SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

September 1, 2022

Useful papers - [SSDK+20],

Contents

1 Introduction

In this paper authors considered generalization of the diffusion models on the continuous space by formulating corresponding stochastic differential equation (SDE). Such generalization gives a lot of interesting and useful features and also beat previous results in terms of FID.

The main idea is as follows: we know discrete form of forward (diffusion) process, then we can write its continuous form in SDE. As we will see it will consist of two parts: deterministic and stochastic. In simple words, deterministic - sample from previous step, stochastic - added noise. And now the most interesting: it turns out that we can write reverse SDE for forward process. It means that we can obtain real sample from noise data. Strictly speaking this is contrary to the laws of physics :). The second law of thermodynamics: the entropy of isolated systems left to spontaneous evolution cannot decrease. BUT in reverse process we decrease the entropy, because we delete noise from data. Maybe here our system is not isolated?

2 SDE framework

Before getting down to the stochastic diffusion models, we need to learn some basics of SDE. We know ordinary differential equation:

$$\begin{gathered} dx = f(x,t)dt \\ x(0) = 0 \end{gathered} \tag{1}$$

Such equations have unique trajectory. Also we can find reverse trajectory (in reverse time). For stochastic equations, we need to add stochastic term:

$$\begin{gathered} dx = f(x,t)dt + g(x,t)dW \\ x(0) \sim p_0(x) \end{gathered} \tag{2}$$

Here we dont have a unique trajectory, but we have a set of trajectories. This is because of the W (Wiener stochastic process). Let's look at it in more detail.

Stochastic process is a random variable but time-dependent, i.e., $X(\omega,t)$, where $\omega-$ random outcome, $t-$ time. So, if we consider a certain time, t_0, then $X(\omega,t_0)$ will be random variable. That‘s why stochastic equation has different trajectories (a random variable has different values at each moment of time). Wiener process has a following properties:

$$\begin{gathered} 1.W(0) = 0 \\ 2.\{W(t_{i+1}) - W(t_i)\}_{\forall i} - \text{set of independent random variables} \\ 3.W(t) - W(s) \sim \mathcal{N}(0, |t-s|), \forall t, s \end{gathered} \tag{3}$$

This process comes from diffusion in physics, where a lot of particles acts on certain particle and its coordinate is described by this process. Also, we know that sum of i.i.e. random variables have normal distribution.

Lets back to the SDE. We understand that we have a set of solutions not some specific one. So, we have to find all possible trajectories? Sounds sad. However, we know that random variable has a distribution. And this distribution changes with x, so it also depends on time, and we can write an equation on this distribution. It turns out that this equation is not stochastic (this called Fokker-Planck Equation). FPE can give us a lot of interesting features. Then lets derive it.

2.1 Derivation of the Fokker-Planck equation

Let $g(x,t) = g(t)$ and $x(0) \sim p_0(x)$:

$$\begin{gathered} dx = f(x,t)dt + g(t)dW \\ x(0) \sim p_0(x) \end{gathered} \tag{4}$$

We first derive deterministic part, i.e., for $g(t) = 0$. Then we will derive stochastic part, i.e., for $f(x,t) = 0$. After that we can put two obtained parts together.

Let $x(t) \sim p(x|t)$. We can try to understand how this distribution, $p(x|t)$, will change, if we change $x(t)$ a bit, i.e., consider $x(t+dt)$.

$$x(t+dt) - x(t) = dx = f(x,t)dt \to \hat{x} = x + f(x,t)dt \tag{5}$$

Our aim is to find $p(\hat{x}|t+dt)$. Reminder of the change of variables formula:

$$\begin{gathered} y = h(x), y \sim p(y), x \sim p(x), \\ p(y) = p(h^{-1}(y)) \left| \frac{dh}{dx} \right|^{-1} \end{gathered} \tag{6}$$

