Generative Models

September 2023

Contents

1 Introduction

We will be looking at two broad classes (not exhaustive) of generative models: one that uses stochastic differential equations (SDE) and the other that uses ordinary differential equations (ODE).

2 Stochastic Calculus

Consider a differential equation

$$dx = f(x)dt \tag{1}$$

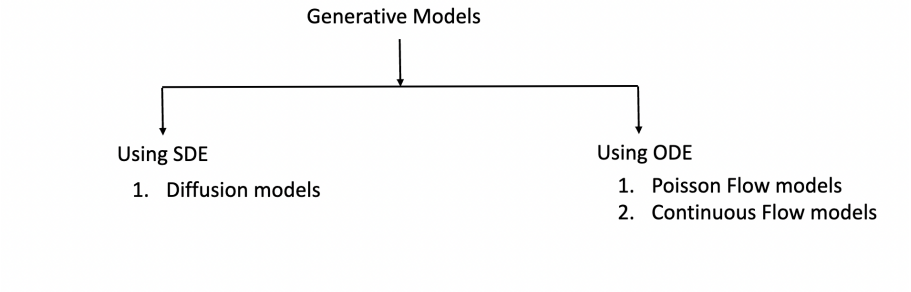

Figure 1: A big picture of the generative models classification

Here $f(x)$ is completely deterministic. We could also add a probabilistic term in the form

$$dx = f(x)dt + g(x)dw \tag{2}$$

where dw is a small displacement that takes random values drawn from a probability distribution $\mathcal{N}(0, dt)$. We will try to build up intuition behind this equation in the subsequent subsections.

2.1 Discrete Random Walk

Consider a particle that starts at position $x_0 = 0$ at time $t = 0$ and thereafter takes step w_i at time $t = i$ drawn from a probability distribution

$$p(w_i) = \begin{cases} 1/2 & w_i = d \\ 1/2 & w_i = -d \end{cases} \tag{3}$$

Then the position of the particle after N steps is

$$x(N) = \sum_{i=1}^{N} w_i. \tag{4}$$

The mean position of the particle after N steps is

$$\mathbb{E}[x(N)] = \sum_{i=1}^{N} \mathbb{E}[w_i] = 0 \tag{5}$$

The variance is

$$\mathrm{Var}[x(N)] = \mathbb{E}[x(N)^2] - \mathbb{E}[x(N)]^2 = d^2 N. \tag{6}$$

Therefore the standard deviation after N steps is $d\sqrt{N}$.

Now let us calculate a few more quantities that will be of interest to us later. For $N_1 > N_2$,

$$\mathbb{E}[x(N_1) - x(N_2)] = \mathbb{E}[x(N_1)] - \mathbb{E}[x(N_2)] = 0 \tag{7}$$

Then,

$$\begin{aligned} \text{Var}[x(N_1) - x(N_2)] &= \mathbb{E}[(x(N_1) - x(N_2))^2] \\ &= \mathbb{E}[x(N_1)^2] + \mathbb{E}[x(N_2)^2] - 2\mathbb{E}[x(N_1).x(N_2)] \\ &= d^2 N_1 + d^2 N_2 - 2\mathbb{E}[x(N_1).x(N_2)] \end{aligned} \tag{8}$$

Now,

$$\begin{aligned} \mathbb{E}[x(N_1).x(N_2)] &= \mathbb{E}\left[\sum_{i=i}^{N_1}\sum_{j=i}^{N_2} w_i.w_j\right] \\ &= \mathbb{E}\left[\sum_{i=i}^{N_2} w_i^2\right] \\ &= d^2 N_2 \end{aligned} \tag{9}$$

Therefore,

$$\text{Var}[x(N_1) - x(N_2)] = d^2(N_1 - N_2) \tag{10}$$

Notice that the variance of $x(N_1) - x(N_2)$ is proportional to the time interval between them.

2.2 Brownian Motion

The continuous version of discrete random walk is called Brownian motion. We will take the small discrete time interval to be Δt and later take the limit $\Delta t \to dt$. But the interval Δt itself can be split into further small time intervals δt. Let

$$N = \frac{\Delta t}{\delta t}. \tag{11}$$

The particle takes N steps in interval Δt and at each step, the probability of displacement is

$$p(w_i) = \begin{cases} 1/2 & w_i = \delta d \\ 1/2 & w_i = -\delta d \end{cases} \tag{12}$$

where δd is a small step taken in a time δt. We want to find out the probability distribution of the particle location after time Δt. In the $N \to \infty$, we can just use central limit theorem.

$$p(x(\Delta t)) = p\left(\lim_{N\to\infty}\sum_{i=1}^{N} w_i\right) = \lim_{N\to\infty}\mathcal{N}(0, N\delta d^2) = \lim_{N\to\infty}\mathcal{N}\left(0, \frac{\Delta t}{\delta t}\delta d^2\right) \tag{13}$$

We will take the limit $N \to \infty$ such that the ration $\delta d^2/\delta t$ remains constant at σ^2. Then

$$p(x(\Delta t)) = \mathcal{N}(0, \Delta t \sigma^2) \tag{14}$$

We would also be interested in another probability distribution

$$p(\Delta x) = p(\ x(t + \Delta t) - x(t)\ ) \tag{15}$$

It should be clear that we can go through the same discretization procedure as before and hence

$$p(\Delta x) = \mathcal{N}(0, \Delta t \sigma^2). \tag{16}$$

Now we take the limit $\Delta t \to dt$ and $\Delta x \to dx$

$$p(dx) = \mathcal{N}(0, \sigma^2 dt) \tag{17}$$

To summarize, what we have seen so far is that in an interval dt, the particle moves by a distance dx, given by a Gaussian distribution of mean 0 and standard deviation $\sigma\sqrt{dt}$.

