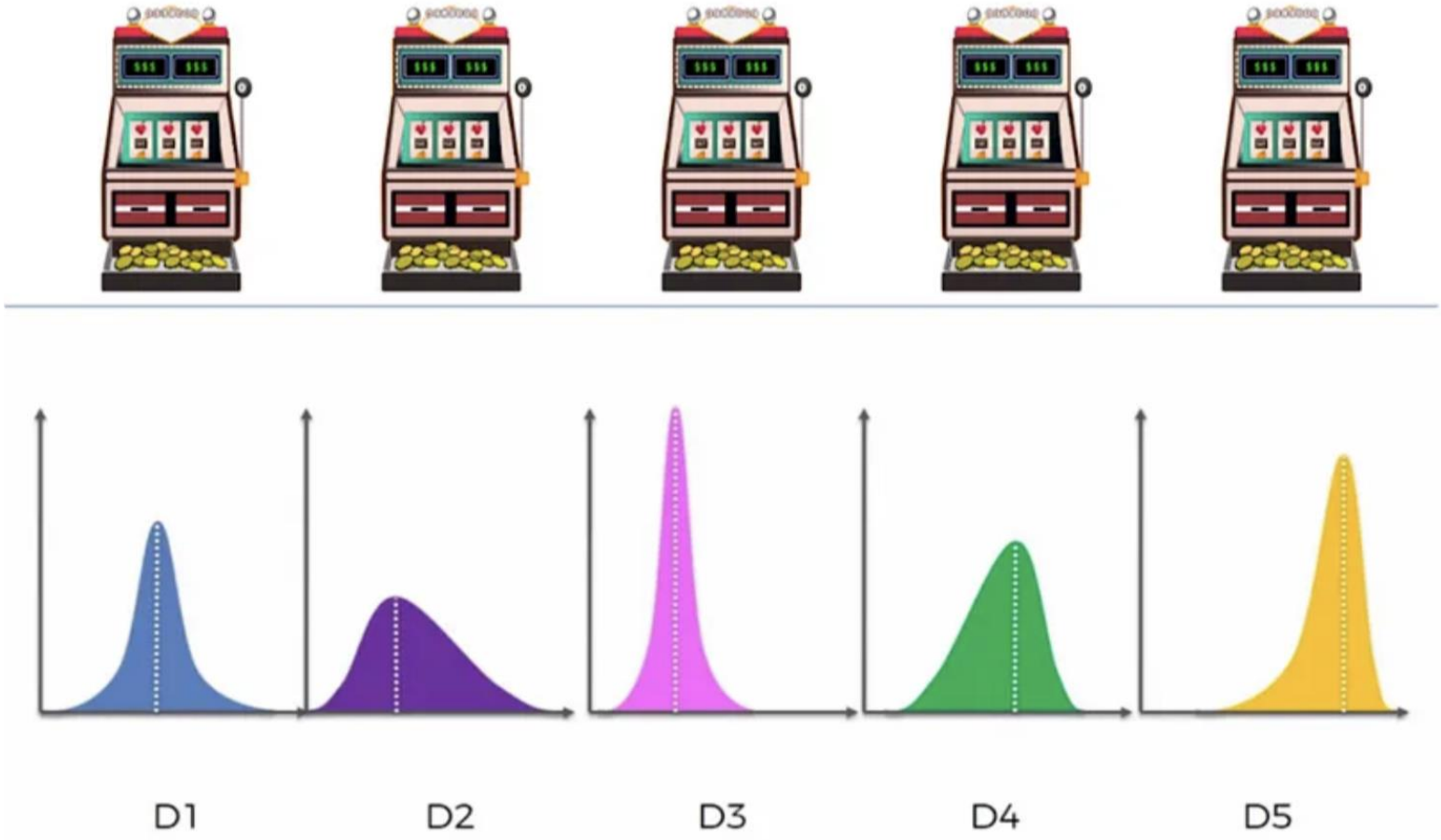
Sutton and Barto definition of RL

- “RL is learning how to map situations to actions so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In the most challenging cases, actions effect not only the immediate reward but also the next situation, and through that, all subsequent rewards. Trial-and-error search and delayed reward are the two most important distinguishing features of RL.”
- Through trial and error, the agent gathers data about the environment. Unlike supervised and unsupervised learning, data collection is part of the RL problem.

