DDPM (Reverse Path)

Who is the learned $p_\theta(x_{t-1}|x_t)$

for clarifying this is learnable.

Assuming the flow is very slow ($\beta_t << 1$), we could approximate it as a Gaussian of the form.

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

Intuition

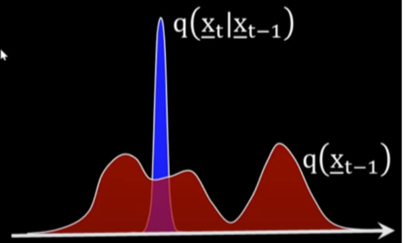
$x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_T$: The migration from x_{t-1} to x_t is essentially an addition of slight noise and consequently, a blur of PDF.

$x_0 \leftarrow \cdots \leftarrow x_{T-1} \leftarrow x_T$: The step from x_t to x_{t-1} should be a delicate denoising of x_t leading to a 'sharpening' effect of the PDF.

- Here is a different perspective that might support our Gaussianity assumption on $p_\theta(\underline{x}_{t-1}|\underline{x}_t)$
- Recall the Bayes relation:

$$q(\underline{x}_{t-1}|\underline{x}_t) = \frac{q(\underline{x}_t|\underline{x}_{t-1})q(\underline{x}_{t-1})}{q(\underline{x}_t)}$$

$$\propto \underbrace{q(\underline{x}_t|\underline{x}_{t-1})}_{\text{A simple and known Gaussian}} \cdot \underbrace{q(\underline{x}_{t-1})}_{\text{An involved \& unknown distribution}}$$



- While $q(\underline{x}_{t-1})$ is unknown, it is expected to be "much wider" than $q(\underline{x}_t|\underline{x}_{t-1})$, and thus can be considered as a constant in this multiplication

Hence we got to know, that it is also **GAUSSIAN**

So to be clear, we shall assume

So, to be clear - we shall assume:

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

and all that remains is to make smart choices regarding the identity of

$$\mu_\theta(x_t, t) \text{ and } \sigma_\theta(x_t, t)$$

→ Many approaches

The formal approach →

The gaussian $q(x_t | x_{t-1})$ define the forward diffusion, given by -

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

The joint probability of the whole path of this forward Markov process is

$$q(x_T, x_{T-1}, \cdots x_1 | x_0) = q(x_T | x_{T-1}) \cdots q(x_2 | x_1) q(x_1 | x_0)$$

Denote this as $\boxed{q(x_{1:T} | x_0)}$

The joint probability of whole path of the Reversed Markov process is

$$p_\theta(x_T, x_{T-1}, \cdots, x_1, x_0) = p_\theta(x_0 | x_1) \cdots p(x_{T-1} | x_T) p(x_T)$$

Denote this as $\boxed{p_\theta(x_{0:T})}$

$$\therefore \; -E_{q(x_0)} \log p_\theta(x_0) \leq E_{q(x_{0:T})} \log \frac{q(x_{1:T} | x_0)}{\cdots}$$

$$q(x_0) \quad \ldots \quad p_\theta(x_{0:T})$$

$$\text{VB (Variational Bound)}$$

goal is to minimize the LHS - the expected negative log likelihood of the true images, so that this probability $p_\theta(x_0)$ is maximal.

instead we minimize the RHS as a proxy (VB) this is closely related to the ELBO use in VAE.

after some massaging of variational bound it comes down to.

$$\boxed{VB = \underbrace{E_{q(0:T)} \log \frac{q(x_T|x_0)}{p_\theta(x_T)}} - \underbrace{E_{q(x_0:T)} \log p_\theta(x_0|x_1)} + \underbrace{E_{q(x_0:T)} \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}}}$$

This expression is zero since x_T is a gaussian and it remains the same even if x_0 is given

we can either neglect this or handle it by a specifically trained gaussian

This is a KL divergence b/w two isotropic gaussians and thus it has a closed form.

$$VB = E_{q(x_0:T)} \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1},x_t)} \cong \sum_{t=2}^{T} KL\left(q(x_{t-1}|x_t,x_0), p_\theta(x_{t-1}|x_t)\right)$$

These two gaussian are given by $p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t,t), \sigma_t^2 I)$

$$q(x_{t-1}|x_t,x_0) = N\left(x_{t-1}; \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\varepsilon_T\right), \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t I\right)$$

We start by setting $\boxed{\sigma_t^2 = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}$ and then these divergences are given by :

$$KL\left(q(x_{t-1}|x_t,x_0), p_\theta(x_{t-1}|x_t)\right) = \frac{1}{\sigma^2}\left\| \frac{1}{\sqrt{1-\beta}}\left(x_t - \frac{\beta_t}{\ldots}\varepsilon_t\right) - \mu_\theta(x_t,t)\right\|^2$$

Thus we set $\boxed{\mu_\theta(x_t, t) = \frac{1}{\sqrt{1-\beta_t}}\,\hat{\epsilon}_t(x_t, t)}$

$$\therefore \boxed{VB = \sum_{t=2}^{T} \boxed{\frac{\beta_t}{(1-\beta_t)(1-\alpha_{t-1})}} \left\| \epsilon_t - \hat{\epsilon}_t(x_t, t) \right\|_2^2}$$

This is the loss for the denoiser design.

Further replaced by 1

F