SAPIENZA
UNIVERSITÀ DI ROMA

Autonomous Networking

Gaia Maselli
Dept. of Computer Science

How can we balance exploration with exploitation?

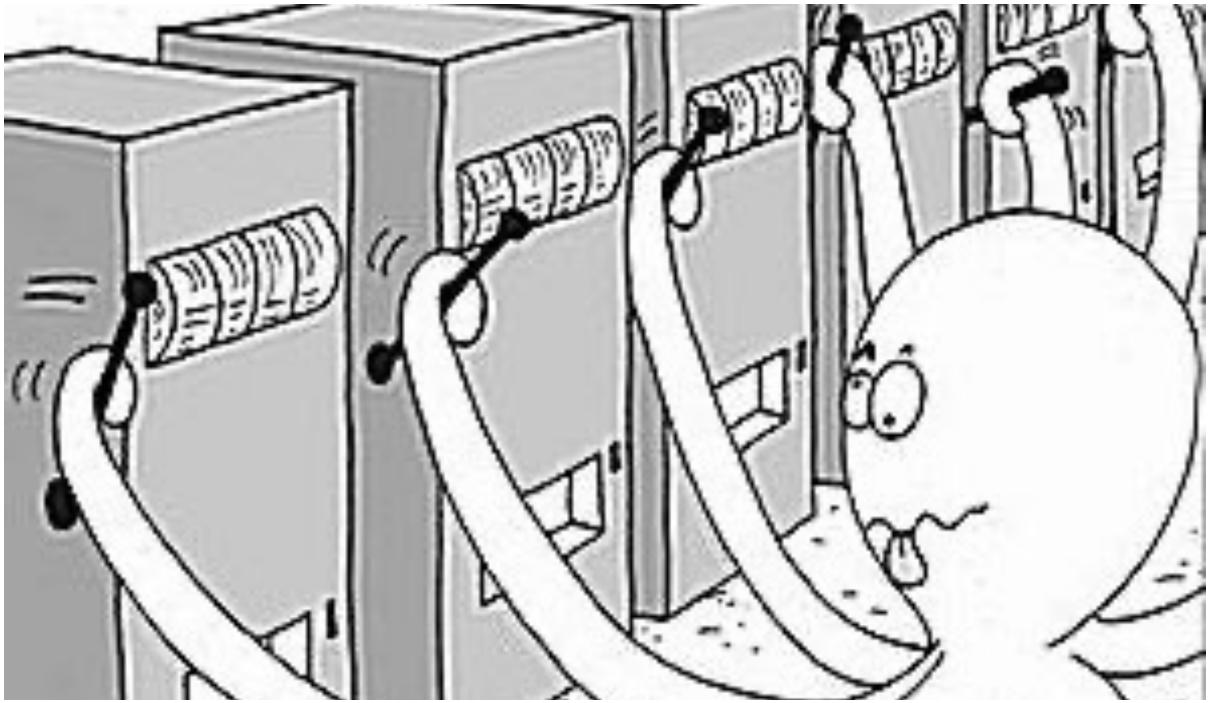
Exploration exploitation trade-off

- Rewards evaluate actions taken (evaluative feedback)
- Evaluative feedback depends on the action taken
- There is need for active exploration (explicit search for good behavior)
- Should the agent explore or exploit?
- Let us consider a simplified version of RL problems

K-armed bandit problem

- Problem: **you are faced repeatedly with a choice among K different options**, or **actions**
- After each choice you receive a numerical reward chosen from a stationary probability distribution that depends on the action you selected
- Objective: maximize the expected total reward over some time period (ex. 1000 action selections, or time-steps)

