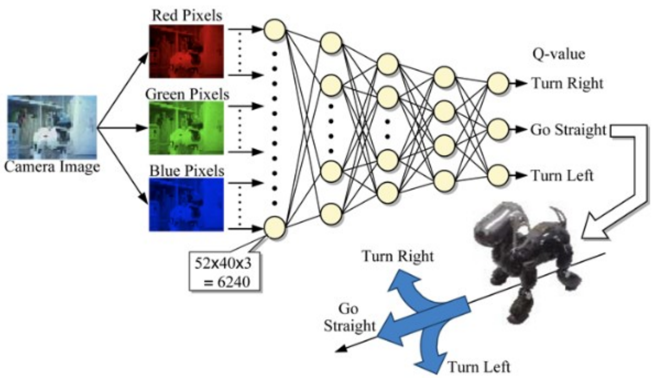
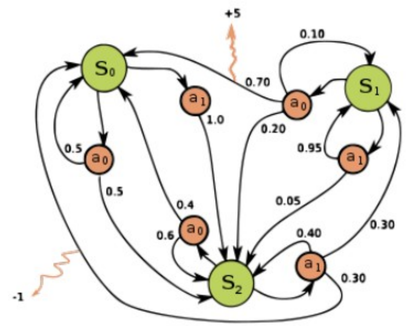
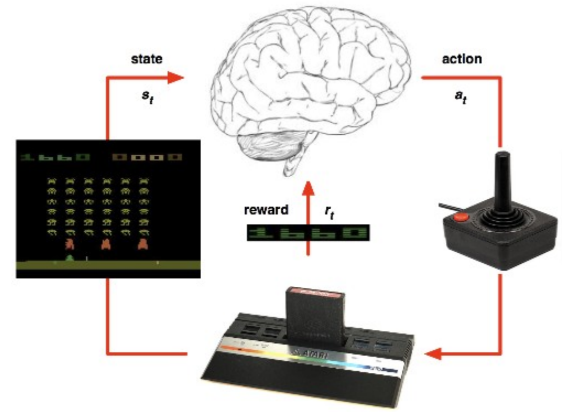
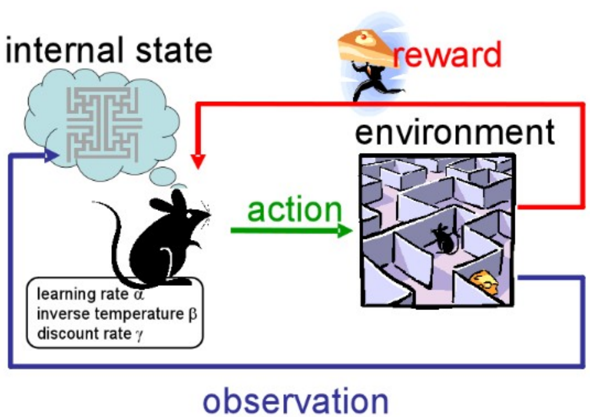
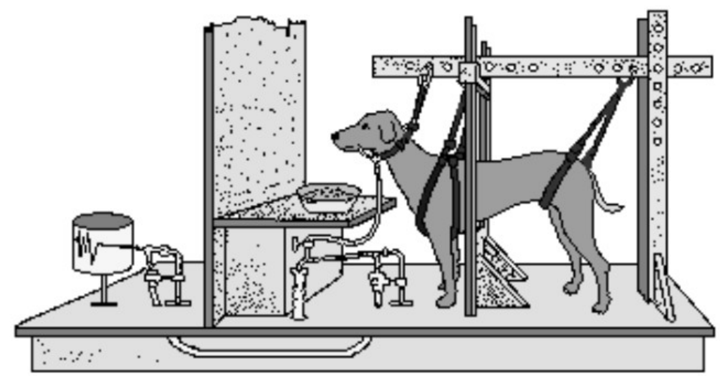
COMP 138: Reinforcement Learning











Instructor: Jivko Sinapov

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

Overview of Syllabus

But first...any questions?

