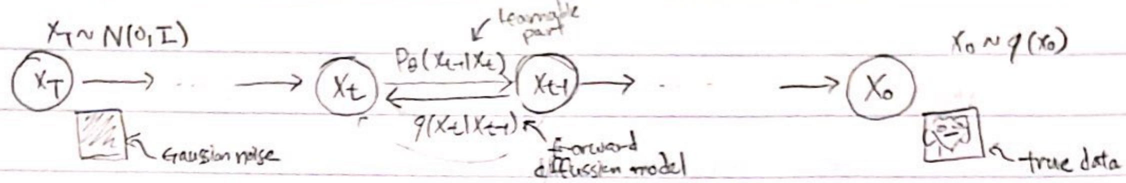
Denoising Diffusion Probabilistic Model (DDPM) for Idiots



이건 아닌듯. → $\cancel{q(x_T) = N(x_T; 0, I)}$

$p(x_T) = N(x_T; 0, I)$ ✓

$\xrightarrow{\quad}$ $q(x_{t-1}|x_t)$

* reverse model은 intractable 하다

DDPM의 objective — DDPM의 목적은 $x_T \sim N(0,I)$ 를 샘플링해서 T번의 forward process $P_\theta(x_{t-1}|x_t)$ 를 거친 결과가 true data distribution $q(x_0)$ 와 비슷하게 만드는 것이다.

구체적으로는 $\underbrace{P_\theta(x_{t-1}|x_t)}_{\text{backward model}}$ 를 학습을 하고, $P_\theta(x_T) = \int \underline{P_\theta(x_0, x_1, x_2, \cdots, x_T)}\, dx_1 dx_2 \cdots dx_T$ 로 model을 학습을 한다. 물론, $P_\theta(x_0, x_1, \cdots, x_T) = P(x_T) \prod_{t=1}^{T} P_\theta(x_{t-1}|x_t)$ 와 같이 정의된다.

Learned backward model — 개별적으로 $P_\theta(x_{t-1}|x_t) \triangleq N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$,

forward diffussion process — $q(x_1, x_2, \cdots, x_T | x_0) \triangleq \prod_{t=1}^{T} q(x_t|x_{t-1})$, $\quad \underline{q(x_t|x_{t-1}) \triangleq N(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)}$

└ 이 x_0 conditioning 이 중요하다.

↳ 우리는 이 forward diffussion process 만 정의를 한다.

Forward diffussion process $q(x_t|x_{t-1})$ 는 정의가 되었지만, 이것으로 induce 되는 backward process 인 $q(x_{t-1}|x_t)$ 는 구할 수가 없다. 하지만 x_0에 condition 된 $q(x_{t-1}|x_t, x_0)$는 구할 수 있다.

Variational Bound — 학습은 $P_\theta(x_0)$의 NLL의 variational bound를 사용한다.

$$\mathbb{E}_{x_0 \sim q(x_0)}\left[-\log P_\theta(x_0)\right] = \mathbb{E}_{x_0 \sim q(x_0)}\left[-\log \int P_\theta(x_0, x_1, \cdots, x_T)\, dx_1 dx_2 \cdots dx_T\right]$$

$$= \mathbb{E}_{x_0 \sim q(x_0)}\left[-\log \int P_\theta(x_0, x_1, \cdots, x_T) \frac{q(x_1, x_2, \cdots, x_T|x_0)}{q(x_1, x_2, \cdots, x_T|x_0)}\, dx_1 dx_2 \cdots dx_T\right]$$

$$= \mathbb{E}_{x_0 \sim q(x_0)}\left[-\log \mathbb{E}_{(x_1, x_2, \cdots, x_T) \sim q(x_1, x_2, \cdots, x_T|x_0)}\left[\frac{P_\theta(x_0, x_1, \cdots, x_T)}{q(x_1, x_2, \cdots, x_T|x_0)}\right]\right]$$

