Exercise Sheet 2 - Solutions*

Omar D. Domingues
omar (dot) darwiche-domingues (at) inria.fr

Last update: January 18, 2019

1 Exercise 1 (Finite-Horizon Problems and Dynamic Programming)

1.1 Notation

A stochastic and non-stationary policy $\pi = (\pi_1, \pi_2, \dots, \pi_{T-1})$ is such that $\pi_t(a|s) = P(A_t = a|S_t = s)$ represents the probability of choosing the action a in state s at time t.

If the policy π_t is deterministic, that is, $\pi_t(\bar{a}_s|s) = 1$ for some action $\bar{a}_s$, we write π_t as a function $\pi_t : \mathcal{S} \to \mathcal{A}$ from the state space $\mathcal{S}$ to the action space $\mathcal{A}$ such that $\pi_t(s) = \bar{a}_s$.

1.2 Bellman Equation

In this exercise, we consider a finite-horizon problem. The agent must choose actions in order to maximize the sum of the rewards up to a fixed time T. The value function at time t is defined as:

$$V_t^\pi(s) = E_\pi\left[\sum_{n=t}^{T-1} r_t(S_n, A_n) + r_T(S_T)\middle| S_t = s\right] \tag{1}$$

and represents how much reward is collected by following the policy $\pi = (\pi_1, \pi_2, \dots, \pi_{T-1})$ starting from time t. Thus, the agent seeks to find a policy π that maximizes V_1^π. Note that the reward function at the final time T does not depend on the action: this is because the interactions between the agent and the environment ends at time T, and the agent's action does not matter once the time is up.

Imagine that π is given and we wish to compute $V_1^\pi(s)$. A very naive computation is to do the following sum:

$$V_1^\pi(s) = \sum_{a_1, a_2, \dots, a_{T-1}} \sum_{s_1, s_2, \dots, s_T} \left(\sum_{n=1}^{T-1} r_t(s_n, a_n) + r_T(s_T)\right) P(s_1, a_1, s_2, a_2, s_3, a_3, \dots s_{T-1}, a_{T-1}, s_T | S_1 = s_1) \tag{2}$$

Let $|\mathcal{S}|$ and $|\mathcal{A}|$ be the number of states and the number of actions, respectively. Each term in the sum above requires $\mathcal{O}(T)$ multiplications (to compute the joint probability from the conditional ones) and we have a sum over $\mathcal{O}(T|\mathcal{S}|^T|\mathcal{A}|^T)$ terms. This gives a total of $\mathcal{O}(T^2|\mathcal{S}|^T|\mathcal{A}|^T)$ operations, which is huge!

We can find a smarter way to compute it by rewriting $V_t^\pi(s)$ as follows:

*Only for the exercises solved during the course.

$$
\begin{aligned}
V_t^\pi(s) &= E_\pi\left[\sum_{n=t}^{T-1} r_t(S_n, A_n) + r_T(S_T)\middle| S_t = s\right] \\
&= E_\pi\left[r_t(S_t, A_t) + \sum_{n=t+1}^{T-1} r_t(S_n, A_n) + r_T(S_T)\middle| S_t = s\right] \\
&= E_\pi\left[r_t(S_t, A_t)\middle| S_t = s\right] + E_\pi\left[E_\pi\left[\sum_{n=t+1}^{T-1} r_t(S_n, A_n) + r_T(S_T)\middle| S_{t+1}, S_t = s\right]\middle| S_t = s\right] \\
&= E_\pi\left[r_t(S_t, A_t)\middle| S_t = s\right] + E_\pi\left[V_{t+1}^\pi(S_{t+1})\middle| S_t = s\right]
\end{aligned} \tag{3}
$$

Finally, we obtain:

$$
V_t^\pi(s) = \sum_a \pi_t(a|s)\left[r_t(s,a) + \sum_{s'} p(s'|s,a) V_{t+1}^\pi(s')\right] \tag{4}
$$

and, by definition, we have:

$$
V_T(s) = E_\pi[r_T(S_T)|S_T = s] = r_T(s) \text{ for all } s \in \mathcal{S} \tag{5}
$$

Equation 4 is called *Bellman Equation* for evaluating the policy π, which is a dynamic programming approach.

Now, we can find V_1^π by the following procedure:

1. Initialization: set $t = T$, $V_T(s) = r_T(s)$;
2. Compute V_{t-1} using equation 4;
3. Set $t \leftarrow t - 1$;
4. If $t = 1$ stop, otherwise go to step 2.

At each iteration t, equation 4 requires $\mathcal{O}(1)$ multiplications and a sum over $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ terms to compute V_t for one state. This gives $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$ operations per iteration. Since we have T iterations, we have a total of $\mathcal{O}(T|\mathcal{S}|^2|\mathcal{A}|)$ operations, which is much less than the naive approach! This is the interest of using dynamic programming.

1.3 Bellman Optimatily Equation - Intuition

In this exercise we want to prove that the optimal value function $V_t^*(s) := \max_\pi V_t^\pi(s)$ satisfies the equations:

$$
\begin{aligned}
&V_t^*(s) = \max_{a \in \mathcal{A}}\left[r_t(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) V_{t+1}^*(s')\right] \text{ for } 1 \le t \le T-1 \\
&V_T^*(s) = r_T(s)
\end{aligned} \tag{6}
$$

First, remember that our policy is non stationary and is written as $\pi = (\pi_1, \pi_2, \ldots, \pi_{T-1})$. Note that $V_t^\pi(s)$ depends only on the actions taken starting from time t. Let's define $\pi_{t:T-1} := (\pi_t, \pi_2, \ldots, \pi_{T-1})$. With a slight abuse of notation, we can write $V_t^\pi = V_t^{\pi_{t:T-1}}$.

