CS294-40 Learning for Robotics and Control Lecture 20 - 11/4/2008

Exploration/Exploitation

Lecturer: Pieter Abbeel *Scribe: Brandon Basso*

Original algorithm, E^3 (explicitly explore or exploit) developed by Kearns and Singh, 1998
Later, easier to implement version, RMax, by Brafman and Tenneholtz.

0.1 Rmax Algorithm (crude sketch):

1. Initialize the MDP $\bar{M} = (S, A, \bar{P}, \bar{R})$ with the (assumed known) set of states S, set of actions A, and with estimates of the transition probabilities and reward function. In particular set $\bar{P}$ such tbar for any given state, the agent will remain in that state no matter which action is taken. Set the reward $\bar{R}$ equal to the maximum reward attainable over all states. (which is presumably known)

2. Compute the optimal policy π for the MDP $\bar{M}$, and start executing.

3. While executing the policy π accumulate statistics about the transitions and the rewards and build up an empirical transition model $\hat{P}$ and empirical average reward function $\hat{R}$. Once we have an accurate model of the reward function or a transition model, i.e., once we have sufficiently many samples for sample complexity bounds to tell us that $|P(s'|s,a) - \hat{P}(s'|s,a)| \leq \epsilon \forall s', a$ with probability $1 - \delta$, we consider that state s "known." We fill the corresponding transition model into $\bar{P}$. (And similarly for the reward function.) We go back to Step 2.

Through the pigeon hole principle, states will become known. Associating the maximum reward with each state (until its known) guarantees sufficient exploration.

Note we only get a statistical estimate of the transition model. Fortunately, it turns out that a good approximation of the true MDP M is sufficient to find a good policy for M. This property is often referred to as the simulation lemma.

Simulation Lemma (crude statement): $M = (S, A, P, R, \gamma)$, $\hat{M} = (S, A, \hat{P}, R, \gamma)$
$\hookrightarrow$ If $\forall s, a$ $d(P(.|s,a), \hat{P}(.|s,a)) \leq \epsilon$ then $d(V_M^\pi, V_{\hat{M}}^\pi) \leq f(\epsilon)$ and $f(\epsilon) \to 0$ as $\epsilon \to 0$

E^3 (Kearns and Singh) is more explicit about exploration versus exploitation.