Policy estimate from training episodes

J. Kostlivá, Z. Straka, P. Švarný

We have:

- unknown grid world of unknown size and structure/shape
- robot/agents moves in unknown directions with unknown parameters

→ We do not know anything

- we only have a few episodes the robot tried

What to do?

A: Run away :-)

B: Examine episodes and learn

C: Guess

D: Try something

Policy estimate from training episodes

J. Kostlivá, Z. Straka, P. Švarný

We have:

- unknown grid world of unknown size and structure/shape
- robot/agents moves in unknown directions with unknown parameters

$\rightarrow$ We do not know anything

- we only have a few episodes the robot tried

What to do?

A: Run away :-)

B: Examine episodes and learn

C: Guess

D: Try something

Policy estimate from training episodes

J. Kostlivá, Z. Straka, P. Švarný

We have:

- unknown grid world of unknown size and structure/shape
- robot/agents moves in unknown directions with unknown parameters

$\rightarrow$ We do not know anything

- we only have a few episodes the robot tried

What to do?

A: Run away :-)

B: Examine episodes and learn

C: Guess

D: Try something

Policy estimate from training episodes

We have:

- unknown grid world of unknown size and structure/shape
- robot/agents moves in unknown directions with unknown parameters

→ We do not know anything

- we only have a few episodes the robot tried

What to do?

A: Run away :-)

B: Examine episodes and learn

C: Guess

D: Try something

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r), known discount factor $\gamma = 1$
Task: for non-terminal states determine the optimal policy. Use model-based learning.

What do we have to learn (model based learning)?

A: policy π

B: state set S, policy π

C: state set S, action set A, transition model $p(s'|s, a)$

D: state set S, action set A, rewards r, transition model $p(s'|s, a)$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r), known discount factor $\gamma = 1$

Task: for non-terminal states determine the optimal policy. Use model-based learning.

What do we have to learn (model based learning)?

A: policy π

B: state set S, policy π

C: state set S, action set A, transition model $p(s'|s, a)$

D: state set S, action set A, rewards r, transition model $p(s'|s, a)$

Example I

Episode 1	**Episode 2**	**Episode 3**	**Episode 4**
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)
Task: for non-terminal states determine the optimal policy

What do we have to learn (model based learning)?

A: policy π

B: state set S, policy π

C: state set S, action set A, transition model $p(s'|s, a)$

D: state set S, action set A, rewards r, transition model $p(s'|s, a)$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

What is the state set S?

A: $S = \{B, C\}$

B: $S = \{A, B, C, D, exit\}$

C: $S = \{A, B, C, D\}$

D: $S = \{A, D\}$

Example I

Episode 1	**Episode 2**	**Episode 3**	**Episode 4**
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

What is the state set S?

A: $S = \{B, C\}$

B: $S = \{A, B, C, D, exit\}$

C: $S = \{A, B, C, D\}$

D: $S = \{A, D\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$

- What are the terminal states?

  A: $\{A, B, C, D\}$
  B: $\{A, D\}$
  C: $\{B, C\}$
  D: $\{A, C, D\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$

- What are the terminal states?

  A: $\{A, B, C, D\}$
  B: $\{A, D\}$
  C: $\{B, C\}$
  D: $\{A, C, D\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$

- What are the terminal states?

  A: $\{A, B, C, D\}$
  B: $\{A, D\}$
  C: $\{B, C\}$
  D: $\{A, C, D\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$

- Terminal states: $\{A, D\}$
- What are the non-terminal states?

  A: $\{A, B, C, D\}$
  B: $\{A, D\}$
  C: $\{B, C\}$
  D: $\{A, B, C\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$

- Terminal states: $\{A, D\}$
- What are the non-terminal states?

  A: $\{A, B, C, D\}$
  B: $\{A, D\}$
  C: $\{B, C\}$
  D: $\{A, B, C\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$

- Terminal states: $\{A, D\}$
- What are the non-terminal states?

  A: $\{A, B, C, D\}$
  B: $\{A, D\}$
  C: $\{B, C\}$
  D: $\{A, B, C\}$

Example I

Episode 1	**Episode 2**	**Episode 3**	**Episode 4**
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$

What is the action set?

