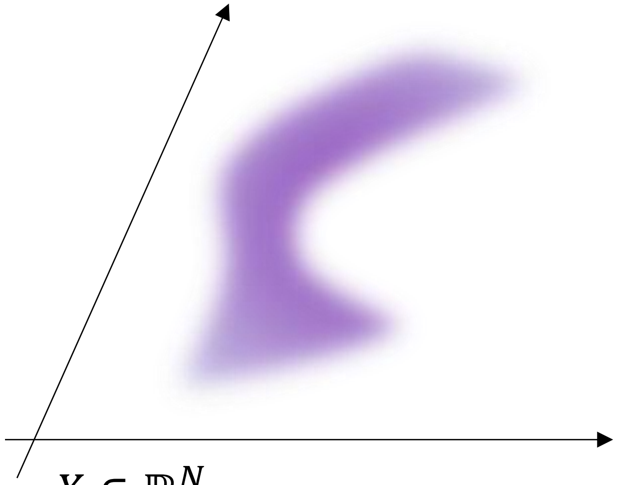
Lecture 5: Denoising Diffusion Probabilistic Models (DDPM)

CSC_52002_EP

Generative Objective: Learn the distribution

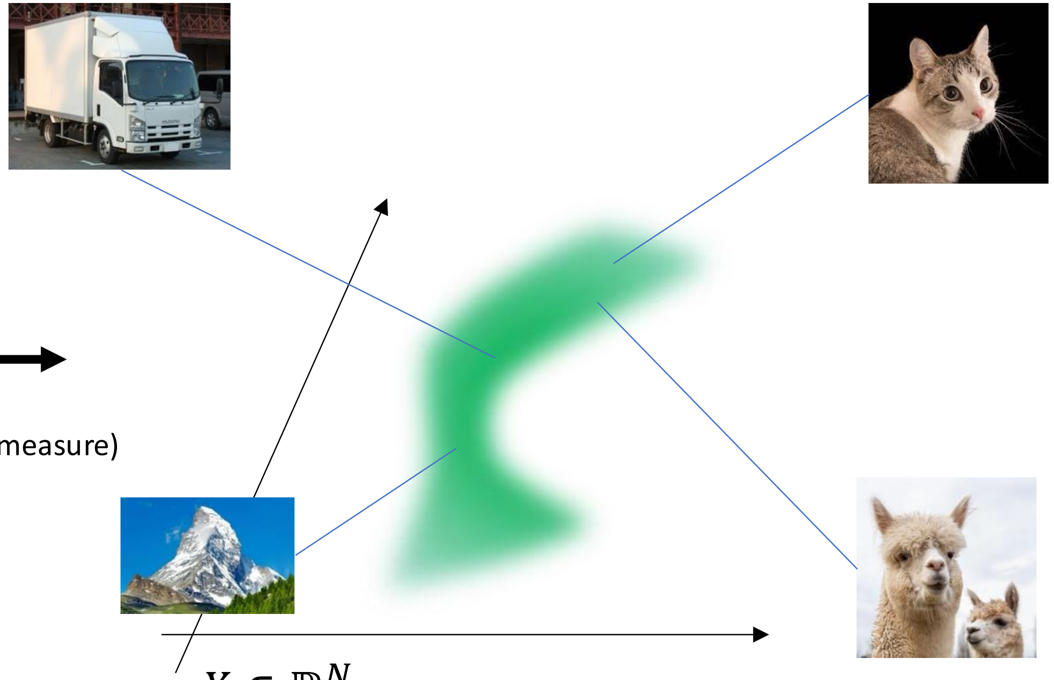


$X \in \mathbb{R}^N$

Distribution of Learnt Data $P_\theta(X)$
with parameter $\theta \in \mathbb{R}^M$

find best θ

s.t. $P_\theta(X) \approx P(X)$ (with a certain measure)

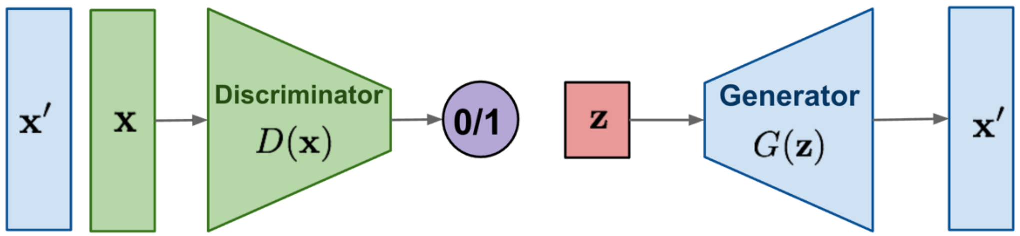


$X \in \mathbb{R}^N$

Distribution of Real Data $P(X)$

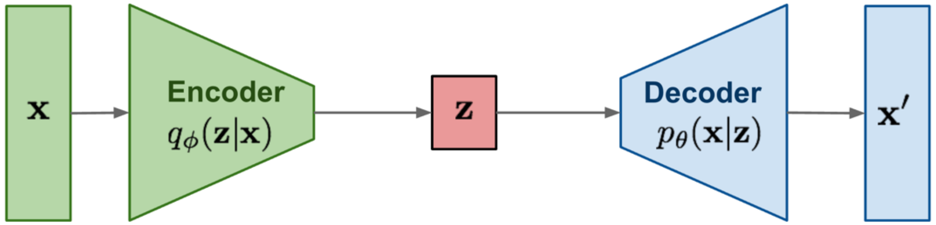
Generative Models

GAN: Adversarial training



$$\min_G \max_D L(D,G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$
$$= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x)]$$

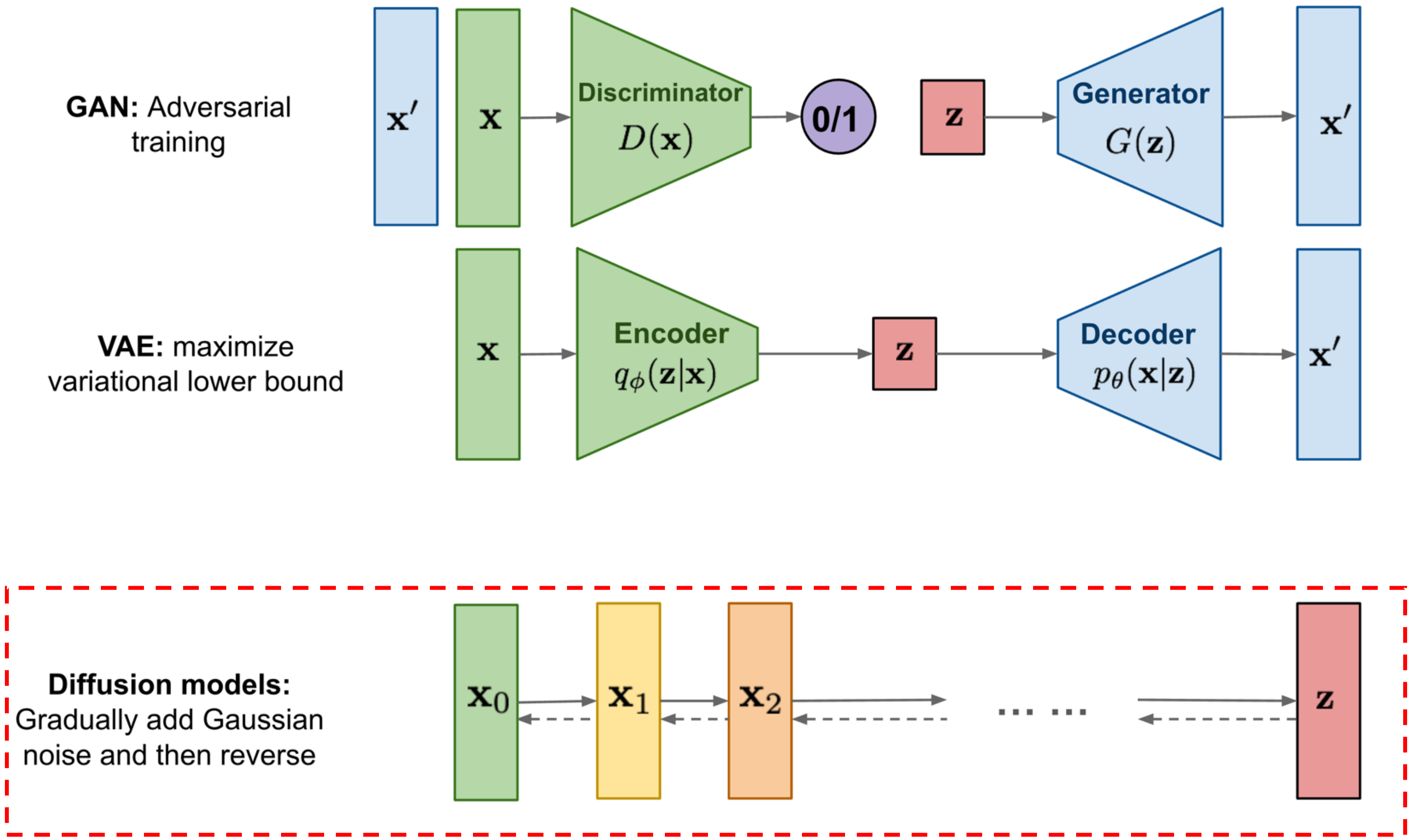
VAE: maximize variational lower bound



$$L_{\text{VAE}}(\theta,\phi) = -\log p_\theta(\mathbf{x}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x}))$$
$$= -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}|\mathbf{z}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$$
$$\theta^*,\phi^* = \arg\min_{\theta,\phi} L_{\text{VAE}}$$

Diffusion models:
Gradually add Gaussian noise and then reverse

Generative Models



$$\begin{aligned}\min_G \max_D L(D,G) &= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \\ &= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1-D(x)]\end{aligned}$$

unstable training

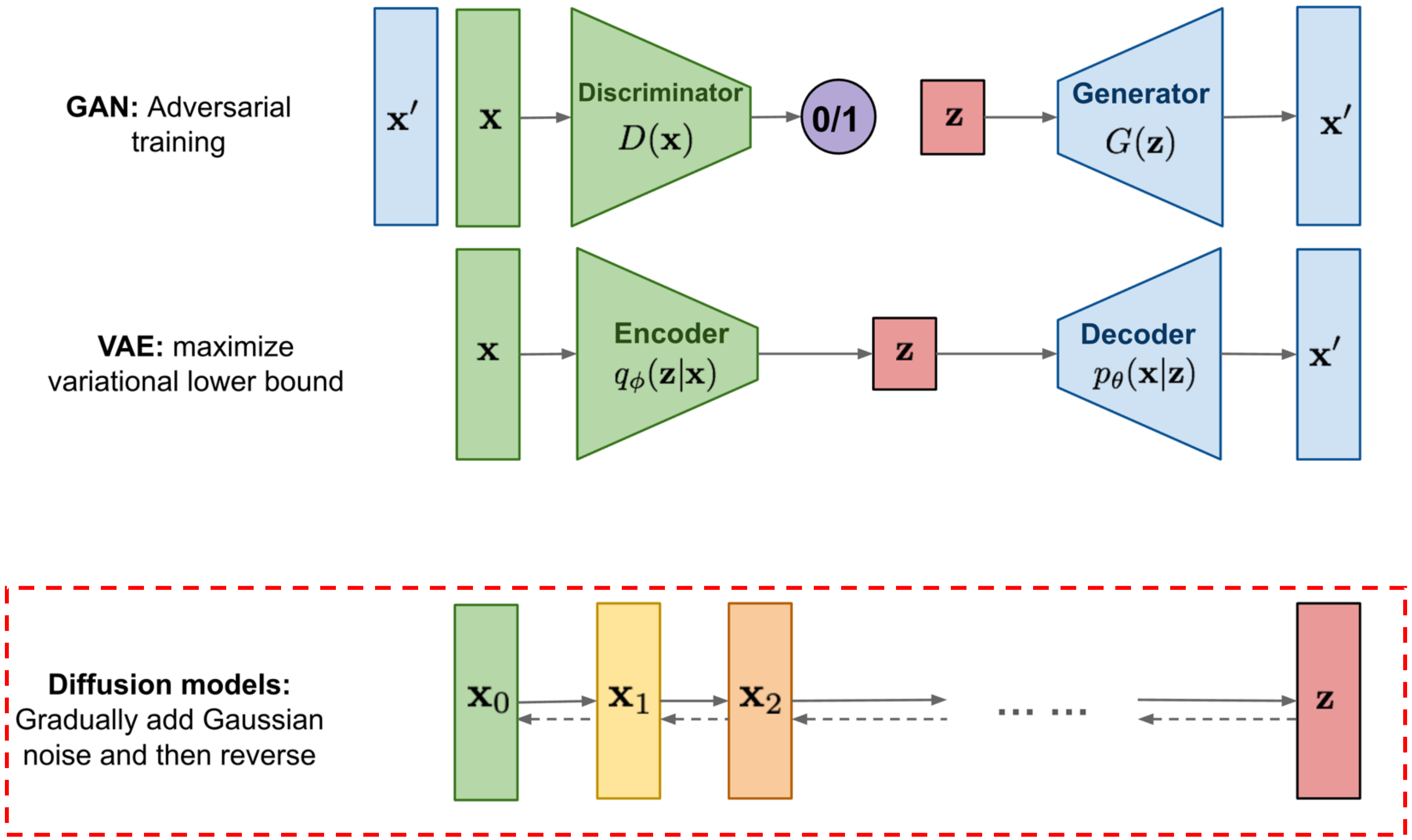
and mode collapse (learning data, instead of distribution)

$$\begin{aligned}L_{\mathrm{VAE}}(\theta,\phi) &= -\log p_\theta(\mathbf{x}) + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x})) \\ &= -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}|\mathbf{z}) + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) \\ \theta^*,\phi^* &= \arg\min_{\theta,\phi} L_{\mathrm{VAE}}\end{aligned}$$

under-represtation of the distribution,

posteriori collapse (Gaussian Priori is not realistic)

Generative Models



$$\min_G \max_D L(D,G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$$
$$= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1-D(x)]$$