Simplest RL problem: Multi-armed Bandit



Multi-armed bandit special case

- “RL is learning how to map situations to actions so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. ~~In the most challenging cases, actions effect not only the immediate reward but also the next situation, and through that, all subsequent rewards.~~ Trial-and-error search ~~and delayed reward~~ are the two most important distinguishing features of RL.”

The k-armed bandit problem

- k options or “actions”
- A_t = action chosen at time step t
- R_t is corresponding reward
  - Assumed random but depends on A_t
- q(a) := E[R_t | A_t = a] (we will not use $q_*(a)$ in class)
- ***Goal:*** find $a \in \{1, \ldots, k\}$ that maximizes q(a)
  - And try to do this in as few steps as possible.

Quiz

1. Suppose we choose action a with probability $\pi(a)$. What is the $E[R_t]$ in terms of q(.) and $\pi(.)$?
2. Suppose the reward is given by $R_t = g(A_t)$ for some fixed deterministic function g(.). What is a good algorithm for finding the optimal action.

Quiz: Action-Value Methods

- Let $Q_t(a)$ is an estimate of q(a) going into time step t
- In words what is the sample average estimate $Q_t(a)$?
- Suppose we choose action a infinitely often. What is the $\lim_t Q_t(a)$? Why?
- What is the ”greedy action”?
- Suppose we first select each of the k actions, then only select the greedy action. Why is the problem with this?
- What is the ε-greedy policy?

