Medical Imaging & Intelligent Reality Lab.

- Understanding Diffusion Models: A Unified Perspective

2023.05.03

Jong Jun Won

울산대학교 UNIVERSITY OF ULSAN
서울아산병원 Asan Medical Center

Preprint

Understanding Diffusion Models: A Unified Perspective

Calvin Luo
Google Research, Brain Team
calvinluo@google.com

August 26, 2022

0. contents

1. Introduction

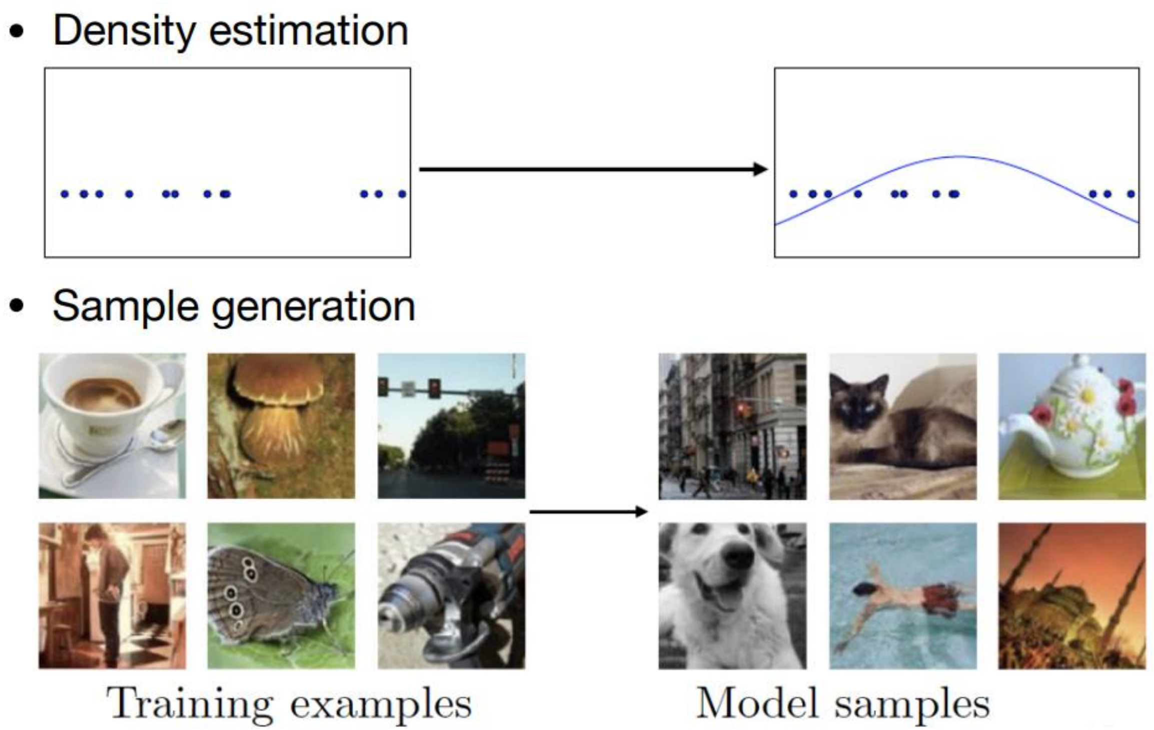
Generative Models

- Given observed samples **x**, the goal is to learn to model its ***true data distribution*** $P(x)$

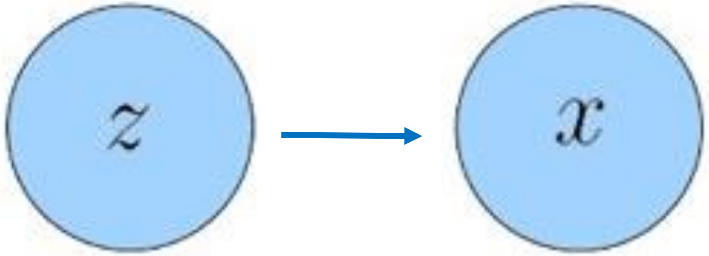
- Data distribution ***: the statistical characteristics and patterns*** present in the real data
  It specifies the probabilities of all events – Introduction to probability

- Once learned, we can generate new samples from our approximate model.

- GAN
- Likelihood-based : **Variational Autoencoders**



2. Background



Latent variable

- We can think of The data(X) as **generated** by an associated unseen latent variable(Z).
- We generally seek to learn **latent representations (low-dimension)** rather than higher-dimensional ones.
- Learning data distribution in lower-dimension is more easier than high-dimension (sparsity, complexity)
- Learning lower-dimensional latents can also be seen as a form of compression
- The best intuition for expressing this idea is through Plato's Allegory of the Cave.

Latent space

2 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

A **latent space**, also known as a **latent feature space** or **embedding space**, is an embedding of a set of items within a manifold in which items resembling each other are positioned closer to one another in the latent space. Position within the latent space can be viewed as being defined by a set of latent variables that emerge from the resemblances from the objects.

2. Background

Latent variable

1. Feature Extraction, Compression from dataset (X -> Z)

2. Generation from latent variable (Z -> X)

(Generated) Object : encapsulate abstract properties (size, shape, and more)

The cave people can never see the hidden objects

They can still reason and draw(generate) inferences about them.

In a similar way, we can approximate latent representations that describe the data we observe.

2. Background

Evidence Lower Bound

- Likelihood-based : to learn a model to maximize the likelihood $P(x)$
- The latent variables and the data as modeled by a joint distribution $P(x, z)$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$$

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}$$

2. Background

Evidence Lower Bound

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$$

- maximizing the likelihood p(x) is difficult because it either involves integrating out all **latent variables z**, which is **intractable** for complex models,

in practice *any* solution takes too many resources to be useful,

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}$$

- it involves having access to a ground truth latent encoder p(z|x)

2. Background

Evidence Lower Bound

$$
\begin{aligned}
\log p(x) &= \log p(x) \int q_\phi(z|x)dz \\
&= \int q_\phi(z|x)(\log p(x))dz \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log p(x)\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{p(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)q_\phi(z|x)}{p(z|x)q_\phi(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] + D_{\mathrm{KL}}(q_\phi(z|x) \parallel p(z|x)) \\
&\geq \boxed{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right]}
\end{aligned}
$$

$q_\emptyset(z|x)$: a parameterizable ϕ model that is learned to estimate the true distribution over latent variables for given x

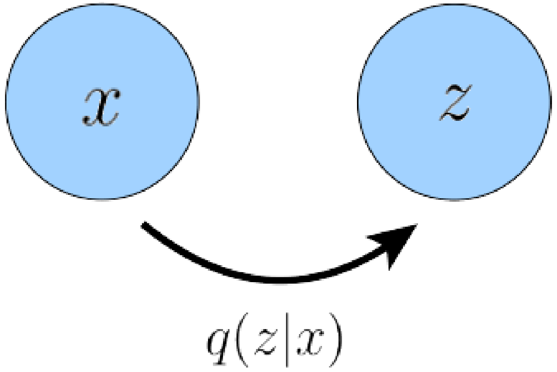
2. Background

Evidence Lower Bound

- the **E**vidence **L**ower **BO**und (ELBO) is a lower bound of the log likelihood of $P(x)$ (the evidence).
- Then, **maximizing the ELBO** becomes a proxy objective with which to **optimize a latent variable model**
- $q_{\emptyset}(z|x)$: parameters ϕ that we seek to optimize true latent variables for given observations x

