Lecture 7 Policy Evaluation Example

Lecture 7 covered policy evaluation including a "forward approach" in which all sample paths were enumerated and a more efficient "backward approach" based on the *policy evaluation* algorithm below.

Definition: expected *value to go*, at stage t , given history h_t, and policy π:

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

Note that this is not the *optimal value to go* since we are evaluating a fixed policy that may or may not be an optimal policy.

Algorithm (Policy Evaluation):

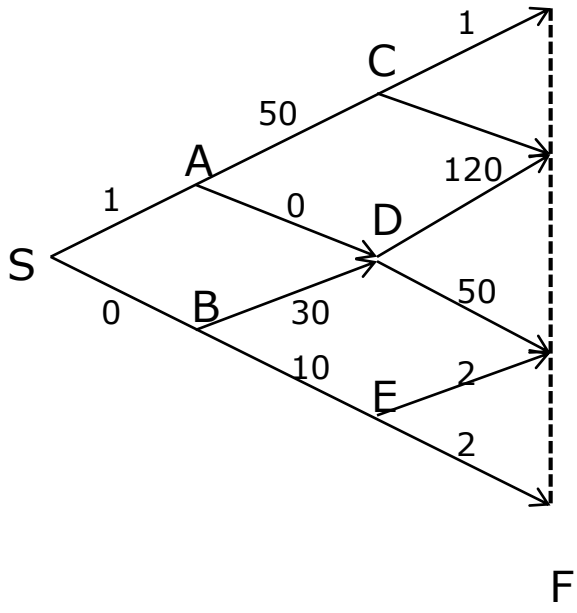
1. Set $\mathrm{t} = \mathrm{N}, v_N^{\pi}(h_N) = r_N(s_N),\ for\ all\ histories\ at\ stage\ N,\ h_N$
2. $If\ t = 1\ stop,\ otherwise\ go\ to\ step\ 3$
3. $Substitute\ t-1\ for\ t\ and\ compute\ v_t^{\pi}(h_t),\ for\ all\ histories\ at\ stage\ t,\ h_t,\ as$:

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

$return\ to\ step\ 2$

Example:

Consider the shortest path across the following network:



Unlike the deterministic shortest path problem, in this case there is uncertainty in the state transitions. At each node the decision maker may choose from two actions: move "up" or "down." However, there is uncertainty in the next state the decision maker will arrive at. The probability of actually going up if the action chosen is up is $\Pr(u|a=u)=0.8$. Similarly, the probability of actually going down if the action chosen is down is $\Pr(d|a=d)=0.7$. The policy we will consider is the following:

Policy π: At vertex S go up. At all future ***states*** go up if you have ever gone up in the past; otherwise go down.

We can now use the policy evaluation algorithm to find the expected cost of traversing the network under this policy, $v_1^\pi(S)$.

Problem setup:

Assume $\lambda = 1$ since it was not specified in the problem statement.

$$r_t(s_t,a_t)=\begin{cases}\Pr(u|a=u)\,L^u+\Pr(d|a=u)\,L^D \; if\; a_t=u\\ \Pr(d|a=d)\,L^d+\Pr(u|a=d)\,L^u \; if\; a_t=d\end{cases}$$

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

Policy Evaluation:

Iteration 1:

Step 1: $t=4, v_4^\pi(h_4)=0,\ for\ all\ h_4$

Step 2: $t\neq 1, therefore\ go\ to\ step\ 3$

Step 3: $t=3, h_3\in\{[S,A,C],\ [S,A,D],\ [S,B,D],\ [S,B,E]\}$

$v_3^\pi([S,A,C])=0.8(1)+(1-0.8)(0)=8$

$v_3^\pi([S,A,D])=0.8(1200)+(1-0.8)(500)=1060$

$v_3^\pi([S,B,D])=0.8(1200)+(1-0.8)(500)=1060$

$v_3^\pi([S,B,E])=0.7(20)+(1-0.7)(20)=20$

Iteration 2:

Step 2: $t\neq 1, therefore\ go\ to\ step\ 3$

Step 3: $t=2, h_2\in\{[S,A],\ [S,B]\}$

$v_2^\pi([S,A])=0.8(500)+(1-0.8)(0)+0.8v_3^\pi([S,A,C])+0.2v_3^\pi([S,A,D])=618.4$

$v_2^\pi([S,B])=0.7(100)+(1-0.7)(300)+0.7v_3^\pi([S,B,E])+0.3v_3^\pi([S,B,D])=492.0$

Iteration 3:

Step 2: $t \neq 1, therefore\ go\ to\ step\ 3$

Step 3: $t = 1, h_1 \in \{[S]\}$

$$v_1^{\pi}([S]) = 0.8(10) + 0.3(0) + 0.8v_2^{\pi}([S,A]) + 0.2v_2^{\pi}([S,B]) = 601.12$$

The algorithm terminates in step 2 of the next iteration since t = 1. The expected path length for policy π is 601.12. This is the mean of a probability distribution define by the cost of all paths that could be followed (see Lecture 7 notes for the forward approach that led to the same answer but with more computation).