

Notes on Diffusion Models

Karthik Balaji O
University of Waterloo
k3balaji@uwaterloo.ca

Derivation of the log likelihood lower bound

One can view a diffusion model as a hierarchical VAE with the following caveats:

- The encoder is not learned but is a pre-defined function, specifically, a normal distribution centered around the latent at the previous timestep.
- The latent dimension is equal to the data dimension.
- The latent distribution approaches the standard normal distribution as the number of timesteps approaches infinity.

The probability of the trajectory $x_{1:T}$ given the initial data point x_0 is

$$q(x_{0:T} | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}) \quad (1)$$

where T is the total number of timesteps. The probability of the reverse trajectory $x_{0:T}$ given a latent x_T sampled from a standard normal is

$$p(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t) \quad (2)$$

We need to maximize the log likelihood of the data point

$$\begin{aligned} \log p(x) &= \log \int_{x_{1:T}} p(x_{0:T}) dx_{1:T} \\ &= \log \int_{x_{1:T}} q(x_{1:T} | x_0) \frac{p(x_{0:T})}{q(x_{1:T} | x_0)} dx_{1:T} \\ &= \log \mathbb{E}_{q(x_{1:T} | x_0)} \left[\frac{p(x_{0:T})}{q(x_{1:T} | x_0)} \right] \\ &\geq \mathbb{E}_{q(x_{1:T} | x_0)} \left[\log \frac{p(x_{0:T})}{q(x_{1:T} | x_0)} \right] && \text{(Jensen's inequality)} \\ &= \mathbb{E}_{q(x_{1:T} | x_0)} \left[\log \frac{p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t)}{\prod_{t=1}^T q(x_t | x_{t-1})} \right] \\ &= \mathbb{E}_{q(x_{1:T} | x_0)} \left[\log \frac{p(x_T) p_\theta(x_0 | x_1) \prod_{t=2}^T p_\theta(x_{t-1} | x_t)}{q(x_1 | x_0) \prod_{t=2}^T q(x_t | x_{t-1})} \right] \end{aligned}$$

Now we need to use a notational trick. We write $q(x_t | x_{t-1})$ above as $q(x_t | x_{t-1}, x_0)$ and use Bayes' rule

$$q(x_t | x_{t-1}, x_0) = \frac{q(x_t | x_0) q(x_{t-1} | x_t, x_0)}{q(x_{t-1} | x_0)} \quad (3)$$

Adding the extra x_0 dependence doesn't change the distribution since the forward and backward diffusion processes are Markovian. But conceptually, we are reducing the variance of the distribution estimate by making x_t depend on x_0 as well,

and it lets us simplify the expression for the log likelihood. We can now write

$$\begin{aligned}
\log p(x) &\geq \mathbb{E}_{q(x_{1:T} | x_0)} \left[\log \frac{p(x_T) p_\theta(x_0 | x_1) \prod_{t=2}^T p_\theta(x_{t-1} | x_t)}{q(x_1 | x_0) \prod_{t=2}^T \frac{q(x_t | x_0) q(x_{t-1} | x_t, x_0)}{q(x_{t-1} | x_0)}} \right] \\
&= \mathbb{E}_{q(x_{1:T} | x_0)} \left[\log \frac{p(x_T) p_\theta(x_0 | x_1)}{q(x_1 | x_0)} \prod_{t=2}^T \frac{p_\theta(x_{t-1} | x_t)}{q(x_{t-1} | x_t, x_0)} \prod_{t=2}^T \frac{q(x_{t-1} | x_0)}{q(x_t | x_0)} \right] \\
&= \mathbb{E}_{q(x_{1:T} | x_0)} \left[\log \frac{p(x_T) p_\theta(x_0 | x_1)}{q(x_T | x_0)} \prod_{t=2}^T \frac{p_\theta(x_{t-1} | x_t)}{q(x_{t-1} | x_t, x_0)} \right] \\
&= \mathbb{E}_{q(x_1 | x_0)} [\log p_\theta(x_0 | x_1)] + \mathbb{E}_{q(x_T | x_0)} \left[\log \frac{p(x_T)}{q(x_T | x_0)} \right] + \sum_{t=2}^T \mathbb{E}_{q(x_{t-1}, x_t | x_0)} \left[\log \frac{p_\theta(x_{t-1} | x_t)}{q(x_{t-1} | x_t, x_0)} \right]
\end{aligned}$$

The second term can go away since it doesn't depend on parameters θ . For the last term, we can bring the expectation over x_{t-1} inside

$$\log p(x) \geq \mathbb{E}_{q(x_1 | x_0)} [\log p_\theta(x_0 | x_1)] - \sum_{t=2}^T \mathbb{E}_{q(x_t | x_0)} [D_{KL}(q(x_{t-1} | x_t, x_0) || p_\theta(x_{t-1} | x_t))] \quad (4)$$

and this is the lower bound of the objective we need to maximize.