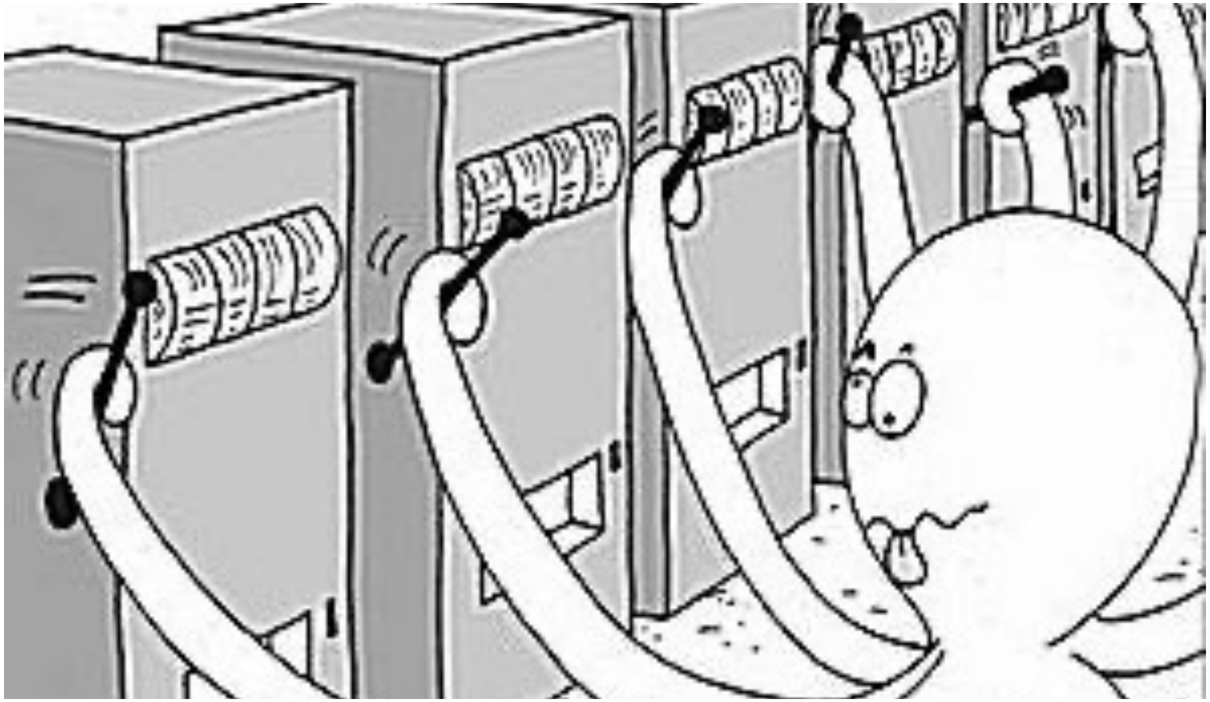
K-armed bandit problem



- Analogy with slot machine "one-armed bandit" except that it has k levers instead of one

- Each action selection is like a play of one of the slot machine's levers, and the rewards are the payoffs for hitting the jackpot
- Through repeated action selections you are to maximize your winning by concentrating your actions on the best levers

K-armed bandit problem



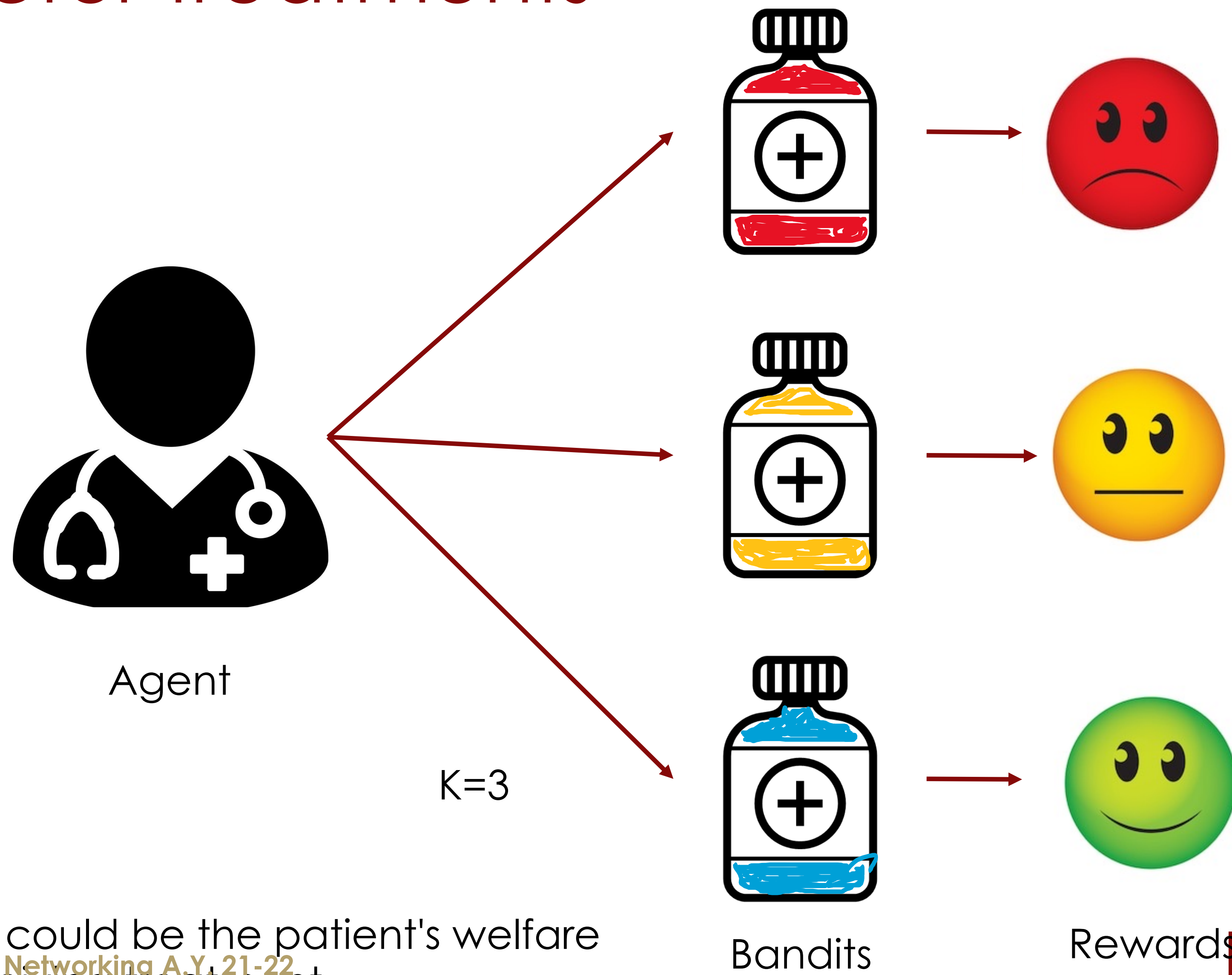
- Analogy with slot machine "one-armed bandit" except that it has k levers instead of one

