CS 497 SML: Week 6 Notes (Feb. 25, 27)

Scribe: M. M. Hassan Mahmud (mmmahmud@uiuc.edu)

1 Stochastic Iterative Algorithm

Last week we finished with an introduction to Reinforcement Learning. We also looked at the **Stochastic Iterative Aglorithm** that allowed us to compute the cost function J in the absence of our knowledge of the state transition probabilities $P(x'|x,u)$. So to compute the cost function J_γ for a strategy γ at state x the algorithm performs the action $\gamma(x)$ in a Monte Carlo simulator of the world and obtains the next state x'. It then updates the the current estimate $\hat{J}_\gamma(x)$ of $J_\gamma(x)$ by blending togather the 1) the current value of $\hat{J}_\gamma(x)$ and the 2) sum of the immediate loss $l(x,\gamma(x))$ obtained from the simulator. Specifically, the following equation is used to update the current estimate $\hat{J}_\gamma(x)$

$$\hat{J}_\gamma(x) = (1-\rho)\hat{J}_\gamma(x) + \rho(l(x,\gamma(x)) + \alpha\hat{J}_\gamma(x')) \tag{1}$$

where the $\rho \in (0,1)$ parameter determines the blending of the two terms. The fact that $\hat{J}_\gamma(x)$ convergest to $\hat{J(x)}$ has been shown by Robbins-Monro. The theorem that proves this is known as the Robbins Monro theorem (Robbins and Monro, 1951).

The version of stocahastic iterative algorithm presented here is actually a more general version of the problem of computing the value of a variable y given the noisy observations $h(y)$ of y. The basic equation follows the same format:

$$\hat{y} = (1-\rho)\hat{y} + \rho h(y) \tag{2}$$

Equation 1 works because the Monte Carlos simulator provides us with the probabilities necessary. So by applying the equation repeatedly, in the limit the equation converges to the actual value of the J_γ

2 Finding Optimal Strategy

So given the Monte-Carlo simulator, how can we find the optimal strategy ? The answer is *Q-leraning*. The basic idea in Q-learning is to find the value or cost to go associated with each state-action pair (x,u) rather than a state. Thus the optimal J^* would correspond to choosing for each state the action that has the lowest Q-value $Q(x,u)$. Thus the optimal Q function, $Q^* : X \times U \to \Re$ is defined as follows:

$$Q^*(x,u) = l(x,u) + \alpha \sum_{x' \in X} (P(x'|x,u) \min_{u' \in U(x')} Q(x',u')) \tag{3}$$

From the above equation we can easily see that $J^*(x) = \min_{u \in U(x')} Q^*(x,u)$ and $\gamma^*(x) = arg\min_{e' \in U(x')} Q^*(x,u)$. Note that in case the the state-action-state transition probabilities are unknown, we can use the stochasitc iterative version of Q-learning:

$$\hat{Q}^*(x,u) = (1-\rho)\hat{Q}^*(x,u) + (\rho)l(x,u) + \alpha \min_{u' \in U(x')} \hat{Q}^*(x',u')) \tag{4}$$

where ρ is the learning rate, as defined above.

3 Imperfect State Information

3.1 States vs. Observation

Now we consider what happens when we do not know what the current state x is, but rather have some observation (which may or may not be of the current state). Examples of such observations include sensor readings for a robot etc. We denote the observations made at step k by y_k. To make things more interesting, we also assume that nature interferes with our observations by performing some kind of *observation action*. That is an action which alters our observation in some way. We denote the set of all such observations actions

by Φ and the observation action at step k by ϕ_k. Given this notation, we can formally state the relationship between the observation, observation action and the state by the following observation equation:

$$y_k = h(x_k, \phi_k) \tag{5}$$

where h is some function. Note that earlier we had said that the observation made may not be of the current state, which seems to contradict the above equation. However, everything works out fine either way so to keep things simple we assume that y_k depends only on the current state.

So from equation 5 we have two kinds of uncertainty.

- Projection which corresponds $y_k = h(x_k)$
- Distrubance which corresponds to $y_k = h(x_k, \phi_k)$.

The first type of uncertainty is called projection, because the original state vector x_k in some space is being projected onto a new vector y_k with possibly different number of elements. The second type of uncertainty is called disturbance because since ϕ_k is unknown. So given this scenario, our aim is to make the right decision. As we shall see this is actually tractable in contrast to the task of recovering x given our lack of knowledge of Φ.

3.2 Making Decisions using Observations : Information Spaces

In order to make decisions using observations, we first need to take tally of the kinds of information we have or allow ourselves to have. These are as follows:

- The initial condition **(IC)** : the inital state x_1 or the set of initial states, $X_1 \subset X$ and the prior over X_1, $P(x_1)$.
- The observation history: $y_1, y_2, \ldots, y_k$
- The action history $u_1, u_2, \ldots, u_{k-1}$.

So given the above information, we can now make decision based on an information state defined by all the information given above. Thus the information state at step k, η_k, is defined as follows:

$$\eta_k = \{IC, u_1, u_2, \ldots, u_{k-1}, y_1, y_2, \ldots, y_k\} \tag{6}$$

Thus η_k contains all the information that could possibly be used to make a decision. We denote the set of all such possible η_ks by N_k and we call N_k the *information space*. The dimension of N_k is given by : $dim(N_k) = k(dim(Y)) + (k-1)(dim(U)) + dim(IC)$. As $k \to \infty$, $dim(N_k) \to \infty$.

3.3 Strategy in an Information Space

We can define strategies for information space states in the same way as in the case of perfect state information.

- Strategy with perfect state informaiton : $\gamma : X \to U$
- Strategy with imperfect state information : $\gamma : N_k \to U$.