2.3 Stochastic Differential Equation

Within the same interval dt, there could also be other sources of displacement, for example by a drift term $f(x,t)$. If only such a drift term is present, then

$$dx = f(x,t)dt \tag{18}$$

If in addition we also have a Brownian motion term, then

$$dx = f(x,t)dt + g(x,t)dw \tag{19}$$

where $p(dw) = \mathcal{N}(0, \sigma^2 dt)$. Note that while we used dx to denote Brownian motion displacement in (17), here we are using dx to denote the most general displacement, a part of which is contributed by a Brownian term dw. It follows that

$$dx \sim\ \mathcal{N}\left(f(x,t)dt, g^2(x,t)\sigma^2 dt\right) \tag{20}$$

Since we know the probability distribution for dw, it is easy to see that

$$\begin{aligned}
\mathbb{E}[dw] &= 0 \\
\mathbb{E}[dw^2] &= \sigma^2 dt \\
\mathbb{E}[dw^3] &= 0 \\
\mathbb{E}[dw^4] &= 3\sigma^4 dt^2 \\
\mathbb{E}[dw^5] &= 0 \\
\mathbb{E}[dw^6] &= 15\sigma^6 dt^3 \\
&\vdots
\end{aligned} \tag{21}$$

2.4 Ito's Lemma

Consider a function $F(x(t), t)$ where the variable $x(t)$ evolves according to the differential equation (19). Then for a small displacement dx and dt, the function $F(x(t), t)$ changes by

$$\begin{aligned} dF(x(t),t) =& \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial t}dt + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}(dx)^2 + \frac{1}{2}\frac{\partial^2 F}{\partial t^2}(dt)^2 + \frac{\partial^2 F}{\partial x \partial t}dxdt \\ & + \mathrm{O}((dx)^n(dt)^m, \quad \text{where} \quad n+m>2) \end{aligned} \tag{22}$$

Now,

$$\begin{aligned} dx &= f(x,t)dt + g(x,t)dw \\ (dx)^2 &= f(x,t)^2dt^2 + g(x,t)^2dw^2 + 2f(x,t)g(x,t)dxdt \\ dxdt &= f(x,t)dt^2 + g(x,t)dwdt \\ (dx)^n(dt)^m &= (f(x,t)dt + g(x,t)dw)^n(dt)^m \end{aligned} \tag{23}$$

Therefore, we get

$$\begin{aligned} dF =& \frac{\partial F}{\partial x}(fdt + gdw) + \frac{\partial F}{\partial t}dt + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}(f^2dt^2 + g^2dw^2 + 2fgdwdt) \\ & + \frac{1}{2}\frac{\partial^2 F}{\partial t^2}(dt)^2 + \frac{\partial^2 F}{\partial x \partial t}(fdt + gdw)dt + ... \end{aligned} \tag{24}$$

In ordinary calculus, we only keep the first order terms for a small change dt and dw. But notice that $\mathbb{E}[dw^2] = \sigma^2 dt$. That is dw^2 is of the order of dt. Similarly $\mathbb{E}[dw^{2n}] \propto dt^n \ \forall n$ and $\mathbb{E}[dw^n] = 0$ for odd n. Therefore, we will make the following substitutions

$$\begin{aligned} dw^2 &\to \sigma^2 dt \\ dw^3 &\to 0 \\ dw^4 &\to 3\sigma^4 dt^2 \\ &\vdots \end{aligned} \tag{25}$$

This is an heuristic argument but happens to be right. The justification for these substitutions is that when we integrate dF, all the Brownian motion random variables dw and their powers can be replaced by their expectation values. So what we get is

$$\begin{aligned} dF &= \frac{\partial F}{\partial x}(fdt + gdw) + \frac{\partial F}{\partial t}dt + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}(g^2\sigma^2 dt) \\ &= \left(\frac{\partial F}{\partial x}f + \frac{\partial F}{\partial t} + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}g^2\sigma^2\right)dt + \frac{\partial F}{\partial x}gdw \end{aligned} \tag{26}$$

This result is called Ito's lemma.

2.5 Fokker-Planck Equation

When we have an ensemble of particles diffusing under the SDE

$$dx = f(x,t)dt + g(x,t)dw \tag{27}$$

one can ask what is the probability density function $p(x,t)$ that a particle is found between positions x and $x+dx$ and a time interval between t and $t+dt$.

Consider a functional $F(x(t))$. The total derivative is

$$dF = \left(\frac{\partial F}{\partial x}f + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}g^2\sigma^2\right)dt + \frac{\partial F}{\partial x}gdw. \tag{28}$$

Taking expectation value at a particular time t

$$\begin{aligned}\mathbb{E}\left[dF\right] &= \mathbb{E}\left[\frac{\partial F}{\partial x}f + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}g^2\sigma^2\right]dt + \mathbb{E}\left[\frac{\partial F}{\partial x}gdw\right] \\ \frac{d}{dt}\mathbb{E}[F] &= \mathbb{E}\left[\frac{\partial F}{\partial x}f + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}g^2\sigma^2\right]\end{aligned} \tag{29}$$

But the expectation value at time t can also written in terms of $p(x,t)$ as

$$\mathbb{E}[h(x)] = \int dx\ p(x,t)h(x). \tag{30}$$

and

$$\frac{d}{dt}\mathbb{E}\left[h\right] = \int dx\frac{dp}{dt}h(x) \tag{31}$$

Therefore,

$$\begin{aligned}\int dx\ \frac{\partial p}{\partial t}(x,t)F &= \int dx\ p(x,t)\frac{\partial F}{\partial x}f + \frac{\sigma^2}{2}\int dx\ p(x,t)\frac{\partial^2 F}{\partial x^2}g^2 \\ \int dx\ \frac{\partial p}{\partial t}(x,t)F &= -\int dx\frac{\partial}{\partial x}(p(x,t)f)F + \frac{\sigma^2}{2}\int dx\frac{\partial^2}{\partial x^2}(p(x,t)g^2)F\end{aligned} \tag{32}$$

In the second line, we have used integration by parts. This gives us the Fokker-Planck equation

$$\frac{\partial p}{\partial t}(x,t) = -\frac{\partial}{\partial x}(p(x,t)f) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}(p(x,t)g^2) \tag{33}$$

After the particles move around for some time, a steady state can be reached and at that point $\frac{\partial p}{\partial t} = 0$. Therefore,

$$\frac{\partial}{\partial x}(p_s(x)f) = \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}(p_s(x)g^2) \tag{34}$$

The steady state solution for

$$f(x,t) = -2\frac{dU(x)}{dx}, \qquad g(x,t) = 1 \tag{35}$$

is

$$p_s(x) = e^{-U(x)/\sigma^2} \tag{36}$$

2.6 Random Walk as Markov Process

Discrete random walk and its continuous counterpart Brownian motion is a Markov process. In a Markov process, the next state depend only on the current state of the system. In random walk, the position of the particle in the next jump is always either $x_i + d$ or $x_i - d$. We can take a very general approach to study Markov processes and see how the Fokker Planck equation emerges. In the language of Markov processes, the Fokker-Planck equation is called the Forward Kolmogorov equation.

