



Học phần RBE3043: Các thuật toán thích nghi

**Buổi 3: Một số khái niệm cần thiết và
Bài toán lựa chọn phương án**

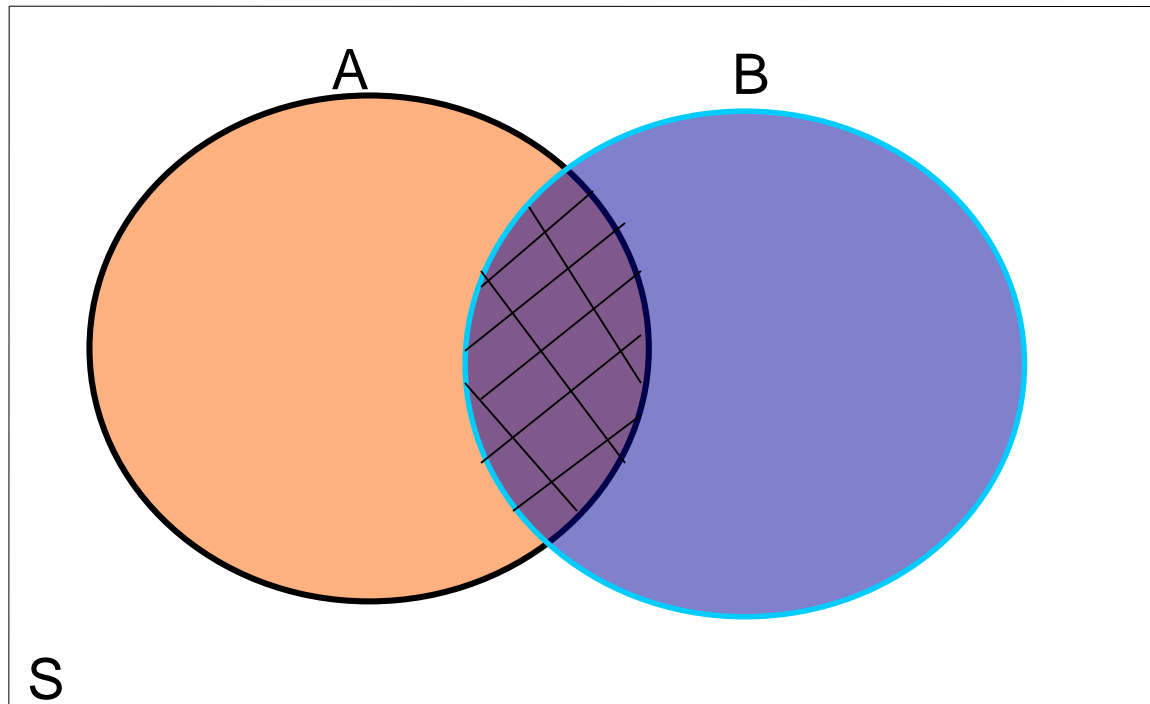
Giảng viên: TS. Nguyễn Thế Hoàng Anh

Hà Nội, ngày 20 tháng 9 năm 2023

Conditional Probability

Def. The
conditional
probability of A
given B is the
probability that
an event, A , will
occur given that
another event, B ,
has occurred

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Conditional Probability

example:

- Toss a balanced die once and record the number on the top face.
- Let E be the event that a 1 shows on the top face.
- Let F be the event that the number on the top face is odd.
 - What is $P(E)$?
 - What is the *Probability of the event E* if we are told that the number on the top face is odd, that is, we know that the event F has occurred?



Conditional Probability

- Key idea: The original sample space no longer applies.

- The new or reduced sample space is

$$S=\{1, 3, 5\}$$

- Notice that the new sample space consists only of the outcomes in F .
- $P(E \text{ occurs given that } F \text{ occurs}) = 1/3$
- Notation: $P(E|F) = 1/3$

Independent Events

If the probability of the occurrence of event A is the same regardless of whether or not an outcome B occurs, then the outcomes A and B are said to be **independent** of one another. Symbolically, if

$$P(A | B) = P(A)$$

then A and B are independent events.

Independent Events

$$P(A \cap B) = P(A | B)P(B)$$

then we can also state the following relationship for independent events:

$$P(A \cap B) = P(A)P(B)$$

if and only if

A and B are independent events.

