

ROB311 Quiz 2

Hanhee Lee

March 13, 2025

Contents

1	Markov	2
1.1	General	2
1.1.1	Random Process	2
1.1.2	Markov Process	2
1.2	Markov Chains (MCs)	2
1.2.1	Bayesian Network	2
1.3	Markov Reward Processes (MRPs)	3
1.3.1	Bayesian Network	3
1.4	Markov Decision Processes (MDPs)	4
1.4.1	Setup	4
1.4.2	Bayesian Network	6
1.4.3	Intuition on Formulae	6
1.5	Canonical Examples	7
1.5.1	Markov Chains	7
1.5.2	Markov Reward Processes	7
1.5.3	Markov Decision Processes	8

Single-Agent Decision Problems: Fully-Observable Single-Agent Decision Algorithms

1 Markov

1.1 General

1.1.1 Random Process

Definition: Time-varying random variables S_0, S_1, S_2, \dots

1.1.2 Markov Process

Definition: Random process + depends on previous time step only (memoryless)

- w.l.o.g. states can contain history of previous states.

1.2 Markov Chains (MCs)

Summary: In a **Markov Chain**, we assume that:

- there are no agents
- state transitions occur automatically
- S_t is the state *after* transition t
- the state transition process is stochastic and memoryless:

$$S_t \perp S_0, \dots, S_{t-2} \mid S_{t-1}$$

- S_t is independent of all previous states given S_{t-1}

Name	Function:
initial state distribution	$p_0(s) := \mathbb{P}[S_0 = s]$
transition distribution	$p(s' s) := \mathbb{P}[S_{t+1} = s' \mid S_t = s]$
Prob. that state of the env. after T transitions is s	$p_T(s) := \mathbb{P}[S_T = s]$ $= \sum_{s'} p_{T-1}(s')p(s s')$
<ul style="list-style-type: none"> • $p_{T-1}(s')$: Prob. s' at $T-1$ (given) <ul style="list-style-type: none"> – $p_0(s)$: Base case • $p(s s')$: Prob. s given s' (from graph) 	

1.2.1 Bayesian Network

Notes: S_0, S_1, S_2, \dots form a **Bayesian Network**:

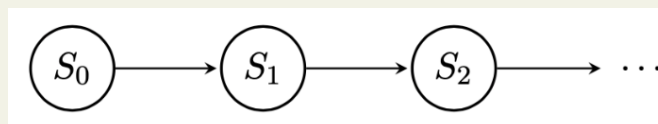


Figure 1

1.3 Markov Reward Processes (MRPs)

Summary: In a **Markov Reward Process**, we assume that:

- there is one agent
- state transitions occur automatically (i.e. agent has no control over actions)
- S_t is the state *after* transition t
- the state transition process is stochastic and memoryless:

$$S_t \perp S_0, \dots, S_{t-2} \mid S_{t-1}$$

- S_t is independent of all previous states given S_{t-1}
- R_t is the reward for transition t , i.e., $(S_{t-1}, \emptyset, S_t)$

Name	Function:
Initial state distribution	$p_0(s) := \mathbb{P}[S_0 = s]$
Transition distribution	$p(s' s) := \mathbb{P}[S_{t+1} = s' S_t = s]$
Reward function	$r(s, s') := \text{reward for transition } (s, \emptyset, s')$
Discount factor	$\gamma \in [0, 1]$
Return after T transitions	$U_T = \sum_{t=1}^T \gamma^{t-1} R_t$ $= U_{T-1} + \gamma^{T-1} R_T$ <ul style="list-style-type: none"> • i.e. The (possibly discounted) sum of the rewards after T transitions (sequence of rewards) • Why? <ul style="list-style-type: none"> – Future rewards are less valuable than immediate rewards. – Won't converge if sum goes to ∞ if $\gamma = 1$.
Expected return after T transitions	$\mathbb{E}[U_T] = \mathbb{E}[U_{T-1}] + \gamma^{T-1} \mathbb{E}[R_T]$ $= \mathbb{E}[U_{T-1}] + \gamma^{T-1} \sum_{s, s'} p_{T-1}(s) p(s' s) r(s, s')$ <ul style="list-style-type: none"> • $p_{T-1}(s)p(s' s)$: Prob. $s \rightarrow s'$ • $r(s, s')$: rwd $s \rightarrow s'$ • $\mathbb{E}[U_0] := 0$: Base case

