Denoising Diffusion Probabilistic Models

Frédéric Richard

M2 MAS - DS, course maths for DS, 2025

Réference. C. Luo, Understanding diffusion models, a unified perspective, 2022.

1 Introduction.

1.1 Generative model and latent variables.

Goal generative models: given observed samples x from a distribution of interest, learn the true data distribution $p(x)$ and be able to generate new samples.

Strategy: observed data are associated to (or generated from an unobserved latent variables z. Learn how to generate x from z.

p may be expressed as the marginal distribution

$$p(x) = \int p(x, z) dz.$$

or, with respect to joint and posterior distribution, as

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

Computing or optimizing $p(x)$ is usually difficult either because the marginal distribution is intractable or there is no ground truth for $p(z|x)$.

1.2 Evidence Lower Bound (ELBO).

A flexible parametrized distribution $q_\phi(z|x)$ is used to approximate $p(z|x)$.

Expression of the evidence

$$\log(p(x)) = \mathbb{E}_{q_\phi(z|x)}\left(\log\left(\frac{p(x, z)}{q_\phi(z|x)}\right)\right) + D_{KL}(q_\phi(z|x)\|p(z|x)),$$

where

$$D_{KL}(q \mid p) = \mathbb{E}_{q(w)}\left(\log\left(\frac{q(w)}{p(w)}\right)\right) = \int \log\left(\frac{q(w)}{p(w)}\right) q(w) dw.$$

ELBO

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

Since $D_{KL} \geq 0$,

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

As the evidence $\log(p(x))$ is constant, the D_{KL} between approximate and true posterior distributions decreases as the ELBO increases. Hence, maximizing the ELBO amounts to minimizing the D_{KL} between approximate and true posterior distributions. Therefore, maximizing the ELBO enables to match the approximate posterior distribution to the true posterior distribution without knowing the true posterior distribution.

2 Variational autoencoder

2.1 Setting

- Encoder : transitions $x \to z$ described through $q_\phi(z|x)$
- Decoder : transitions $z \to x$ described through $p_\theta(x|z)$.

Form of the ELBO

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_{q_\phi(z|x)} \left(\log \left(\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right) \right) \\ &= \mathbb{E}_{q_\phi(z|x)} \left(p_\theta(x|z) \right) - D_{KL} \left(q_\phi(z|x) \| p(z) \right). \end{aligned}$$

In the framework of a VAE, the ELBO is maximized simultaneously in ϕ and θ so as to learn both distributions q_ϕ and p_θ of the encoder and decoder.

The ELBO decomposes into

- a **reconstruction term** $\mathbb{E}_{q_\phi(z|x)} \left(p_\theta(x|z) \right)$. Decreasing this term ensures that the data produced by the decoder from the latent variable can be close to the original data.
- a **prior matching term** $-D_{KL} \left(q_\phi(z|x) \| p(z) \right)$. Decreasing this terms ensures that the learned encoder distribution is as close as possible to a belief prior distribution of the latent variables. It acts as a regularization term in the maximisation of the ELBO.

Usual forms of

- $q_\phi(z|x)$: multivariate Gaussian distribution with mean $\mu_\phi(x)$ and covariance matrix $\sigma^2_\phi(x)I$,
- and $p(z)$: Standard multivariate Gaussian.

Using these forms, the D_{KL} of the matching prior term can be expressed analytically.

The reconstruction term is approximated by an MC estimate $\sum_{l=1}^{L} \log(p_\theta(x|z^{(l)})$, where $z^{(l)}$ are sampled from $q_\phi(z|x)$.

Reparametrization trick:

$$z^{(l)} = \mu_\phi(x) + \sigma^2_\phi(x) \odot \epsilon^{(l)}$$

where $\epsilon^{(l)}$ i.i.d realizations of a $\mathcal{N}(0, I)$.

This trick enables to compute gradients.

2.2 2.2. Hierarchical variational autoencoders.

- A hierarchy $z_{1:T}$ of T latent variables describing a succession of encoding.
- Markovian assumption

$$p(z_t|z_{1:t-1}) = p(z_t|z_{t-1}).$$

3 Diffusion models.

3.1 Definition.

Diffusion model = Hierarchical Markovian variational autoencoder where

- latent variables have the same dimension as x.
- the structure of the latent encoder is not learned but pre-defined,
- the encoder makes the image evolves to a final latent variable T distributed as a standard Gaussian (pure noise).

Let $x_0 = x$ and set $x_t = z_t$. We have

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}),$$

due to the Markovian property.

Modeling of the distribution $x_t|x_{t-1}$ (**variance-preserving scheme**):

$$\mathcal{N}(\sqrt{\alpha_t}x_t, (1-\alpha_t)I)$$

for a fixed or learned schedule ensuring that x_T has a standard Gaussian distribution. The encoding process accounts for a noisification of the image : the image is progressively corrupted until it becomes a pure noise.

