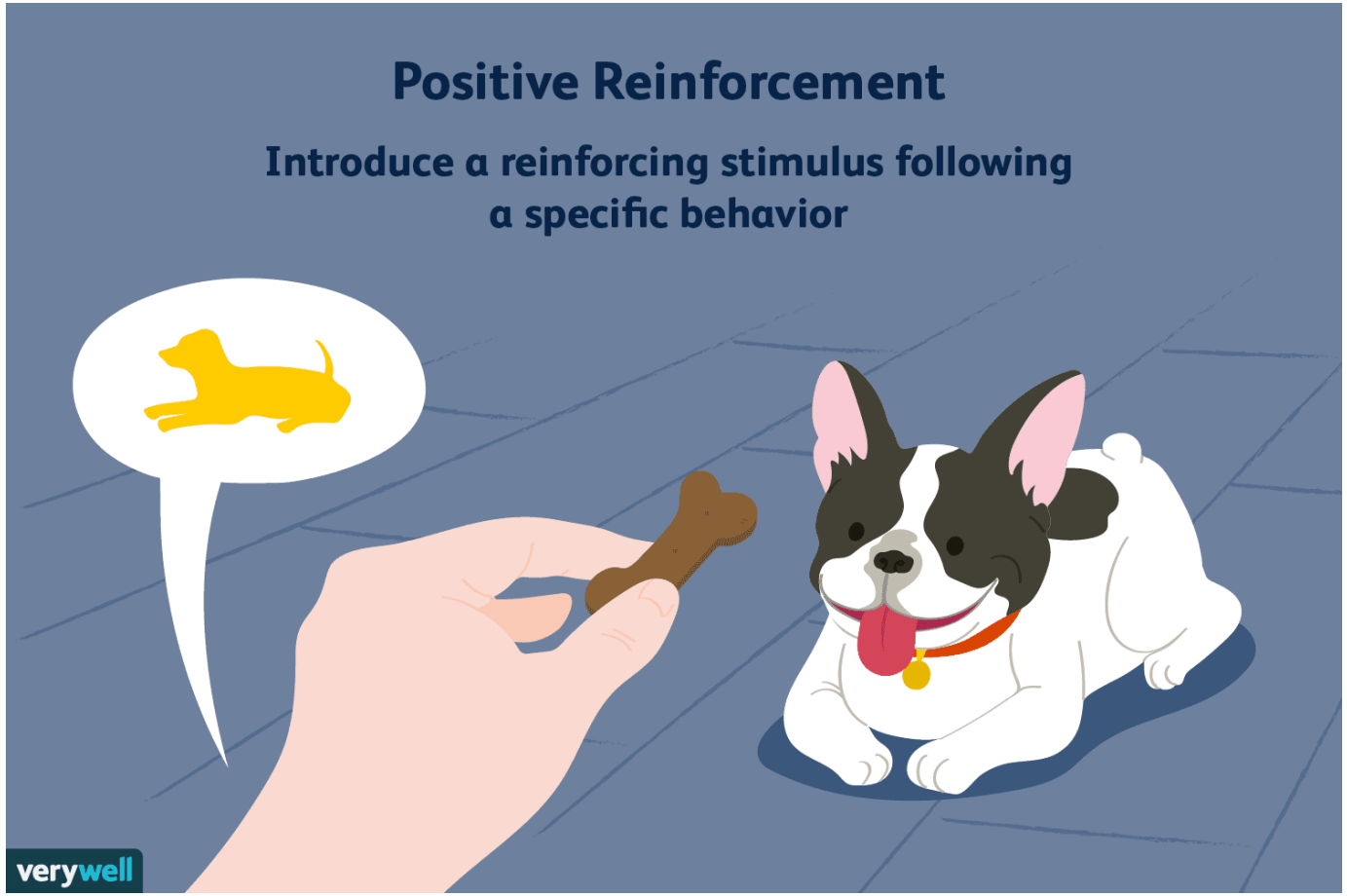
Markov Property & Markov Decision Process

