ELTE | FACULTY OF INFORMATICS

MARKOV DECISION PROCESSES

Deep Reinforcement Learning

Balázs Nagy, PhD

ELTE | IK
DEPARTMENT OF ARTIFICIAL INTELLIGENCE

Finite Markov Decision Processes

- Classical formalization of sequential decision making
- Mathematically idealized form of the RL problem
- Actions influence
  - Immediate reward
  - Subsequent situation (state)
- Delayed reward

Bandit problem

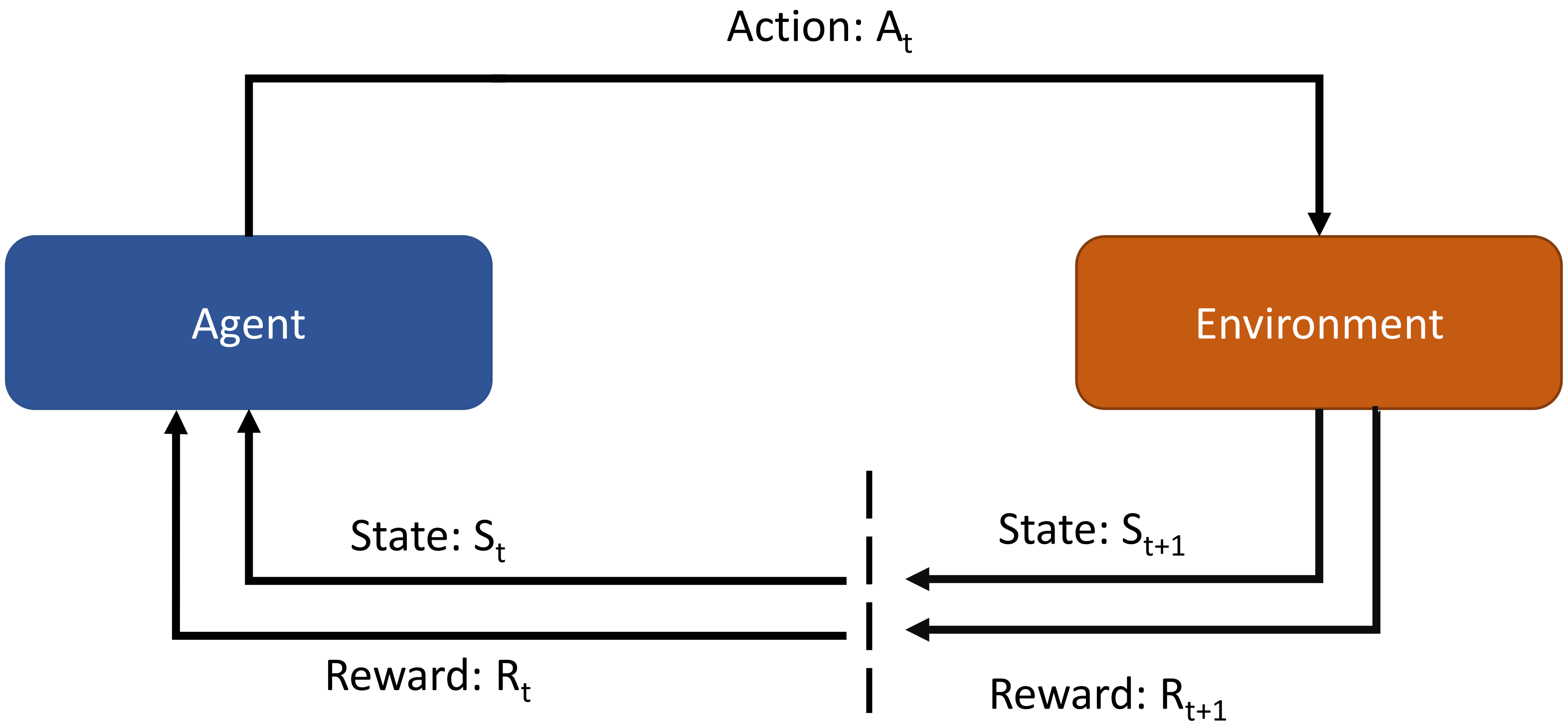
$q_*(a)$

→

MDP

$q_*(s, a)$ or $v_*(s)$

Agent – Environment Interface



Agent: Lerner and decision maker

Environment: The thing the Agent interacts with. Responding to the agent’s action. Presenting new states and rewards

Finite MDP

- MDP + Agent produces a sequence of elements (trajectory)

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$

- Discrete timesteps
  (can be extended later to continuous time space)

Main Elements:
Agent
Environment

$$\left.\begin{array}{ll}\textbf{State} & S_t \\ \textbf{Action} & A_t \\ \textbf{Rewards} & R_t\end{array}\right\} \text{ Finite number of elements}$$

S_t and R_t have a well-defined discrete probability distribution dependent only on the preceding state and action

Dynamic of the MDP

- Transition probability:

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

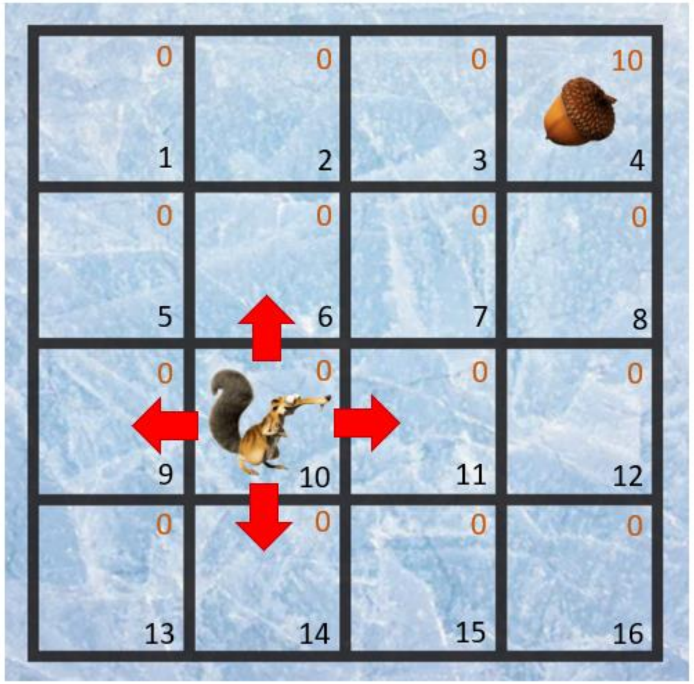
Conditional probability

- Note

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

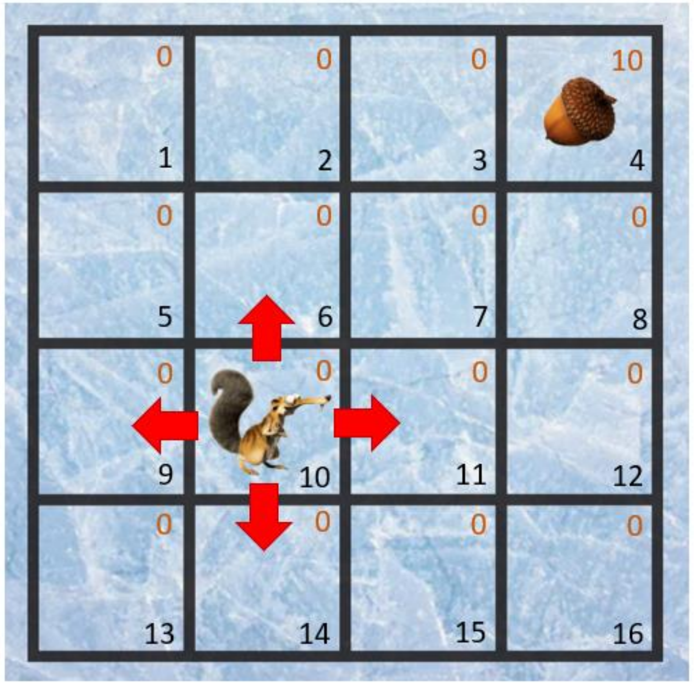
Example

- State space:
- Action space:
- Reward:



Example

- State space: [1,2,...,16]
- Action space: [up, down, right, left]
- Reward: 0 or 10



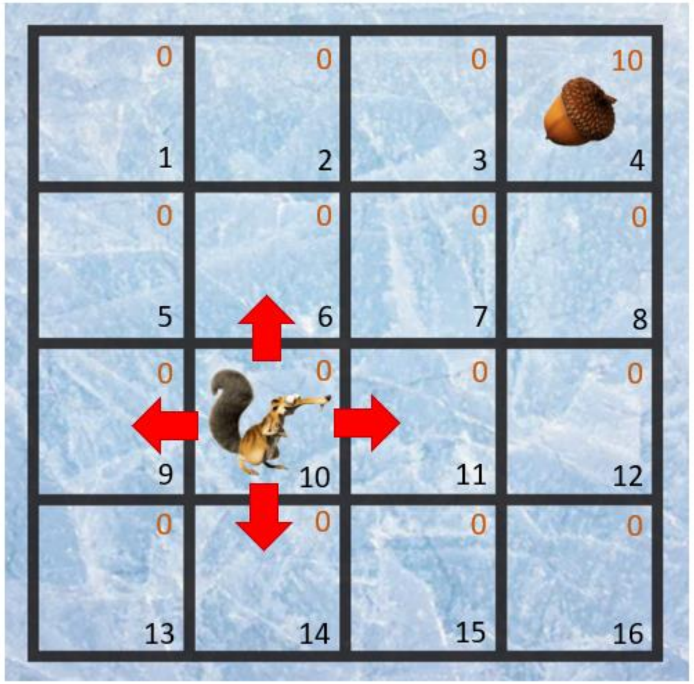
Example

- State space: [1,2,…,16]
- Action space: [up, down, right, left]
- Reward: 0 or 10

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

Deterministic

$p(\ 6, 0 \mid 10, \text{up})$
$p(11, 0 \mid 10, \text{up})$
$p(\ 9, 0 \mid 10, \text{up})$
$p(14, 0 \mid 10, \text{up})$
$p(\ 8, 0 \mid 10, \text{up})$
$p(\ 4, 0 \mid 10, \text{up})$
$p(4, 10 \mid 10, \text{up})$



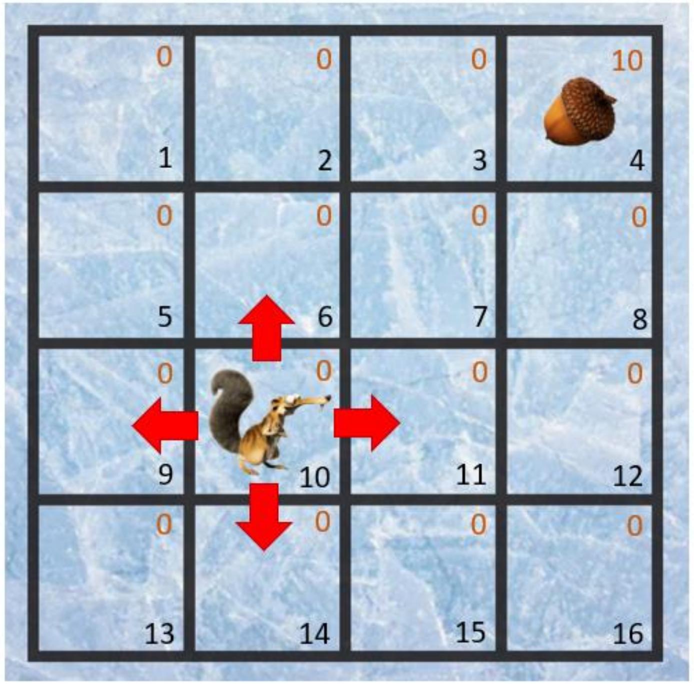
Example

- State space: [1,2,...,16]
- Action space: [up, down, right, left]
- Reward: 0 or 10

