Q6 Exercise 2.8 :

Solution: UCB action selection is given by:

$$A_t = \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

here, c is greater than 0 and controls the degree of exploration.

Initially, $N_t(a) = 0 \quad \forall a$

⇒ for any action a, until it has been selected once $N_t(a)$ remains 0.

⇒ All the actions are explored one by one in initial 10 time steps (in our case), no matter whatever the value of c is.

Once it has explored all the actions, it choose the greedy ~~one~~, hence average reward over 2000 runs increases.

→ One can view the square-root quantity as the variance or the uncertainity in the estimate of a's value as stated in Sutton.

⇒ More is the square root quantity, more the uncertainity in the estimate of an action.

After 11th step, N(a) of the action chosen at 10th step increases and uncertainity decreases. But at the same time, uncertainity for other actions increases as ln t increases. Therefore, UCB algorithm continue to explore at subsequent step, hence average reward decreases.

Now, for different values of c, one may see the different graph.

This is b/c as c increases, ~~uncertainity~~ the actions with higher uncertainity gets selected.

Therefore, if c=1, the spike is less prominent as it choose an action which has a good estimate rather than high uncertainity.

Similarly, when c=4, actions which are not greedy gets selected after 11th step b/c they have higher uncertainity