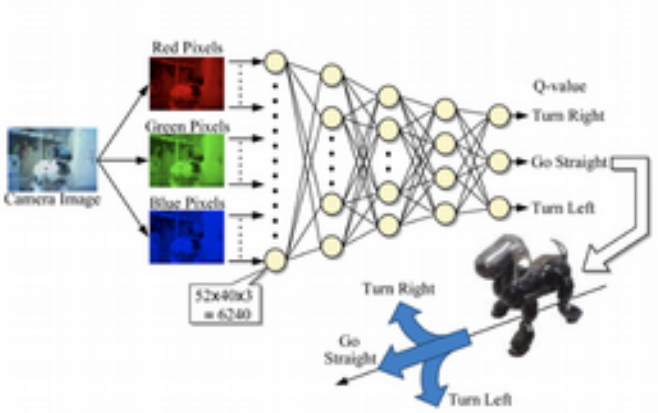
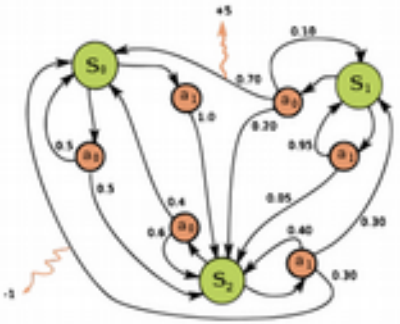
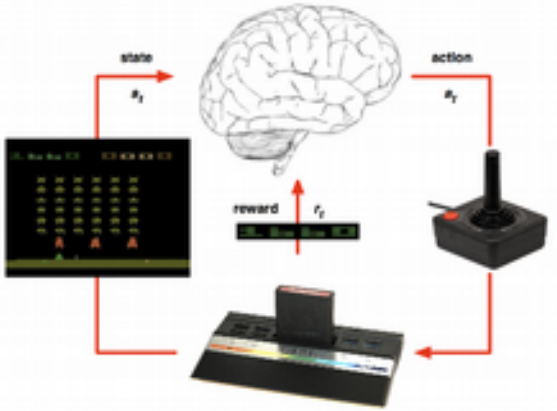
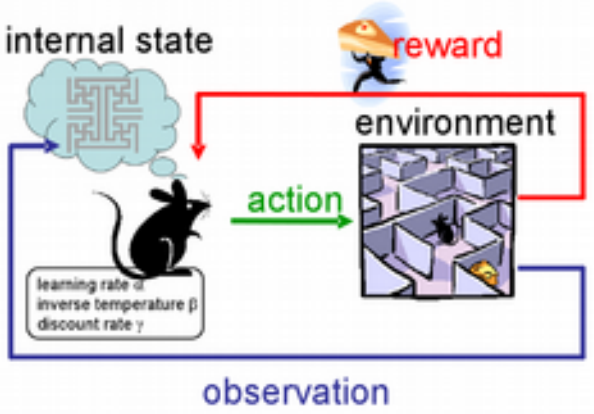
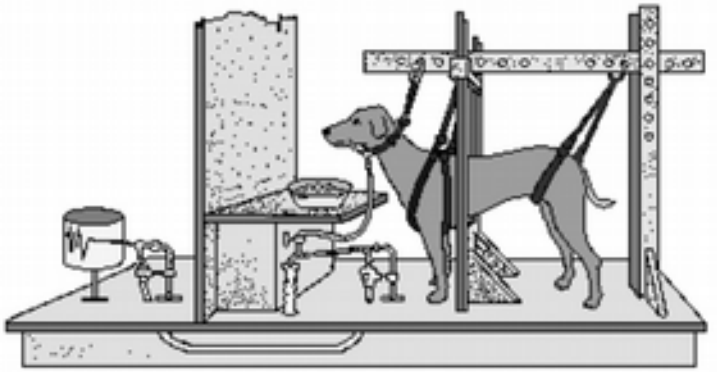
COMP 138: Reinforcement Learning











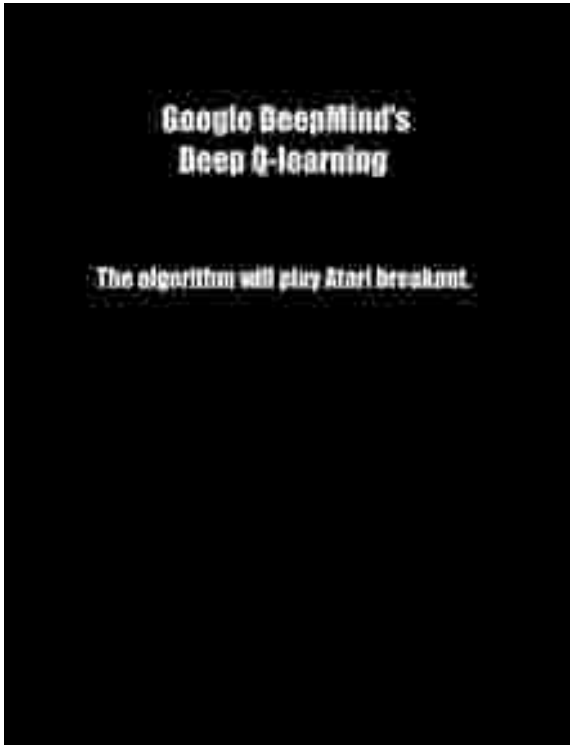
Instructor: Jivko Sinapov
Webpage: https://www.eecs.tufts.edu/~jsinapov/teaching/comp150_RL_Fall2020/

The Multi-Arm Bandit Problem

a.k.a. how to pick between Slot Machines (one-armed bandits) so that you walk out with the most $$$ from the Casino

But first...any questions?

Announcements


Google DeepMind's
Deep Q-learning
The algorithm will play Atari breakout

Discussion Moderation

- Sign up through link on Canvas
- Email me 3 days prior to your session with your discussion plan, notes, and link to any slides you want to use

Reading Assignment

- Chapter 3 of Sutton and Barto

Programming Assignment #1

- First programming assignment due 9/25

The Multi-Arm Bandit Problem

a.k.a. how to pick between Slot Machines (one-armed bandits) so that you walk out with the most $$$ from the Casino

. . . .

Arm 1 Arm 2 Arm k

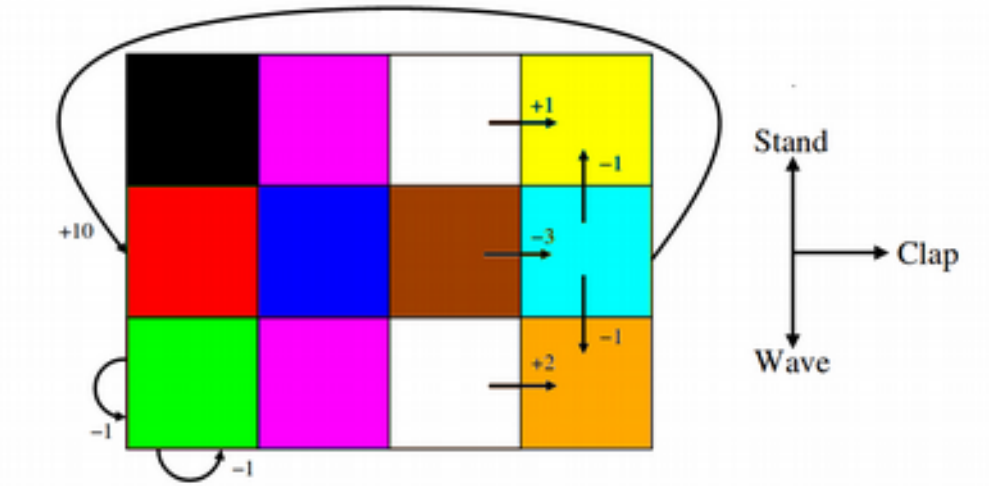
Which lever to pull next?

Which lever to pull next?

0 1 3 0 1

0 0 0 50 0

Discussion: how does MAB relate to RL?



Which lever to pull next?

2 1 3 3 1

0 0 0 50 0

Action-Value Functions

A function that encodes the “value” of performing a particular action (i.e., bandit)

Rewards observed when performing action *a*

$$Q_t(a) = \frac{R_1 + R_2 + \cdots + R_{K_a}}{K_a}.$$

Value function Q

# of times the agent has picked action *a*

Exploitation vs. Exploration

- Greedy: pick the action that maximizes the value function, i.e.,

$$Q_t(A_t^*) = \max_a Q_t(a)$$

- ε-Greedy: with probability ε pick a random action, otherwise, be greedy

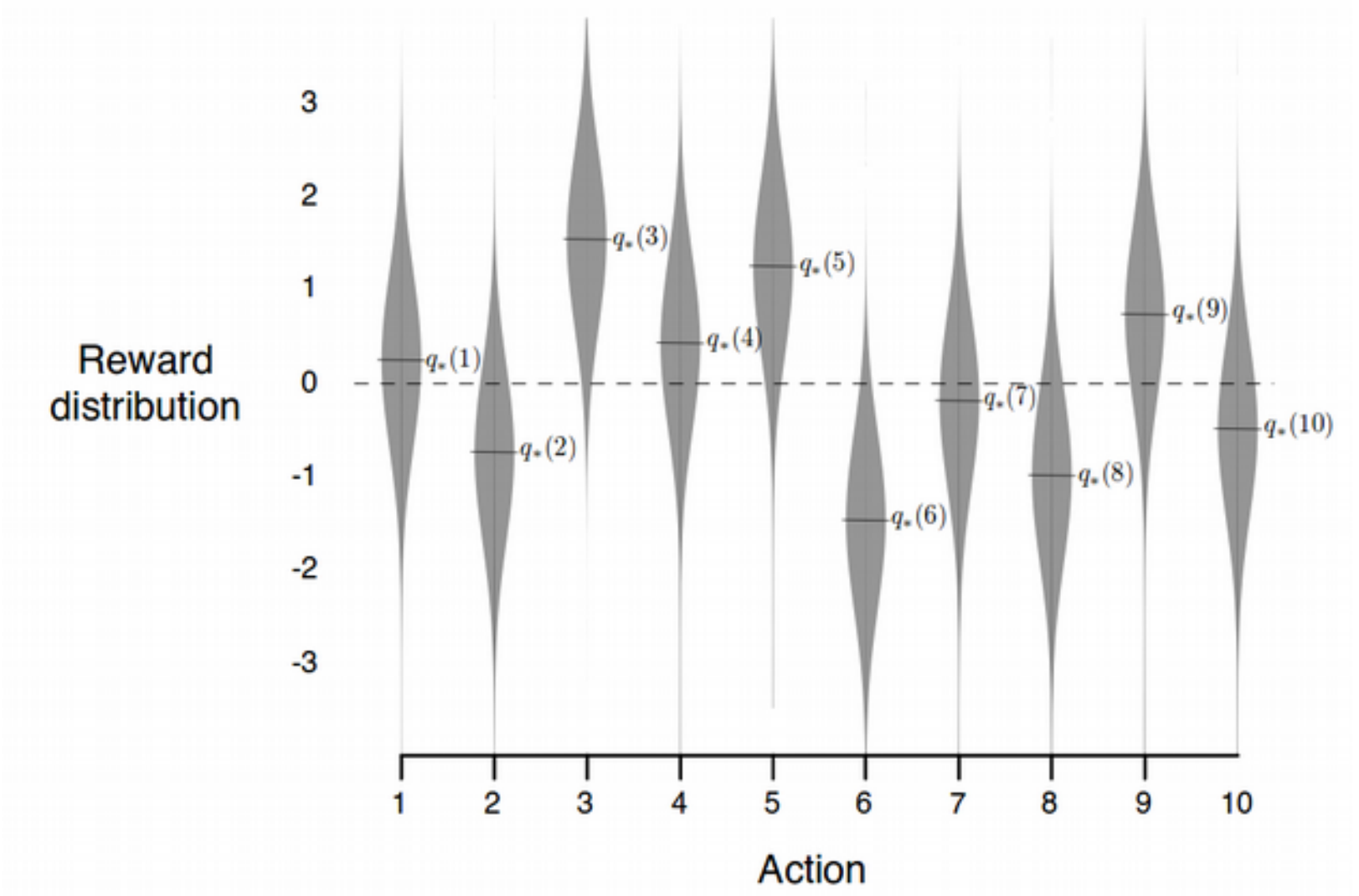
Exercise

Exercise 2.1 In ε-greedy action selection, for the case of two actions and $\varepsilon = 0.5$, what is the probability that the greedy action is selected?

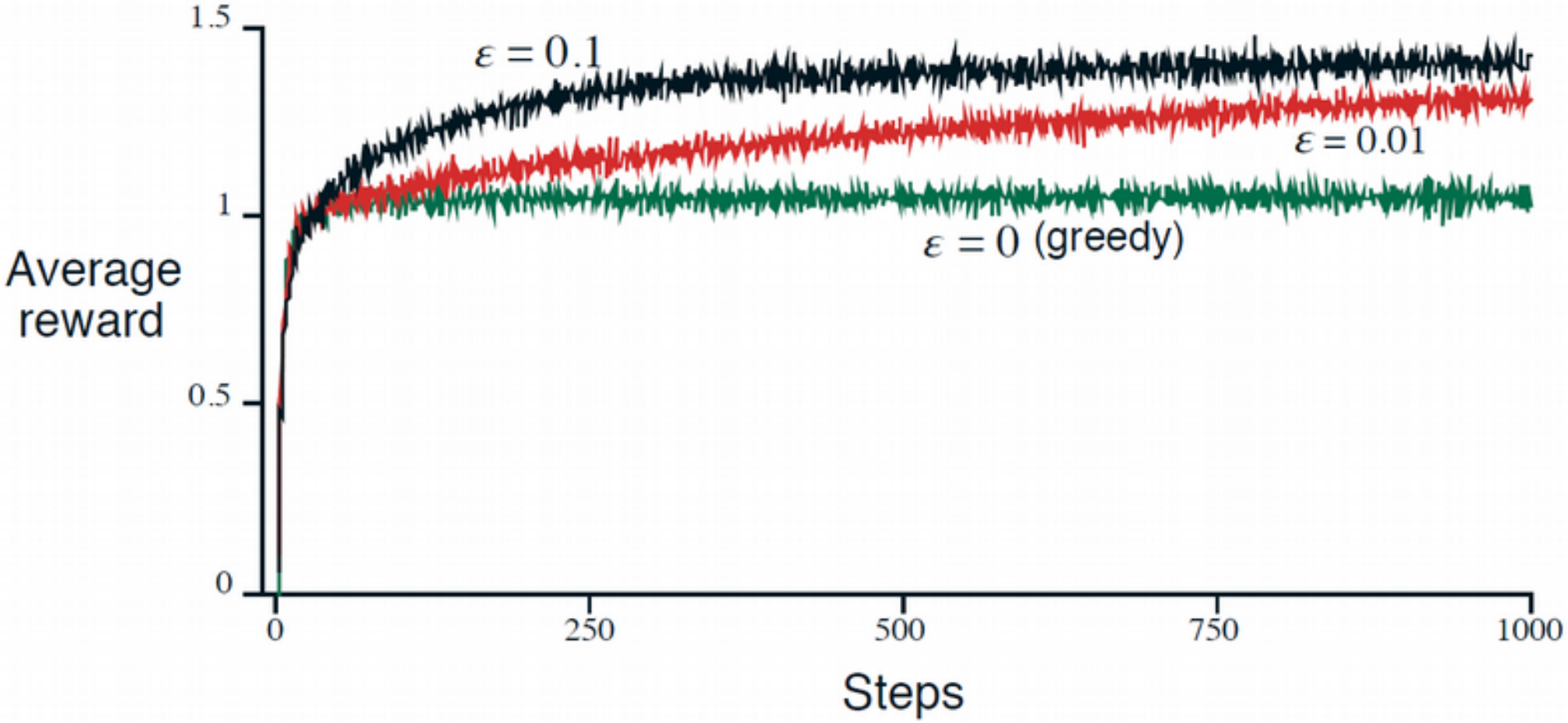
Another one...