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

Deterministic

$p(\ 6, 0 \mid 10, \text{up}) = 1$
$p(11, 0 \mid 10, \text{up}) = 0$
$p(\ 9, 0 \mid 10, \text{up}) = 0$
$p(14, 0 \mid 10, \text{up}) = 0$
$p(\ 8, 0 \mid 10, \text{up}) = 0$
$p(\ 4, 0 \mid 10, \text{up}) = 0$
$p(4, 10 \mid 10, \text{up}) = 0$

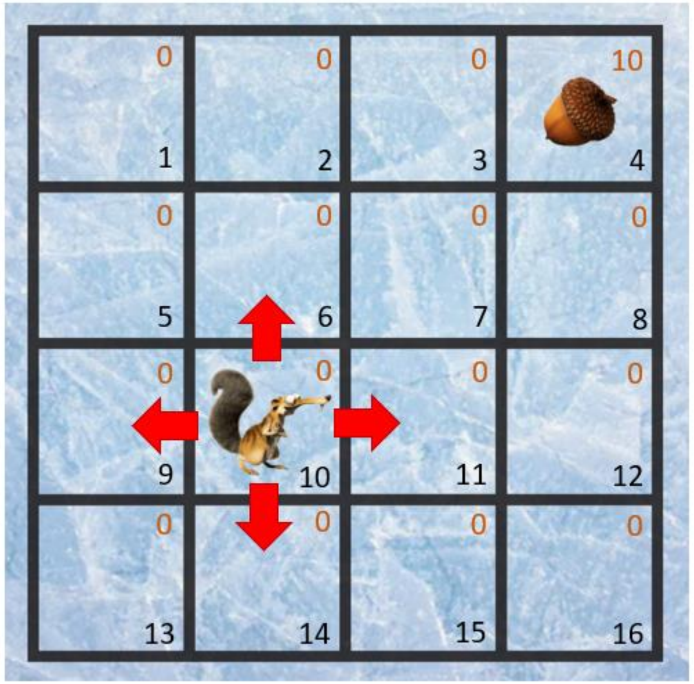


Example

- State space: [1,2,...,16]
- Action space: [up, down, right, left]
- Reward: 0 or 10

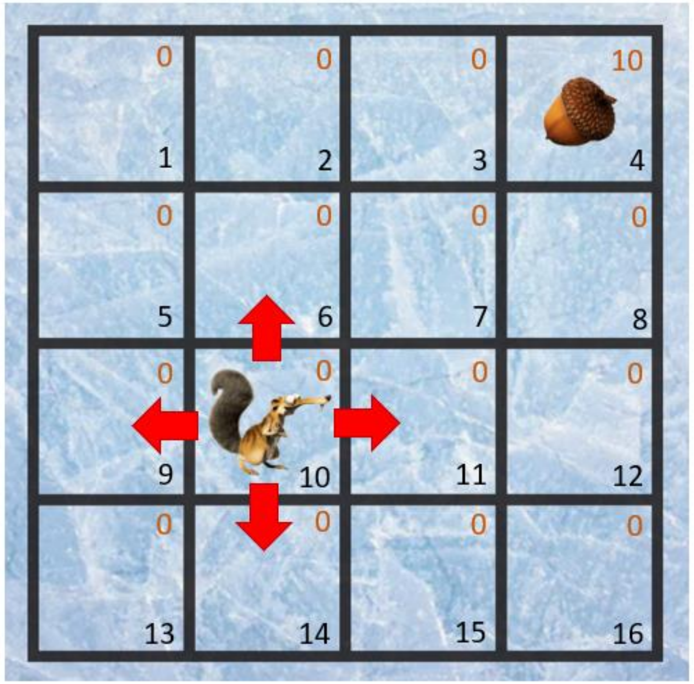
$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

Deterministic		***Stochastic***
$p(6, 0 \mid 10, \text{up})$	$= 1$	$p(6, 0 \mid 10, \text{up})$
$p(11, 0 \mid 10, \text{up})$	$= 0$	$p(11, 0 \mid 10, \text{up})$
$p(9, 0 \mid 10, \text{up})$	$= 0$	$p(9, 0 \mid 10, \text{up})$
$p(14, 0 \mid 10, \text{up})$	$= 0$	$p(14, 0 \mid 10, \text{up})$
$p(8, 0 \mid 10, \text{up})$	$= 0$	$p(8, 0 \mid 10, \text{up})$
$p(4, 0 \mid 10, \text{up})$	$= 0$	$p(4, 0 \mid 10, \text{up})$
$p(4, 10 \mid 10, \text{up})$	$= 0$	$p(4, 10 \mid 10, \text{up})$



Example

- State space: [1,2,...,16]
- Action space: [up, down, right, left]
- Reward: 0 or 10



$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

Deterministic		***Stochastic***	
$p(6, 0 \mid 10, \text{up})$	= 1	$p(6, 0 \mid 10, \text{up})$	= 0.8
$p(11, 0 \mid 10, \text{up})$	= 0	$p(11, 0 \mid 10, \text{up})$	= 0.1
$p(9, 0 \mid 10, \text{up})$	= 0	$p(9, 0 \mid 10, \text{up})$	= 0.1
$p(14, 0 \mid 10, \text{up})$	= 0	$p(14, 0 \mid 10, \text{up})$	= 0.0
$p(8, 0 \mid 10, \text{up})$	= 0	$p(8, 0 \mid 10, \text{up})$	= 0.0
$p(4, 0 \mid 10, \text{up})$	= 0	$p(4, 0 \mid 10, \text{up})$	= 0.0
$p(4, 10 \mid 10, \text{up})$	= 0	$p(4, 10 \mid 10, \text{up})$	= 0.0

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1$$

Formalization

- State transition probability

$$p(s' \mid s, a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r \mid s, a)$$

- Expected immediate reward

$$r(s, a) \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$$

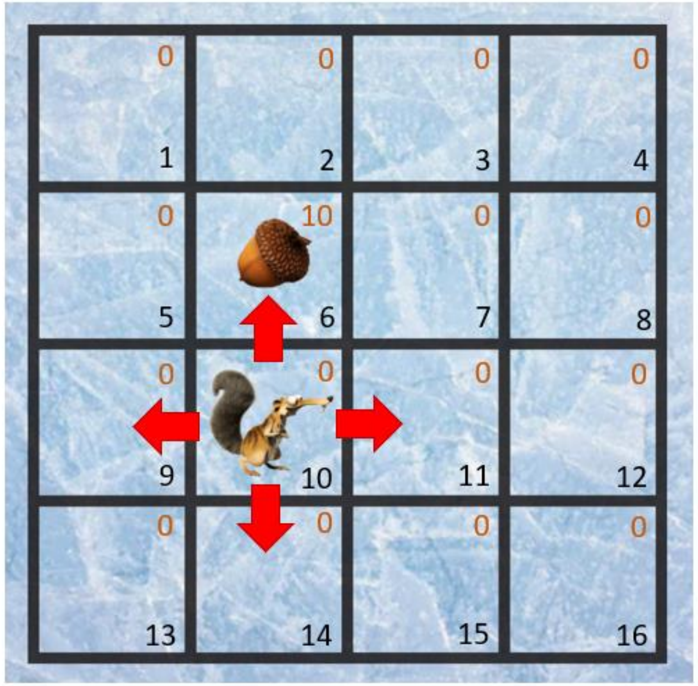
$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

Example

$$r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s',r|s,a)$$

Deterministic

$r(10, \text{up}) =$



Example

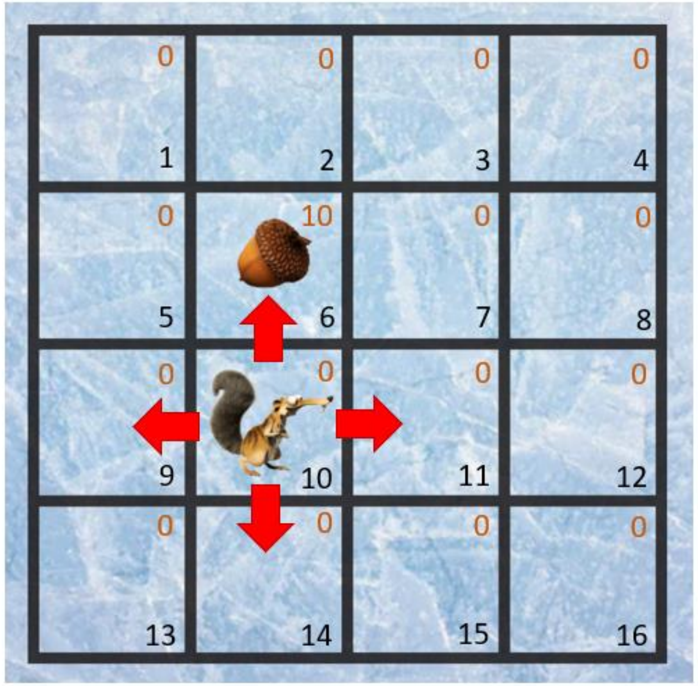
$$r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r \mid s, a)$$

