Introduction to Diffusion

김정훈 2023.07.07

Contents

Generative Models

Diffusion Models

“Diffusion Models Beat GANs on Image Synthesis”
DharIwal & Nichol, OpenAI, 2021

“Cascaded Diffusion Models for High Fidelity Image Generation”
Ho et al., Google, 2021

Generative Models

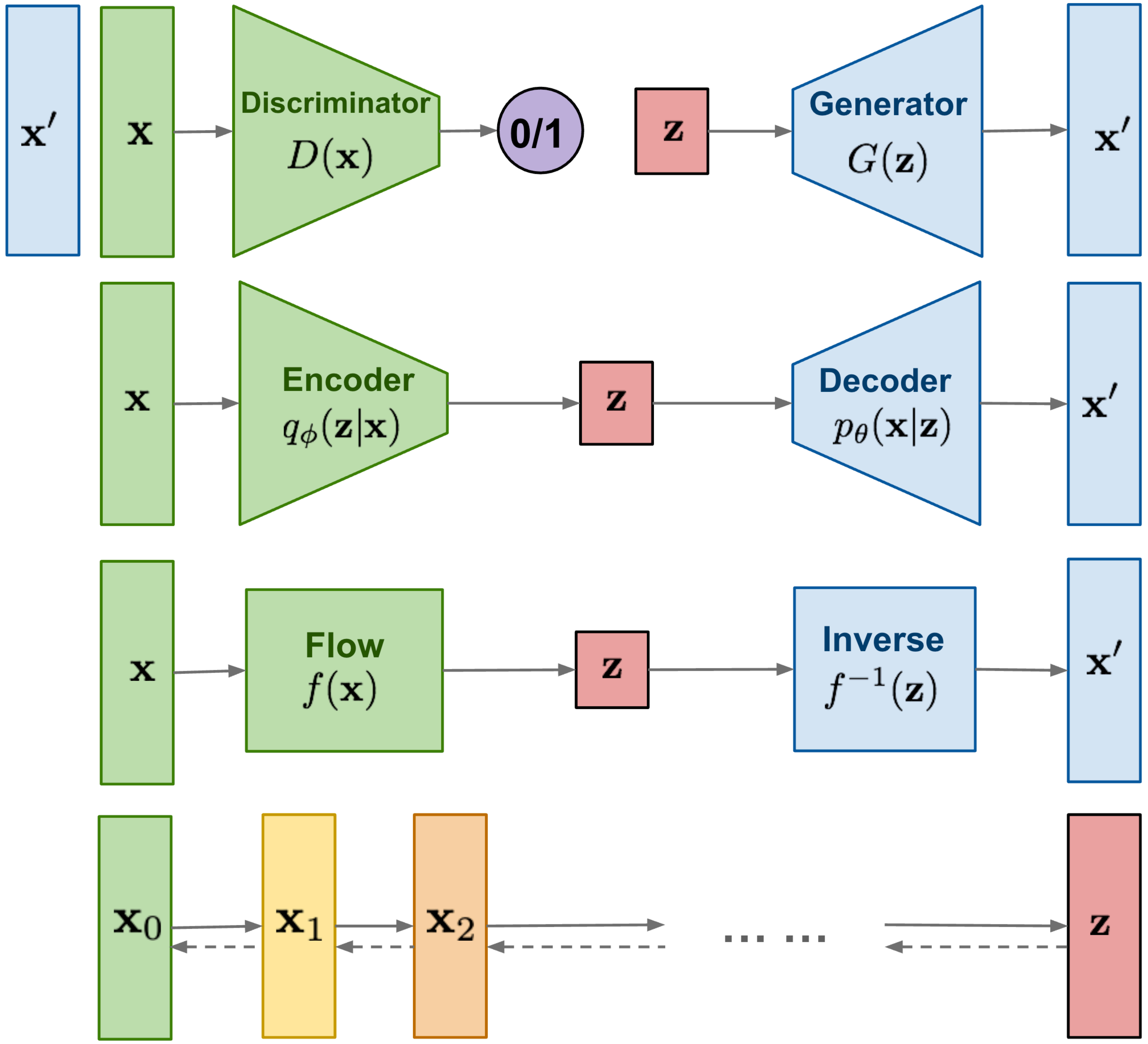
GAN, VAE, Flow-based, Diffusion

GAN: Adversarial training

VAE: maximize variational lower bound

Flow-based models: Invertible transform of distributions

Diffusion models: Gradually add Gaussian noise and then reverse



What is Diffusion?

The donut-shaped smoke will gradually spread out and become even.

What if you could get it back to how it was before it spread out evenly?

Let's use it for deep learning

DDPM

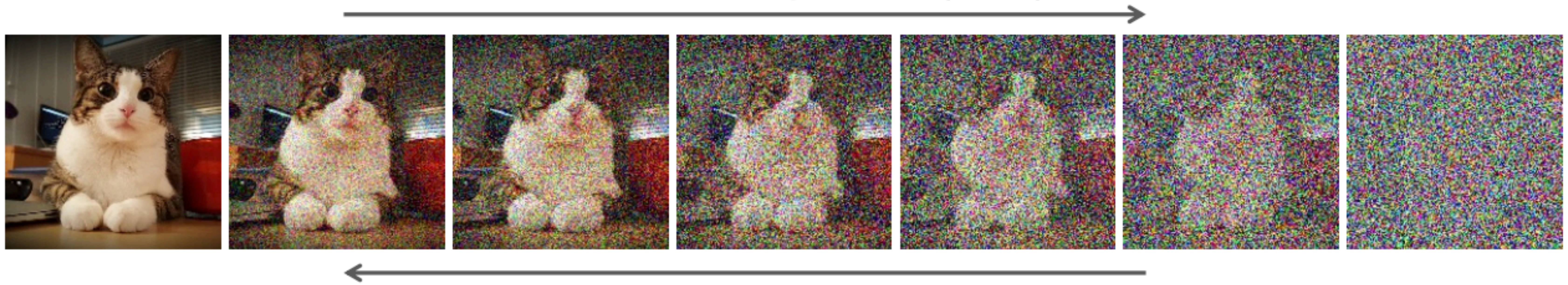
Denoising Diffusion Probabilistic Models

