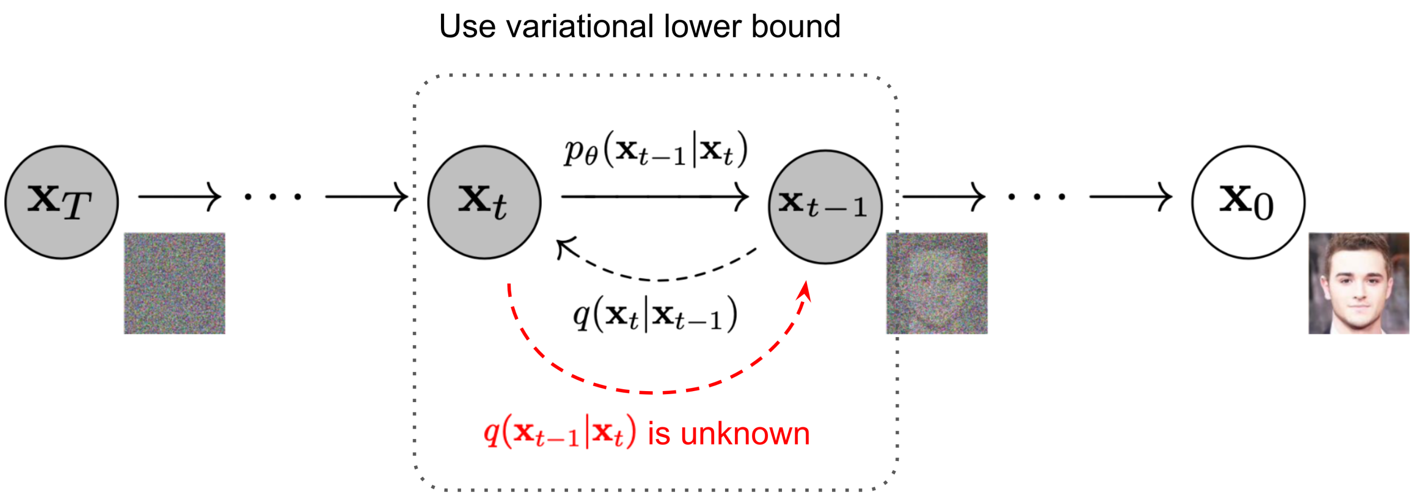
DDPM: Denoising Diffusion Probabilistic Models

0. Architecture



1. Diffusion Process

Based of **base assumption** and **reparameterization trick**:

$$
\begin{aligned}
&x_t \sim \mathcal{N}(\alpha_t x_{t-1}, \beta_t^2 I) \\
&x_t = \alpha_t x_{t-1} + \beta_t \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I)
\end{aligned}
$$

Note that similarly we have $x_{t-1} = \frac{1}{\alpha_t} x_t - \frac{\beta_t}{\alpha_t} \varepsilon_t$, which indicate that each step of **reverse diffusion process** is stiil a gaussion distrubution.
Further, to observe $p(x_t|x_0)$, we have:

$$
\begin{aligned}
x_t &= \alpha_t x_{t-1} + \beta_t \varepsilon_t, && \text{condition on observation of } x_{t-1} \\
&= \alpha_t(\alpha_{t-1} x_{t-2} + \beta_{t-1}\varepsilon_{t-1}) + \beta_t \varepsilon_t \\
&= \alpha_t \alpha_{t-1} x_{t-2} + (\alpha_t \beta_{t-1} \varepsilon_{t-1} + \beta_t \varepsilon_t) \\
&= \alpha_t \alpha_{t-1} x_{t-2} + \sqrt{\alpha_t^2 \beta_{t-1}^2 + \beta_t^2}\, \bar{\varepsilon}_{t:t-1}
\end{aligned}
$$

