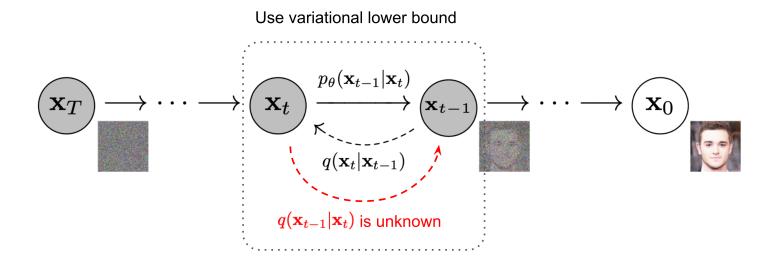
DDPM: Denoising Diffusion Probabilistic Models

0. Architecture



1. Diffusion Process

Based of base assumption and reparameterization trick:

$$egin{aligned} x_t &\sim \mathcal{N}(lpha_t x_{t-1}, eta_t^2 I) \ x_t &= lpha_t x_{t-1} + eta_t arepsilon_t, & arepsilon_t &\sim \mathcal{N}(0, I) \end{aligned}$$

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 still a gaussion distrubution.

Further, to observe $p(x_t|x_0)$, we have:

$$egin{aligned} x_t &= lpha_t x_{t-1} + eta_t arepsilon_t, & ext{condition on observation of } x_{t-1} \ &= lpha_t (lpha_{t-1} x_{t-2} + eta_{t-1} arepsilon_{t-1}) + eta_t arepsilon_t \ &= lpha_t lpha_{t-1} x_{t-2} + (lpha_t eta_{t-1} arepsilon_{t-1} + eta_t arepsilon_t) \ &= lpha_t lpha_{t-1} x_{t-2} + \sqrt{lpha_t^2 eta_{t-1}^2 + eta_t^2} \, ar{arepsilon}_{t:t-1} \end{aligned}$$

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

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

Let
$$ar{lpha}_t=\Pi_{i=1}^Tlpha_i$$
 and $ar{eta}_t=\sqrt{1-\Pi_{i=1}^Tlpha_i^2}$, we still have $ar{lpha}_t^2+ar{eta}_t^2=1$, family we have: $x_t=ar{lpha}_tx_0+ar{eta}_tar{arepsilon}_t.$

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

when $T o\infty$, $arlpha_T o 0$ and $\hateta_T o 1$, 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 :

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

Using Bayes theorem, we have:

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

$$\propto \exp(-\frac{1}{2}(\frac{(x_t - \alpha_t x_{t-1})^2}{\beta_t^2} + \frac{(x_{t-1} - \bar{\alpha}_{t-1} x_0)^2}{\bar{\beta}_{t-1}^2} + \frac{(x_t - \bar{\alpha}_t x_0)^2}{\bar{\beta}_t^2}))$$

$$= \exp(-\frac{1}{2}((\frac{\alpha_t^2}{\beta_t^2} + \frac{1}{\bar{\beta}_{t-1}^2})x_{t-1}^2 - 2(\frac{\alpha_t}{\beta_t^2} x_t + \frac{\bar{\alpha}_t}{\bar{\beta}_{t-1}^2})x_{t-1} + C(x_t, x_0)))$$

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 $\bar{\alpha}_t^2+\bar{\beta}_t^2=1$):

—

$$egin{aligned} ilde{eta}_t^2 &= 1/(rac{lpha_t^2}{eta_t^2} + rac{1}{ar{eta}_{t-1}^2}) = rac{eta_{t-1}^2}{ar{eta}_t^2}eta_t^2 \ ilde{\mu}_t &= (rac{lpha_t}{eta_t^2} x_t + rac{ar{lpha}_t}{ar{eta}_{t-1}^2})/(rac{lpha_t^2}{eta_t^2} + rac{1}{ar{eta}_{t-1}^2}) = rac{lpha_t ar{eta}_{t-1}^2}{ar{eta}_t^2} x_t + rac{ar{lpha}_{t-1} eta_t^2}{ar{eta}_t^2} x_0 \end{aligned}$$

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:

$$egin{aligned} ilde{\mu_t} &= rac{lpha_tar{eta}_{t-1}^2}{ar{eta}_t^2}x_t + rac{ar{lpha}_{t-1}eta_t^2}{ar{eta}_t^2}rac{1}{ar{lpha}_t}(x_t - ar{eta}_tar{arepsilon}_t) \ &= rac{1}{lpha_t}(x_t - rac{eta_t^2}{ar{eta}_t}ar{arepsilon}_t) \end{aligned}$$

