

MDPs, Returns, Value functions, Q-function

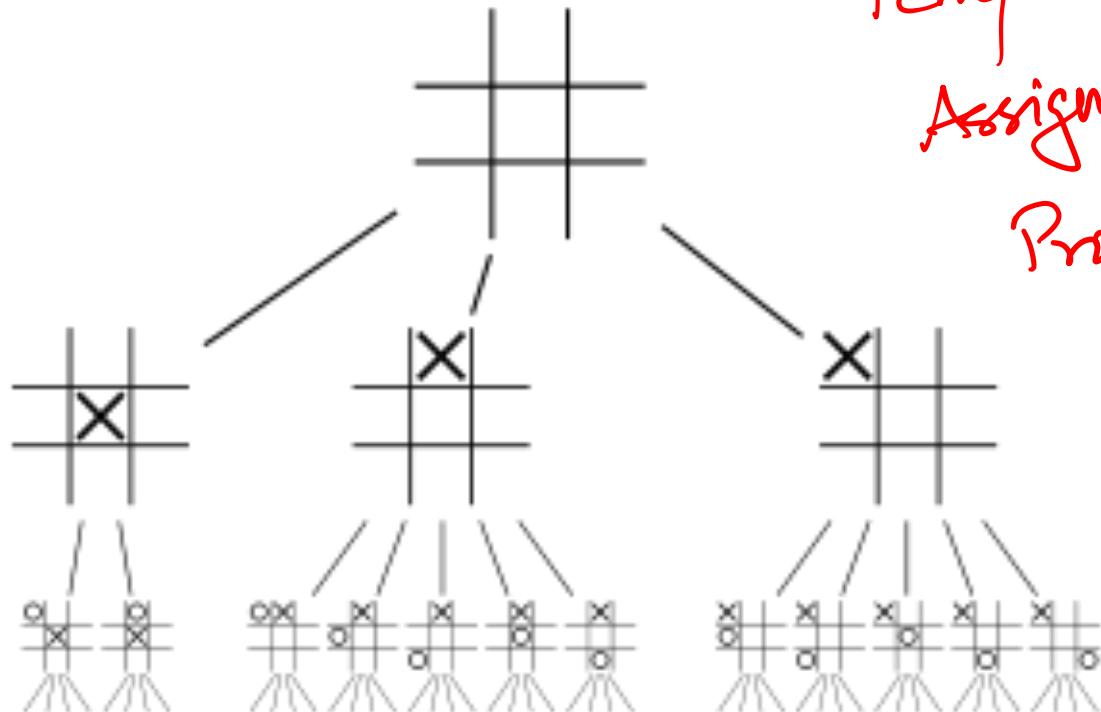
B. Ravindran

Immediate Reinforcement

- The payoff accrues immediately after an action is chosen
- One key question - the dilemma between exploration and exploitation
- *Bandit problems* encapsulate ‘Explore vs Exploit’

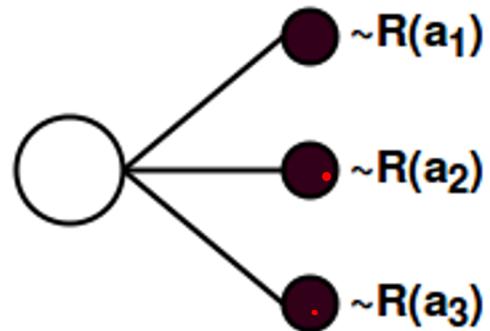


What about Tic-Tac-Toe?



