Probabilistic Aspects of Computer Science: TD Bonus
Average Reward for Unichains

Chargé de TD: Adrien Koutsos

October 25, 2017

A Markov Decision Process $\mathcal{M}$ is said to be a *unichain* if the finite-state Markov chain $\mathcal{M}^\pi$, induced by any deterministic stationary policy $\pi = d^\infty$, has exactly one recurrent strongly connected component plus a possibly empty set of transient states (it is often said to be *recurrent* in case this set of transient states is empty). Otherwise, the MDP is said to be *multichain*. We are interested in the following in studying the limsup and liminf average optimal rewards

$$\mathbf{g}^*_+[s] = \sup_{\pi \in \Pi^{MR}} (\mathbf{g}^\pi_+[s]) \qquad \text{where} \qquad \mathbf{g}^\pi_+ = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{E}^\pi(r(X_i, Y_i))$$

$$\mathbf{g}^*_-[s] = \sup_{\pi \in \Pi^{MR}} (\mathbf{g}^\pi_-[s]) \qquad \text{where} \qquad \mathbf{g}^\pi_- = \liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{E}^\pi(r(X_i, Y_i))$$

and the average optimal reward in case $\mathbf{g}^*_+ = \mathbf{g}^*_-$ (we denote it $\mathbf{g}^*$ in that case).

Exercise 1 (Average rewards in unichains)**.** We consider in this exercise a unichain $\mathcal{M}$.

1. We start by studying a particular deterministic stationary policy d^∞. Let $\mathbf{P}_d$ be the matrix of the finite-state Markov chain $\mathcal{M}^{d^\infty}$. Describe precisely the matrix $\mathbf{P}^*_d$ defined as the Cesaro-limit of the sequence $\{\mathbf{P}^n_d\}_{n\in\mathbb{N}}$. Show that $\mathbf{g}^{d^\infty}$ exists and is a constant vector.

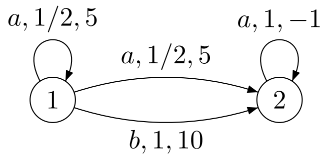
2. We consider the system of equations (E), with variables $g \in \mathbb{R}$ and $\mathbf{h} \in \mathbb{R}^S$:

$$\forall s \in S \qquad g + \mathbf{h}[s] = \max\{r(s,a) + \sum_{s' \in S} p(s' \mid s, a)\mathbf{h}[s'] \mid a \in A_s\}$$

   Using Theorem 3.38, show that

   - let d^∞ be a Blackwell optimal policy. Then $((\mathbf{P}^*_d \mathbf{r}_d)[s_0], \mathbf{D}_d \mathbf{r}_d)$ is a solution of (E), for every state s_0;[1]
   - if $(g, \mathbf{h})$ is a solution of (E), then $g = \mathbf{g}^*[s]$ for every state $s \in S$, and there exists $c \in \mathbb{R}$ such that $\mathbf{h}[s] = (\mathbf{D}_d \mathbf{r}_d)[s] + c$ for every state s, with d^∞ a Blackwell optimal policy.

3. Consider the MDP schematized below:



   Show that it is a unichain, and write the system of equations (E). Solve it. Among all the deterministic stationary policies of this MDP, which one(s) is/are optimal? a Blackwell optimal policy?

[1]We denote by $\mathbf{D}_d$ the deviation matrix defined by $(\mathbf{Id} - \mathbf{P}_d + \mathbf{P}^*_d)^{-1} - \mathbf{P}^*_d$.

4. For $\mathbf{h} \in \mathbb{R}^S$, a decision rule d is $\mathbf{h}$-improving if $\mathbf{r}_d + \mathbf{P}_d\mathbf{h} = \max_{d'}(\mathbf{r}_{d'} + \mathbf{P}_{d'}\mathbf{h})$. Let $(g, \mathbf{h})$ be a solution of (E), and let d be an $\mathbf{h}$-improving decision rule. Show that d^∞ is an optimal policy, i.e., $\mathbf{g}^{d^\infty} = \mathbf{g}^*$.

5. What becomes of the policy iteration presented in the course in this special case?

Exercise 2 (coNP-completeness of deciding if an MDP is a unichain)**.** Show that the problem of deciding whether a given MDP is a unichain is in coNP (equivalently that deciding if it is a multichain is in NP. By reduction of 3-SAT, show that it is indeed a coNP-complete problem. (*Hint: starting from an instance φ of 3-SAT, construct an MDP $\mathcal{M}_\varphi$ such that φ is satisfiable if, and only if, $\mathcal{M}_\varphi$ is a multichain. You may think of using as states of the MDP: one state c_j per clause ($j \in \{1, \dots, m\}$), 4 states s_i, s_i^*, t_i, f_i per litteral ($i \in \{1, \dots, n\}$), and two special states a and b, one encoding a truth assignment, and the other encoding its converse.*)