

Therefore, back to M-step:

$$\max_{\mu, \Sigma, \pi} E_z \log P(x, z)$$

$$= \max_{\mu, \Sigma, \pi} \sum_{i \in [n]} \sum_{k \in [K]} \gamma_{ik} \log(\pi_k N(x_i | \mu_k, \Sigma_k)) \quad \text{s.t.} \quad \sum_k \pi_k = 1$$

$$L(\mu, \Sigma, \pi, \lambda) = \sum_{i \in [n]} \sum_{k \in [K]} \gamma_{ik} \log(\pi_k N(x_i | \mu_k, \Sigma_k)) + \lambda (1 - \sum_k \pi_k)$$

$$\frac{\partial L}{\partial \pi_k} = 0 \Rightarrow \sum_{i \in [n]} \frac{\gamma_{ik}}{\pi_k} - \lambda = 0$$

$$\Rightarrow \pi_k \lambda = \sum_{i \in [n]} \gamma_{ik}$$

$$\Rightarrow \lambda = \lambda \sum_{k \in [K]} \pi_k = \sum_{i \in [n]} \sum_{k \in [K]} \gamma_{ik} = n$$

$$\therefore \pi_k = \frac{\sum_{i \in [n]} \gamma_{ik}}{n}$$

$$N(x_i | \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right)$$

$$\Rightarrow L = \sum_{i \in [n]} \sum_{k \in [K]} \gamma_{ik} \left[\log \pi_k - \frac{d}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_k|) - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right] + \lambda (1 - \sum_k \pi_k)$$

$$\frac{\partial L}{\partial \mu_k} = 0 \Rightarrow \sum_{i \in [n]} \gamma_{ik} \Sigma_k^{-1} (x_i - \mu_k) = 0$$

$$\Sigma_k \times () \Rightarrow \sum_{i \in [n]} \gamma_{ik} x_i = \sum_{i \in [n]} \gamma_{ik} \mu_k$$

$$\therefore \mu_k = \frac{1}{\sum_{i \in [n]} \gamma_{ik}} \sum_{i \in [n]} \gamma_{ik} x_i$$

$$\frac{\partial L}{\partial \Sigma_k} = 0 \Rightarrow \sum_{i \in [n]} \gamma_{ik} \frac{1}{|\Sigma_k|} \frac{\partial |\Sigma_k|}{\partial \Sigma_k} + \sum_{i \in [n]} \gamma_{ik} \underbrace{(x_i - \mu_k)(x_i - \mu_k)^T}_{\text{noti } \frac{\partial (a^T x b)}{\partial x} = ab^T} = 0$$

$$\text{noti } \frac{\partial |X^{-1}|}{\partial X} = -|X^{-1}|(X^{-1})^T$$

$$\Rightarrow \Sigma_k = \frac{\sum_{i \in [n]} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i \in [n]} \gamma_{ik}}$$

Intuitively, it's the empirical estimation of covariance weighted by γ_{ik}

When $\sigma \rightarrow 0$, $\Sigma_k = \sigma^2 I$, $Mof \rightarrow K$ -means

(Variational AutoEncoder)

2024.12.18

* Generative Models : VAE, Diffusion Models

Mof: $\sum_z P(x|z)P(z) = P(x)$ is usually not powerful as Dimension $\uparrow$ and data is no longer clustering.

High dim data: Eg. images

$x \sim P(x)$. In general, sampling $x \sim P(x)$ is difficult.

Only $\left\{ \begin{array}{l} \text{uniform dist. } U[a, b] \\ \text{gaussian dist.} \end{array} \right\}$ are easy to sample from.
else simple one

Idea: First sample z from a simple dist. like Gaussian

then apply a function $f(z; \theta)$ to map z to $\frac{x}{\sigma}$
can be an image

$$P(x) = \int \underbrace{P(x|z; \theta)}_{\downarrow} \underbrace{P(z)}_{\rightarrow N(z|0,1)} dz$$

$N(x|f(z; \theta), \sigma^2 I)$ model noise in real data

根号模型: $\textcircled{2} \rightarrow \textcircled{\otimes} \quad P(x, z) = P(x|z)P(z)$

Use MLE on training data to optimize θ

$$\max_{\theta} \sum_{i \in [n]} \log P(x_i)$$

Then, how to optimize θ ?

Try SGD (Stochastic Gradient Descent)

$$\text{estimate } P(x) \approx \frac{1}{N} \sum_{z_i \sim P(z)} P(x|z_i; \theta)$$

① Need extremely large number of z sampled from $P(z)$ to accurately estimate $P(x)$. And z is high-dim.

