Problem Set no. 3

Given: December 28, 2020
Due: January 10, 2021

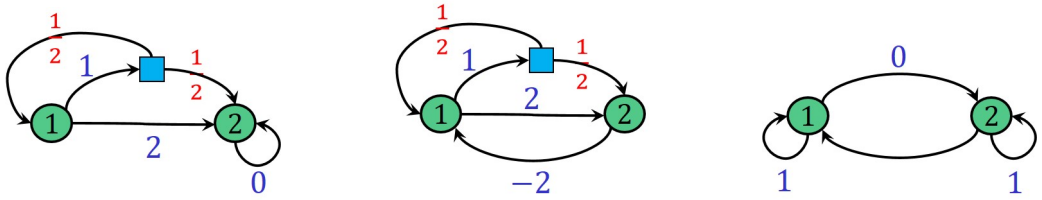
Exercise 3.1 Let $P = \left[\begin{smallmatrix} 1-a & a \\ b & 1-b \end{smallmatrix}\right]$, where $0 < a, b < 1$, and $\mathbf{c} = \left[\begin{smallmatrix} c_1 \\ c_2 \end{smallmatrix}\right]$ be the transition matrix and the cost vector of a Markov Reward Process, i.e., a Markov chain with costs 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.)

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.)

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 $\lambda^* \leq \lambda < 1$.
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 $\mathbf{z}^\pi \leq \mathbf{z}^{\pi'}$
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 $c_a + \sum_j p_{a,j} z_j^\pi \geq y_i^\pi + z_i^\pi$.
4. π is value-optimal: for every (positional) policy π', $\mathbf{y}^\pi \leq \mathbf{y}^{\pi'}$.
5. No action is y-improving for π: for every $i \in S$ and $a \in A_i$, $\sum_j p_{a,j} y_j^\pi \geq y_i^\pi$.

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



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?

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.)

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 $\mathbf{z}^{\sigma,\tau}$.