IIIS 2023 Spring: ATCS - Selected Topics in Learning, Prediction and Optimization
Lecture date: February 27th, 2023
Instructor: Jian Li **Scribe: Yiran Liu**

1 Feyman-Kac Formula

Feyman-Kac formula is an important connection between SDE and parabolic PDE. Consider the following boundary value problem on $[0, T] \times \mathbb{R}$,

$$\begin{cases} \dfrac{\partial F}{\partial t}(t,x) + b(t,x)\dfrac{\partial F}{\partial x}(t,x) + \dfrac{1}{2}\sigma^2(t,x)\dfrac{\partial^2 F}{\partial x^2}(t,x) = 0, \\ F(T,x) = \quad (x). \end{cases} \tag{1}$$

In general, there is no closed form solution of the above PDE. Feyman-Kac formula says that the solution of the above PDE can be written as the expectation of certain stochastic process, de ned by the following SDE de ned over $[t, T]$:

$$\begin{cases} dX_s = b(s, X_s)ds + \sigma(s, X_s)dW_s, \\ X_t = x. \end{cases} \tag{2}$$

Then one can conclude that

$$F(t,x) = \mathbb{E}_{t,x}[\ \ (X_T)], \quad \text{for } t \le T$$

which is the Feyman-Kac formula. Here $\mathbb{E}_{t,x}$ means that the initial point of the stochastic process is $X_t = x$.

Proof: Consider the in nitesimal generator of the SDE:

$$A = b(t,x)\frac{\partial}{\partial x} + \frac{1}{2}\sigma^2(t,x)\frac{\partial^2}{\partial x^2}.$$

So, Eq. (1) can be rewritten as

$$\begin{cases} \dfrac{\partial F}{\partial t}(t,x) + AF(t,x) = 0, \\ F(T,x) = \quad (x). \end{cases} \tag{3}$$

By Ito's Lemma, we have

$$dF = (\frac{\partial F}{\partial t} + AF)ds + \sigma\frac{\partial F}{\partial x}dW_s = \sigma\frac{\partial F}{\partial x}dW_s. \tag{4}$$

$$F(T, X_T) = F(t,x) + \int_t^T \sigma\frac{\partial F}{\partial x}dW_s. \tag{5}$$

Take the expectation, we have $F(t,x) = \mathbb{E}_{t,x}[\ \ (X_T)]$, which is the Feyman-Kac formula. □

Komogorov Backward Equation (KBE): An important corollary of Feyman-Kac formula is that the function $F(t,x) = P(x_{t'} = x' | x_t = x)$ (for xed t', x' and $t' > t$) satis es the rst PDE in Eq equation 1. Namely, the following

$$\frac{\partial P(x_{t'} = x' | x_t = x)}{\partial t}(t,x) + b(t,x)\frac{\partial P(x_{t'} = x' | x_t = x)}{\partial x}(t,x) + \frac{1}{2}\sigma^2(t,x)\frac{\partial^2 P(x_{t'} = x' | x_t = x)}{\partial x^2}(t,x) = 0.$$

This equation is also known as Komogorov Backward Equation (KBE). One way to see this is as follows: for xed t' with $t' = T$, let $F(T,x) = \ (x) = \delta_{x'}(x)$. Hence, we can see that

$$F(t,x) = \mathbb{E}_{t,x}[\ (X_T)] = \mathbb{E}[\delta_{x'}(X_T) \mid X_t = x] = P(x_{t'} = x' | x_t = x).$$

2 Fokker Planck Equation

There is a closely related equation called Kolmogorov Forward Equation, which is also known as Fokker Planck Equation in physics literature.

2.1 Overview

Definition 1 (Hermitian adjoint) *Each linear operator A on a Euclidean vector space defines a Hermitian adjoint (or adjoint) operator A^* on that space according to the rule*

$$\langle Ax, y\rangle = \langle x, A^* y\rangle. \tag{6}$$

where $\langle \cdot, \cdot \rangle$ is the inner product on the vector space.

Definition 2 (Formal adjoint in one variable) *In the functional space of square-integrable functions on a real interval (a,b), the scalar product is defined by*

$$\langle f, g\rangle = \int_a^b \overline{f(x)} g(x) dx,$$

where $f(x)$ denotes the complex conjugate of $f(x)$. We assume that f or g are smooth and their values and derivatives vanish as $x \to a, x \to b$ (such functions are typically called test functions). Consider the differential operator T (T maps a function to a function), defined as follows:

$$Tu = \sum_{k=0}^{n} a_k(x) D^k u.$$

Here D is the differential operator. One can also define the adjoint *of the linear differential operator T as*

$$T^* u = \sum_{k=0}^{n} (-1)^k D^k [\overline{a_k(x)} u]. \tag{7}$$

It is not di cult to prove that Eq. equation 7 is indeed the adjoint (i.e., it satis es De nition 1. We only need to apply integral by part ($\int f \mathrm{d}g = fg - \int g \mathrm{d}f$) repeatedly and we leave it as an exercise.

