HW 1 due tonight

Part way through!

Last Time

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

Stochastic Process $\{x_t\}$ is Markov if

$$P(x_{t+1} \mid x_t, x_{t-1} \ldots x_0) = P(x_{t+1} \mid x_t)$$

Guiding Questions

Guiding Questions

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

Guiding Questions

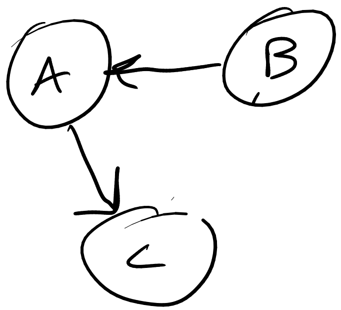
- What is a **Markov decision process**?
- What is a **policy**?

Guiding Questions

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

Decision Networks and MDPs

Bayes Net



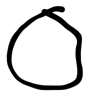
$P(C \mid A, B) = P(C \mid A)$

Decision Networks and MDPs

Decision Network

Decision Networks and MDPs

Decision Network



Decision Networks and MDPs

Decision Network

Chance node

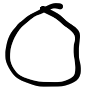
Decision Networks and MDPs

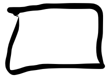
Decision Network

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

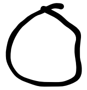
Chance node

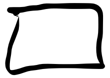
Decision node

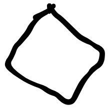
Utility node

Decision Networks and MDPs

Decision Network

 Chance node

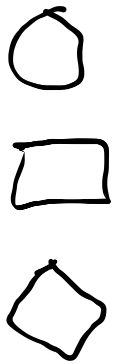
 Decision node

 Utility node

MDP Dynamic Decision Network

Decision Networks and MDPs

Decision Network

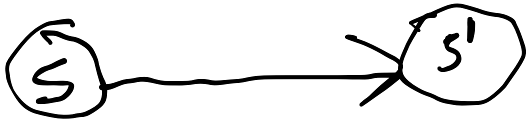


Chance node

Decision node

Utility node

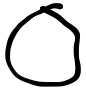
MDP Dynamic Decision Network

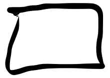


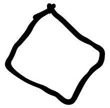
DBN for
a Markov
Stoch. Process

