COMS30007 - Machine Learning

Variational Inference

Carl Henrik Ek

Week 10

Abstract

Now we will move on to a deterministic approximation of the Ising Model. What we will do is to specify a surrogate model and try to fit this model so that it is as close as possible to the actual model. We will first go through the idea of Variational Bayes and then proceed to go through and look at the specific approach we will use for the Ising model.

Math

This derivation is long, I go through the full proof of, variational Bayes, mean-field and mean-field variational Bayes for the Ising model. I tried to be as complete as possible so everything should be self-contained. However, there is likely to be blunders in there so if you find something strange point it out to me. Importantly, you do not need to know any of this, its here if you are really interested in machine learning but none of the questions on the exam will be related to this. If you want to skip the derivations read 1 and then move ahead to 4. However, if I am allowed to say it myself this is quite beautiful so if you enjoy math I suggest you continue reading.

1 Variational Bayes

Inference is the task of fitting our model to some observed data, what we often do is to try to choose the model that maximises the evidence. If we have been given some data $\mathbf{y}$ and have some parameters θ to fit we wish to pick them such that,

$$\hat{\theta} = \text{argmax}_\theta p(\mathbf{y}).$$

As we know the evidence is often intractable to compute but lets see what we can do,

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X} = \log \int p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} \tag{1}$$

$$= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}. \tag{2}$$

The strange thing here is the second row where we have added a distribution $q(\mathbf{X})$ to the equation. This is just a general distribution and because we add it in this form it will not change the integral at all. What we will do now is try to formulate a bound on this integral, in specific we will use something called the Jensen inequality. The Jensen inequality states that a line between two points on the curve will always be above the curve see Figure 1.

$$\lambda f(x_0) + (1-\lambda) f(x_1) \geq f(\lambda x_0 + (1-\lambda) x_1)$$
$$x \in [x_{min}, x_{max}]$$
$$\lambda \in [0, 1]]$$

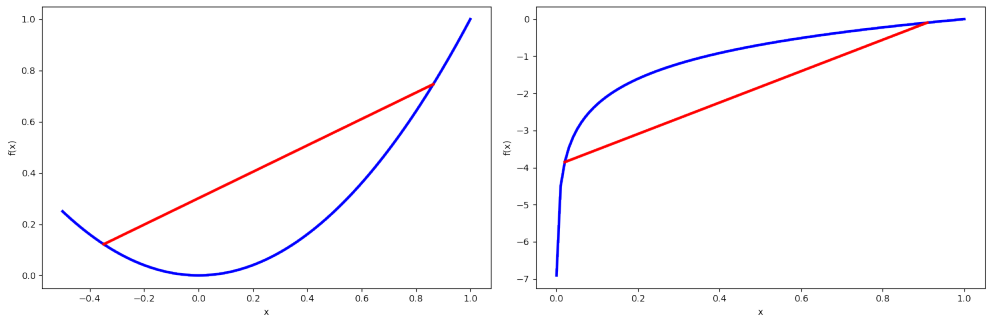

Figure 1: *The plot on the left shows a convex function in blue and line connecting two of the points on the function. The Jensen inequality implies that if you "move" a point along the red line it will always be above the blue. On the right the blue function is a logarithm and we can see that the reverse behaviour is true, the red line will always be below the blue. This means that we can say that, "the red line will always be a lower bound on the blue".*

Even though it might seem trivial it is a very useful property for dealing with probabilities. When we are marginalising variables from our model we are computing expectations, if we are computing an expectation of a convex function, the expectation of the function will always be an upper bound on the function applied to the expectations,

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x]) \tag{3}$$

$$\int f(x)p(x)\mathrm{d}x \geq f\left(\int xp(x)\mathrm{d}x\right) \tag{4}$$

Where this is specifically important is when the function is a logarithm. As a logarithm is a concave function the inequality just flips around as can be seen in Figure **??**. This means that the logarithm of an integral is a upper bound on the log of the integral,

$$\int \log(x)p(x)\mathrm{d}x \leq \log\left(\int xp(x)\mathrm{d}x\right). \tag{5}$$

