Generative Model With Variational Inference
—-VAE and Diffusion Model

Yuzhou Nie(聂宇舟)
School of Statistics, Renmin University of China

Likelihood-based Model

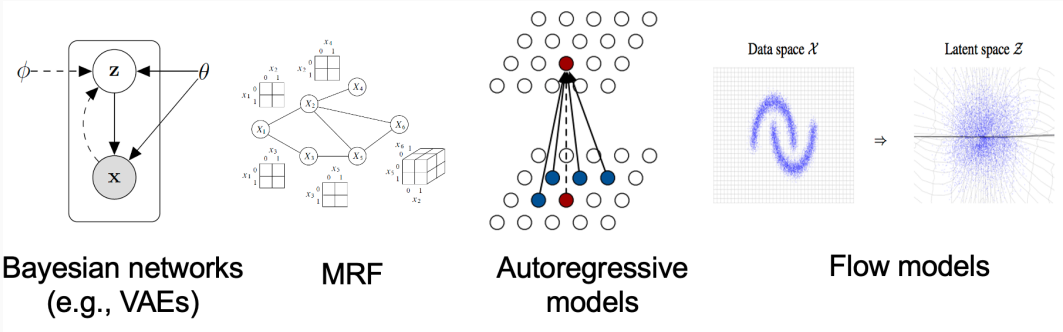

Figure 1: Typical Likelihood-based Models

Likelihood-based Model

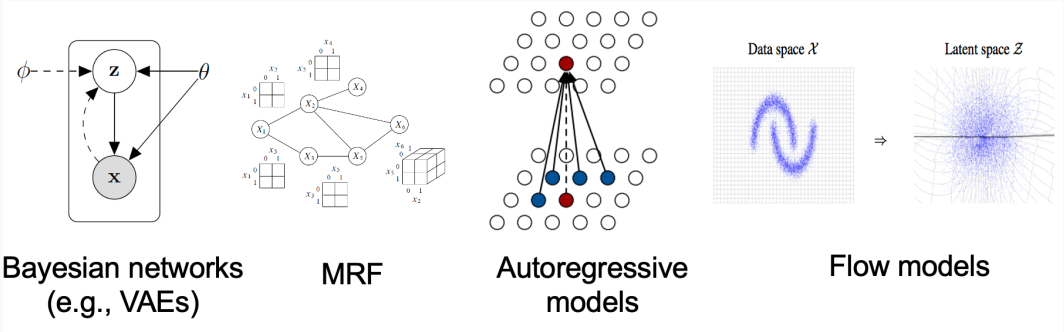

Figure 1: Typical Likelihood-based Models

Variational Inference

Contents

Some Important Notes

- Skip the details of derivation.

Some Important Notes

- Skip the details of derivation.
- Skip the details of basic proposition and conclusion in statistics.

Some Important Notes

- Skip the details of derivation.
- Skip the details of basic proposition and conclusion in statistics.
- Please remember important formulae.

Some Important Notes

- Skip the details of derivation.
- Skip the details of basic proposition and conclusion in statistics.
- Please remember important formulae.
- $p_\theta(x)$ can be regarded as $p(x \mid \theta)$ here.

Start From Maximum Likelihood Estimation

Suppose x are observed variables, $p(x)$ is empirical distribution of x, $p_\theta(x)$ is any parameterized distribution of x.

Start From Maximum Likelihood Estimation

Suppose x are observed variables, $p(x)$ is empirical distribution of x, $p_\theta(x)$ is any parameterized distribution of x.

$$p_\theta(x) \to p(x)$$

Start From Maximum Likelihood Estimation

Suppose x are observed variables, $p(x)$ is empirical distribution of x, $p_\theta(x)$ is any parameterized distribution of x.

$$p_\theta(x) \to p(x)$$

Maximize log likelihood estimation.

$$\theta = \arg\max_\theta \mathbb{E}_{x\sim p(x)}[\log p_\theta(x)dx] = \arg\max_\theta \int p(x) \log p_\theta(x)dx \tag{1}$$

Start From Maximum Likelihood Estimation

Suppose x are observed variables, $p(x)$ is empirical distribution of x, $p_\theta(x)$ is any parameterized distribution of x.

$$p_\theta(x) \to p(x)$$

Maximize log likelihood estimation.

$$\theta = \arg\max_\theta \mathbb{E}_{x\sim p(x)}[\log p_\theta(x) dx] = \arg\max_\theta \int p(x) \log p_\theta(x) dx \tag{1}$$

Minimize KL divergence $\mathbb{KL}(p(x)||p_\theta(x))$:

$$\theta = \arg\min_\theta \mathbb{KL}(p(x)||p_\theta(x)) = \arg\min_\theta \int p(x) \log \left(\frac{p(x)}{p_\theta(x)} \right) dx \tag{2}$$

Start From Maximum Likelihood Estimation

Suppose x are observed variables, $p(x)$ is empirical distribution of x, $p_\theta(x)$ is any parameterized distribution of x.

$$p_\theta(x) \to p(x)$$

Maximize log likelihood estimation.

$$\theta = \arg\max_\theta \mathbb{E}_{x\sim p(x)}[\log p_\theta(x)dx] = \arg\max_\theta \int p(x) \log p_\theta(x)dx \tag{1}$$

Minimize KL divergence $\mathbb{KL}(p(x)||p_\theta(x))$:

$$\theta = \arg\min_\theta \mathbb{KL}(p(x)||p_\theta(x)) = \arg\min_\theta \int p(x) \log \left(\frac{p(x)}{p_\theta(x)}\right) dx \tag{2}$$

Hard to optimize

Variational Inference

Introduce z as unobserved variables, we infer $p(z \mid x)$

Variational Inference

Introduce z as unobserved variables, we infer $p(z \mid x)$

$$p(z \mid x) = \frac{p(x, z)}{\underbrace{\int_z p(x, z) dz}_{\text{intractable}}} \tag{3}$$

Variational Inference

Introduce z as unobserved variables, we infer $p(z \mid x)$

$$p(z \mid x) = \frac{p(x, z)}{\underbrace{\int_z p(x, z) dz}_{\text{intractable}}} \tag{3}$$