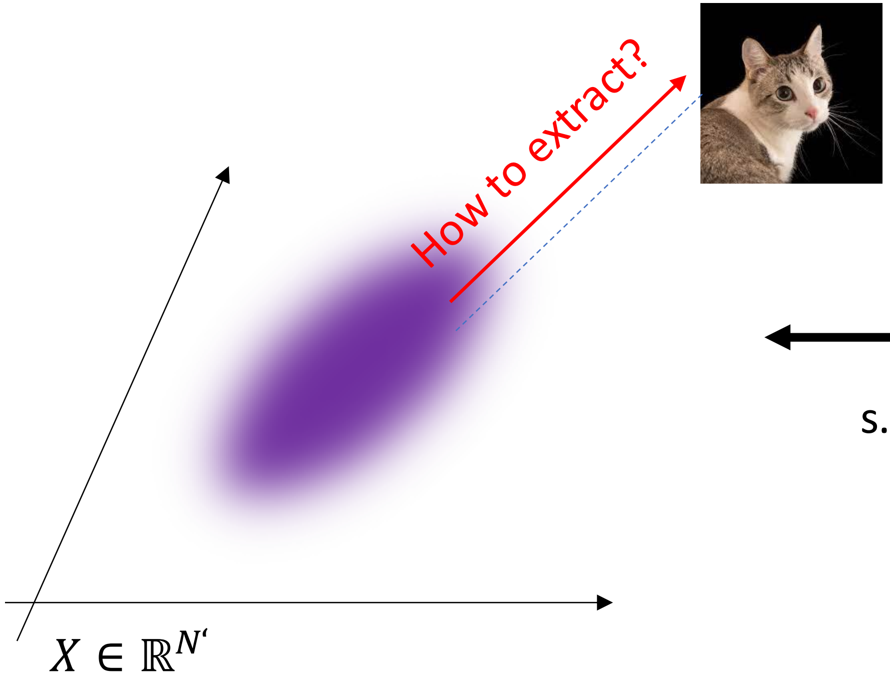
unstable training
and mode collapse (learning data, instead of distribution)

$$L_{\text{VAE}}(\theta,\phi) = -\log p_\theta(\mathbf{x}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x}))$$
$$= -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}|\mathbf{z}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$$
$$\theta^*,\phi^* = \arg\min_{\theta,\phi} L_{\text{VAE}}$$

under-represtation of the distribution,
posteriori collapse (Gaussian Priori is not realistic)

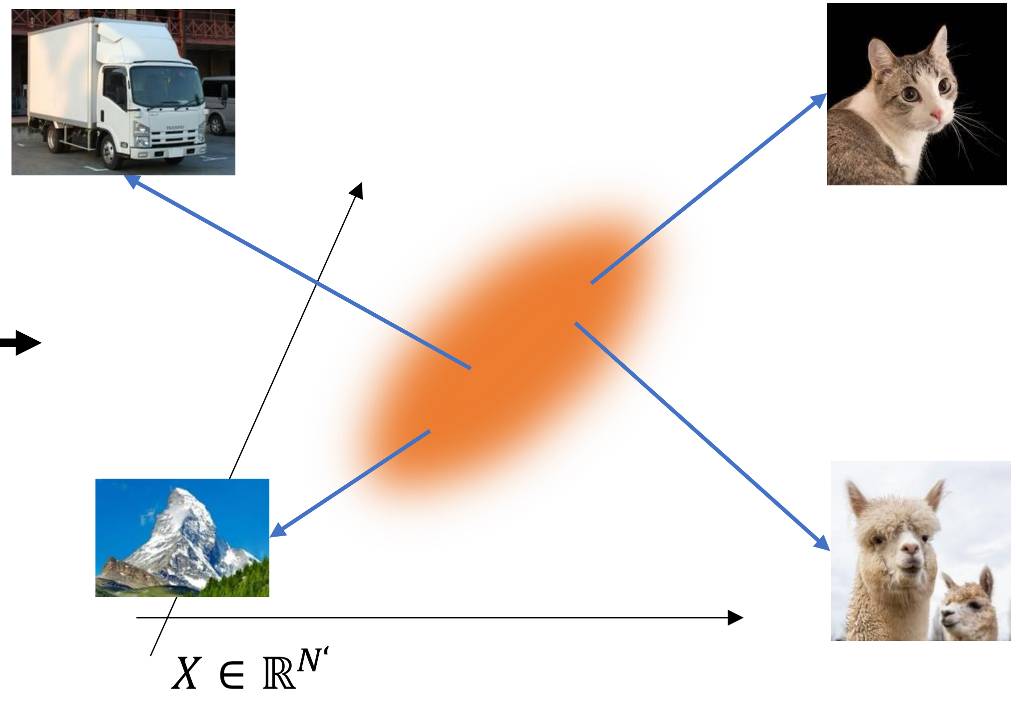
Better representation capacity,
and learn the whole distribution.

Generative Objective: Learn the distribution



find best θ

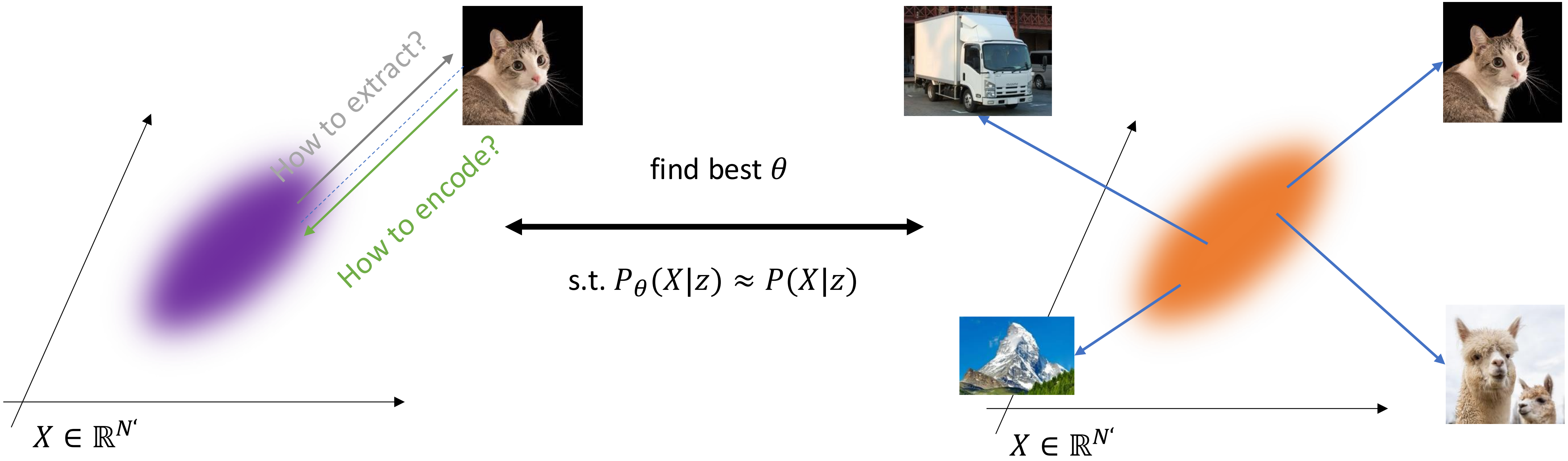
s.t. $P_\theta(X|z) \approx P(X|z)$



Distribution of $P(z)$ and we what to learn $P_\theta(X|z)$ with parameter $\theta \in \mathbb{R}^M$

Learning mapping of Real Data $P(X|z)$

Generative Objective: Learn the distribution

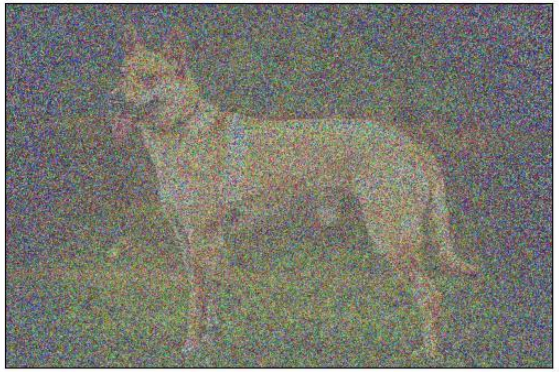


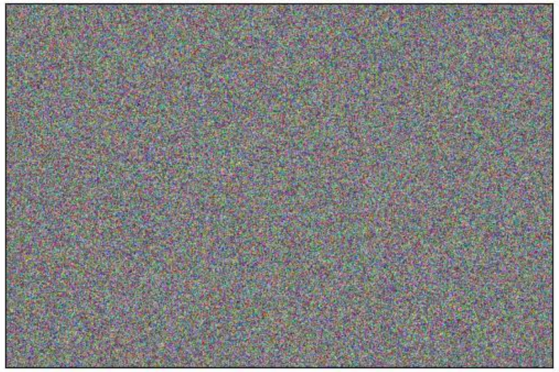
Distribution of $P(z)$ and we what to learn $P_\theta(X|z)$ with parameter $\theta \in \mathbb{R}^M$

Learning mapping of Real Data $P(X|z)$

Generative Objective: Forward Process

Generative Objective: Forward Process

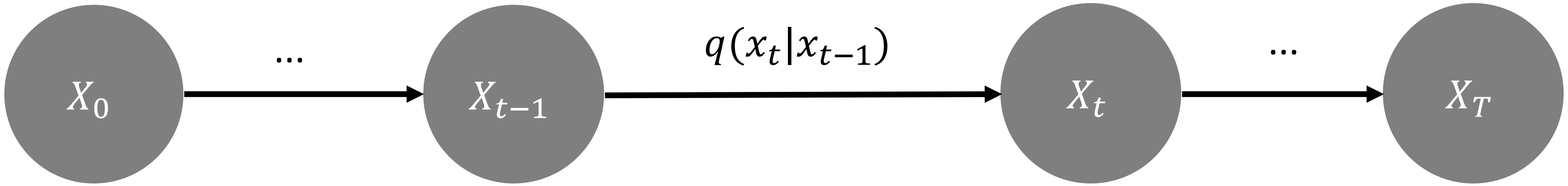




How to push an image to a Gaussian?

Easy, let's add noise !

Generative Objective: Forward Process



Linear Blending:

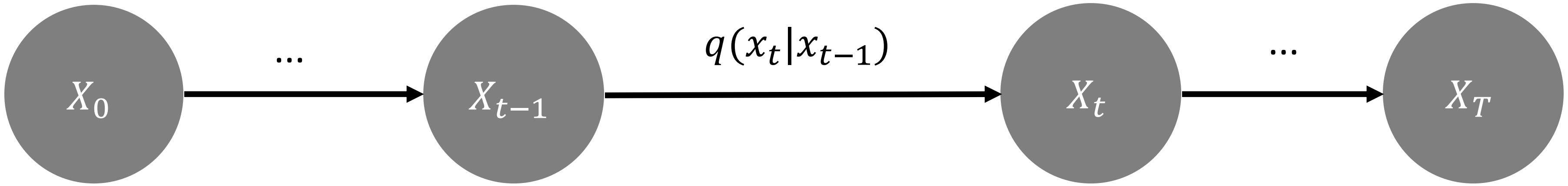
$$\boldsymbol{x}_t = a_t \boldsymbol{x}_{t-1} + b_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

But how to define a and b?

How to push an image to a Gaussian?

Easy, let's add noise !

Generative Objective: Forward Process



Linear Blending:

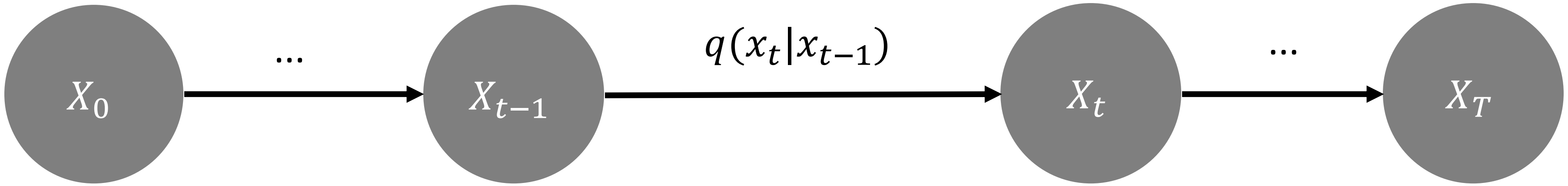
$$\boldsymbol{x}_t = a_t \boldsymbol{x}_{t-1} + b_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

let's assume something nice:

What if I do this from beginning to the end?

$$\begin{aligned}
\boldsymbol{x}_t &= a_t \boldsymbol{x}_{t-1} + b_t \varepsilon_t \\
&= a_t \left(a_{t-1} \boldsymbol{x}_{t-2} + b_{t-1} \boldsymbol{\varepsilon}_{t-1} \right) + b_t \varepsilon_t \\
&= a_t a_{t-1} \boldsymbol{x}_{t-2} + a_t b_{t-1} \boldsymbol{\varepsilon}_{t-1} + b_t \boldsymbol{\varepsilon}_t \\
&= \ldots \\
&= (a_t \ldots a_1) \boldsymbol{x}_0 + (a_t \ldots a_2) b_1 \varepsilon_1 + (a_t \ldots a_3) b_2 \varepsilon_2 + \cdots + a_t b_{t-1} \varepsilon_{t-1} + b_t \varepsilon_t
\end{aligned}$$

Generative Objective: Forward Process



Linear Blending:

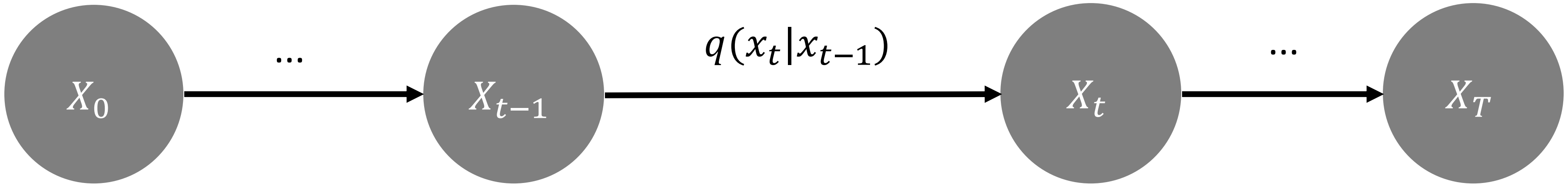
$$\boldsymbol{x}_t = a_t \boldsymbol{x}_{t-1} + b_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

let's assume something nice:

What if I do this from beginning to the end?

$$\begin{aligned}
\boldsymbol{x}_t &= a_t \boldsymbol{x}_{t-1} + b_t \varepsilon_t \\
&= a_t \left(a_{t-1} \boldsymbol{x}_{t-2} + b_{t-1} \boldsymbol{\varepsilon}_{t-1} \right) + b_t \varepsilon_t \\
&= a_t a_{t-1} \boldsymbol{x}_{t-2} + a_t b_{t-1} \boldsymbol{\varepsilon}_{t-1} + b_t \boldsymbol{\varepsilon}_t \\
&= \ldots \\
&= (a_t \ldots a_1) \boldsymbol{x}_0 + (a_t \ldots a_2) b_1 \varepsilon_1 + (a_t \ldots a_3) b_2 \varepsilon_2 + \cdots + a_t b_{t-1} \varepsilon_{t-1} + b_t \varepsilon_t
\end{aligned}$$

Generative Objective: Forward Process



<u>Independent</u>* *Gaussian addition -> Gaussian

$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + (a_t \dots a_2)b_1\underline{\varepsilon_1} + (a_t \dots a_3)b_2\underline{\varepsilon_2} + \cdots + a_t b_{t-1}\underline{\varepsilon_{t-1}} + b_t\underline{\varepsilon_t}$$

Remember this..