② We have n x_i , n is also very large.

So SGD doesn't work.

EM to solve MLE of $P(x)$ with hidden variables z

E step: Compute $P(z|x_i; \theta^{old}) \rightarrow$ tractable in MLE

M step: $\max_{\theta} \int_z P(z|x_i; \theta^{old}) P(x_i, z; \theta) dz$ $\frac{P(x_i|z; \theta) P(z)}{\int_z P(x_i|z; \theta) P(z)}$

$P(z|x_i; \theta^{old})$ is not tractable in VAE

$$P(z|x_i) = \frac{P(x_i|z)P(z)}{\int_z P(x_i|z)P(z)dz} \text{ cannot be computed}$$

EM, GD both don't work!

Essential issue is $P(z|x; \theta)$ not tractable

(隐变量后验计算)

$\Rightarrow$ Find a tractable variational dist. $q(z|x; \theta)$

to approximate $P(z|x; \theta)$

(last class)

$$\log P(x; \theta) = \underbrace{\int_z q(z|x) \log \frac{P(x, z; \theta)}{q(z|x)}}_{\text{ELBO}} - \underbrace{\int_z q(z|x) \log \frac{P(z|x; \theta)}{q(z|x)}}_{+ KL(q(z|x) \| P(z|x; \theta))}$$

Here)

$$\text{ELBO} = \log P(x; \theta) - KL(q(z|x; \theta') \| P(z|x; \theta))$$

$$= \log P(x; \theta) - \int_z q(z|x; \theta') \log \frac{q(z|x; \theta')}{P(z|x; \theta)} dz$$

$$= \int_z q(z|x; \theta') \log P(x; \theta) dz - \int_z q(z|x; \theta') \log \frac{q(z|x; \theta')}{P(z|x; \theta)} dz$$

$$= \int_z q(z|x; \theta') \cdot \log \frac{P(x; \theta) \cdot P(z|x; \theta)}{q(z|x; \theta')} dz$$

$$= \int_z q(z|x; \theta') \log \frac{P(x, z; \theta)}{q(z|x; \theta')} dz$$

下面, 优化此 ELBO

EM optimizes ELBO by alternatively optimize θ and q

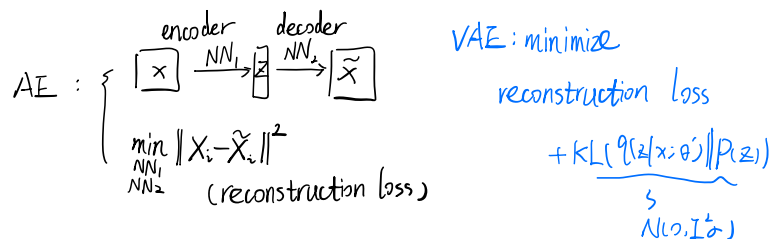
(KL-divergence is 0 $\therefore q$ turns out to be $P(z|x; \theta)$)

VAE optimizes ELBO jointly optimizing θ and $q(\theta')$

(Here KL-divergence may not be 0. so it's just

an approx way)

Before deriving - intuition:



Deriving

$$\text{ELBO} = \int_z q(z|x; \theta') \log P(x|z; \theta) dz$$

$$+ \int_z q(z|x; \theta') \log \frac{P(z)}{q(z|x; \theta')} dz$$

$$= \int_z q(z|x; \theta') \log P(x|z; \theta) dz - KL(q(z|x; \theta') \| P(z))$$

$$= \underbrace{\int_z q(z|x; \theta') \log P(x|z; \theta) dz}_{\text{reconstruction quality}} - \underbrace{KL(q(z|x; \theta') \| P(z))}_{\text{KL-div between encoder dist. \& prior } z}$$

$$\begin{aligned} P(x|z; \theta) &= N(x|f(z; \theta), \sigma^2 I) \\ \log P(x|z; \theta) &= \log \left[\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|f(z; \theta) - x\|^2}{2\sigma^2}\right) \right] \\ &= C - \frac{1}{2\sigma^2} \|f(z; \theta) - x\|^2 \end{aligned}$$

reconstruction loss

Here ELBO for one x_i is derived into:

