Last Time

- What does "Markov" mean in "Markov Decision Process"?

$$P(s_{t+1} \mid s_t, \ldots, s_0) = P(s_{t+1} \mid s_t)$$

$$s_{t+1} \perp s_{t-1} \ldots s_0 \mid s_t$$

Guiding Questions

Guiding Questions

- What is a **Markov decision process**?

Guiding Questions

- What is a **Markov decision process**?
- What is a **policy**?

Guiding Questions

- What is a **Markov decision process**?
- What is a **policy**?
- How do we **evaluate** policies?

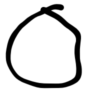
Decision Networks and MDPs

Decision Networks and MDPs

Decision Network

Decision Networks and MDPs

Decision Network



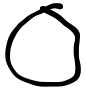
Decision Networks and MDPs

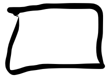
Decision Network

Chance node

Decision Networks and MDPs

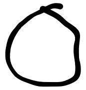
Decision Network

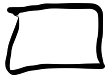
 Chance node



Decision Networks and MDPs

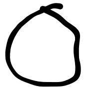
Decision Network

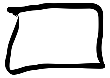
 Chance node

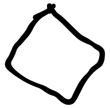
 Decision node

Decision Networks and MDPs

Decision Network

 Chance node

 Decision node



Decision Networks and MDPs

Decision Network

Chance node

Decision node

Utility node

Decision Networks and MDPs

Decision Network

Chance node

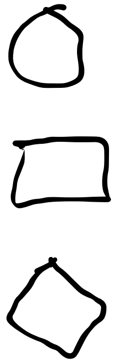
Decision node

Utility node

MDP Dynamic Decision Network

Decision Networks and MDPs

Decision Network

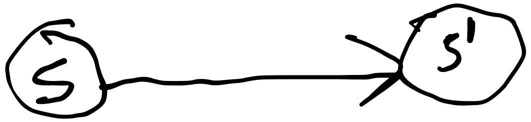


Chance node

Decision node

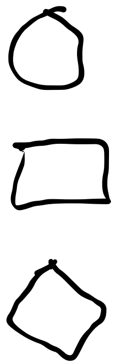
Utility node

MDP Dynamic Decision Network



Decision Networks and MDPs

Decision Network

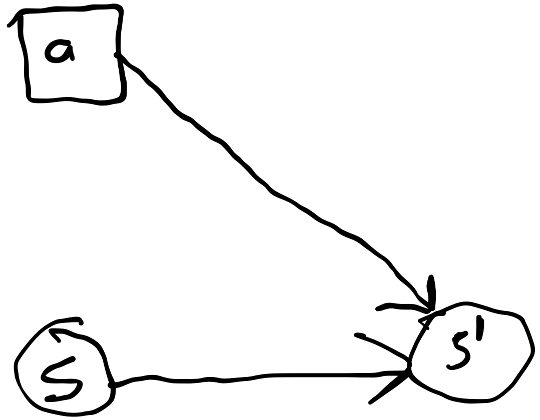


Chance node

Decision node

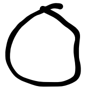
Utility node

MDP Dynamic Decision Network

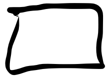


Decision Networks and MDPs

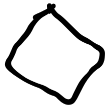
Decision Network



Chance node

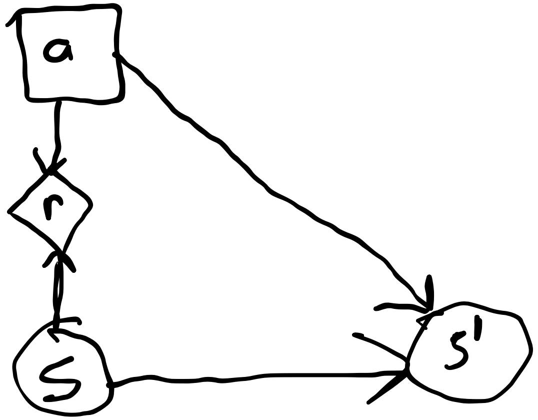