Denoising diffusion models consist of two processes:

- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising

Forward diffusion process (fixed)

Data

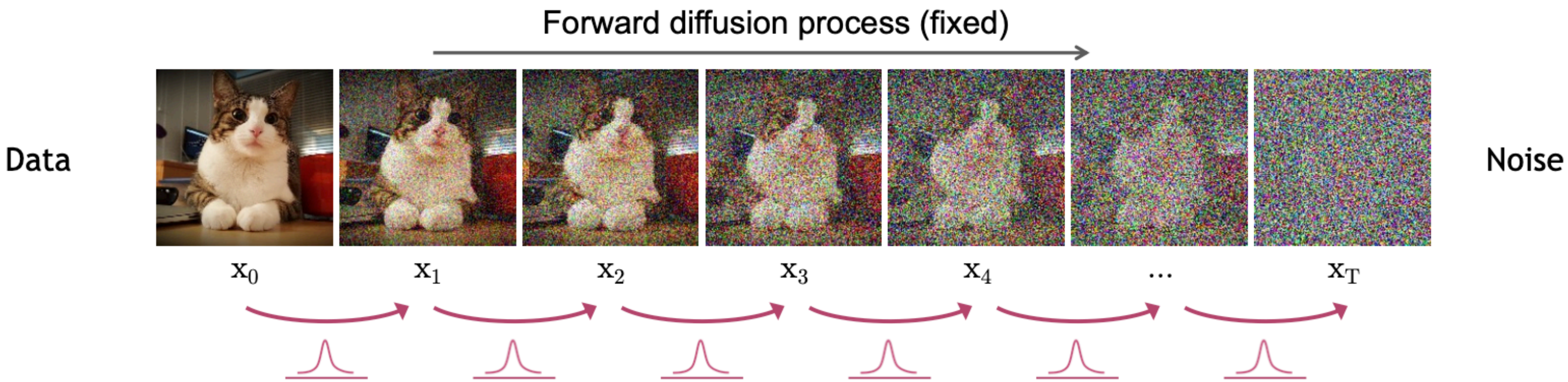


Noise

Reverse denoising process (generative)

Forward Diffusion Process

The formal definition of the forward process in T steps:

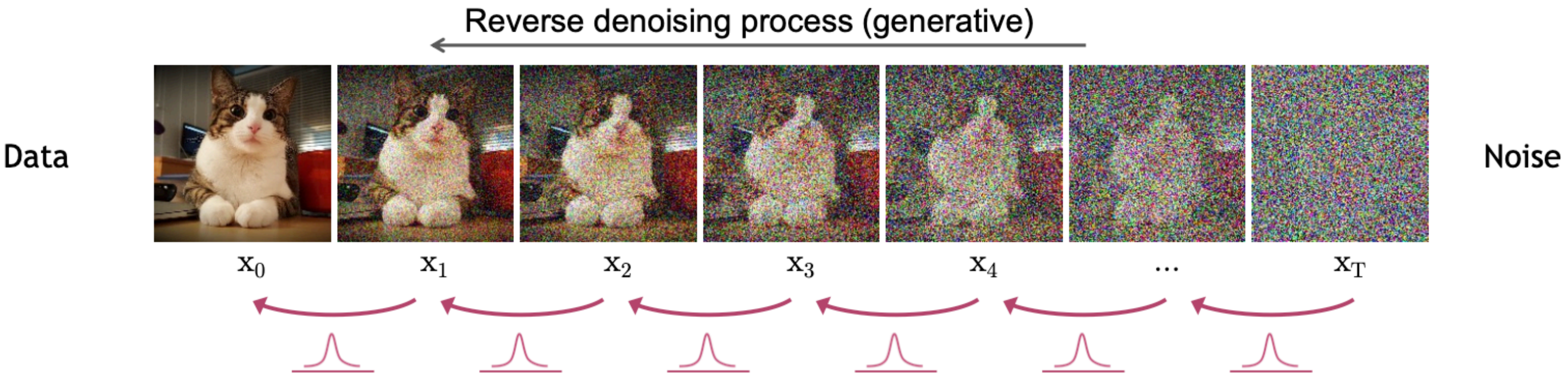


$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I}) \quad \Rightarrow \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

(joint)

Reverse Denoising Process

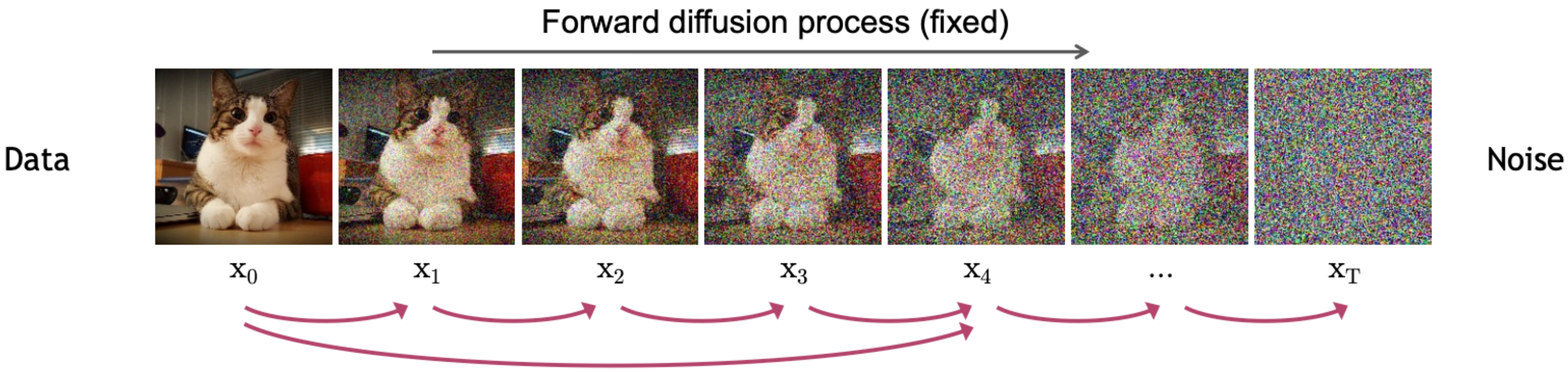
Formal definition of forward and reverse processes in T steps:



$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t)}_{\text{Trainable network}}, \sigma_t^2\mathbf{I}) \quad \Rightarrow \quad p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

Diffusion Kernel



Define $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ ➡ $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}))$ (Diffusion Kernel)

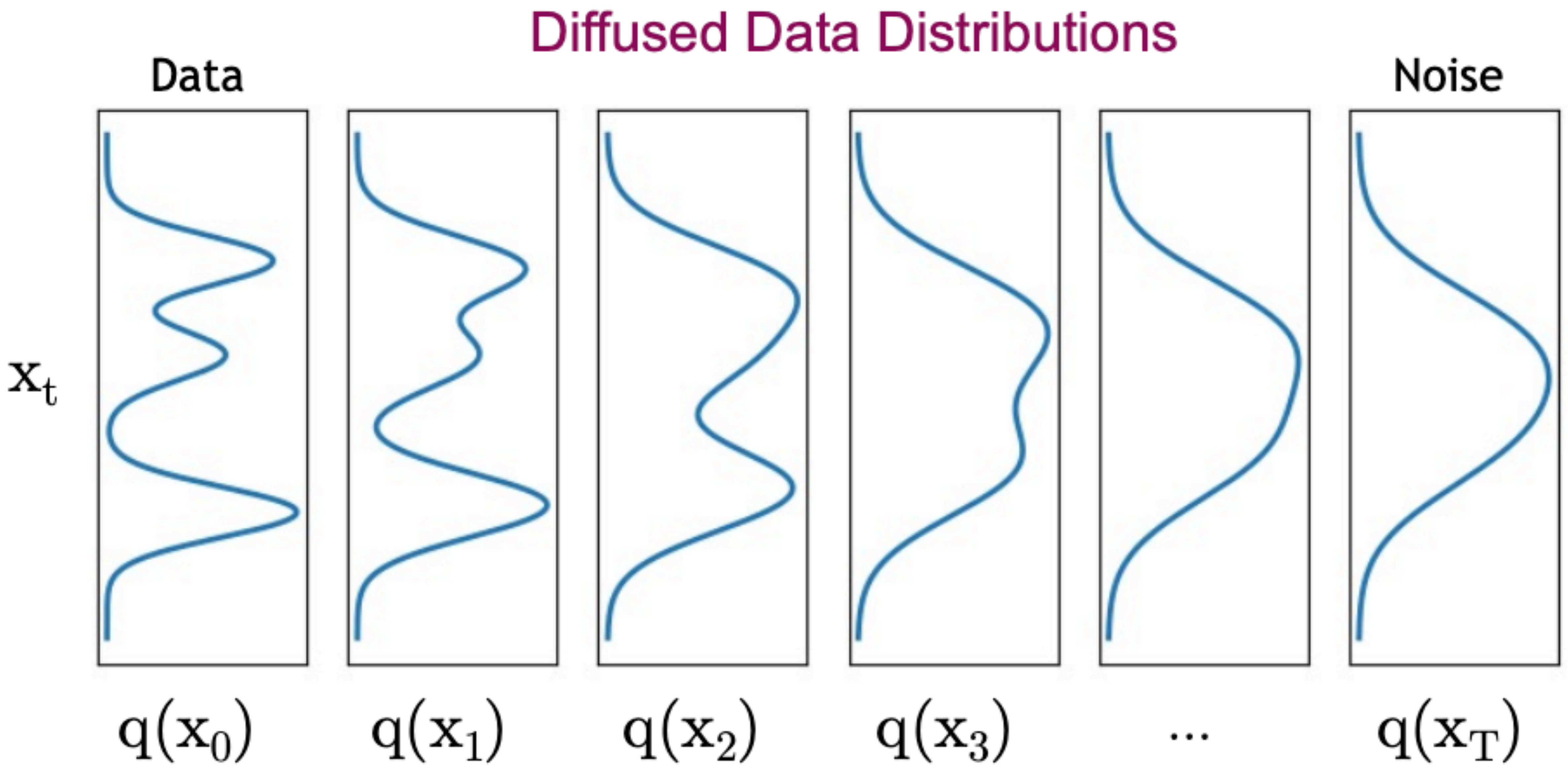
For sampling: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\ \mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\ \epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

β_t values schedule (i.e., the noise schedule) is designed such that $\bar{\alpha}_T \rightarrow 0$ and $q(\mathbf{x}_T|\mathbf{x}_0) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}))$

Diffusion Kernel

So far, we discussed the diffusion kernel $q(\mathbf{x}_t|\mathbf{x}_0)$ but what about $q(\mathbf{x}_t)$?

$$\underbrace{q(\mathbf{x}_t)}_{\substack{\text{Diffused}\\\text{data dist.}}} = \int \underbrace{q(\mathbf{x}_0, \mathbf{x}_t)}_{\substack{\text{Joint}\\\text{dist.}}}\, d\mathbf{x}_0 = \int \underbrace{q(\mathbf{x}_0)}_{\substack{\text{Input}\\\text{data dist.}}}\, \underbrace{q(\mathbf{x}_t|\mathbf{x}_0)}_{\substack{\text{Diffusion}\\\text{kernel}}}\, d\mathbf{x}_0$$