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] \xrightarrow[\text{approximation}]{} p(x) = \frac{p(x,z)}{p(z|x)}$$

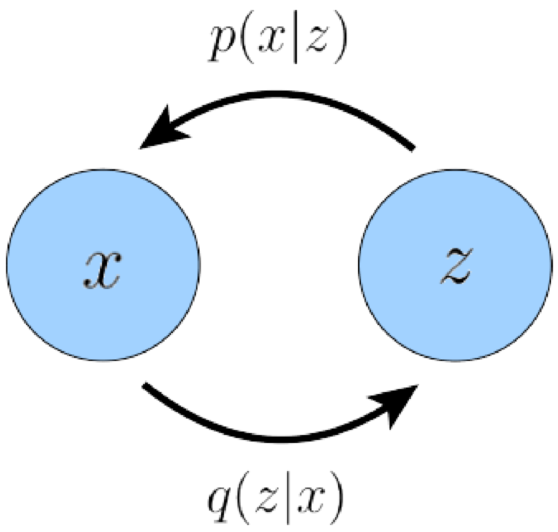
- Our goal is to learn this underlying latent structure that describes our observed data

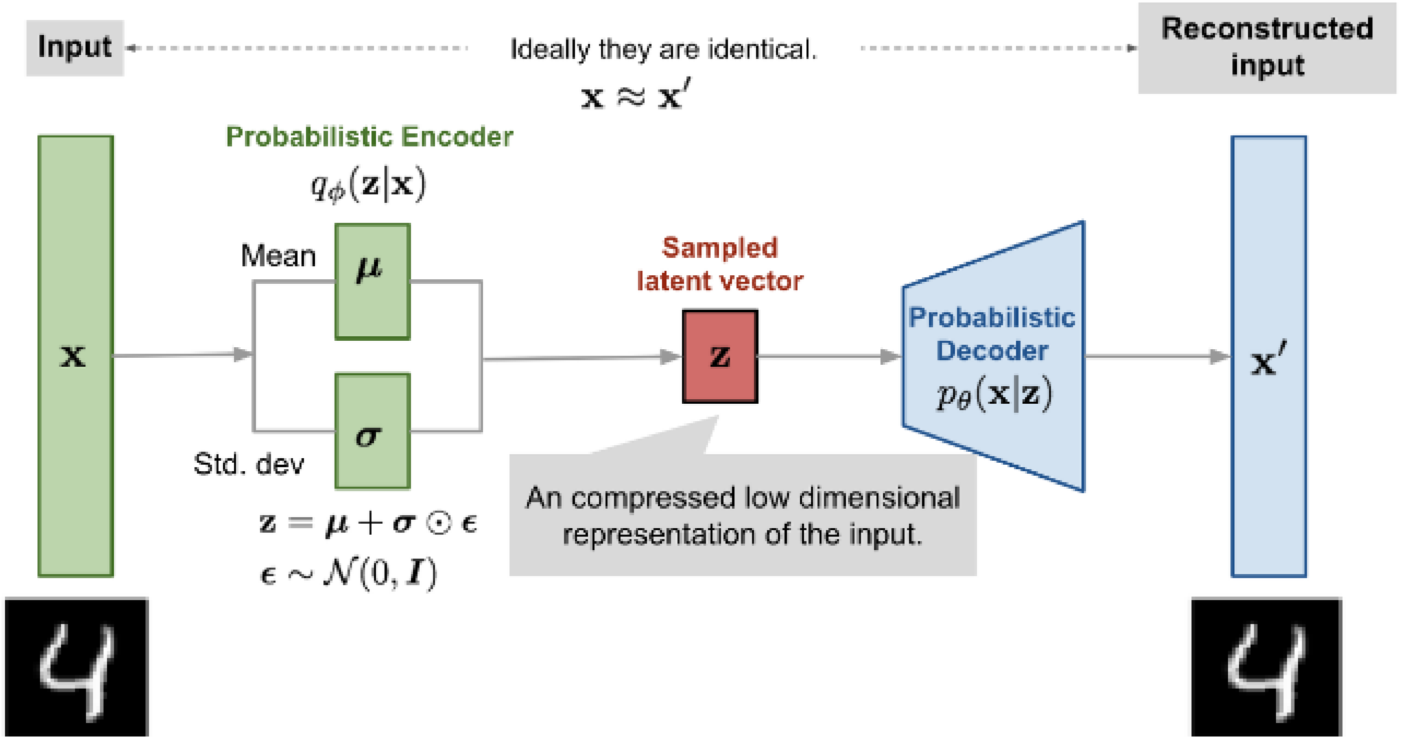


2. Background

Variational Autoencoder

$P(x,z)$



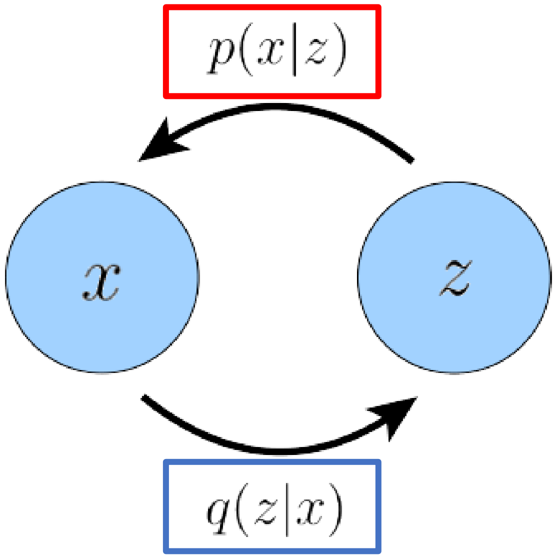


MI2L

Variational Autoencoder

- The ELBO is optimized jointly over parameters ϕ and θ

$$\mathbb{E}_{q_{\phi}(z|x)}\left[\log \frac{p(x,z)}{q_{\phi}(z|x)}\right] = \mathbb{E}_{q_{\phi}(z|x)}\left[\log \frac{p_{\theta}(x|z)p(z)}{q_{\phi}(z|x)}\right]$$
$$= \mathbb{E}_{q_{\phi}(z|x)}\left[\log p_{\theta}(x|z)\right] + \mathbb{E}_{q_{\phi}(z|x)}\left[\log \frac{p(z)}{q_{\phi}(z|x)}\right]$$
$$= \underbrace{\mathbb{E}_{q_{\phi}(z|x)}\boxed{\left[\log p_{\theta}(x|z)\right]}}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(\boxed{q_{\phi}(z|x) \parallel p(z)})}_{\text{prior matching term}}$$