- Each action selection is like a play of one of the slot machine's levers, and the rewards are the payoffs for hitting the jackpot
- Through repeated action selections you are to maximize your winning by concentrating your actions on the best levers

K-armed bandit

- Formalization
- Set of action *A* (or “arms”)
- Reward function *R* that follows an unknown probability distributions
- There is only one state
- At each step *t*, the agent selects an action in *A*
- The environment generates a reward
- The goal is to maximize the cumulative reward

Doctor treatments



Rewards could be the patient's welfare after receiving treatment

Action-value function

- For the doctor to decide which action is best, we must define the value of taking each action.
- We call these values the action values or the action value function
- **Action value**: the value of selecting an action *a* is defined as the expected reward we receive by taking that action

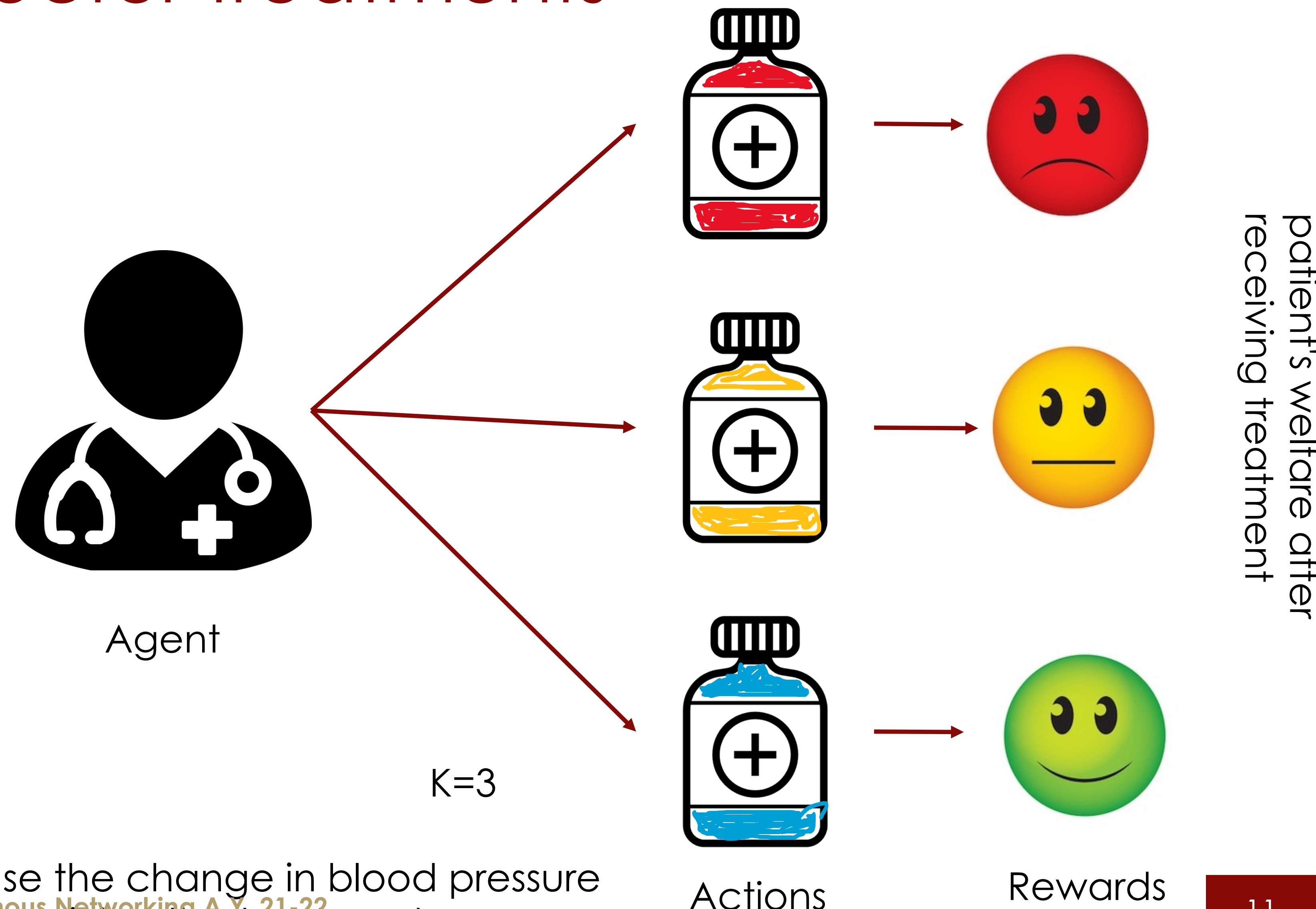
$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

- The goal of the agent is to **maximise** the **expected reward**

How can we estimate action-value?