$$\le \mathbb{E}_{x_0 \sim q(x_0)} \mathbb{E}_{(x_1, x_2, \cdots, x_T) \sim q(x_1, x_2, \cdots, x_T|x_0)}\left[-\log \frac{P_\theta(x_0, x_1, \cdots, x_T)}{q(x_1, x_2, \cdots, x_T|x_0)}\right]$$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log \frac{P_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log \frac{P(x_T) P_\theta(x_{T-1}|x_T) P_\theta(x_{T-2}|x_{T-1}) \cdots P_\theta(x_1|x_2) P_\theta(x_0|x_1)}{q(x_1|x_0) q(x_2|x_1) \cdots q(x_T|x_{T-1})}\right]$$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log P(x_T) - \sum_{t=1}^{T} \log \frac{P_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] \quad \text{—— (18) or (23)}$$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log P(x_T) - \sum_{t=2}^{T} \log \frac{P_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} - \log \frac{P_\theta(x_0|x_1)}{q(x_1|x_0)}\right]$$

$$\underset{(t=2,3,\cdots T)}{q(x_t|x_{t-1})} = q(x_t|x_{t-1}, x_0) \quad \because (x_t \perp x_0 \text{ given } x_{t-1})$$

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

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log P(x_T) - \sum_{t=2}^{T} \log \frac{P_\theta(x_{t-1}|x_t)\, q(x_{t-1}|x_0)}{q(x_{t-1}|x_t, x_0)\, q(x_t|x_0)} - \log \frac{P_\theta(x_0|x_1)}{q(x_1|x_0)}\right]$$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log P(x_T) - \log \prod_{t=2}^{T} \frac{P_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \cdot \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} - \log \frac{P_\theta(x_0|x_1)}{q(x_1|x_0)}\right]$$

$$* \prod_{t=2}^{T} \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} = \frac{q(x_1|x_0)}{q(x_T|x_0)}$$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log P(x_T) - \sum_{t=2}^{T} \log \frac{P_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log \frac{\cancel{q(x_1|x_0)}}{q(x_T|x_0)} - \log \frac{P_\theta(x_0|x_1)}{\cancel{q(x_1|x_0)}}\right]$$

$KL(Q \| P) = -\int Q \log \frac{P}{Q}$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\left[-\log \frac{P(x_T)}{q(x_T|x_0)} - \sum_{t=2}^{T} \log \frac{P_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log P_\theta(x_0|x_1)\right]$$

$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\Big[\underbrace{D_{KL}(q(x_T|x_0) \| P(x_T))}_{:L_T} + \underbrace{\sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) \| P_\theta(x_{t-1}|x_t))}_{:L_{t-1}} \underbrace{- \log P_\theta(x_0|x_1)}_{L_0}\Big] \quad \text{— (5)}$$

(5) : $\mathbb{E}_{x_{0:T}\sim q(x_{0:T})}\Big[\underbrace{D_{KL}(q(x_T|x_0)\,\|\,p(x_T))}_{L_T} + \sum_{t=2}^{T}\underbrace{D_{KL}(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \underbrace{-\log p_\theta(x_0|x_1)}_{L_0}\Big]$

i) L_T (5)에서 L_T는 learnable part 가 없다. $p(x_T)=\mathcal{N}(x_T;0,I)$ 으로 fix 되어있고, 뒤에서 유도하겠지만,
$q(x_t|x_0)=\mathcal{N}(x_t;\sqrt{\bar\alpha_t}\,x_0,(1-\bar\alpha_t)I)$, $\bar\alpha_t=\prod_{s=1}^{t}\alpha_s$, $\alpha_s=1-\beta_s$
와 같이 유도될 수 있다.

ii) L_{t-1} (5)에서 $D_{KL}(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t))$는 learnable $p_\theta(x_{t-1}|x_t)$와 $q(x_{t-1}|x_t,x_0)$의 backward process posterior 의 KLD를 줄이는 방향으로 θ를 학습한다.

iii) L_0 $p_\theta(x_0|x_1)$ 에서 x_0는 true data 이고, x_1은 backward process로 얻어지는 Gaussian random vector이다.
그래서 이것을 논문에서는 reverse process decoder 라고 부른다.

Forward Diffusion Process

$q(x_t|x_{t-1})$ 의 forward diffusion process 만 정의가 되어있다.

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

이 Forward diffusion process의 arbitrary time step의 분포를 closed form 으로 구할 수 있다.

먼저 $x_t|x_{t-1}\sim q(x_t|x_{t-1})$ 을 $z\sim\mathcal{N}(0,I)$를 써서 표현할 수 있다.

$z_1\sim\mathcal{N}(0,\sigma_1^2)$, $z_2\sim\mathcal{N}(0,\sigma_2^2)$
$\downarrow$
$z_1+z_2\sim\mathcal{N}(0,\sigma_1^2+\sigma_2^2)$

$$\begin{aligned}
x_t &= \sqrt{\alpha_t}\,x_{t-1} + \sqrt{1-\alpha_t}\,z_{t-1} \qquad , \ \alpha_t=1-\beta_t \ , \ z_{t-1}\sim\mathcal{N}(0,I)\\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\,x_{t-2}+\sqrt{1-\alpha_{t-1}}\,z_{t-2}\right) + \sqrt{1-\alpha_t}\,z_{t-1}\\
&= \sqrt{\alpha_t\alpha_{t-1}}\,x_{t-2} + \sqrt{\alpha_t-\alpha_t\alpha_{t-1}}\,z_{t-2} + \sqrt{1-\alpha_t}\,z_{t-1}\\
&= \sqrt{\alpha_t\alpha_{t-1}}\,x_{t-2} + \sqrt{\alpha_t-\alpha_t\alpha_{t-1}+1-\alpha_t}\,\bar z_{t-2} \qquad , \ \bar z_{t-2}\sim\mathcal{N}(0,I)\\
&= \sqrt{\alpha_t\alpha_{t-1}}\,x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\,\bar z_{t-2}\\
&= \sqrt{\alpha_t\alpha_{t-1}}\left(\sqrt{\alpha_{t-2}}\,x_{t-3}+\sqrt{1-\alpha_{t-2}}\,z_{t-3}\right) + \sqrt{1-\alpha_t\alpha_{t-1}}\,\bar z_{t-2}\\
&= \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}\,x_{t-3} + \sqrt{\alpha_t\alpha_{t-1}-\alpha_t\alpha_{t-1}\alpha_{t-2}}\,z_{t-3} + \sqrt{1-\alpha_t\alpha_{t-1}}\,\bar z_{t-2}\\
&= \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}\,x_{t-3} + \sqrt{1-\alpha_t\alpha_{t-1}\alpha_{t-2}}\,\bar z_{t-3} \ , \qquad \bar z_{t-3}\sim\mathcal{N}(0,I)\\
&\ \ \vdots\\
&= \sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_1}\,x_0 + \sqrt{1-\alpha_t\alpha_{t-1}\cdots\alpha_1}\,\bar z_0 \ , \qquad \bar z_0\sim\mathcal{N}(0,I)\\
&= \sqrt{\bar\alpha_t}\,x_0 + \sqrt{1-\bar\alpha_t}\,\bar z_0 \ , \quad \bar\alpha_t=\prod_{s=1}^{t}\alpha_s \ , \ \alpha_s=1-\beta_s
\end{aligned}$$

따라서 $q(x_t|x_0)=\mathcal{N}(x_t;\sqrt{\bar\alpha_t}\,x_0,(1-\bar\alpha_t)I)$ 이다.

이 성질을 이용해서 x_0에 condition 된 backward process posterior $q(x_{t-1}|x_t,x_0)$를
구할 수 있다.

this is Gaussian due to the conjugacy of Gaussians

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

$$* \; \mathcal{N}(x|a,b) = \frac{1}{\sqrt{2\pi b}}\exp\left(-\frac{1}{2b}(x-a)^2\right) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{b}x^2 - \frac{2a}{b}x + \frac{a^2}{b}\right)\right)$$

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

$$= \frac{\mathcal{N}(x_t;\sqrt{1-\beta_t}\,x_{t-1},\beta_t I)\,\mathcal{N}(x_{t-1};\sqrt{\bar\alpha_{t-1}}\,x_0,(1-\bar\alpha_{t-1})I)}{\mathcal{N}(x_t;\sqrt{\bar\alpha_t}\,x_0,(1-\bar\alpha_t)I)}$$

