Markov Decision Processes

Last Time

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

Guiding Questions

Guiding Questions

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

Guiding Questions

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

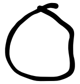
Guiding Questions

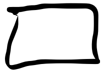
- What is a **Markov decision process**?
- What is a **policy**?
- How do we **evaluate** policies?

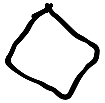
Decision Networks and MDPs

Decision Networks and MDPs

Decision Network

 Chance node

 Decision node

 Utility node

Decision Networks and MDPs

Decision Network

Chance node

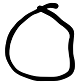
Decision node

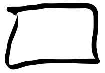
Utility node

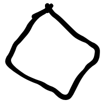
MDP Decision Network

Decision Networks and MDPs

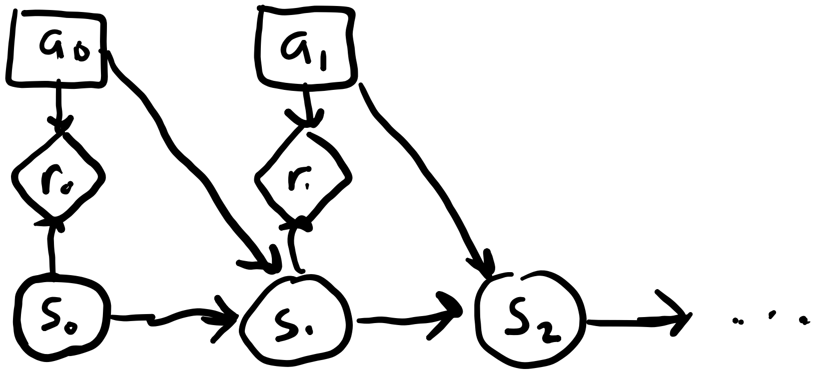
Decision Network


Chance node


Decision node

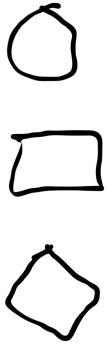

Utility node

MDP Decision Network



Decision Networks and MDPs

Decision Network

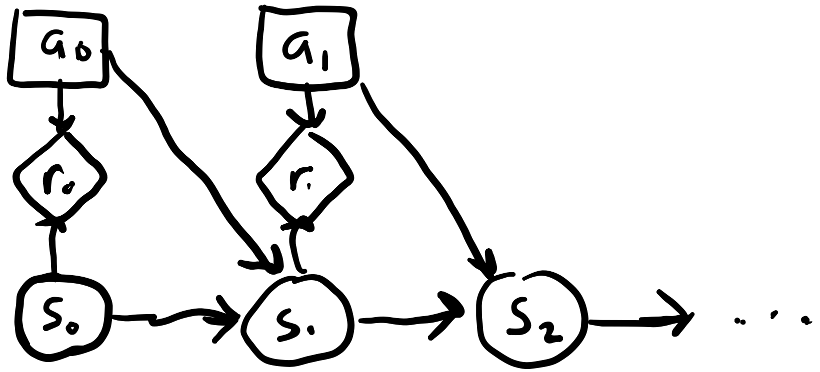


Chance node

