Problem Set no. 3 – With solutions

Given: December, 2020

Credit: Ishay Golinsky

Exercise 3.1 Let $P = \left[\begin{smallmatrix} 1-a & a \\ b & 1-b \end{smallmatrix}\right]$, where $0 < a, b < 1$, and $\mathbf{c} = \left[\begin{smallmatrix} c_1 \\ c_2 \end{smallmatrix}\right]$ be the transition matrix and the cost vector of a Markov Reward Process, i.e., a Markov chain with costs associated with its states. What are the matrices P^* and H? What are the limiting average costs of the two states and what are the biases? (Hint: You can use the fact that $P^n = \frac{1}{a+b}\left[\begin{smallmatrix} b & a \\ b & a \end{smallmatrix}\right] + \frac{(1-a-b)^n}{a+b}\left[\begin{smallmatrix} a & -a \\ -b & b \end{smallmatrix}\right]$, but it is actually easier not to use this exact formula.)

Solution 3.1 Using a result from class, since P is irreducible, all the rows of P^* are the (unique) stationary distribution of P. To find the stationary distribution of P we write

$$\begin{aligned} 0 &= x^T (P - I) \\ &= x^T \begin{pmatrix} -a & a \\ b & -b \end{pmatrix} \\ &= (-ax_1 + bx_2\,,\, ax_1 - bx_2)\,, \end{aligned}$$

which yields $x_2 = \frac{a}{b}x_1$. Normalizing such that $\|x\|_1 = 1$ we get

$$P^* = \frac{1}{a+b}\begin{pmatrix} b & a \\ b & a \end{pmatrix}.$$

The limiting average costs are

$$\mathbf{y} = P^*\mathbf{c} = \frac{ac_2 + bc_1}{a+b}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Using the hint we have

$$\begin{aligned} H &= \sum_{t=0}^{\infty} \left(P^t - P^*\right) \\ &= \sum_{t=0}^{\infty} \frac{(1-a-b)^t}{a+b}\begin{pmatrix} a & -a \\ -b & b \end{pmatrix} \\ &= \frac{1}{1-(1-a-b)}\frac{1}{a+b}\begin{pmatrix} a & -a \\ -b & b \end{pmatrix} \\ &= \frac{1}{(a+b)^2}\begin{pmatrix} a & -a \\ -b & b \end{pmatrix}. \end{aligned}$$

Finally,

$$\mathbf{z} = H\mathbf{c} = \frac{1}{(a+b)^2}\begin{pmatrix} a(c_1 - c_2) \\ b(c_2 - c_1) \end{pmatrix}.$$

Exercise 3.2 Let M be a Markov Reward Process (MRP) with transition matrix P and cost vector $\mathbf{c}$. Let $\mathbf{y}$ and $\mathbf{z}$ be the value and bias vectors as defined in class. For $0 < \lambda < 1$, let $\mathbf{y}_\lambda$ be the discounted value corresponding to discount factor λ. Show that $\mathbf{y}_\lambda = \frac{\mathbf{y}}{1-\lambda} + \mathbf{z} + \boldsymbol{\varepsilon}(\lambda)$, where

$\boldsymbol{\varepsilon}(\lambda) \to 0$ as $\lambda \to 1$. (Note that $\boldsymbol{\varepsilon}(\lambda)$ is a vector.) (Hint: $\mathbf{y}_\lambda = \sum_{t=0}^{\infty} \lambda^t P^t c = \left(\sum_{t=0}^{\infty} \lambda^t P^*\right)\mathbf{c} + \left(\sum_{t=0}^{\infty} \lambda^t (P^t - P^*)\right)\mathbf{c}$. Let $H(\lambda) = \sum_{t=0}^{\infty} \lambda^t (P^t - P^*)$. You may use the fact that $H(\lambda) \to H$, as $\lambda \to 1$, where $H = (I - P + P^*)^{-1} - P^*$ is the deviation matrix defined in class.)

