Generative models: VAE and DM

$p(x, y)$, $p(x, y) \geq 0$ $\int p(x, y)\, dx\, dy = 1$

y - discr.
x - cont. $\sum_y \int p(x, y)\, dx = 1$

product rule: $p(x, y) = p(y|x)\, p(x) = p(x|y)\, p(y)$

sum rule: $\int p(x, y)\, dy = p(x)$

$p(x, y, z)$ x - observed
y - to predict
z - hidden

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{\int p(x, y, z)\, dz}{\int p(x, \hat{y}, \hat{z})\, d\hat{y}\, d\hat{z}}$$

$$\mathcal{N}(\underset{\mathbb{R}}{x} \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mathcal{N}(\underset{\mathbb{R}^d}{x} \mid \underset{\mathbb{R}^d}{\mu}, \underset{\mathbb{R}^{d\times d}}{\Sigma}) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$$\Sigma^T = \Sigma \geq 0$$

$$\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2) \Rightarrow \mathcal{N}(x \mid \mu, \Sigma) = \prod_{i=1}^{d} \mathcal{N}(x_i \mid \mu_i, \sigma_i^2)$$

$$x \sim \mathcal{N}(x \mid \mu, \sigma^2) \qquad z \sim \mathcal{N}(z \mid 0, 1),\ x = \mu + \sigma \cdot z \qquad \Sigma = L \cdot L^T$$

$$x \sim \mathcal{N}(x \mid \mu, \Sigma) \qquad z \sim \mathcal{N}(z \mid 0, I),\ x = \mu + L \cdot z \qquad L = (\Delta 0)$$

f – differentiable

$$\mathbb{E}_{N(x|\mu,\Sigma)} f(x) \to \min_{\mu,\Sigma}$$

Init. μ, L

Iter. until convergence:

$z_1, \dots, z_B \sim N(z|0, I)$

$Loss = \frac{1}{B} \sum_j f(\mu + L z_j)$

$\mu, L = \text{opt\_step}(\nabla_{\mu,L} Loss)$

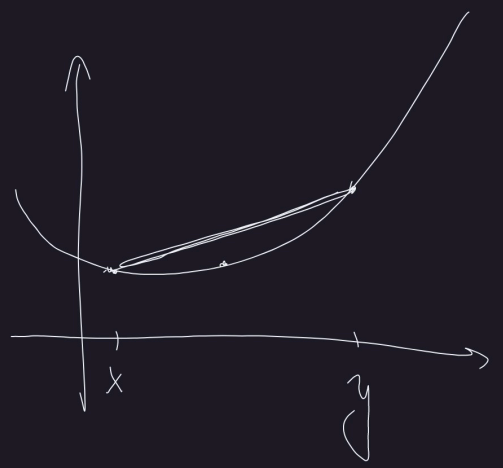
$$\nabla_\mu \mathbb{E}_{N(x|\mu,\Sigma)} f(x) = \nabla_\mu \mathbb{E}_{N(z|0,I)} f(\mu + Lz) =$$

$$= \mathbb{E}_{N(z|0,I)} \nabla_\mu f(\mu + Lz) \approx \frac{1}{B} \sum_{j=1}^{B} \nabla_\mu f(\mu + L z_j),$$

$$z_j \sim N(z|0, I)$$

f - convex $\iff$ $f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y)$

$\forall x, y \quad \forall \alpha \in (0,1)$



Jensen inequality:

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \le \sum_{i=1}^{m} \alpha_i f(x_i)$$

$$\alpha_i \ge 0 \quad \sum_{i=1}^{m} \alpha_i = 1$$

$$f\left(\int \alpha(t) x(t)\, dt\right) \le \int \alpha(t) f(x(t))\, dt$$

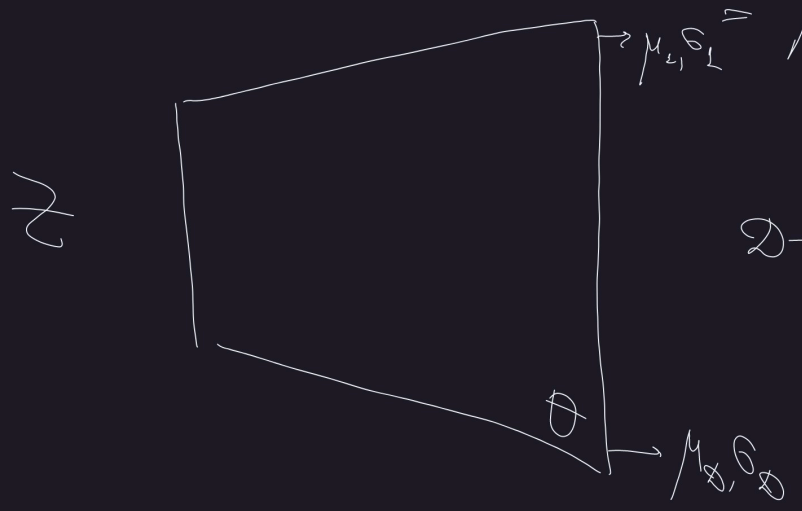
$$\alpha(t) \ge 0 \quad \int \alpha(t)\, dt = 1$$

VAE

$$z \sim p(z) = \mathcal{N}(z \mid 0, I)$$

$$x \mid z \sim p_\theta(x \mid z) = \mathcal{N}(x \mid \mu_\theta(z), \Sigma_\theta(z)) =$$

$$= \mathcal{N}(x \mid \mu_\theta(z), \mathrm{diag}(\log(1+\exp(\sigma_\theta(z)))))$$



D – # of pixels

$$q_{opt}(z) = p_\theta(z|x) = \frac{p_\theta(x,z)}{p(x)} = \frac{p(x,z)}{\int p(x,\tilde{z})d\tilde{z}}$$

$$p_\theta(x,z) = p_\theta(x|z)\cdot p(z)$$

$$\mathbb{E}_q \log \frac{p(x,z)}{p(x|z)} \cdot \frac{p(x)}{p(x,z)} = \log p(x)$$

$$\mathbb{E}_{x\sim p_{data}(x)} \log p_\theta(x) = \mathbb{E}_{x\sim p_{data}(x)} \log \underbrace{\int p_\theta(x,z)dz}_{\text{can't be computed}} \to \max_\theta$$

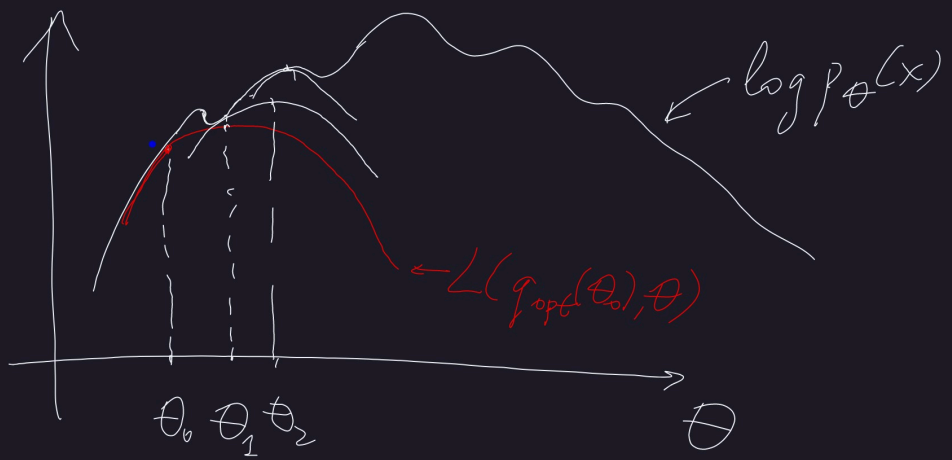
$$\log p_\theta(x) = \log \int p_\theta(x,z)dz \overset{\forall q:\, q(z)\geq 0,\ \int q(z)dz=1}{=} \log \int q(z)\cdot \frac{p_\theta(x,z)}{q(z)}dz \geq$$

Jensen inequality

$$\geq \int q(z)\log\frac{p_\theta(x,z)}{q(z)}dz = \mathbb{E}_{q(z)} \log \frac{p_\theta(x,z)}{q(z)} \to \max_{\theta,q}$$

ELBO

$$= \mathcal{L}(q,\theta)$$



$$q(z|\lambda) \simeq p(z|x)$$

$$\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(x_i) \geq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{q(z_i|x)}\log\frac{p_\theta(x_i,z_i)}{q(z_i)} =$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{q(z_i|x)}\log\frac{p_\theta(x_i|z_i)\,p(z_i)}{q(z_i|x)} =$$

