M2 ENS Lyon: MDP and RL.

Optimality of Gittins Index Policy for Rested Bandits

This proof is inspired from the proof of Whittle. There exits several alternative proof, as such the proof of Weber, Tsitsiklis, ... The proof here is shorter but somehow artificial and seems to come from nowhere.

Recall the following definitions. We consider a multi-arm bandit with n arms. Each arm a is a reward Markov chain with reward $r^a(s_a)$ in global state $s = (s_1, , , , s_a, ..., s_n)$ and transition matrix P^a. The discount factor is denoted λ.

At each point in time, the controller activates one arm $(a(t))$ and gets the reward of the activated arm. The goal is to find a policy that chooses the arms $a(t)$ at time t that maximizes the discounted reward $\mathbb{E} \sum_{t=1}^{\infty} \lambda^{t-1} r^{a(t)}(s_{a(t)}(t))$.

Index definition:

To define the index of all the states in one arm, we focus on a single arm (index a in isolation. We introduce a new parameter, $M \in \mathbb{R}$, the reward of stopping. Now in each state s_a of the arm, a controller decides to continue (action 1) or stop (action 0). More precisely: Under action 1, it gets immediate reward $r^a_1(s_a) := r^a(s_a)$ and moves to the next state according to $P^a_1(s_a, \cdot) := P^a(s_a, \cdot)$. Under action 0, it gets immediate reward $r^a_0(s_a) := M$ and moves to the end state E: $P^a_0(s_a, E) := 1$. In the end state, the arm cannot move anymore and gets a null reward from this time on.

The value operator under policy $\pi = (d, ...)$ is $L^a_d : V \to r^a_d + \lambda P^a_d V$. Here d denotes a decision function $(d : \mathcal{S} \to \mathcal{A})$.

The discounted value vector under policy $\pi = (d, d, \cdots)$ is denoted v^a_π or v^a_d.

The optimal value is denoted $v^{a,*}$.

Question 1: Show that the local Bellman equation of this arm MDP can be written (making the dependence on M explicit): $v^{a,*}(M) = \max(M, L^a_1 v^{a,*}(M))$.

Question 2:

Let $r^a_{\text{min}} := \min_{s_a} r^a(s_a)$ and $r^a_{\text{max}} :== \max_{s_a} r^a(s_a)$.

- Show that is $M \leq r^a_{\text{min}}/(1-\lambda)$ then $v^{a,*}(M) = L_1 v^{a,*}(M)$.

- Show that if $M \geq r^a_{\text{max}}/(1-\lambda)$ then $v^{a,*}(M) = M$.

Show that if $r^a_{\text{min}}/(1-\lambda) < M < r^a_{\text{max}}/(1-\lambda)$ then $v^{a,*}_{s_a}(M)$ is piece-wise affine, increasing and convex, for all s_a.

Let us define the index $I^a(s_a)$ of state s_a of arm a as the smallest value of M such as $v^{a,*}_{s_a}(M) = M$.

Question 3:

Let us go back to the bandit problem (with all the arms). We consider a more general problem as the original one: at any point in time the controler may activate any arm (as in the original case) but can also decide to stop, in which case, it gets a final reward of M.

Show that the optimal value of this bandit problem, $V^*(M)$ is the unique solution that satisfies the following global Bellman equation:

$$V^*(M) = \max(M, \max_a L_1^a V^*(M)). \tag{1}$$

Question 4:

Let us define the function

$$\Phi_s(M) := B - \int_M^B \prod_a \frac{dv_{s_a}^{a,*}(m)}{dm} dm$$

where $B := \max_a r_{\max}^a/(1-\lambda)$.

- Show that $\Phi(M)$ is well defined (the final goal will be to show that $\phi(M)$ satisfies the global Bellman equation above).

In the following we will use the Steiljes integral to be able to integrate against dg when g is not a continuous function.

Basically, $\int f dg$ is well defined and satisfies all the properties of the classical integral if f is continous and g is discontinuous with bounded jumps.

Question 4: Using integration by part, show that using the Steiljes integral definition, for any a,

$$\Phi_s(M) = v_{s_a}^{a,*}(M) Q_s^a(M) + \int_M^\infty v_{s_a}^{a,*}(m) dQ_s^a(m),$$

where

$$Q_s^a(M) := \prod_{b \neq a} \frac{dv_{s_b}^{b,*}(m)}{dm}$$

Question 5: show that $Q_s^a(M)$ is non-negative, non-decreasing and equal to 1 for $M \geq \max_{b\neq a} I^b(s_b)$.

Question 5: In the following the state $s = (s_1, ..., s_a, ..s_n)$ is fixed and not always explicitly written.

Let us define for each arm $\delta^a(M) := v^{a,*}(M) - L_1^a v^{a,*}(M)$. Show that $\delta^a \geq 0$ and equals 0 if $M \leq I^a$.

Question 6: Show that $\Phi_s(M) \geq M$ with equality when $M \geq \max_a I^a(s_a)$.

Question 7: Show that

$$\Phi(M) - L_1^a \Phi(M) = \delta^a(M) Q^a(M) + \int_M^\infty \delta^a(m) dQ^a(m),$$

where implicitly, L_1^a only acts on the a-th components of Φ.

-Show that $\Phi(M) - L_1^a \Phi(M) \geq 0$ with equality if $M \leq I^a$ and $I^a \geq \max_b I^b$.

Question 8: Show that $\phi(M) = V^*(M)$ and that the optimal policy follows the Gittins indexes: the arm with the best index is activated.