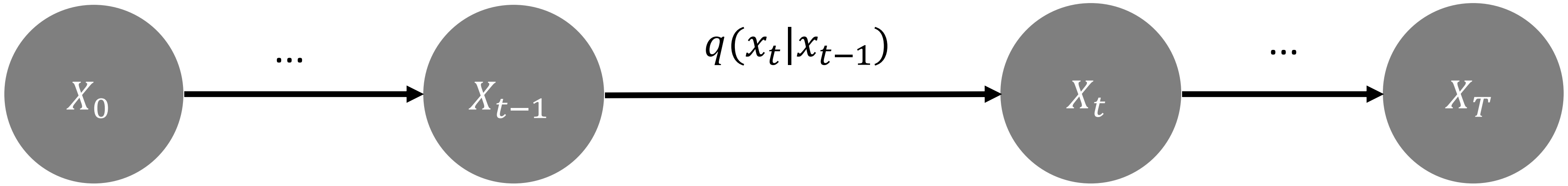
$$\boxed{Z \sim \mathcal{N}\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right)}$$

😮 💨 *pfff...*

$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \cdots + a_t^2 b_{t-1}^2 + b_t^2}\,\underline{\bar{\varepsilon}_t}$$

$$\bar{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Generative Objective: Forward Process



$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}\,\bar{\varepsilon}_t$$

$$\bar{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

we added $(a_t \dots a_1)^2 + \underline{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}$ $-\ (a_t \dots a_1)^2$

$$= (a_t \dots a_2)^2 a_1^2 + (a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2$$

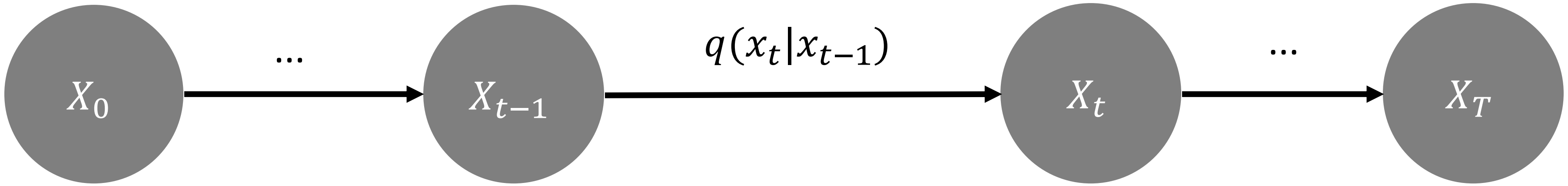
$$= (a_t \dots a_2)^2 \left(a_1^2 + b_1^2\right) + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2$$

$$= (a_t \dots a_3)^2 \left(a_2^2 \left(a_1^2 + b_1^2\right) + b_2^2\right) + \dots + a_t^2 b_{t-1}^2 + b_t^2$$

$$= a_t^2 \left(a_{t-1}^2 \left(\dots \left(a_2^2 \left(a_1^2 + b_1^2\right) + b_2^2\right) + \dots\right) + b_{t-1}^2\right) + b_t^2$$

What the hack? but wait !

Generative Objective: Forward Process



$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}\,\bar{\varepsilon}_t$$

$$\bar{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

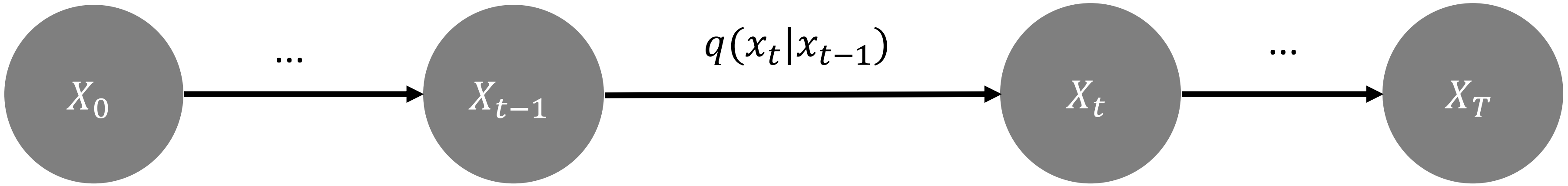
we added $(a_t \dots a_1)^2 + \underline{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}$ $-\ (a_t \dots a_1)^2$

$$\begin{aligned}
&= (a_t \dots a_2)^2 a_1^2 + (a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
&= (a_t \dots a_2)^2 \left(a_1^2 + b_1^2\right) + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
&= (a_t \dots a_3)^2 \left(a_2^2 \left(a_1^2 + b_1^2\right) + b_2^2\right) + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
&= a_t^2 \left(a_{t-1}^2 \left(\dots \left(a_2^2 \left(a_1^2 + b_1^2\right) + b_2^2\right) + \dots\right) + b_{t-1}^2\right) + b_t^2
\end{aligned}$$

what if: $\bar{a}_t = (a_t \dots a_1)^2$

$a_t^2 + b_t^2 = 1$

Generative Objective: Forward Process



$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}\,\bar{\varepsilon}_t$$

$$\bar{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

we added $(a_t \dots a_1)^2 + \underline{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}$ - $(a_t \dots a_1)^2$

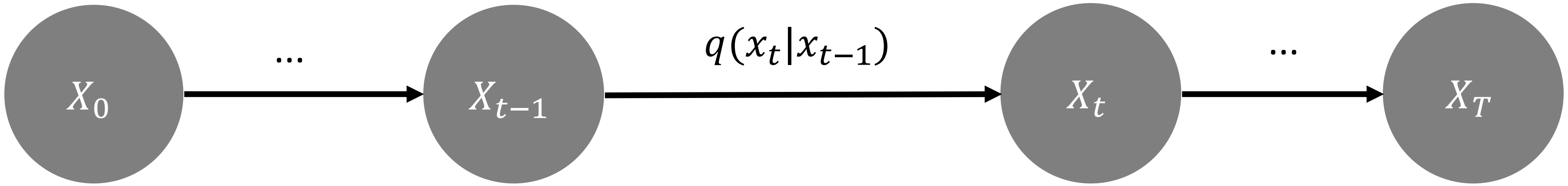
$$
\begin{aligned}
&= (a_t \dots a_2)^2 a_1^2 + (a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
&= (a_t \dots a_2)^2 \left(a_1^2 + b_1^2\right) + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
&= (a_t \dots a_3)^2 \left(a_2^2 \left(a_1^2 + b_1^2\right) + b_2^2\right) + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
&= a_t^2 \left(a_{t-1}^2 \left(\dots \left(a_2^2 \left(a_1^2 + b_1^2\right) + b_2^2\right) + \dots\right) + b_{t-1}^2\right) + b_t^2
\end{aligned}
$$

$= 1 - \bar{a}_t$

what if: $\bar{a}_t = (a_t \dots a_1)^2$

$a_t^2 + b_t^2 = 1$

Generative Objective: Forward Process



$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}\,\bar{\varepsilon}_t$$

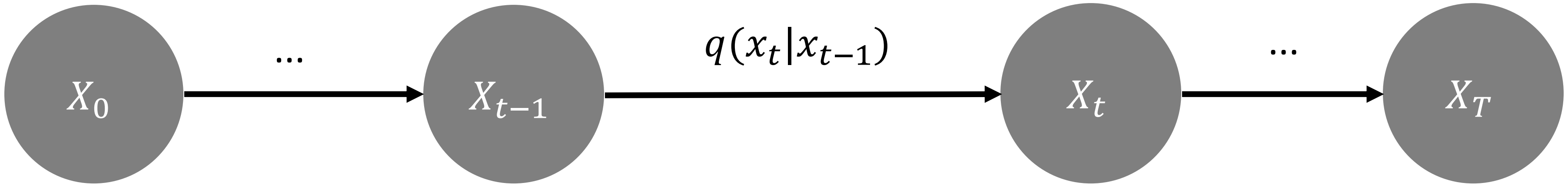
$$\bar{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

if: $a_t^2 + b_t^2 = 1$ $\quad \bar{a}_t = (a_t \dots a_1)^2$

$$\boldsymbol{x}_t = \sqrt{\bar{a}_t}\boldsymbol{x}_0 + \sqrt{1 - \bar{a}_t}\,\bar{\varepsilon}_t, \quad \bar{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$ from x_0

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1 - \alpha_t}\,\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$ from x_t-1

Generative Objective: Forward Process



We call this **<u>a Forward Process.</u>**

- Original image at X_0 and pure noise at X_T
- We repeat the noising T times
- $\beta_t \in (0,1)$ is a noise schedule, ie. linear

$$\alpha_t = 1 - \beta_t$$

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0,\, I)$$

This is also called *variance-preservation*.

Generative Objective: Forward Process

$$X_0 \xrightarrow{\ \dots\ } X_{t-1} \xrightarrow{q(x_t|x_{t-1})} X_t \xrightarrow{\ \dots\ } X_T$$

We call this **a Forward Process.**

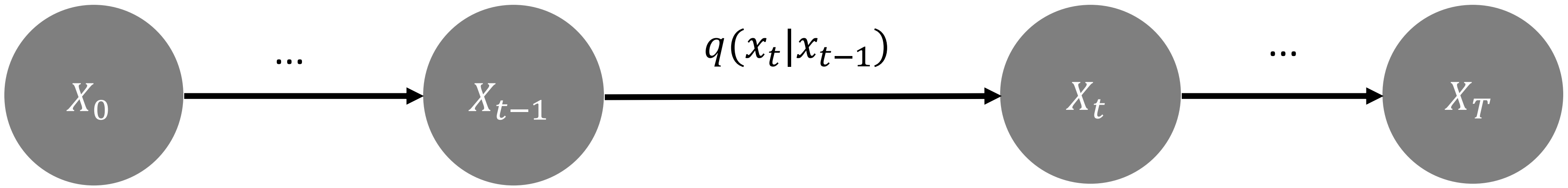
- Original image at X_0 and pure noise at X_T
- We repeat the noising T times
- $\beta_t \in (0,1)$ is a noise schedule, ie. linear

$$x_t = \sqrt{1-\beta_t} x_{t-1} + \sqrt{\beta_t}\epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0,\ I)$$

$$\Downarrow$$

$$q(x_t \mid x_{t-1}) \sim \mathcal{N}(\sqrt{1-\beta_t} x_{t-1},\ \beta_t I)$$

Generative Objective: Forward Process



We call this **<u>a Forward Process.</u>**

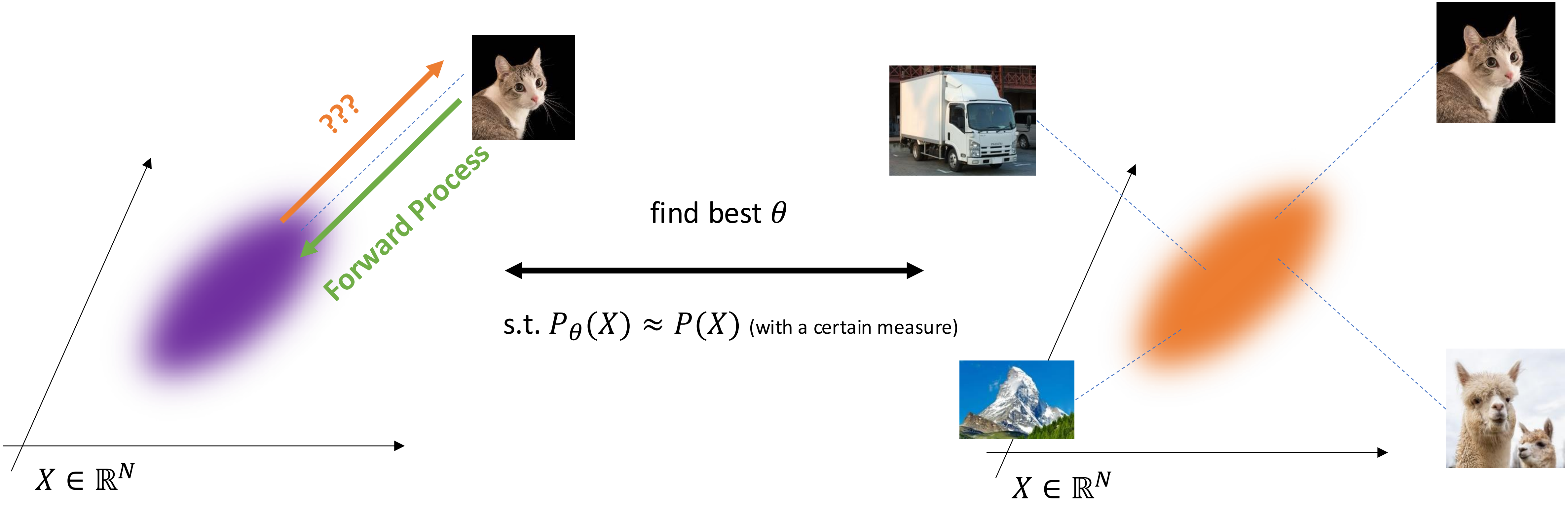
- Original image at X_0 and pure noise at X_T
- We repeat the noising T times
- $\beta_t \in (0,1)$ is a noise schedule, ie. linear

$$q(x_t \mid x_{t-1}) \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

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

This gaussian noising is a Markov chain

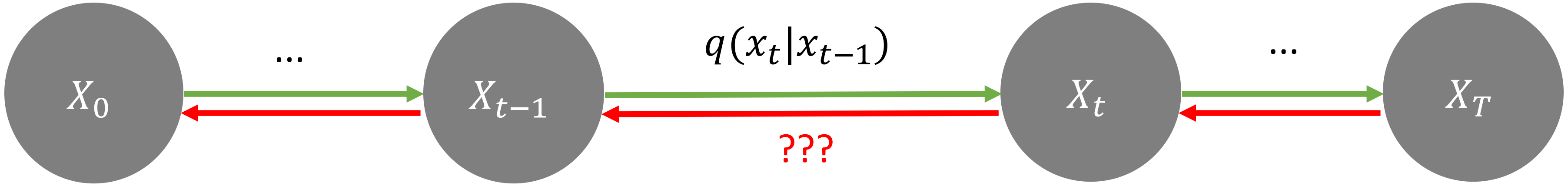
Generative Objective: Learn the distribution



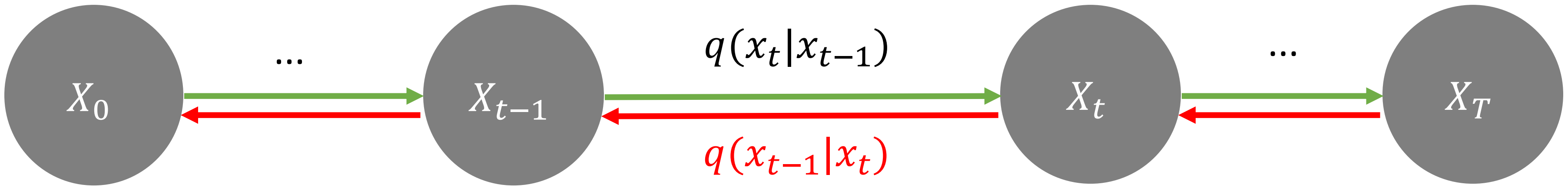
Distribution of Learnt Data $P_\theta(X)$
with parameter $\theta \in \mathbb{R}^M$