Temporal Credit
Assignment
Problem-

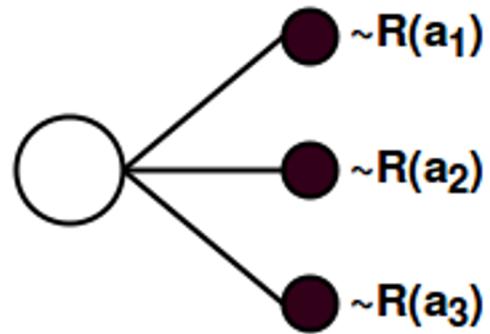
Full RL Problem



Bandit Problem

Monte Carlo

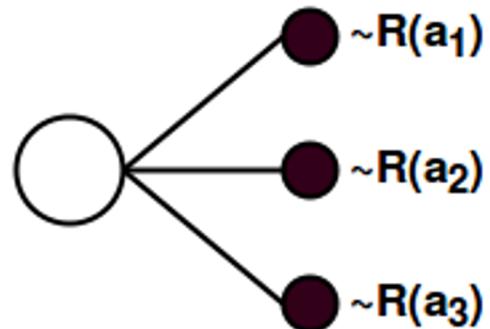
Full RL Problem



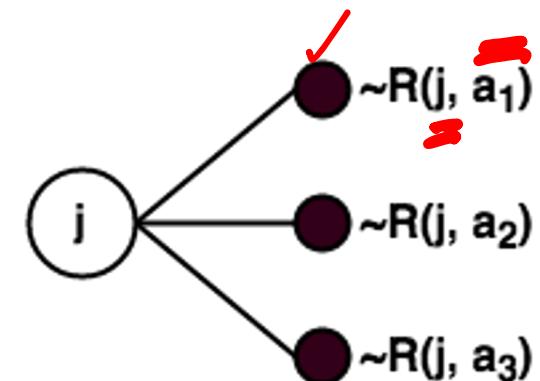
Bandit Problem

$$Q(a_i^*) = \max_i \{Q(a_i)\}$$

Full RL Problem



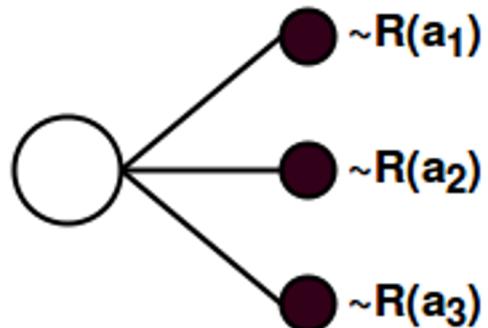
Bandit Problem



Contextual Bandit Problem

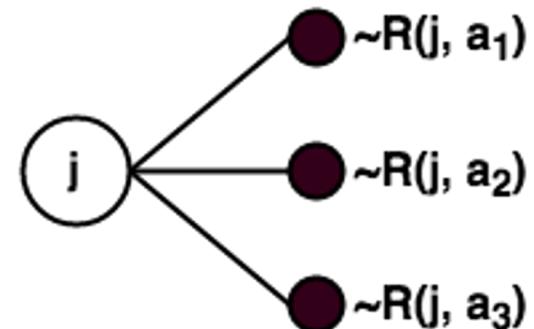
$$Q(a_i^*) = \max_i \{Q(a_i)\}$$

Full RL Problem



Bandit Problem

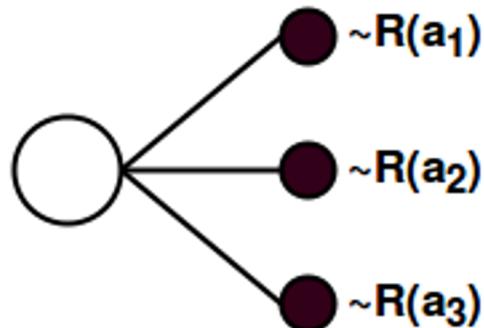
$$Q(a_i^*) = \max_i \{Q(a_i)\}$$



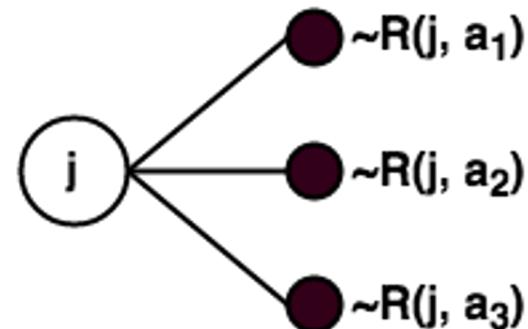
Contextual Bandit Problem

$$Q(\underline{j}, a_i^*) = \max_i \{Q(j, a_i)\}$$

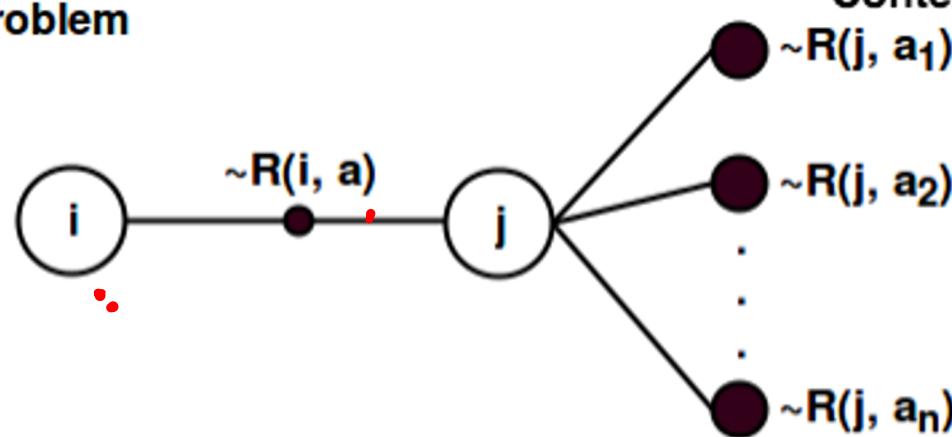
Full RL Problem



Bandit Problem

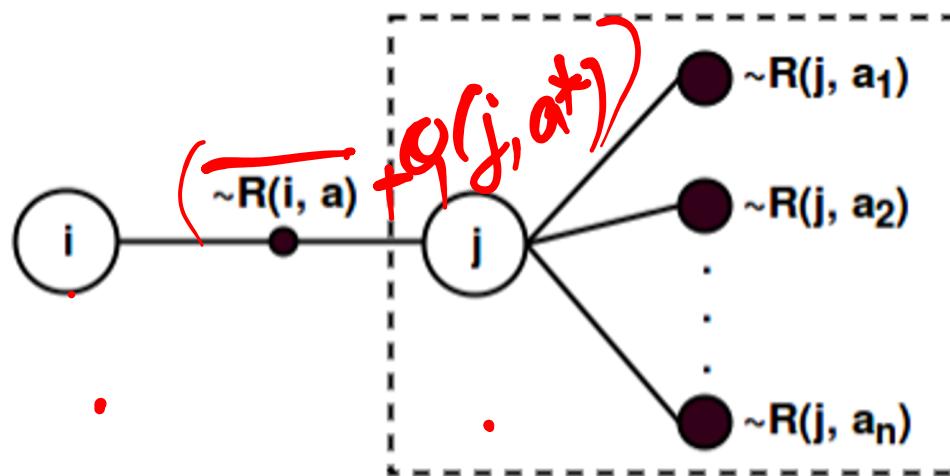
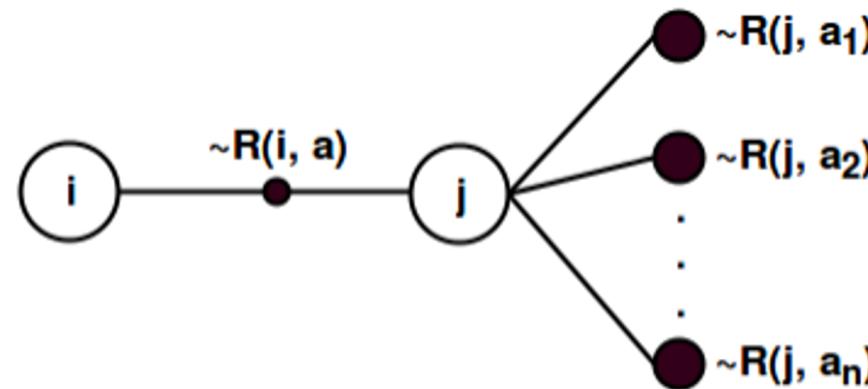


