Department of Mathematics & Applied Mathematics

Notes on active inference, mostly from the book by Parr, Pezzulo and Friston

Associate Professor Jonathan Shock
Department of Mathematics & Applied Mathematics
University of Cape Town
jon.shock@gmail.com

(August 21, 2025)

Contents

Aim of these notes 1

The aim here is to try and fill in some of the mathematical gaps in the book by Friston et al. While there are some superb explanations in the book, there are important details left out in the derivations which may not be that important on a first pass, but I think make things much clearer when you want to understand it in detail.

Starting in chapter 4 2

Out general aim will be to observe things that we expect to observe. We will come on later to the things that we WANT to observe, when we think about preferences, but now we can just think about the things that our model predicts that we are likely to see. That is that we want to have the highest probability of making the observations that we make. ie maximising $P(y)$. y is a measure of our sensory states.

We have a generative model which includes both a prior over external states $P(x)$ as well as $P(y|x)$ - ie. the probability of having a particular sensory measurement given that the world is in a particular state.

From $P(x)$ and $P(y|x)$ (which means that we also know $P(x, y)$, we can, in theory work out $P(y)$ but both of these require us to sum over all possible external states which is generally impossible.

To calculate $P(y)$ we would need to calculate:

$$P(y) = \sum_x P(y|x)P(x) = \mathbb{E}_{P(x)}[P(y|x)] \tag{2.1}$$

The problem here is the expectation over all $P(x)$. Instead we will have some approximation of $P(x)$ which we would like to get as close to $P(x)$ as possible. This is $Q(x)$. We will alter $Q(x)$ so that we can minimise surprise.

We will use Jensen's inequality which says that:

$$\mathbb{E}(\ln(x)) \leq \ln(\mathbb{E}(x)) \tag{2.2}$$

So we can write:

$$\ln P(y) = \ln \sum_x P(y|x)P(x) = \ln \sum_x P(y|x)\frac{P(x)}{Q(x)}Q(x) \tag{2.3}$$

$$= \ln \sum_x \frac{P(y,x)}{Q(x)}Q(x) = \ln \mathbb{E}_{Q(x)}\left[\frac{P(y,x)}{Q(x)}\right] \tag{2.4}$$

Again, we want to minimise this quantity by making our model match the world, and by making $Q(x)$ as close as possible to $P(y|x)$ for a given set of observations y.

We use Jensen now to write:

$$\ln P(y) = \ln \sum_x P(y|x)P(x) = \ln \sum_x P(y|x)\frac{P(x)}{Q(x)}Q(x) \tag{2.5}$$

$$= \ln \sum_x \frac{P(y,x)}{Q(x)}Q(x) = \ln \mathbb{E}_{Q(x)}\left[\frac{P(y,x)}{Q(x)}\right] \tag{2.6}$$

$$\geq \mathbb{E}_{Q(x)}\left[\ln \frac{P(y,x)}{Q(x)}\right] \tag{2.7}$$

We define as the negative of the free energy:

$$-F[Q, y] = \mathbb{E}_{Q(x)}\left[\ln \frac{P(y,x)}{Q(x)}\right] \tag{2.8}$$

So we know that while the free energy will be greater than the surprise. If we can minimise the free energy we are minimising the surprise, or maximising the model evidence. We can minimise the free energy by making sure that our observations match our expectations (changing y, or by updating $Q(x)$.