We can make certain statements about the probability of finding the system in a particular state from transition probabilities. Let the states of a discrete Markov process be denoted as $X_0, X_1, ..., X_{i-1}, X_i, X_{i+1}, ...$ Then the probability

$$P(X_0 = x_0, X_1 = x1, ...) = P(X_0 = x_0) \prod_{t=1,2,3,...} P(X_t = x_t | X_{t-1} = x_{t-1}) \tag{37}$$

The transition probabilities $P(X_t = x_t | X_{t-1} = x_{t-1})$ and the initial probability $P(X_0 = x_0)$ completely determine the system. One can also write the probability of transition from one state to another in terms of an intermediate state. For $t > s$

$$P(X_t = x_t | X_s = x_s) = \sum_k P(X_t = x_t | X_k = x_k) P(X_k = x_k | X_s = x_s) \tag{38}$$

where $t > k > s$. Similarly

$$P(X_t = x_t) = \sum_k P(X_t = x_t | X_k = x_k) P(X_k = x_k) \tag{39}$$

where $t > k$.

All this can be generalized to continuous Markov processes as well. The transition probability in terms of the intermediate state is

$$P(x; t|y; s) = \int dk \; P(x; t|k; m) P(k; m|y; s) \tag{40}$$

where $t > m > s$. Similarly we have

$$p(x; t) = \int dk \; p(x; t|k; m) p(k; m) \tag{41}$$

for $t > m$. We will use these relations to derive the forward and backward Kolmogorov equations.

2.7 Forward Kolmogorov Equation

Consider the transition probability relation

$$P(x; t|y; s) = \int_{-\infty}^{\infty} dk \; P(x; t|k; m) P(k; m|y; s) \tag{42}$$

where $t > m > s$. This relation means that

$$P(x;t+dt|y;s) = \int_{-\infty}^{\infty} dk\ P(x;t+dt|k;t)P(k;t|y;s) \tag{43}$$

This equation can be used to derive an equation that $p(x,t)$ satisfies in general. Let the system be in a state z at time t and state $z+\Delta$ at time $t+dt$. We will write the transition probability $p(z+\Delta;t+dt|z,t)$ as

$$\phi_t(\Delta,z,dt) = p(z+\Delta;t+dt|z,t). \tag{44}$$

So for (58), substitute

$$\begin{aligned} x &= k+\Delta \\ \implies dk &= -d\Delta \\ k = \pm\infty &\implies \Delta = \mp\infty \end{aligned} \tag{45}$$

Therefore, (58) becomes

$$\begin{aligned} P(x;t+dt|y;s) &= \int_{-\infty}^{\infty} dk\ P(x;t+dt|k;t)P(k;t|y;s) \\ &= \int_{-\infty}^{\infty} d\Delta\ P(x;t+dt|x-\Delta;t)P(x-\Delta;t|y;s) \\ &= \int_{-\infty}^{\infty} d\Delta\ \phi_t(\Delta,x-\Delta,dt)P(x-\Delta;t|y;s) \end{aligned} \tag{46}$$

Now Taylor expand the integrand about x.

$$\begin{aligned} P(x;t+dt|y;s) &= \int_{-\infty}^{\infty} d\Delta\ \phi_t(\Delta,x-\Delta,dt)P(x-\Delta;t|y;s) \\ &= \int_{-\infty}^{\infty} d\Delta\ \phi_t(\Delta,x,dt)P(x;t|y;s) \\ &\quad - \int_{-\infty}^{\infty} d\Delta\ \Delta\frac{\partial}{\partial x}\left(\phi_t(\Delta,x,dt)P(x;t|y;s)\right) \\ &\quad + \int_{-\infty}^{\infty} d\Delta\ \frac{\Delta^2}{2}\frac{\partial^2}{\partial x^2}\left(\phi_t(\Delta,x,dt)P(x;t|y;s)\right) \\ &\quad + ... \end{aligned} \tag{47}$$

Using

$$\int_{-\infty}^{\infty} d\Delta\ \phi_t(\Delta,x,dt) = 1 \tag{48}$$

we get

$$\begin{aligned} P(x;t+dt|y;s) - P(x;t|y;s) = &- \frac{\partial}{\partial x}\left(\mathbb{E}_{\Delta\sim\phi_t(\Delta,x,dt)}[\Delta]P(x;t|y;s)\right) \\ &+ \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\mathbb{E}_{\Delta\sim\phi_t(\Delta,x,dt)}[\Delta^2]P(x;t|y;s)\right) \\ &+ ... \end{aligned} \tag{49}$$

We take the first and second moments of $\phi_t(\Delta, x, dt)$ to be of the order of dt and the higher moments to be of the order of $O(dt^n)$ where $n > 1$. This can be motivated by looking at the moments of discrete random walk (with drift term). In the infinitesimal limit of dt, the higher order moments can be eliminated

$$\begin{aligned} \mathbb{E}_{\Delta \sim \phi_t(\Delta,x,dt)}[\Delta] &:= f(x,t)dt \\ \mathbb{E}_{\Delta \sim \phi_t(\Delta,x,dt)}[\Delta^2] &:= g^2(x,t)dt \end{aligned} \tag{50}$$

Therefore we get

$$\frac{\partial}{\partial t}P(x;t|y;s) = -\frac{\partial}{\partial x}\left(f(x,t)P(x;t|y;s)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(g^2(x,t)P(x;t|y;s)\right) \tag{51}$$

This is the forward Kolmogorov equation also called as Fokker-Planck equation. The equations (50) can be used to define the dynamics of a single particle. We can see that the small change in a particle position follows

$$\begin{aligned} dx &\sim \phi_t(\Delta, x, dt) \\ \mathbb{E}[dx] &= f(x,t)dt \\ \mathrm{Var}(dx) &= g^2(x,t)dt - O(dt^2) \approx g^2(x,t)dt \end{aligned} \tag{52}$$

Hence we can write the dynamics as

$$dx = f(x,t)dt + g(x,t)dw \tag{53}$$

with the identification

$$\phi_t(\Delta, x, dt) \sim \mathcal{N}(f(x,t)dt, g^2(x,t)dt). \tag{54}$$

This identification can always be done as using central limit theorem, any sum of random variables drawn from any distribution tends to a Gaussian.

