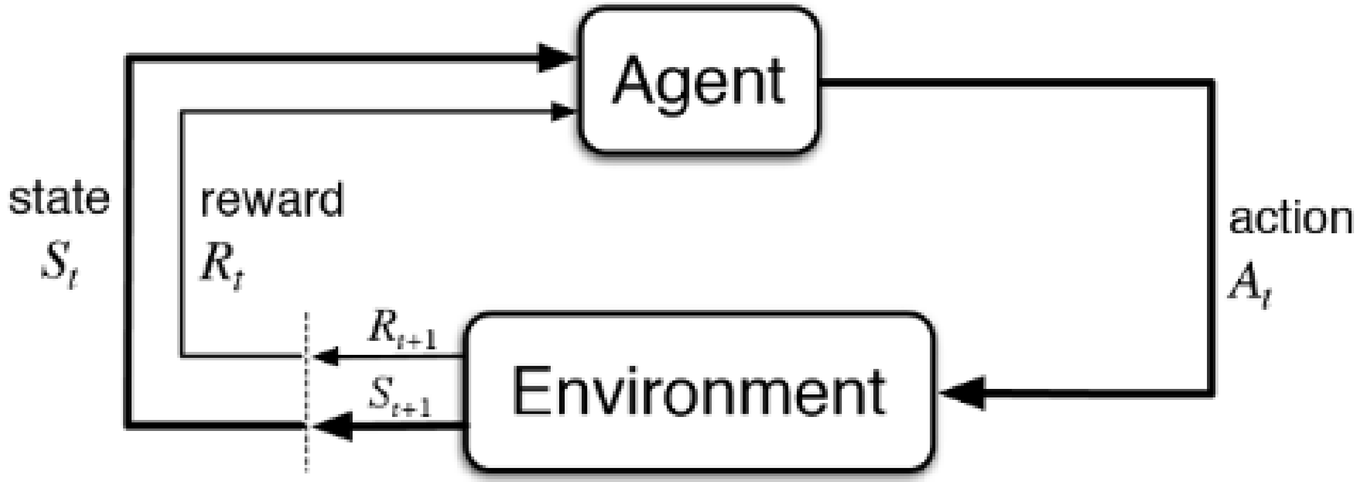
로봇공학개론 (학습기반)
Learning-based Robotics

Lecture 2 – Bellman Equation

Today's Contents

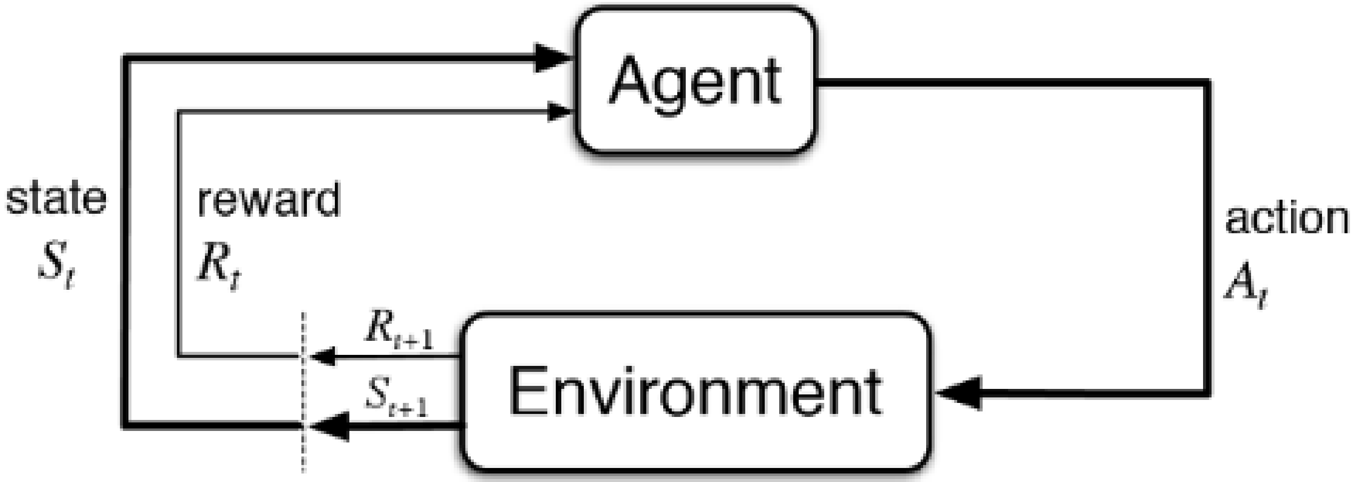
- Revision of MDP
- Bellman Equation
- Maze Example

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)

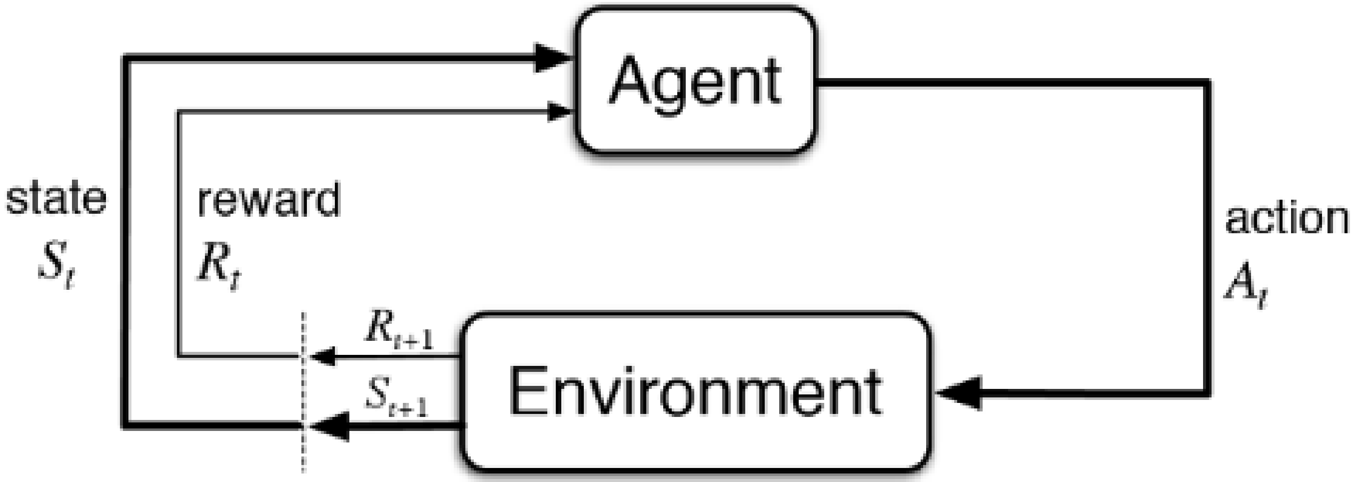


As a result, we collect data in a form of;

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

t : time step index
H : terminal step (also known as Horizon)

Markov Decision Process (MDP)



We define a trajectory as;

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

State Transition

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.

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, H, if exists.