Then

$$\begin{aligned} p(\hat{x}|t+dt) &= p(\hat{x} - f(x,t)dt|t) \left(1 + \frac{\partial f}{\partial x}(x,t)dt\right)^{-1} \\ &\approx p\left(\hat{x} - \left[f(\hat{x}|t) + \frac{\partial f}{\partial x}(x-\hat{x}) + o(x-\hat{x})\right]|t\right)\left(1 + \frac{\partial f}{\partial x}(x,t)dt\right)^{-1} \\ &= p\left(\hat{x} - f(\hat{x}|t)dt + o(dt)|t\right)\left(1 - \frac{\partial f}{\partial x}dt + o(dt)\right) \end{aligned} \tag{7}$$

$$\begin{aligned} &\approx \left(p(\hat{x}|t) + \frac{\partial p}{\partial x}(\hat{x}|t)(\hat{x} - f(\hat{x}|t)dt - \hat{x}) + o(dt)\right)(1 - f(\hat{x},t)dt + o(dt)) \\ &= p(\hat{x}|t) - dt\left(\frac{\partial p(\hat{x}|t)}{\partial x} f(\hat{x},t) + p(\hat{x}|t)f(\hat{x},t)\right) + o(dt) \end{aligned} \tag{8}$$

So, finally

$$\begin{aligned} \frac{p(\hat{x}|t+dt) - p(\hat{x}|t)}{dt} &= -\frac{\partial}{\partial x}(p(\hat{x}|t)f(\hat{x},t)) + o(1) \\ \frac{\partial p(x|t)}{\partial t} &= -\frac{\partial}{\partial x}\left(p(x|t)f(x,t)\right) \end{aligned} \tag{9}$$

Thats fine, lets do the same with stochastic part, i.e. $f(x,t) = 0$. Again change $x(t)$ a bit, i.e., we want to find distribution for $x(t+dt)$

$$x(t+dt) - x(t) = dx = g(t)dW \to \hat{x} = x + g(t)dW \tag{10}$$

We remember that $W(t) - W(s) \sim \mathcal{N}(0, |t-s|)$, but $dW = W(t+dt) - W(t)$, then $dW \sim \mathcal{N}(0, dt)$. Based on this we can write the following:

$$\begin{gathered} x(t+dt) = x(t) + z(t), \\ \text{where } x(t) \sim p(x|t), z(t) \sim \mathcal{N}(0, g^2(t)dt) \end{gathered} \tag{11}$$

So, we have to calculate density of sum of two random independent variables, lets derive formula.

2.1.1 Density of sum of two random variables

$X \sim p(X), Y \sim p(Y); Z = X + Y, p(Z)-?$

$$\begin{gathered} \mathcal{F}(z) = \mathbb{P}(Z < z) = \mathbb{P}(X + Y < z) = \int\int_D p(x,y)dxdy = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} p(x,y)dy \\ p(z) = \frac{d}{dz}F(z) = \int_{-\infty}^{\infty} p(x, z-x)dx = \int_{-\infty}^{\infty} p(z-y, y)dy \end{gathered} \tag{12}$$

If two variables independent:

$$p(z) = \int_{-\infty}^{\infty} p(z-y)p(y)dy = p(x) * p(y) \tag{13}$$

OK, lets back to our case.

$$\begin{gathered} p(\hat{x}|t+dt) = p(x|t) * p(z) = \int p(\hat{x} - z|t)\mathcal{N}(z|0, g^2(t)dt)dz \\ \approx \int \left[p(\hat{x}|t) + \frac{\partial p}{\partial x}(\hat{x}|t)(-z) + \frac{1}{2}z^2\frac{\partial^2 p}{\partial x^2}(\hat{x}|t) + o(z^2) \right] \mathcal{N}(z|0, g^2(t)dt)dz \\ = p(\hat{x}|t) - \frac{\partial p}{\partial x}(\hat{x}|t)\,\mathbb{E}\,z + \frac{\partial^2 p}{\partial x^2}(\hat{x}|t)\mathbb{D}z + o(dt) \\ = p(\hat{x}|t) + \frac{\partial^2 p}{\partial x^2}(\hat{x}|t)g^2(t)dt + o(dt) \\ \frac{\partial p}{\partial t}(x|t) = \frac{1}{2}g^2(t)\frac{\partial^2 p}{\partial x^2}(x|t) \end{gathered} \tag{14}$$

Finally, the Fokker-Planck equation

$$\boxed{\frac{\partial p}{\partial t}(x|t) = -\frac{\partial}{\partial x}\left(p(x|t)f(x,t)\right) + \frac{1}{2}g^2(t)\frac{\partial^2 p}{\partial x^2}(x|t)} \tag{15}$$

So, it gave us the behaviour of the density distribution if x changes with respect to stochastic differential equation. As we can see, there is no stochasticity here.

