Diffusion Model

CMSC 25025 Final Project

Akash Piya

Question 1

As given in the paper, $q_{t|t-1}(x_t|x_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I_d\big)$. Through the reparameterization trick, we can write this as

$$q_{t|t-1}(x_t|x_{t-1}) = \sqrt{\alpha_t}x_{t-1} + \varepsilon_1\sqrt{(1-\alpha_t)}$$

Writing out a few more terms in a similar manner:

$$q_{t-1|t-2}(x_{t-1}|x_{t-2}) = \sqrt{\alpha_{t-1}}x_{t-2} + \varepsilon_2\sqrt{(1-\alpha_{t-1})}$$

$$\Rightarrow q_{t|t-2}(x_t|x_{t-2}) = \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t}\sqrt{1-\alpha_{t-1}}\varepsilon_2 + \sqrt{1-\alpha_t}\varepsilon_1$$

$$q_{t-2|t-3}(x_{t-2}|x_{t-3}) = \sqrt{\alpha_{t-2}}x_{t-3} + \varepsilon_3\sqrt{1-\alpha_{t-2}}$$

$$\Rightarrow q_{t|t-3}(x_t|x_{t-3}) = \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}x_{t-3} + \sqrt{\alpha_t\alpha_{t-1}(1-\alpha_{t-2})}\varepsilon_3 + \sqrt{\alpha_t(1-\alpha_{t-1})}\varepsilon_2 + \sqrt{1-\alpha_t}\varepsilon_1$$

where $\varepsilon_i \in \mathcal{N}(0, I)$. In this form, a clear patterms forms for $q_{t|0}$ (with the substitution that $\overline{\alpha}_t = \prod_{s=1}^{t}\alpha_s$). Generalizing this pattern starting at x_0:

$$q_{t|0}(x_t|x_0) = \sqrt{\overline{\alpha_t}}x_0 + \sum_{i=1}^{t}\beta_i\varepsilon_i \text{ where } \beta_i = \sqrt{(1-\alpha_{t-i+1})\prod_{j=0}^{i-2}\alpha_{t-j}}$$

One can view this expression a "reverse" reparameterization trick. The summation term alone can be interpretated as the sum of several standard normal distributions with variance β_i^2 and mean 0. Since the sum of two normal distributions is a normal distribution with mean and variance equal to the sum of the two underlying means and variance, this summation is equivalent to a single normal distribution with mean 0 and the variance equal to the sum of all the individual variances, β_i^2. Looking at the case when $t = 3$ as an example, we can extract a general pattern.

$$\sum_{i=1}^{3}\beta_i^2 = \alpha_t\alpha_{t-1}(1-\alpha_{t-2}) + \alpha_t(1-\alpha_{t-1}) + (1-\alpha_t) = 1 - \alpha_t\alpha_{t-1}\alpha_{t-2}$$

$$\Rightarrow \sum_{i=1}^{t}\beta_i^2 = 1 - \prod_{s=1}^{t}\alpha_s = 1 - \overline{\alpha_t}$$

By this, we know that the variance of the final distribution is $1 - \overline{\alpha}_t$.

We can then write $q_{t|0}(x_t|x_0) = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon$ where $\varepsilon \sim \mathcal{N}(0, I_d)$ which implies that $q_{t|0} \sim \mathcal{N}\Big(\sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I_d\Big)$ by the reparameterization trick.

Question 2

Since we are trying to maximize log-likelihood, the ELBO inequality (below) provides a tractable upperbound that we can work to maximize.

$$\int_{x_1,\ldots,x_T} \log\left[\frac{\prod_{s=1}^{T}p_{t-1|t}(x_{t-1}|x_t;\theta)p_T(x_T)}{\prod_{t=1}^{T}q_{t|t-1}(x_t|x_{t-1})}\right]\prod_{t=1}^{T}q_{t|t-1}(x_t|x_{t-1})dx_1\ldots dx_T$$

