Stochastic Assignment 4

Andrew Martin

October 16, 2018

Question 1. Marks for mathematical writing

Question 2. Each day, Squirrel the Singer is asked to take on a new singing gig. The gigs are independently distributed over 3 possible types; on a given day, the offered type is i with probability $\alpha_i \in (0,1)$ for $i = 1, \dots, 3$. Upon completion, gigs of type i pay r_i dollars. Once Squirrel has accepted a gig, she may accept no other gigs until that gig is complete. The probability that a gig of type i takes k days is $(1-p_i)^{k-1} p_i$ for $k = 1, 2, \dots$ where $p_i \in (0,1)$

(a) Evaluate the average reward, g_1, of a stationary policy in which Squirrel accepts only gigs of type 1.

Solution Firstly obtain transition probability matrix - let the states 1 and 2 be waiting for a gig and playing the gig respectively.

$$\mathbf{P}_1 = \begin{pmatrix} 1-\alpha_1 & \alpha_1 \\ p_1 & 1-p_1 \end{pmatrix}$$

And the corresponding reward matrix (assume that payment is given after the gig is completed)

$$\mathbf{r}_1 = \begin{pmatrix} 0 & 0 \\ r_1 & 0 \end{pmatrix}$$

Find the equilibrium, π, such that $\pi \mathbf{P}_1 = \pi$, where $\pi \mathbf{1} = 1$
Gives the system of equations:

$$\begin{aligned} (1-\alpha_1)\pi_1 + p_1 \pi_2 &= \pi_1 \\ \pi_1 \alpha + (1-p_1)\pi_2 &= \pi_2 \\ \pi_1 + \pi_2 &= 1 \end{aligned}$$

Which gives

$$\pi_1 + \frac{\alpha_1}{p_1}\pi_1 = 1 \implies \pi_1 = \frac{p_1}{\alpha_1 + p_1}, \quad \pi_2 = \frac{\alpha_1}{\alpha_1 + p_1}$$

Calculate expected reward for each state q_i (where p_{ij} and r_{ij} are the i, j^{th} elements of $\mathbf{P}_1$ and $\mathbf{r}_1$ respectively)

$$\begin{aligned} q_i &= \sum_{j=1}^{2} p_{ij} r_{ij} \\ \implies q_1 &= 0 \\ q_2 &= r_1 p_1 \end{aligned}$$

Lastly, calculate the average reward, g_1

$$\begin{aligned} g_1 &= \sum_{i=1}^{2} \pi_i q_i \\ &= \frac{p_1}{\alpha_1 + p_1} 0 + \frac{\alpha_1}{\alpha_1 + p_1}(r_1 p_1) \\ \therefore g_1 &= \frac{\alpha_1 r_1 p_1}{\alpha_1 + p_1} \end{aligned}$$

As required.

(b) Apply one step of the Policy Improvement Algorithm to determine an improved policy, clearly stating what the improved policy is.
Hint: only have to consider policies for which only one type of gigs is considered - e.g. only accepting gigs of type 2.

Solution

Step 1. Start with the policy to only accept type 1 gigs

Step 2. Solve:

$$\phi(i) + g = \sum_{j=1}^{n} p_{ij} r_{ij} + \sum_{j=1}^{n} p_{ij}\phi(j)$$

And set $\phi(1) = 0$ Which gives

$$g = \alpha_1 \phi(2) \tag{1}$$

$$\phi(2) + g = p_1 r_1 + (1 - p_1)\phi(2) \tag{2}$$

$$\implies \phi(2) = \frac{p_1 r_1}{\alpha_1 + p_1}$$

The right hand side(s) become

$$RHS_1(1) = \frac{p_1 r_1 \alpha_1}{\alpha_1 + p_1}$$

$$RHS_2(2) = p_1 r_1 + \frac{(1 - p_1) p_1 r_1}{\alpha_1 + p_1}$$

Step 3. Try a new policy and see if it improves the right hand side of equations 1 and 2:

$$RHS_2(i) = \sum_{j=1}^{n} p'_{ij} r'_{ij} + \sum_{j=1}^{n} p'_{ij}\phi(j)$$

Note that the new policy is identical apart from replacing α_1 with α_2 and the same for p and r. Hence we will get the gain, g_2 is

$$g_2 = \frac{p_2 r_2 \alpha_2}{\alpha_2 + p_2}$$

Step 4. If this policy is an improvement go to step 2, and set $p = p'$ and $r = r'$. Otherwise stop

This policy is an improvement if

$$g_2 > g_1, \quad \implies \frac{p_2 r_2 \alpha_2}{\alpha_2 + p_2} > \frac{\alpha_1 r_1 p_1}{\alpha_1 + p_1}$$

If this holds, we return to step 2 with the new parameters
If this does not hold, then our current policy is the best policy.