2.2 Useful features from F-P equation

Let $g(t) = 1, f(x,t) = \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)$, then we can obtain Langevin dynamics

$$dx = \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)dt + dW \tag{16}$$

Let put it in F-P equation

$$\frac{\partial p}{\partial t}(x|t) = -\frac{\partial}{\partial x}\left(p(x|t)\frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)\right) + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}(x|t) = 0 \tag{17}$$

It means that Langevin dynamics do not change distribution of random variable.

Lets consider the following equation:

$$dx = f(x,t)dt + g(x,t)dW \rightarrow dx = \left(f(x,t) - \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)g^2(t)\right)dt \tag{18}$$

Then F-P equation:

$$\begin{aligned}\frac{\partial p}{\partial t}(x|t) &= -\frac{\partial}{\partial x}\left(f(x,t) - \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)g^2(t))\right)\\ &= -\frac{\partial}{\partial x}\left(p(x|t)f(x,t)\right) + \frac{1}{2}g^2(t)\frac{\partial^2 p}{\partial x^2}(x|t)\end{aligned} \tag{19}$$

We obtain **very important fact**: if our variable will evolve according to $dx = f(x,t)dt + g(x,t)dW$ then behaviour of its density will be the same if this variable would evolve according to $dx = \left(f(x,t) - \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)g^2(t))\right)dt$, but this equation is not stochastic. Summary:

$$\begin{gathered}dx = f(x,t)dt + g(x,t)dW - \textbf{Forward stochastic equation}\\ dx = \left(f(x,t) - \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)g^2(t)\right)dt - \textbf{Forward deterministic equation}\end{gathered} \tag{20}$$

As you remember in DDPM we also have a reverse process. So, can we obtain reverse stochastic equation? It turns out that yes. Reverse Langevin dynamics has the following form:

$$dx = -\frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)dt + dW \tag{21}$$

Here $dt < 0$, so now we need to move towards the anti-gradient, thats why we put minus. Lets add this dynamics to our deterministic equation, then

$$\begin{gathered}dx = f(x,t)dt + g(t)dW - \textbf{Forward equation}\\ dx = \left(f(x,t) - \frac{\partial}{\partial x}\log p(x|t)g^2(t)\right)dt + g(t)dW - \textbf{Reverse equation}\end{gathered} \tag{22}$$

We obtain **reverse stochastic equation**. It looks incomprehensible. That is, how the reverse equation works? If we obtain some trajectory in forward process, then the reverse equation gives us the same trajectory but in reverse? I think no, reversibility here means that if we start with density p_0 and after evolution obtain p_T, then if we start with p_T but with reverse evolution, then we will obtain p_0. In other words, here we understand the reversibility not according with trajectory, but with densities.

3 Continuous diffusion models

Finally, lets consider our main goal - score based models through stochastic differential equations. We will start with discrete case of diffusion and score based models and then generalize it to continuous one. The main parts of these models are: **forward process, learning** and **reverse process**. Lets write them and then generalize. Just note that we do not need to generalize the reverse process, because if we write continuous of the forward we immediately obtain the reverse, because we have equation for this process.