Decision node

Utility node

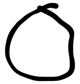
MDP Decision Network

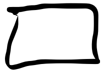


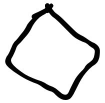
MDP **Dynamic** Decision Network

Decision Networks and MDPs

Decision Network

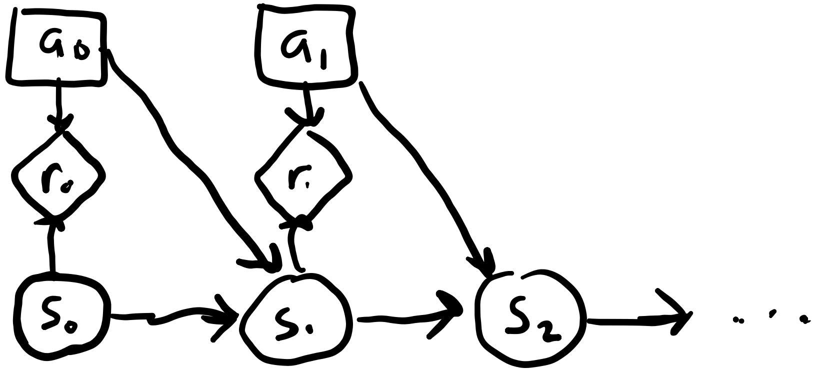

Chance node


Decision node


Utility node

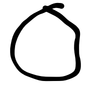
MDP **Dynamic** Decision Network

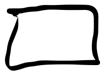
MDP Decision Network

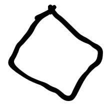


Decision Networks and MDPs

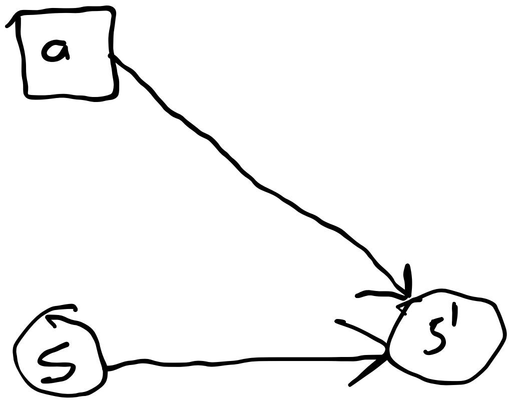
Decision Network


Chance node

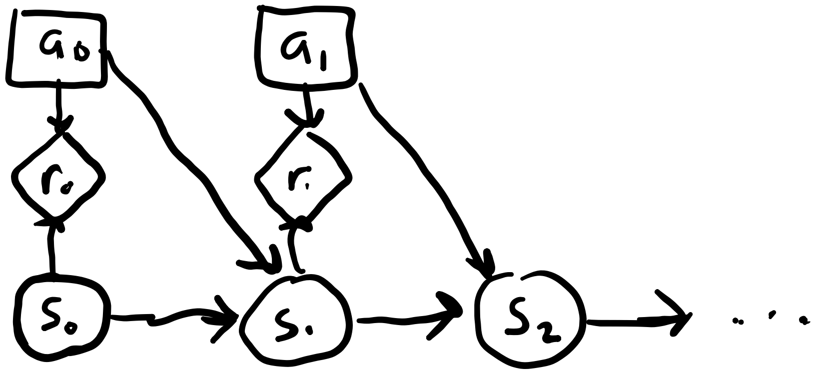

Decision node


Utility node

MDP **Dynamic** Decision Network



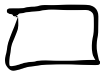
MDP Decision Network

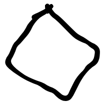


Decision Networks and MDPs

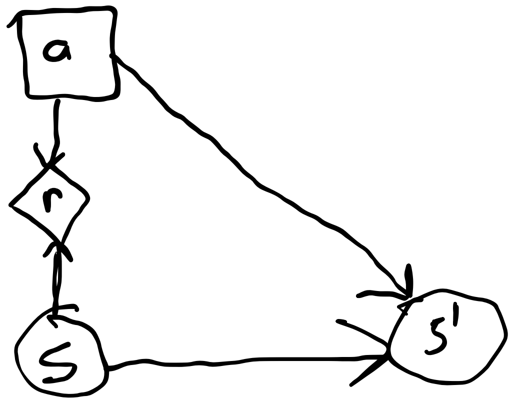
Decision Network

 Chance node

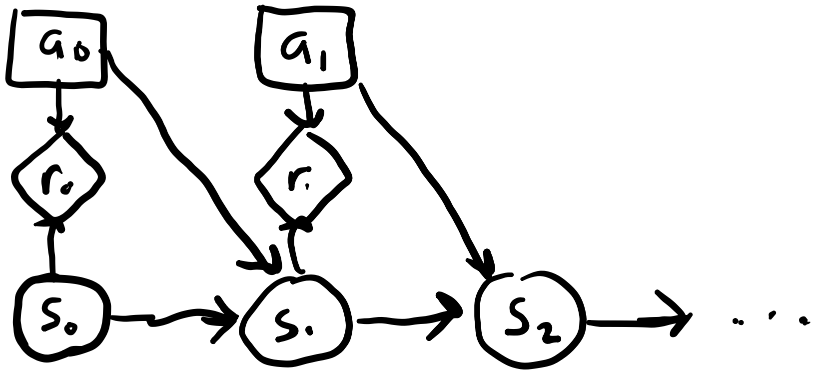
 Decision node

 Utility node

MDP **Dynamic** Decision Network

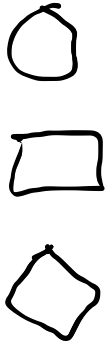


MDP Decision Network



Decision Networks and MDPs

