- Auto-Encoding Variational Bayes
- Diederik P. Kingma, Max Welling

I also read some other very helpful tutorials on this.

Assumptions

Suppose that we have a dataset $X = \{x^{(i)}\}_{i=1}^N$ draw from some distribution $p_\theta(x)$, identically and independently. We assume that the data is produced as follows:

- A value z (latent variable) is first draw from $p_\theta(z)$.
- Given z, x is drawn from $p_\theta(x|z)$

Putting it in a formal way, we have two random variables X and Z, and their joint distribution is defined by

$$p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$$

Where θ is a global parameters. We have a closed form for $p_\theta(x|z)$ and $p_\theta(z)$. **That is, $p_\theta(x|z), p_\theta(z)$ can be written as a function of θ, x, z.** This is all we assume.

The Problem to Solve

This has confused me a lot when first trying to understand VAEs. But now I understand there are actually two problems that need to be addressed:

- **Learning**. Given a dataset $\{x^{(i)}\}_{i=1}^N$, what is the best estimate of θ?
- **Inference**. For a fixed θ, given a single x, which is the distribution $p_\theta(z|x)$ (i.e., can you find a closed form for this?)

Naturally, for learning, we will try things like maximum likelihood. This requires an explicit form of $p_\theta(x)$. Since we only have a closed form for $p_\theta(x|z)$ and $p_\theta(z)$, we will write it as

$$p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz$$

Unfortuately, this is intractable due to integral on z.

For inference, it is also natural to apply the Bayes rule:

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{\int p_\theta(x|z)p_\theta(z)dz}$$

And this is intractable for the same reason.

The way out. Here we will first consider how to address the inference problem. Variational inference reformulates this as an optimization of an approximate posterior $q_\phi(z|x)$, where ϕ is called the variational parameter.

Intuitively, if we can optimize ϕ such that $q_\phi(z|x)$ and $p_\theta(z|x)$ are similary, then we can use $q_\phi(z|x)$ for inference. This means to minimize

$$D_{KL}[q_\phi(z|x) \| p_\theta(z|x)]$$

Given x, θ are fixed. If we expand this, and do some rearrangement, we will find

$$\log p_\theta(x) - D_{KL}[q_\phi(z|x) \| p_\theta(z|x)] = E_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x) \| p_\theta(z)]$$

We will first assume that the RHS is tractable and differentiable w.r.t θ and ϕ (and this is true in some sense). Let's see what will happen if this is true:

- For inference (when θ and x are fixed), we can maximize the RHS to minimize the KL divergence of the two posteriors
- For learning, we can maximize the RHS to maximize $\log p_\theta(x)$. Note the RHS is a lower bound of this quanity. So if $q_\phi(z|x)$ is a reasonble approximation, this is good.

Clearly, if we jointly update the RHS w.r.t. to both θ and ϕ, we will

- Find a good approximate posterior $q_\phi(z|x)$
- Find a good estimate of θ that maximize the log likelihood of the data.

Reparametrization and Batch Training

I will update this next week. But the basic idea is to "move" the stochasticity out of the expectation.

VAE

All that is left is to specify $p_\theta(z)$ and $p_\theta(x|z)$ and $q_\phi(z|x)$. In VAE,

- $p_\theta(z)$ is assumed a centered isotropic multivaraite Gaussian $p_\theta(z) = \mathcal{N}(z; 0, I)$ (independent of θ. This is reasonable. This is addressed in detail in a VAE tutorial.
- $p_\theta(x|z)$ is model as either a Bernoulli or Gaussian (some other reasonable distribution will be OK?).

  $$p_\theta(x|z) = \mathcal{N}(x; \mu(z, \theta), \Sigma(z, \theta))$$

  where μ and Σ are modeled using a neural network.
- $q_\phi(z|x)$ takes a similar form as $p_\theta(x|z)$.

Implementation Details

I will address this part when I have implemented one myself.