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

$$= \exp\Bigg(-\frac{1}{2}\Bigg(\frac{1}{\beta_t}x_t^2 - \frac{2}{\beta_t}\sqrt{1-\beta_t}\,x_t x_{t-1} + \frac{1-\beta_t}{\beta_t}x_{t-1}^2 + \frac{x_{t-1}^2}{(1-\bar\alpha_{t-1})} - \frac{2\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_{t-1})}x_0 x_{t-1} + \frac{\bar\alpha_{t-1}}{(1-\bar\alpha_{t-1})}x_0^2 - \frac{x_t^2}{(1-\bar\alpha_t)} + \frac{2\sqrt{\bar\alpha_t}\,x_0 x_t}{(1-\bar\alpha_t)} - \frac{\bar\alpha_t x_0^2}{(1-\bar\alpha_t)}\Bigg)\Bigg)$$

$$= \exp\left(-\frac{1}{2}\left(\underbrace{\left(\frac{1-\beta_t}{\beta_t}+\frac{1}{(1-\bar\alpha_{t-1})}\right)}_{=1/b}x_{t-1}^2 - \underbrace{\left(\frac{2}{\beta_t}\sqrt{1-\beta_t}\,x_t + \frac{2\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_{t-1})}x_0\right)}_{=\frac{2a}{b}}x_{t-1} + \sim\right)\right)$$

i) $\dfrac{1}{b} = \dfrac{\alpha_t}{\beta_t} + \dfrac{1}{(1-\bar\alpha_{t-1})} = \dfrac{\alpha_t - \alpha_t\bar\alpha_{t-1} + \beta_t}{\beta_t(1-\bar\alpha_{t-1})} = \dfrac{1-\bar\alpha_t}{\beta_t(1-\bar\alpha_{t-1})}$ $\quad \because \alpha_t+\beta_t=1,\ \bar\alpha_t=\prod_{s=1}^{t}\alpha_s,\ \alpha_t\bar\alpha_{t-1}=\bar\alpha_t$

$$b = \frac{\beta_t(1-\bar\alpha_{t-1})}{1-\bar\alpha_t} \quad \leftarrow \text{분산}$$

ii) $\dfrac{2a}{b} = \left(\dfrac{\cancel{2}}{\beta_t}\sqrt{1-\beta_t}\,x_t + \dfrac{\cancel{2}\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_{t-1})}x_0\right)$

$$\boxed{\frac{(1-\bar\alpha_{t-1})}{(1-\bar\alpha_t)}\underbrace{\sqrt{1-\beta_t}}_{=\sqrt{\alpha_t}}\,x_t + \frac{\beta_t\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_t)}x_0}$$

$$a = \boxed{\frac{\beta_t(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\left(\frac{\sqrt{1-\beta_t}\,x_t}{\beta_t} + \frac{\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_{t-1})}x_0\right)} \quad \leftarrow \text{평균.}$$

$$x_t = \sqrt{\bar\alpha_t}\,x_0 + \sqrt{1-\bar\alpha_t}\,z_t$$

$$x_0 = \frac{1}{\sqrt{\bar\alpha_t}}x_t - \sqrt{\frac{1-\bar\alpha_t}{\bar\alpha_t}}\,z_t$$

$$a = \frac{\beta_t(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\left(\frac{\sqrt{1-\beta_t}}{\beta_t}x_t + \frac{\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_{t-1})}\left(\frac{1}{\sqrt{\bar\alpha_t}}x_t - \sqrt{\frac{1-\bar\alpha_t}{\bar\alpha_t}}\,z_t\right)\right)$$

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

$$= \frac{(1-\bar\alpha_{t-1})\sqrt{\alpha_t}}{1-\bar\alpha_t}x_t + \frac{\beta_t}{(1-\bar\alpha_t)}\sqrt{\frac{1}{\alpha_t}}\,x_t - \frac{\beta_t\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_t)}\sqrt{\frac{1-\bar\alpha_t}{\bar\alpha_t}}\,z_t$$

$$= \frac{\alpha_t - \bar\alpha_t + \beta_t}{\sqrt{\alpha_t}(1-\bar\alpha_t)}x_t - \frac{\beta_t\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_t)}\sqrt{\frac{1-\bar\alpha_t}{\bar\alpha_t}}\,z_t$$

$$= \frac{1}{\sqrt{\alpha_t}}x_t - \frac{\beta_t\sqrt{1-\bar\alpha_t}}{(1-\bar\alpha_t)\sqrt{\alpha_t}}z_t$$

$$= \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1}{\sqrt{\alpha_t}}\beta_t\frac{1}{\sqrt{1-\bar\alpha_t}}z_t = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar\alpha_t}}z_t\right) \quad \leftarrow \text{평균}$$