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:

$$L_t = rac{1}{\sigma_t^2} \mathbb{E}\left[||x_{t-1} - ilde{\mu}_{ heta}(x_t, t)||^2
ight]$$

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:

$$L_t = rac{eta_t^2}{\sigma_t^2 lpha_t^2} \mathbb{E}\left[||arepsilon_t - arepsilon_{ heta}(x_t,t)||^2
ight]$$

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

$$egin{aligned} (x_t &= ar{lpha}_t x_0 + ar{eta}_t ar{arepsilon}_t) \ x_t &= lpha_t x_{t-1} + eta_t arepsilon_t \ &= lpha_t ar{(ar{lpha}_{t-1} x_0 + ar{eta}_{t-1} ar{arepsilon}_t) + eta_t arepsilon_t \ &= ar{lpha}_t x_0 + lpha_t ar{eta}_{t-1} ar{arepsilon}_t + eta_t arepsilon_t \end{aligned}$$

Plug into our predict target:

^

$$L_t = rac{eta_t^2}{\sigma_t^2 lpha_t^2} \mathbb{E}\left[||arepsilon_t - arepsilon_ heta(ar{lpha}_t x_0 + lpha_t ar{eta}_{t-1} ar{arepsilon}_t + eta_t arepsilon_t, t)||^2
ight]$$

b. KL divergence

$$egin{aligned} L_t &= D_{KL}(q(x_t|x_{t+1},x_0)||p_{ heta}(x_t|x_{t+1})) \ &= rac{1}{2\sigma_t^2}\mathbb{E}\left[|| ilde{\mu}_t - \mu_{ heta}||^2
ight] \ &= rac{1}{2\sigma_t^2}\mathbb{E}\left[||rac{1}{lpha_t}(x_t - rac{eta_t^2}{ar{eta}_t}ar{arepsilon}_t) - rac{1}{lpha_t}(x_t - rac{eta_t^2}{ar{eta}_t}arepsilon_{ heta}(x_t,t))||^2
ight] \ &= rac{eta_t^4}{2\sigma_t^2ar{eta}_t^2}\mathbb{E}\left[||ar{arepsilon}_t - arepsilon_{ heta}(ar{lpha}_t x_0 + ar{eta}_tar{arepsilon}_t)||^2
ight] \end{aligned}$$

Training and sampling algorithms

Algorithm 1 Training	Algorithm 2 Sampling
1: repeat 2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 3: $t \sim \mathrm{Uniform}(\{1, \dots, T\})$ 4: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 5: Take gradient descent step on $\nabla_{\theta} \ \epsilon - \mathbf{z}_{\theta}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t) \ ^2$ 6: until converged	1: $\mathbf{x}_{T} \sim \mathcal{N}(0, \mathbf{I})$ 2: $\mathbf{for}\ t = T, \dots, 1\ \mathbf{do}$ 3: $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})\ \text{if}\ t > 1$, else $\mathbf{z} = 0$ 4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_{t}}}\left(\mathbf{x}_{t} - \frac{1-\alpha_{t}}{\sqrt{1-\bar{\alpha}_{t}}}\mathbf{z}_{\theta}(\mathbf{x}_{t}, t)\right) + \sigma_{t}\mathbf{z}$ 5: $\mathbf{end}\ \mathbf{for}$ 6: $\mathbf{return}\ \mathbf{x}_{0}$

3. Parameters Setting

• About α_t , T

$$egin{align} lpha_t^2 + eta_t^2 &= 1 \ T &= 1000 \ &lpha_t &= \sqrt{1 - rac{-0.02t}{T}} \ \log ar{lpha}_t &= rac{1}{2} \sum_{t=1}^T \log (1 - rac{0.02t}{T}) < rac{1}{2} \sum_{t=0}^T (-rac{0.02t}{T}) = -0.006(T+1) pprox e^{-5} \ & = -0.006(T+1) = -0.006(T+1) = -0.006(T+1) = 0.006(T+1) = 0.006(T$$

• About σ_t

$$\sigma_t = eta_t$$
 $\sigma_t = rac{ar{eta}_{t-1}}{ar{eta}_t}eta_t$