1.3.1 Bayesian Network

Notes: $S_0, R_1, S_1, R_2, S_2, \dots$ form a **Bayesian Network**:

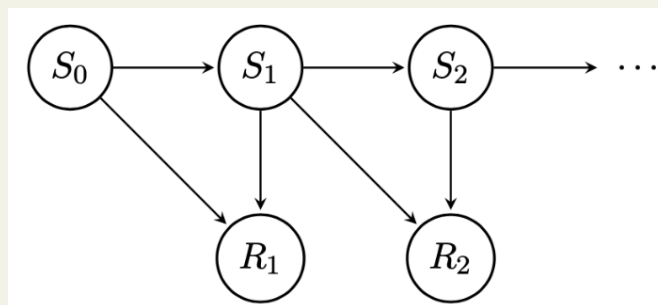


Figure 2

1.4 Markov Decision Processes (MDPs)

1.4.1 Setup

Summary: In a **Markov Decision Process (MDP)**, we assume that:

- there is one agent
- state transitions occur manually (after each action)
- S_t is the state *after* transition t
- A_t is the action inducing transition t
- the state transition process is stochastic and memoryless:

$$S_t \perp S_0, A_1, \dots, S_{t-2}, A_{t-1} \mid S_{t-1}, A_t$$

– S_t is independent of all previous states and actions given S_{t-1} and A_t

- R_t is the reward for transition t , i.e., (S_{t-1}, A_t, S_t)

Name	Function:
initial state distribution	$p_0(s) := \mathbb{P}[S_0 = s]$
transition distribution	$p(s' s, a) := \mathbb{P}[S_t = s' A_t = a, S_{t-1} = s]$
reward function	$r(s, a, s') := \text{reward for transition } (s, a, s')$
a time-invariant policy for choosing actions	$\pi(a s) := \mathbb{P}[A_t = a S_t = s]$
Maximum number of transitions	T_{\max}
<ul style="list-style-type: none"> • A Markov Decision Process can be either: <ul style="list-style-type: none"> – Finite: T_{\max} is finite – Infinite: T_{\max} is infinite * For infinite MDPs, we must have $\gamma < 1$. 	
Prob. that state of the env. after T transitions is s	$p_T(s) = \sum_{a, s'} p_{T-1}(s) \pi(a s') p(s s', a)$ <ul style="list-style-type: none"> • $p_{T-1}(s)$: Prob. s' at $T-1$ • $\pi(a s')$: Action a from s' • $p(s s', a)$: Prob. s given s', a
Expected return after T transitions	$\mathbb{E}_\pi[U_T] = \mathbb{E}_\pi[U_{T-1}] + \gamma^{T-1} \mathbb{E}_\pi[R_t]$ <ul style="list-style-type: none"> • $\mathbb{E}_\pi[R_t] = \sum_{s, a, s'} p_{T-1}(s) \pi(a s) p(s' s, a) r(s, a, s')$ • $\mathbb{E}_\pi[U_0] = 0$: Base case.
Future return after τ transitions	$G_\tau = \sum_{t=\tau+1}^T \gamma^{t-(\tau+1)} R_t$ $= R_{\tau+1} + \gamma G_{\tau+1}$ <ul style="list-style-type: none"> • Starting at $\tau + 1$ for the future return.
Expected future return after τ transitions given $S_\tau = s$	$\mathbb{E}_\pi[G_\tau \mid S_\tau = s] = \mathbb{E}_\pi[R_{\tau+1} \mid S_\tau = s] + \gamma \mathbb{E}_\pi[G_{\tau+1} \mid S_\tau = s]$ $= \sum_{a, s'} \pi(a s) p(s' s, a) (r(s, a, s') + \gamma \mathbb{E}_\pi[G_{\tau+1} \mid S_{\tau+1} = s'])$ <ul style="list-style-type: none"> • $\mathbb{E}_\pi[G_{T_{\max}} \mid S_{T_{\max}} = s] = 0$: Base case.