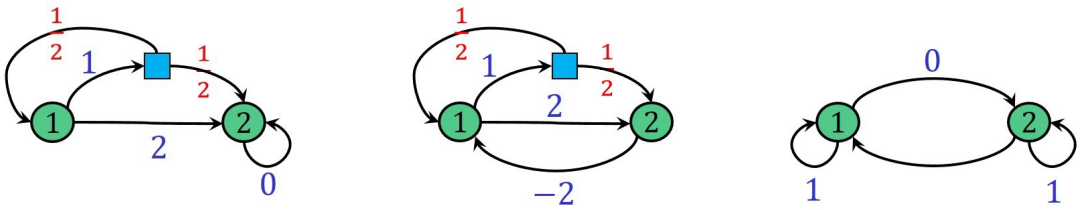
Solution 3.2 Denote $\epsilon(\lambda) = (H(\lambda) - H)\,\mathbf{c} \overset{\lambda \to 1^-}{\longrightarrow} 0$. Following the hint, we have

$$\begin{aligned}
\mathbf{y}_\lambda &= \left(\sum_{t=0}^{\infty} \lambda^t P^t\right)\mathbf{c} \\
&= \left(\sum_{t=0}^{\infty} \lambda^t P^*\right)\mathbf{c} + \left(\sum_{t=0}^{\infty} \lambda^t \left(P^t - P^*\right)\right)\mathbf{c} \\
&= \left(\sum_{t=0}^{\infty} \lambda^t\right) P^*\mathbf{c} + H(\lambda)\mathbf{c} \\
&= \frac{1}{1-\lambda}\mathbf{y} + H\mathbf{c} + (H(\lambda) - H)\,\mathbf{c} \\
&= \frac{\mathbf{y}}{1-\lambda} + \mathbf{z} + \boldsymbol{\epsilon}(\lambda).
\end{aligned}$$

Exercise 3.3 Let π be a positional strategy of an MDP. Show that each one of the following conditions is strictly stronger than the conditions following it. In other words, for $i = 1, 2, 3, 4$, show that condition i implies condition $i+1$, but not vice versa.

1. π is Blackwell-optimal: there exists $0 < \lambda^* < 1$ such that π is λ-optimal, i.e., optimal in the discounted game with discount factor λ, for every $\lambda^* \le \lambda < 1$.
2. π is value-and-bias optimal: for every (positional) policy π', either $\mathbf{y}^\pi < \mathbf{y}^{\pi'}$ or $\mathbf{y}^\pi = \mathbf{y}^{\pi'}$ and $\mathbf{z}^\pi \le \mathbf{z}^{\pi'}$
3. No action is improving for π: for every $i \in S$ and $a \in A_i$, either $\sum_j p_{a,j} y_j^\pi > y_i^\pi$ or $\sum_j p_{a,j} y_j^\pi = y_i^\pi$ but $c_a + \sum_j p_{a,j} z_j^\pi \ge y_i^\pi + z_i^\pi$.
4. π is value-optimal: for every (positional) policy π', $\mathbf{y}^\pi \le \mathbf{y}^{\pi'}$.
5. No action is y-improving for π: for every $i \in S$ and $a \in A_i$, $\sum_j p_{a,j} y_j^\pi \ge y_i^\pi$.

Hint: The separations can be obtained using appropriate policies in one of the following simple MDPs.



