Dry2 - 046203

Zohar Rimon

22.05.2020

1 Properties of the transition matrix

1.1

$$P_{i,j} = p(x_{t+1} = j|x_t = i) \geq 0$$

Thus, obviously, all of the values in the matrix are non-negative and the sum of the rows (considering $|S| = N$):

$$\sum_{j=1}^{N} P_{i,j} = \sum_{j=1}^{N} p(x_{t+1} = j|x_t = i) = 1$$

This is due to the fact that we sum the probability over all the possible values of x_{t+1}, and this is normalized to 1.

1.2

The transition matrix always has an eigenvector with a corresponding eigenvalue of 1 which is the unit vector:

$$P \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{N} 1 \cdot p(x_{t+1} = j|x_t = i) \\ \sum_{j=1}^{N} 1 \cdot p(x_{t+1} = j|x_t = i) \\ \vdots \\ \sum_{j=1}^{N} 1 \cdot p(x_{t+1} = j|x_t = i) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

1.3

Let v an eigenvector with a corresponding eigenvalue of λ:

$$(P \cdot v)_i = \sum_{j=1}^{N} p(x_{t+1} = j|x_t = i) \cdot v_j = \lambda v_i$$

Let $|v_k| = \max\left[|v_1|, |v_2[, \dots |v_N|\right]$, thus:

$$(P \cdot v)_k = \sum_{j=1}^{N} p(x_{t+1} = j|x_t = k) \cdot v_j = \lambda v_k$$

And taking the absolute value:

$$\left|\sum_{j=1}^{N} p(x_{t+1} = j|x_t = k) \cdot v_j\right| = |\lambda|\,|v_k|$$

From the triangle inequality, and the first subsection:

$$|\lambda|\,|v_k| = \left|\sum_{j=1}^{N} p(x_{t+1} = j|x_t = k) \cdot v_j\right| \leq \sum_{j=1}^{N} p(x_{t+1} = j|x_t = k)\,|\cdot|v_j|$$

Dividing by $|v_k|$ we get (for the 0 vector we get the degenerated case of $\lambda = 0$):

$$|\lambda| = \sum_{j=1}^{N} p(x_{t+1} = j|x_t = k) \cdot \frac{|v_j|}{|v_k|} \leq \sum_{j=1}^{N} p(x_{t+1} = j|x_t = k) = 1$$

Thus:

$$|\lambda| \leq 1$$

2 Smart Jack

2.1

A straight forward extension of the bellman equation written in class for the case of random reward:

$$V_k^{\pi}(s) = \left\{ E\left[r_k(s,a)\right] + \sum_{s' \in S} p_k\left(s'|s,a\right) V_{k+1}^{\pi}\left(s'\right) \right\}_{a=\pi_k(s)} , \quad s \in S$$

Where:

$$p(s'|s,a=1) = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 2/3 & 0 & 1/3 \\ 0.75 & 0.25 & 0 \end{pmatrix}$$

and:

$$p(s'|s,a=2) = \begin{pmatrix} 0 & 0.125 & 0.875 \\ 0.5 & 0 & 0.5 \\ 0.75 & 0.25 & 0 \end{pmatrix}$$

We will call Jack's first policy with π_t
Calculating $V_1^{\pi_t}$ with T=3 using backwards recursion (π_t is not dependent on s and only on the time t):

$$V_3^{\pi_t}(s) = E\left[r_3(s, \pi_3 = 2)\right] = \begin{pmatrix} 0.7 \\ 0 \\ 0.5 \end{pmatrix}$$

Calculating $V_2^{\pi_t}(s)$:

$$\begin{aligned} V_2^{\pi_t}(s) = \left\{ E\left[r_2(s,a)\right] + \sum_{s' \in S} p_2\left(s'|s,a\right) V_3^{\pi}\left(s'\right) \right\}_{a=\pi_2=1} &= \begin{pmatrix} 0.2 \\ 1 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0 & 0.5 & 0.5 \\ 2/3 & 0 & 1/3 \\ 0.75 & 0.25 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.7 \\ 0 \\ 0.5 \end{pmatrix} \\ &= \begin{pmatrix} 0.45 \\ 1.6333 \\ 1.025 \end{pmatrix} \end{aligned}$$