① $q(z_i|x) = \mathcal{N}(z_i \mid \tilde{\mu}_i, \mathrm{diag}(\tilde{\sigma}_i))$ $\forall i = 1 \dots N$

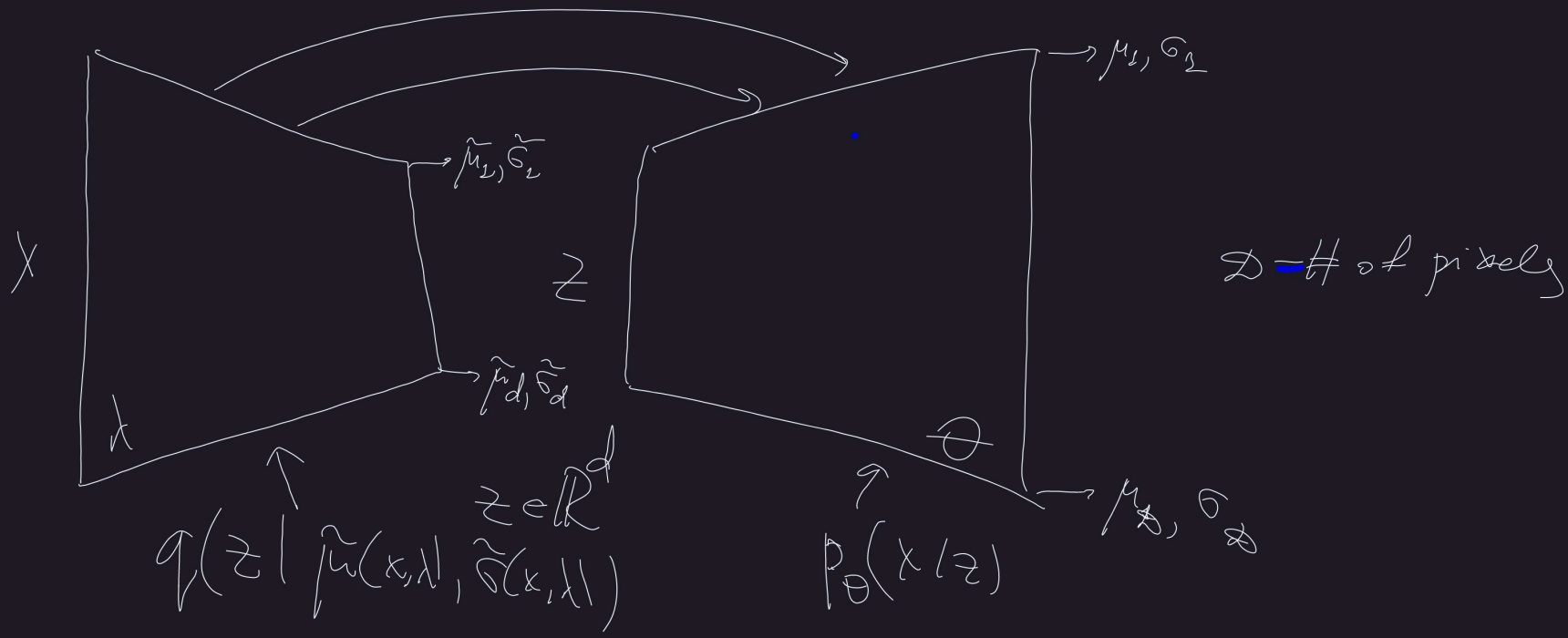
② $q(z_i|x) = \mathcal{N}(z_i \mid \tilde{\mu}(x_i,\lambda), \mathrm{diag}(\tilde{\sigma}(x_i,\lambda)))$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{q(z_i|x)}\log p_\theta(x_i|z_i) + \mathbb{E}_{q(z_i|x)}\log\frac{p(z_i)}{q(z_i)}\right] =$$

reconstruction term

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{q(z_i|x)}\log p_\theta(x_i|z_i) - KL(q(z_i|x)\,||\,p(z_i))\right] \to \max_{\theta,\lambda}$$

regularization



D = # of pixels

Very Deep VAE

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{N(\varepsilon_i|0,I)} \log p_\theta\Big(x_i \,\Big|\, \underbrace{\tilde{\mu}(x_i,\lambda) + \tilde{\sigma}(x_i,\lambda)\cdot\varepsilon_i}_{z_i}\Big) - KL\Big(\tilde{\mu}(x_i,\lambda), \tilde{\sigma}(x_i,\lambda)\Big)\Big] \to \max_{\theta,\lambda}$$