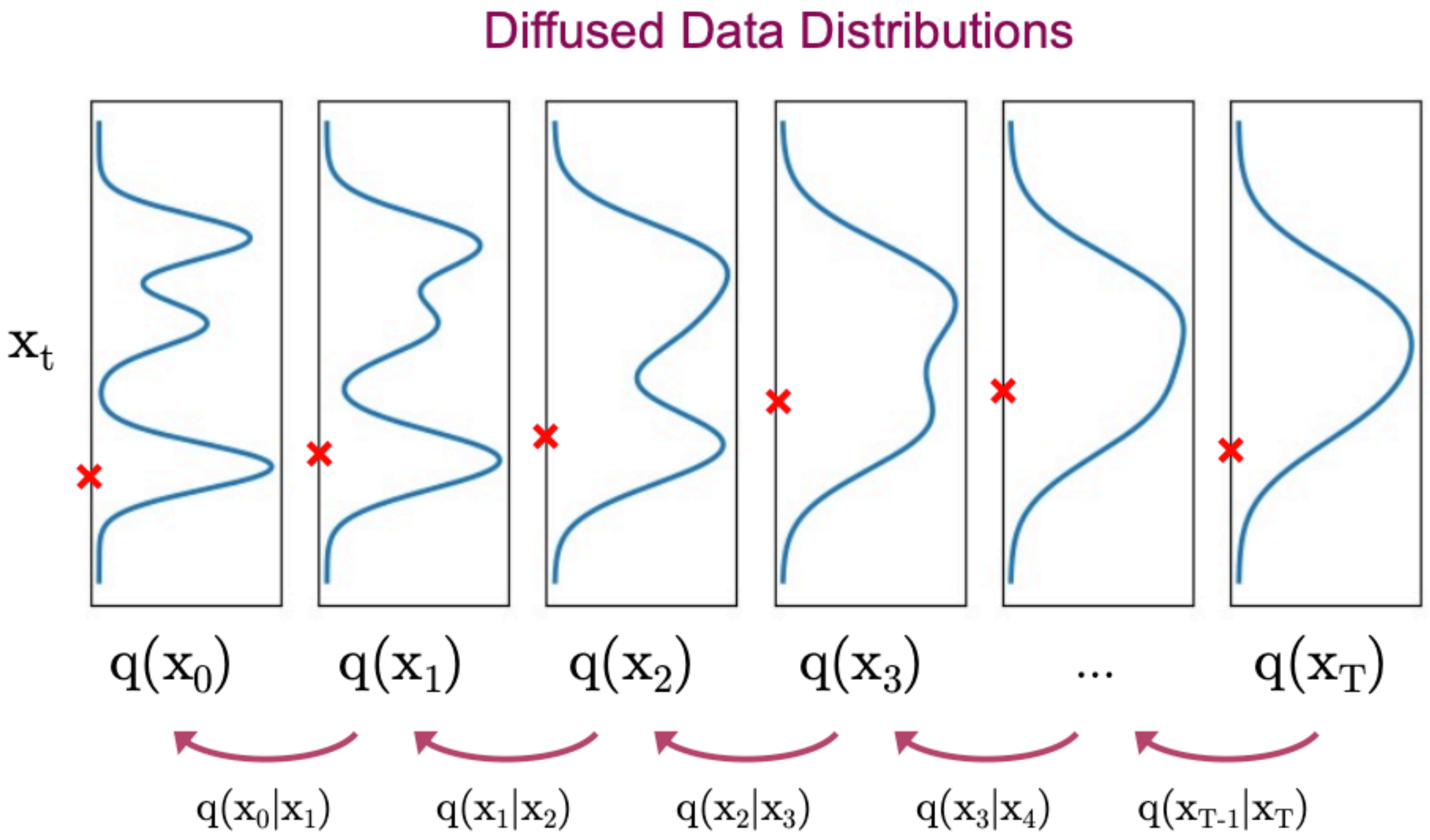


The diffusion kernel is Gaussian convolution.

We can sample $\mathbf{x}_t \sim q(\mathbf{x}_t)$ by first sampling $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and then sampling $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$ (i.e., ancestral sampling).

Generative Learning by Denoising

Recall, that the diffusion parameters are designed such that $q(\mathbf{x}_T) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}))$



Generation:

Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$

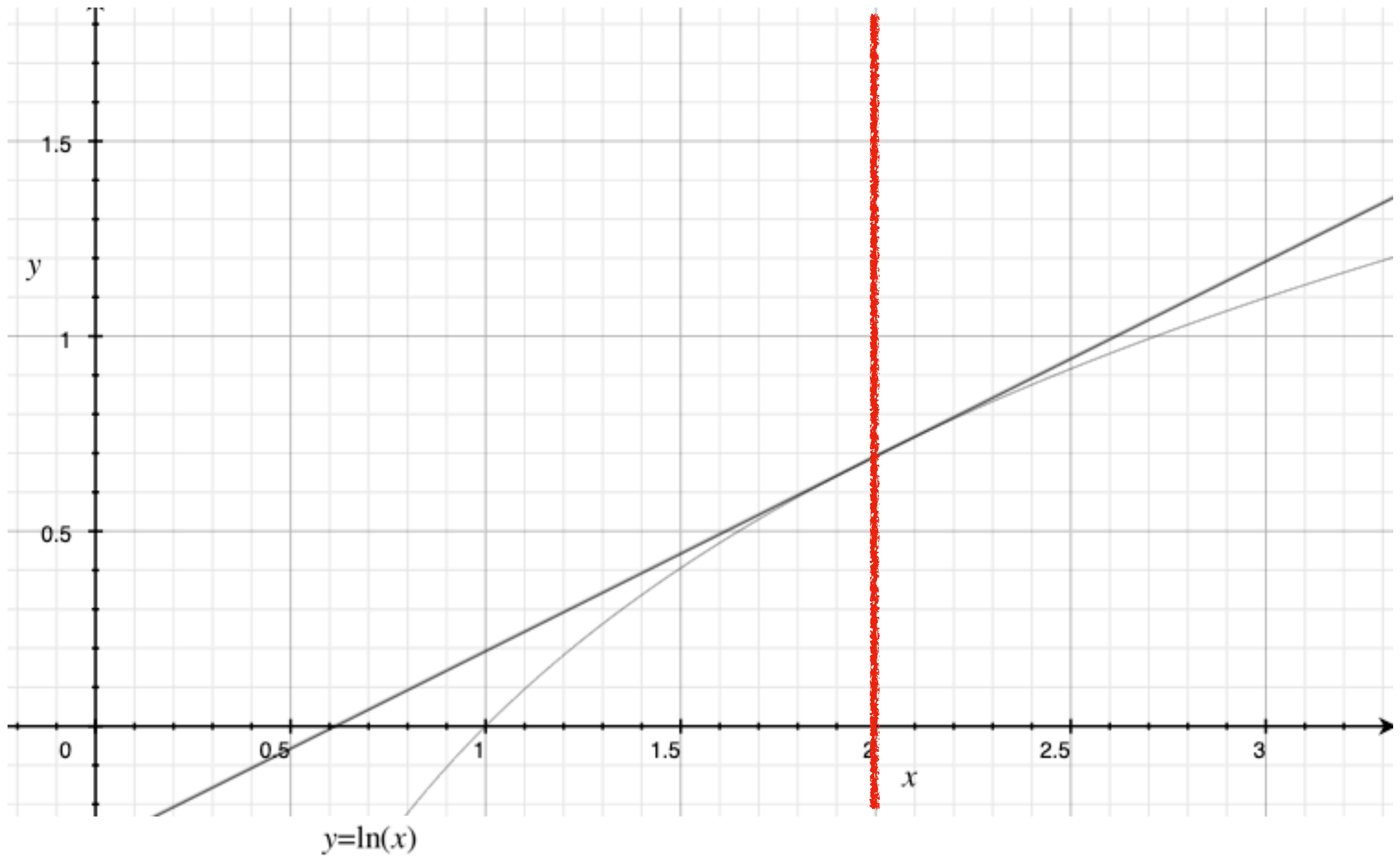
Iteratively sample $\mathbf{x}_{t-1} \sim \underbrace{q(\mathbf{x}_{t-1}|\mathbf{x}_t)}_{\text{True Denoising Dist.}}$

In general, $q(\mathbf{x}_{t-1}|\mathbf{x}_t) \propto q(\mathbf{x}_{t-1})q(\mathbf{x}_t|\mathbf{x}_{t-1})$ is intractable.

Can we approximate $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$? Yes, we can use a Normal distribution if β_t is small in each forward diffusion step.

Variational Inference

Approximate Intractable function



$g(x) = \log(x) \qquad Assume\ it's\ intractable$

$Approximation\ with\ a\ tangent\ linear\ function$

$\Rightarrow\ f(x) = \lambda x + b$

$$Let \quad f^*(\lambda) = \min_x\{\lambda x - f(x)\}\ , \quad for\ given\ \lambda$$

$$Then \quad \lambda x - f^*(\lambda) \geq g(x)\ , \quad for\ all\ \lambda, x$$

$$Let \quad J(x,\lambda) = \lambda x - f^*(\lambda) \quad and$$

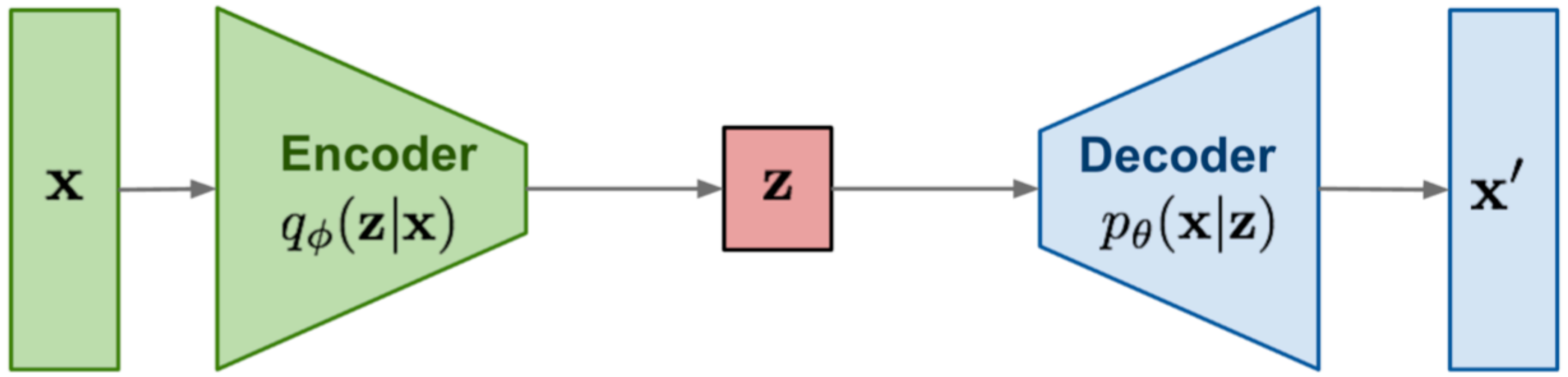
$$\lambda_0 = \arg\min_\lambda\{J(x_0,\lambda)\},\ for\ given\ x_0$$

$$Then \quad \log(x) \approx J(x,\lambda_0) \qquad for\ x\ adjacent\ to\ x_0$$

Variational AutoEncoder

Evidence Lower BOund

$$
\begin{aligned}
& -\log\left(p(\mathbf{x})\right) \\
&= -\log\left(p(\mathbf{x})\right)\int_{-\infty}^{\infty} q(\mathbf{z}|\mathbf{x})\mathbf{dz} && \because \int_{-\infty}^{\infty} q(\mathbf{z}|\mathbf{x})\mathbf{dz} = 1 \\
&= -\int_{-\infty}^{\infty} \log\left(p(\mathbf{x})\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} && \because p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})} \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x},\mathbf{z})\; q(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x})\; p(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} - \int_{-\infty}^{\infty} \log\left(\frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} \\
&\leq -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} && \because D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \geq 0 \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} \\
&= -\int_{-\infty}^{\infty} \log\left(p(\mathbf{x}|\mathbf{z})\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} - \int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}|\mathbf{x})\mathbf{dz} \\
&= -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}\left[\log\left(p(\mathbf{x}|\mathbf{z})\right)\right] - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right)\right] && \because \mathit{definition\ of\ expectation}
\end{aligned}
$$



