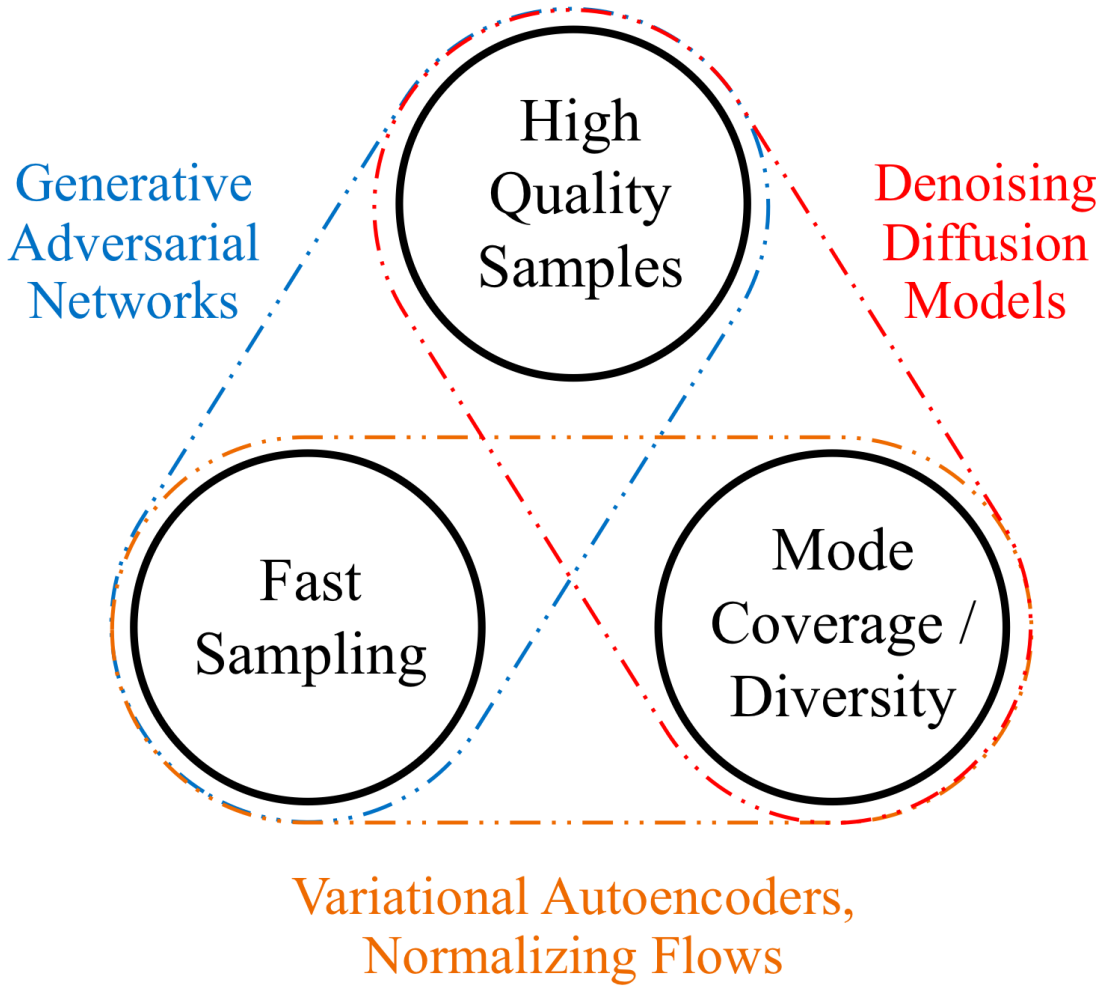
Diffusion models

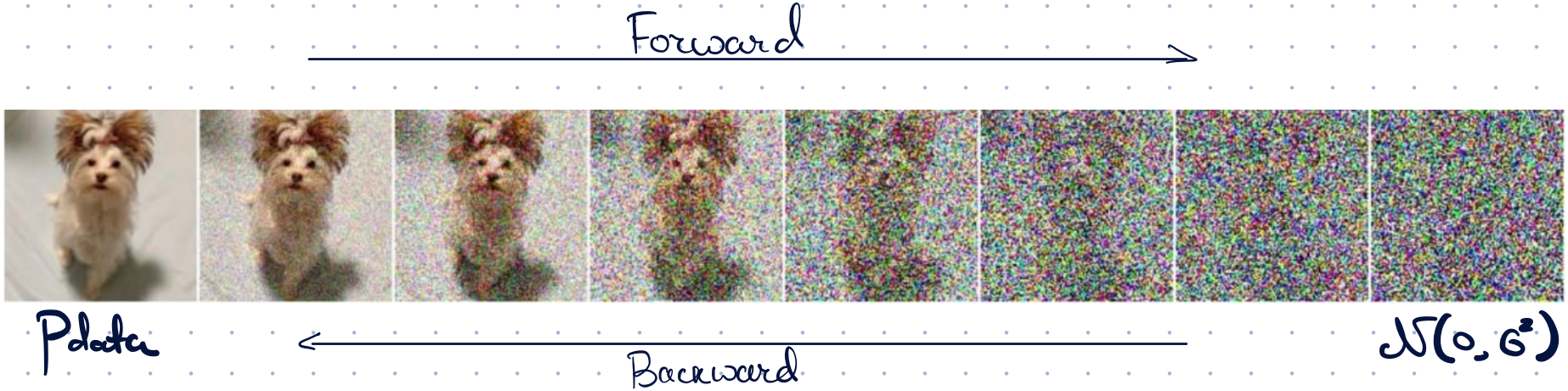
Denis Rakitin

Generative trilemma



https://arxiv.org/abs/2112.07804

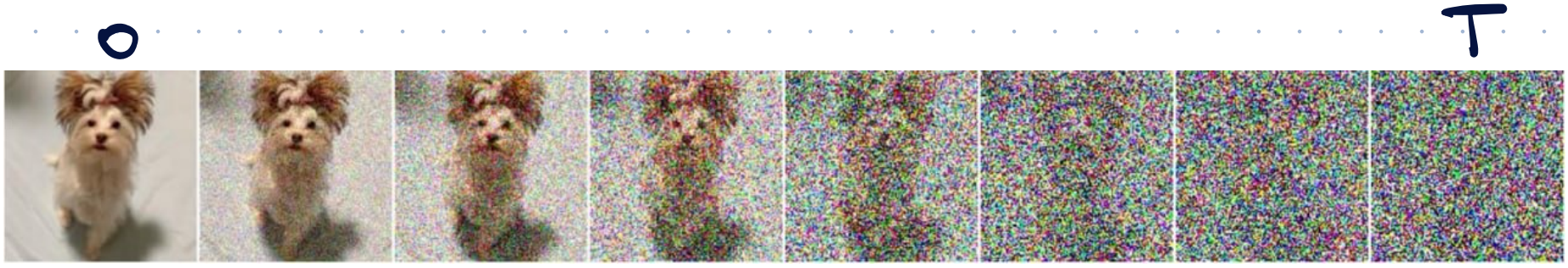
Diffusion



Idea: define forward noising process and try to reverse it.

https://arxiv.org/abs/2011.13456

Forward noising process



$x_0 \sim p_{data}$; $x_t = x_{t-1} + g(t)\varepsilon_t$, $\varepsilon_t \sim N(0, I)$
(Variance exploding) indep. с $x_0 \dots x_{t-1}$.

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \varepsilon_t, \quad \varepsilon_t \sim N(0, I)$$
indep. с $x_0 \dots x_{t-1}$.