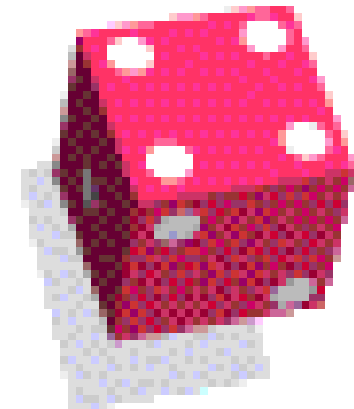
Example

- A coin is tossed and a single 6-sided die is rolled. Find the probability of getting a head on the coin and a 3 on the die.
- Probabilities:

$$P(\text{head}) = 1/2$$

$$P(3) = 1/6$$

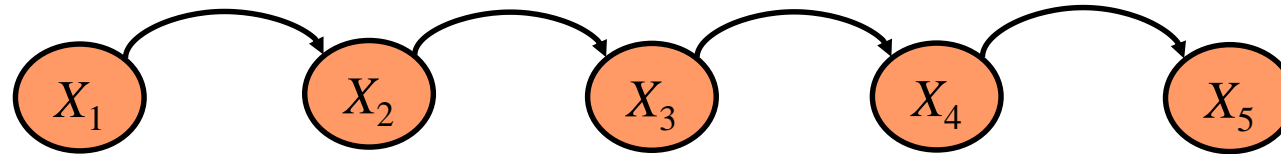
$$P(\text{head and } 3) = 1/2 * 1/6 = 1/12$$



Markov Process

- **Markov Property:** The state of the system at time $t+1$ depends only on the state of the system at time t

$$\Pr[X_{t+1} = x_{t+1} / X_1 \cdots X_t = x_1 \cdots x_t] = \Pr[X_{t+1} = x_{t+1} / X_t = x_t]$$



- **Stationary Assumption:** Transition probabilities are independent of time (t)





$$\Pr[X_{t+1} = b / X_t = a] = p_{ab}$$

Bounded memory transition model

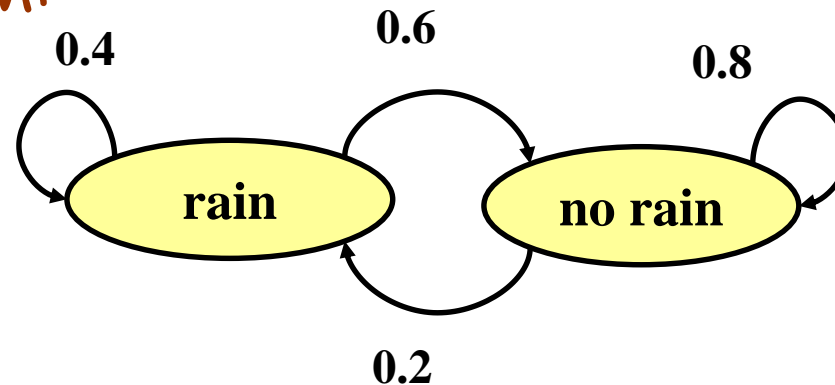
Markov Process

Simple Example

Weather:

- raining today  40% rain tomorrow
 60% no rain tomorrow
- not raining today  20% rain tomorrow
 80% no rain tomorrow

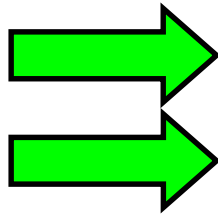
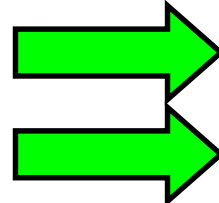
Stochastic FSM:



Markov Process

Simple Example

Weather:

- raining today  40% rain tomorrow
60% no rain tomorrow
- not raining today  20% rain tomorrow
80% no rain tomorrow

The transition matrix:

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix}$$

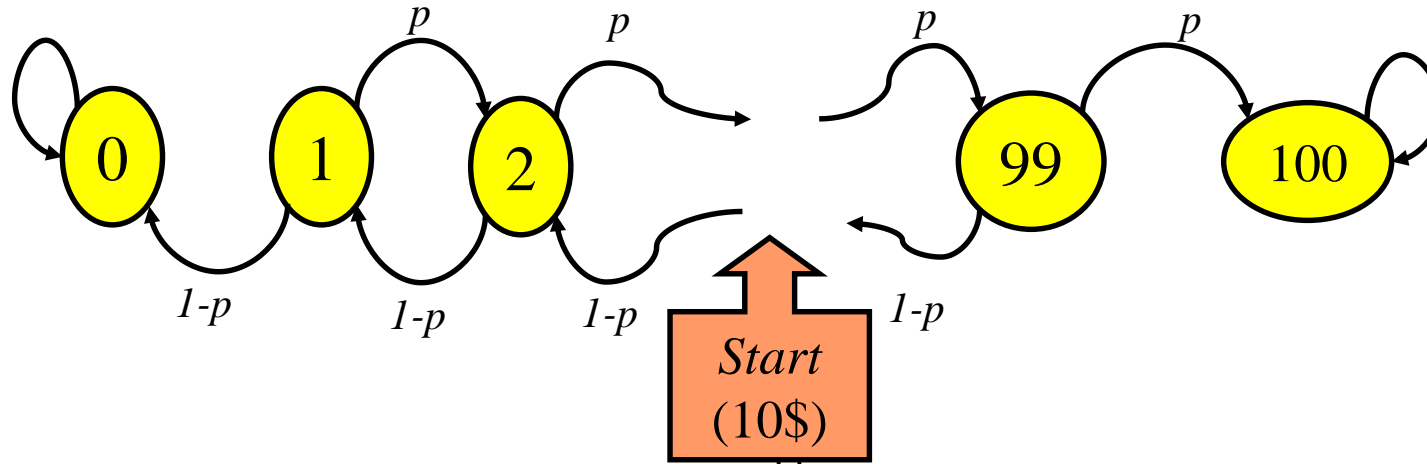
- Stochastic matrix:
Rows sum up to 1
- Double stochastic matrix:
Rows and columns sum up to 1

Markov Process

Gambler's Example

- Gambler starts with \$10
- At each play we have one of the following:
 - Gambler wins \$1 with probability p
 - Gambler loses \$1 with probability $1-p$
- Game ends when gambler goes broke, or gains a fortune of \$100

(Both 0 and 100 are absorbing states)



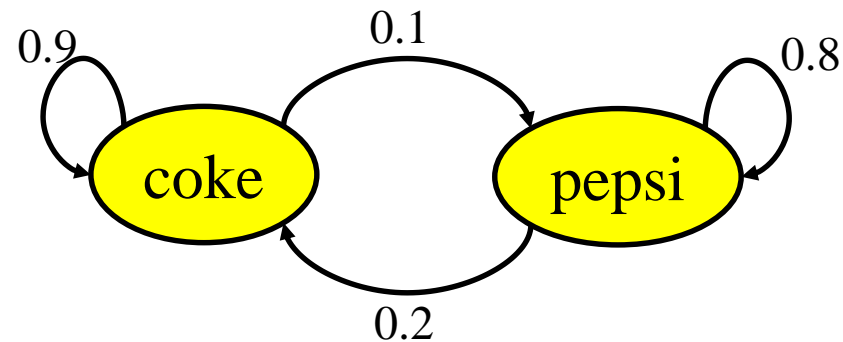
Markov Process

Coke vs. Pepsi Example

- Given that a person's last cola purchase was **Coke**, there is a **90%** chance that his next cola purchase will also be **Coke**.
- If a person's last cola purchase was **Pepsi**, there is an **80%** chance that his next cola purchase will also be **Pepsi**.

transition matrix:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$



Markov Process

Coke vs. Pepsi Example (cont)

Given that a person is currently a **Pepsi** purchaser, what is the probability that he will purchase **Coke** two purchases from now?

$$\Pr[\text{Pepsi} \rightarrow ? \rightarrow \text{Coke}] =$$

$$\Pr[\text{Pepsi} \rightarrow \text{Coke} \rightarrow \text{Coke}] + \Pr[\text{Pepsi} \rightarrow \text{Pepsi} \rightarrow \text{Coke}] =$$

$$0.2 * 0.9 + 0.8 * 0.2 = 0.34$$

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}$$

$\text{Pepsi} \rightarrow ? \quad ? \rightarrow \text{Coke}$

Markov Process

Coke vs. Pepsi Example (cont)

Given that a person is currently a **Coke** purchaser, what is the probability that he will purchase **Pepsi** **three** purchases from now?

$$P^3 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix} = \begin{bmatrix} 0.781 & 0.219 \\ 0.438 & 0.562 \end{bmatrix}$$

Markov Process

Coke vs. Pepsi Example (cont)

- Assume each person makes one cola purchase per week
- Suppose 60% of all people now drink Coke, and 40% drink Pepsi
- What fraction of people will be drinking Coke three weeks from now?

