DDPM Denoising Diffusion Probabilistic Models

Molecular Team Lecture Series
Dong-Hee Shin
06.28.22

Researchers

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Jonathan Ho
Google Research

Ajay Jain
UC Berkeley

Pieter Abbeel
UC Berkeley

What is Diffusion Model?

Latent variable models of the form

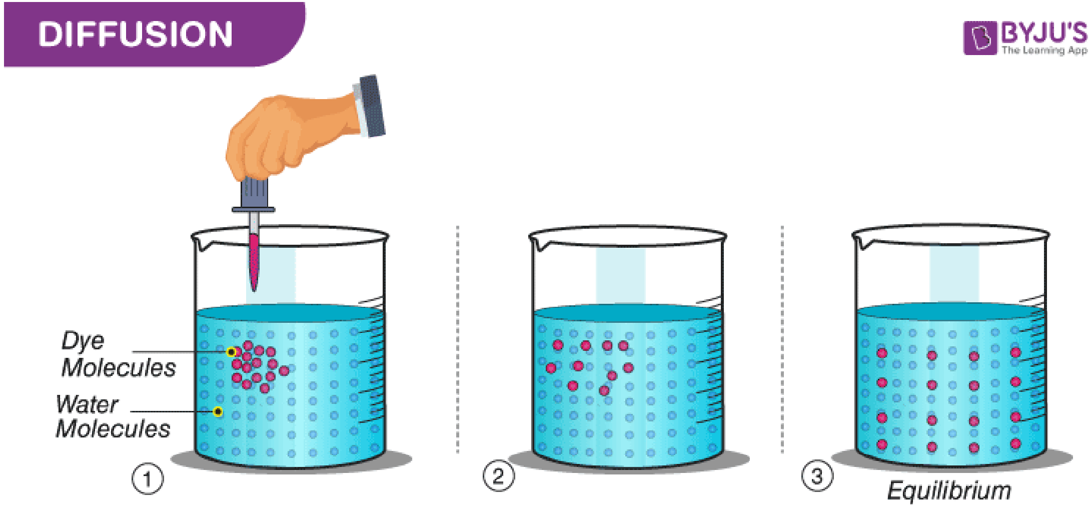
$$p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T},$$

where

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),$$

OK Cool!! What is Diffusion Again?

- **Diffusion:** *gradual* movement of concentration within a body, due to a concentration gradient
- **Diffusion Model:** Markov chain of diffusion steps to *slowly* add random noise to data



Diffusion Model Ancestor

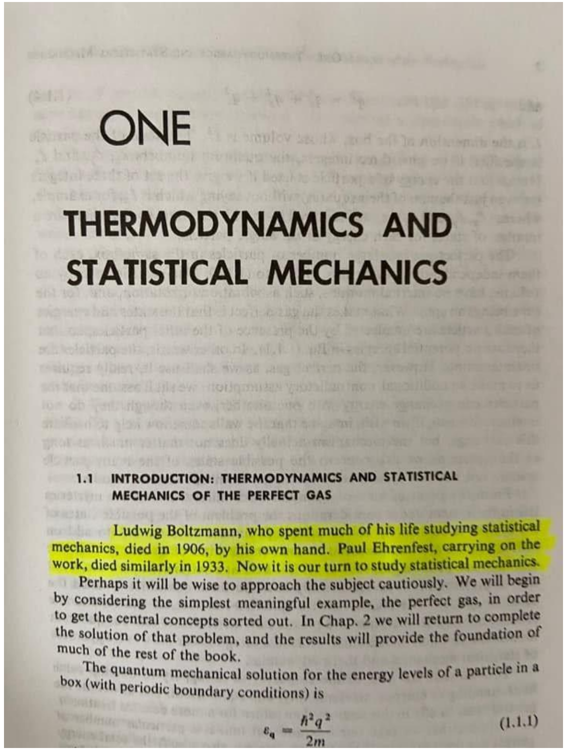
Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein JASCHA@STANFORD.EDU
Stanford University

Eric A. Weiss EAWEISS@BERKELEY.EDU
University of California, Berkeley

Niru Maheswaranathan NIRUM@STANFORD.EDU
Stanford University

Surya Ganguli SGANGULI@STANFORD.EDU
Stanford University



ONE

THERMODYNAMICS AND STATISTICAL MECHANICS

1.1 INTRODUCTION: THERMODYNAMICS AND STATISTICAL MECHANICS OF THE PERFECT GAS

Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics.