VAE training

Init. θ, λ

Iterations until convergence:

$x_1, \dots, x_B \sim p_{data}(x)$

$\varepsilon_1, \dots, \varepsilon_B \sim N(\varepsilon|0, I)$

$$Loss = -\frac{1}{B}\sum_{j=1}^{B}\Big[\log p_\theta(x_j \,|\, \tilde{\mu}(x_j,\lambda) + \tilde{\sigma}(x_j,\lambda)\odot\varepsilon_j) - KL\Big(\tilde{\mu}(x_j,\lambda), \tilde{\sigma}(x_j,\lambda)\Big)\Big]$$

$\theta, \lambda \leftarrow$ optimizer_step$\left(\nabla_{\theta,\lambda} Loss\right)$

Diffusion model

x_0 – image

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

$$q(x_1, x_2 \dots x_T | x_0) = q(x_1|x_0)\, q(x_2|x_1) \dots q(x_T|x_{T-1})$$

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \varepsilon \odot \sqrt{\beta_t}, \quad \varepsilon \sim \mathcal{N}(\varepsilon | 0, I)$$

$$Var(x_{t,i}) = (1-\beta_t) \underbrace{Var(x_{t-1,i})}_{1} + \beta_t \underbrace{Var(\varepsilon_i)}_{1} = 1$$

Linear Gaussian Model

$$x \sim \mathcal{N}(x \mid \mu, \Sigma)$$

$$y \mid x \sim \mathcal{N}(y \mid Ax, \Gamma)$$

$$x, y \sim \mathcal{N}$$

$$x \mid y \sim \mathcal{N}$$

$$y \sim \mathcal{N}(y \mid A\mu, \Gamma + A\Sigma A^T)$$

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

$$\left(q(x_t \mid x_0) = \mathcal{N}(x_t \mid \sqrt{\bar{\alpha}_t}\, x_0, (1 - \bar{\alpha}_t) I \right)$$

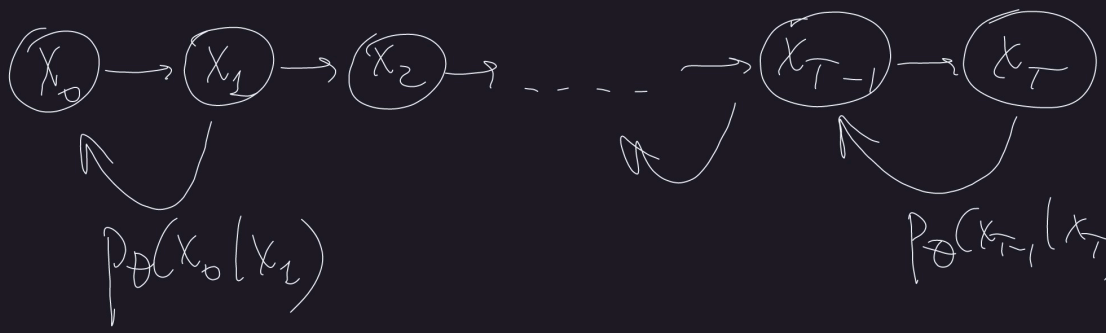
$$\bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$$

$$q(x_t \mid x_0)\, q(x_{T-1} \mid x_T, x_0)\, q(x_{T-2} \mid x_{T-1}, x_0) \cdots q(x_1 \mid x_2, x_0)$$

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

$q(x_{1:T} | x_0)$ $\quad$ $q(x_{t-1} | x_t, x_0)$ $\quad$ $\underset{\|}{N(x_T | 0, I)}$

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



$= p(x_T) \prod_{t=1}^{T} N(x_{t-1} | \mu_\theta(x_t), \Sigma_\theta(x_t))$

$p_\theta(x_0, x_1, \dots x_T)$

$$\log p_\theta(x_0) = \log \int p(x_{0:T})\, dx_{1:T} \geq \mathbb{E}_{q(x_{1:T}|x_0)} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} =$$

ELBO

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

$$= \mathbb{E}_q \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] =$$

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

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

$$\Leftarrow \mathbb{E}_q\left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \cdot \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right] =$$

$$\log q(x_{T-1}|x_0) - \log q(x_T|x_0) + \log q(x_{T-2}|x_0) - \log q(x_{T-1}|x_0) + \dots + \log q(x_1|x_0) - \log q(x_2|x_0)$$