(variance preserving, VP)

$$\mathrm{Var}\, x_t = (1-\beta_t)\mathrm{Var}\, x_{t-1} + \beta_t \cdot I$$

$\mathrm{Var} x_0 = I \Rightarrow \mathrm{Var} x_t = I \ \forall t = 0 \dots T.$

$$q_{t|t-1}(x_t|x_{t-1}) = \mathcal{N}\left(x_t \mid \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\right)$$

$$q_{t|0}(x_t|x_0) = \mathcal{N}(x_t|$$

$$x_t = c_t x_{t-1} + d_t \varepsilon_t = c_0 x_0 + \sum_{s=1}^{t} c_s \cdot \varepsilon_s$$

$$\mathbb{E}_{q_{t|0}(x_t|x_0)} x_t = \mathbb{E}\left(\sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\,\varepsilon_t\right) = \sqrt{1-\beta_t}\, \mathbb{E}_{q_{t-1|0}(x_{t-1}|x_0)} x_{t-1}$$

$$= \left(\prod_{s=1}^{t} \sqrt{1-\beta_s}\right) \cdot \mathbb{E}_{q_{0|0}(x_0|x_0)} x_0 = x_0 \cdot \prod_{s=1}^{t} \sqrt{1-\beta_s}$$

$$\underset{q_{t|0}(x_t|x_0)}{\mathrm{Var}}\, x_t = (1-\beta_t)\underset{q_{t-1|0}(x_{t-1}|x_0)}{\mathrm{Var}}\, x_{t-1} + \beta_t \cdot I$$

$$I - \underset{q_{t|0}(x_t|x_0)}{\mathrm{Var}}\, x_t = I - \beta_t I - (1-\beta_t)\underset{q_{t-1|0}(x_{t-1}|x_0)}{\mathrm{Var}}\, x_{t-1}$$

$$= (1-\beta_t)\left(I - \underset{q_{t-1|0}(x_{t-1}|x_0)}{\mathrm{Var}}\, x_{t-1}\right)$$

$$= \ldots = \left(\prod_{s=1}^{t}(1-\beta_s)\right)\cdot\left(I - \underset{q_{0|0}}{\mathrm{Var}}\, x_0\right) = \left(\prod_{s=1}^{t}(1-\beta_s)\right)\cdot I$$

$$\underset{q_{t|0}(x_t|x_0)}{\mathrm{Var}}\, x_t = I - \left(\prod_{s=1}^{t}(1-\beta_s)\right) I = \left(1 - \prod_{s=1}^{t}(1-\beta_s)\right)\cdot I$$

$$\underset{q_{t|0}(x_t|x_0)}{\mathbb{E}}\, x_t = \left(\prod_{s=1}^{t}\sqrt{1-\beta_s}\right)\cdot x_0 \qquad \underset{q_{t|0}(x_t|x_0)}{\mathrm{Var}}\, x_t = \left(1-\prod_{s=1}^{t}(1-\beta_s)\right) I$$

$$\prod_{s=1}^{t}\sqrt{1-\beta_s} = \sqrt{\alpha_t}\cdot x_0 \qquad \prod_{s=1}^{t}(1-\beta_s) = (1-\alpha_t)\cdot I$$