We can write the free energy as:

$$
\begin{aligned}
F[Q,y] &= -\mathbb{E}_{Q(x)}\left[\ln \frac{P(y,x)}{Q(x)}\right] = -\mathbb{E}_{Q(x)}\left[\ln \frac{P(x|y)P(y)}{Q(x)}\right] && (2.9)\\
&= -\mathbb{E}_{Q(x)}\left[\ln P(x|y)\right] - \mathbb{E}_{Q(x)}\left[\ln P(y)\right] + \mathbb{E}_{Q(x)}\left[\ln Q(x)\right] && (2.10)\\
&= -\mathbb{E}_{Q(x)}\left[\ln P(x|y)\right] - \ln P(y) + \mathbb{E}_{Q(x)}\left[\ln Q(x)\right] && (2.11)\\
&= \mathbb{E}_{Q(x)}\left[\ln Q(x)\right] - \mathbb{E}_{Q(x)}\left[\ln P(x|y)\right] - \ln P(y) && (2.12)\\
&= \mathbb{E}_{Q(x)}\left[\ln Q(x) - \ln P(x|y)\right] - \ln P(y) && (2.13)\\
&= D_{KL}\left[Q(x)||P(x|y)\right] - \ln P(y) && (2.14)
\end{aligned}
$$

So again, minimising the free energy can be done by making $Q(x)$ as close as possible to $P(x|y)$ (when averaged over $Q(x)$ (minimising the divergence), or by minimising the surprise.

We must remember that $-\ln P(y)$ will itself be positive.

By taking:

$$\ln P(x,y) = \ln P(y) + \ln P(x|y) \tag{2.15}$$

we can either take the expectation of this over the distribution $P(x|y)$ ie. the posterior probability, or over the approximation to the posterior $Q(x)$:

$$\mathbb{E}_{P(x|y)}\left[\ln P(x,y)\right] = \ln P(y) + \mathbb{E}_{P(x|y)}\left[\ln(P(x|y)\right] \tag{2.16}$$

or

$$\mathbb{E}_{Q(x)}\left[\ln P(x,y)\right] = \ln P(y) + \mathbb{E}_{Q(x)}\left[\ln(P(x|y)\right] \tag{2.17}$$

but we know from equation 2.13 that

$$\ln P(y) = \mathbb{E}_{Q(x)}\left[\ln Q(x) - \ln P(x|y)\right] - F[Q,y] \tag{2.18}$$

So we have:

$$
\begin{aligned}
\mathbb{E}_{Q(x)}\left[\ln P(x,y)\right] &= \mathbb{E}_{Q(x)}\left[\ln Q(x) - \ln P(x,y)\right] - F[Q,y] + \mathbb{E}_{Q(x)}\left[\ln P(x|y)\right] && (2.19)\\
&= -F[Q,y] + \mathbb{E}_{Q(x)}\left[\ln Q(x) - \ln P(x|y) + \ln P(x|y)\right] && (2.20)\\
&= -F(Q,y) + \mathbb{E}_{Q(x)}\left[\ln Q(x)\right] && (2.21)
\end{aligned}
$$

Comparing equation 2.16 and 2.21:

$$
\begin{aligned}
\mathbb{E}_{P(x|y)}\left[\ln P(x,y)\right] &= \ln P(y) + \mathbb{E}_{P(x|y)}\left[\ln P(x|y)\right] && (2.22)\\
\mathbb{E}_{Q(x)}\left[\ln P(x,y)\right] &= -F(Q,y) + \mathbb{E}_{Q(x)}\left[\ln Q(x)\right] && (2.23)
\end{aligned}
$$

So the free energy approximates the surprise and the $Q(x)$ approximates the posterior, $P(x|y)$.

Chapter 2 details

From chapter 2 we see three ways of writing the Free energy. We start with the equation above:

$$F(Q, y) = -\mathbb{E}_{Q(x)} [\ln P(x, y)] + \mathbb{E}_{Q(x)} [\ln Q(x)] \tag{2.24}$$

and note that the last term is the negative of the entropy of Q. This is a measure of the average surprise of the measurement of $Q(x)$. It's very important to note that this isn't the entropy coming from $P(y)$ which we want to minimise. This is over our approximation of $P(x|y)$.

The flatter is $Q(x)$, the larger will be its entropy.