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
- Only applies to PhD and MS students

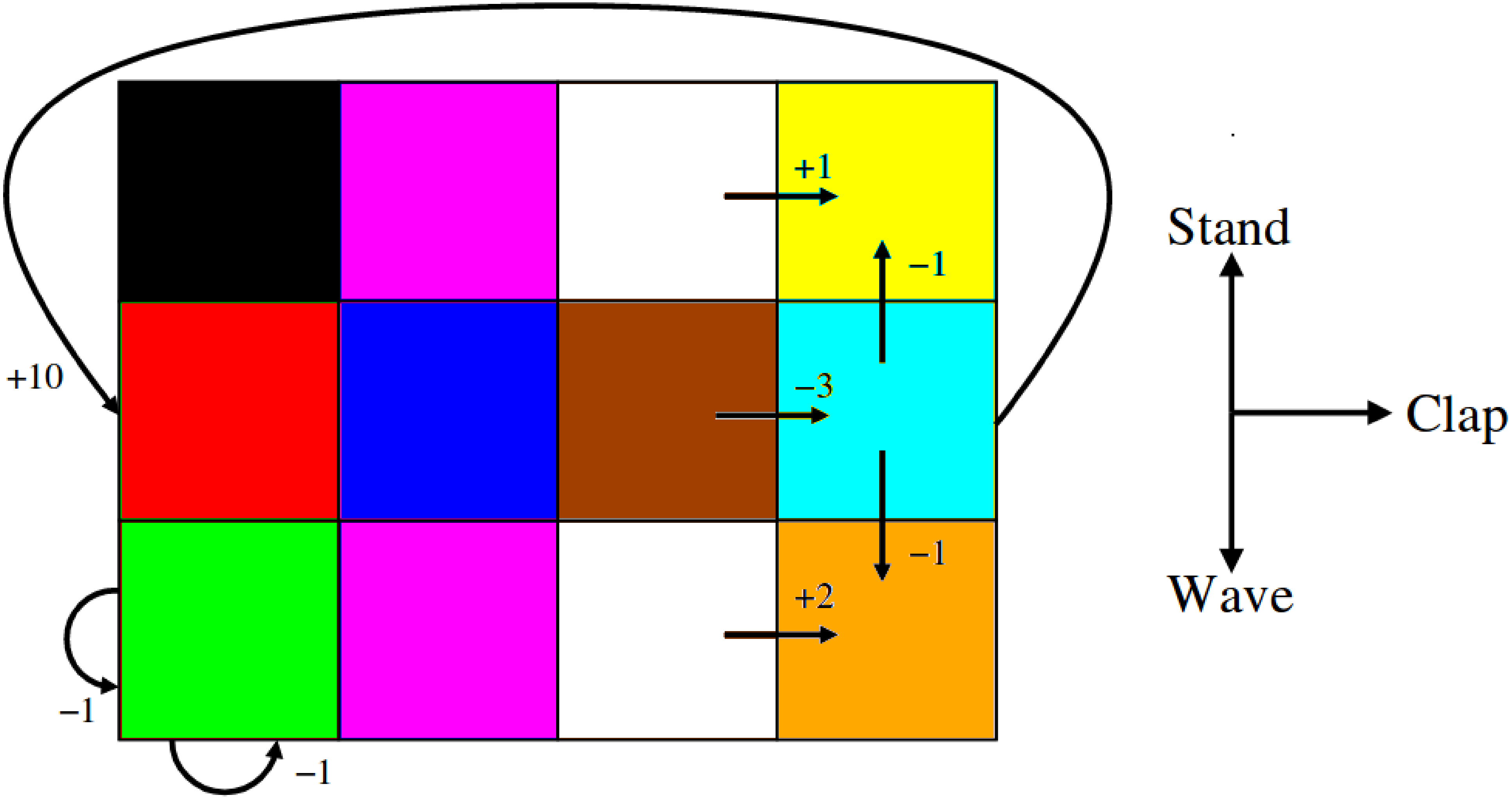
Reading Responses

- First reading response due before class Sep 13th
- Chapter 3 of Sutton and Barto for next week

Programming Assignment #1

- First programming assignment already out

Last time...



Where does RL fall within the field of Artificial Intelligence?

