5.4 On-Policy Monte Carlo Control

How can we avoid the unlikely assumption of exploring starts? The only general way to ensure that all actions are selected infinitely often is for the agent to continue to select them. There are two approaches to ensuring this, resulting in what we call *on-policy* methods and *off-policy* methods. On-policy methods attempt to evaluate or improve the policy that is used to make decisions. In this section we present an on-policy Monte Carlo control method in order to illustrate the idea.

In on-policy control methods the policy is generally *soft*, meaning that $\pi(s,a)>0$ for all $s\in\mathcal{S}$ and all $a\in\mathcal{A}(s)$. There are many possible variations on on-policy methods. One possibility is to gradually shift the policy toward a deterministic optimal policy. Many of the methods discussed in Chapter 2 provide mechanisms for this. The on-policy method we present in this section uses ε*-greedy* policies, meaning that most of the time they choose an action that has maximal estimated action value, but with probability ε they instead select an action at random. That is, all nongreedy actions are given the minimal probability of selection, $\frac{\varepsilon}{|\mathcal{A}(s)|}$, and the remaining bulk of the probability, $1-\varepsilon+\frac{\varepsilon}{|\mathcal{A}(s)|}$, is given to the greedy action. The ε-greedy policies are examples of ε*-soft* policies, defined as policies for which $\pi(s,a)\geq\frac{\varepsilon}{|\mathcal{A}(s)|}$ for all states and actions, for some $\varepsilon>0$. Among ε-soft policies, ε-greedy policies are in some sense those that are closest to greedy.

The overall idea of on-policy Monte Carlo control is still that of GPI. As in Monte Carlo ES, we use first-visit MC methods to estimate the action-value function for the current policy. Without the assumption of exploring starts, however, we cannot simply improve the policy by making it greedy with respect to the current value function, because that would prevent further exploration of nongreedy actions. Fortunately, GPI does not require that the policy be taken all the way to a greedy policy, only that it be moved *toward* a greedy policy. In our on-policy method we will move it only to an ε-greedy policy. For any ε-soft policy, π, any ε-greedy policy with respect to Q^π is guaranteed to be better than or equal to π.

That any ε-greedy policy with respect to Q^π is an improvement over any ε-soft policy π is assured by the policy improvement theorem. Let π' be the ε-greedy policy. The conditions of the policy improvement theorem apply because for any $s\in\mathcal{S}$:

$$\begin{array}{ll}
Q^\pi(s,\pi'(s)) & = \sum_a \pi'(s,a)Q^\pi(s,a) \\
 & = \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) \;+\; (1-\varepsilon)\max_a Q^\pi(s,a) \\
 & \geq \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) \;+\; (1-\varepsilon)\sum_a \frac{\pi(s,a)-\frac{\varepsilon}{|\mathcal{A}(s)|}}{1-\varepsilon}Q^\pi(s, \\
\begin{array}{l}\text{(the sum is a weighted average with nonnegative weights summing to 1, and} \\ \text{as such it must be less than or equal to the largest number averaged)}\end{array} & = \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) \;-\; \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) \;+\; \sum_a \pi(s, \\
 & = V^\pi(s).
\end{array}$$

Thus, by the policy improvement theorem, $\pi'\geq\pi$ (i.e., $V^{\pi'}(s)\geq V^\pi(s)$, for all $s\in\mathcal{S}$). We now prove that equality can hold only when both π' and π are optimal among the ε-soft policies, that is, when they are better than or equal to all other ε-soft policies.

Consider a new environment that is just like the original environment, except with the requirement that policies be ε-soft "moved inside" the environment. The new environment has the same action and state set as the original and behaves as follows. If in state s and taking action a, then with probability $1-\varepsilon$ the new environment behaves exactly like the old environment. With probability ε it repicks the action at random, with equal probabilities, and then behaves like the old environment with the new, random action. The best one can do in this new environment with general policies is the same as the best one could do in the original environment with ε-soft policies. Let $\widetilde{V}^*$ and $\widetilde{Q}^*$ denote the optimal value functions for the new environment. Then a policy π is optimal among ε-soft policies if and only if $V^\pi=\widetilde{V}^*$. From the definition of $\widetilde{V}^*$ we know that it is the unique solution to

$$\begin{aligned}
\widetilde{V}^*(s) &= (1-\varepsilon)\max_a \widetilde{Q}^*(s,a)+\frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a \widetilde{Q}^*(s,a) \\
&= (1-\varepsilon)\max_a \sum_{s'}\mathcal{P}^a_{ss'}\Big[\mathcal{R}^a_{ss'}+\gamma\widetilde{V}^*(s')\Big] \\
&\quad + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a\sum_{s'}\mathcal{P}^a_{ss'}\Big[\mathcal{R}^a_{ss'}+\gamma\widetilde{V}^*(s')\Big].
\end{aligned}$$

When equality holds and the ε-soft policy π is no longer improved, then we also know, from (5.2), that

$$\begin{aligned}
V^\pi(s) &= (1-\varepsilon)\max_a Q^\pi(s,a)+\frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) \\
&= (1-\varepsilon)\max_a \sum_{s'}\mathcal{P}^a_{ss'}\Big[\mathcal{R}^a_{ss'}+\gamma V^\pi(s')\Big] \\
&\quad + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a\sum_{s'}\mathcal{P}^a_{ss'}\Big[\mathcal{R}^a_{ss'}+\gamma V^\pi(s')\Big].
\end{aligned}$$

However, this equation is the same as the previous one, except for the substitution of V^π for $\widetilde{V}^*$. Since $\widetilde{V}^*$ is the unique solution, it must be that $V^\pi=\widetilde{V}^*$.

In essence, we have shown in the last few pages that policy iteration works for ε-soft policies. Using the natural notion of greedy policy for ε-soft policies, one is assured of improvement on every step, except when the best policy has been found among the ε-soft policies. This analysis is independent of how the action-value functions are determined at each stage, but it does assume that they are computed exactly. This brings us to roughly the same point as in the previous section. Now we only achieve the best policy among the ε-soft policies, but on the other hand, we have eliminated the assumption of exploring starts. The complete algorithm is given in Figure 5.6.

```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \leftarrow$ arbitrary
    $Returns(s,a) \leftarrow$ empty list
    $\pi \leftarrow$ an arbitrary $\varepsilon$-soft policy

Repeat forever:
    (a) Generate an episode using $\pi$
    (b) For each pair $s, a$ appearing in the episode:
            $R \leftarrow$ return following the first occurrence of $s, a$
            Append $R$ to $Returns(s,a)$
            $Q(s,a) \leftarrow$ average$(Returns(s,a))$
    (c) For each $s$ in the episode:
            $a^* \leftarrow \arg\max_a Q(s,a)$
            For all $a \in \mathcal{A}(s)$:
            $\pi(s,a) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = a^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq a^* \end{cases}$
```

Figure 5.6: An ε-soft on-policy Monte Carlo control algorithm.

Next Up Previous Contents
Next: 5.5 Evaluating One Policy **Up:** 5. Monte Carlo Methods **Previous:** 5.3 Monte Carlo Control **Contents**

Mark Lee 2005-01-04