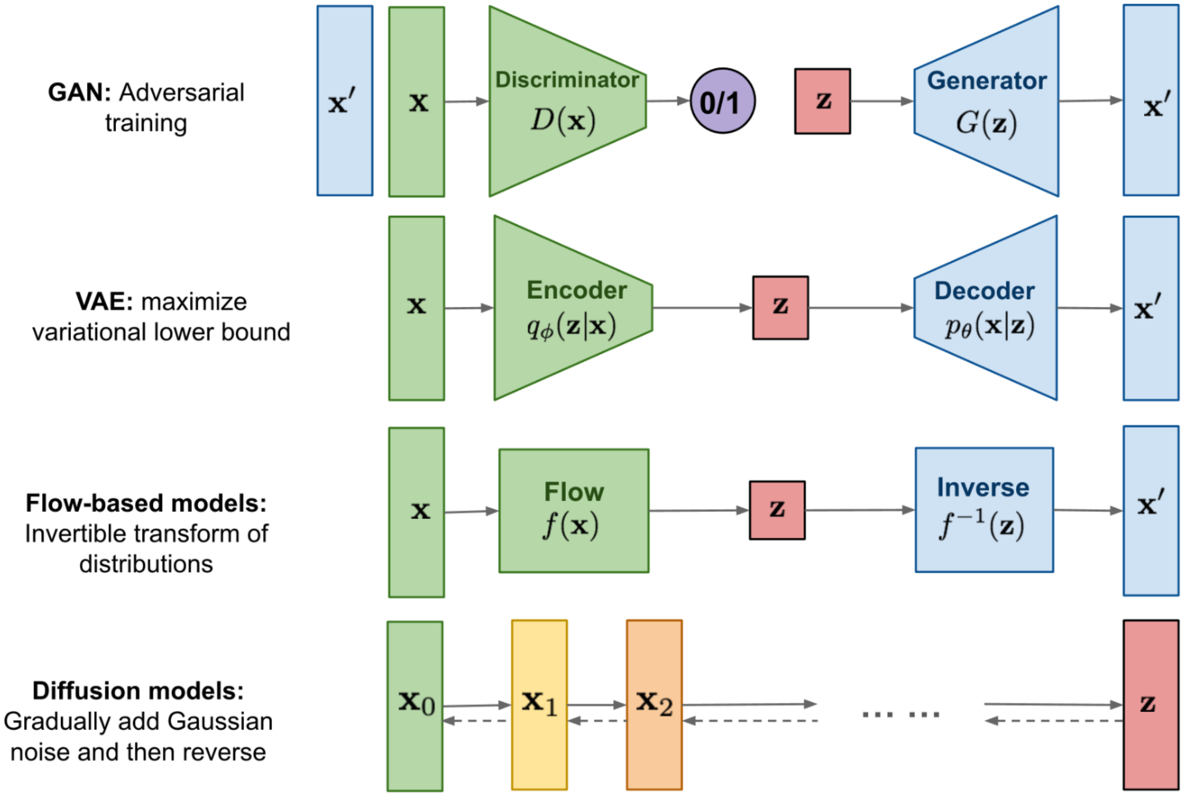
Perhaps it will be wise to approach the subject cautiously. We will begin by considering the simplest meaningful example, the perfect gas, in order to get the central concepts sorted out. In Chap. 2 we will return to complete the solution of that problem, and the results will provide the foundation of much of the rest of the book.

The quantum mechanical solution for the energy levels of a particle in a box (with periodic boundary conditions) is

$$\varepsilon_q = \frac{\hbar^2 q^2}{2m} \quad (1.1.1)$$

- ICML 2015
- Inspired by non-equilibrium statistical physics
- slowly destroy structure in a data distribution and then learn a reverse diffusion process that restores data structure

