max prob $p(x_0)$

marginalize $\int p(x_0, x_{1:T})\, dx_{1:T}$

$x_1, x_2 \ldots x_T$ latent variables

Suppose $x_0 \sim \mathcal{D}$ $\quad$ $\mathcal{D}$ distribution of $\underbrace{3 \times 64 \times 64}_{d = 3 \times 2^{12} = 12{,}288}$ images

$x_T \to x_{T-1} \to \dots \to x_2 \to x_1 \to x_0$

reverse process to learn

denoise

$x_T \sim N(0, I^d)$

$x_0 \sim \mathcal{D}$

given x_0 forward process

add noise per noise schedule

$\beta_1, \beta_2 \cdots \beta_T$

1e-4 ... linear ... .02

posterior distributions, given x_0

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

$x_1 = \sqrt{1-\beta_1}\, x_0 + \sqrt{\beta_1}\, \varepsilon$

$var(x_1) = \beta_1$

$x_T \to N(0,1)$

per noise schedule

not exactly

Not only $q(x_t \mid x_{t-1})$ normal, but

$q(x_t \mid \boxed{x_0})$ normal

So $x_t \sim N(x_t, \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$

and

$\sim N(x_t, \sqrt{\bar{\alpha}_t}\, \boxed{x_0}, (1-\bar{\alpha}_t) I)$

only depends on x_0

$\alpha_t = 1 - \beta_t$

$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

Can apply Bayes' Theorem to

$q(x_t | x_{t-1}, x_0)$ $\quad t > 1$ $\quad$ ($q(x_1 | x_2, x_0)$ and onward)

and it is normal,

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

$$\mathcal{N}\left(x_{t-1} \,\middle|\, \underbrace{\boxed{c_t}}_{\frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}} x_0 + \underbrace{\boxed{\gamma_t}}_{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}} x_t,\ \boxed{\tilde{\beta}_t} I\right)$$

$$\tilde{\beta}_t \rightarrow \left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\right)\beta_t$$

Model reverse process

$$p_\theta(x_{t-1} | x_t) \sim N\left(x_{t-1} \middle| u_\theta(x_t, t), \Sigma_\theta(x_t, t)\right)$$

simplify as $\sim N\left(x_{t-1} \middle| u_\theta(x_t, t), \beta_t I\right)$

assume Markov property

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

likewise for posteriors

$$q(x_{1:T} | x_0) = \prod_{t=1}^{T} q(x_t | x_{t-1})$$

Deriving Loss, reducing Variance part 1

want model to maximize $p_\theta(x_0)$

same as minimize NLL $-\ln p_\theta(x_0)$

marginalize over latent $x_{1:T}$

$$-\ln p_\theta(x_0) = -\ln \int p_\theta(x_{0:T})\, dx_{1:T}$$

bring in posterior

$$= -\ln \int p_\theta(x_{0:T}) \frac{q(x_{1:T} \mid x_0)}{q(x_{1:T} \mid x_0)}\, dx_{1:T}$$

Deriving Loss, reducing variance part 2

$$-\ln p_\theta(x_0) = -\ln \int p_\theta(x_{0:T}) \frac{q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)} dx_{1:T}$$

$$= -\ln \mathop{\mathbb{E}}_{x_{1:T} \sim q(\cdot|x_0)} \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}$$

can sample from $q(\cdot|x_0)$

apply Jensen's inequality

(sim to deriving ELBO)

$$\leq \mathop{\mathbb{E}}_{x_{1:T} \sim q(\cdot|x_0)} -\ln \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}$$

Deriving Loss, reducing Variance part 3

$$\text{loss} = \mathop{\mathbb{E}}_{x_{1:T} \sim q(\cdot|x_0)} -\ln \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}$$

expand using Markov

$$= \mathop{\mathbb{E}}_{x_{1:T} \sim q(\cdot|x_0)} \left(-\ln p_\theta(x_T) - \sum_{t \geq 1} \ln\left(\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right) \right)$$

peel off first term

$$= \mathop{\mathbb{E}}_{x_{1:T} \sim q(\cdot|x_0)} \left(-\ln p_\theta(x_T) - \ln \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} - \sum_{t \geq 2} \ln\left(\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right) \right)$$