$$\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$$

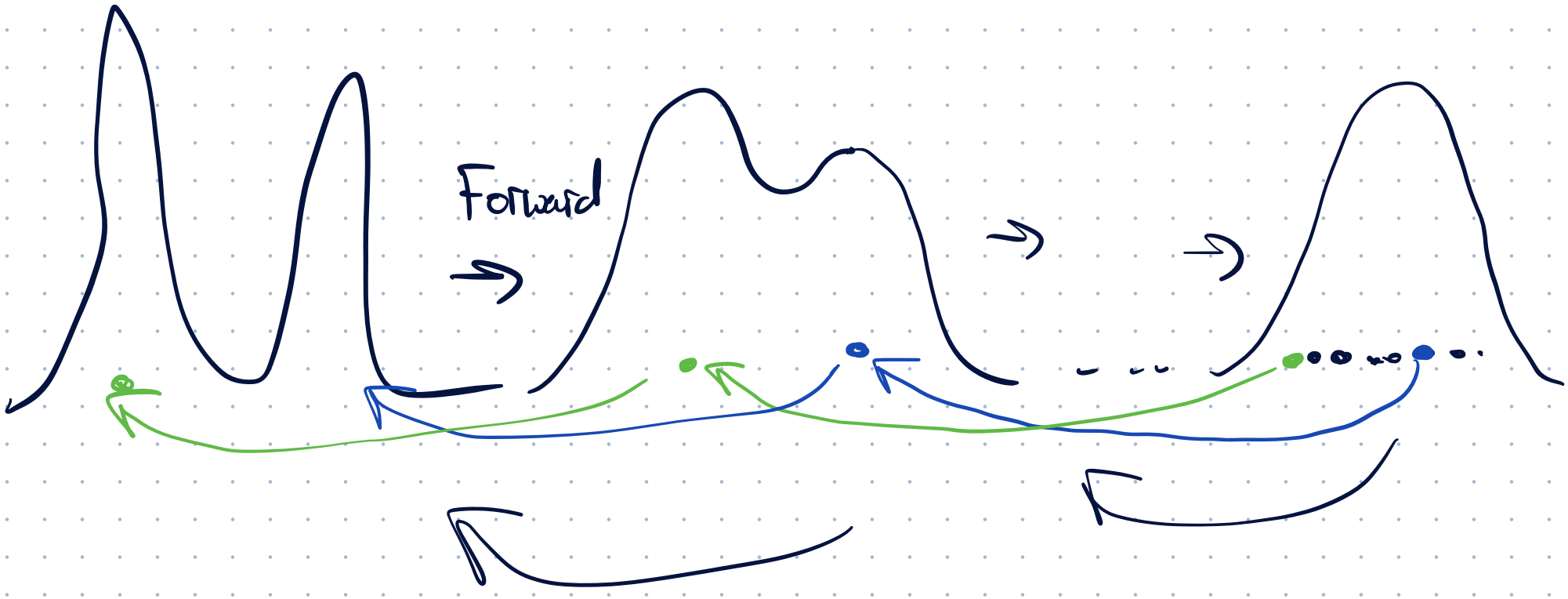
$$q_{t|0}(x_t|x_0) = \mathcal{N}\left(x_t \mid \sqrt{\alpha_t}\, x_0,\ (1-\alpha_t) I\right)$$

$$q_{T|0}(x_T|x_0) \approx \mathcal{N}(x_T \mid 0, I)$$

(если T, β_t : $\prod_{s=1}^{T}(1-\beta_s) \approx 0$)

$$q_T(x_T) = \int \underbrace{q_{T|0}(x_T|x_0)\, q_0(x_0)}_{q_{0,T}(x_0, x_T)}\, dx_0 \approx$$

$$\approx \int \mathcal{N}(x_T | 0, I)\, q_0(x_0)\, dx_0 = \mathcal{N}(x_T | 0, I).$$



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

$$q_{t|0}(x_t|x_0) = \mathcal{N}(x_t|\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$$

$$q_{t-1|t} \quad (x_T \sim \mathcal{N}(0, I) \; x_{t-1} \sim q_{t-1|t}(x_{t-1}|x_t))$$

просто

$$q_{t-1|t}(x_{t-1}|x_t) = \frac{q_{t|t-1}(x_t|x_{t-1})\, q_{t-1}(x_{t-1})}{q_t(x_t)}$$

$$q_t(x_t) = \int q_{t|0}(x_t|x_0)\, p_{data}(x_0)\, dx_0$$

просто

сложно

$q_{t-1|t}(x_{t-1}|x_t)$ — сложное расп. => моделируем нейросетью

θ <- нар-ры нейросети

$$p^{\theta}_{t-1|t}(x_{t-1}|x_t) = \mathcal{N}(x_{t-1} \mid \mu^{\theta}_t(x_t), \sigma_t^2 I)$$