Decision node



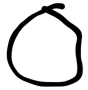
Utility node

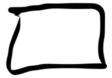
MDP Dynamic Decision Network

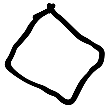


Decision Networks and MDPs

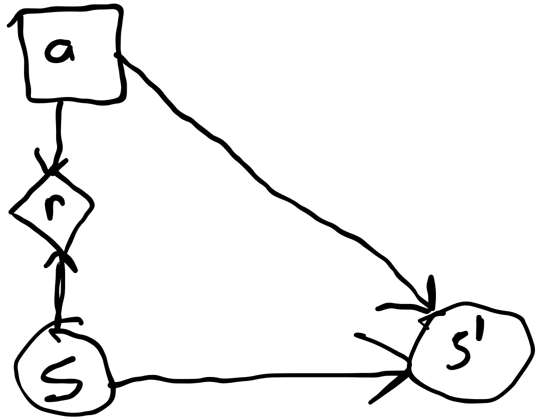
Decision Network

 Chance node

 Decision node

 Utility node

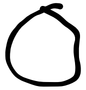
MDP Dynamic Decision Network

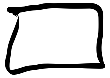


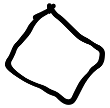
MDP Optimization problem

Decision Networks and MDPs

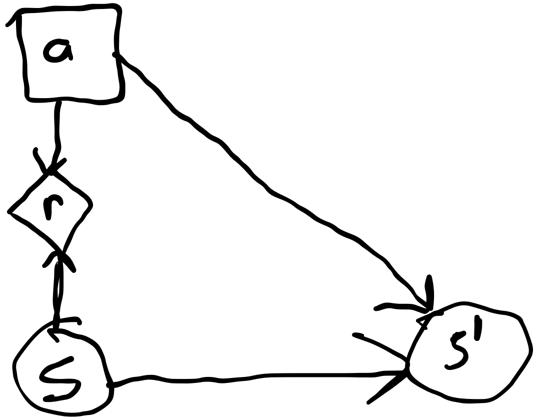
Decision Network

 Chance node

 Decision node

 Utility node

MDP Dynamic Decision Network

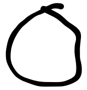


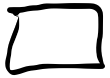
MDP Optimization problem

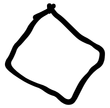
$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=1}^{\infty} r_t\right]$$

Decision Networks and MDPs

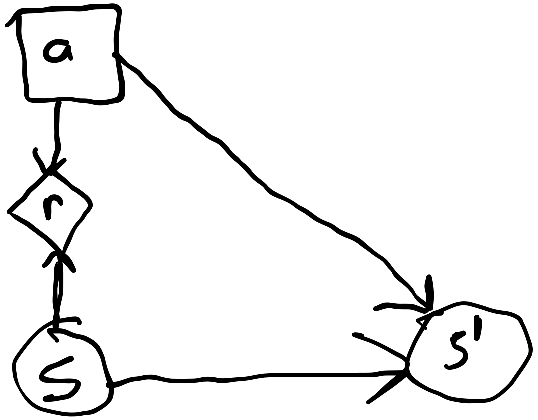
Decision Network

 Chance node

 Decision node

 Utility node

MDP Dynamic Decision Network



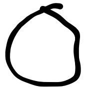
MDP Optimization problem

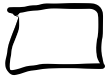
$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=1}^{\infty} r_t\right]$$

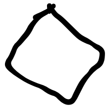
Not well formulated!

Decision Networks and MDPs

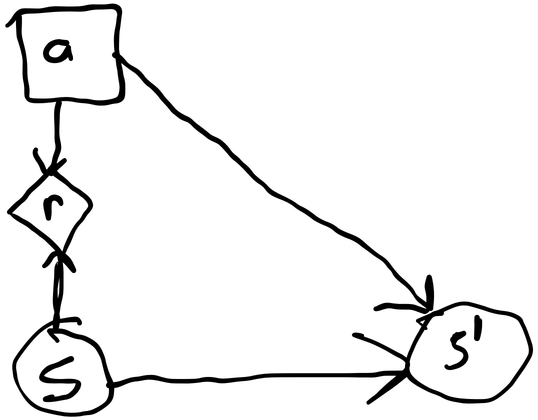
Decision Network

 Chance node

 Decision node

 Utility node

MDP Dynamic Decision Network



MDP Optimization problem

$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=1}^{\infty} r_t\right]$$

Not well formulated!
Infinite

Finite MDP Objectives

Finite MDP Objectives

1. Finite time

Finite MDP Objectives

1. Finite time $\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$

Finite MDP Objectives

1. Finite time $$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

Finite MDP Objectives

1. Finite time

$$\mathrm{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathrm{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