Contextual Bandit Problem



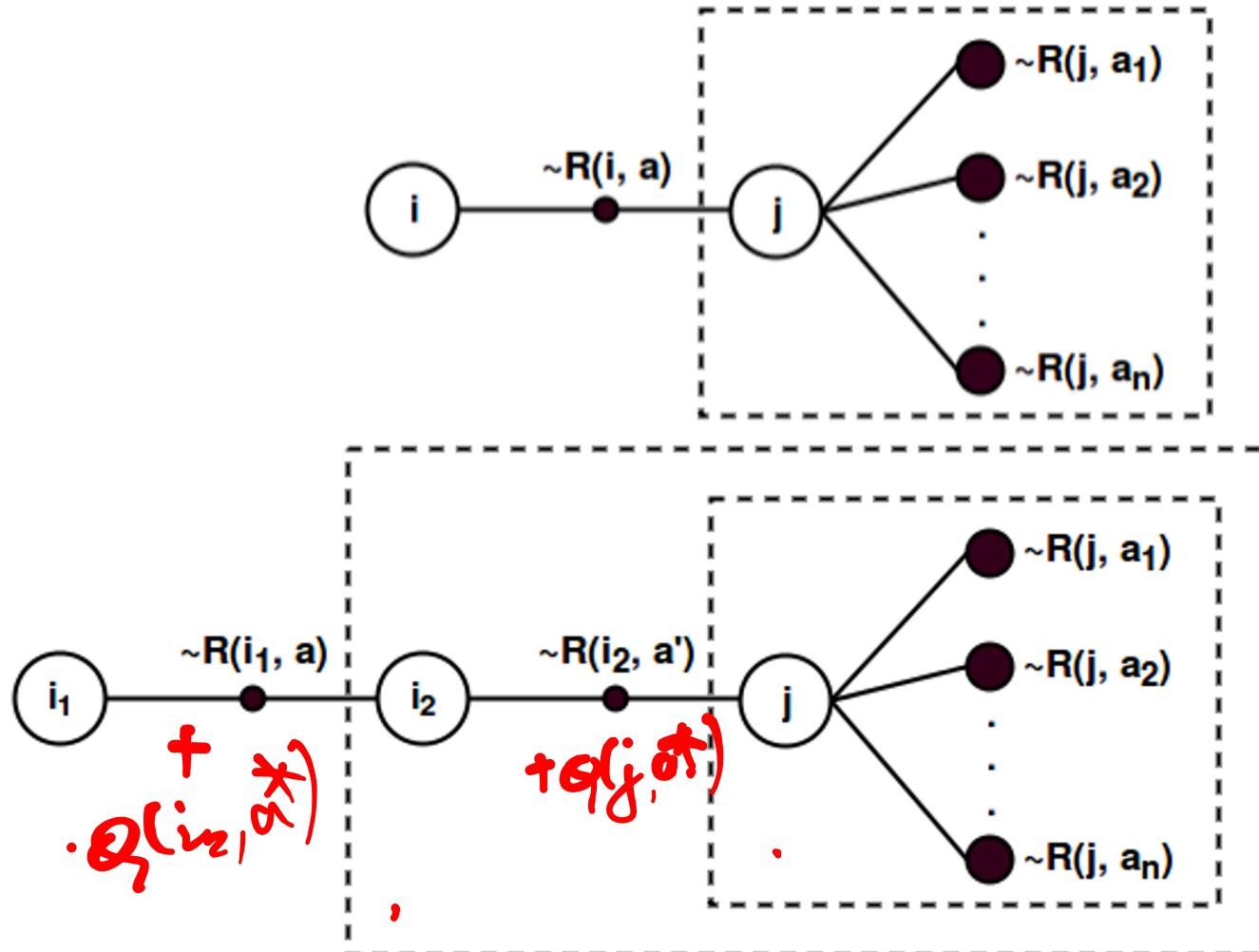
Full Reinforcement Learning Problem

Full RL Problem

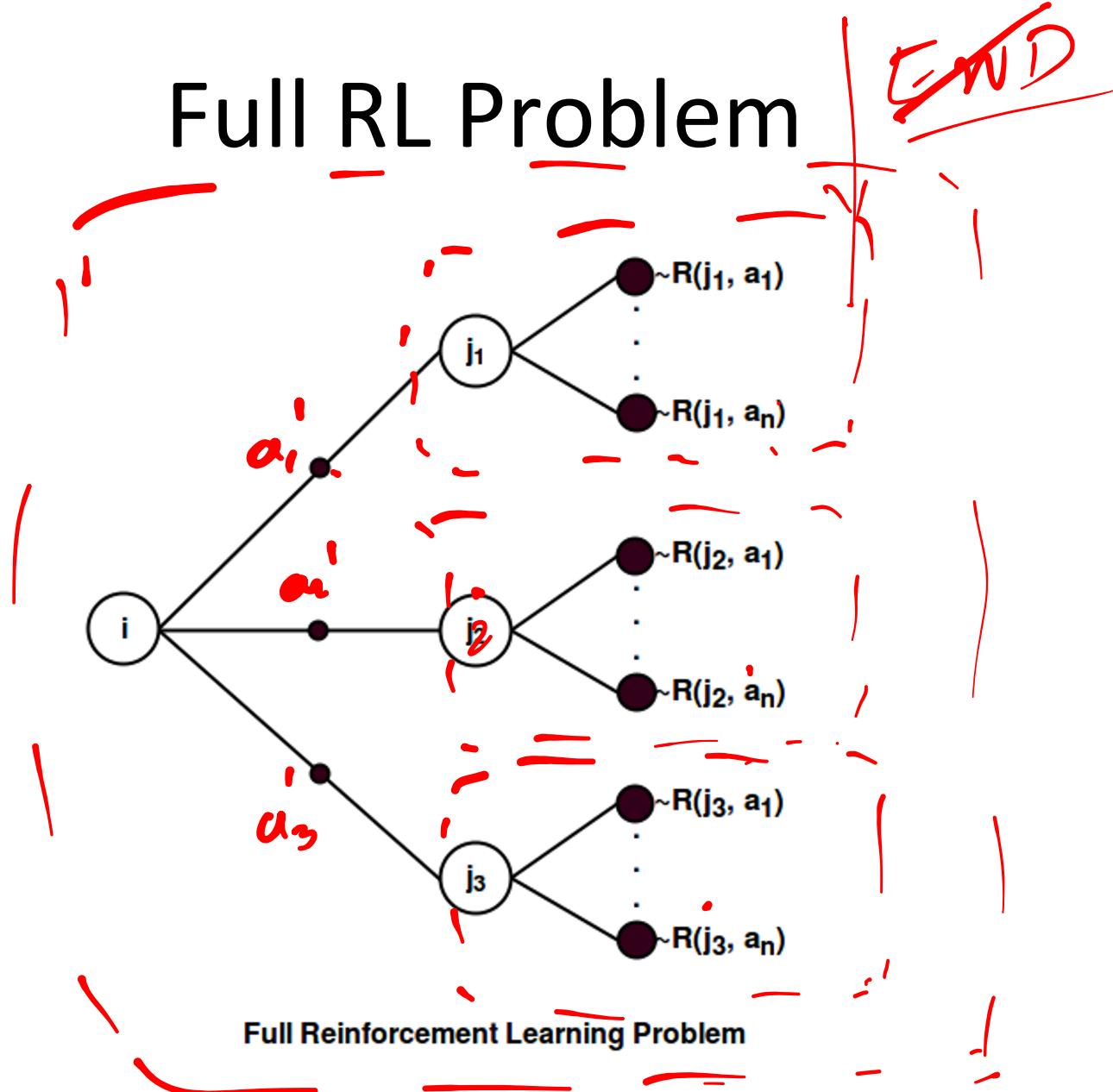


Full Reinforcement Learning Problem