Theorem 3 (Fokker Planck Equation / Kolmogorov Forward Equation) *Assume that $b(x)$ is C^1 and $\sigma(x)$ is C^2. For $\rho \in C^2$, define*

$$\mathcal{L}^*\rho(x) = -\sum_{i=1}^{n} \frac{\partial}{\partial x^i}(b^i(x)\rho(x)) + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{k=1}^{m} \frac{\partial^2}{\partial x^i \partial x^j}(\sigma^{ik}(x)\sigma^{jk}(x)\rho(x)). \tag{8}$$

Suppose that the density $p_t(x)$ exists and is C^1 in t, C^2 in x. Then

$$\frac{\partial}{\partial t}p_t(x) = \mathcal{L}^* p_t(x), t \in [0, T], \tag{9}$$

i.e. the density $p_t(x)$ of X_t must satisfy the Fokker Planck Equation (Kolmogorov Forward Equation).

Proof: Fix an $f \in C_0^2$ (in C^2 and with compact support). By Ito's rule, we obtain

$$f(X_t) = f(X_0) + \int_0^t \mathcal{L}f(X_s)ds + \text{martingale}. \tag{10}$$

(the last term is a martingale as f, and hence its derivatives, have compact support, and thus the integrand is bounded). Taking the expectation and using Fubini's theorem, we obtain

$$\mathbb{E}(f(X_t)) = \mathbb{E}(f(X_0)) + \int_0^t \mathbb{E}(\mathcal{L}f(X_s))ds. \tag{11}$$

Substituting the de nition of $p_t(y)$, integrating by parts, and using Fubini's theorem again, we have

$$\int_{\mathbb{R}^n} f(y)p_t(y)dy = \int_{\mathbb{R}^n} f(y)p_0(y)dy + \int_{\mathbb{R}^n} f(y)\int_0^t \mathcal{L}^* p_s(y)dsdy \tag{12}$$

Now note that this expression holds for any $f \in C_0^2$, so we can conclude that

$$a(y) = p_t(y) - p_0(y) - \int_0^t \mathcal{L}^* p_s(y)ds = 0 \tag{13}$$

for all y, except possibly on some subset with measure zero *w.r.t.* the Lebesgue measure. □

2.2 Ornstein-Uhlenbeck Process

Definition 4 (Ornstein-Uhlenbeck Process) *The Ornstein–Uhlenbeck process x_t is defined by the following stochastic differential equation:*

$$dx_t = -\theta\, x_t\, dt + \sigma\, dW_t \tag{14}$$

where $\theta > 0$ and $\sigma > 0$ are parameters and W_t denotes the Brownian motion process. See some examples in Figure 1. Note that the OU process is a stationary process.

Figure 1: Five simulations with $\quad = 1 ; \quad = 1$

The Ornstein{Uhlenbeck process can also be described in terms of a probability density function, $P(x;t)$, which speci es the probability of nding the process in the state x at time t. This function satis es the Fokker{Planck equation

$$\frac{@P}{@t} = \quad \frac{@}{@x}(xP) + D\frac{@^2P}{@x^2} \tag{15}$$

where $D = \ ^2=2$. This is a linear parabolic partial di erential equation which can be solved by a variety of techniques. The transition probability, also known as the Green's function, $P(x;t \mid x^0;t^0)$ is a Gaussian with mean $x^0e^{\ (t\ t^0)}$ and variance $\frac{D}{\quad}\ 1 \quad e^{\ 2\ (t\ t^0)}$:

$$P(x;t \mid x^0;t^0) = \sqrt{\frac{}{2\ D\ (1 \quad e^{\ 2\ (t\ t^0)})}} \exp\left[\frac{(x \quad x^0e^{\ (t\ t^0)})^2}{2D\ 1 \quad e^{\ 2\ (t\ t^0)}}\right] \tag{16}$$

This gives the probability of the state x occurring at time t given initial state x^0 at time $t^0 < t$. Equivalently, $P(x;t \mid x^0;t^0)$ is the solution of the Fokker{Planck equation with initial condition $P(x;t^0) = \ (x \quad x^0)$.

2.3 Heat Equation

Heat equation on Euclidean space is a special case of Fokker-Planck equation.

