SDE(随机微分方程)学习

第一节（ODE与SDE）：

SDE是由ODE衍生而来，因此讲SDE还是要从ODE开始，对于一个如下的ODE等式而言：

$$y^{''}(t)+qy'(t)+pt=w(t)$$

我们可以将其通过一定的转换变成二维一阶ODE，便是：

令$u=[y'(t),t]^T$,则$u'=\begin{bmatrix}a & b\\ c & d\end{bmatrix}u+\begin{bmatrix}e\\ f\end{bmatrix}w(t)$,解得$a=-q,b=-p,e=1,c=1,d=0,f=0$

因此：任意一个n阶的非齐次常微分方程都可以通过转换变成n维一阶的非齐次常微分方程。这使得对于任何高阶常微分方程我们都可以忽视他而仅仅思考低阶的常微分方程来简化问题。此外，对于上述通过u转化的而言，他属于以下方程的一个特例：

$$\frac{d\mathbf{x}(t)}{dt}=\mathbf{F}(t)\mathbf{x}(t)+\mathbf{L}(t)w(t)$$

如果对应到SDE中，那么$w(t)$则可以看成布朗运动，在diffusion model里面就是一个随机高斯扰动；而$\mathbf{L}(t)$则是diffusion function,$\mathbf{F}(t)$则是drift function。依次对应，接下来我们将讨论如何解该ODE方程。

第二节（常微分方程简单求解）：

对于一个非常简单的常微分方程：$\frac{dx}{x}=Fdt$而言，最简单的方法是两边同时积分，而另一种方法则是将x乘至方程右侧再进行积分，如下：

$$\int_0^t\frac{dx}{dt}dt=\int_0^t Fx(\tau)d\tau$$

$$=>x(t)=x(0)+\int_0^t Fx(\tau)d\tau$$

这个式子很神奇，它的左侧和右侧都包含了$x(\cdot)$,因此，我们将左边对应的值代入到右侧等式的$x(\tau)$中。得到：

$$x(t)=x(0)+\int_0^t F[x(0)+\int_0^\tau Fx(\tau')d\tau']d\tau$$

$$=>x(t)=x(0)+Fx(0)t+\int_0^t\int_0^\tau F^2x(\tau')d\tau'd\tau$$

以此类推，可以得到：

$$x(t)=x(0)+Fx(0)t+F^2x(0)\frac{t^2}{2}+F^3x(0)\frac{t^3}{6}+\cdots$$

$$=>x(t)=x(0)[1+F\frac{t}{1!}+F^2\frac{t^2}{2!}+F^3\frac{t^3}{3!}+\cdots]$$

，我们同样可以得到：

$x(t) = x(0)e^{Ft}$

那么对于x和F是高维向量的情况，该式子也同样适用：

$\mathbf{x}(t) = \mathbf{x}(0)e^{\mathbf{F}t}$,但是矩阵的指数无法直接按位进行指数运算得到，这是矩阵指数。

第三节（一般线性差分方程的解）：

对于前几节提到的SDE：$\frac{dx(t)}{dt} = Fx(t) + Lw(t)$而言，前面假定$F$是一个常数，但如果它是一个时变方程$F(t)$，那么之前的解法便失效了。在这里，我们将$x(t)$ 表述为如下形式：

$x(t) = \psi(t, t_0)x(t_0)$,where $\psi(\tau, t)$满足：

$$\frac{\partial \psi(\tau, t)}{\partial \tau} = F(\tau)\psi(\tau, t)$$

$$\frac{\partial \psi(\tau, t)}{\partial t} = -\psi(\tau, t)F(\tau)$$

$$\psi(\tau, t) = \psi(\tau, s)\psi(s, t)$$

$$\psi(\tau, t) = \psi^{-1}(t, \tau)$$

$$\psi(t, t) = I$$

这时候如果$L(t)w(t)$是0那么以上等式带入便是成立的。如果要完全解出来$x(t)$的表达式，那么它应该是如下形式：

$$x(t) = \psi(t, t_0)x(t_0) + \int_{t_0}^{t} \psi(t, \tau)L(\tau)w(\tau)d\tau$$

验证一下，代入到原式：

$$F(t)\psi(t, t_0)x(t_0) + \frac{d\int_{t_0}^{t} \psi(t, \tau)L(\tau)w(\tau)d\tau}{dt} = F[\psi(t, t_0)x(t_0) + \int_{t_0}^{t} \psi(t, \tau)L(\tau)w(\tau)d\tau] + L(t)w(t)$$

$$=> \frac{d\int_{t_0}^{t} \psi(t, \tau)L(\tau)w(\tau)d\tau}{dt} = F\int_{t_0}^{t} \psi(t, \tau)L(\tau)w(\tau)d\tau + L(t)w(t)$$

