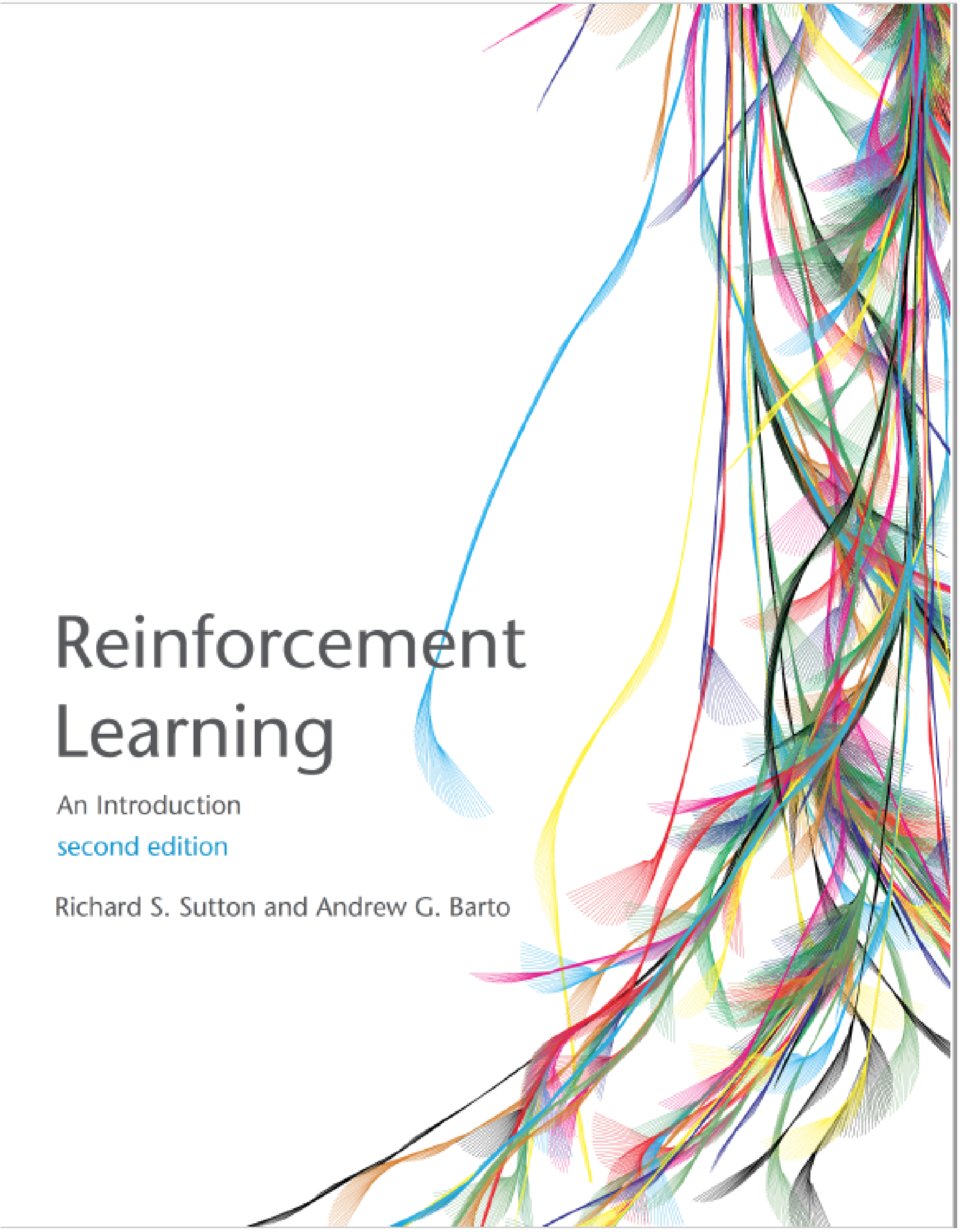
Markov Decision Process

OR: confusion via math!


