). \tag{2}$$

Now let us see some familiar diffusion examples.

Example 1 (Langevin Dynamics)**.** *The SDE is*

$$\mathrm{d}X_t = -f'(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

Substitute μ and σ in Equation (2) with the above equation to obtain

$$\frac{\partial}{\partial x}\left(-f'(x)\pi(x)\right) = \frac{\partial^2}{\partial x^2}\pi(x).$$

Cancel out a differentiation operation:

$$-f'(x)\pi(x) = \pi'(x),$$

which can be solved as

$$\pi(x) = C\cdot e^{-f(x)}. \tag{3}$$

The Ornstein-Uhlenbeck process is a special case of Langevin dynamics, whose stationary distribution is Gaussian.

Example 2 (Ornstein-Uhlenbeck Process)**.** *The SDE is*

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

Substituting $f(x)$ in Equation (3) with $f'(x) = x$, or equivalently, $f(x) = \frac{x^2}{2}$, we obtain

$$\pi(x) = C\cdot e^{-\frac{x^2}{2}}.$$

5 Evolution of High Dimension Diffusion

Consider an n-dimensional diffusion

$$\mathrm{d}X_t = \mu(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t,$$

where $X_t, \mu(X_t), \mathrm{d}B_t \in \mathbb{R}^n$ and $\sigma(X_t) \in \mathbb{R}^{n\times n}$. Now we determine the generator $\mathcal{L}^*$. Similar to the one-dimensional case, we will calculate $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[f(X_t)\right]$ in two ways.

The first ingredient is the following Itô lemma for n-dimensional processes.

Proposition 3 (Itô formula)**.**

$$\mathrm{d}f(X_t) = \langle \boldsymbol{\nabla} f(X_t), \mu(X_t)\rangle \mathrm{d}t + \langle \boldsymbol{\nabla} f(X_t), \sigma(X_t)\mathrm{d}B_t\rangle + \frac{1}{2}\mathsf{Tr}\left(\sigma(X_t)^\top \boldsymbol{\nabla}^2 f(X_t)\sigma(X_t)\right)\mathrm{d}t.$$

Proof. We only give an informal justification, where several steps below follow from the heuristic identity $(\mathrm{d}B_t)_i(\mathrm{d}B_t)_j = \mathbb{1}[i=j]\cdot \mathrm{d}t$.

$$\begin{aligned}
\mathrm{d}f(X_t) &= \langle \boldsymbol{\nabla} f(X_t), \mathrm{d}X_t\rangle + \frac{1}{2}\langle \mathrm{d}X_t, \nabla^2 f(X_t)\mathrm{d}X_t\rangle \\
&= \langle \boldsymbol{\nabla} f(X_t), \mu(X_t)\rangle \mathrm{d}t + \langle \boldsymbol{\nabla} f(X_t), \sigma(X_t)\mathrm{d}B_t\rangle + \frac{1}{2}\langle \sigma(X_t)\mathrm{d}B_t, \boldsymbol{\nabla}^2 f(X_t)\sigma(X_t)\mathrm{d}B_t\rangle \\
&= \langle \boldsymbol{\nabla} f(X_t), \mu(X_t)\rangle \mathrm{d}t + \langle \boldsymbol{\nabla} f(X_t), \sigma(X_t)\mathrm{d}B_t\rangle + \frac{1}{2}\sum_{i,j}\left(\sum_k \sigma(X_t)_{i,k}\sigma(X_t)_{j,k}\right)\left[\boldsymbol{\nabla}^2 f(X_t)\right]_{i,j}\mathrm{d}t \\
&= \langle \boldsymbol{\nabla} f(X_t), \mu(X_t)\rangle \mathrm{d}t + \langle \boldsymbol{\nabla} f(X_t), \sigma(X_t)\mathrm{d}B_t\rangle + \frac{1}{2}\mathsf{Tr}\left(\sigma(X_t)^\top \boldsymbol{\nabla}^2 f(X_t)\sigma(X_t)\right)\mathrm{d}t.
\end{aligned}$$

