Source: https://www.youtube.com/watch?v=IXsA5Rpp25w

We start with Bayes formula, where we have variable X (our data/observations) and want to estimate Z given our data that would have similar properties. Z is latent variable:

$$p_\Theta(Z|X) = \frac{p_\Theta(X|Z)p_\Theta(Z)}{p_\Theta(X)}$$

where:

- $p_\Theta(Z|X)$ is posterior
- $p_\Theta(X|Z)$ is likelihood
- $p_\Theta(Z)$ is prior
- $p_\Theta(X)$ is normalizing constant, evidence

Problem lies in the denominator (which represents continuous random variable) which leads to intractable integral: $\int p_\Theta(x|z)p_\Theta(z)dz$

To get around this problem is to approximate posterior with some other distribution $q_\Theta(Z|X)$

This distribution should be more well-behaved: we can sample from it and calculate posterior $p_\Theta(Z|X) \approx q_\Theta(Z|X)$

But we have to learn parameters for q_Θ. To learn we need to compute dissimilarity with p_Θ

We are going to use KL divergence for that:

$$D_{KL}(q_\Theta||p_\Theta) = \mathbb{E}_{q_\Theta}[log\frac{q_\Theta(Z|X)}{p_\Theta(Z|X)}]$$

Our goal is to minimize KL divergence between p and q. But we can't compute it directly, because of the denominator, which true but intractable posterior. This is a thing we want to approximate!

To tackle this problem, let's first divide problem into pieces and see solve the pieces. First make divison into subtraction because we are using log function:

$$\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] - \mathbb{E}_{q_\Theta}[\log p_\Theta(Z|X)]$$

1. Rewrite second component using Bayes' rule:

$$\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] - \mathbb{E}_{q_\Theta}[\log \frac{p_\Theta(Z,X)}{p_\Theta(X)}]$$

   A little refresher of the Bayes' rule: $p_\Theta(Z|X) = \frac{p_\Theta(X|Z)p_\Theta(Z)}{p_\Theta(X)} = \frac{p_\Theta(Z,X)}{p_\Theta(X)}$

1. Split second component further: $\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] - \mathbb{E}_{q_\Theta}[\log p_\Theta(Z,X)] + \mathbb{E}_{q_\Theta}[p_\Theta(X)]$

1. focus on third term, this is expectation with respect to Z and could expand it to integral form like this:

$$\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] - \mathbb{E}_{q_\Theta}[\log p_\Theta(Z,X)] + \int q_\Theta(Z|X) \log p_\Theta(X) dx$$

1. move the part that does not have Z outside of integral:

$$\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] - \mathbb{E}_{q_\Theta}[\log p_\Theta(Z,X)] + \log p_\Theta(X) \int q_\Theta(Z|X) dx$$

1. final simplification, this is possible because integrating any density function would equate to 1 (area under pdf is 1):

$$D_{KL}(q_\Theta||p_\Theta) = \mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] - \mathbb{E}_{q_\Theta}[\log p_\Theta(Z,X)] + \log p_\Theta(X)$$

Now we have revealed something:

- third component is marginal log likelihood/log evidence (total probability taking into account all hiden variables and parameters): $\log p_\Theta(X)$ We cannot computet it directly because we don't have its analytical form.

But this equation shows that there is a relationship between log-likelihood ka KL divergence. It is interesting because most of the time we are doing log-likelihood estimation:

$$\log p_\Theta(X) = -\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] + \mathbb{E}_{q_\Theta}[\log p_\Theta(Z,X)] + D_{KL}(q_\Theta||p_\Theta)$$

We cannot compute KL divergence. But there is an interesting property that KL divergence possess: it can't be negative. Let's rewrite previous equation and get rid of KL divergence. At minimum we have to get rid of equality:

$$\log p_\Theta(X) \geq -\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] + \mathbb{E}_{q_\Theta}[\log p_\Theta(Z,X)]$$

Right side of the equation we call **ELBO** (evidence lower bound), lower bound on the evidence. Now how could we just get rid of KL divergence? Because by maximizing KL divergence we are indirectly minimizing KL divergence

Let's do final simplification of ELBO

$$ELBO = -\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] + \mathbb{E}_{q_\Theta}[\log p_\Theta(Z,X)]$$

$$ELBO = -\mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] + \mathbb{E}_{q_\Theta}[\log p_\Theta(X|Z)] + \mathbb{E}_{q_\Theta}[\log p_\Theta(Z)]$$

Second term is joint probability of X and latent variable Z can be expanded into two factors. Because we are in log scale we are doing summation instead of multiplication. I have applied prduct rule of probability on joint distribution.

Now move terms around:

$$ELBO = \mathbb{E}_{q_\Theta}[\log p_\Theta(X|Z)] - \mathbb{E}_{q_\Theta}[\log q_\Theta(Z|X)] + \mathbb{E}_{q_\Theta}[\log p_\Theta(Z)]$$

And we'll get a final formula:

$$ELBO = \mathbb{E}_{q_\Theta}[\log p_\Theta(X|Z)] - \mathbb{E}_{q_\Theta}[\log \frac{q_\Theta(Z|X)}{p_\Theta(Z)}]$$

where:

- $\mathbb{E}_{q_\Theta}[\log p_\Theta(X|Z)]$ is expected reconstruction error (we are reconstructing X given Z)
- $\mathbb{E}_{q_\Theta}[\log \frac{q_\Theta(Z|X)}{p_\Theta(Z)}]$ is KL divergence between approximate posterior and prior (don't confuse it with KL divergence we started with, that was KL divergence between approximate posterior and true posterior. We know the prior and we will learn the approximate posterior)

Note we that we chose Q so that it was conditional distribution (it was for Z but conditioned on X). But we couold have picked Q or Z so that it doesn't have to condition. All the derviations still work:

$$ELBO = \mathbb{E}_{q_\Theta}[\log p_\Theta(X|Z)] - \mathbb{E}_{q_\Theta}[\log \frac{q_\Theta(Z)}{p_\Theta(Z)}]$$

Now we ought to find these:

- $q_\Theta(Z|X)$
- $p_\Theta(X|Z)$

There are two algorithms for finding values:

- expectation maximization
- variational autoencoder

Convert to pdf: jupyter nbconvert --to webpdf --allow-chromium-download .\elbo.ipynb

more info: https://alpopkes.com/posts/machine_learning/kl_divergence/

In [ ]: