Using posterior sampling for exploration

These are my thoughts based on reading. Things here can be wrong, I'll try to point out where I might be assuming something that might not be true with footnotes.

Why can we not just use a posterior Q-value directly?

It seems reasonable that sampling the Q-value directly should account for our uncertainty. However, there is an issue with this that originates from the Bellman equation.

$Q^*(s,a|\pi) = R(s,a) + \gamma \sum_{s'} P_{s'|s,a} \sum_{a'} \pi(a'|s') Q_\pi(s',a'|\pi)$

The Q-value is conditioned on a policy π. If we have a constant policy the problem is stationary (1) and all the parametric uncertainty is contained in the posterior.

However, as soon as we start changing our policy, the expected total reward changes making the target non-stationary. Despite this, the function approximation methods used are derived for stationary targets. This means the variance and posterior calculations will underestimate the uncertainty as they do not take into account the uncertainty in the policy of future Q-values. (2)

To deal with this most papers attempt to "propagate" the variance from future Q-values into the current Q-value. This process means we cannot sample the posterior directly. Instead we must assume its distribution to be normal and sample based on our mean and propagated variance estimate.

The way this is dealt with in the UBE is to create a Bellman like equation that relates the variance of the current Q-value to the variance of the next Q-value. This variance can then be learnt in the same way as Q-values are learnt. The downside to this is that the variance is an upper bound and one assumes that the posterior distribution is normal. In addition UBE assumes that Q-value uncertainties are uncorrelated, as seen by the covariance matrix when Thompson sampling.

(This part is based on a quick read, can be wrong) The successor uncertainty paper also does some form of propagation by including the Q value relations in the loss function through regularization terms. This incorporates the Q-value dependencies in contrast to UBE. However it once again assumes a normal posterior.

Directly sampling posterior

There are many ways to sample an approximation to the posterior Q (Dropout, Ensemble Bootstrapping), however most of them do not perform any propagation and thus do not incorporate the policy uncertainty. The randomize prior function paper claims to do so, but I'm having a hard time seeing how it actually does this. (3)

Ideally we want to find a Bellman equation for $Q^*(s,a,\pi)$ where π is the optimal policy such that we can create a model that samples the posterior $Q^*(s,a,\pi)$. I believe that this might give the best possible exploration/exploitation trade-off(4). However, I can't be the first to think of this so there are probably plenty of obstacles stopping this.

Possible direction to move in

My first thought is that we need an Actor-Critic method to be able to have a prior over π. If we start on some simple environments maybe it is possible to find good posterior samples from the policy. If

we denote samples by * we then get $Q^*(s,a) = r(s,a) + Q^*(s', \pi^*(s'))$. Questions surrounding this include:

1. Is this on-policy?
2. How often do we sample π^*? Per Q-value, transition or episode?
3. Can we even get a posterior over π?

Bayesian Reinforcement Learning

Towards the end of the prosjekt oppgave I realized I'm shaky on bayesian reinforcement learning. I find the whole concept a bit confusing. The papers so far all consider a bayesian MDP and that samples values or action-values are supposed to be the correct values for the sampled MDP. I have two issues with this

1. How are we updating the distributions of the reward and transition probabilities? In UBE we just use samples directly from the environment. Is this the equivalent of sampling the posterior reward and transition probabilities?
2. In the original bayesian MDP paper they use value iteration to calculate the correct value function for each sampled MDP. In the UBE we just train the action-value function on the environment and use this over all MDP samples. Why does this work?

I think understanding this will clear up if it is possible to include a policy prior in the mix of things. It might be necessary to build up a new equation to be able to do this.

1: Even with a constant policy our target changes as we learn. However, from the way it is discussed in Sutton this is still considered stationary. I think this is because the return from the environment is stationary. How does this effect our uncertainty measure?

2: Intuitively this isn't very clear and it would be interesting to test on a toy example just to include in the master.

3: Ian Osband wrote this paper and has analyzed posterior sampling from a pretty theoretical perspective. So I think its more likely that I don't see the way it deals with the issue rather than him not dealing with the issue. Either way the paper achieves similar results to the UBE.

4: Based on this relatively unsourced sentence in UBE - "By contrast, for any set of prior beliefs the optimal exploration policy can be computed directly by dynamic programming in the bayesian belief space."