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

$$G_t = \sum_{k=t}^{H} \gamma^{k-t} r_k$$

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:

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

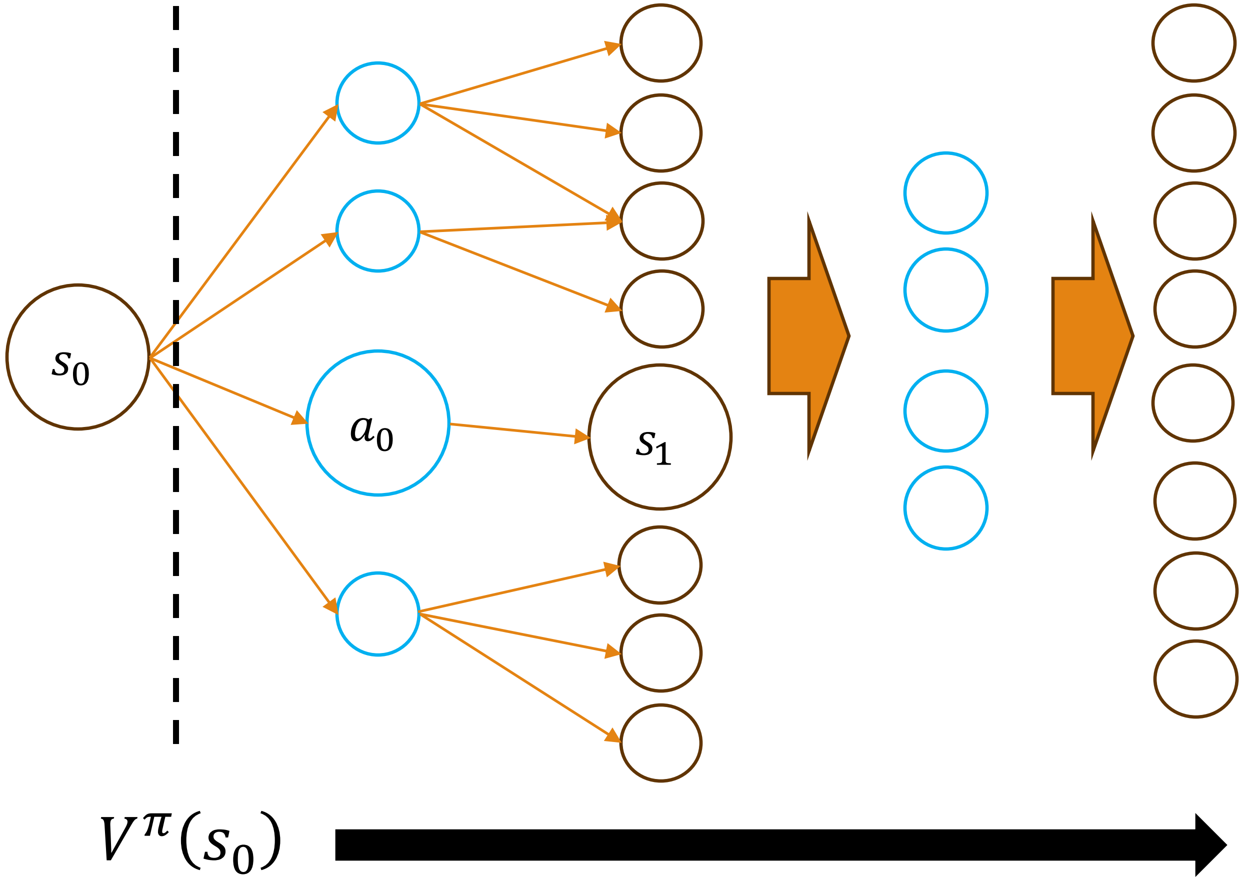
Why Expectation? Because we usually have many different trajectories generated from one policy.

The trajectory, τ, starts with the action, a_t

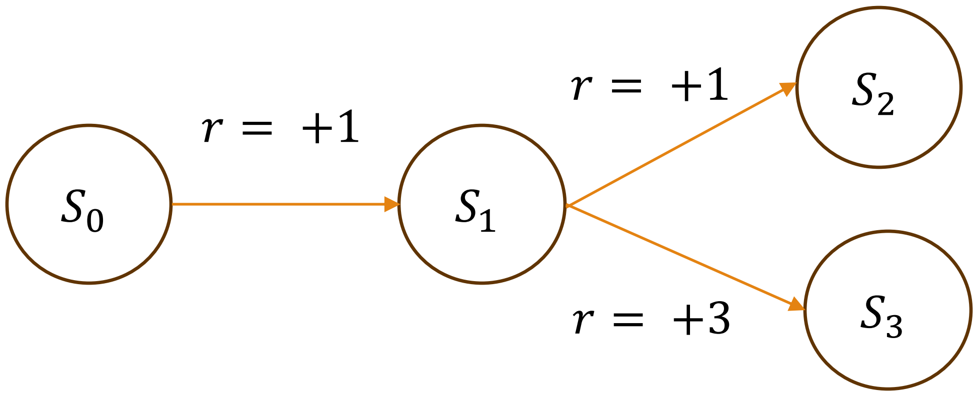
$$\tau_{a_t:a_H} = (a_t, s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \dots, a_H)$$

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



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

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:

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

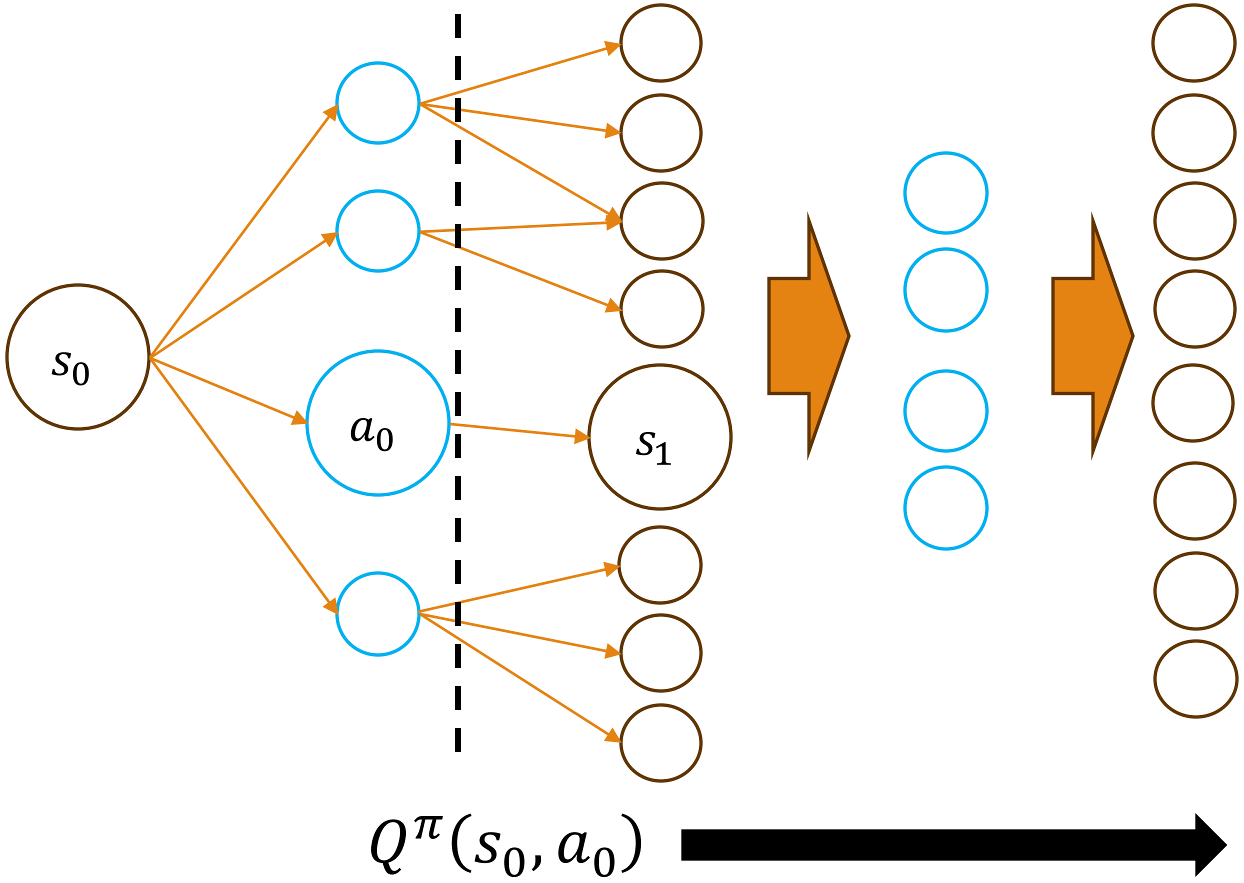
Why Expectation? Because we usually have many different trajectories generated from one policy.