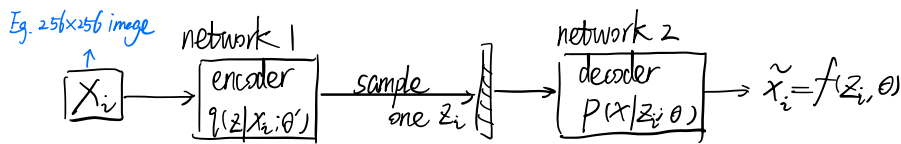
$$- \text{recon. loss} - \beta \cdot KL(q(z|x; \theta') \| P(z))$$

In practice, sample one z_i for each x_i

($\uparrow$ randomness in gradient like SGD $\rightarrow$

alleviate overfitting)

In Conclusion:



Then compute loss of VAE:

$$\frac{1}{n} \sum_{i=1}^n \left[\| \tilde{x}_i - x_i \|^2 + \beta \cdot \text{KL}(q(z|x; \theta') \| p(z)) \right]$$

important

Note: $q(z|x; \theta) = \mathcal{N}(z | \mu(x; \theta), \Sigma(x; \theta))$
 $\hookrightarrow$ usually diagonal

β small: KL contributes less $\rightarrow$ devolves into an AutoEncoder

β big: focus too much on KL

However, BP doesn't work on sampling operation!
 $\downarrow$
 back propagating

Solution: Reparameterization Trick

Instead of sampling from $\mathcal{N}(z | \mu(x), \Sigma(x))$, we sample an $\epsilon \sim \mathcal{N}(0, I)$. $z, \epsilon \in \mathbb{R}^d$

Then let $z = \Sigma(x)^{1/2} \cdot \epsilon + \mu(x) \sim \mathcal{N}(\mu(x), \Sigma(x))$

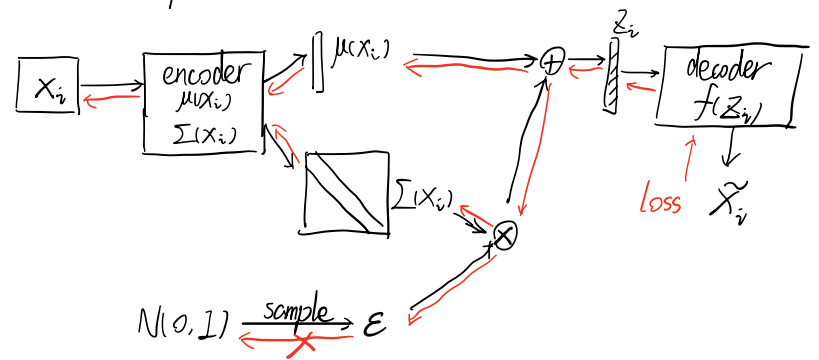
$$\Rightarrow \begin{cases} \mathbb{E}[z] = \mu(x) \\ \text{cov}[z] = \mathbb{E}[(z - \mu(x))(z - \mu(x))^T] \end{cases}$$

$$= \mathbb{E}[\Sigma(x)^{1/2} \epsilon \epsilon^T \Sigma(x)^{1/2}]$$

$$= \Sigma(x)^{1/2} \underbrace{\mathbb{E}[\epsilon \epsilon^T]}_I \Sigma(x)^{1/2} = \Sigma(x)$$

Move sampling process into ϵ

With this reparameterization trick:



There are BPs for Gradient Descent.

Eg. $z \sim \mathcal{U}([a, b])$ repara ~

sample ϵ from $\mathcal{U}([0, 1])$, then $z = \epsilon(b-a) + a$

Derive closed-form of KL

$$\text{KL}(q(z|x; \theta') \| p(z))$$

where $p(z) \sim \mathcal{N}(0, I)$, $q(z|x; \theta') = \mathcal{N}(z | \mu(x, \theta'), \Sigma(x, \theta'))$

(or $\theta' = \theta$: $q(z|x) = \mathcal{N}(z | \mu(x), \Sigma(x))$)

$$\Rightarrow \text{KL}(q(z|x; \theta') \| p(z))$$

$$= \int_z \mathcal{N}(z | \mu(x), \Sigma(x)) \log \frac{\mathcal{N}(z | \mu(x), \Sigma(x))}{\mathcal{N}(z | 0, I)} dz$$

$$= \int_z \mathcal{N}(z | \mu(x), \Sigma(x)) \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma(x)|^{1/2}} \exp(-\frac{1}{2}(z - \mu(x))^T \Sigma(x)^{-1} (z - \mu(x)))}{\frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2} z^T z)}$$

$$\downarrow$$

$$-\frac{1}{2} \log |\Sigma(x)| - \frac{1}{2} (z - \mu(x))^T \Sigma(x)^{-1} (z - \mu(x)) + \frac{1}{2} z^T z$$

