MDP with Random Rewards

Appendix to Part 2 – EL 2805

Alexandre Proutiere

Department of Automatic Control
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

1 Definitions

- Time horizon $T < \infty$;
- State space S, assumed to be finite;
- Action space: let A_s denote the set of possible actions in state s;
- Dynamics: At time t, when the system is in state s and the action is $a \in A_s$, the system evolves to state y with probability $p_t(y|s,a)$;
- Random rewards: at time t, when in state s and when the selected action is $a \in A_s$, the agent receives a reward R_t, sampled from a distribution $q_t(\cdot|s,a)$. We denote by $\mu_t(s,a)$ the average of R_t.

2 Bellman's equation

Let $u_t^\star(s)$ denote the maximal average reward starting from state s at time t. Then we have:

For all s, $u_T^\star(s) = \max_{a \in A_s} \mu_T(s,a)$.

For all $t < T$, for all s,

$$u_t^\star(s) = \max_{a \in A_s} \left(\mu_t(s,a) + \sum_{j \in S} p_t(j|s,a) u_{t+1}^\star(j) \right).$$

In the case of random rewards, policy evaluation and Bellman's equation are obtained by replacing the rewards by their *avergae*.