Using the fact that the product of the conditional distributions $q_{t|t-1}$ is the joint distribution and basic logarithm properties, this can be rewritten as:

$$= \int_{x_1,...,x_T} \left(\sum_{t=1}^{T} \log \left(\frac{p_{t-1|t}(x_{t-1}|x_t)}{q_{t|t-1}(x_t|x_{t-1})} \right) + \log(p_T(x_T)) \right) q(x_1, ..., x_T|x_0) dx_1, ..., dx_T$$

Note that the integral is over all x_i yet if we expand the sum, each term is dependent on only two x_i's so we can integrate over the rest of the marginal variables.

$$= \int_{x_T} \log(p_T(x_T)) q_T(x_T|x_0) dx + \sum_{t=1}^{T} \int_{x_{t-1},x_t} \log \left(\frac{p_{t-1|t}(x_{t-1}|x_t)}{q_{t|t-1}(x_t|x_{t-1})} \right) q(x_{t-1}, x_t|x_0) dx_{t-1} dx_t$$

Question 3

We can now define the loss that tries to maximize the ELBO upper bound by minimizing the negative of the term on the right from the equation above:

$$L(\theta, X_0) = \sum_{t=1}^{T} \int_{x_{t-1},x_t} -\log p(x_{t-1}|x_t; \theta) q_{t-1,t|0}(x_{t-1}, x_t|X_0) dx_{t-1} dx_t \tag{1}$$

Because the forward process: $q(x_t|x_{t-1})$ follows a Gaussian, we assume that the reverse process does as well. The mean is unknown and parameterized in terms of x_t and θ, while we assume the covariance matrix is given by $(1 - \alpha_t)I_d$.

$$p(x_{t-1}|x_t; \theta) = C \exp \left[-\frac{1}{2} (x_{t-1} - \mu(x_t, t; \theta))^T ((1 - \alpha_t) I_d)^{-1} (x_{t-1} - \mu(x_t, t; \theta)) \right]$$

where C is a normalization constant that we can ignore. Of note in this formulation is that the covariance matrix is diagonal with identical entries, $(1 - \alpha_t)$. Hence this matrix scales vectors by a constant and can be treated as a scalar rather than a matrix:

$$-\log p(x_{t-1}|x_t; \theta) = \frac{1}{2(1 - \alpha_t)} \ |x_{t-1} - \mu(x_t, t; \theta)|^2 + C$$

Plugging this into Equation 1, we get that

$$L(\theta, X_0) = \sum_{t=1}^{T} \int_{x_{t-1},x_t} \frac{|x_{t-1} - \mu(x_t, t; \theta)|^2}{2(1 - \alpha_t)} q_{t-1,t|0}(x_{t-1}, x_t|X_0) dx_{t-1} dx_t + C$$

where C is again independent of θ. The term inside the integral looks likes an expectation function where x_t and x_{t-1} are drawn from the distribution $q_{t-1,t|0}(x_{t-1}, x_t|X_0)$. The loss then takes the following form:

$$L(\theta, X_0) = \sum_{t=1}^{T} E_{q_{t-1,t|0}} \left[\frac{|X_{t-1} - \mu(X_t, t; \theta)|^2}{2(1 - \alpha_t)} | X_0 \right] + C$$

Question 4

Note that because q is Markov, $q_{t|t-1,0}(x_t|x_{t-1}, x_0) = q_{t|t-1}(x_t|x_{t-1})$ and that $q_{t-1,t|0}(x_{t-1}, x_t|x_0) = q_{t|t-1,0}(x_t|x_{t-1}, x_0) q_{t-1|0}(x_{t-1}|x_0)$. . This last term $q_{t-1|0}$ has a distribution specified in Question 1 that will be normal. Given this value, we can run the forward process on x_{t-1} to get a distribution over x_t. Using this, we can calculate $q_{t-1,t|0}$ for all values of x_{t-1} and x_t which we can substitute above.