A: $\{\rightarrow, \leftarrow\}$

B: $\{\rightarrow, \leftarrow, \uparrow, \downarrow\}$

C: $\{\rightarrow, \leftarrow, \uparrow\}$

D: $\{\rightarrow, \leftarrow, \downarrow\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$

What is the action set?

A: $\{\rightarrow, \leftarrow\}$

B: $\{\rightarrow, \leftarrow, \uparrow, \downarrow\}$

C: $\{\rightarrow, \leftarrow, \uparrow\}$

D: $\{\rightarrow, \leftarrow, \downarrow\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$

What is the action set?

A: $\{\rightarrow, \leftarrow\}$

B: $\{\rightarrow, \leftarrow, \uparrow, \downarrow\}$

C: $\{\rightarrow, \leftarrow, \uparrow\}$

D: $\{\rightarrow, \leftarrow, \downarrow\}$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

A: deterministic

B: non-deterministic

Let's examine :-)

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

A: deterministic

B: non-deterministic

Let's examine :-)

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

A: deterministic

B: non-deterministic

Let's examine :-)

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- How to compute?
  - A: for each state and action
  - B: for each state, action and new state
  - C: for each state
  - D: for each action and new state

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- How to compute?
  - A: for each state and action
  - B: for each state, action and new state
  - C: for each state
  - D: for each action and new state

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- How to compute?
  1. for each state, action and new state
  2. A: as relative frequencies in one episode
     B: as sum of occurencies in one episode
     C: as relative frequencies in all episodes
     D: as sum of occurencies in all episodes

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- How to compute?
  1. for each state, action and new state
  2. A: as relative frequencies in one episode
     B: as sum of occurencies in one episode
     C: as relative frequencies in all episodes
     D: as sum of occurencies in all episodes

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- How to compute?
  1. for each state, action and new state
  2. as relative frequencies in all episodes
- evaluate $p(C|B, \rightarrow)$

  A: 1
  B: 2/3
  C: 1/2
  D: 1/3

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- How to compute?
  1. for each state, action and new state
  2. as relative frequencies in all episodes
- evaluate $p(C|B, \rightarrow)$

  A: $1 = \frac{\#(B,\rightarrow,C,\cdot)}{\#(B,\rightarrow,\cdot,\cdot)} = 2/2$

  B: $2/3$

  C: $1/2$

  D: $1/3$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- $p(C|B, \rightarrow) = 2/2 = 1$
  $p(A|B, \leftarrow) = 2/2 = 1$
  $p(D|C, \rightarrow) = 2/2 = 1$
  $p(B|C, \leftarrow) = 2/2 = 1$

A: non-deterministic

B: deterministic

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- $p(C|B, \rightarrow) = 2/2 = 1$

  $p(A|B, \leftarrow) = 2/2 = 1$

  $p(D|C, \rightarrow) = 2/2 = 1$

  $p(B|C, \leftarrow) = 2/2 = 1$

A: non-deterministic

B: deterministic

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- $p(C|B, \rightarrow) = 2/2 = 1$

  $p(A|B, \leftarrow) = 2/2 = 1$

  $p(D|C, \rightarrow) = 2/2 = 1$

  $p(B|C, \leftarrow) = 2/2 = 1$

A: non-deterministic

B: deterministic

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$

What is the transition model?

- $p(C|B, \rightarrow) = 2/2 = 1$
  $p(A|B, \leftarrow) = 2/2 = 1$
  $p(D|C, \rightarrow) = 2/2 = 1$
  $p(B|C, \leftarrow) = 2/2 = 1$

A: non-deterministic

B: deterministic

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$

What is the world structure?

A:	A	C	B	D
B:	A	B	C	D
C:	B	A	C	D

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$

What is the world structure?

A:

A	C	B	D

B:

A	B	C	D

C:

B	A	C	D

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$

What is the world structure?

A:

A	C	B	D

B:

A	B	C	D

C:

B	A	C	D

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

What is a correct value for the reward function?