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \quad P^3 = \begin{bmatrix} 0.781 & 0.219 \\ 0.438 & 0.562 \end{bmatrix}$$

$$\Pr[X_3 = \text{Coke}] = 0.6 * 0.781 + 0.4 * 0.438 = 0.6438$$

Q_i - the distribution in week i

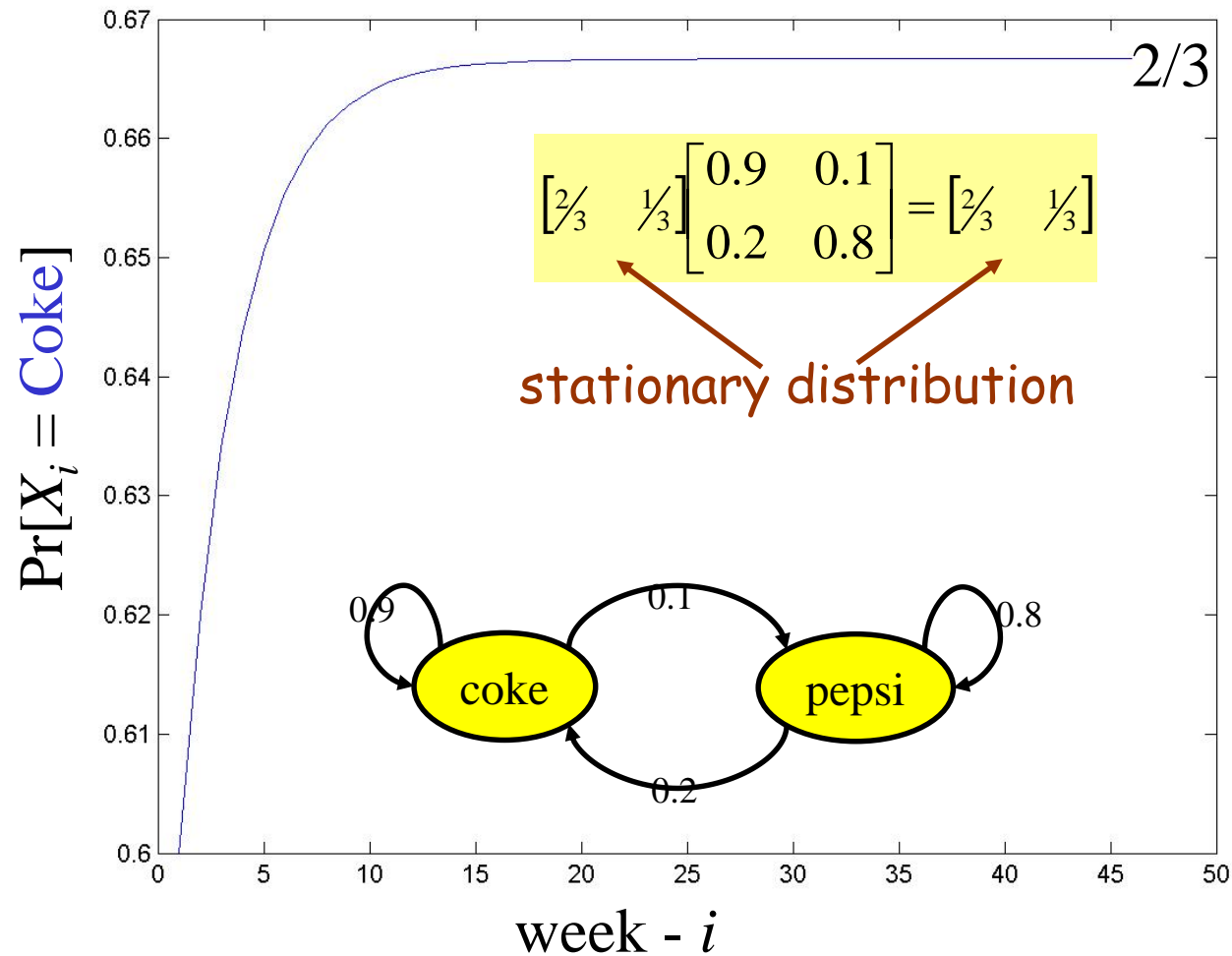
$Q_0 = (0.6, 0.4)$ - initial distribution

$$Q_3 = Q_0 * P^3 = (0.6438, 0.3562)$$

Markov Process

Coke vs. Pepsi Example (cont)

Simulation:



K-armed bandit 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. Your objective is to maximize the expected total reward over some time period, for example, over 1000 action selections, or time steps → k -armed bandit problem.

One arm bandits are the name for the original, mechanical slot machines.