We will now exploit this result to try and find a bound on the intractable evidence,

$$\log\, p(\mathbf{Y}) = \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})\mathrm{d}\mathbf{X} = \tag{6}$$

$$\geq \int q(\mathbf{X})\log\frac{p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})}{q(\mathbf{X})}\mathrm{d}\mathbf{X} \tag{7}$$

$$= \int q(\mathbf{X})\log\frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})}\mathrm{d}\mathbf{X} + \int q(\mathbf{X})\mathrm{d}\mathbf{X}\log\, p(\mathbf{Y}) \tag{8}$$

$$= -\mathrm{KL}\left(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})\right) + \log\, p(\mathbf{Y}). \tag{9}$$

The important part of the computation above is where we move the logarithm inside the integral by exploiting the bound. Importantly the first term after having split the integral into two is what is known as the Kullback-Leibler divergence Ch 1.6.1 [1]. This is a measure of "simiarity" between probability distributions. It is not a metric, its for example not symmetric, but importantly it is only 0 if the two distributions are the same and positive in all other cases. This leads us to an important observation, if $q(\mathbf{X}) = p(\mathbf{X}|\mathbf{Y})$ then the bound is tight. Importantly in order to reach the posterior distribution $p(\mathbf{X}|\mathbf{Y})$ we would have to compute the evidence. So this leads us to the central intuition of Variational Bayes, if we can pick a distribution $q(\mathbf{X})$ such that it is as close as possible to the posterior $p(\mathbf{X}|\mathbf{Y})$ we will have a good surrogate model, if it is exact,

it is the same model. Therefore lets try to minimise the KL-divergence between the two distributions.

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) = \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X} \tag{10}$$

$$= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X},\mathbf{Y})} \mathrm{d}\mathbf{X} + log\ p(\mathbf{Y}) \tag{11}$$

$$= H(q(\mathbf{X})) - \mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X},\mathbf{Y})\right] + log\ p(\mathbf{Y}). \tag{12}$$

What we have done above is to write up the divergence as an expectation over the joint distribution and a term that only dependes on $q(\mathbf{X})$ and the evidence. If we now move the evidence over on the other side of the expression we will get this formulation,

$$log\ p(\mathbf{Y}) = \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))}_{\text{ELBO}} \tag{13}$$

$$\geq \mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X})) = \mathcal{L}(q(\mathbf{X})). \tag{14}$$

As we know the KL-divergence has to be positive the remaining term is a lower-bound on the evidence. This is why this is referred to as the *ELBO-Evidence Lower BOund*.

So that was a whole lot of math but what does it all mean, what does this lead us to, has this actually solved anything? Well if we look at the formula above, what we want to find is $q(\mathbf{X})$ if we do find this, we know that it is an approximation of the true posterior $p(\mathbf{X}|\mathbf{Y})$ this is really useful, further the bound is specified by the computation of an expectation over the *joint* distribution of the data. 1) if we cannot formulate the joint distribution we are in bigger trouble and 2) we are allowed to choose the distribution $q(\mathbf{X})$ that we have to take the expectation over. Clearly this should be simpler to do.

In the next part we will derive a specific family of approximations often referred to as mean-field approximations[1]. They are often not particularly exact but they do work in most cases.