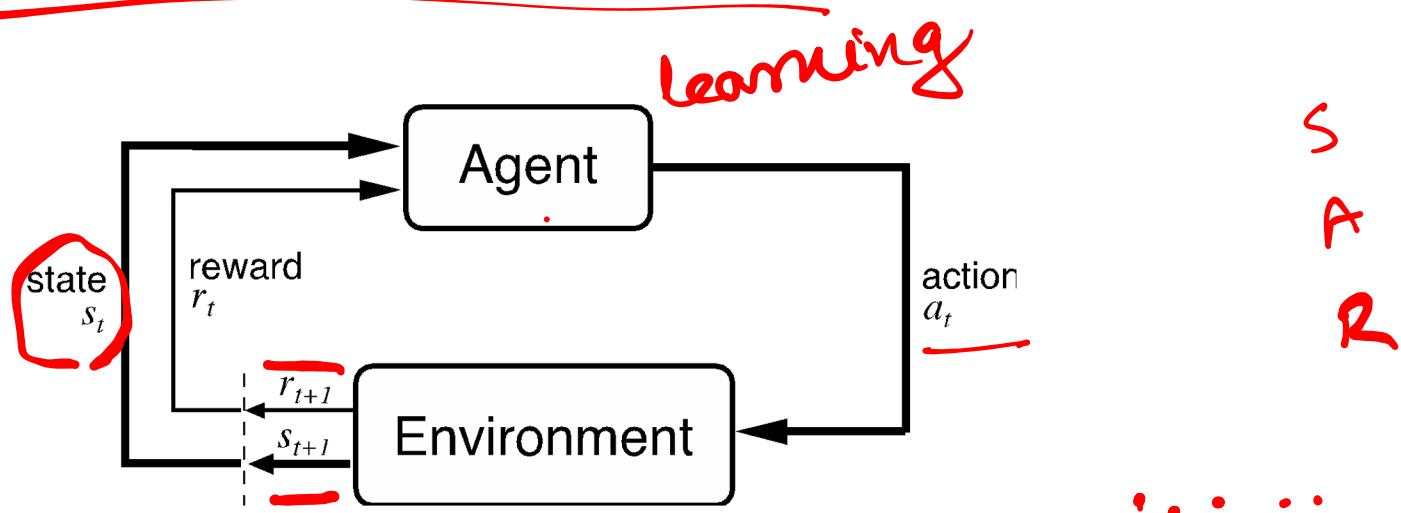
Full RL Problem



Full RL Problem



The Agent-Environment Interface



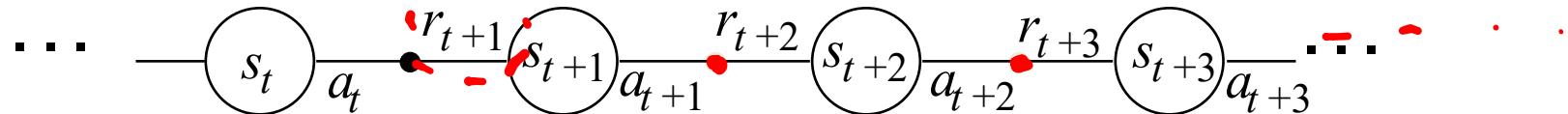
Agent and environment interact at discrete time steps: $t = 0, 1, 2, \dots$

