ELTE | FACULTY OF INFORMATICS

MULTI-ARMED BANDIT

Deep Reinforcement Learning

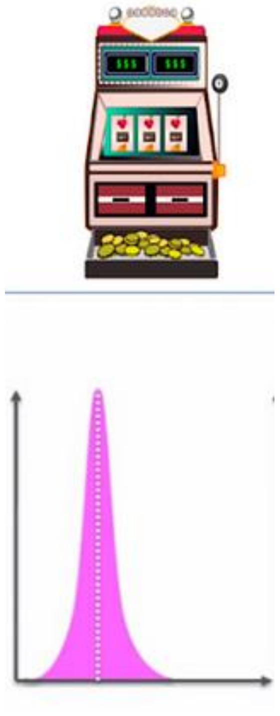
Balázs Nagy, PhD

ELTE | IK

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

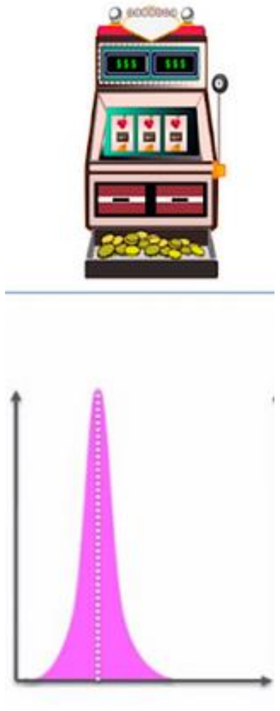
One-Armed Bandit

- **Scenario:**
  If you pull the arm, you get a reward from a stationary probability distribution
- Is it a RL task?



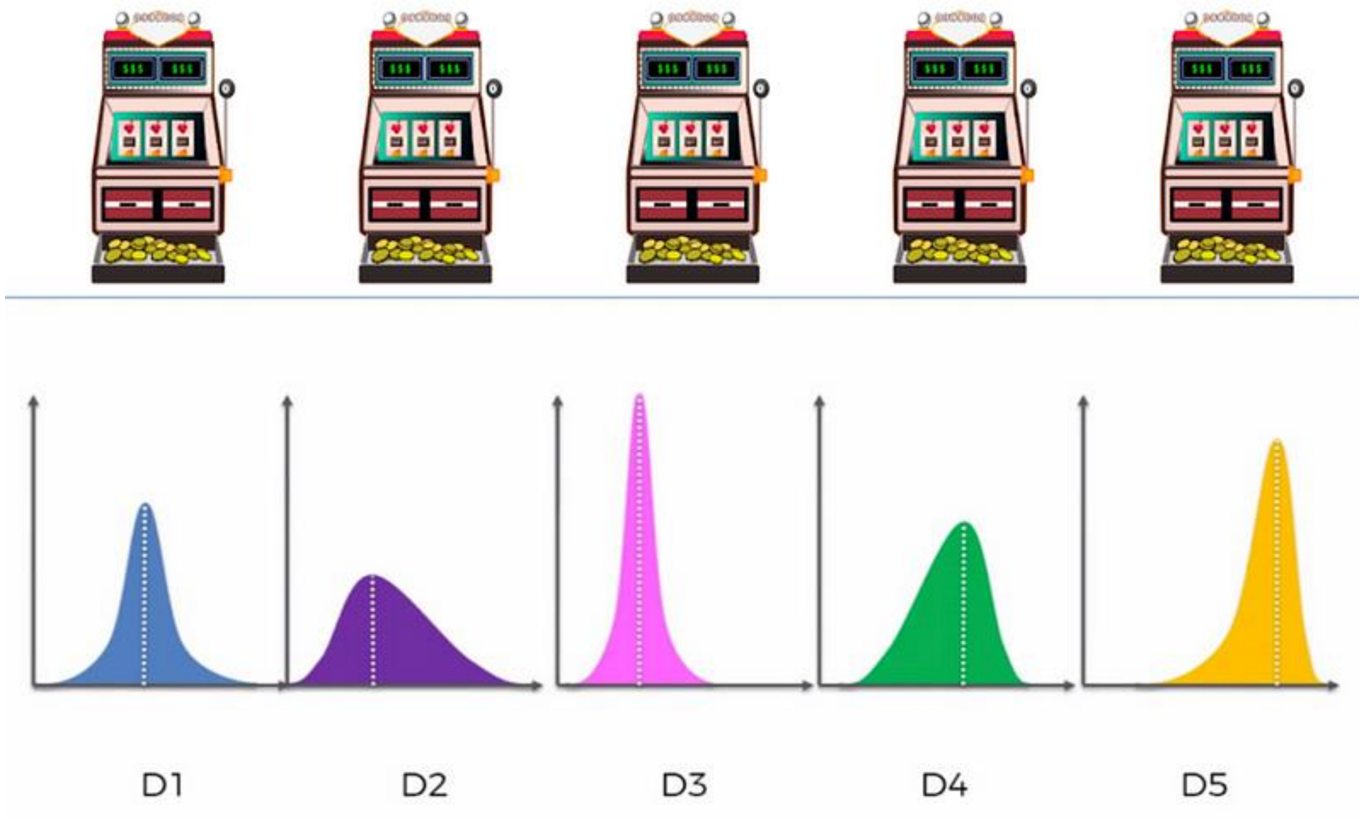
One-Armed Bandit

- **Scenario:**
  If you pull the arm, you get a reward from a stationary probability distribution
- Is it a RL task?
  - **NO**
  - There is only one state and only one action



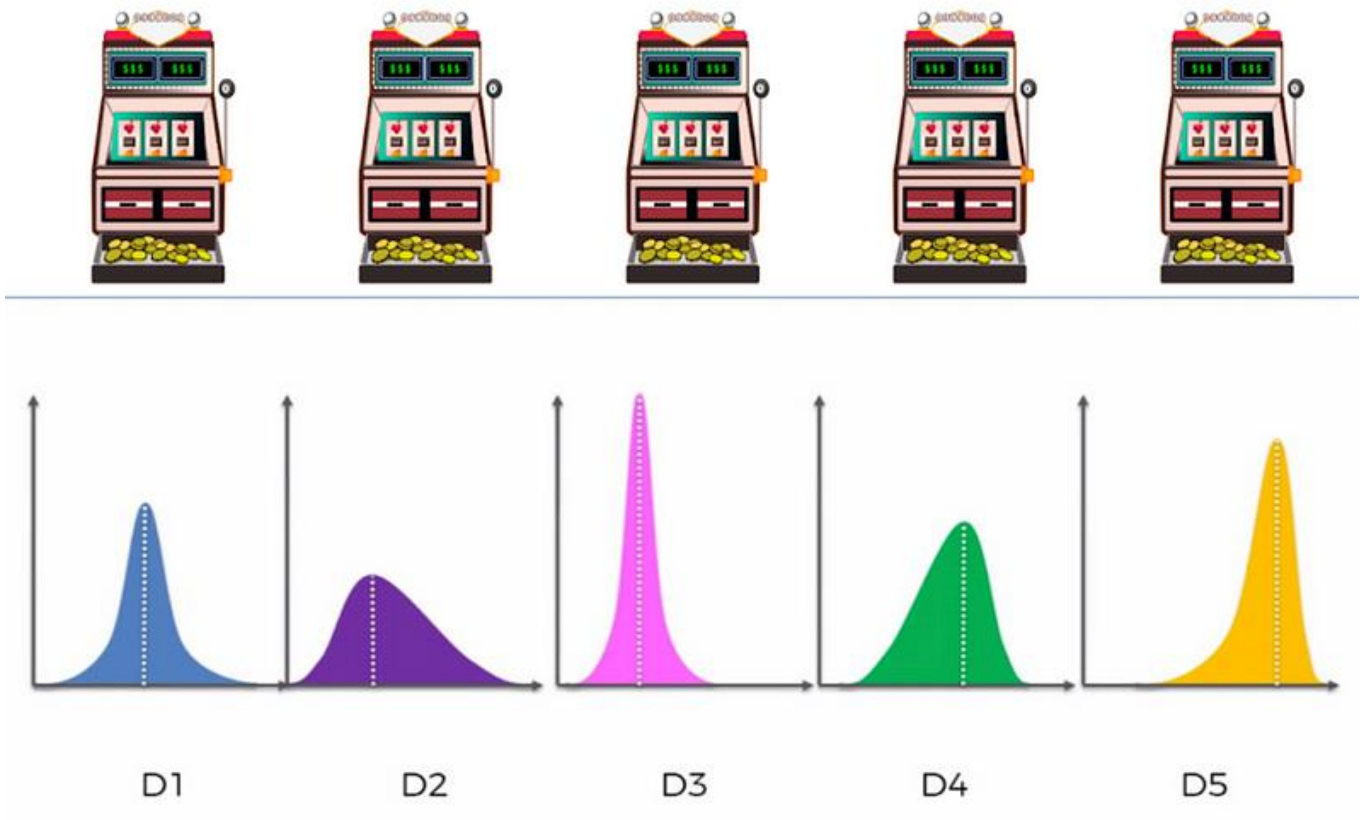
K-Armed Bandit

- **Problem**:
  Given a finite number of pulls (T), how can I optimize my winnings?