As required.

Question 3. Heather receives \$10 for every chess game that she wins. Playing costs her \$$c$ per hour. The total number of chess games that Heather can play is T. The probability of winning one game in the next hour is $\omega(r)$, where $\omega(r)$ is an increasing function of r, the remaining number of games. There is zero probability of winning more than one game in an hour. Heather wants to maximise her net expected profit.
Hint: Heather has zero probability of losing (or drawing) a game.

(a) Specify, with justification, Heathers stopping rule.

Solution Heather should stop after T games or after the expected reward from the next game is non-positive. T games is trivial as Heather can no longer play after that point.
We only care about the reward from the next game since $\omega(r)$ gets smaller as r goes to 0. I.e. if the expected reward from the next game is a, then the expected reward from the game after that is less than a.
We have two options, play or stop.

Stop after the t^{th} game if

$$E[r(t+1)] \leq 0$$

Where $r(t)$ is the reward for the tth game

$$r(t) = 10 - c(\textit{length of game})$$

Note that the time until winning a game is geometrically distributed with parameter $\omega(t)$, and the mean value of $Geom(\omega(t))$ is $\frac{1}{\omega(t)}$

$$E[r(t)] = 10 - \frac{c}{\omega(T-t)}$$

I.e. Heather should stop when either event: number of games played, $t > T$ or $10 - \frac{c}{\omega(T-t)} \leq 0$ occurs. **As required.**

(b) If $T = 12$, $\omega(r) = 1 - e^{-r/5}$ and $c = \$0.5$, determine Heathers expected profit and detail the stopping rule.

Solution Want to stop when $10 - \frac{c}{\omega(T-t)} \leq 0$, i.e. stop for the t^{th} game:

$$\begin{aligned}
10 - \frac{0.5}{1 - e^{-(T-t)/5}} &\leq 0 \\
10(1 - e^{-(T-t)/5}) &\leq 0.5 \\
1 - e^{-(T-t)/5} &\leq 0.05 \\
e^{-(T-t)/5} &\geq 0.95 \\
-(T-t)/5 &\geq \log(0.95) \\
-T + t &\geq 5\log(0.95) \\
t &\geq 5\log(0.95) + T \\
t &\geq 12 \text{ (discrete time)}
\end{aligned}$$

I.e. we do not play 12 or more games

Expected profit after 11 games:

$$\begin{aligned}
E(profit) &= \sum_{t=1}^{11} E(r(t)) \\
&= \sum_{t=1}^{11} \left(10 - \frac{c}{\omega(T-t)}\right) \\
&= 110 - \sum_{t=1}^{11} \frac{0.5}{1 - e^{-(12-t)/5}} \\
&\approx \$99.17
\end{aligned}$$

As required.

Question 4. You are moving overseas soon! Suppose you need to sell your car (a twenty-year-old aqua Mirage) and have 10 weeks in which to advertise and sell it. You receive one offer per week; these offers are independent with a value of j dollars with probability p_j, for $j = 1, \ldots, 10$. Any offer not immediately accepted, can be accepted at a later date. Every week that the Mirage remains unsold, it costs you $\$c$ dollars per week.

The state space is $S = \{1, 2, ..., 10\}$, where state i corresponds to the highest offer to date. There are only two actions you might take when in state i, to either accept the best offer to date with value i or not accept the best offer to date and continue with costs c.

(a) Give the transition probabilities when continuing on and not accepting the best offer to date, with justification.