울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

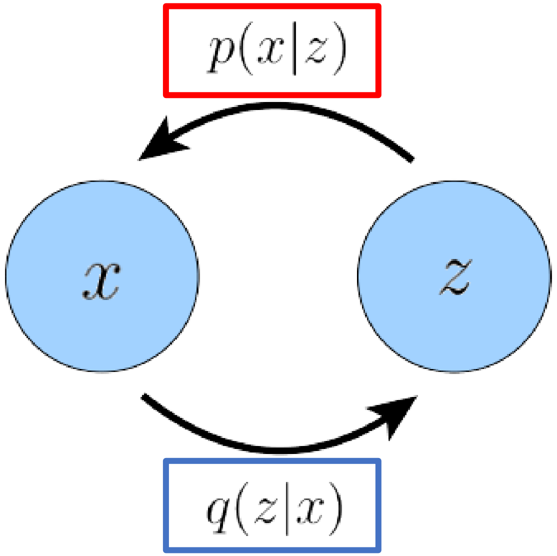
2. Background

MI2L

Variational Autoencoder

- The encoder : a multivariate Gaussian with diagonal covariance
- The latent : a standard multivariate Gaussian

$$\underbrace{\mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\boxed{\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})}\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(\boxed{q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})}\right)}_{\text{prior matching term}}$$



$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \mathbf{I})$$

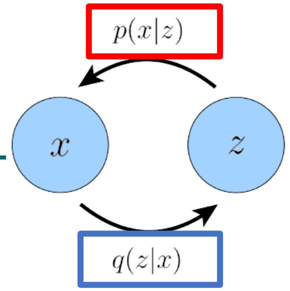
$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \mathbf{I})$$

$$q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\phi}(\boldsymbol{x}), \boldsymbol{\sigma}^2_{\phi}(\boldsymbol{x})\mathbf{I})$$

$$\boldsymbol{z} = \boldsymbol{\mu}_{\phi}(\boldsymbol{x}) + \boldsymbol{\sigma}_{\phi}(\boldsymbol{x}) \odot \boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \mathbf{I})$$

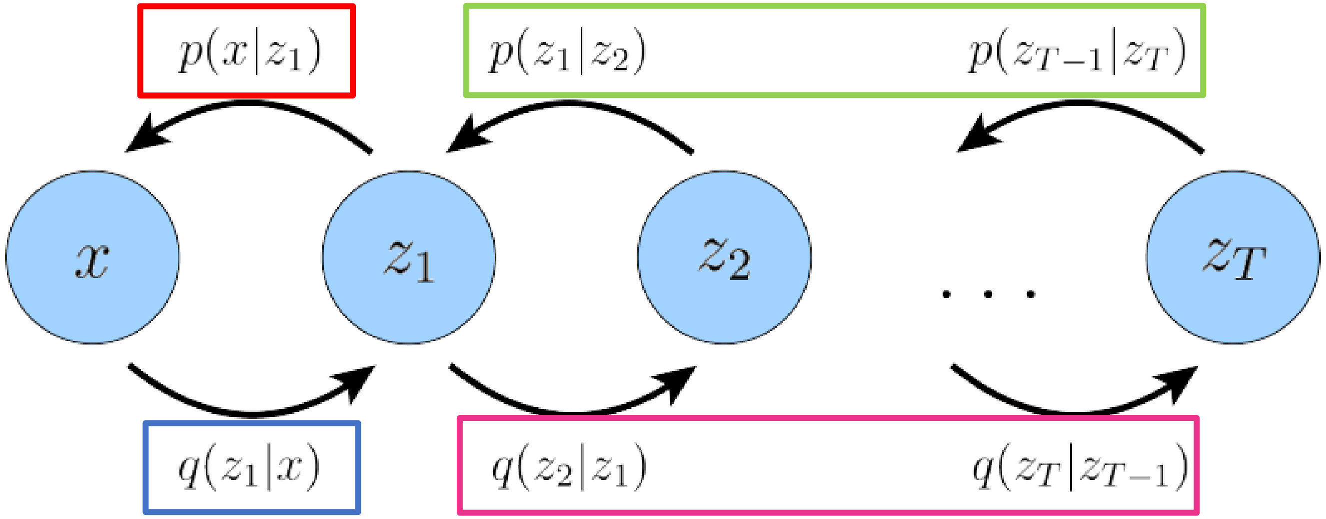
Dot product

2. Background



Hierachical Variational Autoencoder

- A Hierarchical Variational Autoencoder (HAVE) is a generalization of a VAE that extends to multiple hierachies over latent variables.
- Markovian HVAE
- In a MHVAE, the generative process is **a Markov chain**,
- **Each latent Z_t is generated only from the previous latent.**



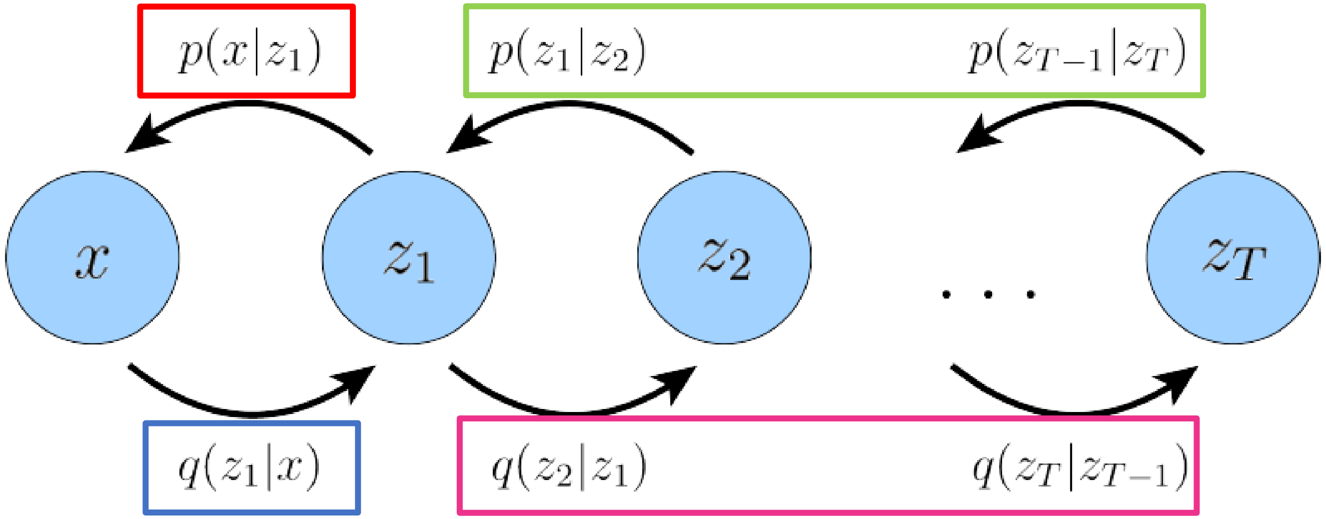
$$\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right]$$

$$p(x, z_{1:T}) = p(z_T)\boxed{p_\theta(x|z_1)}\prod_{t=2}^{T}\boxed{p_\theta(z_{t-1}|z_t)}$$

$$q_\phi(z_{1:T}|x) = \boxed{q_\phi(z_1|x)}\prod_{t=2}^{T}\boxed{q_\phi(z_t|z_{t-1})}$$