Deterministic

$r(10, \text{up}) = 10 \cdot p(6, 10 \mid 10, \text{up}) = 10 \cdot 1 = 10$

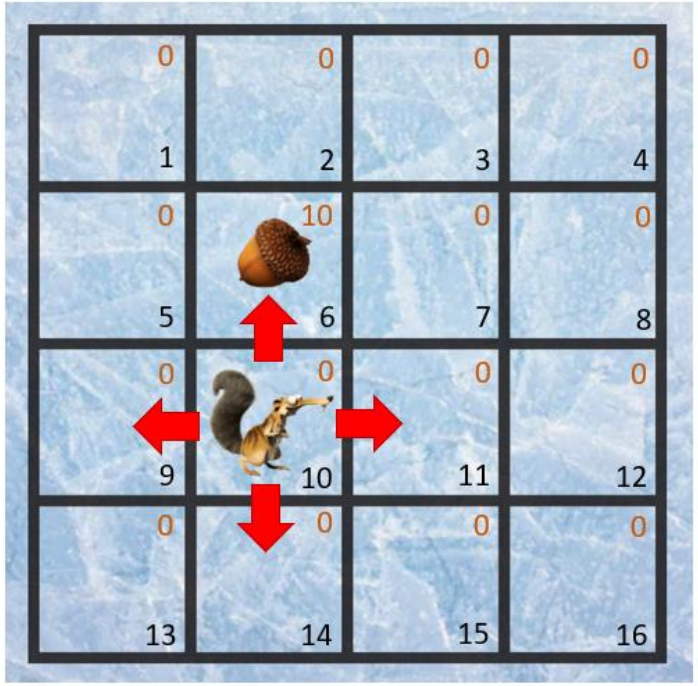
Stochastic

$r(10, \text{up}) =$



Example

$$r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r \mid s, a)$$



Deterministic

$r(10, \text{up}) = 10 \cdot p(6, 10 \mid 10, \text{up}) = 10 \cdot 1 = 10$

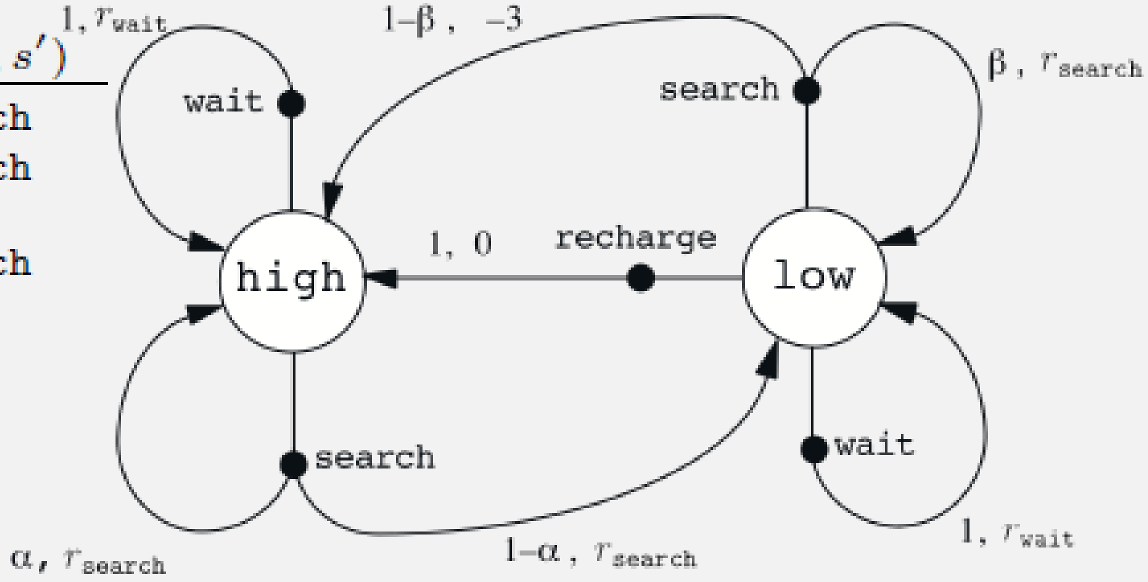
Stochastic

$r(10, \text{up}) = 10 \cdot p(6, 10 \mid 10, \text{up}) +$
$0 \cdot p(9, 0 \mid 10, \text{up}) +$
$0 \cdot p(11, 0 \mid 10, \text{up}) = 10 \cdot 0.8 + 0 \cdot 0.1 + 0 \cdot 0.1 = 8$

Example: Transition graph

- Recycling robot
  - States: [Battery low, Battery high]
  - Actions: [search, wait, recharge]
  - Rewards: Trash collected +1, Rescue -3

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1 - \alpha$	$r_{\texttt{search}}$
low	search	high	$1 - \beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	$r_{\texttt{wait}}$
low	wait	high	0	$r_{\texttt{wait}}$
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	0



General thoughts

- Anything cannot be changed arbitrarily by the agent is considered to be outside of it and thus part of the environment
- Boundary represents the limit of the agent's absolute control, not of its knowledge
- Example: in a human the muscles, skeleton, and sensory system all part of the environment

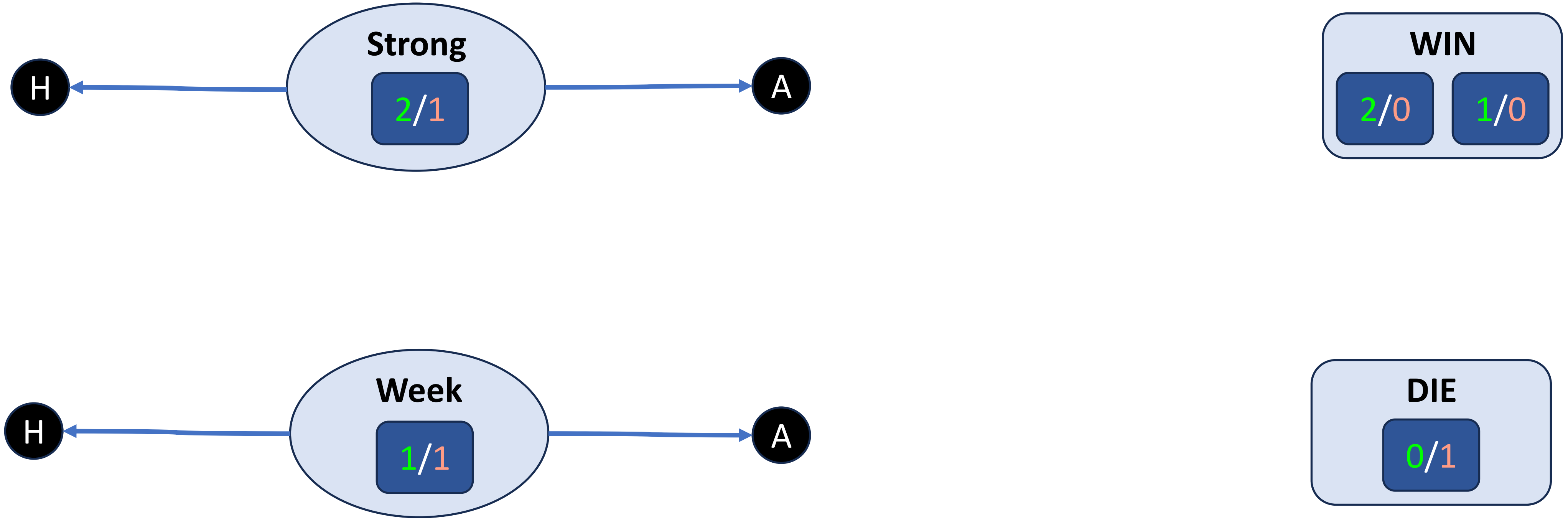
Exercise: Transition graph


You are playing a turn-based fighting game and face an enemy. You have 2 health points (HP), while your enemy has 1 HP. You make one of two moves in a turn, attacking or healing:

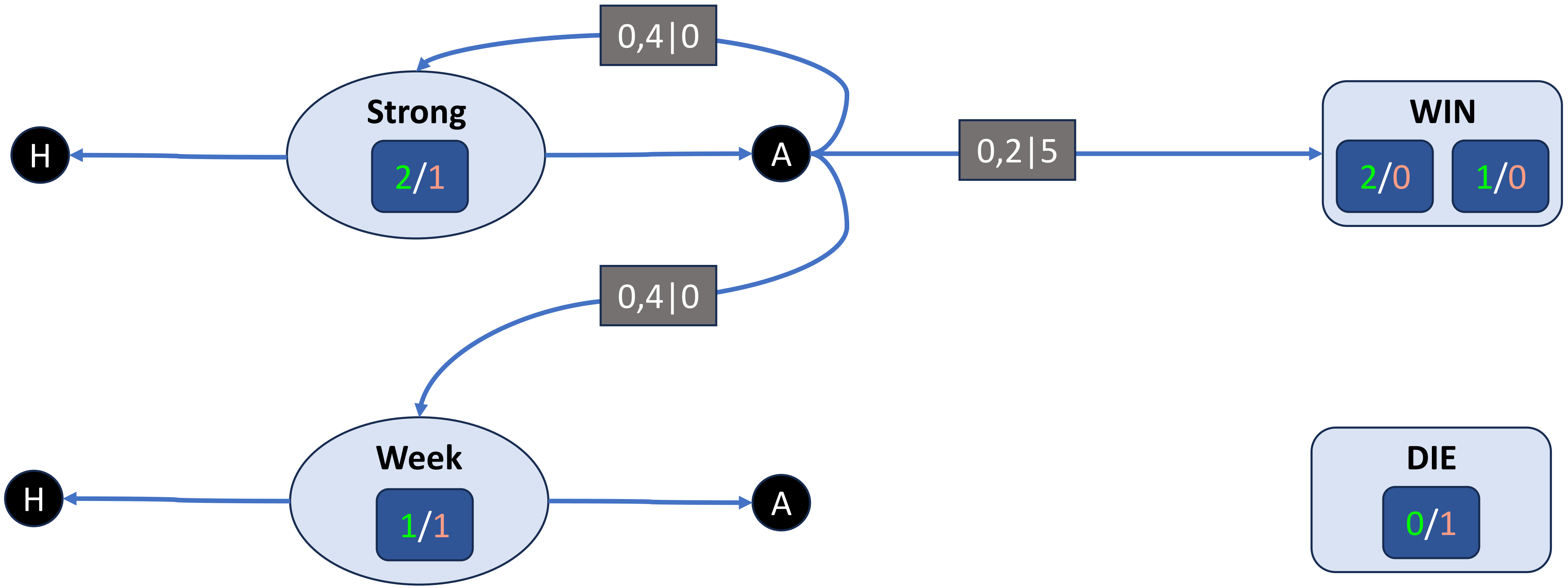
- By attacking, you kill your enemy with a 20% chance.
- By healing, you restore 1 HP with a 60% chance (you can't have more than 2 HP).

After your move, your enemy attacks you, making you lose 1 HP with a 50% chance. You get a reward of 1 for each restored HP, a reward of 5 killing your enemy, and -5 for dying. The game ends if either you or your enemy dies. Construct an MDP that represents this game and draw its transition graph!

