一、条件概率公式与高斯分布的KL散度

1、条件概率的一般形式

$$P(A,B,C)=P(C\mid B,A)P(B,A)=P(C\mid B,A)P(B\mid A)P(A)$$

$$P(B,C|A)=P(B|A)P(C|A,B)$$

2、基于马尔科夫假设的条件概率

如果满足马尔科夫链关系A->B->C，那么有

$$P(A,B,C)=P(C\mid B,A)P(B,A)=P(C\mid B)P(B\mid A)P(A)$$

$$P(B,C|A)=P(B|A)P(C|B)$$

3、高斯分布的KL散度公式

对于两个单一变量的高斯分布p和q而言，它们的KL散度为

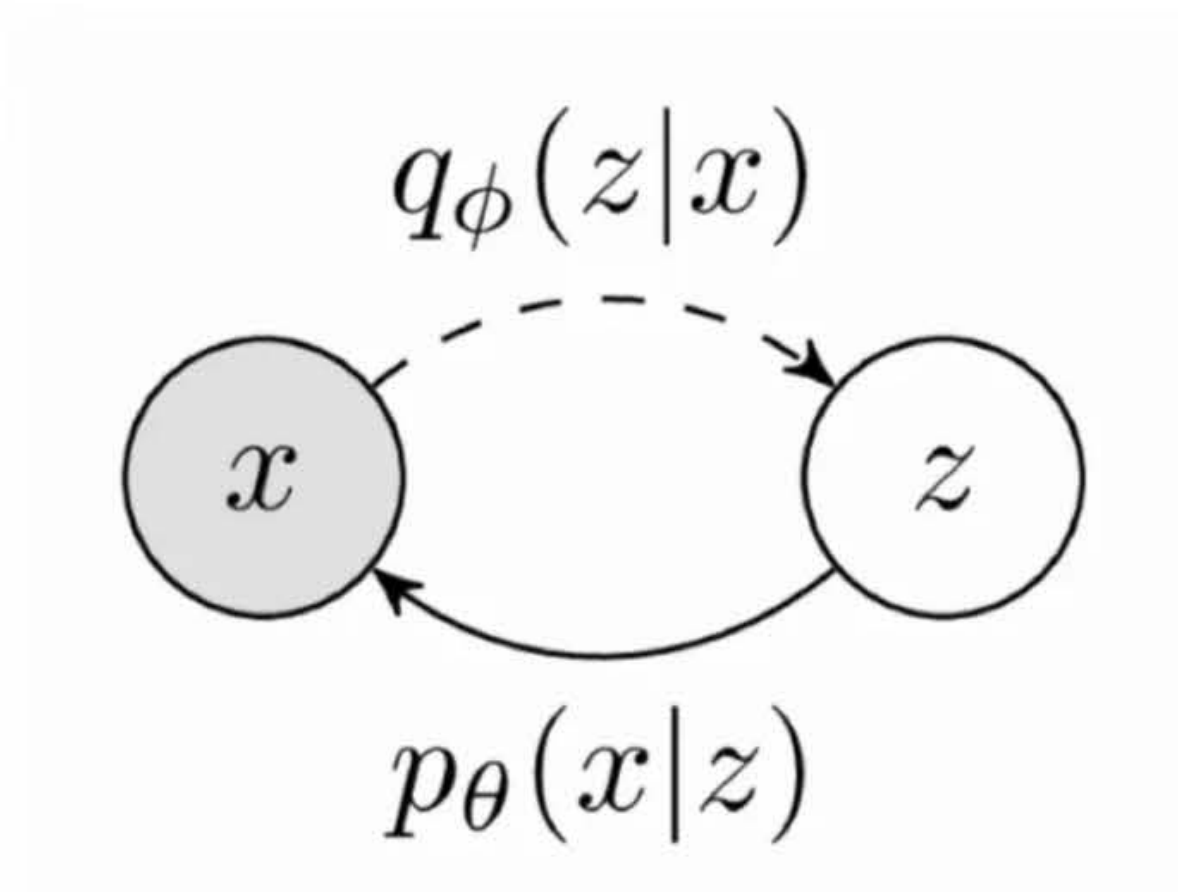
$$KL(p,q)=log\frac{\sigma_2}{\sigma_1}+\frac{\sigma^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}-\frac{1}{2}$$

