1 The Cavity distribution

The cavity distribution is defined as

$$c(\mathbf{f}_i) \propto \int p(\mathbf{f}) \prod_{j \neq i} t(\mathbf{f}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)$$

We'll treat these approximate likelihoods as *pseudo data*. We'll envisage a series of independent normal variables so that

$$p(\tilde{\mathbf{y}}_i|\mathbf{f}_i) = \mathcal{N}(\tilde{\mathbf{y}}_i|\mathbf{f}_i, \beta_i^{-1})$$

The cavity distribution can then be seen as the following conditional density:

$$c(\mathbf{f}_i) = \frac{\int p(\mathbf{f})p(\tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_{\setminus i})\,\mathrm{d}\mathbf{f}_{\setminus i}}{\int p(\mathbf{f})p(\tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_{\setminus i})\,\mathrm{d}\mathbf{f}} = \frac{p(\mathbf{f}_i, \tilde{\mathbf{y}}_{\setminus i})}{p(\tilde{\mathbf{y}}_{\setminus i})} = p(\mathbf{f}_i|\tilde{\mathbf{y}}_{\setminus i})$$

To optimise our approximation, we'll optimise the parameters $\tilde{\mathbf{y}}, \boldsymbol{\beta}$.

2 The variational distribution

Let's define an factorising distribution $q(\mathbf{f})$.

$$q(\mathbf{f}) = \prod_i q_i(\mathbf{f}_i) = \prod_i \frac{p(\mathbf{y}_i|\mathbf{f}_i)c(\mathbf{f}_i)}{\int p(\mathbf{y}_i|\mathbf{f}_i)c(\mathbf{f}_i)\,\mathrm{d}\mathbf{f}_i} = \prod_i \frac{p(\mathbf{y}_i|\mathbf{f}_i)p(\mathbf{f}_i|\tilde{\mathbf{y}}_{\setminus i})}{\int p(\mathbf{y}_i|\mathbf{f}_i)p(\mathbf{f}_i|\tilde{\mathbf{y}}_{\setminus i})\,\mathrm{d}\mathbf{f}_i} = \prod_i p(\mathbf{f}_i|\mathbf{y}_i, \tilde{\mathbf{y}}_{\setminus i})$$

this is clearly just the marginal posterior for $\mathbf{f}_i$, conditioned on the other pseudotargets, and the i^{th} datum $\mathbf{y}_i$. In EP, moments of this distribution are used...

3 EP

EP proceeds by minimising the following KL divergence:

$$\mathrm{KL}[p(\mathbf{f}_i|\mathbf{y}_i, \tilde{\mathbf{y}}_{\setminus i})||p(\mathbf{f}_i|\tilde{\mathbf{y}})]$$

this is done by setting the variables $\tilde{\mathbf{y}}_i, \beta_i$ so that the moments of $p(\mathbf{f}_i|\tilde{\mathbf{y}})$ match those of $p(\mathbf{f}_i|\mathbf{y}_i, \tilde{\mathbf{y}}_{\setminus i})$ nobody seems to know what objective this is minimising, and it doesn't seem to be understood as a variational method.

4 A Variational Bound

Here's Bayes' rule with the usual positions of the marginal and posterior swapped:

$$\log p(\mathbf{y}) = \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})}$$

Introduce $p(\tilde{\mathbf{y}}|\mathbf{f})$ and $q(\mathbf{f})$:

$$\log p(\mathbf{y}) = \log \frac{p(\mathbf{y}|\mathbf{f})}{p(\tilde{\mathbf{y}}|\mathbf{f})} + \log \frac{p(\tilde{\mathbf{y}}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f})} + \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})}$$

and now take the expectation under $q(\mathbf{f})$ to give a variational bound (note that the lhs remains the same):

$$\log p(\mathbf{y}) = \mathbb{E}_{q(\mathbf{f})}\left[\log \frac{p(\mathbf{y}|\mathbf{f})}{p(\tilde{\mathbf{y}}|\mathbf{f})}\right] + \mathbb{E}_{q(\mathbf{f})}\left[\log \frac{p(\tilde{\mathbf{y}}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f})}\right] + \mathrm{KL}[q(\mathbf{f})||p(\mathbf{f}|\mathbf{y})] \quad (1)$$

This standard variational technique gives a natural approximation algorithm: by maximising the first two terms, we naturally minimise the KL divergence between the approximation q and the posterior. Unconventionally, the variational distribution is tied up in the pseudotargets. How can we make sense of this objective, and how can it be maximised?