K-Armed Bandit

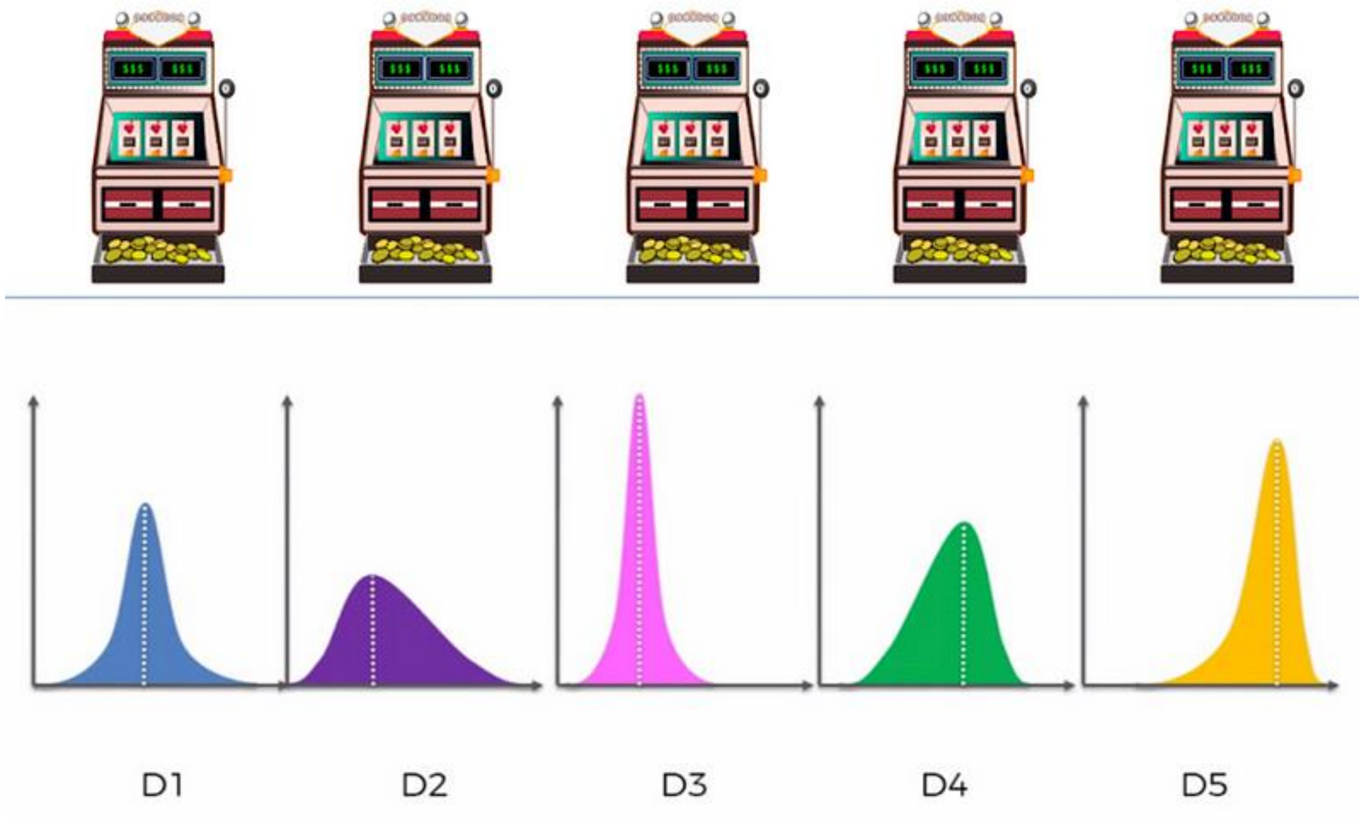
- **Problem**:
  Given a finite number of pulls (T), how can I optimize my winnings?
- **Dilemma**:
  How much should I explore? How much should I exploit?



K-Armed Bandit

- **Problem**:
  Given a finite number of pulls (T), how can I optimize my winnings?
- **Dilemma**:
  How much should I explore? How much should I exploit?

Nonassociative, evaluative feedback problem



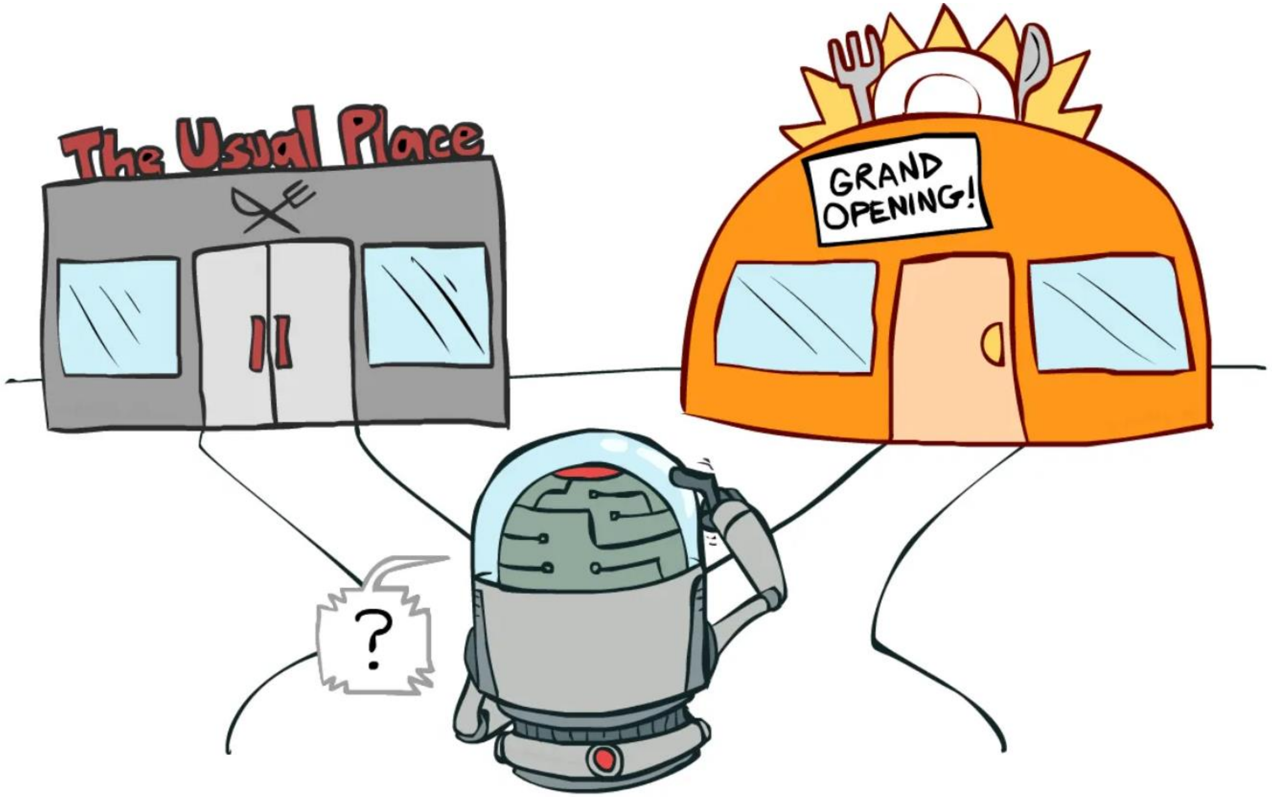
Definitions

- **Evaluative feedback** – dependent of action taken
  Uses training information that evaluates the action taken
- **Instructive feedback** – independent of action taken
  Instructs by giving correct action

- **Nonassociative:**
  Does not involve learning to act in more than one situation
- **Associative:**
  Action are taken in more than one situation

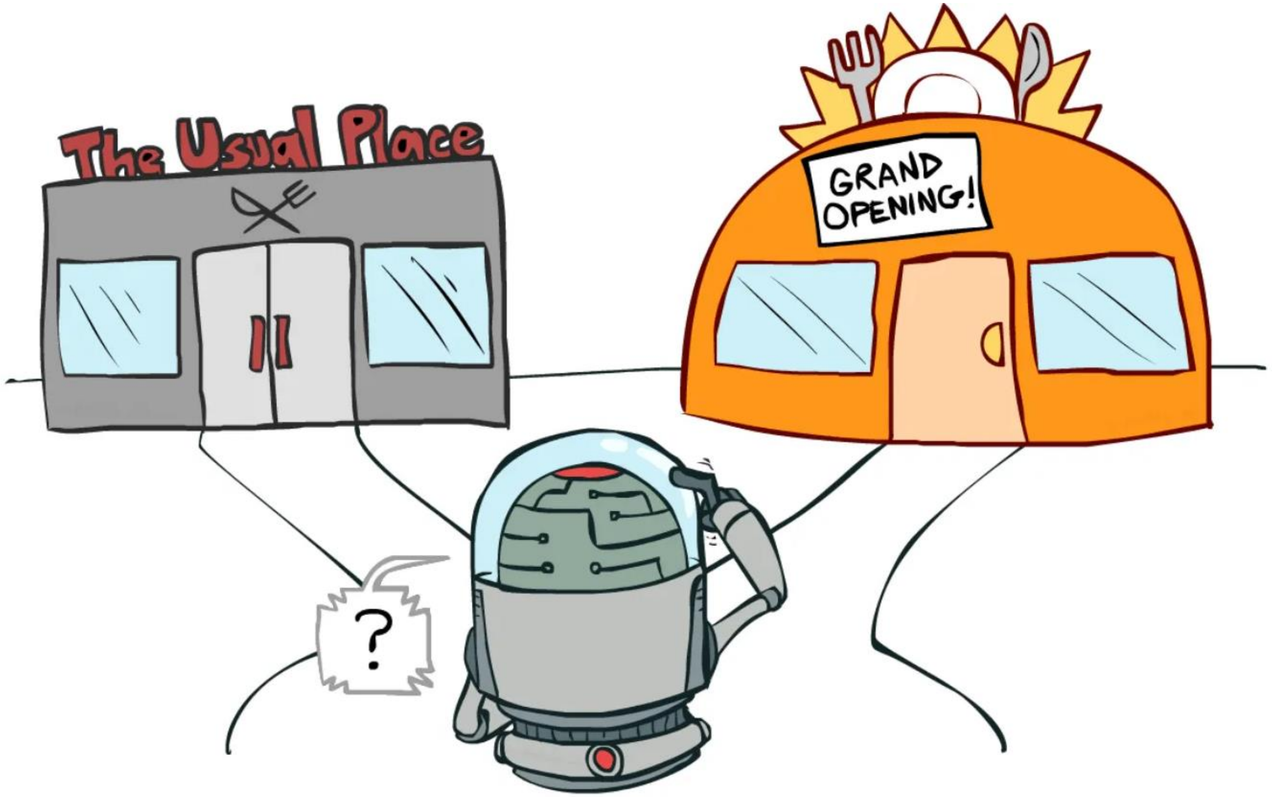
The RL dilemma

- Trade of between:
  - **Exploitation**:
    to obtain a lot of reward the agent must prefer rewarding actions that it tried in the past
  - **Exploration**:
    to discover such actions the agent has to try actions that is not been selected before



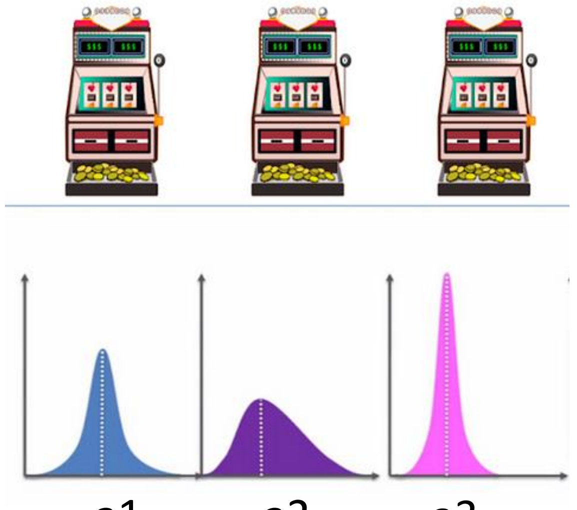
The RL dilemma

- Trade of between:
  - **Exploitation**:
    to obtain a lot of reward the agent must prefer rewarding actions that it tried in the past
  - **Exploration**:
    to discover such actions the agent has to try actions that is not been selected before