Reinforcement Learning Review

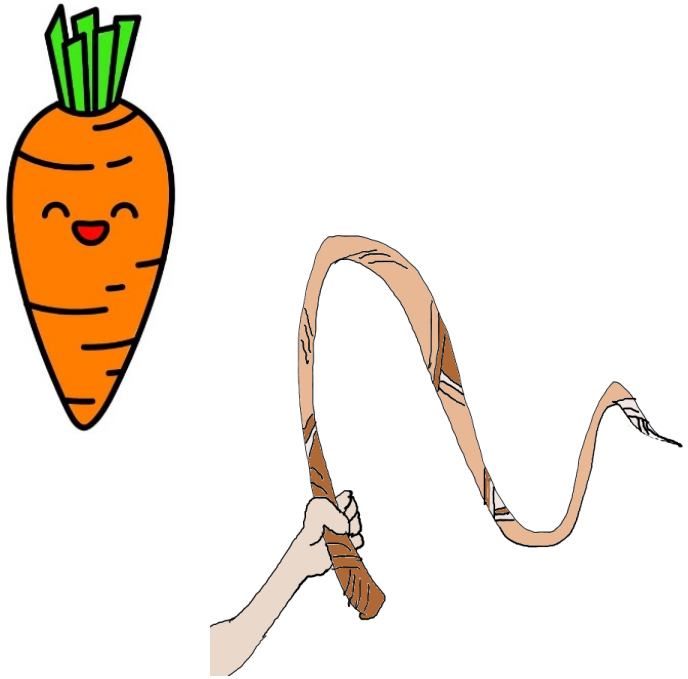
Based on Prof. Oh's Reinforcement Learning Lectures

Suinne Lee

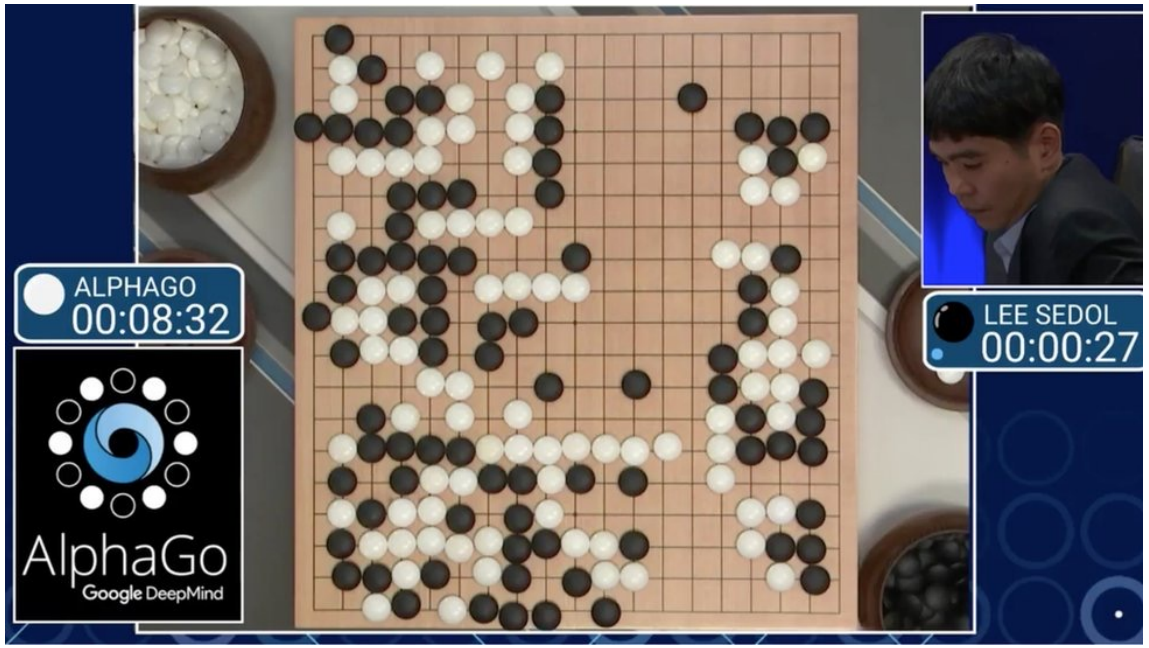
Reinforcement Learning



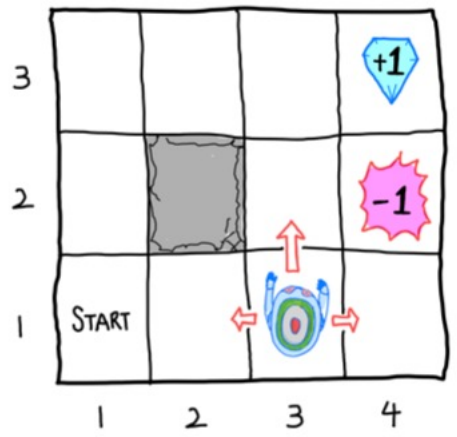
Train something to be good at something through trial and error