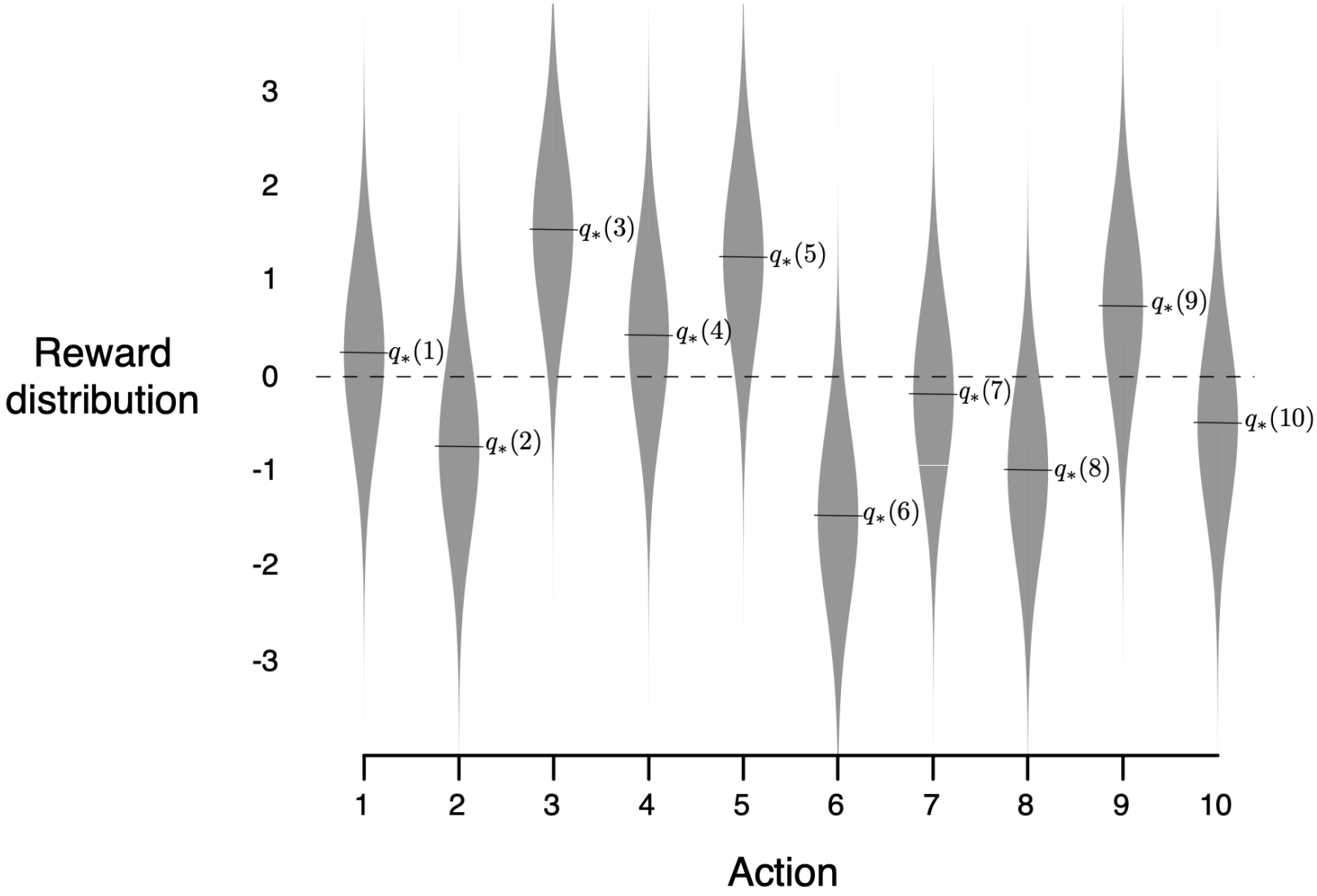
Quiz: ε-greedy policy

- With the ε-greedy policy, for the case k=2 and ε = 0.1, what is the probability that the greed action is selected?
- Generalize this result for arbitrary k and ε.