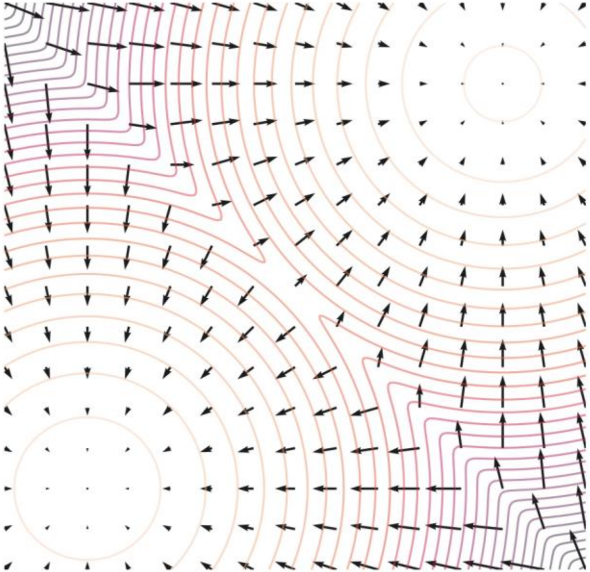
Generative Models

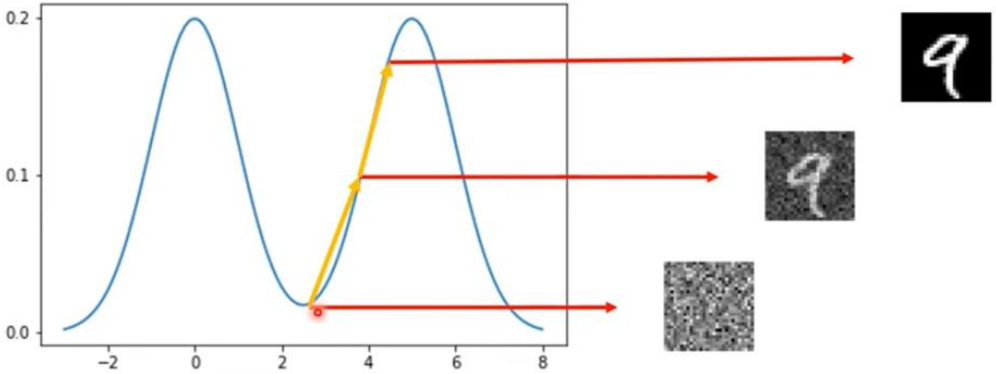


- **GAN:** Instability Training
- **VAE:** Requires ELBO
- **Flow-based:** requires inverse function
- **Diffusion:** latent variable has high dimensinality (same as original data)

Score-based Generative Model

- **Key Idea:** model the gradient of log PDF, a known as (stein) score function
- **Strength:** allow exact likelihood computation by using normalizing flow models





Score Function

$$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}, \longrightarrow p(\mathbf{x}).$$

$$p_\theta(\mathbf{x}) = \frac{e^{-f_\theta(\mathbf{x})}}{Z_\theta},$$

score function $= \nabla_{\mathbf{x}} \log p(\mathbf{x}),$

$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x}) = -\nabla_{\mathbf{x}} f_\theta(\mathbf{x}) - \underbrace{\nabla_{\mathbf{x}} \log Z_\theta}_{=0} = -\nabla_{\mathbf{x}} f_\theta(\mathbf{x}).$$

Multivariate Normal

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

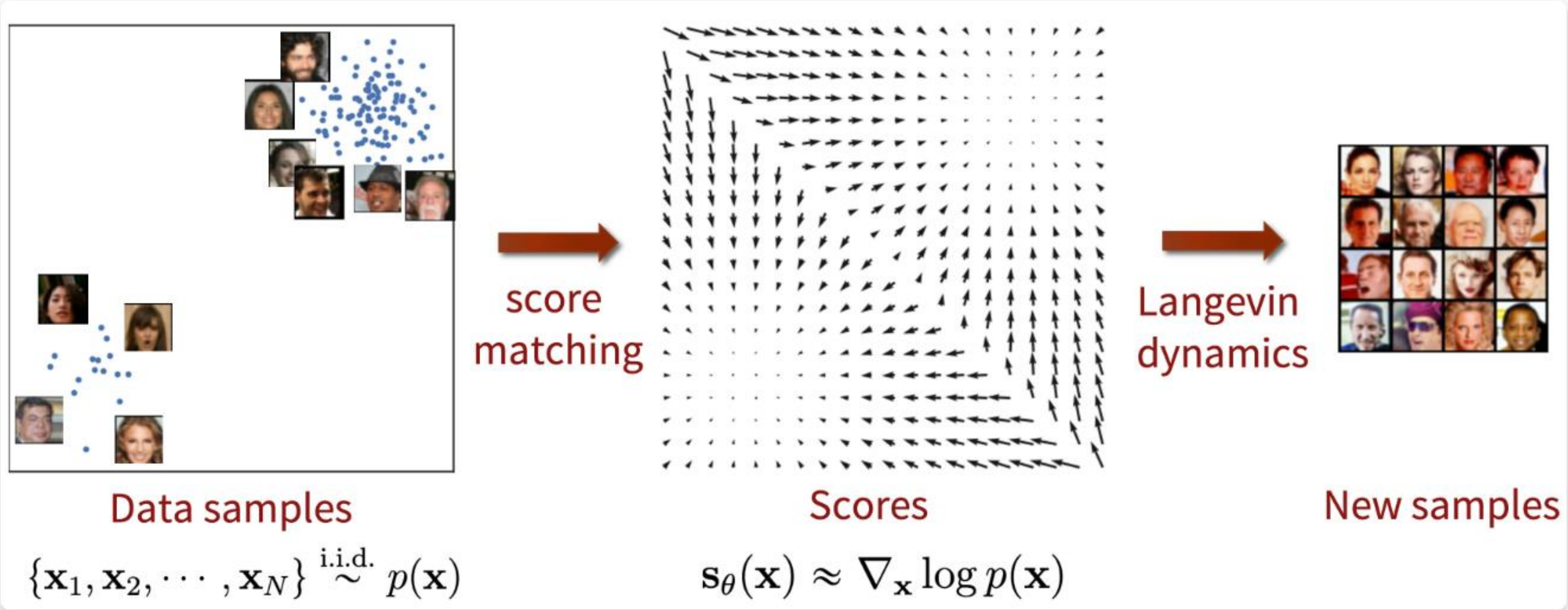
$$\nabla_x logp(x) = -\Sigma^{-1}(x-\mu)$$