Possible solution: $q_\phi(z \mid x) \to p(z \mid x)$

Variational Inference

Introduce z as unobserved variables, we infer $p(z \mid x)$

$$p(z \mid x) = \frac{p(x, z)}{\underbrace{\int_z p(x, z) dz}_{\text{intractable}}} \tag{3}$$

Possible solution: $q_\phi(z \mid x) \to p(z \mid x)$

Minimize KL divergence:

$$\phi = \arg\min_\phi \mathbb{KL}(q_\phi(z \mid x) \| p(z \mid x)) \tag{4}$$

Variational Inference

Introduce z as unobserved variables, we infer $p(z \mid x)$

$$p(z \mid x) = \frac{p(x, z)}{\underbrace{\int_z p(x, z) dz}_{\text{intractable}}} \tag{3}$$

Possible solution: $q_\phi(z \mid x) \to p(z \mid x)$

Minimize KL divergence:

$$\phi = \arg\min_\phi \mathbb{KL}(q_\phi(z \mid x) \| p(z \mid x)) \tag{4}$$

$$\begin{aligned} \mathbb{KL}(q_\phi(z \mid x) \| p(z \mid x)) &= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{q_\phi(z \mid x)}{p(z \mid x)} \right) \right] \\ &= \log p(x) - \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p(x, z)}{q_\phi(z \mid x)} \right) \right] \end{aligned} \tag{5}$$

Variational Inference

$$\min \mathbb{KL}(q_\phi(z \mid x) \| p(z \mid x)) = \max \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p(x,z)}{q_\phi(z \mid x)} \right) \right]}_{ELBO(q_\phi)} \tag{6}$$

Variational Inference

$$\min \mathbb{KL}(q_\phi(z \mid x) \| p(z \mid x)) = \max \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p(x,z)}{q_\phi(z \mid x)} \right) \right]}_{ELBO(q_\phi)} \tag{6}$$

Evidence Lower Bound

Variational Inference

$$\min \mathbb{KL}(q_\phi(z \mid x) \| p(z \mid x)) = \max \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p(x,z)}{q_\phi(z \mid x)} \right) \right]}_{ELBO(q_\phi)} \tag{6}$$

Evidence Lower Bound

Further derive:

$$\begin{aligned} ELBO(q_\phi) &= \mathbb{E}_{q_\phi}[\log p(z)] + \mathbb{E}_{q_\phi}[\log p(x \mid z)] - \mathbb{E}_{q_\phi}[\log q_\phi(z \mid x)] \\ &= \underbrace{\mathbb{E}_{q_\phi}[\log p(x \mid z)]}_{\textit{likelihood}} - \mathbb{KL}(\underbrace{q_\phi(z \mid x)}_{\text{variational}} \| \underbrace{p(z)}_{\text{prior}}) \end{aligned} \tag{7}$$

Variational Inference From Generative Model

$\log p_\theta(x) \to \log p(x)$

Variational Inference From Generative Model

$\log p_\theta(x) \to \log p(x)$

suppose $p_\theta(x) = \int p_\theta(x, z) dz$, we have

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{p_\theta(z \mid x)} \right) \right] \\
&= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q_\phi(z \mid x)} \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \right) \right] \\
&= \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right) \right]}_{ELBO(q_\phi)} + \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \right) \right]}_{\mathbb{KL}\left(q_\phi(z|x) \| p_\theta(z|x)\right)}
\end{aligned}
\tag{8}
$$

Variational Inference From Generative Model

$\log p_\theta(x) \to \log p(x)$

suppose $p_\theta(x) = \int p_\theta(x, z)dz$, we have

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{p_\theta(z \mid x)}\right)\right] \\
&= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z \mid x)}\frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}\right)\right] \\
&= \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right)\right]}_{ELBO(q_\phi)} + \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\left(\frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}\right)\right]}_{\mathbb{KL}\left(q_\phi(z|x)\|p_\theta(z|x)\right)}
\end{aligned}
\tag{8}
$$

$\log p_\theta(x) \geq ELBO(q_\phi)$, equal when $q_\phi(z \mid x) = p_\theta(z \mid x)$

Variational Inference From Generative Model

$\log p_\theta(x) \to \log p(x)$

suppose $p_\theta(x) = \int p_\theta(x, z) dz$, we have

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{p_\theta(z \mid x)} \right) \right] \\
&= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q_\phi(z \mid x)} \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \right) \right] \\
&= \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right) \right]}_{ELBO(q_\phi)} + \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \right) \right]}_{\mathbb{KL}\left(q_\phi(z|x) \| p_\theta(z|x)\right)}
\end{aligned}
\tag{8}
$$

$\log p_\theta(x) \geq ELBO(q_\phi)$, equal when $q_\phi(z \mid x) = p_\theta(z \mid x)$

ELBO is a lower bound for $\log p_\theta(x)$

Variational Inference

For KL divergence:

Variational Inference

For KL divergence:

$$\log p_\theta(x) \geq ELBO(q_\phi) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)})]$$
$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)p(x)})]] \quad (9)$$

Variational Inference

For KL divergence:

$$\log p_\theta(x) \geq ELBO(q_\phi) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)})]$$
$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)p(x)})]] \tag{9}$$

According to $q_\phi(z \mid x) \to p(z \mid x)$, we have $p(x,z) = p(z \mid x)p(x) \approx q_\phi(z \mid x)p(x)$

Variational Inference

For KL divergence:

$$\log p_\theta(x) \geq ELBO(q_\phi) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)})]$$
$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)p(x)})]] \tag{9}$$

According to $q_\phi(z \mid x) \to p(z \mid x)$, we have $p(x,z) = p(z \mid x)p(x) \approx q_\phi(z \mid x)p(x)$

$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)p(x)})]]$$
$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x,z\sim p(x,z)}[\log \frac{p_\theta(x,z)}{p(x,z)}] \tag{10}$$
$$\mathbb{KL}(p(x)\|p_\theta(x)) \leq \mathbb{KL}(p(x,z)\|p_\theta(x,z))$$

Variational Inference

For KL divergence:

$$\log p_\theta(x) \geq ELBO(q_\phi) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)})]$$
$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)p(x)})]] \tag{9}$$

According to $q_\phi(z \mid x) \to p(z \mid x)$, we have $p(x,z) = p(z \mid x)p(x) \approx q_\phi(z \mid x)p(x)$