4、重参数技巧

若希望从高斯分布$N(\mu,\sigma)$中采样，可以先从标准分布$N(0,1)$采样出z，再得到$\sigma*z+\mu$。这样做的好处是将随机性转移到了z这个常量上，而σ和μ则当做仿射变换网络的一部分。

二、 VAE与多层VAE回顾

1、单层VAE的原理公式与置信下界



q加噪, p 去噪

$q_\phi(z|x)$ 表示函数名q, 给定x条件下.z的分布函数. 分布函数的参数是 ϕ

所以机器学习里面的函数都可以看做 $f_\phi(y|x)$ 这种分布.

$$p(x)=\int_z p_\theta(x|z)p(z)$$

→ 因为 $p_\theta(x|z)=\frac{p_\theta(x,z)}{p(z)}$ 所以这个式子是边缘分布的积分.

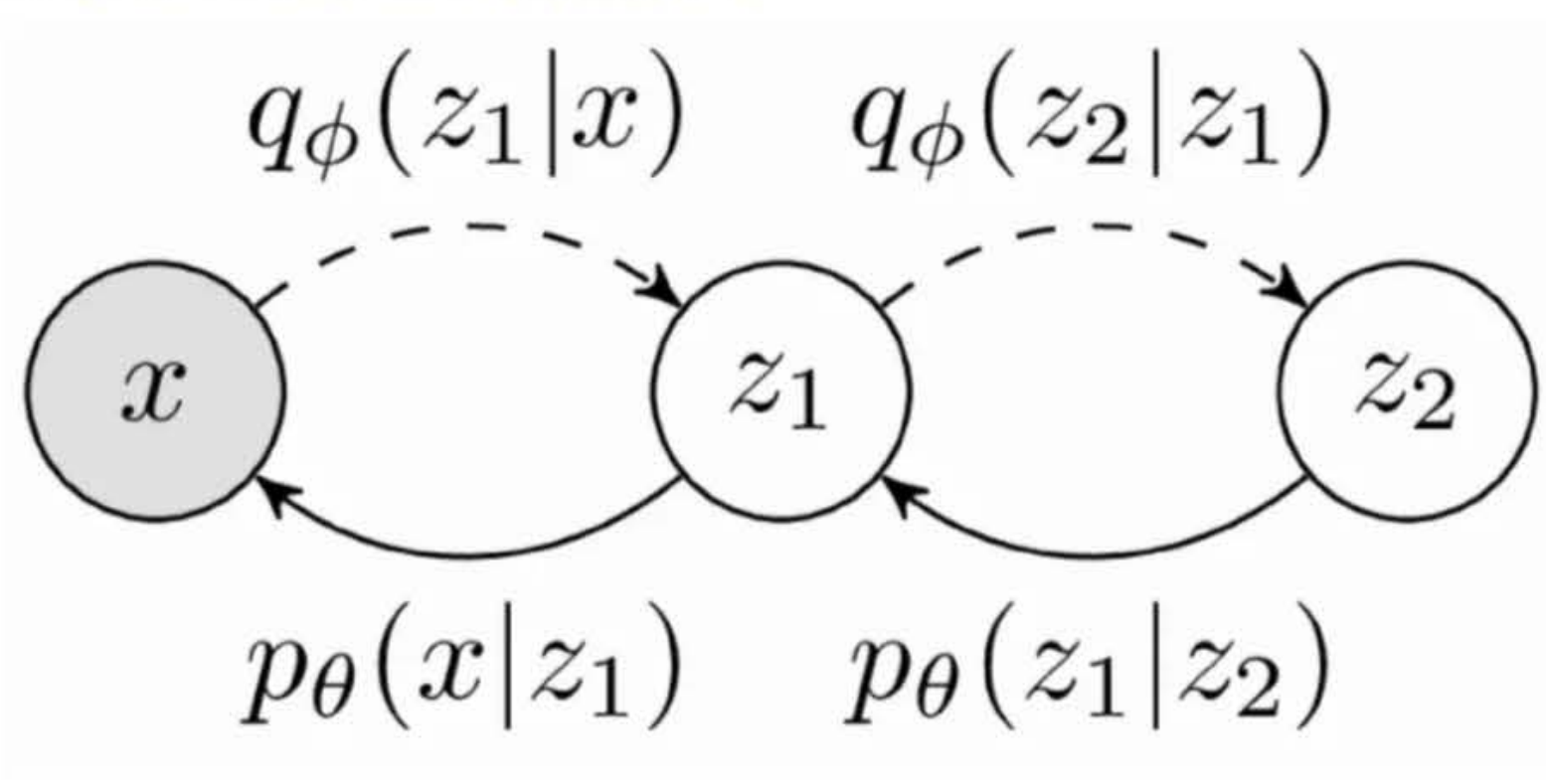
$$p(x)=\int q_\phi(z|x)\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$$

