Variational Inference

For this one, we essentially want to optimize the posterior $p(x|y)$ to get an **approximate posterior** $q(x)$ (parameterised by parameters θ) which is as close to the posterior as possible, and then use Bayes rule to get the evidence.

Setting the parameters θ s.t. $q(\theta)$ is close to $p(x|y)$ directly is hard. Instead, we want to formulate the optimisation as a bound and then minimise the bound so we find an approximation.

To start, we want to find the bound over the evidence $p(Y)$.

Using the evidence term $p(Y)$, we can do some funky maths (along with the Jensen inequality) to end up with the following expression (which will contain our bound):

$$-KL(q(X)||p(X|Y)) + \log p(Y)$$

> **Jensen Inequality**: this basically says that a line which intersects two points of a convex function will be above the function

The first term is the **Kulback-Leiber divergence**, which is a measurement between distributions. In this case, it will measure the 'distance' between our approximation $q(x)$ and the true posterior $p(x|y)$.

It has the interesting property that it is always positive, except when $p(x|y) = q(x)$ in which case it will be 0.

This means that, using the bound above, if the divergence is zero $q(x)$ is the exact posterior. We should probably minimise this measure.

Unfortunately, it is not yet possible to minimise the KL measure as we don't know the true posterior (which is intractable).

Following some other maths we get to:

$$\log p(Y) \geq \mathbb{E}_{q(X)}[\log p(X, Y)] - H(q(X)) = \mathfrak{L}(q(X))$$

This term is referred to as the **Evidence Lower BOund - ELBO**. This is what we want to maximise (i.e. optimise), as by doing so we will automatically minimise the KL measure (as it is bounded by $p(Y)$).

> note: this will not always give us a convex function

We're in a good position now, because

- the log of $\mathbb{E}$ is easier to calculate than $\mathbb{E}$

- we can choose the distribution we want to take the $\mathbb{E}$ over (we have lots of freedom!)
- the term $H(q(X))$ is trivial to calculate

We're now left to pick q.

We could use **Mean-Field Variational Inference**. This chooses a family of distributions which completely factorise over the unknown variables - i.e. we are matching the marginals of the posterior distribution.

Method: (very likely you won't need to know the exact form)

1. Formulate the joint distribution over data and the latent parameters
2. Formulate a fully factorised approximate posterior over the latent variables
3. Fit the marginal approximation by making the bound *tight*
4. Iterate through all variables

- Overall this is very efficient
- But you won't ever get the right answer :-(
- However we can compare how we're doing - if you take two different approximations and get two different bounds you can compare them

The maths is a bit confusing so I would recommend watching the lecture if possible.