Dilemma:
Neither can be pursued exclusively without failing at the task

K-Armed bandit formalization



t	A	R	Q(1)	Q(2)	Q(3)
1					
2					
3					
4					

K-Armed bandit formalization

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

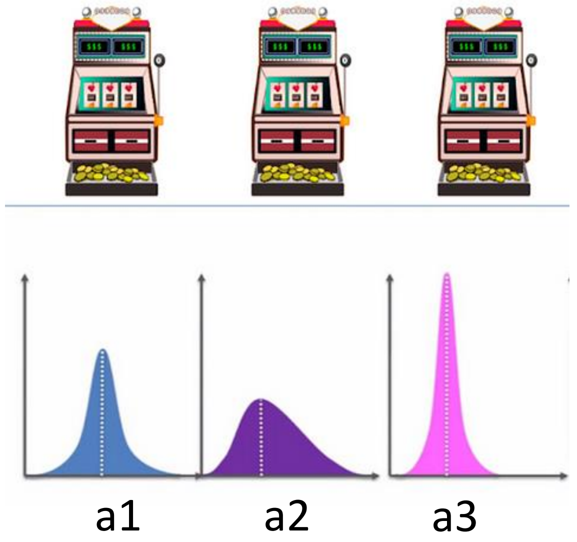
t – timestamp

A_t – action selected at t

R_t – reward given at t

$q_*(a)$ – value of action a

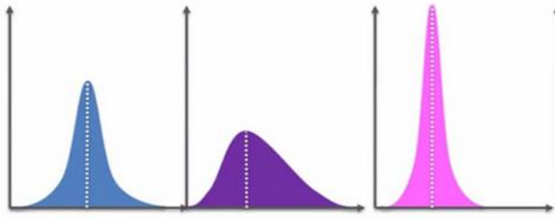
$Q_t(a)$ – estimated value of action a at timestep t



t	A	R	Q(1)	Q(2)	Q(3)
1					
2					
3					
4					

K-Armed bandit formalization

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$



t – timestamp

A_t – action selected at t

R_t – reward given at t

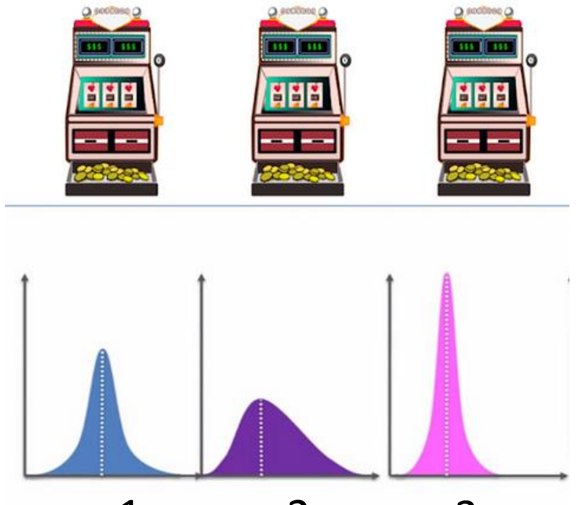
$q_*(a)$ – value of action a

$Q_t(a)$ – estimated value of action a at timestep t

t	A	R	Q(1)	Q(2)	Q(3)
1	1	5			
2					
3					
4					

K-Armed bandit formalization

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$



a1 a2 a3

t – timestamp

A_t – action selected at t

R_t – reward given at t

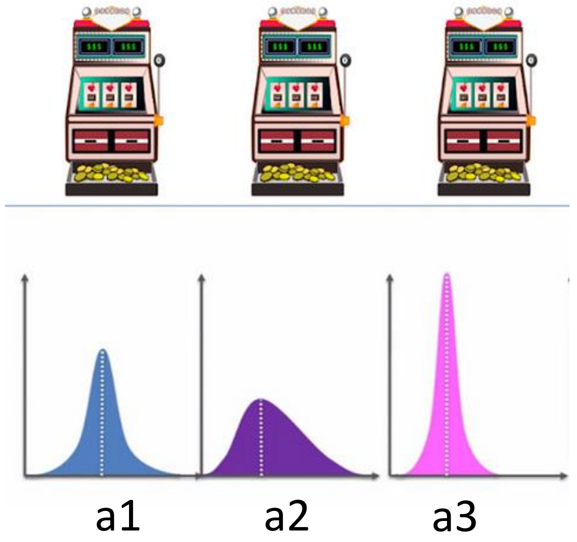
$q_*(a)$ – value of action a

$Q_t(a)$ – estimated value of action a at timestep t

t	A	R	Q(1)	Q(2)	Q(3)
1	1	5	5	0	0
2					
3					
4					

K-Armed bandit formalization

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$



t	A	R	Q(1)	Q(2)	Q(3)
1	1	5	5	0	0
2	2	3			
3	3	4			
4					

t – timestamp

A_t – action selected at t

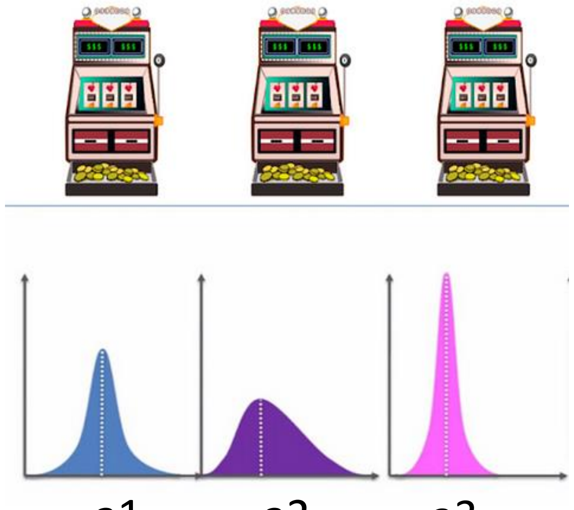
R_t – reward given at t

$q_*(a)$ – value of action a

$Q_t(a)$ – estimated value of action a at timestep t

K-Armed bandit formalization

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$



t – timestamp

A_t – action selected at t

R_t – reward given at t

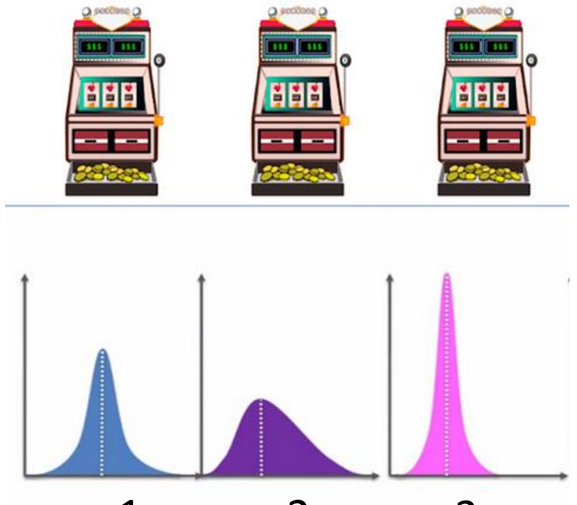
$q_*(a)$ – value of action a

$Q_t(a)$ – estimated value of action a at timestep t

t	A	R	Q(1)	Q(2)	Q(3)
1	1	5	5	0	0
2	2	3	5	3	0
3	3	4	5	3	4
4	1	4			

K-Armed bandit formalization

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$



t – timestamp

A_t – action selected at t

R_t – reward given at t

$q_*(a)$ – value of action a

$Q_t(a)$ – estimated value of action a at timestep t

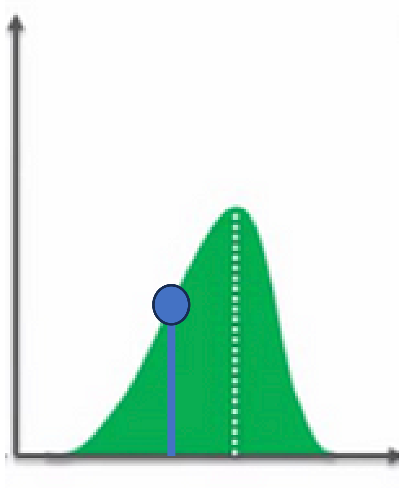
t	A	R	Q(1)	Q(2)	Q(3)
1	1	5	5	0	0
2	2	3	5	3	0
3	3	4	5	3	4
4	1	4	4.5	3	4

Law of Large Numbers (LLN)

In probability theory, the law of large numbers (LLN) is a mathematical theorem that states that the average of the results obtained from a large number of independent and identical random samples converges to the true value, if it exists

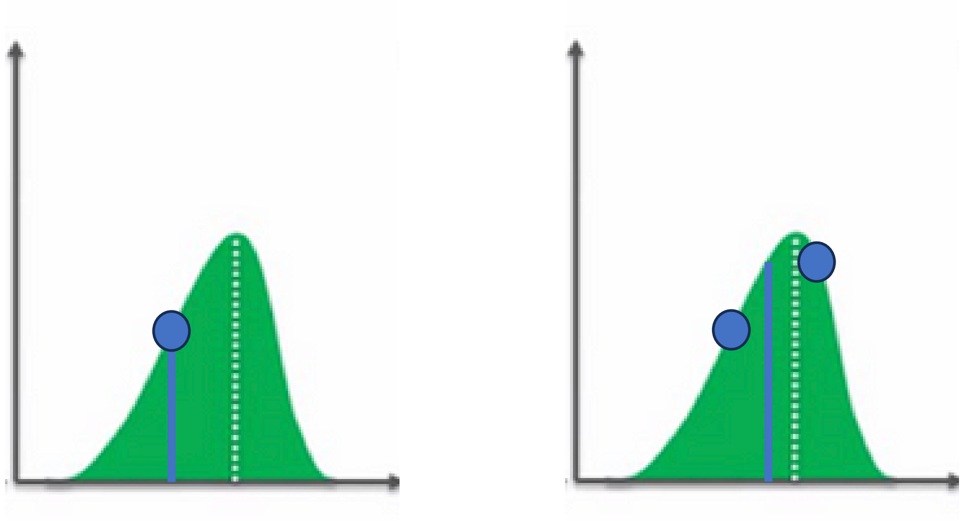
Law of Large Numbers (LLN)

In probability theory, the law of large numbers (LLN) is a mathematical theorem that states that the average of the results obtained from a large number of independent and identical random samples converges to the true value, if it exists



