5.5 Evaluating One Policy While Following Another

So far we have considered methods for estimating the value functions for a policy given an infinite supply of episodes generated using that policy. Suppose now that all we have are episodes generated from a *different* policy. That is, suppose we wish to estimate V^π or Q^π, but all we have are episodes following π', where $\pi' \neq \pi$. Can we learn the value function for a policy given only experience "off" the policy?

Happily, in many cases we can. Of course, in order to use episodes from π' to estimate values for π, we require that every action taken under π is also taken, at least occasionally, under π'. That is, we require that $\pi(s,a) > 0$ implies $\pi'(s,a) > 0$. In the episodes generated using π', consider the ith first visit to state s and the complete sequence of states and actions following that visit. Let $p_i(s)$ and $p_i'(s)$ denote the probabilities of that complete sequence happening given policies π and π' and starting from s. Let $R_i(s)$ denote the corresponding observed return from state s. To average these to obtain an unbiased estimate of $V^\pi(s)$, we need only weight each return by its relative probability of occurring under π and π', that is, by $p_i(s)/p_i'(s)$. The desired Monte Carlo estimate after observing n_s returns from state s is then

$$V(s) = \frac{\sum_{i=1}^{n_s} \frac{p_i(s)}{p_i'(s)} R_i(s)}{\sum_{i=1}^{n_s} \frac{p_i(s)}{p_i'(s)}}. \qquad (5.3)$$

This equation involves the probabilities $p_i(s)$ and $p_i'(s)$, which are normally considered unknown in applications of Monte Carlo methods. Fortunately, here we need only their ratio, $p_i(s)/p_i'(s)$, which *can* be determined with no knowledge of the environment's dynamics. Let $T_i(s)$ be the time of termination of the ith episode involving state s. Then

$$p_i(s_t) = \prod_{k=t}^{T_i(s)-1} \pi(s_k, a_k) \mathcal{P}_{s_k s_{k+1}}^{a_k}$$

and

$$\frac{p_i(s_t)}{p_i'(s_t)} = \frac{\prod_{k=t}^{T_i(s)-1} \pi(s_k, a_k) \mathcal{P}_{s_k s_{k+1}}^{a_k}}{\prod_{k=t}^{T_i(s)-1} \pi'(s_k, a_k) \mathcal{P}_{s_k s_{k+1}}^{a_k}} = \prod_{k=t}^{T_i(s)-1} \frac{\pi(s_k, a_k)}{\pi'(s_k, a_k)}.$$

Thus the weight needed in (5.3), $p_i(s)/p_i'(s)$, depends only on the two policies and not at all on the environment's dynamics.

Exercise 5.3 What is the Monte Carlo estimate analogous to (5.3) for *action* values, given returns generated using π'?

Mark Lee 2005-01-04