2 Mean Field Approximation

The mean-field approximation assumes that the approximative posterior factorises over all the variables as,

$$q(\mathbf{X}) = \prod_i q_i(\mathbf{x}_i).$$

We will proceed by deriving the bound related to this type of approximative distribution.

$$\mathcal{L}(q) = \int q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{X})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} = \int \prod_i q_i(\mathbf{x}_i) \log \frac{p(\mathbf{Y},\mathbf{X})}{\prod_k q_k(\mathbf{x}_k)} \mathrm{d}\mathbf{X} \tag{15}$$

$$= \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y},\mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right) \tag{16}$$

For many types of models we want to update several distributions. What we will do now is to derive a scheme where we update one component in turn. Therefore we would like to re-write $\mathcal{L}(q)$ in such a manner that we can single out a single component,

$$\mathcal{L}(q) = \mathcal{L}(q_j) + \mathcal{L}(q_{\neg j}),$$

where $\neg j$ means all the components except for j.

[1] Again these are models that initially was suggested in physics to model phase transitions first described by Pierre Curie. If you read the literature on Variational Bayes you will often hear the ELBO referred to as the variational free energy, and the first terms the bound as *energy* as the second term corresponds to the *entropy* of the approximating distribution. So don't forget your theromdynamics.

$$\mathcal{L}(q) = \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y}, \mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right) \tag{17}$$

$$= \int_j \int_{\neg j} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left(\log\, p(\mathbf{X}, \mathbf{Y}) - \sum_k \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j \tag{18}$$

$$= \int_j q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \log\, p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}}_{\log\, f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \left(\log\, q_j(\mathbf{x}_j) + \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j \tag{19}$$

$$= \int_j q_j(\mathbf{x}_j) \log\, f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \left(\log\, q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \mathrm{d}\mathbf{x}_{\neg j}}_{=1} + \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \mathrm{d}\mathbf{x}_{\neg j}}_{\text{constant w.r.t. } q_j} \right) \tag{20}$$

$$= \int_j q_j(\mathbf{x}_j) \log\, f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j - \int_j q_j(\mathbf{x}_j) \log\, q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j}_{=1} \tag{21}$$

$$= \int_j q_j(\mathbf{x}_j) \log\, \frac{f_j(\mathbf{x}_j)}{q_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.} \tag{22}$$

$$= - \int_j q_j(\mathbf{x}_j) \log\, \frac{q_j(\mathbf{x}_j)}{f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.} \tag{23}$$

$$= -\mathrm{KL}(q_j(\mathbf{x}_j) || f_j(\mathbf{x}_j)) + \text{const.} \tag{24}$$

The above derivation ends up somewhere rather intuitive, to maximise the lower bound on the evidence with respect to $q_j(\mathbf{x}_j)$ we want to minimise the KL-divergence between the factor $q_j(\mathbf{x}_j$ and the distribution when all *other* factors have been averaged out. Have a look at what the term $f_j(\mathbf{x}_j)$ to see if this makes sense.

As we know that the KL-divergence is always positive and as we are free to choose $q_j(\mathbf{x}_j)$ as we wish we can simply set,

$$q_j(\mathbf{x}_j) = f_j(\mathbf{x}_j) \tag{25}$$

$$\log\, f_j(\mathbf{x}_j) = \int_{\neg j} \underbrace{\prod_{i \neq j} q_i(\mathbf{x}_i)}_{q_{\neg j}(\mathbf{x}_{\neg j})} \log\, p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j} \tag{26}$$

$$= \mathbb{E}_{q_{\neg j}(\mathbf{x}_{\neg j})} \left[\log\, p(\mathbf{Y}, \mathbf{X}) \right] \tag{27}$$

So in order to use the mean-field variational bayes we need to pick the approximate distribution $q(\mathbf{X})$ in such a way that we can compute the expectation above. Now we have derived both variational bayes and the mean field approximation we are ready to move back to our model and work specifically with the Ising model we have defined.