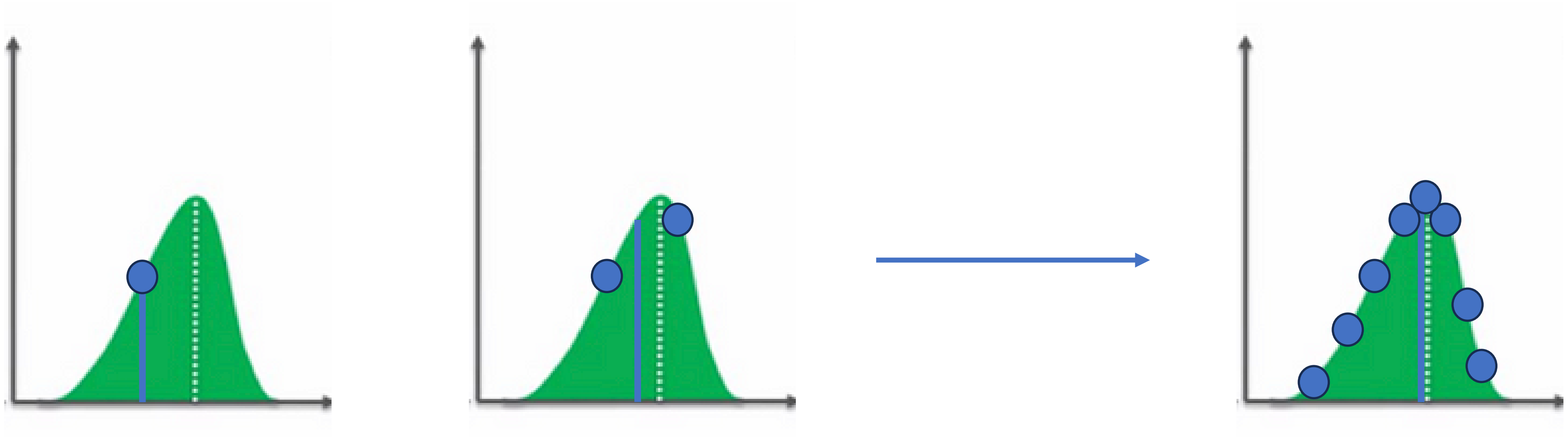
Law of Large Numbers (LLN)

In probability theory, the law of large numbers (LLN) is a mathematical theorem that states that the average of the results obtained from a large number of independent and identical random samples converges to the true value, if it exists



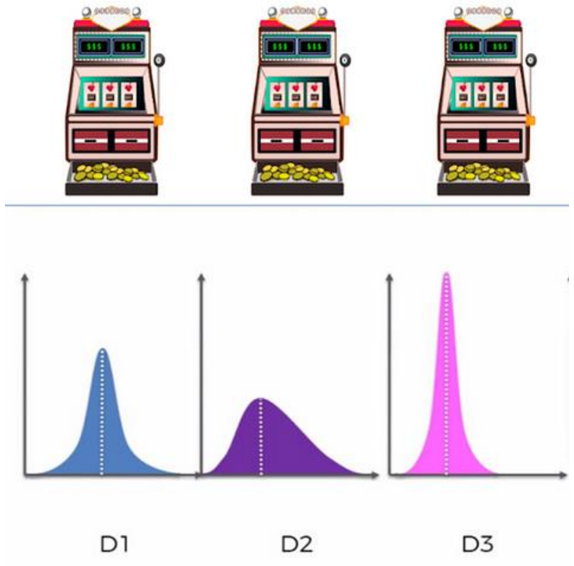
Law of Large Numbers (LLN)

In probability theory, the law of large numbers (LLN) is a mathematical theorem that states that the average of the results obtained from a large number of independent and identical random samples converges to the true value, if it exists



K-Armed bandit formalization

- At any timestep there is at least one action whose estimate value is greatest = **greedy action**

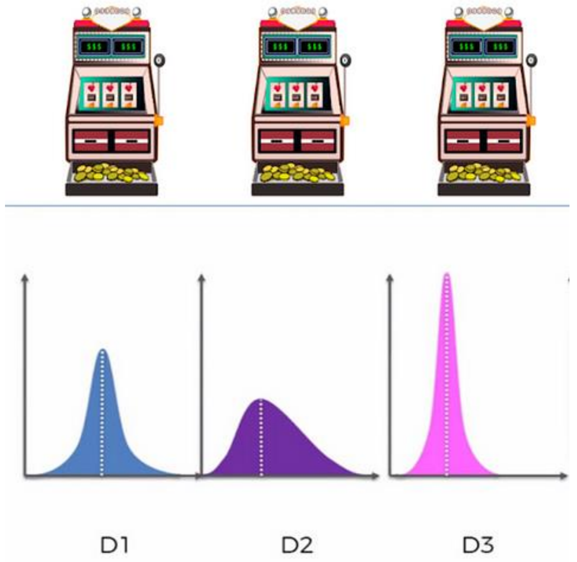


Which *a* selection is greedy?

t	a	r	Q(1)	Q(2)	Q(3)
1	1	5	5	0	0
2	2	3	5	3	0
3	3	4	5	3	4
4	1	4	4.5	3	4

K-Armed bandit formalization

- At any timestep there is at least one action whose estimate value is greatest = **greedy action**



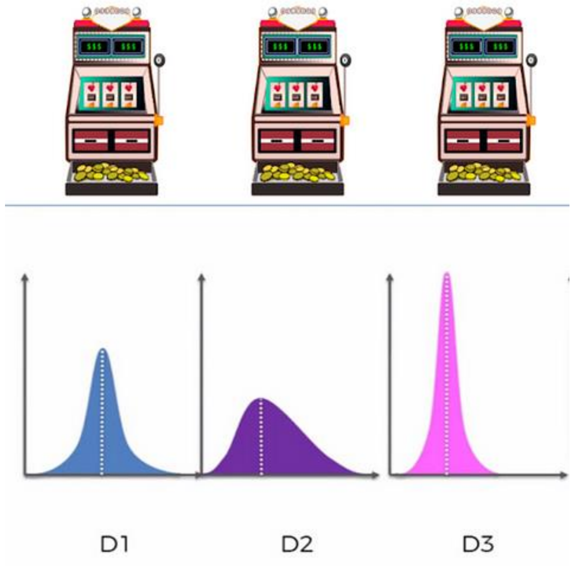
Which *a* selection is greedy?

t	a	r	Q(1)	Q(2)	Q(3)
1	$\boxed{1}$	5	5	0	0
2	2	3	5	3	0
3	3	4	5	3	4
4	$\boxed{1}$	4	4.5	3	4

K-Armed bandit formalization

- At any timestep there is at least one action whose estimate value is greatest = **greedy action**

- If **greedy action** is selected => **Exploiting** current knowledge
- If **non-greedy action** is selected => **Exploring** (enables to improve estimates)

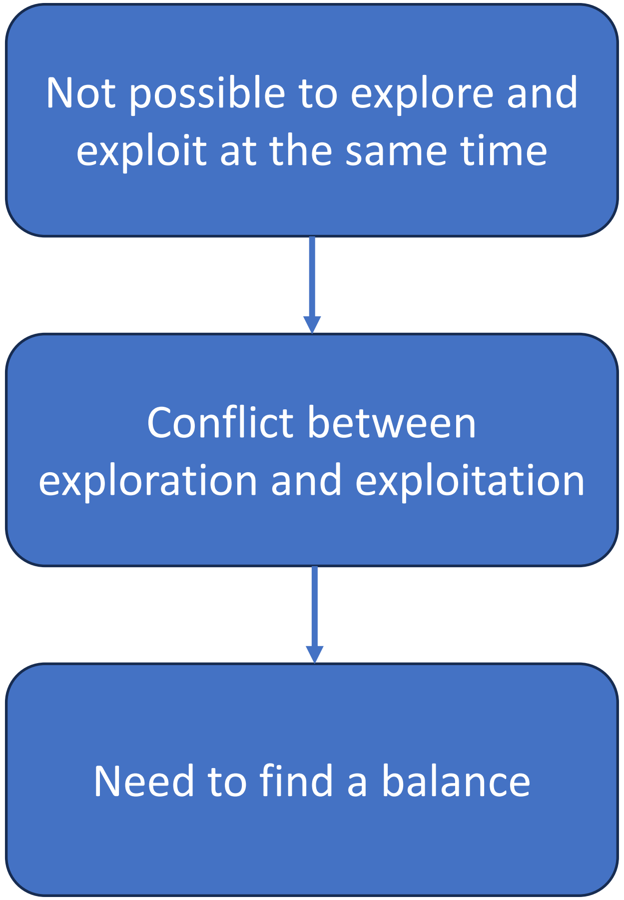


Which *a* selection is greedy?

t	a	r	Q(1)	Q(2)	Q(3)
1	$\boxed{1}$	5	5	0	0
2	2	3	5	3	0
3	3	4	5	3	4
4	$\boxed{1}$	4	4.5	3	4

K-Armed bandit formalization

- At any timestep there is at least one action whose estimate value is greatest = **greedy action**
- If **greedy action** is selected => **Exploiting** current knowledge
- If **non-greedy action** is selected => **Exploring** (enables to improve estimates)



Action-value Methods

- Estimating the values of actions

- Using the estimates to make action selection

Action-value Methods

- Estimating the values of actions

Sample Average Method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

- Using the estimates to make action selection

Action-value Methods

- Estimating the values of actions

Sample Average Method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

If denominator *0*: $Q_t(a)$ defined as a default value
If denominator goes to *Infinity*: $Q_t(a)$ goes to $q_*(a)$

- Using the estimates to make action selection

Action-value Methods

- Estimating the values of actions

Sample Average Method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

If denominator *0*: $Q_t(a)$ defined as a default value
If denominator goes to *Infinity*: $Q_t(a)$ goes to $q_*(a)$

- Using the estimates to make action selection

$$A_t \doteq \arg\max_a Q_t(a)$$

Simplest ('greedy') action selection rule:
select one of the actions with the highest estimated value
(if there are more select among them randomly)

Action-value Methods

- Estimating the values of actions

Sample Average Method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

If denominator *0*: $Q_t(a)$ defined as a default value
If denominator goes to *Infinity*: $Q_t(a)$ goes to $q_*(a)$

- Using the estimates to make action selection

$$A_t \doteq \underset{a}{\arg\max}\, Q_t(a)$$

Simplest ('greedy') action selection rule:
select one of the actions with the highest estimated value
(if there are more select among them randomly)

Greedy action selection always exploits current knowledge to maximise immediate reward

ε-greedy method

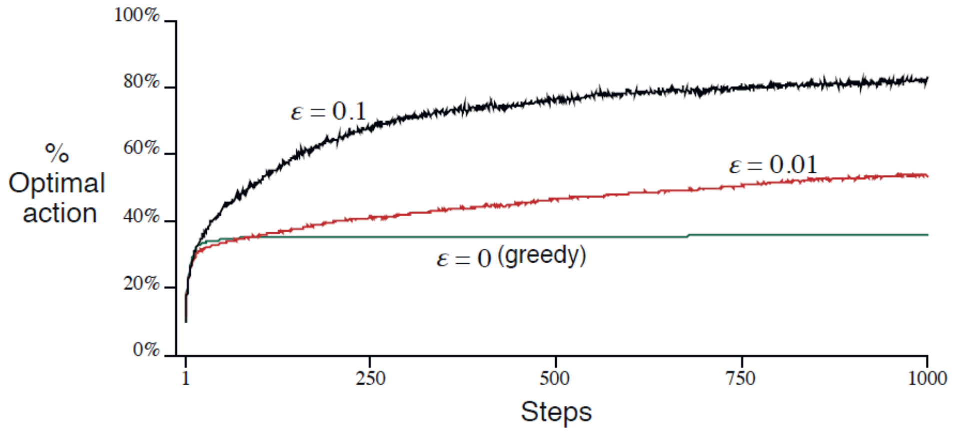
- Behave greedy most of the time but every once in a while, with a small probability (ε), select randomly from all the actions with equal probability, independently of the action-value estimates.