Finite MDP Objectives

1. Finite time
$$\mathrm{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward
$$\lim_{n\to\infty}\mathrm{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty}\mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty}\mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

discount $\gamma \in [0, 1)$

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

discount $\gamma \in [0, 1)$

typically 0.9, 0.95, 0.99

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

discount $\gamma \in [0, 1)$

typically 0.9, 0.95, 0.99

if $\underline{r} \leq r_t \leq \bar{r}$

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

discount $\gamma \in [0, 1)$

typically 0.9, 0.95, 0.99

if $\underline{r} \leq r_t \leq \bar{r}$

then

$$\frac{\underline{r}}{1-\gamma} \leq \sum_{t=0}^{\infty} \gamma^t r_t \leq \frac{\bar{r}}{1-\gamma}$$

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

discount $\gamma \in [0, 1)$

typically 0.9, 0.95, 0.99

if $\underline{r} \leq r_t \leq \bar{r}$

then

$$\frac{\underline{r}}{1-\gamma} \leq \sum_{t=0}^{\infty} \gamma^t r_t \leq \frac{\bar{r}}{1-\gamma}$$

4. Terminal States

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

discount $\gamma \in [0, 1)$

typically 0.9, 0.95, 0.99

if $\underline{r} \leq r_t \leq \bar{r}$

then

$$\frac{\underline{r}}{1-\gamma} \leq \sum_{t=0}^{\infty} \gamma^t r_t \leq \frac{\bar{r}}{1-\gamma}$$

4. Terminal States

Infinite time, but a terminal state (no reward, no leaving) is always reached with probability 1.

Finite MDP Objectives

1. Finite time

$$\mathrm{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathrm{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\boxed{\mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]}$$

discount $\gamma \in [0, 1)$

typically 0.9, 0.95, 0.99

if $\underline{r} \leq r_t \leq \bar{r}$

then

$$\frac{\underline{r}}{1-\gamma} \leq \sum_{t=0}^{\infty} \gamma^t r_t \leq \frac{\bar{r}}{1-\gamma}$$

4. Terminal States

Infinite time, but a terminal state (no reward, no leaving) is always reached with probability 1.

MDP "Tuple Definition"

MDP "Tuple Definition"

(S, A, T, R, γ)

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

$\{1, 2, 3\}$

- S (state space) - set of all possible states

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

$\{1, 2, 3\}$

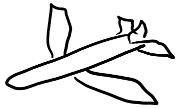
- S (state space) - set of all possible states

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$

MDP "Tuple Definition"

$s \in S = \mathbb{R}^{12}$

$$\begin{bmatrix} x \\ y \\ z \\ u \\ v \\ w \\ \phi \\ \theta \\ \psi \\ p \\ q \\ r \end{bmatrix}$$



(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $\mathbb{R}^2$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

$s \in S$

$\{1, 2, 3\}$ $\quad (x, y) \in \mathbb{R}^2$ $\quad \{0, 1\} \times \mathbb{R}^4$

- S (state space) - set of all possible states

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

- S (state space) - set of all possible states

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

- S (state space) - set of all possible states

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions $\{1, 2, 3\}$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$

$\{\text{test}, \text{wait}, \text{treat}\}$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

  $\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

  $\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

  $\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

  $\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

  $T(s' \mid s, a)$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

$$\underbrace{T(s' \mid s, a)}_{\text{Explicit}}$$

$$\downarrow$$

$$s', r = G(s, a)$$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

$T(s' \mid s, a)$

- R (reward function) - maps each state and action to a reward

$s', r = G(s, a)$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

$T(s' \mid s, a)$

$s' \sim T(s,a)$

$R(s,a) = \underset{s' \sim T(s,a)}{E}\left[R(s,a,s')\right]$

- R (reward function) - maps each state and action to a reward

$R(s, a)$ or $R(s, a, s')$

$s', r = G(s, a)$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

  $\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

  $\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

  $\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

  $\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

  $T(s' \mid s, a)$

- R (reward function) - maps each state and action to a reward

  $R(s, a)$ or $R(s, a, s')$

- γ: discount factor

$s', r = G(s, a)$

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

$T(s' \mid s, a)$

- R (reward function) - maps each state and action to a reward

$R(s, a)$ or $R(s, a, s')$

- γ: discount factor

$s', r = G(s, a)$

- b: initial state distribution

MDP Example

Imagine it's a cold day and you're ready to go to work. You have to decide whether to bike or drive.