Calculating $V_1^{\pi_t}(s)$:

$$\begin{aligned} V_1^{\pi_t}(s) = \left\{ E\left[r_1(s,a)\right] + \sum_{s' \in S} p_1\left(s'|s,a\right) V_2^{\pi}\left(s'\right) \right\}_{a=\pi_1=2} &= \begin{pmatrix} 0.7 \\ 0 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0 & 0.125 & 0.875 \\ 0.5 & 0 & 0.5 \\ 0.75 & 0.25 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.45 \\ 1.6333 \\ 1.025 \end{pmatrix} \\ &= \begin{pmatrix} 1.801 \\ 0.7375 \\ 1.245 \end{pmatrix} \end{aligned}$$

So the expected result is $V_1^{\pi_t}(s_0) = 1.801$ (we also got the expected rewards if would to start from different states)

2.2

Now we have a stochastic policy, so the bellman equation for the value function takes the form:

$$\begin{aligned}
V_k^{\pi}(s) &= E^{\pi,r}\left[r(s,a)\right] + E^{\pi}\left[\sum_{s'\in S} p\left(s'|s,a\right)V_{k+1}^{\pi}\left(s'\right)\}\right], \quad s\in S \\
&= \frac{1}{2}\cdot\left[E^{r}\left[r(s,a=1)\right] + E^{r}\left[r(s,a=2)\right]\right] + \frac{1}{2}\cdot\left[\sum_{s'\in S} p\left(s'|s,a=1\right)V_{k+1}^{\pi}\left(s'\right)\} + \sum_{s'\in S} p\left(s'|s,a=2\right)V_{k+1}^{\pi}\left(s'\right)\}\right] \\
&= \begin{pmatrix} 0.45 \\ 0.5 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0 & 0.3125 & 0.6875 \\ 0.5833 & 0 & 0.4167 \\ 0.75 & 0.25 & 0 \end{pmatrix} \cdot V_{k+1}^{\pi}(s)
\end{aligned}$$

$$V_3^{\pi}(s) = E\left[r_3(s,\pi_3=2)\right] = \begin{pmatrix} 0.45 \\ 0.5 \\ 0.5 \end{pmatrix}$$

Calculating V_2^{π}:

$$V_2^{\pi}(s) = \begin{pmatrix} 0.45 \\ 0.5 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0 & 0.3125 & 0.6875 \\ 0.5833 & 0 & 0.4167 \\ 0.75 & 0.25 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.45 \\ 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.9500 \\ 0.9708 \\ 0.9625 \end{pmatrix}$$

Calculating V_1^{π}:

$$V_1^{\pi}(s) = \begin{pmatrix} 0.45 \\ 0.5 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0 & 0.3125 & 0.6875 \\ 0.5833 & 0 & 0.4167 \\ 0.75 & 0.25 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.9500 \\ 0.9708 \\ 0.9625 \end{pmatrix} = \begin{pmatrix} 1.4151 \\ 1.4552 \\ 1.4552 \end{pmatrix}$$

In order to find the stationary chain we need to calculate a left eigenvector of the transition matrix :

$$p = \begin{pmatrix} 0 & 0.3125 & 0.6875 \\ 0.5833 & 0 & 0.4167 \\ 0.75 & 0.25 & 0 \end{pmatrix}$$

We get (this is the only eigenvector that can define a distribution):

$$\nu = \begin{pmatrix} 0.4076 \\ 0.2204 \\ 0.3720 \end{pmatrix}$$

2.3

The optimal Bellman equation:

$$V_k(s) = \max_a\left\{E\left[r_k(s,a)\right] + \sum_{s'\in S} p_k\left(s'|s,a\right)V_{k+1}\left(s'\right)\right\} \quad s\in S$$

$$\pi_k^{*}(s) = \arg\max_a\left\{E\left[r_k(s,a)\right] + \sum_{s'\in S} p_k\left(s'|s,a\right)V_{k+1}\left(s'\right)\right\} \quad s\in S$$