Distribution of Real Data $P(X)$

Generative Objective: Reverse Process



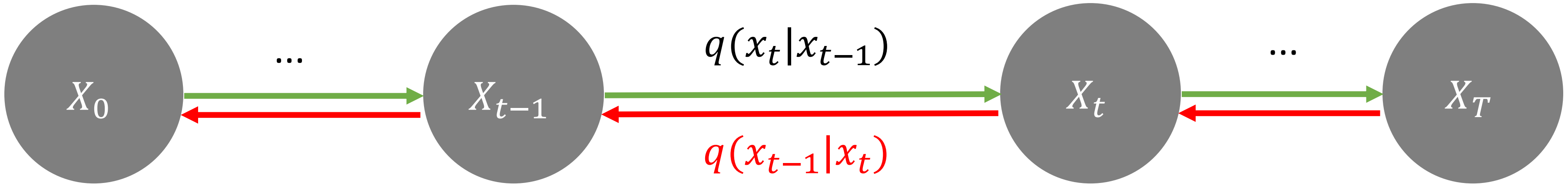
Generative Objective: Reverse Process



Is Bayesian method an efficient method?

$$q(x_{t-1} \mid x_t) = q(x_t \mid x_{t-1})\frac{q(x_{t-1})}{q(x_t)} \longleftarrow \boxed{q(x_t) = \int q(x_t \mid x_{t-1})q(x_{t-1})\mathrm{d}x}$$

Generative Objective: Reverse Process



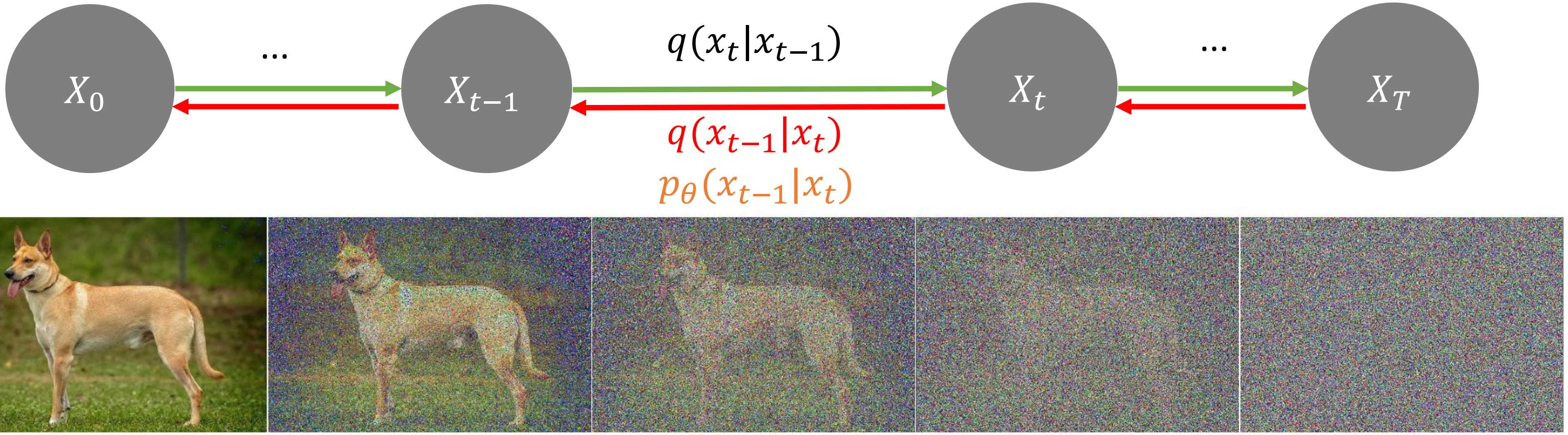
Is Bayesian method an efficient method?

$$q(x_{t-1} \mid x_t) = \underline{q(x_t \mid x_{t-1})}\frac{q(x_{t-1})}{q(x_t)} \longleftarrow \boxed{q(x_t) = \int q(x_t \mid x_{t-1}) q(x_{t-1}) \mathrm{d}x}$$

An exhaustive integration over all $q(x_{t-1})$ 😔

We saw it somewhere before ....

Generative Objective: Reverse Process



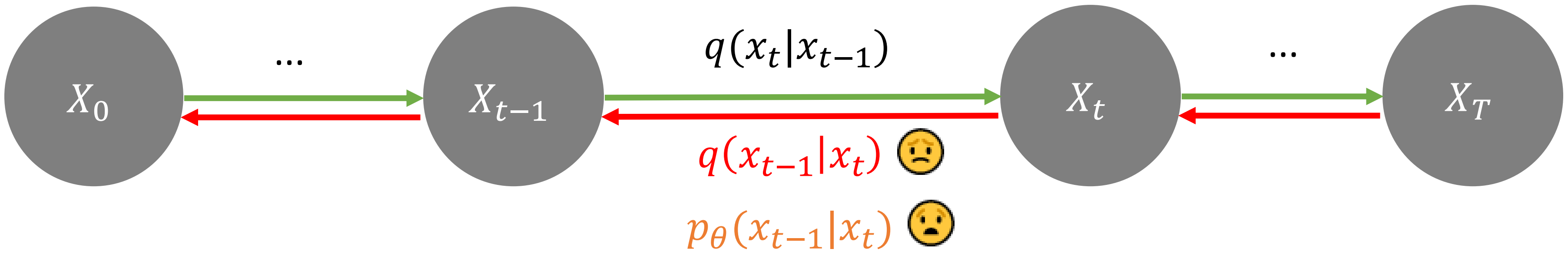
Is Bayesian method an efficient method?

$$q(x_{t-1} \mid x_t) = \underline{\underline{q(x_t \mid x_{t-1})}} \frac{q(x_{t-1})}{q(x_t)} \longleftarrow \boxed{q(x_t) = \int q(x_t \mid x_{t-1}) q(x_{t-1}) \mathrm{d}x}$$

An exhaustive integration over all $q(x_{t-1})$ 😔

We saw it somewhere before in VAE, we learn it!

Generative Objective: Reverse Process

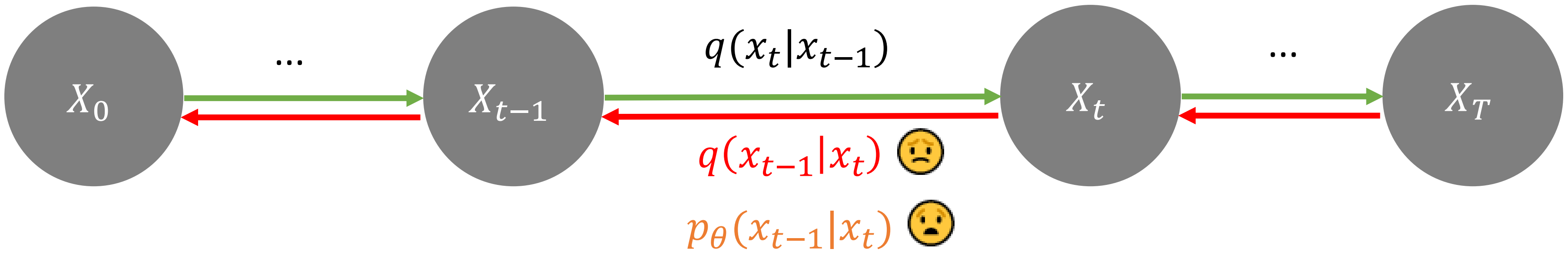


It was easy in VAE case

Why?

$p_\theta(x|z)$ 😀

Generative Objective: Reverse Process

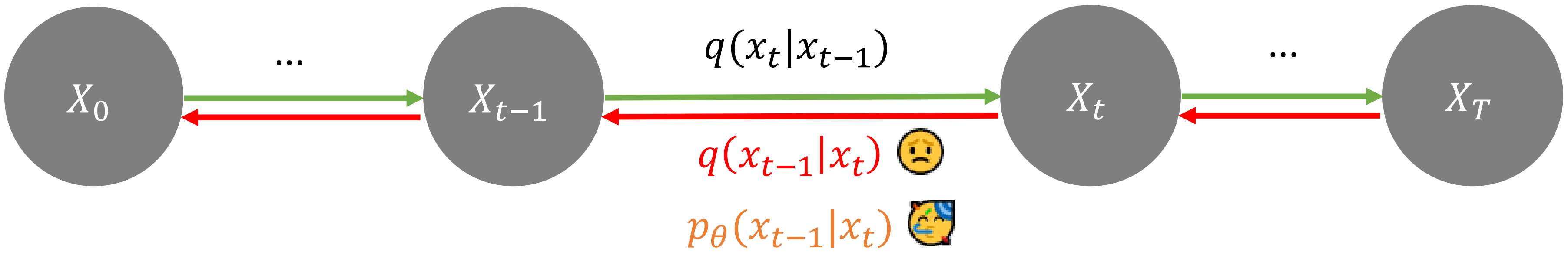


It was easy in VAE case
Why?

$p_\theta(x|z)$

Because you know the image!

Generative Objective: Reverse Process



A very nice property of Gaussian:
if $q(x_t|x_{t-1})$ is a Gaussian with small β (another reason we need many steps!)
Luckily, the $q(x_{t-1}|x_t)$ shall still be a Gaussian,
we therefore learn this Gaussian's mean and variance
by a network approximated $p_\theta(x_{t-1}|x_t)$

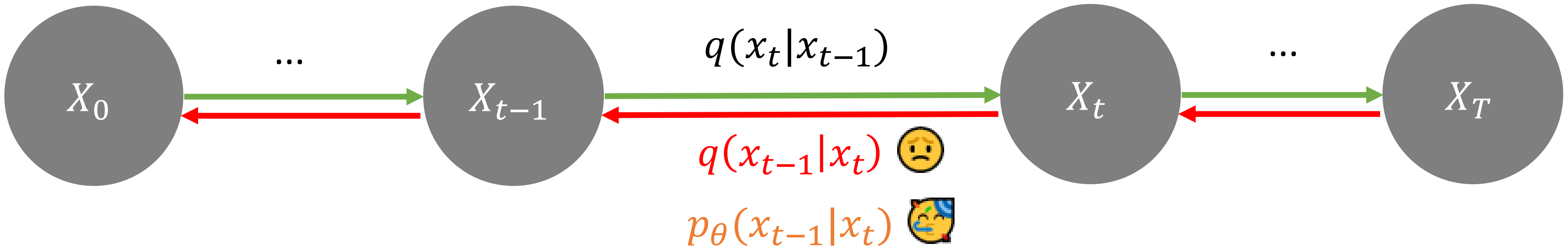
$$q(x_t \mid x_{t-1}) \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1},\ \beta_t I)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underline{\underline{\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)}}, \underline{\underline{\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)}})$$

Learnable parameters

Next question: how to train?

Generative Objective: Reverse Process



$$-L_{\mathrm{VAE}} = \log p_\theta(\mathbf{x}) - \underline{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}))} \leq \log p_\theta(\mathbf{x})$$

In VAE, the training loss is the ELBO.

$$\begin{aligned} -\log p_\theta(\mathbf{x}_0) &\leq -\log p_\theta(\mathbf{x}_0) + D_{\mathrm{KL}}\left(q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \| p_\theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0)\right) \\ &= -\log p_\theta(\mathbf{x}_0) + \sum_{\mathbf{x}_{1:T}} q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \log \underline{\frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}}; \text{ by definition of KL Div} \end{aligned}$$

We can “imagine” diffusion like T step VAE
with z of previous status.
and train a BIG ELBO of all status.

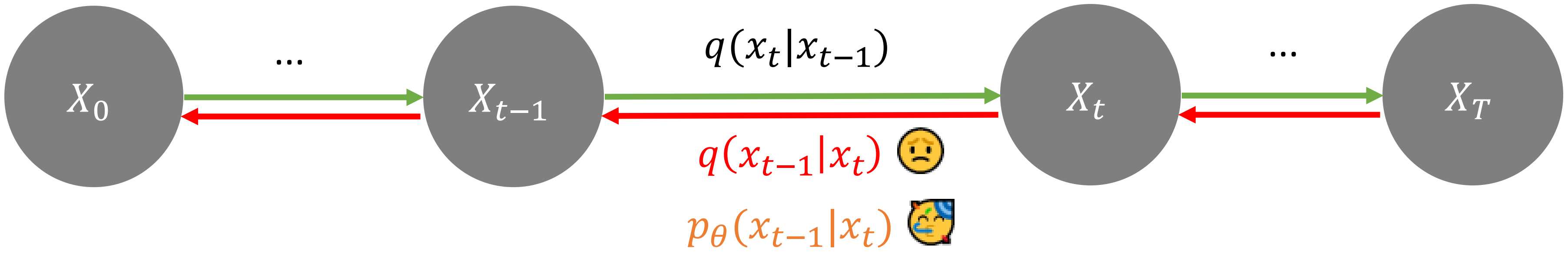
$$\boxed{p(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\, p(\mathbf{x}_{T-1} \mid \mathbf{x}_T)\, p(\mathbf{x}_{T-2} \mid \mathbf{x}_{T-1}, \cancel{\mathbf{x}_T}) \ldots p(\mathbf{x}_0 \mid \mathbf{x}_1, \cancel{\ldots, \mathbf{x}_T}).}$$

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

Markovian

[Sohl-Dickstein et al., 2015]

Generative Objective: Reverse Process



$$-L_{\mathrm{VAE}} = \log p_\theta(\mathbf{x}) - \underline{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x}))} \leq \log p_\theta(\mathbf{x})$$

In VAE, the training loss is the ELBO.

We can "imagine" diffusion like T step VAE with z of previous status.
and train a BIG ELBO of all status.