1. **FORWARD PROCESS**

   - **Score-based models**

$$\begin{aligned} q_{\sigma_i}(x_i|x_0) &= \mathcal{N}(x_i|x_0, \sigma_i^2 I) \\ q_{\sigma_i}(x_i|x_{i-1}) &= \mathcal{N}(x_i|x_{i-1}, (\sigma_i^2 - \sigma_{i-1}^2)I) \\ x_i &= x_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2} z_{i-1}, z_{i-1} \sim \mathcal{N}(z_{i-1}|0, I) \end{aligned} \tag{23}$$

   - **Diffusion models**

$$\begin{aligned} q_{\beta_i}(x_i|x_0) &= \mathcal{N}\left(x_i|\prod_{s=1}^{i}\sqrt{1-\beta_s}x_0, 1 - \prod_{s=1}^{i}(1-\beta_s)\right) \\ q_{\beta_i}(x_i|x_{i-1}) &= \mathcal{N}(x_i|\sqrt{1-\beta_i}x_{i-1}, \beta_i I) \\ x_i &= \sqrt{1-\beta_i}x_{i-1} + \sqrt{\beta_i}z_{i-1}, z_{i-1} \sim \mathcal{N}(z_{i-1}|0, I) \end{aligned} \tag{24}$$

   Here x_i, x_0 - noised and real objects, σ_i, β_i - parameters responsible for the magnitude of the noise.

2. **LEARNING**

   - **Score-based models**

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sigma_i^2 \, \mathbb{E}_{q_{\sigma_i}(x_i|x_0)p(x_0)} \left[||\mathbf{s}_\theta(x_i, \sigma_i) - \nabla_{x_i} \log q_{\sigma_i}(x_i|x_0)||^2\right] \tag{25}$$

   - **Diffusion models**

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} (1-\beta_i) \, \mathbb{E}_{q_{\beta_i}(x_i|x_0)p(x_0)} \left[||\mathbf{s}_\theta(x_i, i) - \nabla_{x_i} \log q_{\beta_i}(x_i|x_0)||^2\right] \tag{26}$$

3. **REVERSE PROCESS**

   - **Score-based models**

$$x_{i-1} = x_i + \epsilon \frac{1}{2} \frac{\sigma_i^2}{\sigma_0^2} \mathbf{s}_\theta(x_i, \sigma_i) + \sqrt{\epsilon \frac{\sigma_i^2}{\sigma_0^2}} z_i \tag{27}$$

   - **Diffusion models**

$$x_{i-1} = \frac{1}{\sqrt{\alpha_i}} \left(x_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}} \mathbf{s}_\theta(x_i, i)\right) + \sqrt{1-\alpha_i} z_i \tag{28}$$

     Note that here we assume that $\sigma_N > \sigma_{N-1} > ...$. But in the previous Song‘s paper we did opposite.

Lets start generalization with the forward process.

3.1 Continuous generalization of the forward process

For score-based models we can write

$$x_i = x_{i-1} + \sqrt{\frac{\sigma_i^2 - \sigma_{i-1}^2}{\Delta i}} z_{i-1} \sqrt{\Delta i} \tag{29}$$

We know that $dW = W(t + dt) - W(t) \sim \mathcal{N}(0, dt)$. So, if $\Delta i \to 0$, then

$$dx = \sqrt{\frac{d\sigma^2(t)}{dt}} dW \tag{30}$$

Lets do the same with diffusion model (we assume that $\beta_i = \beta_i \Delta i$)

$$\begin{aligned} x_i &= \sqrt{1 - \beta_i \Delta i} x_{i-1} + \sqrt{\beta_i \Delta i} z_{i-1} \\ &\approx x_{i-1} - \frac{1}{2} \beta_i \Delta i x_{i-1} + \sqrt{\beta_i} z_{i-1} \Delta i \end{aligned} \tag{31}$$

Then

$$dx = -\frac{1}{2} \beta(t) x dt + \sqrt{\beta(t)} dW \tag{32}$$

So, we obtain stochastic equations for reverse processes. Now, lets consider behaviour of the variance for the x variable.

