

# 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

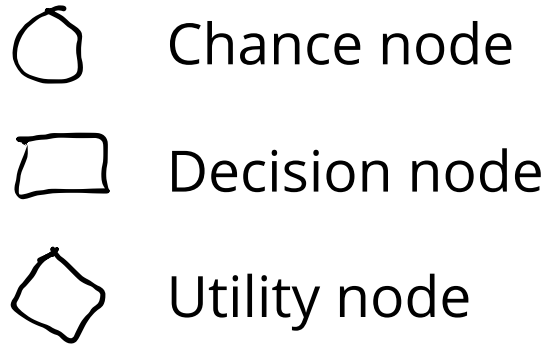


Utility node

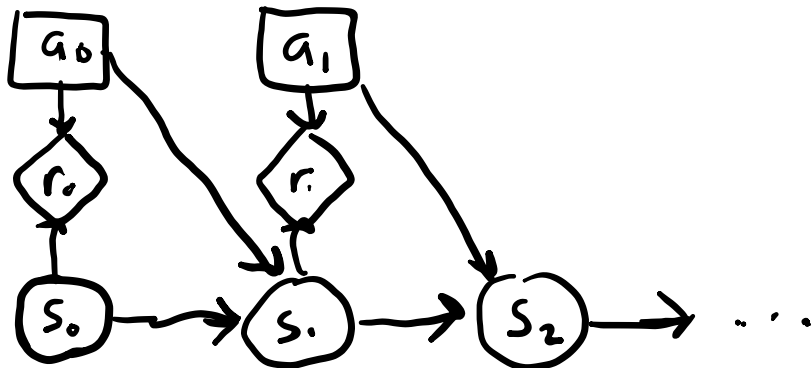
## MDP Decision Network

# Decision Networks and MDPs

## Decision Network

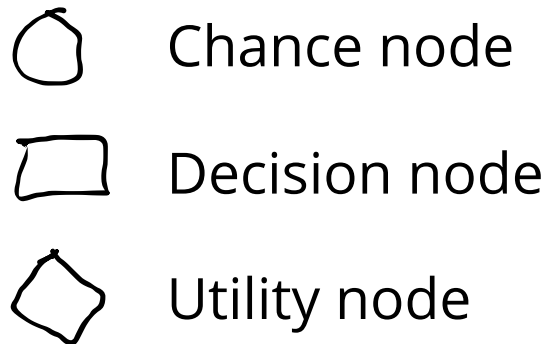


## MDP Decision Network



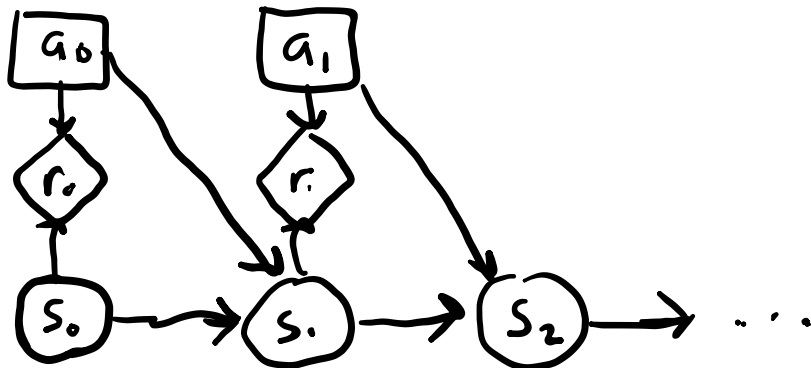
# Decision Networks and MDPs

Decision Network



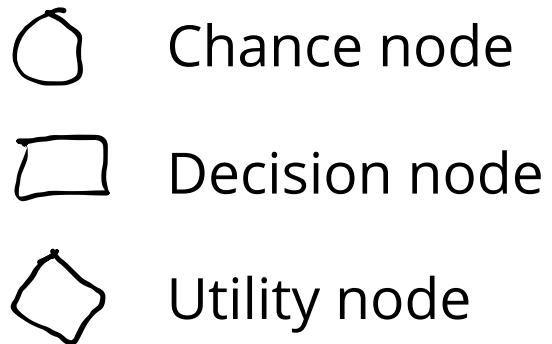