□

Let f be a twice-differentiable function. Similar to the one-dimensional case, on one hand, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[f(X_t)\right] = \langle \frac{\mathrm{d}}{\mathrm{d}t}p_t, f\rangle = \langle \mathcal{L}^* p_t, f\rangle.$$

On the other hand, by Itô formula,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{E}\left[f(X_t)\right] &= \lim_{s\to 0}\mathbf{E}\left[\mathbf{E}\left[\frac{f(X_{t+s})-f(X_t)}{s}\;\middle|\; X_t\right]\right] \\
&= \mathbf{E}\left[\langle \boldsymbol{\nabla} f(X_t), \mu(X_t)\rangle + \frac{1}{2}\mathsf{Tr}\left(\sigma(X_t)^\top \boldsymbol{\nabla}^2 f(X_t)\sigma(X_t)\right)\right].
\end{aligned}$$

We will use the multi-dimensional integration by parts formula:

Proposition 4 (Multi-dimensional integration by parts formula)**.** *For a vector field* $\mathbf{F} = (F_1, \dots, F_n)$ *and a scalar function* u *defined on a region* $\Omega \subseteq \mathbb{R}^n$ *with sufficiently smooth boundary* $\partial\Omega$*.*

$$\int_\Omega \boldsymbol{\nabla}\cdot \mathbf{F}\, u\, \mathrm{d}\mathbf{x} = \int_{\partial\Omega} (\mathbf{F}\cdot\mathbf{n})\, u\, \mathrm{d}S - \int_\Omega \mathbf{F}\cdot\boldsymbol{\nabla} u\, \mathrm{d}\mathbf{x},$$

where $\mathbf{n}$ *is the outward-point unit normal vector on the boundary* $\partial\Omega$*, and* $\mathrm{d}S$ *is the surface measure on* $\partial\Omega$*.*

The *divergence theorem* states that

$$\int_{\Omega} \nabla \cdot \mathbf{F} \, \mathrm{d}\mathbf{x} = \int_{\partial\Omega} \mathbf{F} \cdot \mathbf{n} \, \mathrm{d}S \,.$$

Proof. First note the product rule for divergence:

$$\nabla \cdot (u\mathbf{F}) = u(\nabla \cdot \mathbf{F}) + \mathbf{F} \cdot (\nabla u).$$

We integrate over the region Ω and obtain:

$$\int_{\Omega} \nabla \cdot (u\mathbf{F}) \, \mathrm{d}\mathbf{x} = \int_{\Omega} (u(\nabla \cdot \mathbf{F}) + \mathbf{F} \cdot (\nabla u)) \, \mathrm{d}\mathbf{x} \,. \tag{4}$$

Applying the divergence theorem to the LHS, it becomes to

$$\int_{\partial\Omega} (u\mathbf{F}) \cdot \mathbf{n} \, \mathrm{d}S \,.$$

Plugging into (4) and rearranging, we obtain

$$\int_{\Omega} \nabla \cdot \mathbf{F} \, u \, \mathrm{d}\mathbf{x} = \int_{\partial\Omega} \mathbf{F} \cdot \mathbf{n} \, u \, \mathrm{d}\mathbf{x} - \int_{\Omega} \mathbf{F} \cdot \nabla u \, \mathrm{d}\mathbf{x} \,.$$

□

The integration by parts formula also holds *component-wise*, namely for each $i \in [n]$,

$$\int_{\Omega} \frac{\partial F_i}{\partial x_i} u \, \mathrm{d}\mathbf{x} = \int_{\partial\Omega} F_i \cdot n_i \, u \, \mathrm{d}S - \int_{\Omega} F_i \frac{\partial u}{\partial x_i} \, \mathrm{d}\mathbf{x} \,. \tag{5}$$

The proof is the same as above, using a component-wise divergence theorem.