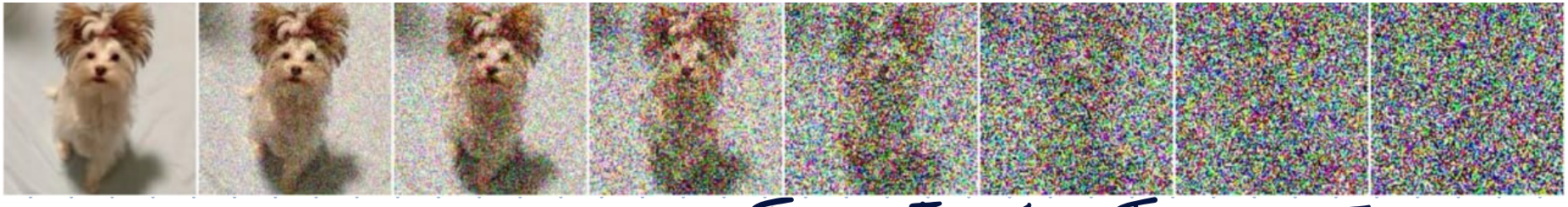
↑ нейронка

$\forall t$

$$x_T \sim \mathcal{N}(0, I) \qquad x_{t-1} = \mu^{\theta}_t(x_t) + \varepsilon_t \cdot \sigma_t$$

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \varepsilon_t$$

$x_0 \sim p_{data}$



$$x_{t-1} = \mu^{\theta}_t(x_t) + \sigma_t \cdot \hat{\varepsilon}_t$$

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

Forward: $q_{0..T}(x_0..x_T) = q_0(x_0) \prod_{t=1}^{T} q_{t|t-1}(x_t|x_{t-1})$

$q_0(x_0) = p_{data}(x_0)$

↖ марк. цепь

Backward $p^{\theta}_{0..T}(x_0..x_T) = \mathcal{N}(x_T|0,I) \prod_{t=1}^{T} p^{\theta}_{t-1|t}(x_{t-1}|x_t)$