A: $r(B) = -1$

B: $r(B, \leftarrow, A) = -4$

C: $r(B) = -3$

D: $r(B, \leftarrow) = -1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

What is a correct value for the reward function?

A: $r(B) = -1$

B: $r(B, \leftarrow, A) = -4$

C: $r(B) = -3$

D: $r(B, \leftarrow) = -1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1$

What is also correct for the reward function?

A: $r(B) = -1$

B: $r(B, \rightarrow) = -3$

C: $r(B) = -3$

D: $r(B, \rightarrow, C) = -1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1$

What is also correct for the reward function?

A: $r(B) = -1$

B: $r(B, \rightarrow) = -3$

C: $r(B) = -3$

D: $r(B, \rightarrow, C) = -1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1$

What is also correct for the reward function?

A: $r(B) = -1$

B: $r(B, \rightarrow) = -3$

C: $r(B) = -3$

D: $r(B, \rightarrow, C) = -1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1$, $r(B, \rightarrow) = -3$

What is also correct for the reward function?

A: $r(C) = -1$

B: $r(C, \leftarrow, B) = -3$

C: None

D: $r(C, \leftarrow) = -1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1, r(B, \rightarrow) = -3$

What is also correct for the reward function?

A: $r(C) = -1$

B: $r(C, \leftarrow, B) = -3$

C: None

D: $r(C, \leftarrow) = -1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1, r(B, \rightarrow) = -3, r(C, \leftarrow) = -1$

What is also correct for the reward function?

A: $r(C) = -1$

B: $r(C, \rightarrow) = -3$

C: $r(C) = -3$

D: $r(C, \rightarrow, D) = -4$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1, r(B, \rightarrow) = -3, r(C, \leftarrow) = -1$

What is also correct for the reward function?

A: $r(C) = -1$

B: $r(C, \rightarrow) = -3$

C: $r(C) = -3$

D: $r(C, \rightarrow, D) = -4$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1, r(B, \rightarrow) = -3, r(C, \leftarrow) = -1$

What is also correct for the reward function?

A: $r(C) = -1$

B: $r(C, \rightarrow) = -3$

C: $r(C) = -3$

D: $r(C, \rightarrow, D) = -4$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

- $r(B, \leftarrow) = -1, r(B, \rightarrow) = -3, r(C, \leftarrow) = -1,\ r(C, \rightarrow) = -3$

Discussion point, do we need more reward values?

A: Yes, for all states and actions.

B: No.

C: Yes, for terminal states.

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$

Add also the terminal state rewards: $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Do we have all we need?

A: Yes

B: No

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Do we have all we need?

A: Yes

B: No

Let's compute the policy.

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Do we have all we need?

A: Yes

B: No

Let's compute the policy.

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Observation: Immediate rewards significantly decrease state value.

A: Best is to go by less decreased path to terminal state

B: We can go to the terminal state arbitrarily

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$

World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Observation: Immediate rewards significantly decrease state value.

A: Best is to go by less decreased path to terminal state

B: We can go to the terminal state arbitrarily

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

A: $q(B, \leftarrow) = 5$

B: $q(B, \leftarrow) = 3$

C: $q(B, \leftarrow) = -1$

D: $q(B, \leftarrow) = -3$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

A: $q(B, \leftarrow) = 5$

B: $q(B, \leftarrow) = 3$

C: $q(B, \leftarrow) = -1$

D: $q(B, \leftarrow) = -3$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \ r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

A: $q(B, \leftarrow) = B \leftarrow A = 6 - 1 = 5$

B: $q(B, \leftarrow) = 3$

C: $q(B, \leftarrow) = -1$

D: $q(B, \leftarrow) = -3$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

- $q(B, \leftarrow) = 5$

(What can we assume about $\pi(C)$?)

A: $q(B, \rightarrow) = 5$

B: $q(B, \rightarrow) = 3$

C: $q(B, \rightarrow) = 1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

- $q(B, \leftarrow) = 5$

(What can we assume about $\pi(C)$?)

A: $q(B, \rightarrow) = 5$

B: $q(B, \rightarrow) = 3$

C: $q(B, \rightarrow) = 1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

- $q(B, \leftarrow) = 5$

(What can we assume about $\pi(C)$?)