Summary:

Name	Function:
Value function	$v_\pi(s, T) := \mathbb{E}_\pi[G_{T_{\max}-T} \mid S_{T_{\max}-T} = s]$ $= \sum_{a, s'} \pi(a \mid s) p(s' \mid s, a) (r(s, a, s') + \gamma v_\pi(s', T-1))$ <ul style="list-style-type: none"> Value of state s under the policy π with T transitions remaining. <ul style="list-style-type: none"> i.e. How good the state is at time T (e.g. If $v(s, T) = 5$, then the expected future return at T is 5). $v(s, 0) = 0$ for all s: Base case
Optimal action	$a^*(s, T) = \arg \max_{a \in \mathcal{A}(s)} \sum_{s'} p(s' \mid s, a) (r(s, a, s') + \gamma v_{\pi^*}(s', T-1))$ $= \arg \max_{a \in \mathcal{A}(s)} q^*(s, a, T)$
Optimal policy	$\pi^*(a \mid s, T) = \arg \max_{\pi(a \mid s, T)} \mathbb{E}_\pi[G_\tau \mid S_\tau = s] = \begin{cases} 1 & \text{if } a = a^*(s, T) \\ 0 & \text{otherwise} \end{cases}$ <ul style="list-style-type: none"> Choose $\pi(\cdot \mid s)$ to maximize the expected future return after T transitions given $S_\tau = s$. Note: Policy always depends on transitions remaining so may omit.
Optimal value function	$v^*(s, T) = \max_a \sum_{s'} p(s' \mid a, s) (r(s, a, s') + \gamma v^*(s', \tau+1))$ <ul style="list-style-type: none"> Assume we use an optimal policy π^*. $v^*(s, 0) = 0$ for all s: Base case.
Q function (quality)	$q_\pi(s, a, T) := \mathbb{E}_\pi[G_{T_{\max}-T} \mid S_{T_{\max}-T} = s, A_{T_{\max}-(T-1)} = a]$ $= \sum_{s'} p(s' \mid s, a) \left(r(s, a, s') + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a', T-1) \right)$ <ul style="list-style-type: none"> Quality of move (s, a) under policy π with T transitions remaining. $q_\pi(s, a, 0) = 0$ for all s, a: Base case.
Optimal Q function	$q^*(s, a, T) = \sum_{s'} p(s' \mid s, a) \left(r(s, a, s') + \gamma \max_{a'} q^*(s', a', T-1) \right)$ <ul style="list-style-type: none"> $q^*(s, a, 0) = 0$ for all s, a: Base case.

1.4.2 Bayesian Network

Notes: $S_0, A_1, R_1, S_1, A_2, R_2, S_2, \dots$ form a **Bayesian Network**:

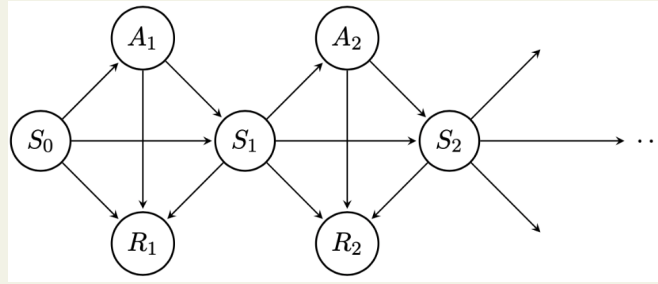


Figure 3

1.4.3 Intuition on Formulae

Notes:

$$\mathbb{E}_\pi[R_{\tau+1} \mid S_\tau = s] = \sum_{a, s'} \pi(a \mid s) p(s' \mid a, s) r(s, a, s')$$

- $\pi(a \mid s) p(s' \mid a, s)$: Prob. of getting to s' from s w/ action a
- $r(s, a, s')$: Reward of getting to s' from s w/ action a

$$\mathbb{E}_\pi[G_{\tau+1} \mid S_\tau = s] = \sum_{a, s'} \pi(a \mid s) p(s' \mid a, s) \mathbb{E}_\pi[G_{\tau+1} \mid S_{\tau+1} = s']$$

- $\pi(a \mid s) p(s' \mid a, s)$: Prob. of getting to s' from s w/ action a
- $\mathbb{E}_\pi[G_{\tau+1} \mid S_{\tau+1} = s']$: Expected future return at $\tau + 1$ from s' at $\tau + 1$.
- $\sum_{a, s'} \tau$: Sum over all possible future states and current actions to get expected future return at $\tau + 1$ from s at τ .

