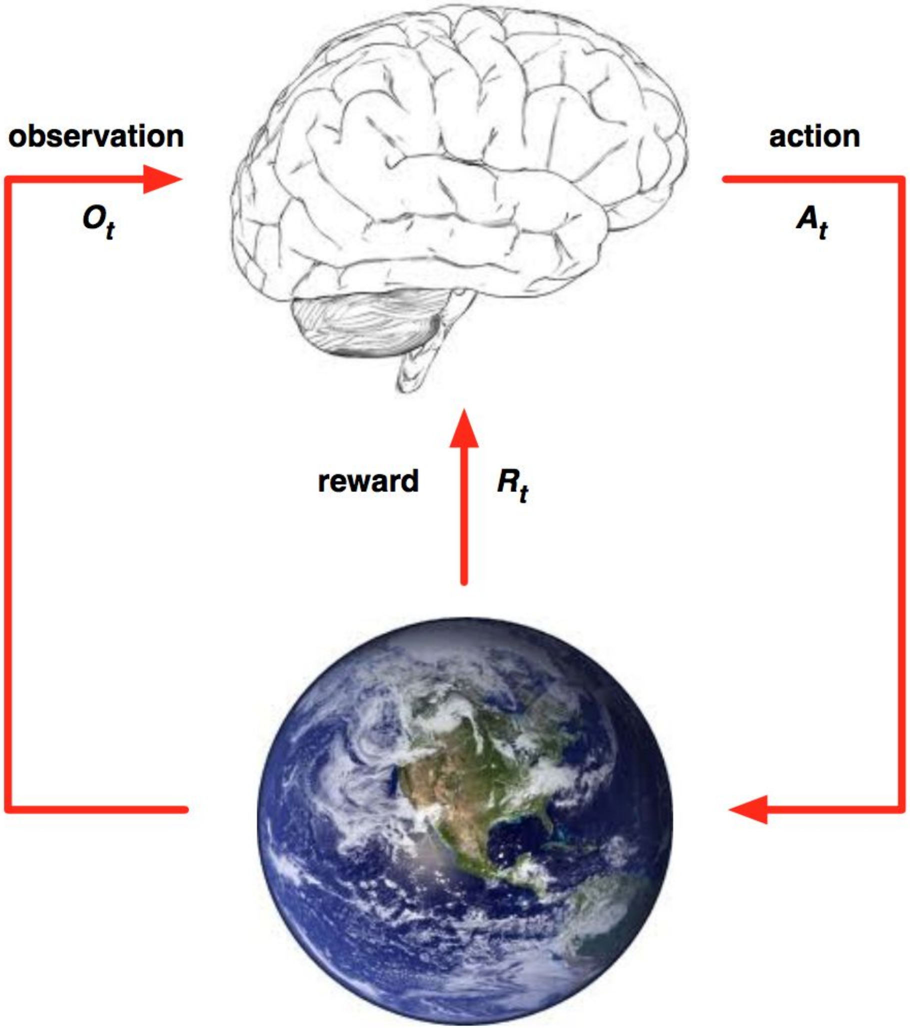
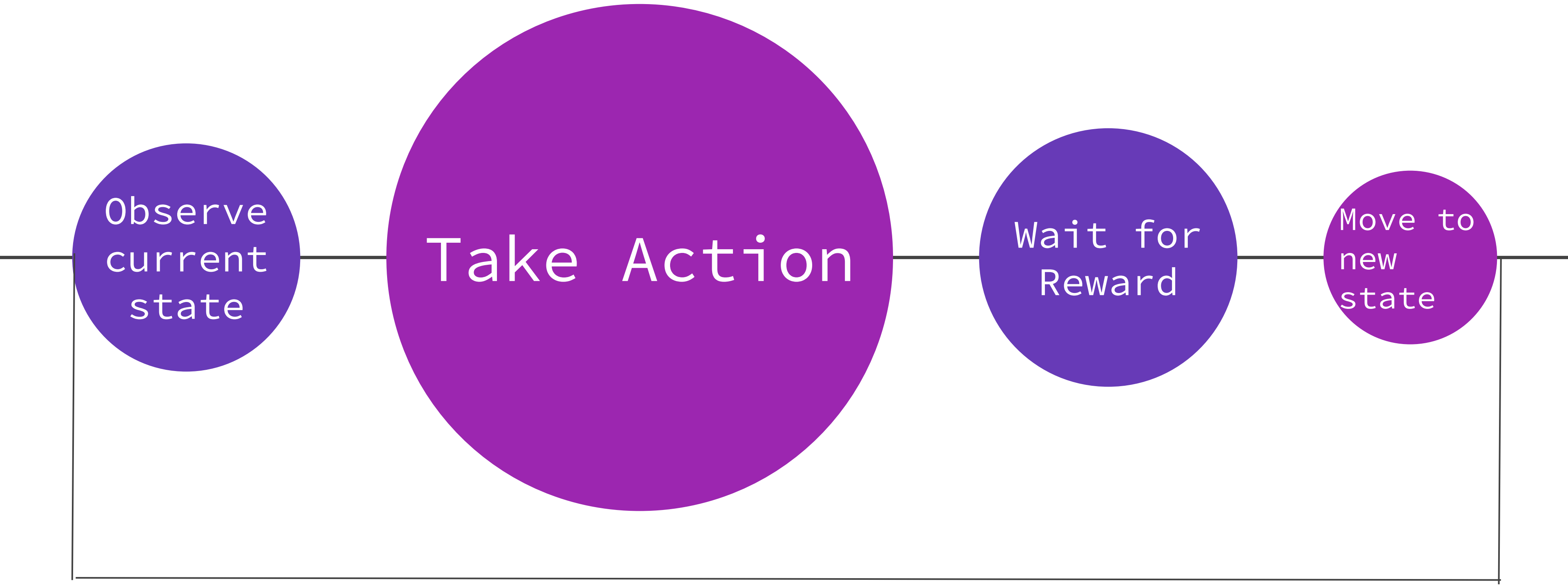
Session 2

