Artificial Intelligence Foundation – JC3001

Lecture 35: Stochastic Planning - II

Prof. Aladdin Ayesh (aladdin.ayesh@abdn.ac.uk)

Dr. Binod Bhattarai (binod.bhattarai@abdn.ac.uk)

Dr. Gideon Ogunniye, (g.ogunniye@abdn.ac.uk)

October 2025

Material adapted from:
Russell and Norvig (AIMA Book): Chapter 17
Sutton and Barto (Reinforcement Learning: An Introduction 2nd ed.)
Sebastian Thrun (Stanford University / Udacity)

UNIVERSITY OF ABERDEEN

Course Progression

- Part 1: Introduction
  1. Introduction to AI ✓
  2. Agents ✓
- Part 2: Problem-solving
  1. Search 1: Uninformed Search ✓
  2. Search 2: Heuristic Search ✓
  3. Search 3: Local Search ✓
  4. Search 4: Adversarial Search ✓
- Part 3: Reasoning and Uncertainty
  1. Reasoning 1: Constraint Satisfaction ✓
  2. Reasoning 2: Logic and Inference ✓
  3. Probabilistic Reasoning 1: BNs ✓
  4. Probabilistic Reasoning 2: HMMs ✓
- Part 4: Planning
  1. Planning 1: Intro and Formalism ✓
  2. Planning 2: Algorithms & Heuristics ✓
  3. Planning 3: Hierarchical Planning ✓
  4. **Planning 4: Stochastic Planning**
- Part 5: Learning
  1. Learning 1: Intro to ML
  2. Learning 2: Regression
  3. Learning 3: Neural Networks
  4. Learning 4: Reinforcement Learning
- Part 6: Conclusion
  1. Ethical Issues in AI
  2. Conclusions and Discussion

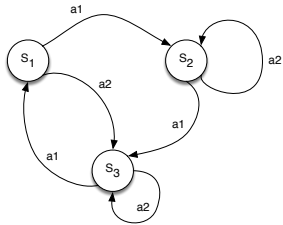
Objectives

- Planning with Utilities
- Stochastic Planning Formalisms
- Dynamic Programming Algorithms for Stochastic Planning

Outline

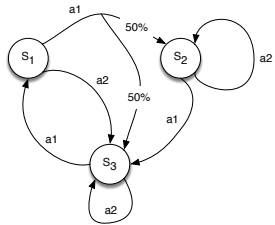
- Markov Decision Processes

- We have examined how a rational agent can represent its preferences over outcomes as a single number, utility
- We briefly examined how multiple attributes can affect the utility evaluation
- How do we decide how to act in sequential decision problems?

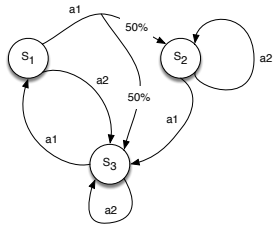


- Such a model, in a fully observable environment, is called a Markov Decision Process

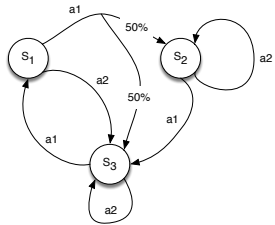
Markov Decision Process (MDP)



- Such a model, in a fully observable environment, is called a Markov Decision Process
- An MDP is defined in terms of
  1. An initial state S_0
  2. A transition model $T(s, a, s')$
  3. A reward function $R(s)$



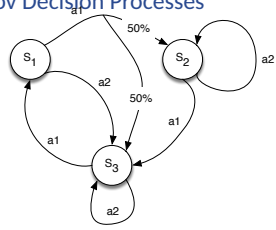
- Such a model, in a fully observable environment, is called a Markov Decision Process
- An MDP is defined in terms of
  1. An initial state S_0
  2. A transition model $T(s, a, s')$ — $P(s' \mid a, s)$ (Markovian)
  3. A reward function $R(s)$ — sometimes expressed as $R(a, s)$



- Such a model, in a fully observable environment, is called a Markov Decision Process
- An MDP is defined in terms of
  1. An initial state S_0
  2. A transition model $T(s, a, s') - P(s' \mid a, s)$ (Markovian)
  3. A reward function $R(s)$ — sometimes expressed as $R(a, s)$
- A solution to a MDP must specify what the agent should do for any state. Such a solution is called a policy