Note that in the case of imperfect state information, it is difficult to say which actions are available since the state is unknown. So in this case we make the simplifying assumption that $U(x) = U \ \forall x \in X$. If we specify $\gamma_k \ \forall k$ we specify exactly what choice the decision maker makes in all possible futures.

4 Manipulating the Information Space

We have 4 possible ways of dealing the size of the information space. These are as follows:

- Limit memory (i.e. forget part of the history)
- Be Bayesian
- Do the non-deterministic equivalence of being Bayesian
- Approximate N_k
- Find equivalence classes in N_k (sometimes defined by the problem itself).

We describe first three of these below.

4.1 Limit Memory

In this we only remember the last i stages. So if $\eta_k = \{IC, u_1, u_2, \ldots, u_{k-1}, y_1, y_2, \ldots, y_k\}$ then if we set i to 1, then $\eta_k = \{IC, u_{k-1}, y_k\}$ or if we go one step further, $\eta_k = y_k$.

4.2 Bayesian Analysis

To manipulate the information space using Bayesian methods, first we make note of the kind of information that we have available or allow ourselves to have. We have the recent observation y, we assume we have $P(\phi)$ and we also know the observation function $y = h(x, \phi)$. From the above three, we can compute $P(y|x)$. Assuming we have the state transitions and the corresponding observations as shown in figure 1, we can the use Baye's rule to obtain the following:

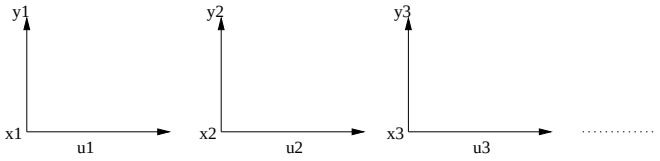

Figure 1: **State Transitions and Observations**

$$P(x_1|y_1) = \frac{P(y_1|x_1)P(x_1)}{\sum_{x_i \in X} P(y_i|x_i)P(x)} \tag{7}$$

We can now apply u_1 to obtain x_2 :

$$P(x_2|y_1, u_1) = \sum_{x_1} P(x_2|x_1, u_1)P(x_1|y_1) \tag{8}$$

Note that in equation 8 x_2 does not depend on y_1, since x_2 is conditionally independent of y_1 given u_1.

Now we would like to determine the more general version of the the probability given in equation 8 which is, given $P(x_k|\eta_k)$ compute $P(x_{k+1}|\eta_{k+1})$, u_k and y_{k+1} where $\eta_{k+1} = \eta_k \cup \{u_k, y_{k+1}\}$

First we handle applicataion of u_k,

$$P(x_{k+1}|\eta_k, u_k) = \sum_{x_k} P(x_{k+1}|\eta_k, x_k, u_k)P(x_k|\eta_k) \tag{9}$$

Assuming x_{k+1} is conditionally independent of η_k given x_k, u_k and x_k is conditionally independent of u_k given η_k, we have:

$$P(x_{k+1}|\eta_k, u_k) = \sum_{x_k} P(x_{k+1}|x_k, u_k)P(x_k|\eta_k) \tag{10}$$

Now we handle y_{k+1}.

$$\begin{aligned} P(x_{k+1}|\eta_k, u_k, y_{k+1}) &= P(x_k|\eta_{k+1}) \\ &= \frac{P(y_{k+1}|x_{k+1})P(x_k+1|\eta_k, u_k)}{\sum_{x_{k+1}} P(y_{k+1}|x_{k+1})P(x_{k+1}|\eta_k, u_k)} \end{aligned} \tag{11}$$

4.3 Non-deterministic approach

In the non-deterministic approach we deal with sets rather than distributions. So instead of dealing with probability of the actual state being a particular state x_k given an observation y_k, η_k, x_k, we deal with the set of all possible states S_k that the actual state can be.

Let $X_1 \subseteq X$ denote our possible set of initial states. Let $S_k(\sim) \subseteq X$ denote the possible set of states given the information $\sim$. Thus in this case $S_k(\sim)$ is the Bayesian equivalent of $P(x_k| \sim)$. Now given $S_k(\eta_k)$ we want to find $S_{k+1}(\eta_{k+1})$.

We start with the initial set of states given by $S(I.C.) = X_1$. Let us also assume that we are given Φ, the set of observations from nature, and from this we can get $S_k(y_k) \subseteq X$, where

$$S_k(y_k) = \{x_k \in X | \exists \phi_k \in \Phi, x_k = h(y_k, \phi_k)\} \tag{12}$$

Now, we recieve y_1 and get $S(IC, y) = S(y_1) \cap X_1$. Applying u_1, we get,

$$S_2(u_1, \eta_1) = \{x_2 \in X | \exists \theta_1 \in \Theta, \exists x_1 \in S_1(\eta_1), x_2 = f(x_1, u_1, \theta_1)\} \tag{13}$$

So the above equation is saying that S_2 the set of all states that can be reached given action u_1 was applied and that nature took action θ_1. Equation 12 can also be written as:

$$S_2(u_1, \eta_1) = \bigcup_{\theta_1 \in \Theta} \bigcup_{x_1 \in S_1(\eta_1)} f(x_1, u_1, \theta_1) \tag{14}$$

It is clear that in general this procedure can be applied to n steps and then the following general form of the update equation is derived. First we apply u_k,

$$S_{k+1}(u_k, \eta_k) = \bigcup_{\theta_k \in \Theta} \bigcup_{x_k \in S_k(\eta_k)} f(x_k, u_k, \theta_k) \tag{15}$$

and recieve y_{k+1}

$$S_{k+1}(\eta_k) = S_{k+1}(\eta_k, u_k, y_{k+1}) \cap S_{k+1}(y_{k+1}) \tag{16}$$