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

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

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



V and Q relationship

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

Then by definition;

V: "How Good this state is"

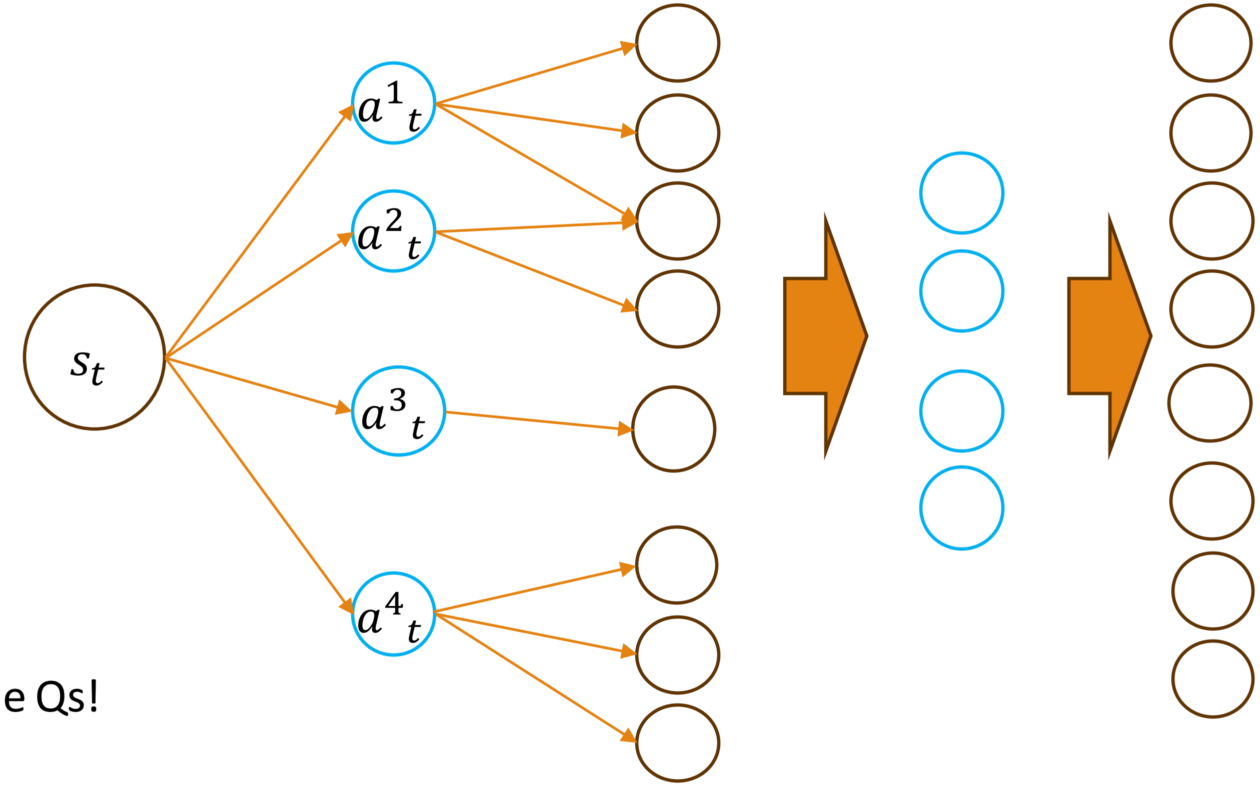
$$V^{\pi}(s_t) = \mathbb{E}_{\tau \sim p_{\pi}(\tau|s_t)}[G_t|s_t]$$

Q: "How Good is taking this action at this state"

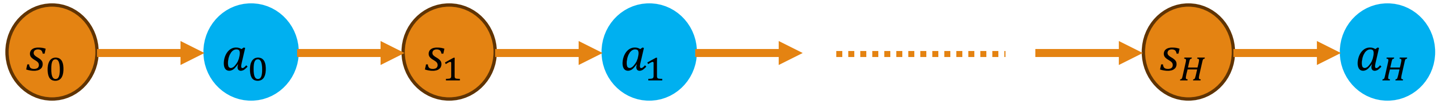
$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\tau \sim p_{\pi}(\tau|s_t, a_t)}[G_t|s_t, a_t]$$

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

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

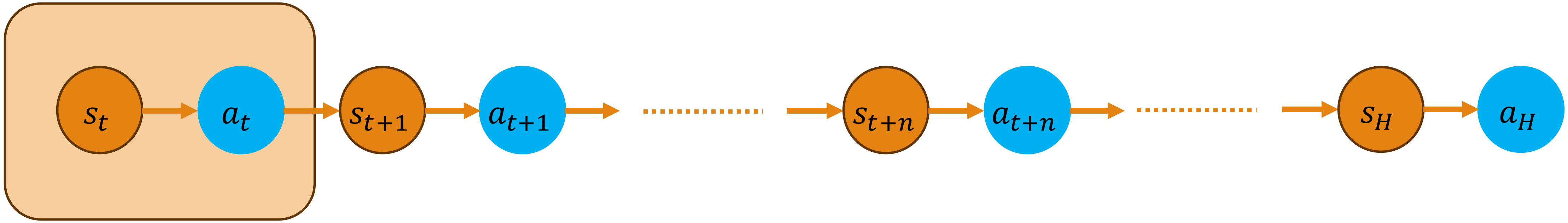


Trajectory Decomposition



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

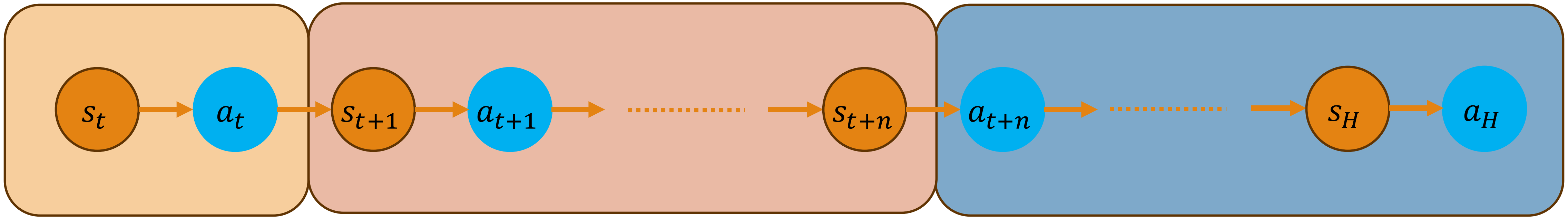
Given current state info



$$\tau = (\, s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \dots, s_{t+n}, a_{t+n}, \dots, s_H, a_H)$$