Now we will solve the optimal bellman equation using backward recursion.
Starting from $V_3(s)$:

$$V_3(s) = \max_a\left\{E\left[r_k(s,a)\right]\right\} \quad s\in S = \begin{pmatrix} 0.7 \\ 1 \\ 0.5 \end{pmatrix}$$

We can see that $\pi_3^*(s_2)$ is arbitrary, thus:

$$\pi_3^*(s) = \arg\max_a \{E[r_k(s,a)]\} \quad s \in S = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

Calculating $V_2(s)$:

$$V_2(s) = \max_a \left\{ E[r(s,a)] + \sum_{s' \in S} p(s'|s,a) V_3(s') \right\} = \max_a \left\{ \begin{pmatrix} 0.45 \\ 1.6333 \\ 1.025 \end{pmatrix}, \begin{pmatrix} 1.1375 \\ 0.6 \\ 1.025 \end{pmatrix} \right\} = \begin{pmatrix} 1.1375 \\ 1.6333 \\ 1.025 \end{pmatrix}$$

Again, we get exactly the same policy as in t=3:

$$\pi_2^*(s) = \arg\max_a \{E[r_k(s,a)]\} \quad s \in S = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

Calculating $V_1(s)$:

$$V_1(s) = \max_a \left\{ E[r(s,a)] + \sum_{s' \in S} p(s'|s,a) V_2(s') \right\} = \max_a \left\{ \begin{pmatrix} 1.5292 \\ 2.1 \\ 1.7615 \end{pmatrix}, \begin{pmatrix} 1.8010 \\ 1.0812 \\ 1.7615 \end{pmatrix} \right\} = \begin{pmatrix} 1.8010 \\ 2.1 \\ 1.7615 \end{pmatrix}$$

We can see that $\pi_3^*(s_2)$ is arbitrary, thus:

$$\pi_1^*(s) = \arg\max_a \{E[r_k(s,a)]\} \quad s \in S = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

Thus the optimal policy is stationary:

$$\pi^*(s) = \arg\max_a \{E[r_k(s,a)]\} \quad s \in S = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

2.4

Now we know that there is β probability to get thrown out of the casino each round. If we will get thrown out, our future reward will be 0. We will denote the random variable of staying in the casino at time step t as c_t Thus, the equation for the cumulative reward in the infinite horizon case is:

$$J^\pi(s) = E^\pi\left[\sum_{t=0}^{\infty} r(s_t, a_t) | s_0 = s\right] = E^\pi\left[\sum_{t=0}^{\infty} E^{c_t}[r(s_t, a_t) | s_0 = s, c_t]\right] = E^\pi\left[\sum_{t=0}^{\infty} r(s_t, a_t) \cdot (1-\beta)^t + 0 | s_0 = s\right]$$
$$= E^\pi\left[\sum_{t=0}^{\infty} r(s_t, a_t) \cdot (1-\beta)^t | s_0 = s\right]$$

Thus, with comparison to the discounted expected return, we can easily see that $\gamma = 1 - \beta$. Thus the γ parameter incorporates the "fear of death" into the agent's policy.

2.5

The optimal Bellman equation:

$$V_k(s) = \max_a \left\{ E[r_k(s,a)] + (1-\beta) \sum_{s' \in S} p_k(s'|s,a) V_{k+1}(s') \right\} \quad s \in S$$

$$\pi_k^*(s) = \arg\max_a \left\{ E[r_k(s,a)] + (1-\beta) \sum_{s' \in S} p_k(s'|s,a) V_{k+1}(s') \right\} \quad s \in S$$

3 The Secretary Problem

3.1

$$g_t(s=0) = p\,(A)$$

Where A is the event that: the current secretary is with the total highest score given that the secretary doesn't have the highest score within the first t (which is impossible)
Thus:

$$g_t(s=0) = 0$$

And the second case:

$$g_t(s=1) = P(B)$$

Where B is the event that: The current secretary is with the total highest score given that the secretary has the highest score within the first t secretaries. This event is equivalent to the event that the optimal secretary is drawn in the first t secretaries, and with the fact that the sampling is uniform we get:

$$g_t(s=1) = P(\text{Best object is in first } t \text{ steps }) = \frac{t}{N}$$

3.2

$$P_t(1|s) = P(C)$$

Where C is the event that: the current secretary is the best within the first t ones. Thus (due to the fact that that the secretaries are sampled uniformly):

$$P_t(1|s) = P(C) = 1/t + 1$$

This is because we want that the current secretary will have the largest score, and we have a total of t+1 secretaries up until now.
Furthermore we know that:

$$P_t(1|s) = 1 - P(C) = t/t + 1$$

both of the results are independent of t.