ε-greedy method

- Behave greedy most of the time but every once in a while, with a small probability (ε), select randomly from all the actions with equal probability, independently of the action-value estimates.

Advantage:
Every action will be sampled an infinite number of times ensuring all $Q_t(a)$ converges to $q_*(a)$

$$\text{Action at time(t)} \begin{cases} \max Q_t(a) & \text{with probability } 1-\epsilon \\ \text{any action (a)} & \text{with probability } \epsilon \end{cases}$$

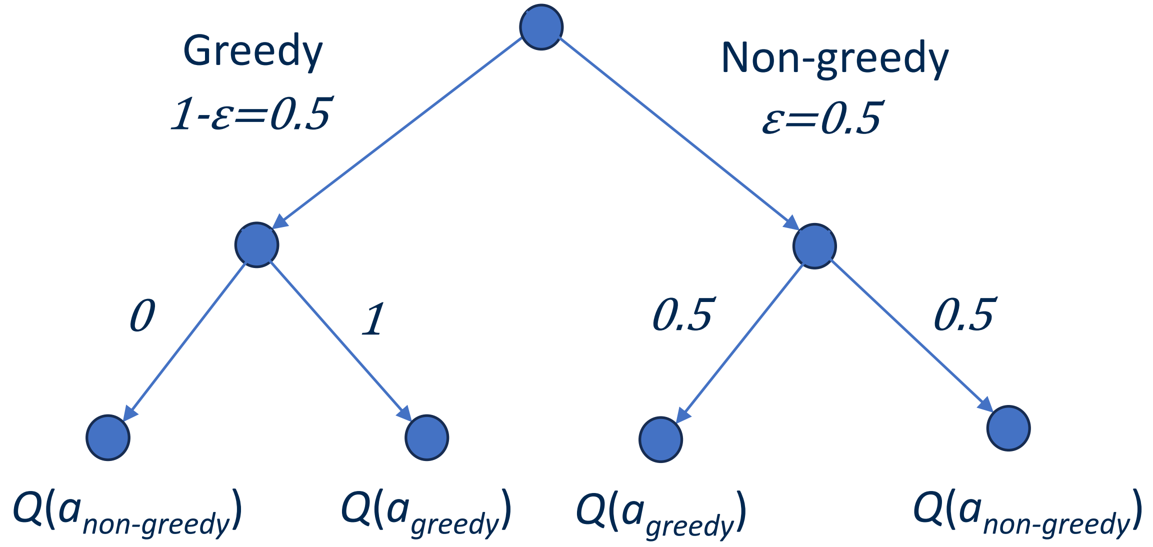


Exercise

In an ε-greedy action selection consider the case of two actions and $\varepsilon=0.5$. What is the probability that the greedy action is selected?

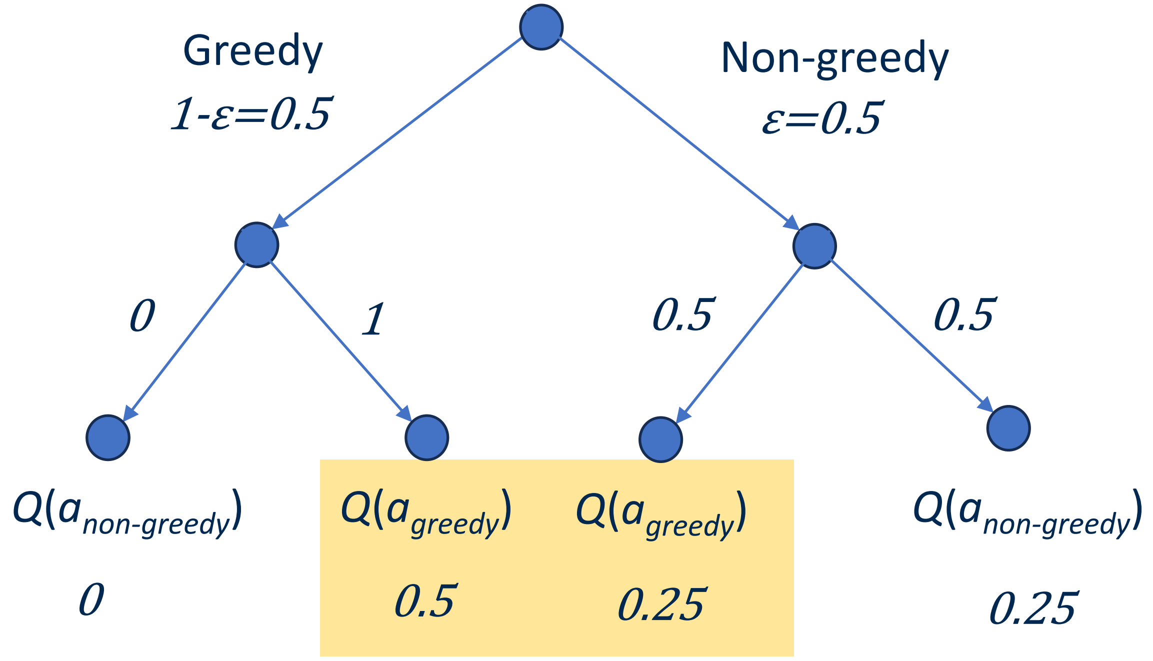
Exercise

In an ε-greedy action selection consider the case of two actions and $\varepsilon=0.5$. What is the probability that the greedy action is selected?



Exercise

In an ε-greedy action selection consider the case of two actions and $\varepsilon=0.5$. What is the probability that the greedy action is selected?



75%

Exercise

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?

Exercise

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?

Init →

t	A_t	R_t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
0	-	-	0	0	0	0
1	1	1				
2	2	1				
3	2	2				
4	2	2				
5	3	0				

Exercise

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?

Init →

t	A_t	R_t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
0	-	-	0	0	0	0
1	1	1				
2	2	1				
3	2	2				
4	2	2				
5	3	0				

Greedy / ε-greedy

Exercise

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?

Init

t	A_t	R_t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
0	-	-	0	0	0	0
1	1	1	1	0	0	0
2	2	1				
3	2	2				
4	2	2				
5	3	0				

Greedy / ε-greedy

Exercise

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?

Init →

t	A_t	R_t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
0	-	-	0	0	0	0
1	1	1	1	0	0	0
2	2	1	1	1	0	0
3	2	2				
4	2	2				
5	3	0				

Greedy / ε-greedy

ε-greedy

Exercise

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?

Init

t	A_t	R_t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
0	-	-	0	0	0	0
1	1	1	1	0	0	0
2	2	1	1	1	0	0
3	2	2	1	1.5	0	0
4	2	2				
5	3	0				

Greedy / ε-greedy

ε-greedy

Greedy / ε-greedy

Exercise

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?

Init

t	A_t	R_t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
0	-	-	0	0	0	0
1	1	1	1	0	0	0
2	2	1	1	1	0	0
3	2	2	1	1.5	0	0
4	2	2	1	1.66	0	0
5	3	0				

Greedy / ε-greedy

ε-greedy

Greedy / ε-greedy

Greedy / ε-greedy

Exercise

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?

	t	A_t	R_t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$	
Init →	0	-	-	0	0	0	0	Greedy / ε-greedy
	1	1	1	1	0	0	0	ε-greedy
	2	2	1	1	1	0	0	Greedy / ε-greedy
	3	2	2	1	1.5	0	0	Greedy / ε-greedy
	4	2	2	1	1.66	0	0	ε-greedy
	5	3	0	1	1.66	0	0	

Incremental implementation

- In the action-value methods all estimated action values are averages of observed rewards

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

- Recording all rewards: memory and computational would grow over time

Incremental implementation

$$
\begin{aligned}
Q_{n+1} &= \boxed{\frac{1}{n}\sum_{i=1}^{n} R_i} \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\boxed{\frac{1}{n-1}\sum_{i=1}^{n-1} R_i}\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \boxed{\frac{1}{n}}\Big[R_n - Q_n\Big]
\end{aligned}
$$

Step size (α)

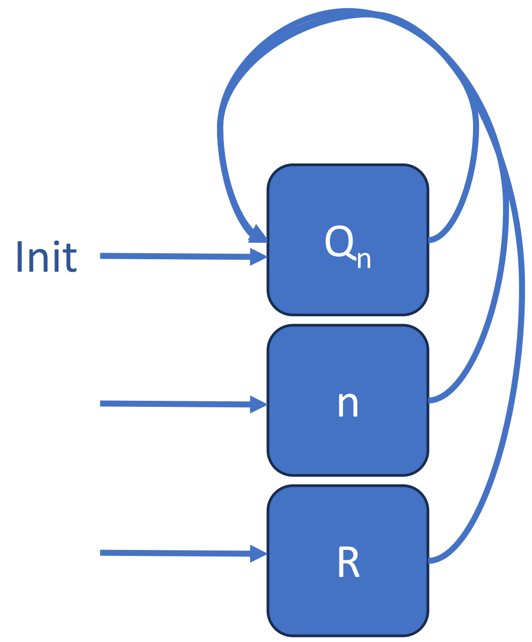
Incremental implementation

$$
\begin{aligned}
Q_{n+1} &= \boxed{\frac{1}{n}\sum_{i=1}^{n} R_i} \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\boxed{\frac{1}{n-1}\sum_{i=1}^{n-1} R_i}\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \boxed{\frac{1}{n}}\Big[R_n - Q_n\Big]
\end{aligned}
$$