Joint distribution of the decoder :

$$p(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t),$$

with x_T standard multivariate Gaussian.

3.2 Expressions of the ELBO.

We have

$$\begin{aligned}
ELBO &= \mathbb{E}_{q(x_{1:T}|x_0)}\left(\log\left(\frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right)\right)\\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left(\log\left(\frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\right)\right)\\
&= \mathbb{E}_{q(x_1|x_0)}\left(\log(p_\theta(x_0|x_1)\right) + \mathbb{E}_{q(x_{T-1},x_T|x_0)}\left(\log\left(\frac{p(x_T)}{q(x_T|x_{T-1})}\right)\right)\\
&\quad + \sum_{t=1}^{T-1}\mathbb{E}_{q(x_{t-1},x_t,x_{t+1}|x_0)}\left(\log\left(\frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right)\right)
\end{aligned}$$

Expression of the ELBO (version 1)

$$
\begin{aligned}
ELBO = &\mathbb{E}_{q(x_1|x_0)}\left(\log(p_\theta(x_0|x_1)\right) - \mathbb{E}_{q(x_{T-1}|x_0)}\left(D_{KL}(q(x_T|x_{T-1})\|p(x_T))\right) \\
&- \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t-1},x_{t+1}|x_0)} D_{KL}\left(q(x_t|x_{t-1})\|p_\theta(x_t|x_{t+1})\right)
\end{aligned}
$$

Interpretation :

- $\mathbb{E}_{q(x_1|x_0)}\left(\log(p_\theta(x_0|x_1)\right)$: **reconstruction term**.
- $-\mathbb{E}_{q(x_{T-1}|x_0)}\left(D_{KL}(q(x_T|x_{T-1})\|p(x_T))\right)$: **prior matching term**,
- $-\mathbb{E}_{q(x_{t-1},x_{t+1}|x_0)} D_{KL}\left(q(x_t|x_{t-1})\|p_\theta(x_t|x_{t+1})\right)$: **consistency terms**. it enforces the denoising distribution of x_t given x_{t+1} to match the noising distribution of x_t given x_{t-1}.

Expectation terms of this expression will be estimated by MC methods. The consistency terms depend on two variables x_{t-1} and x_t, which is sub-optimal. Another decomposition of the ELBO is derived to alleviate this issue.

3.3 Decomposition of the ELBO (version 2).

The main idea is to keep the conditionning on x_0 in the decomposition

$$
q(x_{1:T}|x_0) = q(x_1|x_0) \prod_{t=2}^{T} q(x_t|x_{t-1}, x_0)
$$

of the joint distribution, and write

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

as a function of the backward distribution $\pi(x_{t-1}|x_t, x_0)$ conditioned by the original data.

This leads to

$$
\begin{aligned}
q(x_{1:T}|x_0) &= q(x_1|x_0) \prod_{t=2}^{T} \frac{\pi(x_{t-1}|x_t, x_0) q(x_t|x_0)}{q(x_{t-1}|x_0)} \\
&= q(x_T|x_0) \prod_{t=2}^{T} \pi(x_{t-1}|x_t, x_0).
\end{aligned}
$$

Consently, the ELBO may written as

$$
\begin{aligned}
ELBO &= \mathbb{E}_{q(x_{1:T}|x_0)}\left(\log\left(\frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right)\right) \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left(\log\left(\frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_T|x_0)\prod_{t=2}^{T} \pi(x_{t-1}|x_t, x_0).}\right)\right) \\
&= \mathbb{E}_{q(x_1|x_0)}\left(\log(p_\theta(x_0|x_1)\right) + \mathbb{E}_{q(x_T|x_0)}\left(\log\left(\frac{p(x_T)}{q(x_T|x_0)}\right)\right) \\
&\quad + \sum_{t=2}^{T} \mathbb{E}_{q(x_{t-1},x_t|x_0)}\left(\log\left(\frac{p_\theta(x_{t-1}|x_t)}{\pi(x_{t-1}|x_t, x_0)}\right)\right)
\end{aligned}
$$

Expression of the ELBO (version 2)

$$\begin{aligned} ELBO = & \mathbb{E}_{q(x_1|x_0)}\left(\log(p_\theta(x_0|x_1)\right) - D_{KL}\left(q(x_T|x_0)\|p(x_T)\right) \\ & - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left(D_{KL}\left(\pi(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)\right)\right) \end{aligned}$$

Interpretation :

- $\mathbb{E}_{q(x_1|x_0)}\left(\log(p_\theta(x_0|x_1)\right)$: **reconstruction term**.
- $-D_{KL}(q(x_T|x_0)\|p(x_T))$: **prior matching term**,
- $-\mathbb{E}_{q(x_t|x_0)}\left(D_{KL}\left(\pi(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)\right)\right)$: **denoising matching term**. The learned transition distribution $p_\theta(x_{t-1}|x_t)$ characterizing the denoising of x_t should be as close as possible to the reference transition distribution $\pi(x_{t-1}|x_t, x_0)$.