- AI → ML → RL
- Type of Machine Learning:
  - **Supervised**: learn from labeled examples
  - **Unsupervised**: learn from unlabeled examples
  - **Reinforcement**: learn through interaction

A note on what “supervision” means

- Is the reward function a form of supervision?

A note on what “supervision” means

- Is the reward function a form of supervision?
- **Argument for yes:** it tells the agent whether it did something good or bad
- **Argument for no:** it doesn’t tell the agent whether the action taken was the one that maximizes reward; it doesn’t tell the agent whether the preceding actions during which no reward was observed were good or bad.

A note on what “supervision” means

- Is the reward function a form of supervision?
- ~~**Argument for yes:** it tells the agent whether it did something good or bad~~
- **Argument for no:** it doesn’t tell the agent whether the action taken was the one that maximizes reward; it doesn’t tell the agent whether the preceding actions during which no reward was observed were good or bad.

A note on what “supervision” means

- Supervision IFF the agent is instructed with the correct choice of action
- Supervised ML classifiers have the correct labels for training data points
- RL agents are typically not given data with the correct sequence of actions*

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

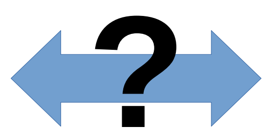
Which lever to pull next?

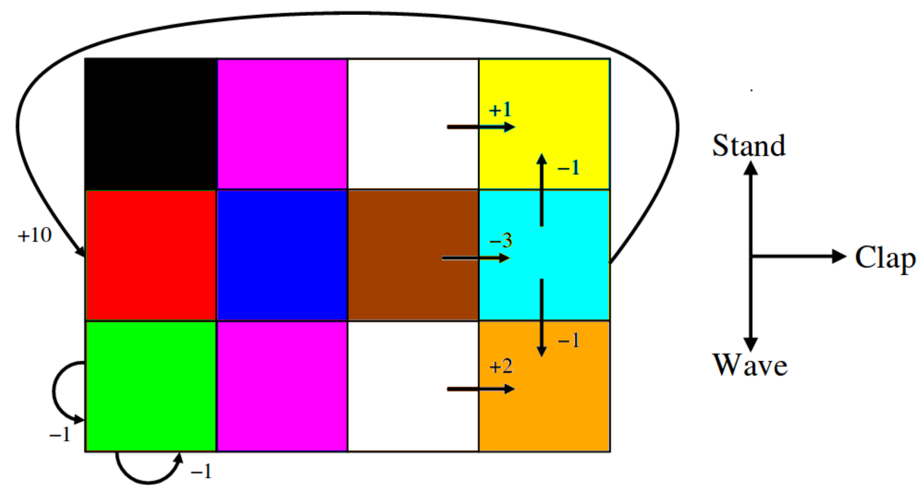
Which lever to pull next?

3 1 3 2 1

0 0 0 50 0

Discussion: how does MAB relate to RL?





Which lever to pull next?