Deriving Loss, reducing Variance part 4

$$\underset{x_{1:T}\sim q(\cdot|x_0)}{E} -\ln p_\theta(x_T) - \ln \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} - \sum_{t\geq 2} \ln\left(\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right)$$

apply Bayes

$$= \underset{x_{1:T}\sim q(\cdot|x_0)}{E} -\ln p_\theta(x_T) - \ln \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} - \sum_{t\geq 2} \ln \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)} \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

Deriving Loss, reducing Variance part 5

$$= \mathop{E}_{x_{1:T} \sim q(\cdot|x_0)} -\ln p_\theta(x_T) - \ln \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} - \sum_{t \geq 2} \ln \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)} \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

these terms form telescoping sum

$t=2$

$$+ \ln q(x_1|x_0) - \ln q(x_1|x_0)$$

$$+ \ln q(x_2|x_0)$$

$t=3$ ...

$t=T$

$$- \ln q(x_{T-1}|x_0)$$

$$\boxed{+ \ln q(x_T|x_0)}$$

only left with this

$$= \mathop{E}_{x_{1:T} \sim q(\cdot|x_0)} -\ln \frac{p_\theta(x_T)}{q(x_T|x_0)} - \ln p_\theta(x_0|x_1) - \sum_{t \geq 2} \ln \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)}$$

Deriving Loss, reducing Variance part 6

$$= \underset{x_{1:T} \sim q(\cdot|x_0)}{E} -\ln \frac{p_\theta(x_T)}{q(x_T|x_0)} - \ln p_\theta(x_0|x_1) - \sum_{t \geq 2} \ln \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t)}$$

$$= \underset{x_1}{E} -\ln p_\theta(x_0|x_1) + \sum_{t \geq 2} KL\left[q(x_{t-1}|x_t) \| p_\theta(x_{t-1}|x_t)\right) + KL\, q(x_T$$

$$= \underbrace{\underset{x_1}{E} -\ln p_\theta(x_0|x_1)}_{\substack{L_0 \text{ loss} \\ \text{term, reconstruction}}} + \sum_{t \geq 2} \underbrace{KL\left[q(x_{t-1}|x_t) \| p_\theta(x_{t-1}|x_t)\right]}_{\substack{\text{two normals} \\ \text{analytic formula} \\ \text{weighted MSE on means}}} + \underbrace{KL\left[q(x_T|x_0) \| p(x_T)\right]}_{\substack{\text{constant} \\ \text{can ignore} \\ \text{for training}}}$$

could do simple decoder

Model Noise

loss $= \mathbb{E}_{x_1} - \ln p_\theta(x_0|x_1) + \underbrace{\sum_{t\geq 2} KL\left[q(x_{t-1}|x_t) \| p_\theta(x_{t-1}|x_t)\right]}$

mean of $q(x_{t-1}|x_t)$

$$L_{t-1} = \mathbb{E}_q \frac{1}{2\sigma_t^2} \left\| \tilde{u}_t(x_t, x_0) - u_\theta(x_t, t) \right\|^2$$

latent $x_t = c_1 x_0 + c_2 \varepsilon$

where $\varepsilon \sim N(0, I)$

so $x_0 = \tilde{c}_1 x_t + \tilde{c}_2 \varepsilon$

and $\tilde{u}_t = \tilde{\tilde{c}}_1 x_t + \tilde{\tilde{c}}_2 \varepsilon$

so set $u_\theta(x_t, t) = \tilde{\tilde{c}}_1 x_t + \tilde{\tilde{c}}_2 \varepsilon_\theta(x_t, t)$

Model Noise

$$\text{loss} = \underbrace{\mathop{E}_{x_1} -\ln p_\theta(x_0|x_1)}_{L_0} + \sum_{t\geq 2} \mathop{E}_{x_0,\epsilon} \underbrace{\omega_t \left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar\alpha_t}\, x_0 + \sqrt{1-\bar\alpha_t}\,\epsilon, t\right)\right\|^2}_{L_{t-1}}$$

$\epsilon \sim N(0, I)$

$$\omega_t = \frac{B_t^2}{2\sigma_t^2 \alpha_t (1-\bar\alpha_t)}$$

simple alg 1

ignore L_0

$\omega_t = 1$