$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{z\sim q_\phi(z|x)}[\log(\frac{p_\theta(x,z)}{q_\phi(z \mid x)p(x)})]]$$
$$\mathbb{E}_{x\sim p(x)}[\log \frac{p_\theta(x)}{p(x)}] \geq \mathbb{E}_{x,z\sim p(x,z)}[\log \frac{p_\theta(x,z)}{p(x,z)}] \tag{10}$$
$$\mathbb{KL}(p(x)\|p_\theta(x)) \leq \mathbb{KL}(p(x,z)\|p_\theta(x,z))$$

$\mathbb{KL}(p(x,z)\|p_\theta(x,z))$ **is an upper bound for** $\mathbb{KL}(p(x)\|p_\theta(x))$

VAE Framework

VAE Framework

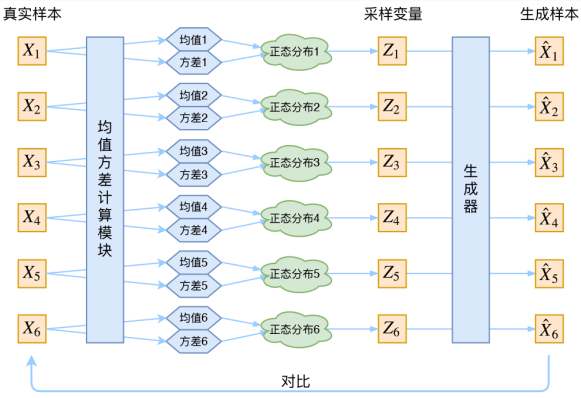

Figure 2: Variational Autoencoder

VAE Framework

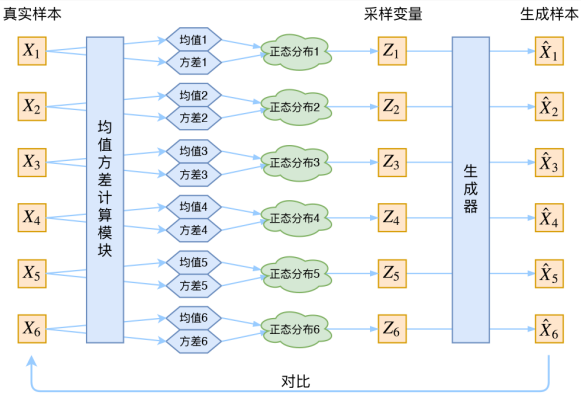

Figure 2: Variational Autoencoder

Objective: loss=reconstruction loss$(X, \hat{X})$ + regularization$(\mathbb{KL}(N(\mu, \sigma), N(0, 1)))$

VAE

$\mathbb{KL}(p(x,z)||p_\theta(x,z))$ is an upper bound for $\mathbb{KL}(p(x)||p_\theta(x))$

VAE

$\mathbb{KL}(p(x,z)\|p_\theta(x,z))$ is an upper bound for $\mathbb{KL}(p(x)\|p_\theta(x))$

We assume

$$p_\theta(x,z) = p_\theta(x \mid z)q(z), p(x,z) = p(x)q_\phi(z \mid x)$$

where $p_\theta(x \mid z), q_\phi(z \mid x)$ are Gaussian distribution with unknown parameters, $q(z)$ is standard Gaussian distribution.

VAE

$\mathbb{KL}(p(x,z)\|p_\theta(x,z))$ is an upper bound for $\mathbb{KL}(p(x)\|p_\theta(x))$

We assume

$$p_\theta(x,z) = p_\theta(x \mid z)q(z), p(x,z) = p(x)q_\phi(z \mid x)$$

where $p_\theta(x \mid z), q_\phi(z \mid x)$ are Gaussian distribution with unknown parameters, $q(z)$ is standard Gaussian distribution.

Remark

With $p_\theta(x \mid z)$ and $q(z)$, the better approximation for $p(z \mid x)$ should be:

$$p(z \mid x) = \frac{p_\theta(x \mid z)q(z)}{p(x)} = \frac{p_\theta(x \mid z)q(z)}{\underbrace{\int p_\theta(x \mid z)q(z)dz}_{intractable}} \tag{11}$$

VAE Loss Function

Objective function:

$$\mathbb{KL}(p(x,z)\|p_\theta(x,z)) = \iint p(x)q_\phi(z \mid x) \log \frac{p(x)q_\phi(z \mid x)}{p_\theta(x \mid z)q(z)} dxdz \tag{12}$$

VAE Loss Function

Objective function:

$$\mathbb{KL}(p(x,z)\|p_\theta(x,z)) = \iint p(x)q_\phi(z \mid x)\log\frac{p(x)q_\phi(z \mid x)}{p_\theta(x \mid z)q(z)}dxdz \tag{12}$$

$\log p(x)$ is irrelevant, we have

$$\mathbb{E}_{x\sim p(x)}\left[-\int q_\phi(z \mid x)\log p_\theta(x \mid z)dz + \mathbb{KL}(q_\phi(z \mid x)\|q(z))\right] \tag{13}$$

VAE Loss Function

Objective function:

$$\mathbb{KL}(p(x,z)\|p_\theta(x,z)) = \iint p(x)q_\phi(z \mid x)\log\frac{p(x)q_\phi(z \mid x)}{p_\theta(x \mid z)q(z)}dxdz \tag{12}$$

$\log p(x)$ is irrelevant, we have

$$\mathbb{E}_{x\sim p(x)}\left[-\int q_\phi(z \mid x)\log p_\theta(x \mid z)dz + \mathbb{KL}(q_\phi(z \mid x)\|q(z))\right] \tag{13}$$

If we generate one sample for $q_\phi(z \mid x)$ each time

$$\mathbb{E}_{x\sim p(x)}[\underbrace{-\log p_\theta(x \mid z)}_{\text{reconstruction}} + \underbrace{\mathbb{KL}(q_\phi(z \mid x)\|q(z))}_{\text{regularization}}] \tag{14}$$

EM Algorithm

Assumption for VAE: $p_\theta(x \mid z), q_\phi(z \mid x)$ are Gaussian distributions.