따라서 $q(x_{t-1}|x_t,x_0) = \mathcal{N}\left(x_{t-1};\dfrac{1}{\sqrt{\alpha_t}}\left(x_t - \dfrac{\beta_t}{\sqrt{1-\bar\alpha_t}}z_t\right), \dfrac{\beta_t(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\right)$

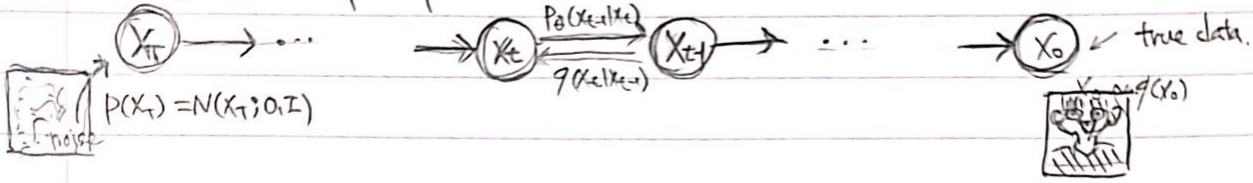
$$= \mathcal{N}\left(x_{t-1};\frac{\beta_t\sqrt{\bar\alpha_{t-1}}}{(1-\bar\alpha_t)}x_0 + \frac{(1-\bar\alpha_{t-1})}{(1-\bar\alpha_t)}\sqrt{\alpha_t}\,x_t, \frac{\beta_t(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\right) \quad \text{— (7)}$$

정리해보자면,

forward process variance

$$\begin{cases} \text{forward diffusion process} & q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I) \\ \text{jump diffusion process} & q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I) \\ \underline{\text{backward process posterior}} & q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_t)} x_0 + \frac{(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\sqrt{\alpha_t}\, x_t, \frac{\beta_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\right) \end{cases}$$

이거 논문에서는 forward process posterior라고 한다



다시 NLL의 variational bound로 돌아와서, 우리가 minimize 하는 loss는 다음과 같다.

$$L = \mathbb{E}_{x_{0:T} \sim q(x_{0:T})}\Big[D_{KL}(q(x_T|x_0) \,\|\, p(x_T)) \qquad : L_T$$
$$+ \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)) \qquad : L_{t-1}$$
$$- \log p_\theta(x_0|x_1) \Big] \qquad : L_0$$

L_T Variational bound의 L_T에는 learnable parameter가 없기때문에 학습에 사용되지는 않는다.

참고로 $\begin{cases} q(x_T|x_0) = \mathcal{N}(x_T; \sqrt{\bar{\alpha}_T}\, x_0, (1-\bar{\alpha}_T) I) \\ p(x_T) = \mathcal{N}(x_T; 0, I) \end{cases}$ 이다.

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

$$\begin{cases} q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_t)} x_0 + \frac{(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\sqrt{\alpha_t}\, x_t, \frac{\beta_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\right) \leftarrow \text{fixed}_{\prime\prime} \\ p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \end{cases}$$

learnable parameters

$\Sigma_\theta(x_t, t)$ 먼저 $\Sigma_\theta(x_t, t)$는 임의의 상수인 $\sigma_t^2 I$로 사용한다. (KLD를 tractable 하게) 이때 σ_t^2은 β_t 나 $\tilde{\beta}_t = \left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\right)$ 모두 사용가능한데, 성능은 비슷했다고 한다. 여기도 tuning의 여지가 있다.