积分里面的z分布函数就可以提出来当期望函数.

$$\log p(x)=\log\mathbb{E}_{z\sim q_\phi(z|x)}\left[\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right]$$

$$\log p(x)\geq\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right]$$

2、多层VAE的原理公式与置信下界



$$p(x)=\int_{z_1}\int_{z_2}p_\theta(x,z_1,z_2)dz_1,dz_2$$

$$p(x)=\int\int q_\phi(z_1,z_2|x)\frac{p_\theta(x,z_1,z_2)}{q_\phi(z_1,z_2|x)}$$

$$p(x)=\mathbb{E}_{z_1,z_2\sim q_\phi(z_1,z_2|x)}\left[\frac{p_\theta(x,z_1,z_2)}{q_\phi(z_1,z_2|x)}\right]$$

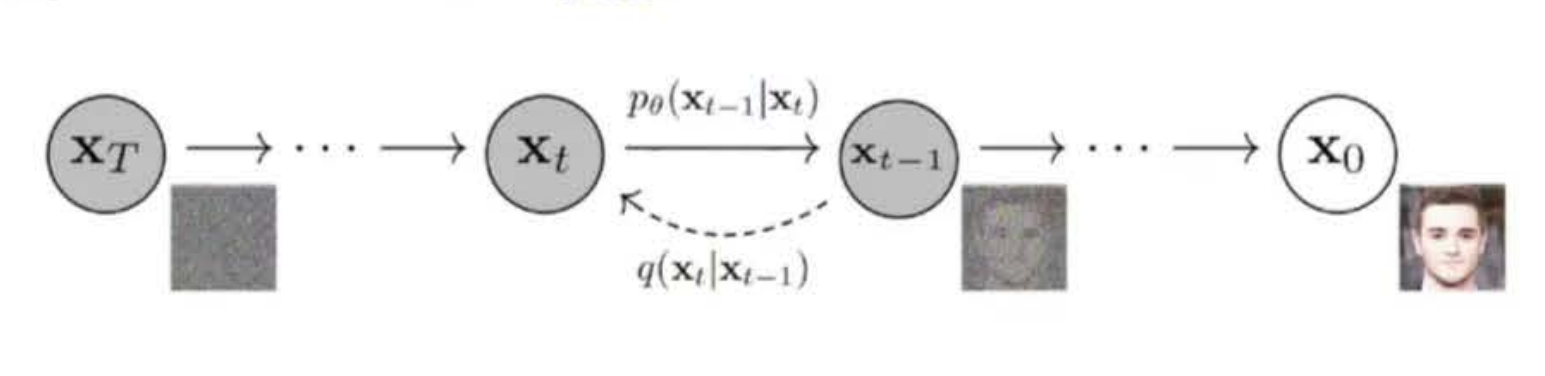
$$\log p(x)\geq\mathbb{E}_{z_1,z_2\sim q_\phi(z_1,z_2|x)}\left[\log\frac{p_\theta(x,z_1,z_2)}{q_\phi(z_1,z_2|x)}\right]$$

$$p(x,z_1,z_2)=p(x|z_1)p(z_1|z_2)p(z_2)$$

$$q(z_1,z_2|x)=q(z_1|x)q(z_2|z_1)$$

$$\mathcal{L}(\theta,\phi)=\mathbb{E}_{q(z_1,z_2|x)}[\log p(x|z_1)-\log q(z_1|x)+\log p(z_1|z_2)-\log q(z_2|z_1)+\log p(z_2)]$$

