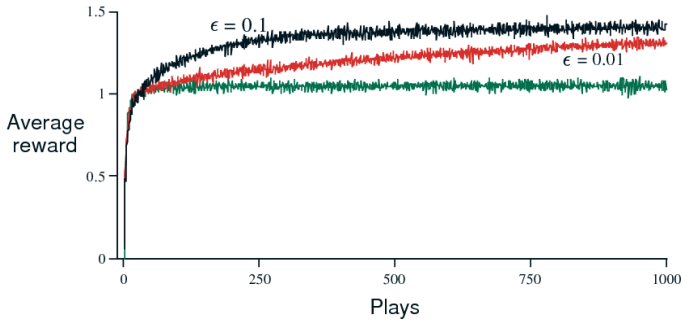
# Reinforcement Learning 2016

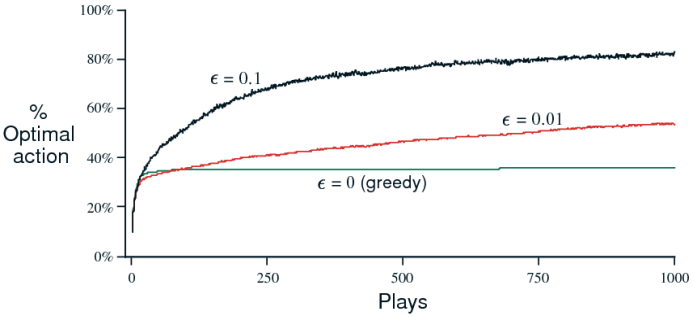
Tutorial 1 (week 3)

## Questions

1. Consider the comparison between ε-greedy methods shown in Figure 2.1 in the Sutton and Barto book. Which method will perform best in the long run in terms of cumulative rewards and cumulative probability of selecting the best action? How much better will it be?

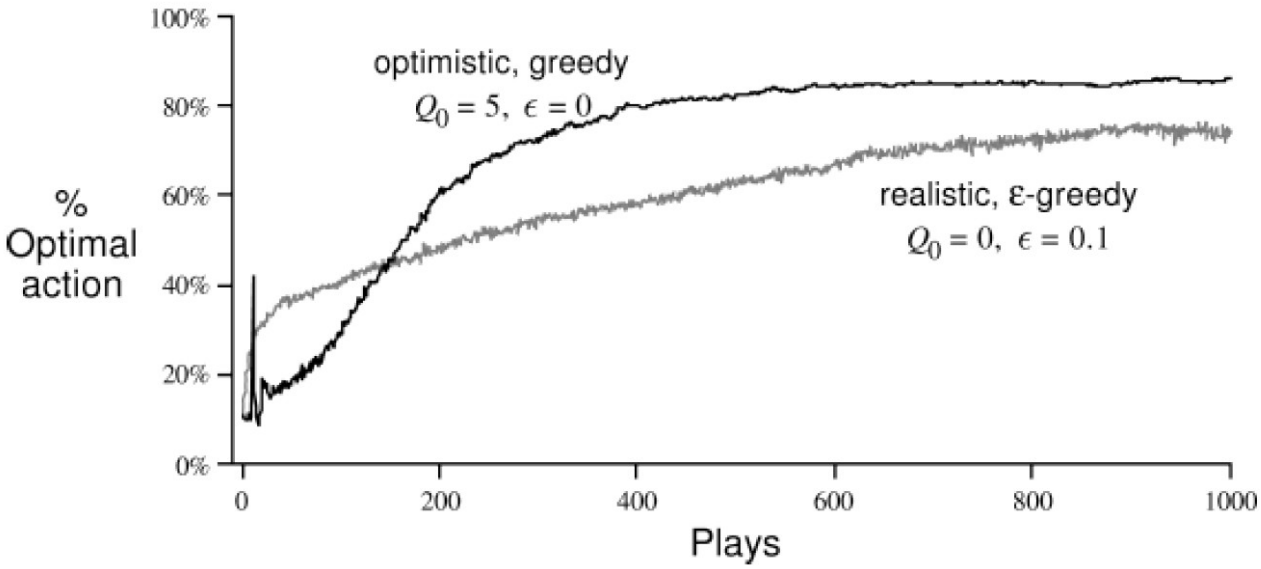



2. Show that in the case of two actions, the softmax operation using the Gibbs distribution becomes the logistic, or sigmoid, function commonly used in artificial neural networks. What effect does the temperature parameter have on the function?




3. In the incremental version of the action value estimation (exponential recency-weighted average, sec 2.6 in Sutton&Barto), if the step size parameter, $\alpha_k(a)$, is not constant, then the estimate $Q_k(a)$ is a weighted average of previously received rewards with a weighting different from the one in sec 2.6. What is the weighting on each prior reward in the general case?

4. Consider the optimistic initial value example, fig. 2.4 in S+B. This represents averages over 2000 individual, randomly chosen 10-armed bandit tasks, so the result should be reliable. How do you explain the oscillations and spikes in the early part of the curve for the optimistic method? What makes this method perform differently on particular early plays?




5. Suppose you face a binary bandit task whose true action values change randomly from play to play. Specifically, suppose that for any play the true values of action 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B). If you are unable to tell which case you face on any given play, what is the best expectation of success you can achieve and how should you behave to achieve it? Now suppose on each play you are told if you are facing case A or B (although you still do not know the true action values), what is the best expectation of success you can achieve in this task, and how should you behave to achieve it?

6. Discuss your solution of the 1D walker problem (see homework in lecture RL3). How do the

- initialisation
- alternative reward definitions
- exploration variants
- parameters and parameter decay schemes
- problem size

influence the solution? Trying out several combinations of the setting of the algorithm should be a group effort. Discuss which variants that have not been tried yet, appear to be worth trying?
What variant of the algorithm turns out (or is likely to be) most efficient?
Is this variant likely to generalise to other problems?

7. Assume that the rewards disappear after having been discovered (e.g. reward being food that is eaten up), so returning to a place where food has been found once might not be a rewarding action, unless the time has passed that is necessary to replenish the food source. Can this problem be solved by reinforcement learning? Explain your answer.