Question 5

By Bayes Theorem,

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

We know that each of these terms can be written as some normal distribution and multiplied together:

$$q(x_{t-1}|x_t, x_0) = C \exp\left(-\frac{1}{2}\frac{|x_t - \sqrt{\alpha_t}x_{t-1}|^2}{1-\alpha_t}\right)$$

$$q(x_{t-1}|x_0) = D \exp\left(-\frac{1}{2}\frac{|x_{t-1} - \sqrt{\overline{\alpha}_{t-1}}x_0|^2}{1-\overline{\alpha}_{t-1}}\right)$$

$$q(x_t|x_0) = E \exp\left(-\frac{1}{2}\frac{|x_t - \sqrt{\overline{\alpha}_t}x_0|^2}{1-\overline{\alpha}_t}\right)$$

where $C \propto \frac{1}{\sqrt{1-\alpha_t}}, D \propto \frac{1}{\sqrt{1-\overline{\alpha}_{t-1}}}, E \propto \frac{1}{\sqrt{1-\overline{\alpha}_t}}$ are all the normalization factors. Plugging these into the initial equation:

$$q(x_{t-1}|x_t, x_0) = \frac{CD}{E} \exp\left(-\frac{1}{2}\left[\frac{|x_t - \sqrt{\alpha}_t x_{t-1}|^2}{1-\alpha_t} + \frac{|x_{t-1} - \sqrt{\overline{\alpha}_{t-1}}x_0|^2}{1-\overline{\alpha}_{t-1}} - \frac{|x_t - \sqrt{\overline{\alpha}_t}x_0|^2}{1-\overline{\alpha}_t}\right]\right)$$

Because the distribution we are looking for is over x_{t-1}, then any other term above can be grouped into a constant that depends on x_t and x_0.

$$= \frac{CD}{E} \exp\left(-\frac{1}{2}\left[\frac{x_t^2 - 2\sqrt{\alpha_t}x_t x_{t-1} + \alpha_t x_{t-1}^2}{1-\alpha_t} + \frac{x_{t-1}^2 - 2\sqrt{\overline{\alpha}_{t-1}}x_0 x_{t-1} + \overline{\alpha}_{t-1}x_0^2}{1-\overline{\alpha}_{t-1}} + C(x_t, x_0)\right]\right)$$

$$= \frac{CD}{E} \exp\left(-\frac{1}{2}\left[x_{t-1}^2\left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\overline{\alpha}_{t-1}}\right) - x_{t-1}\left(\frac{2\sqrt{\alpha}_t x_t}{1-\alpha_t} + \frac{2\sqrt{\overline{\alpha}_{t-1}}x_0}{1-\overline{\alpha}_{t-1}}\right) + C(x_t, x_0)\right]\right)$$

We can complete the square within the exponent:

$$ax^2 + bx = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a}$$

The $\frac{b^2}{4a}$ term can be lumped with C. Making the appropriate substitutions:

$$a = \left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\overline{\alpha}_{t-1}}\right) = \frac{1-\overline{\alpha}_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}$$

$$b = -2\left(\frac{\sqrt{\alpha_t}x_t(1-\overline{\alpha}_{t-1}) + \sqrt{\overline{\alpha}_{t-1}}x_0(1-\alpha_t)}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}\right)$$

$$\frac{b}{2a} = -\left(\frac{\sqrt{\alpha_t}x_t(1-\overline{\alpha}_{t-1}) + \sqrt{\overline{\alpha}_{t-1}}x_0(1-\alpha_t)}{1-\overline{\alpha}_t}\right)$$

In this formulation, we have that

$$q_{t-1|t,0} = \frac{CD}{E} \exp\left(-\frac{1}{2}\left[\frac{1-\overline{\alpha}_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})} \left(x - \frac{\sqrt{\alpha}_t x_t(1-\overline{\alpha}_{t-1}) + \sqrt{\overline{\alpha}_{t-1}} x_0(1-\alpha_t)}{1-\overline{\alpha}_t} \right)^2 + C(x_t, x_9) \right] \right)$$