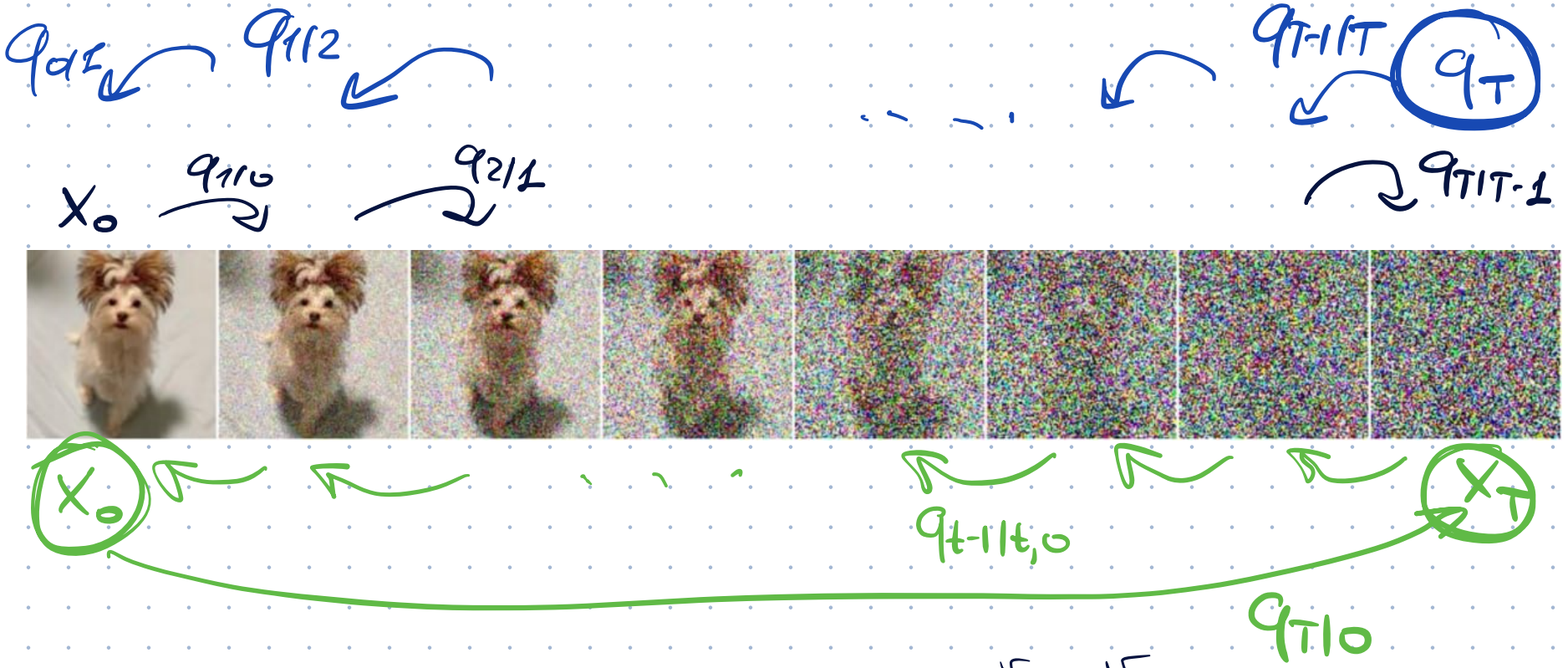
↖ марк. цепь

$$KL\left(q_{0..T}(x_0..x_T) \,\|\, p^{\theta}_{0..T}(x_0..x_T)\right) \to \min_{\theta}$$

$=$

$$\mathbb{E}_{q_{0..T}(x_0..x_T)} \log \frac{q_{0..T}(x_0..x_T)}{p^{\theta}_{0..T}(x_0..x_T)}$$

Model and Training



$$q_{t-1|t} = \frac{q_{t|t-1}\, q_{t-1}}{q_t} \qquad q_{t-1|t,0} = \frac{q_{t|t-1}\, q_{t-1|0}}{q_{t|0}}$$

$$KL\left(q_{0..T}(x_0..x_T) \,\|\, p^{\theta}_{0..T}(x_0..x_T)\right) \to \min_{\theta}$$

$$\mathop{\mathbb{E}}_{q_{0..T}(x_0..x_T)} \log \frac{q_{0..T}(x_0..x_T)}{p^{\theta}_{0..T}(x_0..x_T)} = \mathop{\mathbb{E}}_{q_{0..T}} \log \frac{q_T(x_T)\prod_{t=1}^{T} q_{t-1|t}(x_{t-1}|x_t)}{\mathcal{N}(x_T|0,I)\prod_{t=1}^{T} p^{\theta}_{t-1|t}(x_{t-1}|x_t)}$$

не зависит

$$= \mathop{\mathbb{E}}_{q_{0..T}} \log \frac{q_0(x_0)\, q_{T|0}(x_T|x_0)\prod_{t=2}^{T} q_{t-1|t,0}(x_{t-1}|x_t,x_0)}{\mathcal{N}(x_T|0,I)\, p^{\theta}_{0|1}(x_0|x_1)\prod_{t=2}^{T} p^{\theta}_{t-1|t}(x_{t-1}|x_t)}$$

$\hookrightarrow \min_{\theta}$

$$p^{\theta}_{0|1}(x_0|x_1) \approx \mathcal{S}_{x_1}(x_0)$$

$$\approx \mathop{\mathbb{E}}_{q_{0\ldots T}(x_0 \ldots x_T)} \log \frac{\prod_{t=2}^{T} q_{t-1|t,0}(x_{t-1}|x_t, x_0)}{\prod_{t=2}^{T} p^{\theta}_{t-1|t}(x_{t-1}|x_t)} =$$

$$= \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_{0\ldots T}(x_0 \ldots x_T)} \log \frac{q_{t-1|t,0}(x_{t-1}|x_t, x_0)}{p^{\theta}_{t-1|t}(x_{t-1}|x_t)} =$$

$$= \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_{0,t-1,t}(x_0, x_{t-1}, x_t)} \log \frac{q_{t-1|t,0}(x_{t-1}|x_t, x_0)}{p^{\theta}_{t-1|t}(x_{t-1}|x_t)} =$$