3.3

First, notice that:

$$V_T^\star = V_{t=N}^\star$$

I.e it is a measure which evaluates the final state in [0,1].
Hence, it is expected that we get:

$$V_T^\star(0) = 0$$
$$V_T^\star(1) = 1$$

Notice that in the case of having a secretary chosen, the value function shall be

$$V_t^\star(s) = g_t(s) \tag{1}$$

While on the other hand, if the secretary have not been chosen yet, the value function satisfies the Bellman's Equation.

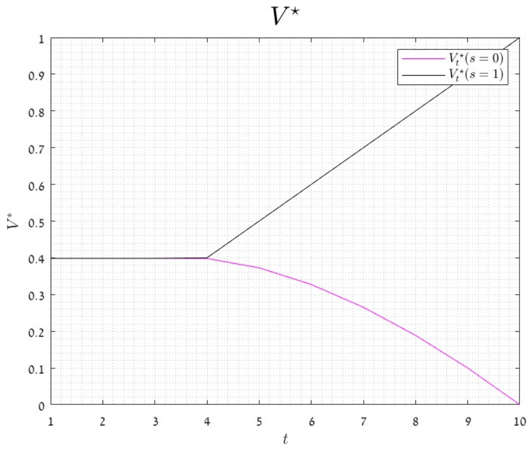

Figure 1: Plotting of V^* vs time

Now, well look to plug it into Bellman's Equation:

$$V_t^\star(s) = \max_{a \in A} \left(r_t(s,a) + \sum_{s'} P_t(s'|s,a) \cdot V_{t+1}^\star(s') \right)$$

$$\longrightarrow_{no\ r\ in\ this\ case} \sum_{s'} P_t(s'|s,a) \cdot V_{t+1}^\star(s') = P_t(0|s)V_{t+1}^\star(0) + P_t(1|s)V_{t+1}^\star(1)$$

Finally, we see that the Bellman's Equation for the value function this case is:

$$V_{t+1}^\star(s) = \max \left[P_t(0|s)V_{t+1}^\star(0) + P_t(1|s)V_{t+1}^\star(1), g_t(s) \right]$$

3.4

Lets start plugging 0 and 1 into the value function we have just found:

$$V_t^\star(s=0) = \max \left[P_t(0|0)V_{t+1}^\star(0) + P_t(1|0)V_{t+1}^\star(1), g_t(0) \right] = \frac{1}{t+1}V_{t+1}^\star(0) + \frac{t}{t+1}V_{t+1}^\star(1)$$

$$V_t^\star(s=1) = \max \left[P_t(0|1)V_{t+1}^\star(0) + P_t(1|1)V_{t+1}^\star(1), g_t(1) \right] = \max \left[\frac{t}{t+1}V_{t+1}^\star(0) + \frac{1}{t+1}V_{t+1}^\star(1), \frac{t}{N} \right]$$

Hence, Finally, we get:

$$V_t^\star(1) = \max \left[V_t^\star(0), \frac{t}{N} \right]$$

Numerical solution can be observed in Figure 1

3.5

The graph implies that for t values smaller than $\tau = 4$ the choice of $\frac{t}{N}$ is not favored since it's value is smaller than $V_t^\star(0)$. We also see that for t greater than $\tau = 4$, once a secretary is recognized as the best $(s = 1)$ we get a value greater than $V_t^\star(0)$ and we'll choose that secretary.