这里如果将$\int_{t_0}^{t} \psi(t, \tau)L(\tau)w(\tau)d\tau$看成一个整体，即函数$y(t)$，那么它就是一阶齐次常微分方程的通解，根据通解公式可以得到，解为：

$y = [\int L(\tau)w(\tau)e^{\int -F(\tau)d\tau}d\tau + C]e^{\int F(\tau)d\tau}$,根据 $\frac{\partial \psi(\tau,t)}{\partial \tau} = F(\tau)\psi(\tau,t)$ 可得，$\frac{\frac{\partial \psi(\tau,t)}{\partial \tau}}{\psi(\tau,t)} = F(\tau)$,因此 $\frac{\partial\, log\, \psi(\tau,t)}{\partial \tau} = F(\tau)$,同理：$\frac{\partial\, log\, \psi(t,\tau)}{\partial \tau} = -F(\tau)$,由于总共积分到时刻$t$,当前为时刻$\tau$,代入得：

$$y = [\int_{t_0}^{t} L(\tau)w(\tau)\psi(t,\tau)d\tau + C] \times [\psi(t,t) - \psi(t_0,t) + \psi(t_0,t)] = \int_{t_0}^{t} L(\tau)w(\tau)\psi(t,\tau)d\tau$$

第四节（傅里叶变换）：

正向的傅里叶变换可以被表示为：$F(g(t)) = \int_{-\infty}^{+\infty} g(t)exp(-iwt)dt$,对应地，逆向傅里叶变换可以被表示为：$F^{-1}(G(iw)) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} G(iw)exp(iwt)dw$,对于SDE方程而言，它的solve可以被表示为：\mathbb{R}^{d\times d}。此外，傅里叶变换还满足如下性质：$F(a(t) * b(t)) = F(a(t))F(b(t))$,其中$a(t)$和$b(t)$是两个时变函数，且 $a(t) * b(t) = \int_{-\infty}^{+\infty} a(t-\tau)b(\tau)d\tau$,也就是*并不是我们平常所认知的乘法，在SDE这本书中，作者将其称为convolution。为什么能用这个解决对应的问题呢？因为能够将高阶微分转化为$(iw)^n$，从而类似于ODE中提到的拉普拉斯变换求解一样，具体可以查看百度，这里不细说。

第五节（SDE可以被转化为ODE，diffusion model相关，即SDE论文的证明，因为不是摘抄自SDE那本书，且十分重要，用蓝色标注）：

SDE可以被表达为如下形式:$dx = \mathbf{f}(x,t)dt + \mathbf{G}(x,t)dw$*,其中*$\mathbf{f}(x,t)$*关于x的函数是*$\mathbb{R}^d \to \mathbb{R}^d$*,而*$\mathbf{G}(x,t)$*关于x的函数则是*$\mathbb{R}^d \to \mathbb{R}^{d\times d}$*，这里为什么是*$\mathbb{R}^{d\times d}$*是因为理论上漂移系数支持任何形式的映射，作者已经从理论证明它的推导可以hold住，所以实际上映射至*$\mathbb{R}^{d\times d}$*是一种较为复杂的形式了。*

*下一步作者借助了*Fokker–Planck方程来形成一个中间表达，并通过中间表达建立起了SDE与ODE等价的桥梁。

Fokker–Planck通过伊藤引理来推导，伊藤引理通过泰勒公式来推导，而Fokker–Planck描述了一个概率密度函数（PDF），这个PDF是关于x的PDF，这个东西其实求出来也没啥用，作者通过形式上的等价做出了SDE到ODE的转化，本质上是通过消去dw来达到的。其中Fokker–Planck方程如下：

$$\frac{\partial p(t,x)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}[\mathbf{f}_i(x)p(t,x)] + \frac{1}{2}\sum_i\sum_j \frac{\partial^2}{\partial x_i \partial x_j}[(GG^T)_{i,j}(x)p(t,x)] \quad (1)$$