Maximize ELBO

Instead of intractable log-likelihood log(p(x))

DDPM

ELBO for DDPM

$$
\begin{aligned}
& -\log\left(p(\mathbf{x}_0)\right) \\
&= -\log\left(p(\mathbf{x}_0)\right)\int_{-\infty}^{\infty} q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \qquad \because \int_{-\infty}^{\infty} q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t = 1 \\
&= -\int_{-\infty}^{\infty} \log\left(p(\mathbf{x}_0)\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x}_0,\mathbf{x}_t)}{p(\mathbf{x}_t|\mathbf{x}_0)}\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \qquad \because p(\mathbf{x}_t|\mathbf{x}_0) = \frac{p(\mathbf{x}_0,\mathbf{x}_t)}{p(\mathbf{x}_0)} \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x}_0,\mathbf{x}_t)\; q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)\; p(\mathbf{x}_t|\mathbf{x}_0)}\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x}_0,\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t - \int_{-\infty}^{\infty} \log\left(\frac{q(\mathbf{x}_t|\mathbf{x}_0)}{p(\mathbf{x}_t|\mathbf{x}_0)}\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \\
&\leq -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x}_0,\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \qquad \because D_{KL}(q(\mathbf{x}_t|\mathbf{x}_0)||p(\mathbf{x}_t|\mathbf{x}_0)) \geq 0 \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x}_0|\mathbf{x}_t)p(\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \\
&= -\int_{-\infty}^{\infty} \log\left(\frac{p(\mathbf{x}_0|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t - \int_{-\infty}^{\infty} \log\left(p(\mathbf{x}_t)\right) q(\mathbf{x}_t|\mathbf{x}_0)\mathbf{d}\mathbf{x}_t \\
&= -\mathbb{E}_{\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)}\left[\log\left(\frac{p(\mathbf{x}_0|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right)\right] - \mathbb{E}_{\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)}\left[\log\left(p(\mathbf{x}_t)\right)\right] \qquad \because \mathit{definition\ of\ expectation}
\end{aligned}
$$

Maximize ELBO

Instead of intractable log-likelihood log(p(x))

DDPM

ELBO for DDPM

$$
\begin{aligned}
& -\log\left(p_\theta(\mathbf{x}_0)\right) \\
&= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\log\left(p_\theta(\mathbf{x}_0)\right)q(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_T|\mathbf{x}_0)\mathbf{dx}_1\mathbf{dx}_2\dots\mathbf{dx}_T \\
&\qquad \because \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}q(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_T|\mathbf{x}_0)\mathbf{dx}_1\mathbf{dx}_2\dots\mathbf{dx}_T = 1 \\
&= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(p_\theta(\mathbf{x}_0)\right)\right] \qquad \because \mathit{definition\ of\ expectation} \\
&= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(\frac{p_\theta(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_T)}{p_\theta(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\dots,\mathbf{x}_T|\mathbf{x}_0)}\right)\right] \qquad \because p_\theta(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\dots,\mathbf{x}_T|\mathbf{x}_0) = \frac{p_\theta(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_t)}{p_\theta(\mathbf{x}_0)} \\
&= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(\frac{p_\theta(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_T)\ q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ p_\theta(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\dots,\mathbf{x}_T|\mathbf{x}_0)}\right)\right] \\
&\leq -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(\frac{p_\theta(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_T)}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right)\right] \qquad \because \mathit{KL\ divergence} \geq 0 \\
&= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(\frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right)\right] \qquad \because \mathit{notation} \\
&= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(\frac{p_\theta(\mathbf{x}_T)\prod_{t=1}^{T}p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\prod_{t=1}^{T}q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right)\right] \qquad \because *_1\ \mathit{and}\ *_2 \\
&= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(p_\theta(\mathbf{x}_T)\right) + \sum_{t=1}^{T}\log\left(\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right)\right]
\end{aligned}
$$

DDPM

ELBO for DDPM

$$= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(p_\theta(\mathbf{x}_T)\right) + \sum_{t=1}^{T}\log\left(\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right)\right]$$

$$= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(p_\theta(\mathbf{x}_T)\right) + \sum_{t=2}^{T}\log\left(\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right) + \log\left(\frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right)\right]$$

$$= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(p_\theta(\mathbf{x}_T)\right) + \sum_{t=2}^{T}\log\left(\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\cdot\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right) + \log\left(\frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right)\right] \quad \because *_3$$

$$= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(p_\theta(\mathbf{x}_T)\right) + \sum_{t=2}^{T}\log\left(\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\right) + \log\left(\prod_{t=2}^{T}\left(\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right)\cdot\frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right)\right]$$

$$= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(p_\theta(\mathbf{x}_T)\right) + \sum_{t=2}^{T}\log\left(\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\right) + \log\left(\frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_T|\mathbf{x}_0)}\right)\right]$$

$$= -\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\left(\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}\right) + \sum_{t=2}^{T}\log\left(\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\right) + \log\left(p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right)\right]$$

$$= \mathbb{E}_q\left[D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) + \sum_{t>1} D_{KL}(\boxed{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) - \log\left(p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right)\right]$$

tractable posterior distribution

Learning Denoising Model

Loss

For training, we can form variational upper bound that is commonly used for training variational autoencoders:

$$\mathbb{E}_{q(\mathbf{x}_0)}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] =: L$$