We can write:

$$F(Q, y) = -\mathbb{E}_{Q(x)} [\ln P(x, y)] - H [Q(x)] \tag{2.25}$$

The two terms on the right are described as the energy and entropy. The energy is the consistency of $Q(x)$ with the generative model.

The latter is the entropy, which, to minimise the free energy, we actually want to maximise over. This sounds surprising, but it means that if we have little evidence, then we should be maximally uncertain.

We can expand this as:

$$\begin{aligned}
F(Q, y) &= -\mathbb{E}_{Q(x)} [\ln (P(y|x)P(x))] + \mathbb{E}_{Q(x)} [\ln Q(x)] && (2.26)\\
&= -\mathbb{E}_{Q(x)} \left[\ln \left(\frac{P(y|x)P(x)}{Q(x)} \right) \right] && (2.27)\\
&= -\mathbb{E}_{Q(x)} \left[\ln \left(\frac{P(x)}{Q(x)} \right) \right] - \mathbb{E}_{Q(x)} [\ln P(y|x)] && (2.28)\\
&= -\mathbb{E}_{Q(x)} \left[\ln \left(\frac{P(x)}{Q(x)} \right) \right] - \mathbb{E}_{Q(x)} [\ln P(y|x)] && (2.29)\\
&= \mathbb{E}_{Q(x)} [\ln Q(x) - \ln P(x)] - \mathbb{E}_{Q(x)} [\ln P(y|x)] && (2.30)\\
&= D_{KL} [Q(x)||P(x)] - \mathbb{E}_{Q(x)} [\ln P(y|x)] && (2.31)
\end{aligned}$$

The first term is the complexity, which we want to minimise. This says that we want $Q(x)$ to be as close as possible to $P(x)$, which means not overfitting - ie. not being too sure about some particular sensory state and thus missing something about states that we haven't seen too many times. This term will help to smooth the difference between $Q(x)$ and $P(x)$. The second term is the negative of the accuracy. We want to maximise the accuracy, which means having a high expectation of the particular observation, given the external state, averaged over the estimated priors of the external states.

The third way that we can split this is the following:

$$F(Q,y) = -\mathbb{E}_{Q(x)}\left[\ln\left(P(y|x)P(x)\right)\right] + \mathbb{E}_{Q(x)}\left[\ln Q(x)\right] \tag{2.32}$$

$$= -\mathbb{E}_{Q(x)}\left[\ln\left(\frac{P(y|x)P(x)}{Q(x)}\right)\right] \tag{2.33}$$

$$= -\mathbb{E}_{Q(x)}\left[\ln\left(\frac{P(y,x)}{Q(x)}\right)\right] \tag{2.34}$$

$$= -\mathbb{E}_{Q(x)}\left[\ln\left(\frac{P(x,y)}{Q(x)}\right)\right] \tag{2.35}$$

$$= -\mathbb{E}_{Q(x)}\left[\ln\left(\frac{P(x|y)P(y)}{Q(x)}\right)\right] \tag{2.36}$$

$$= \mathbb{E}_{Q(x)}\left[\ln Q(x) - \ln P(x|y)\right] - \mathbb{E}_{Q(x)}\left[\ln P(y)\right] \tag{2.37}$$

$$= D_{KL}\left[Q(x)||P(x|y)\right] - \mathbb{E}_{Q(x)}\left[\ln P(y)\right] \tag{2.38}$$

$$= D_{KL}\left[Q(x)||P(x|y)\right] - \ln P(y) \tag{2.39}$$

where here remembering that we want $Q(x)$ to approximate $P(x|y)$, the first term is the divergence between the two which we want to minimise while maximising the model evidence.

I think that it's worth comparing the complexity and the divergence, which are very similar. However, the divergence is dependent on the particular observation, whereas the complexity is only related to the generative model - ie. the prior over external states. The complexity says that we want $Q(x)$ to approximate the prior over external states as well as possible (we don't want to overfit to a given observation, y). The divergence says that we want Q to approximate $P(x|y)$ for a given observation y as well as possible.