Example

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, x = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}$$

$$\nabla_x logp(x) = \begin{bmatrix} 0 \\ -0.5 \end{bmatrix}$$

Score-based Model



Score Matching

$$Loss = \frac{1}{2} E_{p_{data}(x)}[||\overset{\text{True score}}{\boxed{\nabla_x log p(x)}} - \overset{\text{Score Network}}{\boxed{s_\theta(x)}}||_2^2]$$

Score Matching
(2005)

$$E_{p_{data}(x)}\left[tr(\overset{\text{Jacobian Matrix (d*d)}}{\boxed{\nabla_x s_\theta(x)}}) + \frac{1}{2}||s_\theta(x)||_2^2\right]$$

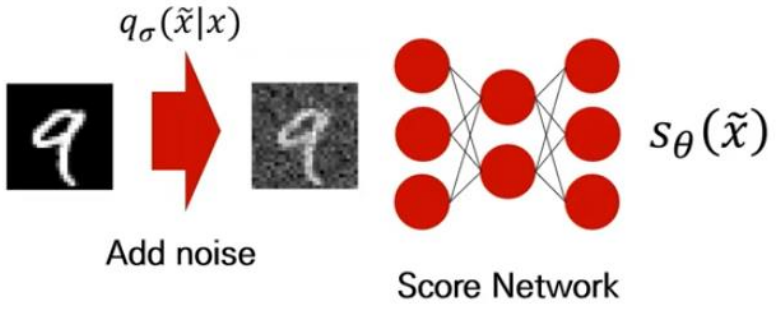
Denoising Score Matching
(2011)

$$\frac{1}{2} E_{q_\sigma(\tilde{x}|x)p_{data}(x)}[||s_\theta(\tilde{x}) - \nabla_x log q_\sigma(\tilde{x}|x)||_2^2]$$

Hyvärinen, A., & Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6(4).
Vincent, P. (2011).A connection between score matching and denoising autoencoders. Neural computation, 23(7), 1661-1674.

Score Network

- Train Score network via denoising score matching
- **Input:** data x
- **Output:** Score for data x
- Note that same dimensionality for input and output



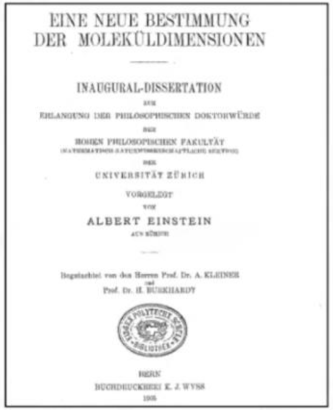
$$\frac{1}{2}E_{q_\sigma(\tilde{x}|x)p_{data}(x)}[||\overset{\text{Network Output}}{\boxed{s_\theta(\tilde{x})}} - \overset{\text{Score}}{\underbrace{\boxed{\nabla_x log q_\sigma(\tilde{x}|x)}}_{-\Sigma^{-1}(\tilde{x}-\mu)}}||_2^2]$$

Background

- Langevin dynamics is modeling of the dynamics of molecular systems
- Langevin dynamics simulations are a kind of Monte Carlo simulation