MDP Example

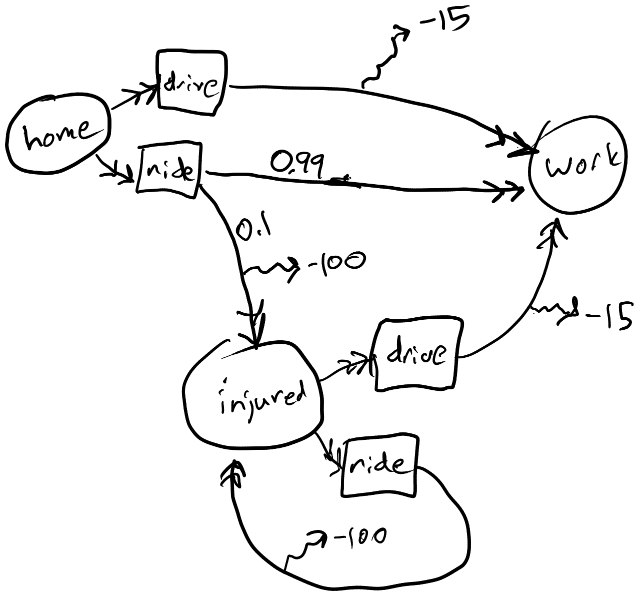
Imagine it's a cold day and you're ready to go to work. You have to decide whether to bike or drive.

- If you drive, you will have to pay $15 for parking; biking is free.

MDP Example

Imagine it's a cold day and you're ready to go to work. You have to decide whether to bike or drive.

- If you drive, you will have to pay $15 for parking; biking is free.
- On 1% of cold days, the ground is covered in ice and you will crash if you bike, but you can't discover this until you start riding. After your crash, you limp home with pain equivalent to losing $100.



$S = \{\text{home}, \text{work}, \text{injured}\}$

$A = \{\text{drive}, \text{ride}\}$

$$T^{\text{drive}} = \begin{array}{c|c} & s' \\ \hline s & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{array} \qquad T^{\text{ride}} = \begin{bmatrix} 0 & 0.99 & 0.01 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R(s,a,s') = \begin{cases} -15 & \text{if } a = \text{drive} \\ -100 & \text{if } s' = \text{injured} \\ 0 & \text{o.w.} \end{cases} \qquad \gamma = 0.99$$

Policies and Simulation

Policies and Simulation

- A *policy*, denoted with π, as in $a_t = \pi(s_t)$ is a function mapping every state to an action.
- When a policy is combined with a Markov decision process, it becomes a Markov stochastic process with

$$P(s' \mid s) = T(s' \mid s, \pi(s))$$

Simulate

Input: (S, A, T, R, γ, b)

Output: $\hat{u}$ (accumulated reward)

$s \leftarrow$ sample(b)

$\hat{u} \leftarrow 0$

for t in $0..T-1$ until $\gamma^t < \varepsilon$

    $a \leftarrow \pi(s)$

    $s', r \leftarrow G(s,a)$ samples from T, R

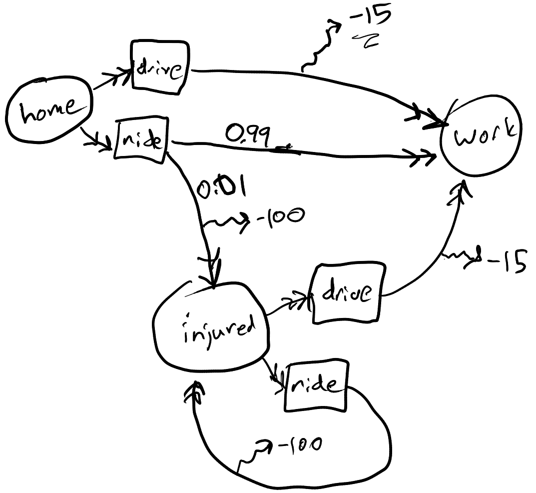
    $\hat{u} \leftarrow \hat{u} + \gamma^t r$