3.1.1 Variance behaviour of x for SBM and DDPM equations

To derive behaviour of the variance we need to use Ito formula. First of all general view of the SDE is:

$$dx = f(x, t) dt + g(t) dW \tag{33}$$

Here $x, dW, f(x, t) \in \mathbb{R}^d$. Let $\phi(x, t) \in \mathbb{R}$ - some function. So, we can write Ito formula:

$$\begin{gathered} d\phi(x, t) = \frac{\partial \phi}{\partial t} dt + \sum_{i=1} \frac{\partial \phi}{\partial x_i} f_i dt + \frac{1}{2} g^2(t) \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j} dt + g(t) \sum_i \frac{\partial \phi}{\partial x_i} dW_i \\ d\,\mathbb{E}\,\phi = \mathbb{E}\,\frac{\partial \phi}{\partial t} dt + \sum_{i=1} \mathbb{E}\left(\frac{\partial \phi}{\partial x_i} f_i dt\right) + \frac{1}{2} g^2(t) \sum_{i,j} \mathbb{E}\left(\frac{\partial^2 f}{\partial x_i \partial x_j} dt\right) + g(t) \sum_i \mathbb{E}\left(\frac{\partial \phi}{\partial x_i} dW_i\right) \\ \frac{d\,\mathbb{E}\,\phi}{dt} = \mathbb{E}\,\frac{\partial \phi}{\partial t} + \sum_{i=1} \mathbb{E}\left(\frac{\partial \phi}{\partial x_i} f_i\right) + \frac{1}{2} g^2(t) \sum_{i,j} \mathbb{E}\left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right) \end{gathered} \tag{34}$$

Let $\phi = x_i$, then

$$\begin{aligned} \frac{d\,\mathbb{E}\,x_i}{dt} &= \mathbb{E}\,f_i \\ \frac{d\,\mathbb{E}\,x}{dt} &= \mathbb{E}\,f(x, t) \end{aligned} \tag{35}$$

If $\phi = x_i^2 - (\mathbb{E}\,x_i)^2$

$$\begin{aligned} \frac{d\mathbb{D}x_i}{dt} &= 2\,\mathbb{E}\left[x_i f_i - f_i\,\mathbb{E}\,x_i\right] + g^2(t) \\ \frac{d\mathbb{D}x}{dt} &= 2\,\mathbb{E}\left[x f(x, t) - f(x, t)\,\mathbb{E}\,x\right] + g^2(t) \end{aligned} \tag{36}$$

Lets start with SBM:

$$\begin{gathered} f(x,t) = 0, g(t) = \sqrt{\frac{d\sigma^2(t)}{dt}} \\ \frac{d\mathbb{D}x}{dt} = \frac{d\sigma^2(t)}{dt} \rightarrow \mathbb{D}x = \sigma^2(t) \end{gathered} \tag{37}$$

As we said earlier, $\{\sigma_i\}$ - increasing sequence, so if $t \to \infty$, then $\mathbb{D}x \to \infty$. This case was called Variance Exploding (VE).
For DDPM:

$$\begin{gathered} f(x,t) = -\frac{1}{2}\beta(t)x, g(t) = \sqrt{\beta(t)} \\ \frac{d\mathbb{D}x}{dt} = \beta(t)\left[I - \mathbb{D}x\right] \rightarrow \mathbb{D}x(t) = I + e^{-\int_0^t \beta(s)ds}\left[\mathbb{D}x(0) - I\right] \end{gathered} \tag{38}$$

As it can be seen, variance $\mathbb{D}x(t)$ is bounded if $\mathbb{D}x(0)$ is bounded. Moreover, if $\mathbb{D}x(0) = I$, then $\mathbb{D}x(t) = I$. It is called Variance Preserving (VP).

Authors also proposed to consider the following forward process:

$$dx = -\frac{1}{2}\beta(t)x + \sqrt{\beta(t)(I - e^{-2\int_0^t \beta(s)ds})}dW \tag{39}$$

Then equation for variance

$$\mathbb{D}x(t) = I + e^{-2\int_0^t \beta(s)ds}I + e^{-\int_0^t \beta(s)ds}\left[\mathbb{D}x(0) - 2I\right] \tag{40}$$

It can be seen that this variance is lower than variance obtained from DDPM, so it is called sub-Variance Preserving.

Lets do summary with continuous generalization of the forward process.