1.5 Canonical Examples

1.5.1 Markov Chains

Example:

1. **Given:** Caveman needs to predict the weather, W , which is either sunny or rainy. Suppose the weather tomorrow depends on the weather today:

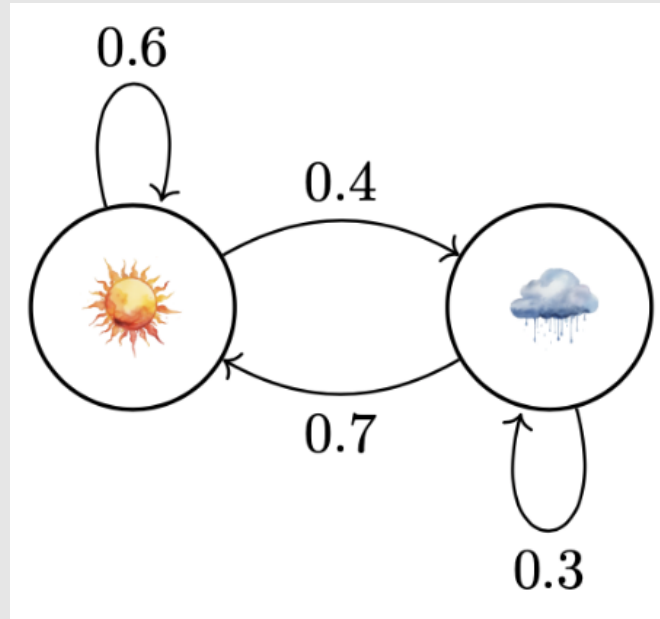


Figure 4

2. **Problem:** Caveman wants to predict the weather on a given day.

1.5.2 Markov Reward Processes

Example:

1. **Given:** Caveman needs to predict the weather, W , which is either sunny or rainy. Suppose the weather tomorrow depends on the weather today:

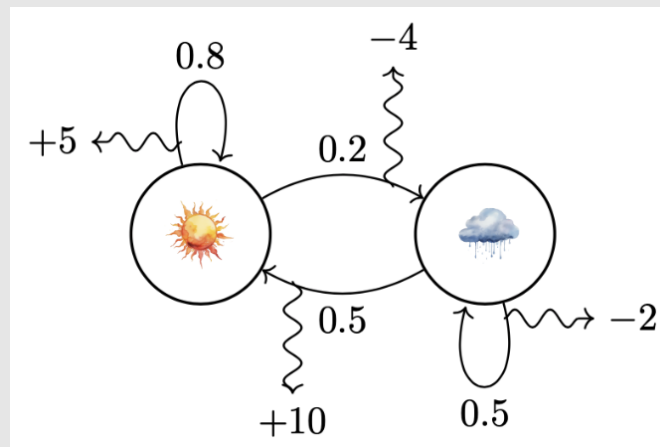


Figure 5

- Depending on the transition, caveman may feel happier/sadder. This is quantified w/ the rewards.
2. **Problem:** Caveman wants to predict the weather on a given day that maximizes his happiness.

1.5.3 Markov Decision Processes

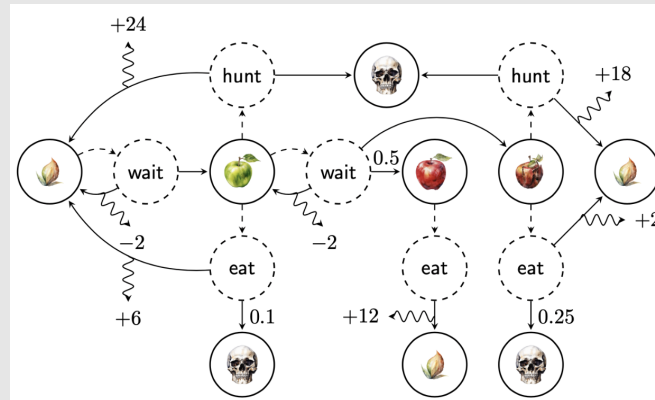
Example:1. **Given:**

Figure 6

- Solid straight line: Outcome of action a from state s .
 - Dotted straight line: Choice of action (policy) from state s .
 - If policy known, then reduced to MRP.
 - Squiggly line: Reward for action a from state s to state s' .
 - Assume uniform probability.
 - Since $\sum p = 1$, therefore count # of arrows going out of s and divide by 1 to get p .
 - Same states have the same connections (i.e. all can use them just to hard to draw)
2. **Problem:** Find the optimal policy for $\gamma = 1$ and $T_{\max} = 5$.
3. **Soln:**

Warning:

- Be careful with the problems. Verify the answers. Go up to at least 2 steps since that tests everything.
- Be able to go through the formula quickly.
- 1st question on the quiz.

