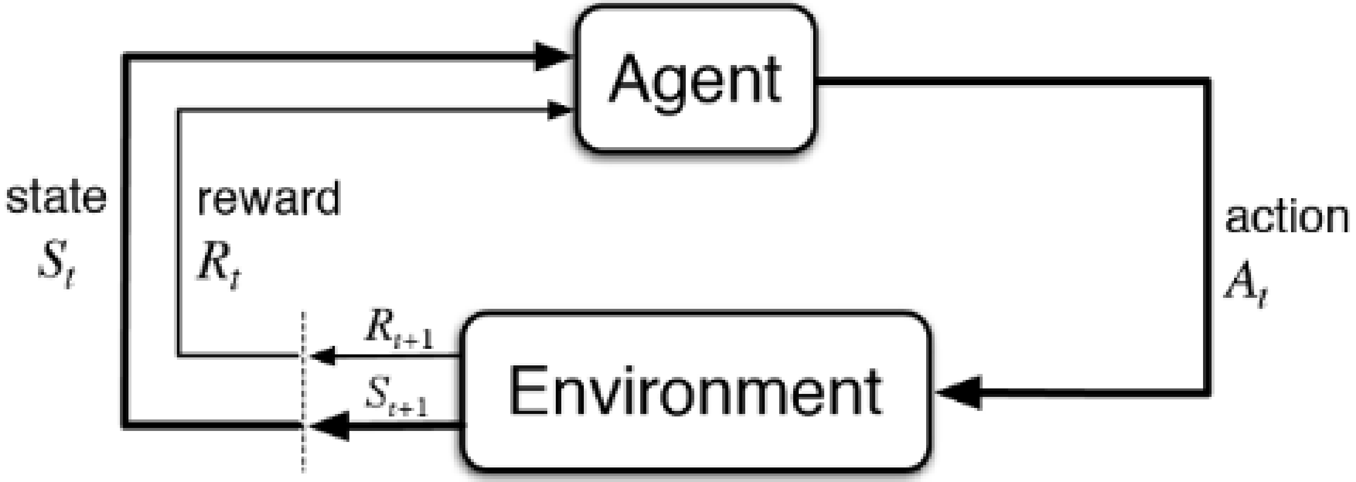
지능 시스템
Intelligent Systems

Lecture 1 – Markov Decision Process

Today's Contents

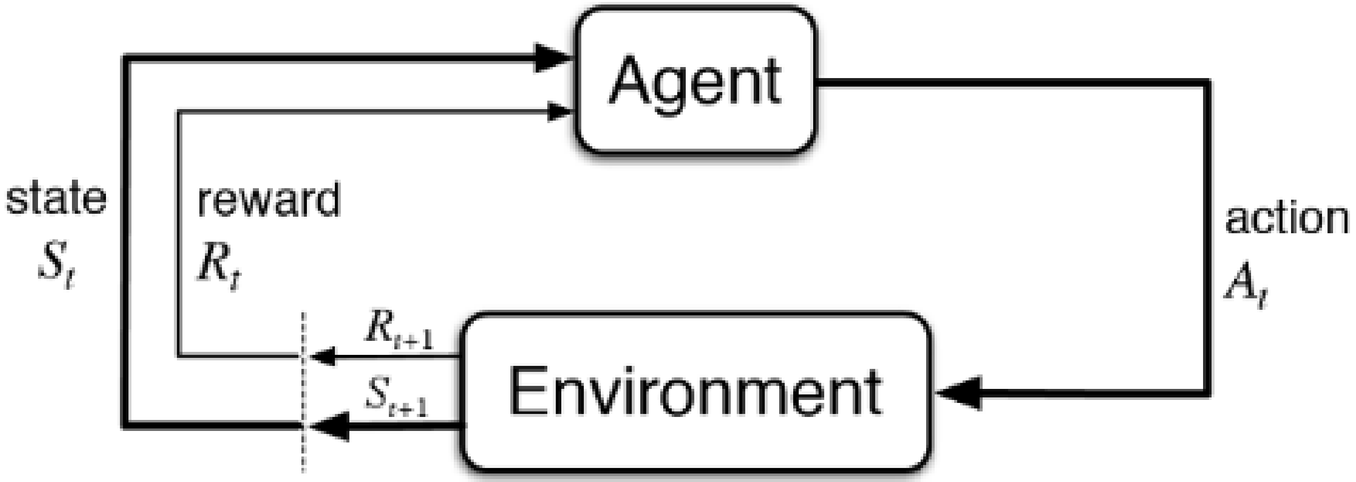
- Markov Decision Process (MDP)
- RL Terminology
- Value Function
- Action-Value Function

Interaction with the Environment



An agent interacts with the environment by taking a relevant action.
The environment gives feedback to the agent with a reward signal.

Markov Decision Process (MDP)



1. The Agent observes the initial Environment State, s_0
2. According to the state, s_0, the Agent performs an action, a_0
3. Due to the action, a_0, the Environment transits the state to its next state, s_1, and gives a reward, r_0, to the Agent.
4. The Agent chooses the next action, a_1 according to the new state, s_1.
5. The above steps are repeated until the Environment terminates. Means reaching to the terminal state, s_T

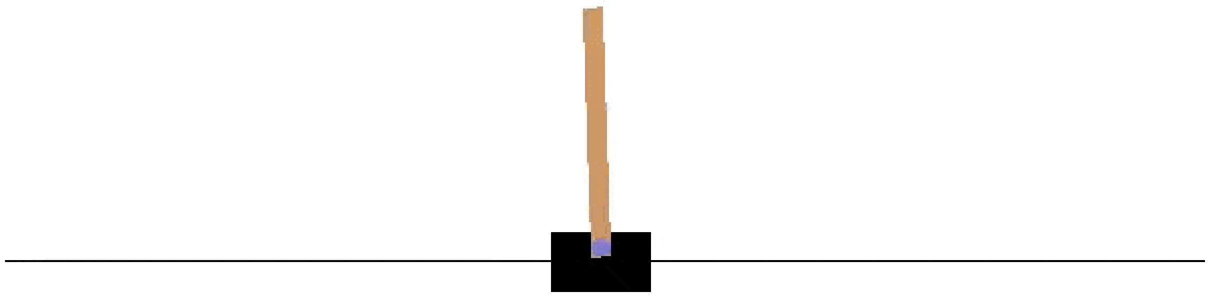
Markov Decision Process (MDP) Example: Cart Pole

Cart Pole / Inverse Pendulum:

State:
- (x,y) coordinate of the pole
- angular velocity of the pole
- velocity of the cart

Action:
- Move Left
- Move Right

Reward:
- +1 for every step



Markov Decision Process (MDP) Example: Atari – Space Invaders

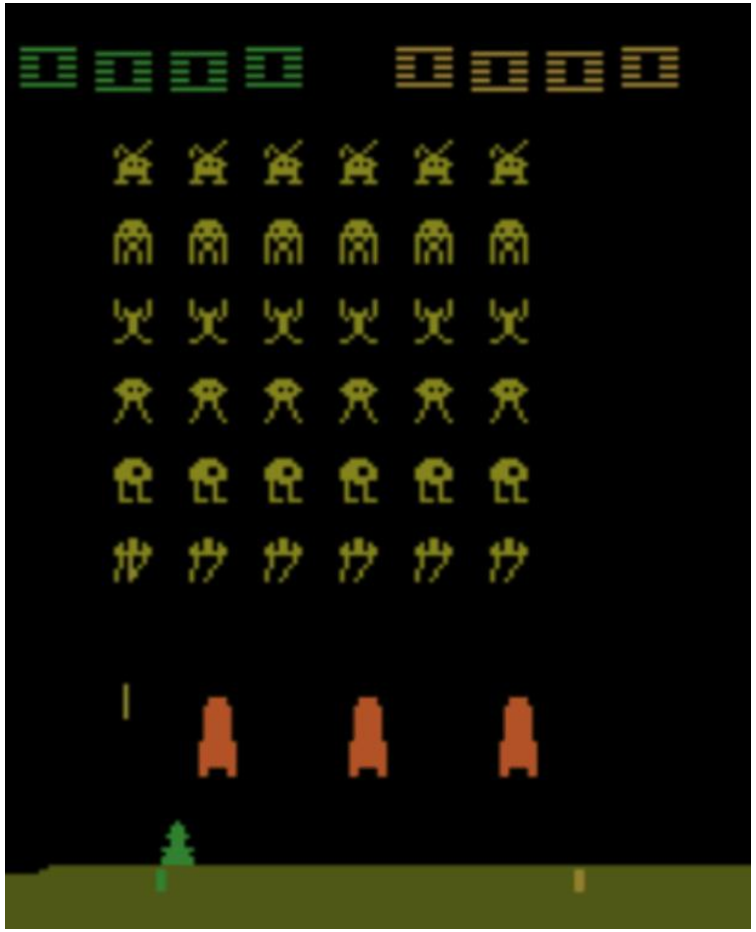
Atari – Space Invaders:

State:
- RAM info that the game engine provides

or
- The game screen image

Action:
- Move Left
- Move Right
- Shoot
- Reset

Reward:
- Game Score



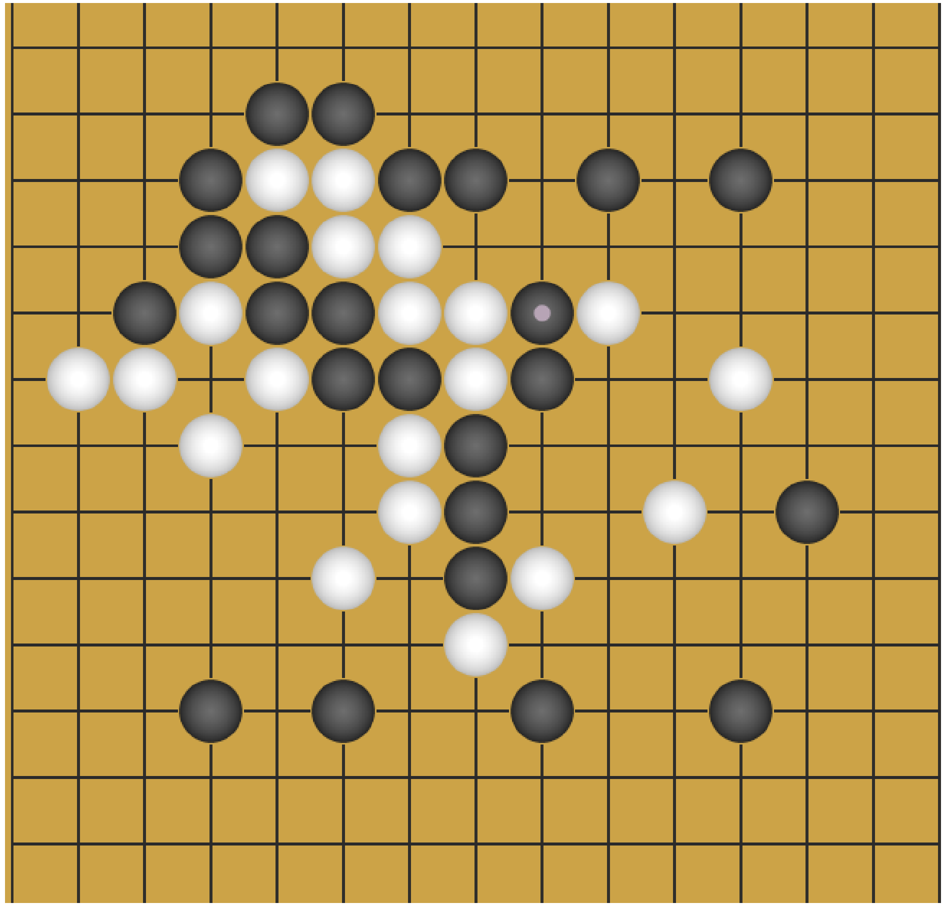
Markov Decision Process (MDP) Example: Go

Go:

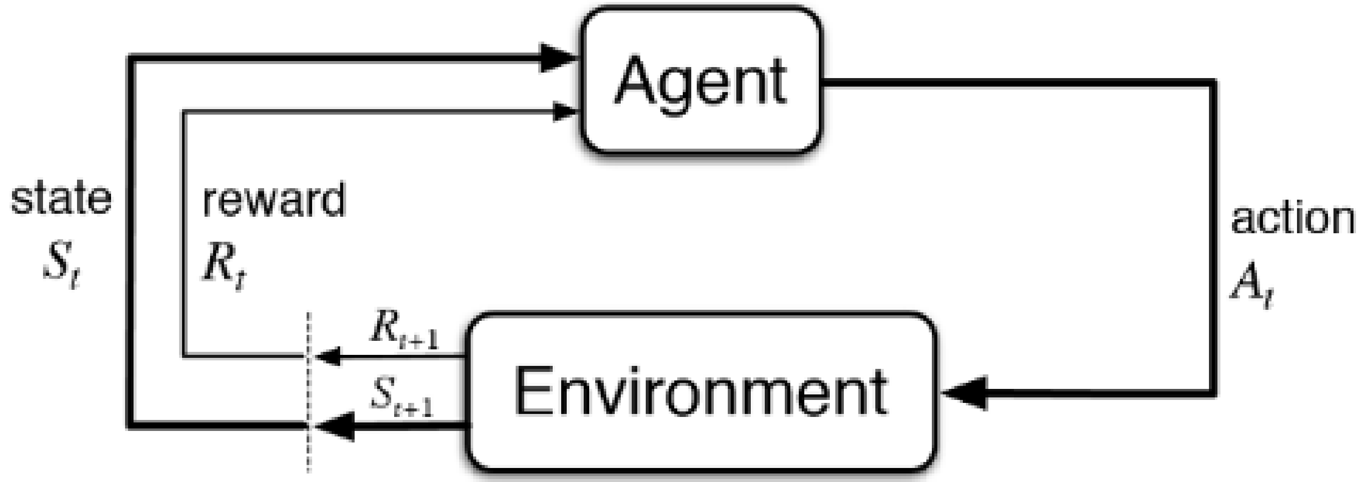
State:
- Location of Stones, empty coordinates

Action:
- The coordinate to place the stone

Reward:
- +1 If Won
- -1 If Lost
- 0 otherwise



Markov Decision Process (MDP)

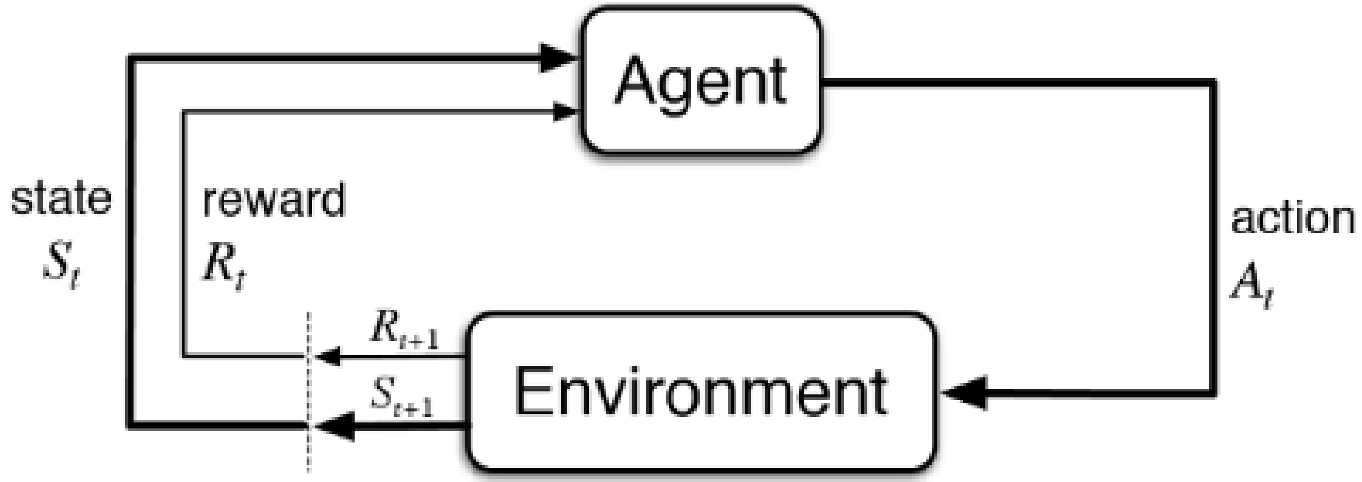


As a result, we collect data in a form of;

$$(s_0, a_0, s_1, r_0), (s_1, a_1, s_2, r_1), \dots, (s_t, a_t, s_{t+1}, r_t), \dots, (s_{T-1}, a_{T-1}, s_T, r_{T-1})$$

t : time step index
T : terminal step

Markov Decision Process (MDP)