A: $q(B, \rightarrow) = 5$

B: $q(B, \rightarrow) = 3$

C: $q(B, \rightarrow) = B \rightarrow C \leftarrow B \leftarrow A = -3 - 1 - 1 + 6 = 1$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

- $q(B, \leftarrow) = 5$
- $q(B, \rightarrow) = 1$

$\rightarrow \; \pi(B) = \leftarrow$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
Compute:

- $q(B, \leftarrow) = 5$
- $q(B, \rightarrow) = 1$

$\rightarrow \; \pi(B) = \leftarrow$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
$\pi(B) = \leftarrow$
Compute now $\pi(C)$:

A: $q(C, \rightarrow) = 5$

B: $q(C, \rightarrow) = 3$

C: $q(C, \rightarrow) = 0$

D: $q(C, \rightarrow) = -3$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
$\pi(B) = \leftarrow$
Compute now $\pi(C)$:

A: $q(C, \rightarrow) = 5$

B: $q(C, \rightarrow) = 3$

C: $q(C, \rightarrow) = 0$

D: $q(C, \rightarrow) = -3$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
$\pi(B) = \leftarrow$
Compute now $\pi(C)$:

A: $q(C, \rightarrow) = 5$

B: $q(C, \rightarrow) = C \rightarrow D = 6 - 3 = 3$

C: $q(C, \rightarrow) = 0$

D: $q(C, \rightarrow) = -3$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
$\pi(B) = \leftarrow$
Compute now $\pi(C)$:

- $q(C, \rightarrow) = 3$

A: $q(C, \leftarrow) = 4$

B: $q(C, \leftarrow) = 3$

C: $q(C, \leftarrow) = 0$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
$\pi(B) = \leftarrow$
Compute now $\pi(C)$:

- $q(C, \rightarrow) = 3$

A: $q(C, \leftarrow) = C \leftarrow B \leftarrow A = 6 - 1 - 1 = 4$

B: $q(C, \leftarrow) = 3$

C: $q(C, \leftarrow) = 0$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
$\pi(B) = \leftarrow$
Compute now $\pi(C)$:

- $q(C, \rightarrow) = 3$
- $q(C, \leftarrow) = 4$

$\rightarrow \; \pi(C) = \leftarrow$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3$ $r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$ go by less decreased path to terminal state
$\pi(B) = \leftarrow$
Compute now $\pi(C)$:

- $q(C, \rightarrow) = 3$
- $q(C, \leftarrow) = 4$

$\rightarrow$ $\pi(C) = \leftarrow$

Example I

Episode 1	Episode 2	Episode 3	Episode 4
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$
			$(B, \leftarrow, A, -1)$
			$(A, \leftarrow, exit, 6)$

each field in table is n-tuple (s, a, s', r)

State set $S = \{A, B, C, D\}$, terminal states: $\{A, D\}$, non-terminal states: $\{B, C\}$
Action set $A = \{\rightarrow, \leftarrow\}$
Deterministic transition model: $p(C|B, \rightarrow) = p(A|B, \leftarrow) = p(D|C, \rightarrow) = p(B|C, \leftarrow) = 2/2 = 1$
World structure:

A	B	C	D

Reward function: $r(\{B, C\}, \leftarrow) = -1, r(\{B, C\}, \rightarrow) = -3 \; r(\{A, D\}, \{\leftarrow, \rightarrow\}) = 6$

Obs.: Immediate rewards significantly decrease state value. $\rightarrow$go by less decreasedpath to terminal state

Solution:

- $\pi(B) = \leftarrow$
- $\pi(C) = \leftarrow$

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

Calculating policy

- state set S,
- action set A,
- rewards r,
- transition model $p(s'|s, a)$
- policy π

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

What is the transition model?

A: deterministic

B: non-deterministic

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

What is a correct transitional probability?

A $p(C|B, \rightarrow) = 0.75$

B $p(A|B, \rightarrow) = 0.75$

C $p(A|B, \leftarrow) = 0.25$

D $p(D|B, \leftarrow) = 0.75$

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

What is a correct transitional probability?