Decision Network

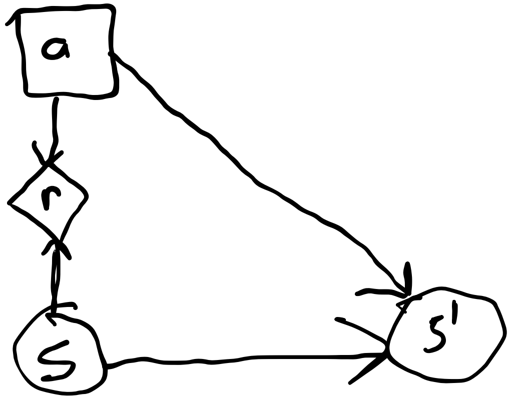


Chance node

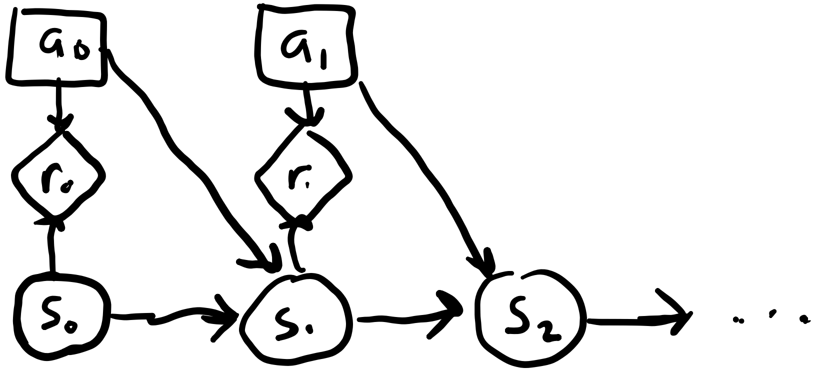
Decision node

Utility node

MDP **Dynamic** Decision Network



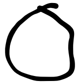
MDP Decision Network

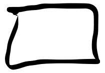


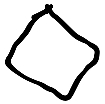
MDP Optimization problem

Decision Networks and MDPs

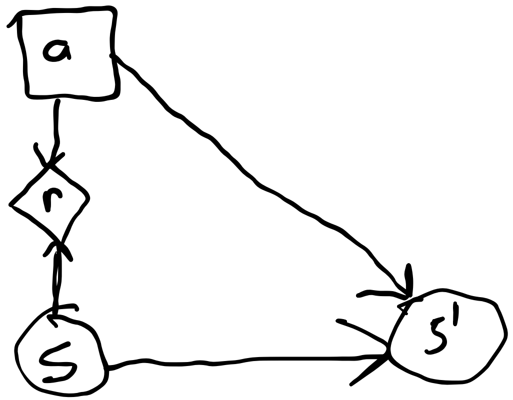
Decision Network


Chance node

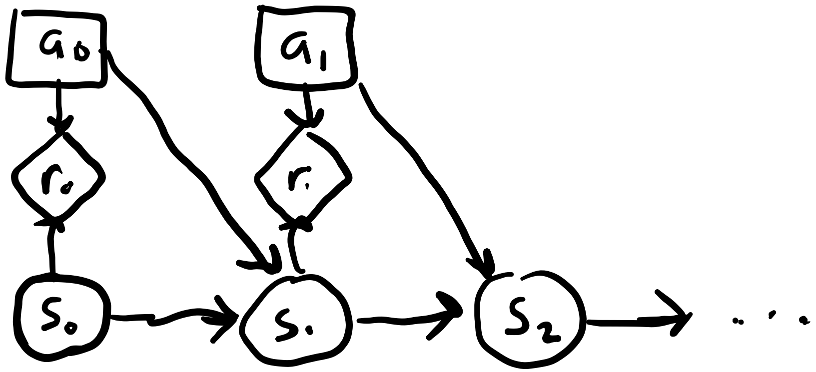

Decision node


Utility node

MDP **Dynamic** Decision Network



MDP Decision Network

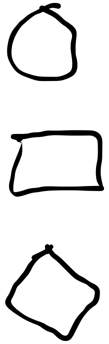


MDP Optimization problem

$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=0}^{\infty} r_t\right]$$

Decision Networks and MDPs

Decision Network

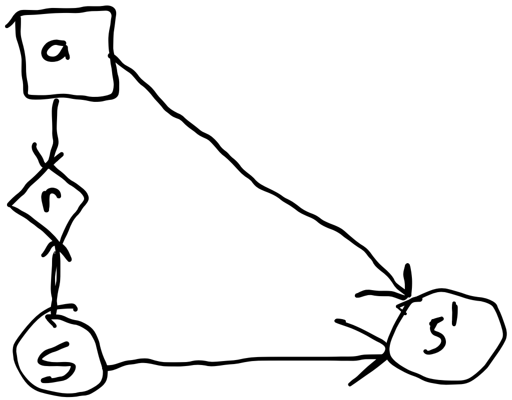


Chance node

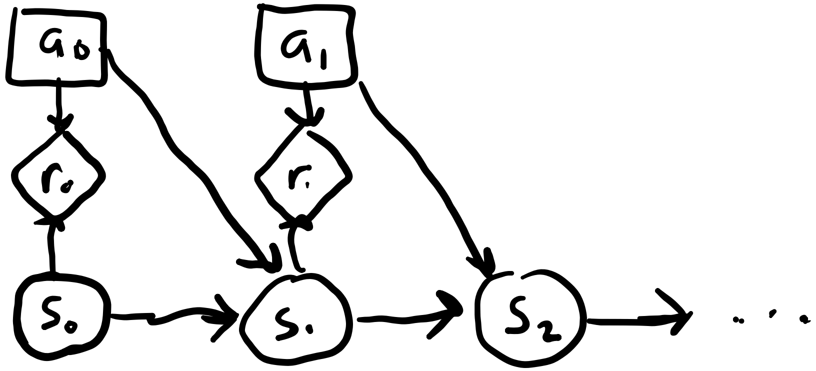
Decision node

Utility node

MDP Dynamic Decision Network



MDP Decision Network



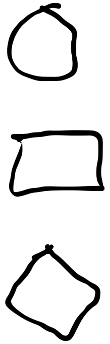
MDP Optimization problem

$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=0}^{\infty} r_t\right]$$

Not well formulated!

Decision Networks and MDPs