Step size (α)

Update rule:

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



Need to store only 3 values
=Memory efficient

Pseudo implementation

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

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

$R \leftarrow bandit(A)$

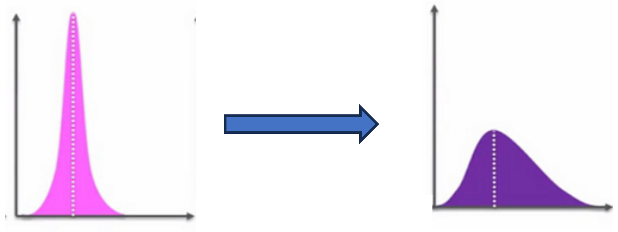
$N(A) \leftarrow N(A) + 1$

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

Tracking a Nonstationary Problem

- If $\alpha = 1/n$
  - good for stationary problems

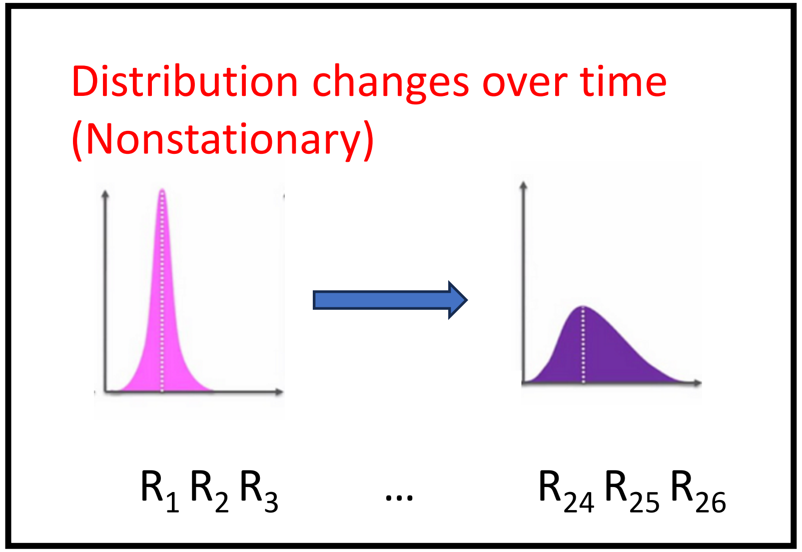
Distribution changes over time
(Nonstationary)



R_1 R_2 R_3 … R_{24} R_{25} R_{26}

Tracking a Nonstationary Problem

- If $\alpha = 1/n$
  - good for stationary problems
- **If $\alpha = constant\,(0, 1]$**
  - **good for nonstationary problems**



Estimated expected reward update:

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

Tracking a Nonstationary Problem

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

Tracking a Nonstationary Problem

$$\boxed{Q_{n+1}} = Q_n + \alpha\Big[R_n - Q_n\Big]$$
$$= \boxed{\alpha R_n + (1-\alpha)Q_n}$$

Tracking a Nonstationary Problem

$$
\begin{aligned}
\boxed{Q_{n+1}} &= Q_n + \alpha\Big[R_n - Q_n\Big] \\
&= \boxed{\alpha R_n + (1-\alpha)\boxed{Q_n}} \\
&= \alpha R_n + (1-\alpha)\boxed{\left[\alpha R_{n-1} + (1-\alpha) Q_{n-1}\right]}
\end{aligned}
$$

Tracking a Nonstationary Problem

$$\begin{aligned}
\boxed{Q_{n+1}} &= Q_n + \alpha\Big[R_n - Q_n\Big] \\
&= \boxed{\alpha R_n + (1-\alpha)\boxed{Q_n}} \\
&= \alpha R_n + (1-\alpha)\boxed{\left[\alpha R_{n-1} + (1-\alpha)Q_{n-1}\right]} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \\
&\qquad \cdots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n \boxed{Q_1}
\end{aligned}$$

Extend

Tracking a Nonstationary Problem

$$
\begin{aligned}
\boxed{Q_{n+1}} &= Q_n + \alpha\Big[R_n - Q_n\Big] \\
&= \boxed{\alpha R_n + (1-\alpha)\boxed{Q_n}} \\
&= \alpha R_n + (1-\alpha)\boxed{\left[\alpha R_{n-1} + (1-\alpha)Q_{n-1}\right]} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \\
&\qquad \cdots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n \boxed{Q_1} \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i
\end{aligned}
$$

Extend

Weighted average

Exponentially weighted average

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

Exponentially weighted average

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = (1-\alpha)^n + (1-\alpha)^n \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} =$$

$$= (1-\alpha)^n (1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i})$$

Exponentially weighted average

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = (1-\alpha)^n + (1-\alpha)^n \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} =$$

$$= \boxed{(1-\alpha)^n (1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i})}$$

$$\boxed{\begin{aligned} &(1-\alpha)^n + (1-\alpha)^{-n} = 1 \\ &(1-\alpha)^{-n} = 1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} \end{aligned}}$$

Exponentially weighted average

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = (1-\alpha)^n + (1-\alpha)^n \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} =$$

$$= \boxed{(1-\alpha)^n (1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i})}$$

$$1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} = 1 + \alpha \boxed{\sum_{i=1}^{n} (\frac{1}{1-\alpha})^{i}} =$$

$$= 1 + \alpha \boxed{\frac{1}{1-\alpha} \frac{(\frac{1}{1-\alpha})^n - 1}{\frac{1}{1-\alpha} - 1}} =$$

$$(1-\alpha)^n + (1-\alpha)^{-n} = 1$$

$$(1-\alpha)^{-n} = 1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i}$$

Geometric series:

$$\sum_{i=1}^{n} x^i = x \frac{x^n - 1}{x - 1}$$

Exponentially weighted average

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = (1-\alpha)^n + (1-\alpha)^n \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} =$$

$$= \boxed{(1-\alpha)^n (1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i})}$$

$$1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} = 1 + \alpha \boxed{\sum_{i=1}^{n} (\frac{1}{1-\alpha})^i} =$$

$$= 1 + \alpha \boxed{\frac{1}{\boxed{1-\alpha \quad \frac{1}{1-\alpha} - 1}} \frac{(\frac{1}{1-\alpha})^n - 1}{}} =$$

$$= 1 + \alpha \frac{1}{\boxed{\alpha}} (\frac{1}{1-\alpha})^n - 1$$

$$= (\frac{1}{1-\alpha})^n = (1-\alpha)^{-n}$$

$$(1-\alpha)^n + (1-\alpha)^{-n} = 1$$

$$(1-\alpha)^{-n} = 1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i}$$

Geometric series:

$$\sum_{i=1}^{n} x^i = x \frac{x^n - 1}{x - 1}$$

$$(1-\alpha)(\frac{1}{1-\alpha} - 1) = (1-\alpha)(\frac{1}{1-\alpha} - \frac{1-\alpha}{1-\alpha}) =$$

$$= 1 - 1 + \alpha = \alpha$$

Exponentially weighted average

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = (1-\alpha)^n + (1-\alpha)^n \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} =$$

$$= \boxed{(1-\alpha)^n (1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i})}$$

$$1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i} = 1 + \alpha \boxed{\sum_{i=1}^{n} (\frac{1}{1-\alpha})^i} =$$

$$= 1 + \alpha \boxed{\frac{1}{\boxed{1-\alpha}} \frac{(\frac{1}{1-\alpha})^n - 1}{\boxed{\frac{1}{1-\alpha} - 1}}} =$$

$$= 1 + \alpha \frac{1}{\boxed{\alpha}} (\frac{1}{1-\alpha})^n - 1$$

$$= (\frac{1}{1-\alpha})^n = \boxed{(1-\alpha)^{-n}}$$

$$(1-\alpha)^n + (1-\alpha)^{-n} = 1$$

$$(1-\alpha)^{-n} = 1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{-i}$$

Geometric series:

$$\sum_{i=1}^{n} x^i = x \frac{x^n - 1}{x - 1}$$

$$(1-\alpha)(\frac{1}{1-\alpha} - 1) = (1-\alpha)(\frac{1}{1-\alpha} - \frac{1-\alpha}{1-\alpha}) =$$

$$= 1 - 1 + \alpha = \alpha$$

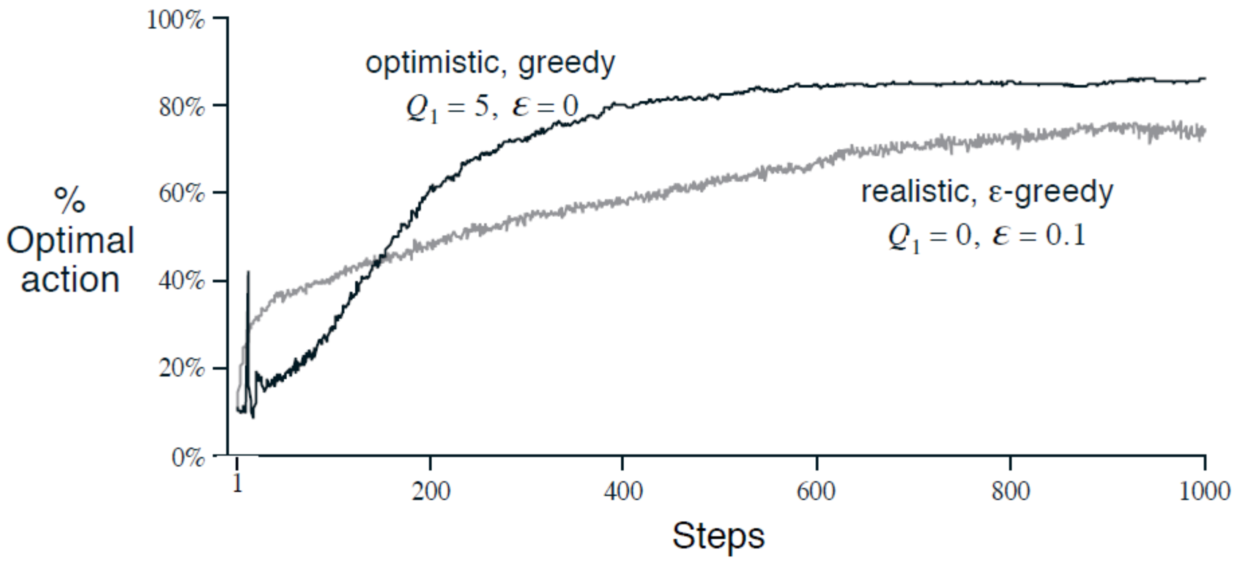
Optimistic Initial Values

- Dependent on initial action-value estimate $Q_1(a)$
  - **Biased** by the initial estimates
  - Usually not a problem
  - Supply prior knowledge
- Optimistic initial values encouraging exploration
  - Good in stationary problems

