Markov Decision Processes

aka MDPs

Markov Processes

- Discrete time : $k = 0, 1, 2 \ldots$
- States : $\mathcal{S}$

$\mathcal{S}$ = { Living room, kitchen, Bedroom ... }

$\mathcal{S}$ = { A, B, C, D, E, ... L }

$\mathcal{S}$ = { (1,1), (1,2), ... (4,4) }

- $$T : \mathcal{S} \times \mathcal{S} \rightarrow \underbrace{[0, 1]}_{\text{probability}}$$

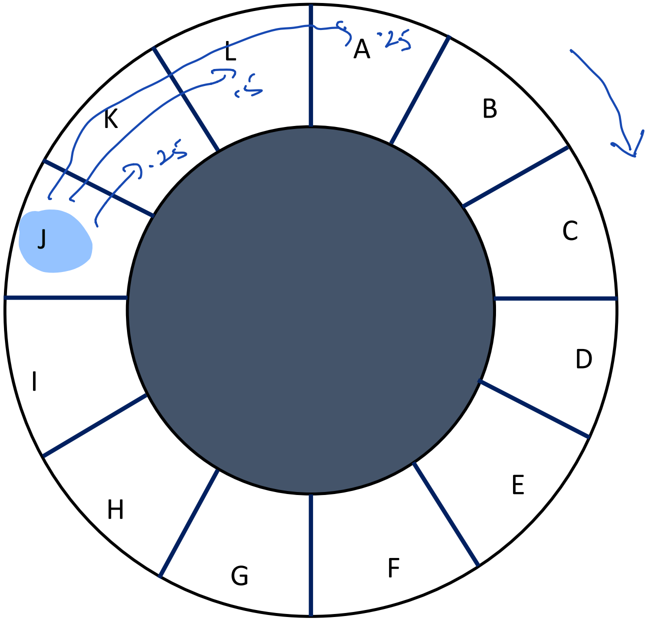
conditioning bar

$$T(s_k, s_{k+1}) = \text{Prob}\{\underline{\text{state } k+1 = s_{k+1}} \;|\; \underline{\text{State } k = s_k}\}$$

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

1,1	1,2	1,3	1,4
2,1			
			4,4

Example: A robot called Sisyphus



robot moves clockwise by d_k steps at stage k.

Let $\begin{cases} P\{d_k = 1\} = 0.25 \\ P\{d_k = 2\} = 0.5 \\ P\{d_k = 3\} = 0.25 \end{cases}$

from this:

$T(A,A) = 0$, $T(A,B) = 0.25$

$T(A,C) = 0.5$ . . .

$T(S_k, S_{k+1})$ can be represented as a table.

S_{k+1}

S_k

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	0.25	0.5	0.25	0	0	0	0	0	0	0	0
B	0	0	0.25	0.5	0.25	0	0	0	0	0	0	0
C	0	0	0	0.25	0.5	0.25	0	0	0	0	0	0
D	0	0	0	0	0.25	0.5	0.25	0	0	0	0	0
E	0	0	0	0	0	0.25	0.5	0.25	0	0	0	0
F	0	0	0	0	0	0	0.25	0.5	0.25	0	0	0
G	0	0	0	0	0	0	0	0.25	0.5	0.25	0	0
H	0	0	0	0	0	0	0	0	0.25	0.5	0.25	0
I	0	0	0	0	0	0	0	0	0	0.25	0.5	0.25
J	0.25	0	0	0	0	0	0	0	0	0	0.25	0.5
K	0.5	0.25	0	0	0	0	0	0	0	0	0	0.25
L	0.25	0.5	0.25	0	0	0	0	0	0	0	0	0

Note

Table does not change as time passes.

Markov Property

At time step k, $T(S_k, S_{k+1})$ is independent of anything that occurs prior to time k.

$$\rightarrow P\{S_{k+1} \mid S_0, S_1, \cdots, S_k\} = P\{S_{k+1} \mid S_k\}$$ ← Markov Property

Example

$$P\{S_3 = E \mid S_0 = A, S_1 = C, \underline{S_2 = D}\} = P\{S_3 = E \mid S_2 = D\}$$

$$P\{S_3 = E \mid S_0 = A, S_1 = B, \underline{S_2 = D}\} = P\{S_3 = E \mid S_2 = D\}$$

Markov *Decision* Processes

Let's give Sisyphus some Free Will:

$$\underbrace{\begin{array}{l}\text{Move Right (counter-clockwise)} \\ \text{Move Left (clockwise)}\end{array}}_{\text{at stage } k\text{, choose an action}}$$

