Review Diffusion model

Problem Formulation: To generate image, we gradually add noise to a given set of images in order to transform it a white noise (Gaussian distribution with mean 0 and variance I).

Forward process: Adding noise to given sample x_0 gradually and expecting that at the end of process $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

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

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

Set $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we have:

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

The special point is that, using that notation gives the ability of sampling x_t at arbitrary timestep t, which is achieved by transformation below:

$$\begin{aligned} x_t &= \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, z_{t-1} && (1)\\ &= \sqrt{\alpha_t}\, (\sqrt{\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_{t-1}}\, z_{t-2}) + \sqrt{1-\alpha_t}\, z_{t-1} && (2)\\ &= \sqrt{\alpha_t \alpha_{t-1}}\, x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\, z_{t-2} + \sqrt{1-\alpha_t}\, z_{t-1} && (3) \end{aligned}$$

Where $z_{t-1},\ z_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Think $\sqrt{\alpha_t(1-\alpha_{t-1})}\, z_{t-2}$ as random variable u and $\sqrt{1-\alpha_t} z_{t-1}$ as random variable v, we have $u \sim \mathcal{N}(\mathbf{0}, \alpha_t(1-\alpha_{t-1})\mathbf{I})$ and $v \sim \mathcal{N}(\mathbf{0}, (1-\alpha_t)\mathbf{I})$. Since u and v are independent, $V[u+v] = V[u]+V[v]$ and we have $(u+v) \sim \mathcal{N}(\mathbf{0}, (1-\alpha_t\alpha_{t-1})\mathbf{I})$. Then equation (3) can be rewrited as follow:

$$\begin{aligned} \mathbf{x_t} &= \sqrt{\alpha_t \alpha_{t-1}}\, x_{t-2} + (u+v) && (4)\\ &= \sqrt{\alpha_t \alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}}\, z && (5) \end{aligned}$$

Where $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, so we have $x_t \sim \mathcal{N}(\sqrt{\alpha_t\alpha_{t-1}}\, x_{t-2}, (1-\alpha_t\alpha_{t-1})\mathbf{I})$. Do this transformation repeatedly and we have closed form for sampling x_t from x_0 as follow:

$$\begin{aligned} \mathbf{x_t} &= \sqrt{\alpha_t \alpha_{t-1} ... \alpha_1}\, x_0 + \sqrt{1-\alpha_t \alpha_{t-1} ... \alpha_1}\, z && (6)\\ &= \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, z && (7) \end{aligned}$$

Where $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, so $x_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\mathbf{I})$.

Background for deriving loss function:

What is ELBO (Evidence Lower Bound)?

Let say that in a latent variable model, we posit that our observed data x is a realization of a random variable X. Moreover, we posit the existence of another random variable Z where X and Z are distributed according to a joint distribution $p(X, Z, \theta)$. The Z variable does not have any observation so remains a latent variable. With that model, there are two predominate tasks which are interesting to dig in:

- Given some fixed value for θ, compute the posterior distribution $p(Z|X, \theta)$.
- Given θ is unknown, find the maximum likelihood estimate of θ: $\text{argmax}_\theta l(\theta)$, where $l(\theta)$ is the log likelihood function:

$$\begin{aligned} l(\theta) &= log\left(p(x, \theta)\right) \\ &= log \int\limits_z p(x, z, \theta) dz \end{aligned}$$

The term "evidence" is just a name given to the likelihood function $l(\theta) = log(x, \theta)$, so "evidence lower bound" is the lower bound of likelihood function, which can be derived as follow:

$$\begin{aligned} log(x, \theta) &= log \int\limits_z p(x, z, \theta) dz \\ &= log \int\limits_z q(z) \frac{p(x, z, \theta)}{q(z)} dz \\ &= log \left(\mathbf{E}_{z \sim q} \left[\frac{p(x, z, \theta)}{q(z)} \right] \right) \\ &\geq \mathbf{E}_{z \sim q} \left[log \left(\frac{p(x, z, \theta)}{q(z)} \right) \right] \end{aligned}$$

