MATH(4/7)406 (Control Theory)
Quiz 2 *Solution* (Unit 3) - September 9, 2014.

Prepared by Yoni Nazarathy

Quiz duration: 40 minutes.

Name:

Student ID:

Consider 3 distinct numbers, x_1, x_2, x_3. These numbers are shuffled randomly and the resulting sequence is presented to you one after the other. E.g you may first get to see x_3 then x_1 and then x_2. Any of the 6 permutations is equally likely.

Your goal is to say "stop" when the smallest number is presented to you. You only get to say "stop" once; after that no more numbers are presented to you. If you say "stop" when presented with the smallest number you don't lose or gain anything. Otherwise (if you say your "stop" at the wrong time) you pay 100 dollars.

Use finite horizon dynamic programming to find a policy that minimizes your expected losses. The policy should be in the form: $\pi = (d_1, d_2, d_3)$ where,

$$d_t : X_t \to \{\text{"stop"}, \text{ "continue"}\},$$

where,

$$X_t = \begin{cases} 1, & \text{if the number presented at time } t \text{ is smallest so far,} \\ 0, & \text{otherwise.} \end{cases}$$

It is obvious that $d_3(1) =$ "stop" and $d_3(0)$ can be either "stop" or "continue" (in any case you'll lose 100 dollars if $X_3 = 0$ and if you haven't said "stop" earlier).

Answer the following (in any order that suits you):

a: Write a precise MDP formulation of the problem (time set, state space, action sets, rewards, transitions, objective).

b: What are optimal decision rules $d_1(\cdot), d_2(\cdot)$?

c: What is the expected cost with the policy that you found?

d: If you wish, verify (e.g. by enumerating all possibilities) that your answers to either (b), (c) or both are correct. This is without credit, but do it if you have time.

Solution:

This problem is almost exactly the "secretary problem" as presented in Section 4.6.4 of [Put94] and handled in HW3, Problem 4. Here the "best candidate" has value $\min\{x_1, x_2, x_3\}$.

There are two outcomes to this problem: "Stop correctly" and "Stop incorrectly". For a given a policy, π, denote the probability of "stopping correctly" by p^π, then the expected losses are, $100(1 - p^\pi)$. Minimization of this quantity is like maximization of p^π. Hence from this point on, the problem is exactly as in Section 4.6.4 with $N = 3$.

To model the problem as an MDP, we simply consider the problem of maximizing p^π. An alternative solution would be to maximize $-100(1 - p^\pi)$. The difference will be with rewards.

(a) (20pts):

This is a "possible solution". There are others too.

$$T = \{1, 2, 3\}.$$

$$S = \{0, 1, \Delta\}.$$

Here $s = 0$ implies the current candidate is not the best observed so far; $s = 1$ implies that the current candidate is the best observed so far; $s = \Delta$ implies we have stopped (game over).

For the action sets, there are several options. Here is one:

$$A_0 = A_1 = \{\text{stop, continue}\}, \qquad A_\Delta = \{\text{don't care}\}.$$

The (expected rewards) are generally time-dependent. First, $r_t(0, a) = r_t(\Delta, a) = 0$ for $t = 1, 2, 3$ and any action a. Further, $r_t(1, \text{continue}) = 0$ for $t = 1, 2, 3$. Finally,

$$r_1(1, \text{stop}) = \frac{1}{3}, \quad r_2(1, \text{stop}) = \frac{2}{3}, \quad r_3(1, \text{stop}) = 1.$$

Note that the above rewards are for the problem of maximizing p^π.

As in any stopping problem, the transition probabilities between states 0 and 1 do not depend on the action. For $t = 1, 2$:

$$p_t(0|0) = \frac{t}{t+1}, \quad p_t(1|0) = \frac{1}{t+1}.$$

Further (for $t = 1, 2$):

$$p_t(0|1) = \frac{t}{t+1}, \quad p_t(1|1) = \frac{1}{t+1}.$$

Finally, $p_t(\Delta|s, a) = 1$ if $a = \text{stop}$ and is 0 otherwise.