EM Algorithm

~~Assumption for VAE: $p_\theta(x \mid z), q_\phi(z \mid x)$ are Gaussian distributions.~~

Directly optimize $\mathbb{KL}(p(x, z) \| p_\theta(x, z))$

EM Algorithm

Directly optimize $\mathbb{KL}(p(x,z)\|p_\theta(x,z))$

Objective function:

$$\mathbb{KL}(p(x,z)\|p_\theta(x,z)) = \iint p(x)q_\phi(z \mid x)\log\frac{p(x)q_\phi(z \mid x)}{p_\theta(x \mid z)q(z)}dxdz$$

Alternating Optimization

1. Fix $q_\phi(z \mid x)$, optimize $p_\theta(x \mid z)$
   We have

$$p_\theta(x \mid z) = \arg\max_{p_\theta(x|z)} \mathbb{E}_{x\sim p(x)}\left[\int q_\phi(z \mid x)\log p(x,z)dz\right] \tag{15}$$

2. ...

Alternating Algorithm

$$\mathbb{KL}(p(x,z)||p_\theta(x,z)) = \int\int p(x)q_\phi(z \mid x) \log \frac{p(x)q_\phi(z|x)}{p_\theta(x|z)q(z)} dxdz$$

Alternating Algorithm

$$\mathbb{KL}(p(x,z)\|p_\theta(x,z)) = \iint p(x)q_\phi(z \mid x) \log \frac{p(x)q_\phi(z|x)}{p_\theta(x|z)q(z)} dxdz$$

Alternating Optimization

1. ...
2. Fix $p_\theta(x \mid z)$, so $p(x,z)$ is fixed, then optimize $q_\phi(z \mid x)$
   define $p_\theta(x \mid z)q(z) = p_\theta(z \mid x)p_\theta(x)$

$$p_\theta(x) = \int p_\theta(x \mid z)q(z)dz, \quad p_\theta(z \mid x) = \frac{p_\theta(x \mid z)q(z)}{p_\theta(x)} \tag{16}$$

Then we have

$$\begin{aligned} q_\phi(z \mid x) &= \arg\min_{q_\phi(z|x)} \mathbb{E}_{x\sim p(x)} \left[\int q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)p_\theta(x)} dz \right] \\ &= \arg\min_{q_\phi(z|x)} \mathbb{E}_{x\sim p(x)} [\mathbb{KL}(q_\phi(z \mid x)\|p_\theta(z \mid x)) - \log p_\theta(x)] \end{aligned} \tag{17}$$

Alternating Algorithm

$q_\phi(z \mid x)$ have optimal solution :

Alternating Algorithm

$q_\phi(z \mid x)$ have optimal solution:

$$q_\phi(z \mid x) = p_\theta(z \mid x) = \frac{p_\theta(x \mid z)q(z)}{\int p_\theta(x \mid z)q(z)dz} \tag{18}$$

Alternating Algorithm

$q_\phi(z \mid x)$ have optimal solution:

$$q_\phi(z \mid x) = p_\theta(z \mid x) = \frac{p_\theta(x \mid z)q(z)}{\int p_\theta(x \mid z)q(z)dz} \tag{18}$$

$$\begin{aligned} p_\theta(x \mid z) &= \arg\max_{p_\theta(x|z)} \mathbb{E}_{x\sim p(x)} \left[\int q_\phi(z \mid x) \log p(x,z) dz \right] \\ &= \arg\max_{\theta} \mathop{\mathbb{E}}_{z\sim p_\theta(z|x)} [\log p(x,z)] \end{aligned}$$

Alternating Algorithm

$q_\phi(z \mid x)$ have optimal solution:

$$q_\phi(z \mid x) = p_\theta(z \mid x) = \frac{p_\theta(x \mid z)q(z)}{\int p_\theta(x \mid z)q(z)dz} \tag{18}$$

$$\begin{aligned} p_\theta(x \mid z) &= \arg\max_{p_\theta(x|z)} \mathbb{E}_{x \sim p(x)} \left[\int q_\phi(z \mid x) \log p(x, z) dz \right] \\ &= \arg\max_{\theta} \mathop{\mathbb{E}}_{z \sim p_\theta(z|x)} [\log p(x, z)] \end{aligned}$$

We get EM Algorithm by alternately computing these two formulae.

Diffusion Model: Autoregressive VAE

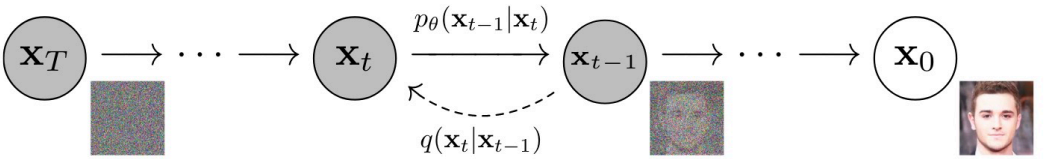

Figure 2: The directed graphical model considered in this work.

- Forward process: From right to left, model $x_0, x_1, \cdots, x_T$ as a Markov Chain: $x_t = \alpha_t x_{t-1} + \beta_t \quad \text{where} \quad \alpha_t^2 + \beta_t^2 = 1$
- Reverse process: From left to right, learn $p_\theta(x_{t-1} \mid x_t)$ to fit $q(x_{t-1} \mid x_t)$.

Diffusion Model: Autoregressive VAE

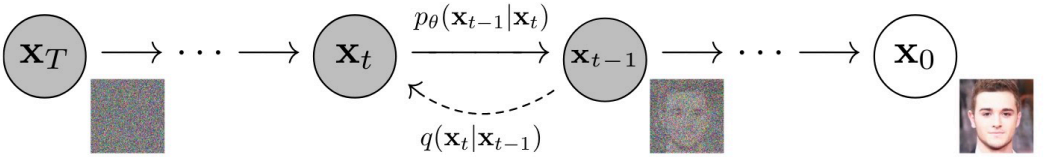

Figure 2: The directed graphical model considered in this work.

- Forward process: From right to left, model $x_0, x_1, \cdots, x_T$ as a Markov Chain: $x_t = \alpha_t x_{t-1} + \beta_t \quad \text{where} \quad \alpha_t^2 + \beta_t^2 = 1$
- Reverse process: From left to right, learn $p_\theta(x_{t-1} \mid x_t)$ to fit $q(x_{t-1} \mid x_t)$.

