Problem Set no. 1 – With solutions

Given: November 2, 2020

Exercise 1.1 Let π be an arbitrary randomized and history-dependent policy for a total reward MDP Γ that satisfies the stopping condition. Argue directly, without relying on any of the results proved in class, that there is a deterministic, possibly history-dependent, policy π' such that $Val^{\pi'}(\Gamma, i) \leq Val^{\pi}(\Gamma, i)$, for every $i \in S$.

Solution 1.1
We begin with a discussion of the nature of a general randomized strategies.
A randomized strategy is a probability distribution over deterministic, possibly history-dependent, strategies. This definition is not as 'innocent' as it looks, as it involves a probability distribution over the uncountable set of history-dependent deterministic strategies. (This is similar to the definition continuous random variables.) A precise definition of this notion requires some knowledge of *measure theory* that we do not want to assume in this course.
At a very high level, if π is a randomized strategy, then $Val^{\pi}(\Gamma, i) = \mathbb{E}_{\pi' \sim \pi}[\, Val^{\pi'}(\Gamma, i)]$. (A rigorous definition of this expectation again requires some care.) A basic property of expectations is that there is always a specific, not random, choice that attains a value that is at most the expectation. This says that there is always a deterministic strategy π' such that $Val^{\pi'}(\Gamma, i) \leq Val^{\pi}(\Gamma, i)$, as required.
Below, we just use the intuitive fact that if π is a probability distribution over deterministic strategies, then $\mathbb{P}_{\pi' \sim \pi}[\pi'(h) = a]$, the probability of choosing action a given a finite history h, when a deterministic strategy π' is chosen according to the distribution given by π, is well-defined for every finite history h and every action a.
A *behavior strategy* is a (seemingly) more restricted version of a randomized strategy. For every finite history h which ends in state i, it assigns a probability distribution over the actions of A_i. Actually, for MDP strategies, there is essentially no difference between general randomized strategies and behavior strategies. If π is a general strategy, we simply let $\bar{\pi}$ be the behavior strategy such that for $\mathbb{P}[\bar{\pi}(h) = a] = \mathbb{P}_{\pi' \sim \pi}[\pi'(h) = a]$. Then, the strategies π and π' are indistinguishable. The only difference is that π makes all random choices in advance, which greatly complicates things, while $\bar{\pi}$ makes the random choices only when needed.
This might also be an appropriate place to formally define histories. A *history* is a sequence $(s_0, a_0, s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$. The interpretation is that in time $i = 0, 1, \ldots, t-1$, the process was in state s_i and action $a_i \in A_i$ was taken. At time t the process is in state s_t. (As the sets A_i, for $i \in S$ are assumed to be disjoint, we can actually define the history to be $(a_0, \ldots, a_{t-1}, s_t)$.)
We now give a more detailed proof for behavior strategies. Let π be a behavior strategy. Then,

$$Val^{\pi}(\Gamma, i, h) = \sum_{a \in \pi(i,h)} Pr(\pi(i,h) = a) \cdot \left(c_a + \sum_{j \in S} p_{a,j}\, Val^{\pi}(\Gamma, i, h.(a,j)) \right),$$

for every $i \in S$ and history h. We define a deterministic policy π' as follows. For every $i \in S$ and history h, let $\pi'(i,h) = a'$ where $a' = argmin_{a \in \pi(i,h)} c_a + \sum_{j \in S} p_{a,j}\, Val^{\pi}(\Gamma, i, h.(a,j))$. Let us define the intermediate policies $\pi^{(r)}$, $r = 0, 1, ...$ as the behavioral policies that coincide with π' for the first r steps and then coincide with π. Clearly $\pi^{(0)} = \pi, \lim_{r \to \infty} Val^{\pi^{(r)}}(\Gamma, i) = Val^{\pi'}(\Gamma, i)$. The following claim ends the proof.