Expected Free Energy from section 2.8

Here we have not just states, x and observations y but sequences of each, which we label $\tilde{x}$ and $\tilde{y}$. We also have a policy π, and a preference parameter C, which tells us the observations that we would most like to see.

We now have our approximation to the posterior path through state space given by:

$$Q(\tilde{x}|\pi) \tag{2.40}$$

This is a simple extension to the Q we had before.

However, we now introduce also:

$$Q(\tilde{x},\tilde{y}|\pi) = Q(\tilde{x}|\pi)P(\tilde{y}|\tilde{x}) \tag{2.41}$$

which is now an approximation to the joint probability of being on trajectory $\tilde{x}$ and making observations $\tilde{y}$. Note that we don't know the posterior $P(\tilde{x}|\tilde{y})$ but we do know the likelihood of observing $\tilde{y}$ given that we take trajectory $\tilde{x}$.

It seems in equation 2.6 that we also have a quantity $Q(\tilde{y}|\pi)$, but to calculate this, do we not need to sum over all $\tilde{x}$, ie:

$$Q(\tilde{y}|\pi) = \sum_{\tilde{x}} Q(\tilde{x},\tilde{y}|\pi) \tag{2.42}$$

This sum over the external trajectories seems like something that we wanted to avoid in the first place, so it's not clear to me how we get $Q(\tilde{y}|\pi)$

Thoughts on the vectorised expected free energy 3

In chapter 4 we are given the following:

$$\vec{\pi}_o = \sigma(-\vec{G}) \tag{3.1}$$

which says that the probability of using each policy is related to the expected free energy of that policy normalised via a sigma function.

This means that $\vec{\pi}_o$ has the dimension of the number of different policies, and $\vec{G}$ is a vector of length $|\Pi|$, where $\Pi = \{\pi_1, \pi_2, \pi_3 ...\}$. The components of $\vec{G}$ can thus be indexed by the particular policy, π. We are next told that in vectorised form:

$$\vec{G}_\pi = \vec{H} \cdot \vec{S}_{\pi\tau} + \vec{O}_{\pi\tau} \cdot \vec{\zeta}_{\pi\tau} \tag{3.2}$$

While some of these quantities make sense to be vectorised as they all are in the text, it's not clear why they all are. In particular, $\vec{H}$ and $\vec{S}_{\pi\tau}$ live in the vector space of states, but when we dot them together, we should get a scalar in that space. Thus $\vec{G}_\pi$ would seem to be a scalar component, indexed simply by the policy.

Now we have to work out what we are taking dot products over in the above.

$$\vec{S}_{\pi\tau} \tag{3.3}$$

is the probability of being in a particular state at a given time τ given that we are using policy π. This means then that $\vec{S}_{\pi\tau}$ is a vector of states, and that is indexed by π and τ.

> We have to be very careful here because it looks like we've lost the approximations that we made using Q. However, Q is now a vector over the discrete states. For example:
> Taking the vector $\vec{S}_{\pi\tau}$ which is a vector of probabilities that we will be in a given state, we note that this is really under our generative model, and so:
>
> $$Q(s_\tau|\pi) = cat(\vec{S}_{\pi\tau})$$
>
> So in a sense we can think that Q inherits the discrete indices of the states and observations at discrete timesteps.