$$\Rightarrow \text{KL} = \mathbb{E}_{z \sim \mathcal{N}(\mu(x), \Sigma(x))} \left[-\frac{1}{2} \log |\Sigma(x)| - \frac{1}{2} (z - \mu(x))^T \Sigma(x)^{-1} (z - \mu(x)) + \frac{1}{2} z^T z \right]$$

$$\textcircled{1} \mathbb{E}_z [(z - \mu)^T \Sigma^{-1} (z - \mu)]$$

noti $(z - \mu)^T \Sigma^{-1} (z - \mu)$ is a scalar

trace trick

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$$

$$\Rightarrow \mathbb{E}_z [\text{tr}((z - \mu)^T \Sigma^{-1} (z - \mu))]$$

$$= \text{tr}(\underbrace{\mathbb{E}_z [(z - \mu)(z - \mu)^T]}_{\Sigma} \Sigma^{-1}) = \text{tr}(I) = d$$

$$\begin{aligned}
\Rightarrow E_z[z^T z] &= E_z[(z - \mu + \mu)^T (z - \mu + \mu)] \\
&= E_z[(z - \mu)^T (z - \mu)] + 2 E_z[(z - \mu)^T] \cdot \mu + E_z[\mu^T \mu] \\
&= \text{tr}(E_z[(z - \mu)(z - \mu)^T]) + \mu^T \mu \\
&\quad (\text{trace trick}) \quad \Sigma \\
&= \text{tr}(\Sigma) + \mu^T \mu
\end{aligned}$$

$$\Rightarrow KL = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} d + \frac{1}{2} \text{tr}(\Sigma) + \frac{1}{2} \mu^T \mu$$

$\sigma(x)^2$ is a d-dim vector

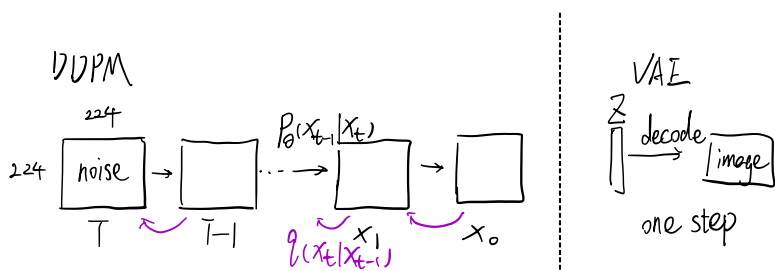
$$\text{if } \Sigma(x) = \sigma(x)^2 I, \quad KL \rightarrow \frac{1}{2} \sum_{j \in [d]} \log(\sigma(x)_j^2) - \frac{d}{2} + \frac{1}{2} \sum_{j \in [d]} \sigma(x)_j^2 + \frac{1}{2} \mu(x)^T \mu(x)$$

We can directly let NN output $\log \sigma(x)^2$
then take exp to get $\sigma(x)^2$

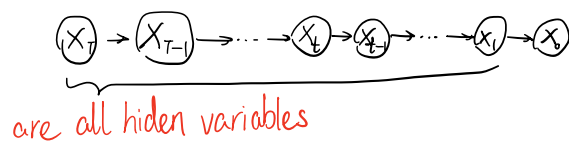
* Diffusion Model: Denoising Diffusion

Probabilistic Model (DDPM)

Principle: use T steps to gradually transform a Gaussian noise to a real image



Graphic Model of DDPM



Learn θ , use $p_\theta(x_{t-1}|x_t)$ as generative model

$$P(x_T) = N(x_T | 0, I)$$

Goal: $\max_\theta \log P_\theta(x_0)$

$$\begin{aligned}
P_\theta(x_0) &= \int P_\theta(x_0, x_1, \dots, x_T) dx_1 \dots dx_T \\
&= \int P(x_T) \prod_{t \in [T]} P_\theta(x_{t-1}|x_t) dx_1 \dots dx_T
\end{aligned}$$

But $x_1 \dots x_T$ are hidden variables $\therefore$ intractable

Use EM?

$$\text{E-Step: } P_\theta(x_{1:T}|x_0) = \frac{P_\theta(x_0, \dots, x_T)}{\underbrace{P_\theta(x_0)}_{\text{intractable}}}$$