Decision Networks and MDPs

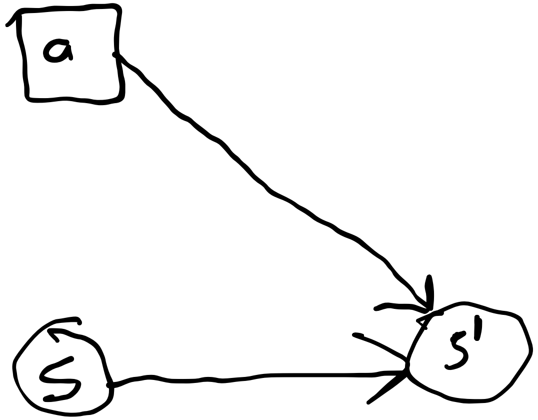
Decision Network

 Chance node

 Decision node

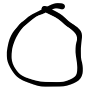
 Utility node

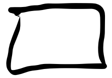
MDP Dynamic Decision Network

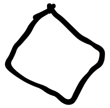


Decision Networks and MDPs

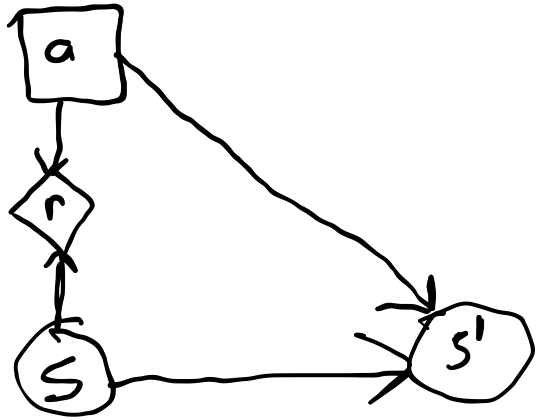
Decision Network

 Chance node

 Decision node

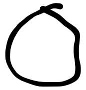
 Utility node

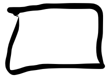
MDP Dynamic Decision Network

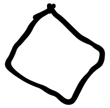


Decision Networks and MDPs

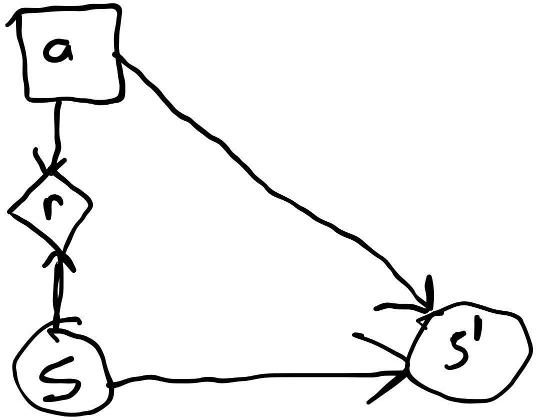
Decision Network

 Chance node

 Decision node

 Utility node

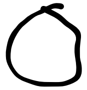
MDP Dynamic Decision Network

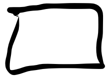


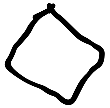
MDP Optimization problem

Decision Networks and MDPs

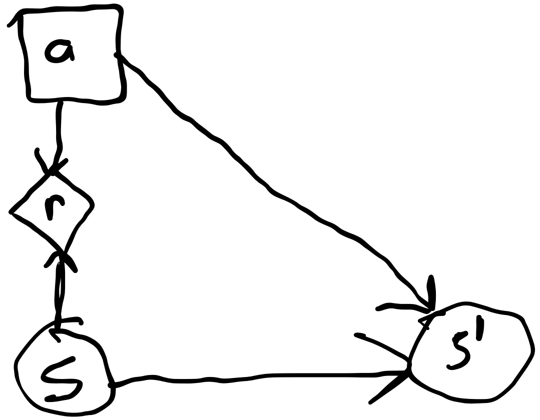
Decision Network

 Chance node

 Decision node

 Utility node

MDP Dynamic Decision Network

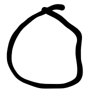


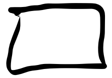
MDP Optimization problem

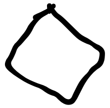
$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=1}^{\infty} r_t\right]$$

Decision Networks and MDPs

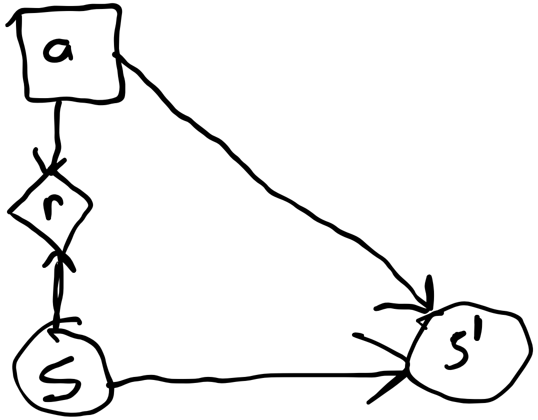
Decision Network

 Chance node

 Decision node

 Utility node

MDP Dynamic Decision Network



MDP Optimization problem

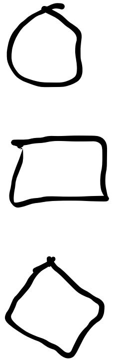
$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=1}^{\infty} r_t\right]$$

Not well formulated!