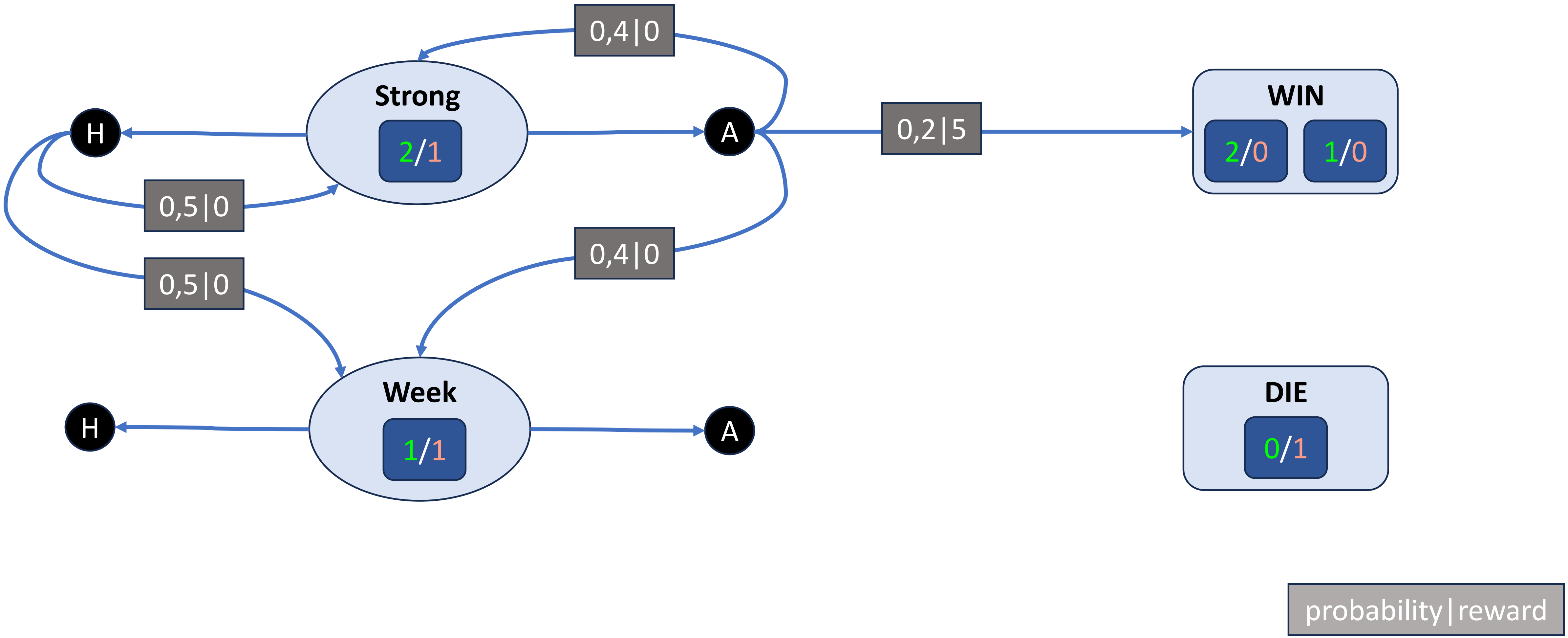
Exercise: Transition graph



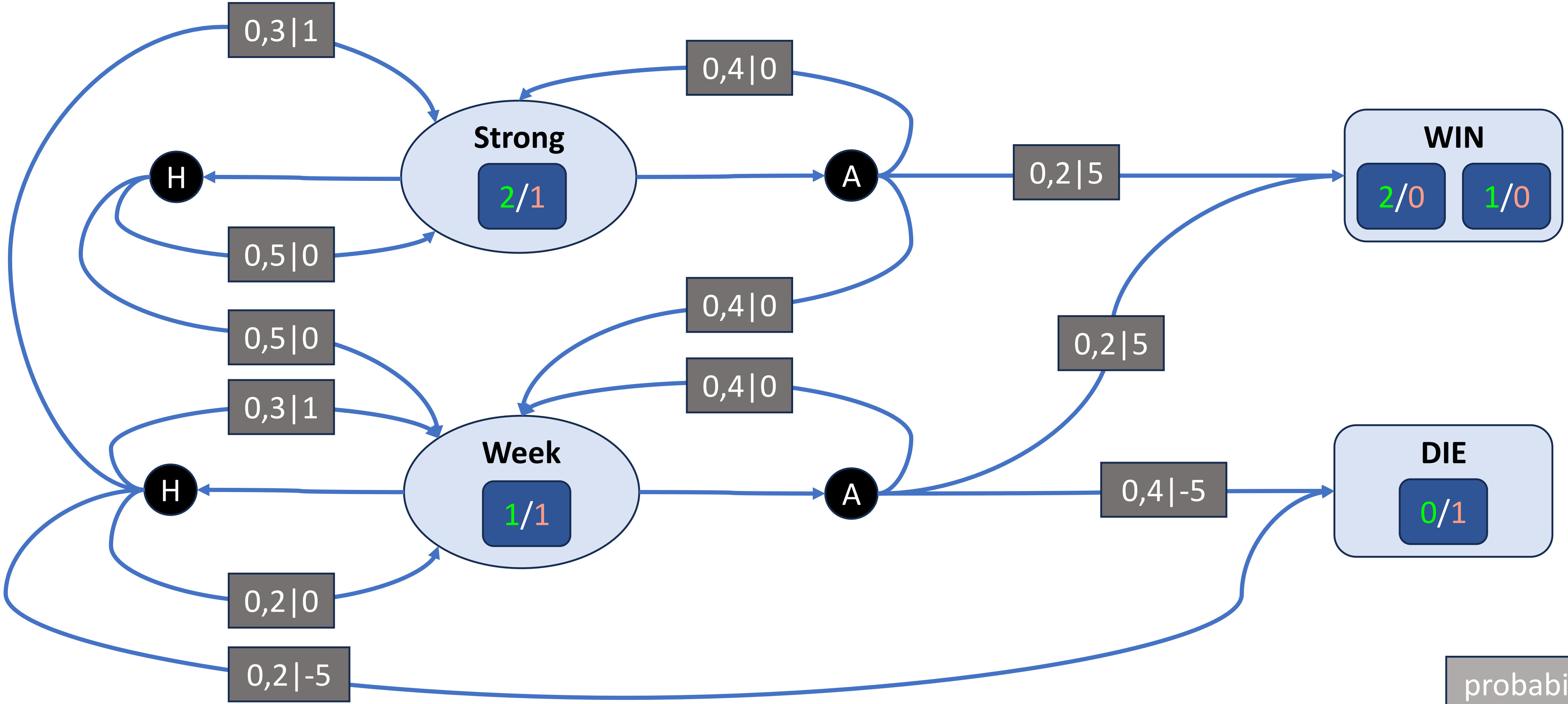
Exercise: Transition graph



Exercise: Transition graph



Exercise: Transition graph

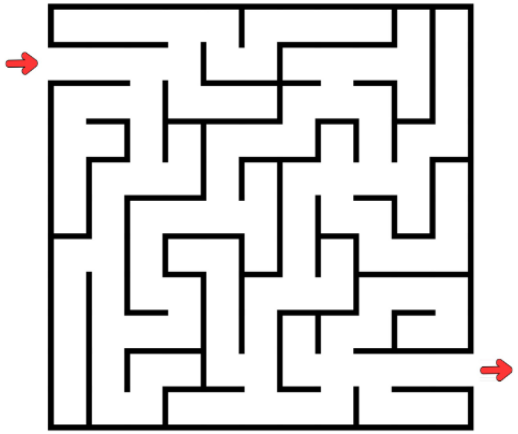


probability|reward

Goal and Rewards

- The goal of the agent is formalized with a reward signal
- Passing from the environment to the agent

The goal of the agent is to maximise the expected value of the cumulative sum of a received scalar signal (reward)



A, Reward: +1 for each step, +10 for reaching the exit (terminal state)

Goal:

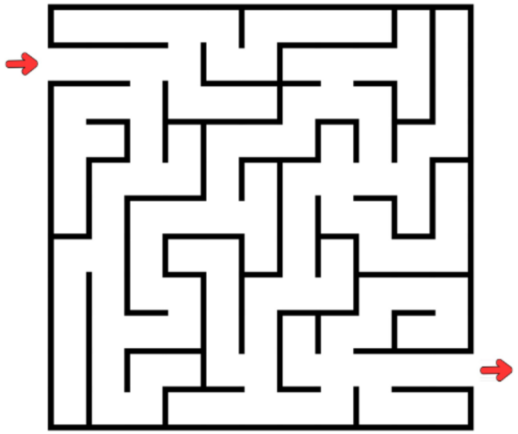
B, Reward: -1 for each step, +10 for reaching the exit (terminal state)

Goal:

Goal and Rewards

- The goal of the agent is formalized with a reward signal
- Passing from the environment to the agent

The goal of the agent is to maximise the expected value of the cumulative sum of a received scalar signal (reward)



A, Reward: +1 for each step, +10 for reaching the exit (terminal state)

Goal: Stay in the Maze

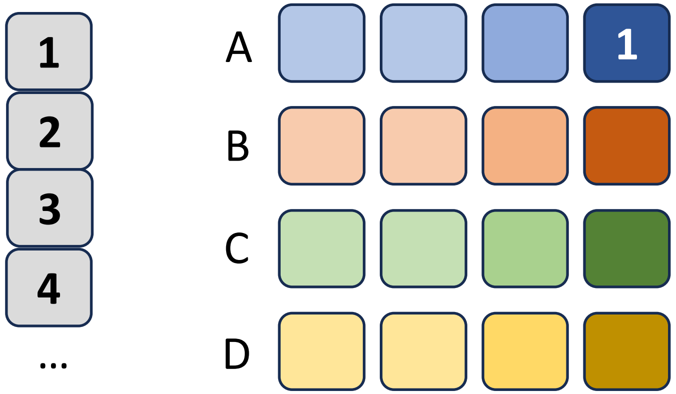
B, Reward: -1 for each step, +10 for reaching the exit (terminal state)

Goal: Leave the Maze

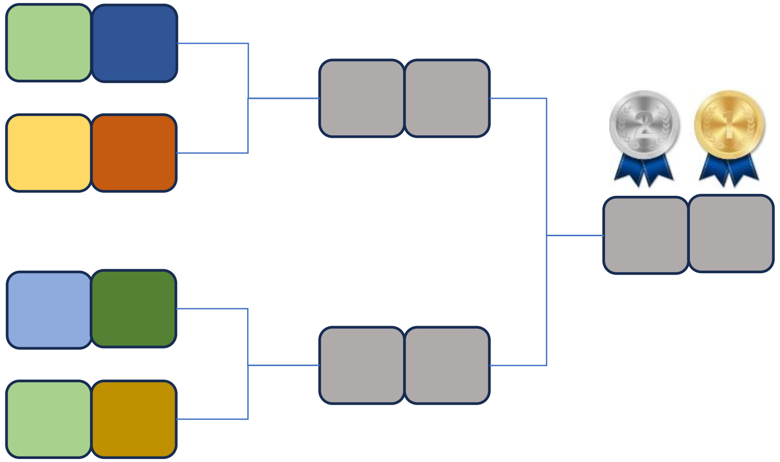
Goal and Rewards

- Game Theory Analogy
  - Game theory is the study of mathematical models of strategic interactions among **rational agents**
  - How to design an environment and reward system to get the desired behaviour

Badminton



Group round



Elimination round

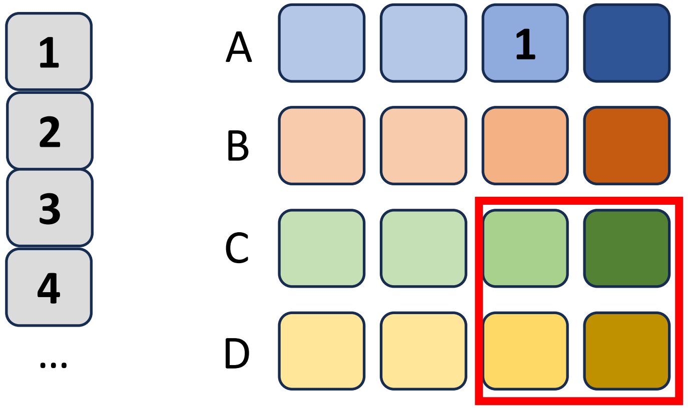
Goal and Rewards

- Game Theory Analogy
  - Game theory is the study of mathematical models of strategic interactions among rational agents
  - How to design an environment and reward system to get the desired behaviour

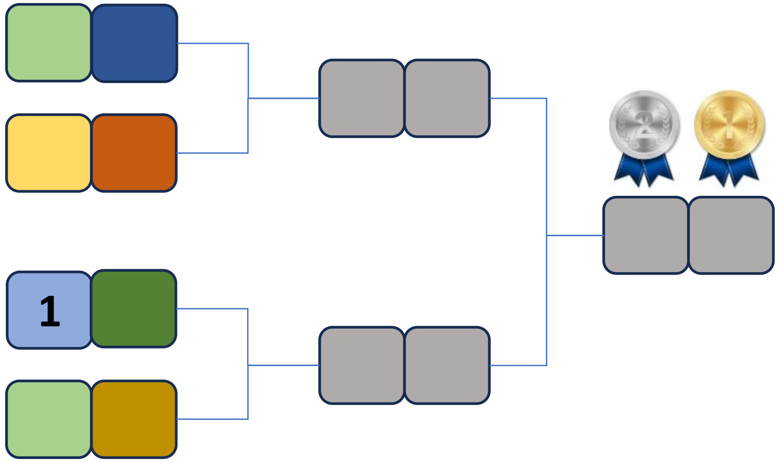
LONDON 2012

Badminton

After an upset win...