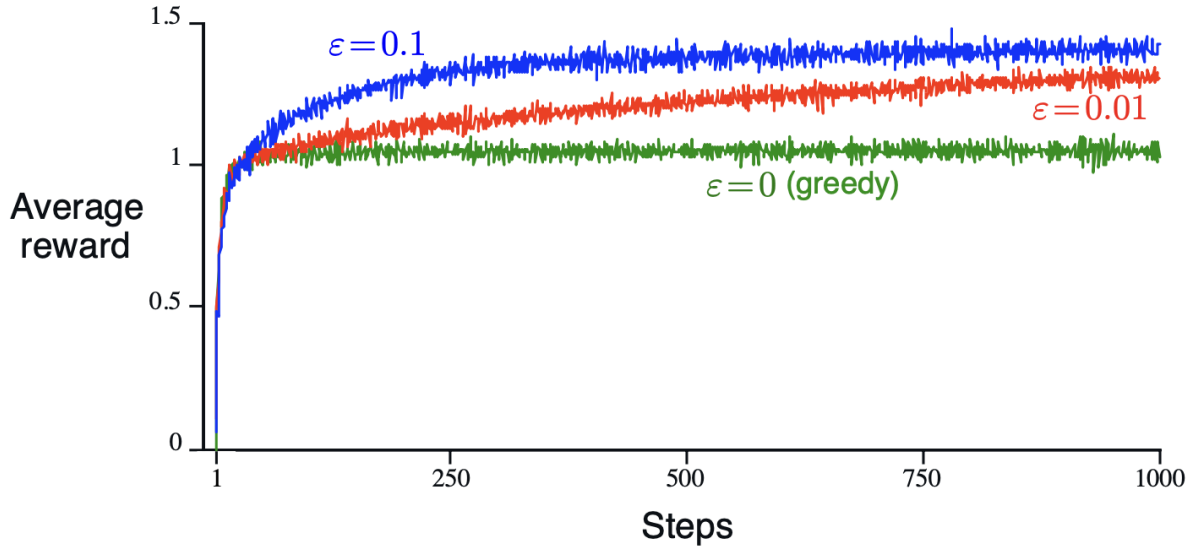
10-arm testbed

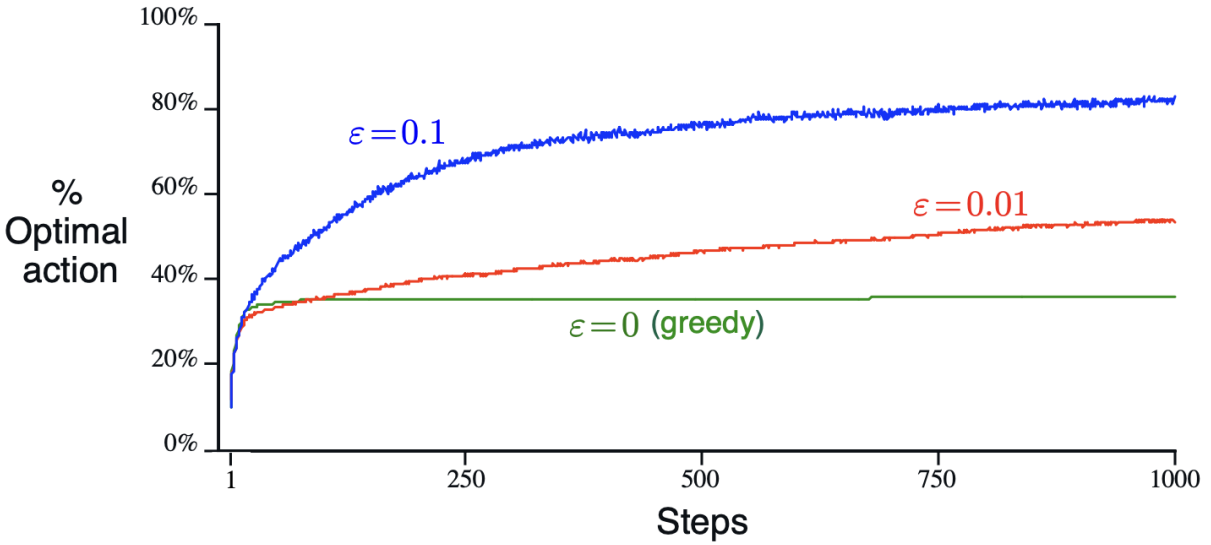
2000 randomly generated problems.

What is a randomly generated problem?



Let's discuss these figures!





Quiz: Bandit example

Exercise 2.2: Bandit example Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ε-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $R_1 = -1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = -2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these time steps the ε case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred? □

Quiz

- In supervised learning, we are given labeled data. In unsupervised learning we are given unlabeled data. What is the data in the multi-arm bandit problem?

- Can we calculate Q_{n+1} incrementally as a function of Q_n, R_n, and n?

- In the ε-greedy policy, what is exploration? What is exploitation?