Trajectory Decomposition

Given current state info



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

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

$$\tau_{a_{t+n}:\, a_H} = (a_{t+n}, s_{t+n+1}, a_{t+n+1}, \ldots, s_H, a_H)$$

Return Decomposition

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

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

$$G_t = \sum_{k=t}^{H} \gamma^{k-t} r_k$$

Now consider the returns in form of:

$$G_{t:t+1} = r_t + \gamma r_{t+1}$$

$$G_{t:t+2} = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2}$$

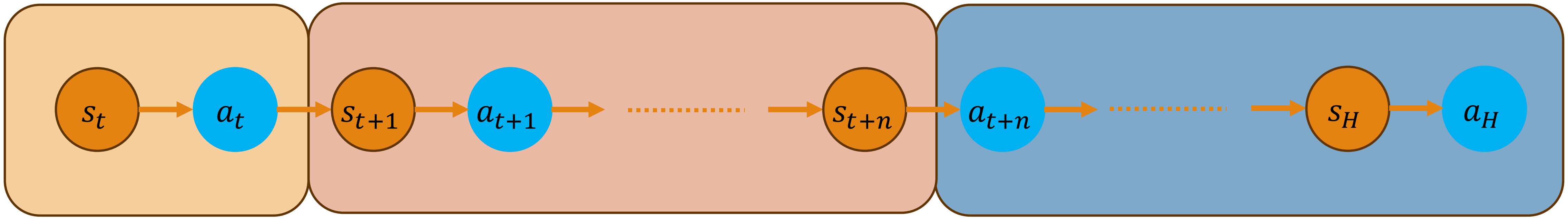
$$G_{t:t+3} = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3}$$

$$G_{t:t+n} = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \ldots + \gamma^n r_{t+n}$$

$$G_{t:t+n} = \sum_{k=t}^{t+n} \gamma^{k-t} r_k$$

Return Decomposition

Given current state info



$$\tau = (s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \dots, s_{t+n}, a_{t+n}, \dots, s_H, a_H)$$

$$G_{t:t+n-1} = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^{n-1} r_{t+n-1} = \sum_{k=t}^{t+n-1} \gamma^{k-t} r_k$$

$$G_{t+n:H} = \gamma^n r_{t+n} + \gamma^{n+1} r_{t+n+1} + \cdots + \gamma^{H-t} r_H = \sum_{k=t+n}^{H} \gamma^{k-t} r_k$$

Q Decomposition

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

$$Q^{\pi}(s_t, a_t) = Q_1 + Q_2$$

$$= \int_{\tau_{s_{t+1}:\, a_H}} G_{t:t+n-1} p(\tau_{s_{t+1}:\, a_H} | s_t, a_t) d\tau_{s_{t+1}:\, a_H} \Rightarrow Q_1$$

$$+ \int_{\tau_{s_{t+1}:\, a_H}} G_{t+n:H} p(\tau_{s_{t+1}:\, a_H} | s_t, a_t) d\tau_{s_{t+1}:\, a_H} \Rightarrow Q_2$$

Q Decomposition

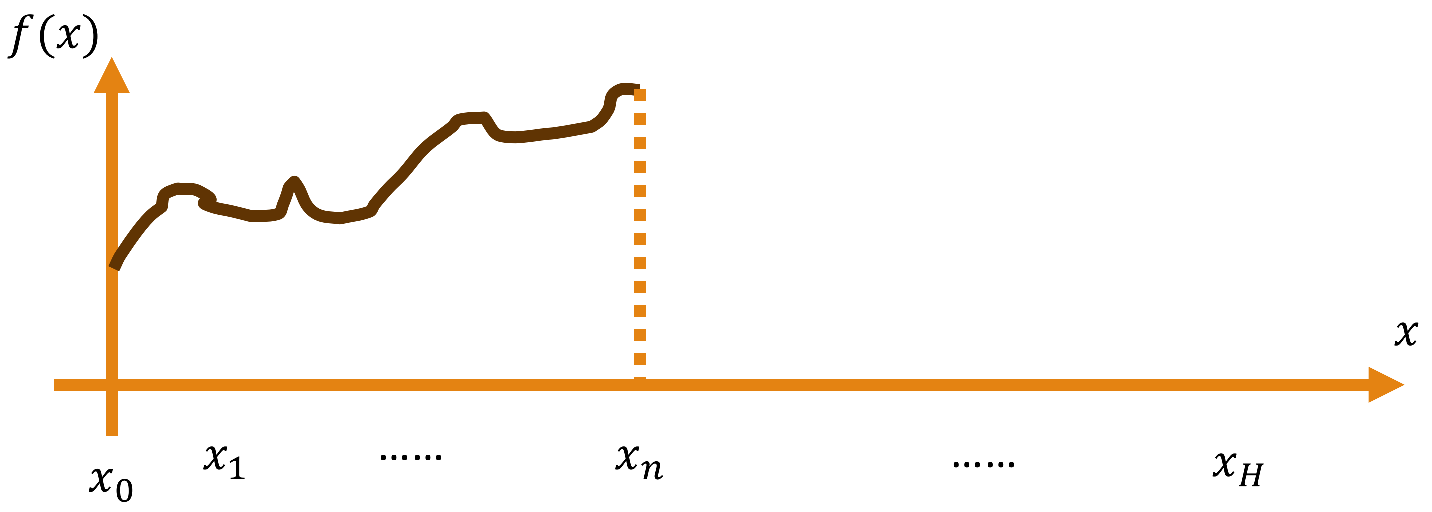
$$Q_1 = \int_{\tau_{s_{t+1}:\, a_H}} G_{t:t+n-1} p\left(\tau_{s_{t+1}:\, a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, a_H}$$

Chain Rule of Probability. Assuming a Markov Sequence.

$$p\left(\tau_{s_{t+1}:\, a_H} \middle| s_t, a_t\right) = p\left(\tau_{s_{t+1}:\, s_{t+n}}\ ,\ \tau_{a_{t+n}:\, a_H} \middle| s_t, a_t\right)$$

$$= p\left(\tau_{s_{t+1}:\, s_{t+n}} \middle| s_t, a_t\right) p\left(\tau_{a_{t+n}:\, a_H} \middle| s_{t+n}\right)$$

Integration Illustration



$$\int_{x_0}^{x_{0:H}} f(x_{0:H})dx_{0:H} = \int_{x_0}^{x_{0:n}} f(x_{0:n})dx_{0:n}$$

Q Decomposition

$$Q_1 = \int_{\tau_{s_{t+1}:\, a_H}} G_{t:t+n-1}\, p\left(\tau_{s_{t+1}:\, a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, a_H}$$

$$= \int_{\tau_{s_{t+1}:\, s_{t+n}}} G_{t:t+n-1}\, p\left(\tau_{s_{t+1}:\, s_{t+n}} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, s_{t+n}}$$

Q Decomposition

$$Q_2 = \int_{\tau_{s_{t+1}:a_H}} G_{t+n:H}\, p\left(\tau_{s_{t+1}:a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:a_H}$$

$$= \int_{\tau_{s_{t+1}:s_{t+n}}} \int_{\tau_{a_{t+n}:a_H}} \left[\sum_{k=t+n}^{H} \gamma^{k-t} r_k\right] p\left(\tau_{s_{t+1}:s_{t+n}} \middle| s_t, a_t\right) p\left(\tau_{a_{t+n}:a_H} \middle| s_{t+n}\right) d\tau_{s_{t+1}:s_{t+n}}\, d\tau_{a_{t+n}:a_H}$$

$$= \int_{\tau_{s_{t+1}:s_{t+n}}} \gamma^n \left[\int_{\tau_{a_{t+n}:a_H}} \left[\sum_{k=t+n}^{H} \gamma^{k-t-n} r_k\right] p\left(\tau_{a_{t+n}:a_H} \middle| s_{t+n}\right) d\tau_{a_{t+n}:a_H}\right] p\left(\tau_{s_{t+1}:s_{t+n}} \middle| s_t, a_t\right) d\tau_{s_{t+1}:s_{t+n}}$$

Q Decomposition

$$Q_2 = \int_{\tau_{s_{t+1}:s_{t+n}}} \gamma^n \left[\int_{\tau_{a_{t+n}:a_H}} \left[\sum_{k=t+n}^{H} \gamma^{k-t-n} r_k \right] p(\tau_{a_{t+n}:a_H} | s_{t+n}) d\tau_{a_{t+n}:a_H} \right] p(\tau_{s_{t+1}:s_{t+n}} | s_t, a_t) d\tau_{s_{t+1}:s_{t+n}}$$

Remember the definition of Value function?