Group round



Elimination round

Returns and Episodes

- Agent's goal is to maximize the cumulative reward
- Formalize the **Expected return** as $\boldsymbol{G_t}$

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

Returns and Episodes

- Agent's goal is to maximize the cumulative reward
- Formalize the **Expected return** as $\boldsymbol{G_t}$

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

T – terminal state

Episodic task

T finite

Continuing task

$T = \infty$

Problematic calculation!

Expected discounted return

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

γ – discount rate [0,1]

- The discount rate determines the present value of future rewards
- Updated goal: maximise the sum of the discounted reward

Expected discounted return

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \boxed{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}}$$

γ – discount rate [0,1]

Infinite sum has a finite value if R_k is bounded

γ=0 ”myopic” agent: maximising immediate reward

- The discount rate determines the present value of future rewards
- Updated goal: maximise the sum of the discounted reward

Expected discounted return considerations

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- Recursive calculation:

$$\begin{aligned}
G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\
&= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots \right) \\
&= R_{t+1} + \gamma G_{t+1}
\end{aligned}$$

Expected discounted return considerations

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

If R_t=1:

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

Proof:

Expected discounted return considerations

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

If R_t=1:

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

Proof:

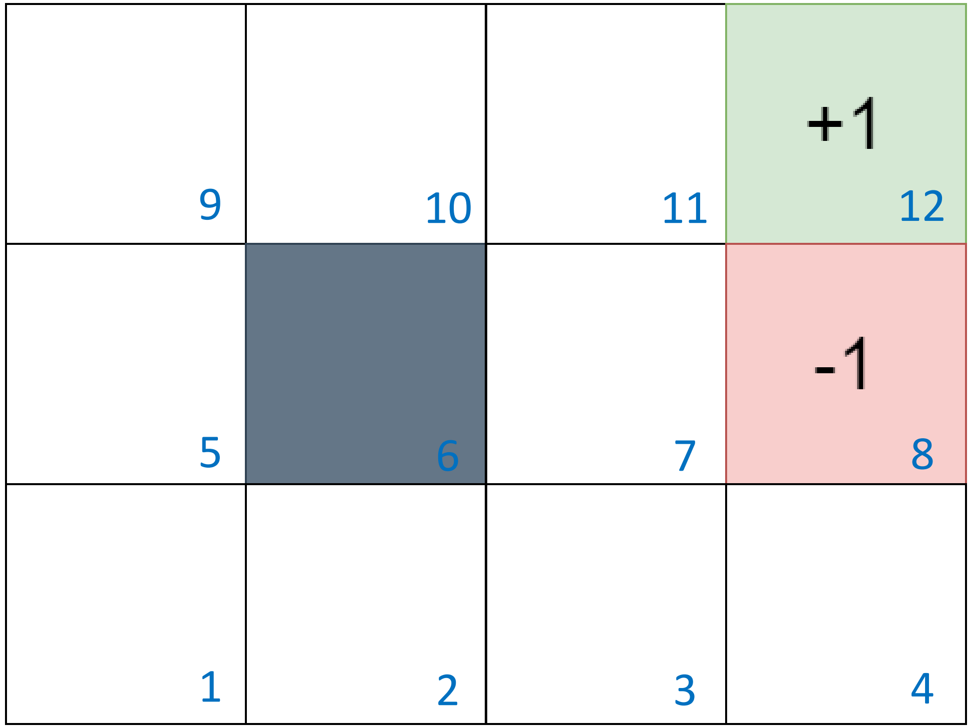
(1) $$G_t = \sum_{k=0}^{\infty} \gamma^k = 1 + \gamma + \gamma^2 + \ldots + \gamma^n + \ldots$$

γ(1)=(2) $$\gamma G_t = \gamma + \gamma^2 + \ldots + \gamma^{n+1} + \ldots$$

(1)-(2) $$(1-\gamma) G_t = 1 - 0$$

$$G_t = \frac{1}{(1-\gamma)}$$

MDP summary



[up, down, left, right]

Markovian property:
- Only present matters
- Stationary (rules do not change)

States: S

Model: $T(s, a, s') \sim Pr(s'|s, a)$

Actions: $A(s), A$

Reward: $R(s), R(s, a), R(s, a, s')$

MDP

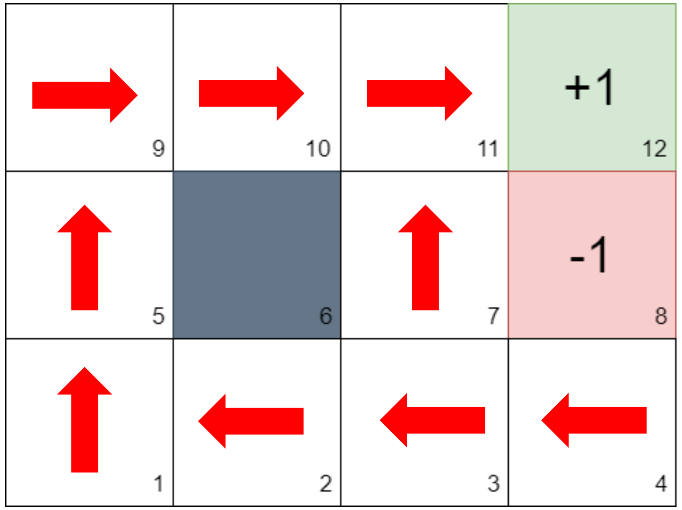
Problem definition

Policy: $\pi(s) \to a$

Solution

π^* **Optimal policy:** maximises the long term expected reward

Example: MDP optimal policy

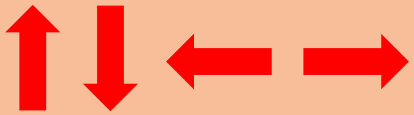


Rules:

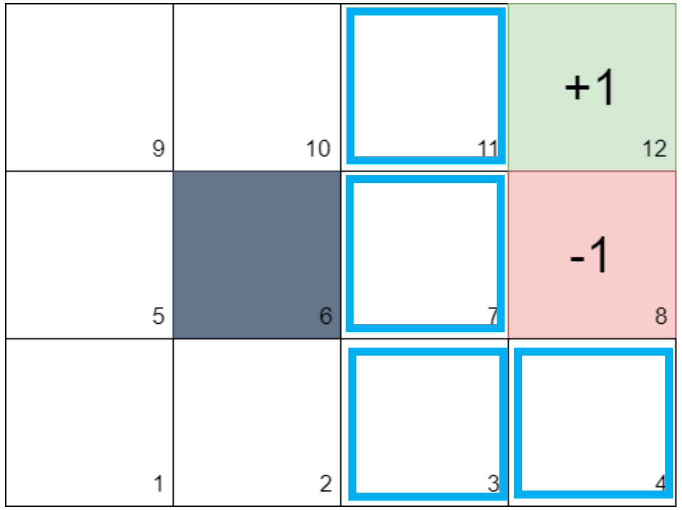
- **Stochastic**
- **Rewards given**

$$R(s) = -0.04$$

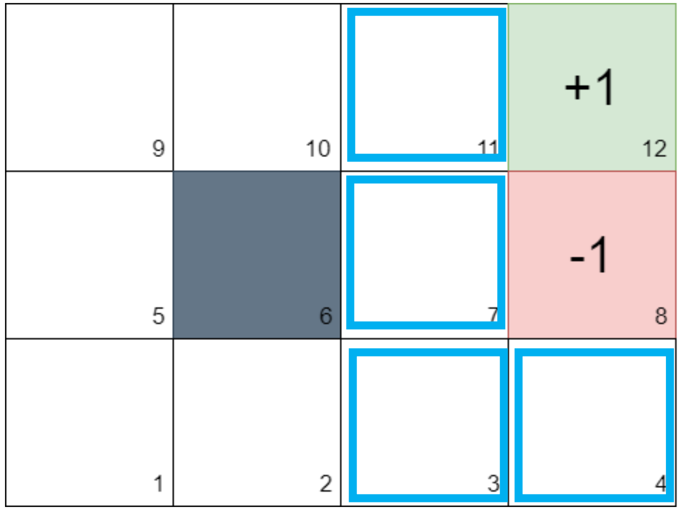
Action Space:



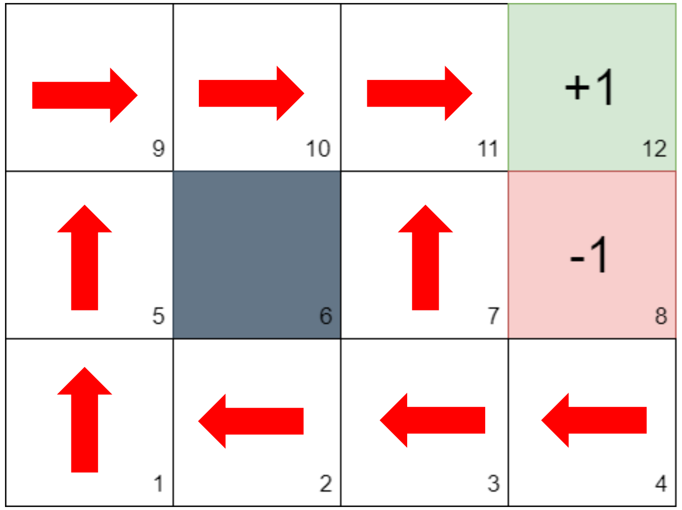
$$R(s) = +2$$



$$R(s) = -2$$



Example: MDP optimal policy



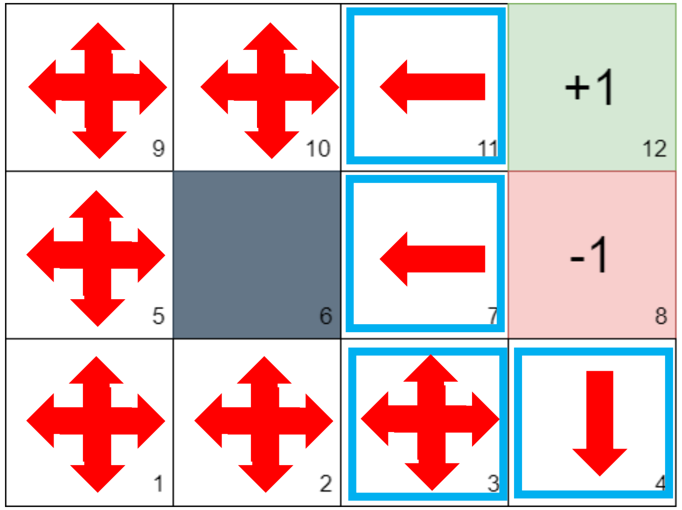
Rules:

- **Stochastic**
- **Rewards given**

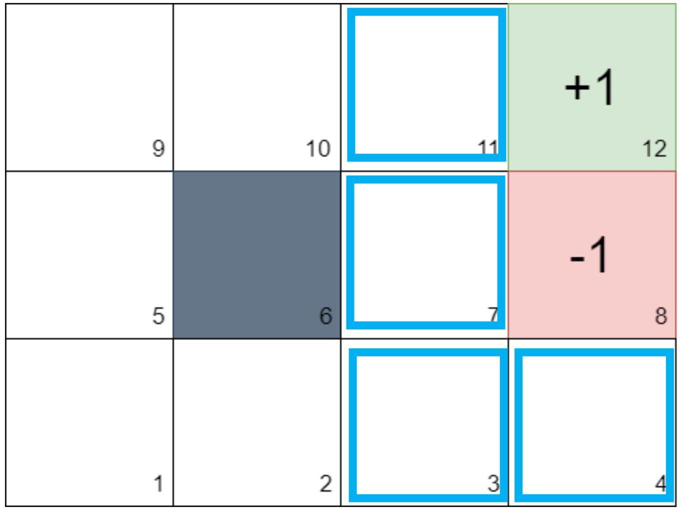
$$R(s) = -0.04$$

Action Space: ↑ ↓ ← →

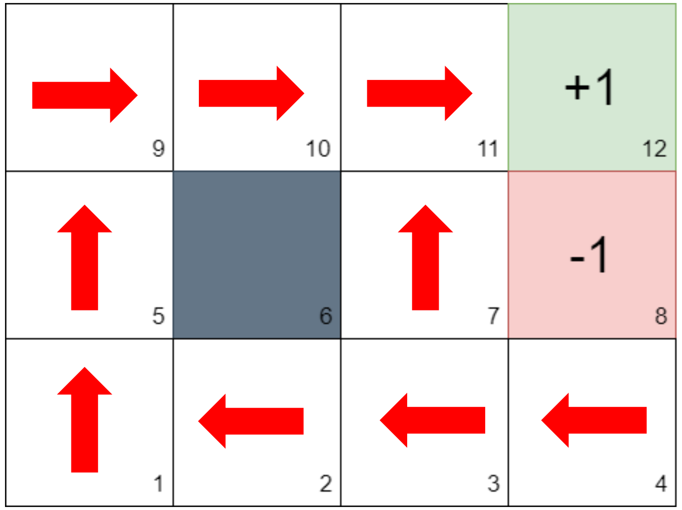
$$R(s) = +2$$



$$R(s) = -2$$



Example: MDP optimal policy



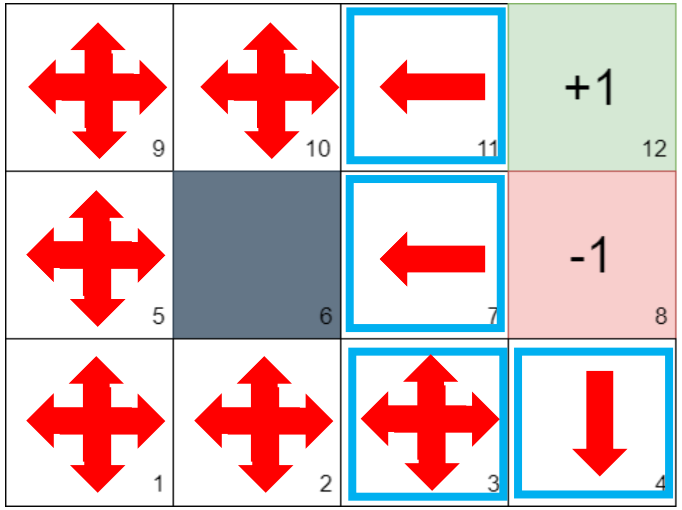
Rules:

- **Stochastic**
- **Rewards given**

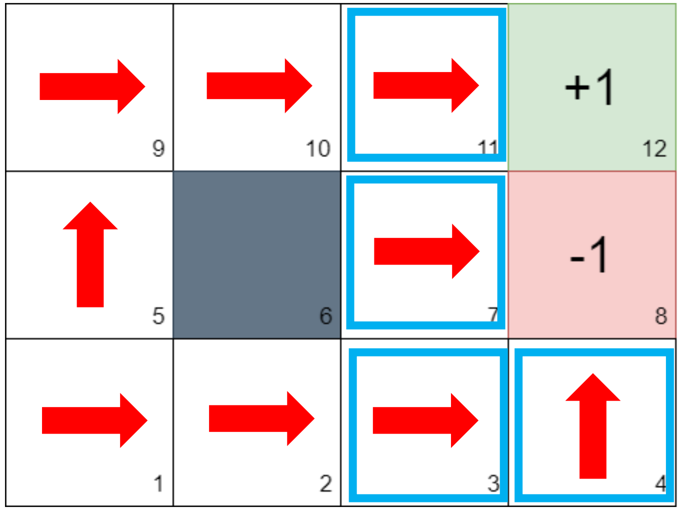
$$R(s) = -0.04$$

Action Space: ↑ ↓ ← →

$$R(s) = +2$$



$$R(s) = -2$$



Exercise

Suppose γ=0.5 and the following sequence of rewards is received R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, and R_5 = 2, with T = 5. What are G_0, G_1, ... , G_5?

Exercise

Suppose γ=0.5 and the following sequence of rewards is received R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, and R_5 = 2, with T = 5. What are G_0, G_1, … , G_5?

t	R_t	Gt
0	-	
1	-1	
2	2	
3	6	
4	3	
5 (T)	2	

Work backwards

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$G_T = 0$$

Exercise

Suppose γ=0.5 and the following sequence of rewards is received R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, and R_5 = 2, with T = 5. What are G_0, G_1, … , G_5?

t	R_t	Gt
0	-	
1	-1	
2	2	
3	6	
4	3	
5 (T)	2	**0**

Work backwards

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$G_T = 0$$

Exercise

Suppose γ=0.5 and the following sequence of rewards is received R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, and R_5 = 2, with T = 5. What are G_0, G_1, … , G_5?

t	R_t	Gt
0	-	
1	-1	
2	2	
3	6	
4	3	2+0.5·0 = **2**
5 (T)	2	**0**

Work backwards

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$G_T = 0$$

Exercise

Suppose γ=0.5 and the following sequence of rewards is received R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, and R_5 = 2, with T = 5. What are G_0, G_1, … , G_5?

t	R_t	Gt
0	-	
1	-1	
2	2	
3	6	3+0.5·2 = **4**
4	3	2+0.5·0 = **2**
5 (T)	2	**0**

Work backwards

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$G_T = 0$$

Exercise

Suppose $\gamma=0.5$ and the following sequence of rewards is received R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, and R_5 = 2, with T = 5. What are G_0, G_1, … , G_5?

t	R_t	Gt
0	-	-1+0.5·6 = **2**
1	-1	2+0.5·8 = **6**
2	2	6+0.5·4 = **8**
3	6	3+0.5·2 = **4**
4	3	2+0.5·0 = **2**
5 (T)	2	**0**

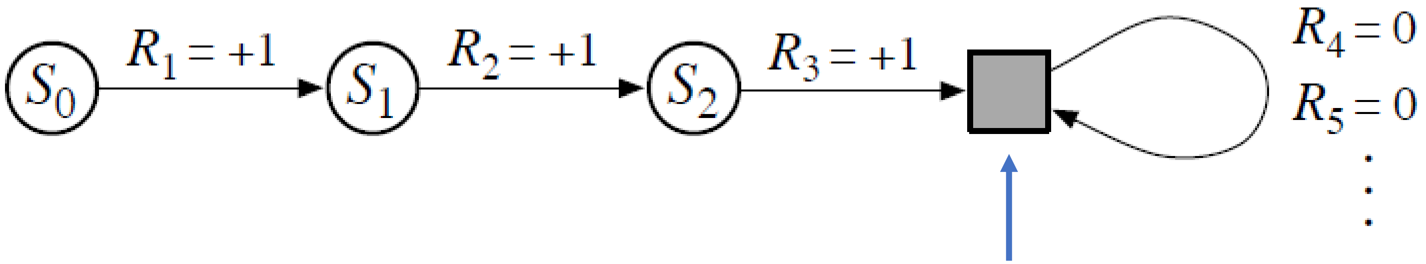
Work backwards

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$G_T = 0$$

Episodic to Continuous task

- Unified notation to treat Episodic and Continuous tasks the same



Absorbing state
instead of **Terminal state**

- Rewritten expected discounted return

$$G_t \doteq \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

$$T = \infty \text{ or } \gamma = 1$$

But not both at the same time

Policies and Value Functions

- **Value functions *v()***: functions of states (or of state-action pairs) that estimate how good it is for the agent to be in a given state (or how good it is to perform a given action in a given state)
- **Policy *π()***: a mapping from states to probabilities of selecting each possible action

Policies and Value Functions

States - *s*

9	10	11	12
5	6	7	8
1	2	3	4

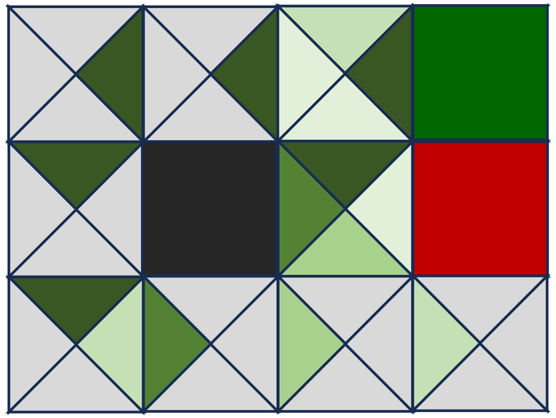
Rewards – *r*(*s*)

0	0	0	+5
0	0	0	-5
0	0	0	0

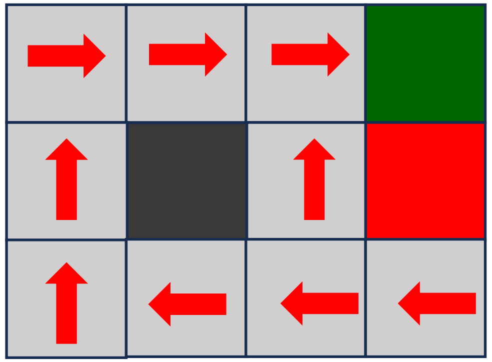
State - value functions *v(s)*

2	3	4	+5
1	0	2	-5
0	1	1	2

Action - value functions *q(s,a)*



Policy - *π*(*s*)



Value functions

- **State-value function for policy π** : the value function of a state *s* under a policy *π*, denoted $v_\pi(s)$, is the expected return when starting in *s* and following *π* thereafter

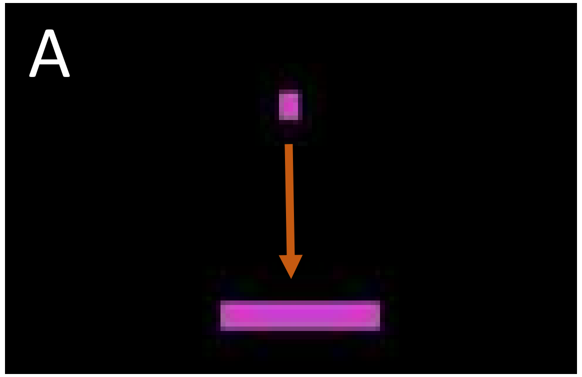
$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \text{ for all } s \in \mathcal{S}$$