3 Mean Field Variational Bayes in Ising Model

Now let us formulate the mean field approximation for the Ising model, lets first remind ourselves of the model. We specified a prior of the form,

$$p(\mathbf{x}) = \frac{1}{Z_0} e^{E_0(\mathbf{X})} \tag{28}$$

$$E_0(\mathbf{X}) = \sum_i^N \sum_{j \in (i)} w_{ij} x_i x_j \tag{29}$$

If we look at the term $w_{ij}x_ix_j$ we can see that it will be postive if the latent values are the same and negative otherwise. The larger the value the higher the probability which fits well with our Ising model. The other term we need is the likelhood,

$$p(\mathbf{Y}|\mathbf{X}) = \prod_i p(y_i|x_i) = \frac{1}{Z_1} \prod_i e^{L_i(x_i)}, \tag{30}$$

where the function L_i should give a large value if it is likely that x_i have generated y_i. Now we are ready to formulate our approximate distribution. We will use a full mean-field approximation so we will assume that the approximate distribution over each latent variable is independent,

$$q(\mathbf{x}) = \prod_i q(x_i, \mu_i),$$

where we have introduced μ_i as a variational parameter that parametrises this distribution. In specific μ_i will be $\mathbb{E}_{q_i}[x_i]$. The first thing we need to get is the joint distribution of the model. We will through out the task work in log-space which gives us,

$$\log p(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) \tag{31}$$

$$= \log \left(\prod_i e^{L_i(x_i)} \frac{1}{Z_0} e^{\sum_{j \in \mathcal{N}(i)} w_{ij} x_i x_j} \right) \tag{32}$$

$$= \sum_i \left(L_i(x_i) + \sum_{j \in \mathcal{N}(i)} w_{ij} x_i x_j \right) + \text{const.} \tag{33}$$

As we have choosen a fully factorised approximative distribution $q(\mathbf{x})$ we will compute each expectation in the bound in turn so we want to write up the joint distribution where we only consider one variable. This means,

$$\log p(\mathbf{x}, \mathbf{y}) = L_i(x_i) + x_i \sum_{j \in \mathcal{N}(i)} w_{ij} x_j + \text{const.},$$

where we have included all the term over the remaining latent variables in the constant term. We are now ready to compute the expectation to get the approximative posterior,

$$\log q_i(x_i) = \log f_i(x_i) = \int \prod_{j \neq i} q_j(x_j) \log p(\mathbf{x}, \mathbf{y}) \mathrm{d}x_{\neg i} \tag{34}$$

$$= \int \prod_{j \neq i} q_j(x_j) (L_i(x_i) + \sum_{k \in \mathcal{N}(i)} w_{ik} x_i x_k + \text{const.}) \mathrm{d}x_{\neg i} \tag{35}$$

$$= \underbrace{\int \prod_{j \neq i} q_j(x_j) \mathrm{d}x_{\neg i}}_{=1} L_i(x_i) + \int \prod_{j \neq i} q_j(x_j) \sum_{k \in \mathcal{N}(i)} w_{ik} x_i x_k \mathrm{d}x_{\neg i} + \text{const.} \tag{36}$$

The first integral will compute to one as $q_j(x_j)$ is a distribution. The second term is a bit trickier to deal with so we are going to deal with it on its own.

$$\int \prod_{j\neq i} q_j(x_j) \sum_{k\in\mathcal{N}(i)} w_{ik} x_i x_k \mathrm{d}x_{\neg i} = \int \prod_{j\neq i} q_j(x_j) x_i \sum_{k\in\mathcal{N}(i)} w_{ik} x_k \mathrm{d}x_{\neg i} \tag{37}$$

$$= \int x_i \sum_{k\in\mathcal{N}(i)} w_{ik} \left(\prod_{j\neq i} q_j(x_j)\right) x_k \mathrm{d}x_{\neg i} = x_i \sum_{k\in\mathcal{N}(i)} w_{ik} \int \left(\prod_{j\neq i} q_j(x_j)\right) x_k \mathrm{d}x_{\neg i}. \tag{38}$$