We know that this function is well-normalized. The $+C$ term above can be remove from the exponential and amalgamated with $\frac{CD}{E}$. Regardless, the expression above is a gaussian with mean $\tilde{\mu}_t$ and variance ρ_t where:

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

$$\rho_t = \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}$$

Question 6

We now examine a single term in the loss function (Question 3). Given the initial expression:

$$\int \frac{|x_{t-1} - \mu(x_t, t; \theta)|^2}{2(1-\alpha_t)} q_{t-1,t|0}(x_{t-1}, x_t | X_0) dx_{t-1} dx_t$$

We can use the fact that $p(a,b) = p(a|b)p(b)$ on q.

$$= \int \frac{|x_{t-1} - \mu(x_t, t; \theta)|^2}{2(1-\alpha_t)} q_{t-1|t,0}(x_{t-1} | x_t, X_0) q_{t|0}(x_t | X_0) dx_{t-1} dx_t$$

By Question 5, we know that $q_{t-1|t,0}$ is Gaussian so defining $\tilde{\mu}_t = \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}} x_0 + (1-\overline{\alpha}_{t-1})\sqrt{\alpha_t} x_t}{1-\overline{\alpha}_t}$ and $\rho_t = \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}$

$$= \int \frac{|x_{t-1} - \mu(x_t, t; \theta)|^2}{2(1-\alpha_t)} C \exp\left[-\frac{1}{2} \frac{|x_{t-1} - \tilde{\mu}_t|^2}{\rho_t} \right] q_{t|0}(x_t | X_0) dx_{t-1} dx_t$$

where $C = \frac{1}{\sqrt{2\pi\rho_t}}$. We can integrate over x_{t-1}.

$$= \frac{C}{2(1-\alpha_t)} \int \int_{-\infty}^{\infty} |x_{t-1} - \mu(x_t, t; \theta)|^2 \exp\left[-\frac{1}{2} \frac{|x_{t-1} - \tilde{\mu}_t|^2}{\rho_t} \right] dx_{t-1} q_{t|0}(x_t | X_0) dx_t \tag{2}$$

Let's make the substitution that $y = x_{t-1} - \tilde{\mu}_t$. The integral in question then becomes:

$$= \int_{-\infty}^{\infty} |y + \tilde{\mu}_t - \mu_t|^2 \exp\left(-\frac{1}{2\rho_t} |y|^2 \right) dy$$

$$\begin{aligned} = \int_{-\infty}^{\infty} |y|^2 \exp\left(-\frac{1}{2\rho_t} |y|^2 \right) dy + \int_{-\infty}^{\infty} 2y(\tilde{\mu}_t - \mu_t) \exp\left(-\frac{1}{2\rho_t} |y|^2 \right) dy \\ + \int_{-\infty}^{\infty} (\tilde{\mu}_t - \mu_t)^2 \exp\left(-\frac{1}{2\rho_t} |y|^2 \right) dy \end{aligned} \tag{3}$$

The third term is a gaussian integral if one makes the substitution that $u = \frac{y}{\sqrt{2\rho_t}}$ which is well-known to equal $\sqrt{\pi}$. Hence this last term is $(\tilde{\mu}_t - \mu_t)^2 \sqrt{2\rho_t \pi}$. The second term is 0 because the exponential term is an even function and the multiplicative factor is linear, hence odd. The resulting function is odd and hence the integral equals 0. To calculate the first term, note that:

$$I_l = \int_{-\infty}^{\infty} \exp(-lx^2)dx \Rightarrow I_l^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-l(x^2+y^2))dydx$$

Setting $x = r\cos(\theta)$ and $y = r\sin(\theta)$ and switching to polar coordinates, which adds an r factor

$$I_l^2 = \int_0^{2\pi} d\theta \int_0^{\infty} r\exp(-lr^2)dr = \frac{\pi}{l} \Rightarrow I_l = \sqrt{\frac{\pi}{l}}$$

Now $\frac{d}{dl}I_l$ evaluated at $l=1$ equals $\int_{-\infty}^{\infty} -x^2\exp(-x^2)dx = -\frac{\sqrt{\pi}}{2}$. Refocusing our attention on the first term of Equation 3, we can make the substitution $u = \frac{y}{\sqrt{2\rho_t}}$ and then use the result we derived. The first term then equals $\sqrt{2\pi\rho_t^3}$.

