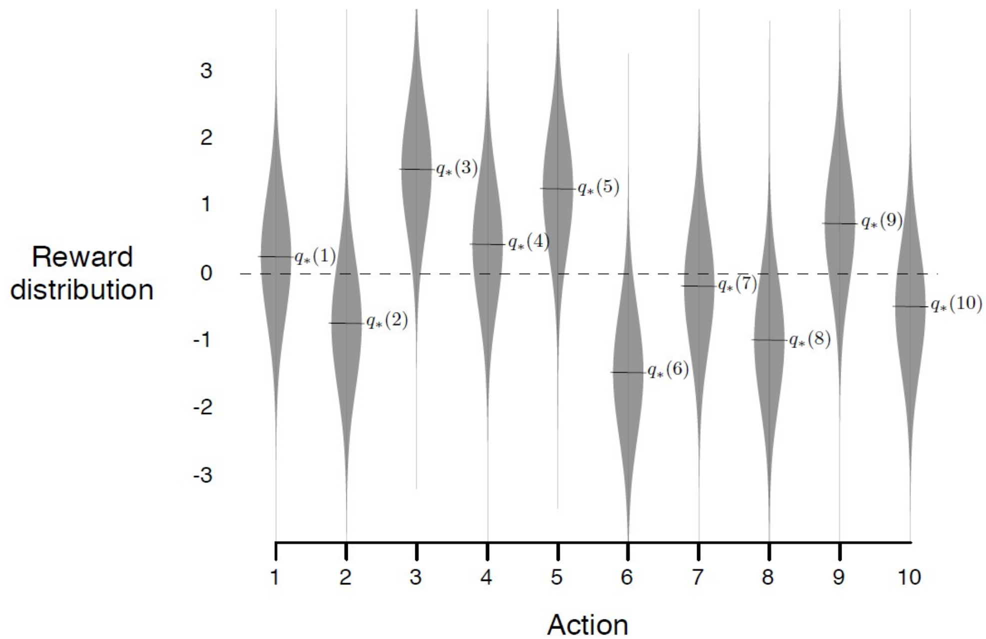
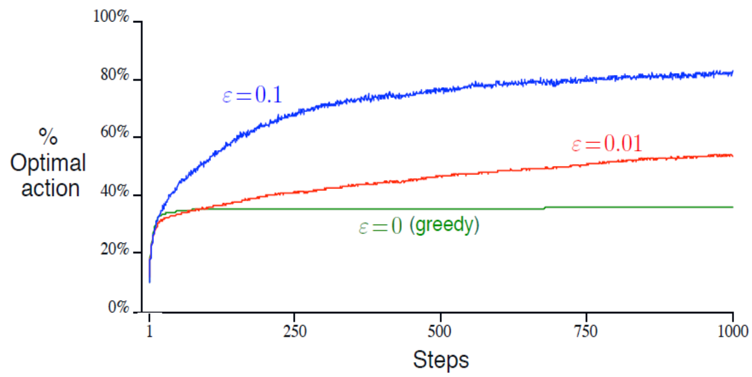
Exercise: k-armed bandits

We use the 10-armed bandit testbed from Sutton & Barto 2020

- The true value $q_*(a)$ of each of the ten actions a is selected according to a normal distribution with mean 0 and unit variance
- The actual rewards are selected according to a mean $q_*(a)$, and are also unit-variance normal distributed, as suggested by these gray distributions.
- Action-value estimates using the sample-average technique (with an initial estimate of 0)
  1. Run 1000 time steps for the generated 10-armed bandit problems and action-value algorithms
  2. Use 2000 randomly generated 10-armed bandit problems of this type.
     Run 1000 time steps for each of the 2000 randomly generated 10-armed bandit problems and action-value algorithms and average the results for each time step.



Task 1) Why does the value of the greedy algorithm's performance (on average) is above 10% optimal (which is what you would expect by just selecting randomly and sticking to this action? How could you exploit this behavior?



Task 2) Write a program that can reproduce the results of Sutton and Barto for the greedy and the ε-greedy algorithm (ε = 0.1 and ε=0.01) as shown in the lecture.
Use the simple bandit psydocode algorithm:

A simple bandit algorithm

Initialize, for $a = 1$ to k:
$Q(a) \leftarrow 0$
$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$
$N(A) \leftarrow N(A) + 1$
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

Average the return for each time step over all 2000 runs for the final evaluation of: step vs. average return.

Task 3) Change the maximum number of steps from 1000 to 2000.
Do the ε-greedy algorithms (ε = 0.1 and ε=0.01) converge?

Task 4) Change the true value $q_*(a)$ of 5 randomly selected actions (out of the 10 actions) at time step 1000 (i.e. now we have constructed a non-stationary problem). Run for 2000 time steps.
Can you explain the behavior of the greedy and ε-greedy algorithms?

Task 5) Add the results of a weighted average method with α = 0.9 and ε-greedy (ε = 0.01) action selection to your final plot. Change the true value $q_*(a)$ as in task 4. What do you observe, especially comparing the sample average to the weighted average method of ε-greedy (ε = 0.01) action selection? Can you explain the behavior?

$$Q_{n+1} \doteq Q_n + \alpha\Big[R_n - Q_n\Big]$$

Task 6) Add Upper-confident-bound action selection (UCB) c = 1 to your final plot. Change the true value $q_*(a)$ as in task 4.

Note that: If $N_t(a) == 0$, then a is considered to be a maximizing action

Note that: lim t→0 log(t) → -inf

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