$q(x_{t-1} \mid x_t, x_0)$ **can be represented by** $q(x_t \mid x_0)$ **and** $q(x_t \mid x_{t-1})$**.**

$q(x_{t-1} \mid x_t, x_0)$ **can be used to learn** $p_\theta(x_{t-1} \mid x_t)$

Diffusion Model: Autoregressive VAE

VAE

- encode: $x \to z$
- decode: $z \to x$

Diffusion Model: Autoregressive VAE

VAE

- encode: $x \to z$
- decode: $z \to x$

Diffusion Model

- encode: $x = x_0 \to x_1 \to x_2 \to \cdots \to x_{T-1} \to x_T = z$
- decode: $z = x_T \to x_{T-1} \to x_{T-2} \to \cdots \to x_1 \to x_0 = x$

Diffusion Model: Autoregressive VAE

VAE

- encode: $x \to z$
- decode: $z \to x$

Diffusion Model

- encode: $x = x_0 \to x_1 \to x_2 \to \cdots \to x_{T-1} \to x_T = z$
- decode: $z = x_T \to x_{T-1} \to x_{T-2} \to \cdots \to x_1 \to x_0 = x$

Drawback for VAE: encoder and decoder are both Gaussian distributions, the model is not complicated enough.

Diffusion Model: Autoregressive VAE

VAE

- encode: $x \to z$
- decode: $z \to x$

Diffusion Model

- encode: $x = x_0 \to x_1 \to x_2 \to \cdots \to x_{T-1} \to x_T = z$
- decode: $z = x_T \to x_{T-1} \to x_{T-2} \to \cdots \to x_1 \to x_0 = x$

Drawback for VAE: encoder and decoder are both Gaussian distributions, the model is not complicated enough.

By modeling each step as a Gaussian distribution, the model becomes more complicated intuitively.

Model the Joint Distributions

For forward process, we define $q\left(x_t \mid x_{t-1}\right)$

For reverse process, we define $p_\theta\left(x_{t-1} \mid x_t\right)$

Model the Joint Distributions

For forward process, we define $q\left(x_t \mid x_{t-1}\right)$

For reverse process, we define $p_\theta\left(x_{t-1} \mid x_t\right)$

For Markov Chain assumption, we have

$$
\begin{aligned}
q\left(x_0, x_1, x_2, \cdots, x_T\right) &= q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_2 \mid x_1\right) q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \\
p_\theta\left(x_0, x_1, x_2, \cdots, x_T\right) &= p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-2} \mid x_{T-1}\right) p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right)
\end{aligned} \tag{19}
$$

Model the Joint Distributions

For forward process, we define $q\left(x_t \mid x_{t-1}\right)$

For reverse process, we define $p_\theta\left(x_{t-1} \mid x_t\right)$

For Markov Chain assumption, we have

$$
\begin{aligned}
q\left(x_0, x_1, x_2, \cdots, x_T\right) &= q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_2 \mid x_1\right) q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \\
p_\theta\left(x_0, x_1, x_2, \cdots, x_T\right) &= p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-2} \mid x_{T-1}\right) p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right)
\end{aligned} \tag{19}
$$

We use the KL divergence between two joint distributions as our objective.

$$
\mathbb{KL}(q \| p_\theta) = \int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log \frac{q\left(x_T \mid x_{T-1}\right) \cdots p\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right)}{p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right)} dx_0 \cdots dx_T \tag{20}
$$

Derive the KL divergence

$$
\begin{aligned}
&\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log \frac{q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right)}{p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right)} dx_0 dx_1 \cdots dx_T \\
&= -\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right) dx_0 dx_1 \cdots dx_T \\
&= -\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \left[\log p_\theta\left(x_T\right) + \sum_{t=1}^{T} \log p_\theta\left(x_{t-1} \mid x_t\right)\right] dx_0 dx_1 \cdots dx_T
\end{aligned} \tag{21}
$$

Derive the KL divergence

$$
\begin{aligned}
&\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log \frac{q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right)}{p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right)} dx_0 dx_1 \cdots dx_T \\
=&-\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right) dx_0 dx_1 \cdots dx_T \\
=&-\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \left[\log p_\theta\left(x_T\right)+\sum_{t=1}^T \log p_\theta\left(x_{t-1} \mid x_t\right)\right] dx_0 dx_1 \cdots dx_T
\end{aligned} \tag{21}
$$

Typically, $p_\theta(x_T)$ is defined as standard Gaussian distribution.

For each $p_\theta(x_{t-1} \mid x_t)$, we have

$$
-\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_1 \cdots dx_T \tag{22}
$$

Derive the KL divergence

$$
\begin{aligned}
&\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log \frac{q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right)}{p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right)} dx_0 dx_1 \cdots dx_T \\
=&-\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_0 \mid x_1\right) \cdots p_\theta\left(x_{T-1} \mid x_T\right) p_\theta\left(x_T\right) dx_0 dx_1 \cdots dx_T \\
=&-\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right)\left[\log p_\theta\left(x_T\right)+\sum_{t=1}^T \log p_\theta\left(x_{t-1} \mid x_t\right)\right] dx_0 dx_1 \cdots dx_T
\end{aligned} \tag{21}
$$

Typically, $p_\theta(x_T)$ is defined as standard Gaussian distribution.

For each $p_\theta(x_{t-1} \mid x_t)$, we have

$$
-\int q\left(x_T \mid x_{T-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_1 \cdots dx_T \tag{22}
$$

$p_\theta(x_{t-1}, x_t)$ is only related to $x_0 \cdots x_t$, so

$$
=-\int q\left(x_t \mid x_{t-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_1 \cdots dx_t \tag{23}
$$

Derive the KL divergence

$$= -\int q\left(x_t \mid x_{t-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_1 \cdots dx_t$$

Derive the KL divergence

$$= -\int q\left(x_t \mid x_{t-1}\right) \cdots q\left(x_1 \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_1 \cdots dx_t$$

According to the following proposition, we have

$$= -\int q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1} \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_{t-1} dx_t \tag{24}$$

Derive the KL divergence

$$= -\int q(x_t \mid x_{t-1}) \cdots q(x_1 \mid x_0)\, \tilde{q}(x_0) \log p_\theta(x_{t-1} \mid x_t)\, dx_0 dx_1 \cdots dx_t$$

According to the following proposition, we have

$$= -\int q(x_t \mid x_{t-1})\, q(x_{t-1} \mid x_0)\, \tilde{q}(x_0) \log p_\theta(x_{t-1} \mid x_t)\, dx_0 dx_{t-1} dx_t \tag{24}$$