We weight each state with $\vec{H}$ which is:

$$\vec{H} = -\mathsf{diag}\left(\vec{A}^T \cdot ln(\vec{A})\right) \tag{3.4}$$

meaning that $\vec{H}$ is a row vector of the diagonals of the above. Note that I have made the first $\vec{A}$ into $\vec{A}^T$. $\vec{A}$ is a matrix of the probability of making a given observation given that you are in a particular state. This is part of the generative model. ie.

$$\vec{A} = \begin{pmatrix} \mathsf{p}(o_1|s_1) & \mathsf{p}(o_1|s_2) & ... \\ \mathsf{p}(o_2|s_1) & \mathsf{p}(o_2|s_2) & ... \\ ... & ... & ... \end{pmatrix} \tag{3.5}$$

In a similar vein as for $\vec{S}_{\pi\tau}$, we have:

$$Q(o_\tau|\pi) = Cat(\vec{A}\vec{s}_{\pi\tau}) = Cat(\vec{o}_{\pi\tau})$$

which says that from our observation matrix, which is the dot product between the generative model state probabilities and the generative model relationship between states and observations, we form our approximation of what we expect that observations to be.

We note that in writing this, it looks like we have marginalised out over the states. However, we are not marginalising out over the true states, but those states captured by the generative model only. This is an important distinction. Given a distribution over policies (ie. the probability of being in a given policy), a distribution over states at a given time, under a given policy and the generative model of how states give rise to observations, we can form:

$$Q(s_\tau) = Cat(\sum_\pi \pi_\pi \vec{S}_{\pi\tau}]) \tag{3.6}$$

$$Q(s_\tau|\pi) = Cat(\vec{S}_{\pi\tau}) \tag{3.7}$$

$$Q(o_\tau|\pi) = Cat(\vec{A}\vec{S}_{\pi\tau}) = Cat(\vec{O}_{\pi\tau}) \tag{3.8}$$

$$\tag{3.9}$$

Therefore $\vec{H}$ seems to be a vector living in the space of states. It makes sense then to take the dot product between $\vec{H}$ and $\vec{S}_{\pi\tau}$ but it does mean that we are left with both the π AND the τ indices. It seems that the tau indices somehow need to be summed over to give us our $\vec{G}$.

Now looking at the second term:

$$\vec{O}_{\pi\tau} = \vec{A}\vec{S}_{\pi\tau} \tag{3.10}$$

$\vec{A}$ is a matrix indexed over states and observations, and we are using the matrix multiplication with the vector of the probability of being in a given state at a particular time under a particular policy. So this vector quantity $\vec{o}$ is essentially the probability of making a particular observation given that we are timestep τ and we are under policy π.

Finally the quantity:

$$\vec{\zeta}_{\pi\tau} = \ln \vec{o}_{\pi\tau} - ln\vec{C}_\tau \tag{3.11}$$

$\vec{C}$ should be independent of the policy and is related to our preference (or indeed prior) over observations.

So the second term in equation 3.12 has a dot product over the observation space whereas the first term is a dot product over the state space.

Writing the whole thing out, component-wise, I believe we get:

$$G_\pi = \left[\sum_s \left(-\mathsf{diag}\left(\sum_o \left(A^T\right)_{so} (ln(A))_{os}\right)(S_{\pi\tau})_s\right)\right] + \left[\sum_o \left(\left(\sum_s \left(A^T\right)_{os} (S_{\pi\tau})_s\right)((\ln o_{\pi\tau})_o - (lnC_\tau)_o)\right)\right] \tag{3.12}$$

The question then is what happens to τ. I think that there should really be a τ index on G_π as well.

More details on the expected free energy 4

We have looked in the last section at one way of writing the expected free energy, given in chapter 4. This is:

$$\begin{aligned} G_{\tau\pi} &= \vec{H} \cdot \vec{S}_{\pi\tau} + \vec{O}_{\pi\tau} \cdot \vec{\zeta}_{\pi\tau} \\ &= \left[\sum_s \left(-\mathsf{diag}\left(\sum_o \left(A^T\right)_{so} (ln(A))_{os}\right)(S_{\pi\tau})_s\right)\right] + \left[\sum_o \left(\left(\sum_s \left(A^T\right)_{os} (S_{\pi\tau})_s\right)((\ln o_{\pi\tau})_o - (lnC_\tau)_o)\right)\right] \end{aligned} \tag{4.1}$$

Where we have put the τ index back on G. How does this expression relate to any of the others in the previous chapters?

