Concept Notes: The Variational Auto-Encoder

Luis Oala

October 7, 2018

1 Setting: The Situation

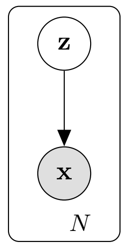


Figure 1: We can observe, i.e. actually see, N samples of the random variable $\mathbf{x}$. We assume that $\mathbf{x}$ depends on another variable $\mathbf{z}$, which is latent , i.e. we cannot observe it. Graph adapted from (Kingma and Welling, 2013).

Imagine a situation as depicted in Figure 1. Our situation includes the following ingredients, assumptions and constraints:

- A size N observed sample $\mathbf{X}$ of random variable $\mathbf{x}$. We assume that $\mathbf{x}$ follows the *likelihood* $p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z})$.
- A latent variable $\mathbf{z}$ which we assume to follow the *prior* $p_{\boldsymbol{\theta}^*}(\mathbf{z})$. It is called latent because we cannot see any of the $\mathbf{z}$ values.
- We assume that $p_{\boldsymbol{\theta}^*}(\mathbf{z})$ and $p_{\boldsymbol{\theta}^*}(\mathbf{x}|bz)$ belong to the parametric family of distributions $p_{\boldsymbol{\theta}}(\mathbf{z})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.
- We do not know anything about $\boldsymbol{\theta}^*$ or $\mathbf{z}$.

Basic probability rules reference

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \quad (1)$$

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z})d\mathbf{z} \quad (2)$$

$$\text{Bayes' Theorem: } \overbrace{p(\mathbf{z}|\mathbf{x})}^{\textit{posterior}} = \frac{\overbrace{p(\mathbf{x}|\mathbf{z})}^{\textit{likelihood}}\ \overbrace{p(\mathbf{z})}^{\textit{prior}}}{\underbrace{p(\mathbf{x})}_{\textit{evidence}}} \tag{3}$$

2 The goal

We want to learn what sensible values for our unknown, latent variable $\mathbf{z}$ would be given the information we have available. Bayes' formula gives us a straightforward way to do so. We are interested in the *posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$.

3 The approach

3.1 Getting started

For a given $\boldsymbol{\theta}$ we have all the ingredients to calculate the *posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ using Bayes' formula (Equation 3):

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \tag{4}$$

Remember that per our initial assumption we have access to $p_{\boldsymbol{\theta}}(\mathbf{z})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ for some $\boldsymbol{\theta}$. We just do not know the true $\boldsymbol{\theta}^*$. Via Equation 2 we have, in theory, access to the *evidence* $p_{\boldsymbol{\theta}}(\mathbf{x})$ as well.

So we can calculate, again in theory, the *posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ ...

3.2 Problem 1: Intractable $p_{\boldsymbol{\theta}}(\mathbf{x})$

... but the integration over all possible values of $\mathbf{z}$ in Equation 2 is computationally not feasible (TODO: why actually not?). That means we cannot calculate $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ simply using Bayes after all.

3.3 Idea for Solution to Problem 1: What about using $D_{\mathbf{KL}}$?

Okay, so we cannot simply calculate $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ with the information available at our disposal. What about this idea: We try to approximate $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ using a parametric distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ by doing

$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}} D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \tag{5}$$

The $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ with the problematic $p_{\boldsymbol{\theta}}(\mathbf{x})$ is still part Equation 5, but maybe with some wishful thinking and function magic the D_{KL} is somehow getting rid of it? Let us see:

$$D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})] - \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})] + \log p_{\boldsymbol{\theta}}(\mathbf{x}) \tag{6}$$

TODO: include the derivation for Equation 6 from page 12 of your notes. This also covers the ELBO derivation part

Unfortunately, the answer is no. $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ is still part of the expression. No magic, we are back to square one.

3.4 Actual Solution to Problem 1: Use your ELBO

But our efforts in the previous step, using the D_{KL}, were not in vain. Upon closer examination of Equation 6 the first two terms turn out to be part of a familiar expression: the evidence lower bound (ELBO). Viewed as a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ the ELBO is defined as:

$$\mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})] \tag{7}$$

Note that the ELBO is a function of $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, all things we have access to and can compute with. That sounds promising! We can reformulate Equation 6 by inserting the $\mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta})$ as such:

$$D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}) \tag{8}$$

In Equation 8 it also becomes apparant why the ELBO is called ELBO. When rearranging Equation 8 we get

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \underbrace{D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}_{\geq 0} + \mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}) \tag{9}$$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}) \tag{10}$$

Thus we can see that the ELBO indeed is a lower bound to the likelihood of the *evidence*. We can also rewrite Equation 7 as: (TODO: include detailed derivation and clean point about this single datapoint vs. all decomposition of the objective)

$$\mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{z})) \tag{11}$$

Let us take a step back and recap what we have seen so far. Our initial goal was to learn about plausible values for $\mathbf{z}$, a latent variable that we cannot observe. Bayes' Theorem gives us a straightforward way to reason about about $\mathbf{z}$ by using the information we have on $p_{\boldsymbol{\theta}}(\mathbf{z})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. Unfortunately we run into problems approximating the *posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ because the computation of the *evidence* $p_{\boldsymbol{\theta}}(\mathbf{x})$ is intractable. We thought about using a parametric model $q_{\boldsymbol{\phi}}((\mathbf{z}|\mathbf{x}))$ to minimize the D_{KL} to the *posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, but saw that the intractable $p_{\boldsymbol{\theta}}(\mathbf{x})$ is still in the expression. So far so good.