Given that $\alpha_t^2 + \beta_t^2 = 1$, the we have:

$$
\begin{aligned}
x_t &= \alpha_t\alpha_{t-1}x_{t-2} + \sqrt{\alpha_t^2\beta_{t-1}^2 + \beta_t^2}\,\bar{\varepsilon}_{t:t-1} \\
&= \alpha_t\alpha_{t-1}x_{t-2} + \sqrt{\alpha_t^2(1-\alpha_{t-1}^2) + (1-\alpha_t^2)}\,\bar{\varepsilon}_{t:t-1} \\
&= \alpha_t\alpha_{t-1}x_{t-2} + \sqrt{1-\alpha_t^2\alpha_{t-1}^2}\,\bar{\varepsilon}_{t:t-1} \\
&= \cdots \\
&= \Pi_{i=1}^T\alpha_i x_0 + \sqrt{1-\Pi_{i=1}^T\alpha_i^2}\,\bar{\varepsilon}_t, && \text{condition on observation of } x_0
\end{aligned}
$$

Let $\bar{\alpha}_t = \Pi_{i=1}^T\alpha_i$ and $\bar{\beta}_t = \sqrt{1-\Pi_{i=1}^T\alpha_i^2}$, we still have $\bar{\alpha}_t^2 + \bar{\beta}_t^2 = 1$, fanilly we have:

$$
x_t = \bar{\alpha}_t x_0 + \bar{\beta}_t\bar{\varepsilon}_t.
$$

Note that usually $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_T < 1$, and α_t is closed to 1.

when $T \to \infty$, $\bar{\alpha}_T \to 0$ and $\hat{\beta}_T \to 1$, further we have $x_T \sim \mathcal{N}(0, I)$.

2. Generation Process (Reverse Process)

A single step of reverse process

It is noteworthy that the reverse conditional probability is tractable when conditional on x_0:

$$
q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\tilde{\mu}(x_t, x_0), \tilde{\beta}_t^2 I)
$$

Using Bayes theorem, we have:

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

Following the standard Gaussian density function, the mean and variance can be parameterized as follow (recall that $\alpha_t^2 + \beta_t^2 = 1$ and $\bar{\alpha}_t^2 + \bar{\beta}_t^2 = 1$):

$$\tilde{\beta}_t^2 = 1/(\frac{\alpha_t^2}{\beta_t^2} + \frac{1}{\bar{\beta}_{t-1}^2}) = \frac{\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}\beta_t^2$$

$$\tilde{\mu}_t = (\frac{\alpha_t}{\beta_t^2}x_t + \frac{\bar{\alpha}_t}{\bar{\beta}_{t-1}^2})/(\frac{\alpha_t^2}{\beta_t^2} + \frac{1}{\bar{\beta}_{t-1}^2}) = \frac{\alpha_t \bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}x_t + \frac{\bar{\alpha}_{t-1}\beta_t^2}{\bar{\beta}_t^2}x_0$$

Further, we can represent $x_0 = \frac{1}{\bar{\alpha}_t}(x_t - \bar{\beta}_t \bar{\varepsilon}_t)$ and plug into the above equation and obtaion:

$$\begin{aligned}
\tilde{\mu}_t &= \frac{\alpha_t \bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}x_t + \frac{\bar{\alpha}_{t-1}\beta_t^2}{\bar{\beta}_t^2}\frac{1}{\bar{\alpha}_t}(x_t - \bar{\beta}_t \bar{\varepsilon}_t) \\
&= \frac{1}{\alpha_t}(x_t - \frac{\beta_t^2}{\bar{\beta}_t}\bar{\varepsilon}_t)
\end{aligned}$$

Analysis of predict target

Recall taht we need to learning a neural network to approximate the conditioned probability distributions in the reverse process, $x_{t-1} = \mu_\theta(x_t, t)x_t + \sigma\tilde{\varepsilon}_t$

- **a. Euclidean distance**

After the above derivation, we now analysis optimize target. A natural thinking is predict x_{t-1} directly and minimize Euclidean distance:

$$L_t = \frac{1}{\sigma_t^2}\mathbb{E}\left[||x_{t-1} - \tilde{\mu}_\theta(x_t, t)||^2\right]$$