Sohl-Dickstein et al. ICML 2015 and Ho et al. NeurIPS 2020 show that:

$$L = \mathbb{E}_q\left[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T))}_{L_T} + \sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\right]$$

where $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ is the tractable posterior distribution:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0), \tilde{\beta}_t\mathbf{I}),$$

$$\text{where } \tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \text{ and } \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \qquad \because *_4$$

Learning Denoising Model

Loss

Since both $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ and $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ are Normal distributions, the KL divergence has a simple form:

$$L_{t-1} = D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}||\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)||^2\right] + C$$

$$\because KL(p, q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

Recall that $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\ \mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\ \epsilon$. Ho et al. NeurIPS 2020 observe that:

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) \qquad \because *_5$$

They propose to represent the mean of the denoising model using a *noise-prediction* network:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\ \epsilon_\theta(\mathbf{x}_t, t)\right)$$

With this parameterization

$$L_{t-1} = \mathbb{E}_{\mathbf{x}_0\sim q(\mathbf{x}_0), \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\frac{\beta_t^2}{2\sigma_t^2(1-\beta_t)(1-\bar{\alpha}_t)}||\epsilon - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\ \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\ \epsilon}_{\mathbf{x}_t}, t)||^2\right] + C$$

Learning Denoising Model

Loss

$$L_{t-1} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\underbrace{\frac{\beta_t^2}{2\sigma_t^2 (1-\beta_t)(1-\bar{\alpha}_t)}}_{\lambda_t} ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, t)||^2 \right]$$

The time dependent λ_t ensures that the training objective is weighted properly for the maximum data likelihood training.

However, this weight is often very large for small t's.

Ho et al. NeurIPS 2020 observe that simply setting $\lambda_t = 1$ improves sample quality. So, they propose to use:

$$L_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(1,T)} \left[||\epsilon - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon}_{\mathbf{x}_t}, t)||^2 \right]$$

For more advanced weighting see Choi et al., Perception Prioritized Training of Diffusion Models, CVPR 2022.

Summary

Training and Sample Generation

Algorithm 1 Training

repeat

$\quad \mathbf{x}_0 \sim q(\mathbf{x}_0)$

$\quad t \sim \mathrm{Uniform}(\{1, \ldots, T\})$

$\quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\quad$ Take gradient descent step on

$$\boxed{\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2}$$

until converged

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

for $t = T, \ldots, 1$ **do**

$\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$

$$\boxed{\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}}$$

end for

return $\mathbf{x}_0$

Summary

Training and Sample Generation

Algorithm 1 Training

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

```python
# train batch
for batch_idx in range(0, x_0.size(0), BATCH_SIZE):
    indices = permutation[batch_idx : batch_idx + BATCH_SIZE]
    batch = x_0[indices]
    epsilon = torch.randn_like(batch).to(DEVICE)

    t = torch.randint(1, N_STEPS + 1, size=(BATCH_SIZE,), device=DEVICE)

    x_t = \
        alpha_bars[t - 1].sqrt().view(-1, 1) * batch + \
        (1 - alpha_bars[t - 1]).sqrt().view(-1, 1) * epsilon
    epsilon_theta = model(x_t, t - 1)

    loss = (epsilon - epsilon_theta).square().mean()

    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
```

Summary

Training and Sample Generation

Algorithm 2 Sampling

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

```python
def p_sample(x_t, model, t_minus_1, betas, alphas, alpha_bars):
    beta_t = betas[t_minus_1]
    alpha_t = alphas[t_minus_1]
    alpha_bar_t = alpha_bars[t_minus_1]
    alpha_bar_t_minus_1 = alpha_bars[torch.clamp_min(t_minus_1 - 1, 0)]

    mean = 1 / alpha_t.sqrt() * (x_t - beta_t / (1 - alpha_bar_t).sqrt() * model(x_t, t_minus_1))
    std = ((1 - alpha_bar_t_minus_1) / (1 - alpha_bar_t) * beta_t).sqrt()
    z = torch.randn_like(x_t).to(DEVICE)

    x_t_minus_1 = mean + std * z

    return x_t_minus_1
```

Code

Tutorial using swiss roll

https://github.com/fidabspd/mywiki/blob/master/tutorial_and_demo/diffusion/diffusion_tutorial.ipynb

References

- arXiv:2006.11239
- https://www.youtube.com/watch?v=cS6JQpEY9cs&ab_channel=ArashVahdat
- https://www.youtube.com/watch?v=_JQSMhqXw-4&ab_channel=%EA%B3%A0%EB%A0%A4%EB%8C%80%ED%95%99%EA%B5%90%EC%82%B0%EC%97%85%EA%B2%BD%EC%98%81%EA%B3%B5%ED%95%99%EB%B6%80DSBA%EC%97%B0%EA%B5%AC%EC%8B%A4
- https://www.youtube.com/watch?v=uFoGaIVHfoE&t=13s&ab_channel=%EB%94%94%ED%93%A8%EC%A0%84%EC%98%81%EC%83%81%EC%98%AC%EB%A0%A4%EC%95%BC%EC%A7%80
- https://developers-shack.tistory.com/8
- https://modulabs.co.kr/blog/variational-inference-intro/

Appendix

Mathematical expression

*1

$$
\begin{aligned}
& p_\theta(\mathbf{x}_{0:T}) \\
=& p_\theta(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) \\
=& \frac{p_\theta(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)}{p_\theta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)} \cdot \frac{p_\theta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)}{p_\theta(\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_T)} \cdot \ldots \cdot \frac{p_\theta(\mathbf{x}_{T-1}, \mathbf{x}_T)}{p_\theta(\mathbf{x}_T)} \cdot p_\theta(\mathbf{x}_T) \\
=& p_\theta(\mathbf{x}_0|\mathbf{x}_1) \cdot p_\theta(\mathbf{x}_1|\mathbf{x}_2) \cdot \ldots \cdot p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T) \cdot p_\theta(\mathbf{x}_T) \\
=& p_\theta(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)
\end{aligned}
$$