3 1 3 2 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

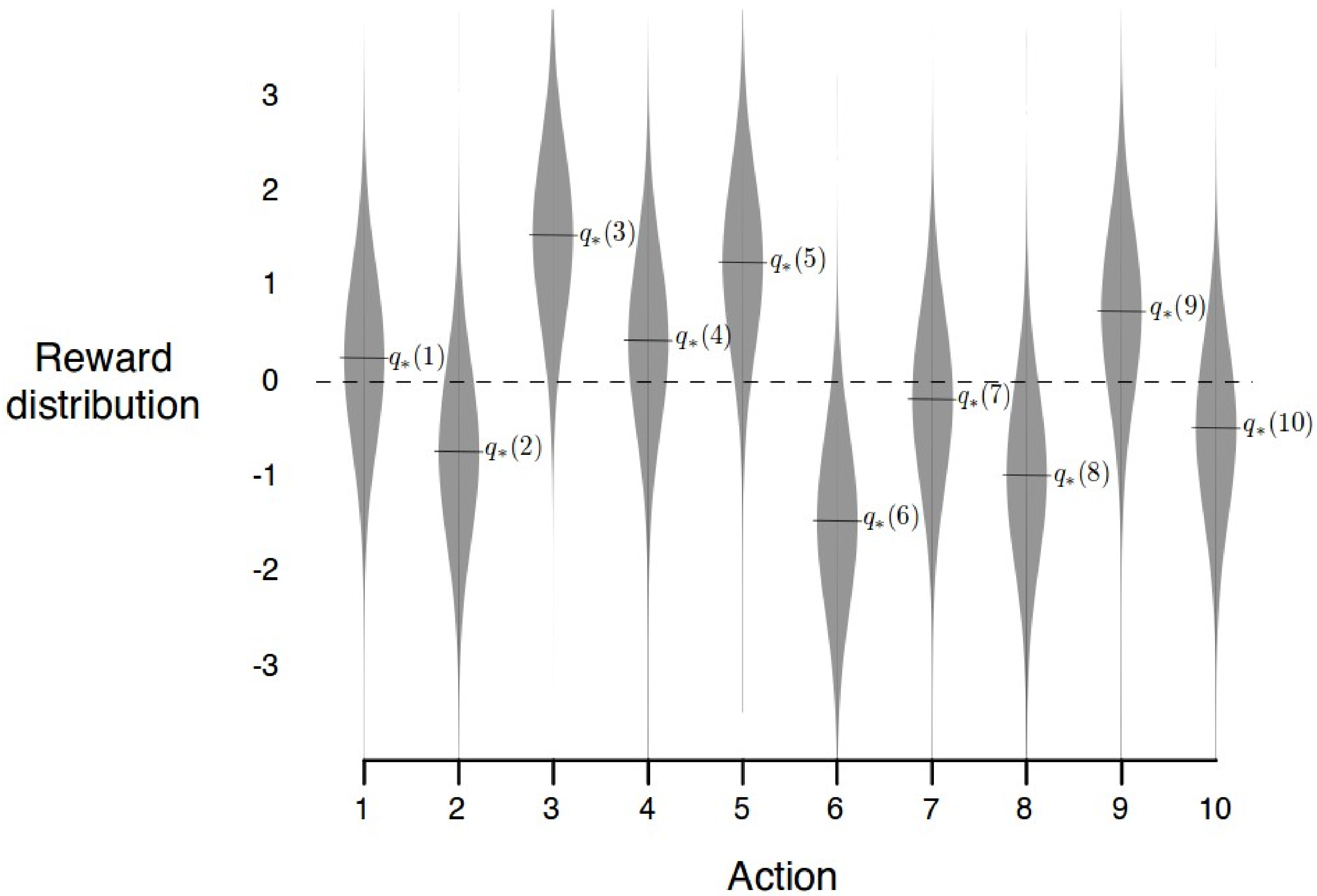
Exercise

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?