Now, let's see the intuition behind equation 6. We have, by using equation 4:

$$
\begin{aligned}
V_t^*(s) &= \max_{\pi} V_t^{\pi}(s) \\
&= \max_{\pi} \sum_a \pi_t(a|s) \left[r_t(s,a) + \sum_{s'} p(s'|s,a) V_{t+1}^{\pi}(s') \right] \\
&= \max_{(\pi_t, \pi_{t+1:T-1})} \sum_a \pi_t(a|s) \left[r_t(s,a) + \sum_{s'} p(s'|s,a) V_{t+1}^{\pi}(s') \right] \\
&= \max_{\pi_t} \sum_a \pi_t(a|s) \left[r_t(s,a) + \max_{\pi_{t+1:T-1}} \sum_{s'} p(s'|s,a) V_{t+1}^{\pi}(s') \right] \\
&= \max_{\pi_t} \sum_a \pi_t(a|s) \left[r_t(s,a) + \max_{\pi_{t+1:T-1}} \sum_{s'} p(s'|s,a) V_{t+1}^{\pi_{t+1:T-1}}(s') \right] \\
&= \max_{a} \left[r_t(s,a) + \max_{\pi_{t+1:T-1}} \sum_{s'} p(s'|s,a) V_{t+1}^{\pi_{t+1:T-1}}(s') \right]
\end{aligned}
\tag{7}
$$

where the second line comes from equation 4 and the last line comes from the fact that $\sum_a \pi_t(a|s) f(a) \leq \sum_a \pi_t(a|s) \max_a f(a) = \max_a f(a)$ for any function f and equality can be achieved by choosing $\pi_t(a|s)$ that puts probability 1 for one action $\bar{a}$ such that $\bar{a} \in \operatorname{argmax}_a f(a)$.

Observe that:

$$
V_{t+1}^*(s') = \max_{\pi_{t+1:T-1}} V_{t+1}^{\pi_{t+1:T-1}}(s') \tag{8}
$$

Thus, if we can invert the max and the sum operator in equation 7, we prove our desired result, which is equation 6. But can we do this formally?

1.4 Bellman Optimality Equation - Formal Proof

Let $\bar{V}_t$ be such that:

$$
\begin{aligned}
&\bar{V}_t(s) = \max_{a \in \mathcal{A}} \left[r_t(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) \bar{V}_{t+1}^*(s') \right] \text{ for } 1 \leq t \leq T-1 \\
&\bar{V}_T(s) = r_T(s)
\end{aligned}
\tag{9}
$$

We will prove that $\bar{V}_t(s) = V_t^*(s)$ for all $s \in \mathcal{S}$ and for all t in two steps:

1. Step 1: prove that $V_t^*(s) \leq \bar{V}_t(s)$
2. Step 2: prove that $\bar{V}_t(s) \leq V_t^*(s)$;

1.4.1 Step 1

Let's proceed by induction on t:

- For $t = T$, we have $\bar{V}_T(s) = V_T^*(s) = r_T(s)$;
- Assume that $\bar{V}_n^*(s) \leq \bar{V}_n(s)$ for all $n \in \{t+1, \ldots, T\}$ and for all s.

We have, for an arbitrary policy π and all s:

$$
\begin{aligned}
V_t^\pi(s) &= \sum_a \pi_t(a|s)\left[r_t(s,a) + \sum_{s'} p(s'|s,a)V_{t+1}^\pi(s')\right] \\
&\leq \max_a \left[r_t(s,a) + \sum_{s'} p(s'|s,a)V_{t+1}^\pi(s')\right] \\
&\leq \max_a \left[r_t(s,a) + \sum_{s'} p(s'|s,a)V_{t+1}^*(s')\right] \\
&\leq \max_a \left[r_t(s,a) + \sum_{s'} p(s'|s,a)\bar{V}_{t+1}(s')\right] \\
&= \bar{V}_t(s)
\end{aligned}
\tag{10}
$$

Since π is arbitrary, we have $\forall\pi,\ V_t^\pi(s) \leq \bar{V}_t(s) \implies \max_\pi V_t^\pi(s) \leq \bar{V}_t(s)$ which is what we wanted to prove.

1.4.2 Step 2

Define a deterministic policy $\pi' = (\pi'_1, \pi'_2, \ldots, \pi'_{T-1})$ such that:

$$
\begin{aligned}
\pi'_{T-1}(s) &\in \operatorname*{argmax}_a \left[r_{T-1}(s,a) + \sum_{s'} p(s'|s,a)r_T(s')\right] \\
\pi'_t(s) &\in \operatorname*{argmax}_a \left[r_t(s,a) + \sum_{s'} p(s'|s,a)V_{t+1}^{\pi'}(s')\right]
\end{aligned}
\tag{11}
$$

It can be proven by induction that $V_t^{\pi'}(s) = \bar{V}_t(s)$ for all s. Finally:

$$
\bar{V}_t(s) = V_t^{\pi'}(s) \leq \max_\pi V_t^\pi(s) = V_t^*(s) \text{ for all } s \tag{12}
$$