2. Background

Hierachical Variational Autoencoder



$$\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right]$$

$$p(x, z_{1:T}) = p(z_T)\boxed{p_\theta(x|z_1)}\prod_{t=2}^{T}\boxed{p_\theta(z_{t-1}|z_t)}$$

$$q_\phi(z_{1:T}|x) = \boxed{q_\phi(z_1|x)}\prod_{t=2}^{T}\boxed{q_\phi(z_t|z_{t-1})}$$

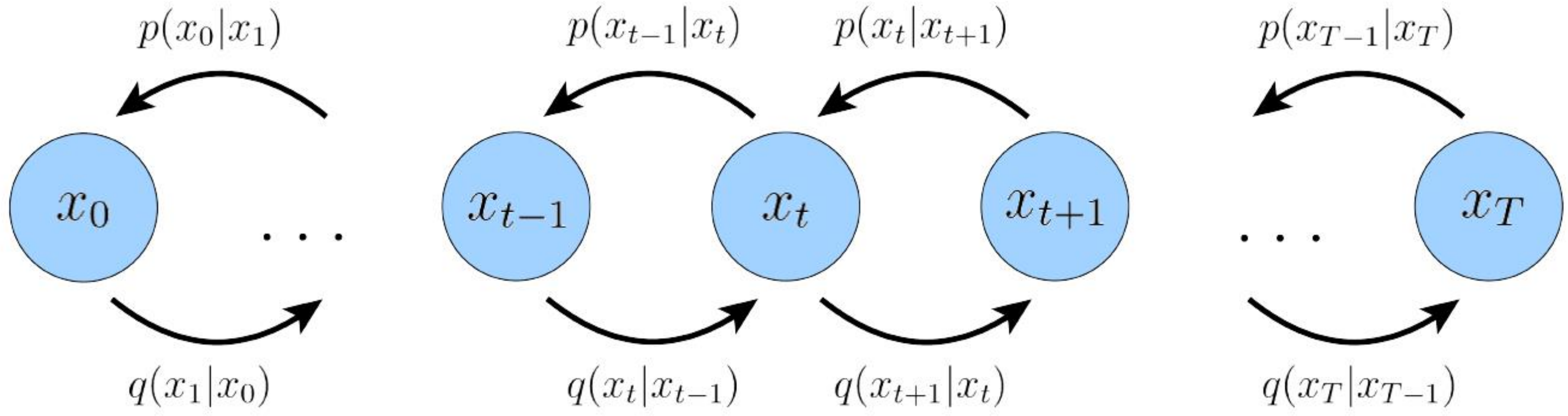
$$\mathbb{E}_{q_\phi(z_{1:T}|x)}\left[\log \frac{p(x,z_{1:T})}{q_\phi(z_{1:T}|x)}\right] = \mathbb{E}_{q_\phi(z_{1:T}|x)}\left[\log \frac{p(z_T)p_\theta(x|z_1)\prod_{t=2}^{T}p_\theta(z_{t-1}|z_t)}{q_\phi(z_1|x)\prod_{t=2}^{T}q_\phi(z_t|z_{t-1})}\right]$$

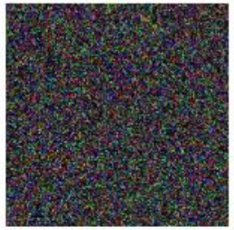
3. Variational Diffusion Models

VDM is as a Markovian HVAE with three key restrictions

3. Variational Diffusion Models

First, The latent dimension is exactly equal to the data dimension



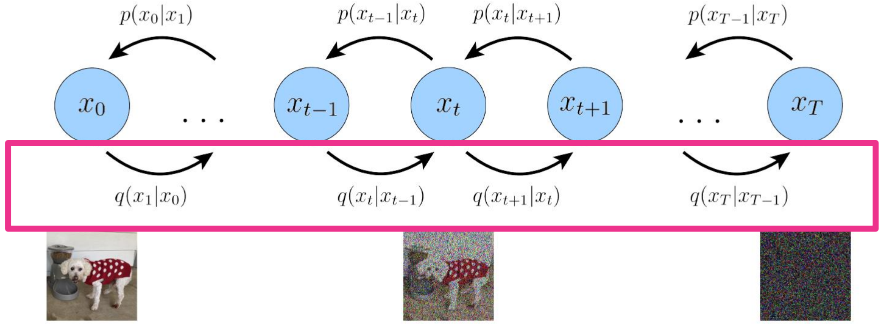


512*512 = 512*512

512*512 = 512*512

3. Variational Diffusion Models

Second, The structure of the latent encoder at each timestep **is not learned**



$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$

- it is pre-defined as a linear Gaussian model.

- It is a Gaussian distribution centered around the output of the previous timestep.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

- α_t is coefficient that can vary with the hierachical depth t

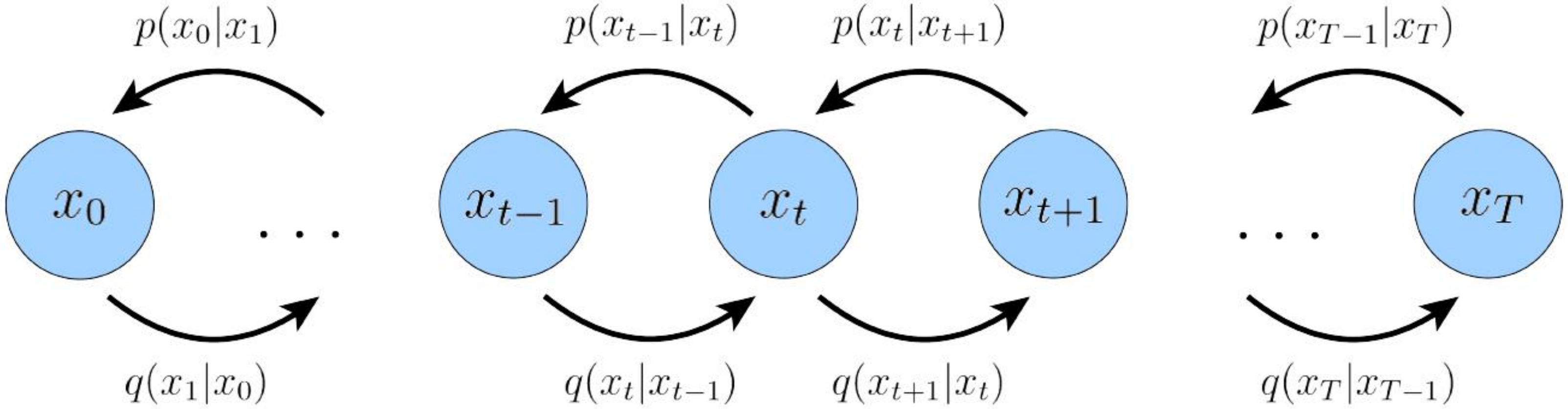
- Signal to noise ratio(SNR) must monotonically decrease over time

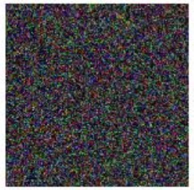
3. Variational Diffusion Models

Third, The Gaussian parameters of the latent encoders vary over time

- In such a way that the distribution of the latent at final timestep T is a standard Gaussian

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \boxed{\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}})$$





#. HOW Diffusion, WHY Gaussian?

