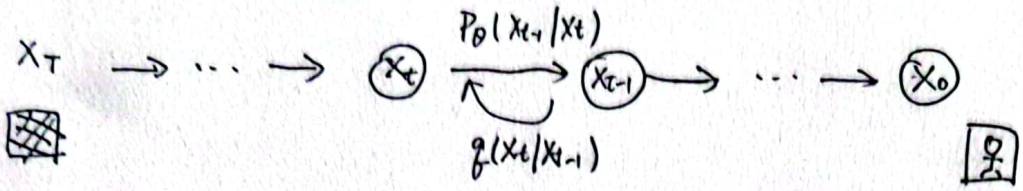

a) Diffusion Process

For input data: $X_0 \sim q(X_0)$

$$q(X_t|X_{t-1}) = N(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I)$$

$\{\beta_t\}_{t=1}^T$ is variance & $\beta_i \in [0,1)$ & $\beta_1 < \beta_2 < \cdots < \beta_T$ (variance schedule)

if the T is huge enough, then $X_T \sim N(0,1)$. For every step:

$$q(X_{1:T}|X_0) := \prod_{t=1}^{T} q(X_t|X_{t-1}) \quad \text{(Markov chain)}$$

And the process is fixed, we need to define a variance schedule. (eg. DDPM - linear)

Def: $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Through reparameterization trick

$$X_t = \sqrt{1-\beta_t}\,X_{t-1} + \sqrt{\beta_t}\,\varepsilon_{t-1} \quad \text{, where } \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots \sim N(0,I)$$

$$= \sqrt{\alpha_t}\,X_{t-1} + \sqrt{1-\alpha_t}\,\varepsilon_{t-1}$$

$$= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}X_{t-2} + \sqrt{1-\alpha_{t-1}}\,\varepsilon_{t-2}\right) + \sqrt{1-\alpha_t}\,\varepsilon_{t-1}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\,X_{t-2} + \sqrt{\left(\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\right)^2 + \left(\sqrt{1-\alpha_t}\right)^2}\,\bar{\varepsilon}_{t-2}$$

, where $\bar{\varepsilon}_{t-2}$ merges two Gaussians (*)

$$= \cdots$$

$$= \sqrt{\bar{\alpha}_t}\,X_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon$$

(*) $\varepsilon_1 \sim N(0,\sigma_1^2 I)$, $\varepsilon_2 \sim N(0,\sigma_2^2 I)$

$\Downarrow \oplus$

$\varepsilon \sim N(0,(\sigma_1^2+\sigma_2^2)I)$.

through cumi-reparameterization:

$$\cdot\ q(X_t|X_0) = N(X_t; \sqrt{\bar{\alpha}_t}\,X_0, (1-\bar{\alpha}_t)I)$$

$\sqrt{\bar{\alpha}_t}$: signal_~~noise~~ rate

$\sqrt{1-\bar{\alpha}_t}$: noise_rate

② b) reverse denoising process (generative)

define the reverse denoising process as a Markov chain. by Networks.

To calculate $q(x_{t-1}|x_t)$, we could use a paramerizaed Gaussion distribution

$$P_\theta(x_{0:T}) = p(x_T)\prod_{i=1}^{T} P_\theta(x_{t-1}|x_T)$$

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

where $p(x_T) = \mathcal{N}(x_T; 0, I)$

the $P_\theta(x_{t-1}|x_t)$ is paramerizaed Gaussian distribution, their mean & variance are got by trainning networks $\mu_\theta(x_t,t)$, $\Sigma_\theta(x_t,t)$. In fact, the diffusion model aims to get these trained networks cause they are part of the final diffusion model.

Though the $q(x_t|x_{t-1})$ can't be calculate, the perior distribution $q(x_{t-1}|x_t, x_0)$ could be done. Theres

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

1. Baysian condition equation:

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

2. Cause diffusion process Markov chain, we know that

$$q(x_t|x_{t-1}, x_0) = q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

and,

$$q(x_{t-1}|x_0) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1-\bar{\alpha}_{t-1})I),$$

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

so,

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

$$\propto \exp\left(-\frac{1}{2}\left(\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t}\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left(\frac{x_t^2 - 2\sqrt{\alpha_t}x_t x_{t-1} + \alpha_t x_{t-1}^2}{\beta_t} + \frac{x_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}x_{t-1}x_0 + \bar{\alpha}_{t-1}x_0^2}{1-\bar{\alpha}_{t-1}} - \frac{x_t^2 - 2\sqrt{\bar{\alpha}_t}x_t x_0 + \bar{\alpha}_t x_0^2}{1-\bar{\alpha}_t}\right)\right)$$

$$= \exp\left[-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)x_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t}x_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0\right)x_{t-1} + C(x_t, x_0)\right)\right]$$

There $C(X_t, X_0)$ has no relation with X_{t-1}. $q(X_{t-1}|X_t, X_0)$

According to Gaussian distribution definitions, we could get perior distribution mean & variance.