Let's take the first term first. We'll assume that the likelihood in the nominator factorises, and we've defined the denominator in a similar fashion. This means we can write the whole thing as a sum:

$$\mathbb{E}_{q(\mathbf{f})}\left[\log \frac{p(\mathbf{y}|\mathbf{f})}{p(\tilde{\mathbf{y}}|\mathbf{f})}\right] = \sum_i \mathbb{E}_{q(\mathbf{f}_i)}\left[\log \frac{p(\mathbf{y}_i|\mathbf{f}_i)}{p(\tilde{\mathbf{y}}_i|\mathbf{f}_i)}\right]$$

Now multiply top and bottom of the fraction by $p(\mathbf{f}_i|\tilde{\mathbf{y}}_{\setminus i})$, and normalise.

$$\mathbb{E}_{q(\mathbf{f})}\left[\log \frac{p(\mathbf{y}|\mathbf{f})}{p(\tilde{\mathbf{y}}|\mathbf{f})}\right] = \sum_i \mathbb{E}_{q(\mathbf{f}_i)}\left[\log \frac{q(\mathbf{f}_i)}{p(\mathbf{f}_i|\tilde{\mathbf{y}})}\right] + \log \frac{p(\mathbf{y}_i|\tilde{\mathbf{y}}_{\setminus i})}{p(\tilde{\mathbf{y}}_i|\tilde{\mathbf{y}}_{\setminus i})}$$

Okay, that's now a sum of KL divergences, plus some constants. EP proceeds by minimising these one at a time, and this procedure is intuitive: we'll consecutively make the marginals of the EP approximation ($p(\mathbf{f}_i|\tilde{\mathbf{y}})$) closer to the variational approximation.

Since the KL divergences are all positive, we can write a bound on the marginal likelihood by removing them from the expression, as well as the KL divergence between the variational distribution q and the posterior. We have

$$\log p(y) \geq \mathbb{E}_{q(\mathbf{f})}\left[\log \frac{p(\tilde{\mathbf{y}}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f})}\right] + \sum_i \log \frac{p(\mathbf{y}_i|\tilde{\mathbf{y}}_{\setminus i})}{p(\tilde{\mathbf{y}}_i|\tilde{\mathbf{y}}_{\setminus i})} \quad (2)$$

and maximising this expression minimises the KL divergence from the variational distribution to the posterior, and from the marginals of the EP approximation ($p(\mathbf{f}_i|\tilde{\mathbf{y}})$) to the variational distribution.

5 Iterative procedures

Now, does the EP method really optimise this bound? We'll split out one of the terms (the i^{th}): write $p(\mathbf{f})$ as $p(\mathbf{f}_i|\mathbf{f}_{\setminus i})p(\mathbf{f}_{\setminus i})$, and remember that both our approximation q and the pseudotargets factorize.

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{f})}\left[\log\frac{p(\tilde{\mathbf{y}}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f})}\right] &= \mathbb{E}_{q(\mathbf{f}_i)q(\mathbf{f}_{\setminus i})}\left[\log\frac{p(\tilde{\mathbf{y}}_i|\mathbf{f}_i)p(\tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_{\setminus i})p(\mathbf{f}_i|\mathbf{f}_{\setminus i})p(\mathbf{f}_{\setminus i})}{q(\mathbf{f}_i)q(\mathbf{f}_{\setminus i})}\right] \\
&= \mathbb{E}_{q(\mathbf{f})}\left[\log\frac{p(\tilde{\mathbf{y}}_i|\mathbf{f}_i)p(\mathbf{f}_i|\mathbf{f}_{\setminus i})}{q(\mathbf{f}_i)}\right] + \mathbb{E}_{q(\mathbf{f}_{\setminus i})}\left[\log\frac{p(\tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_{\setminus i})p(\mathbf{f}_{\setminus i})}{q(\mathbf{f}_{\setminus i})}\right]
\end{aligned}
\tag{3}
$$