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? □

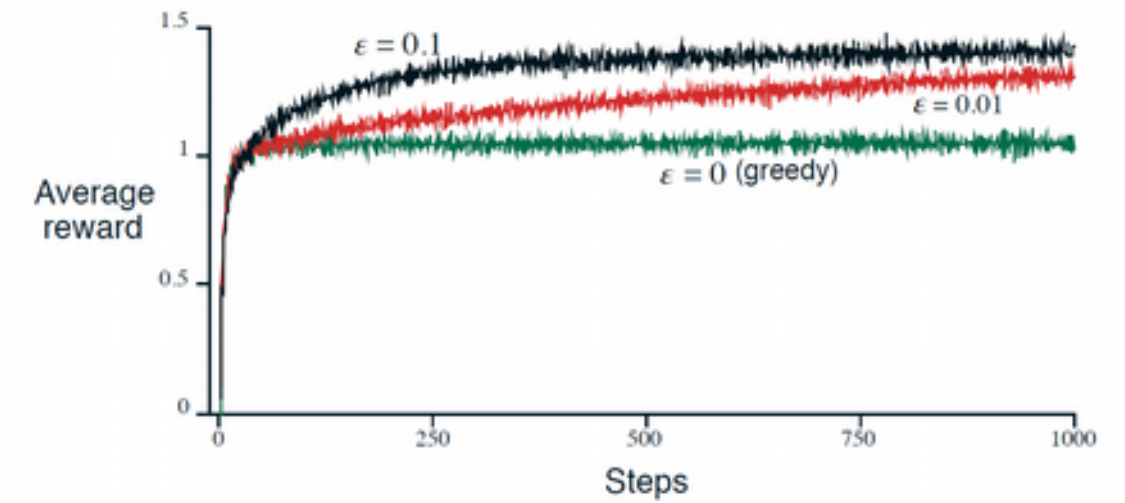
10-armed example



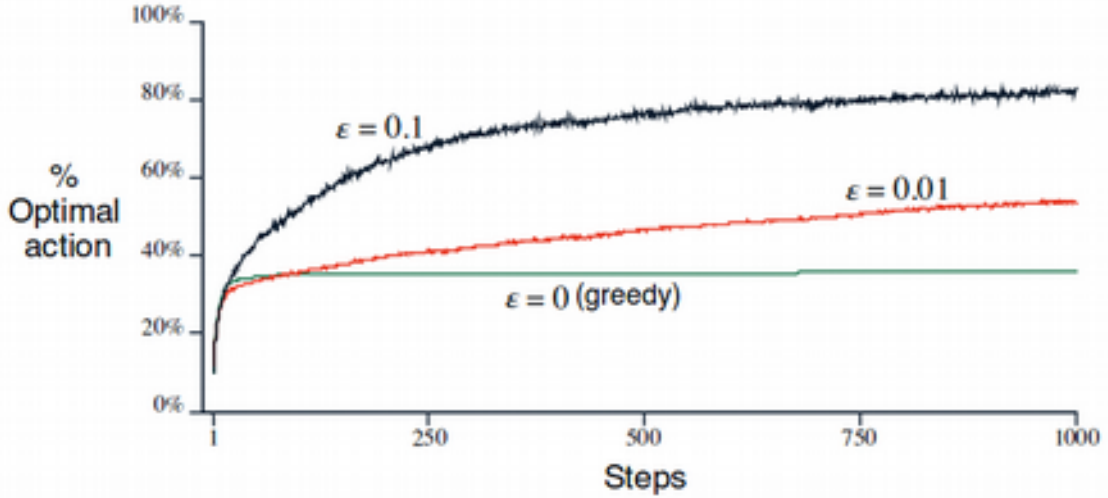
10-armed example

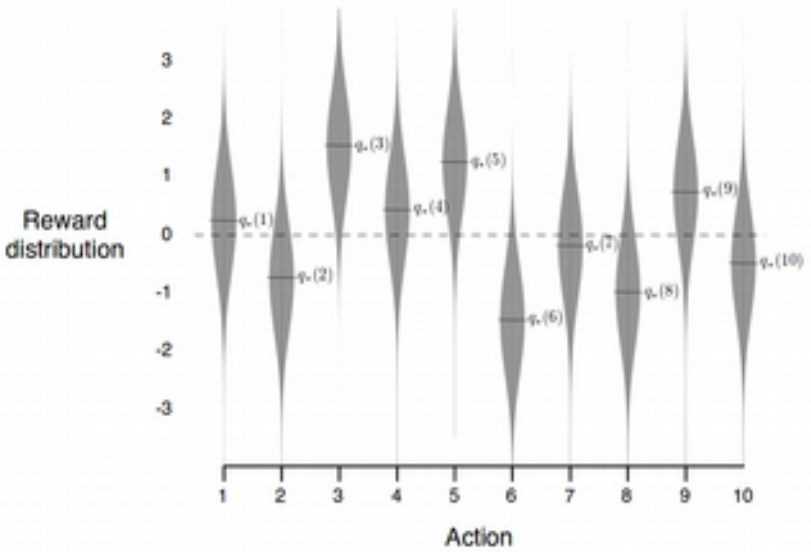


10-armed example exercise



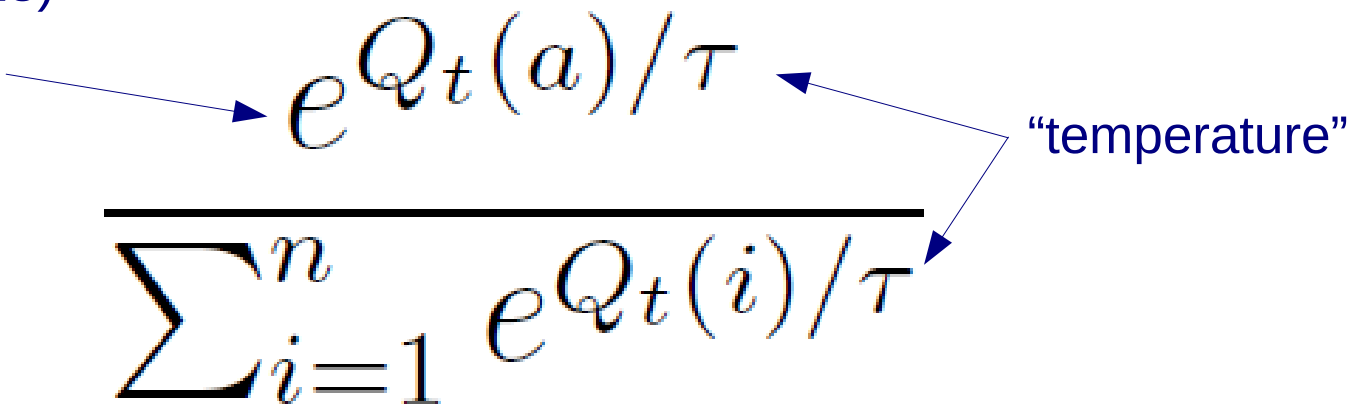
In the comparison shown in Figure 2.2, **which method will perform best in the long run in terms of cumulative reward** and **probability of selecting the best action?** How much better will it be?





Soft-Max Action Selection

Exponent of natural logarithm (~ 2.718)



As temperature goes up, all actions become nearly equally likely to be selected; as it goes down, those with higher value function outputs become more likely

Updating $Q_t(a)$ after observing R

Batch:

$$Q_t(a) = \frac{R_1 + R_2 + \cdots + R_{K_a}}{K_a}$$

Incremental:

$$\begin{aligned}
Q_{k+1} &= \frac{1}{k}\sum_{i=1}^{k} R_i \\
&= \frac{1}{k}\left(R_k + \sum_{i=1}^{k-1} R_i\right) \\
&= \frac{1}{k}\Big(R_k + (k-1)Q_k + Q_k - Q_k\Big) \\
&= \frac{1}{k}\Big(R_k + kQ_k - Q_k\Big) \\
&= Q_k + \frac{1}{k}\Big[R_k - Q_k\Big],
\end{aligned}$$

Updating $Q_t(a)$ after observing R

$$NewEstimate \leftarrow OldEstimate + StepSize \left[Target - OldEstimate \right]$$

A Simple Bandit Algorithm

How do we construct a value function at the start (before any actions have been taken)

How do we construct a value function at the start (before any actions have been taken)

