Markov Decision Processes

10-601 Machine Learning Fall 2012
Recitation November 27, 2012
Selen Uguroglu

Outline

1. MDP overview
2. Value Iteration
3. Policy Iteration
4. Previous Exam Questions

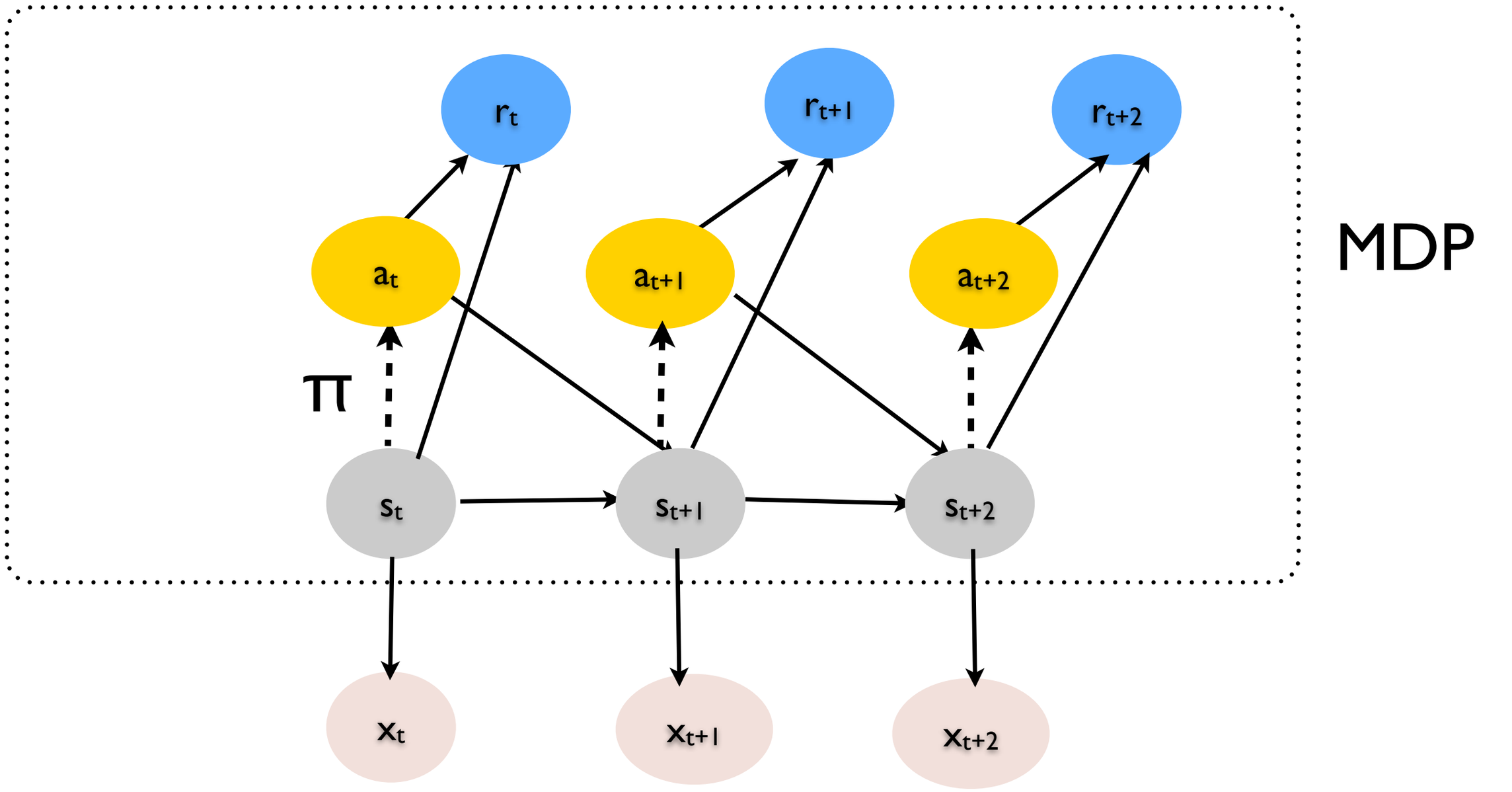
Outline

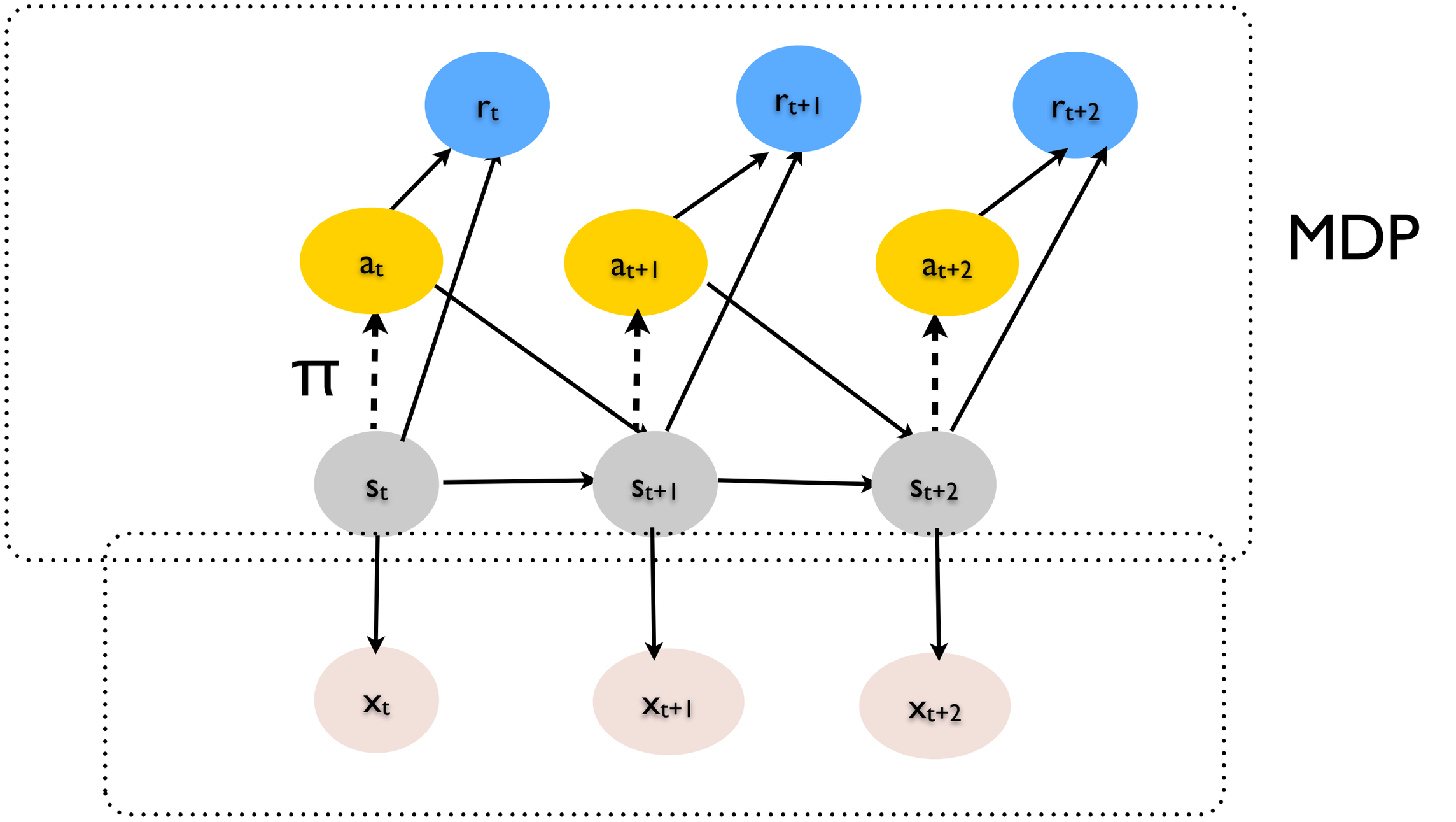
1. MDP overview
2. Value Iteration
3. Policy Iteration
4. Previous Exam Questions