Reparametrization trick

The variance follows a schedule over T timesteps - β_1, \dots, β_T . We let $\alpha = 1 - \beta$. At each timestep, we have

$$x_t \sim \mathcal{N}(x_t | \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t) \mathbf{I}) \quad (5)$$

Sampling is of course, not differentiable. So we have to make use of the reparametrization trick. Define $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Then

$$\begin{aligned}
x_t &= \sqrt{\alpha_t} x_{t-1} + (1 - \alpha_t) \epsilon_{t-1}^* \\
&= \sqrt{\alpha_t} (\sqrt{\alpha_{t-1}} x_{t-2} + (1 - \alpha_{t-1}) \epsilon_{t-2}^*) + (1 - \alpha_t) \epsilon_{t-1}^* \\
&= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \epsilon_{t-2} \quad (\text{using } \mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)) \\
&= \dots \\
&= \sqrt{\prod_{i=1}^t \alpha_i} x_0 + \sqrt{1 - \prod_{i=1}^t \alpha_i} \epsilon_0 \\
&= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_0 \quad (\epsilon_0 \sim \mathcal{N}(0, \mathbf{I}))
\end{aligned}$$

Forward diffusion process posterior

How do we actually compute the KL divergence term in the sum in equation (4)? We first need to know what $q(x_{t-1} | x_t, x_0)$ is. Using Bayes' rule

$$q(x_{t-1} | x_t, x_0) = \frac{q(x_t | x_{t-1}, x_0) q(x_{t-1} | x_0)}{q(x_t | x_0)} \quad (6)$$

We have defined q as a Gaussian. Using the reparametrization trick, we know that

$$\begin{aligned}
q(x_t | x_0) &= \mathcal{N}(x_t | \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I}) \\
q(x_{t-1} | x_0) &= \mathcal{N}(x_{t-1} | \sqrt{\bar{\alpha}_{t-1}} x_0, (1 - \bar{\alpha}_{t-1}) \mathbf{I})
\end{aligned}$$

and according to equation (5), we have

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

Substituting these into equation (6) and simplifying, we get

$$q(x_{t-1} | x_t, x_0) = \frac{\mathcal{N}(x_t | \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t) \mathbf{I}) \mathcal{N}(x_{t-1} | \sqrt{\bar{\alpha}_{t-1}} x_0, (1 - \bar{\alpha}_{t-1}) \mathbf{I})}{\mathcal{N}(x_t | \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I})} \quad (7)$$

We now look at the coefficients. Let's consider the product of coefficients of the Gaussians in the numerator by the coefficient of the Gaussian in the denominator

$$(2\pi\tilde{\beta})^{-\frac{d}{2}} = \frac{(2\pi(1 - \alpha_t))^{-\frac{d}{2}} (2\pi(1 - \bar{\alpha}_{t-1}))^{-\frac{d}{2}}}{(2\pi(1 - \bar{\alpha}_t))^{-\frac{d}{2}}} \quad (8)$$

So the overall variance $\tilde{\beta}$ is

$$\tilde{\beta} = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \quad (9)$$

The d in the exponent comes from the fact that we're dealing with a multivariate Gaussian and is the data dimensionality. We now consider the desired functional form (that of a Gaussian), which is on the left, and the expression for the forward posterior on the right. We're ignoring the constant in front of the Gaussian distribution expression for now. We have

$$\begin{aligned} \exp\left\{-\frac{1}{2\tilde{\beta}}(x_{t-1} - \tilde{\mu})^2\right\} &= \exp\left\{-\left[\frac{(x_t - \sqrt{\alpha_t} x_{t-1})^2}{2(1 - \alpha_t)} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0)^2}{2(1 - \bar{\alpha}_{t-1})} - \frac{(x_t - \sqrt{\bar{\alpha}_t} x_0)^2}{2(1 - \bar{\alpha}_t)}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} x_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}}{1 - \alpha_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} x_0\right) x_{t-1} + f(x_t, x_0)\right]\right\} \end{aligned}$$