$$\boxed{V^\pi(s_t) = \mathbb{E}_{\tau_{a_t:a_H} \sim p_\pi(\tau | s_t)}[G_t | s_t] = \int_{\tau_{a_{t+1}:a_H}} \left[\sum_{k=t}^{H} \gamma^{k-t} r_k \right] p(\tau_{a_{t+1}:a_H} | s_t) d\tau_{a_{t+1}:a_H}}$$

Q Decomposition

$$Q_2 = \int_{\tau_{s_{t+1}:s_{t+n}}} \gamma^n \underbrace{\left[\int_{\tau_{a_{t+n}:a_H}} \left[\sum_{k=t+n}^{H} \gamma^{k-t-n} r_k\right] p\left(\tau_{a_{t+n}:a_H} \middle| s_{t+n}\right) d\tau_{a_{t+n}:a_H}\right]}_{\boxed{V^{\pi}(s_{t+n})}} p\left(\tau_{s_{t+1}:s_{t+n}} \middle| s_t, a_t\right) d\tau_{s_{t+1}:s_{t+n}}$$

$$\boxed{V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:a_H} \sim p_{\pi}(\tau|s_t)}[G_t | s_t] = \int_{\tau_{a_{t+1}:a_H}} \left[\sum_{k=t}^{H} \gamma^{k-t} r_k\right] p\left(\tau_{a_{t+1}:a_H} \middle| s_t\right) d\tau_{a_{t+1}:a_H}}$$

Q Decomposition

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

$$Q^{\pi}(s_t, a_t) = Q_1 + Q_2$$

$$= \int_{\tau_{s_{t+1}:\, s_{t+n}}} G_{t:t+n-1} p(\tau_{s_{t+1}:\, s_{t+n}} | s_t, a_t) d\tau_{s_{t+1}:\, s_{t+n}} \Rightarrow Q_1$$

$$+ \int_{\tau_{s_{t+1}:\, s_{t+n}}} \gamma^n V^{\pi}(s_{t+n}) p(\tau_{s_{t+1}:\, s_{t+n}} | s_t, a_t) d\tau_{s_{t+1}:\, s_{t+n}} \Rightarrow Q_2$$

Q Decomposition – Before

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\tau \sim p_{\pi}(\tau|s_t,a_t)}[G_t | s_t, a_t] = \int_{\tau_{s_{t+1}:\,a_H}} G_t p\left(\tau_{s_{t+1}:\,a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\,a_H}$$

$$Q^{\pi}(s_t, a_t) = Q_1 + Q_2$$

$$= \int_{\boxed{\tau_{s_{t+1}:\,a_H}}} G_{t:t+n-1} p\left(\tau_{s_{t+1}:\,a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\,a_H} \Rightarrow Q_1$$

$$+ \int_{\boxed{\tau_{s_{t+1}:\,a_H}}} G_{t+n:H} p\left(\tau_{s_{t+1}:\,a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\,a_H} \Rightarrow Q_2$$

Q Decomposition – After

$$Q^\pi(s_t, a_t) = \mathbb{E}_{\tau \sim p_\pi(\tau | s_t, a_t)}[G_t | s_t, a_t] = \int_{\tau_{s_{t+1}:\, a_H}} G_t p\left(\tau_{s_{t+1}:\, a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, a_H}$$

$$Q^\pi(s_t, a_t) = Q_1 + Q_2$$

$$= \int_{\boxed{\tau_{s_{t+1}:\, s_{t+n}}}} G_{t:t+n-1} p\left(\tau_{s_{t+1}:\, s_{t+n}} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, s_{t+n}} \quad \Longrightarrow \quad Q_1$$

$$+ \int_{\boxed{\tau_{s_{t+1}:\, s_{t+n}}}} \gamma^n V^\pi(s_{t+n}) p\left(\tau_{s_{t+1}:\, s_{t+n}} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, s_{t+n}} \quad \Longrightarrow \quad Q_2$$

Bellman Equation – Q

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\tau \sim p_{\pi}(\tau | s_t, a_t)}[G_t | s_t, a_t] = \int_{\tau_{s_{t+1}:\, a_H}} G_t p\left(\tau_{s_{t+1}:\, a_H} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, a_H}$$

$$Q^{\pi}(s_t, a_t) = \int_{\tau_{s_{t+1}:\, s_{t+n}}} [G_{t:t+n-1} + \gamma^n V^{\pi}(s_{t+n})] p\left(\tau_{s_{t+1}:\, s_{t+n}} \middle| s_t, a_t\right) d\tau_{s_{t+1}:\, s_{t+n}}$$

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

Bellman Equation – Q

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

For $n = 1$

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim p_{\pi}(s_{t+1} | s_t, a_t)}[r_t + \gamma V^{\pi}(s_{t+1}) | s_t, a_t]$$

For $n = 2$

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\tau_{s_{t+1}:\, s_{t+2}} \sim p_{\pi}(\tau | s_t, a_t)}[r_t + \gamma r_{t+1} + \gamma^2 V^{\pi}(s_{t+2}) | s_t, a_t]$$

Bellman Equation – V

$$V^{\pi}(s_t) = \mathbb{E}_{k\sim\pi}[Q^{\pi}\left(s_t, {a^k}_t\right)] = \int_{a_t} \pi(a_t|s_t)Q^{\pi}(s_t, a_t)da_t$$

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

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

Bellman Equation – V

$$V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:\,s_{t+n}} \sim p_{\pi}(\tau|s_t,a_t)}[G_{t:t+n-1} + \gamma^n V^{\pi}(s_{t+n})|s_t]$$

For $n = 1$

$$V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi(a_t|s_t), s_{t+1} \sim p(s_{t+1}|s_t,a_t)}[r_t + \gamma V^{\pi}(s_{t+1})|s_t]$$

For $n = 2$

$$V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t:\,s_{t+2}} \sim p_{\pi}(\tau|s_t,a_t)}[r_t + \gamma r_{t+1} + \gamma^2 V^{\pi}(s_{t+2})|s_t]$$

Bellman Equation

Value for $n = 1$

$$V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi(a_t|s_t), s_{t+1} \sim p(s_{t+1}|s_t, a_t)}[r_t + \gamma V^{\pi}(s_{t+1})|s_t]$$

Q for $n = 1$

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim p_{\pi}(s_{t+1}|s_t, a_t)}[r_t + \gamma V^{\pi}(s_{t+1})|s_t, a_t]$$

??

Q for $n = 1$

$$Q^{\pi}(s_t, a_t) = r_t + \mathbb{E}_{a_t \sim \pi(a_t|s_t), s_{t+1} \sim p(s_{t+1}|s_t, a_t)}[\gamma Q^{\pi}(s_{t+1}, a_{t+1})|s_t, a_t]$$

Bellman Optimal Equation

Optimal Value / Action-Value is their value at optimal policy.

$$V^*(s_t) = \max_\pi V^\pi(s_t)$$

$$Q^*(s_t, a_t) = \max_\pi Q^\pi(s_t, a_t)$$

Bellman Optimal Equation

Optimal Value / Action-Value is their value at optimal policy.

$$V^*(s_t) = \max_{\pi} V^{\pi}(s_t)$$

Bellman Equation

$$\boxed{V^*(s_t) = \max_{a_t}\left[r + \mathbb{E}_{s_{t+1}\sim p_\pi(s_{t+1}|s_t,a_t)}[\gamma V^*(s_{t+1})]\right]}$$

$$Q^*(s_t, a_t) = \max_{\pi} Q^{\pi}(s_t, a_t)$$

Bellman Equation

$$\boxed{Q^*(s_t, a_t) = r + \mathbb{E}_{s_{t+1}\sim p_\pi(s_{t+1}|s_t,a_t)}[\gamma V^*(s_{t+1})]}$$

$$\boxed{V^*(s_t) = \max_{a_t} Q^*(s_t, a_t)}$$

Bellman Optimal Equation

Optimal Value / Action-Value is their value at optimal policy.