We define a trajectory as;

$$\tau = (\, s_0, a_0, s_1, a_1, s_2, a_2, \dots, s_T, a_T)$$

State Transition

$$s_t, a_t \longrightarrow \boxed{\textbf{Environment}} \longrightarrow S_{t+1}$$

The Environment provides a state transition function.

$$p(s_{t+1}|s_t, a_t)$$

State transition Probability: Probability of reaching to the next state given the current state and action.

Markov Sequence

The sequence of states provided by MDP is assumed to be Markov.
That is;

$$p(s_{t+1}|s_t, s_{t-1}, s_{t-2}, \dots, s_0, a_t, a_{t-1}, a_{t-2}, \dots, a_0) = p(s_{t+1}|s_t, a_t)$$

State transition Probability of a Markov Sequence:

You only need just one step previous information to extract next information. Further history is not required.

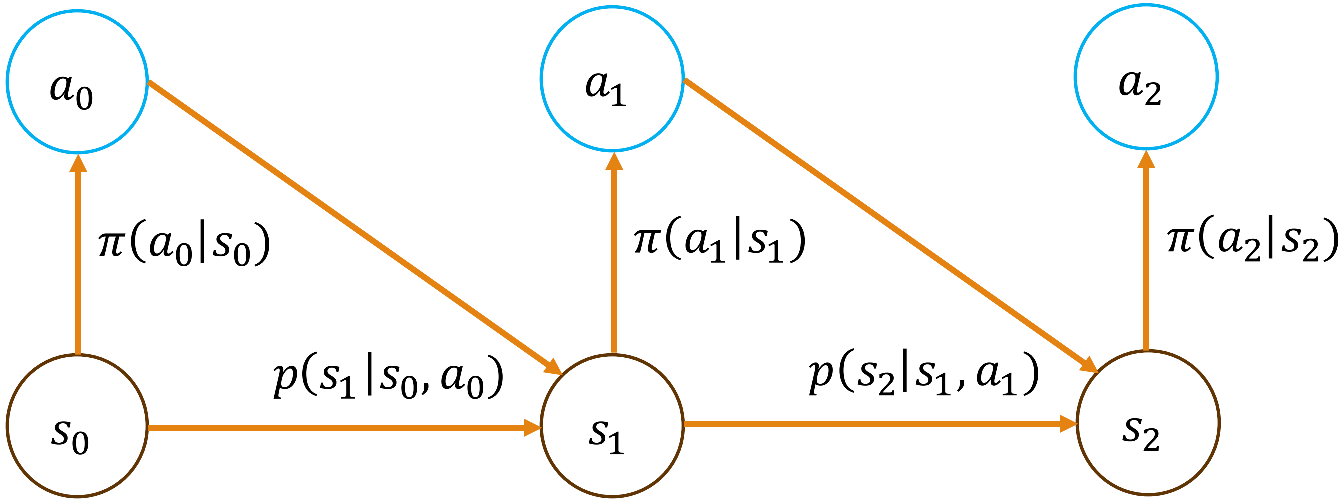
Policy

Policy is a function that maps a state to an action.

$$\pi(a_t|s_t) = p(a_t|s_t)$$

Policy: a probability of selecting a certain action, given the state.

MDP Procedure



$$\tau = (s_0, a_0, s_1, a_1, s_2, a_2, \dots)$$

Return

Reward, r_t : The instantaneous signal that the agent receives at each time step.

Return, G_t: The sum of rewards from the current time step, t, to the terminal time step, T, if exists.

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \ldots + \gamma^{T-t} r_T$$

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} r_t$$

Discount Rate / Discount Factor, $\gamma \in [0, 1]$: A constant value to indicate that future rewards are worth less than instant rewards. It also prevents Return from being infinite.

Value Function

Value refers to "How Good this state is".

It is defined as:

$$\boxed{V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_{\pi}(\tau|s_t)}[G_t|s_t]}$$

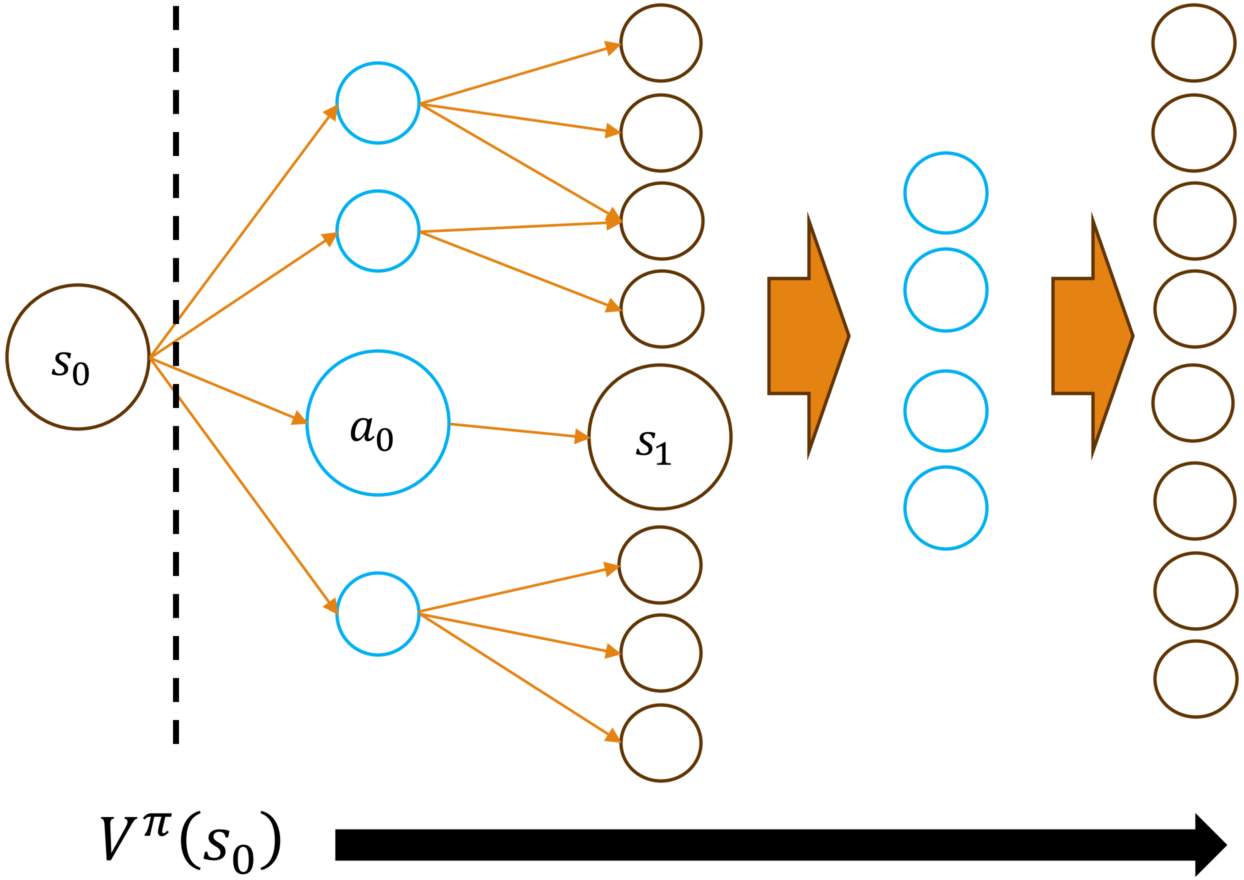
Why Expectation? Because we usually have many different trajectories generated from one policy.