Note that $\tilde{\mu}_\theta(x_t, t)$ is not a good predict target, and meanwhile $x_{t-1} = \frac{1}{\alpha_t}x_t - \frac{\beta_t}{\alpha_t}\varepsilon_t$ and $\tilde{\mu}_\theta(x_t, t) = \frac{1}{\alpha_t}x_t - \frac{\beta_t}{\alpha_t}\varepsilon_\theta(x_t, t)$, then we have:

$$L_t = \frac{\beta_t^2}{\sigma_t^2\alpha_t^2}\mathbb{E}\left[||\varepsilon_t - \varepsilon_\theta(x_t, t)||^2\right]$$

Further, the prediction $\varepsilon_\theta(x_t, t)$ have not based on the observation of x_0. Since,

$$\begin{aligned}
(x_t &= \bar{\alpha}_t x_0 + \bar{\beta}_t \bar{\varepsilon}_t) \\
x_t &= \alpha_t x_{t-1} + \beta_t \varepsilon_t \\
&= \alpha_t(\bar{\alpha}_{t-1}x_0 + \bar{\beta}_{t-1}\bar{\varepsilon}_t) + \beta_t\varepsilon_t \\
&= \bar{\alpha}_t x_0 + \alpha_t\bar{\beta}_{t-1}\bar{\varepsilon}_t + \beta_t\varepsilon_t
\end{aligned}$$

Plug into our predict target:

$$L_t = \frac{\beta_t^2}{\sigma_t^2 \alpha_t^2}\mathbb{E}\left[||\varepsilon_t - \varepsilon_\theta(\bar{\alpha}_t x_0 + \alpha_t \bar{\beta}_{t-1}\bar{\varepsilon}_t + \beta_t \varepsilon_t, t)||^2\right]$$

- **b. KL divergence**

$$\begin{aligned} L_t &= D_{KL}(q(x_t|x_{t+1}, x_0)||p_\theta(x_t|x_{t+1})) \\ &= \frac{1}{2\sigma_t^2}\mathbb{E}\left[||\tilde{\mu}_t - \mu_\theta||^2\right] \\ &= \frac{1}{2\sigma_t^2}\mathbb{E}\left[||\frac{1}{\alpha_t}(x_t - \frac{\beta_t^2}{\bar{\beta}_t}\bar{\varepsilon}_t) - \frac{1}{\alpha_t}(x_t - \frac{\beta_t^2}{\bar{\beta}_t}\varepsilon_\theta(x_t, t))||^2\right] \\ &= \frac{\beta_t^4}{2\sigma_t^2\bar{\beta}_t^2}\mathbb{E}\left[||\bar{\varepsilon}_t - \varepsilon_\theta(\bar{\alpha}_t x_0 + \bar{\beta}_t \bar{\varepsilon}_t)||^2\right] \end{aligned}$$

Training and sampling algorithms

Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Take gradient descent step on
    $\nabla_\theta \left\| \boldsymbol{\epsilon} - \mathbf{z}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
until converged

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{z}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

3. Parameters Setting

- About α_t, T

$$\begin{aligned} &\alpha_t^2 + \beta_t^2 = 1 \\ &T = 1000 \\ &\alpha_t = \sqrt{1 - \frac{-0.02t}{T}} \\ &\log\bar{\alpha}_t = \frac{1}{2}\sum_{t=1}^{T}\log(1 - \frac{0.02t}{T}) < \frac{1}{2}\sum_{t=0}^{T}(-\frac{0.02t}{T}) = -0.006(T+1) \approx e^{-5} \end{aligned}$$

- About σ_t

$$\sigma_t = \beta_t$$

$$\sigma_t = \frac{\bar{\beta}_{t-1}}{\bar{\beta}_t}\beta_t$$