This comes from the non-vectorised expression given by:

$$G(\pi) = \mathbb{E}_{\tilde{Q}}\left[H\left[P(\tilde{o}|\tilde{s})\right]\right] + D_{KL}\left[Q(\tilde{o}|\pi)||P(\tilde{o}|C)\right] \tag{4.2}$$

which is described as the expected ambiguity plus the risk.

This now makes some more sense, as we have an expectation over $\tilde{Q}$ which is captured via the categorical variable in $S_{\pi\tau}$ of the entropy of $P(\tilde{o}|\tilde{s})$ which is captured in the $\vec{A}$ matrix. So this first term makes sense.

We should also note that when it says $\tilde{Q}$ it really means $Q(\tilde{s}|\pi)$

For the second term we have, in the non-vectorised version the difference in the distributions between what we want/expect to observe, and what we are actually observing. In the vectorised version we have the difference between the probability of observations matrix and the preference distribution matrix. Remembering that the definition of the KL-divergence is the expectation over the difference in logs of the distribution, this term also makes sense.

Can we be a bit clearer about why we call them "ambiguity" and "risk"?

If the matrix of $P(\tilde{o}|\tilde{s})$ is sparse, or at best diagonal, then we have a really strong belief in what observations will come from what states. This means that we see something, and we know what it is. This is in contrast to a matrix which is non-sparse, which would mean that a given state could give rise to many different observations. This would clearly have a high entropy.

For the KL divergence we call this the risk term because it measures how far away our distribution over expected observations given the policy is from our preference distribution.

4.1 Different ways of writing the expected free energy

Now let's revisit the different ways that we can write the expected free energy.

The ambiguity+risk descriptions seems clear. We want a policy which will make our observations unambiguous while also having a small risk. We don't want to be really certain about the fact that we are on fire. Nor do we want to think that in the future we will be at the perfect temperature, but not be sure about what our observations are really telling us.

Just as we we did for the free energy, we will see the different ways to split up this expression. The

best way to do this is to expand it out as much as possible first:

$$G(\pi) = \mathbb{E}_{Q(\tilde{s}|\pi)}\left[H\left[P(\tilde{o}|\tilde{s})\right]\right] + D_{KL}\left[Q(\tilde{o}|\pi)||P(\tilde{o}|C)\right] \tag{4.3}$$

$$= -\sum_{\tilde{s}}\left[Q(\tilde{s}|\pi)\sum_{\tilde{o}}P(\tilde{o}|\tilde{s})\ln P(\tilde{o}|\tilde{s})\right] + \sum_{\tilde{o}}Q(\tilde{o}|\pi)\left(\ln Q(\tilde{o}|\pi) - \ln P(\tilde{o}|C)\right) \tag{4.4}$$

where again we should remember that $P(\tilde{o}|\tilde{s})$ are not dependent on the policy, but they are known within our generative model.

The other way of writing this in chapter 2 is (though converting all to s and o from x and y):

$$G(\pi) = -\mathbb{E}_{Q(\tilde{s},\tilde{o}|\pi)}\left[D_{KL}[Q(\tilde{s}|\tilde{o},\pi)||Q(\tilde{s}|\pi)]\right] - \mathbb{E}_{Q(\tilde{o}|\pi)}\left[\ln P(\tilde{o}|C)\right] \tag{4.5}$$