Solution 3.3 $1 \implies 2$: Assuming 1, for every $\lambda \in [\lambda^*, 1)$ we have $\mathbf{y}_\lambda^\pi \le \mathbf{y}_\lambda^{\pi'}$. Using the result of exercise 3.2 it follows that

$$\frac{\mathbf{y}^\pi}{1-\lambda} + \mathbf{z}^\pi + \boldsymbol{\epsilon}^\pi(\lambda) \le \frac{\mathbf{y}^{\pi'}}{1-\lambda} + \mathbf{z}^{\pi'} + \boldsymbol{\epsilon}^{\pi'}(\lambda).$$

Rearranging, we get

$$\mathbf{y}^{\pi} - \mathbf{y}^{\pi'} \leq (1-\lambda)\left(\boldsymbol{\epsilon}^{\pi'}(\lambda) + \mathbf{z}^{\pi'} - \boldsymbol{\epsilon}^{\pi}(\lambda) - \mathbf{z}^{\pi}\right) \overset{\lambda \to 1^-}{\longrightarrow} 0,$$

yielding $\mathbf{y}^{\pi} \leq \mathbf{y}^{\pi'}$. Now, in case $\mathbf{y}^{\pi} = \mathbf{y}^{\pi'}$, using exercise 3.2 again we have

$$\mathbf{z}^{\pi} + \boldsymbol{\epsilon}^{\pi}(\lambda) \leq \mathbf{z}^{\pi'} + \boldsymbol{\epsilon}^{\pi'}(\lambda).$$

Rearranging, we get

$$\mathbf{z}^{\pi} - \mathbf{z}^{\pi'} \leq \boldsymbol{\epsilon}^{\pi'}(\lambda) - \boldsymbol{\epsilon}^{\pi}(\lambda) \overset{\lambda \to 1^-}{\longrightarrow} 0,$$

yielding $\mathbf{z}^{\pi} \leq \mathbf{z}^{\pi'}$.

$2 \implies 3$: This was shown in class under *improvement→progress*. We've seen that if π' is obtained from π by an improving switch, then

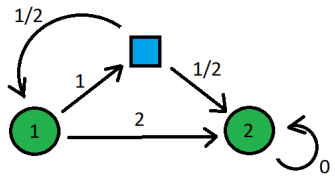
$$(\mathbf{y}^{\pi}, \mathbf{z}^{\pi}) > \left(\mathbf{y}^{\pi'}, \mathbf{z}^{\pi'}\right),$$

which means $\neg 3 \implies \neg 2$.

$3 \implies 4$: This was shown in class under *no improvement→value-optimality*.

$4 \implies 5$: This was also shown in class - case 1 in the *improvement→progress* proof shows that if π' is obtained from π by a y-improving switch, then $\mathbf{y}^{\pi} > \mathbf{y}^{\pi'}$, contradicting π's value-optimality.

$2 \not\implies 1$: Consider the following game:



Denote by π_1, π_2 the two possible positional policies (corresponding with a choice of action in state 1). It is straight forward to verify that both have $\mathbf{y}^{\pi_i} = \mathbf{0}$ and $\mathbf{z}^{\pi_i} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$. In particular, both are value-and-bias-optimal. However, in the discounted game we have

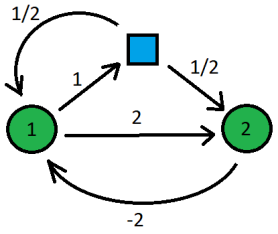
$$y_1^{\pi_1} = 1 + \lambda\left(\frac{1}{2}y_1^{\pi_1} + \frac{1}{2}y_2^{\pi_1}\right) = 1 + \frac{\lambda}{2}y_1^{\pi_1},$$

therefore

$$y_1^{\pi_1} = \frac{2}{2-\lambda} < 2 = y_1^{\pi_2},$$

meaning π_2 is not λ-optimal.