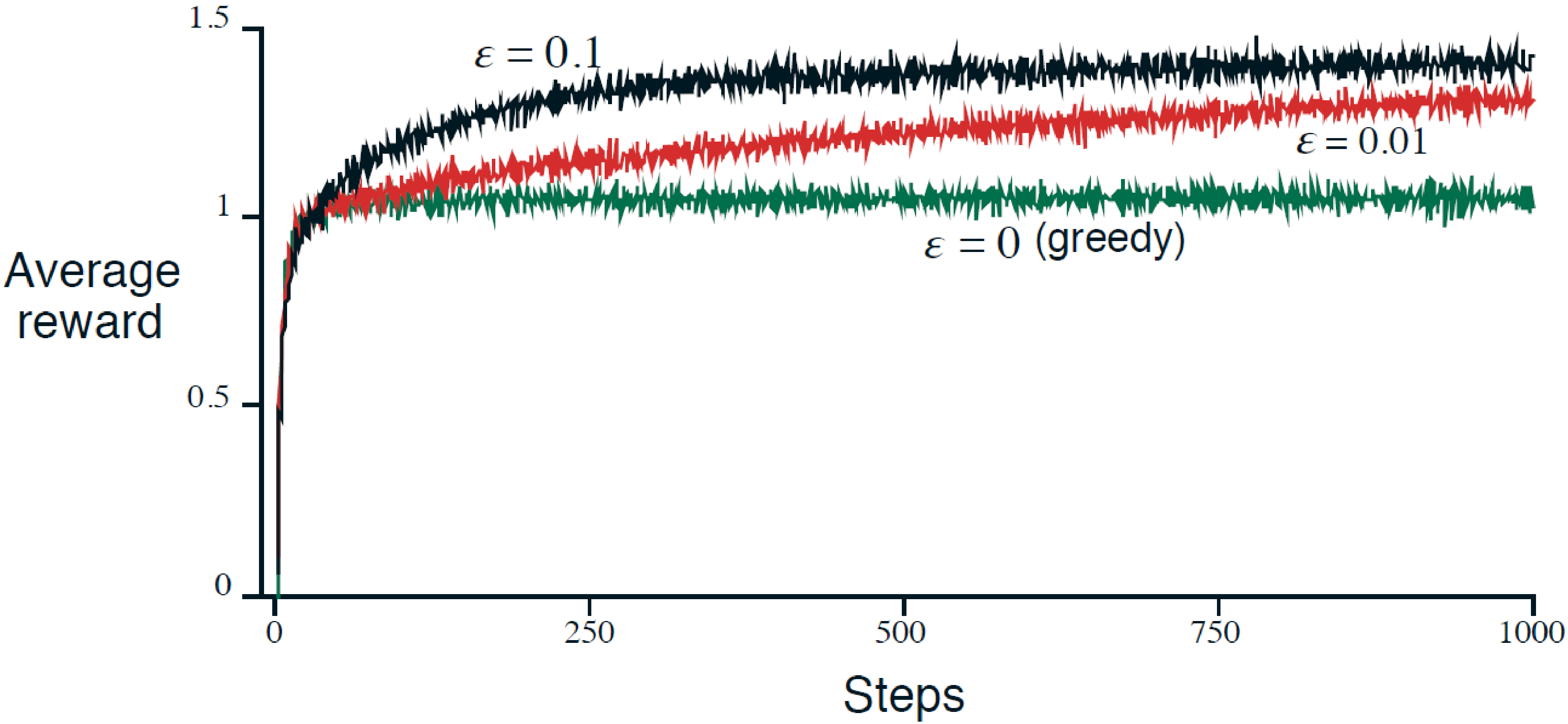
In-Class Small Group Exercise

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? $\square$

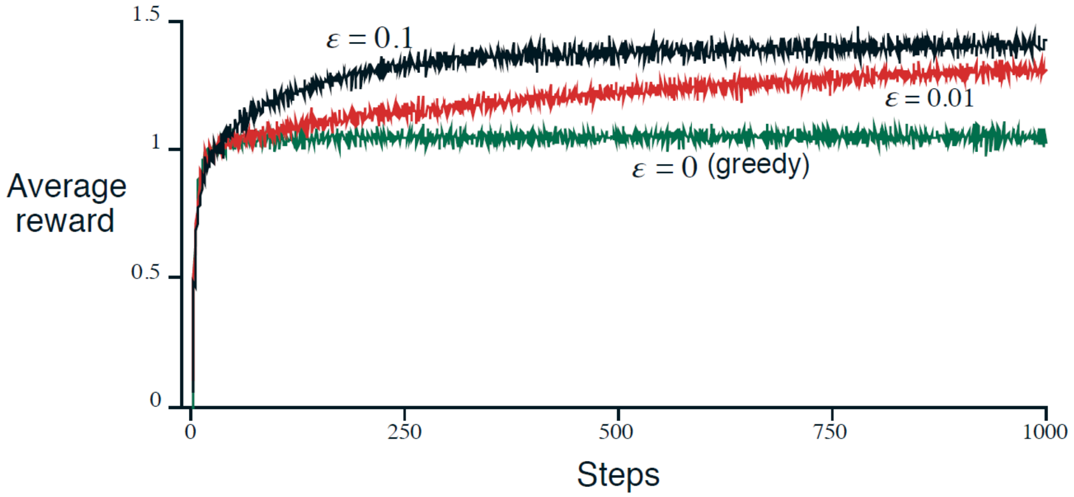
10-armed example



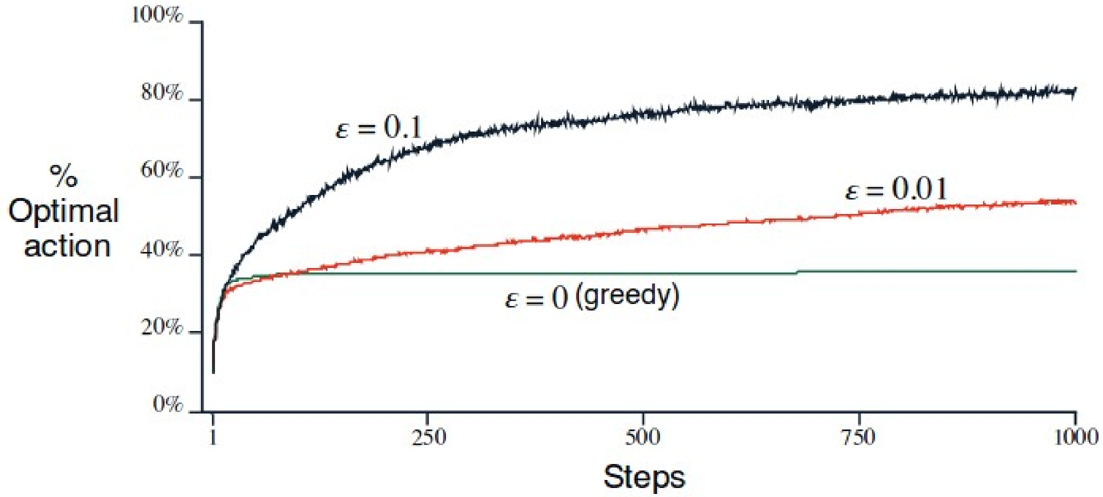
10-armed example