这里的$p(t,x)$是PDF，也就是要求解的目标，而其余$f_i, (GG^T)_{i,j}$其实是对应向量和矩阵中的标量。这个公式推导很复杂，可以看知乎的相关文章：https://zhuanlan.zhihu.com/p/535688931。那通过这个公式，其实我们可以得到如下公式：

$$\frac{\partial p_t(x)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}[\mathbf{f}_i(x,t)p_t(x)] + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}[\sum_{k=1}^{d} G_{ik}(x,t)G_{jk}(x,t)p_t(x)] \quad (2)$$

这个公式和上面倒数第二个公式没什么区别，也就是把$p(t,x)$表示为$p_t(x)$，其余表达也是等同地替换。之所以写这个公式是因为Song原文是这个公式，那么它下一步其实是分别二阶偏导数并提取出公因子$\frac{\partial}{\partial x_i}$来进行合并，因此该式子可以被写成：

$$\frac{\partial p_t(x)}{\partial t} = -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}[\mathbf{f}_i(x,t)p_t(x)] + \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial x_i}[\sum_{j=1}^{d}\frac{\partial}{\partial x_j}[\sum_{k=1}^{d}G_{ik}(x,t)G_{jk}(x,t)p_t(x)]] \quad (3)$$

作者下一步是对$[\sum_{j=1}^{d}\frac{\partial}{\partial x_j}[\sum_{k=1}^{d}G_{ik}(x,t)G_{jk}(x,t)p_t(x)]]$进行操作来消去这些求和符号，但这里其实我看原文作者得出的结果是$p_t(x)\nabla\cdot[\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + p_t(x)\mathbf{G}(x,t)\mathbf{G}(x,t)^T\nabla_x\ log\ p_t(x)$,这个式子中加粗的G应该是矩阵，然而我并不清楚$\nabla\cdot$是什么？，但是对于这个结果的第二项，是可以了解出来的是一个向量，其中这个向量是关于公式（3）中以累加i为index的向量，当然累加到最后应该是标量。而$\sum_{k=1}^{d}$其实是一个点积操作，在结果中直接用$\mathbf{G}(x,t)\mathbf{G}(x,t)^T$进行代替。当然这个推导还没展示，作者使用的是乘法求导法则：

$$\sum_{j=1}^{d}\frac{\partial}{\partial x_j}[\sum_{k=1}^{d}G_{ik}(x,t)G_{jk}(x,t)p_t(x)] = \sum_{j=1}^{d}\frac{\partial}{\partial x_j}[\sum_{k=1}^{d}G_{ik}(x,t)G_{jk}(x,t)]p_t(x) + \sum_{j=1}^{d}[\sum_{k=1}^{d}G_{ik}(x,t)G_{jk}(x,t)p_t(x)\frac{\partial}{\partial x_j}log\,p_t(x)$$

$$=>p_t(x)\nabla\cdot[\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + p_t(x)\mathbf{G}(x,t)\mathbf{G}(x,t)^T\nabla_x\ log\ p_t(x) \quad (4)$$

其实是非常简单的变换，一个乘法求导法则，第二个就是通过将$\sum_{k=1}^{d}$转化为了点积，然后把公式4代回到公式3中，那么就可以得到：

$$-\sum_{i=1}^{d}\frac{\partial}{\partial x_i}[\mathbf{f}_i(x,t)p_t(x)] + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2}{\partial x_i\partial x_j}[\sum_{k=1}^{d}G_{ik}(x,t)G_{jk}(x,t)p_t(x)]$$

$$=> -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}[\mathbf{f}_i(x,t)p_t(x)] + \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial x_i}[p_t(x)\nabla\cdot[\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + p_t(x)\mathbf{G}(x,t)\mathbf{G}(x,t)^T\nabla_x\,log\,p_t(x)] \quad (5)$$

其实第二项挺对我这种新手不友好的，由于上面说了转化后得出来的结果是向量，而原来是标量，所以代入到公式4其实是存在出入的，然而由于$\frac{\partial}{\partial x_j}$的存在，因此向量$p_t(x)\nabla\cdot[\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + p_t(x)\mathbf{G}(x,t)\mathbf{G}(x,t)^T\nabla_x\ log\ p_t(x)$中其实只有一个值求导后不是0，而其他都是0，因此结果是没有差异的。

我们知道求导是可以加法结合的，因此最后公式5可以转化为：

$$=> -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left([f_i(x,t)p_t(x)] - \frac{1}{2}[p_t(x)\nabla\cdot[\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + p_t(x)\mathbf{G}(x,t)\mathbf{G}(x,t)^T\nabla_x\,log\,p_t(x)]\right) \quad (6)$$