4 Property From The Lecture

4.1

Show that:

$$
\begin{aligned}
V^{\pi}(s) &\triangleq E^{\pi}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_t,a_t\right)|s_0=s\right)\\
&=E^{\pi}\left(\sum_{t=1}^{\infty}\gamma^{t-1}r\left(s_t,a_t\right)|s_1=s\right)
\end{aligned}
$$

We will prove this Lemma under the assumption that π is stationary.
Changing the sum index:

$$
E^{\pi}\left(\sum_{t=1}^{\infty}\gamma^{t-1}r\left(s_t,a_t\right)|s_1=s\right)=E^{\pi}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_{t+1},a_{t+1}\right)|s_1=s\right)
$$

Using the stationary assumption:

$$
E^{\pi}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_{t+1},a_{t+1}\right)|s_1=s\right)=E^{\pi}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_{t+1},\pi(s_{t+1})\right)|s_1=s\right)=E^{\pi}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_t,\pi(s_t)\right)|s_0=s\right)=V^{\pi}(s)
$$

5 Second Moment And Variance Of Return

5.1

The definition of the second moment of the discounted return:

$$
M^{\pi}(s)=E^{\pi,s}\left(\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2\right)
$$

Assuming π is deterministic, we can see that:

$$
\begin{aligned}
M^{\pi}(s)=E^{\pi,s}\left(\left(r\left(s_0,a_0\right)+\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2\right)&=E^{\pi,s}\left(\left(r\left(s,\pi(s)\right)+\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2\right)\\
&=E^{\pi,s}\left(r^2\left(s,\pi(s)\right)+\left(\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2+2\cdot\left(\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)\cdot r\left(s,\pi(s)\right)\right)\\
&=r^2\left(s,\pi(s)\right)+E^{\pi,s}\left(\left(\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2\right)+2\cdot r\left(s,\pi(s)\right)\cdot E^{\pi,s}\left(\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)
\end{aligned}
$$

We saw in class that:

$$
E^{\pi,s}\left(\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)=\gamma\cdot E^{\pi,s}\left(V(s')\right)
$$

We can do a similar derivation for the quadratic part:

$$
E^{\pi,s}\left(\left(\sum_{t=1}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2\right)=\gamma^2E^{\pi,s}\left(\left(\sum_{t=1}^{\infty}\gamma^{t-1}r\left(s_t,a_t\right)\right)^2\right)
$$

Using the law of total expectation (on the random variable s'), we get:

$$
\begin{aligned}
\gamma^2E^{\pi,s}\left(\left(\sum_{t=1}^{\infty}\gamma^{t-1}r\left(s_t,a_t\right)\right)^2\right)&=\gamma^2E^{\pi}\left(E^{\pi}\left(\left(\sum_{t=1}^{\infty}\gamma^{t-1}r\left(s_t,a_t\right)\right)^2\mid s_0=s,s_1=s'\right)\mid s_0=s'\right)\\
&=\gamma^2E^{\pi}\left(E^{\pi}\left(\left(\sum_{t=1}^{\infty}\gamma^{t-1}r\left(s_t,a_t\right)\right)^2\mid s_1=s'\right)\mid s_0=s'\right)
\end{aligned}
$$

Just like section (4), here we also assume that π is stationary:

$$E^{\pi}\left(\left(\sum_{t=1}^{\infty}\gamma^{t-1}r\left(s_t,a_t\right)\right)^2 \mid s_1=s'\right)=E^{\pi}\left(\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2 \mid s_0=s'\right)=M^{\pi}(s')$$

Plugging back all the results:

$$M^{\pi}(s)=r^2\left(s,\pi(s)\right)+\gamma^2E^{\pi,s}\left(M^{\pi}(s')\right)+2\cdot r\left(s,\pi(s)\right)\cdot\gamma\cdot E^{\pi,s}\left(V(s')\right)$$

5.2

We now have |S| equations for M^{π} (one for each state in S). But in order to solve this equations we first need to know $V^{\pi}(s)$ for each s in S. Thus, in total we have $2\cdot|S|$ equations to solve in order to know M for each s in S.

5.3

We know the basic property about variances:

$$W^{\pi}(s)=\text{Var}^{\pi,s}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)=E^{\pi,s}\left(\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)^2\right)-\left(E^{\pi,s}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_t,a_t\right)\right)\right)^2$$

But those are exactly V and M, thus:

$$W^{\pi}(s)=M^{\pi}(s)-\left(V^{\pi}(s)\right)^2$$