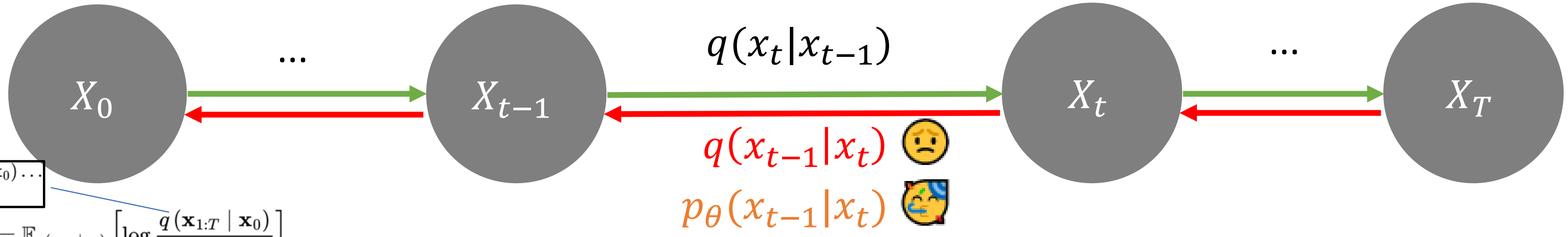
$$\begin{aligned}
-\log p_\theta(\mathbf{x}_0) &\leq -\log p_\theta(\mathbf{x}_0) + D_{\mathrm{KL}}(q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \| p_\theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0)) \\
&= -\log p_\theta(\mathbf{x}_0) + \sum_{\mathbf{x}_{1:T}} q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}; \text{ by definition of KL Div} \\
&= -\log p_\theta(\mathbf{x}_0) + \mathbb{E}_{\mathbf{x}_{1:T} \sim q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T}) / p_\theta(\mathbf{x}_0)}\right]; \text{ by definition of expectation} \\
&= -\log p_\theta(\mathbf{x}_0) + \mathbb{E}_q\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})} + \log p_\theta(\mathbf{x}_0)\right]; \ p_\theta(x_0) \text{ in independent to } q \\
&= -\log p_\theta(\mathbf{x}_0) + \mathbb{E}_q\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] + \log p_\theta(\mathbf{x}_0) \\
&= \mathbb{E}_q\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right]
\end{aligned} \tag{6}$$

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

VLB (variational lower bound)

Sohl-Dickstein et al., 2015

Generative Objective: Reverse Process



$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = q(\mathbf{x}_1 \mid \mathbf{x}_0) q(\mathbf{x}_2 \mid \mathbf{x}_1, \mathbf{x}_0) \ldots q(\mathbf{x}_T \mid \mathbf{x}_{T-1}, \ldots, \mathbf{x}_0)$$

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

$$
\begin{aligned}
L_{\mathrm{VLB}} &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \\
&= \mathbb{E}_q\left[\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)}\right] \\
&= \mathbb{E}_q\left[-\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)}\right] \\
&= \mathbb{E}_q\left[-\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)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \boxed{\left(\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)}\right)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \boxed{\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)}} + \boxed{\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)}\right] \\
&= \mathbb{E}_q\left[\underline{-\log p_\theta(\mathbf{x}_T)} + \boxed{\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)}} + \boxed{\underline{\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{\cancel{q(\mathbf{x}_1|\mathbf{x}_0)}}}} + \log \frac{\cancel{q(\mathbf{x}_1|\mathbf{x}_0)}}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_q\left[\underline{\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)\right] \\
&= \mathbb{E}_q[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T} + \sum_{t=2}^T \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{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}]
\end{aligned}
$$

Bayesian, then x_0 due to conditional independence of Markov process.

$$\mathbb{E}_q\left[\log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}\right]$$

To next step: *only the t relevant*

$$= \mathbb{E}_{q(x_T|x_0)}\left[\log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}\right]$$

To next step: *Definition of expectation*

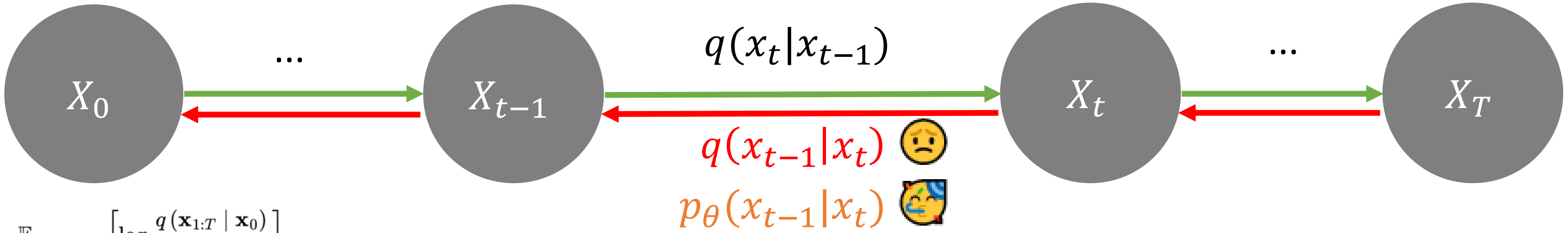
$$= \sum_{\mathbf{x}_T}\left[q(\mathbf{x}_T|\mathbf{x}_0) \log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}\right]$$

To next step: *Definition of KL Div*

$$= D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \| p_\theta(\mathbf{x}_T))$$

[Sohl-Dickstein et al., 2015]

Generative Objective: Reverse Process



$$
\begin{aligned}
L_{\mathrm{VLB}} &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \\
&= \mathbb{E}_q\left[\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)}\right] \quad \textit{expand each step} \\
&= \mathbb{E}_q\left[-\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)}\right] \\
&= \mathbb{E}_q\left[-\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)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \boxed{\left(\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)}\right)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \boxed{\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)}} + \boxed{\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)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \boxed{\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)}} + \boxed{\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{\cancel{q(\mathbf{x}_1|\mathbf{x}_0)}}} + \log \frac{\cancel{q(\mathbf{x}_1|\mathbf{x}_0)}}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_q\left[\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)\right] \\
&= \mathbb{E}_q[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T} + \sum_{t=2}^T \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{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}]
\end{aligned}
$$

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

x_0 due to conditional independence of Markov process.

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

To next step: *only the t relevant*

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

To next step:

using a skill that $\mathbb{E}_{X,Y} f(\cdot) = \sum_x \sum_y p(x,y) f(x,y) = \sum_x p(x) \sum_y p(y|x) f(x,y) = \mathbb{E}_X \mathbb{E}_{Y|X} f(x,y)$

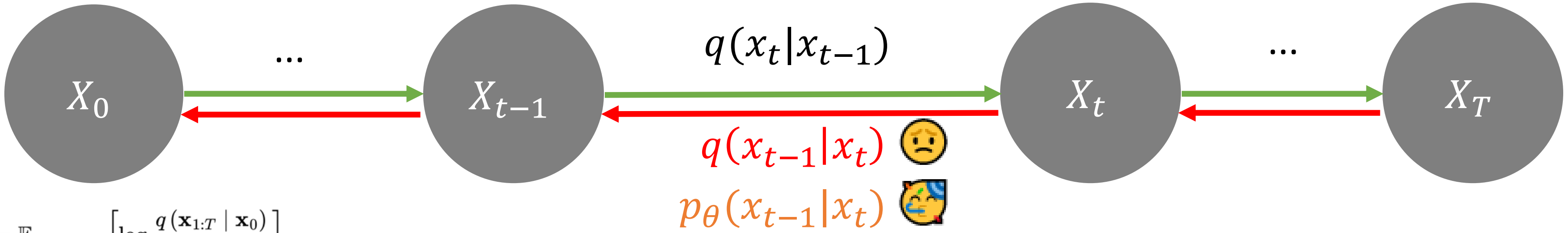
$$= \sum_{t=2}^T \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \mathbb{E}_{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\left[\log \frac{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}\right]$$

To next step: recall that $\mathbb{E}_{q(x)}\left[\log \frac{q(x)}{p(x)}\right] = \sum_x q(x) \log \frac{q(x)}{p(x)} = D_{KL}(q(x), p(x))$

$$= \sum_{t=2}^T \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \| p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right)\right] \quad (3)$$

Sohl-Dickstein et al., 2015

Generative Objective: Reverse Process



$$
\begin{aligned}
L_{\mathrm{VLB}} &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \\
&= \mathbb{E}_q\left[\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)}\right] \quad \textit{expand each step} \\
&= \mathbb{E}_q\left[-\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)}\right] \\
&= \mathbb{E}_q\left[-\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)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \boxed{\left(\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)}\right)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \boxed{\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)}} + \boxed{\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)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_T) + \boxed{\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)}} + \boxed{\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{\cancel{q(\mathbf{x}_1|\mathbf{x}_0)}}} + \log \frac{\cancel{q(\mathbf{x}_1|\mathbf{x}_0)}}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_q\left[\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)\right] \\
&= \mathbb{E}_q[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T} + \sum_{t=2}^T \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{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}]
\end{aligned}
$$

x_0 due to conditional independence of Markov process.

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

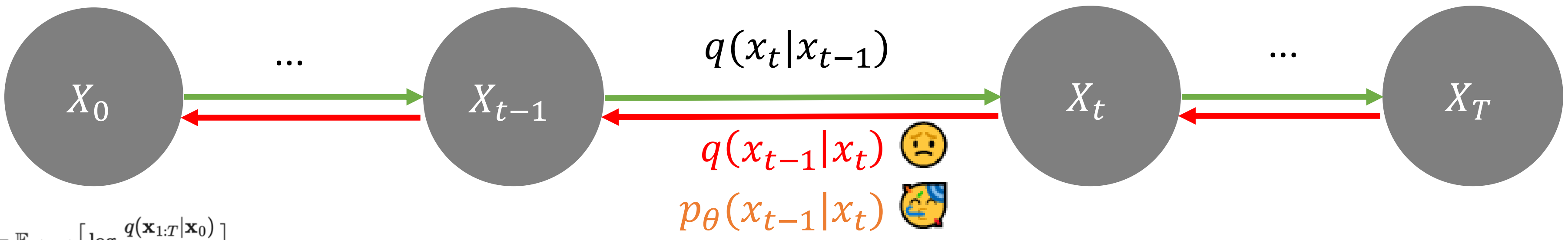
$$\mathbb{E}_q\left[\log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_T)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)\right]$$

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

To next step: $\mathbb{E}_{x,y} f(x) = \mathbb{E}_x f(x)$ recall also that $\mathbb{E}_q$ is $\mathbb{E}_{q(x_{1:T}|x_0)}$

$$= \mathbb{E}_q\left[D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \| p_\theta(\mathbf{x}_T)) + \sum_{t=2}^T D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \| p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)) + \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)\right] \quad (5)$$

Generative Objective: Reverse Process



$$
\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] \quad \textit{expand each step} \\
&= \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 \boxed{\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) + \boxed{\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)}} + \boxed{\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[\underline{-\log p_\theta(\mathbf{x}_T)} + \boxed{\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)}} + \boxed{\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{\cancel{q(\mathbf{x}_1|\mathbf{x}_0)}}} + \log \frac{\cancel{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}
$$

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

x_0 due to conditional independence of Markov process.

$$
\begin{aligned}
L_{\text{VLB}} &= L_T + L_{T-1} + \cdots + L_0 \\
\text{where } L_T &= D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T)) \\
L_t &= D_{\text{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t+1},\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \text{ for } 1 \leq t \leq T-1 \\
L_0 &= -\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)
\end{aligned}
$$

L_T not trainable constant,
L_0 separate discrete decoder

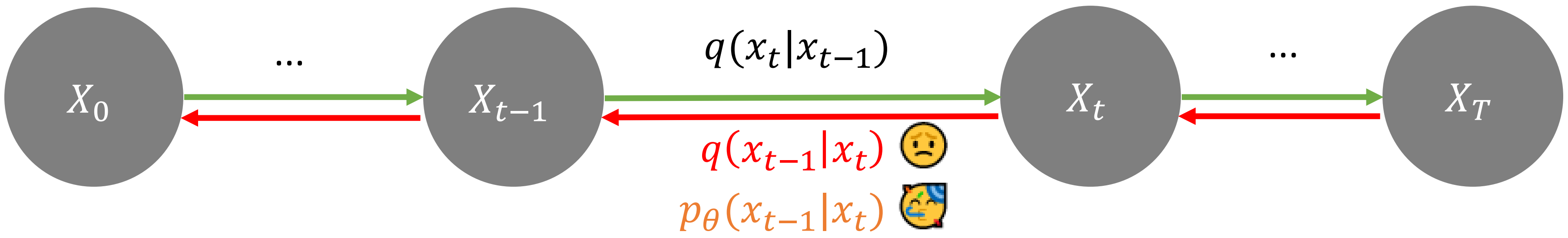
For L_t the minimizing objective can be achieved by minimizing the Gaussian distribution between $q(x_{t-1}|x_t, x_0)$ and $p_\theta(x_{t-1}|x_t)$

But .... what is $q(x_{t-1}|x_t, x_0)$? conditioning on x_0 is similar to the VAE invert sampling.

Need to understand the difference between $q(x_{t-1}|x_t)$ and $q(x_{t-1}|x_t, x_0)$

[Sohl-Dickstein et al., 2015](#)

Generative Objective: Reverse Process



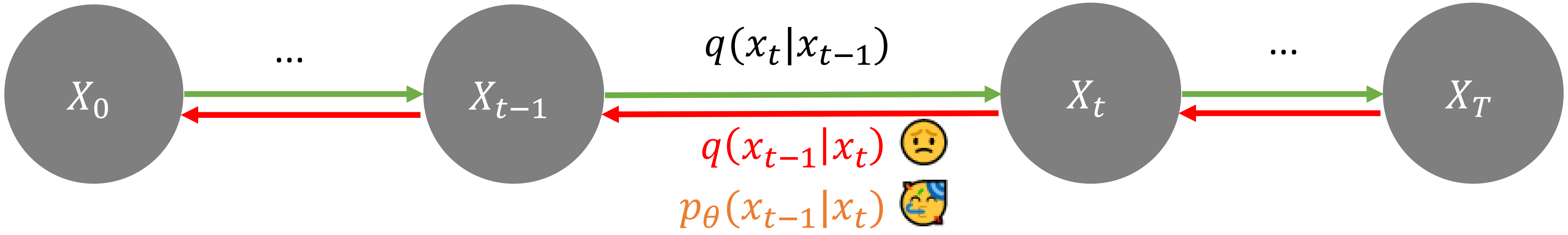
Before looking at $q(x_{t-1}|x_t, x_0)$.

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} && \text{;where } \boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_{t-2}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} && \text{;where } \bar{\boldsymbol{\epsilon}}_{t-2} \text{ merges two Gaussians (*).} \\
&= \dots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \\
q(\mathbf{x}_t|\mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})
\end{aligned}$$

$\boxed{\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i}$ $\boxed{\alpha_t = 1 - \beta_t}$

[Sohl-Dickstein et al., 2015]

Generative Objective: Reverse Process



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

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

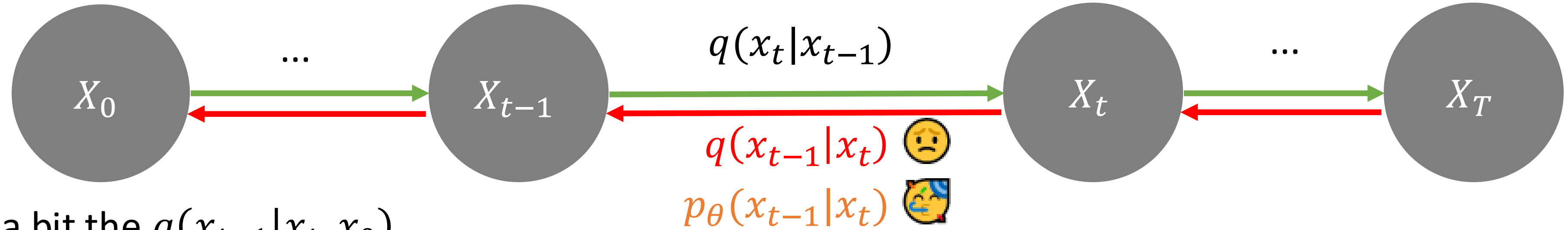
vs.

$$q(x_{t-1} \mid x_t) = q(x_t \mid x_{t-1})\frac{q(x_{t-1})}{q(x_t)}$$

$$\boxed{q(x_t) = \int q(x_t \mid x_{t-1})q(x_{t-1})\mathrm{d}x}$$

[Sohl-Dickstein et al., 2015]

Generative Objective: Reverse Process



Let's transform a bit the $q(x_{t-1}|x_t, x_0)$.