- Sample-average method

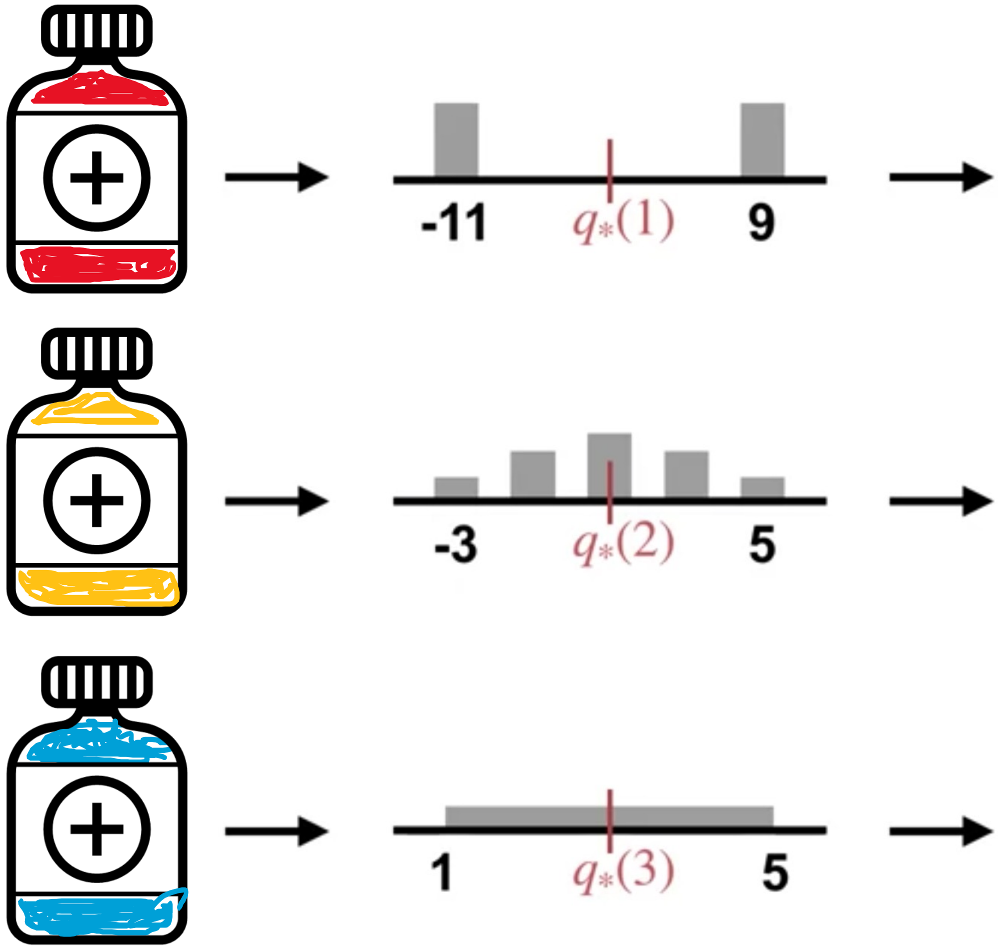
Doctor treatments



Let’s use the change in blood pressure after receiving the treatment.

Calculating $q_*(a)$

Each treatment may yield rewards following different probability distributions



Bernoulli

$$q_*(a) = .5 \times -11 + .5 \times 9$$

$$q_*(a) = 1$$

Binomial

$$q_*(a) = 3$$

Uniform

Q_* is the mean of the distributions for each action

Sample-average estimate

- The reward distribution is not known → the doctor will run many trial to learn about each treatment
- The estimated value for action *a* is the sum of rewards observed when taking action *a* divided by the total number of times action *a* has been taken