Refered From

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

J Sohl-Dickstein 저술 · 2015

- inspired by **non-equilibrium statistical physics**, is to systematically and slowly destroy structure in a data distribution through an ***iterative forward diffusion process***.

- We then ***learn a reverse diffusion process*** that restores structure in data, yielding a highly flexible and tractable generative model of the data

		Gaussian	*Binomial*
Well behaved (analytically tractable) distribution	$\pi\left(\mathbf{x}^{(T)}\right) =$	$\mathcal{N}\left(\mathbf{x}^{(T)}; \mathbf{0}, \mathbf{I}\right)$	$\mathcal{B}\left(\mathbf{x}^{(T)}; 0.5\right)$
Forward diffusion kernel	$q\left(\mathbf{x}^{(t)} \vert \mathbf{x}^{(t-1)}\right) =$	$\mathcal{N}\left(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}\sqrt{1-\beta_t}, \mathbf{I}\beta_t\right)$	$\mathcal{B}\left(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}\left(1-\beta_t\right) + 0.5\beta_t\right)$
Reverse diffusion kernel	$p\left(\mathbf{x}^{(t-1)} \vert \mathbf{x}^{(t)}\right) =$	$\mathcal{N}\left(\mathbf{x}^{(t-1)}; \mathbf{f}_\mu\left(\mathbf{x}^{(t)}, t\right), \mathbf{f}_\Sigma\left(\mathbf{x}^{(t)}, t\right)\right)$	$\mathcal{B}\left(\mathbf{x}^{(t-1)}; \mathbf{f}_b\left(\mathbf{x}^{(t)}, t\right)\right)$

#. HOW Diffusion, WHY Gaussian?

Refered From

ON THE THEORY OF STOCHASTIC PROCESSES, WITH PARTICULAR REFERENCE TO APPLICATIONS

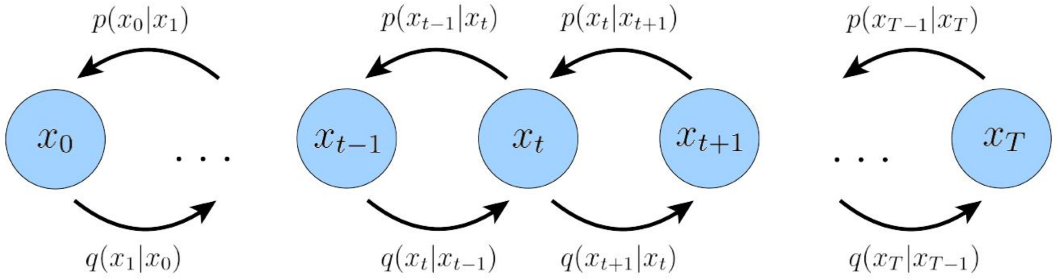
W. FELLER

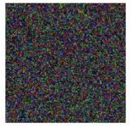
CORNELL UNIVERSITY

(ONR. Project for Research in Probability)

W Feller 저술 · 1949

HOW?
Markov process : differential
Forward, Backward의 **확률 분포가 같다.**
-> 원본 분포의 추정이 가능하다.





Markov processes leading to ordinary differential equations

Under these conditions[21] $u(\tau,\xi;t,x)$ satisfies the "forward equation"

$$u_t(\tau,\xi;t,x) = \tfrac{1}{2}\,[a(t,x)u(\tau,\xi;t,x)]_{xx} + [b(t,x)u(\tau,\xi;t,x)]_x,$$

and the "backward equation"

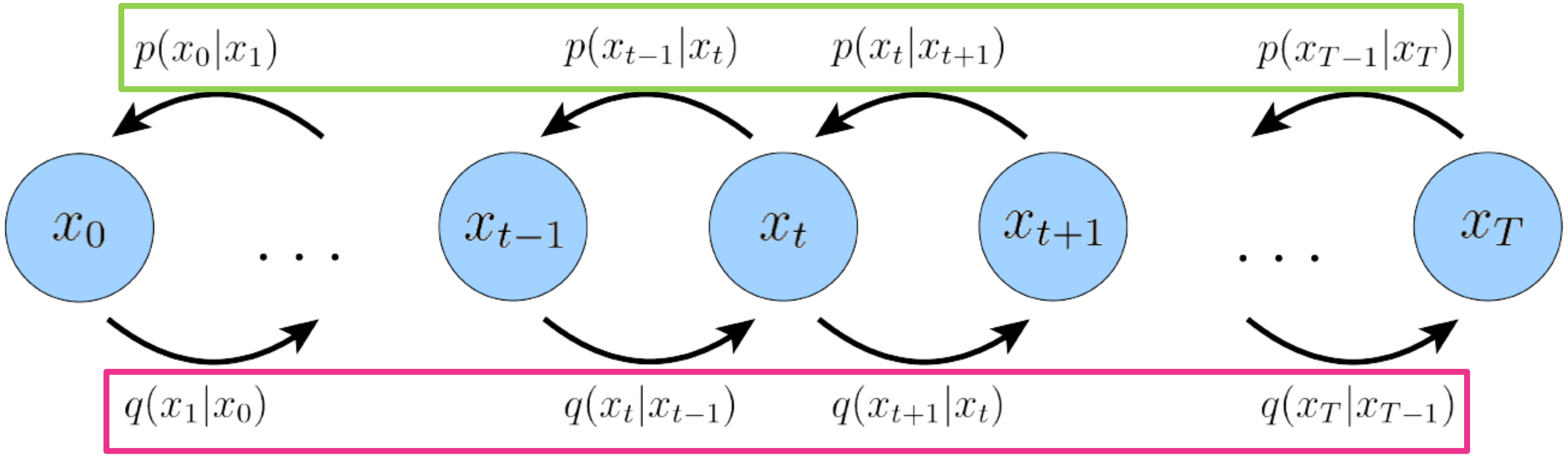
$$u_\tau(\tau,\xi;t,x) = \tfrac{1}{2}\,a(\tau,\xi)u_{\xi\xi}(\tau,\xi;t,x) + b(\tau,\xi)u_\xi(\tau,\xi;t,x);$$