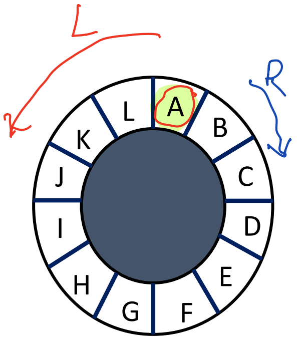
Assume symmetry

$$\text{For } L: \quad P\{d_k\} = \begin{cases} .25 & d_k = 1 \\ .5 & d_k = 2 \\ .25 & d_k = 3 \end{cases} \quad \text{For } R$$

$$T_L(A, B) = 0.25$$

$$T_R(B, A) = 0.25$$

By choose R or L, Sisyphus choose which State transition Matrix is operational!



Start in State A

$T_L(S_k, S_{k+1})$

k\k+1	A	B	C	D	E	F	G	H	I	J	K	L
A	0	0.25	0.5	0.25	0	0	0	0	0	0	0	0
B	0	0	0.25	0.5	0.25	0	0	0	0	0	0	0
C	0	0	0	0.25	0.5	0.25	0	0	0	0	0	0
D	0	0	0	0	0.25	0.5	0.25	0	0	0	0	0
E	0	0	0	0	0	0.25	0.5	0.25	0	0	0	0
F	0	0	0	0	0	0	0.25	0.5	0.25	0	0	0
G	0	0	0	0	0	0	0	0.25	0.5	0.25	0	0
H	0	0	0	0	0	0	0	0	0.25	0.5	0.25	0
I	0	0	0	0	0	0	0	0	0	0.25	0.5	0.25
J	0.25	0	0	0	0	0	0	0	0	0	0.25	0.5
K	0.5	0.25	0	0	0	0	0	0	0	0	0	0.25
L	0.25	0.5	0.25	0	0	0	0	0	0	0	0	0

$T_R(S_k, S_{k+1})$

k\k+1	A	B	C	D	E	F	G	H	I	J	K	L
A	0	0	0	0	0	0	0	0	0	.25	.5	.25
B												
C												
D												
E												
F												
G												
H												
I												
J												
K												
L												

Rewards

Assign a reward value to each state.

$$R : S \to \mathbb{R}$$

Example: power station in state E.

$R(E) = +1$ (charge up)

$R(s) = -0.2 \quad s \neq E$ (waste power)

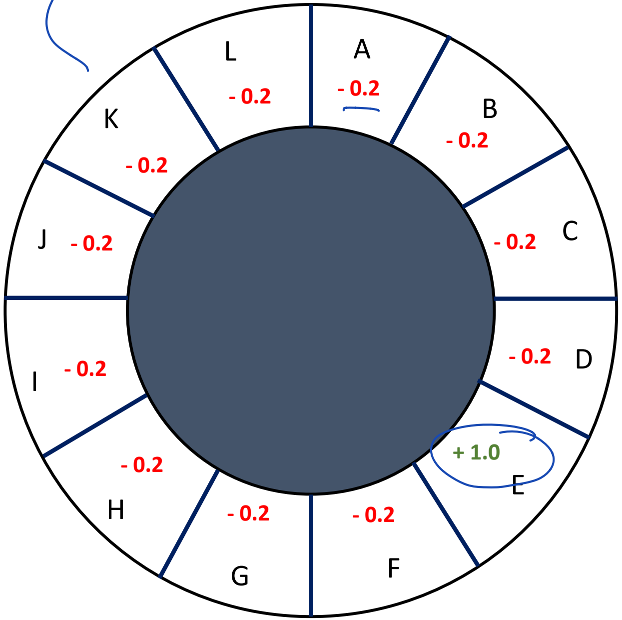
<u>Define</u> h-stage return for a sequence $(s_0, \cdots s_h)$

$$r_h(s_0, s_1, \cdots s_h) = \sum_{i=0}^{h} R(s_i)$$

$r_2(A,B,C) = R(A) + R(B) + R(C)$

$= -0.6$

→ Note: This sequence is deterministic.



Markov Decision Processes

aka MDPs

Part 2: Expectation

Expectation

Suppose a random variable, X, takes values from a set $\{c_1, c_2, \ldots, c_n\}$, $c_i \in \mathbb{R}$, $i = 1, \ldots, n$, the expected value of X is

$$\boxed{E[X]} = \sum_{i=0}^{n} c_i \, P\{X = c_i\}$$

Example roll a die, X = # of dots on top face

$X \in \{1, 2, 3, 4, 5, 6\}$, $P\{X = i\} = 1/6$ for $i = 1, 2, 3, 4, 5, 6$,

$$E[X] = \sum_{i=1}^{6} \frac{1}{6} \times i = 3.5$$

Intuition

~~If~~ we roll a die many times, the average # of dots will tend to $E[X]$, i.e., 3.5.