$$Q^*(s_t, a_t) = \max_{\pi} Q^{\pi}(s_t, a_t)$$

$$Q^*(s_t, a_t) = r + \mathbb{E}_{s_{t+1} \sim p_{\pi}(s_{t+1} | s_t, a_t)}[\gamma V^*(s_{t+1})]$$ 1

$$V^*(s_t) = \max_{a_t} Q^*(s_t, a_t)$$ 2

$$\boxed{Q^*(s_t, a_t) = r + \mathbb{E}_{s_{t+1} \sim p_{\pi}(s_{t+1} | s_t, a_t)}[\gamma \max_{a_t} Q^*(s_{t+1}, a_{t+1})]}$$ 1 + 2

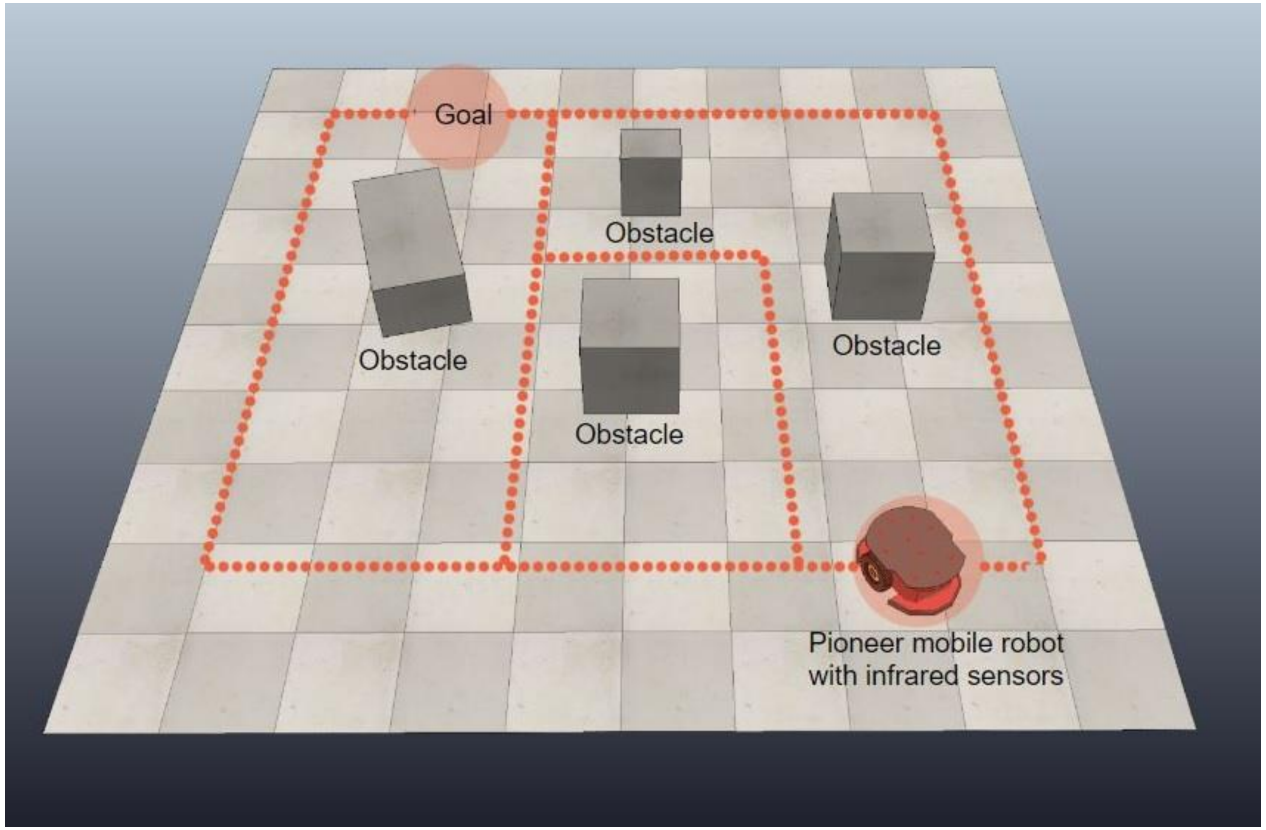
Reinforcement Learning Objective

Find Optimal Policy!!

A policy that selects an action with the highest Q value at all states.

$$\boxed{\pi^*(s_t) = \underset{a_t}{\operatorname{argmax}}\, Q^*(s_t, a_t)}$$

Robot Path Planning Example

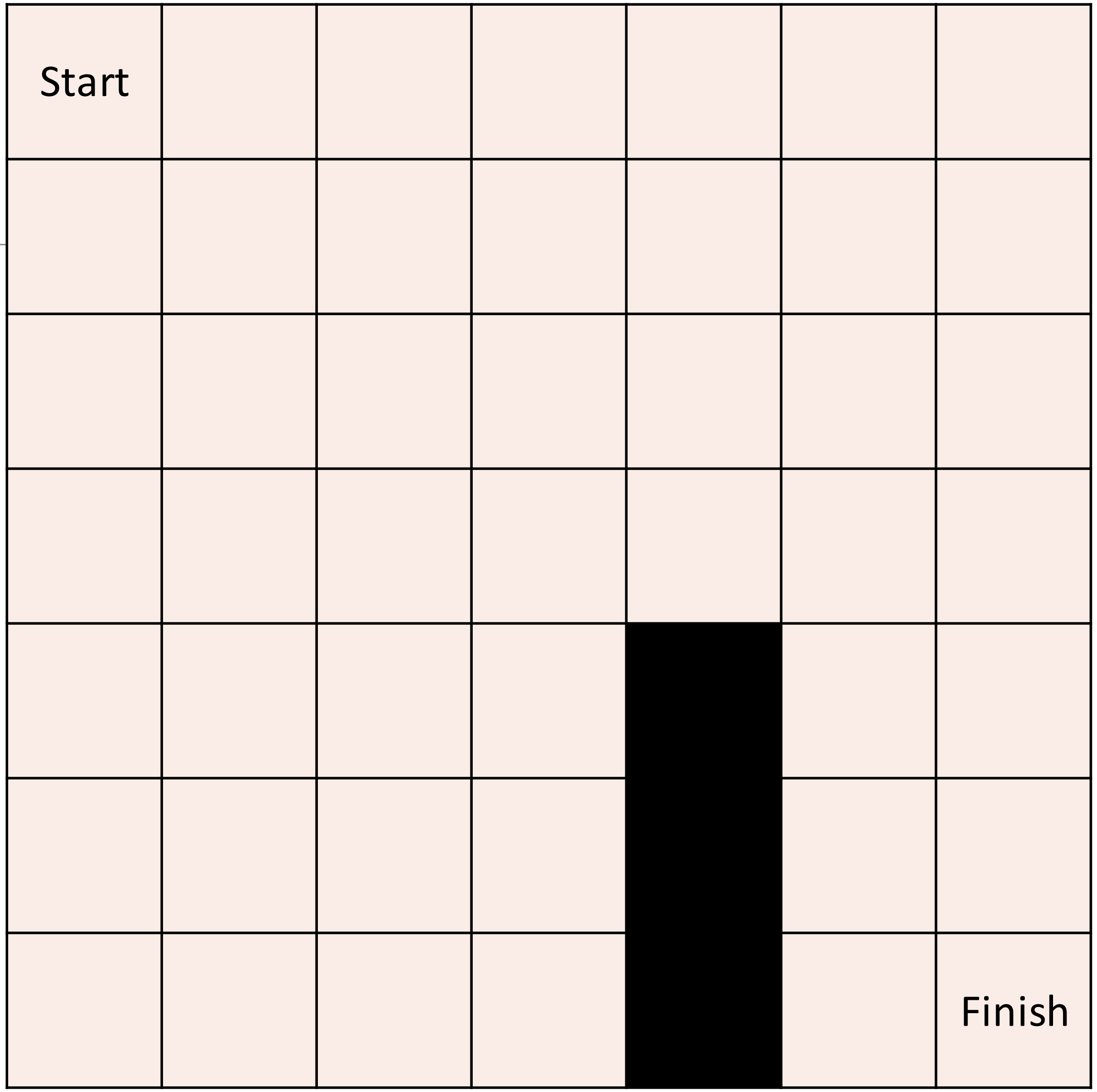


Maze Example

- An episode always Start in the Start cell.
- An episode terminates if the agent reaches the Final cell.