Solution This gives a 10×10 matrix. In short, I will always take the better deal out of my current stored deal and the next offer. This gives the transition probabilities:

$$p_{ij} = \begin{cases} p_j, & j > i \\ \sum_{k=1}^{i} p_k, & j = i \\ 0, & j < i \end{cases}$$

Or in matrix form:

$$P = \begin{pmatrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} \\ 0 & \sum_{i=1}^2 p_i & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} \\ 0 & 0 & \sum_{i=1}^3 p_i & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} \\ 0 & 0 & 0 & \sum_{i=1}^4 p_i & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} \\ 0 & 0 & 0 & 0 & \sum_{i=1}^5 p_i & p_6 & p_7 & p_8 & p_9 & p_{10} \\ 0 & 0 & 0 & 0 & 0 & \sum_{i=1}^6 p_i & p_7 & p_8 & p_9 & p_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & \sum_{i=1}^7 p_i & p_8 & p_9 & p_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sum_{i=1}^8 p_i & p_9 & p_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sum_{i=1}^9 p_i & p_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sum_{i=1}^{10} p_i \end{pmatrix}$$

The interpretation is if I already have an offer for $\$i$, with probability p_j, I will get an offer for j dollars. My options are to either stick with my $\$i$ offer, or to accept the new $\$j$ offer. I will accept it if $j > i$, which will occur with probability p_j. This is the transition p_{ij} for $j > i$.

If $j \le i$ I want to stay with my current offer, which is the transition p_{ii}. The probability of this occurring is the sum of all the probabilities corresponding to all of the values lower than my current offer, i.e. $\sum_{j=1}^i p_j$.

Since I will never move to a worse offer than my current offer, all of the p_{ij}, $j < i$ are 0, as i will never make these transitions. **As required.**

(b) What is the optimal policy for selling your car?

Solution Want to maximise expected profit.

On any given step we have $\$k$ (possibly after a week) we want to get

$$\max\{k, E[\textit{value next week}] - c\}$$

I.e. the optimal policy is to find a threshold k to accept. I will denote

$$v := E[\textit{value next week}]$$

And note $v \ge k$, since for any offer $< k$, we keep k. This gives:

$$v = \sum_{i=1}^{k} p_i k + \sum_{i=k+1}^{10} p_i i$$

If our first offer is $\$k$, we want to accept k if:

$$\begin{aligned} k &\ge v - c \\ k &\ge \sum_{i=1}^{k} p_i k + \sum_{i=k+1}^{10} p_i i - c \\ k &\ge k - \sum_{i=k+1}^{10} p_i k + \sum_{i=k+1}^{10} p_i i - c \quad \text{(LOTP)} \\ 0 &\ge \sum_{i=k+1}^{10} p_i (i-k) - c \\ c &\ge \sum_{i=k+1}^{10} p_i (i-k) \end{aligned}$$

I.e. we should accept the offer \$$k$ if

$$c \geq \sum_{i=k+1}^{10} p_i(i-k)$$

Or in an expanded form:

$$c \geq \begin{cases} p_2 + 2p_3 + 3p_4 + 4p_5 + 5p_6 + 6p_7 + 7p_8 + 8p_9 + 9p_{10}, & k = 1 \\ p_3 + 2p_4 + 3p_5 + 4p_6 + 5p_7 + 6p_8 + 7p_9 + 8p_{10}, & k = 2 \\ p_4 + 2p_5 + 3p_6 + 4p_7 + 5p_8 + 6p_9 + 7p_{10}, & k = 3 \\ p_5 + 2p_6 + 3p_7 + 4p_8 + 5p_9 + 6p_{10}, & k = 4 \\ p_6 + 2p_7 + 3p_8 + 4p_9 + 5p_{10}, & k = 5 \\ p_7 + 2p_8 + 3p_9 + 4p_{10}, & k = 6 \\ p_8 + 2p_9 + 3p_{10}, & k = 7 \\ p_9 + 2p_{10}, & k = 8 \\ p_{10}, & k = 9 \\ 0, & k = 10 \end{cases}$$

On the second week *all* offers (including that which we received) are effectively reduced by \$$c$. So this formula will still work I.e. for the next week, we are choosing

$$\max\{k - c, E[valuenextweek] - 2c\} = \max\{k, E[valuenextweek] - c\} - c$$

So the relationship between k and c still holds in general

As required.