$\Rightarrow$ Introduce variational dist. $q(x_1 \dots x_T | x_0)$ to approximate true $P_\theta(x_1 \dots x_T | x_0)$

$$\text{Then ELBO} = \log P_\theta(x_0) - KL(q(x_1 \dots x_T | x_0) || P_\theta(x_1 \dots x_T | x_0))$$

(至此与VAE全同, except that there are T hidden variables)

What q to use

$$\text{Factorize } q(x_{1:T}|x_0) = q(x_T|x_{T-1}) \dots q(x_1|x_0)$$

called **Forward Diffusion Process**

$$q(x_t|x_{t-1}) = N(x_t | \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t)I) \quad ?$$

hyperparameter $\alpha_t \in (0,1)$ is a **noise scheduler**

(Usually, t small $\rightarrow \alpha_t$ is large; $t \uparrow \rightarrow \alpha_t$ gets smaller)

Reparameterization trick!

$$\rightarrow x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} \epsilon_{t-1} \quad \epsilon_{t-1} \sim N(0, I)$$

$$= \sqrt{\alpha_t} (\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}} \epsilon_{t-2}) + \sqrt{1-\alpha_t} \epsilon_{t-1}$$

$$= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \underbrace{\sqrt{\alpha_t - \alpha_t \alpha_{t-1}} \epsilon_{t-2}} + \sqrt{1-\alpha_t} \epsilon_{t-1}$$

Addition of 2 gaussian with mean 0, variance $\sqrt{\alpha_t - \alpha_t \alpha_{t-1}}$ and $\sqrt{1-\alpha_t}$

produces $\sqrt{1-\alpha_t \alpha_{t-1}} \epsilon_{t-2}$

$$\therefore x_t = \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}} \epsilon_{t-2}$$

$$\Rightarrow x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t} \epsilon \quad \alpha_t = \prod_{i \in [t]} \alpha_i$$

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

$\therefore$ 从 x_0 随机 sample $t \rightarrow x_t$, 逆向训练参数

Use of noise scheduler:

$$\textcircled{1} \text{ when } t \rightarrow T, 1-\alpha_t \rightarrow 1 \quad \therefore \alpha_t \rightarrow 0$$

$$q(x_T|x_0) \rightarrow N(x_T | 0, I)$$

x_T becomes independent of x_0

$\textcircled{2}$ no need to compute $q(x_t|x_{t-1})$ step by step

The Reverse Denoising Process (Generation)

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1} | \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$\begin{aligned} \text{ELBO} &= \int_{x_{1:T}} q(x_{1:T}|x_0) \log p_\theta(x_0) dx_{1:T} \\ &\quad - \int_{x_{1:T}} q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)} dx_{1:T} \end{aligned}$$

$$= \int_{x_{1:T}} q(x_{1:T}|x_0) \log \left[\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] dx_{1:T}$$

$$= \int_{x_{1:T}} q(x_{1:T}|x_0) \log \left[\frac{p(x_T) \prod_{t \in [T]} p_\theta(x_{t-1}|x_t)}{\prod_{t \in [T]} q(x_t|x_{t-1})} \right] dx_{1:T}$$

mismatch

$\therefore$ it's similar to KL, but is not.

Using bayes' rule

$$q(x_{t-1}|x_t) = \frac{q(x_t|x_{t-1}) q(x_{t-1})}{q(x_t)} \quad \text{but they are intractable}$$

DDPM's beautiful observation

Though $q(x_{t-1}|x_t)$ intractable,

$q(x_{t-1}|x_t, x_0)$ is tractable

$$\text{Because: } q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)}{q(x_t|x_0)} \quad \text{not } q(x_t|x_{t-1})$$

$$\text{not } x_t \perp x_0 | x_{t-1}$$

Now, we have $q(x_t|x_{t-1}) = N(x_t | \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t)I)$

$$\begin{cases} q(x_t|x_0) = N(x_t | \sqrt{\alpha_t} x_0, (1-\alpha_t)I) \\ t=T: \alpha_t \rightarrow 0, q(x_T|x_0) \approx N(x_T | 0, I) \end{cases}$$