Example:

T	s	a	$q^*(s, a, T) = \sum_{s'} p(s' s, a) \left(r(s, a, s') + \gamma \max_{a'} q^*(s', a', T - 1) \right)$
0	-	-	0
<ul style="list-style-type: none"> Best Action: $a^*(s, 0) = \text{NA}$ 			
1	seed	wait	$q^*(\text{seed}, \text{wait}, 1) = \underbrace{0.5(-2 + 0)}_{s' = \text{seed}} + \underbrace{0.5(0 + 0)}_{s' = \text{ga}} = -1$
<ul style="list-style-type: none"> Best Action: $a^*(\text{seed}, 1) = \text{wait}$ 			
1	ga	wait	$q^*(\text{ga}, \text{wait}, 1) = \underbrace{0.25(-2 + 0)}_{s' = \text{ga}} + \underbrace{0.5(0 + 0)}_{s' = \text{rea}} + \underbrace{0.25(0 + 0)}_{s' = \text{roa}} = -0.5$
1	ga	eat	$q^*(\text{ga}, \text{eat}, 1) = \underbrace{0.1(0 + 0)}_{s' = \text{dead}} + \underbrace{0.9(6 + 0)}_{s' = \text{seed}} = 5.4$
1	ga	hunt	$q^*(\text{ga}, \text{hunt}, 1) = \underbrace{0.5(24 + 0)}_{s' = \text{seed}} + \underbrace{0.5(0 + 0)}_{s' = \text{dead}} = 12$
<ul style="list-style-type: none"> Best Action: $a^*(\text{ga}, 1) = \text{hunt}$ 			
1	rea	eat	$q^*(\text{rea}, \text{eat}, 1) = \underbrace{1(12 + 0)}_{s' = \text{seed}} = 12$
<ul style="list-style-type: none"> Best Action: $a^*(\text{rea}, 1) = \text{eat}$ 			
1	roa	eat	$q^*(\text{roa}, \text{eat}, 1) = \underbrace{0.25(0 + 0)}_{s' = \text{dead}} + \underbrace{0.75(2 + 0)}_{s' = \text{seed}} = 1.5$
1	roa	hunt	$q^*(\text{roa}, \text{hunt}, 1) = \underbrace{0.5(0 + 0)}_{s' = \text{dead}} + \underbrace{0.5(18 + 0)}_{s' = \text{seed}} = 9$
<ul style="list-style-type: none"> Best Action: $a^*(\text{roa}, 1) = \text{hunt}$ 			
1	dead	-	$q^*(\text{dead}, -, 1) = \underbrace{1(0 + 0)}_{s' = \text{end}} = 0$
<ul style="list-style-type: none"> Best Action: $a^*(s, 1) = -$ 			
<ul style="list-style-type: none"> Optimal Policy w/ 1 Transition Remaining: $\pi^*(a s, 1) = \begin{cases} 1 & \text{if } a = a^*(s, 1) \\ 0 & \text{otherwise} \end{cases}$ 			

Example:

T	s	a	$q^*(s, a, T) = \sum_{s'} p(s' s, a) \left(r(s, a, s') + \gamma \max_{a'} q^*(s', a', T - 1) \right)$
2	seed	wait	$q^*(\text{seed}, \text{wait}, 2) = \underbrace{0.5(-2 - 1)}_{s'=\text{seed}} + \underbrace{0.5(0 + 12)}_{s'=\text{ga}} = 4.5$
• Best Action: $a^*(\text{seed}, 2) = \text{wait}$			
2	ga	wait	$q^*(\text{ga}, \text{wait}, 2) = \underbrace{0.25(-2 + 12)}_{s'=\text{ga}} + \underbrace{0.5(0 + 12)}_{s'=\text{rea}} + \underbrace{0.25(0 + 9)}_{s'=\text{roa}} = 10.75$
2	ga	eat	$q^*(\text{ga}, \text{eat}, 2) = \underbrace{0.1(0 + 0)}_{s'=\text{dead}} + \underbrace{0.9(6 - 1)}_{s'=\text{seed}} = 4.5$
2	ga	hunt	$q^*(\text{ga}, \text{hunt}, 2) = \underbrace{0.5(24 - 1)}_{s'=\text{seed}} + \underbrace{0.5(0 + 0)}_{s'=\text{dead}} = 11.5$
• Best Action: $a^*(\text{ga}, 2) = \text{hunt}$			
2	rea	eat	$q^*(\text{rea}, \text{eat}, 2) = \underbrace{1(12 - 1)}_{s'=\text{seed}} = 11$
• Best Action: $a^*(\text{rea}, 2) = \text{eat}$			
2	roa	eat	$q^*(\text{roa}, \text{eat}, 2) = \underbrace{0.25(0 + 0)}_{s'=\text{dead}} + \underbrace{0.75(2 - 1)}_{s'=\text{seed}} = 0.75$
2	roa	hunt	$q^*(\text{roa}, \text{hunt}, 2) = \underbrace{0.5(0 + 0)}_{s'=\text{dead}} + \underbrace{0.5(18 - 1)}_{s'=\text{seed}} = 8.5$
• Best Action: $a^*(\text{roa}, 2) = \text{hunt}$			
2	dead	-	$q^*(\text{dead}, -, 2) = \underbrace{1(0 + 0)}_{s'=\text{end}} = 0$
• Best Action: $a^*(s, 2) = -$			
• Optimal Policy w/ 2 Transitions Remaining: $\pi^*(a s, 2) = \begin{cases} 1 & \text{if } a = a^*(s, 2) \\ 0 & \text{otherwise} \end{cases}$			