Markov Decision Process (MDP) - In Sutton and Barton's Book

1 Markov Decision Processes



- Such a model, in a fully observable environment, is called a Markov Decision Process
- An MDP is defined in terms of:
  1. A finite set of states $\mathcal{S}$
  2. A finite set of actions $\mathcal{A}$
  3. A markovian transition model $T(s, a, s') = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a)$
  4. A reward function $R(s) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
  5. A discount factor $\gamma \in [0, 1]$
- A solution to a MDP must specify what the agent should do for any state. Such a solution is called a policy

Definition

The return G_t is the total discounted reward from time-step t.

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- The discount $\gamma \in [0, 1]$ is the present value of future rewards
- The value of receiving reward R after $k + 1$ time-steps is $\gamma^k R$.
- This values immediate reward above delayed reward.
  - γ close to 0 leads to "myopic" evaluation
  - γ close to 1 leads to "far-sighted" evaluation

Most Markov reward and decision processes are discounted. Why?

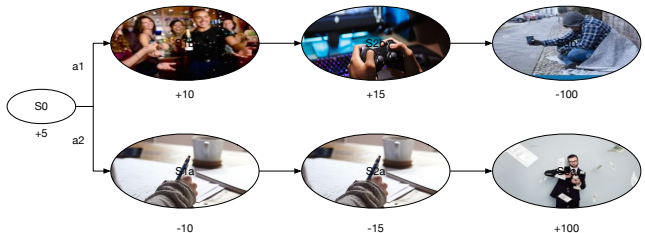
Why discount?

Most Markov reward and decision processes are discounted. Why?

- Mathematically convenient to discount rewards
- Avoids infinite returns in cyclic Markov processes
- Uncertainty about the future may not be fully represented
- If the reward is financial, immediate rewards may earn more interest than delayed rewards
- Animal/human behaviour shows preference for immediate reward
- It is sometimes possible to use undiscounted Markov reward processes (i.e. $\gamma = 1$), e.g. if all sequences terminate.

Intuition on sequential decision making

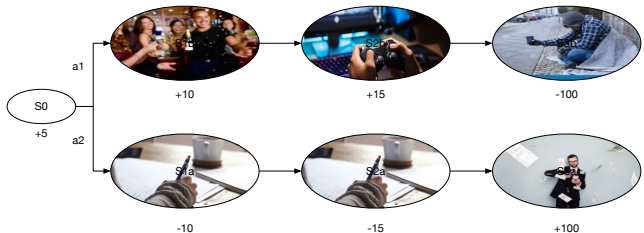
$V(s_0) =$ Greater possible value when we start from s_0



$$V(s_0) = 5 + \begin{cases} a_1 : 10 + 15 + (-100) \\ a_2 : -10 + (-15) + 100 \end{cases}$$
$$= 5 + (-10) + (-15) + 100 = 80$$

Intuition on sequential decision making

$V(s_0) =$ Greater possible value when we start from s_0

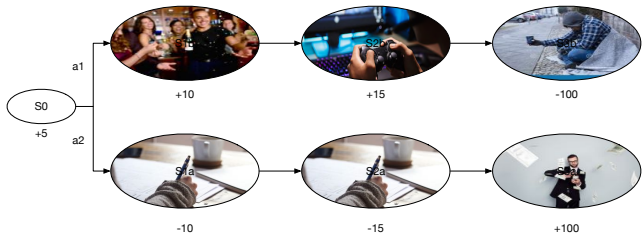


$$V(s_0) = 5 + \begin{cases} a_1 : 10 + 15 + (-100) \\ a_2 : -10 + (-15) + 100 \end{cases}$$
$$= 5 + (-10) + (-15) + 100 = 80$$

If these rewards take ten years to achieve, are they worth the same?