De nition 5 (Heat Equation) Given an open subset $U \quad R^n$ and a subinterval $I \quad R$, one says that a function $u : U \quad I \ ! \ R$ is a solution of the heat equation if

$$\frac{@u}{@t} = \quad r^{\ 2}u = \quad u \tag{17}$$

where $\quad$ is the Laplacian operator.

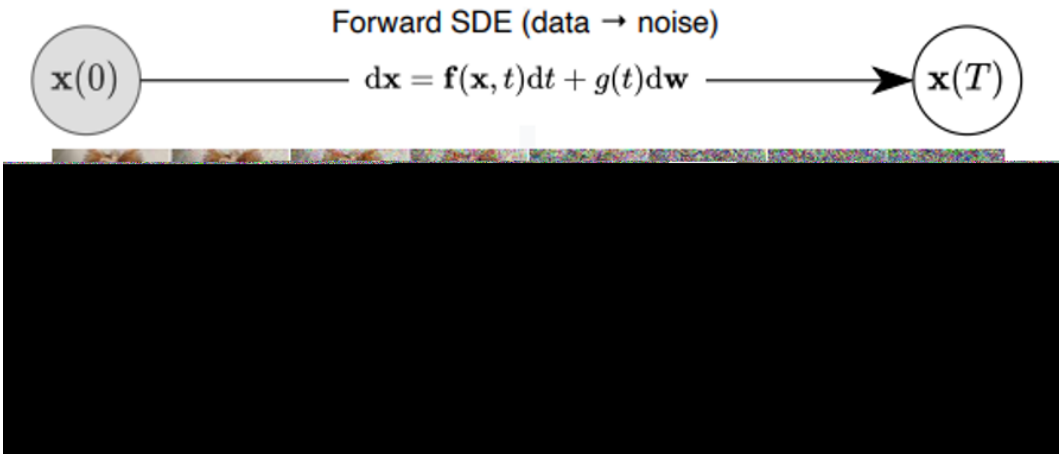

Figure 2: Di usion models based on the reverse process. Figure from [5].

We train a Noise Conditional Score Network (NCSN), denoted by $s_\theta(x, \sigma)$, with a weighted sum of denoising score matching objectives:

$$\theta^* = \arg\min \sum_{1}^{N} \sigma^2 \mathbb{E}_{p_{data}(x)} \mathbb{E}_{p_{\sigma_i}(\boldsymbol{x}|x)} [||s_\theta(\boldsymbol{x}, \sigma) - \nabla_{\boldsymbol{x}} \log p_{\sigma_i}(\boldsymbol{x}|x)||_2^2] \tag{25}$$

Given su cient data and model capacity, the optimal score-based model $s_{\theta^*}(x, \sigma_i)$ matches $\nabla_{\boldsymbol{x}} \log p_{\sigma_i}(\boldsymbol{x}|x)$ almost everywhere. For sampling, we run M steps Langevin MCMC to get a sample for each $p_{\sigma_i}(x)$ sequentially:

$$x_i^m = x_i^{m-1} + \epsilon_i s_{\theta^*}(x, \sigma_i) + \sqrt{2\epsilon_i} z_i^m \tag{26}$$

where $\epsilon_i > 0$is the step size, and z_i^m is standard normal. This can be seen as a discretization of the reversed SDE. The above is repeated for $i = N, N-1, \cdots, 1$ in turn with x_N^0 and $x \sim \mathcal{N}(x; 0, \sigma_{\text{max}}^2 I)$ and $x_i^0 = x_{i+1}^M$ when $i < N$. As $M \to \infty$ and $\epsilon_i \to 0$ for all i, x_1^M becomes an exact sample from $p_{\sigma_{\text{min}}} \approx p_{data}(x)$ under some regularity conditions.

4.2 Denoising Di usion Probabilistic Models (DDPM) [4]

For each training data point $x_0 \sim p_{data}(x)$, a discrete Markov chain $x_0, x_1, \cdots, x_N$ is constructed such that

$$p(x_i|x_{i-1}) = \mathcal{N}(x_i; \sqrt{1-\beta_i} x_{i-1}, \beta_i I),$$

and therefore $p(x_i|x_0) = \mathcal{N}(x_i; \sqrt{\alpha_i} x_0, (1-\alpha_i) I)$, where $0 < \beta_1, \cdots, \beta_N < 1, \alpha_i = \prod_{j=1}^{i} (1-\beta_j)$. In fact, this corresponds to a discretization of the following SDE (OU process)

