Stochastic differential equations in generative modeling. Diffusion models

Alexander Lobashev

Skolkovo Institute of Science and Technology

21 November 2022

Stochastic differential equations

Stochastic differential equation (SDE) is an integral equation of the form:

$$X_t = X_0 + \int_0^t \mu(X_s, s)ds + \int_0^t \sigma(X_s, s)dw_s \tag{1}$$

where X_t is an unknown stochastic process and

- $\int_0^t \mu(X_s, s)ds$ - Riemann integral
- $\int_0^t \sigma(X_s, s)dw_s$ - Ito integral

There is a short notation for the integral equation above:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dw_t \tag{2}$$

Stochastic differential equations

If the initial condition X_0 is a random variable with PDF $p(X|t=0)$ and we apply the SDE

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dw_t \tag{3}$$

then the probability density will evolve according to the corresponding Fokker-Planck equation

$$\frac{\partial p(x|t)}{\partial t} = -\frac{\partial}{\partial x}\left(\mu(x,t)p(x|t)\right) + \frac{\partial^2}{\partial x^2}\left(\frac{\sigma(x,t)^2}{2}p(x|t)\right) \tag{4}$$

Stochastic differential equations

Suppose that we have some distribution if initial conditions $X_0 \sim p(x|t=0)$ and we are adding noise to it via the SDE

$$dX_t = \mu(X_t, t)dt + dw_t \tag{5}$$

Which drift $\mu(X_t, t)$ should we choose to stay in the same initial distribution $p_0(x) = p(x|t=0)$?

- Example 1: a night our smartphone camera takes noisy photos, but we want to get sharp and high quality ones.
- Example 2: we want to reduce the noise in the measurement of some physical quantity.

Stochastic differential equations

$$dX_t = \mu(X_t, t)dt + dw_t \tag{6}$$

Which drift $\mu(X_t, t)$ should we choose to stay in the same initial distribution $p_0(x) = p(x|t = 0$?

- It is obvious that we need to somehow predict the noise to stay in the same distribution, but how exactly?
- Let's write Fokker-Planck equation:

$$\frac{\partial p(x|t)}{\partial t} = -\frac{\partial}{\partial x}\left(\mu(x,t)p(x|t)\right) + \frac{\partial^2}{\partial x^2}\left(\frac{\sigma(x,t)^2}{2}p(x|t)\right) \tag{7}$$

We have here $\sigma(x, t) = 1$ and also we want to get stationary density so $\frac{\partial p(x|t)}{\partial t} = 0$. Then

$$0 = -\frac{\partial}{\partial x}\left(\mu(x,t)p(x|t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(p(x|t)\right) \tag{8}$$

Stochastic differential equations

$$dX_t = \mu(X_t, t)dt + dw_t \tag{9}$$

Which drift $\mu(X_t, t)$ should we choose to stay in the same initial distribution $p_0(x) = p(x|t = 0)$?
We have here $\sigma(x, t) = 1$ and also we want to get stationary density so $\frac{\partial p(x|t)}{\partial t} = 0$. Then

$$0 = -\frac{\partial}{\partial x}\left(\mu(x, t)p(x|t) + \frac{1}{2}\frac{\partial}{\partial x}p(x|t)\right) \tag{10}$$

$$C(t) = -\mu(x, t)p(x|t) + \frac{1}{2}\frac{\partial}{\partial x}p(x|t) \tag{11}$$

Since C(t) is arbitrary let's set it to 0 for simplicity. Then

$$\mu(x, t) = \frac{1}{2p(x|t)}\frac{\partial}{\partial x}p(x|t) = \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t) \tag{12}$$

Stochastic differential equations

$$dX_t = \mu(X_t, t)dt + dw_t \tag{13}$$

Which drift $\mu(X_t, t)$ should we choose to stay in the same initial distribution $p_0(x) = p(x|t = 0)$?
Our answer is

$$\mu(x, t) = \frac{1}{2p(x|t)} \frac{\partial}{\partial x} p(x|t) = \frac{1}{2} \frac{\partial}{\partial x} \log p(x|t) \tag{14}$$

Our SDE with the noise compensation is then becomes:

$$dX_t = \frac{1}{2} \frac{\partial}{\partial x} \log p_0(X_t)dt + dw_t \tag{15}$$

Stochastic differential equations

Now we have and SDE which allows us to stay in the same initial distribution $p_0(x)$

$$dX_t = \frac{1}{2}\frac{\partial}{\partial x}\log p_0(X_t)dt + dw_t \tag{16}$$

Intuition: here we perform gradient optimization by minimizing negative log-likelihood to compensate added noise

$$dX_t = -\frac{1}{2}\nabla_x L(X_t)dt + dw_t \tag{17}$$

$L(x) = -\log p_0(x)$ - negative log-likelihood loss.