Particle random walk: $$dx_t = \mu dt + \sigma dW_t$$

Particle density diffusion: $$\frac{\partial \rho}{\partial t} = c\Delta\rho$$


EINE NEUE BESTIMMUNG
DER MOLEKÜLDIMENSIONEN

INAUGURAL-DISSERTATION

ERLANGUNG DER PHILOSOPHISCHEN DOKTORWÜRDE

HOHEN PHILOSOPHISCHEN FAKULTÄT

UNIVERSITÄT ZÜRICH

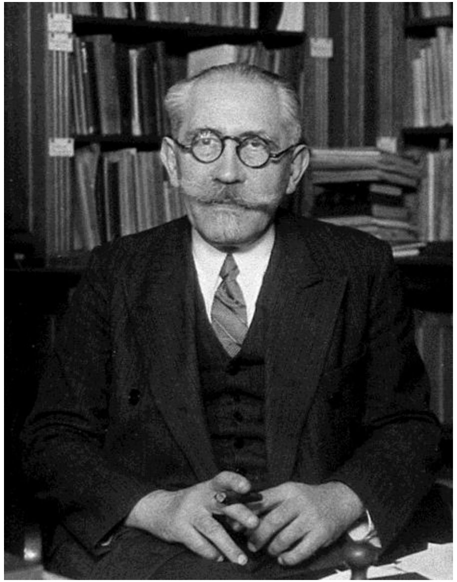
VORGELEGT

ALBERT EINSTEIN

Begutachtet von den Herren Prof. Dr. A. KLEINER
Prof. Dr. H. BURKHARDT

BERN
BUCHDRUCKEREI K. J. WYSS

Albert Einstein



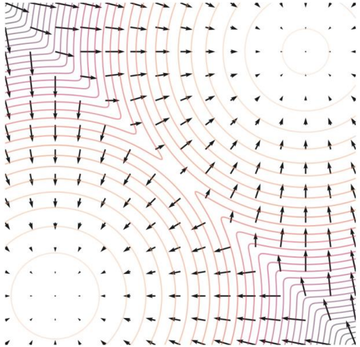
Paul Langevin

Lorentz, Einstein and Langevin in 1927

Langevin Dynamics

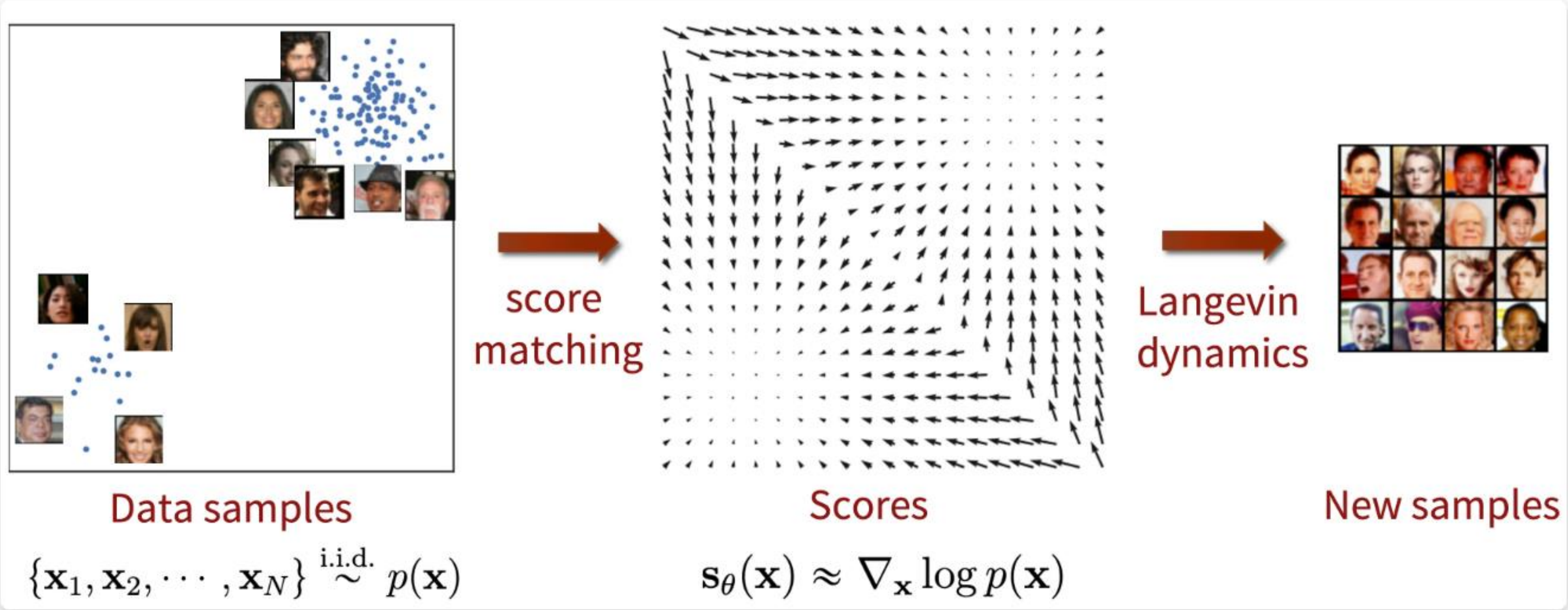
- After training score network, then we can use this neural network to estimate score in all data distribution
- Langevin dynamics provides an MCMC procedure to sample from a distribution p(x) using only its score function

$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathbf{z}_i, \quad i = 0, 1, \cdots, K,$$



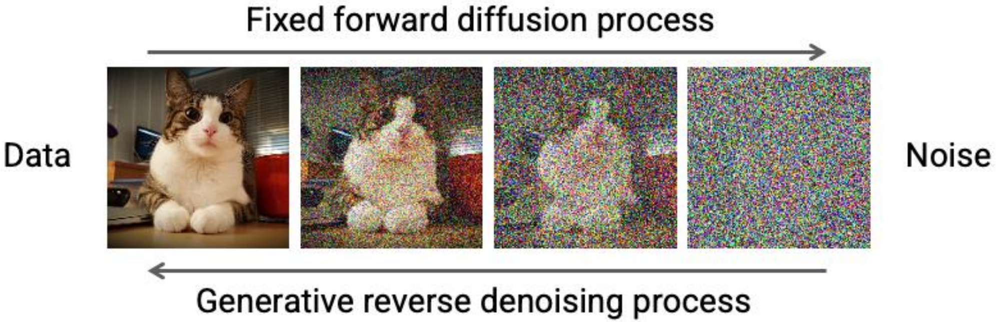
$$\mathbf{z}_i \sim \mathcal{N}(0, I).$$