三、 Diffusion Model图示



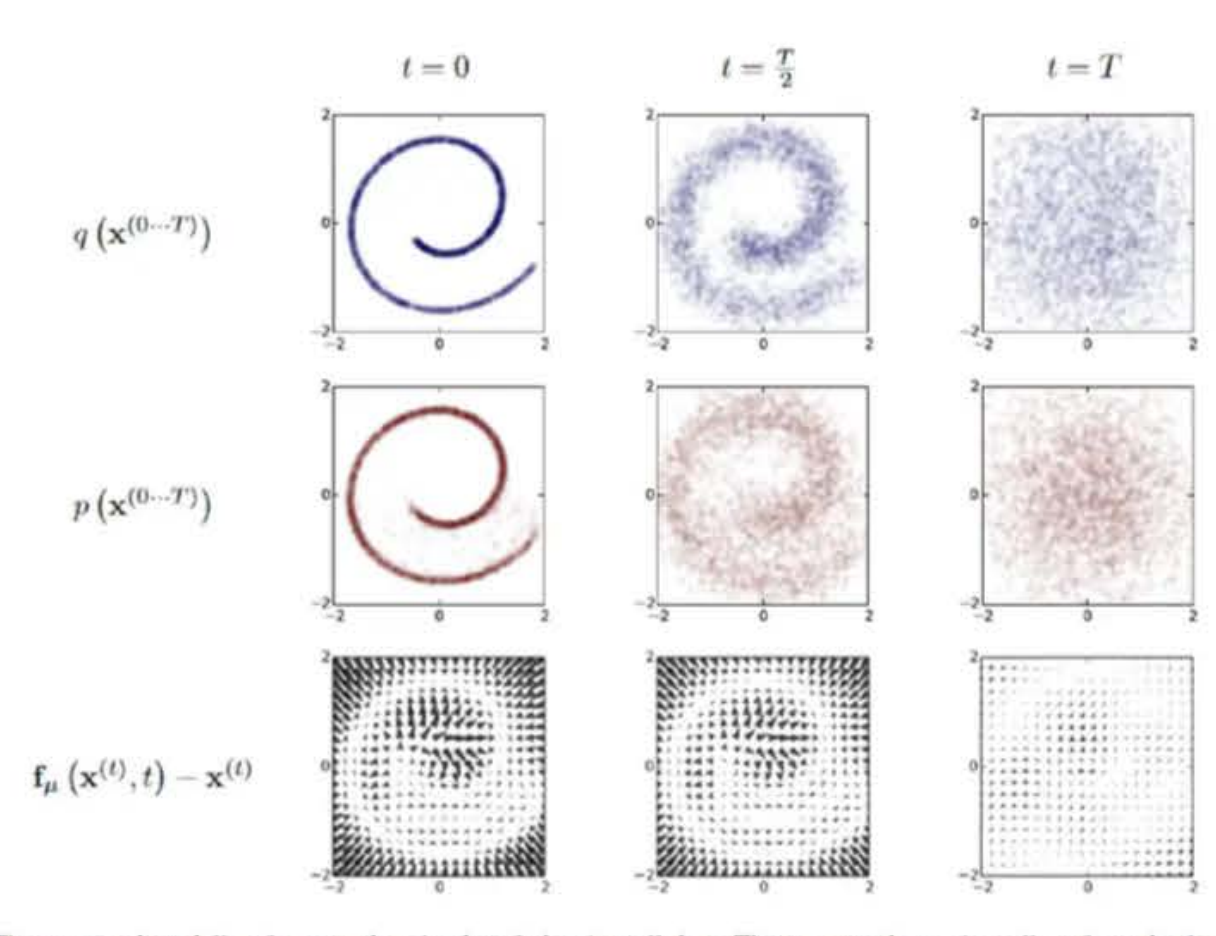


Figure 1. The proposed modeling framework trained on 2-d swiss roll data. The top row shows time slices from the forward trajectory $q\left(\mathbf{x}^{(0\cdots T)}\right)$. The data distribution (left) undergoes Gaussian diffusion, which gradually transforms it into an identity-covariance Gaussian (right). The middle row shows the corresponding time slices from the trained reverse trajectory $p\left(\mathbf{x}^{(0\cdots T)}\right)$. An identity-covariance Gaussian (right) undergoes a Gaussian diffusion process with learned mean and covariance functions, and is gradually transformed back into the data distribution (left). The bottom row shows the drift term, $\mathbf{f}_\mu\left(\mathbf{x}^{(t)},t\right)-\mathbf{x}^{(t)}$, for the same reverse diffusion process.