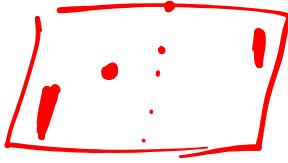
Agent observes state at step t : $s_t \in S$

produces action at step t : $a_t \in A(s_t)$

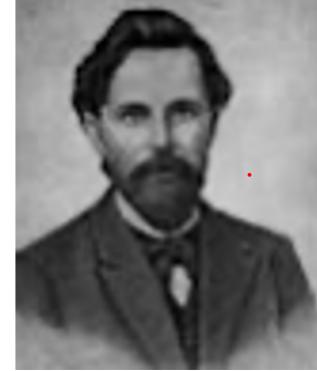
gets resulting reward: $r_{t+1} \in \mathcal{R}$

and resulting next state: s_{t+1}





The Markov Property



- “the state” at step t , means whatever information is available to the agent at step t about its environment.
- The state can include immediate “sensations”, highly processed sensations, and structures built up over time from sequences of sensations.
- Ideally, a state should summarize past sensations so as to retain all “essential” information, i.e., it should have the

(Markov Property):

$$\Pr \left\{ s_{t+1} = s', r_{t+1} = r \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, \dots, r_1, s_0, a_0 \right\} = \Pr \left\{ s_{t+1} = s', r_{t+1} = r \mid s_t, a_t \right\}$$

for all s' , r , and histories $s_t, a_t, r_t, s_{t-1}, a_{t-1}, \dots, r_1, s_0, a_0$.

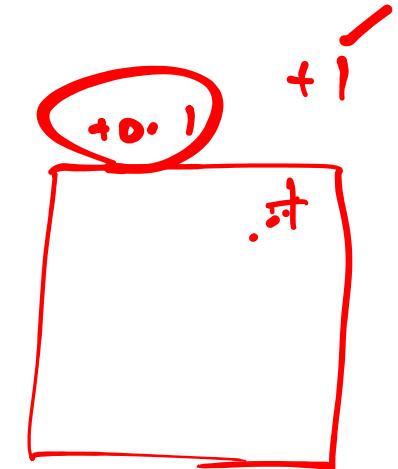
Markov Decision Processes

- MDP, M , is the tuple: $M = \langle S, A, p, r \rangle$
 - S : set of states.
 - A : set of actions.
 - $p : S \times A \times S \rightarrow [0, 1]$: probability of transition.
 - $r : S \times A \times S \rightarrow \mathbb{R}$: expected reward.
 - Policy: $\pi : S \times A \rightarrow [0, 1]$ (can be deterministic)
 - Maximize total expected reward
 - Learn an optimal policy π^*
- $p(s_{t+1}, a_{t+1} | s_t, a_t)$
- $r = E[r_{t+1}]$
- $\pi(s) = a$
- $\pi(a|s)$
- Stochastic
- $\cancel{\pi(s, a | s)}$