Examples:

- Doctor choosing an experimental description
- Choosing a slot machine levers to pull



Greedy selection rule

$$A_t \doteq \operatorname{argmax}_a Q_t(a)$$

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

ϵ -Greedy action selection

- Discard $\operatorname{argmax}_a Q_t(a)$ with probability ϵ and re-sample another action with uniform distribution
- As the number of steps increases, every action will be sampled and infinite number of times, ensuring that $Q_t(a)$ converges to $q_*(a)$

The 10-armed testbed

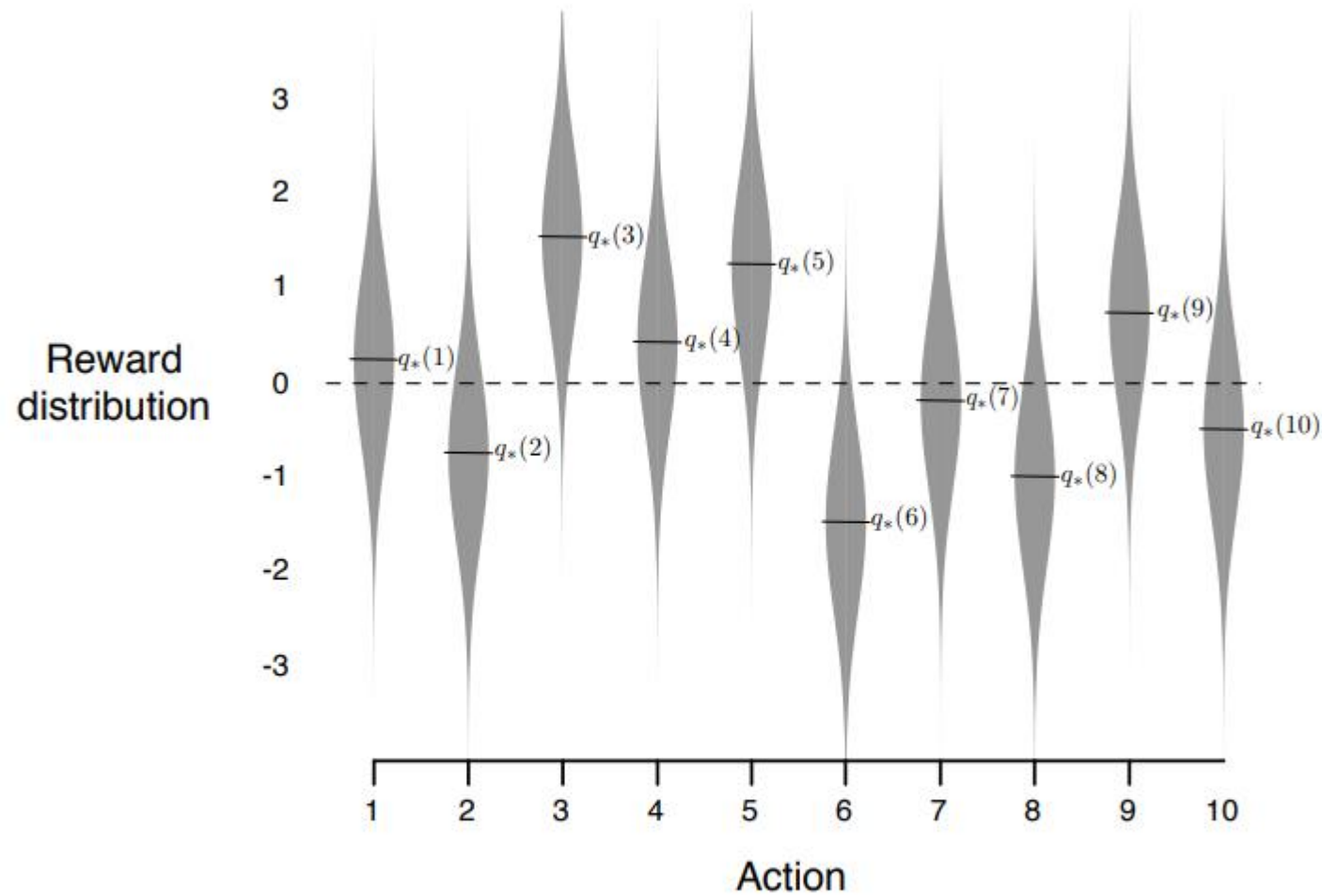
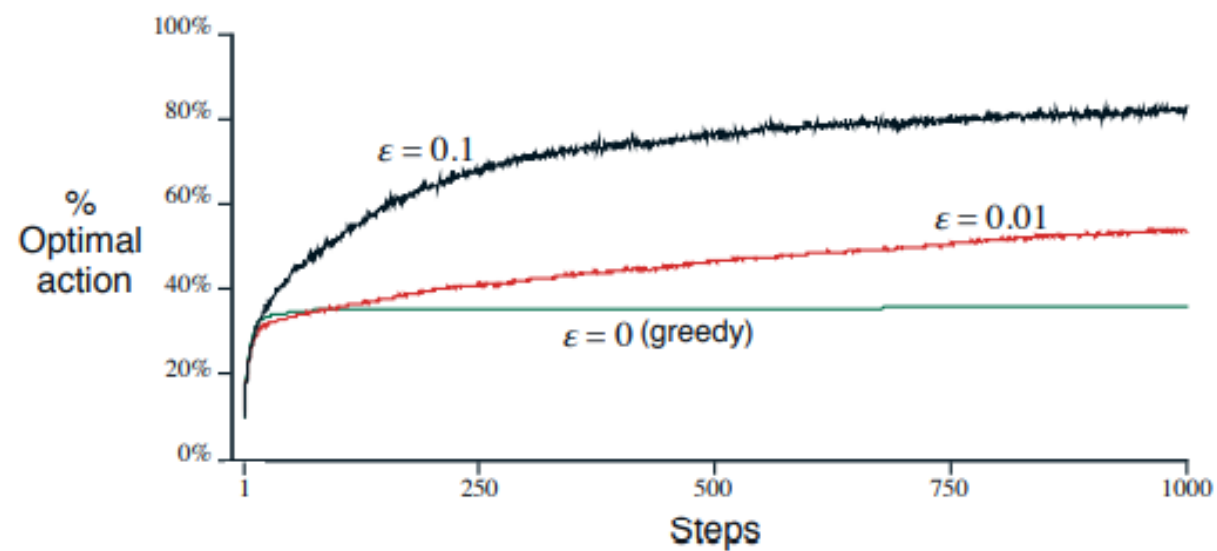
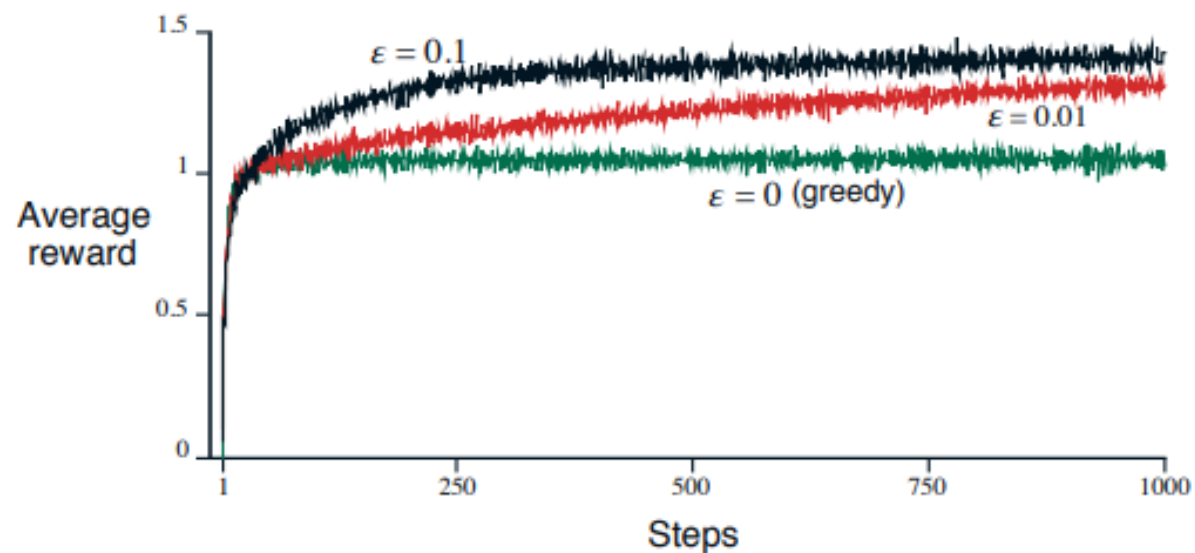


Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution, as suggested by these gray distributions.