Expected 1-stage return

$$E[R(S_k)] = \sum_{s \in S} R(s)\, P(s)$$

1-stage return from $S_0 = D$, action $a_1 = L$.

$$E[R(S_1) \mid \underbrace{S_0 = D, a_1 = L}_{\text{explicit conditions}}]$$

$$= R(E)\, P\{S_1 = E \mid S_0 = D, a_1 = L\}$$

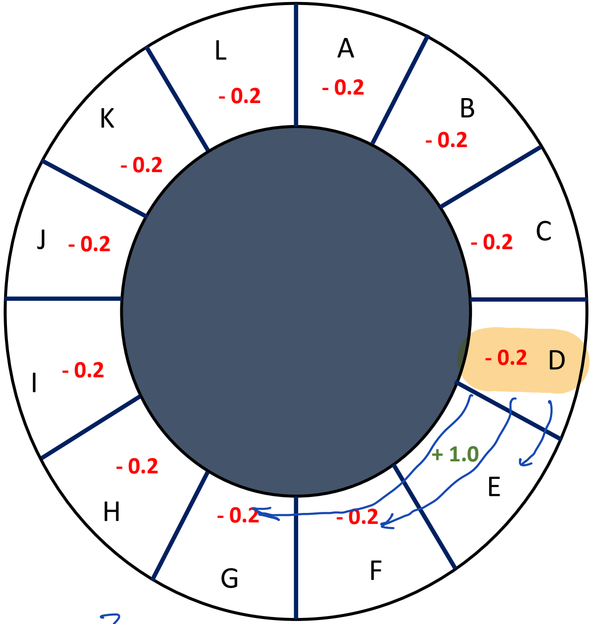
1.0×0.25

$$+ R(F)\, P\{S_1 = F \mid S_0 = D, a_1 = L\}$$

$+ -0.2 \times 0.5$

$$+ R(G)\, P\{S_1 = G \mid S_0 = D, a_1 = L\}$$

$+ -0.2 \times 0.25 = .1$



Generalizing to expected h-stage return

Suppose $s_0 = A$, $a_1 = L$, $a_2 = L$.

Compute $E\left[\underline{r_2(s_0, s_1, s_2)} \mid s_0 = A, a_1 = L, a_2 = L\right]$

$= \underline{R(A)} + E\left[R(s_1) + R(s_2) \mid s_0 = A, a_1 = L, a_2 = L\right]$

$\underline{R(A)} + \sum_{\substack{s_1 \in S \\ s_2 \in S}} \underbrace{\left(R(s_1) + R(s_2)\right)}_{1} P\{s_2, s_1 \mid s_0 = A, a_1 = L, a_2 = L\}$

Def of conditional probability

$\underline{P\{s_2, s_1 \mid s_0 = A, a_1 = L, a_2 = L\}} = P\{s_2 \mid s_1, s_0 = A, a_1 = L, a_2 = L\} \, P\{s_1 \mid s_0 = A, a_1 = L, a_2 = L\}$

$= \underbrace{\underbrace{P\{s_2 \mid s_1, a_2 = L\}}_{\text{in table}} \; \underbrace{P\{s_1 \mid s_0 = A, a_1 = L\}}_{\text{table}}}_{2}$

For each possible sequence

1. compute specific return
2. compute prob. of sequence
 - Tabulate results

Example: Expected reward for two "Left" actions, starting from state A

Sequence	d_1, d_2	Probability	Reward
ABC	1,1	0.25×0.25	-0.6
ABD	1,2	0.25×0.5	-0.6
ABE	1,3	0.25×0.25	$+0.6$
ACD	2,1		
ACE	2,2		
ACF	2,3		
ADE	3,1		
ADF	3,2		
ADG	3,3		