    $s \leftarrow s'$

return $\hat{u}$

Break

- Suggest a policy that you think is optimal for the icy day problem



$S = \{\text{home}, \text{work}, \text{injure}$

$A = \{\text{drive}, \text{ride}\}$

$$T^{\text{drive}} = \begin{array}{c c} & s' \\ s & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{array} \qquad T^{\text{ride}} = \begin{bmatrix} 0 & 0.99 & 0.01 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R(s,a,s') = \begin{cases} -15 & \text{if } a = \text{drive} \\ -100 & \text{if } s' = \text{injured} \\ 0 & \text{o.w.} \end{cases} \qquad \gamma = 0.99$$

$$0.99 \cdot 0 + 0.01(-100 - 15) = -1.15$$

$$\pi(s) = \begin{cases} & \text{if } s = \text{home} \\ \text{drive} & \text{if } s = \text{injured} \\ & \end{cases}$$

P

U

Utility

$A \succ B$: prefer A to B

$A \sim B$: indifferent

$A \succeq B$: prefer A or indifferent

Lottery: $[S_1 : p_1 ; S_2 : p_2 ; \ldots S_n : p_n]$

Completeness: Exactly 1 holds:

$A \succ B \quad B \succ A \quad A \sim B$

Transitivity: If $A \succeq B$ and $B \succeq C$ the $A \succeq C$

Continuity: If $A \succeq C \succeq B$ then $\exists p$ s.t.

$[A : p ; B : 1-p] \sim C$

Independence: If $A \succ B$ then

$[A : p ; C : 1-p] \succeq [B : p ; C : 1-p]$

$\exists$ U s.t.

$U(A) > U(B)$ iff $A \succ B$

$U(A) = U(B)$ iff $A \sim B$

$\underbrace{U([S_1 : p_1 ; \ldots S_n : p_n])}$ $= \sum_i p_i S_i$

Slide not on Exam

Policy Evaluation

$$U(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid a_t = \pi(s_t)\right]$$

Naive:

marginal of s_t given π is being executed

$$U(\pi) = \sum_{t=0}^{\infty} \gamma^t P^{\pi}(s_t) R(s_t, \pi(s_t))$$

$$P^{\pi}(s_t) = \sum_{s_{t-1}} T(s_t \mid s_{t-1}, \pi(s_{t-1})) P^{\pi}(s_{t-1})$$

Naive Policy Evaluation not on Exam

Value Function-Based Policy Evaluation

$$U(\pi) = \underset{\tau}{E}\left[\sum \gamma^t R(s_t, a_t) \right|$$

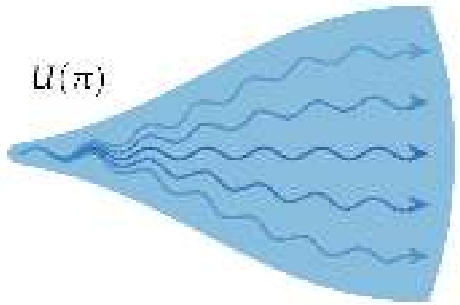
Monte Carlo Policy Evaluation

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

Monte Carlo Policy Evaluation

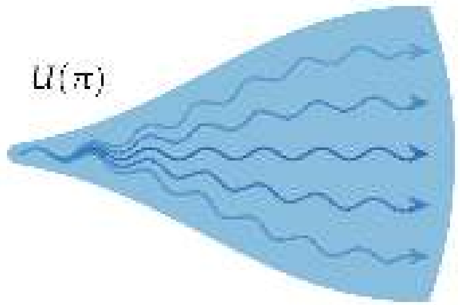
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Monte Carlo Policy Evaluation

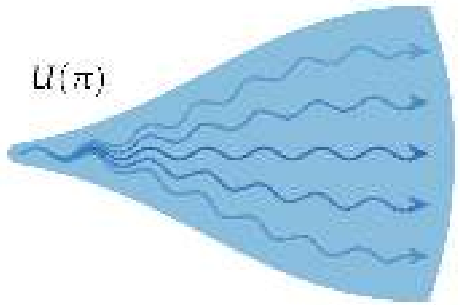
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP



Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

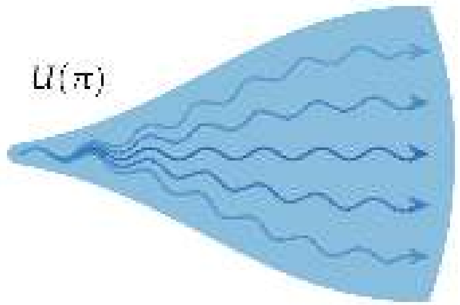


Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Let $\tau = (s_0, a_0, r_0, s_1, \dots, s_T)$ be a *trajectory* of the MDP