Optimistic Initial Values

- Dependent on initial action-value estimate $Q_1(a)$
  - **Biased** by the initial estimates
  - Usually not a problem
  - Supply prior knowledge
- Optimistic initial values encouraging exploration
  - Good in stationary problems

- Set the values optimistically high
- Greedy action selection will choose a high value
- New reward will be smaller
- Greedy action selection will choose a high value from the high initial values



Upper-Confidence-Bound Action Selection

- Greedy – selects the best-looking action
- ε-greedy – non greedy actions are tried with no preference
- UCB - select among the non-greedy actions according to their potential

$$A_t \doteq \underset{a}{\arg\max} \left[Q_t(a) + c \underbrace{\sqrt{\frac{\ln t}{N_t(a)}}}_{\text{Measure of uncertainty} \sim \text{variance}} \right]$$

Measure of uncertainty
~variance

$\ln t$ – natural logarithm of time
$N_t(a)$ – number of times action a is selected
$c>0$ – degree of exploration

Upper-Confidence-Bound Action Selection

- Greedy – selects the best-looking action
- ε-greedy – non greedy actions are tried with no preference
- UCB - select among the non-greedy actions according to their potential

$$A_t \doteq \underset{a}{\arg\max}\left[Q_t(a) + c\underbrace{\sqrt{\frac{\ln t}{N_t(a)}}}_{\text{Measure of uncertainty variance}}\right]$$

$\ln t$ – natural logarithm of time
$N_t(a)$ – number of times action a is selected
$c>0$ – degree of exploration

Con:
-Not good in large state spaces
- $N_t(a)$ needs to be stored
- not good in nonstationary cases

Gradient Bandit Algorithm

- Learn a numerical preference for each action – $H_t(a)$
- Large preference = action taken more often
- Preference ≠ Reward
- Only relative preference of an action over another

Soft-max distribution:

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

$\pi_t(a)$ - the probability of taking action a at time t

Gradient Bandit Algorithm

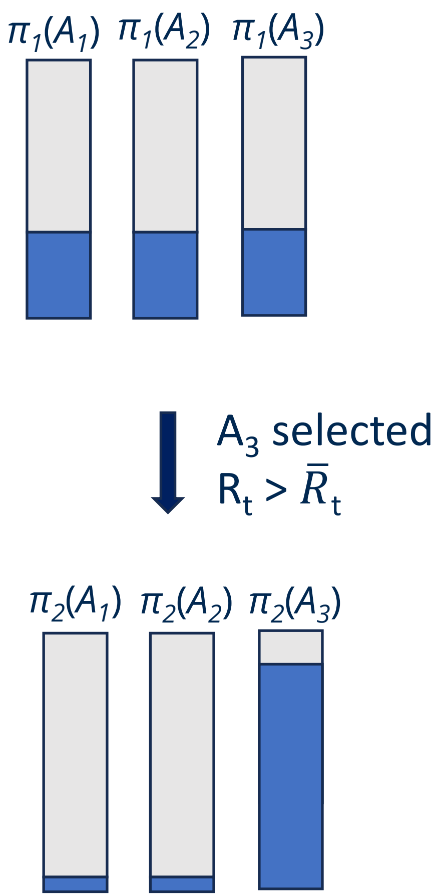
$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

Update based on stochastic gradient:

If positive preference will increase

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\boxed{\left(R_t - \bar{R}_t\right)}\left(1 - \pi_t(A_t)\right), \qquad \text{and}$$

$$H_{t+1}(a) \doteq H_t(a) - \alpha\left(R_t - \bar{R}_t\right)\pi_t(a), \qquad \text{for all } a \neq A_t$$

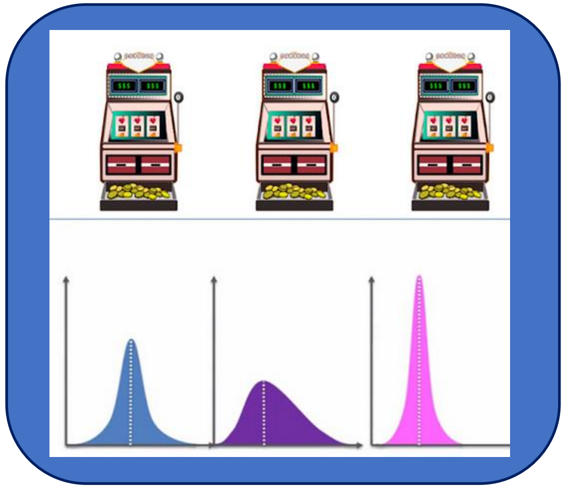


$\alpha > 0$ – step size parameter
k – number of actions
$H_t(a)$ – the preference of action a at time t
$\pi_t(a)$ - the probability of taking action a at time t
$\bar{R}$ - the average of all the rewards up through and including time t
(base line reward, not depend on selected action)

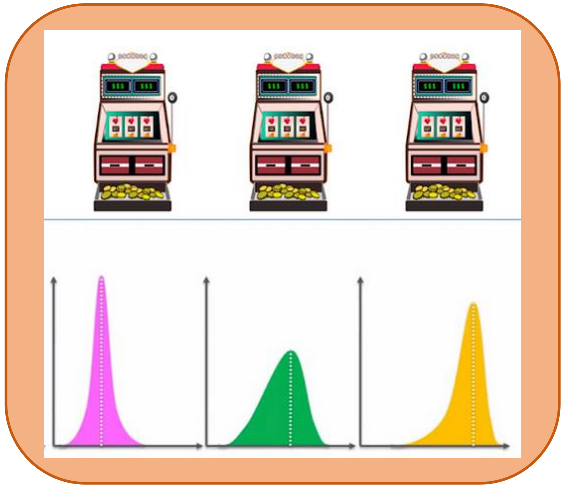
Associative search

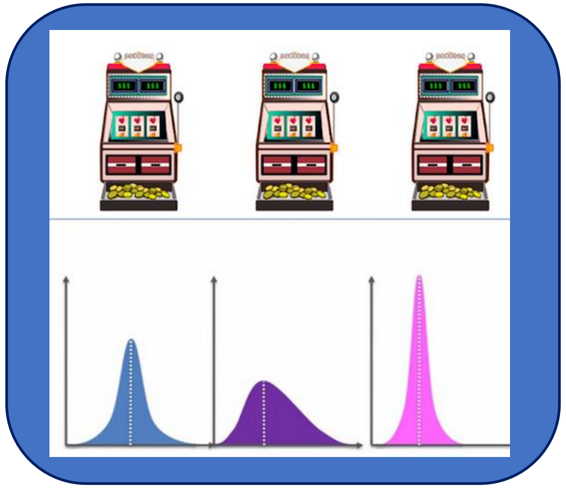
So far, we dealt with**:**

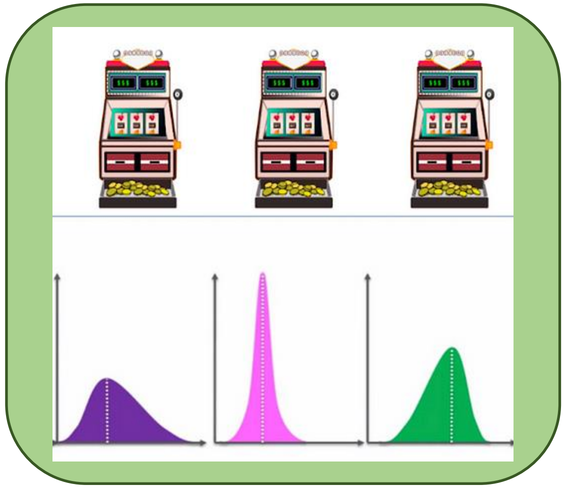
- **Nonassociative** task
- Find the best actions in **one situation**
- **Stationary** or **Nonstationary**



Associative search

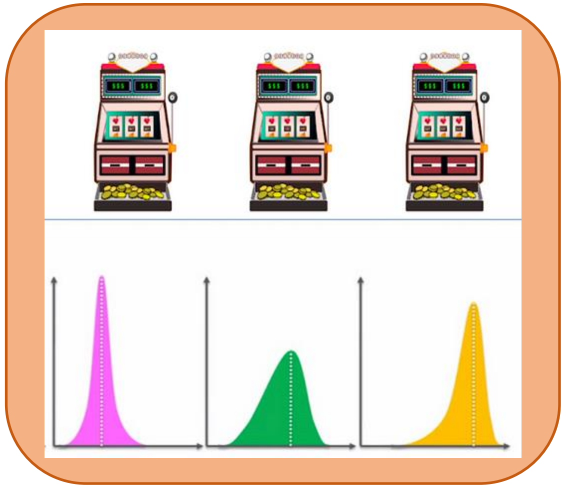


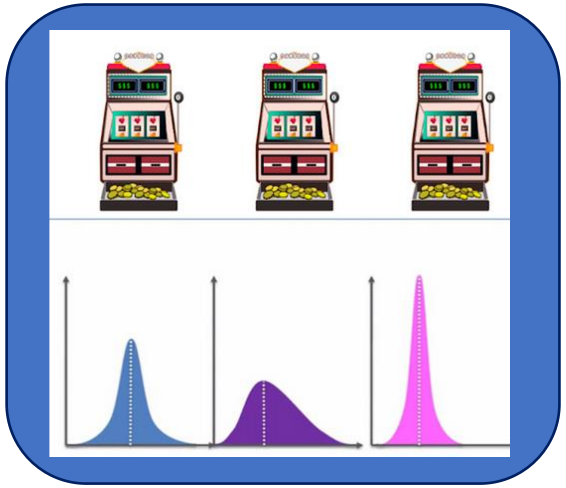


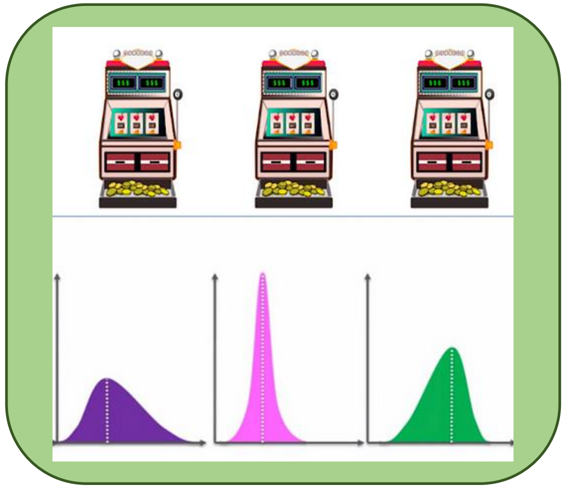


What if we have multiple situations?