Notice that we could also have used the relation

$$p(x;t) = \int dk\; p(x;t|k;m)p(k;m) \tag{55}$$

to derive an equation for $p(x,t)$. In this case, the corresponding Fokker-Planck equation will be

$$\frac{\partial}{\partial t}P(x;t) = -\frac{\partial}{\partial x}\left(f(x,t)P(x;t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(g^2(x,t)P(x;t)\right). \tag{56}$$

2.8 Backward Kolmogorov Equation

Consider the transition probability relation

$$P(x;t|y;s) = \int_{-\infty}^{\infty} dk\; P(x;t|k;m)P(k;m|y;s) \tag{57}$$

where $t > m > s$. This relation means that

$$P(x;t|y;s-ds) = \int_{-\infty}^{\infty} dk\ P(x;t|k;s)P(k;s|y;s-ds) \tag{58}$$

In a small time interval ds, let the system evolve from a state y to $y+\Delta$. Then

$$\begin{aligned} k &= y+\Delta \\ \implies dk &= d\Delta \\ k = \pm\infty &\implies \Delta = \pm\infty \end{aligned} \tag{59}$$

Therefore

$$\begin{aligned} P(x;t|y;s-ds) &= \int_{-\infty}^{\infty} dk\ P(x;t|k;s)P(k;s|y;s-ds) \\ &= \int_{-\infty}^{\infty} d\Delta\ P(x;t|y+\Delta;s)P(y+\Delta;s|y;s-ds) \\ &= \int_{-\infty}^{\infty} d\Delta\ \phi_{s-ds}(\Delta,y,ds)P(x;t|y+\Delta;s) \\ &= \int_{-\infty}^{\infty} d\Delta\ \phi_{s-ds}(\Delta,y,ds)P(x;t|y;s) \\ &\qquad + \int_{-\infty}^{\infty} d\Delta\ \phi_{s-ds}(\Delta,y,ds)\Delta\frac{\partial}{\partial y}P(x;t|y;s) \\ &\qquad + \int_{-\infty}^{\infty} d\Delta\ \phi_{s-ds}(\Delta,y,ds)\frac{\Delta^2}{2}\frac{\partial^2}{\partial y^2}P(x;t|y;s) \\ &\qquad + ... \end{aligned} \tag{60}$$

We get

$$P(x;t|y;s-ds)-P(x;t|y;s) = f(y,s-ds)ds\frac{\partial}{\partial y}P(x;t|y;s)+\frac{1}{2}g^2(y,s-ds)ds\frac{\partial^2}{\partial y^2}P(x;t|y;s)$$

In the limit $ds \to 0$, we get

$$-\frac{\partial}{\partial s}P(x;t|y;s) = f(y,s)\frac{\partial}{\partial y}P(x;t|y;s) + \frac{1}{2}g^2(y,s)\frac{\partial^2}{\partial y^2}P(x;t|y;s). \tag{61}$$

This is called the backward Kolmogorov equation.

2.9 Reversing the Diffusion process

We would like to reverse the diffusion process i.e. we want to find $P(y;s|x;t)$ where $s < t$. If we could find a partial differential equation that is satisfied by $P(y;s<t|x;t)$, then maybe we could also find the particle dynamics that leads to reverse diffusion. The idea is to start with the joint probability density $P(y;s,x;t)$

$$P(y;s,x;t) = P(y;s)P(x;t|y;s) \tag{62}$$

We want a differential equation where time s is a variable (and time t is fixed). Therefore, differentiate it with s

$$\frac{\partial}{\partial s}P(y;s,x;t) = P(y;s)\frac{\partial}{\partial s}P(x;t|y;s) + P(x;t|y;s)\frac{\partial}{\partial s}P(y;s) \tag{63}$$

Next we apply equations (56) and (61) to the above expression

$$\begin{aligned}
\frac{\partial}{\partial s}P(y;s,x;t) =& -P(y;s)\left(f(y,s)\frac{\partial}{\partial y}P(x;t|y;s) + \frac{1}{2}g^2(y,s)\frac{\partial^2}{\partial y^2}P(x;t|y;s)\right) \\
&+ P(x;t|y;s)\left(-\frac{\partial}{\partial y}\left(f(y,s)P(y;s)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s)\right)\right) \\
=& -P(y;s)\left(f(y,s)\frac{\partial}{\partial y}\frac{P(y;s,x;t)}{P(y;s)} + \frac{1}{2}g^2(y,s)\frac{\partial^2}{\partial y^2}\frac{P(y;s,x;t)}{P(y;s)}\right) \\
&+ \frac{P(y;s,x;t)}{P(y;s)}\left(-\frac{\partial}{\partial y}\left(f(y,s)P(y;s)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s)\right)\right)
\end{aligned}$$

Divide by $P(x;t)$. We get

$$\begin{aligned}
\frac{\partial}{\partial s}P(y;s|x;t) =& -P(y;s)\left(f(y,s)\frac{\partial}{\partial y}\frac{P(y;s|x;t)}{P(y;s)} + \frac{1}{2}g^2(y,s)\frac{\partial^2}{\partial y^2}\frac{P(y;s|x;t)}{P(y;s)}\right) \\
&+ \frac{P(y;s|x;t)}{P(y;s)}\left(-\frac{\partial}{\partial y}\left(f(y,s)P(y;s)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s)\right)\right)
\end{aligned}$$

Now the equation is in terms of $P(y;s|x;t)$ and $P(y;s)$. Now if we could massage this equation and recast it as a Fokker-Planck equation, then we can find the reverse time dynamics.

First notice that the 1st and the 3rd term can be combined as a single derivative term.

$$\begin{aligned}
\frac{\partial}{\partial s}P(y;s|x;t) =& -\frac{\partial}{\partial y}\left(f(y,s)P(y;s|x;t)\right) \\
&- \frac{1}{2}g^2(y,s)P(y;s)\frac{\partial^2}{\partial y^2}\frac{P(y;s|x;t)}{P(y;s)} \\
&+ \frac{1}{2}\frac{P(y;s|x;t)}{P(y;s)}\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s)\right)
\end{aligned} \tag{64}$$

Next we use

$$\begin{aligned}
\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s|x;t)\right) =& g^2(y,s)P(y;s)\frac{\partial^2}{\partial y^2}\frac{P(y;s|x;t)}{P(y;s)} \\
&+ \frac{P(y;s|x;t)}{P(y;s)}\frac{\partial^2}{\partial y^2}g^2(y,s)P(y;s) \\
&+ 2\frac{\partial}{\partial y}\left(g^2(y,s)P(y;s)\right)\frac{\partial}{\partial y}\left(\frac{P(y;s|x;t)}{P(y;s)}\right)
\end{aligned} \tag{65}$$

Therefore, equation (64) becomes

$$\begin{aligned}\frac{\partial}{\partial s}P(y;s|x;t) =& -\frac{\partial}{\partial y}\left(f(y,s)P(y;s|x;t)\right)\\ &+\frac{P(y;s|x;t)}{P(y;s)}\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s)\right)\\ &-\frac{1}{2}\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s|x;t)\right)\\ &+\frac{\partial}{\partial y}\left(g^2(y,s)P(y;s)\right)\frac{\partial}{\partial y}\left(\frac{P(y;s|x;t)}{P(y;s)}\right)\end{aligned} \tag{66}$$

Finally, we get

$$\begin{aligned}\frac{\partial}{\partial s}P(y;s|x;t) =& -\frac{\partial}{\partial y}\left(f(y,s)P(y;s|x;t) - \frac{P(y;s|x;t)}{P(y;s)}\frac{\partial}{\partial y}\left(g^2(y,s)P(y;s)\right)\right)\\ &-\frac{1}{2}\frac{\partial^2}{\partial y^2}\left(g^2(y,s)P(y;s|x;t)\right)\end{aligned} \tag{67}$$

Now this equation is in the form of Fokker-Planck equation or forward Kolmogorov equation. Now we can write the dynamical equation for the reverse diffusion process.

$$dy = \left(f(y,s) - \frac{1}{P(y;s)}\frac{\partial}{\partial y}\left(g^2(y,s)P(y;s)\right)\right)ds + g(y,s)dw \tag{68}$$

Notice that here we take $dw \to -ds$ for the reverse diffusion. Now let us try to make sense of the above dynamical equation. It should be clear that as such it is difficult to reverse a diffusion process since over a period of time, all the initial information is lost. But what if we provide the initial probability information $P(y,s)$. Then of course, it should be possible to reverse the process. This is something that we should expect.