MDP **Dynamic** Decision Network

MDP Decision Network



# Decision Networks and MDPs

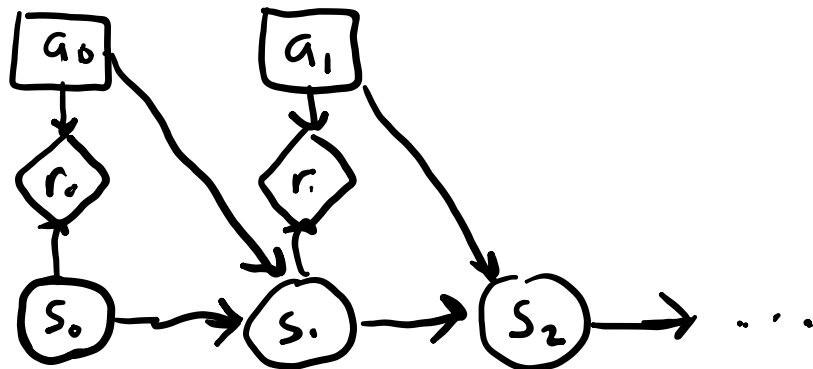
Decision Network



MDP **Dynamic** Decision Network

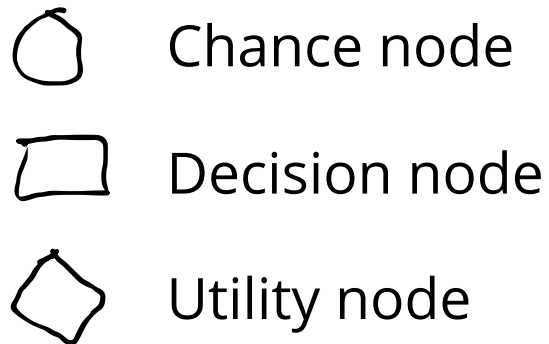


MDP Decision Network

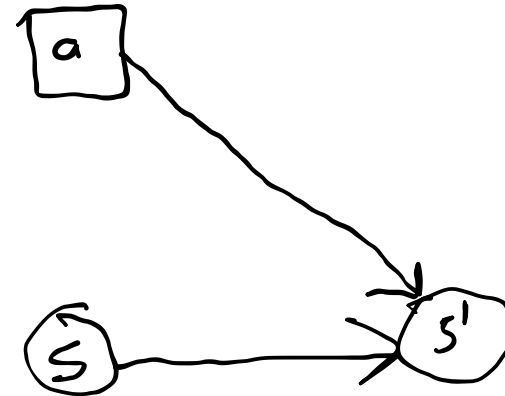


# Decision Networks and MDPs

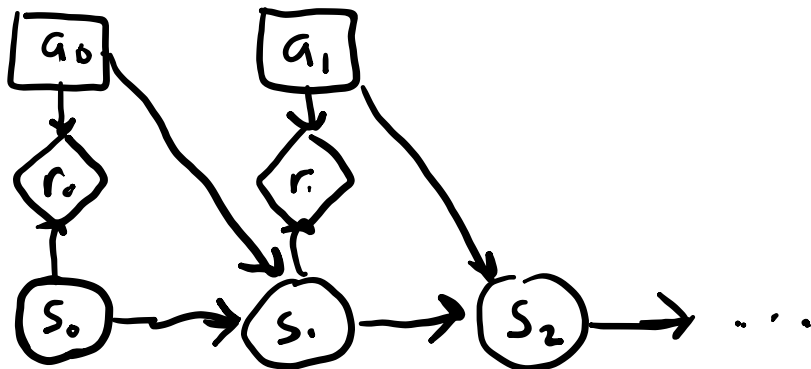
Decision Network



MDP **Dynamic** Decision Network