3.4 Expression of the reference backward distribution \pi.

We have

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

Due to Markovian property, $q(x_t|x_{t-1}, x_0) = q(x_t|x_{t-1})$.

We seek for an expression of $q(x_{t-1})$ under the variance-preserving scheme :

$$x_t|x_{t-1} \sim \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$$

Including the parametrization trick, we have

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1},$$

where ϵ_{t-1} is standard multivariate Gaussian.

Notice that, for $t \geq 2$,

$$\begin{aligned} x_t = & \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-2}\right) + \sqrt{1-\alpha_t}\epsilon_{t-1}, \\ = & \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-1}, \\ = & \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1} + (1-\alpha_t)}\epsilon^*_{t-1}, \\ = & \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon^*_{t-1}, \end{aligned}$$

for some standard multivariate Gaussian variables ϵ^*_{t-1}. By induction, it follows that, for $t > 0$,

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon^*_0,$$

with

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s.$$

In other words, $x_t|x_0 \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$.

Then, we can deduce

$$\begin{aligned}
\pi(x_{t-1}|x_t, x_0) &= \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)}, \\
&\propto \exp\Big\{ -\frac{1}{2(1-\alpha_t)} \left|x_t - \sqrt{\alpha_t}x_{t-1}\right|^2 \\
&\quad - \frac{1}{2(1-\bar{\alpha}_{t-1})} \left|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\right|^2) \\
&\quad + \frac{1}{2(1-\bar{\alpha}_t)} \left|x_t - \sqrt{\bar{\alpha}_t}x_0\right|^2) \Big\} \\
&\propto \exp\left(-\frac{1}{2\sigma_\pi^2(t)} \left|x_{t-1} - \mu_\pi(x_t, x_0)\right|^2 \right),
\end{aligned}$$

where

$$\begin{cases} \mu_\pi(x_t, x_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_t)x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t} \\ \sigma_\pi^2(t) &= \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \end{cases}$$

3.5 Expression of the denoising matching term (version 1).

To approximate the reference distribution π, the $p_\theta(x_{t-1}|x_t)$ is also defined as a multivariate Gaussian distribution of means $\mu_\theta(x_t, t)$ and covariance matrix $\sigma_\pi^2(t)I$.

The KL Divergence between two multivariate Gaussian distributions $\mathcal{N}(\mu_x, \Sigma_x)$ and $\mathcal{N}(\mu_y, \Sigma_y)$ can be analytically expressed as

$$\frac{1}{2}\left(\log\frac{|\Sigma_y|}{|\Sigma_x|} - d + \mathrm{trace}(\Sigma_y^{-1}\Sigma_x) + (\mu_x - \mu_y)^T\Sigma_y^{-1}(\mu_x - \mu_y) \right)$$

Hence,

$$D_{KL}\left(\pi(x_{t-1}|x_t, x_0)\right) \| p_\theta(x_{t-1}|x_t)) = \frac{1}{2\sigma_\pi^2(t)} \left|\mu_\theta(x_t) - \mu_\pi(x_t, x_0)\right|^2$$

According to the expression of $\mu_\pi(x_t, x_0)$, it is relevant to express $\mu_\theta(x_t)$ in the form

$$\mu_\theta(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_t)x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{x}_\theta(x_t, t)}{1-\bar{\alpha}_t},$$

where $\hat{x}_\theta(x_t, t)$ stands for an estimate of x_0 from x_t.

Using this form, the KL divergence write

$$D_{KL}\left(\pi(x_{t-1}|x_t, x_0)\right) \| p_\theta(x_{t-1}|x_t)) = \frac{1}{2\sigma_\pi^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \left|\hat{x}_\theta(x_t, t) - x_0\right|^2.$$

3.6 Expression of the denoising matching term (version 2).

Recall that

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0^*,$$

so that

$$x_0 = \frac{x_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon_0^*,$$

Plugging x_0 in the expression of μ_π, it can be shown that

$$\mu_\pi(x_t, x_0) = \frac{x_t}{\sqrt{\alpha_t}} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_0^*.$$

Expressing $\mu_\theta(x_t, t)$ in the form

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

the KL divergence can be written

$$D_{KL}\left(\pi(x_{t-1}|x_t, x_0)\right)\|p_\theta(x_{t-1}|x_t)) = \frac{1}{2\sigma_\pi^2(t)}\frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\left|\hat{\epsilon}_\theta(x_t, t) - \epsilon_0^*\right|^2.$$