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

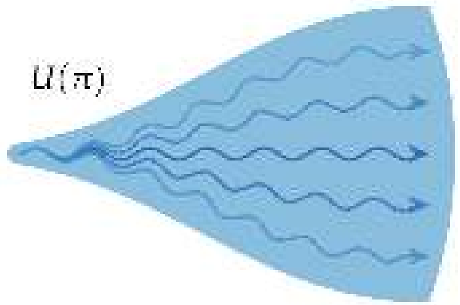
random variable

$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

where $\hat{u}^{(i)}$ is generated by a rollout simulation

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

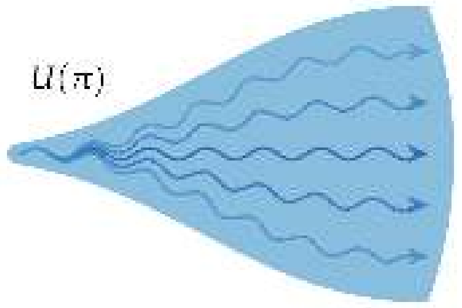
where $\hat{u}^{(i)}$ is generated by a rollout simulation

How can we quantify the accuracy of $\bar{u}_m$?

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP



$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

also an R.V.

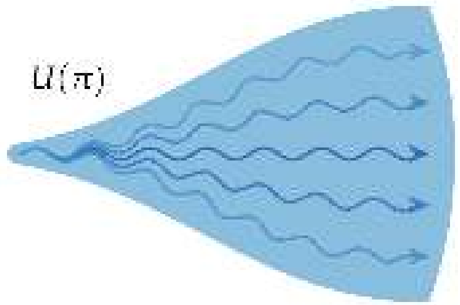
$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

How can we quantify the accuracy of $\bar{u}_m$?

where $\hat{u}^{(i)}$ is generated by a rollout simulation

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

also an R.V.

$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

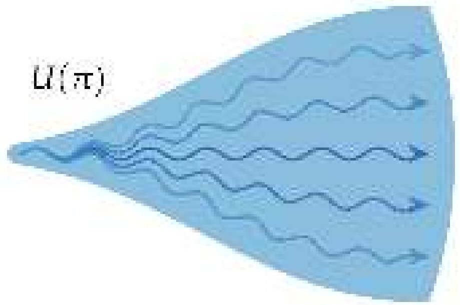
How can we quantify the accuracy of $\bar{u}_m$?

C.L.T.

where $\hat{u}^{(i)}$ is generated by a rollout simulation

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

also an R.V.

$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

where $\hat{u}^{(i)}$ is generated by a rollout simulation

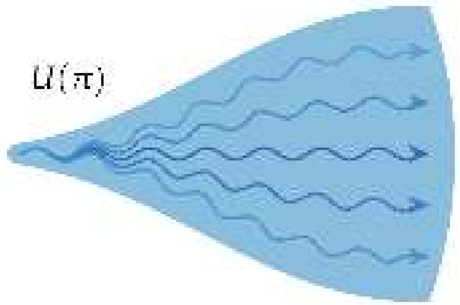
How can we quantify the accuracy of $\bar{u}_m$?

C.L.T. $\quad \dfrac{\bar{u}_m - U(\pi)}{\sigma_m / \sqrt{m}} \xrightarrow{d} \mathcal{N}(0, 1)$

CLT not on exam

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

also an R.V.

$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

where $\hat{u}^{(i)}$ is generated by a rollout simulation

How can we quantify the accuracy of $\bar{u}_m$?

C.L.T. $\underbrace{\dfrac{\bar{u}_m - U(\pi)}{\sigma_m/\sqrt{m}} \xrightarrow{d} \mathcal{N}(0,1)}$

CLT not on exam

$$\text{s.e.m.} = \frac{\text{std}(\hat{u})}{\sqrt{m}}$$

Standard Error of the Mean

Guiding Questions

Guiding Questions

- What is a **Markov decision process**?

Guiding Questions

- What is a **Markov decision process**?
- What is a **policy**?

Guiding Questions

- What is a **Markov decision process**?
- What is a **policy**?
- How do we **evaluate** policies?