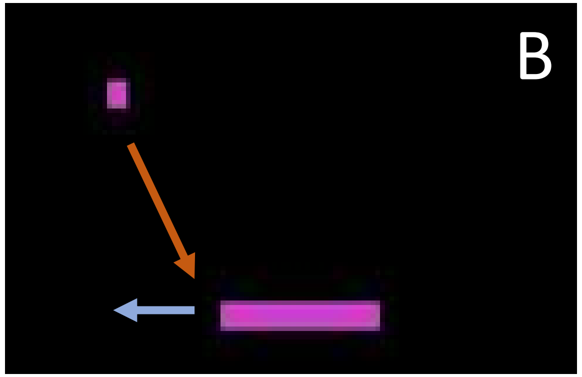
- **Action-value function for policy π :** the value of taking action *a* in state *s* under a policy *π*, denoted $q_\pi(s,a)$, as the expected return starting from *s*, taking the action *a*, and thereafter following policy *π*

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

Q-value intuition

It can be very difficult for humans to accurately estimate *Q*-values

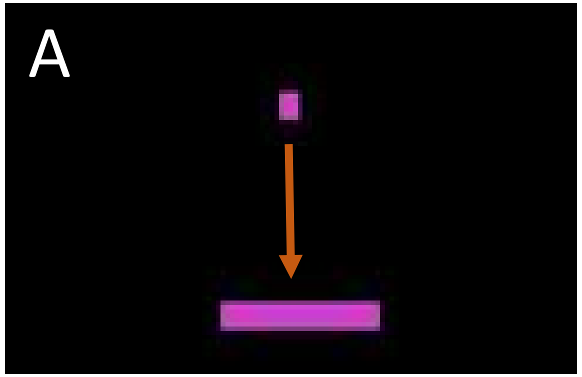


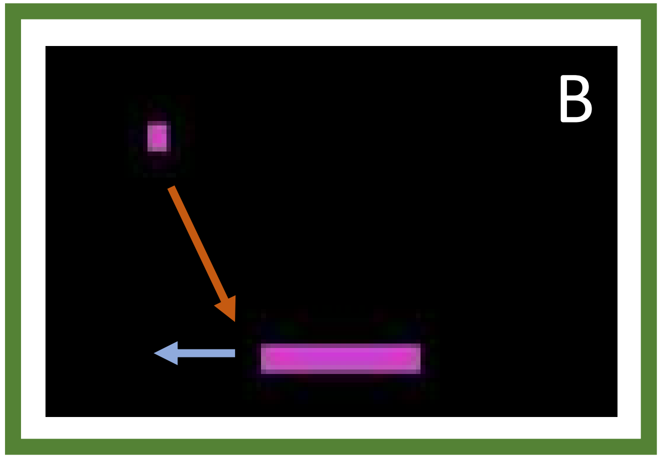


Which (s, a) pair has a higher *Q*-value?

Q-value intuition

It can be very difficult for humans to accurately estimate *Q*-values





Which (*s*, *a*) pair has a higher *Q*-value?

Recursive value iteration

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] \Big] \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \Big[r + \gamma v_\pi(s') \Big], \quad \text{for all } s \in \mathcal{S} \end{aligned}$$

Recursive value iteration

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] \Big] \\
&= \boxed{\sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \Big[r + \gamma v_\pi(s') \Big]}, \quad \text{for all } s \in \mathcal{S}
\end{aligned}$$

Bellman equation for v_π

Example: Value iteration

T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	T

k=0 (init)

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

Rules:

Actions: [up, down, right, left]

Random action selection

R = -1 for each step

Deterministic world

No discount used

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right]$$

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a) \left[r + \gamma v_\pi(s') \right]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	0 T

Deterministic

No discount

Random policy

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	0 T

Deterministic

p(1, -1 | 1, U) = 1
p(5, -1 | 1, D) = 1
p(2, -1 | 1, R) = 1
p(T, -1 | 1, L) = 1

p(6, -1 | 1, L) = 0
…

No discount

$\gamma = 1$

Random policy

π(U | 1) = π(D | 1) = π(R | 1) = π(L | 1) = 0.25

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	$\boxed{-1 \;\; 1}$	2	3
4	5	6	7
8	9	10	11
12	13	14	0 T

$$\begin{aligned} v(1) = {} & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] = 4 \cdot 0.25 \cdot -1 = -1 \end{aligned}$$

Deterministic

$p(1, -1 \mid 1, U) = 1$
$p(5, -1 \mid 1, D) = 1$
$p(2, -1 \mid 1, R) = 1$
$p(T, -1 \mid 1, L) = 1$

$p(6, -1 \mid 1, L) = 0$
…

No discount

$\gamma = 1$

Random policy

$\pi(U \mid 1) = \pi(D \mid 1) = \pi(R \mid 1) = \pi(L \mid 1) = 0.25$

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	$\boxed{-1 \quad 1}$	-1 2	-1 3
-1 4	-1 5	-1 6	1 7
-1 8	-1 9	-1 10	-1 11
-1 12	-1 13	-1 14	0 T

$$\begin{aligned} v(1) = {} & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] = 4 \cdot 0.25 \cdot -1 = -1 \end{aligned}$$

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	-1 1	-1 2	-1 3
-1 4	-1 5	-1 6	1 7
-1 8	-1 9	-1 10	-1 11
-1 12	-1 13	-1 14	0 T

k=2

0 T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	0 T

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	-1 1	-1 2	-1 3
-1 4	-1 5	-1 6	1 7
-1 8	-1 9	-1 10	-1 11
-1 12	-1 13	-1 14	0 T

k=2

0 T	-1.75 1	2	3
4	5	6	7
8	9	10	11
12	13	14	0 T

$$\begin{aligned} v(1) = {} & 0.25 \cdot 1 \cdot [-1+1 \cdot 0] \; + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot -1] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot -1] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot -1] = 3 \cdot 0.25 \cdot -2+1 \cdot 0.25 \cdot -1 = -1.75 \end{aligned}$$

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	-1 1	-1 2	-1 3
-1 4	-1 5	-1 6	1 7
-1 8	-1 9	-1 10	-1 11
-1 12	-1 13	-1 14	0 T

k=2

0 T	-1.75 1	2	3
4	5	6	7
8	9	10	11
12	13	14	0 T

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	-1 1	-1 2	-1 3
-1 4	-1 5	-1 6	1 7
-1 8	-1 9	-1 10	-1 11
-1 12	-1 13	-1 14	0 T

k=2

0 T	-1.75 1	2	3
4	5	$\boxed{-2}$ 6	7
8	9	10	11
12	13	14	0 T

$$\begin{aligned} v(6) = {} & 0.25 \cdot 1 \cdot [-1+1 \cdot -1] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot -1] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot -1] + \\ & 0.25 \cdot 1 \cdot [-1+1 \cdot -1] = 4 \cdot 0.25 \cdot -2 = -2 \end{aligned}$$

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=0

0 T	0 1	0 2	0 3
0 4	0 5	0 6	0 7
0 8	0 9	0 10	0 11
0 12	0 13	0 14	0 T

k=1

0 T	-1 1	-1 2	-1 3
-1 4	-1 5	-1 6	1 7
-1 8	-1 9	-1 10	-1 11
-1 12	-1 13	-1 14	0 T

k=2

0 T	-1.75 1	-2 2	-2 3
-1.75 4	-2 5	-2 6	-2 7
-2 8	-2 9	-2 10	-1.75 11
-2 12	-2 13	-1.75 14	0 T

v(6) = 0.25 · 1 · [-1+1 · -1] +
0.25 · 1 · [-1+1 · -1] +
0.25 · 1 · [-1+1 · -1] +
0.25 · 1 · [-1+1 · -1] = 4 · 0.25 · -2 = -2

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=10

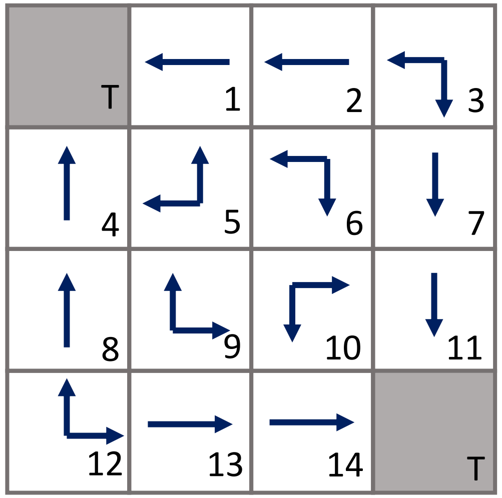
0 (T)	-6.1 (1)	-8.4 (2)	-9.0 (3)
-6.1 (4)	-7.7 (5)	-8.4 (6)	-8.4 (7)
-8.4 (8)	-8.4 (9)	-7.7 (10)	-6.1 (11)
-9.0 (12)	-8.4 (13)	-6.1 (14)	0 (T)

Example: Value iteration

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

k=10

0 T	-6.1 1	-8.4 2	-9.0 3
-6.1 4	-7.7 5	-8.4 6	-8.4 7
-8.4 8	-8.4 9	-7.7 10	-6.1 11
-9.0 12	-8.4 13	-6.1 14	0 T



Optimal policy π*

Optimal policy

- There is always at least one optimal policy

- Optimal state-value function

$$v_*(s) \doteq \max_\pi v_\pi(s)$$

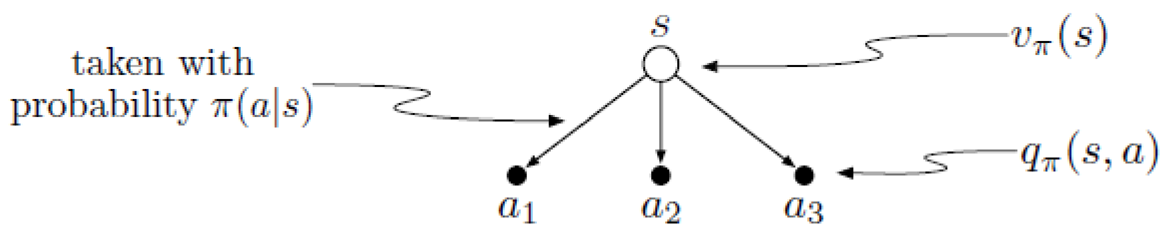
- Optimal action-value function

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a)$$

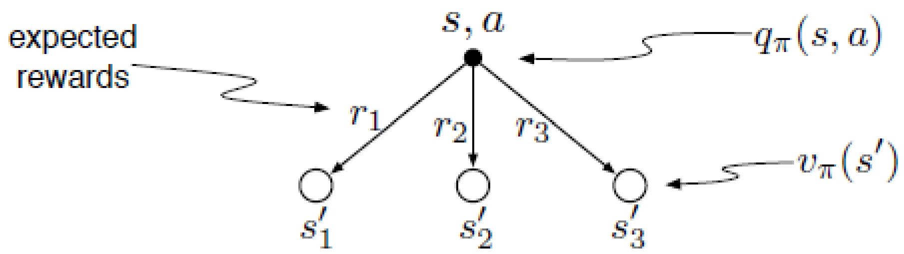
- q_* in terms of v_*

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$$

Backup diagram

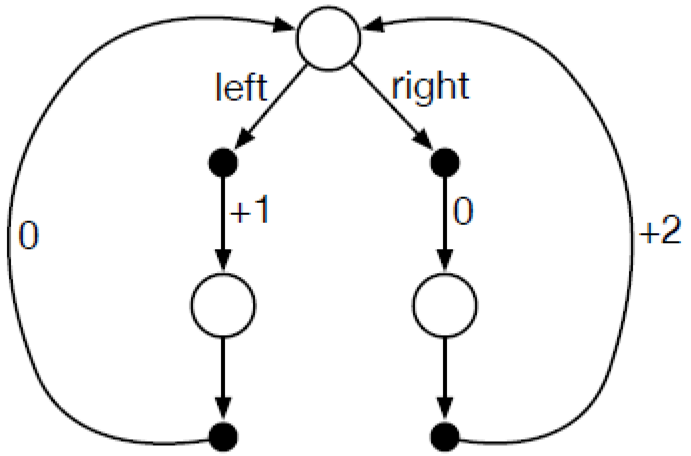


Each open circle represents a state
Each solid circle represents a state-action pair



Exercise

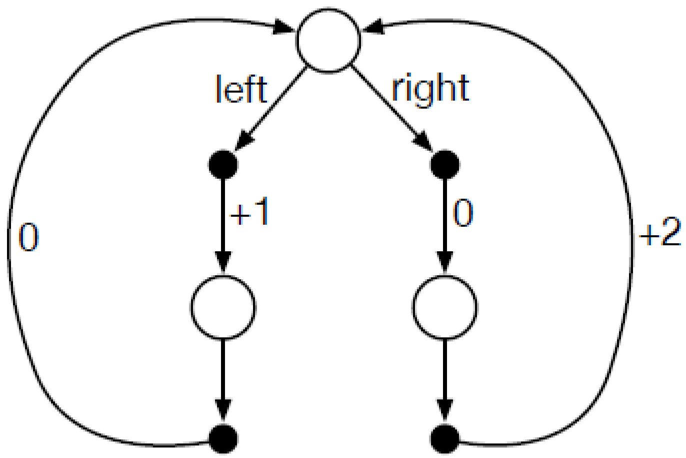
Consider the following continuing MDP. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, left and right. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?