$3 \not\implies 2$: Consider the game:



Denote by π_1, π_2 the two possible positional policies (corresponding with the cost of the action from state 1). It is straight forward to verify that $\mathbf{y}^{\pi_1} = \mathbf{y}^{\pi_2} = \mathbf{0}$ and that $\mathbf{z}^{\pi_2} = (1, -1)^T$. Next we compute $\mathbf{z}^{\pi_1}$: note that

$$P_{\pi_1} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$

is irreducible, therefore $P^*_{\pi_1}$'s rows are the stationary distribution of P_{π_1},

$$P^*_{\pi_1} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}.$$

We have

$$\begin{aligned} H^{\pi_1} &= \left(I - P_{\pi_1} + P^*_{\pi_1}\right)^{-1} - P^*_{\pi_1} \\ &= \begin{pmatrix} \frac{2}{9} & \frac{-2}{9} \\ \frac{-4}{9} & \frac{4}{9} \end{pmatrix}. \end{aligned}$$

Therefore

$$\mathbf{z}^{\pi_1} = H_{\pi_1} c = \begin{pmatrix} \frac{2}{3} \\ \frac{-4}{3} \end{pmatrix}.$$

Now, we have $\mathbf{z}^{\pi_1} < \mathbf{z}^{\pi_2}$ and $\mathbf{y}^{\pi_1} = \mathbf{y}^{\pi_2}$, therefore π_2 is not value-and-bias-optimal. It remains to show that there is no improving action wrt to π_2. The only possible switch wrt to π_2 has

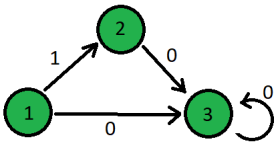
$$\sum_{j=1}^{2} p_{a,j} y_j^{\pi_2} = 0 = y_1^{\pi_2}$$

and

$$c_a + \sum_{j=1}^{2} p_{a,j} z_j^{\pi_2} = 1 + \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-1) = 1 = y_1^{\pi_2} + z_1^{\pi_2},$$

that is, not improving.

$4 \not\Longleftrightarrow 3$: Consider the game:



Let π_1 be the policy that goes from state 1 directly to state 3, and let π_2 be the policy that passes through state 2. It is straight forward to verify that $\mathbf{y}^{\pi_1} = \mathbf{y}^{\pi_2} = \mathbf{z}^{\pi_1} = \mathbf{0}$ and that $\mathbf{z}^{\pi_2} = (1, 0, 0)^T$.

In particular, π_2 is value-optimal. It remains to show that there is an improving action wrt to π_2. The only possible switch wrt to π_2 has

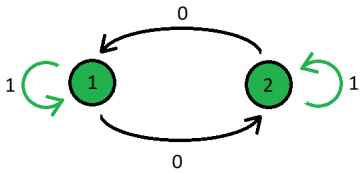
$$\sum_{j=1}^{3} p_{a,j} y_j^{\pi_2} = 0 = y_1^{\pi_2}$$

and

$$c_a + \sum_{j=1}^{3} p_{a,j} z_j^{\pi_2} = 0 + z_3^{\pi_2} = 0 < 1 = y_1^{\pi_2} + z_1^{\pi_2},$$

that is, improving.

$5 \not\Longrightarrow 4$: Let π be the positional policy indicated by the green arrows:



We have $\mathbf{y}^\pi = \mathbf{1}$. Clearly π is not value optimal since $\mathbf{y} = \mathbf{0}$ can be achieved. However, as mentioned in class, there is no y-improving switch wrt π.

Exercise 3.4 Extend the policy iteration algorithm for MDPs with the limiting average reward objective given in class to TBSGs with the limiting average reward objective. When computing optimal counter-strategies, which of the five notions of optimality listed in the previous exercise should be used?

Solution 3.4 This question was removed.