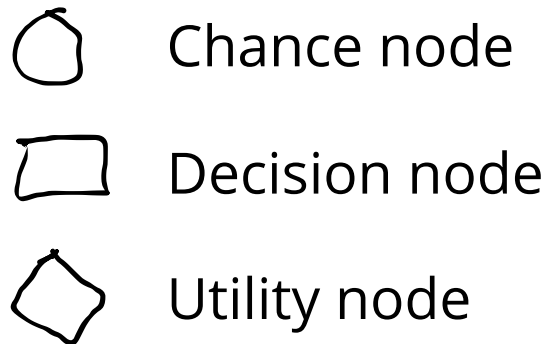


MDP Decision Network

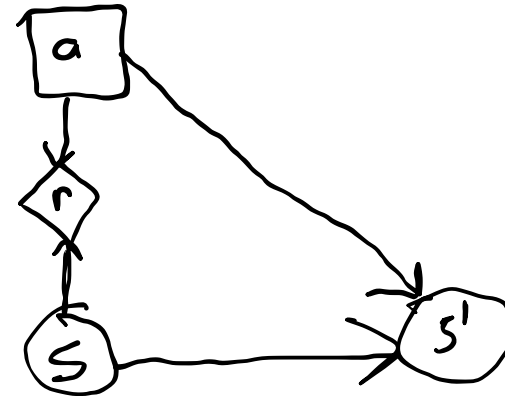


# Decision Networks and MDPs

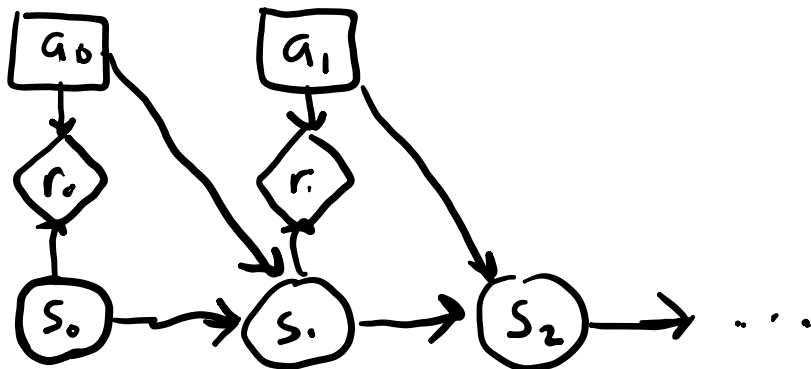
Decision Network



MDP **Dynamic** Decision Network

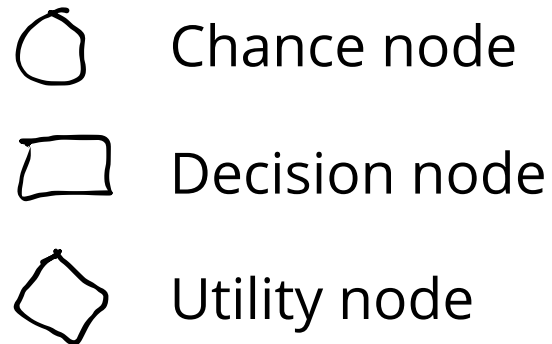


MDP Decision Network

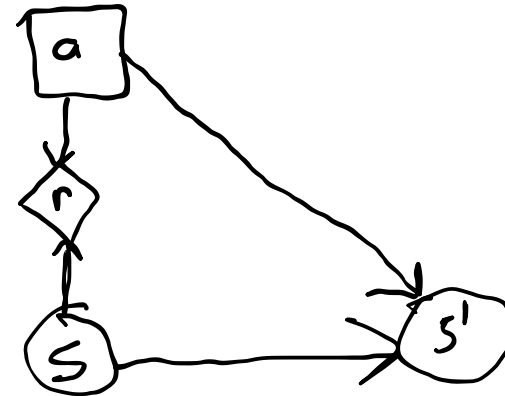


# Decision Networks and MDPs

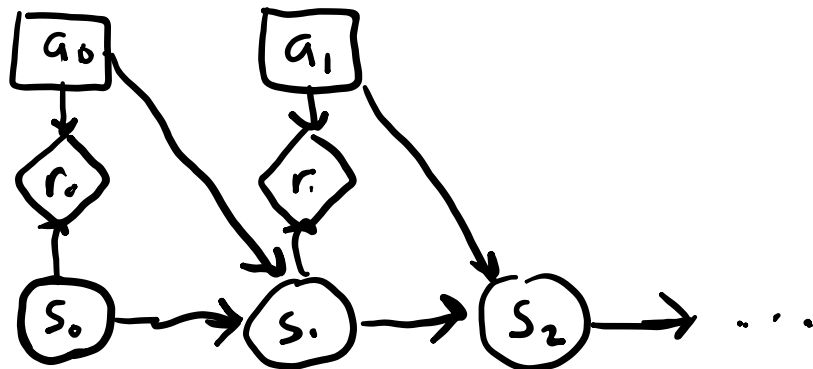