More suited for something where the agent actively engages with the environment



Example: Grid World

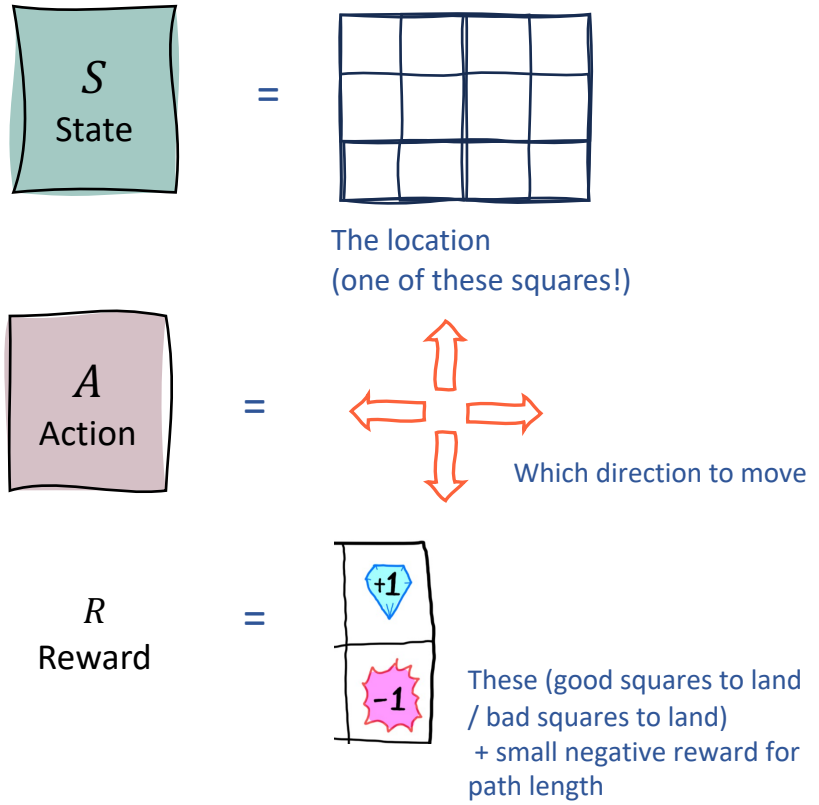


The Very Basics

- Agent
  - The player? Robot? Something that takes action!
- State:
  - Environment? Setting? Surrounding?
  - Capital S: notation for the set of all possible states, or *state space*
- Action
  - What is done
- Reward
  - Something as a measure of how good the state/action was.

Goal: maximize reward!

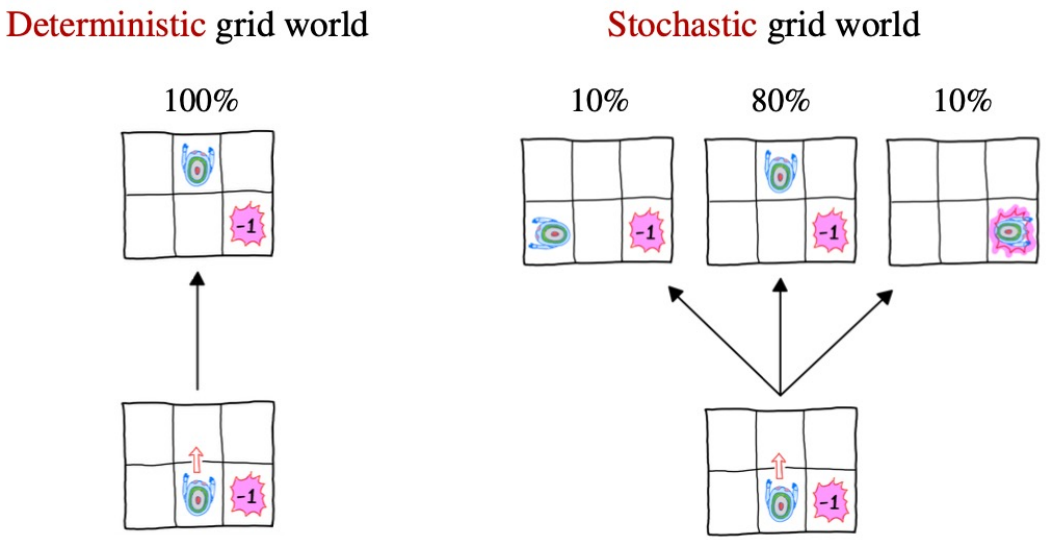
In the Grid world Example:



- State → Choose Action. (Policy)
- Action → Next State (state transition, given action)

Each of these steps may be noisy (stochastic)

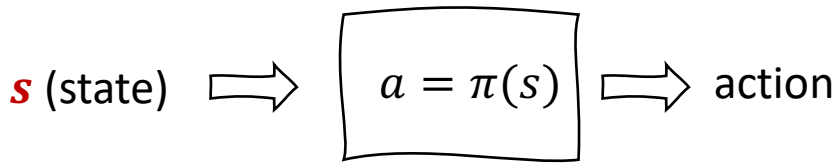
- State → Next State (state transition)



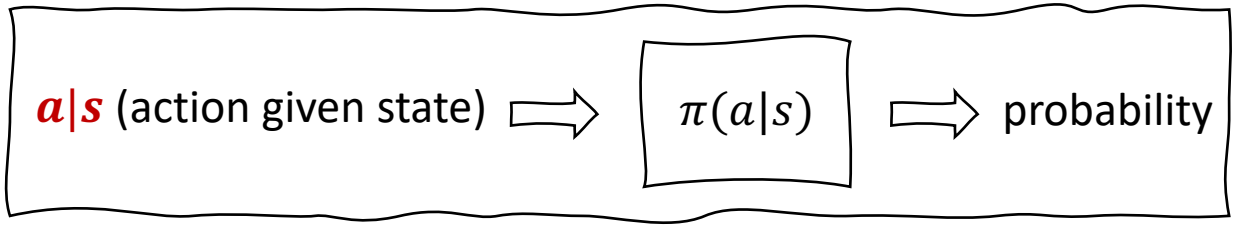
So we also establish the following concept:

- Policy
  - Deterministic: $\pi(s) = a$
    - Function of **state**. Gives a determined function value: which is the action.
  - Stochastic $\pi(a|s)$
    - **Probability distribution** of **state**. Which means the notation $\pi(a|s)$ itself means a completely different thing than $\pi(s) = a$: $\pi(a|s)$ in itself means the **probability mass or density**, as a function of the **variable $a|s$** (action given state)

Deterministic:



Stochastic:

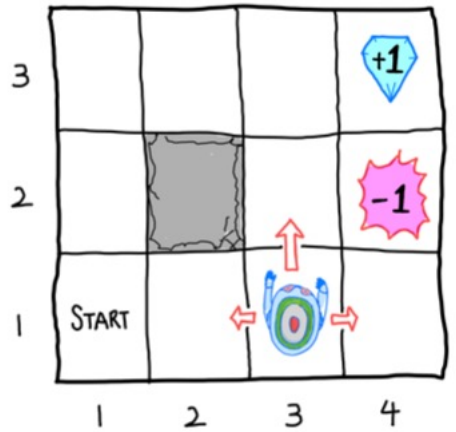


As a whole, represents action taken given a state in the form of a probability distribution.

Back to the Grid World Example

$$\text{Episode } (1,1) \xrightarrow{\text{north}} (1,2) \xrightarrow[\text{wall}]{\text{east}} (1,2) \xrightarrow{\text{north}} (1,3) \xrightarrow[\text{east } 10\%]{\text{south}} (2,3) \xrightarrow{\text{east}} (3,3) \xrightarrow{\text{east}} (4,3)$$

with the total reward $5c+1$



Episode: one full game. 한 판.

(a bit of a vague term, but we can also say that if the state, state transitions (and thus rewards) are identical tor two episodes, they are the same episode.)

Taking this into account, depending on whether the state transition is deterministic or stochastic, a single policy may bring about one or multiple episodes.

Markov Property

If we were to consider the state transition probability in a general model, the future state would depend on all past states as well as the present state. But the Markov Property restricts this to a specific case where it doesn't.

- $P(S_{t+1}|S_0, S_1, \dots S_t) = P(S_{t+1}|S_t)$
- The future state does not depend on the past states (independent of them), only on the present state.
- In information theory:

- Random variables X, Y, Z are said to form a **Markoc chain** in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X
- Specifically, $X \to Y \to Z$ if

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

Some simple consequences

- $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

- $X \to Y \to Z$ implies $Z \to Y \to X$
- If $Z = f(Y)$, then $X \to Y \to Z$

Markov Process (Markov Chain)

- Stochastic process:

  - def: collection of random variables indexed by time (can be generalized to a continuous case)

  $S_0, S_1, \ldots, S_{t-1}, S_t, S_{t+1}, \ldots$

- The Markov Process is a process where the Markov Property holds.
- In RL, we assume the processes to be Markov.