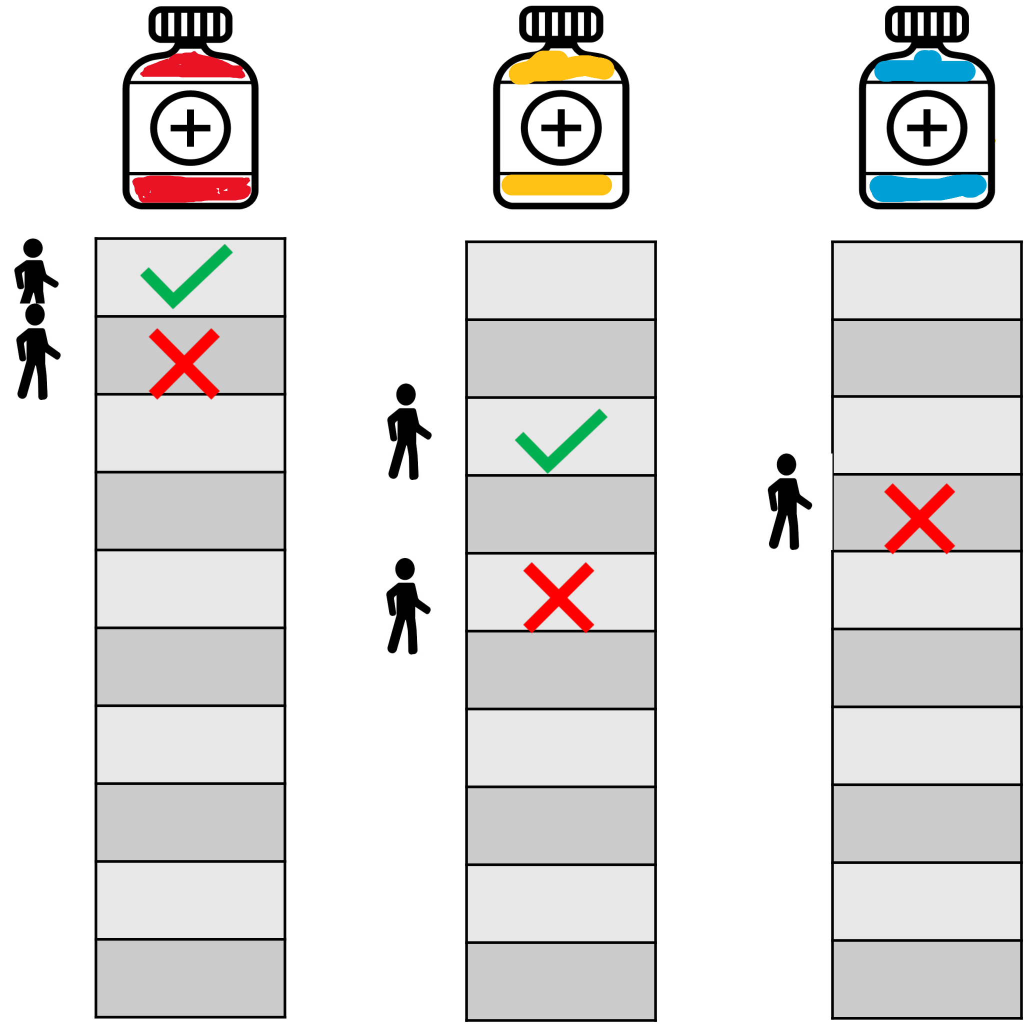
$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}},$$

- Where $1_{predicate}$ denotes the random variable that is 1 if predicate is true and 0 if it is not.
- If the denominator is zero, then we define $Q_t(a)=0$

