Generative Models - DRE7053
Lecture 4

NORA Summer School 2024

Rogelio A Mancisidor
Assistant Professor
Department of Data Science and Analytics
BI Norwegian Business School

June 10-14, 2024

Outline

1. Diffusion Models
   - Denoising Diffusion Models
   - Forward Diffusion Process
   - Backward Diffusion Process
2. The Objecttive Function
   - Deriving the ELBO

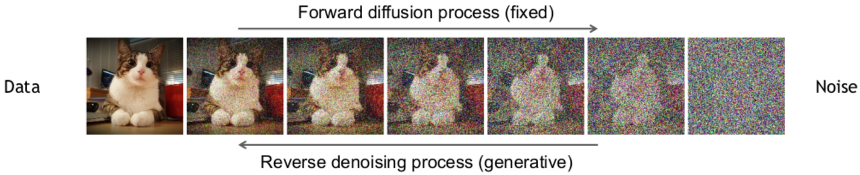
Outline

1. Diffusion Models
   - Denoising Diffusion Models
   - Forward Diffusion Process
   - Backward Diffusion Process

2. The Objecttive Function
   - Deriving the ELBO

Forward and Backward steps

- Denoissing diffusion models has two steps:
  1. The forward diffusion process
  2. The backward process, which learns to generate the data



The Joint Distribution of the Diffusion Process

- At each step in the forward diffusion process we have

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

- Hence, the joint probability of the entire process from $\boldsymbol{x}_0$ (original sample) to $\boldsymbol{x}_t$ is

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

  where the variance β_t increases with t
- Note that each step depends only on the its previous state (a Markovian process!)
- Which is similar to the inference model in hierarchical VAEs

The Joint Distribution of the Diffusion Process

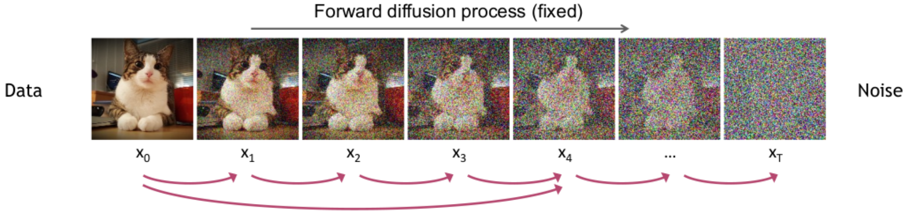
- At each step, we use the same sampling mechanism as in VAEs:

- Hence we can sample at each t from

$$\boldsymbol{x}_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1 - \bar{\alpha}_t)\boldsymbol{I}) \tag{2}$$

- β_t schedule is designed such that $\bar{\alpha}_T \to 0$, so $q(\boldsymbol{x}_T|\boldsymbol{x}_0) \approx \mathcal{N}(\mathbf{0}, \mathbf{1})$

Forward Diffusion Process



- Note that
  - In the forward step we do not train any neural network
  - The distribution $q(\boldsymbol{x}_T|\boldsymbol{x}_0)$ is the same as the prior distribution in VAEs

Intuition

- The forward diffusion process is an iterative procedure that transforms the input data $\boldsymbol{x}_0$
- We add *noise* to $\boldsymbol{x}_0$ such that at time T

$$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

- That is, the mean and variance of $\boldsymbol{x}_0$ moves gradually towards $\mathbf{0}$ and $\mathbf{1}$
- The unknown distribution of $\boldsymbol{x}_0$ should be more complex than $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$

Backward Diffusion Process - I

- We want to *undo* the forward diffusion process
- Ideally, we want to estimate $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ recursively
- However,

$$\begin{aligned} q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) =& \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})q(\boldsymbol{x}_{t-1})}{q(\boldsymbol{x}_t)} \\ \propto& q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \int q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)q(\boldsymbol{x}_0)d\boldsymbol{x}_0 \end{aligned}$$

it is intractable (for similar reasons as in VAEs)

Backward Diffusion Process - II

- We approximate $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ directly and create an iterative noise-reduction process $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \approx q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$
- The approximation holds if β_t is small in the forward diffusion step
- We learn the paramaters of $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, using neural networks (just as we did in VAEs), i.e.

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \boldsymbol{\sigma}^2 = f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))$$

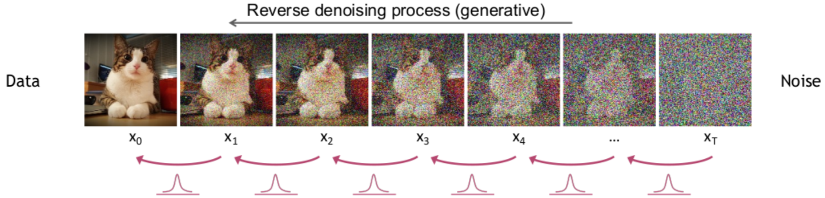
- $f_{\boldsymbol{\theta}}(\cdot)$ is commonly a U-net

Backward Diffusion Process - III

- The joint probability of the reverse process is

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \tag{3}$$

  where $p(\boldsymbol{x}_T) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
- Note that this is similar to the generative process in hierarchical VAEs



Outline

1 Diffusion Models
- Denoising Diffusion Models
- Forward Diffusion Process
- Backward Diffusion Process

2 The Objecttive Function
- Deriving the ELBO

Deriving the (last!) ELBO

- We start by defining an objective function that maximizes the likelihood of $\boldsymbol{x}_0$ assign by the model $p_{\boldsymbol{\theta}}$
- Specifically, we optimize

$$\log p(\boldsymbol{x}_0) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{4}$$

which circumvents the marginal distributions that are intractable.

Easier ELBO Formulation

- Note that the expectation of a function of a subset of variables is given by taking the expectation only over the subset!

$$\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}[f(\boldsymbol{x}_{a:b})] = \mathbb{E}_{q(\boldsymbol{x}_{a:b}|\boldsymbol{x}_0)}[f(\boldsymbol{x}_{a:b})]$$