3 Diffusion Models

For diffusion models the kind of SDEs we are interested in are

$$d\mathbf{x} = f(\mathbf{x},t)dt + g(t)d\mathbf{w} \tag{69}$$

We have generalized to a multidimensional case. Here $d\mathbf{w}$ is a noise from spherical Gaussian. Then the reverse time SDE is

$$\begin{aligned}d\mathbf{x} &= \left(f(\mathbf{x},t) - \frac{g^2(t)}{P(\mathbf{x},t)}\nabla_{\mathbf{x}}P(\mathbf{x},t)\right) + g(t)d\mathbf{w}\\ &= \left(f(\mathbf{x},t) - g^2(t)\nabla_{\mathbf{x}}\log P(\mathbf{x},t)\right) + g(t)d\mathbf{w}\end{aligned} \tag{70}$$

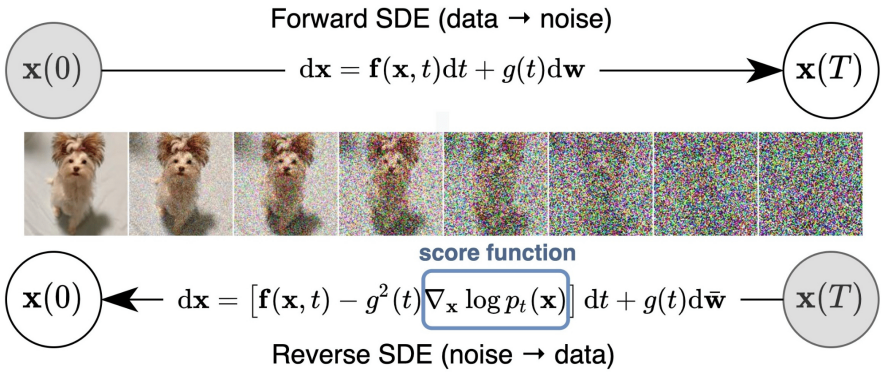

Figure 2: Diffusion models use forward SDE to noise the initial image and reverse or backward SDE to denoise the image.

Usually for the diffusion process, we restrict ourselves to SDEs of the form

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}dt + \sqrt{\beta(t)}d\mathbf{w} \tag{71}$$

Therefore the reverse process is

$$d\mathbf{x} = \left(-\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\nabla_{\mathbf{x}}\log P(\mathbf{x}, t)\right)dt + \sqrt{\beta(t)}d\mathbf{w} \tag{72}$$

In actual implementation, we discretize the SDEs. The forward process is then

$$\mathbf{x}_{t+dt} - \mathbf{x}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t dt + \sqrt{\beta(t)dt}\,\epsilon \tag{73}$$

Therefore

$$\mathbf{x}_{t+dt} = \left(1 - \frac{1}{2}\beta(t)dt\right)\mathbf{x}_t + \sqrt{\beta(t)dt}\,\epsilon \tag{74}$$

Note that the above expression is valid only for infinitesimal interval dt. To generalize to a bigger interval of time Δt (but still small enough in our computers), we can think of the above expression as

$$\mathbf{x}_{t+\Delta t} = \sqrt{1 - \beta(t)\Delta t}\,\mathbf{x}_t + \sqrt{\beta(t)\Delta t}\,\epsilon \tag{75}$$

which generalizes to (74) in the small Δt limit.

For later purpose, we will need the conditional probability $P(\mathbf{x}_t, t|\mathbf{x}_0, 0)$. Here

we will do a rough calculation of such a density function.

$$\begin{aligned}
\mathbf{x}(\Delta t) &= \left(1 - \frac{1}{2}\beta(0)\Delta t\right)\mathbf{x}(0) + \sqrt{\beta(0)\Delta t}\,\epsilon \\
\mathbf{x}(2\Delta t) &= \left(1 - \frac{1}{2}\beta(\Delta t)\Delta t\right)\mathbf{x}(\Delta t) + \sqrt{\beta(\Delta t)\Delta t}\,\epsilon \\
&= \left(1 - \frac{1}{2}\beta(\Delta t)\Delta t\right)\left(1 - \frac{1}{2}\beta(0)\Delta t\right)\mathbf{x}(0) \\
&\quad + \left(1 - \frac{1}{2}\beta(\Delta t)\Delta t\right)\sqrt{\beta(0)\Delta t}\,\epsilon + \sqrt{\beta(\Delta t)\Delta t}\,\epsilon
\end{aligned} \tag{76}$$

Continuing, we get

$$\begin{aligned}
\mathbf{x}(n\Delta t) =& \mathbf{x}(0)\prod_{i=0}^{n-1}\left(1 - \frac{1}{2}\beta(i\Delta t)\Delta t\right) \\
&+ \sum_{i=0}^{n-1}\sqrt{\beta(i\Delta t)\Delta t}\epsilon\prod_{j=i+1}^{n-1}\left(1 - \frac{1}{2}\beta(j\Delta t)\Delta t\right)
\end{aligned} \tag{77}$$

What we have in mind is that the above expression can be somehow written as some sort of an integral in the limit $n \to \infty$ (such that $n\Delta \to t$). But we have product of terms in the above expression. In order to convert products into sums, lets consider logarithms of products (and later take exponentials). Consider

$$\begin{aligned}
\log\prod_{i=0}^{n-1}\left(1 - \frac{1}{2}\beta(i\Delta t)\Delta t\right) &= \sum_{i=0}^{n-1}\log\left(1 - \frac{1}{2}\beta(i\Delta t)\Delta t\right) \\
&= -\sum_{i=0}^{n-1}\frac{1}{2}\beta(i\Delta t)\Delta t \\
&= -\frac{1}{2}\int_0^t \beta(u)du
\end{aligned} \tag{78}$$