Example

A reward of 1 if treatment succeeds, 0 otherwise

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i}{t-1}$$



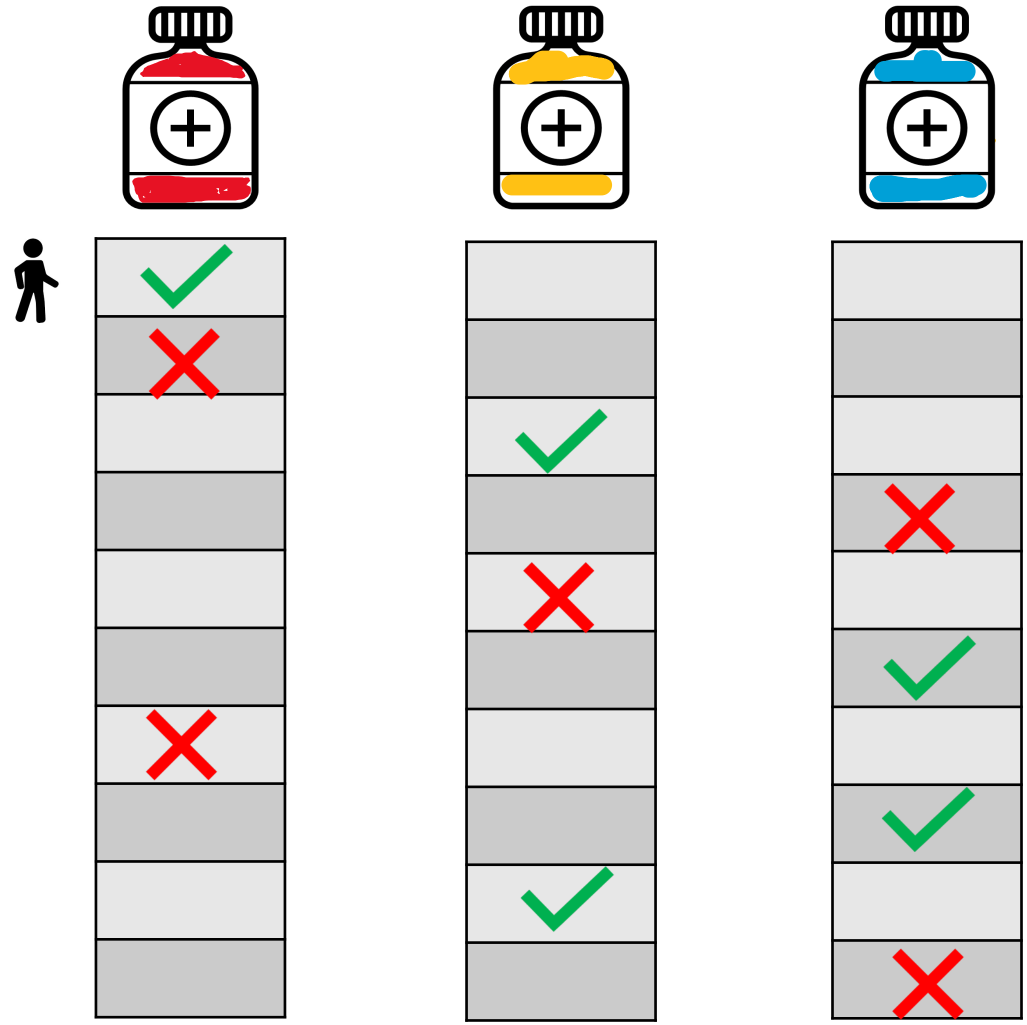
$Q_1(R)=0.5$ $Q_1(Y)=0.5$ $Q_1(B)=0$

Example

A reward of 1 if treatment succeeds, 0 otherwise

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i}{t-1}$$

Actions are selected randomly



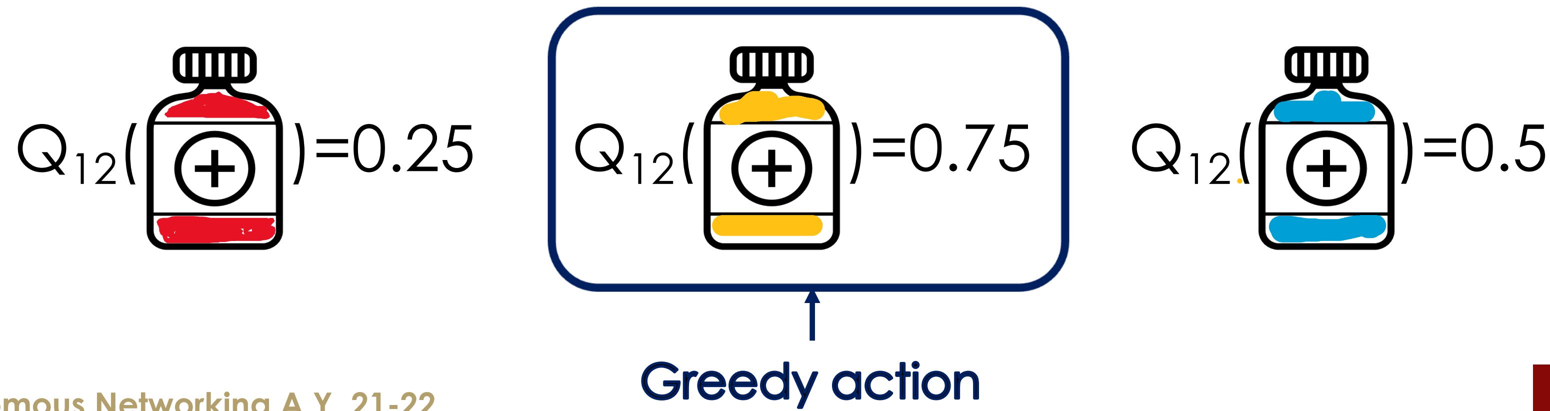
$Q_1(R)=0.33$ $Q_1(Y)=0.66$ $Q_1(B)=0.5$

How to select an action

- Random (previous example)
- Greedy
- ε-greedy

Greedy action

- In reality, our doctors would not randomly assign treatments to their patients.
- Instead, they would probably assign the treatment that they currently think is the best (trying to get the most reward he can right now.)
- We call this method of choosing actions **greedy**.
- The greedy action is the action that currently has the largest estimated value.



Greedy action

- Selecting the greedy action means the agent is exploiting its current knowledge.
- We can compute the greedy action by taking the argmax of our estimated values

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

- Greedy action selection always exploits current knowledge to maximize immediate reward
- It spends no time at all sampling apparently inferior actions to see if they might be better
  - Get stuck on suboptimal actions (no exploration)

ε-greedy

- How do we choose when to exploit and when to explore?
- Behave greedly most of time, but every once in a while (**with small probability ε**), instead **select randomly from among all actions** with equal probability, independently of the action value estimates

$$A_t \leftarrow \begin{cases} \underset{a}{\operatorname{argmax}}\ Q_t(a) & \textbf{with probability } 1-\epsilon \\ a \sim Uniform(\{a_1 \ldots a_k\}) & \textbf{with probability } \epsilon \end{cases}$$

- As the number of step increases $Q_t(a)$ converges to $q_*(a)$

Exercise 1

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

Exercise 2

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

$A1 = 1$ $R1 = 1$

$A2 = 2$ $R2 = 1$

$A3 = 2$ $R3 = 2$

$A4 = 2$ $R4 = 2$

$A5 = 3$ $R5 = 0$

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?