We are however going to rewrite this slightly as it turns out that the KL divergence in the first term is not needed, and we can simply write it as the difference in logs:

$$G(\pi) = -\mathbb{E}_{Q(\tilde{s},\tilde{o}|\pi)}\left[\ln Q(\tilde{s}|\tilde{o},\pi) - \ln Q(\tilde{s}|\pi)\right] - \mathbb{E}_{Q(\tilde{o}|\pi)}\left[\ln P(\tilde{o}|C)\right] \tag{4.6}$$

$$= \sum_{\tilde{s}}\sum_{\tilde{o}} -Q(\tilde{s},\tilde{o}|\pi)(\ln Q(\tilde{s}|\tilde{o},\pi) - \ln Q(\tilde{s}|\pi)) - \sum_{\tilde{o}}Q(\tilde{o}|\pi)\ln P(\tilde{o}|C) \tag{4.7}$$

$$= \sum_{\tilde{s}}\sum_{\tilde{o}} Q(\tilde{s},\tilde{o}|\pi)(\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}|\tilde{o},\pi)) - \sum_{\tilde{o}}Q(\tilde{o}|\pi)\ln P(\tilde{o}|C) \tag{4.8}$$

$$= \sum_{\tilde{s}}\sum_{\tilde{o}} P(\tilde{o}|\tilde{s})Q(\tilde{s}|\pi)(\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}|\tilde{o},\pi)) - \sum_{\tilde{o}}Q(\tilde{o}|\pi)\ln P(\tilde{o}|C) \tag{4.9}$$

where in the last line we used the definition that:

$$Q(\tilde{o},\tilde{s}|\pi) = P(\tilde{o}|\tilde{s})Q(\tilde{s}|\pi)$$

but we can also write that:

$$Q(\tilde{o},\tilde{s}|\pi) = Q(\tilde{o}|\tilde{s})Q(\tilde{s}|\pi)$$

meaning that really $Q(\tilde{o}|\tilde{s},\pi) = P(\tilde{o}|\tilde{s})$ which just says that we are just using our generative model for how an observation is dependent on a state.

So the last terms in equations 4.4 and 4.9 are clearly equal. Now we need to show that the rest are equal. We can take everything inside the double sum in equation 4.4 by writing everything but the last term as:

$$-\sum_{\tilde{s}}\left[Q(\tilde{s}|\pi)\sum_{\tilde{o}}P(\tilde{o}|\tilde{s})\ln P(\tilde{o}|\tilde{s})\right] + \sum_{\tilde{o}}Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi) \tag{4.10}$$

$$= -\sum_{\tilde{s}}\left[Q(\tilde{s}|\pi)\sum_{\tilde{o}}P(\tilde{o}|\tilde{s})\ln P(\tilde{o}|\tilde{s})\right] + \sum_{\tilde{s}}Q(\tilde{s}|\pi)\sum_{\tilde{o}}Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi) \tag{4.11}$$

$$= \sum_{\tilde{s}}\sum_{\tilde{o}}Q(\tilde{s}|\pi)(Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi) - P(\tilde{o}|\tilde{s})\ln P(\tilde{o}|\tilde{s})) \tag{4.12}$$

So what we need to be able to show is that:

$$P(\tilde{o}|\tilde{s})Q(\tilde{s}|\pi)(\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}|\tilde{o},\pi)) = Q(\tilde{s}|\pi)(Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi) - P(\tilde{o}|\tilde{s})\ln P(\tilde{o}|\tilde{s})) \tag{4.13}$$

ie.

$$P(\tilde{o}|\tilde{s})(\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}|\tilde{o},\pi)) = (Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi) - P(\tilde{o}|\tilde{s})\ln P(\tilde{o}|\tilde{s})) \tag{4.14}$$

We can use Bayes' to write:

$$Q(\tilde{s}|\tilde{o},\pi) = Q(\tilde{o}|\tilde{s},\pi)\frac{Q(\tilde{s}|\pi)}{Q(\tilde{o}|\pi)}$$

Again $Q(\tilde{o}|\tilde{s},\pi)$ is actually not dependent on the policy, and is equal to $P(\tilde{o}|\tilde{s})$. So:

$$Q(\tilde{s}|\tilde{o},\pi) = P(\tilde{o}|\tilde{s})\frac{Q(\tilde{s}|\pi)}{Q(\tilde{o}|\pi)}$$