Proposition

Given by [Ho, Neurips 2020], we have

$$q(x_t \mid x_0) = \int q(x_t \mid x_{t-1}) \cdots q(x_1 \mid x_0)\, dx_1 \cdots dx_{t-1} = \mathcal{N}\left(x_t; \bar{\alpha}_t x_0, \bar{\beta}_t^2 I\right) \tag{25}$$

where $x_t = \alpha_t x_{t-1} + \beta_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, I), \quad \bar{\alpha}_t = \alpha_1 \cdots \alpha_t, \quad \bar{\beta}_t = \sqrt{1 - (\alpha_t \cdots \alpha_1)^2}$

Derive the KL divergence

$$= -\int q(x_t \mid x_{t-1}) \cdots q(x_1 \mid x_0)\, \tilde{q}(x_0) \log p_\theta(x_{t-1} \mid x_t)\, dx_0 dx_1 \cdots dx_t$$

According to the following proposition, we have

$$= -\int q(x_t \mid x_{t-1})\, q(x_{t-1} \mid x_0)\, \tilde{q}(x_0) \log p_\theta(x_{t-1} \mid x_t)\, dx_0 dx_{t-1} dx_t \tag{24}$$

Proposition

Given by [Ho, Neurips 2020], we have

$$q(x_t \mid x_0) = \int q(x_t \mid x_{t-1}) \cdots q(x_1 \mid x_0)\, dx_1 \cdots dx_{t-1} = \mathcal{N}\left(x_t; \bar{\alpha}_t x_0, \bar{\beta}_t^2 I\right) \tag{25}$$

where $x_t = \alpha_t x_{t-1} + \beta_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, I), \quad \bar{\alpha}_t = \alpha_1 \cdots \alpha_t, \quad \bar{\beta}_t = \sqrt{1 - (\alpha_t \cdots \alpha_1)^2}$

So $p_\theta(x_{t-1} \mid x_t)$ is irrelevant to $x_1, \cdots, x_{t-2}$ and we have (24).

Objective Function

$$= -\int q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1} \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_{t-1} dx_t$$

1. Ignore those constants irrelevant to optimization objective, $-\log p_\theta\left(x_{t-1} \mid x_t\right)$ is $\frac{1}{2\sigma_t^2}\left\|x_{t-1} - \mu\left(x_t\right)\right\|^2$

Objective Function

$$= - \int q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1} \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_{t-1} dx_t$$

1. Ignore those constants irrelevant to optimization objective, $-\log p_\theta\left(x_{t-1} \mid x_t\right)$ is $\frac{1}{2\sigma_t^2}\left\|x_{t-1} - \mu\left(x_t\right)\right\|^2$
2. With $x_{t-1} = \frac{1}{\alpha_t}\left(x_t - \beta_t \varepsilon_t\right)$, we can parameterized $\mu\left(x_t\right)$ as $\mu\left(x_t\right) = \frac{1}{\alpha_t}\left(x_t - \beta_t \epsilon_\theta\left(x_t, t\right)\right.$

Objective Function

$$= - \int q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1} \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_{t-1} dx_t$$

1. Ignore those constants irrelevant to optimization objective, $-\log p_\theta\left(x_{t-1} \mid x_t\right)$ is $\frac{1}{2\sigma_t^2}\left\|x_{t-1} - \mu\left(x_t\right)\right\|^2$
2. With $x_{t-1} = \frac{1}{\alpha_t}\left(x_t - \beta_t \varepsilon_t\right)$, we can parameterized $\mu\left(x_t\right)$ as $\mu\left(x_t\right) = \frac{1}{\alpha_t}\left(x_t - \beta_t \epsilon_\theta\left(x_t, t\right)\right.$
3. According to our proposition, $q\left(x_{t-1} \mid x_0\right)$ is $x_{t-1} = \bar{\alpha}_{t-1} x_0 + \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1}$

Objective Function

$$= - \int q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1} \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_{t-1} dx_t$$

1. Ignore those constants irrelevant to optimization objective, $-\log p_\theta\left(x_{t-1} \mid x_t\right)$ is $\frac{1}{2\sigma_t^2}\left\|x_{t-1} - \mu\left(x_t\right)\right\|^2$
2. With $x_{t-1} = \frac{1}{\alpha_t}\left(x_t - \beta_t \varepsilon_t\right)$, we can parameterized $\mu\left(x_t\right)$ as $\mu\left(x_t\right) = \frac{1}{\alpha_t}\left(x_t - \beta_t \epsilon_\theta\left(x_t, t\right)\right.$
3. According to our proposition, $q\left(x_{t-1} \mid x_0\right)$ is $x_{t-1} = \bar{\alpha}_{t-1} x_0 + \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1}$
4. $q\left(x_t \mid x_{t-1}\right)$ is $x_t = \alpha_t x_{t-1} + \beta_t \varepsilon_t$, where $\bar{\varepsilon}_{t-1}, \varepsilon_t \sim \mathcal{N}(0, I)$

Objective Function

$$= -\int q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1} \mid x_0\right) \tilde{q}\left(x_0\right) \log p_\theta\left(x_{t-1} \mid x_t\right) dx_0 dx_{t-1} dx_t$$

1. Ignore those constants irrelevant to optimization objective, $-\log p_\theta\left(x_{t-1} \mid x_t\right)$ is $\frac{1}{2\sigma_t^2}\left\|x_{t-1} - \mu\left(x_t\right)\right\|^2$
2. With $x_{t-1} = \frac{1}{\alpha_t}\left(x_t - \beta_t \varepsilon_t\right)$, we can parameterized $\mu\left(x_t\right)$ as $\mu\left(x_t\right) = \frac{1}{\alpha_t}\left(x_t - \beta_t \epsilon_\theta\left(x_t, t\right)\right.$
3. According to our proposition, $q\left(x_{t-1} \mid x_0\right)$ is $x_{t-1} = \bar{\alpha}_{t-1} x_0 + \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1}$
4. $q\left(x_t \mid x_{t-1}\right)$ is $x_t = \alpha_t x_{t-1} + \beta_t \varepsilon_t$, where $\bar{\varepsilon}_{t-1}, \varepsilon_t \sim \mathcal{N}(0, I)$

$$= \frac{\beta_t^2}{\alpha_t^2 \sigma_t^2} \mathbb{E}_{\bar{\varepsilon}_{t-1}, \varepsilon_t \sim \mathcal{N}(0,I), x_0 \sim \tilde{q}(x_0)}\left[\left\|\varepsilon_t - \epsilon_\theta\left(\bar{\alpha}_t x_0 + \alpha_t \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1} + \beta_t \varepsilon_t, t\right)\right\|^2\right] \tag{26}$$