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

The nice property of reverted Gaussian is still a Gaussian

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

Yes, we know how to deal with this!
Just noise the image from beginning! 😀

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

recall also:

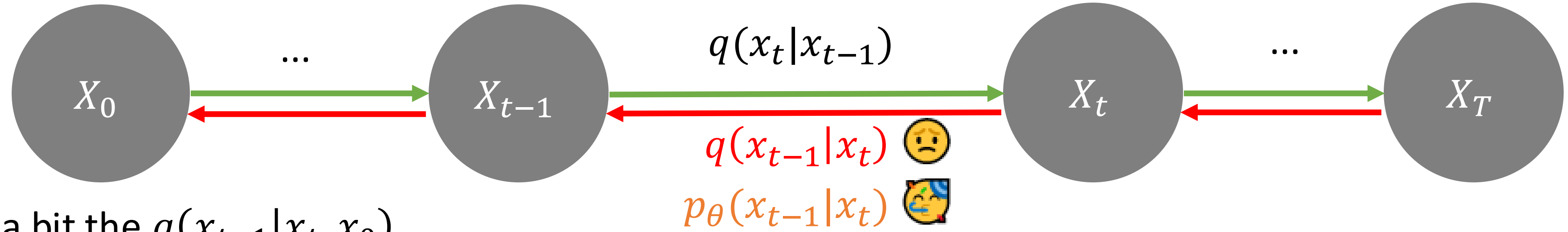
$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$

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

$$= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \mathcal{N}\left(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I}\right)}{\mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)}$$

Sohl-Dickstein et al., 2015

Generative Objective: Reverse Process



Let's transform a bit the $q(x_{t-1}|x_t, x_0)$.

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

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

The nice property of reverted Gaussian is still a Gaussian

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

Yes, we know how to deal with this!
Just noise the image from beginning! 😀

$$\boxed{\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i}$$

recall also:

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

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

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

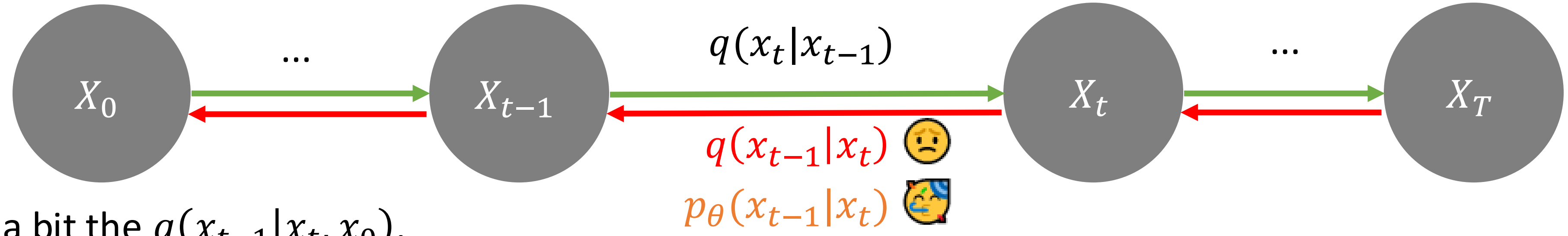
$$= \exp\Big(-\frac{1}{2}\big((\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\mathbf{x}_{t-1}^2 - (\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0)\mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0)\big)\Big)$$

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

Gaussian

[Sohl-Dickstein et al., 2015](#)

Generative Objective: Reverse Process



Let’s transform a bit the $q(x_{t-1}|x_t, x_0)$.

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

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

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

$$\boxed{\bar{\alpha}_t = \prod_{i=1}^{T}\alpha_i}$$

$$\boxed{\alpha_t = 1-\beta_t}$$

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

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

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

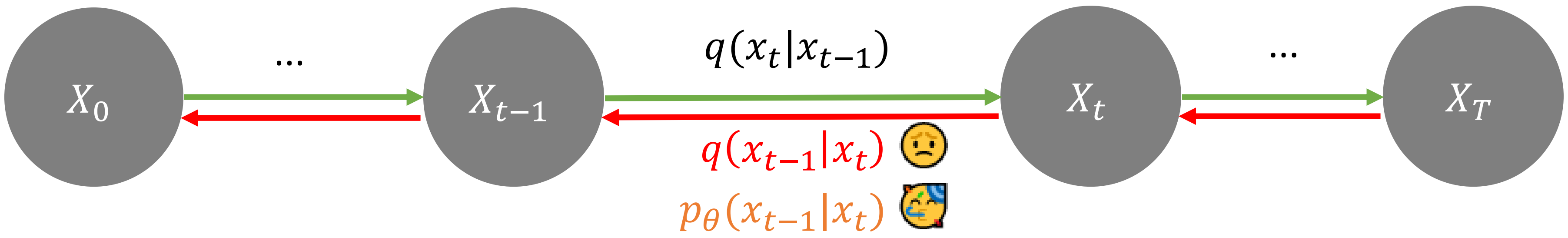
remove x_0 by $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t)$

Estimated image from t

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_t &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t) \\
&= \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\Big)
\end{aligned}$$

[Sohl-Dickstein et al., 2015](#)

Generative Objective: Reverse Process



Let's transform a bit the $q(x_{t-1}|x_t, x_0)$

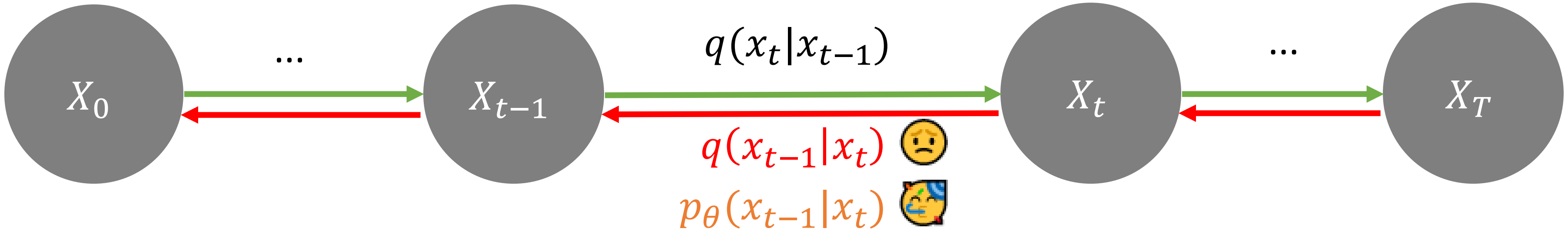
$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \cancel{\mathbf{x}_0}), \tilde{\beta}_t \mathbf{I})$$

$$\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right) \qquad \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$$

How to understand this?
$q(x_{t-1}|x_t, x_0)$ is tractable (when x_0 is known), and written as a Gaussian format dependent on noise scheduler (α and β) and x_t .

Sohl-Dickstein et al., 2015

Generative Objective: Reverse Process



Let's organize our hints, we want to minimize:

$$
\begin{aligned}
L_{\text{VLB}} &= L_T + L_{T-1} + \cdots + L_0 \\
\text{where } L_T &= D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T)) \\
&\boxed{L_t = D_{\text{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \text{ for } 1 \leq t \leq T-1}
\end{aligned}
$$

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

$$\boxed{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underline{\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)}, \underline{\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)})}$$

Learnable parameters

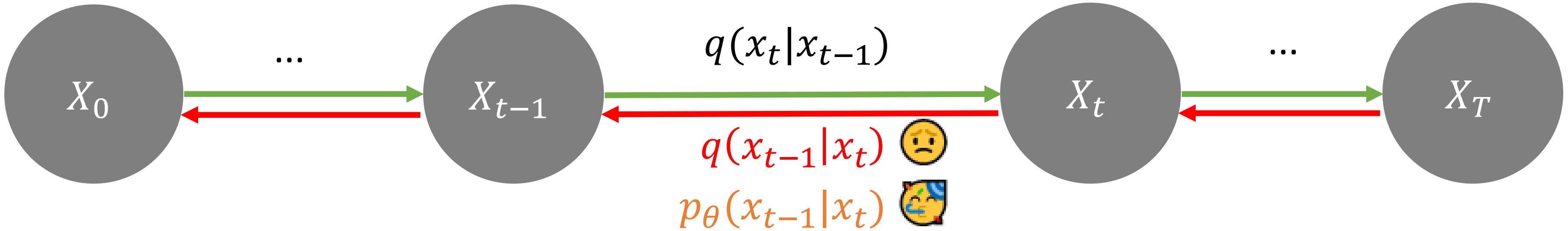
$$\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right)$$

$$\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$$

Nice! We need to compute the KL Divergence of two Gaussians then!

[Sohl-Dickstein et al., 2015]

Generative Objective: Reverse Process



Let's organize our hints, we want to minimize:

$$
\begin{aligned}
L_{\text{VLB}} &= L_T + L_{T-1} + \cdots + L_0 \\
\text{where } L_T &= D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T)) \\
&\boxed{L_t = D_{\text{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \text{ for } 1 \le t \le T-1}
\end{aligned}
$$

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

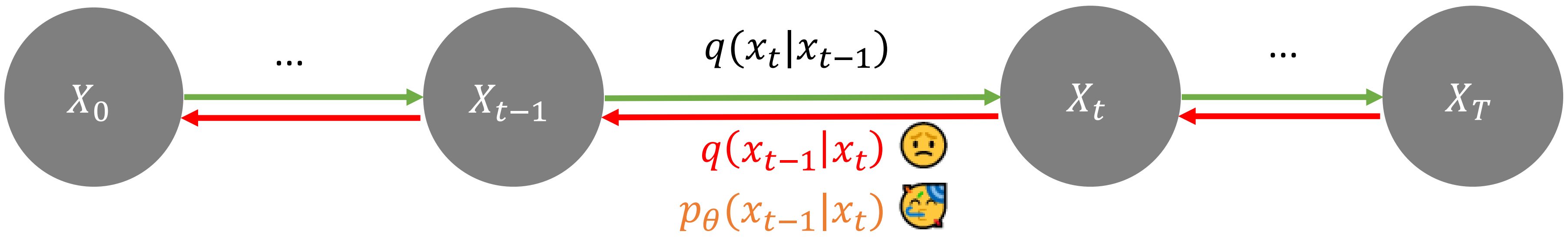
$$\boxed{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))}$$

Learnable parameters

$$
\begin{aligned}
L_t &= \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \Big[\frac{1}{2\|\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\|_2^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 \Big] \\
&= \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \Big[\frac{1}{2\|\boldsymbol{\Sigma}_\theta\|_2^2} \Big\| \frac{1}{\sqrt{\alpha_t}} \Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_t \Big) - \frac{1}{\sqrt{\alpha_t}} \Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \Big) \Big\|^2 \Big] \\
&= \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \Big[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\boldsymbol{\Sigma}_\theta\|_2^2} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 \Big] \\
&= \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \Big[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\boldsymbol{\Sigma}_\theta\|_2^2} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t)\|^2 \Big]
\end{aligned}
$$

Sohl-Dickstein et al., 2015

Generative Objective: Reverse Process



Let's organize our hints, we want to minimize:

$$
\begin{aligned}
L_{\text{VLB}} &= L_T + L_{T-1} + \cdots + L_0 \\
\text{where } L_T &= D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T)) \\
&\boxed{L_t = D_{\text{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \text{ for } 1 \le t \le T-1}
\end{aligned}
$$

$$
\begin{aligned}
L_t &= \mathbb{E}_{\mathbf{x}_0,\epsilon}\Big[\frac{1}{2\|\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t)\|_2^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\|^2\Big] \\
&= \mathbb{E}_{\mathbf{x}_0,\epsilon}\Big[\frac{1}{2\|\boldsymbol{\Sigma}_\theta\|_2^2}\Big\|\frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\Big) - \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\Big)\Big\|^2\Big] \\
&= \mathbb{E}_{\mathbf{x}_0,\epsilon}\Big[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\boldsymbol{\Sigma}_\theta\|_2^2}\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\|^2\Big] \\
&= \mathbb{E}_{\mathbf{x}_0,\epsilon}\Big[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\boldsymbol{\Sigma}_\theta\|_2^2}\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\underline{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t}, t)\|^2\Big]
\end{aligned}
$$

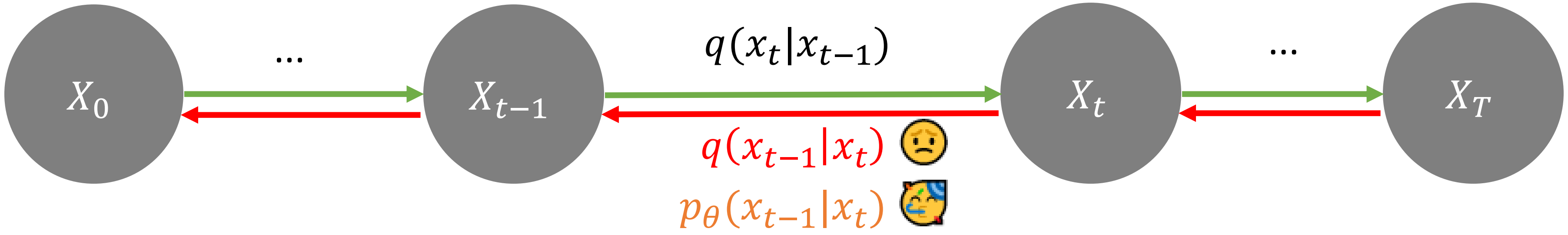
Ho et al. (2020) Find we can safely ignore the weighting

$$
\begin{aligned}
L_t^{\text{simple}} &= \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\Big[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\|^2\Big] \\
&= \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\Big[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t}_{x\_t}, t)\|^2\Big]
\end{aligned}
$$

Network to predict the noise at each step

Training data

Generative Objective: Reverse Process



Minimizing Variational Lower Bound:

$$
\begin{aligned}
L_{\mathrm{VLB}} &= L_T + L_{T-1} + \cdots + L_0 \\
\text{where } L_T &= D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T)) \\
&\boxed{L_t = D_{\mathrm{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \text{ for } 1 \le t \le T-1}
\end{aligned}
$$