Claim. $Val^{\pi^{(r+1)}}(\Gamma, i, h) \leq Val^{\pi^{(r)}}(\Gamma, i, h)$ for every $i \in S$ and history h.
Proof. By induction on r. Notice that if the history is of length $> r$ then $Val^{\pi^{(r+1)}}(\Gamma, i, h) = Val^{\pi^{(r)}}(\Gamma, i, h)$. We continue by backward induction on $|h| = r, ..., 0$.
Base case. Assume $|h| = r$.

$$Val^{\pi^{(r+1)}}(\Gamma, i, h) = c_{a'} + \sum_{j \in S} p_{a',j}\, Val^{\pi}(\Gamma, i, h.(a', j)) \leq Val^{\pi}(\Gamma, i, h) = Val^{\pi^{(r)}}(\Gamma, i, h).$$

Inductive step. $|h| < r$ and therefore

$$Val^{\pi^{(r+1)}}(\Gamma, i, h) = c_{a'} + \sum_{j \in S} p_{a',j}\, Val^{\pi^{(r+1)}}(\Gamma, i, h.(a, j)) \leq c_{a'} + \sum_{j \in S} p_{a',j}\, Val^{\pi^{(r)}}(\Gamma, i, h.(a, j)) =$$

$Val^{\pi^{(r)}}(\Gamma, i, h)$.

Exercise 1.2 (a) Describe a simple linear time algorithm for finding all the states from which a Markov chain stops with probability 1. (The running time should be linear in the number of non-zero transition probabilities.) (Hint: Start by finding all the states from which the Markov chain stops with a positive probability.)

(b) Extend the algorithm to find all states of a Markov Decision Process (MDP) from which the process ends with probability 1, no matter what the controller does. The running time should still be linear.

Solution 1.2 (a) Convert the Markov chain into a directed graph $G = (V \cup \{s\}, E)$, where V is the set of states of the Markov chain and s is a sink. For every $i, j \in V$, $(i, j) \in E$ if and only if $p_{i,j} > 0$. Similarly, $(i, s) \in E$ if and only if $\sum_j p_{i,j} < 1$. Let S be the set of vertices from which there is a directed path to s. The set S can be found by a BFS from s on the graph obtained by reversing the direction of the edges of G. Let $T = V \setminus S$ be the set of vertices from which there is no directed path to s. Clearly from all states of T the chain never stops, i.e., stops with probability 0. Now, let $S_0 \subseteq S$ be the subset of vertices of S from which there is a directed path to a vertex in T. The set S_0 can again be found in linear time. The set S_0 contains all vertices from which the chain stops with a probability larger than 0 but strictly smaller than 1. To see this note that if $i \in S_0$, then there is a directed path from i to a vertex in T and also a directed path from i to s. The chain moves along each one of these paths with positive probabilities. Finally, $S_1 = S \setminus S_0$ is the set of vertices from which the chain stops with probability 1. Note that S_1 is a closed set, i.e., if $i \in S_1$ and $(i, j) \in S_1$, then $j \in S_1$. Thus, if the chain starts in a state from S_1 it can never leave S_1. For each $i \in S_1$ there is a path P_i from i to s. The length of the path is at most $|S_1| \leq n$, where n is the number of states of the chain. Let α_i be the product of the probabilities on P_i. Thus, starting from i there is a probability of at least α_i of reaching the sink, i.e., stopping, after at most n steps. Let $\alpha = \min_{i \in S_1} \alpha_i$. Then, from every vertex $i \in S_1$ there is a probability of at least α of stopping after at most n steps. Thus, the probability of not stopping after kn steps is at most $(1-\alpha)^k$. As k tends to infinity, this tends to 0, thus the probability of stopping, from every vertex of S_1, is indeed 1.

