Exercises - Chapter 17

Carl Fredriksson, c@msp.se

Exercise 17.1

This section has presented options for the discounted case, but discounting is arguably inappropriate for control when using function approximation (Section 10.4). What is the natural Bellman equation for a hierarchical policy, analogous to (17.4), but for the average reward setting (Section 10.3)? What are the two parts of the option model, analogous to (17.2) and (17.3), for the average reward setting?

My answer:

I believe the general form of the equation stays the same, but the two parts of the option model changes. For the reward part $r(s, \omega)$, we subtract $r(\pi)$ (the average reward under π) every time step and remove the discounting:

$$r(s, \omega) \doteq \mathbb{E}\big[R_1 - r(\pi) + R_2 - r(\pi) + R_3 - r(\pi) + \cdots + R_\tau - r(\pi) | S_0 = s, A_{0:\tau-1} \sim \pi_\omega, \tau \sim \gamma_\omega\big]$$
$$r(s, \omega) = \mathbb{E}\big[R_1 + R_2 + R_3 + \cdots + R_\tau - r(\pi)\tau | S_0 = s, A_{0:\tau-1} \sim \pi_\omega, \tau \sim \gamma_\omega\big]$$

For the state-transition part $p(s'|s, \omega)$ we simply have to remove the discounting:

$$p(s'|s, \omega) \doteq \sum_{k=1}^{\infty} \Pr\{S_k = s', \tau = k | S_0 = s, A_{0:k-1} \sim \pi_\omega, \tau \sim \gamma_\omega\}$$