Reinforcement
Learning
An Introduction
second edition
Richard S. Sutton and Andrew G. Barto

Chapter 3

Markov Chain

no actions, no rewards

Markov Reward Porcess

rewards

Markov Decision Process

actions + rewards

Markov Decision Process

actions + rewards + observations

Markov Chain

Markov Property:
next state only dependent on previous state

Example

Example

state $=$ order of cards

Example

state = order of cards

dynamics = riffle shuffle

Example

state $=$ order of cards

dynamics $=$ riffle shuffle

distributions $= d_0, d_1, d_2, d_3, \ldots$

Example

state $=$ order of cards

dynamics $=$ riffle shuffle

distributions $= d_0, d_1, d_2, d_3, \ldots$

$d_0 =$ all weight on one state

Example

state $=$ order of cards

dynamics $=$ riffle shuffle

distributions $= d_0, d_1, d_2, d_3, \ldots$

$d_0 =$ all weight on one state

$d =$ stationary distribution

Example

state $=$ order of cards

dynamics $=$ riffle shuffle

distributions $= d_0, d_1, d_2, d_3, \ldots$

$d_0 =$ all weight on one state

$d =$ stationary distribution

THM $\log n$ steps to d.

Markov Chain

Markov Chain

states: $S = \{1, \ldots, n\}$

Markov Chain

states: $S = \{1, \ldots, n\}$

trajectory: $s_1, s_2, s_3, \ldots$

$$\boxed{s_1 \longrightarrow s_2 \longrightarrow s_3 \longrightarrow s_4}$$

Markov Chain

states: $S = \{1, \ldots, n\}$

trajectory: $s_1, s_2, s_3, \ldots$

$$\boxed{s_1 \to s_2 \to s_3 \to s_4}$$

transition matrix: $M \qquad M_{ij} = p(i|j)$

Markov Chain

states: $S = \{1, \ldots, n\}$

trajectory: $s_1, s_2, s_3, \ldots$

$$\boxed{s_1 \longrightarrow s_2 \longrightarrow s_3 \longrightarrow s_4}$$

transition matrix: $M \qquad M_{ij} = p(i|j)$

dynamics: $p(s' = s_{t+1} | s_t = s)$

Markov Chain

states: $S = \{1, \ldots, n\}$

trajectory: $s_1, s_2, s_3, \ldots$

$$\boxed{s_1 \longrightarrow s_2 \longrightarrow s_3 \longrightarrow s_4}$$

transition matrix: $M \qquad M_{ij} = p(i|j)$

dynamics: $p(s' = s_{t+1} | s_t = s)$

distributions: $d_0, d_1, d_2, d_3, \ldots$

Markov Chain

states: $S = \{1, \ldots, n\}$

trajectory: $s_1, s_2, s_3, \ldots$

$$\boxed{s_1 \longrightarrow s_2 \longrightarrow s_3 \longrightarrow s_4}$$

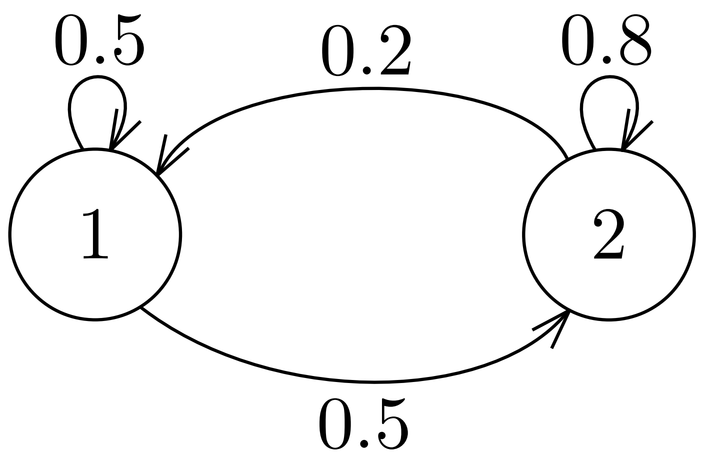
transition matrix: $M \qquad M_{ij} = p(i|j)$

dynamics: $p(s' = s_{t+1} | s_t = s)$

distributions: $d_0, d_1, d_2, d_3, \ldots$

transition: $d_{i+1} = M d_i$

Assignment 3

1 

Use a random number generator and produce a short trajectory.

