5.3



5.4



5.5

First-visit estimator is 10

Every-visit estimator is 10/9

5.6



5.7

If the behavior is inconsistent with the target policy, whose importance ratio is zero, it will contribute neither to the numerator nor denominator of (5.6). For the first several episodes, the numerator will increase more quickly for most of the importance ratio will be zero. Then, the denominator will increase forever but the numerator will decrease since the rewards can be -1.

5.8

When we use every-visit MC method, the variance of the estimator is still infinite.

5.9

Input: a policy pi to be evaluated.

Initialize:

, arbitrarily, for all 

 <- an empty list, for all 

The number of the s state being visited. 

Loop forever ( for each episode):

Generate an episode following   
 

Loop for each step of episodic, 



Unless appears in 





5.10



5.11

In off-policy MC control, the next step action is the greedy action which is deterministic, so the probability is 1.