Therefore

$$\lim_{n\to\infty}\prod_{i=0}^{n-1}\left(1 - \frac{1}{2}\beta(i\Delta t)\Delta t\right) = \exp\left(-\frac{1}{2}\int_0^t \beta(u)du\right). \tag{79}$$

The term

$$\sum_{i=0}^{n-1}\sqrt{\beta(i\Delta t)\Delta t}\epsilon\prod_{j=i+1}^{n-1}\left(1 - \frac{1}{2}\beta(j\Delta t)\Delta t\right) \tag{80}$$

is a sum of Normal distributions. This can be written as a single normal distribution, whose variance at time $i\Delta t$ is

$$\sigma^2(i\Delta t) = \beta(i\Delta t)\Delta t\prod_{j=i+1}^{n-1}\left(1 - \frac{1}{2}\beta(j\Delta t)\Delta t\right)^2 \tag{81}$$

In the limit $n \to \infty$, we get

$$\sigma^2(u) = \beta(u)du \exp\left(-\int_u^t \beta(v)dv\right). \tag{82}$$

Therefore

$$\mathbf{x}(t) = \mathbf{x}(0) \exp\left(-\frac{1}{2}\int_0^t \beta(u)du\right) + \eta_t \tag{83}$$

where

$$\eta_t \sim \mathcal{N}\left(0, \int_0^t \beta(u)du \exp\left(-\int_u^t \beta(v)dv\right)\mathbf{I}\right). \tag{84}$$

All this can be written compactly as

$$P(\mathbf{x}_t, t|\mathbf{x}_0, 0) = \mathcal{N}\left(\mathbf{x}(0) \exp\left(-\frac{1}{2}\int_0^t \beta(u)du\right), \int_0^t \beta(u)du \exp\left(-\int_u^t \beta(v)dv\right)\mathbf{I}\right).$$

This can be further simplified as

$$P(\mathbf{x}_t, t|\mathbf{x}_0, 0) = \mathcal{N}\left(\mathbf{x}(0) \exp\left(-\frac{1}{2}\int_0^t \beta(u)du\right), \mathbf{I} - \exp\left(-\int_0^t \beta(u)du\right)\mathbf{I}\right). \tag{85}$$

The function $\beta(t)$ is chosen such that its integral $\int_0^t \beta(u)du \to \infty$ as $t \to \infty$. So after a long time we get $P(\mathbf{x}_t, t|\mathbf{x}_0, 0) \sim \mathcal{N}(0, \mathbf{I})$.

4 Score Matching

To use the reverse SDE, all we need is a way to learn $\nabla_{\mathbf{x}} \log P(\mathbf{x}_t, t)$. For this we use a neural network (called score function) $\mathbf{s}_\theta(\mathbf{x}_t, t)$ to learn the gradient of the $P(\mathbf{x}_t, t)$. The loss function would be

$$\mathcal{L}(\theta) = \mathbb{E}_{p(\mathbf{x}_t,t)}[||\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}} \log P(\mathbf{x}_t, t)||^2] \tag{86}$$

It can be shown that when two probability distributions agree on the gradient, then they must be equal. The main advantage of learning the score function is that we don't need to deal with the problem of intractable normalization constant. But still we have another problem. How do we determine $\nabla_{\mathbf{x}} \log P(\mathbf{x}_t, t)$? The trick is to trade $P(\mathbf{x}_t, t)$ for $P(\mathbf{x}_t, t|\mathbf{x}_0, 0)$. Notice that we already know the closed form expression for $P(\mathbf{x}_t, t|\mathbf{x}_0, 0)$ given in equation (85).

Notice that the data that we have $\mathcal{D}$ are from the distribution $P(\mathbf{x}, 0)$. Now we will massage the loss function in a form that is convenient for implementation. Expanding the loss function

$$\begin{aligned}\mathcal{L}(\theta) =& \mathbb{E}_{p(\mathbf{x}_t,t)}[||\mathbf{s}_\theta(\mathbf{x}_t, t)||^2] + \mathbb{E}_{p(\mathbf{x}_t,t)}[||\nabla_{\mathbf{x}} \log P(\mathbf{x}_t, t)||^2] \\ & - 2\mathbb{E}_{p(\mathbf{x}_t,t)}[\mathbf{s}_\theta(\mathbf{x}_t, t)^T \nabla_{\mathbf{x}} \log P(\mathbf{x}_t, t)]\end{aligned} \tag{87}$$

The 2nd term can be ignored as it doesn't depend on θ. Now the 3rd can be rewritten as

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x}_t,t)}[\mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}\log P(\mathbf{x}_t,t)] &= \int P(\mathbf{x}_t,t)\mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}\log P(\mathbf{x}_t,t)dV \\
&= \int \mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}P(\mathbf{x}_t,t)dV \\
&= \int \mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}\left(\int d\mathbf{x}_0 P(\mathbf{x}_t,t|\mathbf{x}_0,0)P(\mathbf{x}_0,0)\right)dV \\
&= \int\int d\mathbf{x}_0 P(\mathbf{x}_0,0)\mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}P(\mathbf{x}_t,t|\mathbf{x}_0,0)dV \\
&= \mathbb{E}_{p(\mathbf{x}_0,0)}\left[\int \mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}P(\mathbf{x}_t,t|\mathbf{x}_0,0)dV\right] \\
&= \mathbb{E}_{p(\mathbf{x}_0,0)}\mathbb{E}_{p(\mathbf{x}_t,t|\mathbf{x}_0,0)}[\mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}\log P(\mathbf{x}_t,t|\mathbf{x}_0,0)]
\end{aligned}$$

Notice that we could do the 4th step because the gradient acts on $\mathbf{x}_t$ and not on $\mathbf{x}_0$. Plugging this into equation (87) and adding a θ independent term (2nd term below) we get

$$\begin{aligned}
\mathcal{L}(\theta) =& \mathbb{E}_{p(\mathbf{x}_t,t)}[||\mathbf{s}_\theta(\mathbf{x}_t,t)||^2] + \mathbb{E}_{p(\mathbf{x}_0,0)}\mathbb{E}_{p(\mathbf{x}_t,t|\mathbf{x}_0,0)}[||\nabla_{\mathbf{x}}\log P(\mathbf{x}_t,t|\mathbf{x}_0,0)||^2] \\
&- 2\mathbb{E}_{p(\mathbf{x}_0,0)}\mathbb{E}_{p(\mathbf{x}_t,t|\mathbf{x}_0,0)}[\mathbf{s}_\theta(\mathbf{x}_t,t)^T\nabla_{\mathbf{x}}\log P(\mathbf{x}_t,t|\mathbf{x}_0,0)] \\
=& \mathbb{E}_{p(\mathbf{x}_0,0)}\mathbb{E}_{p(\mathbf{x}_t,t|\mathbf{x}_0,0)}[||\mathbf{s}_\theta(\mathbf{x}_t,t) - \nabla_{\mathbf{x}}\log P(\mathbf{x}_t,t|\mathbf{x}_0,0)||^2]
\end{aligned} \tag{88}$$