Variance Exploding equation

$$\boxed{dx = \sqrt{\frac{d\sigma^2(t)}{dt}}dW} \tag{41}$$

Variance Preserving equation

$$\boxed{dx = -\frac{1}{2}\beta(t)xdt + \sqrt{\beta(t)}dW} \tag{42}$$

sub-Variance Preserving equation

$$\boxed{dx = -\frac{1}{2}\beta(t)x + \sqrt{\beta(t)\left(I - e^{-2\int_0^t \beta(s)ds}\right)}dW} \tag{43}$$

3.2 Continuous generalization of the learning

Continuous generalization of the learning can be written as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{p(t)}\left[\lambda(t)\, \mathbb{E}_{q(x(t)|x(0),t)p(x(0))}\left[||\mathbf{s}_\theta(x(t),t) - \nabla_{x(t)}\log q(x(t)|x(0),t)||^2\right]\right] \tag{44}$$

Here $\lambda(t)$ - positive weighting function, $p(t)$ - uniform distribution, $\mathbb{U}([0,T])$. We need to understand the form of $q(x(t)|x(0),t)$. Previously, we derived formulas for calculation the evolution of mean and variance. Thus, we can write the following.
Variance Exploding case:

$$\begin{gathered}\frac{d}{dt}\,\mathbb{E}\,x(t) = 0 \rightarrow \mathbb{E}\,x(t) = \mathbb{E}\,x(0) \\ \frac{d}{dt}\mathbb{D}x(t) = \frac{d}{dt}\sigma^2(t) \rightarrow \mathbb{D}x(t) = \mathbb{D}x(0) + \sigma^2(t) - \sigma^2(0)\end{gathered} \tag{45}$$

But at the zero moment of time we have experimental distribution, i.e, dataset, $x(0) \sim \delta(x(0))$, then $\mathbb{E}\,x(0) = x(0), \mathbb{D}x(0) = 0$. Finally,

$$q(x(t)|x(0),t) = \mathcal{N}(x(t)|x(0), \left[\sigma^2(t) - \sigma^2(0)\right] I) \tag{46}$$

For other cases we can do same. Lets write transition kernel for all cases:

$$\boxed{q(x(t)|x(0),t) = \begin{cases} \mathcal{N}(x(t)|x(0), [\sigma^2(t) - \sigma^2(0)]\, I), & \textbf{(VE)} \\ \mathcal{N}\left(x(t)|x(0)e^{-\frac{1}{2}\int_0^t \beta(s)ds}, I - Ie^{-\int_0^t \beta(s)ds}\right) & \textbf{(VP)} \\ \mathcal{N}\left(x(t)|x(0)e^{-\frac{1}{2}\int_0^t \beta(s)ds}, \left[1 - e^{-\int_0^t \beta(s)ds}\right]^2 I\right) & \textbf{(sub-VP)} \end{cases}} \tag{47}$$

3.3 Continuous generalization of the reverse process

So, now we understood how to generalize forward process and training objective. If these steps is performed we need to make samples from the model. We remember that if the forward SDE is known then it is possible to write the reverse one:

$$\begin{gathered}dx = f(x,t)dt + g(t)dW - \textbf{Forward equation} \\ dx = \left(f(x,t) - \frac{\partial}{\partial x}\log p(x|t) g^2(t)\right) dt + g(t)dW - \textbf{Reverse equation}\end{gathered} \tag{48}$$

In other words, if we know equation for the forward process and score then we immediately obtain reverse equation for making samples. Lets write them for our three cases:
Variance Exploding equation

$$\boxed{\begin{gathered}dx = \sqrt{\frac{d\sigma^2(t)}{dt}}dW - \textbf{Forward} \\ dx = -\mathbf{s}_\theta(x(t),t)\frac{d\sigma^2(t)}{dt}dt + \sqrt{\frac{d\sigma^2(t)}{dt}}dW - \textbf{Reverse}\end{gathered}} \tag{49}$$