Intuition on sequential decision making

$V(s_0) =$ Greater possible value when we start from s_0



$$V(s_0) = 5 + \begin{cases} a_1 : 10 + 15 + (-100) \\ a_2 : -10 + (-15) + 100 \end{cases}$$

$$= 5 + (-10) + (-15) + 100 = 80$$

$$a_1 : 10 \quad + \gamma 15 \quad + \gamma^2(-100)$$

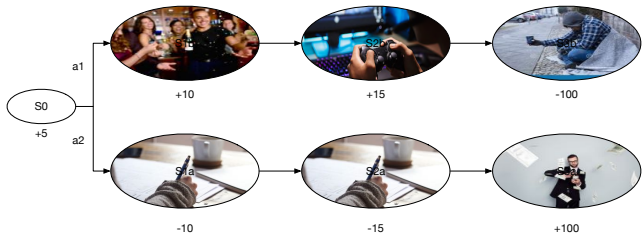
$$a_2 : -10 \quad + \gamma(-15) \quad + \gamma^2 100$$

If these rewards take ten years to achieve, are they worth the same?

$\gamma = 0$

Intuition on sequential decision making

$V(s_0) =$ Greater possible value when we start from s_0



$$V(s_0) = 5 + \begin{cases} a_1 : 10 + 15 + (-100) \\ a_2 : -10 + (-15) + 100 \end{cases}$$

$$= 5 + (-10) + (-15) + 100 = 80$$

$$a_1 : 10 \quad + \gamma 15 \quad + \gamma^2(-100)$$

$$a_2 : -10 \quad + \gamma(-15) \quad + \gamma^2 100$$

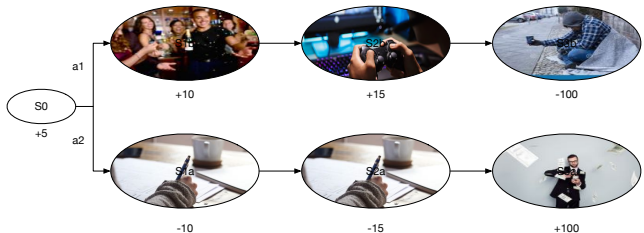
If these rewards take ten years to achieve, are they worth the same?

$\gamma = 0$

$\gamma = 1$

Intuition on sequential decision making

$V(s_0) =$ Greater possible value when we start from s_0



$$V(s_0) = 5 + \begin{cases} a_1 : 10 + 15 + (-100) \\ a_2 : -10 + (-15) + 100 \end{cases}$$

$$= 5 + (-10) + (-15) + 100 = 80$$

$$a_1 : 10 \quad + \gamma 15 \quad + \gamma^2(-100)$$

$$a_2 : -10 \quad + \gamma(-15) \quad + \gamma^2 100$$

If these rewards take ten years to achieve, are they worth the same?

$\gamma = 0$

$\gamma = 1$

$\gamma = (0, 1]$

- We denote a policy by π
- $\pi(s)$ identifies the action recommended by the policy in state s
- A complete policy tells an agent what to do no matter the outcome of an action
  - It explicitly represents the agent function and therefore describes a simple reflex agent, computed from the information used for a utility based agent

- We denote a policy by π
- $\pi(s)$ identifies the action recommended by the policy in state s
- A complete policy tells an agent what to do no matter the outcome of an action
  - It explicitly represents the agent function and therefore describes a simple reflex agent, computed from the information used for a utility based agent
  - The quality of a policy is measured by the expected utility of the possible environment histories generated by executing that policy

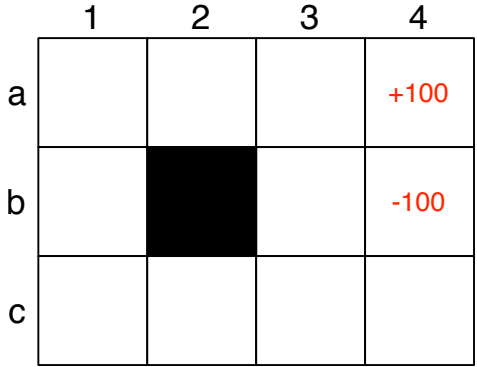
- We denote a policy by π
- $\pi(s)$ identifies the action recommended by the policy in state s
- A complete policy tells an agent what to do no matter the outcome of an action
  - It explicitly represents the agent function and therefore describes a simple reflex agent, computed from the information used for a utility based agent
  - The quality of a policy is measured by the expected utility of the possible environment histories generated by executing that policy
  - An optimal policy is one that yields the highest expected utility, and is denoted π^*