Example:

$$T \quad s \quad a \quad q^*(s, a, T) = \sum_{s'} p(s' | s, a) \left(r(s, a, s') + \gamma \max_{a'} q^*(s', a', T-1) \right)$$

$$3 \quad \text{seed} \quad \text{wait} \quad q^*(\text{seed}, \text{wait}, 3) = \underbrace{0.5(-2 + 4.5)}_{s'=\text{seed}} + \underbrace{0.5(0 + 11.5)}_{s'=\text{ga}} = 7$$

- Best Action: $a^*(\text{seed}, 3) = \text{wait}$

$$3 \quad \text{ga} \quad \text{wait} \quad q^*(\text{ga}, \text{wait}, 3) = \underbrace{0.25(-2 + 11.5)}_{s'=\text{ga}} + \underbrace{0.5(0 + 11)}_{s'=\text{rea}} + \underbrace{0.25(0 + 8.5)}_{s'=\text{roa}} = 10$$

$$3 \quad \text{ga} \quad \text{eat} \quad q^*(\text{ga}, \text{eat}, 3) = \underbrace{0.1(0 + 0)}_{s'=\text{dead}} + \underbrace{0.9(6 + 4.5)}_{s'=\text{seed}} = 9.45$$

$$3 \quad \text{ga} \quad \text{hunt} \quad q^*(\text{ga}, \text{hunt}, 3) = \underbrace{0.5(24 + 4.5)}_{s'=\text{seed}} + \underbrace{0.5(0 + 0)}_{s'=\text{dead}} = 14.25$$

- Best Action: $a^*(\text{ga}, 3) = \text{hunt}$

$$3 \quad \text{rea} \quad \text{eat} \quad q^*(\text{rea}, \text{eat}, 3) = \underbrace{1(12 + 4.5)}_{s'=\text{seed}} = 16.5$$

- Best Action: $a^*(\text{rea}, 3) = \text{eat}$

$$3 \quad \text{roa} \quad \text{eat} \quad q^*(\text{roa}, \text{eat}, 3) = \underbrace{0.25(0 + 0)}_{s'=\text{dead}} + \underbrace{0.75(2 + 4.5)}_{s'=\text{seed}} = 4.875$$

$$3 \quad \text{roa} \quad \text{hunt} \quad q^*(\text{roa}, \text{hunt}, 3) = \underbrace{0.5(0 + 0)}_{s'=\text{dead}} + \underbrace{0.5(18 + 4.5)}_{s'=\text{seed}} = 11.25$$

- Best Action: $a^*(\text{roa}, 3) = \text{hunt}$

$$3 \quad \text{dead} \quad - \quad q^*(\text{dead}, -, 3) = \underbrace{1(0 + 0)}_{s'=\text{end}} = 0$$

- Best Action: $a^*(s, 3) = -$

- Optimal Policy w/ 3 Transitions Remaining: $\pi^*(a | s, 3) = \begin{cases} 1 & \text{if } a = a^*(s, 3) \\ 0 & \text{otherwise} \end{cases}$

Example:

$$T \quad s \quad a \quad q^*(s, a, T) = \sum_{s'} p(s' | s, a) \left(r(s, a, s') + \gamma \max_{a'} q^*(s', a', T-1) \right)$$

$$4 \quad \text{seed} \quad \text{wait} \quad q^*(\text{seed}, \text{wait}, 4) = \underbrace{0.5(-2 + 7)}_{s'=\text{seed}} + \underbrace{0.5(0 + 14.25)}_{s'=\text{ga}} = 9.625$$

