M2 ENS Lyon: MDP and RL.

1 Optimality for all large discounts

Let us consider a MDP $\mathcal{M} = (S, A, r, P)$, with state space S, action space A, transitions P, rewards r. A stationary policy π is a function from the state space to the action space: $\pi(s) \in A$ is the action taken by policy π in state s. Under discount β, $(0 < \beta < 1)$ the discounted value of policy π starting in s at time 0 is:

$$V_\beta^\pi(s) = \mathbb{E} \sum_{t=0}^{\infty} \beta^t r(X_t, \pi(X_t)).$$

The undiscounted gain of π is

$$g^\pi(s) = \mathbb{E} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} r(X_t, \pi(X_t)),$$

where, in both equations, X_t is the state of $\mathcal{M}$ at time t under π.

Let $r^\pi(s) := r(s, \pi(s))$ denote the reward under π in state s and P^π the transition matrix under π: The probability to go from state i to j is

$$P^\pi(i, j) := P(j|i, \pi(i)).$$

1.1

Explain why the matrix $(I - \beta P^\pi)$ is always invertible for $0 < \beta < 1$.

1.2

The Cramer formula for the inverse of a matrix is $M^{-1} = \frac{1}{\det(M)} \begin{pmatrix} C_{i,j} \end{pmatrix}_{i,j}$. The coefficients of this matrix are $C_{i,j} := (-1)^{i+j} \det(M \setminus \{i, j\})$ where $M \setminus \{i, j\}$ is the matrix M where row i and column j are removed.

By using this formula, show that $V_\beta^\pi(s)$ is a rational function: $V_\beta^\pi(s) = \frac{F(\beta)}{G(\beta)}$, where F is a polynomial function of degree $\leq n - 1$ and G is a polynomial function of degree $\leq n$, and G is never null on the open interval $(0, 1)$.

1.3

Let π' be another policy, show that $V_\beta^\pi(s) - V_\beta^{\pi'}(s)$ is also a rational function of β with a non-null denominator. What is the maximal degree of the numerator? What is the maximal number of values for β in the open interval $(0, 1)$ where this function can be equal to 0 (if it is not the null function).

1.4

Show that there exists $\beta^o < 1$ such that for all $\beta \in (\beta^o, 1)$, the discounted values of any pair of policies and any state s always compare in the same way.

1.5

Show that there exists a policy π^o that is discount optimal for all $\beta \in (\beta^o, 1)$.

1.6

Do you think that policy π^o is gain optimal (also maximizes the gain g^π)? Do you think that any gain optimal policy π^* is also discount optimal for all $\beta \in (\beta^o, 1)$? Explain your answers (no formal proof is required here).