(b) The algorithm can be extended to work for MDPs as follows. We begin by finding the set of states S from which there is a positive probability of stopping, no matter what the controller does. The set S can be found by working backward from s. Initialize $S \leftarrow \{s\}$. As long as there is a state $i \notin S$ such that for every action $a \in A_i$ there exists $j \in S$ with $p_{a,j} > 0$, add i to S. It is possible to implement this process in linear time. When the process ends, S is the required set. Now $T = V \setminus S$ is the set of states from which the controller has a policy under which the MDP

never stops. Next, we find the subset $S_0 \subseteq S$ from which the controller has a policy of reaching a state of T. Then, $S_1 = S \setminus S_0$ is the set of states from which the MDP stops with probability 1 no matter what the controller does. The proof that this is indeed the case is given in the solution to the next exercise.

Exercise 1.3 Show that if an MDP on n states is *stopping*, i.e., stops from each initial state with probability 1, no matter what the controller does, then there exists $\varepsilon > 0$ such that from each state the probability that the process stops after at most n steps is at least ε. (Hint: rely on the algorithm developed in the previous exercise.)

Solution 1.3 Let S_0, S_1 and T be as in the solution of Exercise 1.1(b). (If the MDP is stopping, then $S_1 = V$ and $S_0, T = \emptyset$. We consider the more general situation.) Partition S_1 into *layers* in the following way. Let $L_0 = \{s\}$. For every $k > 0$, $i \in L_k$ if and only if $i \in S_1$, $i \notin \cap_{r=0}^{k-1} L_r$ and for every action $a \in A_i$, there exists $j \in \cap_{r=0}^{k-1} L_r$ such that $p_{a,j} > 0$. The set $L_k \subseteq S_1$ is the set of states from which the MDP stops with positive probability within k steps, no matter what the controller does, and from which the process never leaves S_1, no matter what the controller does. Let m be the maximal k for which $L_k \neq \emptyset$. Then, $m \leq n$ and $S_1 = \cup_{k=1}^{m} L_k$. For $i \in L_k$, and $a \in A_i$, let $q_a = \min\{p_{a,j} \mid j \in \cup_{r=1}^{k-1} L_r\}$. By the definition of L_k we have $q_a > 0$. Let $q_i = \min\{q_a \mid a \in A_i\}$, and finally $\alpha_k = \min\{q_i \mid i \in L_k\}$. Note that $\alpha_k > 0$. If $i \in L_k$, then there is a probability of at least $\prod_{r=1}^{k} \alpha_r$ that the MDP stops within the next k steps, no matter what the controller does. We can thus take $\varepsilon = \prod_{r=1}^{m} \alpha_r$.

Exercise 1.4 Show that if P is the probability transition matrix of a stopping Markov chain, then $(I - P)^{-1} = \sum_{r \geq 0} P^r$. (Note that $P^0 = I$.) (Prove that the infinite sum exists and is equal to the inverse.) In particular, $(I - P)^{-1}$ exists, all its entries are non-negative, and all entries on the diagonal are at least 1. (Hint: rely on the previous exercise to show that the sum is bounded by a geometric series.)

Solution 1.4 By Exercise 1.2, there exists $\varepsilon > 0$ such that the from each starting state, the chain has a probability of at least ε of stopping within n steps. Recall that $P^r_{i,j}$ is the probability of moving from state i to state j in r steps. Thus, for every $k \geq 1$, we have $P^{kn}_{i,j} \leq (1-\varepsilon)^k$, and consequently $P^r_{i,j} \leq (1-\varepsilon)^{\lfloor r/n \rfloor}$ for every $r \geq 0$. Thus, $\sum_{r \geq 0} P^r_{i,j} \leq n \sum_{k \geq 0} (1-\varepsilon)^k < \infty$. Hence, $\sum_{r \geq 0} P^r$ exists. Now $\left(\sum_{r \geq 0} P^r\right)(I - P) = \sum_{r \geq 0} P^r - \sum_{r \geq 1} P^r = P^0 = I$, hence $\sum_{r \geq 0} P^r = (I - P)^{-1}$, as claimed. As all entries in P are non-negative, all entries in P^r and hence in $(I - P)^{-1}$ are non-negative. Finally $P^0 = I$ and hence all the entries on the diagonal of $(I - P)^{-1}$ are at least 1.