Objective Function

$$\frac{\beta_t^2}{\alpha_t^2 \sigma_t^2} \mathbb{E}_{\bar{\varepsilon}_{t-1}, \varepsilon_t \sim \mathcal{N}(0,I), x_0 \sim \tilde{q}(x_0)} \left[\left\| \varepsilon_t - \epsilon_\theta \left(\bar{\alpha}_t x_0 + \alpha_t \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1} + \beta_t \varepsilon_t, t \right) \right\|^2 \right]$$

Objective Function

$$\frac{\beta_t^2}{\alpha_t^2 \sigma_t^2} \mathbb{E}_{\bar{\varepsilon}_{t-1}, \varepsilon_t \sim \mathcal{N}(0,I), x_0 \sim \tilde{q}(x_0)} \left[\left\| \varepsilon_t - \epsilon_\theta \left(\bar{\alpha}_t x_0 + \alpha_t \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1} + \beta_t \varepsilon_t, t \right) \right\|^2 \right]$$

For this equation, we have $\bar{\varepsilon}_{t-1}, \varepsilon_t$ to randomly generate, we use a trick [Ho, Neurips 2020] to replace them with one random variable ε.

Proposition

$\alpha_t \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1} + \beta_t \varepsilon_t$ is equivalent to $\bar{\beta}_t \varepsilon \mid \varepsilon \sim \mathcal{N}(0, I)$

$\beta_t \bar{\varepsilon}_{t-1} - \alpha_t \bar{\beta}_{t-1} \varepsilon_t$ is equivalent to $\bar{\beta}_t \omega \mid \omega \sim \mathcal{N}(0, I)$

It can be verified that $\mathbb{E}\left[\varepsilon \omega^\top\right] = 0$

Objective Function

$$\frac{\beta_t^2}{\alpha_t^2\sigma_t^2}\mathbb{E}_{\bar{\varepsilon}_{t-1},\varepsilon_t\sim\mathcal{N}(0,I),x_0\sim\tilde{q}(x_0)}\left[\left\|\varepsilon_t-\epsilon_\theta\left(\bar{\alpha}_t x_0+\alpha_t\bar{\beta}_{t-1}\bar{\varepsilon}_{t-1}+\beta_t\varepsilon_t,t\right)\right\|^2\right]$$

For this equation, we have $\bar{\varepsilon}_{t-1},\varepsilon_t$ to randomly generate, we use a trick [Ho, Neurips 2020] to replace them with one random variable ε.

Proposition

$\alpha_t\bar{\beta}_{t-1}\bar{\varepsilon}_{t-1}+\beta_t\varepsilon_t$ *is equivalent to* $\bar{\beta}_t\varepsilon \mid \varepsilon\sim\mathcal{N}(0,I)$

$\beta_t\bar{\varepsilon}_{t-1}-\alpha_t\bar{\beta}_{t-1}\varepsilon_t$ *is equivalent to* $\bar{\beta}_t\omega \mid \omega\sim\mathcal{N}(0,I)$

It can be verified that $\mathbb{E}\left[\varepsilon\omega^\top\right]=0$

We use ε and ω to represent ε_t

$$\varepsilon_t=\frac{\left(\beta_t\varepsilon-\alpha_t\bar{\beta}_{t-1}\omega\right)\bar{\beta}_t}{\beta_t^2+\alpha_t^2\bar{\beta}_{t-1}^2}=\frac{\beta_t\varepsilon-\alpha_t\bar{\beta}_{t-1}\omega}{\bar{\beta}_t} \tag{27}$$

Objective Function

$$\varepsilon_t = \frac{\left(\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega\right) \bar{\beta}_t}{\beta_t^2 + \alpha_t^2 \bar{\beta}_{t-1}^2} = \frac{\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega}{\bar{\beta}_t}$$

Objective Function

$$\varepsilon_t = \frac{\left(\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega\right) \bar{\beta}_t}{\beta_t^2 + \alpha_t^2 \bar{\beta}_{t-1}^2} = \frac{\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega}{\bar{\beta}_t}$$

Plug this equation into our objective.

$$\begin{aligned} & \frac{\beta_t^2}{\alpha_t^2 \sigma_t^2} \mathbb{E}_{\bar{\varepsilon}_{t-1}, \varepsilon_t \sim \mathcal{N}(0, I), x_0 \sim \tilde{q}(x_0)} \left[\left\| \varepsilon_t - \epsilon_\theta \left(\bar{\alpha}_t x_0 + \alpha_t \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1} + \beta_t \varepsilon_t, t \right) \right\|^2 \right] \\ = & \frac{\beta_t^2}{\alpha_t^2 \sigma_t^2} \mathbb{E}_{\omega, \varepsilon \sim \mathcal{N}(0, I), x_0 \sim \tilde{q}(x_0)} \left[\left\| \frac{\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega}{\bar{\beta}_t} - \epsilon_\theta \left(\bar{\alpha}_t x_0 + \bar{\beta}_t \varepsilon, t \right) \right\|^2 \right] \end{aligned} \tag{28}$$

Objective Function

$$\varepsilon_t = \frac{\left(\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega\right) \bar{\beta}_t}{\beta_t^2 + \alpha_t^2 \bar{\beta}_{t-1}^2} = \frac{\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega}{\bar{\beta}_t}$$

Plug this equation into our objective.