The above loss function corresponds to a particular time t. To get a loss average over time, we use

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim U(0,T)}\mathbb{E}_{p(\mathbf{x}_0,0)}\mathbb{E}_{p(\mathbf{x}_t,t|\mathbf{x}_0,0)}[||\mathbf{s}_\theta(\mathbf{x}_t,t) - \nabla_{\mathbf{x}}\log P(\mathbf{x}_t,t|\mathbf{x}_0,0)||^2] \tag{89}$$

This provides a convenient loss function that can be used for training. First we sample a time t uniformly from 0 to T. We then get a sample from our dataset, and then get a sample at time t conditioned on our data point using equation (85) to evaluate the loss function. Notice that for this whole thing to work we have relied on the closed form expression for $P(\mathbf{x}_t,t|\mathbf{x}_0,0)$ that is given in equation (85).

Once we know the score function we can use the reverse time diffusion equation to generate samples.

$$d\mathbf{x} = \left(-\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\mathbf{s}_\theta(\mathbf{x}_t,t)\right)dt + \sqrt{\beta(t)}d\mathbf{w} \tag{90}$$

Write down the discretized version of the above equation.

5 Poisson Flow Generative Models

Poisson flow model is based on learning an ODE (as opposed to SDE). This method is based on considering data points as electric charges. The presence

of electric charges generate electric fields. When we know the electric fields we can "flow" along them either in the "forward" direction (adding noise) or in the "backward" direction (generating samples).

Electric potential due to charge distribution
Given a charge distribution $\rho(x)$, the electric potential due to it is given is determined from Poisson equation

$$\nabla^2\phi(x) = -\rho(x) \tag{91}$$

Electric field from electric potential
Given an electric potential, one can determine the electric field as

$$\mathbf{E}(x) = -\nabla\phi(x) \tag{92}$$

Forward ODE from electric field
Once we know the elctric field we can flow in the forward direction (noising) using

$$\frac{d\mathbf{x}}{dt} = \mathbf{E}(x). \tag{93}$$

This equation can be shown to a noising process. All particles moving along the electric field will move outward and at long distances, the electric field loses all the "memory" of the structure of the data.

One can also write down the corresponding Fokker-Plack equation:

$$\frac{dp(x,t)}{dt} = -\nabla.(p(x,t)\mathbf{E}(x)) \tag{94}$$

Backward ODE from electric field
The backward flow or denoising is done using

$$\frac{d\mathbf{x}}{dt} = -\mathbf{E}(x) \tag{95}$$

Now all that is required to address the question of "how do we learn the electric field generated by the data?"

Finding electric field from data distribution Given the data distribution, the electric field can be found using

$$\mathbf{E}(x) = \frac{1}{S_N(1)} \int \frac{\mathbf{x}-\mathbf{y}}{||\mathbf{x}-\mathbf{y}||^N} p(\mathbf{y}) d\mathbf{y} \tag{96}$$

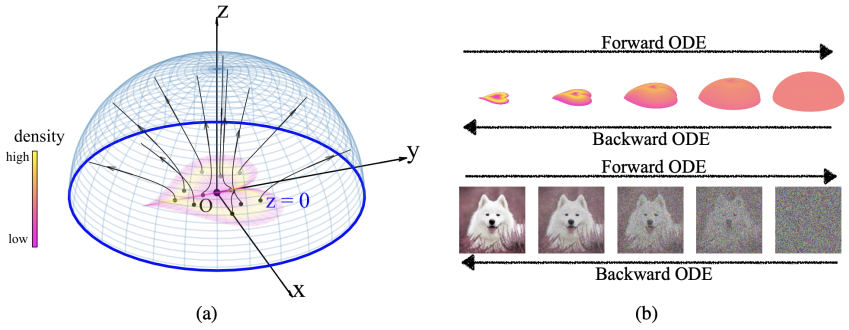

Figure 3: **a**) The data distribution is taken as charge distribution. The charge distribution gives rise to electric field. **b**) Forward and Reverse ODE processes.

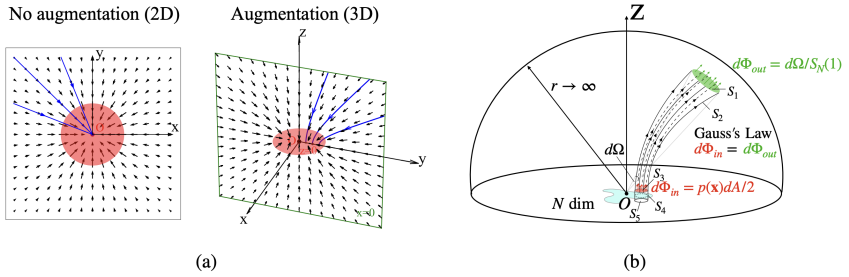

Figure 4: **a**) Mode collapse problem is avoided by working with one extra dimension. **b**) We start with particles on the hemisphere and use backward ODE to flow to $z = 0$ plane to generate samples.

This is just Coulomb's law generalized to higher dimensional space in which the data lives. For example when $N = 3$, we get the usual Coulomb formula.

Mode collapse problem In the backward process, when we just follow the electric field, the samples will all converge to a point. The way to deal with this problem is to add an extra dimension - z. Let $\tilde{\mathbf{x}} = (\mathbf{x}, z)$. With this the new electric field is

$$\mathbf{E}(\tilde{\mathbf{x}}) = \frac{1}{S_N(1)} \int \frac{\tilde{\mathbf{x}} - \tilde{\mathbf{y}}}{||\tilde{\mathbf{x}} - \tilde{\mathbf{y}}||^{N+1}} p(\tilde{\mathbf{y}}) d\tilde{\mathbf{y}} \tag{97}$$

Similarly, the forward and backward equations become $d\tilde{\mathbf{x}}/dt = \mathbf{E}(\tilde{\mathbf{x}})$ and $d\tilde{\mathbf{x}}/dt = -\mathbf{E}(\tilde{\mathbf{x}})$.

Theorem: When particles are uniformly sampled on the the upper hemisphere ($z > 0$) of radius r, and evolved backward using $d\tilde{\mathbf{x}}/dt = -\mathbf{E}(\tilde{\mathbf{x}})$, then in the

limit $r \to \infty$ the process generates samples from $p(\tilde{\mathbf{x}})$ The proof for the theorem follows from the Gauss law.