四、扩散过程 (Diffusion Process)

1、给定初始数据分布$x_0\sim q(x)$，可以不断地向分布中添加高斯噪声，该噪声的标准差是以固定值β_t而确定的，均值是以固定值β_t和当前t时刻的数据x_t决定的。这个过程是一个马尔科夫链过程.

→ 最终目标为了刻画q(X)

2、随着t的不断增大，最终数据分布x_T变成了一个各向独立的高斯分布。

Given a data point sampled from a real data distribution $\mathbf{x}_0\sim q(\mathbf{x})$, let us define a *forward diffusion process* in which we add small amount of Gaussian noise to the sample in T steps, producing a sequence of noisy samples $\mathbf{x}_1,\ldots,\mathbf{x}_T$. The step sizes are controlled by a variance schedule $\{\beta_t\in(0,1)\}_{t=1}^{t}$.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1})=\mathcal{N}(\mathbf{x}_t;\sqrt{1-\beta_t}\mathbf{x}_{t-1},\beta_t\mathbf{I})\quad q(\mathbf{x}_{1:T}|\mathbf{x}_0)=\prod_{t=1}^{T}q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

The data sample $\mathbf{x}_0$ gradually loses its distinguishable features as the step t becomes larger. Eventually when $T\to\infty$, $\mathbf{x}_T$ is equivalent to an isotropic Gaussian distribution.

3、任意时刻的$q(x_t)$推导也可以完全基于x_0和β_t来计算出来，而不需要做迭代

注意这里，两个正态分布$X\sim N(\mu_1,\sigma_1)$和$Y\sim N(\mu_2,\sigma_2)$的叠加后的分布$aX+bY$的均值为$a\mu_1+b\mu_2$，方差为$a^2\sigma_1^2+b^2\sigma_2^2$。所以$\sqrt{\alpha_t-\alpha_t\alpha_{t-1}}z_{t-2}+\sqrt{1-\alpha_t}z_{t-1}$可以重参数化成只含一个随机变量$z$构成的$\sqrt{1-\alpha_t\alpha_{t-1}}z$的形式。

A nice property of the above process is that we can sample $\mathbf{x}_t$ at any arbitrary time step t in a closed form using reparameterization trick. Let $\alpha_t=1-\beta_t$ and $\bar{\alpha}_t=\prod_{i=1}^{T}\alpha_i$:

$$\begin{aligned}\mathbf{x}_t&=\sqrt{\alpha_t}\mathbf{x}_{t-1}+\sqrt{1-\alpha_t}\mathbf{z}_{t-1} &&;\text{where }\mathbf{z}_{t-1},\mathbf{z}_{t-2},\cdots\sim\mathcal{N}(\mathbf{0},\mathbf{I})\\&=\sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2}+\sqrt{1-\alpha_t\alpha_{t-1}}\bar{\mathbf{z}}_{t-2} &&;\text{where }\bar{\mathbf{z}}_{t-2}\text{ merges two Gaussians (*).}\\&=\ldots\\&=\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\mathbf{z}\\q(\mathbf{x}_t|\mathbf{x}_0)&=\mathcal{N}(\mathbf{x}_t;\sqrt{\bar{\alpha}_t}\mathbf{x}_0,(1-\bar{\alpha}_t)\mathbf{I})\end{aligned}$$

(*) Recall that when we merge two Gaussians with different variance, $\mathcal{N}(\mathbf{0},\sigma_1^2\mathbf{I})$ and $\mathcal{N}(\mathbf{0},\sigma_2^2\mathbf{I})$, the new distribution is $\mathcal{N}(\mathbf{0},(\sigma_1^2+\sigma_2^2)\mathbf{I})$. Here the merged standard deviation is $\sqrt{(1-\alpha_t)+\alpha_t(1-\alpha_{t-1})}=\sqrt{1-\alpha_t\alpha_{t-1}}$.