(b + c) (40pts + 40pts):

As in [Put94], Denote $u_t^*(1)$ to be the maximum probability of choosing the best candidate when $X_t = 1$. Further denote $u_t^*(0)$ to be the maximum probability of choosing the best candidate when $X_t = 0$.

Hence $u_3^*(1) = 1$ and $u_3^*(0) = 0$. The optimality equations for $t = 1, 2$ are,

$$u_t^*(1) = \max\big\{\frac{t}{3},\ \frac{1}{t+1}u_{t+1}^*(1) + \frac{t}{t+1}u_{t+1}^*(0)\big\}, \tag{1}$$

$$u_t^*(0) = \max\big\{0,\ \frac{1}{t+1}u_{t+1}^*(1) + \frac{t}{t+1}u_{t+1}^*(0)\big\}. \tag{2}$$

By noticing that the right element in the maximum of $u_t^*(0)$ is non-negative, these simplify to:

$$u_t^*(0) = \frac{1}{t+1}u_{t+1}^*(1) + \frac{t}{t+1}u_{t+1}^*(0),$$
$$u_t^*(1) = \max\big\{\frac{t}{3},\ u_t^*(0)\big\}.$$

Now that all that remains is to find the solution using backward induction. First for $t = 2$:

$$u_2^*(0) = \frac{1}{3}u_3^*(1) + \frac{2}{3}u_3^*(0) = \frac{1}{3},$$
$$u_2^*(1) = \max\big\{\frac{2}{3},\ u_2^*(0)\big\} = \frac{2}{3},$$

moving to $t = 1$:

$$u_1^*(0) = \frac{1}{2}u_2^*(1) + \frac{1}{2}u_2^*(0) = \frac{1}{2},$$
$$u_1^*(1) = \max\big\{\frac{1}{3},\ u_1^*(0)\big\} = \frac{1}{2}.$$

One way to see the decision rule is to look at the original equations (1) and (2) with the values substituted in and see which action is maximizing. For $t = 1$:

$$u_1^*(1) = \max\big\{\frac{1}{3},\ \frac{1}{2}u_2^*(1) + \frac{1}{2}u_2^*(0)\big\} = \max\big\{\frac{1}{3},\ \frac{1}{2}\big\},$$
$$u_1^*(0) = \max\big\{0,\ \frac{1}{2}u_2^*(1) + \frac{1}{2}u_2^*(0)\big\} = \max\big\{0,\ \frac{1}{2}\big\}.$$

Hence $d_1(1)$ = "continue" and $d_1(0)$ = "continue".

For $t = 2$:

$$u_2^*(1) = \max\big\{\frac{2}{3},\ \frac{1}{3}u_3^*(1) + \frac{2}{3}u_3^*(0)\big\} = \max\big\{\frac{2}{3},\ \frac{1}{3}\big\},$$
$$u_2^*(0) = \max\big\{0,\ \frac{1}{3}u_3^*(1) + \frac{2}{3}u_3^*(0)\big\} = \max\big\{0,\ \frac{1}{3}\big\}.$$

Hence $d_2(1)$ = "stop" and $d_3(0)$ = "continue".

So an optimal policy is to continue on the first number, then if the second number is greater than the first number stop, otherwise wait for the third number.

The value of the MDP is $u_1^*(1) = \frac{1}{2}$. I.e. $p^{\pi^*} = \frac{1}{2}$.

So the expected cost is 50**.**

(d) (0pts):

This can be checked simply by looking at all permutations assume the numbers are $x_i = i$.

$$(1,2,3) \to \text{lose}$$
$$(1,3,2) \to \text{lose}$$
$$(2,1,3) \to \text{win}$$
$$(2,3,1) \to \text{win}$$
$$(3,1,2) \to \text{win}$$
$$(3,2,1) \to \text{lose}$$

Indeed we win in half of the cases.