Solve for $q(x_{t-1}|x_t, x_0)$:

$$= \frac{\frac{1}{(2\pi)^{d/2} (1-\alpha_t)^{d/2}} \exp\left(-\frac{\|x_t - \sqrt{\alpha_t} x_{t-1}\|^2}{2(1-\alpha_t)}\right)}{\frac{1}{(2\pi)^{d/2} (1-\alpha_t)^{d/2}} \exp\left(-\frac{\|x_t - \sqrt{\alpha_t} x_0\|^2}{2(1-\alpha_t)}\right)}$$

$$= \frac{1}{(2\pi)^{d/2} \left[\frac{(1-\alpha_t)(1-\alpha_{t-1})}{(1-\alpha_t)} \right]^{d/2}} \exp \left\{ -\frac{1}{2} \left[\left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\alpha_{t-1}} \right) \|x_{t-1}\|^2 - \left(\frac{2\sqrt{\alpha_t} x_t^T}{1-\alpha_t} + \frac{2\sqrt{\alpha_{t-1}} x_0^T}{1-\alpha_{t-1}} \right) x_{t-1} + C(x_0, x_t) \right] \right\}$$

$$= \dots = \frac{1}{(2\pi)^{d/2} \left[\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \right]^{d/2}} \exp \left[-\frac{\left\| x_{t-1} - \left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} x_0 \right) \right\|^2}{2 \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}} \right]$$

In conclusion.

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1} | \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} x_0, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} I)$$

Using reparam~ trick: $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t} \epsilon_t$

$$\Rightarrow x_0 = \frac{x_t - \sqrt{1-\alpha_t} \epsilon_t}{\sqrt{\alpha_t}}$$

Then the mean becomes

$$\mu = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} x_t - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}} \epsilon_t$$

$$\Rightarrow \mu = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_t)$$

Now we have

$$q(x_{t-1}|x_0, x_t) \sim N(\frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_t), \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} I)$$

$$\text{Back to ELBO} = \mathbb{E}_{x_{1:T} \sim q(x_{1:T}|x_0)} \left[\log p(x_T) + \sum_{t=2}^T \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p(x_T) + \sum_{t=2}^T \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$(\text{因 } q(x_t|x_{t-1}) = \frac{q(x_{t-1}|x_t, x_0) q(x_t|x_0)}{q(x_{t-1}|x_0)})$$

$$\therefore \sum_{t=2}^T \log \left(\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \right) = \log q(x_1|x_0) - \log q(x_T|x_0)$$

$$\Rightarrow \text{ELBO} = \mathbb{E}_{q(x_{1:T}|x_0)} \left[\underbrace{\log \frac{p(x_T)}{q(x_T|x_0)}}_{\text{SS}} + \log p_\theta(x_0|x_1) + \sum_{t=2}^T \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \right]$$

$$\begin{aligned} \Rightarrow & \int q(x_{1:T}|x_0) \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} dx_{1:T} \\ &= \int_{x_t, x_{t-1}} q(x_{t-1}|x_t, x_0) q(x_t|x_0) \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} dx_t dx_{t-1} \\ &= \mathbb{E}_{x_t \sim q(x_t|x_0)} \text{KL}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t)) \end{aligned}$$

Given $N(\mu_1, \Sigma_1) \cdot N(\mu_2, \Sigma_2)$.

$$\text{KL}(\cdot \| \cdot) = -\frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{d}{2} + \frac{1}{2} \text{tr}(\Sigma_2^{-1} \Sigma_1) + \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2)$$

$$x: p_\theta(x_{t-1}|x_t) = N(x_{t-1} | \underbrace{\frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t))}_{\mu_\theta(x_t, t)}, \underbrace{\sigma_t^2 I}_{\text{let unlearned}})$$

In practice $\sigma_t^2 = 1 - \alpha_t$

$$\Rightarrow L_t = \frac{(1-\alpha_t)^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \|\epsilon_t - \epsilon_\theta(x_t, t)\|^2$$

$$\mathbb{E}_{\epsilon_t \sim N(0, I)} \left[\left(\frac{(1-\alpha_t)^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \right) \|\epsilon_t - \epsilon_\theta(x_t, t)\|^2 \right]$$

in practice, omit weights!

$$\text{DDPM loss: } \mathbb{E}_{x_0, \epsilon_t \sim N(0, I), t \sim \text{uniform}[1, T]} \left\| \epsilon_t - \underbrace{\epsilon_\theta(x_t, t)}_{\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t} \epsilon_t} \right\|^2$$

Sampling: $x_T \sim N(0, I)$

for $t=T, \dots, 1$ do

$z=0$, if $t=1$ else $z \sim N(0, I)$

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t)) + \alpha_t \cdot z$$

return x_0

Side Remarks:

1. often use U-Net for denoising network

2. To speed up, use latent diffusion:

use VAE to map x_0 to latent space (隐空间)

then do Diffusion Model on latent space

3. Denoising Diffusion Implicit Models

4. Consistency Models