Decision Network



MDP **Dynamic** Decision Network



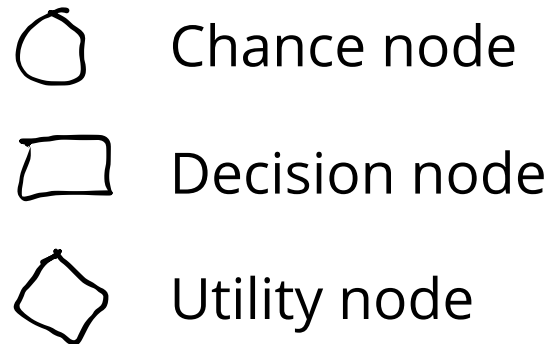
MDP Decision Network



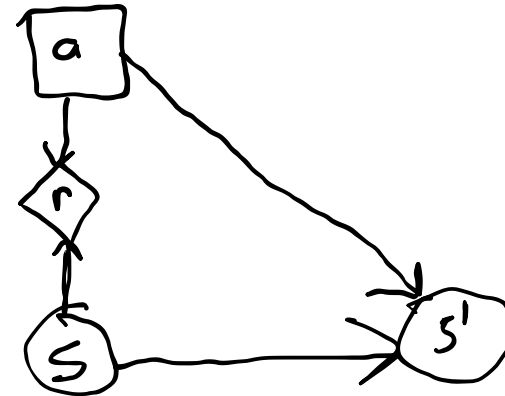
MDP Optimization problem

# Decision Networks and MDPs

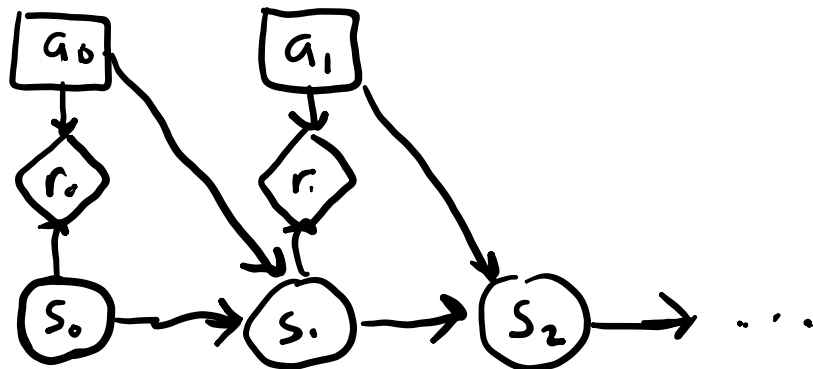
Decision Network



MDP **Dynamic** Decision Network



MDP Decision Network



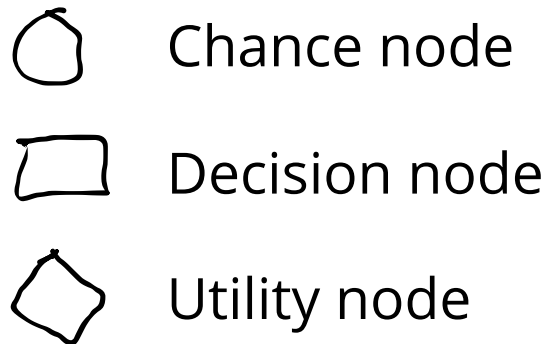
MDP Optimization problem