Exercise 3.5 Let $\Gamma = (S_0, S_1, (A_i)_{i \in S_0 \cup S_1}, p, c)$ be a TBSG. Let $S = S_0 \cup S_1$ and $A = \cup_{i \in S} A_i$. Let $0 < \lambda < 1$ be a discount factor and let $s \in S_0 \cup S_1$ be a selected state. Let $\Gamma_{s,\lambda} = (S_0, S_1, (A_i)_{i \in S_0 \cup S_1}, p', c)$ be a game obtained by changing the transition probabilities as follows: For every $a \in A$, let $p'_{a,j} = \lambda p_{a,j}$, if $j \neq s$, and $p'_{a,s} = \lambda p_{a,s} + (1 - \lambda)$. What is the relation between $Val_\lambda(\Gamma, s)$ and $Val_{AVG}(\Gamma_{s,\lambda}, s)$? Here, $Val_\lambda(\Gamma, s)$ is the value of a game played on Γ starting from s with the λ-discounted criterion, while $Val_{AVG}(\Gamma_{s,\lambda}, s)$ is the value of a game played on $\Gamma_{s,\lambda}$ starting from s with the limiting average criterion. (Note that this may be viewed as a reduction from discounted games to limiting average games.)

Solution 3.5 We will show that for every positional profile π it holds that

$$\mathrm{Val}^\pi_{\mathrm{AVG}}(\Gamma_{s,\lambda}, s) = (1 - \lambda) \cdot \mathrm{Val}^\pi_\lambda(\Gamma, s),$$

and in particular,

$$\mathrm{Val}_{\mathrm{AVG}}(\Gamma_{s,\lambda}, s) = (1 - \lambda) \cdot \mathrm{Val}_\lambda(\Gamma, s).$$

Denote by J_s the matrix that has $\mathbf{1}$ in the s^{th} column and 0 everywhere else. We have

$$P'_\pi = \lambda P_\pi + (1 - \lambda) J_s.$$

We start by finding the stationary distribution of P'_π (which is unique since P'_π has one recurrent class). For $x \geq 0$ such that $\|x\|_1 = 1$ we have

$$x^T J_s = \left(\sum_{i=1}^{n} x_i \right) e_s^T = e_s^T,$$

therefore

$$\begin{aligned} & x^T = \lambda x^T P_\pi + (1-\lambda) x^T J_s \\ \iff & x^T \left(I - \lambda P_\pi \right) = (1-\lambda) e_s^T \\ \iff & x^T = (1-\lambda) e_s^T \left(I - \lambda P_\pi \right)^{-1}. \end{aligned}$$

Next, we have

$$\begin{aligned} (P'_\pi)^* &= \begin{pmatrix} - & x^T & - \\ & \vdots & \\ - & x^T & - \end{pmatrix} \\ &= (1-\lambda) \begin{pmatrix} - & e_s^T \left(I - \lambda P_\pi \right)^{-1} & - \\ & \vdots & \\ - & e_s^T \left(I - \lambda P_\pi \right)^{-1} & - \end{pmatrix} \\ &= (1-\lambda) \begin{pmatrix} - & e_s^T & - \\ & \vdots & \\ - & e_s^T & - \end{pmatrix} \left(I - \lambda P_\pi \right)^{-1} \\ &= (1-\lambda) J_s \left(I - \lambda P_\pi \right)^{-1}. \end{aligned}$$

Finally, we have

$$\mathrm{Val}_\lambda \left(\Gamma \right) = \left(I - \lambda P_\pi \right)^{-1} \mathbf{c},$$

therefore

$$\begin{aligned} \mathrm{Val}^\pi_{\mathrm{AVG}} \left(\Gamma_{s,\lambda} \right) &= (P'_\pi)^* \mathbf{c} \\ &= (1-\lambda) J_s \mathrm{Val}_\lambda \left(\Gamma \right) \\ &= (1-\lambda) \mathrm{Val}_\lambda \left(\Gamma, s \right) \mathbf{1}. \end{aligned}$$

Taking only the s^{th} entry, we get

$$\mathrm{Val}^\pi_{\mathrm{AVG}} \left(\Gamma_{s,\lambda}, s \right) = (1-\lambda) \mathrm{Val}_\lambda \left(\Gamma, s \right).$$

Exercise 3.6 Let $G = (V_0, V_1, E, c)$ be an MPG and let σ and τ be positional strategies of the two players. Describe a linear time algorithm for computing $\mathbf{y}^{\sigma,\tau}$ and $\mathbf{z}^{\sigma,\tau}$.