The trajectory, τ, starts with the action, a_t

$$\tau_{a_t:a_T} = (a_t, s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \dots, a_T)$$

Given several trajectories, we use them to compute Returns for each trajectory at each time step. Taking the mean leads to a Value function.

Value Function



Value Function – Example 1

$$S_0 \xrightarrow{r\,=\,+1} S_1 \xrightarrow{r\,=\,+1} S_2 \xrightarrow{r\,=\,+1} S_3$$

S_0: Initial State
S_3: Terminal State
Assume $\gamma = 1$

$$\boxed{V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_{\pi}(\tau|s_t)}[G_t|s_t]}$$

Reward of +1 for each step and transition.
Consider the values at each state;

$$V(S_2) = 1$$
$$V(S_1) = 2$$
$$V(S_0) = 3$$

Value Function – Example 1

$$S_0 \xrightarrow{r\,=\,+1} S_1 \xrightarrow{r\,=\,+1} S_2 \xrightarrow{r\,=\,+1} S_3$$

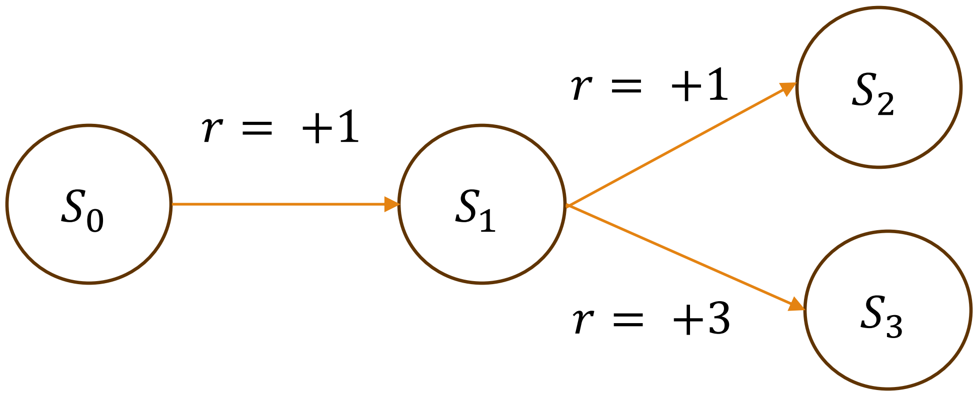
S_0: Initial State
S_3: Terminal State
Assume $\boxed{\gamma = 0.9}$

$$\boxed{V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_{\pi}(\tau | s_t)}[G_t | s_t]}$$

Reward of +1 for each step and transition.
Consider the values at each state;

$$V(S_2) = 1$$
$$V(S_1) = 1.9$$
$$V(S_0) = 2.71$$

Value Function – Example 2



$p(S_2|S_1) = 0.7$
$p(S_3|S_1) = 0.3$

S_0: Initial State
S_2, S_3: Terminal State
Assume $\gamma = 1$

$$\boxed{V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_{\pi}(\tau|s_t)}[G_t|s_t]}$$

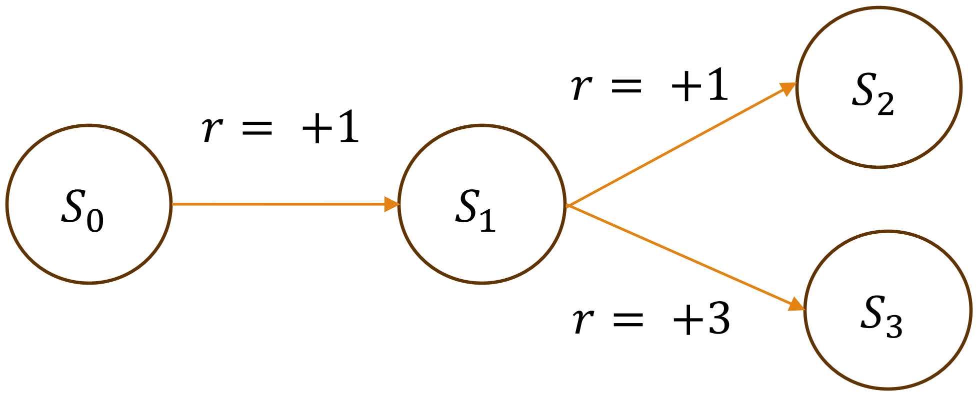
Consider the values at each state;

$$V(S_1) = p(S_2|S_1) * (+1) + p(S_3|S_1) * (+3) = 1.6$$

$$V(S_0) = (+1) + V(S_1) = 2.6$$

This is known as a Bellman Equation. We will learn more about this later.

Value Function – Example 2



$p(S_2|S_1) = 0.7$
$p(S_3|S_1) = 0.3$

S_0: Initial State
S_2, S_3: Terminal State
Assume $\boxed{\gamma = 0.9}$

$$\boxed{V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_{\pi}(\tau|s_t)}[G_t|s_t]}$$

Consider the values at each state;

$$V(S_1) = p(S_2|S_1) * (+1) + p(S_3|S_1) * (+3) = 1.6$$

$$V(S_0) = (+1) + \gamma * V(S_1) = 2.44$$

This is known as a Bellman Equation. We will learn more about this later.

Action-Value Function

Action-Value refers to "How Good is taking this action at this state".

It is defined as:

$$\boxed{Q^{\pi}(s_t, a_t) = \mathbb{E}_{\tau_{s_{t+1}:a_T} \sim p_{\pi}(\tau | s_t)}[G_t | s_t, a_t]}$$

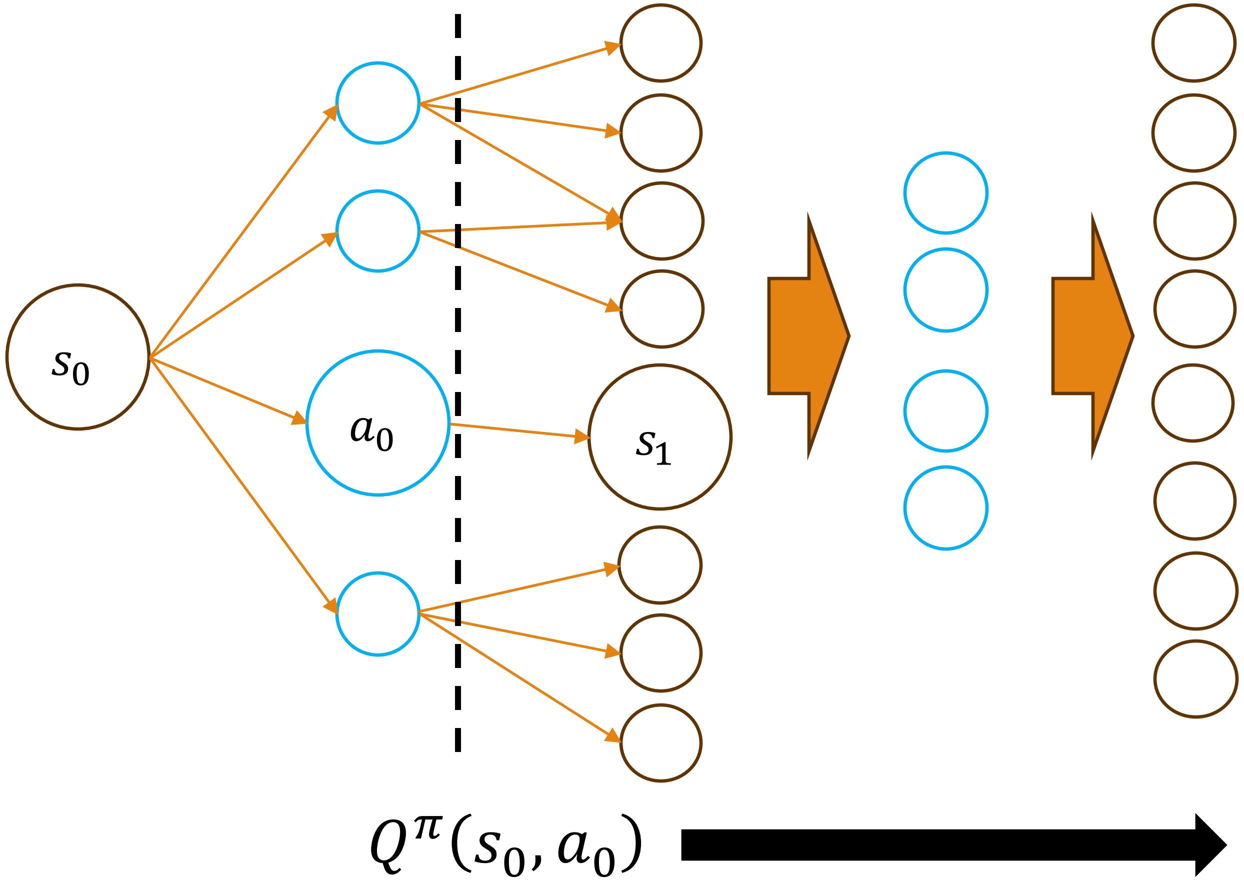
Why Expectation? Because we usually have many different trajectories generated from one policy.