Exercise 1.5 (a) Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be the *value iteration operator* of a discounted MDP, i.e., $T(\mathbf{y})_i = \min_{a \in A_i} c_a + \lambda \sum_{j \in S} p_{a,j} y_j$, for $\mathbf{y} = (y_i) \in \mathbb{R}^n$, $i \in S$, where $0 < \lambda < 1$ is the *discount factor*. Prove that T is a *contraction*, i.e., that $||T(\mathbf{x}) - T(\mathbf{y})||_\infty \leq \lambda ||\mathbf{x} - \mathbf{y}||_\infty$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

(b) Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be the *value iteration operator* of a *stopping* MDP on n states, i.e., $T(\mathbf{y})_i = \min_{a \in A_i} c_a + \sum_{j \in S} p_{a,j} y_j$, for $\mathbf{y} \in \mathbb{R}^n$, $i \in S$. Prove that $T^{(n)}$, i.e., T iterated n times, is a *contraction*, i.e., there exists $0 < \lambda < 1$ such that $||T^{(n)}(\mathbf{x}) - T^{(n)}(\mathbf{y})||_\infty \leq \lambda ||\mathbf{x} - \mathbf{y}||_\infty$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

(c) In the previous item, prove that T is *not* necessarily a contraction.

Solution 1.5 (a) Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and let $i \in [n]$. Let $a \in A_i$ be the action for which $T(\mathbf{y})_i = c_a + \lambda \sum_{j \in S} p_{a,j} y_j$. Then $T(\mathbf{x})_i \leq c_a + \lambda \sum_{j \in S} p_{a,j} x_j$. Thus,

$$T(\mathbf{x})_i - T(\mathbf{y})_i \;\leq\; \lambda \sum_{j \in S} p_{a,j}(x_j - y_j) \;\leq\; \lambda ||\mathbf{x} - \mathbf{y}||_\infty \,.$$

Similarly, we can show that $T(\mathbf{y})_i - T(\mathbf{x})_i \leq \lambda||\mathbf{x}-\mathbf{y}||_\infty$. Hence $|T(\mathbf{x})_i - T(\mathbf{y})_i| \leq \lambda||\mathbf{x}-\mathbf{y}||_\infty$ and $||T(\mathbf{x}) - T(\mathbf{y})||_\infty \leq \lambda||\mathbf{x}-\mathbf{y}||_\infty$.

(b) As the MDP is stopping, there exists $\varepsilon > 0$ for which the probability that the process stops within n steps is at least ε, no matter what the controller does. As we saw, $T^{(n)}(\mathbf{x})_i$ is the value of an n-step game that starts at i with $\mathbf{x}$ as a terminal cost vector. Let π be an optimal policy for the controller in this finite game. (We know that π is deterministic and that it can be found using backward induction.) Let $p^\pi_{i,j}$ be the probability that the chain moves from i to j in n steps when the controller uses policy π. Then, $T^{(n)}(\mathbf{x})_i = c_\pi + \sum_{j\in S} p^\pi_{i,j} x_j$, where c_π is a constant that depends on π but not on $\mathbf{x}$. (Can you write down an explicit expression for c_π?) Now $T^{(n)}(\mathbf{y})_i \leq c_\pi + \sum_{j\in S} p^\pi_{i,j} y_j$. Thus,

$$T^{(n)}(\mathbf{y})_i - T^{(n)}(\mathbf{x})_i \;\leq\; \sum_{j\in S} p^\pi_{i,j}(y_j - x_j) \;\leq\; \left(\sum_{j\in S} p^\pi_{i,j}\right) ||\mathbf{y}-\mathbf{x}||_\infty \;\leq\; (1-\varepsilon)\,||\mathbf{y}-\mathbf{x}||_\infty\,.$$

