Forward Process: $q(x_t|x_{t-1})$

Reverse Process: $p(x_{t-1}|x_t)$

Forward Process

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$$

Let $a_t = 1-\beta_t$

$\bar{a}_t = \prod_{s=1}^{t} a_s$

Then, $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$

$= \sqrt{1-\beta_t}\, x_{t-1} + \beta_t \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$

$= \sqrt{a_t}\, x_{t-1} + \sqrt{1-a_t}\, \varepsilon$

$= \sqrt{a_t a_{t-1}}\, x_{t-2} + \sqrt{1-a_t a_{t-1}}\, \varepsilon$

$\vdots$

$= \sqrt{\bar{a}_t}\, x_0 + \sqrt{1-\bar{a}_t}\, \varepsilon \Rightarrow q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{a}_t}\, x_0, (1-\bar{a}_t) I)$

Loss function: $-\log(p_\theta(x_0))$

isn't computable

ELBO: $-\log(p_\theta(x_0)) \le -\log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0) \| p_\theta(x_{1:T}|x_0))$

$$: \log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}\right)$$

$$p_\theta(x_{1:T}|x_0) \overset{\text{Bayes' rule}}{=} \frac{p_\theta(x_0|x_{1:T})\, p_\theta(x_{1:T})}{p_\theta(x_0)}$$

$$= \frac{p_\theta(x_0, x_{1:T})}{p_\theta(x_0)} = \frac{p_\theta(x_{0:T})}{p_\theta(x_0)}$$

Bayes' rule:

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}$$

Then, $$\log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}\right) = \log\left(\frac{q(x_{1:T}|x_0)}{\frac{p_\theta(x_{0:T})}{p_\theta(x_0)}}\right) = \log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} \cdot p_\theta(x_0)\right)$$

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

Then our ELBO $: -\log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0) \| p_\theta(x_{1:T}|x_0))$

$$= \cancel{-\log(p_\theta(x_0))} + \log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right) + \cancel{\log(p_\theta(x_0))}$$

Therefore, $$-\log(p_\theta(x_0)) \le \log\left(\frac{\overset{\text{forward process}}{q(x_{1:T}|x_0)}}{\underset{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{p_\theta(x_{0:T})}}\right)$$

$$= \log\left(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}\right)$$

$$= -\log(p(x_T)) + \log\left(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}\right)$$

Since $\log(a\cdot b) = \log(a) + \log(b)$

$$= -\log(p(x_T)) + \sum_{t=1}^{T} \log\left(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}\right)$$

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

By applying Bayes' rule to $q(x_t|x_{t-1}) \Rightarrow \dfrac{q(x_{t-1}|x_t)\, q(x_t)}{q(x_{t-1})}$

Problem: All of these quantities have high variance since we don't know where we started from

Solution: Conditioning on x_0. enable model to have lower possible candidate.

$$q(x_t|x_{t-1}) \Rightarrow \frac{q(x_{t-1}|x_t)\, q(x_t)}{q(x_{t-1})}$$

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

Then,

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

$$\Rightarrow -\log(P(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)\, q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)\, q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

Why we need this part?

$$q(x_1|x_0) = \frac{q(x_0|x_1)\, q(x_1)}{q(x_0)}$$

$$\Rightarrow \frac{q(x_0|x_1, x_0)\, q(x_1|x_0)}{q(x_0|x_0)}$$: doesn't make sense

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

$$\Downarrow$$

$$\log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right)$$

Why? e.g. let $T=4$, then

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

$$\Rightarrow \log \prod_{t=2}^{4} \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

$$\Rightarrow \log\left(\frac{q(x_4|x_0)\, \cancel{q(x_3|x_0)}\, \cancel{q(x_2|x_0)}}{\cancel{q(x_3|x_0)}\, \cancel{q(x_2|x_0)}\, q(x_1|x_0)}\right)$$

$$\Rightarrow \log\left(\frac{q(x_4|x_0)}{q(x_1|x_0)}\right)$$

Therefore,

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

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

$$= \log(q(x_T|x_0)) - \cancel{\log(q(x_1|x_0))} + \cancel{\log(q(x_1|x_0))} - \log(p_\theta(x_0|x_1))$$

$$= \log(q(x_T|x_0)) - \log(p_\theta(x_0|x_1))$$

Then, our loss function :

