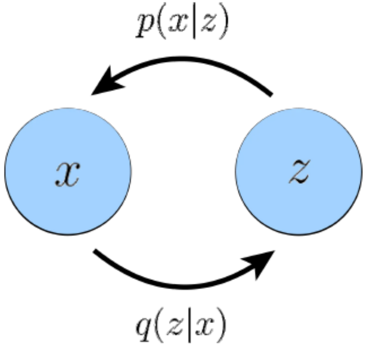
Fundamentals of Machine Learning

DIFFUSION MODELS

Amit K Roy-Chowdhury

Acknowledgments: https://calvinyluo.com/2022/08/26/diffusion-tutorial.html#evidence-lower-bound
https://arxiv.org/abs/2403.18103

Standard VAE



$$\ln p(\mathbf{x}) \geq \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction Error}} - \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln q_\phi(\mathbf{z}|\mathbf{x}) - \ln p(\mathbf{z})\right]}_{\text{KL Divergence}}$$

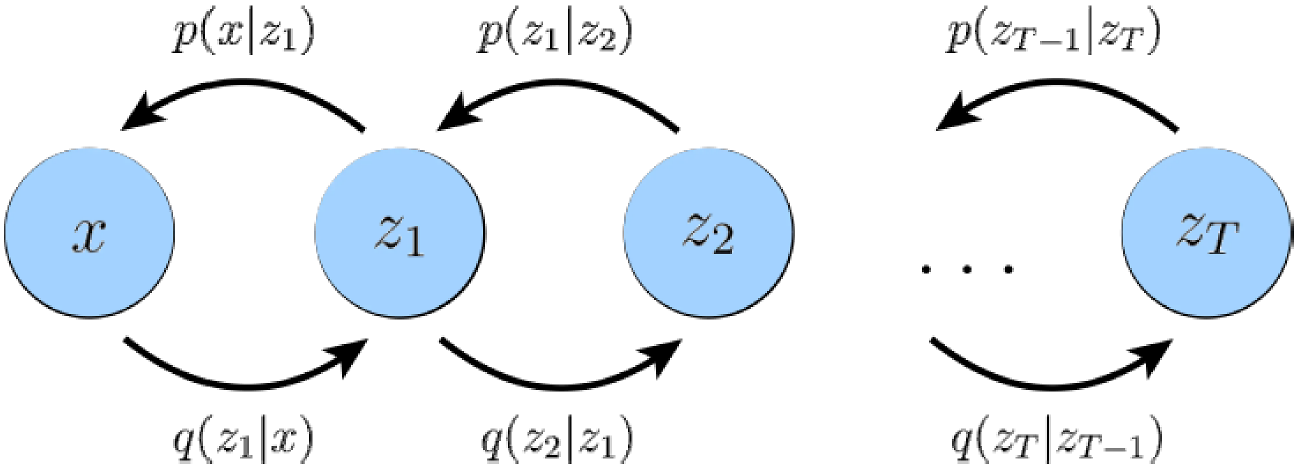
Objective: $$\underset{\phi,\theta}{\arg\max}\, \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \mathcal{D}_{\mathrm{KL}}(q_\phi(z \mid x) \,||\, p(z))$$

Markov Chain

A discrete-time Markov chain $\{X_n | n = 0, 1, \ldots\}$ *is a discrete-time, discrete-value random sequence such that given* $X_0, \ldots, X_n$, *the next random variable* X_{n+1} *depends only on* X_n *through the transition probability*

$$P\left[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\right] = P\left[X_{n+1} = j | X_n = i\right] = P_{ij}.$$

Hierarchical VAE

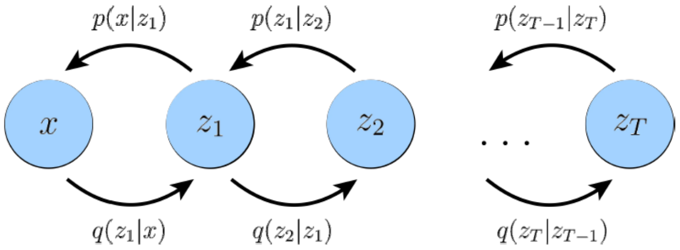


Markovian Hierarchical VAE: each z_t conditions only on the previous latent variable.