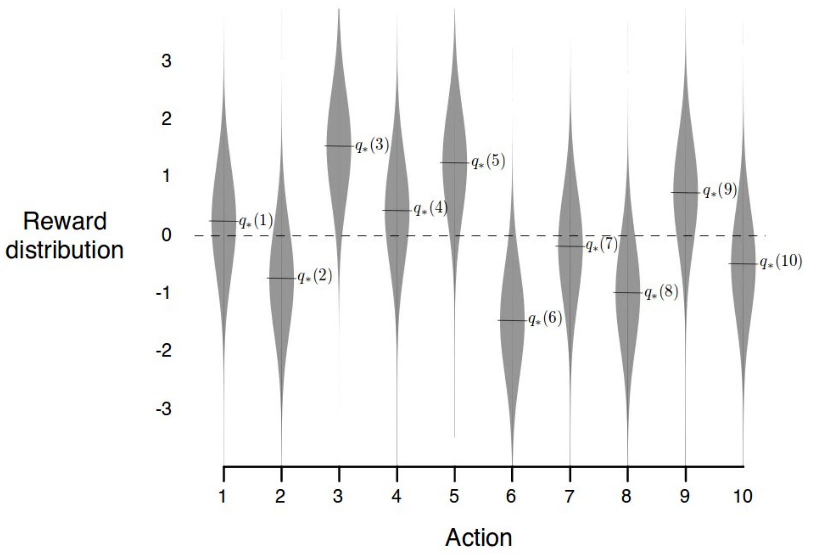


10-armed example exercise



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?**





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

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Repeat forever:

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

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

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

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

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

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

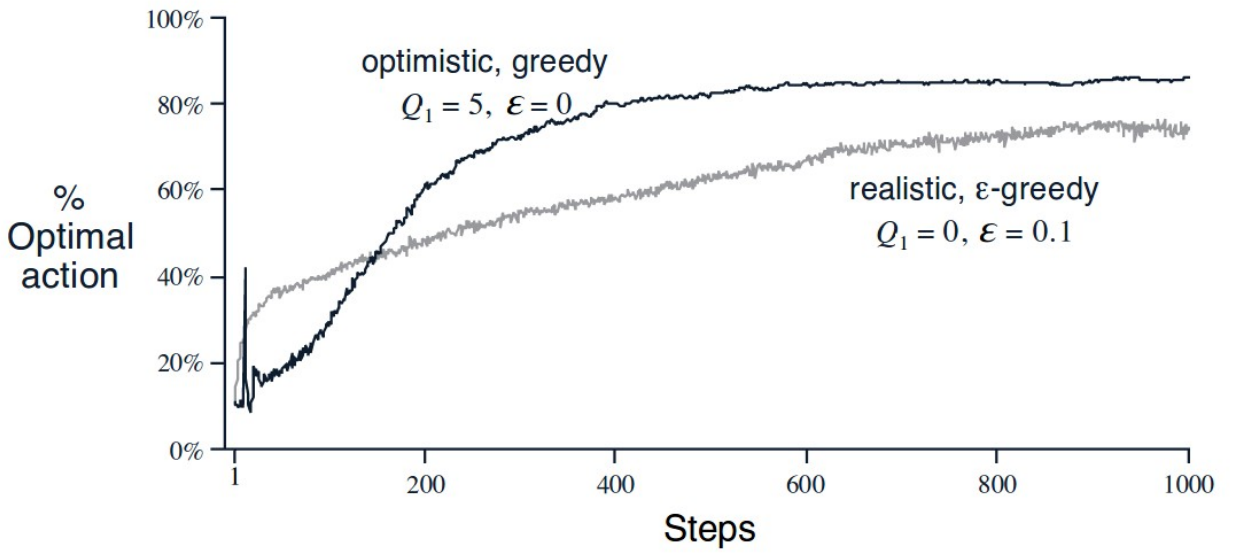
Arm 1

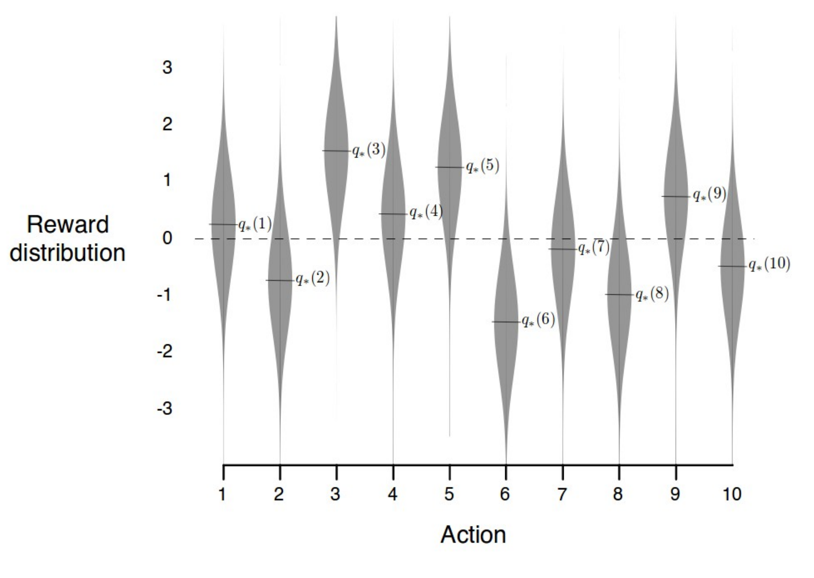
Arm 2

. . . .

Arm k

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





Soft-Max Action Selection

Exponent of natural logarithm (~ 2.718)

$$\frac{e^{Q_t(a)/\tau}}{\sum_{i=1}^{n} e^{Q_t(i)/\tau}}$$

“temperature”

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

Breakout (10 min, if time permits)

- Form a group of 3-4 with your neighbors
- Introduce yourself (why are you taking this class?)
- Brainstorm – what are some practical applications of MABs? What are some limitations of the formulation so far that make it difficult to apply in practice?

The Multi-Arm Bandit Problem

The casino always wins...so why is this problem important?

Arm 1

Arm 2

. . . .

Arm k

MAB TF library introduction

https://www.youtube.com/watch?v=7QFSziiAnxI

MAB Research

- Identifying outlier arms in MAB problems (NeurIPS 2017): https://www.youtube.com/watch?v=ecgNELgTzS8
- Finding Structure in MABs: https://www.youtube.com/watch?v=Thmh_--kVmg
- Budgeted Combinatorial MABs (AAMAS 2022): https://www.youtube.com/watch?v=gEigPlsmJ0M
- Contextual Bandits in Healthcare: https://www.youtube.com/watch?v=gbZHmPJaU0I

Next time...

- Upper Confidence Bound and Gradient-based algorithms for action selection