Associative search



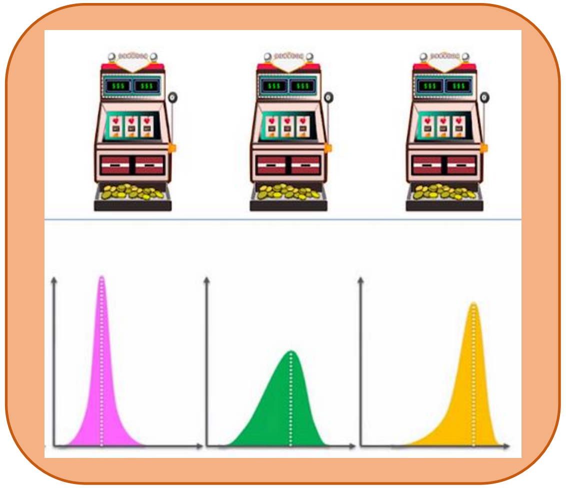


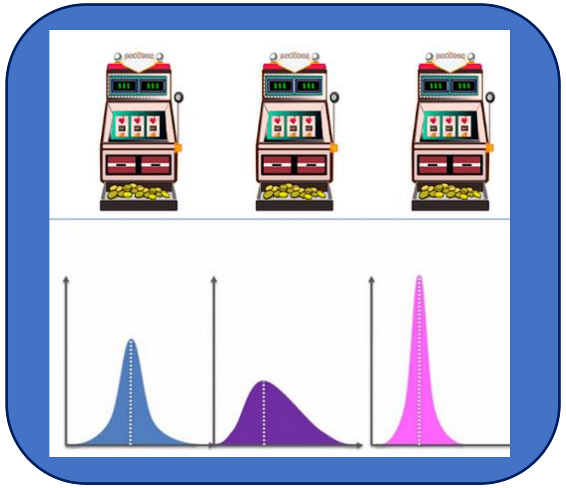


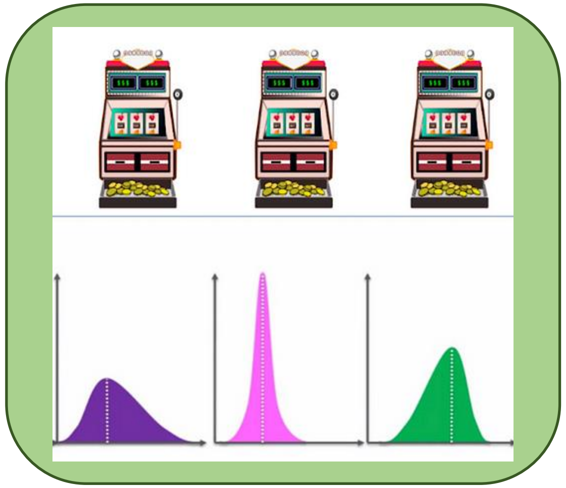
What if we have multiple situations?

And face one randomly each timestep.

Associative search



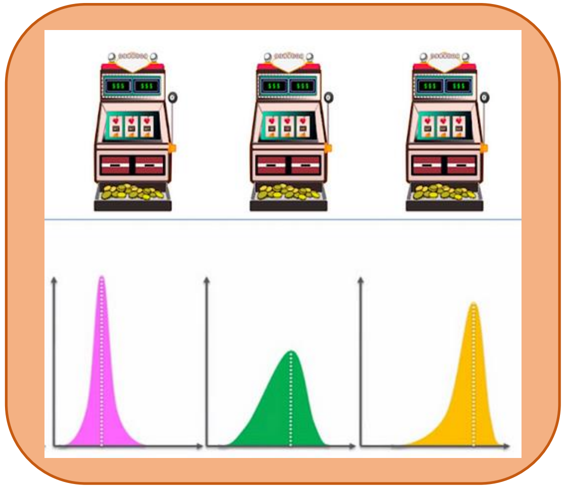


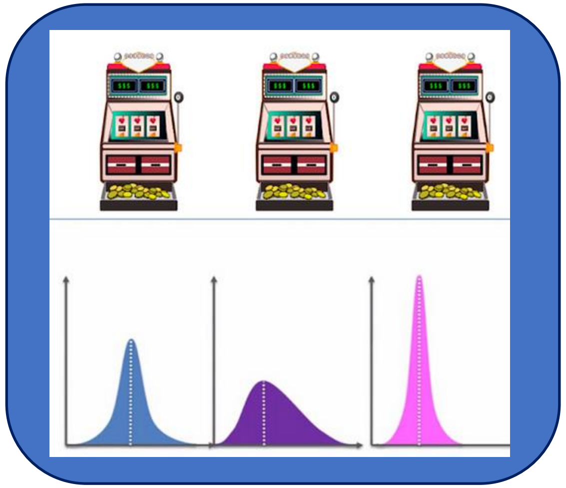


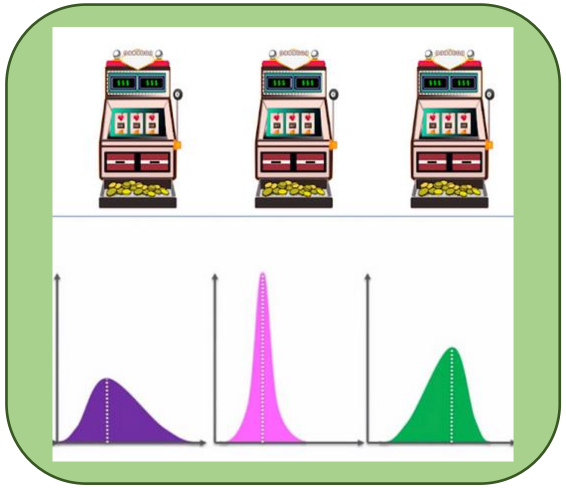
Solution:

Learn distinct policies for each situation

Associative search







Solution:

Learn distinct policies for each situation

If actions are allowed to affect the next situation as well as the reward

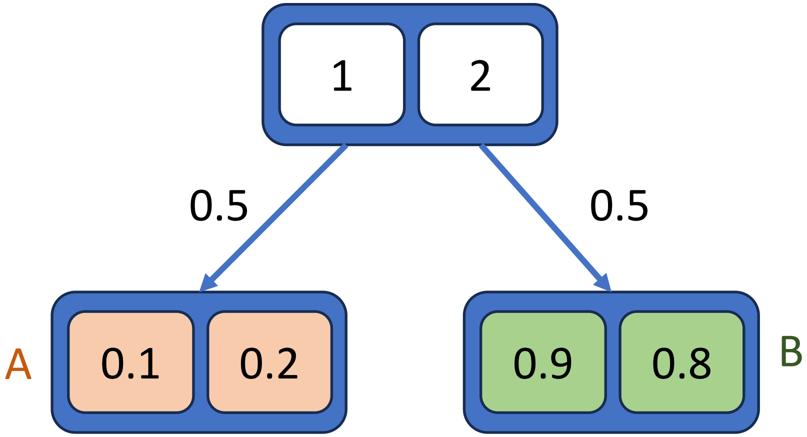
= **Full Reinforcement Learning Task**

Exercise

Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specially, suppose that, for any time step, the true values of actions 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).

a, If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it?

b, Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?

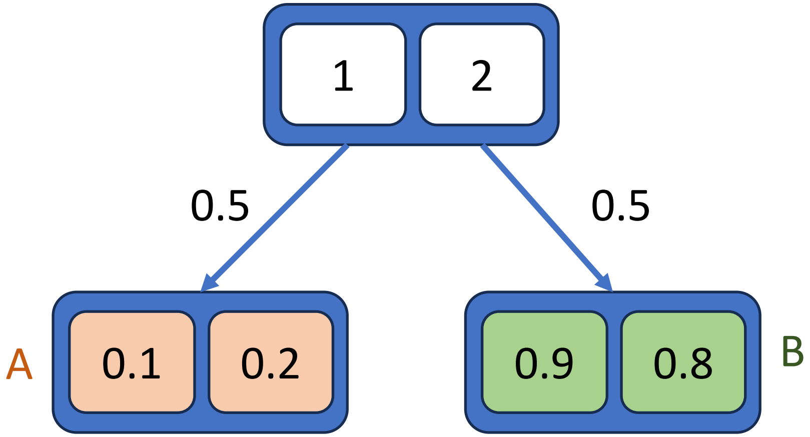


Exercise

Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specially, suppose that, for any time step, the true values of actions 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).

a, If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it?

b, Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?



a, You do not know:

Action 1: $0.5 \cdot 0.1 + 0.5 \cdot 0.9 = 0.5$

Action 2: $0.5 \cdot 0.2 + 0.5 \cdot 0.8 = 0.5$

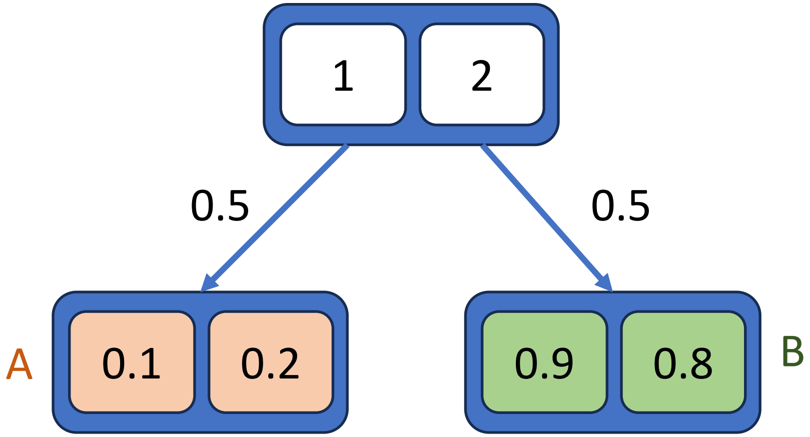
Max = 0.5

Exercise

Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specially, suppose that, for any time step, the true values of actions 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).

a, If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it?

b, Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?



a, You do not know:

Action 1: $0.5 \cdot 0.1 + 0.5 \cdot 0.9 = 0.5$

Action 2: $0.5 \cdot 0.2 + 0.5 \cdot 0.8 = 0.5$

Max = 0.5

b, You do know:

Action Best: $0.5 \cdot 0.2 + 0.5 \cdot 0.9 = 0.55$

Max = 0.55

ELTE | FACULTY OF INFORMATICS

Thank you for your attention!