Reinforcement Learning


observation
O_t
action
A_t
reward
R_t

Chain of events



Environment: You are in state 65. You have 4 possible actions. (**Observe**)
Agent:I'll take action 2. (**Action**)
Environment: You received a reinforcement of 7 units. (**Reward**)

Environment: You are now in state 15. You have 2 possible actions.
Agent:I'll take action 1.
Environment: You received a reinforcement of -4 units.

Environment: You are now in state 65. You have 4 possible actions.
Agent:I'll take action 2.
Environment: You received a reinforcement of 5 units.

Environment: You are now in state 44. You have 5 possible actions.

⋮ ⋮

Grid World

---

a) Find the shortest possible path to reach the goal.
b) What is the reliability of the path given the conditions?
   i) Action Desired - 0.8
   ii) Right angle to desired action - 0.1 & 0.1

Markov Decision Process

- **States** - s
- **Actions** - a(s), a
- **Model** - Pr(s'|s,a)
- **Rewards** - R(a), R(s), R(s,a)

- **Policy** - π(s) -> a

Underlying Markovian Properties

- Present
- Stationary

States

History

History is the sequence of observations, actions and rewards till that point. Observe, take action, receive reward.

H(t) = o(1),A(1),R(1),o(2),A(2),R(2),o(3),A(3),....o(t-1),A(t-1),R(t-1),o(t),A(t),R(t).

What happens next depends on the history:

>The agent selects the action.

>The environment selects observation/rewards.

State

State is the information used to determine what happens next. Formally it is a function of the history.

$$S_t = f(H_t)$$

2 Types: Environment State and Agent State.

Environment State

The environment state S_t^a(subscript t, superscript e) is the environments private representation. It isn't usually visible to the agent. Even if it is, it may contain irrelevant information.

Based on the environment state, the agent's next observation and reward are computed.

Agent State

The agent state S_t^a(subscript t, superscript a) is the agent's internal representation. Agent uses this to compute next action. This information is used by RL algorithms. It is usually a function of History.

Markov State

An information state (a.k.a. Markov state) contains all useful information from the history.