Designing an RL solution

- States

- Enough information to take decisions
 - Raw inputs often not sufficient



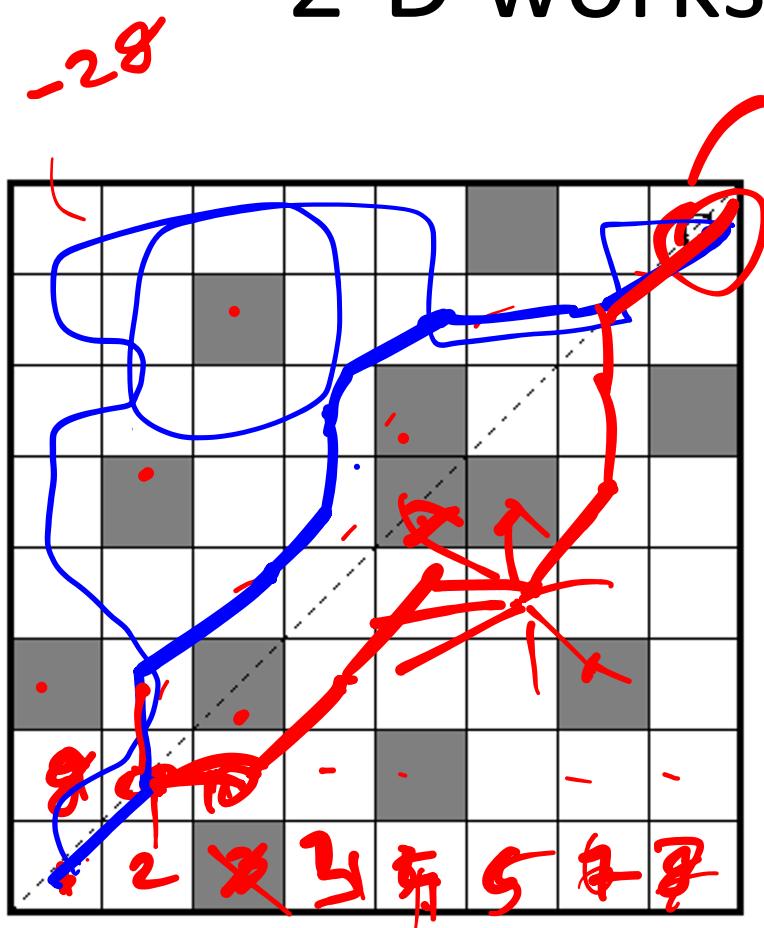
- Actions

- The control variables
 - Discrete – items to recommend, moves in a game
 - Continuous – torque to a motor, rate of mixing

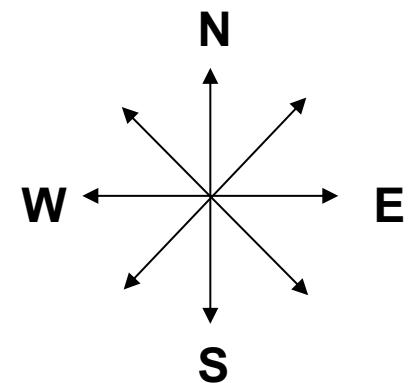
- Rewards

- Define the goal of the problem

Example 2-D workspace



$$M = \langle S, A, p, r, \cdot \rangle$$



Robot Control

- **Input** consists of the reading of the sonars, the bump sensors, the camera, the arm position, and wheel encoder
 - State** is typically a short history of the sensor readings
- **Actions** are the torques to the motors
- **Positive rewards** on achieving the goal; **Negative rewards** for bumping into obstacles



MDP Example

The product that a computer manufacturing company sells can either be hot-selling, or selling poorly. If the product is hot-selling, the company can advertise its product to get an immediate reward of 4 units. On doing so, with probability 0.8, the product continues to sell well. However, with probability 0.2, the product starts selling poorly. The company can alternately decide not to advertise a hot-selling product yielding an immediate reward of 6 units with a 0.5 probability that the product continues to remain hot-selling.