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

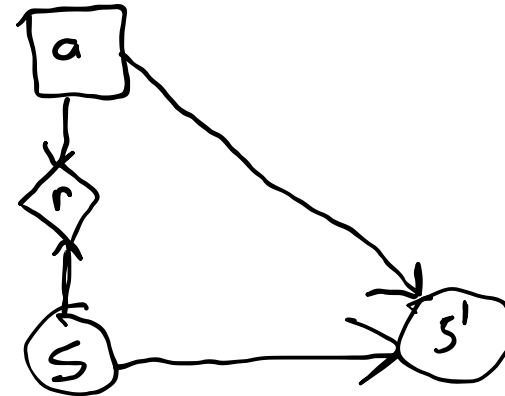


# Decision Networks and MDPs

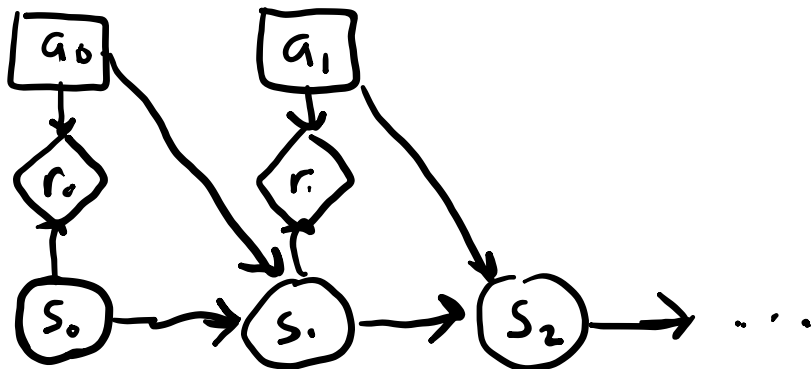
Decision Network



MDP **Dynamic** Decision Network



MDP Decision Network



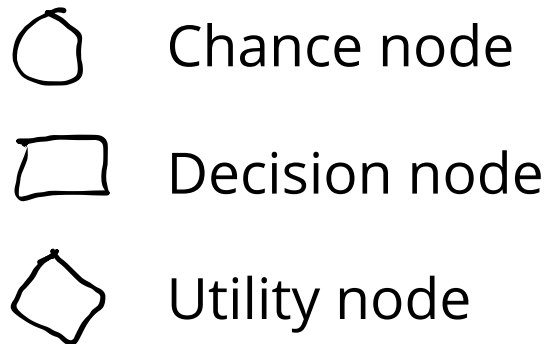
MDP Optimization problem

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

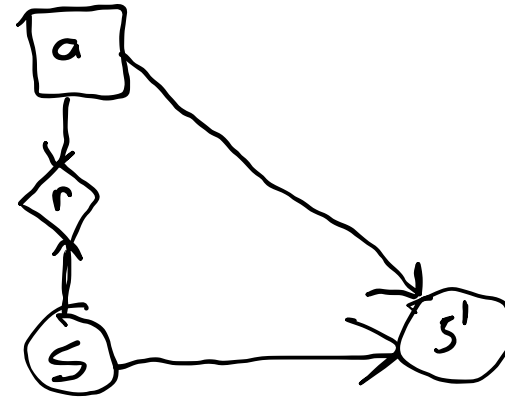
Not well formulated!