Problems with conventional planning in stochastic environments

- Branching factor

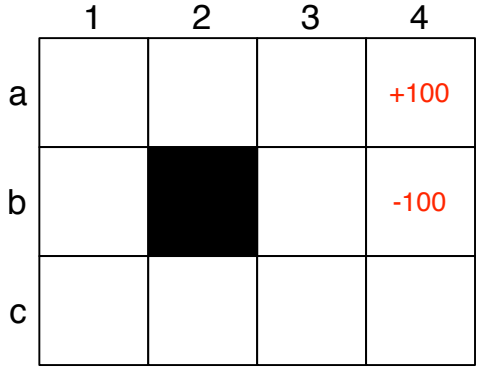
Problems with conventional planning in stochastic environments

- Branching factor
- Tree Depth



What is the optimal action when you are in:

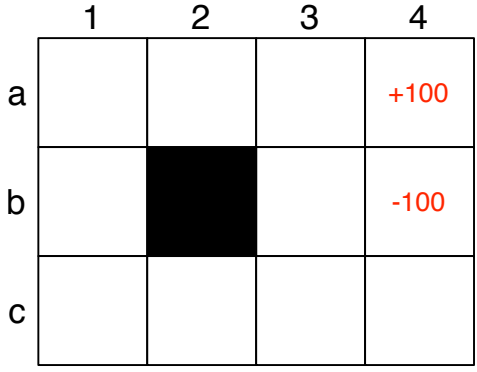
a1: N S W E



What is the optimal action when you are in:

a1: N S W E $\rightarrow$ E

c1: N S W E

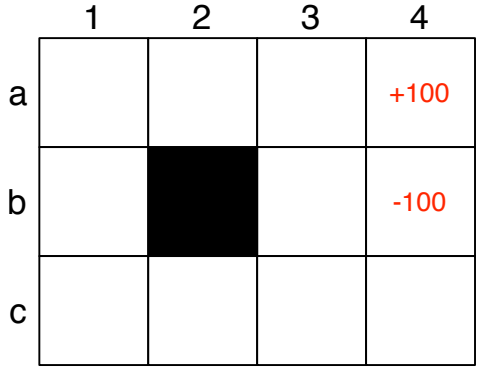


What is the optimal action when you are in:

a1: N S W E $\rightarrow$ E

c1: N S W E $\rightarrow$ N

c4: N S W E



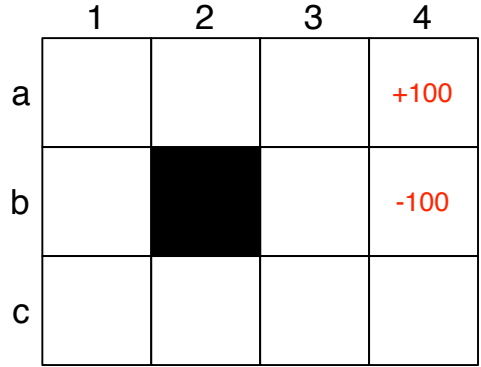
What is the optimal action when you are in:

a1: N S W E $\rightarrow$ E

c1: N S W E $\rightarrow$ N

c4: N S W E $\rightarrow$ S

b3: N S W E



This is somewhat unintuitive, why?

What is the optimal action when you are in:

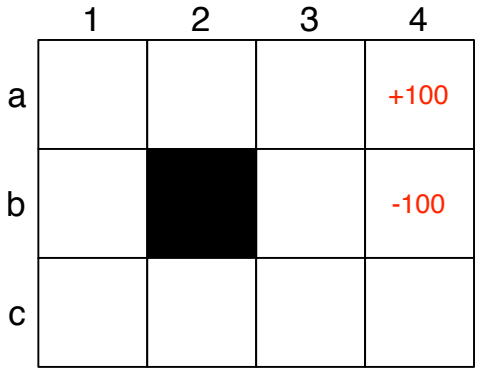
a1: N S W E $\rightarrow$ E

c1: N S W E $\rightarrow$ N

c4: N S W E $\rightarrow$ S

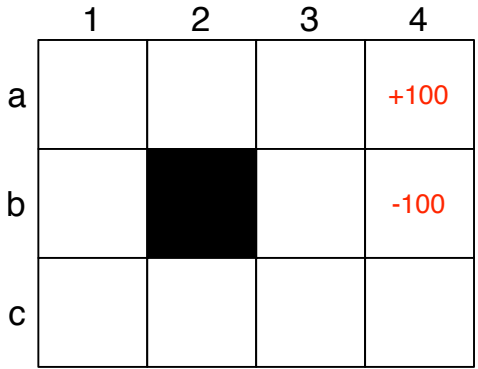
b3: N S W E $\rightarrow$ W

MDPs and Costs



$$R(s) = \begin{cases} +100 & a4 \\ -100 & b4 \\ -3 & other\ states \end{cases}$$

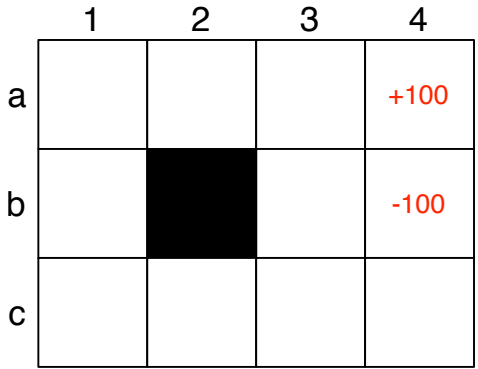
MDPs and Costs



$$R(s) = \begin{cases} +100 & a4 \\ -100 & b4 \\ -3 & other\ states \end{cases}$$

$$\sum_{t=0}^{\infty} R_t$$

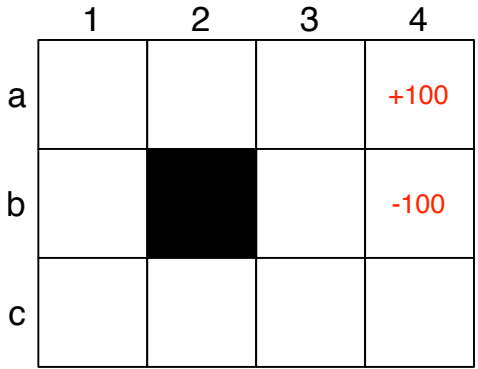
MDPs and Costs



$$R(s) = \begin{cases} +100 & a4 \\ -100 & b4 \\ -3 & other\ states \end{cases}$$

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

MDPs and Costs

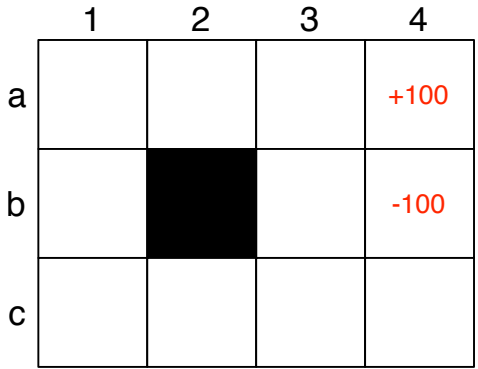


$$R(s) = \begin{cases} +100 & a4 \\ -100 & b4 \\ -3 & other\ states \end{cases}$$

$$\mathbb{E}\left[\sum_{t=0}^{\infty} R_t\right] \rightarrow max$$

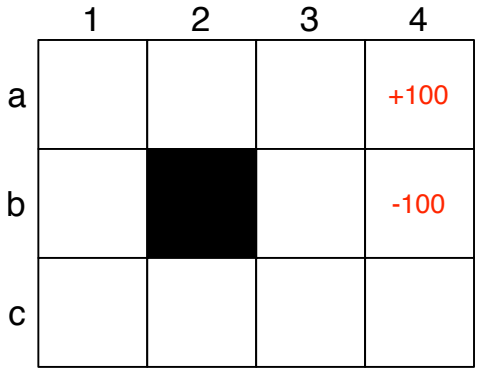
MDPs and Costs

1 Markov Decision Processes



$$R(s) = \begin{cases} +100 & a4 \\ -100 & b4 \\ -3 & other\ states \end{cases}$$