Therefore, Equation 3 becomes

$$= \sqrt{2\pi\rho_t}(|\tilde{\mu}_t - \mu_t|^2 + \rho_t)$$

Plugging this result into Equation 2, we get

$$= \frac{C}{2(1-\alpha_t)} \int \sqrt{2\pi\rho_t}(|\tilde{\mu}_t - \mu_t|^2 + \rho_t) q_{t|0}(x_t|x_0)dx_t$$

Recall that C was some normalization parameter that equals $\frac{1}{\sqrt{2\pi\rho_t}}$. Then, Equation 2 equals

$$\int \frac{|\tilde{\mu}_t(x_t, x_0) - \mu(x_t, t; \theta)|^2 + \rho_t}{2(1-\alpha_t)} q_{t|0}(x_t|x_0)dx_t$$

But this expression can be further simplified by noting that this is an expected value sampling from $q_{t|0}$. This equals:

$$\mathbb{E}_{q_{t|0}}\left[\frac{|\tilde{\mu}_t(x_t, x_0) - \mu(x_t, t; \theta)|^2 + \rho_t}{2(1-\alpha_t)} \mid X_0\right]$$

Question 7

Recall that

$$\tilde{\mu}_t = \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}x_0 + (1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}x_t}{1-\overline{\alpha}_t} \text{ and } x_0 = \frac{x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon_t}{\sqrt{\overline{\alpha}_t}}$$

We then get that

$$\tilde{\mu}_t = \frac{\frac{(1-\alpha_t)\left(x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon_t\right)}{\sqrt{\alpha_t}} + (1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}x_t}{1-\overline{\alpha}_t} = \frac{1}{\sqrt{\alpha}_t(1-\overline{\alpha}_t)}\left[x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon_t - \alpha_t x_t + \alpha_t\sqrt{1-\overline{\alpha}_t}\varepsilon_t + \alpha_t x_t - \overline{\alpha}_t x_t\right]$$

$$= \frac{1}{\sqrt{\alpha}_t(1-\overline{\alpha}_t)}\left[x_t(1-\overline{\alpha}_t) - \sqrt{1-\overline{\alpha}_t}\varepsilon_t(1-\alpha_t)\right] = \frac{1}{\sqrt{\alpha_t}}\left[x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_t\right]$$

Question 8

By our work in Question 6, we know that the loss function can be written as

$$L(\theta, x_0) = \sum_{t=1}^{T} \mathbb{E}_{\varepsilon_t}\left[\frac{|\tilde{\mu}(x_t, x_0) - \mu(x_t, t; \theta)|^2}{2(1-\alpha_t)} \mid x_0\right]$$

If we define $\mu(x_t, t; \theta) = \frac{1}{\sqrt{\alpha_t}}\left[x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} e_t(x_t, t; \theta)\right]$ and use the $\tilde{\mu}$ found in Question 7, the loss function simplifies to

$$= \sum_{t=1}^{T} \mathbb{E}_{\varepsilon_t}\left[\frac{1}{2(1-\alpha_t)} \left| \frac{1}{\sqrt{\alpha_t}} \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} (e_t(x_t, t; \theta) - \varepsilon_t) \right|^2\right]$$

$$L(\theta, X_0) = \sum_{t=1}^{T} \mathbb{E}_{\varepsilon_t}\left[\frac{1-\alpha_t}{2\alpha_t(1-\overline{\alpha}_t)} \left|\varepsilon_t - e_t(x_t, t; \theta)\right|^2\right]$$

Therefore our network only needs to determine the predict the noise added given some x_t and time step t.