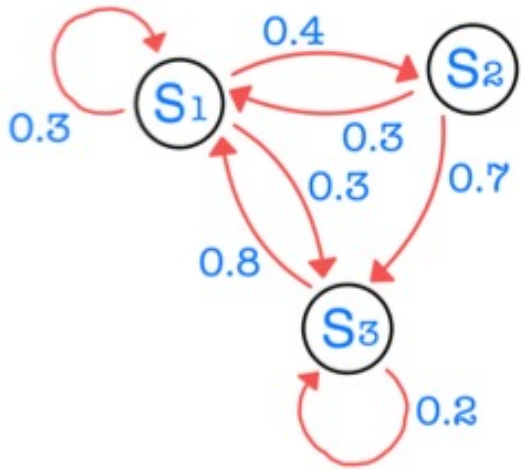
Markov Process: Mathematical Definition

- Defined as the tuple (S, P)
  - S: States. (Variables, not specific values!)
  - P: Transition matrix

    $P_{ij} = P_{s_i s_j} = p(s_j \mid s_i) = P(S_{t+1} = s_j \mid S_t = s_i)$

    - Naturally, the sum of elements in each row (outgoing from state s_i) is 1.

Markov Process



	S1	S2	S3
S1	0.3	0.4	0.3
S2	0.3	0.0	0.7
S3	0.8	0.0	0.2

Markov Decision Process (MDP)

- Elementary probabilistic model for implementing RL

Markov Decision Process (MDP)

- S: State space
- A: Action space

State and action spaces need not be finite.

- P: transition probability, from s to s' given a

$$P^{a}_{ss'} = p(s' \mid s, a) = P(S_{t+1} = s' \mid S_t = s, A_t = a)$$

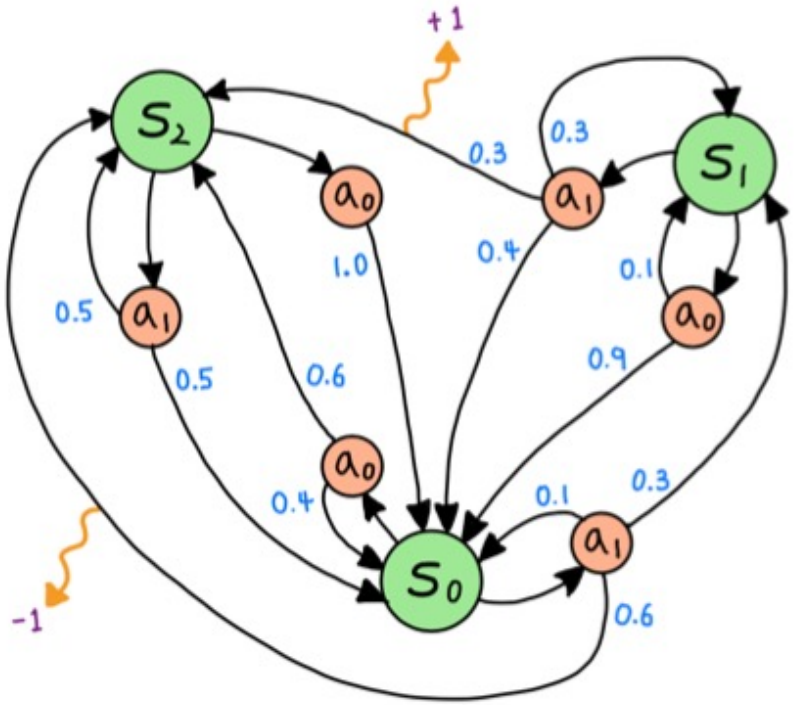
- R: Reward function

$R^{a}_{ss'}$ (possibly 3 different types: R_s, R^{a}_{s}, $R^{a}_{ss'}$)

The reward may depend on
- the current state
- the current state and action taken
- the current state, action, next state achieved

- Discount factor $\gamma \in [0, 1]$: to be explained (has to do with prioritizing immediate rewards)

Markov Decision Process (MDP)



3 states: S_0, S_1, S_2
2 actions: a_0, a_1
2 rewards: +1, -1

$$P^{a_1}_{s_0 s_1} = ?$$

$$R^{a_1}_{s_1 s_2} = ?$$

Example → next talk!

Glossary

- Agent
- State
- Action
- Reward
- Policy
- Episode
- Markov Property
- Markov Chain / Markov Process