$r_t = 1$

Decision Networks and MDPs

Decision Network

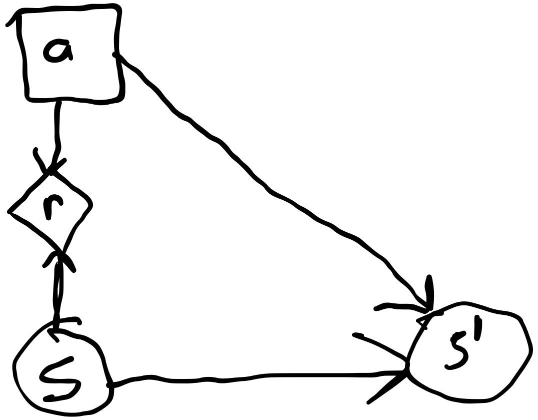


Chance node

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

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

2. Average reward

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

3. Discounting

Finite MDP Objectives

1. Finite time

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

2. Average reward

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

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

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

3. Discounting

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

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

$\gamma = 0.9$

$\gamma^0 = 1$

Finite MDP Objectives

1. Finite time

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

2. Average reward

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

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

$\gamma^0 = 1$ $\gamma^1 = 0.99$

$\gamma^2 = 0.99^2$

Finite MDP Objectives

1. Finite time

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

2. Average reward

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

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 $\underline{r} \le r_t \le \bar{r}$

Finite MDP Objectives

1. Finite time

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

2. Average reward

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

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

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

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

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

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

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

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

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

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.

MDP "Tuple Definition"

MDP "Tuple Definition"

$$(S, A, T, R, \gamma)$$

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)

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

$\{1, 2, 3\}$

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

MDP "Tuple Definition"

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

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

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

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

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)

state space

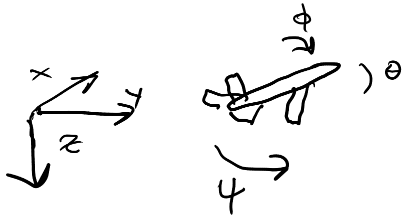
$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

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

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

MDP "Tuple Definition"

$(x, y, z, u, v, w, \phi, \theta, \psi, p, q, r) \in \mathbb{R}^{12}$



(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)

- 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

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 $\{1, 2, 3\}$

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

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

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

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

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

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

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

MDP "Tuple Definition"

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

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

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

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

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

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

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

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

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

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

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

  $R(s, a)$

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

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

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

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

$R(s, a)$

- γ: discount factor

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

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

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

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

$R(s, a)$

- γ: discount factor

$S = \{1,2,3\}$ $A = \{1,2\}$

s'	$T(s' \mid s=1, a=1)$
1	0.5
2	0.2
3	0.3

- b: initial state distribution

$s_0 \sim b$

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.