Decision Network

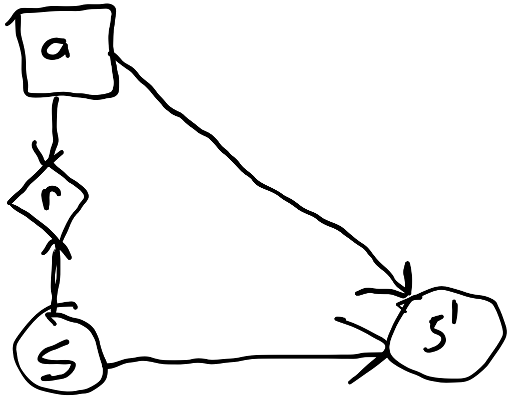


Chance node

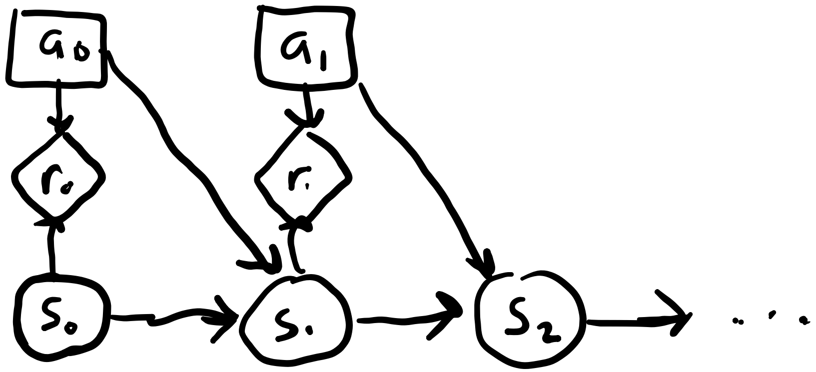
Decision node

Utility node

MDP Dynamic Decision Network



MDP Decision Network



MDP Optimization problem

$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=0}^{\infty} r_t\right]$$

Not well formulated!
Infinite

Finite MDP Objectives

Finite MDP Objectives

1. Finite time

Finite MDP Objectives

1. Finite time

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

Finite MDP Objectives

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

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

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

2. Average reward

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

3. Discounting

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

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

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

2. Average reward

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

3. Discounting

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

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.