where we expanded the terms and separated the x_{t-1} terms. $f(x_t, x_0)$ is a constant w.r.t x_{t-1} and we're using it to denote the rest of the expansion. Note that the coefficient of x_{t-1}^2 is the inverse of the variance $\tilde{\beta}$. We factor out $\frac{1}{\tilde{\beta}}$

$$\exp\left\{-\frac{1}{2\tilde{\beta}}(x_{t-1} - \tilde{\mu})^2\right\} = \exp\left\{-\frac{1}{2\tilde{\beta}}\left[x_{t-1}^2 - 2\underbrace{\left(\frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} x_0\right)}_{\tilde{\mu}} x_{t-1} + \underbrace{\tilde{\beta} f(x_t, x_0)}_{\tilde{\mu}^2}\right]\right\}$$

The $\tilde{\beta} f(x_t, x_0)$ term simplifies to $\tilde{\mu}^2$ above, and so we have a complete square in the exponent by setting

$$\tilde{\mu}(x_t, x_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} x_0 \quad (10)$$

So we arrive at the forward diffusion process posterior distribution

$$\begin{aligned} q(x_{t-1} | x_t, x_0) &= \mathcal{N}(x_{t-1} | \tilde{\mu}(x_t, x_0), \tilde{\beta}) \\ &= \mathcal{N}\left(x_{t-1} | \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} x_0, \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\right) \end{aligned} \quad (11)$$

The loss function

We now simplify the KL divergence in equation (4). We need to consider the KL divergence between two multivariate Gaussians. The KL divergence between two Gaussians is given by

$$\mathcal{KL}(\mathcal{N}(x | \mu_1, \Sigma_1) \| \mathcal{N}(x | \mu_2, \Sigma_2)) = \frac{1}{2} \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right] \quad (11)$$

Our covariance matrices are diagonal matrices. For a particular timestep t in the sum in equation (4), the mean of the forward posterior is given by equation (10), denoted by $\mu_q(x_t, x_0)$. We define the functional form of the mean predicted by the neural net for the reverse distribution as

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \hat{x}_\theta(x_t, t) \quad (12)$$

The KL divergence at timestep t - after all this derivation - simplifies to

$$D_{KL}(q(x_{t-1} | x_t, x_0) \| p_\theta(x_{t-1} | x_t, x_0)) = \frac{1}{\beta_t} \frac{\bar{\alpha}_{t-1}(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^2} \|\hat{x}_\theta(x_t, t) - x_0\|^2 \quad (13)$$

If we solve for x_0 in equation we derived for reparametrization, we'll get

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} x_t - \frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \epsilon_0 \quad (14)$$

If we substitute this in equation (10), simplify, and reframe the functional form of the mean of the reverse distribution by replacing $\hat{x}_\theta(x_t, t)$ with $\hat{\epsilon}_\theta(x_t, t)$, we'll get

$$D_{KL}(q(x_{t-1} | x_t, x_0) \| p_\theta(x_{t-1} | x_t, x_0)) = \frac{1}{\beta_t} \frac{(1 - \alpha)^2}{(1 - \bar{\alpha}_t)\alpha_t} \|\hat{\epsilon}_\theta(x_t, t) - \epsilon_0\|^2 \quad (15)$$

Ho et. al. found that ignoring the constant in front gave better results. They also chose to not add noise for the first timestep, so they dropped the reconstruction term in (4) as well. Substituting (15) in equation (4), our loss function is

$$\mathcal{L}(\theta) = \sum_{t=2}^T \mathbb{E}_{q(x_t | x_0)} [\|\epsilon_\theta(x_t, t) - \epsilon_0\|^2] \quad (16)$$

The incredible thing is after all this derivation, the loss function is just the sum of mean squared error between the predicted noise and the true noise over all timesteps. The training algorithm just falls out the loss function: until convergence, we randomly choose an image, randomly choose a timestep, sample ϵ_0 from a standard normal, sample x_t using the reparametrization trick, compute $\|\epsilon_\theta(x_t, t) - \epsilon_0\|^2$, and backprop.