然后，作者将$[\mathbf{f}_i(x,t)p_t(x)] - \frac{1}{2}[p_t(x)\nabla\cdot[\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + p_t(x)\mathbf{G}(x,t)\mathbf{G}(x,t)^T\nabla_x\ log\ p_t(x)]$看成了一个整体即$\hat{\mathbf{f}}_i(x,t)p_t(x)$，而上面这些项中都有$p_t(x)$（所以这里能看出为什么上面乘法法则求导时第二项要分出来一个$log\ p_t(x)$），因此可以消去$p_t(x)$可以得到：

$$\hat{\mathbf{f}}_i(x,t) = \mathbf{f}_i(x,t) - \frac{1}{2}[\nabla \cdot [\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + \mathbf{G}(x,t)\mathbf{G}(x,t)^T \nabla_x \, log \, p_t(x)]$$

现在，如果对于一个ODE方程，便是$\mathbf{G}(x,t) = \mathbf{0}$,即没有随机项，那么我们可以将这种情况下的$\mathbf{G}(x,t)$定义为$\widehat{\mathbf{G}}(x,t)$,如果是这样，那么推导得出的Fokker–Planck方程的PDF应该满足：

$$\frac{\partial}{\partial t} p_t(x) = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}([\mathbf{f}_i(x,t)p_t(x)]) \quad (7)$$

那么只要这个时候的$f_i(x,t) = \widehat{f}_i(x,t)$就可以了。所以，原来的$f_i(x,t)$和$\mathbf{G}(x,t)$只要通过
$\hat{\mathbf{f}}_i(x,t) = \mathbf{f}_i(x,t) - \frac{1}{2}[\nabla \cdot [\mathbf{G}(x,t)\mathbf{G}(x,t)^T] + \mathbf{G}(x,t)\mathbf{G}(x,t)^T \nabla_x \, log \, p_t(x)]$的转化，便可满足ODE方程：
$dx = \hat{\mathbf{f}}(x,t)dt$

第六节（非马尔科夫链也能hold住gaussian case且可以用ELBO推导，diffusion model相关，即DDIM论文的证明，因为不是摘抄自SDE那本书，且十分重要，用蓝色标注）：

首先按照ddpm的前向，那么x_0应该是一个图像，但是图像也是二维向量也可以被拉成一维向量。作者定义这个x_0存在着K个正数，在这个基础上，作者定义了$q(x_t|x_0)$,其表示为如下形式：

$$q(x_t|x_0) = \text{Cat}\left(\alpha_t x_0 + (1-\alpha_t)\mathbf{1}_K\right)$$

这里的$\mathbf{1}_K$代表了一个形状和x_0相同的且值皆为$1/K$的向量。这里的α定义会更加严格，需保证：
$1 = \alpha_o \geq \alpha_1 \geq \cdots \geq \alpha_t = 0$这个形式，这里的$\text{Cat}$应该是描述了一个分布，是categorical的缩写。类比到Gaussian case，作者是希望能够通过这种形式来hold住这种情况。然后，作者定义了$q(x_{t-1}|x_t, x_0)$:

$$q(x_{t-1}|x_t, x_0) = \text{Cat}(\sigma_t x_t + (\alpha_{t-1} - \sigma_t\alpha_t)x_0 + ((1-\alpha_{t-1}) - (1-\alpha_t)\sigma_t)\mathbf{1}_K)$$

这个公式看上去很奇怪，但实际上却是和DDPM的定义相同，查看博客：https://spaces.ac.cn/archives/9164 其中对于$q(x_{t-1}|x_t, x_0)$的定义为 $\text{Cat}(\frac{\alpha_t\overline{\beta}_{t-1}^2}{\overline{\beta}_t^2}x_t + \frac{\overline{\alpha}_{t-1}\beta_t^2}{\overline{\beta}_t^2}x_0 + \frac{\overline{\beta}_{t-1}^2\beta_t^2}{\overline{\beta}_t^2}\mathbf{1}_K)$,虽然形式上差异巨大，但当我们进行简单的换元，即 $\sigma_t = \frac{\alpha_t\overline{\beta}_{t-1}^2}{\overline{\beta}_t^2}$ 后，那么这两个式子就是同一个结果，但实际上，这边作者在定义中没有明确保证$\sigma_t = \frac{\alpha_t\overline{\beta}_{t-1}^2}{\overline{\beta}_t^2}$ 。因此，可以理解常见的DDPM其实是作者定义情况下的一个子集，毕竟作者定义在这里的σ_t是未知的。然后，对于x_0来说，自然还是未知的，因此作者采用了可学习的模型进行预测，因此分布定义为：

$$p_\theta(x_{t-1}|x_t) = \text{Cat}(\sigma_t x_t + (\alpha_{t-1} - \sigma_t\alpha_t)f_\theta^{(t)}(x_t) + ((1-\alpha_{t-1}) - (1-\alpha_t)\sigma_t)\mathbf{1}_K)$$

