5.7 Incremental Implementation

Monte Carlo methods can be implemented incrementally, on an episode-by-episode basis, using extensions of techniques described in Chapter 2. They use averages of *returns* just as some of the methods for solving n-armed bandit tasks described in Chapter 2 use averages of *rewards*. The techniques in Sections 2.5 and 2.6 extend immediately to the Monte Carlo case. They enable Monte Carlo methods to process each new return incrementally with no increase in computation or memory as the number of episodes increases.

There are two differences between the Monte Carlo and bandit cases. One is that the Monte Carlo case typically involves multiple situations, that is, a different averaging process for each state, whereas bandit problems involve just one state (at least in the simple form treated in Chapter 2). The other difference is that the reward distributions in bandit problems are typically stationary, whereas in Monte Carlo methods the return distributions are typically nonstationary. This is because the returns depend on the policy, and the policy is typically changing and improving over time.

The incremental implementation described in Section 2.5 handles the case of simple or arithmetic averages, in which each return is weighted equally. Suppose we instead want to implement a *weighted* average, in which each return R_n is weighted by w_n, and we want to compute

$$V_n = \frac{\sum_{k=1}^{n} w_k R_k}{\sum_{k=1}^{n} w_k}. \qquad (5.4)$$

For example, the method described for estimating one policy while following another in Section 5.5 uses weights of $w_n(s) = p_n(s)/p'_n(s)$. Weighted averages also have a simple incremental update rule. In addition to keeping track of V_n, we must maintain for each state the cumulative sum W_n of the weights given to the first n returns. The update rule for V_n is

$$V_{n+1} = V_n + \frac{w_{n+1}}{W_{n+1}}\Big[R_{n+1} - V_n\Big] \qquad (5.5)$$

and

$$W_{n+1} = W_n + w_{n+1}$$

where $W_0 = 0$.

Exercise 5.5 Modify the algorithm for first-visit MC policy evaluation (Figure 5.1) to use the incremental implementation for stationary averages described in Section 2.5.

Exercise 5.6 Derive the weighted-average update rule (5.5) from (5.4). Follow the pattern of the derivation of the unweighted rule (2.4) from (2.1).

Exercise 5.7 Modify the algorithm for the off-policy Monte Carlo control algorithm (Figure 5.7) to use the method described above for incrementally computing weighted averages.

Next Up Previous Contents

Next: 5.8 Summary **Up:** 5. Monte Carlo Methods **Previous:** 5.6 Off-Policy Monte Carlo **Contents**

Mark Lee 2005-01-04