# Decision Networks and MDPs

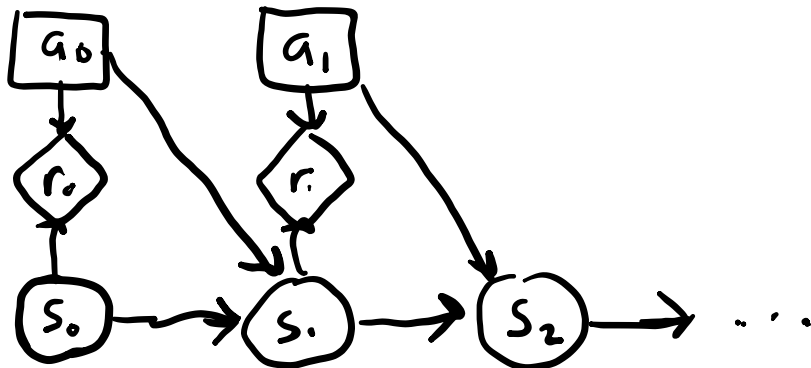
Decision Network



MDP **Dynamic** Decision Network



MDP Decision Network



MDP Optimization problem

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

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

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

2. Average reward

# Finite MDP Objectives

1. Finite time

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

2. Average reward

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

# Finite MDP Objectives

1. Finite time

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

2. Average reward

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

3. Discounting



# Finite MDP Objectives

1. Finite time

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

2. Average reward

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

3. Discounting

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

# Finite MDP Objectives

1. Finite time

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

2. Average reward

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

3. Discounting

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

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

# Finite MDP Objectives

1. Finite time

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

2. Average reward

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

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

# Finite MDP Objectives

1. Finite time

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

2. Average reward

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

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

# Finite MDP Objectives

1. Finite time

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

2. Average reward

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

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

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

2. Average reward

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

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

4. Terminal States

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

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

2. Average reward

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

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

4. Terminal States

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

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

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

2. Average reward

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

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

4. Terminal States

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

then

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



# MDP "Tuple Definition"

# MDP "Tuple Definition"

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

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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



# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

- $S$  (state space) - set of all possible states  $\{1, 2, 3\}$   
 $\{\text{healthy, pre-cancer, cancer}\}$
- $A$  (action space) - set of all possible actions  $\{1, 2, 3\}$   
 $\{\text{test, wait, treat}\}$
- $T$  (transition distribution) - explicit or implicit ("generative")  
model of how the state changes

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

- $S$  (state space) - set of all possible states  $\{1, 2, 3\}$   
 $\{\text{healthy, pre-cancer, cancer}\}$
- $A$  (action space) - set of all possible actions  $\{1, 2, 3\}$   
 $\{\text{test, wait, treat}\}$
- $T$  (transition distribution) - explicit or implicit ("generative")  
model of how the state changes  $T(s' \mid s, a)$

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

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

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

- $S$  (state space) - set of all possible states  $\{1, 2, 3\}$   
 $\{\text{healthy, pre-cancer, cancer}\}$
- $A$  (action space) - set of all possible actions  $\{1, 2, 3\}$   
 $\{\text{test, wait, treat}\}$
- $T$  (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  $R(s, a, s')$
- $\gamma$ : discount factor

"Generative Model": Alternative to  $T$  and  $R$

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

# MDP "Tuple Definition"

$(S, A, T, R, \gamma)$  (and  $b$  and/or  $S_T$  in some contexts)

- $S$  (state space) - set of all possible states  
 $\{1, 2, 3\}$   
 $\{\text{healthy, pre-cancer, cancer}\}$
  - $A$  (action space) - set of all possible actions  
 $\{1, 2, 3\}$   
 $\{\text{test, wait, treat}\}$
  - $T$  (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  
 $R(s, a, s')$
  - $\gamma$ : 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.

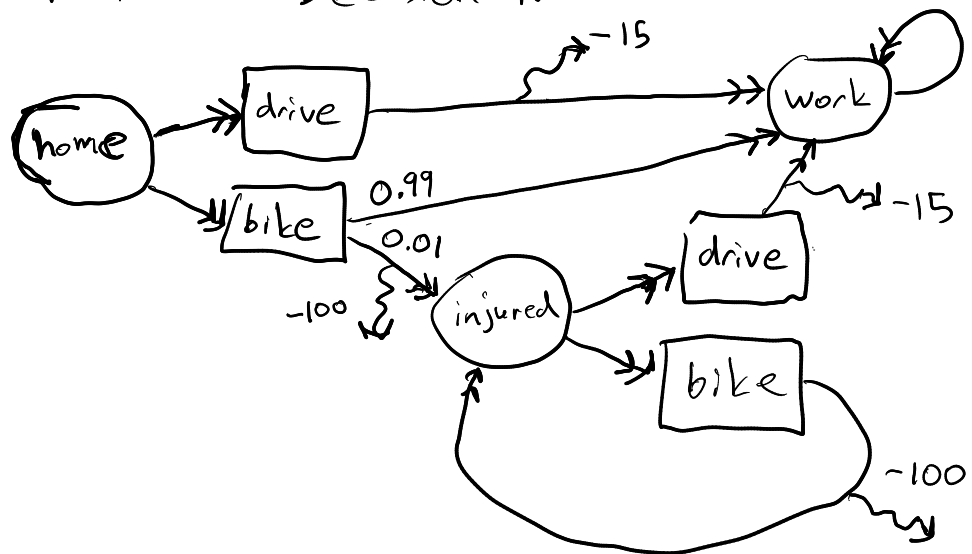
- If you drive, you will have to pay \$15 for parking; biking is free.