Variance Preserving equation

$$\boxed{\begin{gathered}dx = -\frac{1}{2}\beta(t)xdt + \sqrt{\beta(t)}dW - \textbf{Forward} \\ dx = -\left(\frac{1}{2}x + \mathbf{s}_\theta(x(t),t)\right)\beta(t) + \sqrt{\beta(t)}dW - \textbf{Reverse}\end{gathered}} \tag{50}$$

sub-Variance Preserving equation

$$\boxed{\begin{gathered} dx = -\frac{1}{2}\beta(t)x + \sqrt{\beta(t)\left(I - e^{-2\int_0^t \beta(s)ds}\right)}dW - \textbf{Forward} \\ dx = -\left(\frac{1}{2}x + \mathbf{s}_\theta(x(t),t)\left(I - e^{-2\int_0^t \beta(s)ds}\right)\right)\beta(t) + \sqrt{\beta(t)\left(I - e^{-2\int_0^t \beta(s)ds}\right)}dW - \textbf{Reverse} \end{gathered}} \tag{51}$$

Looks good but we have a problem: to obtain samples we have to solve the reverse equation. Note that in the forward process we do not need to solve forward equation, it's just enough for us to know expectation and variance behaviour of the variable that evolve according to forward equation. But here everything is more complicated.

So, as we know numerical solution of the DE assumes discretization. In the paper authors consider Euler-Maruyama method, that can be written as follows:

$$\begin{gathered} dx = \left(f(x,t) - \mathbf{s}(x,t)g^2(t)\right)dt + g(t)dW \\ x_i - x_{i+1} = -\left(f_{i+1}(x_{i+1}) - s_{i+1}(x_{i+1})g_{i+1}^2\right)\Delta t + g_{i+1}\sqrt{\Delta t}z_{i+1}, \Delta t = i + 1 - i \\ x_i = x_{i+1} - f_{i+1}(x_{i+1}) + s_{i+1}(x_{i+1})g_{i+1}^2 + g_{i+1}z_{i+1} \end{gathered} \tag{52}$$

Thus, we can solve reverse equation and obtain samples as a consequence. But we remember that we can apply MCMC methods. In other words, we can correct solution of the reverse SDE using Langevin MCMC. In the paper it is called Corrector.

3.3.1 Deterministic representation

We remember that it is possible to write deterministic variant of the stochastic reverse equation that gives the same evolution of the density as SDE:

$$dx = \left(f(x,t) - \frac{1}{2}\frac{\partial}{\partial x}\log p(x|t)g^2(t)\right)dt \tag{53}$$

This means that diffusion models it is normalizing flows but with fixed $f(x,t)$. That is, in normalizing flows we have $f_\theta(x,t)$ parameterized by neural network. There are some works that also learn $f_\theta(x,t)$ for diffusion models (normalizing diffusion models paper).

4 Summary

So, we have considered continuous generalization of the diffusion models. For forward process we obtained: VE, VP, sub-VP SDE. For each of them we can write continuous learning objective and reverse SDE. To solve reverse SDE we use Euler-Maruyama as predictor and Langevin MCMC as corrector. In the paper the best results was shown by VE SDE. So, lets write end-to-end algorithm for the latter.

1. Training VE SDE

$$\begin{gathered} \mathcal{L}(\theta) = \mathbb{E}_{p(t)}\left[\lambda(t)\,\mathbb{E}_{q(x(t)|x(0),t)p(x(0))}\left[||\mathbf{s}_\theta(x(t),\sigma(t)) - \nabla_{x(t)}\log q(x(t)|x(0),t)||^2\right]\right] \\ q(x(t)|x(0),t) = \mathcal{N}\left(x(t)|x(0), \sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t I\right) \end{gathered} \tag{54}$$

We have chosen $\sigma^2(t)$ as follows: we remember that $\{\sigma_i\}_{i=1}^N$ - geometric sequence. Then we can write $\sigma_i = \sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^{\frac{i-1}{N-1}} \rightarrow_{N\rightarrow\infty} \sigma(t) = \sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t, t \in [10^{-5}, 1], \sigma_{min} = 0.01, \sigma_{max}$ depends on the dataset. We typically choose
$\lambda(t) \sim 1/\,\mathbb{E}_{q(x(t)|x(0),t)}\left[||\nabla_{x(t)}\log q(x(t)|x(0),t)||^2\right]$