Now we can manipulate the first term.

$$\begin{aligned} \mathbf{E}\left[\langle \nabla f(X_t), \mu(X_t) \rangle\right] &= \int_{\mathbb{R}^n} \langle \nabla f(\mathbf{x}), \mu(\mathbf{x}) \rangle p_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \sum_{i \in [n]} \int_{\mathbb{R}^n} \frac{\partial f(\mathbf{x})}{\partial x_i} \mu(\mathbf{x})_i p_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} \,. \end{aligned}$$

For each $i \in [n]$, applying eq. (5) and noticing that the probability $p_t(\mathbf{x})$ vanishes at infinity, we obtain

$$\int_{\mathbb{R}^n} \frac{\partial f(\mathbf{x})}{\partial x_i} \mu(\mathbf{x})_i p_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} = -\int_{\mathbb{R}^n} f(\mathbf{x}) \frac{\partial}{\partial x_i} (\mu(\mathbf{x})_i p_t(\mathbf{x})) \, \mathrm{d}\mathbf{x} \,.$$

Summing over all $i \in [n]$, we obtain

$$\mathbf{E}\left[\langle \nabla f(X_t), \mu(X_t) \rangle\right] = -\int_{\mathbb{R}^n} f(\mathbf{x}) \nabla \cdot (\mu(\mathbf{x}) p_t(\mathbf{x})) \, \mathrm{d}\mathbf{x} = -\langle f, \nabla \cdot (\mu \cdot p_t) \rangle.$$

For the second term, we can similarly have

$$\begin{aligned} \mathbf{E}\left[\mathsf{Tr}\left(\sigma(X_t)^\top \nabla^2 f(X_t) \sigma(X_t)\right)\right] &= \mathbf{E}\left[\sum_{i,j \in [n]} \frac{\partial^2 f(X_t)}{\partial x_i \partial x_j} \left[\sigma(X_t)\sigma(X_t)^\top\right]_{i,j}\right] \\ &= \sum_{i,j \in [n]} \int_{\mathbb{R}^n} \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \cdot \left[\sigma(\mathbf{x})\sigma(\mathbf{x})^\top\right]_{i,j} \mathrm{d}\mathbf{x} \,. \end{aligned}$$

For fixed i and j, integrating by parts twice, we obtain

$$\begin{aligned}\int_{\mathbb{R}^n} \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \cdot \left[\sigma(\mathbf{x})\sigma(\mathbf{x})^\top\right]_{i,j} p_t(\mathbf{x})\,\mathrm{d}\mathbf{x} &= -\int_{\mathbb{R}^n} \frac{\partial f(\mathbf{x})}{\partial x_j} \cdot \frac{\partial}{\partial x_i}\left([\sigma(\mathbf{x})\sigma(\mathbf{x})^\top]_{i,j} \cdot p_t(\mathbf{x})\right)\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^n} f(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j}\left([\sigma(\mathbf{x})\sigma(\mathbf{x})^\top]_{i,j} \cdot p_t(\mathbf{x})\right)\mathrm{d}\mathbf{x} \\ &= \langle f, \frac{\partial^2}{\partial x_i \partial x_j}\left([\sigma\sigma^\top]_{i,j} \cdot p_t\right)\rangle.\end{aligned}$$

Combining above, we obtain

$$\langle \mathcal{L}^* p_t, f\rangle = -\langle \boldsymbol{\nabla} \cdot (\mu \cdot p_t), f\rangle + \langle f, \frac{\partial^2}{\partial x_i \partial x_j}\left([\sigma\sigma^\top]_{i,j} \cdot p_t\right)\rangle.$$

This implies that

$$\mathcal{L}^* f(\mathbf{x}) = -\sum_{i\in[n]} \frac{\partial}{\partial x_i}(\mu(\mathbf{x}) \cdot f(\mathbf{x})) + \frac{1}{2}\sum_{i,j\in[n]} \frac{\partial^2}{\partial x_i \partial x_j}\left([\sigma(\mathbf{x})\sigma(\mathbf{x})^\top]_{i,j} \cdot f(\mathbf{x})\right).$$

References