So we have:

$$P(\tilde{o}|\tilde{s})(\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}|\tilde{o},\pi)) \tag{4.15}$$
$$= P(\tilde{o}|\tilde{s})(\ln Q(\tilde{s}|\pi) - \ln\left(P(\tilde{o}|\tilde{s})\frac{Q(\tilde{s}|\pi)}{Q(\tilde{o}|\pi)}\right)) \tag{4.16}$$
$$= P(\tilde{o}|\tilde{s})(\ln Q(\tilde{o}|\pi) - \ln\left(P(\tilde{o}|\tilde{s})\right)) \tag{4.17}$$

We have to be a little careful here as we are not equating the terms above, but really the expectation over these terms, so for the two expressions to be the same we need that:

$$\sum_{\tilde{s}}\sum_{\tilde{o}} Q(\tilde{s}|\pi)(P(\tilde{o}|\tilde{s})\ln Q(\tilde{o}|\pi) - P(\tilde{o}|\tilde{s})\ln\left(P(\tilde{o}|\tilde{s})\right)) = \sum_{\tilde{s}}\sum_{\tilde{o}} Q(\tilde{s}|\pi)(Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi) - P(\tilde{o}|\tilde{s})\ln P(\tilde{o}|\tilde{s})) \tag{4.18}$$

One of these terms matches. Can we now show that:

$$\sum_{\tilde{s}}\sum_{\tilde{o}} Q(\tilde{s}|\pi)(P(\tilde{o}|\tilde{s})\ln Q(\tilde{o}|\pi)) = \sum_{\tilde{s}}\sum_{\tilde{o}} Q(\tilde{s}|\pi)(Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi)) \tag{4.19}$$

We note that on the right hand side, the only $\tilde{s}$ is in $Q(\tilde{s}|\pi)$ and therefore we can remove that sum immediately. So we have:

$$\sum_{\tilde{s}}\sum_{\tilde{o}} Q(\tilde{s}|\pi)(P(\tilde{o}|\tilde{s})\ln Q(\tilde{o}|\pi)) = \sum_{\tilde{o}}(Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi)) \tag{4.20}$$

We can also rewrite $P(\tilde{o}|\tilde{s}) = Q(\tilde{o}|\tilde{s},\pi)$

$$\sum_{\tilde{s}}\sum_{\tilde{o}} Q(\tilde{s}|\pi)Q(\tilde{o}|\tilde{s},\pi)\ln Q(\tilde{o}|\pi) = \sum_{\tilde{o}}(Q(\tilde{o}|\pi)\ln Q(\tilde{o}|\pi)) \tag{4.21}$$

and finally we see that:

$$\sum_{\tilde{s}} Q(\tilde{s}|\pi)Q(\tilde{o}|\tilde{s},\pi) = Q(\tilde{o},\pi)$$

which is exactly what we need...and the two sides are the same!

OK, so having convinced ourselves that these are the same, we can feel confident in analysing what the secondary way of writing it means. Now we see why writing 4.5 in terms of the KL divergence may be smart. We write it again here for reference:

$$G(\pi) = -\mathbb{E}_{Q(\tilde{s},\tilde{o}|\pi)}\left[D_{KL}[Q(\tilde{s}|\tilde{o},\pi)||Q(\tilde{s}|\pi)]\right] - \mathbb{E}_{Q(\tilde{o}|\pi)}\left[\ln P(\tilde{o}|C)\right] \tag{4.22}$$

We want to minimize this expression. There are two terms with - signs in front, so we want to maximise each of these. Maximising the first means making $Q(\tilde{s}|\tilde{o}, \pi)$ as different from $Q(\tilde{s}|\pi)$ as possible, which means that we seek observations which give us knew information. That is, if you thought that you were in a given state, but you make an observation, you want it to give you maximal new information about which state you are really in. The second terms says that you want to make observations which are as close to preferred observations as possible.