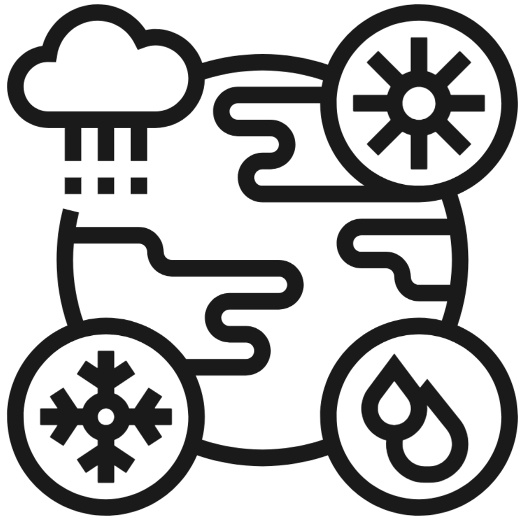
Post on Teams

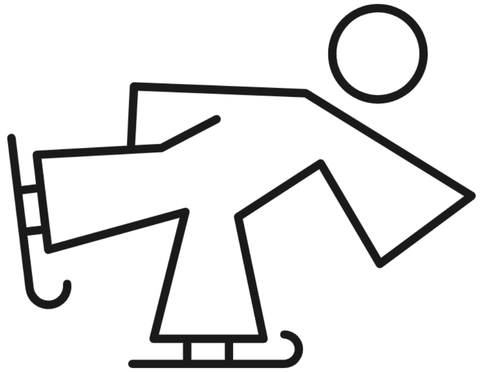
2 Write down the transition matrix.
Compute the distributions d_1, d_2, d_3.
$d_0 = (1, 0)$

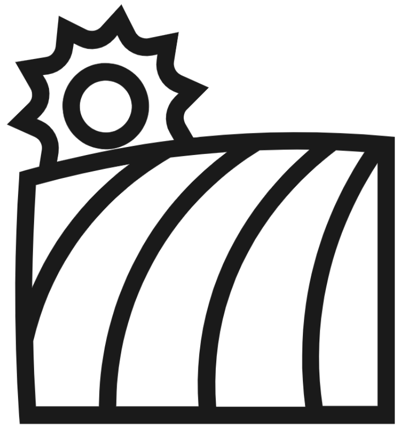
3 Find one example related to nature.

Example Markov Reward Process

The weather in the entire world. The weather tomorrow is a probabilistic function of the weather today. Weather gives rewards in many ways plants grow, freezing temperatures allow us to go skating. :)







Markov Reward Process

Markov Reward Process

states: $S = \{1, \ldots, n\}$

Markov Reward Process

states: $S = \{1, \ldots, n\}$

rewards: $R \subseteq \mathbb{R}$

Markov Reward Process

states: $S = \{1, \ldots, n\}$

rewards: $R \subseteq \mathbb{R}$

trajectory: $s_1, r_1, s_2, r_2, s_3, r_3, \ldots$

$$
\begin{tikzcd}
s_1 \arrow[r] \arrow[d] & s_2 \arrow[r] \arrow[d] & s_3 \arrow[r] \arrow[d] & s_4 \arrow[d] & \cdots \\
r_1 & r_2 & r_3 & r_4 &
\end{tikzcd}
$$

Markov Reward Process

states: $S = \{1, \ldots, n\}$

rewards: $R \subseteq \mathbb{R}$

trajectory: $s_1, r_1, s_2, r_2, s_3, r_3, \ldots$

$$\begin{tikzcd}
s_1 \arrow[r] \arrow[d] & s_2 \arrow[r] \arrow[d] & s_3 \arrow[r] \arrow[d] & s_4 \arrow[d] & \cdots \\
r_1 & r_2 & r_3 & r_4 &
\end{tikzcd}$$

dynamics: $p(s' = s_{t+1}, r = r_{t+1} | s_t = s) \quad p(s', r | s)$

Markov Reward Process

states: $S = \{1, \dots, n\}$

rewards: $R \subseteq \mathbb{R}$

trajectory: $s_1, r_1, s_2, r_2, s_3, r_3, \dots$

$$\begin{tikzcd}
s_1 \arrow[r] \arrow[d] & s_2 \arrow[r] \arrow[d] & s_3 \arrow[r] \arrow[d] & s_4 \arrow[d] & \cdots \\
r_1 & r_2 & r_3 & r_4 &
\end{tikzcd}$$

dynamics: $p(s' = s_{t+1}, r = r_{t+1} | s_t = s) \quad p(s', r | s)$

horizon $=$ steps till terminal state

episodic vs continuing

Markov Reward Process

states: $S = \{1, \ldots, n\}$

rewards: $R \subseteq \mathbb{R}$

trajectory: $s_1, r_1, s_2, r_2, s_3, r_3, \ldots$

$$\begin{tikzcd} s_1 \arrow[r] \arrow[d] & s_2 \arrow[r] \arrow[d] & s_3 \arrow[r] \arrow[d] & s_4 \arrow[d] & \cdots \\ r_1 & r_2 & r_3 & r_4 & \end{tikzcd}$$

dynamics: $p(s' = s_{t+1}, r = r_{t+1} | s_t = s) \qquad p(s', r | s)$

horizon $=$ steps till terminal state

episodic vs continuing

return $G(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$

discount factor $0 < \gamma \leq 1$.