- Best Action: $a^*(\text{seed}, 4) = \text{wait}$

$$4 \quad \text{ga} \quad \text{wait} \quad q^*(\text{ga}, \text{wait}, 4) = \underbrace{0.25(-2 + 14.25)}_{s'=\text{ga}} + \underbrace{0.5(0 + 16.5)}_{s'=\text{rea}} + \underbrace{0.25(0 + 11.25)}_{s'=\text{roa}} = 14.125$$

$$4 \quad \text{ga} \quad \text{eat} \quad q^*(\text{ga}, \text{eat}, 4) = \underbrace{0.1(0 + 0)}_{s'=\text{dead}} + \underbrace{0.9(6 + 7)}_{s'=\text{seed}} = 11.7$$

$$4 \quad \text{ga} \quad \text{hunt} \quad q^*(\text{ga}, \text{hunt}, 4) = \underbrace{0.5(24 + 7)}_{s'=\text{seed}} + \underbrace{0.5(0 + 0)}_{s'=\text{dead}} = 15.5$$

- Best Action: $a^*(\text{ga}, 4) = \text{hunt}$

$$4 \quad \text{rea} \quad \text{eat} \quad q^*(\text{rea}, \text{eat}, 4) = \underbrace{1(12 + 7)}_{s'=\text{seed}} = 19$$

- Best Action: $a^*(\text{rea}, 4) = \text{eat}$

$$4 \quad \text{roa} \quad \text{eat} \quad q^*(\text{roa}, \text{eat}, 4) = \underbrace{0.25(0 + 0)}_{s'=\text{dead}} + \underbrace{0.75(2 + 7)}_{s'=\text{seed}} = 6.75$$

$$4 \quad \text{roa} \quad \text{hunt} \quad q^*(\text{roa}, \text{hunt}, 4) = \underbrace{0.5(0 + 0)}_{s'=\text{dead}} + \underbrace{0.5(18 + 7)}_{s'=\text{seed}} = 12.5$$

- Best Action: $a^*(\text{roa}, 4) = \text{hunt}$

$$4 \quad \text{dead} \quad - \quad q^*(\text{dead}, -, 4) = \underbrace{1(0 + 0)}_{s'=\text{end}} = 0$$

- Best Action: $a^*(s, 4) = -$

- Optimal Policy w/ 4 Transitions Remaining: $\pi^*(a | s, 4) = \begin{cases} 1 & \text{if } a = a^*(s, 4) \\ 0 & \text{otherwise} \end{cases}$

Example:

T	s	a	$q^*(s, a, T) = \sum_{s'} p(s' s, a) \left(r(s, a, s') + \gamma \max_{a'} q^*(s', a', T-1) \right)$
5	seed	wait	$q^*(\text{seed}, \text{wait}, 5) = \underbrace{0.5(-2 + 9.625)}_{s'=\text{seed}} + \underbrace{0.5(0 + 15.5)}_{s'=\text{ga}} = 11.5625$
• Best Action: $a^*(\text{seed}, 5) = \text{wait}$			
5	ga	wait	$q^*(\text{ga}, \text{wait}, 5) = \underbrace{0.25(-2 + 15.5)}_{s'=\text{ga}} + \underbrace{0.5(0 + 19)}_{s'=\text{rea}} + \underbrace{0.25(0 + 12.5)}_{s'=\text{roa}} = 16$
5	ga	eat	$q^*(\text{ga}, \text{eat}, 5) = \underbrace{0.1(0 + 0)}_{s'=\text{dead}} + \underbrace{0.9(6 + 9.625)}_{s'=\text{seed}} = 14.0625$
5	ga	hunt	$q^*(\text{ga}, \text{hunt}, 5) = \underbrace{0.5(24 + 9.625)}_{s'=\text{seed}} + \underbrace{0.5(0 + 0)}_{s'=\text{dead}} = 16.8125$
• Best Action: $a^*(\text{ga}, 5) = \text{hunt}$			
5	rea	eat	$q^*(\text{rea}, \text{eat}, 5) = \underbrace{1(12 + 9.625)}_{s'=\text{seed}} = 21.625$
• Best Action: $a^*(\text{rea}, 5) = \text{eat}$			
5	roa	eat	$q^*(\text{roa}, \text{eat}, 5) = \underbrace{0.25(0 + 0)}_{s'=\text{dead}} + \underbrace{0.75(2 + 9.625)}_{s'=\text{seed}} = 8.71875$
5	roa	hunt	$q^*(\text{roa}, \text{hunt}, 5) = \underbrace{0.5(0 + 0)}_{s'=\text{dead}} + \underbrace{0.5(18 + 9.625)}_{s'=\text{seed}} = 13.8125$
• Best Action: $a^*(\text{roa}, 5) = \text{hunt}$			
5	dead	-	$q^*(\text{dead}, -, 5) = \underbrace{1(0 + 0)}_{s'=\text{end}} = 0$
• Best Action: $a^*(s, 5) = -$			
• Optimal Policy w/ 5 Transitions Remaining: $\pi^*(a s, 5) = \begin{cases} 1 & \text{if } a = a^*(s, 5) \\ 0 & \text{otherwise} \end{cases}$			