We will now expand the integration over each term and find something rather beautiful,

$$x_i \sum_{k\in\mathcal{N}(i)} w_{ik} \int \left(\prod_{j\neq i} q_j(x_j)\right) x_k \mathrm{d}x_{\neg i} = \tag{39}$$

$$x_i \sum_{k\in\mathcal{N}(i)} w_{ik} \int \left(q_1(x_1)q_2(x_2)\cdot\ldots\cdot q_N(x_N)\right) x_k \mathrm{d}x_1 \mathrm{d}x_2 \cdot\ldots\cdot \mathrm{d}x_N \tag{40}$$

$$= x_i \sum_{k\in\mathcal{N}(i)} w_{ik} \underbrace{\int q_1(x_1)\mathrm{d}x_1}_{=1} \underbrace{\int q_2(x_2)\mathrm{d}x_2}_{=1} \cdot\ldots\cdot \int q_k(x_k)x_k\mathrm{d}x_k \cdot\ldots\cdot \underbrace{\int q_N(x_N)\mathrm{d}x_N}_{=1} \tag{41}$$

$$= x_i \sum_{k\in\mathcal{N}(i)} w_{ik} \int q_k(x_k)x_k\mathrm{d}x_k = x_i \sum_{k\in\mathcal{N}(i)} w_{ik} \mathbb{E}_{q_k(x_k)}[x_k] = \tag{42}$$

$$= x_i \sum_{k\in\mathcal{N}(i)} w_{ik}\mu_k \tag{43}$$

Now we are ready to tidy things up and write out the approximative posterior for x_i by combining our terms,

$$\log\, q_i(x_i) = \log\, f_i(x_i) = L_i(x_i) + x_i \underbrace{\sum_{k\in\mathcal{N}(i)} w_{ik}\mu_k}_{m_i} + \text{const.} \tag{44}$$

$$= L_i(x_i) + x_i \cdot m_i + \text{const.} \tag{45}$$

The expression above does make sense, we have one term which relates to the observations at x_i and one term which relates to the prior term relating the expectations of the nearby latent locations. This means that we can write our approximative distribution as,

$$q(\mathbf{x}) \propto \prod_i q_i(x_i) \propto e^{x_i m_i + L_i(x_i).}$$

We need to make sure that the approximation that we have is an actual distribution therefore making sure that it integrates to 1. In this case this is really simply as $x_i \in [1, -]$ and therefore we can write it up as,

$$\hat{q}_i(x_i = 1) = \frac{1}{q(x_i = 1) + q(x_i = -1)} q(x_i = 1) = \frac{e^{m_i + L_i(1)}}{e^{m_i + L_i(1)} + e^{-m_i + L_i(-1)}} \tag{46}$$

$$= \left\{ \begin{array}{c} \text{Simplification:} \\ \frac{e^a}{e^a + e^b} = \frac{1}{e^{-a}(e^a + e^b)} = \frac{1}{1 + e^{b-a}} \end{array} \right\} \tag{47}$$

$$= \frac{1}{1 + e^{-2m_i - L_i(1) + L_i(-1)}} = \frac{1}{1 + e^{-2\underbrace{(m_i + \frac{1}{2}L_i(1) - \frac{1}{2}L_i(-1))}_{a_i}}} \tag{48}$$

$$= \frac{1}{1 + e^{-2a_i}} = \text{sigm}(2a_i). \tag{49}$$

As the probability for x_i taking value 1 is equal to a sigmoid the probability for the other case is trivial,

$$q_i(x_i = -1) = \text{sigm}(-2a_i)$$

Importantly the proposal distribution is completely defined by its expected value μ_i as a last step we now want to find a way to update this parameter. This is easy to do by going through its definition,

$$\mu_i = \mathbb{E}_{q_i(x_i)}[x_i] = \sum_{x_i \in [1,-1]} x_i q_i(x_i) = (+1)q_i(x_i = 1) + (-1)q_i(x_i = 1) \tag{50}$$

$$= \frac{1}{1 + e^{-2a_i}} - \frac{1}{1 + e^{2a_i}} = \frac{e^{a_i}}{e^{a_i} + e^{-a_i}} - \frac{e^{-a_i}}{e^{-a_i} + e^{a_i}} \tag{51}$$

$$= \frac{e^{a_i} - e^{-a_i}}{e^{a_i} + e^{-a_i}} = \tanh(a_i) = \tanh\left(m_i + \frac{1}{2}(L_i(1) - L_i(-1))\right). \tag{52}$$