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

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

$$= D_{KL}(q(x_T|x_0)\,\|\,p(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

this part can be ignored completely since q (forward process) has no learnable parameters and $p(x_T)$ is just a random normal noise sample

Therefore, our loss function:

$$\sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

Since $p_\theta(x_{t-1}|x_t)$ is a reverse process, we can write

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

$$= \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \beta_t I)$$

variance is fixed

$$q(x_{t-1}|x_t,x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t,x_0), \tilde{\beta}_t I)$$

And

$$\tilde{\mu}_t(x_t,x_0) = \frac{\sqrt{1-a_t}\,(1-\bar{a}_{t-1})}{1-\bar{a}_t}x_t + \frac{\sqrt{\bar{a}_{t-1}}\,\beta_t}{1-\bar{a}_t}x_0, \qquad \tilde{\beta}_t = \frac{1-\bar{a}_{t-1}}{1-\bar{a}_t}\cdot\beta_t$$

☆ from forward process, we know $x_t = \sqrt{\bar{a}_t}\,x_0 + \sqrt{1-\bar{a}_t}\,\varepsilon$

Then,

$$x_t - \sqrt{1-\bar{a}_t}\,\varepsilon = \sqrt{\bar{a}_t}\,x_0$$

$$x_0 = \frac{1}{\sqrt{\bar{a}_t}}\cdot(x_t - \sqrt{1-\bar{a}_t}\,\varepsilon)$$

Therefore,

$$\tilde{\mu}_t = \frac{\sqrt{1-a_t}\,(1-\bar{a}_{t-1})}{1-\bar{a}_t}x_t + \frac{\sqrt{\bar{a}_{t-1}}\,\beta_t}{1-\bar{a}_t}\cdot\frac{1}{\sqrt{\bar{a}_t}}(x_t - \sqrt{1-\bar{a}_t}\,\varepsilon)$$

$$\vdots$$

$$= \frac{1}{\sqrt{a_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{a}_t}}\varepsilon\right)$$

Then, our Loss function goes to

$$L_t = \frac{1}{2\sigma_t^2} \| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, x_0) \|^2$$

$$= \frac{1}{2\sigma_t^2} \left\| \frac{1}{\sqrt{a_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{a}_t}} \varepsilon \right) - \frac{1}{\sqrt{a_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{a}_t}} \varepsilon_\theta(x_t, t) \right) \right\|^2$$

$$\Rightarrow \underbrace{\frac{\beta_t^2}{2\sigma_t^2 a_t (1-\hat{a}_t)}}_{\text{ignore}} \| \varepsilon - \varepsilon_\theta(x_t, t) \|^2 \Rightarrow \| \varepsilon - \varepsilon_\theta(x_t, t) \|^2$$

Reverse Process: Sampling

$$: \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Since $\mu_\theta(x_t, t) = \frac{1}{\sqrt{a_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{a}_t}} \varepsilon_\theta(x_t, t) \right)$

and $\Sigma_\theta(x_t, t)$ can be replaced by β_t

$$\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) = \mathcal{N}\left(x_{t-1}; \frac{1}{\sqrt{a_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{a}_t}} \varepsilon_\theta(x_t, t) \right), \beta_t \right)$$

Applying reparameterization trick

$$\Rightarrow x_{t-1} = \frac{1}{\sqrt{a_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{a}_t}} \varepsilon_\theta(x_t, t) \right) + \sqrt{\beta_t}\, \varepsilon$$