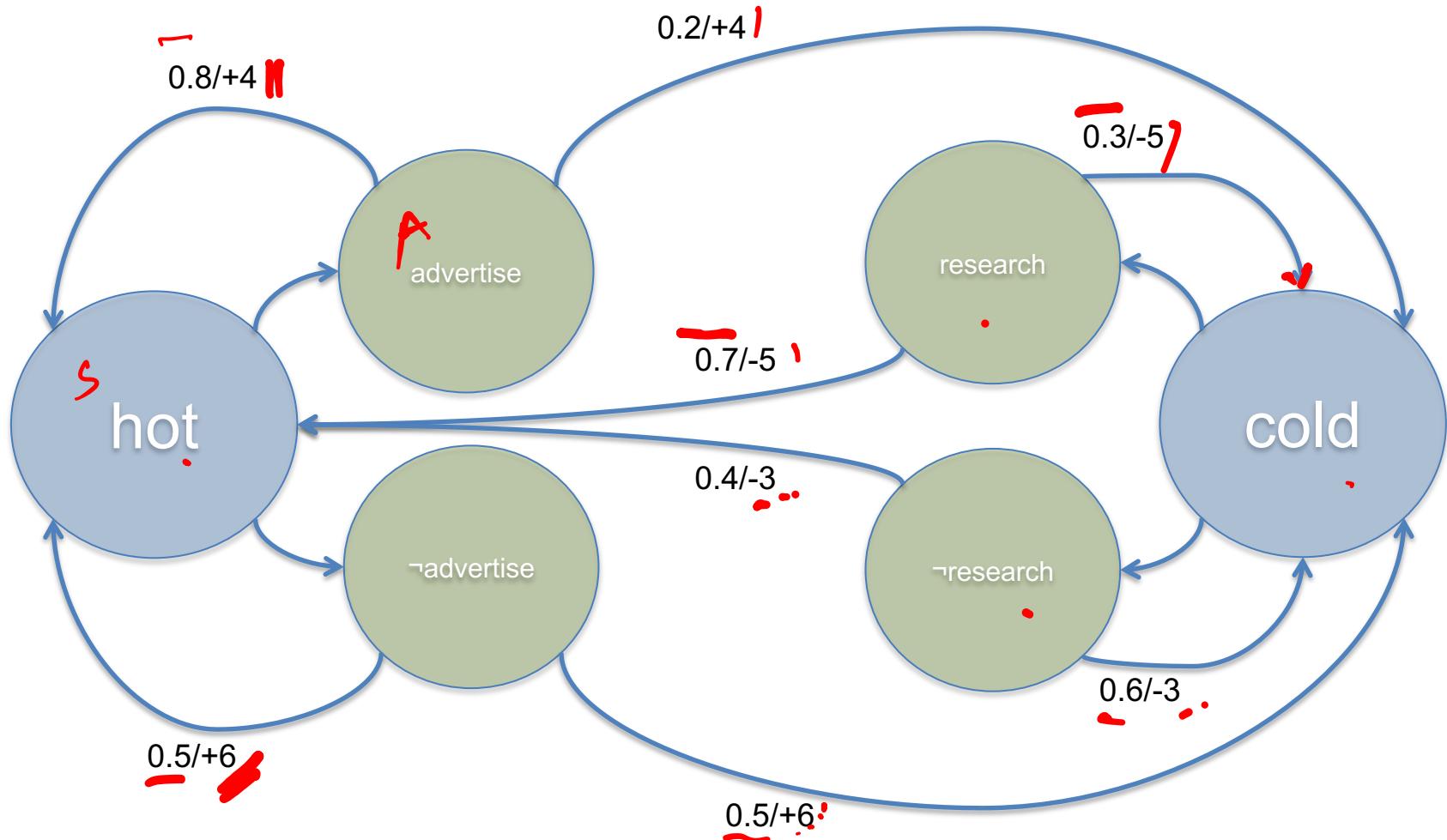
In case the product is selling poorly, the company has the option to perform research to improve its product. This results in an immediate reward of -5 units and a 0.7 probability of the research leading to the product becoming a hot-selling one. Finally, if the company decides not to perform research for a poorly selling product, it receives an immediate reward of -3, with the product continuing to sell poorly with a probability of 0.6.

Formulate this problem as a Markov Decision Process.

Formulating MDP

- What are the choices (actions) that can be made in this problem?
 - Actions: $A = \{\text{advertise}, \neg\text{advertise}, \text{research}, \neg\text{research}\}$
- What are the states in which these choices can be made?
 - advertise or \neg advertise for hot-selling product
 - research or \neg research for poorly-selling product
 - States: $S = \{\text{hot}, \text{cold}\}$
- Applicable actions in each state:
 - $A(\text{hot}) = \{\text{advertise}, \neg\text{advertise}\}$
 - $A(\text{cold}) = \{\text{research}, \neg\text{research}\}$
- Finally, consider the transition probabilities and rewards

Formulating MDP (Contd.)



The Agent Learns a Policy

T₁

Policy at step t , $\boxed{\pi_t :}$

a mapping from states to action probabilities

$\pi_t(s, a) =$ probability that $a_t = a$ when $s_t = s$

- Reinforcement learning methods specify how the agent changes its policy as a result of experience.
- Roughly, the agent's goal is to get as much reward as it can over the long run.

Returns

Suppose the sequence of rewards after step t is :

$$r_{t+1}, r_{t+2}, r_{t+3}, \dots$$

What do we want to maximize?

We want to maximize the **return**, G_t , for each step t .

Episodic tasks: interaction breaks naturally into episodes, e.g., plays of a game, trips through a maze.

$$G_t = \overbrace{r_{t+1} + r_{t+2} + \dots + r_T}^{\text{rewards}} + \gamma_{T+1} \mathbb{E} [c_{t+1}] \dots$$

where T is a final time step at which a **terminal state** is reached, ending an episode.

Returns for Continuing Tasks

Continuing tasks: interaction does not have natural episodes.

Discounted return:

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1},$$

where $\gamma, 0 \leq \gamma \leq 1$, is the **discount rate**.

$$0 \leq \gamma < 1$$

shortsighted $0 \leftarrow \gamma \rightarrow 1$ farsighted

In general,

we want to maximize the **expected return**, $E\{G_t\}$, for each step t .