Recall


score
matching
Langevin
dynamics
Data samples
$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$
Scores
$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x})$
New samples

Connecting to Diffusion Models

Forward Diffusion Process



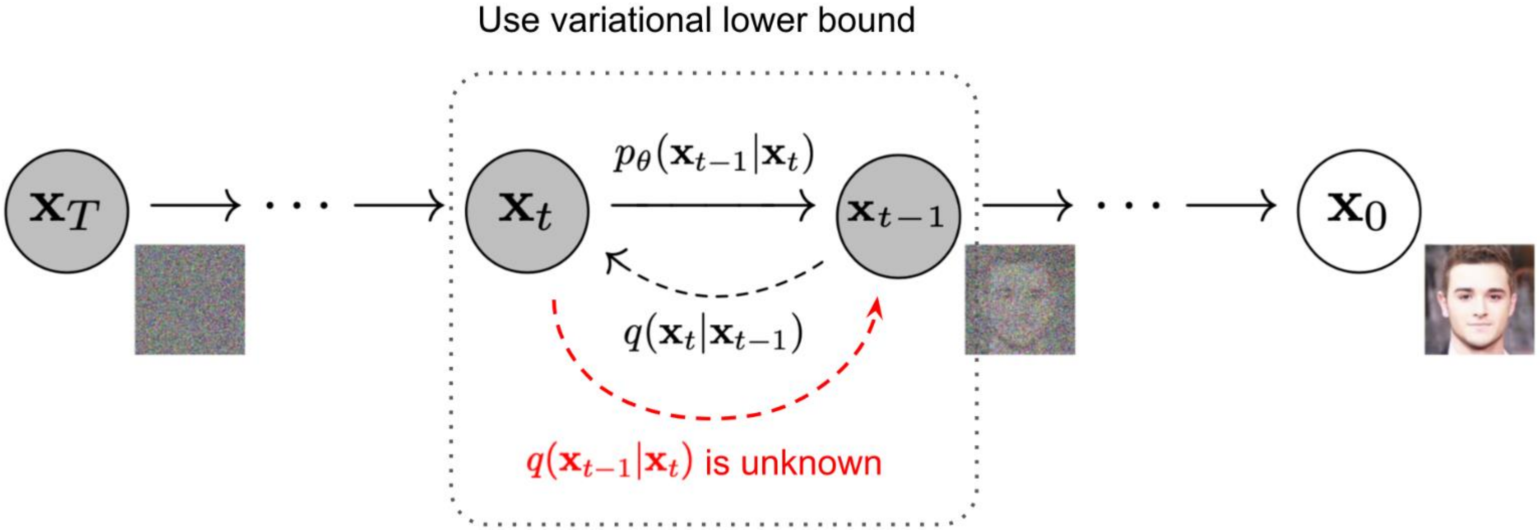
$$\mathbf{x}_0 \sim q(\mathbf{x}),$$

$$\mathbf{x}_1, \ldots, \mathbf{x}_T.$$

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

A nice property of the above process is that we can sample $\mathbf{x}_t$ at any arbitrary time step t in a closed form using reparameterization trick. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$:

Reverse Diffusion Process (1)



$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Loss Function (1)

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

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

$$\text{where}\quad\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0):=\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1\to\bar{\alpha}_t}\mathbf{x}_0+\frac{\sqrt{\alpha_t}(1\to\bar{\alpha}_{t-1})}{1\to\bar{\alpha}_t}\mathbf{x}_t\quad\text{and}\quad\tilde{\beta}_t:=\frac{1\to\bar{\alpha}_{t-1}}{1\to\bar{\alpha}_t}\beta_t$$

$$\begin{aligned}q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)&=q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}\\&\propto\exp\Big(-\frac{1}{2}\big(\frac{(\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t}+\frac{(\mathbf{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(\mathbf{x}_t-\sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t}\big)\Big)\\&=\exp\Big(-\frac{1}{2}\big((\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})\mathbf{x}_{t-1}^2-(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{2\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{x}_0)\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0)\big)\Big)\end{aligned}$$

Loss Function (2)