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 and/or S_T in some contexts)

MDP "Tuple Definition"

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

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

MDP "Tuple Definition"

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

$\{1, 2, 3\}$

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

MDP "Tuple Definition"

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

$\{1, 2, 3\}$

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

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

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

MDP "Tuple Definition"

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

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

$\{1, 2, 3\}$

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

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

$\{1, 2, 3\}$

MDP "Tuple Definition"

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

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

  $\{1, 2, 3\}$

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

- 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 and/or S_T in some contexts)

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

  $\{1, 2, 3\}$

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

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

  $\{1, 2, 3\}$

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

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

MDP "Tuple Definition"

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

- S (state space) - set of all possible states $\{1, 2, 3\}$

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

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

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

- 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 and/or S_T in some contexts)

- S (state space) - set of all possible states $\{1, 2, 3\}$ $\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$
- A (action space) - set of all possible actions $\{1, 2, 3\}$ $\{\text{test}, \text{wait}, \text{treat}\}$
- 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

MDP "Tuple Definition"

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

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

  $\{1, 2, 3\}$

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

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

  $\{1, 2, 3\}$

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

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

MDP "Tuple Definition"

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

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

  $\{1, 2, 3\}$

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

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

  $\{1, 2, 3\}$

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

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

"Generative Model": Alternative to T and R

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

MDP "Tuple Definition"

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

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

  $\{1, 2, 3\}$

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

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

  $\{1, 2, 3\}$

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

- 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

"Generative Model": Alternative to T and R

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

MDP "Tuple Definition"

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

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

$\{1, 2, 3\}$

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

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

$\{1, 2, 3\}$

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

- 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
- b: initial state distribution
- S_t: set of terminal states

"Generative Model": Alternative to T and R

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

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.

Policies and Simulation

cf. Alg. 9.1, p. 184

Policies and Simulation

- A *policy*, denoted with $\pi(a_t \mid s_t)$, is a conditional distribution of actions given states.

cf. Alg. 9.1, p. 184

Policies and Simulation

- A *policy*, denoted with $\pi(a_t \mid s_t)$, is a conditional distribution of actions given states.
- $a_t = \pi(s_t)$ is used as shorthand when a policy is deterministic.

cf. Alg. 9.1, p. 184

Policies and Simulation

- A *policy*, denoted with $\pi(a_t \mid s_t)$, is a conditional distribution of actions given states.
- $a_t = \pi(s_t)$ is used as shorthand when a policy is deterministic.
- When a policy is combined with an MDP, it becomes a Markov stochastic process with

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

  or

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

cf. Alg. 9.1, p. 184

Policies and Simulation

- A *policy*, denoted with $\pi(a_t \mid s_t)$, is a conditional distribution of actions given states.
- $a_t = \pi(s_t)$ is used as shorthand when a policy is deterministic.
- When a policy is combined with an MDP, it becomes a Markov stochastic process with

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

or

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

MDP Objective:

$$\text{maximize}\, U(\pi) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$$

cf. Alg. 9.1, p. 184

Policies and Simulation

- A *policy*, denoted with $\pi(a_t \mid s_t)$, is a conditional distribution of actions given states.
- $a_t = \pi(s_t)$ is used as shorthand when a policy is deterministic.
- When a policy is combined with an MDP, it becomes a Markov stochastic process with

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

or

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

MDP Objective:

$$\text{maximize } U(\pi) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$$

<u>Algorithm: Rollout Simulation</u>

Inputs: MDP (S, A, R, T, γ, b) (only need generative model, G), Policy π, horizon H