Similarly we can show that $T^{(n)}(\mathbf{x})_i - T^{(n)}(\mathbf{y})_i \leq (1-\varepsilon)\,||\mathbf{y}-\mathbf{x}||_\infty$ and hence $|T^{(n)}(\mathbf{x})_i - T^{(n)}(\mathbf{y})_i| \leq (1-\varepsilon)\,||\mathbf{y}-\mathbf{x}||_\infty$ and thus $||T^{(n)}(\mathbf{x}) - T^{(n)}(\mathbf{y})||_\infty \leq (1-\varepsilon)\,||\mathbf{y}-\mathbf{x}||_\infty$. Thus $T^{(n)}$ is a contraction with $\lambda = 1-\varepsilon$.

(c) Consider the following (deterministic) MDP, where the shaded circle represents the sink: Then, $T(x_1, x_2) = (x_2, 0)$ and this is not a contraction. On the other hand $T^{(2)}(x_1, x_2) = (0, 0)$ which is a contraction.

Exercise 1.6 (a) Write down the primal and dual LPs that correspond to an MDP with a discount factor $0 < \lambda < 1$. (b) Show directly, without relying on fact that the optimality equations for discounted MDPs have a solution, that the dual LP has at least one feasible solution. (c) Show directly that the dual LP for discounted MDPs is bounded. (d) Show that the dual LP for discounted MDPs has an optimal solution which is also a solution of the optimality equations for discounted MDPs.
(Bonus: Solve Exercise 1.6 for non-discounting MDPs that satisfy the stopping condition.)

Solution 1.6 (a) Dual:

$$\begin{aligned}
&\max \sum_{i\in S} y_i \\
&y_i \leq c_a + \lambda \cdot \sum_{j\in S} p_{a,j} \cdot y_j, \;\; a \in A_i
\end{aligned}$$

Primal:

$$\begin{aligned}
&\min \sum_{a\in A} c_a \cdot x_a \\
&\sum_{a\in A_i} x_a - \lambda \cdot \sum_{a\in A} p_{a,i} \cdot x_a = 1, && i \in S \\
&x_a \geq 0 && a \in A
\end{aligned}$$

(b) Let $c_{min} = \min_{a \in A} c_a$. If $c_{min} \geq 0$, then $y = 0$ is a feasible solution. Otherwise, let $y = x \cdot (1, 1, ..., 1)$, where $x = c_{min}/(1 - \lambda)$. Thus

$$y_i - \lambda \cdot \sum_{j \in S} p_{a,j} \cdot y_j = x \cdot \left(1 - \lambda \cdot \sum_{j \in S} p_{a,j} \right) \leq x \cdot (1 - \lambda) = c_{min} \leq c_a,$$

for every $i \in S, a \in A_i$.
(c) Let y be a feasible solution. Let $c_{max} = \max_{a \in A} c_a$. WLOG $y_1 = \max_i(y_i)$, it is enough to provide a bound on y_1. Let $a \in A_1$,

$$y_1 \ \leq \ c_a + \lambda \cdot \sum_{j \in S} p_{a,j} \cdot y_j \ \leq \ c_{max} + \lambda \cdot \sum_{j \in S} p_{a,j} \cdot y_1 \ \leq \ c_{max} + \lambda \cdot y_1 \ ,$$

and therefore $y_1 \leq c_{max}/(1 - \lambda)$.
(d) By (b)+(c) the dual LP has an optimal solution y. If y is not a solution to the optimality equations then $y_i < \min_{a \in A_i} c_a + \lambda \cdot \sum_{j \in S} p_{a,j} \cdot y_j \equiv T$. Notice that increasing y_i to T (without changing any of the other y values) keeps the solution feasible, a contradiction to y being optimal solution to the dual LP.