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

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

$$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\mathbf{z}_t)$$

$$\tilde{\boldsymbol{\mu}}_t = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\mathbf{z}_t)$$

$$= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{z}_t\right)$$

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

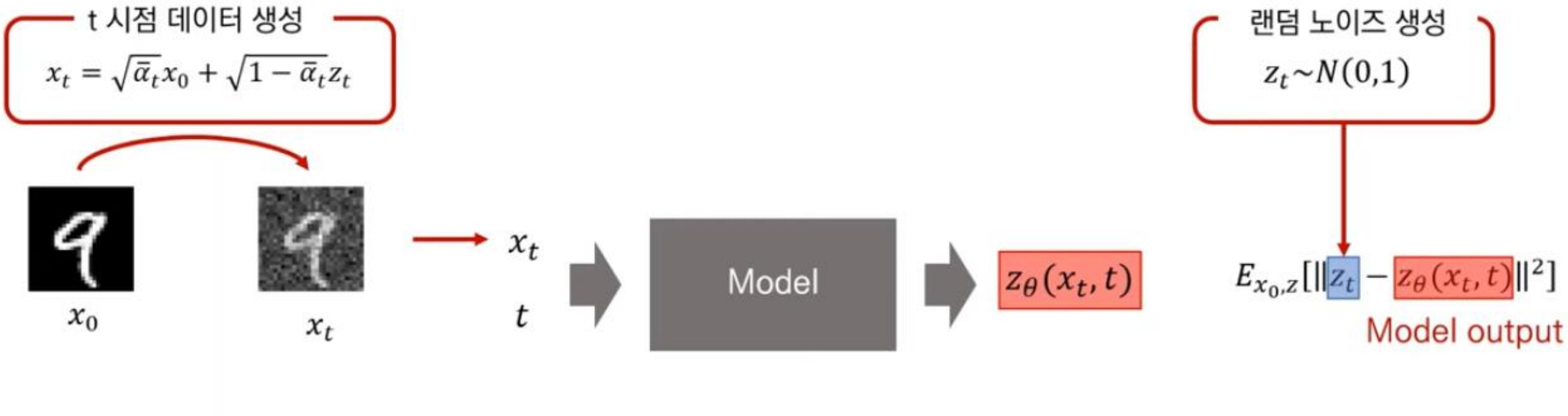
Loss Function (3)

$$
\begin{aligned}
-\log p_\theta(\mathbf{x}_0) &\le -\log p_\theta(\mathbf{x}_0) + D_{\text{KL}}(q(\mathbf{x}_{1:T}|\mathbf{x}_0) \| p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)) \\
&= -\log p_\theta(\mathbf{x}_0) + \mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})/p_\theta(\mathbf{x}_0)}\Big] \\
&= -\log p_\theta(\mathbf{x}_0) + \mathbb{E}_q\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})} + \log p_\theta(\mathbf{x}_0)\Big] \\
&= \mathbb{E}_q\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big]
\end{aligned}
$$

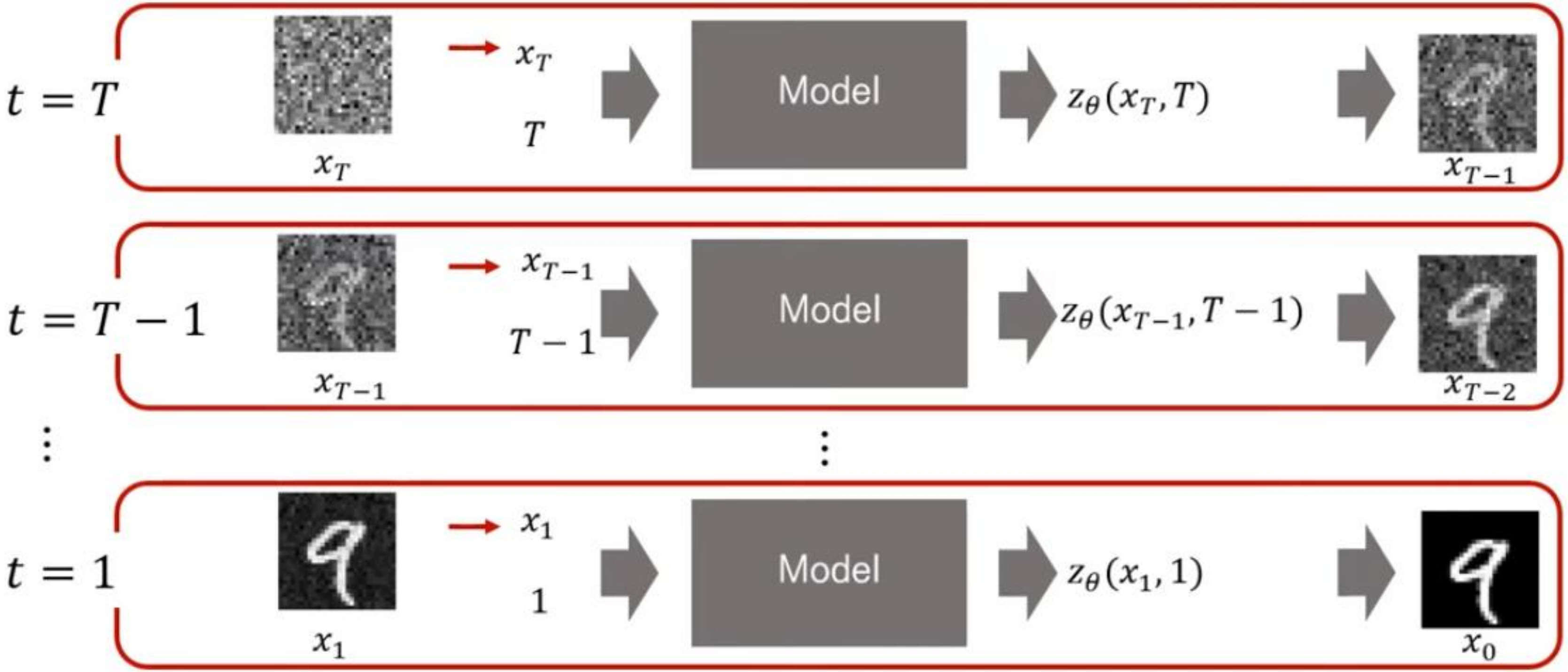
$$
\text{Let } L_\text{VLB} = \mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big] \ge -\mathbb{E}_{q(\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0)
$$

