M2 ENS Lyon: MDP and RL.

1 Complexity of Policy Iteration

Recall the following definitions:

We consider a MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P)$ with state space $\mathcal{S}$ of size S, action space $\mathcal{A}$ of size A, rewards $r(s,a)$ are all assumed to be non-negative and transition probabilities are $P(y|s,a)$. The discount factor is denoted λ.

The discounted value over an infinite horizon of policy $\pi = (d, d, \cdots)$ is denoted V_π or V_d. Here d denotes a decision function ($d : \mathcal{S} \to \mathcal{A}$). The optimal value is denoted V^*. The optimal policy is denoted $\pi^* = (d^*, \cdots)$. This means that $V^* = V_{\pi^*} = V_{d^*}$.

In the following, we mostly use d instead of π because here, all policies are stationary: $\pi = (d, d, d \cdots)$.

Let V be any vector in $\mathbb{R}^S$. The norm $||V||_\infty = \max_i |V_i|$. The norm of a matrix in $\mathbb{R}^S \times \mathbb{R}^S$ is $||M||_\infty = \max_i \sum_j |M_{ij}|$.

The Bellman operator is $\mathcal{L} : V \to \max_d (r_d + \lambda P_d V)$.

The value operator under policy $\pi = (d, d, \cdots, d)$ is $L_d : V \to r_d + \lambda P_d V$.

Recall *Policy Iteration (PI)*:

d_0: arbitrary decision function.
Repeat
- $V_k := L_{d_k} V_k$
- $d_{k+1} := argmax(\mathcal{L} V_k)$
Until $d_{k+1} = d_k$.

For simplicity we denote L_{d_k} as L_k in the following.

Question 1: Show that $L_k V_k \geq L_k V_{k-1}$ and $L_k V_{k-1} \geq L_{d^*} V_{k-1}$. Explain why this implies $V_k \geq L_{d^*} V_{k-1}$.

Question 2: We denote by V_k (as in the algorithm) the value under policy (d_k) and V^* the optimal value. Show that $||V^* - V_k||_\infty$ is λ-contracting: $||V^* - V_k||_\infty \leq \lambda ||V^* - V_{k-1}||_\infty$.

Hint: use $V^* = \mathcal{L} V^* = L_{d^*} V^*$ and use $V_k \geq L_{d^*} V_{k-1}$.

Question 3: We introduce the gap function: $\Delta(d, d') = V_d - L_{d'} V_d$. It "measures" how much d is better that d' under value V_d.

Show the following identities:
$V_{d'} - V_d = (I - \lambda P_{d'})^{-1}(-\Delta(d, d'))$ and $V_{d'} - V_d = (I - \lambda P_d)^{-1} \Delta(d', d)$.

Hint: Use $V_{d'} = (I - \lambda P_{d'})^{-1} r_{d'}$ and $V_d = (I - \lambda P_{d'})^{-1}(I - \lambda P_{d'}) V_d$.

Question 4: Show that $\Delta(d^*, d_k) \leq V^* - V_k$ and $\Delta(d^*, d_k) \geq 0$.

Question 5: Show that $||I - \lambda P_d||_\infty = \frac{1}{1-\lambda}$ for any d.

Question 6: Show that $||\Delta(d^*, d_k)||_\infty \leq \frac{\lambda^k}{1-\lambda}||\Delta(d^*, d_0)||_\infty$.

Question 7: Let s_0 be the state acheiving the infinite norm in $||\Delta(d^*, d_0)||_\infty$. Show that $\Delta(d^*, d_k)(s_0) \leq \frac{\lambda^k}{1-\lambda}\Delta(d^*, d_0)(s_0)$.

Question 8:

Explain why if d_0 is not optimal then $\Delta(d^*, d_0)(s_0) > 0$.

Define $k_\lambda = \lceil \frac{\log(1-\lambda)}{\log(\lambda)} \rceil$. Then for all $k > k_\lambda$, $\frac{\lambda^k}{1-\lambda} < 1$.
(*) Explain why action $d(s_0)$ is never taken in all the policies d_k , $k \geq k_\lambda$.

Question 9: Show that Policy iteration takes at most $S(A-1)k_\lambda$ steps.