# 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.

State transition diagram  
Not a Decision Network



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

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

$$T^{\text{drive}} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$\begin{matrix} h \\ i \\ w \end{matrix}$ 
 $\begin{matrix} h \\ i \\ w \end{matrix}$

$$T^{\text{bike}} = \begin{bmatrix} 0 & 0.01 & 0.99 \\ 0 & 1.0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$\begin{matrix} h \\ i \\ w \end{matrix}$ 
 $\begin{matrix} h \\ i \\ w \end{matrix}$

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

$$\gamma = 0.99$$

$$S_{\text{term}} = \{\text{work}\}$$

$$b(s) = \begin{cases} 1 & \text{if } s = \text{home} \\ 0 & \text{o.w.} \end{cases}$$

# Policies and Simulation

# Policies and Simulation

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

# 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.



# Policies and Simulation

- A *policy*, denoted with  $\pi(a_t | s_t)$ , 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

$$\underline{P(s' | s)} = \sum_a T(s' | s, a) \pi(a | s)$$

or

$$P(s' | s) = \cancel{\sum_a} T(s' | s, \pi(s))$$

# 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:

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



# 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:

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

## Algorithm: Rollout Simulation

Inputs: MDP  $(S, A, R, T, \gamma, b)$  (only need generative model,  $G$ ), Policy  $\pi$ , horizon  $H$