$S_0 = A, S_1 = B, S_2 = C$

Example: Expected reward for two "Left" actions, starting from state A

Sequence	d_1, d_2	Probability	Reward
ABC	1,1	$0.25 \times 0.25 = 0.0625$	-0.6
ABD	1,2	$0.25 \times 0.5 = 0.125$	-0.6
ABE	1,3	$0.25 \times 0.25 = 0.0625$	$+0.6$
ACD	2,1	$0.5 \times 0.25 = 0.125$	-0.6
ACE	2,2	$0.5 \times 0.5 = 0.25$	$+0.6$
ACF	2,3	$0.5 \times 0.25 = 0.125$	-0.6
ADE	3,1	$0.25 \times 0.25 = 0.0625$	$+0.6$
ADF	3,2	$0.25 \times 0.5 = 0.125$	-0.6
ADG	3,3	$0.25 \times 0.25 = 0.0625$	-0.6

$$
\begin{aligned}
E[R_0 + R_1 + R_2] = {} & (0.0625 \times -0.6) + (0.125 \times -0.6) + (0.0625 \times 0.6) \\
& + (0.125 \times -0.6) + (0.25 \times 0.6) + (0.125 \times -0.6) \\
& + (0.0625 \times 0.6) + (0.125 \times -0.6) + (0.0625 \times -0.6) = \mathbf{-0.225}
\end{aligned}
$$

Discounted Reward

Suppose Sisyphus runs forever... $E[r_\infty] \rightsquigarrow \pm\infty$

Discounted Reward:

$$r_h = \sum_{i=0}^{h} \gamma^i R(s_i) \quad \text{for } 0 < \gamma < 1$$

$\to \gamma$ = discount factor

as $h \to \infty$

$$\lim_{h \to \infty} r_h = \sum_{i=0}^{\infty} \gamma^i R(s_i) \leq \sum_{i=0}^{\infty} \gamma^i R_{max} = \frac{R_{max}}{1-\gamma}$$

because $\sum \gamma^i = \frac{1}{1-\gamma}$ for $0 < \gamma < 1$.

To make decisions, we'll use Expected discounted reward

$$\underbrace{E[r_h] = E\left[\sum_{i=0}^{h} \gamma^i R(s_i) \,\middle|\, a_1, a_2, \dots a_h\right]}_{\text{use this for decision-making}}$$

Probability of a Sequence

Use the definition of conditional probability for this: $\boldsymbol{P(x,y) = P(x|y)P(y)}$ ***[See example on next slide]***

This relationship holds for arbitrary conditioning events, as long as all terms are conditioned on the same event:

$$P(x,y|\,ANYTHING\,) = P(x|y, ANYTHING)P(y|ANYTHING)$$

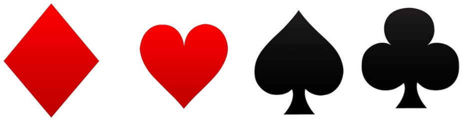
For a sequence of actions executed from an initial state, we have

$$P\{s_2, s_1 \mid S_0 = A, a_1 = L, a_2 = L\} = P\{s_2, \mid s_1, S_0 = A, a_1 = L, a_2 = L\}P\{s_1 \mid S_0 = A, a_1 = L, a_2 = L\}$$

And applying the Markov property (i.e., the transition from $\mathrm{k} = 1$ to $k = 2$ does not depend on history) we obtain:

$$P\{s_2, s_1 \mid S_0 = A, a_1 = L, a_2 = L\} = P\{s_2, \mid s_1, a_2 = L\}P\{s_1 \mid S_0 = A, a_1 = L\}$$

Example of Joint/Conditional Probability: What is the probability of drawing at random a red ace?



Four suits:

- ***Hearts, Diamonds, Clubs, Spades***

Each suit has 13 cards

- ***Ace, King, Queen, Jack, 10, … 2***

<u>Two Possible Strategies:</u>

- Directly compute the probability by counting:

$$P(red, ace) = \frac{\#\, red\ aces}{\#\, of\ cards} = \frac{2}{52} = \frac{1}{26}$$

- Use joint/conditional probability relationship:

$$P(red, ace) = P(ace\,|red)\ P(red) = \frac{2}{26} \times \frac{1}{2} = \frac{1}{26}$$

Markov Decision Processes

aka MDPs

Part 3: Policies, and the Value Function

Policies and Expected Return under policy π

$$E[r_h] = E\left[\sum_{i=0}^{h} \gamma^i R(s_i) \,\middle|\, a_1, a_2, \ldots a_h\right]$$

Def A policy $\pi : S \to A$, where A = set of actions s.t.
$\pi(s) \to a$, a = action to be taken from/in state s.