2. Sampling VE SDE

Algorithm 1 Corrector-Predictor VE SDE

```
Require: $r = 0.16, N = 1000, \{\sigma_i\}_{i=1}^N, x_N \sim \mathcal{N}(x_N|0, \sigma_N I)$
for $i = N-1$ to $1$ do
    $x_i' = x_{i+1} + \left(\sigma_{i+1}^2 - \sigma_i^2\right)\mathbf{s}_\theta(x_{i+1}, \sigma_{i+1}) + \sqrt{\sigma_{i+1}^2 - \sigma_i^2}\,z, z \sim \mathcal{N}(z|0, I)$
    $\epsilon_i = 2r\left(||z||/||\mathbf{s}_\theta(x_i', \sigma_i)||\right)^2, z \sim \mathcal{N}(z|0, I)$
    $x_i = x_i' + \epsilon_i \mathbf{s}_\theta(x_i', \sigma_i) + \sqrt{2\epsilon_i}\,z$
return $x_1$
```

3. Architecture details for $\mathbf{s}_\theta(x(t), \sigma(t))$ (see code ncsnpp.py)

First of all it is necessary to add: gradient clipping, learning rate warm-up schedules, EMA (exponential moving average). The latter has a significant impact on performance and it can be written as follows:

$$\theta' = m\theta' + (1-m)\theta_i \tag{55}$$

where θ_i - parameters after the i-th training iteration, θ' - be an independent copy of the parameters. In the paper authors propose $m = 0.999$.
Neural Network architecture:

- Mostly based on UNet from the DDPM paper and StyleGAN2 features.
- Instead of positional embedding using Fourier feature embedding:

$$\begin{gathered}\sigma_{emb}(t) = 2\pi(\sigma(t)\cdot W);\\ \sigma(t) \in \mathbb{R}^{b\times 1}, W \in \mathbb{R}^{1\times emb}, \sigma_{emb}(t) \in \mathbb{R}^{b\times emb}\\ \sigma_{emb}(t) = \mathbf{CAT}\left[\cos(\sigma_{emb}(t)), \sin(\sigma_{emb}(t))\right]\end{gathered} \tag{56}$$

- Upsampling and downsampling images with anti-aliasing based on Finite Impulse Response (FIR) [Zha19]. It is known that convolution is shift-equivariant and max pooling is shift invariant. But because of the latter we lose shift-equivariance. This approach is aim to solve this problem.
- Rescaling all skip connections in residual blocks by $1/\sqrt{2}$ (o_o):

$$x + h \rightarrow (x+h)/\sqrt{2} \tag{57}$$

- Replacing classical residual block with BigGAN residual block 1 .Instead of BatchNorm authors used GroupNorm.
- Increasing the number of residual blocks per resolution from 2 to 4 and to 8 (called deep)
- Use residual progressive for input.

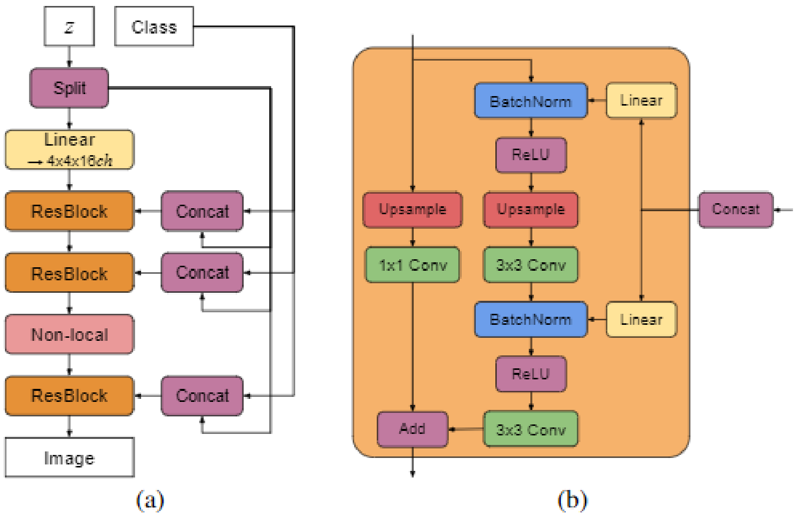

Figure 1: Residual block in BigGAN

References

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Zha19] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.