3. Variational Diffusion Models

Basic Formula

$$\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{\boxed{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}}{\boxed{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}}\right]$$

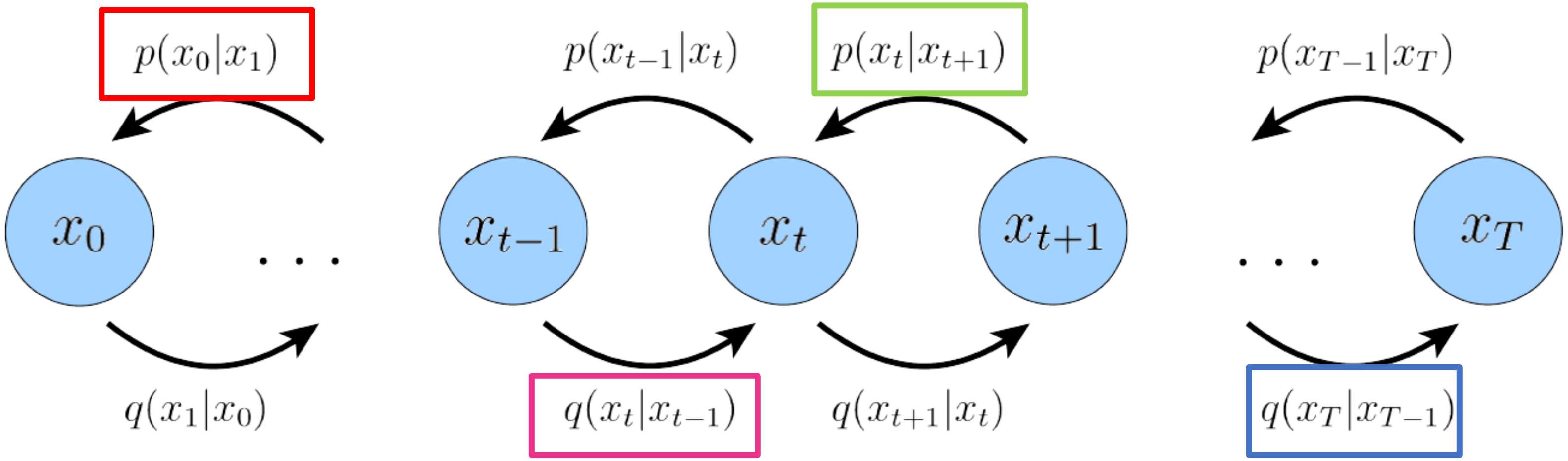


3. Variational Diffusion Models

Basic Formula

$$\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

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



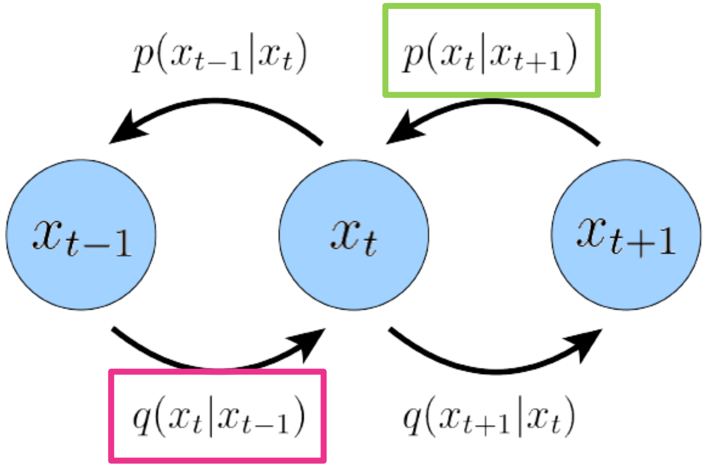
$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T))\right]}_{\text{prior matching term}}$$

$$- \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}$$

3. Variational Diffusion Models

Basic Formula

$$\sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(x_{t-1},x_{t+1}|x_0)}\left[D_{\mathrm{KL}}(\boxed{q(x_t|x_{t-1})} \parallel \boxed{p_\theta(x_t|x_{t+1})})\right]}_{\text{consistency term}}$$



However, actually optimizing the ELBO using the terms we just derived might be suboptimal; because the consistency term is computed as an expectation over two random variables.

the variance of its Monte Carlo estimate could potentially be higher than a term that is estimated using only one random variable per timestep.

Let us try to derive a form for our ELBO where each term is coputed as an expectation over only one random variable at a time

3. Variational Diffusion Models

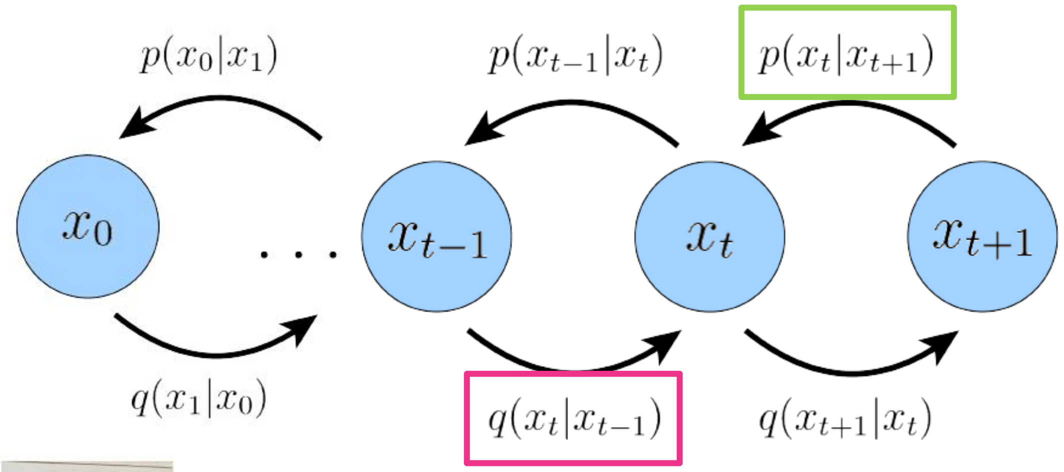
Basic Formula

The key insight is that we can rewrite encoder transitions

$$
\begin{aligned}
& q(x_t|x_{t-1}) \\
&= q(x_t|x_{t-1}, x_{t-2}) \\
&= q(x_t|x_{t-1}, x_{t-2}, x_1) \\
&= q(x_t|x_{t-1}, x_1, x_0) \\
&= q(x_t|x_{t-1}, x_0)
\end{aligned}
$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}$$