How to estimate action-value

- Sample-average method
- Incremental

Incremental formula for action-value

- When we perform many trials we have many values to calculate average action-value
- Can we compute it incrementally?
- To simplify notation we concentrate on a single action.
- Let R_i now denote the reward received after the i^{th} selection of this action, and let Q_n denote the estimate of its action value after it has been selected n-1 times

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

Incremental formula for action-value

- Given Q_n and the n^{th} reward, R_n, the new average of all n rewards can be computed by

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

which holds even for n=1, obtaining $Q_2=R_1$

Incremental formula for action-value

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

- The general form for the incremental update rule is

$$NewEstimate \leftarrow OldEstimate \;+\; StepSize\,\Big[Target - OldEstimate\Big]$$

- [Target – OldEstimate] is an *error* in the estimate
- StepSize changes from time step to time step

Pseudocode for bandit

- The function bandit(a) is assumed to take an action and return a corresponding reward.

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

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

$R \leftarrow bandit(A)$

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

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

How to estimate action-value

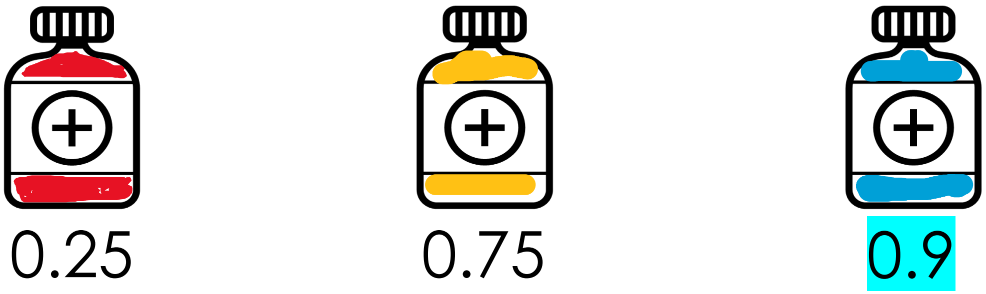
- Sample-average method
- Incremental
  - Stationary problems
  - Nonstationary problems

Nonstationary problem

- The averaging methods discussed so far are appropriate for **stationary** bandit problems, that is, for bandit problems in which the **reward probabilities do not change over time**
- There are often nonstationary problems, in which reward probabilities change over time

Example

- Doctor trials
- What if one of the treatments was more effective under certain conditions? Specifically, let's say the treatment B is more effective during the winter months.



- The distribution of rewards changes with time
- The doctor is unaware of this change but would like to adapt to it

Tracking a nonstationary problem

- One option is to use a fixed step size.
- If step-size parameter is constant then the most recent rewards affect the estimate more than older rewards.

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

- Where $\alpha \in (0,1]$

Tracking a nonstationary problem

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

How to select an action

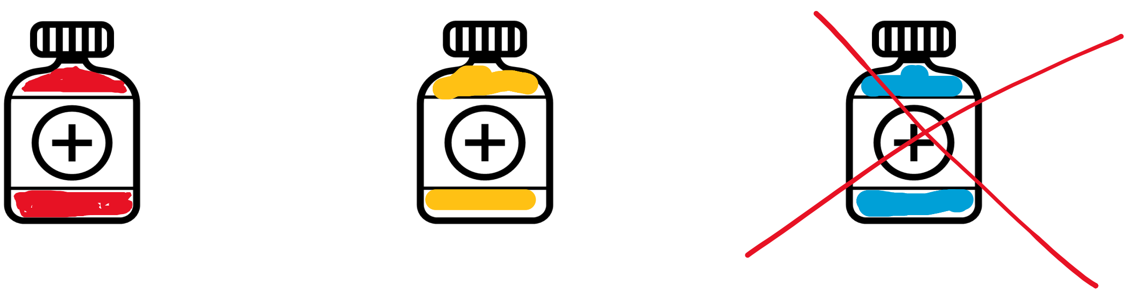
- Random
- Greedy
- ε-greedy
- Optimistic initial values

Optimistic initial values

- All methods seen so far are dependent to some extent on the initial action-value estimates, $Q_1(a)$
- **Initial action values can be used as a simple way to encourage exploration**
- What if our doctor performing medical trials was initially very optimistic about the outcome of each treatment?
- Perhaps the doctor starts with the assumption that each treatment is 100% effective, until shown otherwise.