Zeros:	0	0	0
Random:	-0.23	0.76	-0.9
Optimistic:	+5	+5	+5

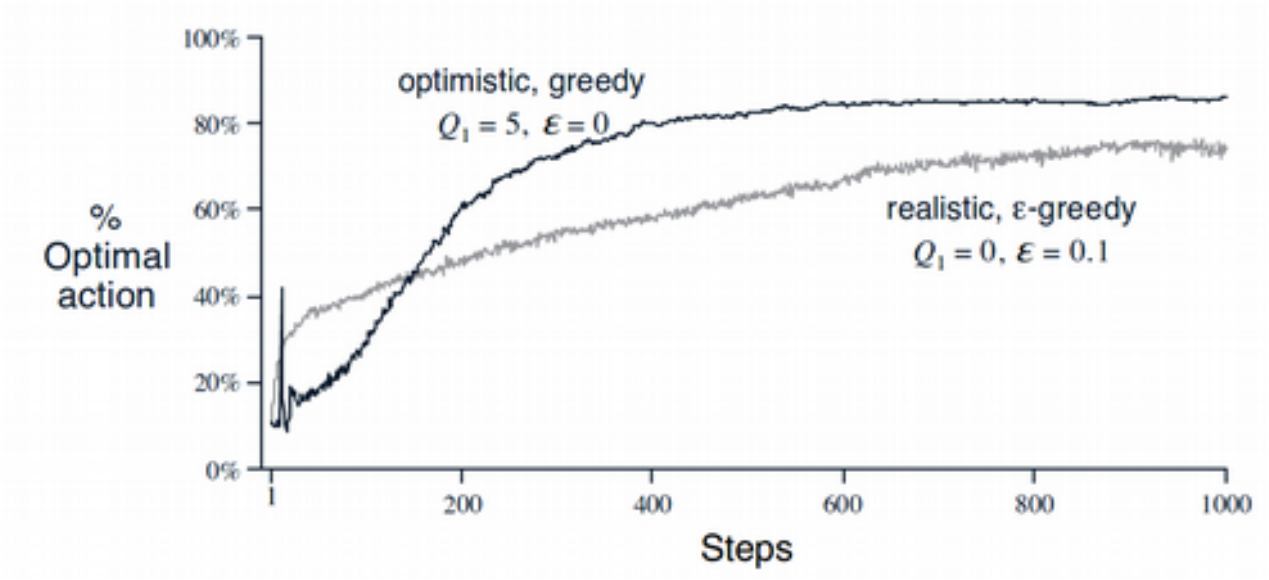
Arm 1

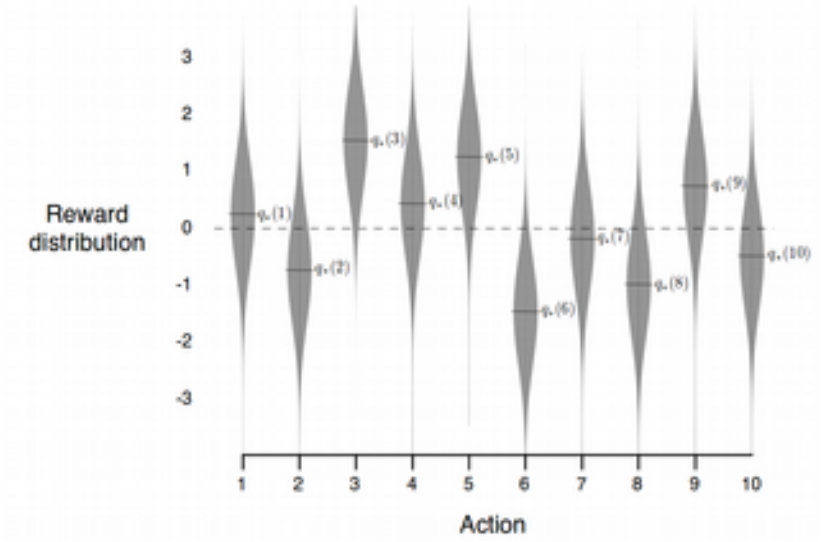
Arm 2

. . . .