Experiment results



A simple bandit algorithm

Initialize, for $a = 1$ to k :

$$Q(a) \leftarrow 0$$

$$N(a) \leftarrow 0$$

Loop forever:

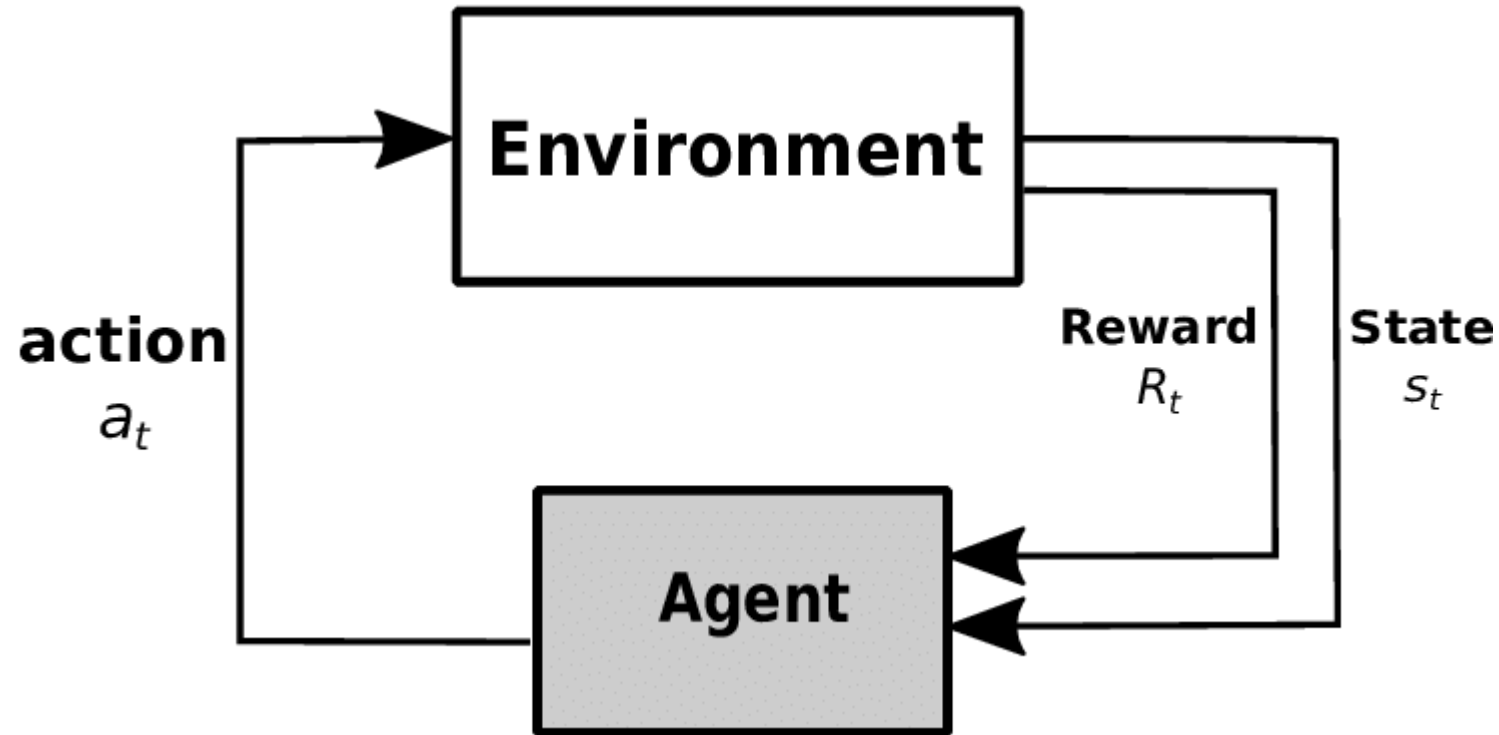
$$A \leftarrow \begin{cases} \arg \max_a Q(a) & \text{with probability } 1 - \varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases} \quad (\text{breaking ties randomly})$$

$$R \leftarrow \text{bandit}(A)$$

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

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

Reinforcement learning definition



Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal.