The last inequality follows from Jensen inequality. The gap between evidence and ELBO is KL divergence:

$$\begin{aligned} log\left(p(x, \theta)\right) - \mathbf{E}_{z \sim q} \left[log \left(\frac{p(x, z, \theta)}{q(z)} \right) \right] &= \mathbf{E}_{z \sim q} \left[log\left(p(x, \theta)\right) \right] - \mathbf{E}_{z \sim q} \left[log \left(\frac{p(x, z, \theta)}{q(z)} \right) \right] \\ &= \mathbf{E}_{z \sim q} \left[log\left(p(x, \theta)\right) - log \left(\frac{p(x, z, \theta)}{q(z)} \right) \right] \\ &= \mathbf{E}_{z \sim q} \left[log \left(p(x, \theta) \frac{q(z)}{p(x, z, \theta)} \right) \right] \\ &= \mathbf{E}_{z \sim q} \left[log \frac{q(z)}{p(z|x, \theta)} \right] \\ &= \mathbf{D}_{\text{KL}}(q(z) \,||\, p(z|x, \theta) \end{aligned}$$

Training is procedure of maximizing negative log likelihood, which can be achieved approximately by maximizing ELBO (variational bound).

$$\mathbf{E}[-\log p_\theta(x_0)] \leq \mathbf{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

Set $L = \mathbf{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$ and make a few transformation, we have:

$$\begin{aligned}
L &= \mathbf{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] \\
&= \mathbf{E}_q \left[-\log p(x_T) - \sum_{t\geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] \\
&= \mathbf{E}_q \left[-\log p(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} - \log(\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}) \right] \\
&= \mathbf{E}_q \left[-\log p(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} . \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} - \log(\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}) \right] \text{(apply bayes rule)} \\
&= \mathbf{E}_q \left[-\log p(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \sum_{t>1} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} - \log(\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}) \right] \\
&= \mathbf{E}_q \left[-\log p(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log \frac{1}{q(x_T|x_0)} - \log(\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}) \right] \\
&= \mathbf{E}_q \left[-\log \frac{p(x_T)}{q(x_T|x_0)} - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log(\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}) \right]
\end{aligned}$$

Let $L_t = \mathbf{E}_q \left[-\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \right]$ as an example.

$$\begin{aligned}
\mathbf{E}_q \left[-\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \right] &= \int q(x_{0:T}) \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} dx_{0:T} \\
&= \int q(x_{t-1}|x_t, x_0) q(x_t, x_0) q(x_{1:t-2}, x_{t+1:T}|x_0, x_{t-1}, x_t) \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} dx_{0:T} \\
&= \int \left[\int q(x_{t-1}|x_t, x_0) \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} dx_{t-1} \right] q(x_{0:t-2}, x_{t:T}|x_{t-1}) dx_{0:t-2,t:T} \\
&= \int [\mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))]\, q(x_{0:t-2}, x_{t:T}|x_{t-1}) dx_{0:t-2,t:T} \\
&= \mathbf{E}_{q(x_{0:t-2}, x_{t:T}|x_{t-1})} [\mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))]
\end{aligned}$$

One noticeable property is that with given x_{t-1}, x_0, $q(x_{t-1}|x_0, x_t)$ and $p_\theta(x_{t-1}|x_t)$ are completely defined and $\mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))$ is known, or can be set as a constant. Thus, we have the following equation:

$$\mathbf{E}_{q(x_{t-1})}[\mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))] = \mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))$$

From that observation, we finally get the form of L_t:

$$\begin{aligned}
\mathbf{E}_q\left[-\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right] &= \mathbf{E}_{q(x_{0:t-2}, x_{t:T}|x_{t-1})}[\mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))] \\
&= \mathbf{E}_{q(x_{0:t-2}, x_{t:T}|x_{t-1})}\mathbf{E}_{q(x_{t-1})}[\mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))] \\
&= \mathbf{E}_{q(x_{0:T})}[\mathrm{D_{KL}}(q(x_{t-1}|x_0, x_t)||p_\theta(x_{t-1}|x_t))]
\end{aligned}$$

Applying above transformation to the general L, we get the final form of L:

$$\begin{aligned}
L &= \mathbf{E}_q\left[-\log\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] \\
&= \mathbf{E}_q\left[\mathrm{D_{KL}}(q(x_T|x_0)||\, p(x_T)) + \sum_{t>1}\mathrm{D_{KL}}(q(x_{t-1}|x_t, x_0)||\, p_\theta(x_{t-1}|x_t)) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})\right]
\end{aligned}$$