- Reward of -1 is given at every time step.

- The agent can move in all 4 directions.
- If the agent moves towards the borders of the maze or the wall, the agent stays in the current cell.



Maze Example

- Assume Optimal Policy, V, and Q values.
- Assume the discount rate is 1.0
- The optimal policy at the red cell is;

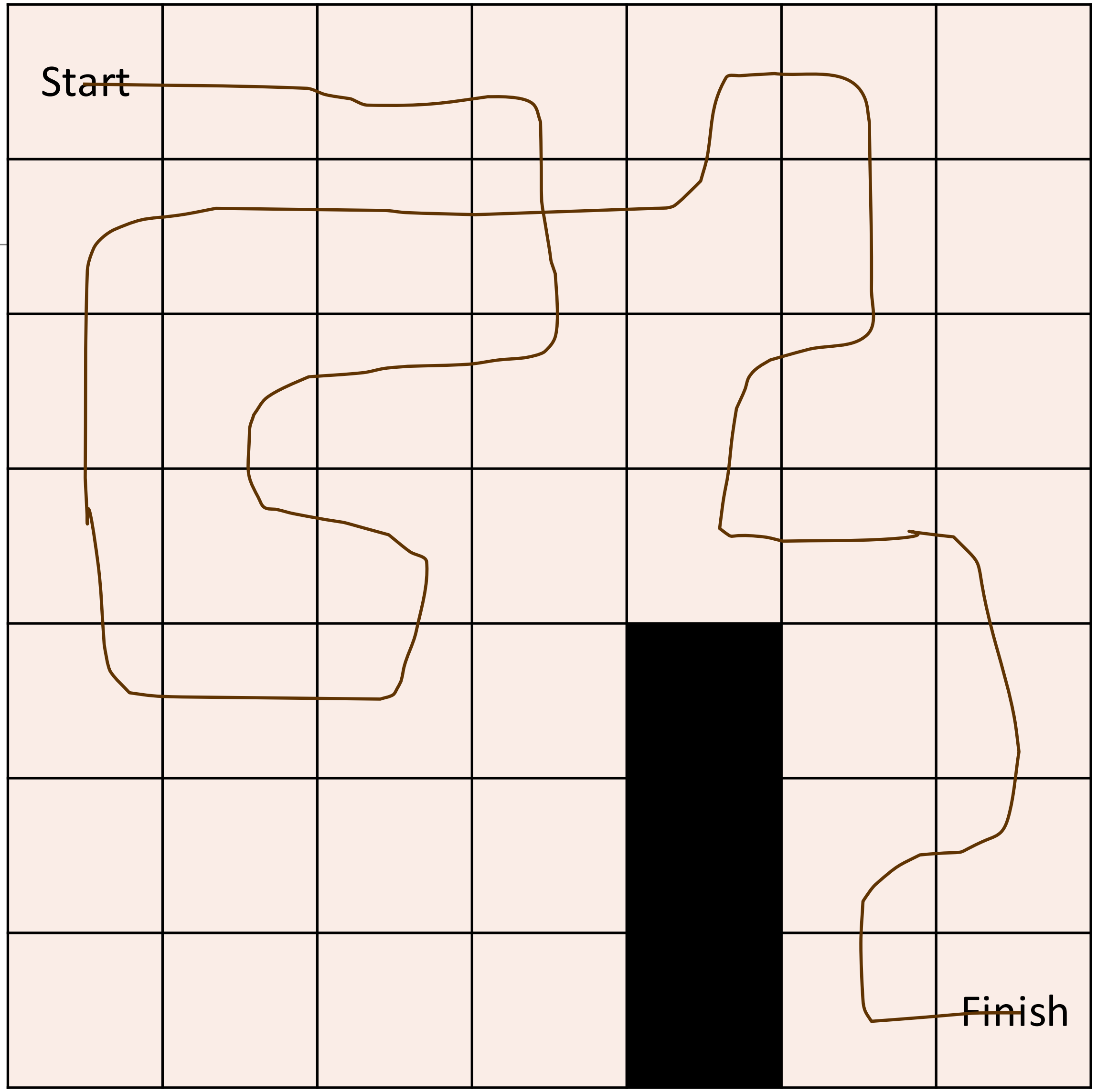
$$\pi^*(s_t) = \{\mathrm{UP: 0\%, Down: 50\%, LEFT: 0\%, RIGHT: 50\%}\}$$

Start	-11	-10	-9	-8	-7	-6
-11	-10	-9	-8	-7	-6	-5
-10	-9	-8	-7	-6	-5	-4
-9	-8	-7	-6	-5	-4	-3
-10	-9	-8	-7		-3	-2
-11	-10	-9	-8		-2	-1
-12	-11	-10	-9		-1	Finish

Maze Example

- Assume random policy.
- All actions are sampled with an equal probability of 25%