Reinforcement learning – Unsupervised learning – Supervised learning

- Supervised learning methods: Artificial Neural Network (ANN), Deep belief net, Support Vector Machine (SVM), Convolutional Neural network (CNN), Naives Bayes, K-nearest neighbor
- Unsupervised learning methods: Independent component analysis (ICA), Principal component analysis (PCA), K-mean clustering, Gaussian mixture model (GMM)

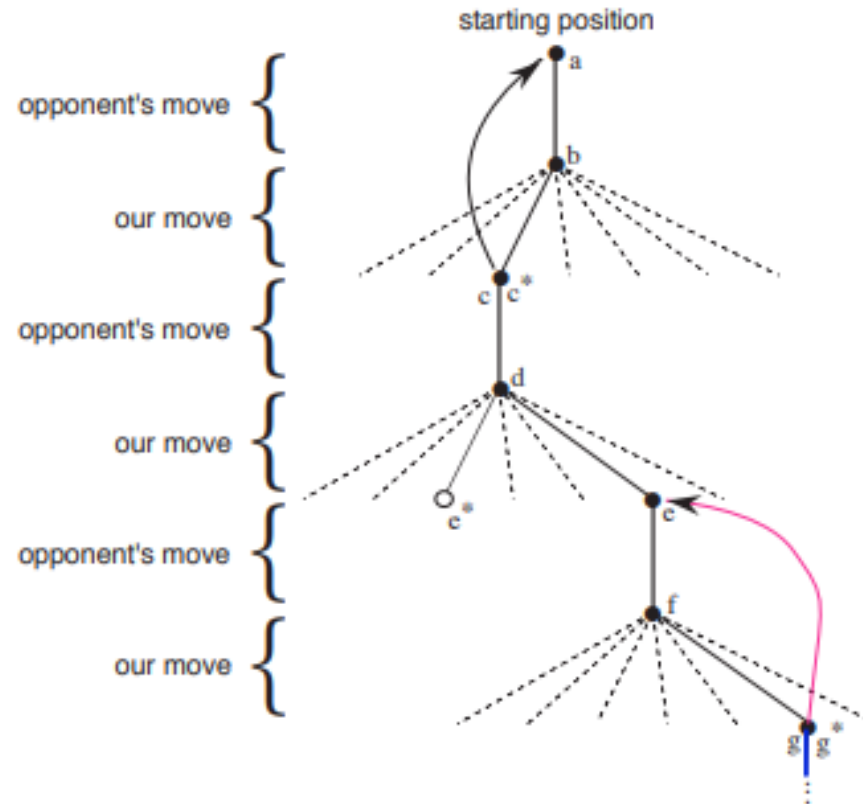
Reinforcement learning – Unsupervised learning – Supervised learning

Reinforcement learning is different from *supervised learning*, the kind of learning studied in most current research in the field of machine learning. Supervised learning is learning from a training set of labeled examples provided by a knowledgeable external supervisor. Each example is a description of a situation together with a specification—the label—of the correct action the system should take to that situation, which is often to identify a category to which the situation belongs. The object of this kind of learning is for the system to extrapolate, or generalize, its responses so that it acts correctly in situations not present in the training set. This is an important kind of learning, but alone it is not adequate for learning from interaction. In interactive problems it is often impractical to obtain examples of desired behavior that are both correct and representative of all the situations in which the agent has to act. In uncharted territory—where one would expect learning to be most beneficial—an agent must be able to learn from its own experience.

Reinforcement learning is also different from what machine learning researchers call *unsupervised learning*, which is typically about finding structure hidden in collections of unlabeled data. The terms supervised learning and unsupervised learning would seem to exhaustively classify machine learning paradigms, but they do not. Although one might be tempted to think of reinforcement learning as a kind of unsupervised learning because it does not rely on examples of correct behavior, reinforcement learning is trying to maximize a reward signal instead of trying to find hidden structure. Uncovering structure in an agent's experience can certainly be useful in reinforcement learning, but by itself does not address the reinforcement learning problem of maximizing a reward signal. We therefore consider reinforcement learning to be a third machine learning paradigm, alongside supervised learning and unsupervised learning and perhaps other paradigms as well.

An example: Tic-tac-toe

X	O	O
O	X	X
		X



A sequence of tic-tac-toe moves

- State's value: the whole table is the learned value function. State A has higher value than state B if the current estimate of prob. of a win for A is higher than for B.
- Greedily: selecting the move that leads to the state with greatest value or highest estimated prob. of winning.
- Exploratory: randomly moves

Formula for a temporal different learning method

The update to the estimated value of s followed the rule

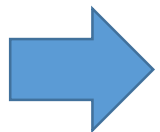
$$V(s) \leftarrow V(s) + \alpha [V(s') - V(s)],$$

s denotes the state before the greedy move

s' denote the state after the move

$V(s)$ estimated value of s

α step-size parameter or learning rate



The update is a temporal-different learning method

Nhớ giữ sức khỏe! Take care!

Thank you!
Q&A