Our initial motivation for using the D_{KL} was to have an objective that we can minimize to learn about the *posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. If we now stare at Equation 8 for a while we can realize the following. At the optimal $\boldsymbol{\phi}^*$ we have that $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}^*}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = 0$. Additionally we know that $\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta})$ from Equation 10. Since only the ELBO is a function of $\boldsymbol{\phi}$ (see RHS of Equation 8) our objective of minimizing the D_{KL} (LHS of Equation 8) is equivalent to maximizing the ELBO w.r.t. $\boldsymbol{\phi}$. And this is what we will do for *variational inference*. Finally we have a tractable approach to approximate the *posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ and learn about plausible values for $\mathbf{z}$ given our available data $\mathbf{x}$: we maximize the ELBO!

3.5 While we are at it: Let us learn about $p_{\boldsymbol{\theta}}(\mathbf{x})$, too!

We learned that *evidence* $p_{\boldsymbol{\theta}}(\mathbf{x})$ is major culprit for all our problems. It is the reason why we have to take the tiresome detour via the ELBO to find a tractable optimization problem that satisfies our goal of learning about $\mathbf{z}$. Along the way we learned that ELBO is called ELBO because it is a lower bound to the likelihood of the *evidence*. We said previously that we will maximize the ELBO w.r.t. the variational parameters $\boldsymbol{\phi}$ because this is equivalent to minimizing the D_{KL}. But as we know from Equation 10 (and from its name) the ELBO lower bounds the likelihood of the *evidence* $p_{\boldsymbol{\theta}}(\mathbf{x})$. Thus if we want to learn and model the data $\mathbf{x}$ as well we can do so by maximizing the ELBO w.r.t. $\boldsymbol{\theta}$, too! And this is what happens in Variational Auto-Encoders: next to the *variational inference* on the parameters $\boldsymbol{\phi}$ underlying variable $\mathbf{z}$ we also perform *variational expectation maximization* on the parameters $\boldsymbol{\theta}$ underlying variable $\mathbf{x}$.

3.6 Keeping the concepts apart, clear and concise

Just to reiterate and manifest the terminology (to me at least this is very important to keep the concepts ordered in my head).

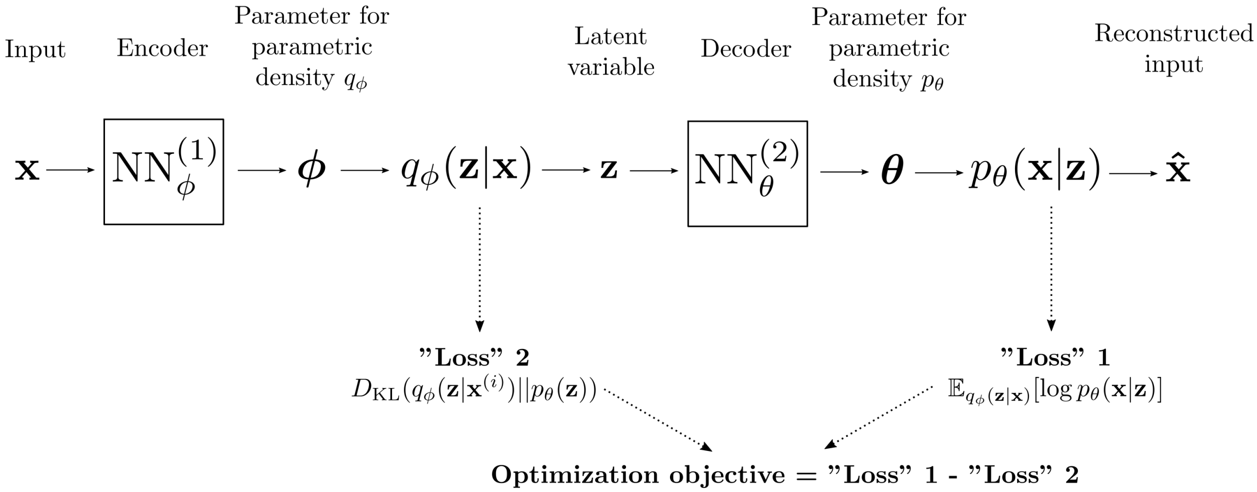
- *Variational inference*
  - Keep $\boldsymbol{\theta}$ fixed and do $\underset{\boldsymbol{\phi}}{\operatorname{argmax}}\ \mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta})$
  - Allows us to learn about the latent variable $\mathbf{z}$ via the approximated, parametric posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$
- *Variational EM (expectation maximization)*
  - Keep $\boldsymbol{\phi}$ fixed and do $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathbf{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta})$
  - Allows us to learn about the observable variable $\mathbf{x}$ via the available, parametric densities $p_{\boldsymbol{\theta}}(\mathbf{z})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.

3.7 Putting it all together

So we learned that in Variational Auto-Encoders we do not just perform *variational inference* on the parameters of the latent variable but also learn a generative model of the observable data at the same time. Let us put all these parts together into a graphical representation for better overview:

Finally, a few short notes on how we can, in practice, evaluate the individual terms in the ELBO objective (Equation 11):

- $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]$
  - Sampling based (sample several $\mathbf{z}$ for a particular $\mathbf{x}^{(i)}$ and take empirical agerage)
- $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{z}))$



- Sampling based (sample several $\mathbf{z}$ for a particular $\mathbf{x}^{(i)}$ and take empirical agerage)
- Analytically (e.g. this is possible if $q_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\theta}}$ are both Gaussian[1])

For practical recommendations regarding the sampling procedure please check the original paper by (Kingma and Welling, 2013). They have a number of tips.

[1] See the appendix for such an example