其中$f_\theta^{(t)}(x_t)$是一个可学习的映射，能够将x_t映射到一个K维度的向量。对于随机项$\mathbf{1}_K$,它的系数$((1-\alpha_{t-1}) - (1-\alpha_t)\sigma_t)$如果不断变小，那么这个基于SDE的反向采样会逐渐丧失随机性而变成ODE，在这种情况下，使用$f_\theta^{(t)}(x_t)$预测的概率会不断变高，那么对于一个KL散度：

$$D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$$

就可以被很好的定义。这是因为我们只去选择了中间那一项$(\alpha_{t-1}-\sigma_t\alpha_t)f_\theta^{(t)}(x_t)$来作为预测结果。因此，它满足如下的放缩公式（作者声称之所以满足这个是因为KL散度是凸的）：

$$D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) \leq (\alpha_{t-1}-\sigma_t\alpha_t)D_{KL}(\mathrm{Cat}(x_0)||\mathrm{Cat}(f_\theta^{(t)}(x_t)))$$

对于这个upper bound来说，右侧只是一个多分类损失，因此我们可以得出类似的论点，即σ_t的变化如何影响目标（直到重新加权）。但我可能看不出来这个放缩是如何推出来的，因此我决定自己理解一遍。简单来说，不论是分布中$p_\theta(x_{t-1}|x_t)$还是$q(x_{t-1}|x_t,x_0)$,其中都有公共项：$\sigma_t x_t+((1-\alpha_{t-1})-(1-\alpha_t)\sigma_t)\mathbf{1}_K$

因此我们只需要证明：$D_{KL}(a+c||b+c)\leq D_{KL}(a||b)$即可，而由于$D_{KL}$是凸函数，因此，其满足$D_{KL}(a+c||b+c)\leq D_{KL}(a||b)+D_{KL}(c||c)$即可，而由于$D_{KL}(c||c)$等于0，因此得证。

第七节（如果对$q_\sigma(x_t|x_0)$满足$\mathcal{N}(\sqrt{\alpha_t}x_0,(1-\alpha_t)\mathbf{I})$,那么$q_\sigma(x_{t-1}|x_0)$满足$\mathcal{N}(\sqrt{\alpha_{t-1}}x_0,(1-\alpha_{t-1})\mathbf{I})$，即DDIM论文的证明，因为不是摘抄自SDE那本书，且十分重要，用蓝色标注）：

我们先有如下条件（这里为了方便省略σ）：

$$q(x_{t-1}|x_t,x_0)=\mathcal{N}(\sqrt{\alpha_{t-1}}x_0+\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t-\sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}},\sigma_t^2\mathbf{I})$$

$$q(x_t|x_0)=\mathcal{N}(\sqrt{\alpha_t}x_0,(1-\alpha_t)\mathbf{I})$$

我们需要根据上述条件，得出$q(x_{t-1}|x_0)$是什么。而$q(x_{t-1}|x_0)$来自于积分，即：

$$q(x_{t-1}|x_0)=\int_{-\infty}^{+\infty}q(x_t|x_0)q(x_{t-1}|x_t,x_0)dx_t$$

这个公式要求解需要根据一本书：PRML：github:PRML 中的2.115公式求解，这是一个以高斯分布为条件分布，然后求积分的公式，如下：

给定$\boldsymbol{x}$的一个边缘高斯分布，以及在给定$\boldsymbol{x}$的条件下$\boldsymbol{y}$的条件高斯分布，形式为

$$p(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}\mid\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\boldsymbol{y}\mid\boldsymbol{x})=\mathcal{N}(\boldsymbol{y}\mid\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b},\boldsymbol{L}^{-1}) \tag{2.114}$$

$\boldsymbol{y}$的边缘分布以及给定$\boldsymbol{y}$的条件下$\boldsymbol{x}$的条件分布为

$$p(\boldsymbol{y})=\mathcal{N}(\boldsymbol{y}\mid\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b},\boldsymbol{L}^{-1}+\boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T) \tag{2.115}$$

$$p(\boldsymbol{x}\mid\boldsymbol{y})=\mathcal{N}(\boldsymbol{x}\mid\boldsymbol{\Sigma}\{\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\},\boldsymbol{\Sigma}) \tag{2.116}$$