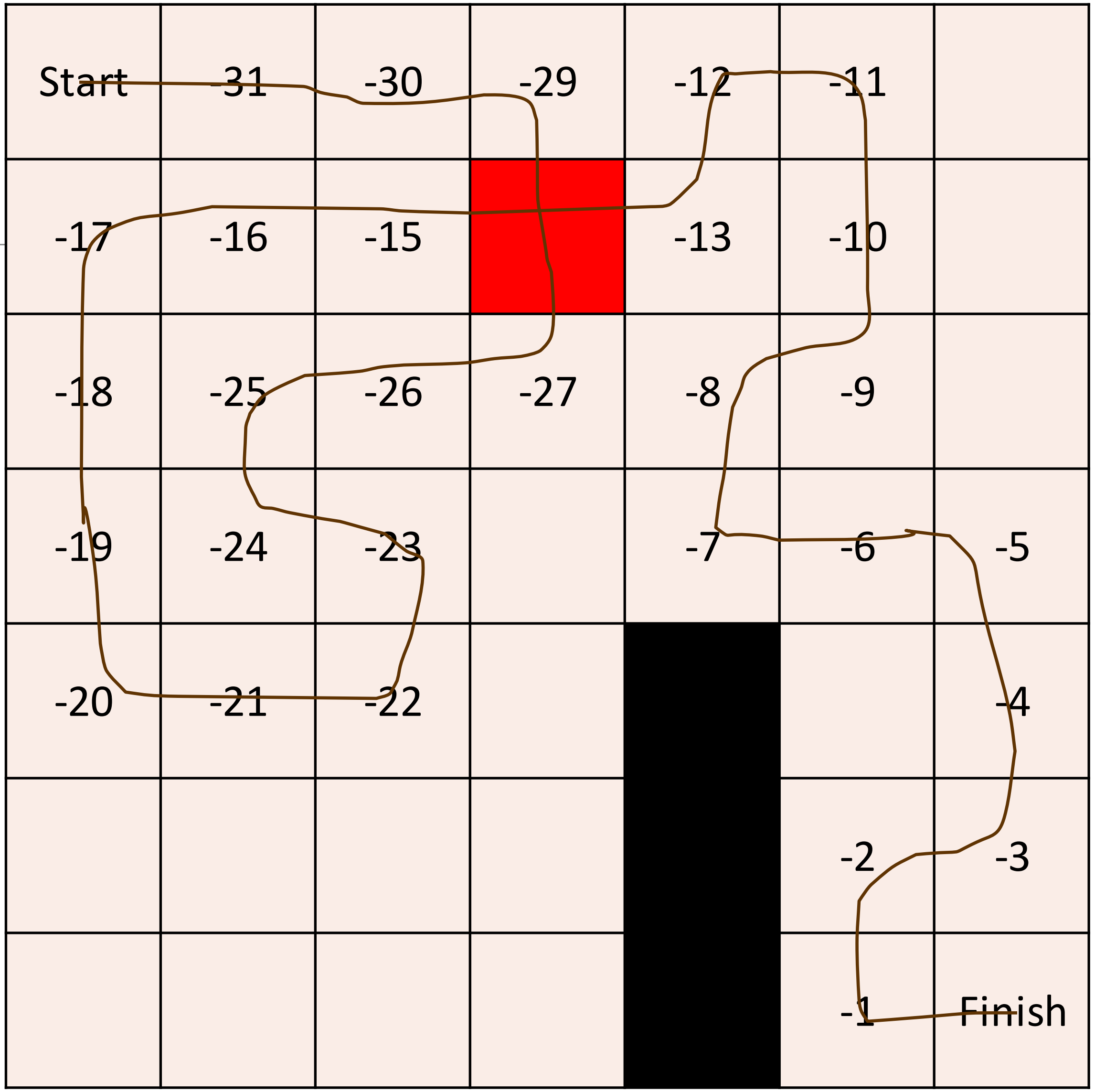
- Assume the discount rate is 1.0



Maze Example

- Assume random policy.
- All actions are sampled with an equal probability of 25%
- Assume the discount rate is 1.0
- How will the agent set the value for the red cell?

-14 or -28?

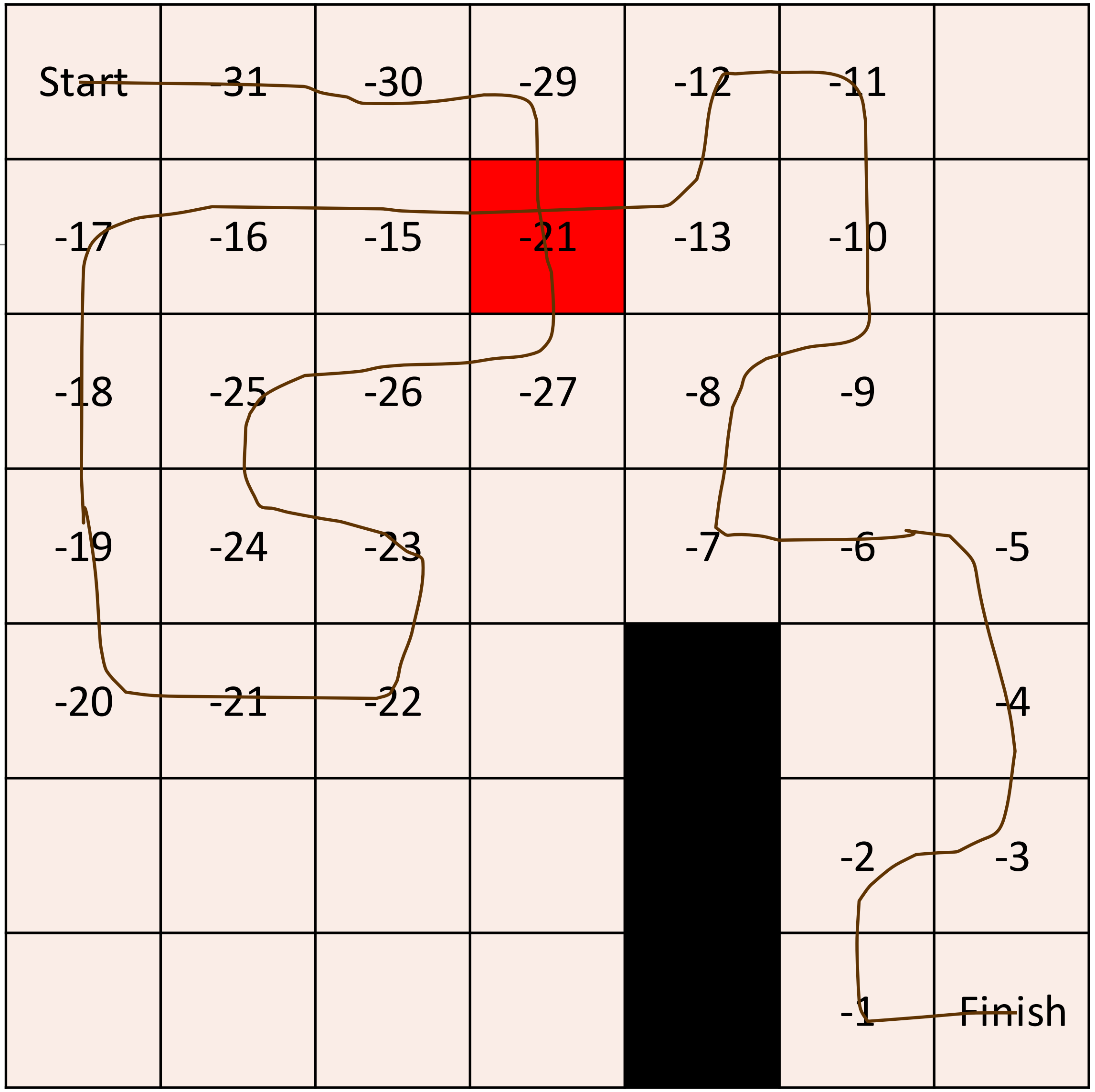


Maze Example

- Assume random policy.
- All actions are sampled with an equal probability of 25%
- Assume the discount rate is 1.0
- How will the agent set the value for the red cell?

-14 or -28?

ANS: At the same rollout stage, the agent usually takes an **average** of all experiences. But often takes the **highest** value!



Maze Example

- Assume random policy.
- All actions are sampled with an equal probability of 25%
- Assume the discount rate is 1.0
- Problem!!!
- The optimal policy now will oscillate!
- Solution??

Start	-31	-30	-29	-12	-11	
-17	-16	-15	-21	-13	-10	
-18	-25	-26	-27	-8	-9	
-19	-24	-23		-7	-6	-5
-20	-21	-22				-4
					-2	-3
					-1	Finish

Maze Example

- Assume random policy.
- All actions are sampled with an equal probability of 25%
- Assume the discount rate is 1.0
- Consider this case!
- The optimal solution now makes sense!
- How about the Empty cells?

Start	-31	-30	-29	-12	-11	
-17	-16	-15	-14	-13	-10	
-18	-25	-26	-27	-8	-9	
-19	-24	-23		-7	-6	-5
-20	-21	-22				-4
					-2	-3
					-1	Finish

Maze Example

- Assume random policy.
- All actions are sampled with an equal probability of 25%
- Assume the discount rate is 1.0
- Case where we initialize values to 0.
- The agent enters to state that it has not seen before!
- Solution?

Start	-31	-30	-29	-12	-11	0
-17	-16	-15	-14	-13	-10	0
-18	-25	-26	-27	-8	-9	0
-19	-24	-23	0	-7	-6	-5
-20	-21	-22	0		0	-4
0	0	0	0		-2	-3
0	0	0	0		-1	Finish

Maze Example

- Assume random policy.
- All actions are sampled with an equal probability of 25%

- Assume the discount rate is 1.0

- Case where we initialize values to the lowest value.

- Problem??

Start	-31	-30	-29	-12	-11	-99
-17	-16	-15	-14	-13	-10	-99
-18	-25	-26	-27	-8	-9	-99
-19	-24	-23	-99	-7	-6	-5
-20	-21	-22	-99		-99	-4
-99	-99	-99	-99		-2	-3
-99	-99	-99	-99		-1	Finish