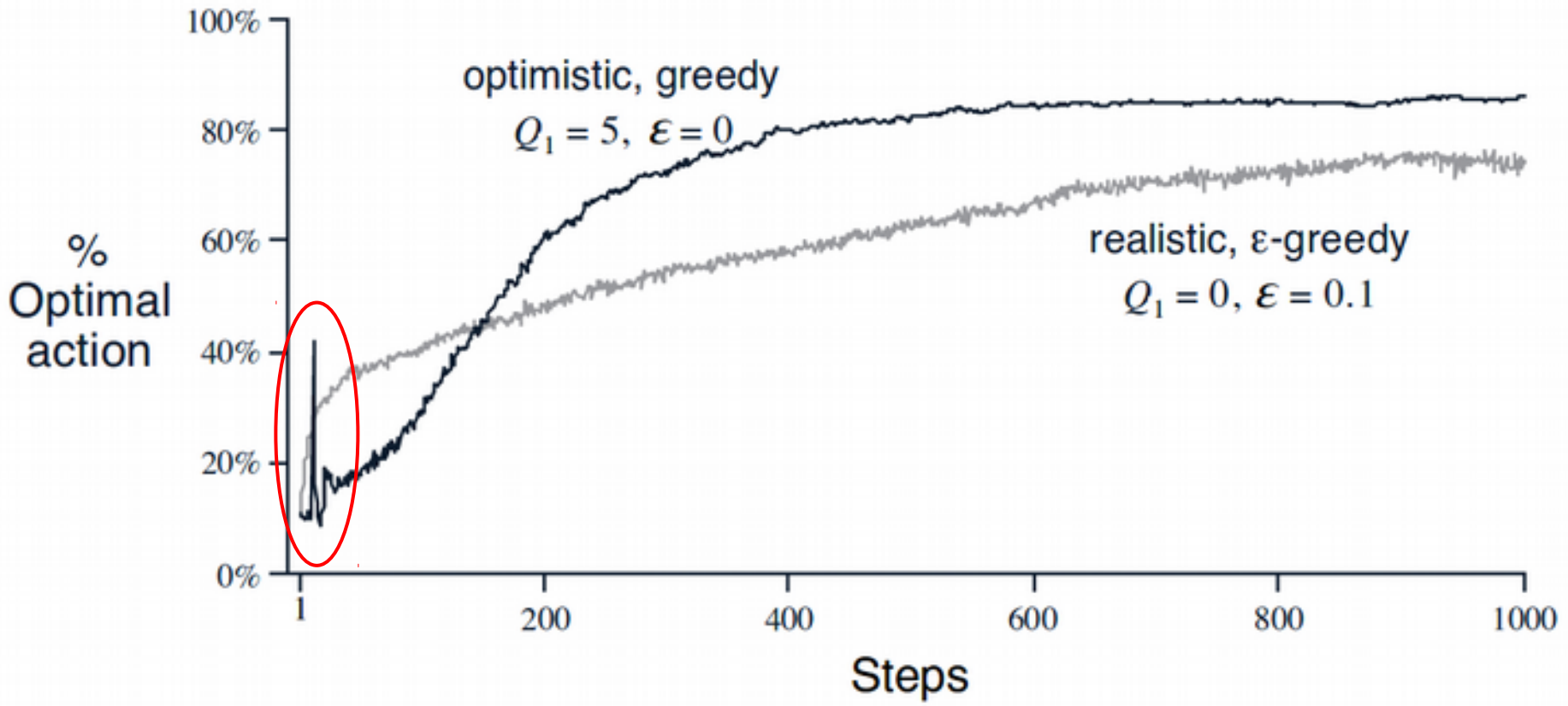
Arm k

How do we construct a value function at the start (before any actions have been taken)





Mysterious Spikes (Ex. 2.6)



Tracking a non-stationary problem

- What does it mean for a problem to be non-stationary?
- What are some ways to address non-stationary problems?

What happens when the payout of a bandit is changing over time?

$$Q_t(a) = \frac{R_1 + R_2 + \cdots + R_{K_a}}{K_a}$$

$$Q_k + \frac{1}{k}\Big[R_k - Q_k\Big]$$

What happens when the payout of a bandit is changing over time?

$$Q_{k+1} = Q_k + \alpha \Big[R_k - Q_k \Big]$$

instead of

$$Q_k + \frac{1}{k} \Big[R_k - Q_k \Big]$$

(section 2.5)

Exploration

- What’s wrong with epsilon-greedy exploration?

Exploration

- What’s wrong with epsilon-greedy exploration?
- What are some ways we can explore in a more intelligent manner?

Upper Confidence Bound Action Selection

Upper Confidence Bound Action Selection

UCB Spikes In Figure 2.4 the UCB algorithm shows a distinct spike in performance on the 11th step. Why is this? Note that for your answer to be fully satisfactory it must explain both why the reward increases on the 11th step and why it decreases on the subsequent steps. Hint: if c=1, then the spike is less prominent.

Reading Responses Discussion

“One question that I had throughout the reading was how a stationary problem versus a non-stationary problem would look in real life?”

- Amy

Reading Responses Discussion

"What would we expect to see with epsilon-greedy when rewards are non-stationary? If the rewards change over time, how should we determine the epsilon to use for exploration? Would the epsilon be related to the distribution of the reward function?"

- Eric

Reading Responses Discussion

“How can [we] know that the environment has changed except through rewards?”

- Xiaohui

Reading Responses Discussion

“Exploration vs. exploitation boils down to the precise values of the estimates, uncertainties, and number of remaining steps. If we desire a machine where we need to minimize error as much as possible, such as self-driving cars, that must reach a point where they should only exploit, at what point is it right to make that sort of decision? For machines that can affect everyday people, is it ethical to have a machine keep exploring and perhaps find better approaches to problems when a misstep could potentially cause harm?”

- Andrew

Reading Responses Discussion

“For the initial values, what exactly does “optimistic” mean? How does the optimistic method more encouraged to explore more when ε=0? How exactly is it able to surpass the realistic method? The optimistic method seems to level out, but the realistic method continues to increase. Is the superiority of one method over the other based on the limit of steps the machine can conduct? “

- Andrew

Reading Responses Discussion

“What's the difference between H(a) and Q(a)? They both use reward to update themselves.”