s	a	$q^*(s, a, 0)$	$q^*(s, a, 1)$	$q^*(s, a, 2)$	$q^*(s, a, 3)$	$a^*(s, 2)$	$\pi^*(a \mid s, 2)$
empty	hunt	0	2	2	2	seed	0
empty	seed	0	0	1	2.1		1
<ul style="list-style-type: none"> • $q^*(\text{empty}, \text{hunt}, 1) = 0.5(4 + 0) + 0.5(0 + 0) = 2$ • $q^*(\text{empty}, \text{seed}, 1) = 1(0 + 0) = 0$ • $q^*(\text{empty}, \text{hunt}, 2) = 0.5(4 + 0) + 0.5(0 + 0) = 2$ • $q^*(\text{empty}, \text{seed}, 2) = 1(0 + 1 \cdot 1) = 1$ • $q^*(\text{empty}, \text{hunt}, 3) = \underbrace{0.5(4 + 0)}_{s'=\text{full}} + \underbrace{0.5(0 + 0)}_{s'=\text{dead}} = 2$ • $q^*(\text{empty}, \text{seed}, 3) = \underbrace{1(0 + 1 \cdot 2.1)}_{s'=\text{ga}} = 2.1$ 							
ga	grab	0	1	1	1	clock	0
ga	clock	0	0	2.1	2.1		1
<ul style="list-style-type: none"> • $q^*(\text{ga}, \text{grab}, 1) = 1(1 + 0) = 1$ • $q^*(\text{ga}, \text{clock}, 1) = 0.7(0 + 0) + 0.3(0 + 0) = 0$ • $q^*(\text{ga}, \text{grab}, 2) = 1(1 + 0) = 1$ • $q^*(\text{ga}, \text{clock}, 2) = 0.7(0 + 3) + 0.3(0 + 0) = 2.1$ • $q^*(\text{ga}, \text{grab}, 3) = \underbrace{1(1 + 0)}_{s'=\text{full}} = 1$ • $q^*(\text{ga}, \text{clock}, 3) = \underbrace{0.7(0 + 3)}_{s'=\text{ra}} + \underbrace{0.3(0 + 0)}_{s'=\text{roa}} = 2.1$ 							
dead	-	0	0	0	0	-	1
<ul style="list-style-type: none"> • $q^*(\text{dead}, -, 1) = 1(0 + 0) = 0$ • $q^*(\text{dead}, -, 2) = 1(0 + 0) = 0$ • $q^*(\text{dead}, -, 3) = 1(0 + 0) = 0$ 							
full	-	0	0	0	0	-	1
<ul style="list-style-type: none"> • $q^*(\text{full}, -, 1) = 1(0 + 0) = 0$ • $q^*(\text{full}, -, 2) = 1(0 + 0) = 0$ • $q^*(\text{full}, -, 3) = 1(0 + 0) = 0$ 							
ra	grab	0	3	3	3	grab	1
<ul style="list-style-type: none"> • $q^*(\text{ra}, \text{grab}, 1) = 1(3 + 0) = 3$ • $q^*(\text{ra}, \text{grab}, 2) = 1(3 + 0) = 3$ • $q^*(\text{ra}, \text{grab}, 3) = \underbrace{1(3 + 0)}_{s'=\text{full}} = 3$ 							
roa	grab	0	0	0	0	grab	1
<ul style="list-style-type: none"> • $q^*(\text{roa}, \text{grab}, 1) = 1(0 + 0) = 0$ • $q^*(\text{roa}, \text{grab}, 2) = 1(0 + 0) = 0$ • $q^*(\text{roa}, \text{grab}, 3) = \underbrace{1(0 + 0)}_{s'=\text{dead}} = 0$ 							