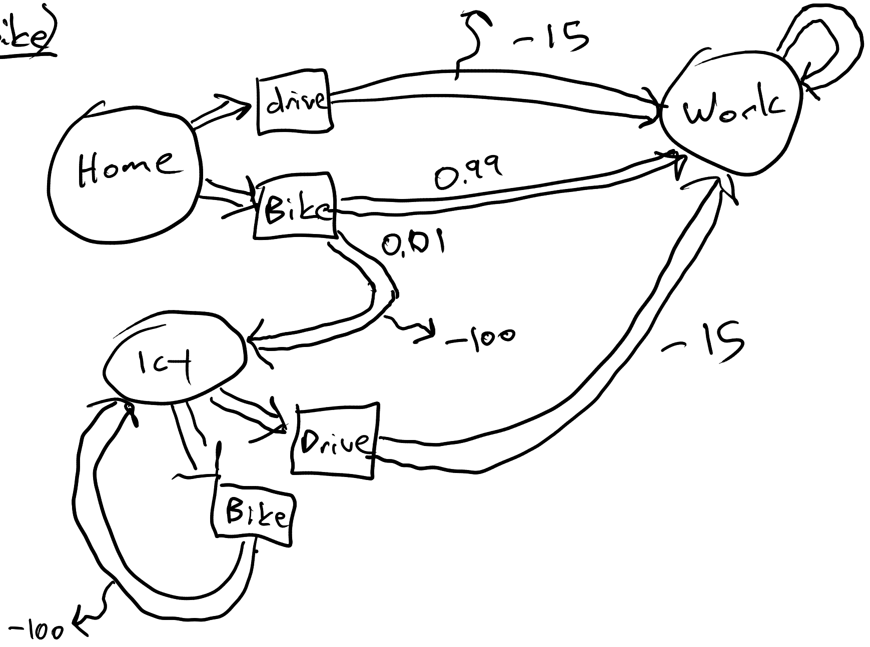
MDP Example

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.

$$\begin{cases} S = \{\text{home}, \text{icy}, \text{work}\} \\ A = \{\text{drive}, \text{bike}\} \\ R(s,a,s') = \begin{cases} -15 & \text{if } a=\text{drive} \\ -100 & \text{if } a=\text{bike},\ s'=\text{icy} \\ 0 & \text{o.w.} \end{cases} \\ \gamma = 0.99 \\ T(s' \mid s,a) = \begin{cases} 1 & \text{if } s=\text{home},\ a=\text{drive},\ s'=\text{work} \\ 0.99 & \text{if } s=\text{home},\ a=\text{bike},\ s'=\text{work} \\ \vdots & \\ 0 & \text{if } a=\text{drive},\ s'=\text{icy} \end{cases} \end{cases}$$

s'	$T(s' \mid s=\text{home}, a=\text{bike})$
work	0.99
icy	0.01
home	0



Policy

Policy

- A *policy*, denoted with, π as in $a_t = \pi(s_t)$ is a function mapping every state to an action.

Policy

- A *policy*, denoted with, π as in $a_t = \pi(s_t)$ is a function mapping every state to an action.
- An optimal policy, π^*, maximizes the sum of expected rewards:

$a_t = \pi(s_t)$

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \, \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t \, R\left(s_t, \pi(s_t)\right)\right]$$

Breakout Rooms

- Name, "I think Colorado winters are ______"
- Suggest a policy that you think is optimal for the icy day problem

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

MDP Simulation

MDP Simulation

<u>Algorithm: Rollout Simulation</u>

Given: MDP (S, A, R, T, γ, b)

$s \leftarrow \text{sample}(b)$

$\hat{u} \leftarrow 0$

for t in $0 \ldots T-1$

  $s', r \leftarrow G(s, a)$

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

  $s \leftarrow s'$

return $\hat{u}$

$\gamma^T < \varepsilon$

$T(s'|s,a)$

$R(s,a)$

Policy Evaluation

Policy Evaluation

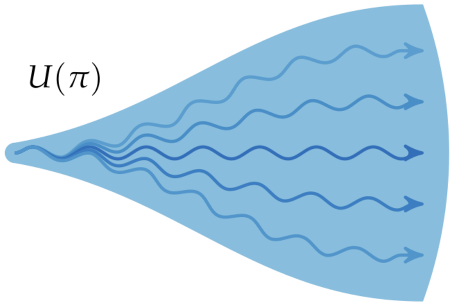
- *Evaluating* a policy means determining the expected reward.

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

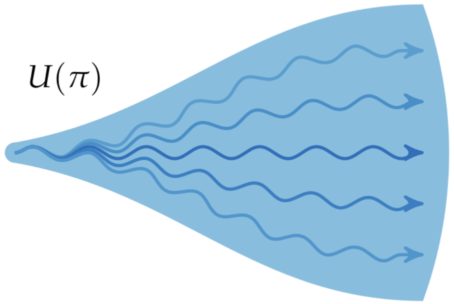
Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Policy Evaluation