Def $V^\pi(s)$ = expected return for executing policy π from state s.

$$V^\pi(s) = E[r_\infty(s) \mid \pi]$$

$$= E\left[\sum_{i=0}^{\infty} \gamma^i R(s_i) \,\middle|\, \pi, s_0 = s\right]$$

↳ at $i=0$, nothing random - initial state pull out $i=0$ term

$s = s_0$

$$= R(s) + E\left[\sum_{i=1}^{\infty} \gamma^i R(s_i) \,\middle|\, \pi\right]$$

$$= R(s) + \gamma E\left[\sum_{i=1}^{\infty} \gamma^{i-1} R(s_i) \,\middle|\, \pi\right]$$

Factor out γ^1, a constant

$$\Downarrow$$

$$\boxed{\text{Let } j = i-1 \implies j+1 = i}$$

$$\Downarrow$$

$$= R(s) + \gamma \underbrace{E\left[\sum_{j=0}^{\infty} \gamma^j R(s_{j+1}) \,\middle|\, \pi\right]}_{\text{Expected return under } \pi \text{ from state } s_{j+1}}$$

$T_a(s, s')$

Notation

$T(s, a, s')$ is transition probability for executing action a in state s & arriving to state s'

$$= R(s) + \gamma E\left[V^{\pi}(s') \,\middle|\, \pi\right]$$

we don't know this value

$$= R(s) + \gamma \sum_{s' \in S'} T(s, \pi(s), s') V^{\pi}(s')$$

an action, chosen by policy π.

All possible next states from s by executing $\pi(s)$

Optimal policies and the Value Function

Let π^* denote the optimal policy, $\pi^* = \arg\max_{\pi} V^{\pi}(s)$

Def The value function V^* $(= V^{\pi^*})$ gives the maximum expected future return for each state s:

$$V^* : S \longrightarrow \mathbb{R}$$

Given V^*, it's simple to compute the optimal action from state s:

Optimal policies and the Value Function (cont)

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \sum_{s' \in S} \underbrace{T(s,a,s')}_{\text{prob of next state under action } a} \underbrace{V^*(s')}_{\text{Value fn for next state}}$$

Thus, V^* satisfies

$$\boxed{V^*(s) = R(s) + \gamma \max_{a \in \mathcal{A}} \sum_{s'} T(s,a,s')\, V^*(s')}$$

$s \in \{A, B, C, \cdots L\}$

$\Leftrightarrow$ Bellman Equation

The Bellman Equation

— Richard Bellman

If we have N_s states, then for each $s \in \mathcal{S}$ we construct a specific instance of Bellman Eqn.

IF Bellman eqn were linear, we would be done ⟶ merely solve linear system.

<u>But</u> Bellman is not linear... <u>max</u>

Recap...

An MDP is defined by:

S = set of states

A = set of actions

$T: S \times A \times S \to [0,1]$

$R: S \to \mathbb{R}$

γ : discount factor (maybe)

Problem Find $\pi^* = \arg\max_{\pi} \boxed{E\left[\sum_{k=0}^{\infty} \gamma^k R(s_k) \mid \pi\right]}$

↳ Transform to Bellman

Value Iteration

$$V^*(s) = R(s) + \gamma \max_a \sum T(s,a,s') V^*(s) \leftarrow \text{Truth}$$

Define V^k as approximation to V^* at k^{th} iteration. $V^0(s) =$ arbitrary or $R(s)$

Idea V^{k+1} improves estimate V^k, and $V^k \xrightarrow[k \to \infty]{} V^*$

$$V^{k+1}(s) = \underbrace{R(s)}_{\text{Truth}} + \underbrace{\gamma \max_a \sum_{s'} T(s,a,s') \underbrace{V^k(s)}_{\text{Best guess at iteration } k}}_{\text{Best guess at } k \text{ for exp. future return under optimal action}}$$

Value Iteration (cont)

Example: Expected reward for two "Left" actions, starting from state A

State	V^0	V^1	V^2
A	1	0.3	-0.05
B	1	0.3	-0.05
C	1	0.3	0.1
D	1	0.3	0.25
E	1	1.5	1.15
F	1	0.3	0.1
G	1	0.3	0.25
H	1	0.3	-0.05
I	1	0.3	-0.05
J	1	0.3	-0.05
K	1	0.3	-0.05
l	1	0.3	-0.05

$$V^1(s) = R(s) + 0.5 \max_{\mathrm{a}} \sum T(s,a,s')V^0(s)$$

$a \in \{L, R\}$

Move left Move right

$$V^1(s) = R(s) + 0.5 \max\{0.25 \times 1 + 0.5 \times 1 + 0.25 \times 1, 0.25 \times 1 + 0.5 \times 1 + 0.25 \times 1\}$$

For $s \in \{A, B, C, D, F, G, H, I, J, K, L\}$

$$V^1(s) = -0.2 + 0.5(1) = 0.3$$

For $s = E$

$$V^1(s) = 1.0 + 0.5(1) = 1.5$$

$\boxed{\text{Initial Guess } V^0(s) = 1 \text{ for all } s}$