$$= \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_{0,t}(x_0, x_t)} \mathop{\mathbb{E}}_{q_{t-1|0,t}(x_{t-1}|x_0, x_t)} \log \frac{q_{t-1|t,0}(x_{t-1}|x_t, x_0)}{p^{\theta}_{t-1|t}(x_{t-1}|x_t)} =$$

$$= \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_{0,t}(x_0, x_t)} KL\left(q_{t-1|t,0}(x_{t-1}|x_t, x_0) \,\|\, p^{\theta}_{t-1|t}(x_{t-1}|x_t)\right)$$

$$p^{\theta}_{t-1|t}(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu^{\theta}_t(x_t), \gamma_t^2 I) \Leftarrow$$

$$q_{t-1|t,0}(x_{t-1}|x_t, x_0) = \frac{q_{t|t-1}(x_t|x_{t-1})\, q_{t-1|0}(x_{t-1}|x_0)}{q_{t|0}(x_t|x_0)}$$

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

$$q_{t|0} = \mathcal{N}(x_t|\sqrt{\alpha_t}\, x_0, (1-\alpha_t) I)$$

$$q_{t-1|0} = \mathcal{N}(x_{t-1}|\sqrt{\alpha_{t-1}}\, x_0, (1-\alpha_{t-1}) I)$$

$$\mathcal{N}(z|\mu, \sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|z-\mu\|^2\right)$$

$$= \text{const} \cdot \exp(-k b \cdot \varphi - e(z, \mu))$$

$$q_{t-1|t,0}(x_{t-1}|x_t, x_0) =$$

$$\frac{\exp(-kb.(x_{t-1}, x_t))\exp(-kb.(x_{t-1}, x_0))}{\exp(-kb(x_t, x_0))} \cdot \text{const}$$

$$= \text{const} \cdot \exp(-kb.(x_{t-1}, x_t, x_0))$$

$$= \mathcal{N}(x_{t-1}|A_t x_t + B_t x_0, C_t^2 I).$$

$$= \sum_{t=2}^{T} \mathbb{E}_{q_{0,t}(x_0, x_t)} KL\left(q_{t-1|t,0}(x_{t-1} | x_t, x_0) \,\|\, p^{\theta}_{t-1|t}(x_{t-1}|x_t) \right) =$$

$$p^{\theta}_{t-1|t}(x_{t-1}|x_t) = \mathcal{N}(x_{t-1} | \mu^{\theta}_t(x_t), C_t^2 I)$$

$$q_{t-1|t,0}(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1} | A_t x_t + B_t x_0, C_t^2 I\right).$$

$$= \sum_{t=2}^{T} \mathbb{E}_{q_{0,t}(x_0, x_t)} w_t \left\| \mu^{\theta}_t(x_t) - (A_t x_t + B_t x_0) \right\|^2 =$$

$$\mu^{\theta}_t(x_t) = A_t x_t + B_t D^{\theta}_t(x_t)$$

$$= \sum_{t=2}^{T} \mathbb{E}_{q_{0,t}(x_0, x_t)} \hat{w}_t \left\| D^{\theta}_t(x_t) - x_0 \right\|^2 \longrightarrow \min_{\theta}$$

$$q_{0,t}(x_0, x_t) = p_{data}(x_0)\, q_{t|0}(x_t|x_0)$$

Шаг обучения : batch_size B

- $X_0^{(1)} \dots X_0^{(B)} \sim p_{data}$
- $t^{(1)} \dots t^{(B)} \sim$ семплим мом. времени (напр. $U\{2, \dots, T\}$)
- $\varepsilon^{(1)} \dots \varepsilon^{(B)}$ — нез. $\mathcal{N}(0, I)$
- $\alpha^{(1)} \dots \alpha^{(B)}$: $\alpha^{(i)} = \alpha_{t^{(i)}} = \prod_{s=1}^{t^{(i)}} (1 - \beta_s)$
- $X_{noisy}^{(i)} = \sqrt{\alpha^{(i)}}\, X_0^{(i)} + \sqrt{1-\alpha^{(i)}}\, \varepsilon^{(i)}$.
- $\frac{1}{B} \sum_{i=1}^{B} \| D_{t^{(i)}}^{\theta}(X_{noisy}^{(i)}) - X_0^{(i)} \|^2$. backward()