We've managed to split out the i^{th} factor into the first term, which looks like an unnormalised KL divergence. We could go ahead and start minimising that term, but that wouldn't be EP (will come back to that idea later). We need the top of this to contain the cavity distribution. Some manipulation of the cavity distribution reveals that

$$
\log p(\mathbf{f}_i|\mathbf{f}_{\setminus i}) = \log p(\mathbf{f}_i|\tilde{\mathbf{y}}_{\setminus i}) + \log\frac{p(\mathbf{f}_{\setminus i}, \tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_i)}{p(\tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_{\setminus i})p(\mathbf{f}_{\setminus i})}
$$

We can substitute this into the above to get (some terms cancel):

$$
\begin{aligned}
&= \mathbb{E}_{q(\mathbf{f}_i)}\left[\log\frac{p(\tilde{\mathbf{y}}_i|\mathbf{f}_i)p(\mathbf{f}_i|\tilde{\mathbf{y}}_{\setminus i})}{q(\mathbf{f}_i)}\right] + \mathbb{E}_{q(\mathbf{f})}\left[\log\frac{p(\mathbf{f}_{\setminus i}, \tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_i)}{q(\mathbf{f}_{\setminus i})}\right] \\
&= -\text{KL}[p(\mathbf{f}_i|\tilde{\mathbf{y}})||q(\mathbf{f}_i)] + \log p(\tilde{\mathbf{y}}_i|\tilde{\mathbf{y}}_{\setminus i}) + \mathbb{E}_{q(\mathbf{f})}\left[\log\frac{p(\mathbf{f}_{\setminus i}, \tilde{\mathbf{y}}_{\setminus i}|\mathbf{f}_i)}{q(\mathbf{f}_{\setminus i})}\right]
\end{aligned}
\tag{4}
$$

So minimising this KL wrt $\tilde{\mathbf{y}}_i, \beta_i$ appears to maximise our bound. But does it? The term on the right is also affected by these variables! This happens in two ways: first, the nominator is conditioned on $\mathbf{f}_i$; second, $q(\mathbf{f}_{\setminus i})$ are dependent on $\tilde{\mathbf{y}}_i, \beta_i$.

The first of these can only be resolved by making $p(\mathbf{f}_{\setminus i}|\mathbf{f}_i) = p(\mathbf{f}_{\setminus i})$, i.e making the i^{th} point independent. This leads to a rather uninteresting Gaussian Process, so is impracticable.

The second can be more easily avoided by not changing the $q(\mathbf{f}_{\setminus i})$ when we change $\tilde{\mathbf{y}}_i, \beta_i$, leaving each of the qs dependent on *old* values of $\tilde{\mathbf{y}}, \boldsymbol{\beta}$ until we update them. We'll need this minor trick in our variational method...

6 A convergent variational method.

How about just taking the bound and optimising it by gradient methods? We'll need some properties of the sigmoid-Gaussian distribution.

7 The sigmoid-Gaussian

Let's define a distribution for x which is made by multiplying a Gaussian by a sigmoid. (We'll use x instead of $\mathbf{f}_i$) here for simplicity.

$$
q(x) = \phi(x)\mathcal{N}(x|m, s)/z,
$$

with z defined as

$$z = \int \phi(x)\mathcal{N}(x|m, s)\,\mathrm{d}x$$

Some properties of this distribution are given by [**?**], namely

$$\begin{aligned}
z &= \phi(m') \\
\mathbb{E}_q[x] &= m + \frac{s\mathcal{N}(m')}{\phi(m')\sqrt{1+s}} \\
\mathbb{E}_q[x^2] &= 2m\mathbb{E}_q[x] - m^2 + s - \frac{s^2 m'\mathcal{N}(m')}{\phi(m')(1+s)} \\
\mathbb{E}_q[(x - \mathbb{E}_q[x])^2] &= s - \frac{s^2 m'\mathcal{N}(m')}{\phi(m')(1+s)}(m' + \frac{\mathcal{N}(m')}{\phi(m')})
\end{aligned}$$

with $m' = \frac{m}{\sqrt{1+s}}$. We also need the entropy, and the derivative of the entropy wrt m, s. unfortunately, the entropy turns out to be intractable