=

Minimizing predicted noise based on data:

$$
\begin{aligned}
L_t^{\text{simple}} &= \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\Big[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\Big] \\
&= \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\Big[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar\alpha_t}\mathbf{x}_0 + \sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}_t, t)\|^2\Big]
\end{aligned}
$$

In code:

Algorithm 1 Training

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

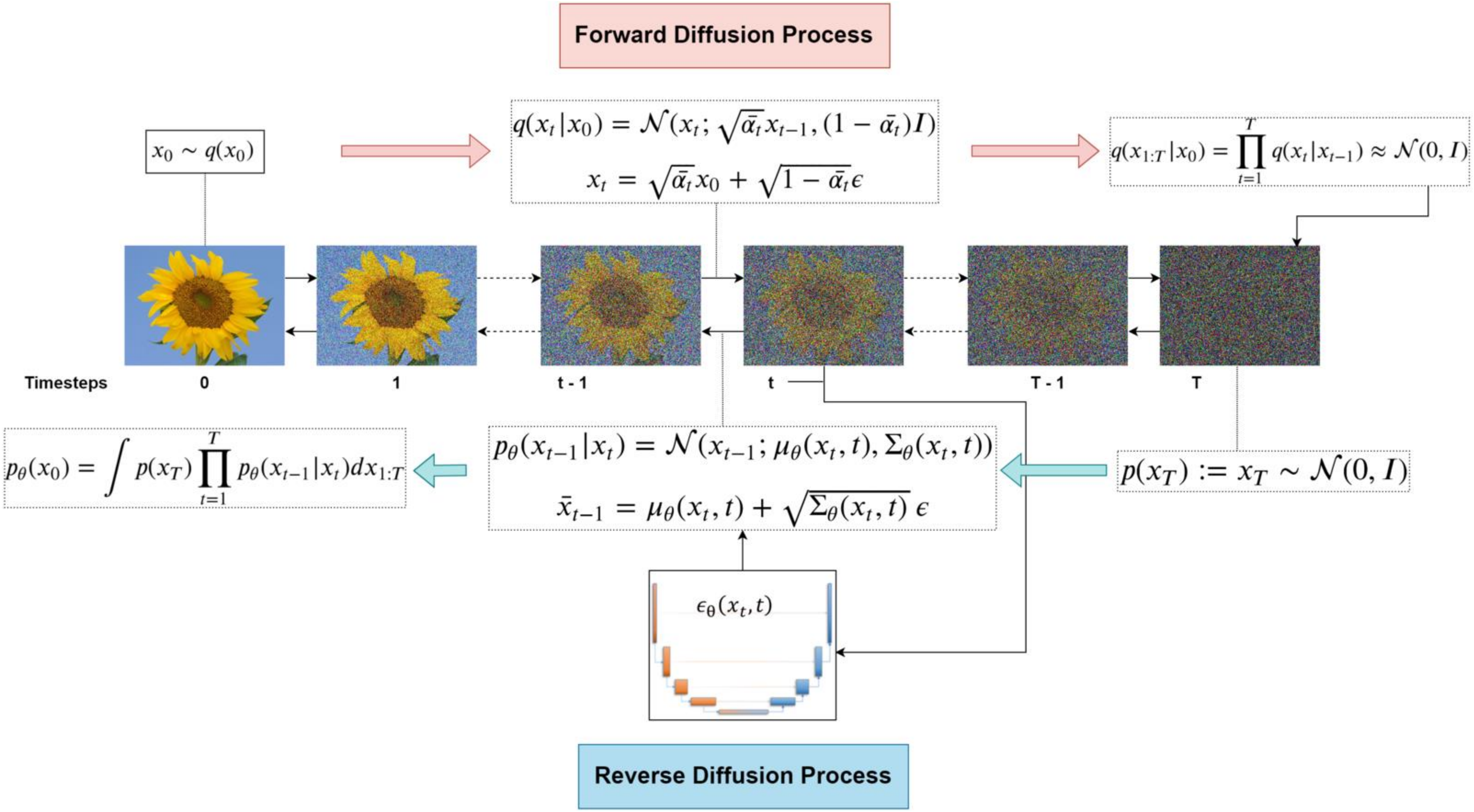
Algorithm 2 Sampling

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

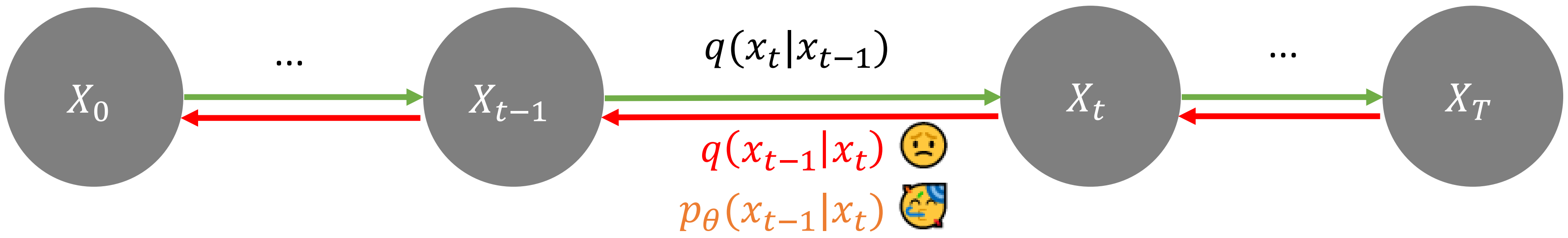
$$\boxed{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde\beta_t \mathbf{I})}$$

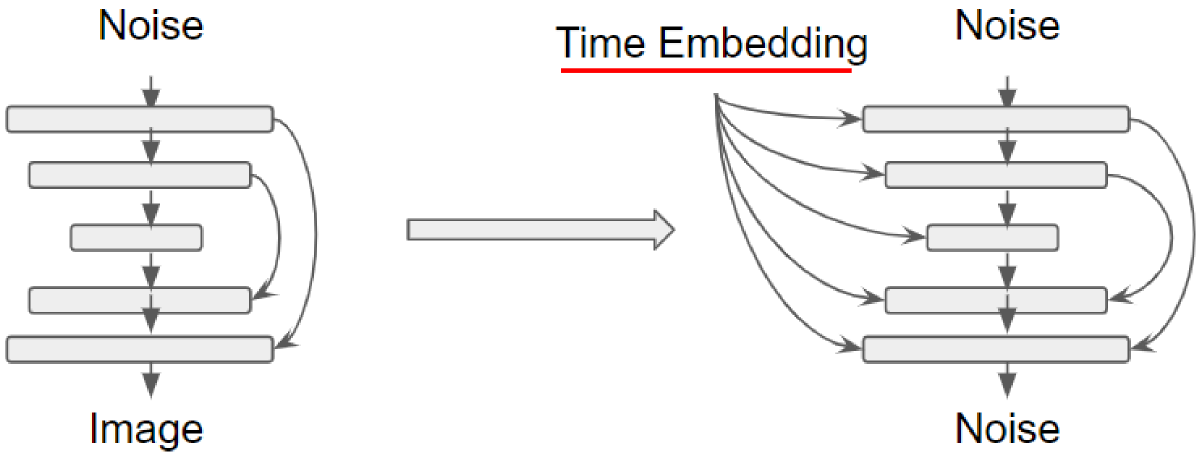
$$\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\varepsilon_t\right) \qquad \frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t} \cdot \beta_t$$

Denoising Diffusion Probabilistic Model (DDPM)



Generative Objective: Reverse Process





A standard U-Net to predict noise
from previous noise and time information.

U-Net for standard encoding

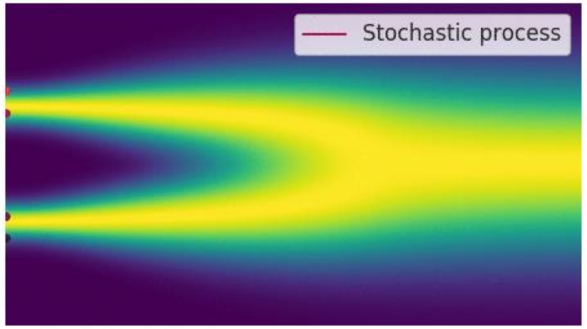
A (very simple) U-Net for diffusion Model

Generative Objective: Reverse Process

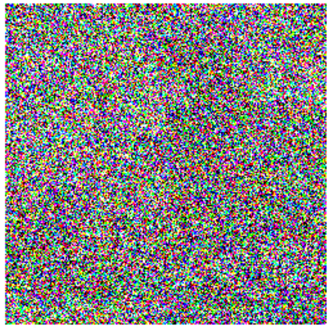
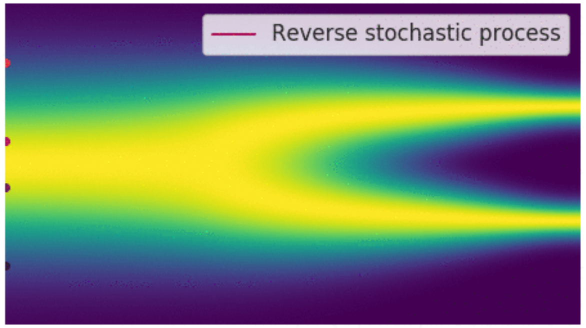
X_0 ... X_{t-1} $\xrightarrow{q(x_t|x_{t-1})}$ X_t ... X_T

$q(x_{t-1}|x_t)$ 😟

$p_\theta(x_{t-1}|x_t)$ 🥳

Stochastic noising

Stochastic denoising

Results: DDPM

Sampled results:
LSUN Church Dataset

Sampled results:
LSUN Church Bedroom

Results: DDPM

Sampled results:
CelebA-HQ Dataset

Generative Objective: Reverse Process
Convergence

- GANs: if the Discriminator can successfully differentiate between real/fake then we stop the training

- VAE: reconstruction loss (meaningful)

- Diffusion models: complicated: we need to measure the distance between two distributions → FID

DDPM or Score-Matching?

Denoising formula:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \underbrace{\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}_{\text{Look like a gradient!}}\right) + \sigma_t \mathbf{z}$$

From Tweedie's Formula for any $z \sim \mathcal{N}(z; \mu_z, \Sigma_z)$

$$\mu_z = z + \Sigma_z \nabla \log p(z) \qquad (1)$$

$$\nabla \log p_\theta(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p_\theta(\boldsymbol{x})$$

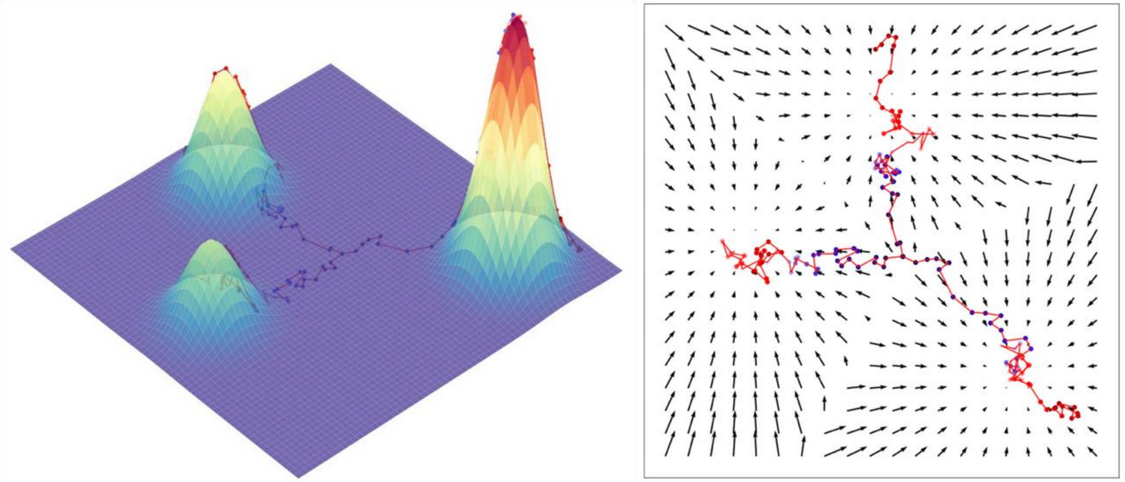
We also know:

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right) \qquad (2)$$

Together (1)(2) we have:

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

$$\nabla \log p(\boldsymbol{x}_t) = -\frac{\varepsilon_t}{\sqrt{1-\bar{\alpha}_t}} \quad \text{score-function}$$



Score Vector towards the data

Luo, Calvin. "Understanding diffusion models: A unified perspective." arXiv preprint arXiv:2208.11970 (2022).

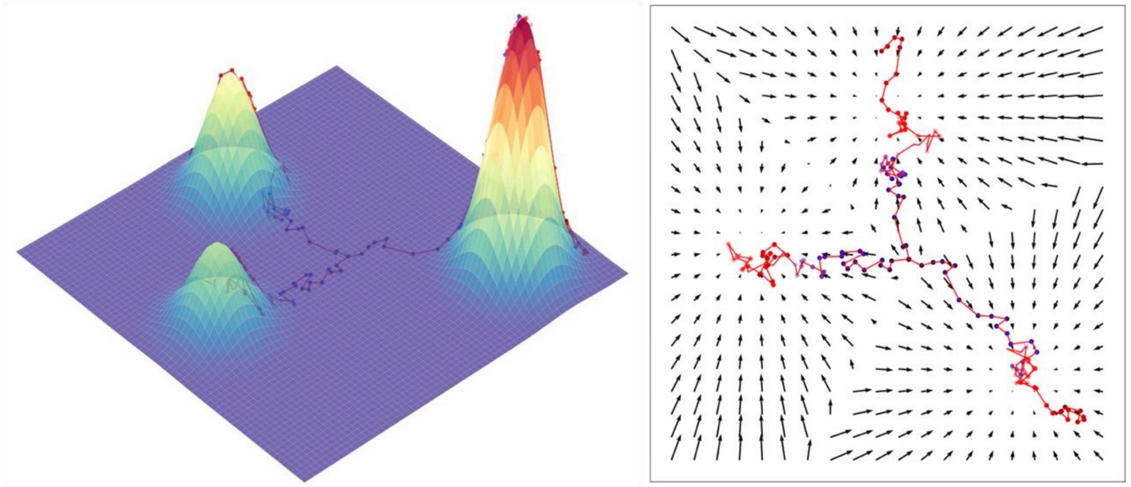
Forward Sampling is:

$$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t + \delta \nabla \log p(\boldsymbol{x}_t) + \sqrt{2\delta}\varepsilon_t, \quad t = 0, 1, \cdots, T$$

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

sigma is scheduler

DDPM or Score-Matching?



Score Vector towards the data

Luo, Calvin. "Understanding diffusion models: A unified perspective." arXiv preprint arXiv:2208.11970 (2022).