Sampling

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

$$X_{t-1} \sim p^{\theta}_{t-1|t}(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu^{\theta}_t(x_t), C_t^2 I)$$

$$= A_t x_t + B_t D^{\theta}_t(x_t)$$

$$X_{t-1} = A_t X_t + B_t D^{\theta}_t(x_t) + C_t \varepsilon_t$$

DDPM

$$\sum_{t=2}^{T} \omega_t \mathop{\mathbb{E}}_{q_{0,t}(x_0,x_t)} \| D_t^\theta(x_t) - x_0 \|^2 =$$

$$q_{0,t}(x_0,x_t) = q_0(x_0)\, \mathcal{N}(\varepsilon|0,I)$$

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \varepsilon$$

$$x_t =$$

$$= \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_0(x_0)\mathcal{N}(\varepsilon|0,I)} \omega_t \| D_t^\theta(\sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \varepsilon) - x_0 \|^2 =$$

$$= \sum_{t} \mathop{\mathbb{E}}_{q_0(x_0)\mathcal{N}(\varepsilon|0,I)} \omega_t \left\| D_t^\theta(x_t) - \frac{x_t - \sqrt{1-\alpha_t}\,\varepsilon}{\sqrt{\alpha_t}} \right\|^2 =$$

$$D_t^\theta(x_t) = \frac{x_t - \sqrt{1-\alpha_t}\, \varepsilon_t^\theta(x_t)}{\sqrt{\alpha_t}}$$

$$\ominus \sum_t \mathbb{E}_{q_0(x_0)\mathcal{N}(\varepsilon|0,I)} \frac{1}{\omega_t} \left\| \varepsilon_t^\theta\left(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\,\varepsilon\right) - \varepsilon \right\|^2$$

$$L_{simple} = \sum_{t=1}^{T} \mathbb{E}_{q_0(x_0)\mathcal{N}(\varepsilon|0,I)} \left\| \varepsilon_t^\theta\left(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\,\varepsilon\right) - \varepsilon \right\|^2 .$$

Forward transitions

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

Show $q_{t|0}(x_t|x_0) = \mathcal{N}(x_t|\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$,

$$\alpha_t = \prod_{s=1}^{t}(1-\beta_s).$$

ε - prediction

Let $\sum_{t=1}^{T} \omega_t \mathbb{E}_{q_{0,t}(x_0, x_t)} \| D_t^\theta(x_t) - x_0 \|^2 \to \min_\theta$

be the training functional of DDPM.

Show that it is equivalent to

$$\sum_{t=1}^{T} \hat{\omega}_t \mathbb{E}_{q_0(x_0)\mathcal{N}(\varepsilon|0,I)} \| \varepsilon_t^\theta(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\varepsilon) - \varepsilon \|^2 \to \min_\theta$$

KL between Gaussians

Let $x \sim \underbrace{\mathcal{N}(\mu_1, \sigma_1^2)}_{p_1}$; $y \sim \underbrace{\mathcal{N}(\mu_2, \sigma_2^2)}_{p_2}$; $x, y \in \mathbb{R}$.

Show that $KL(p_1 \| p_2) = \frac{1}{2}\left(\log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} - 1\right)$.

In particular, if $\hat{x} \sim \mathcal{N}(\mu_1, \sigma_1^2 I)$, $\hat{y} \sim \mathcal{N}(\mu_2, \sigma_2^2 I)$,

then $KL(p_1 \| p_2) = C \cdot \|\mu_1 - \mu_2\|^2 + f(\sigma_1, \sigma_2)$.