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



3. Variational Diffusion Models

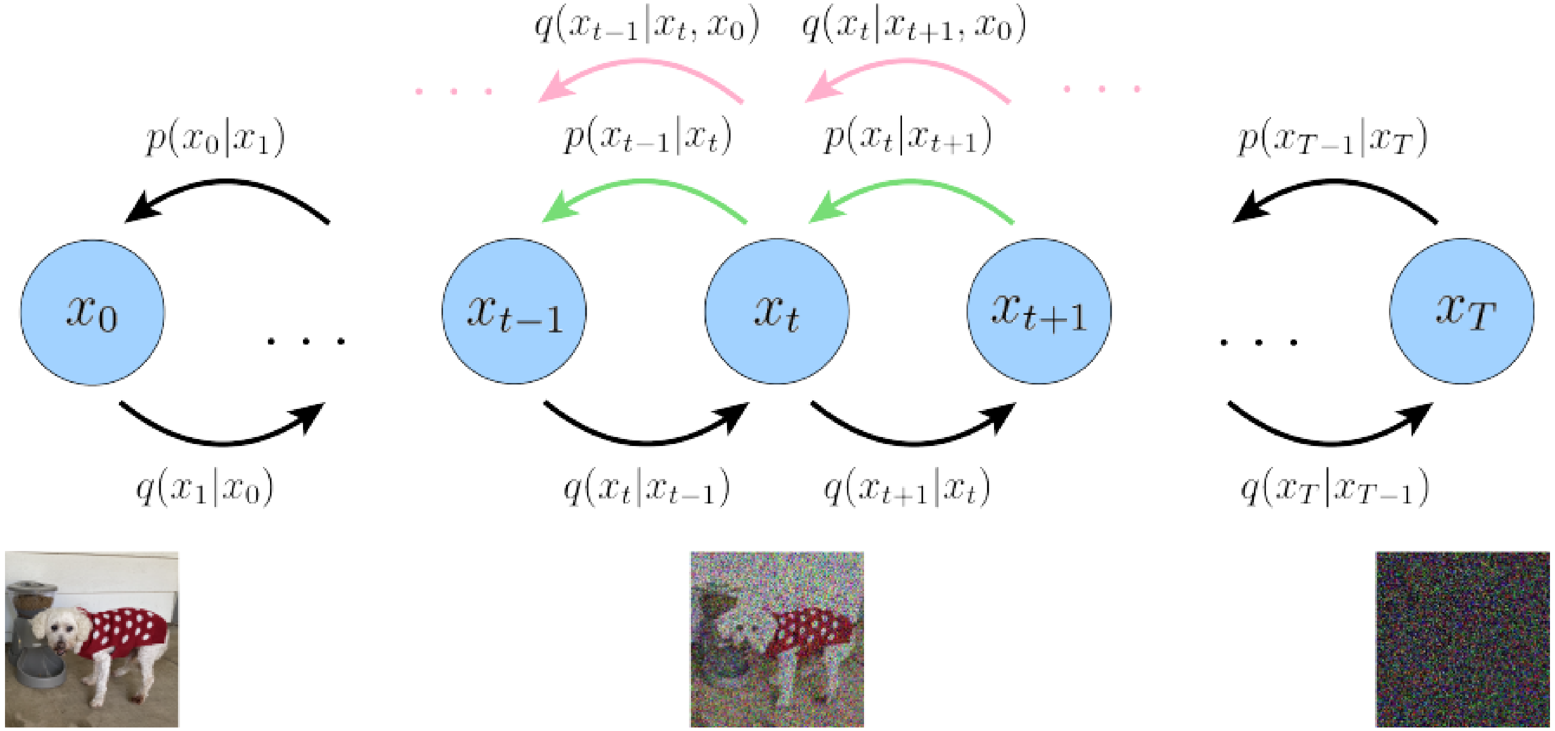
Basic Formula

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}$$

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

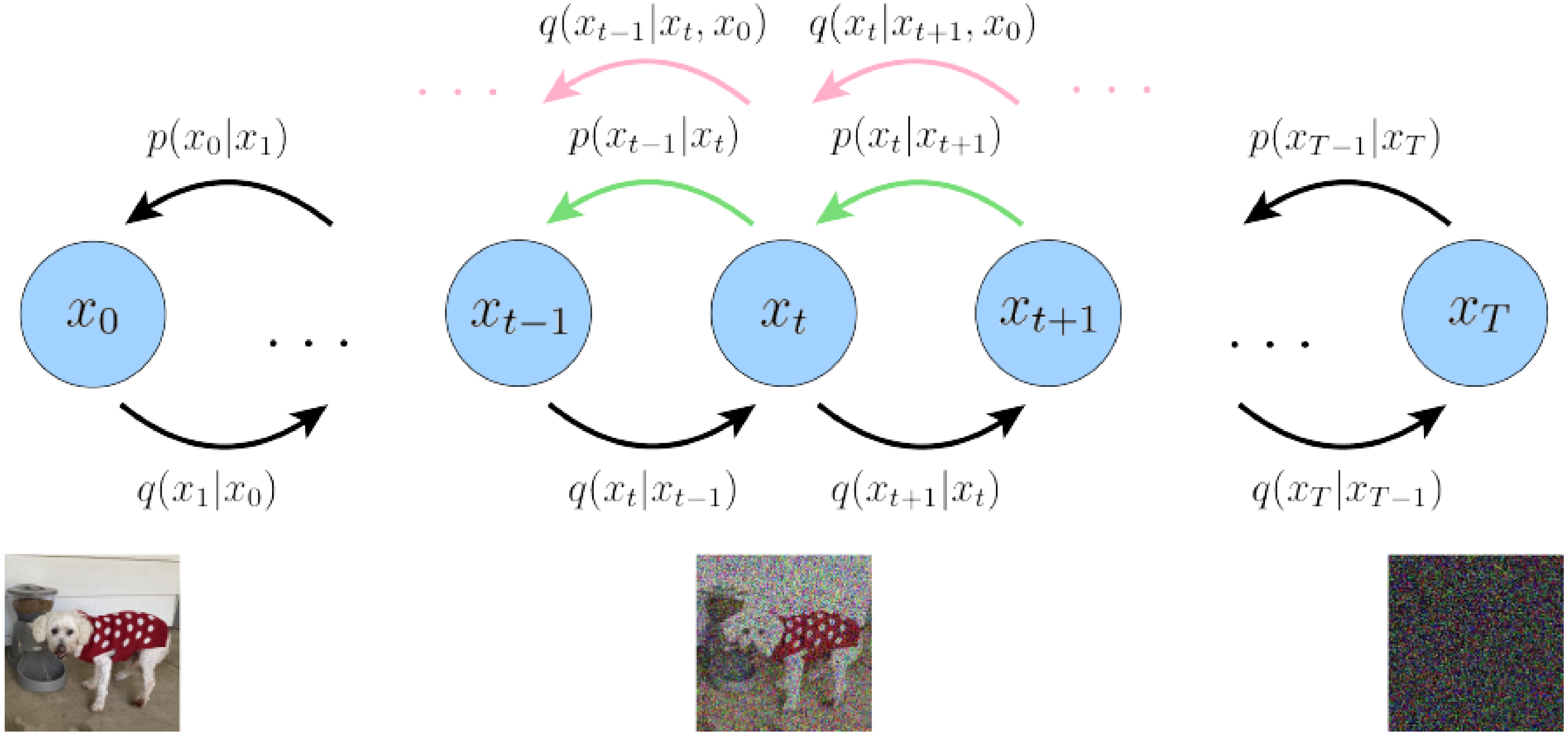
$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}}$$

$$- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(\boxed{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} \parallel \boxed{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)})\right]}_{\text{denoising matching term}}$$