Now we have and SDE which allows us to stay in the same initial distribution $p_0(x)$

$$dX_t = \frac{1}{2}\frac{\partial}{\partial x}\log p_0(X_t)dt + dw_t \tag{18}$$

Example 1: if $p_0(x) = Ce^{-x^2}$ is Gaussian, then our SDE becomes Ornstein–Uhlenbeck process:

$$dX_t = \frac{1}{2}\frac{\partial}{\partial x}\log Ce^{-X_t^2}dt + dw_t =$$

$$= -\frac{1}{2}\frac{\partial}{\partial x}X_t^2 dt + dw_t$$

and then

$$dX_t = -X_t dt + dw_t$$

Stochastic differential equations

Now we have and SDE which allows us to stay in the same initial distribution $p_0(x)$

$$dX_t = \frac{1}{2}\frac{\partial}{\partial x}\log p_0(X_t)dt + dw_t \tag{19}$$

Example 2: if $p_0(x) = Ce^{-U(x)}$ is a particle in a force field $U(x)$.

$$dX_t = \frac{1}{2}\frac{\partial}{\partial x}\log Ce^{-U(X_t)}dt + dw_t =$$

$$= -\frac{1}{2}\frac{\partial}{\partial x}U(X_t)dt + dw_t$$

Note that we don't need to compute the partition function

$$C = \int e^{-U(x)}dx$$

to be able to sample from this distribution.
Note also that potential field equals to the negative log-likelihood loss.

Stochastic differential equations

Now we have and SDE which allows us to stay in the same initial distribution $p_0(x)$

$$dX_t = \frac{1}{2}\frac{\partial}{\partial x}\log p_0(X_t)dt + dw_t \tag{20}$$

The gradient $\frac{1}{2}\frac{\partial}{\partial x}\log p_0(X_t)$ equals to the added noise (not exactly, the denoised sample only must be from the same distribution).

- What if we only have samples from $p_0(x)$ and we don't know the density?

We could train a machine learning algorithm to denoise our samples. This could be done by using the loss function:

$$L(\theta) = \int p_0(X)\int p(\epsilon)||\epsilon_\theta(X+\epsilon) - \epsilon||^2 d\epsilon dX \tag{21}$$

Stochastic differential equations

- What if we only have samples from $p_0(x)$ and we don't know the density?

We could train a machine learning algorithm to denoise our samples. This could be done by using the loss function:

$$L(\theta) = \int p_0(X) \int p(\epsilon) ||\epsilon_\theta(X+\epsilon) - \epsilon||^2 d\epsilon dX \tag{22}$$

Here $\epsilon_\theta(x)$ is a neural network with parameters θ which tries to predict noise from the noisy samples $X + \epsilon$.

$$\epsilon_\theta(x) \approx -\frac{1}{2}\frac{\partial}{\partial x} \log p_0(X_t) \tag{23}$$

And we could sample new data from the same distribution by using our SDE

$$dX_t = \frac{1}{2}\frac{\partial}{\partial x} \log p_0(X_t)dt + dw_t \tag{24}$$

Probability flow ODE

We have considered an SDE

$$dX_t = \mu(X_t, t)dt + dw_t \tag{25}$$

Let's generalize our result to the case if $\sigma(x, t) = \sigma(t)$

$$dX_t = \mu(X_t, t)dt + \sigma(t)dw_t \tag{26}$$

From the Fokker-Planck equation we obtain

$$\mu(x, t) = \frac{1}{p(x|t)} \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} (p(x|t)) = \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \log p(x|t) \tag{27}$$

and the generalized density-conserving SDE becomes

$$dX_t = \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \log p(X_t|t)dt + \sigma(t)dw_t \tag{28}$$

Probability flow ODE

Consider an arbitrary SDE of the form

$$dX_t = \mu(X_t, t)dt + \sigma(t)dw_t \tag{29}$$

Density-conserving SDE

$$dX_t = \frac{\sigma(t)^2}{2}\frac{\partial}{\partial x}\log p(X_t|t)dt + \sigma(t)dw_t \tag{30}$$

Now let's consider an ordinary differential equation

$$dX_t = \left[\mu(X_t, t) - \frac{\sigma(t)^2}{2}\frac{\partial}{\partial x}\log p(X_t|t)\right]dt \tag{31}$$

and try to write a Fokker-Planck equation for it

$$\frac{\partial p(x|t)}{\partial t} = -\frac{\partial}{\partial x}\left(\left[\mu(X_t, t) - \frac{\sigma(t)^2}{2}\frac{\partial}{\partial x}\log p(X_t|t)\right]p(x|t)\right) \tag{32}$$