Exercise

Consider the following continuing MDP. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, left and right. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

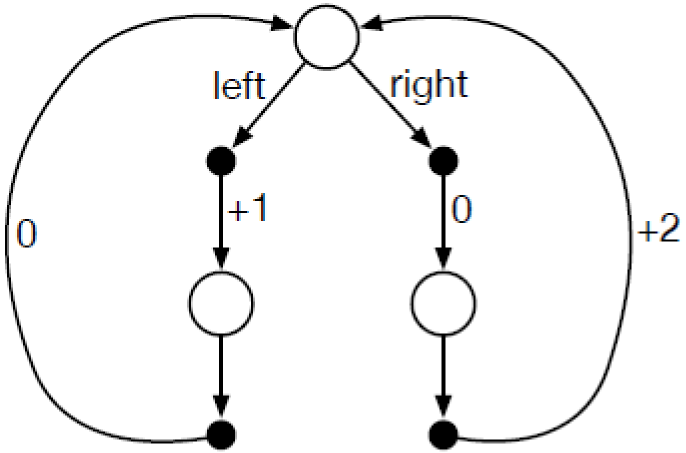
$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s]$$



Exercise

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s]$$

$\gamma = 0$



$\gamma = 0.5$

$\gamma = 0.9$

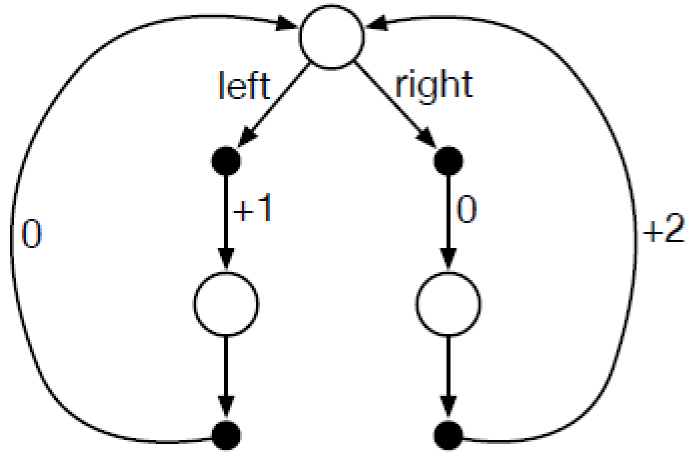
Exercise

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s]$$

$\gamma = 0$

$$\boxed{v_{\pi_{left}}(s)} = \mathbb{E}_\pi[1 + 0 \cdot G_{t+1}] = 1$$

$$v_{\pi_{right}}(s) = \mathbb{E}_\pi[0 + 0 \cdot G_{t+1}] = 0$$



$\gamma = 0.5$

$\gamma = 0.9$

Exercise

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s]$$

$\gamma = 0$

$$\boxed{v_{\pi_{left}}(s)} = \mathbb{E}_\pi[1 + 0 \cdot G_{t+1}] = 1$$

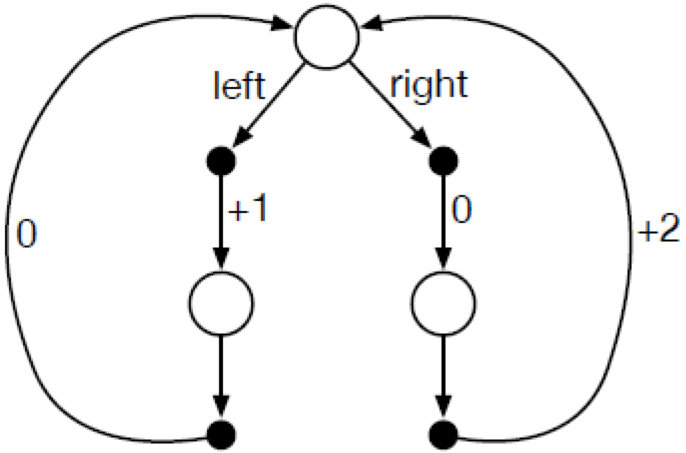
$$v_{\pi_{right}}(s) = \mathbb{E}_\pi[0 + 0 \cdot G_{t+1}] = 0$$

$\gamma = 0.5$

$$\boxed{v_{\pi_{left}}(s)} = \mathbb{E}_\pi[1 + 0.5 \cdot \boxed{0}] = 1$$

$$\boxed{v_{\pi_{right}}(s)} = \mathbb{E}_\pi[0 + 0.5 \cdot \boxed{2}] = 1$$

$\gamma = 0.9$



Just look only the first loop
as an approximate solution

Exercise

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s]$$

$\gamma = 0$

$$\boxed{v_{\pi_{left}}(s)} = \mathbb{E}_\pi[1 + 0 \cdot G_{t+1}] = 1$$

$$v_{\pi_{right}}(s) = \mathbb{E}_\pi[0 + 0 \cdot G_{t+1}] = 0$$

$\gamma = 0.5$

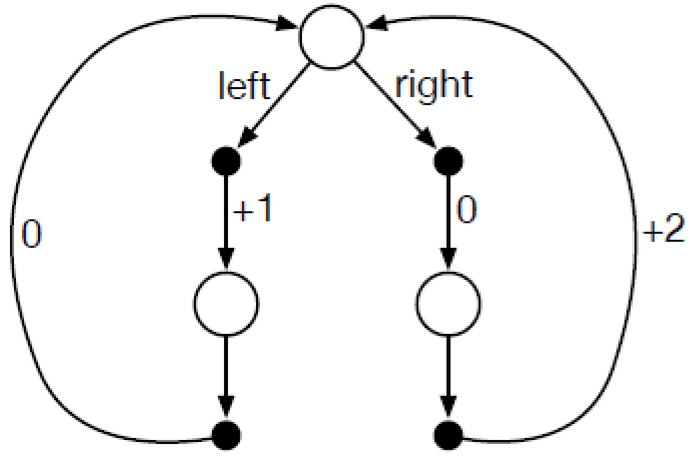
$$\boxed{v_{\pi_{left}}(s)} = \mathbb{E}_\pi[1 + 0.5 \cdot 0] = 1$$

$$\boxed{v_{\pi_{right}}(s)} = \mathbb{E}_\pi[0 + 0.5 \cdot 2] = 1$$

$\gamma = 0.9$

$$v_{\pi_{left}}(s) = \mathbb{E}_\pi[1 + 0.9 \cdot 0] = 1$$

$$\boxed{v_{\pi_{right}}(s)} = \mathbb{E}_\pi[0 + 0.9 \cdot 2] = 1.8$$



Just look only the first loop
as an approximate solution

Exercise

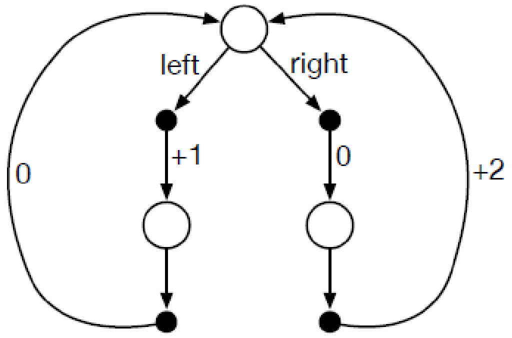
Answer: At any time step t, if we are in the top state, the discounted return will be the following:
If $\pi = \pi_{left}$:

$$G_t = 1 + 0\gamma + \gamma^2 + 0\gamma^3 + \ldots = \sum_{k=0}^{\infty} \gamma^{2k} = \frac{1}{1-\gamma^2}$$

If $\pi = \pi_{right}$:

$$G_t = 0 + 2\gamma + 0\gamma^2 + 2\gamma^3 + \ldots = \sum_{k=0}^{\infty} 2\gamma^{2k+1} = \frac{2\gamma}{1-\gamma^2}$$

At $\gamma = 0.5$ both policies are optimal. If $\gamma < 0.5$, π_{left} is optimal, and if $\gamma > 0.5$, π_{right} is optimal.



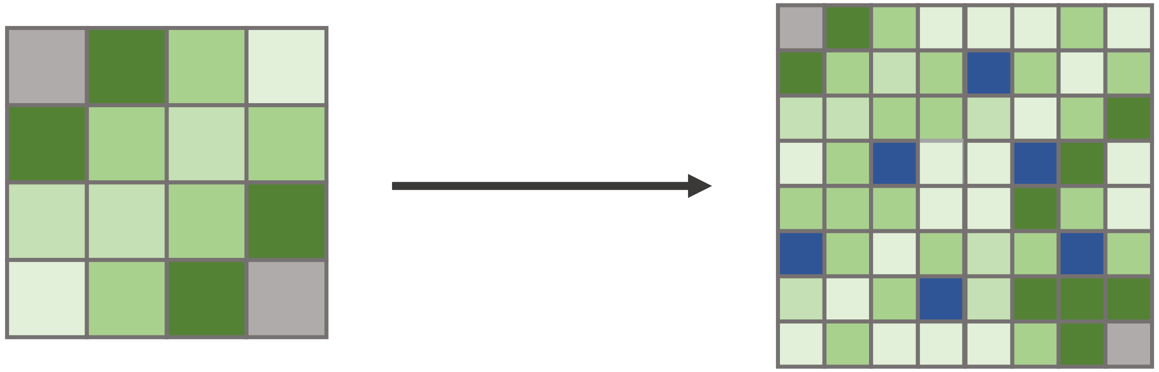
Problems

- Optimal policies work well, but in practice hard to achieve
- Works well on small problems
- Computational heavy on large problems
- Memory needy

Problems

- Optimal policies work well, but in practice hard to achieve
- Works well on small problems
- Computational heavy on large problems
- Memory needy

Curse of dimensionality describes the phenomenon where the feature space becomes increasingly sparse for an increasing number of dimensions



Function
Approximation
Needed

ELTE | FACULTY OF INFORMATICS

Thank you for your attention!