- Xu

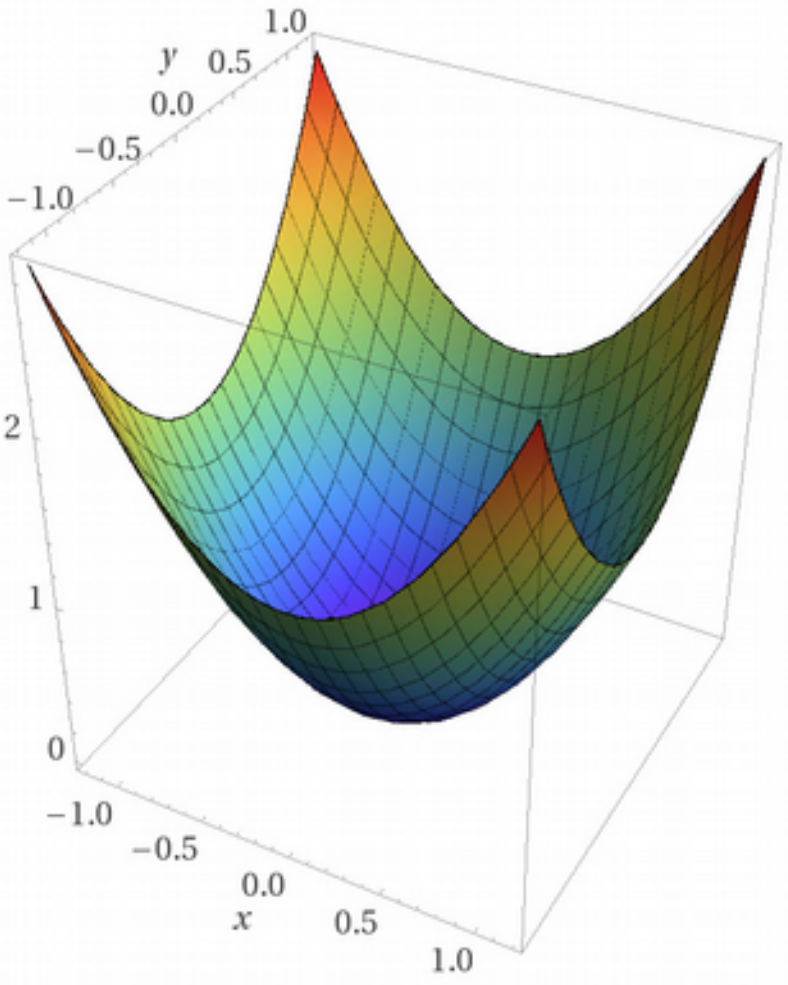
Reading Responses Discussion

"In which case should we choose evolutionary methods? I think if the space of policy is too big, it will be difficult to explore all policy."

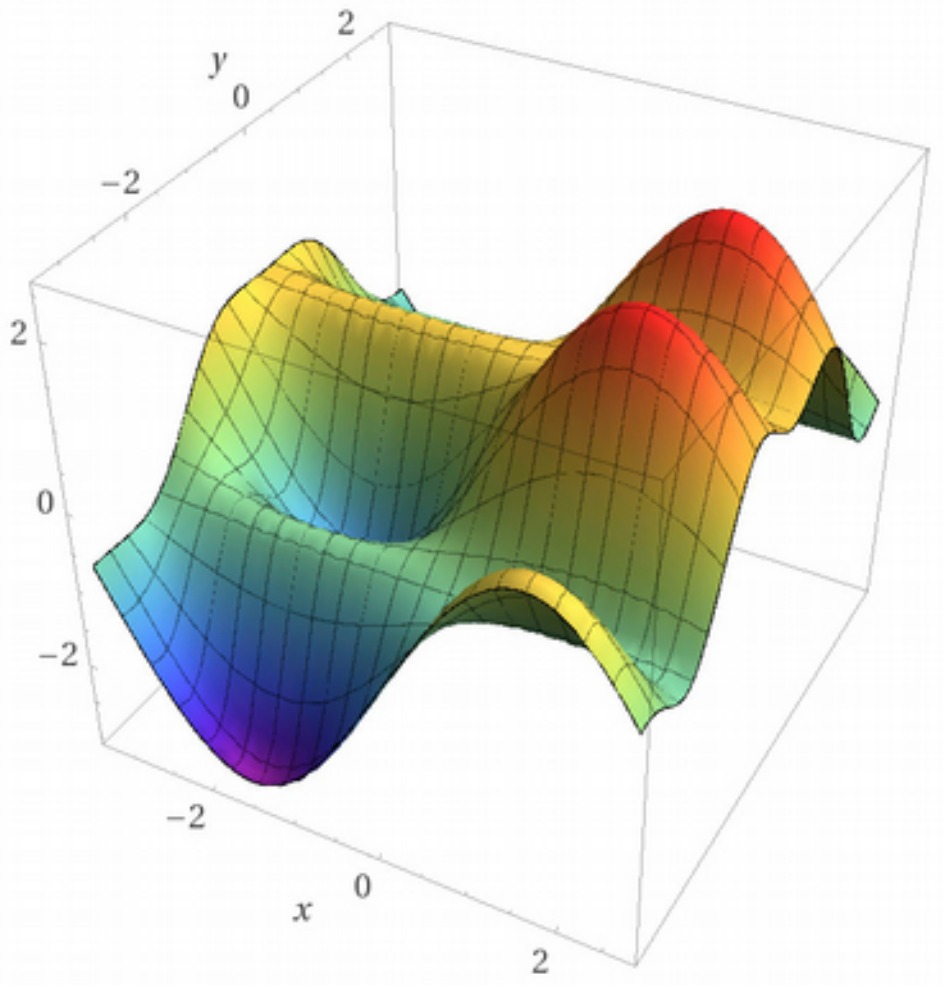
- Xiaohui

"Why do evolutionary methods have advantages over value-based methods on problems in which the learning agent cannot sense the complete state of the environment? If that is the case, how do we identify if a state is completely observable or partially observable?"

- Yash



Computed by WolframAlpha



Computed by WolframAlpha

So…which algorithm is best?

- e-greedy
- UCB
- Greedy with optimistic initialization
- Gradient badit
- ?