3. Variationa

Basic Formula



Denoising

$$\sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(\boxed{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \parallel \boxed{p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})})\right]}_{\text{consistency term}}$$

$$\sum_{t=2} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(\boxed{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \parallel \boxed{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)})\right]}_{\text{denoising matching term}}$$

3. Variational Diffusion Models

Formula

$$[D_{\mathrm{KL}}(\boxed{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \parallel \boxed{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)})$$

$$\boxed{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)} = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}$$

$$\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$$

$$\boxed{p\left(\mathrm{x}^{(t-1)}|\mathrm{x}^{(t)}\right)} = \mathcal{N}\left(\mathrm{x}^{(t-1)}; \mathrm{f}_\mu\left(\mathrm{x}^{(t)}, t\right), \mathrm{f}_\Sigma\left(\mathrm{x}^{(t)}, t\right)\right)$$

$$u(\tau, \xi; t, x) = \frac{1}{2\{\pi(t-\tau)\}^{\frac{1}{2}}} \exp\left\{-\frac{(x-\xi)^2}{4(t-\tau)}\right\}$$

Gaussian distribution.

3. Variational Diffusion Models

$$u(\tau,\xi;t,x) = \frac{1}{2\{\pi(t-\tau)\}^{\frac{1}{2}}} \exp\left\{-\frac{(x-\xi)^2}{4(t-\tau)}\right\}$$

Gaussian distribution.

Formula

$$[D_{\mathrm{KL}}(\boxed{q(x_{t-1}|x_t, x_0)} \parallel \boxed{p_{\theta}(x_{t-1}|x_t)})$$

$$p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right) = \mathcal{N}\left(\mathbf{x}^{(t-1)}; \mathbf{f}_{\mu}\left(\mathbf{x}^{(t)}, t\right), \mathbf{f}_{\Sigma}\left(\mathbf{x}^{(t)}, t\right)\right)$$

$$\propto \mathcal{N}(x_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}}_{\mu_q(x_t, x_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$$

$$D_{\mathrm{KL}}(\mathcal{N}(x;\mu_x,\Sigma_x) \parallel \mathcal{N}(y;\mu_y,\Sigma_y)) = \frac{1}{2}\left[\log\frac{|\Sigma_y|}{|\Sigma_x|} - d + \mathrm{tr}(\Sigma_y^{-1}\Sigma_x) + (\mu_y-\mu_x)^T\Sigma_y^{-1}(\mu_y-\mu_x)\right]$$

$$\arg\min_{\theta} \frac{1}{2\sigma_q^2(t)}\left[\left\|\mu_{\theta}-\mu_q\right\|_2^2\right]$$

μ_{θ} as shorthand for $\mu_{\theta}(x_t, t)$

μ_q as shorthand for $\mu_q(x_t, x_0)$

3. Variational Diffusion Models

Formula

$$[D_{\mathrm{KL}}(\boxed{q(x_{t-1}|x_t, x_0)} \,\|\, \boxed{p_{\theta}(x_{t-1}|x_t)})$$

$$\arg\min_{\theta} \frac{1}{2\sigma_q^2(t)} \left[\|\mu_{\theta} - \mu_q\|_2^2 \right]$$

μ_{θ} as shorthand for $\mu_{\theta}(x_t, t)$

μ_q as shorthand for $\mu_q(x_t, x_0)$

$$\boxed{\mu_q(x_t, x_0)} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}$$

$$\boxed{\mu_{\theta}(x_t, t)} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{x}_{\theta}(x_t, t)}{1-\bar{\alpha}_t}$$

$\hat{x}_{\theta}(x_t, t)$ is parameterized by a neural network that seeks to predict original image from noisy image and time index.

3. Variational Diffusion Models

MI2L

Formula

$$\mu_q(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})x_t + \sqrt{\bar\alpha_{t-1}}(1-\alpha_t)x_0}{1-\bar\alpha_t}$$

μ_q as shorthand for $\mu_q(x_t, x_0)$

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})x_t + \sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\hat{x}_\theta(x_t, t)}{1-\bar\alpha_t}$$

μ_θ as shorthand for $\mu_\theta(x_t, t)$

$\hat{x}_\theta(x_t, t)$ is parameterized by a neural network that seeks to predict original image from noisy image and time index.

$$[D_{\mathrm{KL}}(\boxed{q(x_{t-1}|x_t, x_0)} \,\|\, \boxed{p_\theta(x_{t-1}|x_t)})$$

$$= \arg\min_\theta \frac{1}{2\sigma_q^2(t)}\left[\|\mu_\theta - \mu_q\|_2^2\right]$$

$$= \arg\min_\theta \frac{1}{2\sigma_q^2(t)} \frac{\bar\alpha_{t-1}(1-\alpha_t)^2}{(1-\bar\alpha_t)^2}\left[\|\hat{x}_\theta(x_t, t) - x_0\|_2^2\right] \rightarrow$$

Optimizing VDM boils down to **learning a neural network to predict the original ground truth image** from an arbitrarily noisified version of it

4. Three equivalent Interpretations

Formula

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\right\|_2^2\right]$$

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

$$\boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$$

$$= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$$

$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is parameterized by a neural network that learns to predict the source noise

$$\left[D_{\mathrm{KL}}\left(\boxed{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \,\|\, \boxed{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}\right)\right.$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\left[\left\|\boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right\|_2^2\right] \longrightarrow$$

learning a VDM by predicting the original image is equivalent to learning to predict the noise

울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

4. Three equivalent Interpretations

Formula

$$\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla \log p(\boldsymbol{x}_t)$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$$

$$\underline{\left[D_{\mathrm{KL}}\left(\boxed{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \parallel \boxed{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}\right)\right.}$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{\alpha_t}\left[\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla \log p(\boldsymbol{x}_t)\right\|_2^2\right]$$

Here, $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is a neural network that learns to predict the score function $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$, which is the gradient of $\boldsymbol{x}_t$ in data space, for any arbitrary noise level t.

4. Three equivalent Interpretations

Formula

Mathematically, for a Gaussian variable $z \sim \mathcal{N}(z; \mu_z, \Sigma_z)$

Tweedie's Formula states that : $\mathbb{E}[\mu_z|z] = z + \Sigma_z \nabla_z \log p(z)$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

$$\mathbb{E}[\boldsymbol{\mu}_{x_t}|\boldsymbol{x}_t] = \boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$$

$$\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 = \boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t)$$

$$\therefore \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}}$$

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla \log p(\boldsymbol{x}_t)$$

5. Conclusion

Formula

Three equivalent objectives to optimize a VDM:

1. To predict the original image
2. To predict the source noise
3. To predict $\nabla \log p(x_t)$ at an arbitrary noise level