A $p(C|B, \rightarrow) = 0.75$, see the episodes
$(B, \rightarrow)$ occurs 4 times, three of which lead to C, one case to A thus also $p(A|B, \rightarrow) = 0.25$

B $p(A|B, \rightarrow) = 0.75$

C $p(A|B, \leftarrow) = 0.25$

D $p(D|B, \leftarrow) = 0.75$

Transition model: Similarly for other probabilities. Agent follows the direction given with probability 0.75. Otherwise, it goes the other direction.

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

What is the reward function?

A $r(B, \rightarrow, C) = -3$

B $r(B, \rightarrow, A) = -3$

C $r(B, \leftarrow, A) = -3$

D $r(B, \leftarrow, C) = -3$

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

What is the reward function?

A $r(B, \rightarrow, C) = -3$

B $r(B, \rightarrow, A) = -3$

C $r(B, \leftarrow, A) = -3$

D $r(B, \leftarrow, C) = -3$

$\Rightarrow$ $r(\cdot)$ depends on s, s' only, not the action. Similarly for other possibilities of $r(\cdot)$.

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

What is the reward function?

A $r(B, \rightarrow, C) = -3$

B $r(B, \rightarrow, A) = -3$

C $r(B, \leftarrow, A) = -3$

D $r(B, \leftarrow, C) = -3$

$\Rightarrow r(\cdot)$ depends on s, s' only, not the action. Similarly for other possibilities of $r(\cdot)$.

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$	$(B, \leftarrow, A, -1)$	$(C, \rightarrow, D, -3)$	$(C, \leftarrow, B, -1)$	$(B, \leftarrow, C, -3)$	$(B, \rightarrow, A, -1)$	$(C, \rightarrow, B, -1)$	$(C, \rightarrow, D, -3)$
$(C, \rightarrow, D, -3)$	$(A, \rightarrow, exit, 6)$	$(D, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$	$(C, \leftarrow, B, -1)$	$(A, \rightarrow, exit, 6)$	$(B, \rightarrow, C, -3)$	$(D, \rightarrow, exit, 6)$
$(D, \leftarrow, exit, 6)$			$(C, \leftarrow, B, -1)$	$(B, \leftarrow, A, -1)$		$(C, \leftarrow, D, -3)$	
			$(B, \leftarrow, A, -1)$	$(A, \leftarrow, exit, 6)$		$(D, \leftarrow, exit, 6)$	
			$(A, \leftarrow, exit, 6)$				

Result:

- States: $S = \{A, B, C, D\}$, terminal$= \{A, D\}$, nonterminal$= \{B, C\}$
- Action set: $\{\leftarrow, \rightarrow\}$
- Rewards:
  $r(B, \{\leftarrow, \rightarrow\}, C) = -3$, $r(B, \{\leftarrow, \rightarrow\}, A) = -1$,
  $r(C, \{\leftarrow, \rightarrow\}, B) = -1$, $r(C, \{\leftarrow, \rightarrow\}, D) = -3$
- World structure:

A	B	C	D

- Transition model: Agent follows the direction given with probability 0.75. Otherwise, it goes the other direction.
- Policy: $\pi(B) =?$, $\pi(C) =?$

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

Policy evaluation:

$\leftarrow, \rightarrow$ $q(B, \leftarrow) = ?, q(C, \rightarrow) = ?$

$\rightarrow, \rightarrow$ $q(B, \rightarrow) = ?, q(C, \rightarrow) = ?$

$\rightarrow, \leftarrow$ $q(B, \rightarrow) = ?, q(C, \leftarrow) = ?$

$\leftarrow, \leftarrow$ $q(B, \leftarrow) = ?, q(C, \leftarrow) = ?$

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

A single policy computation:

$\leftarrow, \rightarrow$ $q(B, \leftarrow) = ?, q(C, \rightarrow) = ?$

A $q(B, \leftarrow) = 0.5 \cdot (-1) + 0.5 \cdot (-3),$
$q(C, \rightarrow) = 0.5 \cdot (-1) + 0.5 \cdot (-3)$