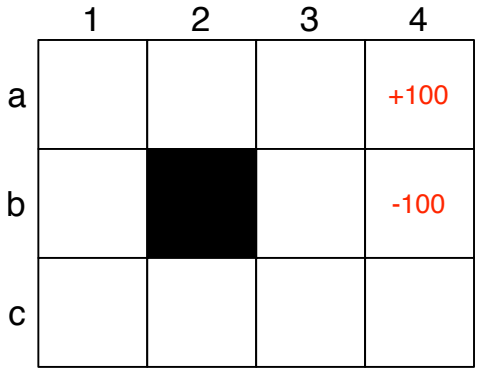
$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] \rightarrow max$$



$$R(s) = \begin{cases} +100 & a4 \\ -100 & b4 \\ -3 & other\ states \end{cases}$$

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

$\gamma =$ discount factor

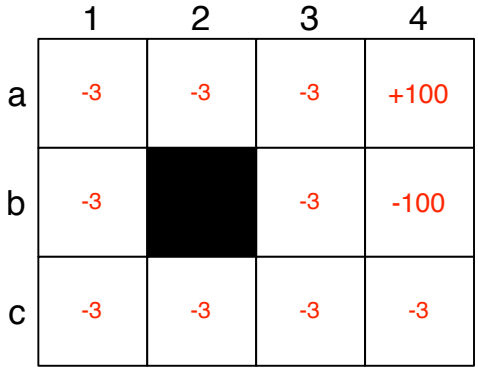


$$R(s) = \begin{cases} +100 & a4 \\ -100 & b4 \\ -3 & other\ states \end{cases}$$

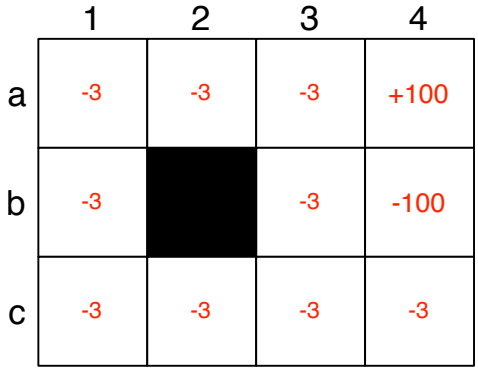
$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] \rightarrow max$$

$\gamma =$ discount factor $\gamma = 0.9$

Value Functions



$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t) \mid s_0 = s\right]$$



$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t) \mid s_0 = s\right]$$

Planning = Calculate Value Functions

Value Iteration

	1	2	3	4
a	0	0	0	+100
b	0		0	-100
c	0	0	0	0

$$V(a3, E) = ?$$

Value Iteration

	1	2	3	4
a	0	0	0	+100
b	0		0	-100
c	0	0	0	0

$$V(a3, E) = ?$$

	1	2	3	4
a	0	0	77	+100
b	0		0	-100
c	0	0	0	0

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

$$V(s) \leftarrow$$

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

$$V(s) \leftarrow V(s')$$

Value Iteration

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

$$V(s) \leftarrow \sum_{s'} P(s' \mid s, a) * V(s')$$

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

$$V(s) \leftarrow \left[\sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

$$V(s) \leftarrow \left[\gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

Value Iteration

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

$$V(a3, E) = 0.8 * 100 - 3 = 77$$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

	1	2	3	4
a	85	89	93	+100
b	81		68	-100
c	77	73	70	47

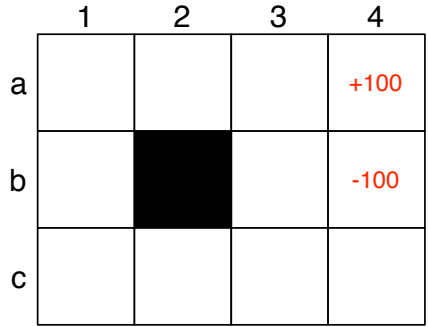
$$V(a3, E) = 0.8 * 100 - 3 = 77$$

$$V(s) \leftarrow \begin{cases} R(s) \; if \; terminal \\ [\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s')] + R(s) \end{cases}$$

Back up function

When $V(s) = [\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s')] + R(s)$ (converges)
it becomes the **Bellman Equation**