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

$$\tilde{\mu}_t(X_t, X_0) = (\frac{\sqrt{\alpha_t}}{\beta_t}X_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}X_0)/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})$$

$$= (\frac{\sqrt{\alpha_t}}{\beta_t}X_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}X_0)\cdot\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}X_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}X_0$$

we could see that $\tilde{\beta}_t$ is constant, and the $\tilde{\mu}_t$ is function rely on X_t & X_0.

c). optimization:

if we consider the X_t as a latent variable, and the diffusion model could be a hierarchical VAEs. so, we could get ELBO (variational lower bound) through variational inference. there

$$\log P_\theta(X_0) = \log \int P_\theta(X_{0:T})\, dX_{1:T}$$

$$= \log \int \frac{P_\theta(X_{0:T})\, q(X_{1:T}|X_0)}{q(X_{1:T}|X_0)} dX_{1:T} \geq E_{q(X_{1:T}|X_0)}[\log \frac{P_\theta(X_{0:T})}{q(X_{1:T}|X_0)}]$$

$\downarrow$

$$-\log P_\theta(X_0) \leq -\log P_\theta(X_0) + D_{KL}(q(X_{1:T}|X_0)\,||\,P_\theta(X_{1:T}|X_0))$$

$$= -\log P_\theta(X_0) + E_{X_{1:T}\sim q(X_{1:T}|X_0)}[\log \frac{q_\theta(X_{1:T}|X_0)}{P_\theta(X_{0:T})/P_\theta(X_0)}]$$

$$= -\log P_\theta(X_0) + E_q[\log \frac{q(X_{1:T}|X_0)}{P_\theta(X_{0:T})} + \log P_\theta(X_0)]$$

$$= E_q[\log \frac{q(X_{1:T}|X_0)}{P_\theta(X_{0:T}|X_0)}]$$

For networks training, we put negative infront of VLB:

$$L = -L_{VLB} = E_{q(X_{1:T}|X_0)}[-\log \frac{P_\theta(X_{0:T})}{q(X_{1:T}|X_0)}]$$

$$= E_q[\log \frac{q(X_{1:T}|X_0)}{P_\theta(X_{0:T})}]$$

④

and then:

$$L = E_{q(X_{1:T}|X_0)}\left[\log \frac{q(X_{1:T}|X_0)}{P_\theta(X_{0:T})}\right] = E_q\left[\log \frac{\prod_{t=1}^{T} q(X_t|X_{t-1})}{P_\theta(X_T)\prod_{t=1}^{T} P_\theta(X_{t-1}|X_t)}\right]$$

$$= E_q\left[-\log P_\theta(X_T) + \sum_{t=1}^{T} \log \frac{q(X_t|X_{t-1})}{P_\theta(X_{t-1}|X_t)}\right]$$

$$= E_q\left[-\log P_\theta(X_T) + \sum_{t=2}^{T} \log \frac{q(X_t|X_{t-1})}{P_\theta(X_{t-1}|X_t)} + \log \frac{q(X_1|X_0)}{P_\theta(X_0|X_1)}\right]$$

$$= E_q\left[-\log P_\theta(X_T) + \sum_{t=2}^{T} \log \frac{q(X_t|X_{t-1}, X_0)}{P_\theta(X_{t-1}|X_t)} + \log \frac{q(X_1|X_0)}{P_\theta(X_0|X_1)}\right] \quad \text{(add prior } X_0\text{)}$$

$$= E_q\left[-\log P_\theta(X_T) + \sum_{t=2}^{T} \log \left[\frac{q(X_{t-1}|X_t, X_0)}{P_\theta(X_{t-1}|X_t)} \cdot \frac{q(X_t|X_0)}{q(X_{t-1}|X_0)}\right] + \log \frac{q(X_1|X_0)}{P_\theta(X_0|X_1)}\right]$$