The trajectory, τ, starts with the next state, s_{t+1}

$$\tau_{s_{t+1}:a_T} = (s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \ldots, a_T)$$

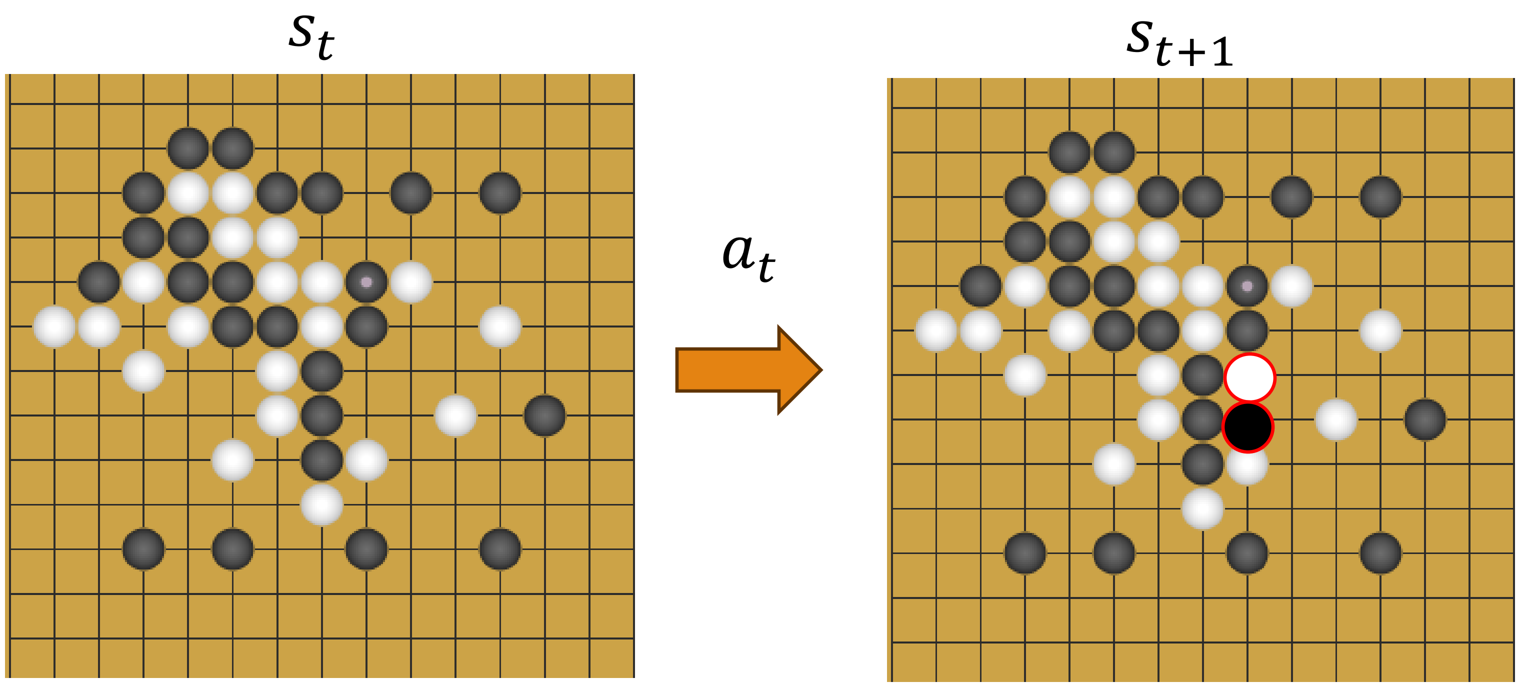
Given several trajectories, we use them to compute Returns for each trajectory at each time step.
Taking the mean leads to a Value function.