Outputs: Utility estimate  $\hat{u}$

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

$\hat{u} \leftarrow 0$

for  $t$  in  $0 \dots H - 1$

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

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

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

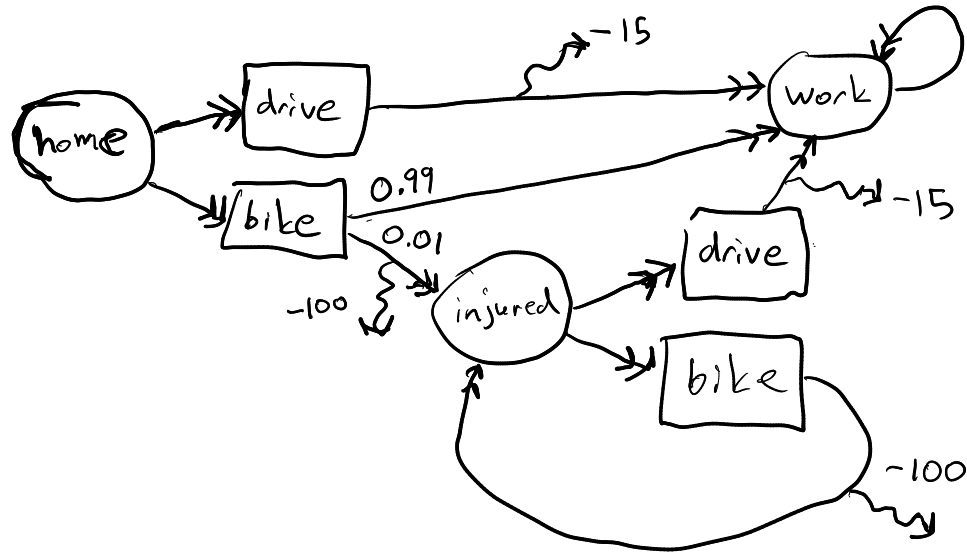
$s \leftarrow s'$

return  $\hat{u}$

# Break

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

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



$$U(\text{bike from home}) = 0.99 \times 0 + 0.01(-100 + -15) = -1.15$$

$$U(\text{drive from home}) = -15$$

# Naive Policy Evaluation

Naive Policy Evaluation not on Exam

# Naive Policy Evaluation

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

/   
  $a_t = \pi(s_t)$

Naive Policy Evaluation not on Exam

# Naive Policy Evaluation

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

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

# Naive Policy Evaluation

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

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

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

Naive Policy Evaluation not on Exam

# Naive Policy Evaluation

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

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

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

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

Naive Policy Evaluation not on Exam

# Monte Carlo Policy Evaluation

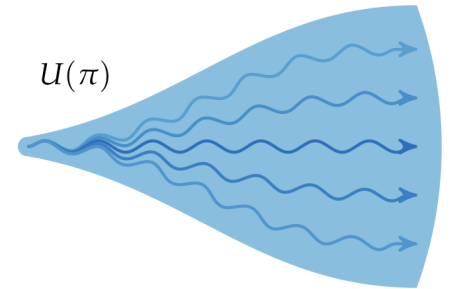


# Monte Carlo Policy Evaluation

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

# Monte Carlo Policy Evaluation

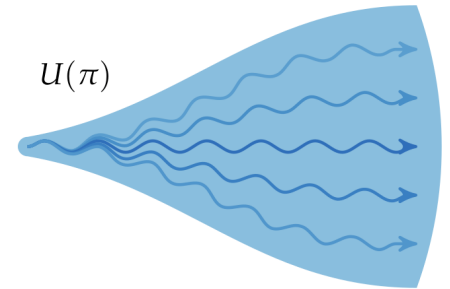
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



# Monte Carlo Policy Evaluation

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

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

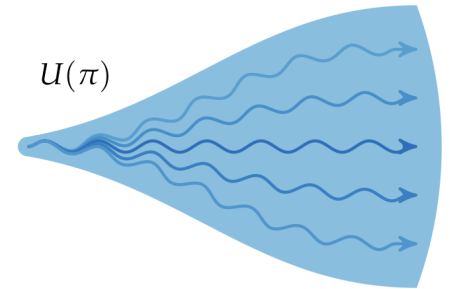


# Monte Carlo Policy Evaluation

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

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

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



# Monte Carlo Policy Evaluation

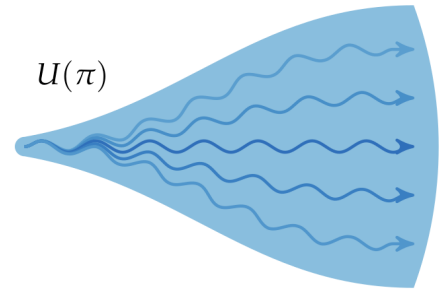
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

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

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

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

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



# Monte Carlo Policy Evaluation

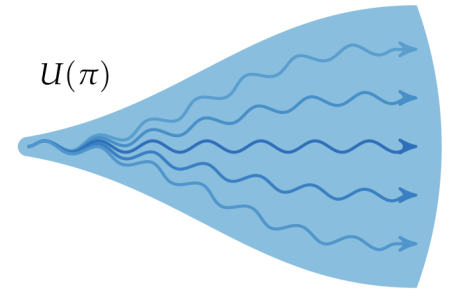
- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*

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

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

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

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



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

$$\begin{aligned} \text{Var}(\bar{u}_m) &= \text{Var}\left(\frac{1}{m} \sum_i \hat{u}^{(i)}\right) \\ &= \frac{1}{m^2} \text{Var}\left(\sum_i \hat{u}^{(i)}\right) \\ &= \frac{1}{m^2} \sum_{i=1}^m \text{Var}(\hat{u}^{(i)}) \end{aligned}$$

$$\text{Var}(\bar{u}_m) = \frac{1}{m} \text{Var}(\hat{u}^{(i)})$$

$$\text{Standard Error of Mean } SEM \equiv \frac{1}{\sqrt{m}} \text{std}(\hat{u})$$

# Monte Carlo Policy Evaluation

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

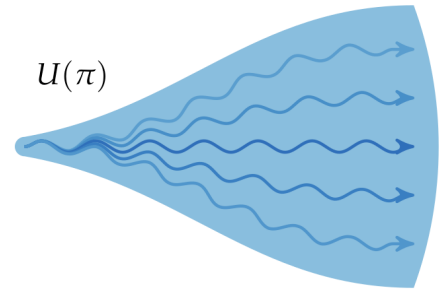
Let  $\tau = (s_0, a_0, r_0, s_1, \dots, s_T)$  be a *trajectory* of the MDP

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

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?