Usually, we can afford a larger update step when the sample gets noisier, so $\beta_1<\beta_2<\cdots<\beta_T$ and therefore $\bar{\alpha}_1>\cdots>\bar{\alpha}_T$.

→ 因为beta1距离目标越远所以越需要降低学习难度. $\beta_1<\beta_2\cdots<\beta_T$

五、逆扩散过程 (Reverse Process)

逆过程是从高斯噪声中恢复原始数据，我们可以假设它也是一个高斯分布，但是无法逐步地去拟合分布，所以需要构建一个参数分布来去做估计。逆扩散过程仍然是一个马尔科夫链过程。

If we can reverse the above process and sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, we will be able to recreate the true sample from a Gaussian noise input, $\mathbf{x}_T\sim\mathcal{N}(\mathbf{0},\mathbf{I})$. Note that if β_t is small enough, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ will also be Gaussian. Unfortunately, we cannot easily estimate $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ because it needs to use the entire dataset and therefore we need to learn a model p_θ to approximate these conditional probabilities in order to run the *reverse diffusion process*.

$$p_\theta(\mathbf{x}_{0:T})=p(\mathbf{x}_T)\prod_{t=1}^{T}p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)=\mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta(\mathbf{x}_t,t),\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t))$$

六、后验的扩散条件概率$q(x_{t-1}|x_t,x_0)$分布是可以用公式表达的

多加一个x0,的话有解析式.

也就是说，给定x_t和x_0，我们是可以计算出x_{t-1}

注意：高斯分布的概率密度函数是$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

注意：$ax^2+bx=a(x+\frac{b}{2a})^2+C$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)=\mathcal{N}(\mathbf{x}_{t-1};\tilde{\boldsymbol{\mu}}(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{I})$$

Using Bayes' rule, we have:

$$\begin{aligned}q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)&=q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}\\&\propto\exp\Big(-\frac{1}{2}\big(\frac{(\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t}+\frac{(\mathbf{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(\mathbf{x}_t-\sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t}\big)\Big)\\&=\exp\Big(-\frac{1}{2}\big((\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})\mathbf{x}_{t-1}^2-(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{2\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{x}_0)\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0)\big)\Big)\end{aligned}$$

← 上面公式带入得到这个.

where $C(\mathbf{x}_t,\mathbf{x}_0)$ is some function not involving $\mathbf{x}_{t-1}$ and details are omitted. Following the standard Gaussian density function, the mean and variance can be parameterized as follows:

$$\tilde{\beta}_t=1/(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t$$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0)=(\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{x}_0)/(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})=\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0$$

变换成一个高斯函数.

!!!紧急重要更新：大家注意，上面蓝色公式的第二项x_0系数中的分子分母中的所有$\bar{\alpha}_t$应该写成$\bar{\alpha}_{t-1}$。

估计是因为 α_t, α_{t-1} 差别很小.所以这里面都记做 α_t 了.

根据前面x_0与x_t之间的关系式，我们可以知道

$$\mathbf{x}_0=\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\mathbf{z}_t)$$

将x_0的表达式代入到$q(x_{t-1}|x_t,x_0)$的分布中，可以重新给出该分布的均值表达式，这个时候表达式中不再含有x0，并且多了噪声项，这为后面我们设计神经网络提供了基础。也就是说，在给定x_0的条件下，后验

条件高斯分布的均值计算只与x_t和z_t有关。z_t是t时刻的随机正态分布变量，源自重参数化。

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_t&=\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\mathbf{z}_t)\\&=\frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{z}_t\Big)\end{aligned}$$

$$=\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}+\frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\right)x_t-\frac{\beta_t z_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}$$

$$=\left(\frac{\alpha_t-\bar{\alpha}_t+\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\right)x_t-\frac{\beta_t z_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}$$