Outputs: Utility estimate $\hat{u}$

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

$\hat{u} \leftarrow 0$

for t in $0 \ldots H-1$

    $a \leftarrow \text{sample}(\pi(a \mid s))$

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

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

    $s \leftarrow s'$

return $\hat{u}$

cf. Alg. 9.1, p. 184

Break

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

Naive Policy Evaluation

Naive Policy Evaluation not on Exam

Naive Policy Evaluation

$$U(\pi) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$$

Naive Policy Evaluation not on Exam

Naive Policy Evaluation

$$U(\pi) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$$

Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP, and $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$

Naive Policy Evaluation not on Exam

Naive Policy Evaluation

$$U(\pi) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$$

Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP, and $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$

$$U(\pi) = \mathrm{E}\left[R(\tau) \mid \pi\right] = \sum_{\tau} R(\tau) P(\tau \mid \pi)$$

Naive Policy Evaluation not on Exam

Naive Policy Evaluation

$$U(\pi) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$$

Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP, and $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$

$$U(\pi) = \mathrm{E}\left[R(\tau) \mid \pi\right] = \sum_{\tau} R(\tau) P(\tau \mid \pi)$$

$$P(\tau \mid \pi) = b(s_0) \prod_{t=0}^{\infty} T(s_{t+1} \mid s_t, \pi(t))$$

Naive Policy Evaluation not on Exam

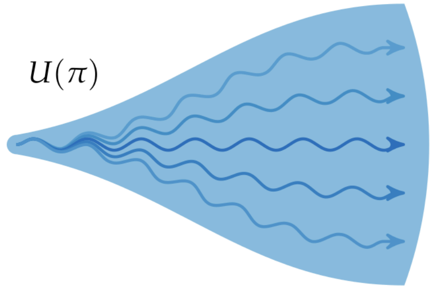
Monte Carlo Policy Evaluation

Monte Carlo Policy Evaluation

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

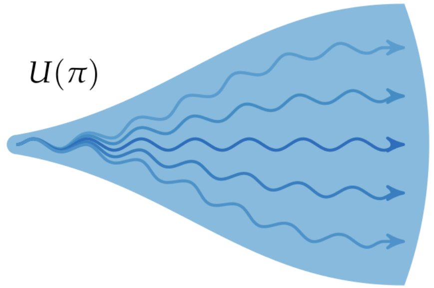
Monte Carlo Policy Evaluation

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



Monte Carlo Policy Evaluation

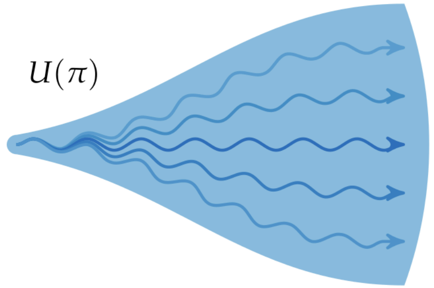
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



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*

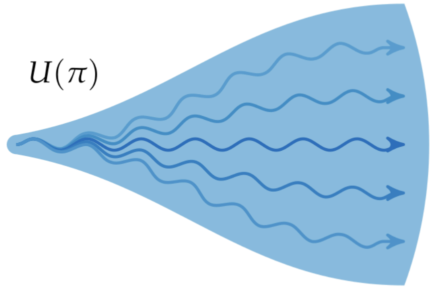


Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

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

Monte Carlo Policy Evaluation

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



Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ 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

Monte Carlo Policy Evaluation

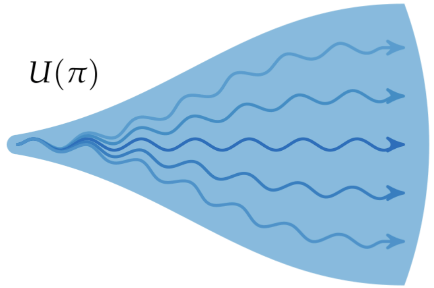
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

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

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

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

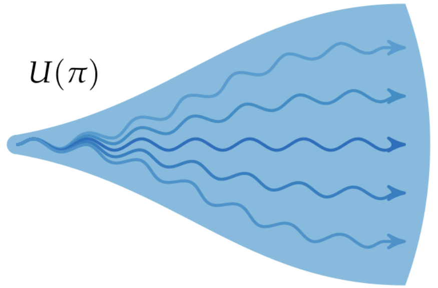
where $\hat{u}^{(i)}$ is generated by a rollout simulation



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.

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

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

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

Value Function-Based Policy Evaluation

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?