$\mu_\theta(x_t, t)$ 앞서 $\Sigma_\theta(x_t, t)$를 θ에 independent 한 상수, σ_t^2, 로 놓았기 때문에

$$L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t))$$
$$= \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2\right] + C$$

로 표현이 가능하다. $\tilde{\mu}_t$는 $q(x_{t-1}|x_t, x_0)$의 평균이다. $* \; \tilde{\mu}_t(x_t, x_0) = \left(\frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_t)} x_0 + \frac{(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\sqrt{\alpha_t}\, x_t\right)$

$$L_{t-1} - C = \mathbb{E}_{x_t \sim q(x_t|x_0)}\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2\right]$$

$x_t \sim q(x_t|x_0) \rightarrow x_t = \sqrt{\bar{\alpha}_t}\, \underbrace{x_0}_{\text{given}} + \sqrt{1-\bar{\alpha}_t}\, \underbrace{\varepsilon}_{\text{randomness}}$, $\varepsilon \sim N(0, I)$ 와 같이 표현할수 있다.

Plugging $x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} x_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\varepsilon$ into $\tilde{\mu}_t(x_t, x_0)$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_t)} \cdot \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon) + \frac{(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\sqrt{\alpha_t}\, x_t$$
$$= \frac{1}{(1-\bar{\alpha}_t)} \cdot \left(\frac{\beta_t}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon) + (1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}\, x_t\right)$$
$$= \frac{1}{(1-\bar{\alpha}_t)}\left(\frac{\beta_t}{\sqrt{\alpha_t}} + \frac{\alpha_t(1-\bar{\alpha}_{t-1})}{\sqrt{\alpha_t}}\right) x_t - \frac{\beta_t\sqrt{1-\bar{\alpha}_t}}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\varepsilon$$
$$= \frac{\alpha_t + \beta_t - \bar{\alpha}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} x_t - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\varepsilon$$

따라서 x_t 와 ϵ 으로 reparametrize 된 $\tilde{\mu}_t(x_t, x_0)$ 는

$$\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$ 이 된다.

현재 ($L_{t-1}-C$)를 계산할 때, x_t는 given 이다.

$$\rightarrow L_{t-1} - C = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim N(0,I)}\left[\frac{1}{2\sigma_t^2}\left\| \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) - \mu_\theta(x_t, t)\right\|^2\right] \quad (10)$$

(10) 이 말하는 것은 $\mu_\theta(x_t, t)$ 가 $\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$ 을 예측하게 하는 것이다.

따라서 $\mu_\theta(x_t, t)$ 를 다음과 같이 parametrize 하자.

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \quad \text{—(11)}$$

└ 이게 parametrized function

우리의 learned backward model $x_{t-1} \sim p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2)$ 이다.

따라서 다음과 같이 sampling 할 수 있다.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z_t, \qquad z_t \sim N(0, I)$$

그리고 (11)을 (10)에 넣으면,

$$L_{t-1} - C = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim N(0,I)}\left[\frac{1}{2\sigma_t^2}\left\| \frac{1}{\sqrt{\alpha_t}}\left(\cancel{x_t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) - \frac{1}{\sqrt{\alpha_t}}\left(\cancel{x_t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)\right\|^2\right]$$

$$= \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim N(0,I)}\left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)}\left\| \epsilon - \epsilon_\theta(x_t, t)\right\|^2\right]$$

또한 x_0가 주어졌을 때 x_t는 $q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$ 이다.

$\rightarrow \quad x_t|x_0 = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim N(0, I)$

Denoising Objective

$$= \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim N(0,I)}\left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)}\left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \quad \text{— (12)}$$

(12)는 denoising score matching 과 비슷하다.

논문에서는 (11)이 Langevin dynamics 의 variational bound 와 비슷하다고 한다.

Langevin dynamics

어떤 posterior $p(\theta|X) \propto p(\theta)\prod_{i=1} p(x_i|\theta)$ 가 있을 때, sampling 을 다음과 같이 gradient 정보만 가지고도 할 수 있다.