Appendix

Mathematical expression

*2

$$
\begin{aligned}
& q(\mathbf{x}_{1:T}|\mathbf{x}_0) \\
=& \frac{q(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_T)}{q(\mathbf{x}_0)} \\
=& \frac{q(\mathbf{x}_1, \mathbf{x}_0)}{q(\mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_0)}{q(\mathbf{x}_1, \mathbf{x}_0)} \cdot ... \cdot \frac{q(\mathbf{x}_T, ..., \mathbf{x}_0)}{q(\mathbf{x}_{T-1}, ..., \mathbf{x}_0)} \\
=& \frac{q(\mathbf{x}_1, \mathbf{x}_0)}{q(\mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_2, \mathbf{x}_1)}{q(\mathbf{x}_1)} \cdot ... \cdot \frac{q(\mathbf{x}_T, \mathbf{x}_{T-1})}{q(\mathbf{x}_{T-1})} \qquad \because \textit{Markov chain property} \\
=& q(\mathbf{x}_1|\mathbf{x}_0) \cdot q(\mathbf{x}_2|\mathbf{x}_1) \cdot ... \cdot q(\mathbf{x}_T|\mathbf{x}_{T-1}) \\
=& \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})
\end{aligned}
$$

Appendix

Mathematical expression

*3

$$
\begin{aligned}
&q(\mathbf{x}_t|\mathbf{x}_{t-1}) \\
=&q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0) \qquad \because \textit{Markov chain property} \\
=&\frac{q(\mathbf{x}_t,\mathbf{x}_{t-1},\mathbf{x}_0)}{q(\mathbf{x}_{t-1},\mathbf{x}_0)} \\
=&\frac{q(\mathbf{x}_t,\mathbf{x}_{t-1},\mathbf{x}_0)\ q(\mathbf{x}_t,\mathbf{x}_0)}{q(\mathbf{x}_t,\mathbf{x}_0)\ q(\mathbf{x}_{t-1},\mathbf{x}_0)} \\
=&q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\cdot\frac{\frac{q(\mathbf{x}_t,\mathbf{x}_0)}{q(\mathbf{x}_0)}}{\frac{q(\mathbf{x}_{t-1},\mathbf{x}_0)}{q(\mathbf{x}_0)}}
\end{aligned}
$$

Appendix
Mathematical expression

*4

$$
\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) &= q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)\,\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \qquad \because \mathit{Bayes'\ rule} \\
&= q(\mathbf{x}_t|\mathbf{x}_{t-1})\,\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \qquad \because \mathit{Markov\ chain\ property} \\
&\propto \exp\Bigg(-\frac{1}{2}\Bigg(\frac{(\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t}+\frac{(\mathbf{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(\mathbf{x}_t-\sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t}\Bigg)\Bigg) \\
&\qquad \because \mathit{Gaussian\ PDF} = \frac{1}{\sigma\sqrt{2\pi}}\exp\Bigg(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Bigg) \\
&= \exp\Bigg(-\frac{1}{2}\Bigg(\frac{\mathbf{x}_t^2-2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1}+\alpha_t\mathbf{x}_{t-1}^2}{\beta_t}+\frac{\mathbf{x}_{t-1}^2-2\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_{t-1}\mathbf{x}_0+\bar{\alpha}_{t-1}\mathbf{x}_0^2}{1-\bar{\alpha}_{t-1}}-\frac{(\mathbf{x}_t-\sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t}\Bigg)\Bigg) \\
&= \exp\Bigg(-\frac{1}{2}\Bigg(\Big(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\Big)\mathbf{x}_{t-1}^2-\Big(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\Big)\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0)\Bigg)\Bigg)
\end{aligned}
$$

Appendix

Mathematical expression

$*_4$

$$= \exp\Bigg(-\frac{1}{2}\Bigg(\Big(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\Big)\mathbf{x}_{t-1}^2 - \Big(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\Big)\mathbf{x}_{t-1} + C(\mathbf{x}_t,\mathbf{x}_0)\Bigg)\Bigg)$$

$$\mathit{Let}\quad \frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}} = A,\quad \frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0 = B,\quad C(\mathbf{x}_t,\mathbf{x}_0) = C$$

$$= \exp\Bigg(-\frac{1}{2}\Big(A\mathbf{x}_{t-1}^2 - B\mathbf{x}_{t-1} + C\Big)\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}A\Big(\mathbf{x}_{t-1}^2 - \frac{B}{A}\mathbf{x}_{t-1} + \Big(\frac{B}{2A}\Big)^2 - \Big(\frac{B}{2A}\Big)^2 + \frac{C}{A}\Big)\Bigg)$$

$$\propto \exp\Bigg(-\frac{1}{2}\Bigg(\frac{\Big(\mathbf{x}_{t-1} - \frac{B}{2A}\Big)^2}{\frac{1}{A}}\Bigg)\Bigg)$$

$$\therefore\ \tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0) := \frac{B}{2A}\quad \mathit{and}\quad \tilde{\beta}_t := \frac{1}{A}$$

Appendix

Mathematical expression

$*_4$

$$
\begin{aligned}
\tilde{\beta}_t &= 1/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) \\
&= 1/(\frac{\alpha_t - \bar{\alpha}_t + \beta_t}{1-\bar{\alpha}_{t-1}} \cdot \frac{1}{\beta_t}) \\
&= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0)/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) \\
&= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0) \cdot \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0
\end{aligned}
$$

Appendix

Mathematical expression

*5

$$
\begin{aligned}
\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \cdot \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon) \\
&\qquad \because \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \\
&\qquad\quad \mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon) \\
&= \left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\right)\mathbf{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\sqrt{1-\bar{\alpha}_t}}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\epsilon \\
&= \left(\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} + \frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\right)\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon \\
&= \left(\frac{1}{\sqrt{\alpha_t}} \cdot \frac{\alpha_t - \bar{\alpha}_t + \beta_t}{1-\bar{\alpha}_t}\right)\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)
\end{aligned}
$$