We can formulate form of posterior $q(x_{t-1}|xt, x_0)$ base on Bayes rule.

$$\begin{aligned}
q(x_{t-1}|xt, x_0) &= \frac{q(x_t, x_{t-1}|x_0)}{q(x_t|x_0)} && (8) \\
&= q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} && (9) \\
&= q(x_t|x_{t-1})\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} && (10) \\
&\propto \exp\left(-\frac{1}{2}(\frac{(x_t - \sqrt{\alpha_t}\, x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\, x_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t})\right) && (11)
\end{aligned}$$

(Shout out for Cuong Pham)

$$= \exp\left(-\frac{1}{2}((\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})x_{t-1}^2 - (\frac{2\sqrt{\alpha_t}}{\beta_t}x_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0)x_{t-1} + \mathbf{C}(x_t, x_0))\right) \quad (12)$$

Where $C(x_t, x_0)$ is a term not relating to x_t. Equation **??** can be derived to Gassian form. Making a simple transformation to do that.

$$
\begin{aligned}
&(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})x_{t-1}^2-(\frac{2\sqrt{\alpha_t}}{\beta_t}x_t+\frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0)x_{t-1}+\mathbf{C}(x_t,x_0) &(13)\\
&=(\frac{\alpha_t(1-\bar{\alpha}_{t-1})+\beta_t}{\beta_t(1-\bar{\alpha}_{t-1})})\,x_{t-1}^2-2(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t+\sqrt{\bar{\alpha}_{t-1}}\beta_t x_0}{\beta_t(1-\bar{\alpha}_{t-1})})x_{t-1}+\mathbf{C}(x_t,x_0) &(14)\\
&=\frac{\alpha_t-\alpha_t\bar{\alpha}_{t-1}+1-\alpha_t}{\beta_t(1-\bar{\alpha}_{t-1})}x_{t-1}^2-2(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\beta_t(1-\bar{\alpha}_{t-1})}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\beta_t(1-\bar{\alpha}_{t-1})}x_0)x_{t-1}+\mathbf{C}(x_t,x_0) &(15)\\
&=\frac{1-\bar{\alpha}_t}{\beta_t(1-\bar{\alpha}_{t-1})}x_{t-1}^2-2(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\beta_t(1-\bar{\alpha}_{t-1})}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\beta_t(1-\bar{\alpha}_{t-1})}x_0)x_{t-1}+\mathbf{C}(x_t,x_0) &(16)\\
&=\frac{1-\bar{\alpha}_t}{\beta_t(1-\bar{\alpha}_{t-1})}\left(x_{t-1}^2-2(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0)x_{t-1}\right)+\mathbf{C}(x_t,x_0) &(17)
\end{aligned}
$$

From here, we have $\mathbf{E}_{q(x_{t-1}|x_t,x_0)}[x_{t-1}]=\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0$ and $\mathbf{V}_{q(x_{t-1}|x_t,x_0)}[x_{t-1}]=\frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$.

Training now is to find $p_\theta(x_{t-1}|x_t)$ that approximates $q(x_{t-1}|x_t,x_0)$. It is straight to form the reverse probability as a Gaussian distribution: $\mathcal{N}(x_{t-1},\mu_\theta(x_t,t),\Sigma_\theta(x_t,t))$.

According to DDPM, the term Σ_θ can be fixed and the model can even create high-quality samples. In detail, it is fixed to $\sigma_t^2\mathbf{I}=\tilde{\beta}\mathbf{I}$. The critical part is to define an effective way to approximate the μ_θ. As the posterior and the approximate reverse distribution has the same form, we can rewrite the loss as follow.

$$
\begin{aligned}
\mathrm{D_{KL}}(q(x_{t-1}|x_0,x_t)||p_\theta(x_{t-1}|x_t)) &= \frac{1}{2}\log\frac{\sigma_t}{\tilde{\sigma}_t}+\frac{\tilde{\sigma}_t^2+\parallel\tilde{\mu}-\mu_\theta\parallel^2}{2\sigma_t^2}-\frac{1}{2}\\
&=\frac{\parallel\tilde{\mu}-\mu_\theta\parallel^2}{2\sigma_t^2}+C\\
&\text{(Due to the assumption of } \sigma_t^2\mathbf{I}=\tilde{\beta}\mathbf{I})
\end{aligned}
$$