Value Iteration - Deterministic Quiz

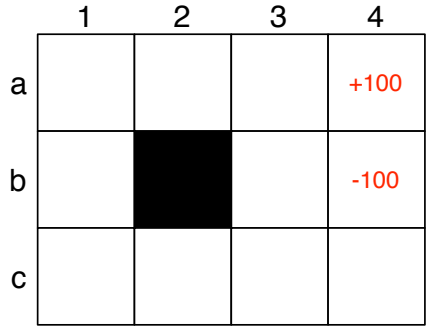


$R(s) = -3$ and $\gamma = 1$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

What are the values of the following states, given a **deterministic** T? $V(a3) = ?$

Value Iteration - Deterministic Quiz

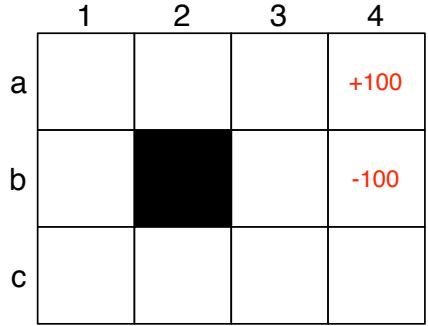


$R(s) = -3$ and $\gamma = 1$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

What are the values of the following states, given a **deterministic** T? $V(a3) = ?$

Value Iteration - Deterministic Quiz

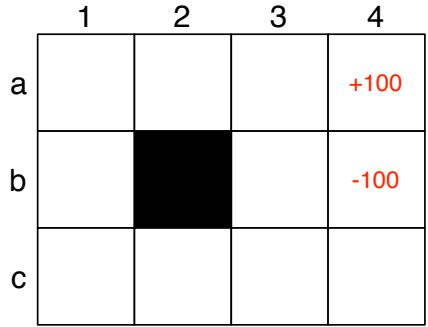


$R(s) = -3$ and $\gamma = 1$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

What are the values of the following states, given a **deterministic** T? $V(a3) = 97$
$V(b3) = ?$

Value Iteration - Deterministic Quiz



$R(s) = -3$ and $\gamma = 1$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

What are the values of the following states, given a **deterministic** T? $V(a3) = 97$

$V(b3) = 94$

$V(c1) =?$

Value Iteration - Deterministic Quiz

	1	2	3	4
a	91	94	97	+100
b	88		94	-100
c	85	88	91	

$R(s) = -3$ and $\gamma = 1$

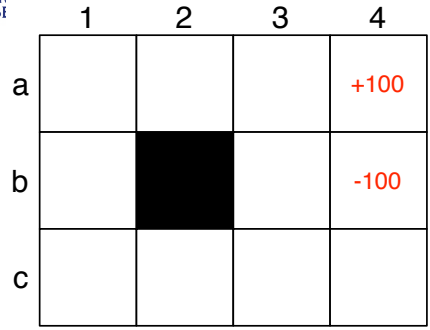
$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

What are the values of the following states, given a **deterministic** T? $V(a3) = 97$

$V(b3) = 94$

$V(c1) = 85$

Value Iteration - Stochastic Quiz



$R(s) = -3$ and $\gamma = 1$

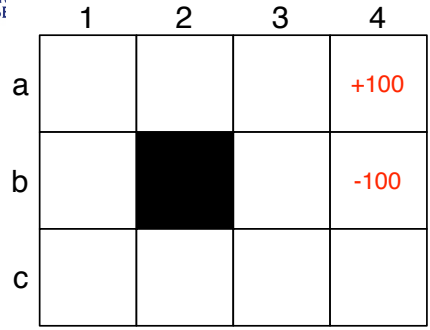
$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

What are the values of the states below, given:

- **stochastic** T, $p = 0.8$
- $V(s)$ originally 0

$V(a3) = ?$

Value Iteration - Stochastic Quiz



$R(s) = -3$ and $\gamma = 1$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

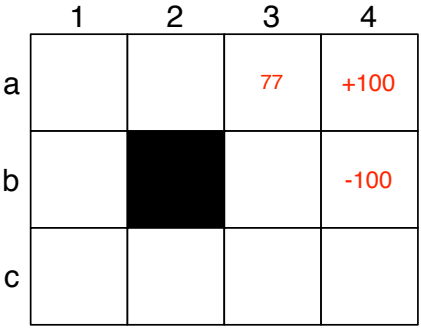
What are the values of the states below, given:

- **stochastic** T, $p = 0.8$
- $V(s)$ originally 0

$V(a3) = ?$

Value Iteration - Stochastic Quiz

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$



$R(s) = -3$ and $\gamma = 1$

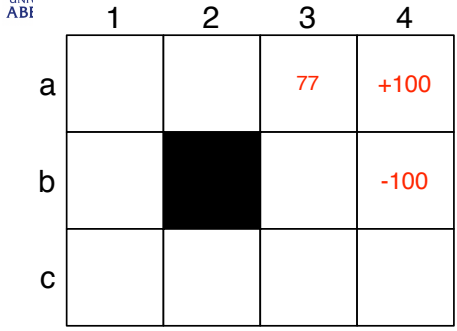
What are the values of the states below, given:

- **stochastic** T, $p = 0.8$
- $V(s)$ originally 0

$V(a3) = 77$
$V(b3) = ?$

Value Iteration - Stochastic Quiz

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s')\right] + R(s)$$



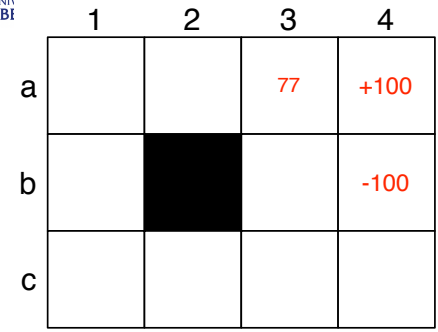
$R(s) = -3$ and $\gamma = 1$

What are the values of the states below, given:

- **stochastic** T, $p = 0.8$
- $V(s)$ originally 0

$V(a3) = 77$
$V(b3) = 48.6$

Value Iteration - Stochastic Quiz



$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s')\right] + R(s)$$

What are the values of the states below, given:

- **stochastic** T, $p = 0.8$
- $V(s)$ originally 0

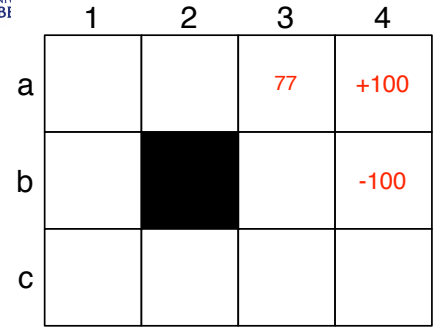
$V(a3) = 77$
$V(b3) = 48.6$

$R(s) = -3$ and $\gamma = 1$

N: $[0.8 * 77 + 0.1 * (-100) + 0.1 * 0] - 3 = 48.6$

Value Iteration - Stochastic Quiz

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s')\right] + R(s)$$



What are the values of the states below, given:

- **stochastic** T, $p = 0.8$
- $V(s)$ originally 0

$V(a3) = 77$
$V(b3) = 48.6$

$R(s) = -3$ and $\gamma = 1$

N: $[0.8 * 77 + 0.1 * (-100) + 0.1 * 0] - 3 = 48.6$

W: $[0.1 * 77 + 0.8 * 0 + 0.1 * 0] - 3 = 4.7$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

$$V(s) \leftarrow \left[\max_a \gamma \sum_{s'} P(s' \mid s, a) * V(s') \right] + R(s)$$

$$\pi(s) = \arg\max_a \sum_{s'} P(s' \mid s, a) * V(s')$$

- In order to specify a utility/value function, we must answer several questions
  - Is there a finite or infinite horizon for decision making
  - Finite horizon: after N time steps have passed, no additional utility can be gained, $U_h([s_0, s_1, \ldots, s_{N+k}]) = U_h([s_0, s_1, \ldots, s_N])$
  - Given the grid world above, starting at (3,1) the agent must head for +1 if $N = 3$, so optimal decision is "Up". With a longer horizon, the safe route can be taken instead
  - Given a finite horizon, the optimal action in a given state could change over time; such an optimal policy is non-stationary
  - If the optimal action depends only on the current state, then the optimal policy is stationary

Stochastic Planning Summary

- Stochastic Planning
- Utilities and the Expectation Operator
- Markov Decision Processes and Partially Observable MDPs
  - Formalism
  - Policies

Any Questions.