Diffusion Models (Loss Function)

Dr. Alireza Aghamohammadi

Predicting the Noise

- A useful modification to improve result quality is to change the neural network's objective: instead of predicting the denoised image, it predicts the total noise component added to the image.
- To illustrate, we start with the expression:

$$z_t = \sqrt{\alpha_t}\, x + \sqrt{1 - \alpha_t}\, \epsilon_t,$$

  where z_t represents the noisy image at time step t, x is the original image, α_t is a noise scale factor, and ϵ_t is a noise sample.
- We can rearrange this to isolate x:

$$x = \frac{1}{\sqrt{\alpha_t}} z_t - \frac{\sqrt{1 - \alpha_t}}{\sqrt{\alpha_t}} \epsilon_t.$$

- This formulation allows us to express the mean $m_t(x, z_t)$ of the reverse conditional distribution $q(z_{t-1} \mid z_t, x)$ as:

$$m_t(x, z_t) = \frac{1}{\sqrt{1 - \beta_t}} \left(z_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_t \right),$$

- Instead of using a neural network $f(z_t, \phi, t)$ to predict the denoised image, we introduce a neural network $g(z_t, \phi, t)$ that predicts the total noise ϵ_t added to x.
- The relationship between these two network functions can be expressed as:

$$f(z_t, \phi, t) = \frac{1}{\sqrt{1-\beta_t}} \left(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} g(z_t, \phi, t) \right).$$

- The KL divergence between the reverse conditional distribution $q(z_{t-1} \mid z_t, x)$ and the model distribution $P_r(z_{t-1} \mid z_t, \phi)$ is given by:

$$\begin{aligned} & D_{KL}\left(q(z_{t-1} \mid z_t, x) \,\|\, P_r(z_{t-1} \mid z_t, \phi)\right) \\ &= \frac{\beta_t}{2(1-\alpha_t)(1-\beta_t)} \|g(z_t, \phi, t) - \epsilon_t\|^2 + \text{const} \\ &= \frac{\beta_t}{2(1-\alpha_t)(1-\beta_t)} \|g(\sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon_t, \phi, t) - \epsilon_t\|^2 + \text{const}. \end{aligned}$$

- The reconstruction loss is:

$$\begin{aligned} \log P_r(x \mid z_1, \phi) &= -\frac{1}{2\beta_1} \|x - f(z_1, \phi, 1)\|^2 + \text{const} \\ &= -\frac{1}{2(1-\beta_1)} \|g(z_1, \phi, 1) - \epsilon_1\|^2 + \text{const}. \end{aligned}$$

- Thus, the KL divergence and reconstruction loss terms can be combined for a unified objective.

The Loss Function

- The final loss function is defined as:

$$\mathcal{L}(\phi) = -\sum_{t=1}^{T} \|g(\sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon_t, \phi, t) - \epsilon_t\|^2,$$

  where $g(\cdot)$ predicts the noise component at each step t.
- This loss has an intuitive interpretation: for each time step t and each training data point x:
  - We sample a noise vector ϵ_t to create a noisy version $z_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon_t$.
  - The loss function then measures the squared difference between the predicted noise $g(z_t, \phi, t)$ and the actual noise ϵ_t.