七、目标数据分布的似然函数

我们可以在负对数似然函数的基础上加上一个KL散度，于是就构成了负对数似然的上界了，上界越小，负对数似然自然也就越小，那么对数似然就越大了。

$$\begin{aligned}-\log p_\theta(\mathbf{x}_0)&\leq-\log p_\theta(\mathbf{x}_0)+D_{\mathrm{KL}}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0))\\&=-\log p_\theta(\mathbf{x}_0)+\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})/p_\theta(\mathbf{x}_0)}\Big]\\&=-\log p_\theta(\mathbf{x}_0)+\mathbb{E}_q\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}+\log p_\theta(\mathbf{x}_0)\Big]\\&=\mathbb{E}_q\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big]\\\text{Let }L_{\mathrm{VLB}}&=\mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big]\geq-\mathbb{E}_{q(\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0)\end{aligned}$$

→ 因为KL散度横大于0

进一步可以写出如上公式的交叉熵的上界，接下来，我们可以对交叉熵的上界进行化简

$$\begin{aligned}L_{\mathrm{VLB}}&=\mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big]\\&=\mathbb{E}_q\Big[\log\frac{\prod_{t=1}^{T}q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_T)\prod_{t=1}^{T}p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\Big]\\&=\mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T)+\sum_{t=1}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\Big]\\&=\mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T)+\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}+\log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big]\\&=\mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T)+\sum_{t=2}^{T}\log\Big(\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\cdot\frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}\Big)+\log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big]\\&=\mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T)+\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}+\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}+\log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big]\\&=\mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T)+\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}+\log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{q(\mathbf{x}_1|\mathbf{x}_0)}+\log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big]\\&=\mathbb{E}_q\Big[\log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}+\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\Big]\\&=\mathbb{E}_q[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_T))}_{L_T}+\sum_{t=2}^{T}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}]\end{aligned}$$

变量是theta

这里论文将$p_\theta(x_{t-1}|x_t)$分布的方差设置成一个与β相关的常数，因此可训练的参数只存在于其均值中

→ 这个非常关键.

对于两个单一变量的高斯分布p和q而言，它们的KL散度为
$KL(p,q)=log\frac{\sigma_2}{\sigma_1}+\frac{\sigma^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}-\frac{1}{2}$

$$L_{t-1}=\mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0)-\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\|^2\right]+C$$

既然这里的loss是从KL divergence出发的，或者说是与分布有关的，那我们可以设计一个黑箱子神经网络，把它称之为D_θ网络。对于D_θ网络，输入是x_t和时间编码t，对于输出是什么，取决于我们的建模目标。

而我们可以有多种建模目标：【1】直观的做法是让D_θ网络的输出等于前向过程中的后验分布均值$\tilde{\mu}(x_t,x_0)$，这种建模方法俗称预测后验分布的期望值；【2】根据$\tilde{\mu}(x_t,x_0)$表达式，它里面的x_0对于D_θ网络是未知的，因此第二种做法是让D_θ网络的输出等于x_0，这种做法即直接预测原始数据。有人问，既然可以通过D_θ网络直接预测x_0了，那是不是采样过程就直接计算$D_\theta(x_T,T)$的输出即可认为是生成了样本了呢？答案是直接一步到位，质量会比较差，还是需要通过马尔科夫高斯条件迭代而获得最终高质量的生成样本；【3】当我们把$\tilde{\mu}(x_t,x_0)$中的x_0用x_t去表示的时候，$\tilde{\mu}(x_t,x_0)$就变成了如下只包含x_t和随机变量ϵ的式子。其中x_t对于D网络是已知的，而ϵ是未知的，因此这个时候，我们可以选择建模目标是让D_θ网络的输出等于ϵ了

$$\begin{aligned}L_{t-1}-C&=\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}),\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon})-\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon})\right)-\boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}),t)\right\|^2\right]\\&=\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon})-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)-\boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}),t)\right\|^2\right]\end{aligned}$$

在DDPM论文中，作者选择了方案【3】，即让D_θ网络的输出等于ϵ，预测噪音法。于是，新的逆向条件分布的均值可以表示成(下式中的ϵ_θ相当于我们定义的广义的D_θ网络的具体目标形式):

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)=\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t,\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t))\right)=\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right)$$

于是L_{t-1}可以化简成如下表达式

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right]$$

DDPM作者又发现，干脆将系数丢掉，训练更加稳定质量更好，于是有了下面的L_{simple}

$$L_{\mathrm{simple}}(\theta):=\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right]$$