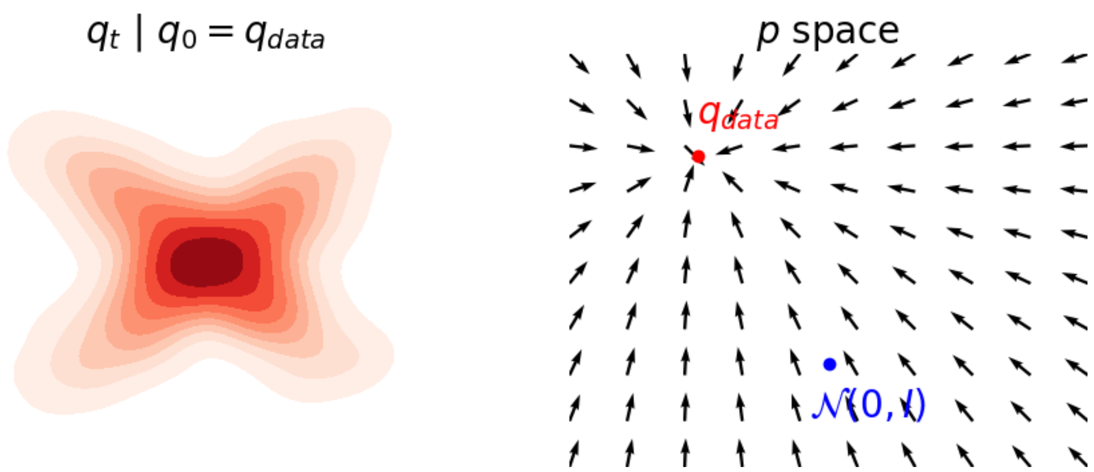
Forward Sampling is:

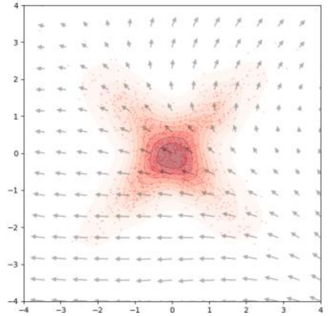
$$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t + \delta \nabla \log p\left(\boldsymbol{x}_t\right) + \sqrt{2\delta}\varepsilon_t, \quad t = 0, 1, \cdots, T$$

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

sigma is scheduler

Idea is to have the reverse, it can be computed by modelising this process a SDE and a reverse SDE.

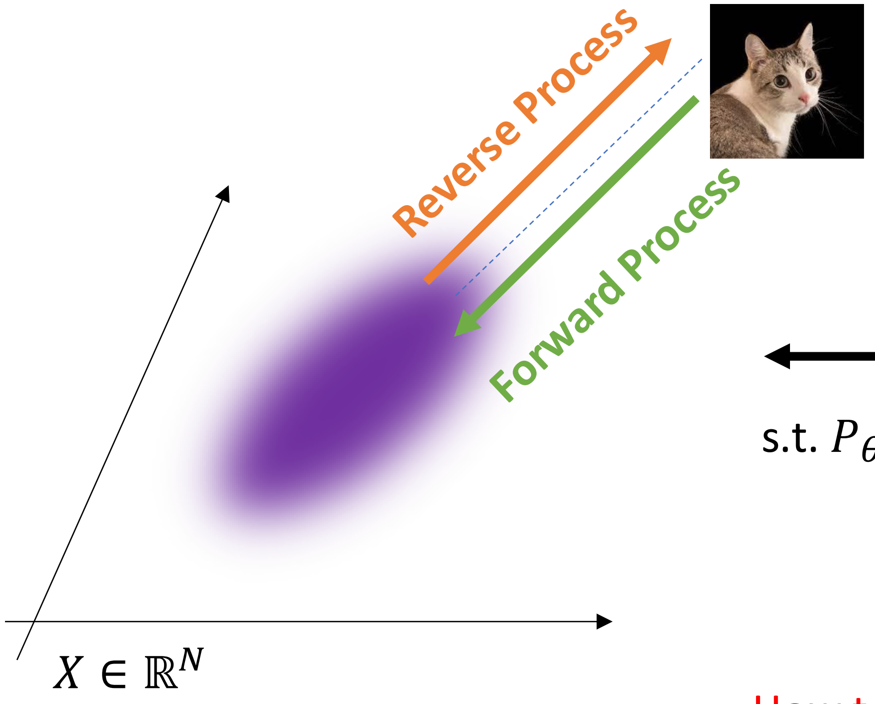




Network learns a sth like a gradient field (score vector)

$$s(x) = \nabla_x \log p(x)$$

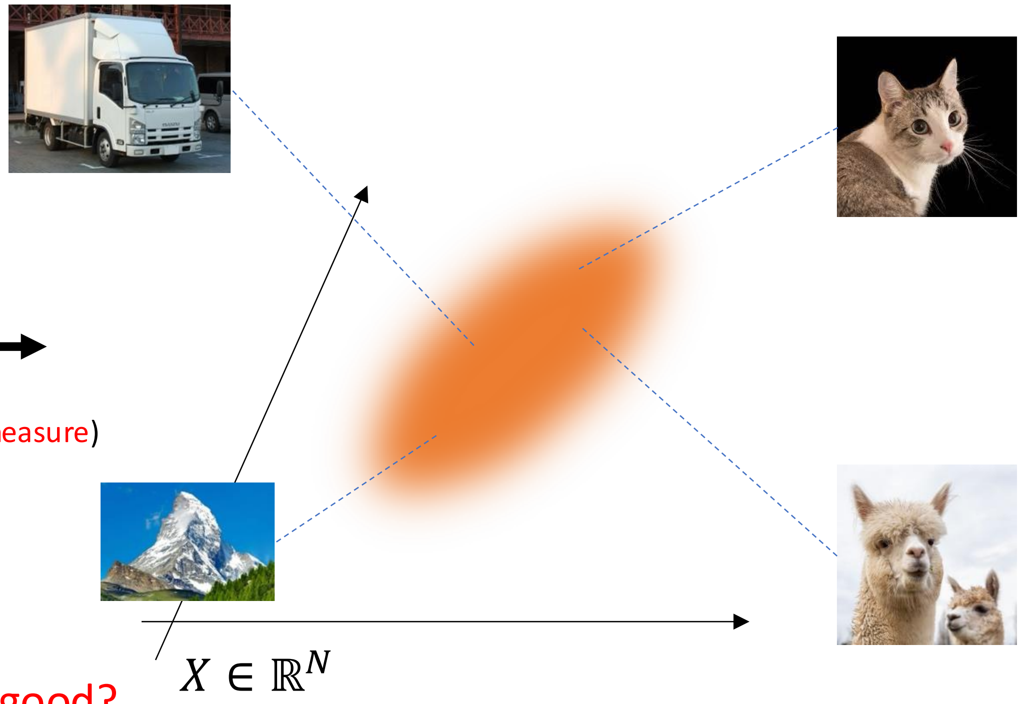
Metric: FID (Fréchet inception distance)



find best θ

s.t. $P_\theta(X) \approx P(X)$ (with a certain measure)

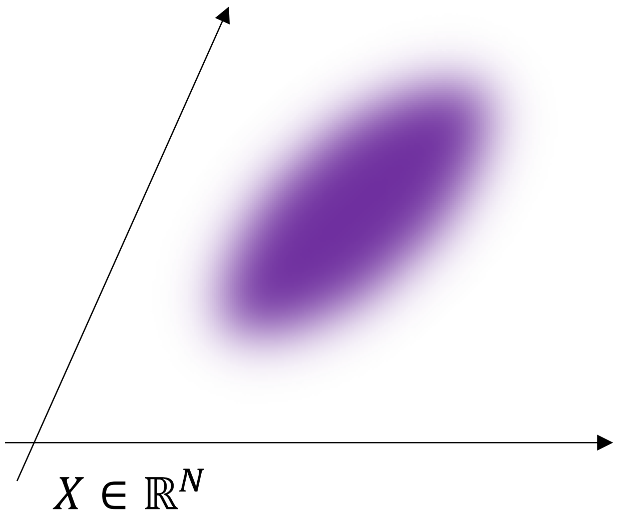
How to know if the results are good?



Distribution of Learnt Data $P_\theta(X)$
with parameter $\theta \in \mathbb{R}^M$

Distribution of Real Data $P(X)$

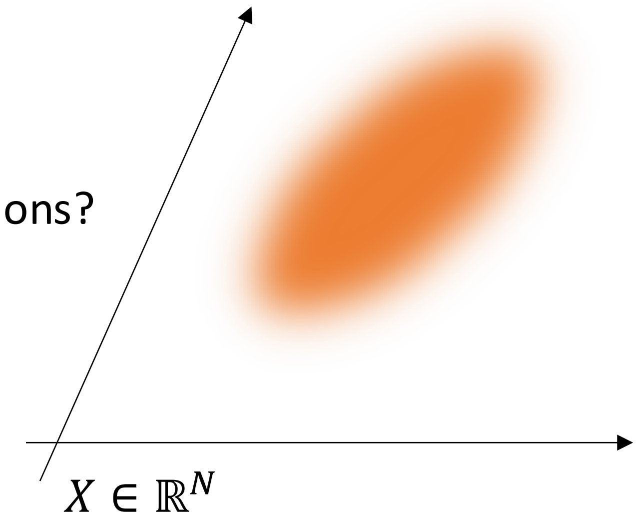
Metric: FID (Fréchet inception distance)



Distribution of Learnt Data $P_\theta(X)$
with parameter $\theta \in \mathbb{R}^M$

How to know if the results are good?

Measure distance between two distributions?



Distribution of Real Data $P(X)$

Metric: FID (Fréchet inception distance)

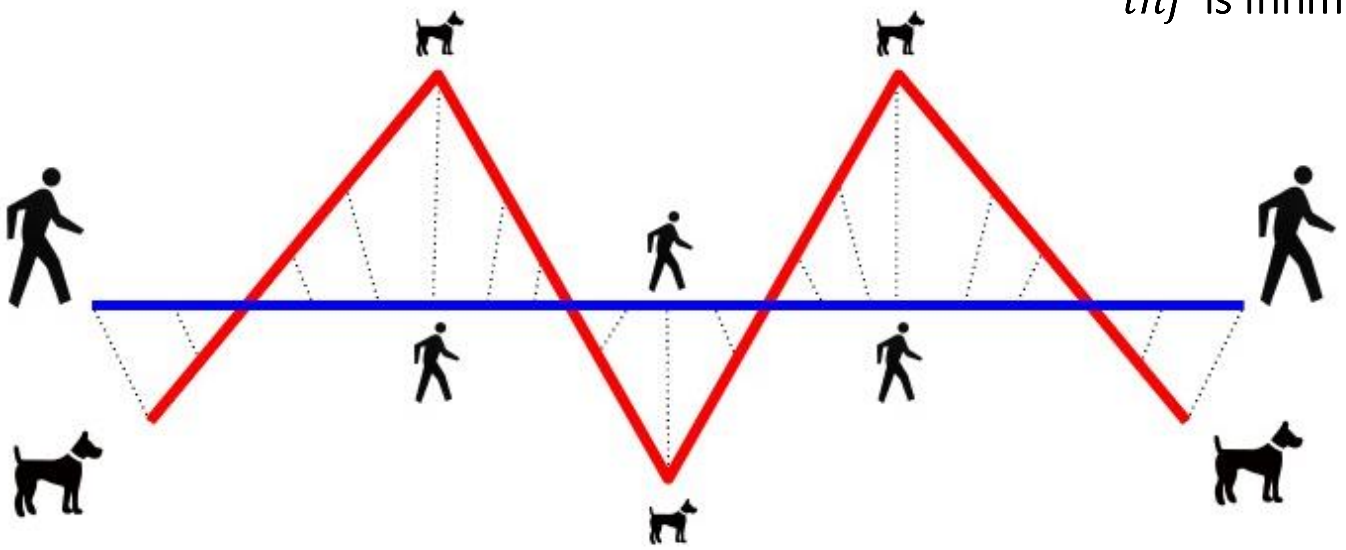
Fréchet Distance: How to walk your dog?

$$F(A,B) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \left\{ d\Big(A\big(\alpha(t)\big),\, B\big(\beta(t)\big)\Big) \right\}$$

α, β are *all* reparameterizations (different velocity possibilities)

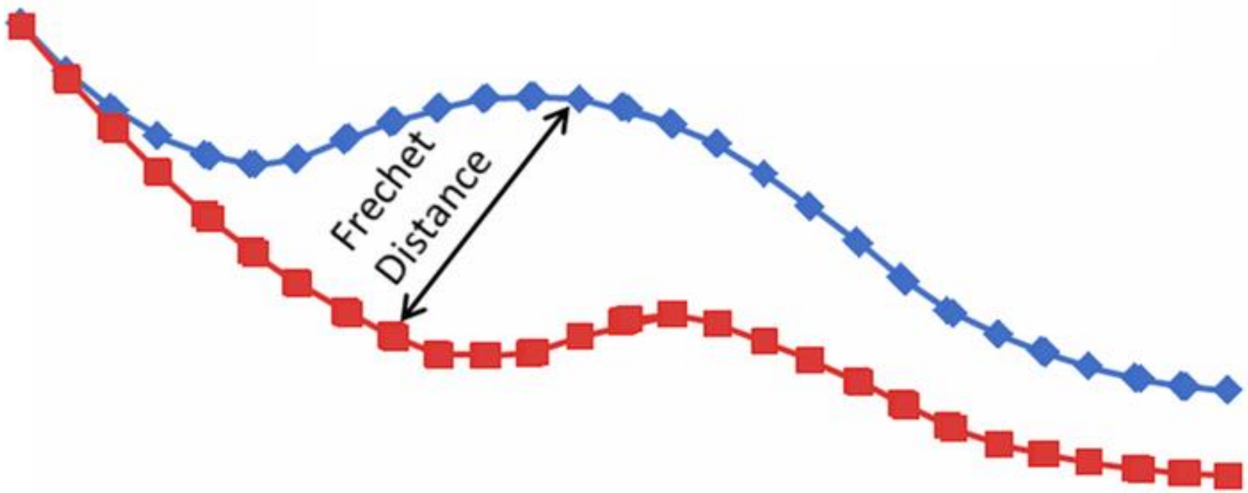
d is a distance measure

inf is infimum, i.e., *greatest lower bound*



Metric: FID (Fréchet inception distance)

$$d_F(\mu,\nu) := \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^n\times\mathbb{R}^n} \|x-y\|^2 \, \mathrm{d}\gamma(x,y)\right)^{1/2}$$



Fréchet Distance: How to walk your dog?

Metric: FID (Fréchet inception distance)

$$d_F(\mu,\nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \, \mathrm{d}\gamma(x,y) \right)^{1/2}$$

If they are multidimensional gaussians:

$$d_F(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\mu',\Sigma'))^2 = \|\mu - \mu'\|_2^2 + \mathrm{tr}\left(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{\frac{1}{2}}\right)$$

Instead of computing image pixel information,
we compute the deepest layer of Inceptionv3 network trained on ImageNet dataset