So now we are there, we have the approximative distribution for the mean field approximation of an Ising model and we have equation that tells us how to update the parameters of this distribution. Now before we implement this, let us see if it makes sense.

$$q_i(x_i = 1|\mu_i) = \text{sigm}(2a_i) = \text{sigm}\left(2(m_i + \frac{1}{2}(L_i(1) - L_i(-1)))\right) \tag{53}$$

$$\mu_i = \tanh(a_i) = \tanh\left(m_i + \frac{1}{2}(L_i(1) - L_i(-1))\right) \tag{54}$$

$$m_i = \sum_{j \in \mathcal{N}(i)} w_{ij}\mu_j \tag{55}$$

We are in effect trying to find the probability of a latent variable that can take two different values, this means that a sigmoid function makes perfect sense as an approximative distribution. The update of the variational parameter μ_i is updated as a $\tanh(2ai)$ function this also makes sense as we have $x_i \in [-1, 1]$ we now have an updated which is bounded between those two values. Below we have plotted the two functions,

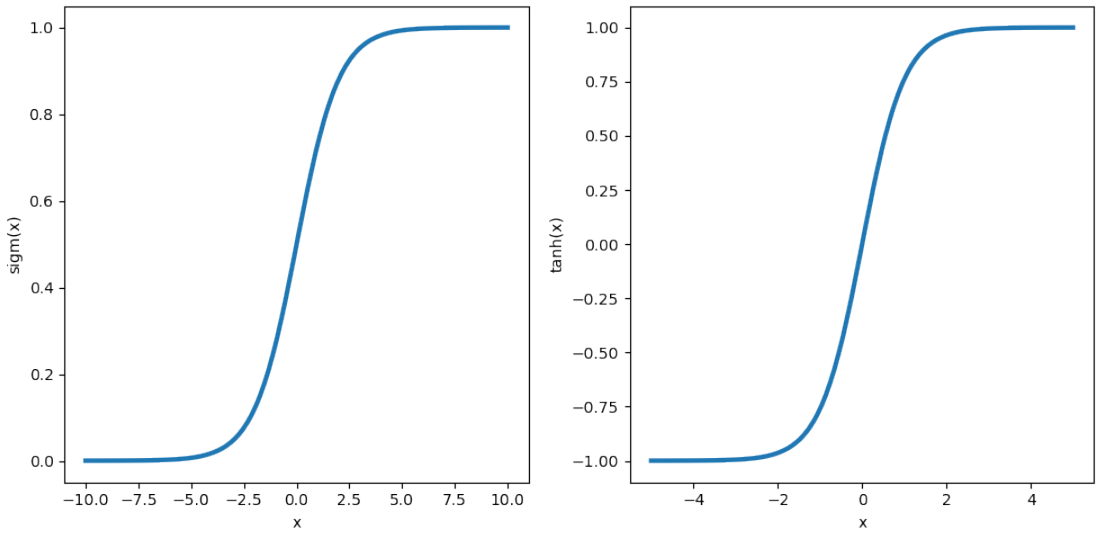

Figure 2: *The above plot shows on the left a sigmoid function and on the right a hyperbolicus tangent function. They look very similar but observe that the sigmoid has its asymptots at 0 and 1 while the tanh is at -1 and 1.*