$$
\begin{aligned}
L_\text{VLB} &= \mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big] \\
&= \mathbb{E}_q\Big[\log \frac{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_T)\prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=1}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \Big(\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \cdot \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}\Big) + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\Big] \\
&= \mathbb{E}_q[\underbrace{D_\text{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T} + \sum_{t=2}^T \underbrace{D_\text{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}]
\end{aligned}
$$

DDPM Training

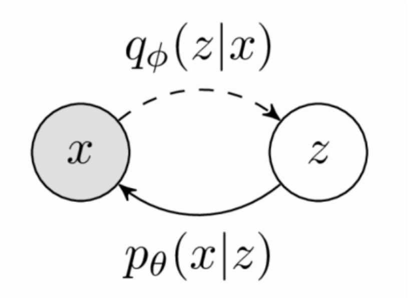


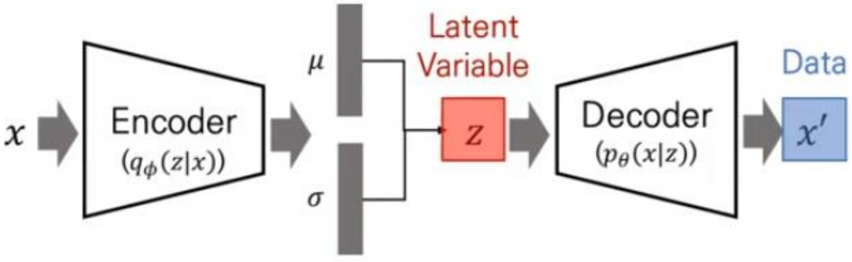
DDPM Test/Generation



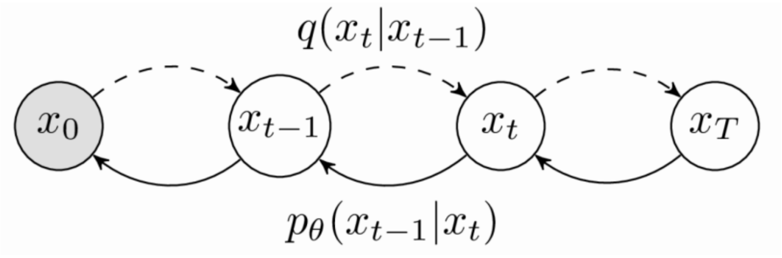
VAE vs. DDPM

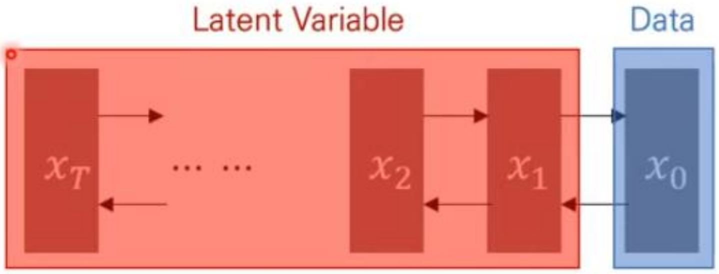
- **VAE:** fixed small dimensionality for latent variable (z)
- **Diffusion Model:** Latent variable is modeled as Markov chain





VAE





DDPM

Summary

1) Extreme flexibility in model structure
2) Exact sampling tractability
3) Not using surrogate function, but directly optimize negative log-likelihood
4) Incorporate Markov chain property into generative model

Thank you

Dong-Hee Shin
(dongheeshin@korea.ac.kr)