Markov Reward Process

states: $S = \{1, \ldots, n\}$

rewards: $R \subseteq \mathbb{R}$

trajectory: $s_1, r_1, s_2, r_2, s_3, r_3, \ldots$

$$\begin{tikzcd} s_1 \arrow[r] \arrow[d] & s_2 \arrow[r] \arrow[d] & s_3 \arrow[r] \arrow[d] & s_4 \arrow[d] & \cdots \\ r_1 & r_2 & r_3 & r_4 & \end{tikzcd}$$

dynamics: $p(s' = s_{t+1}, r = r_{t+1} | s_t = s)$ $\quad p(s', r | s)$

horizon $=$ steps till terminal state

episodic vs continuing

return $G(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$

discount factor $0 < \gamma \leq 1$.

value: $v(s) = \mathbb{E}\, G(s)$

Assignment 4

Give an example of a **continuing** Markov reward process related to the real world.

- Describe the statespace.
- Describe the rewards.
- Give a short example trajectory.
- Describe the dynamics.
- Argue that you satisfy the Markov property.
- What discount factor seems useful?
- Compute the return of your example trajectory.

Post on Teams

Example Markov Decision Process

Quiz Game, 10 levels, random question, win and go to the next level or loose everything, actions: continue or quit.

Markov Decision Process

Markov Decision Process

states S, rewards R, actions A

Markov Decision Process

states S, rewards R, actions A

trajectory: $s_1, r_1, a_1, s_2, r_2, a_2, s_3, r_3, a_3, \ldots$

$$\begin{tikzcd} s_1 a_1 \arrow[r] \arrow[d] & s_2 a_2 \arrow[r] \arrow[d] & s_3 a_3 \arrow[r] \arrow[d] & s_4 a_4 \cdots \arrow[d] \\ r_1 & r_2 & r_3 & r_4 \end{tikzcd}$$

Markov Decision Process

states S, rewards R, actions A

trajectory: $s_1, r_1, a_1, s_2, r_2, a_2, s_3, r_3, a_3, \ldots$

$$\begin{tikzcd}
s_1 a_1 \arrow[r] \arrow[d] & s_2 a_2 \arrow[r] \arrow[d] & s_3 a_3 \arrow[r] \arrow[d] & s_4 a_4 \cdots \arrow[d] \\
r_1 & r_2 & r_3 & r_4
\end{tikzcd}$$

dynamics: $p(s', r | s, a)$

Markov Decision Process

states S, rewards R, actions A

trajectory: $s_1, r_1, a_1, s_2, r_2, a_2, s_3, r_3, a_3, \ldots$

$$\begin{tikzcd} s_1 a_1 \arrow[r] \arrow[d] & s_2 a_2 \arrow[r] \arrow[d] & s_3 a_3 \arrow[r] \arrow[d] & s_4 a_4 \arrow[d] & \cdots \\ r_1 & r_2 & r_3 & r_4 & \end{tikzcd}$$

dynamics: $p(s', r | s, a)$

policy $\pi : S \to A \qquad \pi : S \to d(A)$

Markov Decision Process

states S, rewards R, actions A

trajectory: $s_1, r_1, a_1, s_2, r_2, a_2, s_3, r_3, a_3, \ldots$

$$\begin{tikzcd}
s_1 a_1 \arrow[r] \arrow[d] & s_2 a_2 \arrow[r] \arrow[d] & s_3 a_3 \arrow[r] \arrow[d] & s_4 a_4 \cdots \arrow[d] \\
r_1 & r_2 & r_3 & r_4
\end{tikzcd}$$

dynamics: $p(s', r|s, a)$

policy $\pi : S \to A \quad \pi : S \to d(A)$

return: $G(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$
(trajectory)

Markov Decision Process

states S, rewards R, actions A

trajectory: $s_1, r_1, a_1, s_2, r_2, a_2, s_3, r_3, a_3, \dots$

$$\begin{tikzcd}
s_1 a_1 \arrow[r] \arrow[d] & s_2 a_2 \arrow[r] \arrow[d] & s_3 a_3 \arrow[r] \arrow[d] & s_4 a_4 \cdots \arrow[d] \\
r_1 & r_2 & r_3 & r_4
\end{tikzcd}$$

dynamics: $p(s', r | s, a)$

policy $\pi : S \to A \quad \pi : S \to d(A)$

return: $G(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots$
(trajectory)

value: $v_\pi(s) = \mathbb{E}_\pi \; G(s)$