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

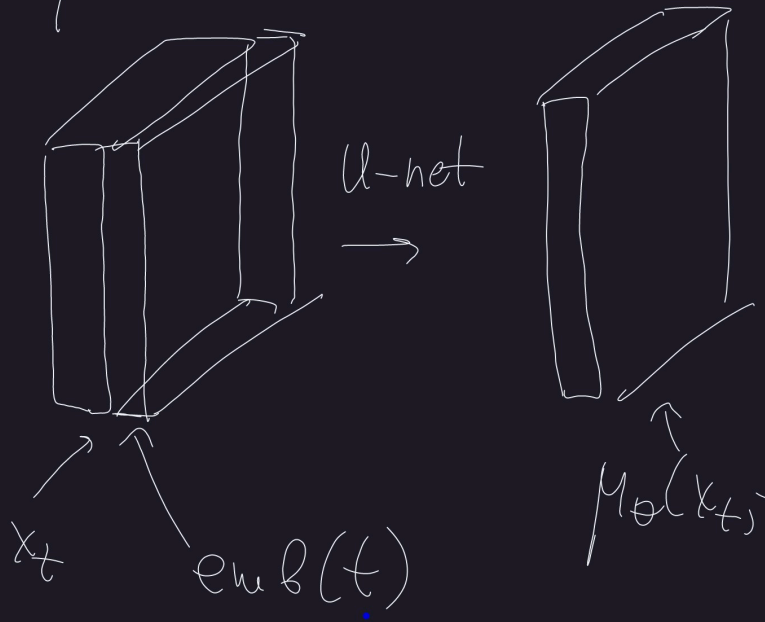
$$= -KL(q(x_T|x_0) \,||\, p(x_T)) - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} KL(q(x_{t-1}|x_t, x_0) \,||\, p_\theta(x_{t-1}|x_t)) + \mathbb{E}_{q(x_1|x_0)} \log p_\theta(x_0|x_1)$$

regul. term — const. term

$\rightarrow \max_\theta$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t,t),\sigma_t^2 I)$$

$$\sigma_t^2 = \beta_t \quad (\text{fixed!})$$



$$KL\left(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t)\right) =$$

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

$$= \text{const} + \frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t(x_t,x_0) - \mu_\theta(x_t,t)\right\|^2$$

$$\mathbb{E}_{q(x_t|x_0)} KL\ldots = \mathbb{E}_{\mathcal{N}(\varepsilon|0,I)} \frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t\left(\underbrace{\sqrt{\bar{\alpha}_t}x_0 + \varepsilon\sqrt{1-\bar{\alpha}_t}}_{=x_t(x_0,\varepsilon)}, x_0\right) - \mu_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \varepsilon\sqrt{1-\bar{\alpha}_t}, t\right)\right\|^2 \ominus$$

$$\Leftarrow \mathbb{E}_{\mathcal{N}(\varepsilon|0,I)} \frac{1}{2\sigma_t^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(x_t(x_0,\varepsilon) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon \right) - \mu_\theta(x_t(x_0,\varepsilon), t) \right\|^2 =$$

$$\mu_\theta(x_t, t) := \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon_\theta(x_t, t) \right)$$

$$= \mathbb{E}_{\mathcal{N}(\varepsilon|0,I)} \frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \left\| \varepsilon - \varepsilon_\theta\left(\underbrace{\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \cdot \varepsilon}_{x_t}, t \right) \right\|^2$$

$$L_{simple}(\theta) = \mathbb{E}_{t, x_0, \varepsilon} \left\| \varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, t \right) \right\|_2^2$$

DDPM training:

Init. θ

Iter. until convergence:

$x_0 \sim p_{data}(x_0)$

$t \sim \text{Uniform}(1, 2, \ldots, T)$

$\varepsilon \sim \mathcal{N}(\varepsilon \mid 0, I)$

$\text{Loss} = \| \varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon, t) \|^2$

$\theta = \text{opt\_step}(\nabla_\theta \text{Loss})$

Sampling:

$x_T \sim \mathcal{N}(x_T \mid 0, I)$

for $t = T, T-1, \ldots, 1$:

$z \sim \mathcal{N}(z \mid 0, I)$

$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon_\theta(x_t, t)\right) + z \cdot \sigma_t$

output: x_0