# 1 Properties of Bellman Operators

Let $\mathcal{M} := (\mathcal{S}, A, r, P, \lambda)$ be a MDP. We denote by $L_\pi$ the linear operator for a static deterministic policy $\pi$. For any vector $V$:

$$L_\pi V := r_\pi + \lambda P_\pi V$$

$L$ is the bellman operator:

$$LV := max_\pi L_\pi V$$

1. Observe $L$ and $L_\pi$ are monotonous:
Let $V$, $V'$ such that $V \leq V'$ componentwise. Let's fix some policy $\pi$.
As $P_\pi$ is a stochastic matrix and $\lambda > 0$:

$$V > V' \Leftrightarrow r_\pi + \lambda P_\pi V > r_\pi + \lambda P_\pi V' \Leftrightarrow L_\pi V > L_\pi V'$$

This being true for any fixed policy; it is also true for the best one:

$$V > V' \Leftrightarrow LV > LV'$$

2. $L$ and $L_\pi$ are homogenous: Let $c \in \mathbb{R}$, and let $\pi$ be any policy.
Observe that, as $P_\pi$ is a stochastic matrix, $P_\pi \mathbb{I} = \mathbb{I}$; $\mathbb{I}$ being the unit vector.
Hence:

$$L_\pi(V + c\mathbb{I}) = r_\pi + \lambda P_\pi V + \lambda P_\pi c\mathbb{I} = L_\pi V + \lambda c\mathbb{I}$$

This being true for any fixed policy, it is also true for the best one:

$$L(V + c\mathbb{I}) = LV + \lambda c\mathbb{I}$$

3.

# 2 Multiarmed Bandits

We consider a bandit as a finite set of $n$ arms. Each arm $i$ has $S$ possible states, a reward vector $r_i$ and a transition probability matrix $P_i$ . Here is the evolution of the bandit used by a player:
At each round $t \in \mathbb{N}$, the arms are in states, say $s = (s_1(t), s_2(t), ..., s_n(t))$.
The player decides to use one arm among the $S$ possible arms (say arm $i$). He gets a reward $\lambda^t r_i(s_i(t))$ and arm $i$ moves to a new state $s_i'$ with probability

$P_i(s_i'|s_i)$. The other arms stay in their current state. The player wants to maximize the sum of its rewards over an infinite horizon.

1. Let $\mathcal{S} := \{s = (s_{1,j}, .., s_{n,j}) | j \in [[1, S]]\}$ be the set of states.
Let $\mathcal{A}_s := \{j | j \in [[1, S]]\}$ be the action set; which do not depends on $s$. For all arm $i$, let $(P_i)_{l,c} := P_i(s_l|s_c)$, the probability to pass from state $s_c$ to state $s_l$ when using arm $i$, and let $r_i(s) = r_i(s_i)$ be the reward vector. If $\lambda \in ]0, 1[$; then $(\mathcal{S}, \mathcal{A}_s, P_i, r_i)$ defines a Markov Decision Process discounted by $\lambda$.

2. Now let's consider a particular arm $i_0$. In the next few questions assume the dropping of indexes. Consider a new game where the controller has the choice at each step to stop and earn $M$[1]; or action the arm, move to a new state according to the probability matrix associated earn his reward [1], and start a new step.

Let $W(s, M)$ be the optimal gain expected to earn over an infinite horizon, starting in state $s$.

It is clear that it is equal either to $M$, either to the gain of the arm in state $s$, plus $W(s', M)$; where $s'$ is a state reached from state $s$; discounted by $\lambda$. In other words:

$$W(s, M) = max(M, r(s) + \lambda \sum_{s'} P(s'|s) W(s', M))$$

We can write the $W(s, M)$ inside a vector $W_M := (W(s, M))_{s \in [[1,S]]}$ and $R := (r(s))_{s \in [[1,S]]}$ so that $W_M$ verifies:

$$W_M = max(M, R + \lambda P W)$$

3. Let $M* := R + \lambda P W$.

- Suppose $M < M*$. Then $W(s, M) = M*$. Consequently, if $M_1 < M_2 < M*$, then $M* = W(s, M_1) \leq W(s, M_2) = M*$.
- If $M* < M$, then $W(s, M) = M$. Consequently, if $M_1 > M_2 > M*$, then $M1 = W(s, M_1) \geq W(s, M_2) = M2$.
- Finally, if $M_1 < M* < M_2$, then $M* = W(s, M_1) \leq W(s, M_2) = M_2$.

We conclude that $W(s, M)$ is increasing in M.

[1] being discounted by $\lambda$, of course.

Suspose $M < \frac{r_{min}}{1-\lambda}$, where $r_{min} = min_s(R)$.
Imagine a scenario in which we always use the lever, and always earn the minimal reward of the machine. The total gain over an infinite horizon would be: $\sum_{t=0}^{\infty} \lambda^t r_{min} = \frac{r_{min}}{1-\lambda}$. Logically, if $M$ induce a lower gain than the one in the worst scenario possible, then it is never a good choice to stop to earn $M$. So we must have $W = R + \lambda PW = LW$ where $L$ is the Bellman operator. Hence $W$ is the fixed point of $L$.

Suppose $M > \frac{r_{max}}{1-\lambda}$, where $r_{max} = max_s(R)$.
Similarty; the best scenario possible would give us $\frac{r_{max}}{1-\lambda}$; so it is always a better choice to stop and take $M$. Hence $W(s, M) = M$.