Action-Value Function



Action-Value – Example 1

Assume the Agent plays White

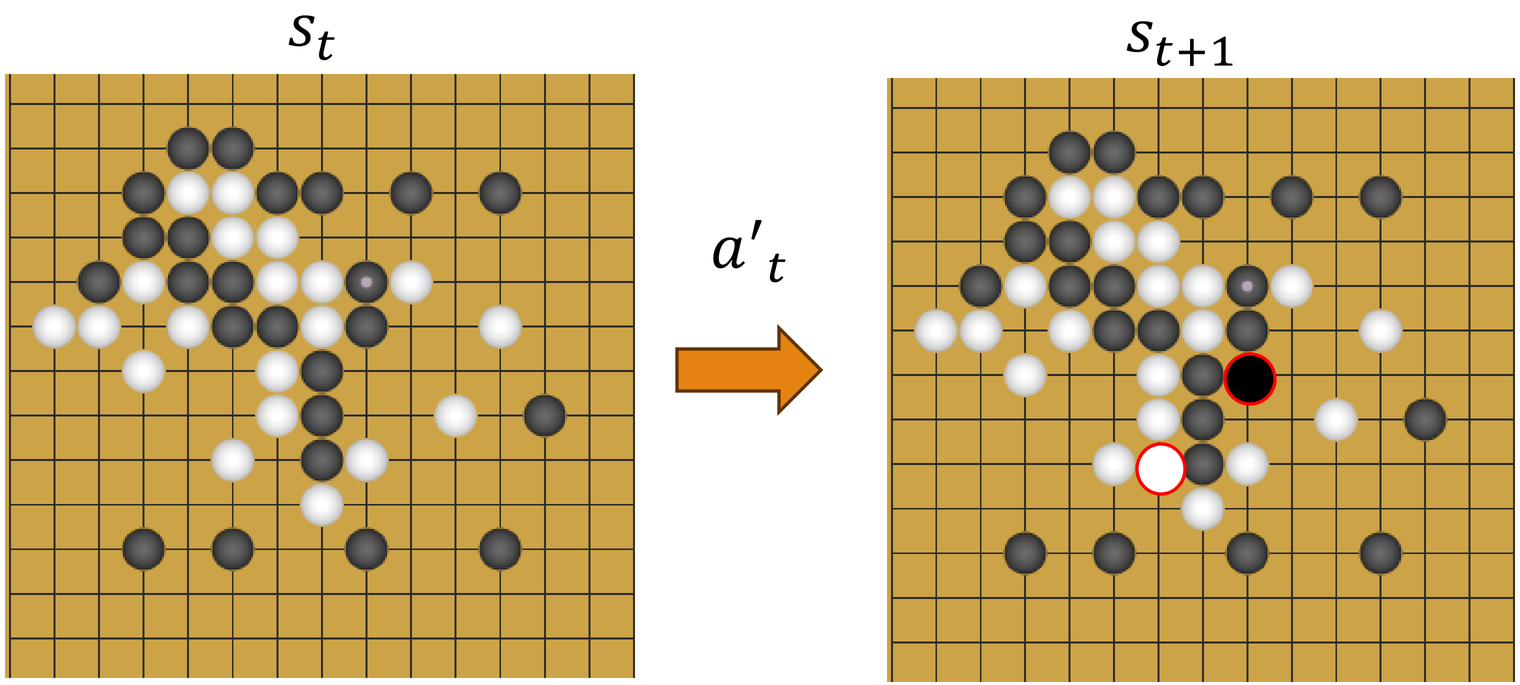


$V^\pi(s_t) = 0.5$

$V^\pi(s_{t+1}) = 0.6$

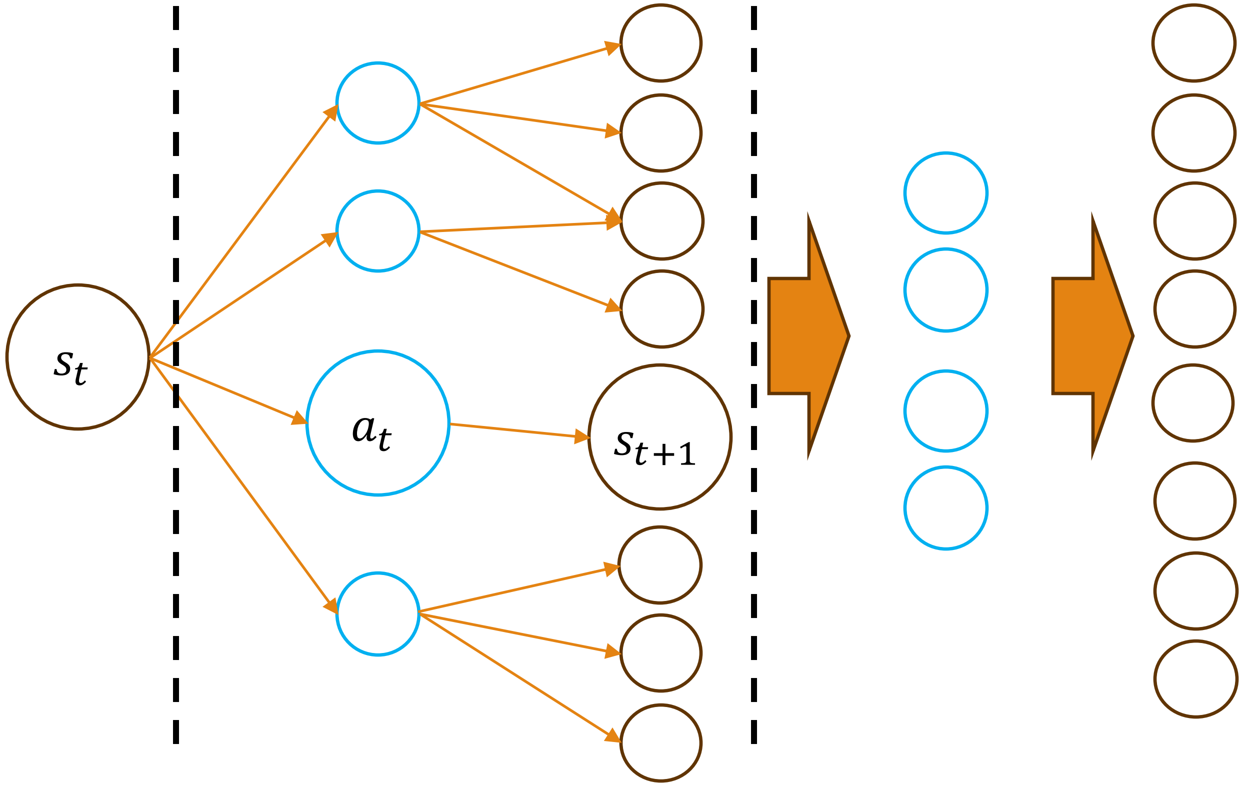
Action-Value – Example 1

Assume the Agent plays White



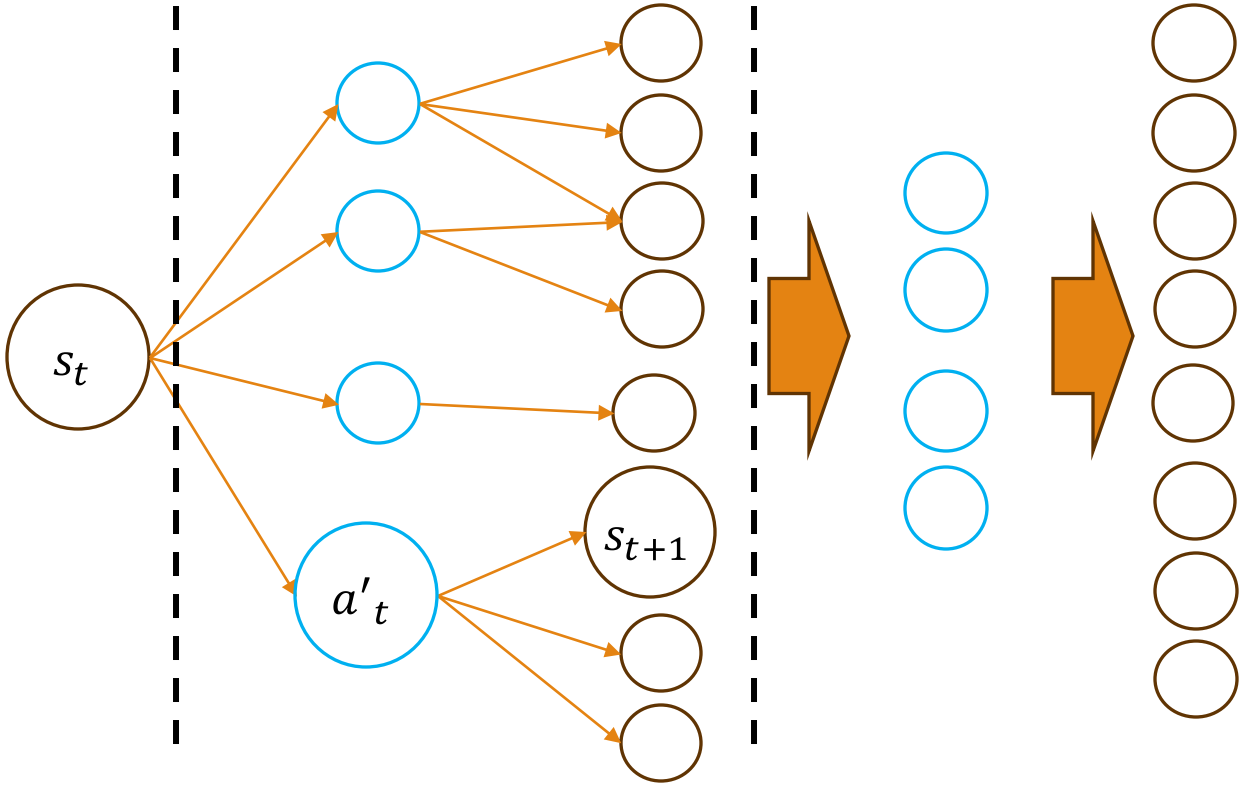
$V^\pi(s_t) = 0.5$ $\qquad$ $V^\pi(s_{t+1}) = 0.4$