Learning the electric field from data
Given data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ i.i.d sampled from data distribution $p(\mathbf{x})$, the electric field is

$$\mathbf{E}_{\mathcal{D}}(\tilde{\mathbf{x}}) = \sum_{i=1}^{n} \frac{\tilde{\mathbf{x}} - \tilde{\mathbf{x}_i}}{||\tilde{\mathbf{x}} - \tilde{\mathbf{x}_i}||^{N+1}}. \tag{98}$$

Notice that we used summation instead of integration as we have a discrete set of data. Define normalized electric field as

$$\mathbf{v} = -\sqrt{N}\frac{\mathbf{E}}{||\mathbf{E}||_2} \tag{99}$$

We will use the normalized electric field instead of electric field itself. In practice we use mini-batch data $\mathcal{B} = \{\mathbf{x}_i\}_i^{|\mathcal{B}|}$. Now we train a neural network $\mathbf{f}_\theta$ using the loss function

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} ||\mathbf{f}_\theta(\tilde{\mathbf{y}}_i) - \mathbf{v}_{\mathcal{B}}(\tilde{\mathbf{y}}_i)||^2 \tag{100}$$

Once we have learnt the normalized electric field we can use the backward ODE to generate new samples. Since we have two parameters t and z, both can be combined into one in the backward process as follows

$$\begin{aligned} d\tilde{\mathbf{x}} &= (d\mathbf{x}, dz) \\ &= \left(\frac{d\mathbf{x}}{dt}\frac{dt}{dz}dz, dz\right) \\ &= \left(\mathbf{v}_{\mathbf{x}}\mathbf{v}_{\mathbf{x}}^{-1}dz, dz\right) \end{aligned} \tag{101}$$

Now we just choose some large $z = z_{max}$ and flow to $z = 0$. What we have in essence done is instead of using the hemisphere $r = z_{max}$ to define our initial distribution, we are using $z = z_{max}$ hyperplane.

6 Continuous Flow Models

Continuous flow models use an ODE to generate samples. Think of a flow defined by an ODE

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t) \tag{102}$$

The solution to the above differential equation is

$$\mathbf{x}(T) = \mathbf{x}(0) + \int_0^T \mathbf{f}(\mathbf{x}, t)dt \tag{103}$$

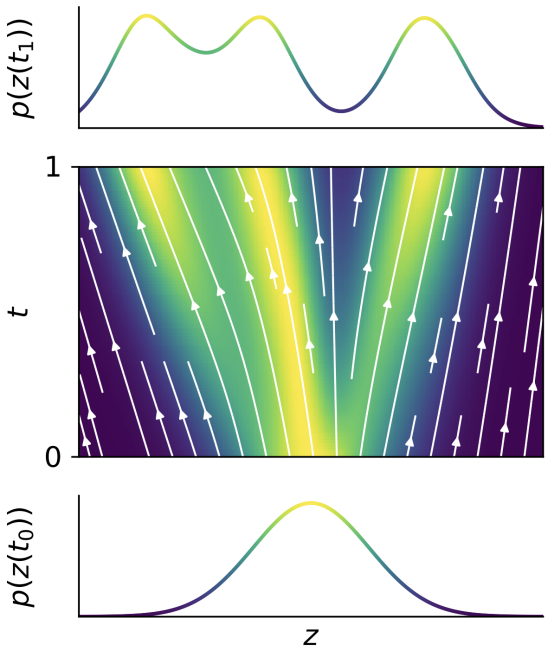

Figure 5: In continuous flow models we find vector fields that induce a transformation on from a normal distribution to the data distribution.

One a different note, consider a change of variables from $\mathbf{x}_0$ to $\mathbf{x}_1$. Then the change in corresponding probability density is given by

$$p(\mathbf{x}_0) = p(\mathbf{x}_1) \left| \frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0} \right| \tag{104}$$

or

$$\log(p(\mathbf{x}_1)) = \log(p(\mathbf{x}_0)) - \log \left(\left| \frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0} \right| \right) \tag{105}$$

What we want to find is a differential equation for the probability density under a small change in variable $\mathbf{x}$. So we parametrize $\mathbf{x}_1$ as

$$\mathbf{x}_1 = \mathbf{x}_0 \; + \; \epsilon \, \mathbf{f}(\mathbf{x}_0) \tag{106}$$

Therefore

$$\frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0} = \mathbf{I} \; + \; \epsilon \, \frac{\partial \mathbf{f}(\mathbf{x}_0)}{\partial \mathbf{x}_0} \tag{107}$$

and the determinant is given by

$$\left| \frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0} \right| = 1 \; + \; \epsilon \, \mathrm{Tr} \left(\frac{\partial \mathbf{f}(\mathbf{x}_0)}{\partial \mathbf{x}_0} \right) \; + \; O(\epsilon^2) \tag{108}$$

It is a simple exercise to see that the above relation holds. Therefore

$$\begin{aligned}\log\left(\left|\frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0}\right|\right) &= \log\left(1 \ + \ \epsilon \ \mathrm{Tr}\left(\frac{\partial \mathbf{f}(\mathbf{x}_0)}{\partial \mathbf{x}_0}\right) \ + \ O(\epsilon^2)\right) \\ &= \epsilon \ \mathrm{Tr}\left(\frac{\partial \mathbf{f}(\mathbf{x}_0)}{\partial \mathbf{x}_0}\right) \ + \ O(\epsilon^2)\end{aligned} \tag{109}$$

We finally get

$$\log(p(\mathbf{x}_1)) - \log(p(\mathbf{x}_0)) = -\epsilon \ \mathrm{Tr}\left(\frac{\partial \mathbf{f}(\mathbf{x}_0)}{\partial \mathbf{x}_0}\right) \ + \ O(\epsilon^2) \tag{110}$$

In the limit $\epsilon \to 0$, we get an ODE

$$\frac{\partial \log\, p(\mathbf{x}(t))}{\partial t} = -\mathrm{Tr}\left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right) \tag{111}$$

This equation gives the flow of log probability under the flow of the variable $\mathbf{x}$ given by

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}) \tag{112}$$

Notice that equation (111) can be considered as the Fokker-Planck equation for log probability density when the underlying dynamics is given by equation (112).

Learning the flow
We use a neural network $\mathbf{f}_\theta(\mathbf{x})$ for the flow. Integrating equation (111), we get

$$\log(p(\mathbf{x}(t_1))) = \log(p(\mathbf{x}(t_0))) - \int_{t_0}^{t_1} dt \ \mathrm{Tr}\left(\frac{\partial \mathbf{f}_\theta(\mathbf{x})}{\partial \mathbf{x}}\right) \tag{113}$$

Both the flow and the Fokker-Planck equation can be simultaneously used to learn the function $\mathbf{f}_\theta(\mathbf{x})$ using neural ODE.