The University of Vermont

CS253A QR: Reinforcement Learning: Assignment №3

Ayat Ospanov

September 17, 2018

Contents

1 Exercise 2.6: Mysterious Spikes

When given optimistic initial values, $Q_n(a) > R_n, \forall a$ at the first steps. This means, that

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n) \leq Q_n$$

Therefore, all next step values are less (or equal) than the values at the previous step. As we initialized all starting values with a priori big values, there are lots of equal maximum values. Consequently, there is a big rate of optimal actions (optimal by algorithm) corresponding to those values. When all Q_ns drop under real high values of distributions, the curve stabilizes. This effect was also mentioned in the Exercise 2.5.

2 Exercise 2.7: Unbiased Constant-Step-Size Trick

As we have shown in the exercise 2.4, Q_{n+1} for variable α is:

$$Q_{n+1} = \prod_{i=1}^{n}(1-\alpha_i)Q_1 + \sum_{i=1}^{n}\left(\alpha_i \prod_{j=i+1}^{n}(1-\alpha_j)R_i\right)$$

Here the weight given to R_i is $\alpha_i \prod_{j=i+1}^{n}(1-\alpha_j)$. This is exponential. As we have $\alpha_i = \beta_i$, we have to show that $\beta_i \in [0,1]$ to prove that Q_n is an exponential recency-weighted average. Given $\bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1})$,

$$\beta_n = \frac{\alpha}{\alpha + (1-\alpha)\bar{o}_{n-1}}$$

By definition, $\bar{o}_n \geq 0, \forall n \geq 0$ and $\alpha \in [0,1]$. Thus, $\alpha + (1-\alpha)\bar{o}_{n-1} \geq \alpha$. Knowing this, it is obvious that $\beta_n \in [0,1]$.

Now, let's show that Q_n doesn't have the initial bias. To do this, have a look at Q_2:

$$Q_2 = Q_1 + \beta_1(R_1 - Q_1)$$

Since $\bar{o}_0 = 0$,

$$\beta_1 = \frac{\alpha}{\alpha + (1-\alpha)\bar{o}_0} = \frac{\alpha}{\alpha + (1-\alpha)0} = 1$$

This means, that $Q_2 = Q_1 + R_1 - Q_1 = R_1$.

We have proven that Q_n is an exponential recency-weighted average without initial bias.

3 Exercise 3.1

Examples	States	Actions	Rewards
Self-driving car	Surroundings: e.g. distance to objects, etc.	Steering, accelerating, braking, ...	Getting to destination without crashes, overall distance to objects, etc.
Towers of Hanoi	All possible disk states	Move the top disk to another tower	Negative (e.g. -1) for a move, so we do less steps
Checkers/Chess	Positions of pieces	Moves of each piece	Winning the game

4 Exercise 3.2

No. There are cases when it is impossible to use (classic) MDP: when there are infinite number of states or actions; when they are continuous; when the environment is non-stationary.

5 Exercise 3.3

It depends on the reliability of the actions. If you are sure the body does all actions (accelerator, steering wheel, and brake) (almost) without failures, we can select these actions. As the higher level actions depend on lower level ones, we need to be sure they are effective and faultless. So in terms of driving a car, if we simulate a driver, we can reliably accelerate, brake and etc. a car because they are easy to simulate, thus we can take these actions as main actions. When it is possible to simulate the motion of a car, we can choose 'choices of where to drive' as actions.

6 Exercise 3.4

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
high	wait	high	r_{wait}	1
low	search	low	r_{search}	β
low	search	high	-3	$1-\beta$
low	wait	low	r_{wait}	1
low	recharge	high	0	1