$$dx = -\frac{1}{2}\beta(t) x dt + \sqrt{\beta(t)} dw.$$

Similar to SMLD, we can denote the perturbed data distribution as $p_{\alpha_i}(\boldsymbol{x}) = \int p_{data}(x) p_{\alpha_i}(\boldsymbol{x}|x) dx$. The noise scales are prescribed such that x_N is approximately distributed according to $\mathcal{N}(0, I)$. A

variational Markov chain in the reverse direction is parameterized with $p_\theta(x_{i-1}|x_i) = \mathcal{N}(x_{i-1}; \frac{1}{\sqrt{1-\beta_i}}(x_i + \beta_i s_\theta(x_i, i)), \beta_i I$, and trained with a re-weighted variant of the evidence lower bound (ELBO):

$$\theta^* = \arg\min \sum_{1}^{N} (1-\alpha_i)\mathbb{E}_{p_{data}(x)}\mathbb{E}_{p_{\alpha_i}(\boldsymbol{x}|x)}[||s_\theta(\boldsymbol{x}, i) - \nabla_{\boldsymbol{x}} \log p_{\alpha_i}(\boldsymbol{x}|x)||_2^2] \tag{27}$$

After solving the optimal model $s_{\theta^*}(\boldsymbol{x}, i)$, samples can be generated by starting from $x_N \sim \mathcal{N}(0, I)$ and following the estimated reverse Markov chain as below

$$x_{i-1} = \frac{1}{\sqrt{1-\beta_i}}(x_i + \beta_i s_{\theta^*}(\boldsymbol{x}, i)) + \sqrt{\beta_i} z_i \tag{28}$$

We call this method ancestral sampling, since it amounts to performing ancestral sampling from the graphical model $\prod_{i=1}^{N} p_\theta(x_{i-1}|x_i)$. The objective is also a weighted sum of denoising score matching objectives, which implies that the optimal model, $s_{\theta^*}(\boldsymbol{x}, i)$, matches the score of the perturbed data distribution, $\nabla_{\boldsymbol{x}} \log p_{\alpha_i}(\boldsymbol{x}|x)$. Notably, the weights of the i-th summand, namely σ_i^2 and α_i, are related to corresponding perturbation kernels in the same functional form: $\sigma_i^2 \propto \mathbb{E}[\nabla_{\boldsymbol{x}} \log p_{\sigma_i}(\boldsymbol{x}|x)]$ and $\alpha_i \propto \mathbb{E}[\nabla_{\boldsymbol{x}} \log p_{\alpha_i}]$.

4.3 ESTIMATING SCORES FOR THE SDE [5]

By starting from samples of $x_T \sim p_T$ and reversing the process, we can obtain samples $x_0 \sim p_0$. A remarkable result from Anderson[1] states that the reverse of a diffusion process is also a diffusion process, running backwards in time and given by the reverse-time SDE:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\overline{\mathbf{w}} \tag{29}$$

where $\overline{\mathbf{w}}$ is a standard Wiener process when time flows backwards from T to 0, and dt is an infinitesimal negative timestep. Once the score of each marginal distribution, $\nabla_x \log p_t(x)$, is known for all t, we can derive the reverse diffusion process from Eq. 29 and simulate it to sample from p_0.

The score of a distribution can be estimated by training a score-based model on samples with score matching. To estimate $\nabla_x \log p_t(x)$, we can train a time-dependent score-based model $s_\theta(x, t)$ via a continuous generalization to Eqs. 25 and 27:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_t\left\{\lambda(t)\mathbb{E}_{\mathbf{x}(0)}\mathbb{E}_{\mathbf{x}(t)|\mathbf{x}(0)}\left[\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0))\right]_2^2\right\} \tag{30}$$

Here $\lambda : [0, T] \to R_{>0}$ is a positive weighting function, t is uniformly sampled over $[0, T]$], $x(0) \sim p_0$ and $x_t \sim p_{0t}(x(t)|x(0))$. With sufficient data and model capacity, score matching ensures that the optimal solution to Eq. 30, denoted by $s_{\theta^*}(x, t)$, equals $\nabla_x \log p_t(x)$ for almost all x and t. As in SMLD and DDPM, we can typically choose

$$\lambda \propto 1/\mathbb{E}\left[\nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0))\right]_2^2. \tag{31}$$

References

1. Anderson B D O. Reverse-time diffusion equation models[J]. Stochastic Processes and their Applications, 1982, 12(3): 313-326.

2. https://www.vanillabug.com/posts/sde/

3. Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution[J]. Advances in neural information processing systems, 2019, 32.

4. Ho J, Jain A, Abbeel P. Denoising di usion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

5. Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic di erential equations[J]. arXiv preprint arXiv:2011.13456, 2020.