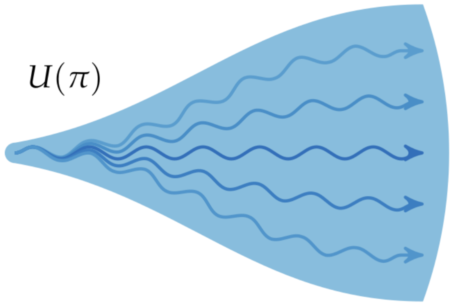
- *Evaluating* a policy means determining the expected reward.
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



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

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

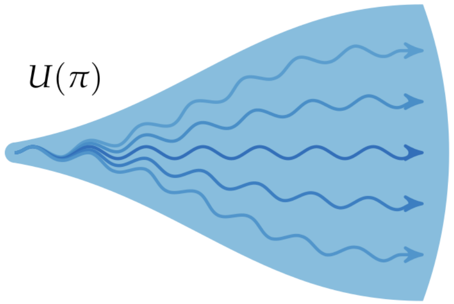


Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

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

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- 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

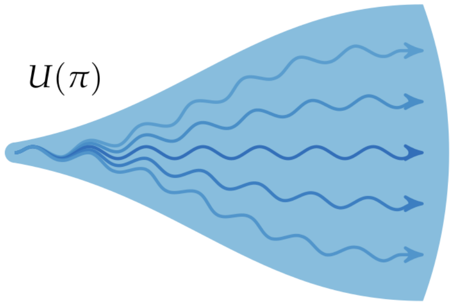
$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

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

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

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- 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

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

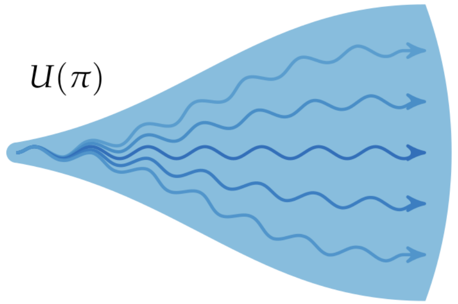
How can we quantify the accuracy of $\bar{u}_m$?

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

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

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- 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.

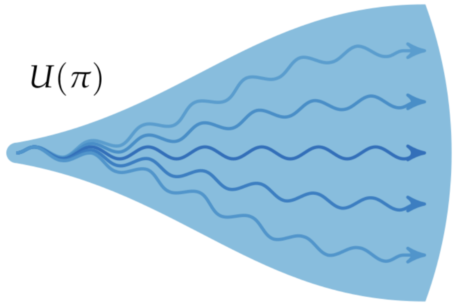
$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

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

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

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- 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

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

also an R.V.

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

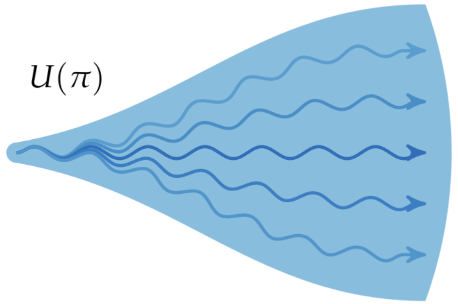
where $\hat{u}^{(i)}$ is generated by a rollout simulation

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

C.L.T.

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- 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.

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

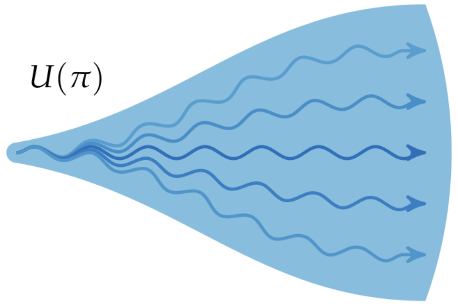
where $\hat{u}^{(i)}$ is generated by a rollout simulation

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

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

Policy Evaluation

- *Evaluating* a policy means determining the expected reward.
- 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

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

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

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

Guiding Questions

Guiding Questions

- What is a **Markov decision process**? (S, A, T, R, γ)

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?