Probability flow ODE

Now let's consider an ordinary differential equation

$$dX_t = \left[\mu(X_t, t) - \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \log p(X_t|t) \right] dt \tag{33}$$

and try to write a Fokker-Planck equation for it

$$\frac{\partial p(x|t)}{\partial t} = -\frac{\partial}{\partial x} \left(\left[\mu(X_t, t) - \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \log p(X_t|t) \right] p(x|t) \right) =$$

$$= -\frac{\partial}{\partial x} \left[\mu(X_t, t) p(x|t) \right] + \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \left[p(x|t) \frac{\partial}{\partial x} \log p(X_t|t) \right],$$

$$p(x|t) \frac{\partial}{\partial x} \log p(X_t|t) = p(x|t) \frac{1}{p(X_t|t)} \frac{\partial}{\partial x} p(X_t|t) = \frac{\partial}{\partial x} p(X_t|t)$$

Probability flow ODE

Now let's consider an ordinary differential equation

$$dX_t = \left[\mu(X_t, t) - \frac{\sigma(t)^2}{2}\frac{\partial}{\partial x}\log p(X_t|t)\right] dt \tag{34}$$

We have obtained the Fokker-Planck equation

$$\frac{\partial p(x|t)}{\partial t} = -\frac{\partial}{\partial x}\left[\mu(X_t, t)p(x|t)\right] + \frac{\sigma(t)^2}{2}\frac{\partial^2}{\partial x^2}p(X_t|t)$$

Note that the same Fokker-Planck equations correspond to the SDE

$$dX_t = \mu(X_t, t)dt + \sigma(t)dw_t$$

Probability flow ODE

This ODE is called the probability flow ODE

$$dX_t = \left[\mu(X_t, t) - \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \log p(X_t|t)\right] dt \tag{35}$$

If we start from an initial distribution $p_0(x)$ then the density evolution will be the same as for the SDE

$$dX_t = \mu(X_t, t)dt + \sigma(t)dw_t \tag{36}$$

It means that if we want to numerically solve SDE we instead solve ODE using some advanced high-order solver.

Reverse time SDE

Let's add probability conserving SDE:

$$dX_t = \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \log p(X_t|t)dt + \sigma(t)dw_t \tag{37}$$

to the probability flow ODE:

$$dX_t = \left[\mu(X_t, t) - \frac{\sigma(t)^2}{2} \frac{\partial}{\partial x} \log p(X_t|t) \right] dt \tag{38}$$

We will obtain the original forward time SDE:

$$dX_t = \mu(X_t, t)dt + \sigma(t)dw_t \tag{39}$$

Reverse time SDE

Now consider $dt < 0$. In the probability conserving SDE we need to change the sign before dt. So probability-conserving SDE in reverse time becomes

$$dX_t = -\frac{\sigma(t)^2}{2}\frac{\partial}{\partial x}\log p(X_t|t)dt + \sigma(t)dw_t \tag{40}$$

If $dt < 0$ then probability flow ODE stays the same, since ODEs are reversible in time.

Now let's add reverse time ODE and SDE:

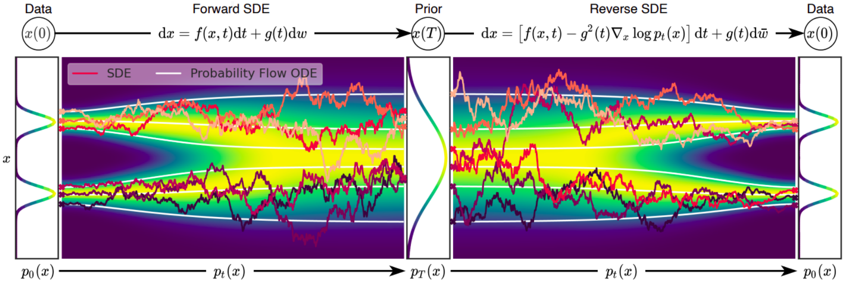
$$dX_t = \left[\mu(X_t, t) - \sigma(t)^2\frac{\partial}{\partial x}\log p(X_t|t)\right]dt + \sigma(t)dw \tag{41}$$

This is the SDE which corresponds the the reverse time probability flow ODE

$$dX_t = \left[\mu(X_t, t) - \frac{\sigma(t)^2}{2}\frac{\partial}{\partial x}\log p(X_t|t)\right]dt, \quad dt < 0 \tag{42}$$

Diffusion models

- Diffusion models leverage reverse time SDE to generate data from a given set of samples.



Good introduction to diffusion models "Understanding Diffusion Models: A Unified Perspective". https://arxiv.org/pdf/2208.11970.pdf