Action-Value – Example 1



$V^\pi(s_t) = 0.5$ $V^\pi(s_{t+1}) = 0.6$

Action-Value – Example 1



$V^{\pi}(s_t) = 0.5$

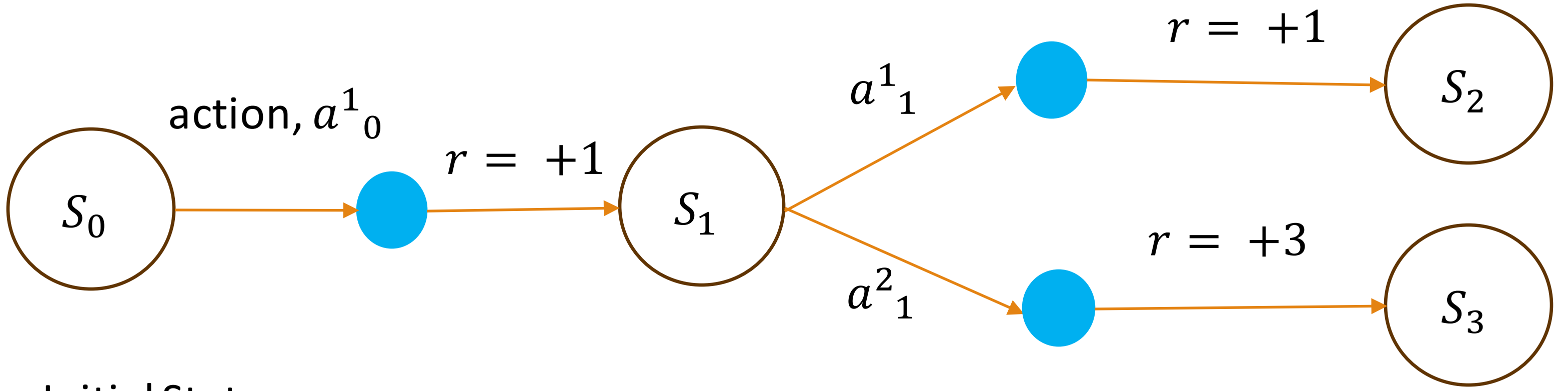
$V^{\pi}(s_{t+1}) = 0.4$

Action-Value – Example 1

Taking a_t leads to a better state, with higher value. Therefore,

$$Q^{\pi}(s_t, a_t) \; > \; Q^{\pi}(s_t, a'_t)$$

Action-Value – Example 2



$\pi(a^1{}_0|S_0) = 1$
$\pi(a^1{}_1|S_1) = 0.5$
$\pi(a^2{}_1|S_1) = 0.5$

S_0: Initial State
S_2, S_3: Terminal State
Assume $\gamma = 1$

$$\boxed{V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_{\pi}(\tau|s_t)}[G_t|s_t]}$$

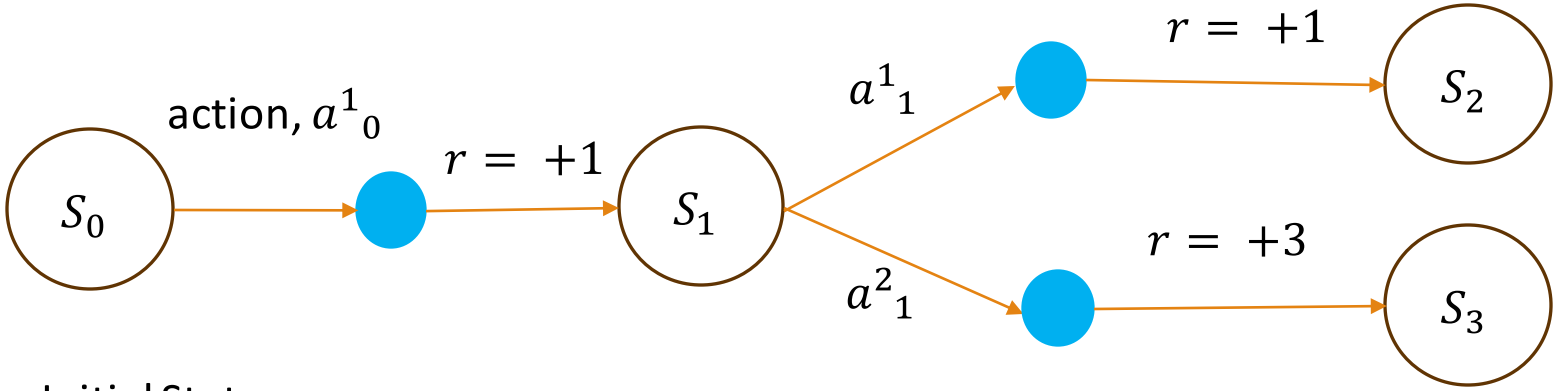
$$\boxed{Q^{\pi}(s_t, a_t) = \mathbb{E}_{\tau_{s_{t+1}:a_T} \sim p_{\pi}(\tau|s_t)}[G_t|s_t, a_t]}$$

Consider the values at each state;

$$V(S_1) = \pi(a^1{}_1|S_1) * (+1) + \pi(a^2{}_1|S_1) * (+3) = 2$$

$$V(S_0) = \pi(a_0|S_0) * (+1) + V(S_1) = 3$$

Action-Value – Example 2



$$\pi(a^1{}_0|S_0) = 1$$
$$\pi(a^1{}_1|S_1) = 0.5$$
$$\pi(a^2{}_1|S_1) = 0.5$$

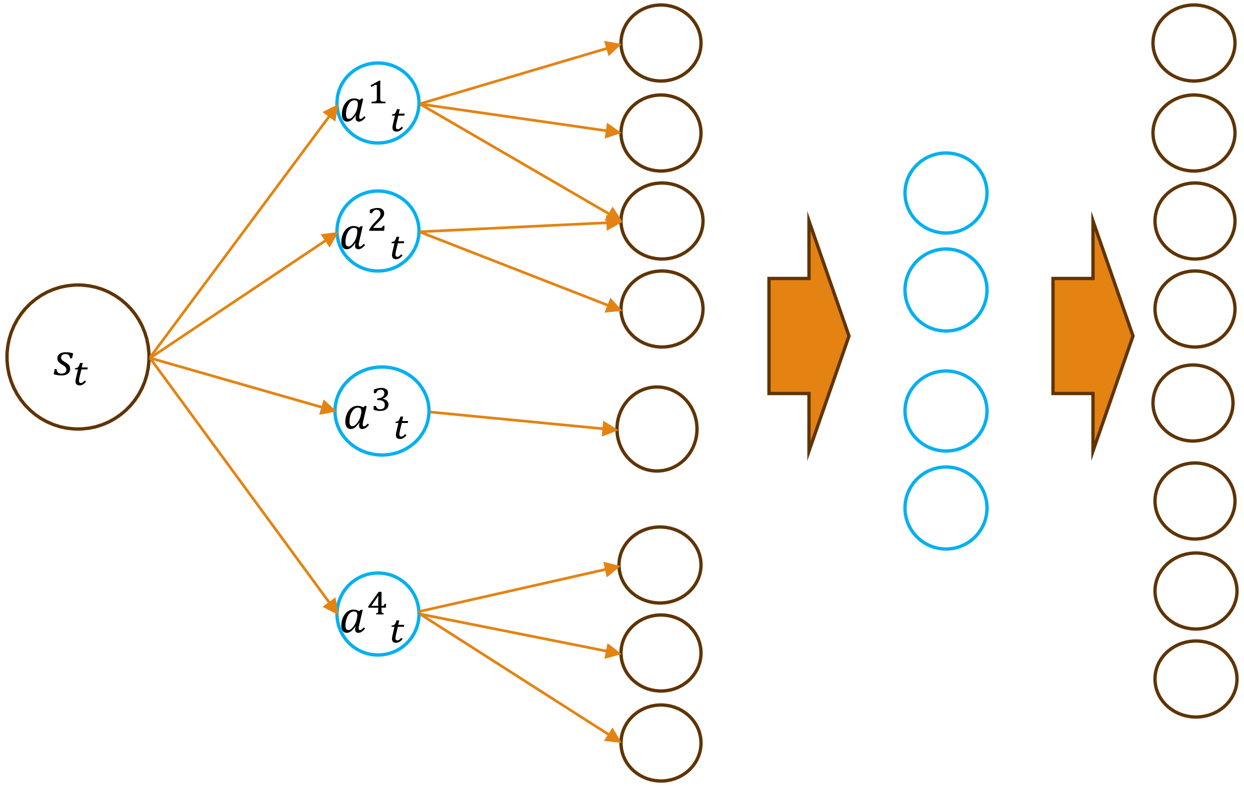
S_0: Initial State
S_2, S_3: Terminal State
Assume $\gamma = 1$

$$\boxed{V^\pi(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_\pi(\tau|s_t)}[G_t|s_t]}$$
$$\boxed{Q^\pi(s_t, a_t) = \mathbb{E}_{\tau_{s_{t+1}:a_T} \sim p_\pi(\tau|s_t)}[G_t|s_t, a_t]}$$

Consider the action-values at each state/action pair;

$$V(S_1) = \pi(a^1{}_1|S_1) * (+1) + \pi(a^2{}_1|S_1) * (+3) = 2$$
$$Q(a^1{}_1, S_1) = 1$$
$$Q(a^2{}_1, S_1) = 3$$

V and Q relationship



V and Q relationship

Assume we have only 4 possible actions at state, s_t.

