Lecture 7 Notes

In Lecture 7 we covered *policy evaluation* for stochastic dynamic programs. This is an important first step toward finding optimal policies for stochastic dynamic programs. We first define the expected value to go at stage t , given history h_t, and policy π as follows:

$$v_t^\pi(h_t) = E_{h_t}^\pi[\sum_{n=t}^{N-1} \lambda^{n-1} r_n(X_n,\, Y_n) + r_N(X_N)]$$

Next, the following algorithm was proposed to compute $v_t^\pi(h_t)$ for all t and h_t:

Algorithm (Policy Evaluation):

1. Set $\mathrm{t} = \mathrm{N}, v_N^\pi(h_N) = r_N(s_N),\ for\ all\ h_N$
2. $If\ t = 1\ stop,\ otherwise\ go\ to\ step\ 3$
3. $Substitute\ t-1\ for\ t\ and\ compute\ v_t^\pi(h_t),\ for\ all\ h_t,\ as$:

$$v_t^\pi(h_t) = \mathrm{r_t}\big(\mathrm{s_t},\, \mathrm{d_t(h_t)}\big) + \lambda \sum_{j \in S} p_t\big(j \big| s_t,\, d_t(h_t)\big) v_{t+1}^\pi(h_{t+1})$$

$return\ to\ step\ 2$

The appropriateness of the algorithm relies on whether $v_t^\pi(h_t)$ generated by the algorithm does in fact equal the expected future reward to go at stage t, state h_t, under policy π. The following theorem is similar to Theorem 4.2.1 of Puterman.

Theorem (≈4.2.1 Puterman): For $\pi \in \Pi^{HD}$ at each stage t the policy evaluation algorithm generates $\mathrm{v_t^\pi(h_t)}$ for all h_t and $v_1^\pi(s) = u_N^\pi(s)$ for all $\mathrm{s} \in S$.

Proof: The proof can be completed by induction. For $t = N\ v_N(h_N) = r_N(s_N)$ for all s_N by definition. This represents the *base case* for the induction proof. Now, for *the induction hypothesis*, assume that the policy evaluation algorithm generates $v_n^\pi(h_n)$ for all h_n at stages $n = t+1, t+2, \dots, N$, then it follows that

$v_t^\pi(h_t) = r_t\big(s_t, d_t(h_t)\big) + E_{h_t}^\pi[E_{h_{t+1}}^\pi[\sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N)]]$ (by the induction hypothesis)

$= r_t\big(s_t, d_t(h_t)\big) + E_{h_t}^\pi[\sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N)]]$ (since $H_{t+1} \subset H_t$)

$= E_{h_t}^\pi[\sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N)]$ (since s_t and d_t are known at t)

Note: The above proof implicitly assumed linear additive rewards.