Hierarchical VAE

$$p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t)$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}_1 \mid \boldsymbol{x})\prod_{t=2}^{T} q_{\boldsymbol{\phi}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1})$$



ELBO for MHVAE

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) d\boldsymbol{z}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})}{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} d\boldsymbol{z}_{1:T} \\
&= \log \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} \left[\frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} \right] \\
&\geq \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} \right]
\end{aligned}
$$

ELBO for VAE

$$
\begin{aligned}
\ln p(\mathbf{x}) &= \ln \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \ln \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \ln \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})} \left[\frac{p(\mathbf{x}|\mathbf{z}) p(\mathbf{z})}{q_\phi(\mathbf{z})} \right] \\
&\geq \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})} \ln \left[\frac{p(\mathbf{x}|\mathbf{z}) p(\mathbf{z})}{q_\phi(\mathbf{z})} \right] \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})} \left[\ln p(\mathbf{x}|\mathbf{z}) + \ln p(\mathbf{z}) - \ln q_\phi(\mathbf{z}) \right] \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})} \left[\ln p(\mathbf{x}|\mathbf{z}) \right] - \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})} \left[\ln q_\phi(\mathbf{z}) - \ln p(\mathbf{z}) \right]
\end{aligned}
$$

Plugging in from previous slide:

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

Variational Diffusion Models (VDMs)

Latent dimension is the same as data dimension: both data and latent variables as $\boldsymbol{x}_t$

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

Structure of the latent encoder at each step is a linear Gaussian model

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

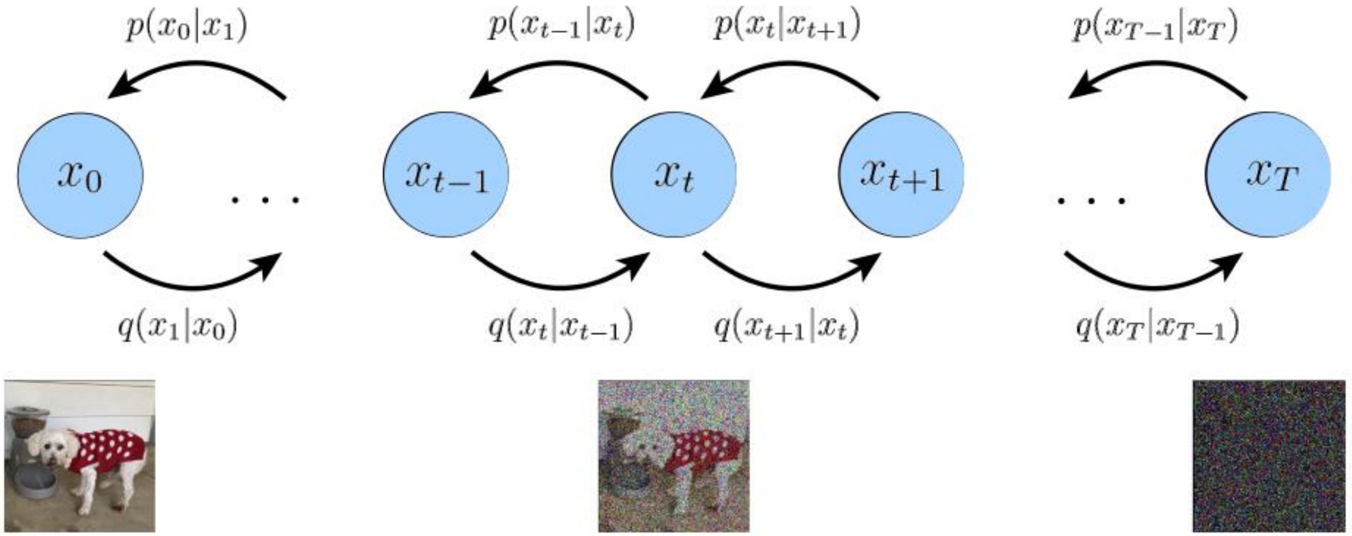
α_t evolves such that the latent at the final timestep is a standard Gaussian.