Definition

A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, ..., S_t]$$

- "The future is independent of the past given the present"

$$H_{1:t} \rightarrow S_t \rightarrow H_{t+1:\infty}$$

- Once the state is known, the history may be thrown away
- i.e. The state is a sufficient statistic of the future
- The environment state S_t^e is Markov
- The history H_t is Markov

2 kinds of environments

Fully Observable: Agent will directly observe the environment state. Hence, Agent state = Environment State = Information State.

$$O_t = S_t^a = S_t^e$$

Formally, this is a MDP. Markov Decision Process.

2 kinds of environments

Partially Observable: Agent will indirectly observe the environment state.

- A robot with camera vision isn't told its absolute location
- A trading agent only observes current prices
- A poker playing agent only observes public cards

Now agent state != environment state. Formally this is a partially observable MDP. Or a POMDP.

And that makes all the difference

- Agent must construct its own state representation S_t^a, e.g.
  - Complete history: $S_t^a = H_t$
  - Beliefs of environment state: $S_t^a = (\mathbb{P}[S_t^e = s^1], ..., \mathbb{P}[S_t^e = s^n])$
  - Recurrent neural network: $S_t^a = \sigma(S_{t-1}^a W_s + O_t W_o)$

Some examples

Agent Type	*Performance measure*	*Environment*	*Actuators*	*Sensors*
Medical diagnosis system	Healthy patient, minimize costs	Patient, hospital	Questions, tests, treatments	Symptoms, findings, patient's answers
Satellite image analysis system	Correct categorization	Satellite link	Print a categorization of scene	Pixels of varying intensity, color
Part-picking robot	% parts in correct bins	Conveyor belt with parts	Pick up parts and sort into bins	Pixels of varying intensity
Interactive English tutor	Student's score on test	Set of students; testing agency	Print exercises, suggestions, corrections	Keyboard input

From (Russell and Norvig, 2003)

Environment properties

- Fully vs. partially observable: whether agent's can obtain complete and accurate information about the environment
- deterministic vs. stochastic: whether the next state of the environment is fully determined by the current state and action performed by the agent
- episodic vs. sequential: whether agent's next action depends only on the current state of the environment (episodic), or on assessment of past environment states (sequential)
- static vs. dynamic: whether the environment changes independently of the agent's actions
- discrete vs. continuous: whether the possible actions and percepts on an environment are finite (discrete environment) or not (continuous environment)
- single vs. multiple agents

Types of environments

Environment	Observable	Deterministic	Episodic	Static	Discrete	Agents
Crossword puzzle	fully	yes	sequential	static	yes	single
Chess w/ clock	fully	strategic	sequential	semi	yes	multi
Poker	partially	strategic	sequential	static	discrete	multi
Backgammon	fully	stochastic	sequential	static	discrete	multi
Car driving	partially	stochastic	sequential	dynamic	continuous	multi
Medical diagnosis	partially	stochastic	sequential	dynamic	continuous	single
Image analysis	fully	deterministic	episodic	semi	continuous	single
Robot arm	partially	stochastic	episodic	dynamic	continuous	single
English tutor	partially	stochastic	sequential	dynamic	discrete	multi
Plant controller	partially	stochastic	sequential	dynamic	continuous	single

From (Russell and Norvig, 2003)

Next Time on Reinforcement Learning: Major Components of an RL Agent

An RL agent includes one or more of these components.

- **Policy**: Agent’s behaviour function.
- **Value** function: How good is each state and/or action.
- **Model**: Agent’s representation of the environment.
- **Rewards**

Policy

- A policy is the agent’s **behaviour**
- It is a map from **state** to **action**, e.g.
- **Deterministic** policy: $a = \pi(s)$
- **Stochastic** policy: $\pi(a|s) = P[A_t = a|S_t = s]$

Value function

- Value function is a prediction of future reward.
- Used to evaluate the goodness/badness of states.
- And there to select between actions, e.g.,

$$v_{\pi}(s)=E_{\pi}\ R_{t}+1+\gamma R_{t}+2+\gamma 2R_{t}+3+...|S_{t}\ =s$$

Model

- A model predicts what the environment will do next P predicts the next state
- R predicts the next (immediate) reward, e.g.

$$P_{ss'}^{a'} = P[S_{t+1} = s' \mid S_t = s, A_t = a]$$

$$R^a_s = E[R_t + 1 \mid S_t = s, A_t = a]$$