Solution 3.6 Let H be the sub-graph of G that is obtained by removing all edges not played by (σ, τ). Denote by P the transition matrix of the process obtained by playing (σ, τ). We start by decomposing H to *weak* connected components (by that we mean connected components of the graph obtained by replacing all edges of H with undirected edges), then proceed to solve $\mathbf{y}$ and $\mathbf{z}$ for each component separately. Such decomposition can be done in $O(n)$, therefore from here we assume wlog that H is weakly connected. The high level algorithm is as follows.

1. find a cycle C in H
2. set $\mathbf{y} \leftarrow \frac{1}{|C|}\sum_{i\in C} c_i \cdot \mathbf{1}$
3. solve for $\mathbf{z}$: $\begin{cases}(I-P)\mathbf{z} = \mathbf{c}-\mathbf{y}\\ \sum_{i\in C} z_i = 0\end{cases}$
4. output $\mathbf{y}, \mathbf{z}$

The only stage that is not trivially implemented in linear time is stage 3. Note that since H is weakly connected and $|E(H)| = |V(H)|$, it follows that there exist exactly one cycle C in H. Since every play eventually reaches C, all states have the value $\frac{1}{|C|}\sum_{i\in C} c_i$, as set in stage 2. The following two claims show all that remains.

Claim 1. $\mathbf{z}$ is correctly computed**.**

Proof of claim 1. We've seen in class that $\mathbf{z}$ is characterized by the following system of equations

$$\begin{cases}(I-P)\mathbf{z} = \mathbf{c}-\mathbf{y}\\ P^*\mathbf{z} = 0\end{cases},$$

therefore it remains to show that

$$P^*\mathbf{z} = 0 \iff \sum_{i\in C} z_i = 0.$$

Assume wlog that $C = (1, 2, \ldots, k)$. Then P has the form

$$P = \begin{pmatrix}P_1 & 0\\ Q_1 & Q_2\end{pmatrix},$$

where

$$P_1 = \begin{pmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & & \vdots\\ \vdots & & \ddots & \ddots & 0\\ 0 & & & 0 & 1\\ 1 & 0 & \cdots & & 0\end{pmatrix}_{k\times k}.$$

It is straight forward to verify that $\mathbf{1}^T P_1 = \mathbf{1}^T$, therefore

$$P_1^* = \frac{1}{k}\begin{pmatrix}1 & \cdots & 1\\ \vdots & \ddots & \vdots\\ 1 & \cdots & 1\end{pmatrix}_{k\times k}.$$

We have

$$\begin{aligned}P^* &= \begin{pmatrix}P_1^* & 0\\ Q_1' & 0\end{pmatrix}\\ &= \frac{1}{k}\begin{pmatrix}\mathbf{1}_{n\times k} & \mathbf{0}_{n\times(n-k)}\end{pmatrix},\end{aligned}$$

using that the rows of Q_1' are convex combinations of the rows of P_1^*. From here the required equivalence is easy to see.

Claim 2. Stage 3 can be implemented in linear time.

Proof of claim 2. Assume wlog again that $C = (1, 2, \ldots, k)$. The equation $(I - P)\mathbf{z} = \mathbf{c} - \mathbf{y}$ for indices $i \in C$ reads

$$z_i - z_{i+1 (mod\, k)} = c_i - y_i.$$

We set $z_1 = 0$ and solve for $i = 2, \ldots, k$ in order, using the equation above. Then we normalize by

$$z_i \leftarrow z_i - \frac{1}{|C|} \sum_{j \in C} z_j$$

so that $\sum_{i \in C} z_i = 0$ holds as well. For $i \notin C$, the required equation reads

$$z_i = c_i - y_i + z_j,$$

where j is such that an edge from i to j exists. We preform topological sort on the vertices in $V \setminus C$ and solve for z_i using the equation above, in reverse topological order.