(Baysian method)

$$= E_q\left[-\log P_\theta(X_T) + \sum_{t=2}^{T} \log \frac{q(X_{t-1}|X_t, X_0)}{P_\theta(X_{t-1}|X_t)} + \log \frac{q(X_T|X_0)}{q(X_1|X_0)} + \log \frac{q(X_1|X_0)}{P_\theta(X_0|X_1)}\right]$$

$$= E_q\left[\log \frac{q(X_1|X_0)}{P_\theta(X_T)} + \sum_{t=2}^{T} \log \frac{q(X_{t-1}|X_t, X_0)}{P_\theta(X_{t-1}|X_t)} - \log P_\theta(X_0|X_1)\right]$$

$$= E_{q(X_T|X_0)}\left[\log \frac{q(X_1|X_0)}{P_\theta(X_T)}\right] + \sum_{t=2}^{T} E_{q(X_t, X_{t-1}|X_0)}\left[\log \frac{q(X_{t-1}|X_t, X_0)}{P_\theta(X_{t-1}|X_t)}\right] - E_{q(X_1|X_0)}[\log P_\theta(X_0|X_1)]$$

$$= E_{q(X_T|X_0)}\left[\log \frac{q(X_1|X_0)}{P_\theta(X_T)}\right] + \sum_{t=2}^{T} E_{q(X_t|X_0)}\left[q(X_{t-1}|X_t, X_0) \log \frac{q(X_{t-1}|X_t, X_0)}{P_\theta(X_{t-1}|X_t)}\right] - E_{q(X_1|X_0)}[\log P_\theta(X_0|X_1)]$$

$$= \underbrace{D_{KL}(q(X_T|X_0)\,||\,P_\theta(X_T))}_{L_T} + \underbrace{\sum_{t=2}^{T} E_{q(X_t|X_0)}[D_{KL}(q(X_{t-1}|X_t, X_0)\,||\,P_\theta(X_{t-1}|X_t))}_{L_{t-1}} \underbrace{- E_{q(X_1|X_0)}\log P_\theta(X_0|X_1)}_{L_0}$$

L_0: ~~the~~ rebuild original data:

$$P_\theta(X_0|X_1) = \prod_{i=1}^{D} \int_{\delta_-(X_0^i)}^{\delta_+(X_0^i)} \mathcal{N}(X_0; \mu_\theta^i(X_1, 1), \Sigma_\theta^i(X_1, 1))\, dx$$

$$\delta_+(x) = \begin{cases} \infty & x = 1 \\ x + \frac{1}{255} & x < 1 \end{cases}, \quad \delta_-(x) = \begin{cases} -\infty & x = -1 \\ x - \frac{1}{255} & x > -1 \end{cases}$$

⑤

L_T: to calculate the KL converge between noise distribution and prior distribution.

$p(X_T) = \mathcal{N}(0, I)$, $q(X_T|X_0) = \mathcal{N}(0,1)$, the $L_T \to 0$ with non trained-parameterization

L_{t-1}: to calculate the KL between $p_\theta(X_{t-1}|X_t)$ and $q(X_{t-1}|X_t, X_0)$. The $q(X_{t-1}|X_t, X_0)$ is a Gaussian distribution, and that's why we def $p_\theta(X_{t-1}|X_t)$ as a Gaussian distribution by networks parameterization.

For both L_0 & L_{t-1}, we want to get trained-well networks $\mu_\theta(X_t, t)$ & $\Sigma_\theta(X_t, t)$ (to L_0, $t=1$). There, DDPM simplified the object:

$$\Sigma_\theta(X_t, t) = \sigma_t^2 I. \quad (\sigma_t^2 = \beta_t \text{ or } \tilde{\beta}_t, \text{ or trainble variance})$$

there we suppose $\sigma_t^2 = \tilde{\beta}_t$. so,

$$q(X_{t-1}|X_t, X_0) = \mathcal{N}(X_{t-1}; \tilde{\mu}(X_t, X_0), \sigma_t^2 I)$$

$$p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \sigma_t^2 I)$$

To calculate the KL ~~beta~~ with two Gaussian distribution:

$$KL(p_1||p_2) = \frac{1}{2}\left[tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_1^{-1}(\mu_2-\mu_1) - n + \log\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right]$$

so, we have

$$D_{KL}[q(X_{t-1}|X_t, X_0)\,||\,p_\theta(X_{t-1}|X_t)] = D_{KL}[\mathcal{N}(X_{t-1}; \tilde{\mu}(X_t, X_0), \sigma_t^2 I)\,||\,\mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \sigma_t^2 I)]$$