Algorithm 1 Variational Bayes for Ising Model

procedure MEAN FIELD VARIATIONAL BAYES
  $\boldsymbol{\mu} \leftarrow$ initialise variational distributions
  $\boldsymbol{x} \leftarrow$ initialise latent variables
  for $\tau = 1 \dots T$ **do**
    for $i = 1 \dots N$ **do**
      $m_i^{\tau+1} = \sum_{j \in \mathcal{N}(i)} w_{ij} \mu_j^{\tau} \leftarrow$ compute parameter
      $\mu_i^{\tau+1} = \tanh\left(m_i + \frac{1}{2}(L_i(1) - L_i(-1))\right) \leftarrow$ update variational parameter
  return $q(\mathbf{x})$

Except for the asymptotic vaules, does the posterior and the variational update make sense, both are functions of a_i as,

$$a_i = m_i + \frac{1}{2}(L_i(1) - L_i(-1)) = \sum_{j \in \mathcal{N}(i)} w_{ij}\mu_j + \frac{1}{2}(L_i(1) - L_i(-1)).$$

The first term in the above expression relates to the nodes that are neighbours of i. In effect it is a weighting of the expected values of the posteriors of these nodes where the weights w_{ij} are what encodes our prior assumption in how neighbours interact. As the weights w_{ij} are all positive if all neighbours are in agreement, i.e. have the same sign, then the first term will "push" towards either -1 or $+1$. For simplicity lets assume that all neighours have $\mu_j = 1$ then m_i will be a positive value and vice versa. If it is close to zero that means that the neighbours are all in disagreement **or** that we are very uncertain of their values, i.e. μ_j is close to zero. The second term is a difference between that comes from the likelihood terms, if it is positive this means that it is much more likely that $x_i = 1$ describes the data y_i compared to $x_i = -1$. So for the extremes, if all neighbours are $\mu_j = 1$ and $L_i(1) - L_i(-1) > 0$ then a_i will be a large positive value, i.e. the posterior $q_i(x_i)$ will be large and μ_i will get a value close to one. So these equations do make sense.

4 Implementation

Now we are ready to finally implement our variational Bayes inference scheme. What we will do is to start off with some initial value for our variational parameter and then we will update each of the distributions in turn. You can try and start with different values and see how it changes the way we reach a solution. The algorithm we should implement is outlined in Algorithm 1.

5 Summary

This lab implemented a deterministic approximation to approximate inference. Importantly, with this approach we will never reach the true solution, but importantly we will get to an approximate solution really quickly. Also because we now have a parametrised form of our posterior we can do lots of interesting things with it as it is just like any other distribution we can just treat it as it is the true posterior. Variational methods are really useful in practice but they do take a bit more effort to formulate compared to sampling and I believe that this task gave you a good example of this.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

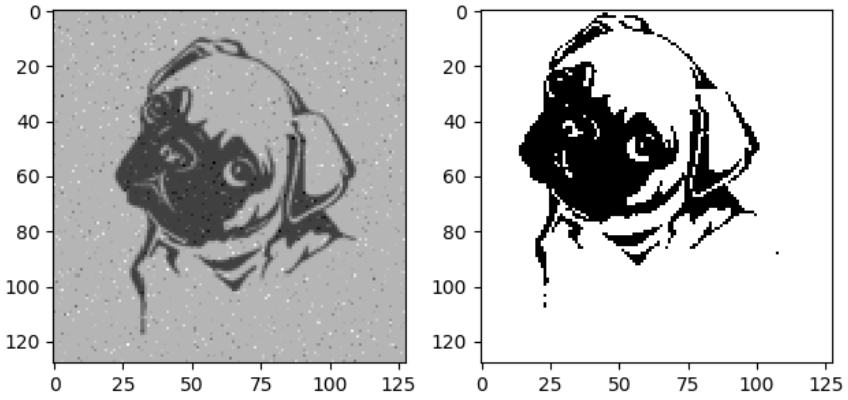

Figure 3: *This figure shows the result of my implementation of variational inference for the image denoising example. As you can see the ising prior cleans up the image rather nicely and we are left with a lovely black and white pug*