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

where,

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

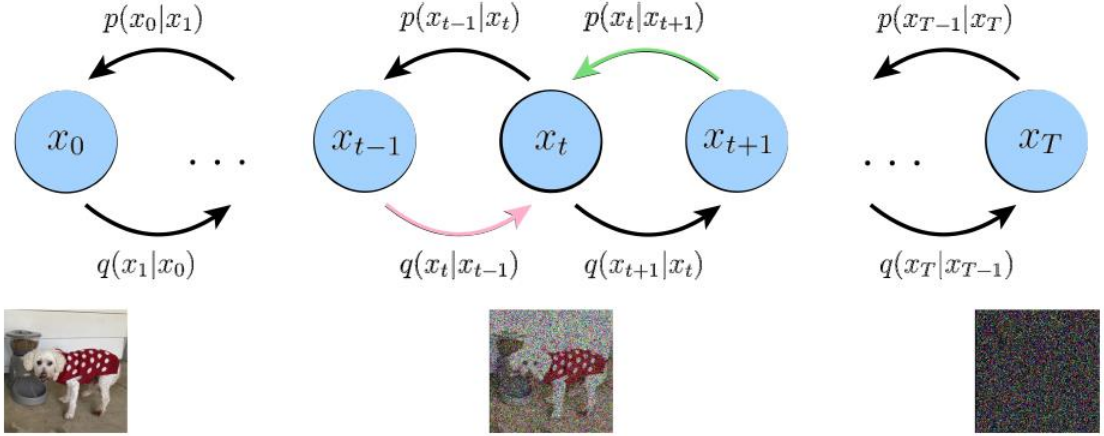
Variational Diffusion Models



- Learn $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$
- Sample Gaussian noise from $p(\boldsymbol{x}_T)$ and denoise based on $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ for T steps

ELBO for VDMs

$$
\begin{aligned}
\log p(\boldsymbol{x}) &\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}\left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \,||\, p(\boldsymbol{x}_T))\right]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \,||\, p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}
\end{aligned}
$$



VDM Distribution

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$$

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

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I})$$

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I})$$

Conditional Distribution:

$$q_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t \mid \sqrt{\overline{\alpha}_t}\mathbf{x}_0, \ (1-\overline{\alpha}_t)\mathbf{I}),$$

$$\text{where } \overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i.$$

VDM Distribution - Proof

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \underbrace{\sqrt{\alpha_t}\sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}}_{\mathbf{w}_1}.
\end{aligned}$$

Sum of two Gaussians is a Gaussian

Mean is zero

Covariance:

$$\begin{aligned}
\mathbb{E}[\mathbf{w}_1\mathbf{w}_1^T] &= [(\sqrt{\alpha_t}\sqrt{1-\alpha_{t-1}})^2 + (\sqrt{1-\alpha_t})^2]\mathbf{I} \\
&= [\alpha_t(1-\alpha_{t-1}) + 1 - \alpha_t]\mathbf{I} = [1-\alpha_t\alpha_{t-1}]\mathbf{I}
\end{aligned}$$

Thus:

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}\mathbf{x}_{t-3} + \sqrt{1-\alpha_t\alpha_{t-1}\alpha_{t-2}}\boldsymbol{\epsilon}_{t-3} \\
&= \vdots \\
&= \left(\sqrt{\prod_{i=1}^{t}\alpha_i}\right)\mathbf{x}_0 + \left(\sqrt{1-\prod_{i=1}^{t}\alpha_i}\right)\boldsymbol{\epsilon}_0.
\end{aligned}$$

Define $\overline{\alpha}_t = \prod_{i=1}^{t}\alpha_i$

Then: $\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}_0.$ $\Longleftrightarrow$ $\mathbf{x}_t \sim q_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t \,|\, \sqrt{\overline{\alpha}_t}\mathbf{x}_0,\ (1-\overline{\alpha}_t)\mathbf{I})$