其中

$$\boldsymbol{\Sigma}=(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1} \tag{2.117}$$

对应地，新的均值为 $\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\sqrt{\alpha_t}x_0-\sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}} = \sqrt{\alpha_{t-1}}x_0$,而新的方差为：

$$\sigma_t^2 I + \left(\frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}}\right)^2 (1-\alpha_t)\mathbf{I} = (1-\alpha_{t-1})\mathbf{I}$$

。因此，$q(x_{t-1}|x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0, (1-\alpha_{t-1})\mathbf{I})$。

所以归纳法可证明所有$x_t, t\in[0,T]$都成立。

第八节（VP-SDE和VE-SDE,属于文章Score-Based Generative Modeling through SDE的推导）：

VE-SDE:

$$x_{i+1} = x_i + \sqrt{\sigma_{i+1}^2 - \sigma_i^2}z_i$$

$$\text{if} \quad x_{i+1}-x_i \ll \delta, \quad s.t. \quad \delta\to 0, \quad \text{then}:$$

$$=> x(t+\delta) = x(t) + \sqrt{\sigma^2(t+\delta)-\sigma^2(t)}z(t)$$

$$=> x(t+\delta) = x(t) + \sqrt{\sigma^2(t)+\delta\nabla_t\sigma^2(t)+O(t^2)-\sigma^2(t)}z(t)$$

$$\approx> x(t+\delta) = x(t) + \sqrt{\delta\nabla_t\sigma^2(t)}z(t)$$

$$=> \delta\nabla x(t) = \sqrt{\nabla_t\sigma^2(t)}\cdot\sqrt{\delta}z(t)$$

$$=> dx = \sqrt{\nabla_t\sigma^2(t)}\cdot\sqrt{\delta}z(t)$$

$$\approx> dx = \sqrt{\frac{d[\sigma^2(t)]}{dt}}dw$$

VP-SDE:

$$x_{i+1} = \sqrt{1-\beta_{i+1}}x_i + \sqrt{\beta_{i+1}}z_i$$

$$=> x(t+\delta) = \sqrt{1-\beta(t+\delta)}x(t) + \sqrt{\beta(t+\delta)}z(t), \quad (a)$$

, where $\beta(t)$is a function that $0<\beta(t_1)<\beta(t_2)<\cdots<\beta(t_N)<1$, s.t. $0\le t_1<t_2<\cdots<t_N\le 1$

$$=> x(t) = \sqrt{1-\beta(t)}x(t-\delta) + \sqrt{\beta(t)}z(t-\delta), \quad (b)$$

则

为了简单起见，我们令$x(t)\approx x(t-\delta), z(t)\approx z(t-\delta)$, 并计算$(a)-(b)$得：

$$=> dx = [\sqrt{1-\beta(t+\delta)} - \sqrt{1-\beta(t)}]x(t) + [\sqrt{\beta(t+\delta)} - \sqrt{\beta(t)}]z(t)$$

$$=> dx = [\sqrt{1-\beta(t)} + \delta\nabla_t\sqrt{1-\beta(t)} - \sqrt{1-\beta(t)}]x(t) + [\sqrt{\beta(t)} + \delta\nabla_t\sqrt{\beta(t)} - \sqrt{\beta(t)}]z(t)$$

$$=> dx = [\delta \nabla_t \sqrt{1-\beta(t)}]x(t) + [\delta \nabla_t \sqrt{\beta(t)}]z(t)$$

$$=> dx = [-\delta \frac{1}{2} \frac{1}{\sqrt{1-\beta(t)}} \nabla_t \beta(t)]x(t) + [\delta \nabla_t \sqrt{\beta(t)}]z(t)$$

$$=> dx = [-\frac{1}{2} \frac{1}{\sqrt{1-\beta(t)}} d\beta]x(t) + [d\sqrt{\beta}]z(t)$$

由于$\beta(a)$的值域接近于0，所以：

$$\approx> dx = [-\frac{1}{2} d\beta]x(t) + [d\sqrt{\beta}]z(t)$$ 或者 $$dx = [-\frac{1}{2} \nabla_t \beta(t)]x(t)dt + [\nabla_t \sqrt{\beta(t)}]dw$$