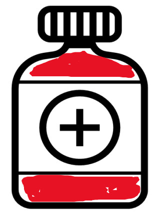
Example

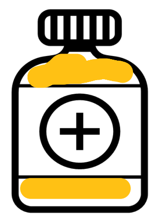
- Our doctor would begin prescribing treatments at random, until one of the treatments fails to cure a patient
- The doctor might then choose from the other two treatments at random
- Again, the doctor would continue until one of these treatments fails to work



Example

- Previously the initial estimated values were assumed to be 0, which is not necessarily optimistic
- Now, our doctor optimistically assumes that each treatment is highly effective before running the trial.
- let's make the initial value for each action 2

$Q_1($  $)=\cancel{0}2$

$Q_1($  $)=\cancel{0}2$

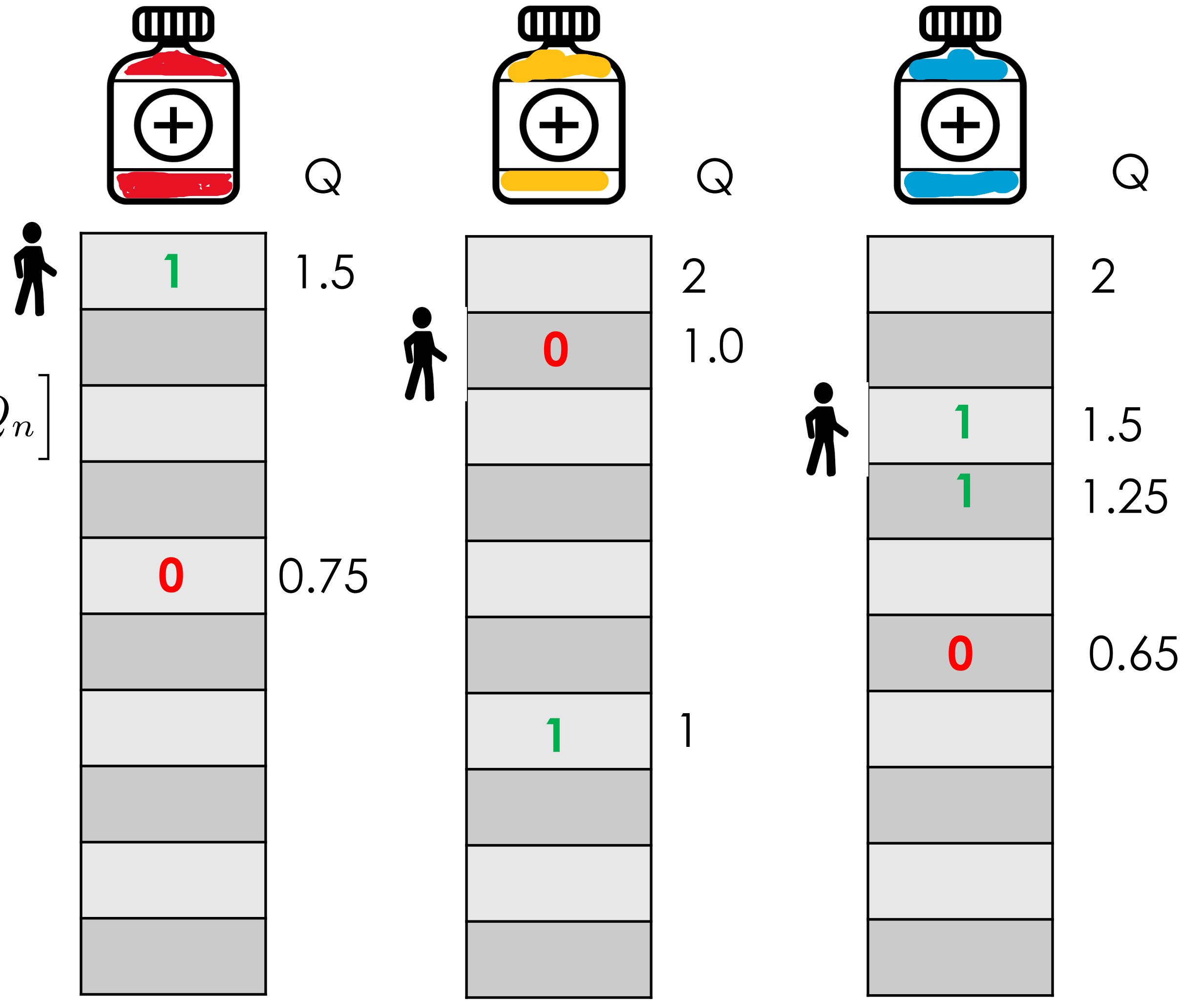
$Q_1($  $)=\cancel{0}2$

Example

- A reward of 1 if treatment succeeds, 0 otherwise

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

Let α=0.5
$Q_1(a)$=2



Optimistic initial values

- All actions are tried several times before the value estimates converge
- The system does a fair amount of exploration even if greedy actions are selected all the time

Conclusions

- Methods to evaluate action values
  - Sample average
  - Incrementally
    - Stationary problems
    - Nonstationary problems
- Strategies for action selection
  - Random
  - Greedy
  - ε-greedy
  - Optimistic initial values