B $q(B, \leftarrow) = 0.25 \cdot (6 - 1) + 0.75 \cdot (-3 + V(C)),$
$q(C, \rightarrow) = 0.25 \cdot (-1) + 0.75 \cdot (-3 + V(B))$

C $q(B, \leftarrow) = 0.75 \cdot (6 - 1) + 0.25 \cdot (-3 + V(C)),$
$q(C, \rightarrow) = 0.75 \cdot (-3 + 6) + 0.25 \cdot (-1 + V(B))$

D $q(B, \leftarrow) = 0.75 \cdot (6 - 1) + 0.25 \cdot (-3),$
$q(C, \rightarrow) = 0.5 \cdot (-1) + 0.25 \cdot (-3)$

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

A single policy computation:

$\leftarrow, \rightarrow$ $q(B, \leftarrow) =?, q(C, \rightarrow) =?$

A $q(B, \leftarrow) = 0.5 \cdot (-1) + 0.5 \cdot (-3),$
$q(C, \rightarrow) = 0.5 \cdot (-1) + 0.5 \cdot (-3)$

B $q(B, \leftarrow) = 0.25 \cdot (6 - 1) + 0.75 \cdot (-3 + V(C)),$
$q(C, \rightarrow) = 0.25 \cdot (-1) + 0.75 \cdot (-3 + V(B))$

C $q(B, \leftarrow) = 0.75 \cdot (6 - 1) + 0.25 \cdot (-3 + V(C)),$
$q(C, \rightarrow) = 0.75 \cdot (-3 + 6) + 0.25 \cdot (-1 + V(B))$

D $q(B, \leftarrow) = 0.75 \cdot (6 - 1) + 0.25 \cdot (-3),$
$q(C, \rightarrow) = 0.5 \cdot (-1) + 0.25 \cdot (-3)$

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

A single policy computation. As the policy is fixed $V(B) = q(B, \leftarrow), V(C) = q(C, \rightarrow)$:

- $q(B, \leftarrow) = 0.75 \cdot (6 - 1) + 0.25 \cdot (-3 + q(C, \rightarrow))$
- $q(C, \rightarrow) = 0.75 \cdot (-3 + 6) + 0.25 \cdot (-1 + q(B, \leftarrow))$

Therefore:

- $q(B, \leftarrow) = 0.75 \cdot 5 + 0.25 \cdot (-3 + .75 \cdot 3 + 0.25 \cdot (-1 + q(B, \leftarrow)) = ... \approx 3.72$
- $q(C, \rightarrow) = 0.75 \cdot 3 + 0.25 \cdot (-1 + 3.72) \approx 2.93$

And we calculate for the remaining policies.

Example II

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5	Episode 6	Episode 7	Episode 8
$(B, \rightarrow, C, -3)$ $(C, \rightarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(B, \leftarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$	$(C, \leftarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \leftarrow, C, -3)$ $(C, \leftarrow, B, -1)$ $(B, \leftarrow, A, -1)$ $(A, \leftarrow, exit, 6)$	$(B, \rightarrow, A, -1)$ $(A, \rightarrow, exit, 6)$	$(C, \rightarrow, B, -1)$ $(B, \rightarrow, C, -3)$ $(C, \leftarrow, D, -3)$ $(D, \leftarrow, exit, 6)$	$(C, \rightarrow, D, -3)$ $(D, \rightarrow, exit, 6)$

$\leftarrow, \rightarrow$ $q(B, \leftarrow) \approx 3.73,$
$q(C, \rightarrow) \approx 2.93$

$\rightarrow, \rightarrow$ $q(B, \rightarrow) \approx 0.62,$
$q(C, \rightarrow) \approx 2.15$

$\rightarrow, \leftarrow$ $q(B, \rightarrow) \approx -2.29,$
$q(C, \leftarrow) \approx -1.71$

$\leftarrow, \leftarrow$ $q(B, \leftarrow) \approx 3.70,$
$q(C, \leftarrow) \approx 2.77$

And we can determine the best policy: $\pi(B) = \leftarrow, \pi(C) = \rightarrow$