Notes while reading Reinforcement learning an introduction (Sutton/Barto)

Willem

January 2013

Contents

Chapter 1

Introduction

Chapter 2

Mutli-armed bandits

Chapter 3

Finite Markov Decision Process

Chapter 4

Dynamic Programming

4.1 Exercise 4.8

The reward is only obtained when the capital is above 99. When the capital is at 50, there is a 50% chance you can win the game. So this obviously is the optimal policy. When you reach 51: it would be rather odd to bet the entire capital, as you don't need to risk it all to reach 100. Bigger downside, but same upside. So the best course of action is to bet with 1, see if you can grow this above 50. If you lose it, you still have a 50% chance to win by betting it all.

Chapter 5

Monte Carlo Methods

5.1 Exercises

5.1.1 Exercise 5.1 page 94

The last 2 rows in the rear means you either have 21, or 20, which means the odd's are very good you will win. (hence high value function)

The last row on the left means the dealer has an ace, so it's at an advantage to get a higher score.

The front row's are higher on the upper diagram, as there is a usuable ace. Which means that if you get a bad hit that put's you over 21. It can count as 1.

5.1.2 Exercise 5.2 page 94

As this is Markov process eg. The cards drawn are not exhaustible. The odds of winning on the second time your in the same state is just as good as the first time.

5.1.3 Exercise 5.4 page 99

The "Append G to Returns (S_t, A_t) would be replaced by increasing a count and added it as running average to some table.

5.1.4 Exercise 5.5 page 105

10 Steps means 9 towards the non-terminal, and one towards the terminal. The rewards are all-way's the same so the final cost=10.

If $\gamma = 1$ then $G = G + \gamma R_{k+1}$ in every iteration.

In case of all visit the complete horizon counts 10 times in the non-terminal state, as the 10th time we leave the non-terminal state for good and enter the terminal state. (1+2+3+4+5+6+7+8+9+10)/10 = 55/10 = 5.5 So the value is 5.

In case of the first-visit, we only count the first visit which has a reward of 1.

5.1.5 Exercise 5.6 page 108

Q(s, a) is similar to V(s), it takes the V(s) given a certain step was taken first.

$$Q(s,a) = \frac{\sum_{t \in J(s,a)} \rho_{t+1:T(t)-1} G_t}{\sum_{t \in J(s,a)} \rho_{t+1:T(t)-1}} \tag{5.1}$$

5.1.6 Exercise 5.7 page 108

If there are but a few samples, the bias will be the dominating error. And it will increase as more and more samples are added. Until there are so many samples, it starts to disappear.

5.1.7 Exercise 5.8 page 108

A first Visit MC has less terms then a every Visit MC. All terms have a positive value, so it would also go to infinite.

5.1.8 Exercise 5.11 page 111

If the target policy is a greedy deterministic policy, and the loop is broken off if $\pi(S_t) \neq A_t$. Then $\pi(A_t|S_t) = 1$ by definition.