Upper Confidence Bound (UCB) Policy

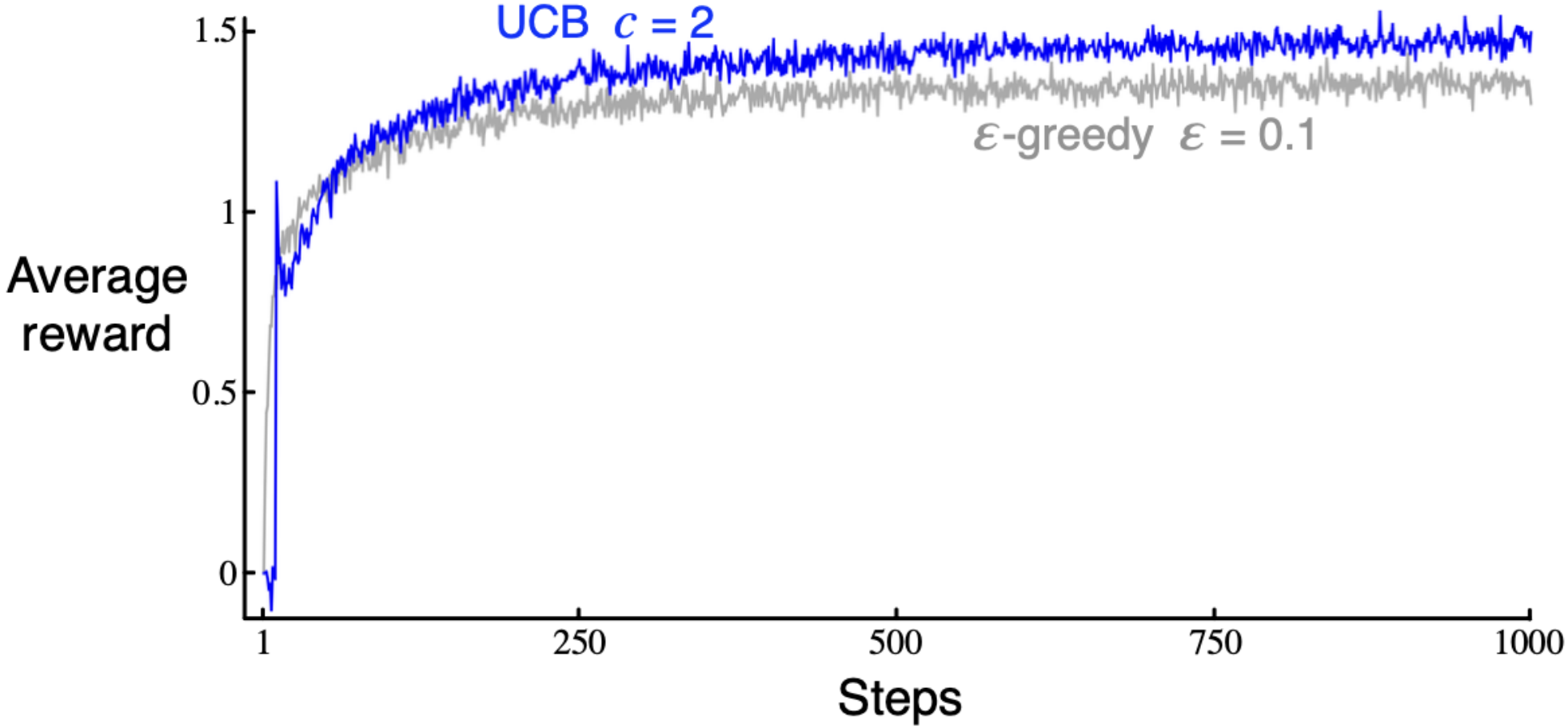
$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

- where $N_t(a)$ is the number of times action a selected prior to time t
- UCB algorithm is an alternative to the ε-greedy policy
- The term $c\sqrt{\frac{\ln t}{N_t(a)}}$ is a bonus term, and provides exploration
- **Question:** if action a' has been chosen less often than a'', which of the two actions will have a larger bonus?
- The bonus term is a measure of the uncertainty of our estimate: the larger the $N_t(a)$, the more certain we are about our estimate $Q_t(a)$.

Where does name “upper confidence bound” come from?

- P(q(a) $\leq$ $Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}}$) is close to 1.
- So $Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}}$ is essentially an upper bound on q(a)
- We would like to choose the action that maximizes q(a). But we don’t know the q(a)’s. The algorithm instead chooses the action that maximizes the upper bounds, which we do know:

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


UCB $c = 2$
ε-greedy $\varepsilon = 0.1$
Average reward
1.5
1
0.5
0
1
250
500
750
1000
Steps