$$= \frac{1}{2}\left[n + \frac{1}{\sigma_t^2}\left[||\mu_\theta(X_t, t) - \tilde{\mu}(X_0, X_0)||^2 - n + \log 1\right]\right]$$

$$= \frac{1}{2\sigma_t^2}\,||\tilde{\mu}_t(X_t, t) - \mu_\theta(X_t, t)||^2.$$

and, L_{t-1}:

$$L_{t-1} = E_{q(X_t|X_0)}\left[\frac{1}{2\sigma_t^2}\,||\tilde{\mu}(X_t, t) - \mu_\theta(X_t, t)||^2\right]$$

According to the equation, we want ~~to get~~ the difference between $\tilde{\mu}(X_t, t)$ and $\mu_\theta(X_t, t)$ become smaller.

⑥ But DDPM found that if we predict ε_θ rather than $\mu_\theta(x_t,t)$ it could be better.

According to diffusion process:

$$x_t(x_0,\varepsilon)=\sqrt{\bar\alpha_t}\,x_0+\sqrt{1-\bar\alpha_t}\,\varepsilon \quad \text{, where } \varepsilon\sim N(0,1)$$

put it into L_{t-1}:

$$L_{t-1}=E_{x_0}\left(E_{q(x_t|x_0)}\left[\frac{1}{2\sigma_t^2}\|\tilde\mu_t(x_t,x_0)-\mu_\theta(x_t,t)\|^2\right]\right)$$

$$=E_{x_0,\varepsilon\sim N(0,I)}\left[\frac{1}{2\sigma_t^2}\|\tilde\mu(x_t(x_0,\varepsilon),\frac{1}{\sqrt{\bar\alpha_t}}(x_t(x_0,\varepsilon)-\sqrt{1-\bar\alpha_t}\,\varepsilon))-\mu_\theta(x_t(x_0,\varepsilon),t)\|^2\right]$$

$$=E_{x_0,\varepsilon\sim N(0,I)}\left[\frac{1}{2\sigma_t^2}\|\left(\frac{\sqrt{\bar\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t(x_0,\varepsilon)+\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}\cdot\frac{1}{\sqrt{\bar\alpha_t}}(x_t(x_0,\varepsilon)-\sqrt{1-\bar\alpha_t}\,\varepsilon)\right)-\mu_\theta(x_t(x_0,\varepsilon)t)\|^2\right]$$

$$=E_{x_0,\varepsilon\sim N(0,I)}\left[\frac{1}{2\sigma_t^2}\|\frac{1}{\sqrt{\alpha_t}}\left(x_t(x_0,\varepsilon)-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon\right)-\mu_\theta(x_t(x_0,\varepsilon),t)\|^2\right].$$

then, we reparamerizaded $\mu_\theta(x_t(x_0,\varepsilon),t)$:

$$\mu_\theta(x_t(x_0,\varepsilon),t)=\frac{1}{\sqrt{\alpha_t}}\left[x_t(x_0,\varepsilon)-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon_\theta(x_t(x_0,\varepsilon),t)\right]$$

there ε_θ is a trained function by network. That means we ~~use ε_θ~~ aims to predict ε rather than $\mu_\theta(x_t,t)$. Then,

$$L_{t-1}=E_{x_0,\varepsilon\sim N(0,I)}\left[\frac{1}{2\sigma_t^2}\|\frac{1}{\sqrt{\alpha_t}}(x_t(x_0,\varepsilon)-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon)-\frac{1}{\sqrt{\alpha_t}}(x_t(x_0,\varepsilon)-\frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon_\theta)\|^2\right]$$

$$=E_{x_0,\varepsilon\sim N(0,I)}\left[\frac{\beta_t^2}{2\sigma_t^2\cdot\alpha_t(1-\bar\alpha_t)}\|\varepsilon-\varepsilon_\theta(x_t(x_0,\varepsilon),t)\|^2\right]$$

$$=E_{x_0,\varepsilon\sim N(0,I)}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar\alpha_t)}\|\varepsilon-\varepsilon_\theta(\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\,\varepsilon,t)\|^2\right]$$

DDPM simplified training objective function:

$$L_{t-1}^{simple}=E_{x_0,\varepsilon\sim N(0,I)}\left[\|\varepsilon-\varepsilon_\theta(\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\,\varepsilon,t)\|^2\right]$$