Then by definition;

V: “How Good this state is”

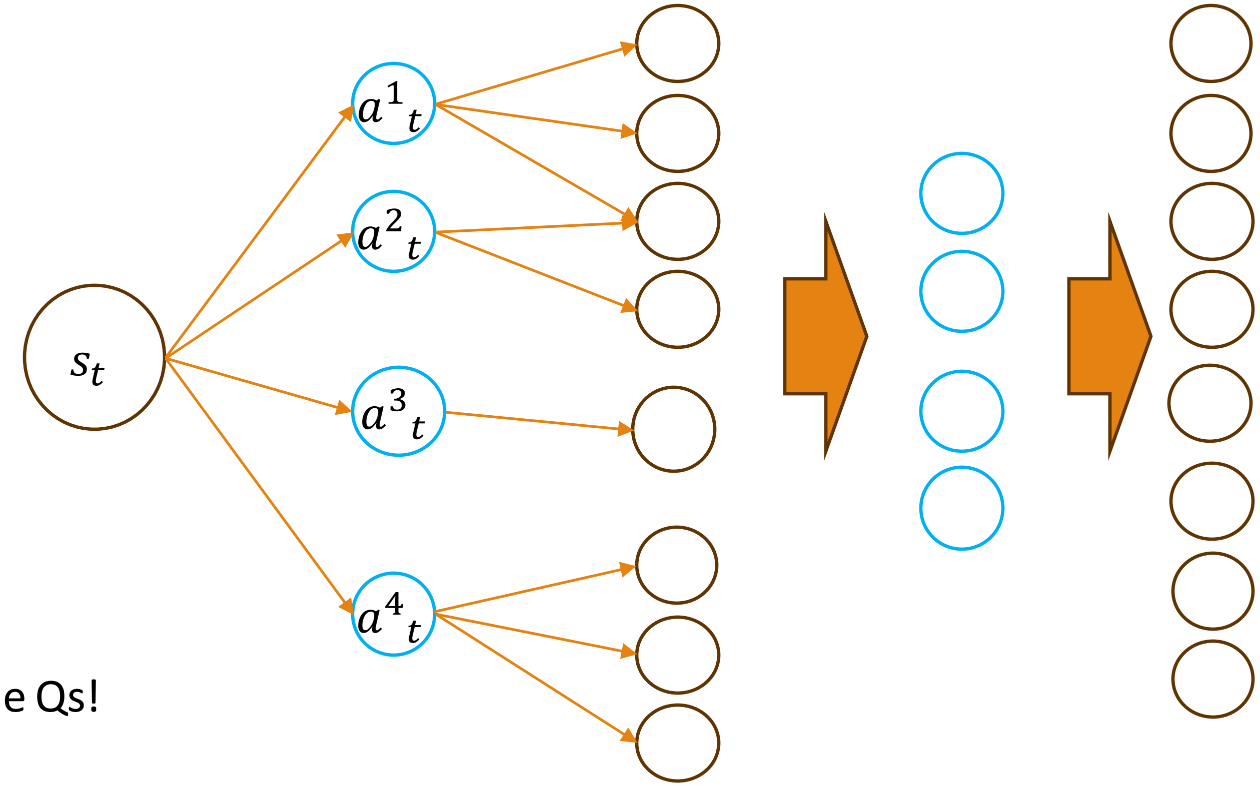
$$V^{\pi}(s_t) = \mathbb{E}_{\tau \sim p_{\pi}(\tau|s_t)}[G_t | s_t]$$

Q: “How Good is taking this action at this state”

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\tau \sim p_{\pi}(\tau|s_t)}[G_t | s_t, a_t]$$

Value turns out to be **weighted mean** value of all possible Qs!

$$V^{\pi}(s_t) = \mathbb{E}_{k \sim \pi}[Q^{\pi}(s_t, {a^k}_t)]$$



V and Q relationship – Proof

Consider the Trajectories:

$$\tau_{a_t:a_T} = (a_t, s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \dots, a_T)$$

$$\tau_{s_{t+1}:a_T} = (s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \dots, a_T)$$

$$\tau_{a_t:a_T} = (a_t) \cup (s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \dots, a_T)$$

$$\tau_{a_t:a_T} = (a_t) \cup \tau_{s_{t+1}:a_T}$$

V and Q relationship – Proof

Consider the Value function by definition;

$$V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_T} \sim p_{\pi}(\tau|s_t)}[G_t|s_t]$$

$$V^{\pi}(s_t) = \int_{\tau_{a_t:a_T}} G_t \, p(\tau_{a_t:a_T}|s_t) \, d\tau_{a_t:a_T}$$

$$\tau_{a_t:a_T} = (a_t) \cup \tau_{s_{t+1}:a_T}$$

Now consider the probability function;

$$p(\tau_{a_t:a_T}|s_t) = p(a_t|s_t) p(\tau_{s_{t+1}:a_T}|s_t, a_t)$$

$$p(\tau_{a_t:a_T}|s_t) = \pi(a_t|s_t) p(\tau_{s_{t+1}:a_T}|s_t, a_t)$$

$$V^{\pi}(s_t) = \int_{a_t} \int_{\tau_{s_{t+1}:a_T}} G_t \, p(\tau_{s_{t+1}:a_T}|s_t, a_t) \, \pi(a_t|s_t) \, d\tau_{s_{t+1}:a_T} d_{a_t}$$

V and Q relationship – Proof

$$V^{\pi}(s_t) = \int_{a_t} \int_{\tau_{s_{t+1}:a_T}} G_t \, p\left(\tau_{s_{t+1}:a_T} \middle| s_t, a_t\right) \pi(a_t|s_t) \, d\tau_{s_{t+1}:a_T} d_{a_t}$$

$$V^{\pi}(s_t) = \int_{a_t} \left[\int_{\tau_{s_{t+1}:a_T}} G_t \, p\left(\tau_{s_{t+1}:a_T} \middle| s_t, a_t\right) d\tau_{s_{t+1}:a_T}\right] \pi(a_t|s_t) da_t$$

$$V^{\pi}(s_t) = \int_{a_t} Q^{\pi}(s_t, a_t) \, \pi(a_t|s_t) da_t$$

$$V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi(a_t|s_t)}[Q^{\pi}(s_t, a_t)|s_t]$$

Exercise

α and β are the only possible actions.
S_x are states for all x.
Rewards are given only at terminal states.
Policy is uniform. That is; $\pi(\alpha|S_x) = \pi(\beta|S_x) = 0.5$ for all states.
Transition probabilities are as follows;

$$p(S_D|S_B,\alpha) = 0.2$$
$$p(S_E|S_B,\alpha) = 0.8$$
$$p(S_F|S_B,\beta) = 0.1$$
$$p(S_G|S_B,\beta) = 0.9$$
$$p(S_H|S_C,\alpha) = 1.0$$
$$p(S_I|S_C,\beta) = 0.05$$
$$p(S_K|S_C,\beta) = 0.95$$

Q1: Compute values at state S_A, S_B, and S_C
Q2: Compute $Q(S_C,\alpha)$ and $Q(S_C,\beta)$. Then show their relationship with the value at S_C.