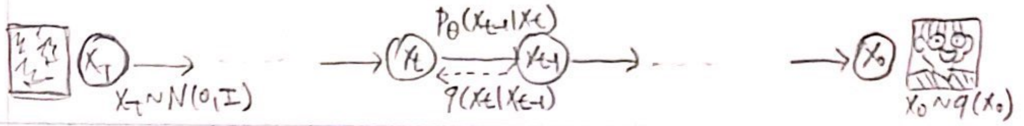
prior

log likelihood

$$\Delta\theta = \frac{\epsilon_t}{2}\left(\underbrace{\nabla p(\theta)}_{\text{prior}} + \frac{N}{n}\sum_{i=1}^{n}\nabla \log p(x_i|\theta)\right)$$

* n은 subset $\{x_i\}_{i=1}^{n}$ 의 크기, N은 전체데이터의 크기.

$\sum_{t=1}^{\infty}\epsilon_t = \infty, \quad \sum_{t=1}^{\infty}\epsilon_t^2 < \infty$ 이어야하고, 대표적인 예로는 $\epsilon_t = \frac{1}{t}$ 가 있다.



L_0 마지막으로 남은 term은 $L_0 = \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} [-\log p_\theta(x_0|x_1)]$ ↓

Reverse process decoder 즉 x_T에서부터 x_1까지 sampling 된 reverse process의 결과로 나온 Gaussian을 따르는 x_1을 실제 데이터의 분포인 $q(x_0)$의 Domain으로 보내줘야 한다. (decoder)

이 모델는 independent discrete decoder로 모델링한다.

$$p_\theta(x_0|x_1) = \mathcal{N}(x_0; \mu_\theta(x_1, t=1), \sigma_1^2 I)$$
$$= \prod_{i=1}^{D} \int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} \mathcal{N}(x; \mu_\theta^i(x_1, 1), \sigma_1^2)\, dx \quad —(13)$$

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

(13)은 실제로 쓰진 않고, 근사를 한다. (Gaussian pdf times the bin width, ignoring σ_1^2)

Simplified objective 실제 학습에 $L_{t-1} - C = \mathbb{E}_{x_0 \sim q(x_0), \varepsilon \sim N(0,I)} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \| \varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t) \|^2 \right]$

에서 $\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)}$는 무시하고,

reconstruct noise

$$\mathbb{E}_{x_0 \sim q(x_0),\, \varepsilon \sim N(0,I),\, t \sim \mathrm{Unif}(\{1, \dots, T\})} \left[\| \varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t) \|^2 \right] \quad —(14)$$

learnable function　　Input data

를 사용한다.

실제 실험에서 $T = 1000$을 사용했고, $\beta_1 = 10^{-4}$, $\beta_T = 0.02$이고, 그 중간은 interpolate 한다. 그리고 reverse process는 Unet 구조를 사용하고, t는 Transformer sinusoidal position embedding을 사용한다.

Progressive coding $\underbrace{\mathbb{E}_{x_0 \sim q(x_0)} [-\log p_\theta(x_0)]}_{\text{MLE objective}} \le \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} \left[-\log p(x_T) - \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]$ —(18), (23)

↳ variational bound

(18)의 variational bound는 다음과 같이 유도될 수 있다. (tractable 하지는 않다.)

$$L = \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} \left[-\log p(x_T) - \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]$$
$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} \left[-\log p(x_T) - \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)\, q(x_{t-1})}{q(x_{t-1}|x_t)\, q(x_t)} \right] \qquad \frac{q(x_0)}{q(x_T)}$$
$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} \left[-\log p(x_T) - \log \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)} \cdot \frac{q(x_{t-1})}{q(x_t)} \right]$$
$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} \left[-\log p(x_T) - \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)} - \log \frac{q(x_0)}{q(x_T)} \right]$$
$$= \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} \left[-\log \frac{p(x_T)}{q(x_T)} - \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)} - \log q(x_0) \right] \qquad H_q(x_0) = -\int q \log q$$

$$= D_{KL}(q(x_T) \| p(x_T)) + \mathbb{E}_{x_{0:T} \sim q(x_{0:T})} \left[\sum_{t=1}^{T} D_{KL}(q(x_{t-1}|x_t) \| p_\theta(x_{t-1}|x_t)) \right] + H_q(x_0)$$

Intractable　$p(x_T) \sim N(0, I)$　　Intractable　　learnable reverse process