$$\begin{aligned} &\frac{\beta_t^2}{\alpha_t^2 \sigma_t^2} \mathbb{E}_{\bar{\varepsilon}_{t-1}, \varepsilon_t \sim \mathcal{N}(0,I), x_0 \sim \tilde{q}(x_0)} \left[\left\| \varepsilon_t - \epsilon_\theta \left(\bar{\alpha}_t x_0 + \alpha_t \bar{\beta}_{t-1} \bar{\varepsilon}_{t-1} + \beta_t \varepsilon_t, t \right) \right\|^2 \right] \\ =&\frac{\beta_t^2}{\alpha_t^2 \sigma_t^2} \mathbb{E}_{\omega, \varepsilon \sim \mathcal{N}(0,I), x_0 \sim \tilde{q}(x_0)} \left[\left\| \frac{\beta_t \varepsilon - \alpha_t \bar{\beta}_{t-1} \omega}{\bar{\beta}_t} - \epsilon_\theta \left(\bar{\alpha}_t x_0 + \bar{\beta}_t \varepsilon, t \right) \right\|^2 \right] \end{aligned} \tag{28}$$

Notice that ω can be separated in this equation, we have

$$\frac{\beta_t^4}{\bar{\beta}_t^2 \alpha_t^2 \sigma_t^2} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,I), x_0 \sim \tilde{p}(x_0)} \left[\left\| \varepsilon - \frac{\bar{\beta}_t}{\beta_t} \epsilon_\theta \left(\bar{\alpha}_t x_0 + \bar{\beta}_t \varepsilon, t \right) \right\|^2 \right] \tag{29}$$

Score-based Methods for Diffusion Model

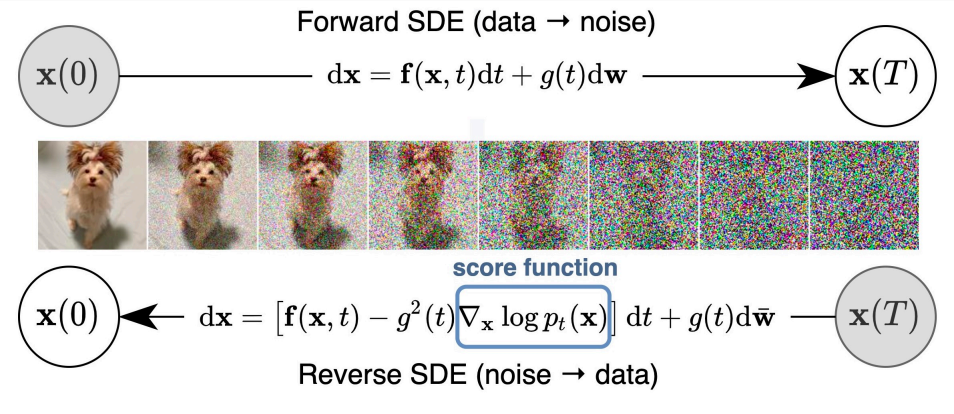

Figure 3: Solving a reverse SDE yields a score-based generative model.

note: $p_t(x) = p(x_t)$, where $x_t = func(x, t)$

Connection to DDPM

DDPM:

$$x_i = \sqrt{1-\beta_i}x_{i-1} + \sqrt{\beta_i}z_{i-1}, \quad i = 1, \cdots, N \tag{30}$$

Connection to DDPM

DDPM:

$$x_i = \sqrt{1-\beta_i}x_{i-1} + \sqrt{\beta_i}z_{i-1}, \quad i = 1, \cdots, N \tag{30}$$

Corresponding Stochastic Differential Equation:

$$\mathrm{d}x = -\frac{1}{2}\beta(t)x\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}w \tag{31}$$

Connection to DDPM

DDPM:

$$x_i = \sqrt{1-\beta_i}x_{i-1} + \sqrt{\beta_i}z_{i-1}, \quad i = 1, \cdots, N \tag{30}$$

Corresponding Stochastic Differential Equation:

$$\mathrm{d}x = -\frac{1}{2}\beta(t)x\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}w \tag{31}$$

DDPM is one of the discretization version of SDE, with specific coefficients.

Solve the Reverse SDE

Reverse SDE:

$$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\overline{w} \tag{32}$$

Solve the Reverse SDE

Reverse SDE:

$$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\overline{w} \tag{32}$$

Unknown distributions:

- the terminal distribution $p_T(x) \approx \pi(x)$
- the score function $\nabla_x \log p_t(x)$

In order to estimate $\nabla_x \log p_t(x)$, we train a Time-Dependent Score-Based Model $s_\theta(x,t)$, such that $s_\theta(x,t) \approx \nabla_x \log p_t(x)$

$$\mathbb{E}_{t\in\mathcal{U}(0,T)}\mathbb{E}_{p_t(x)}\left[\lambda(t)\left\|\nabla_x \log p_t(x) - s_\theta(x,t)\right\|_2^2\right] \tag{33}$$

Solve the Reverse SDE

Reverse SDE:

$$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\overline{w} \tag{32}$$

Unknown distributions:

- the terminal distribution $p_T(x) \approx \pi(x)$
- the score function $\nabla_x \log p_t(x)$

In order to estimate $\nabla_x \log p_t(x)$, we train a Time-Dependent Score-Based Model $s_\theta(x,t)$, such that $s_\theta(x,t) \approx \nabla_x \log p_t(x)$

$$\mathbb{E}_{t\in\mathcal{U}(0,T)}\mathbb{E}_{p_t(x)}\left[\lambda(t)\left\|\nabla_x \log p_t(x) - s_\theta(x,t)\right\|_2^2\right] \tag{33}$$

There exists a family of methods called score matching [Hyvarinen JMLR 2005;Song, UAI 2020] that minimize the divergence without knowledge of the ground-truth score function.

Connection to Likelihood Estimation

Proposition

Given by [Song, ICLR 2021]:

When $\lambda(t) = g(t)^2$, we have

$$\mathbb{KL}\left(p_0(x) \| p_\theta(x)\right) \leq \frac{T}{2} \mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}_{p_t(x)} \left[\lambda(t) \left\| \nabla_x \log p_t(x) - s_\theta(x,t) \right\|_2^2\right] + \mathbb{KL}\left(p_T \| \pi\right) \tag{34}$$

According to this, we can train score-based generative models to achieve very high likelihoods.

Take Home Message

- **Variational inference can be used to derive many likelihood-based generative models including VAE, DDPM, etc.**
- **ELBO is a lower bound for** $\log p_\theta(x)$

  $\Updownarrow$

  $\mathbb{KL}(p(x,z)||p_\theta(x,z))$ **is an upper bound for** $\mathbb{KL}(p(x)||p_\theta(x))$
- **DDPM is an autoregressive version of VAE and one of the discretization version of SDE, with specific coefficients.**

Thanks!