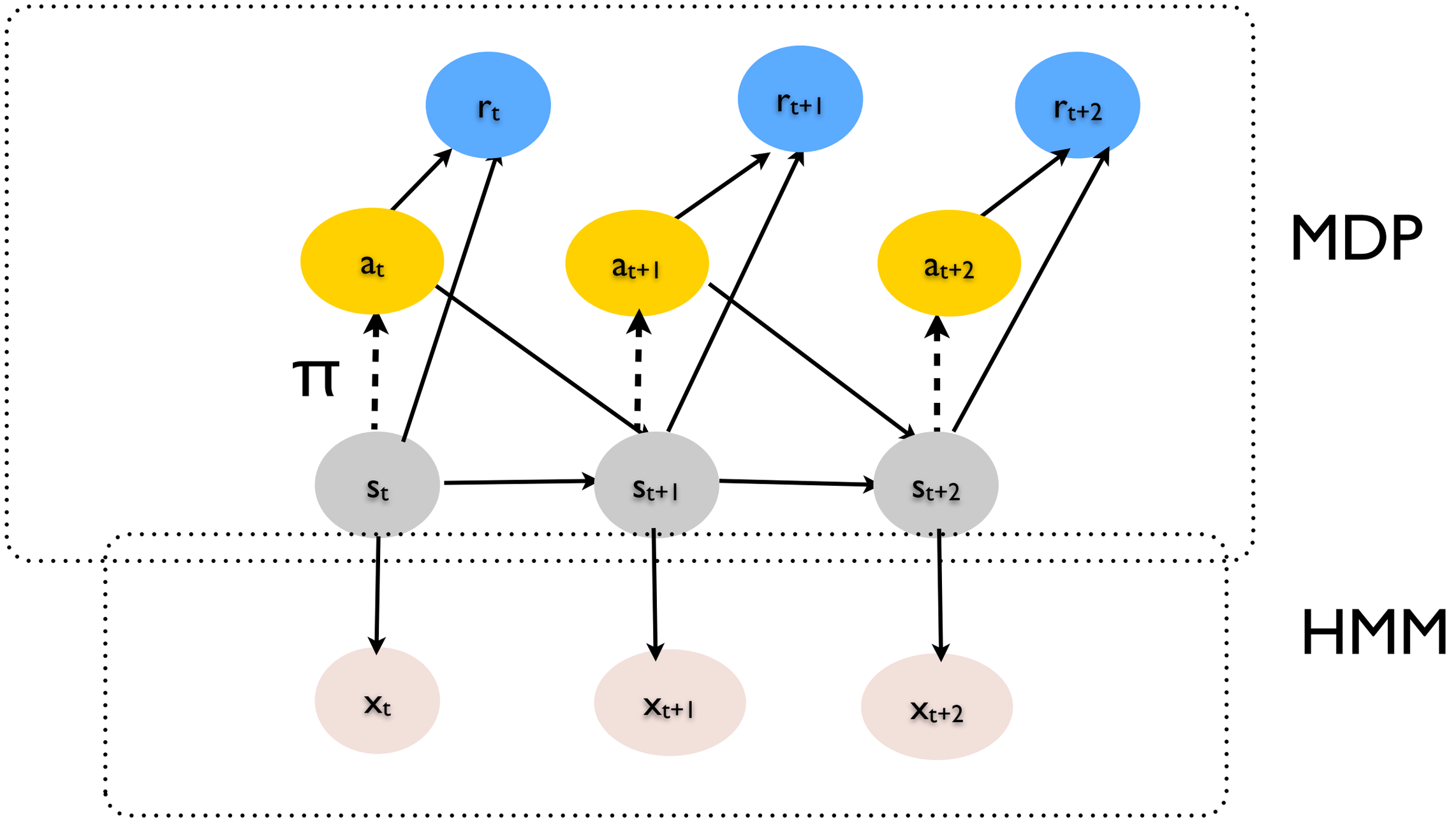
Markov Decision Processes

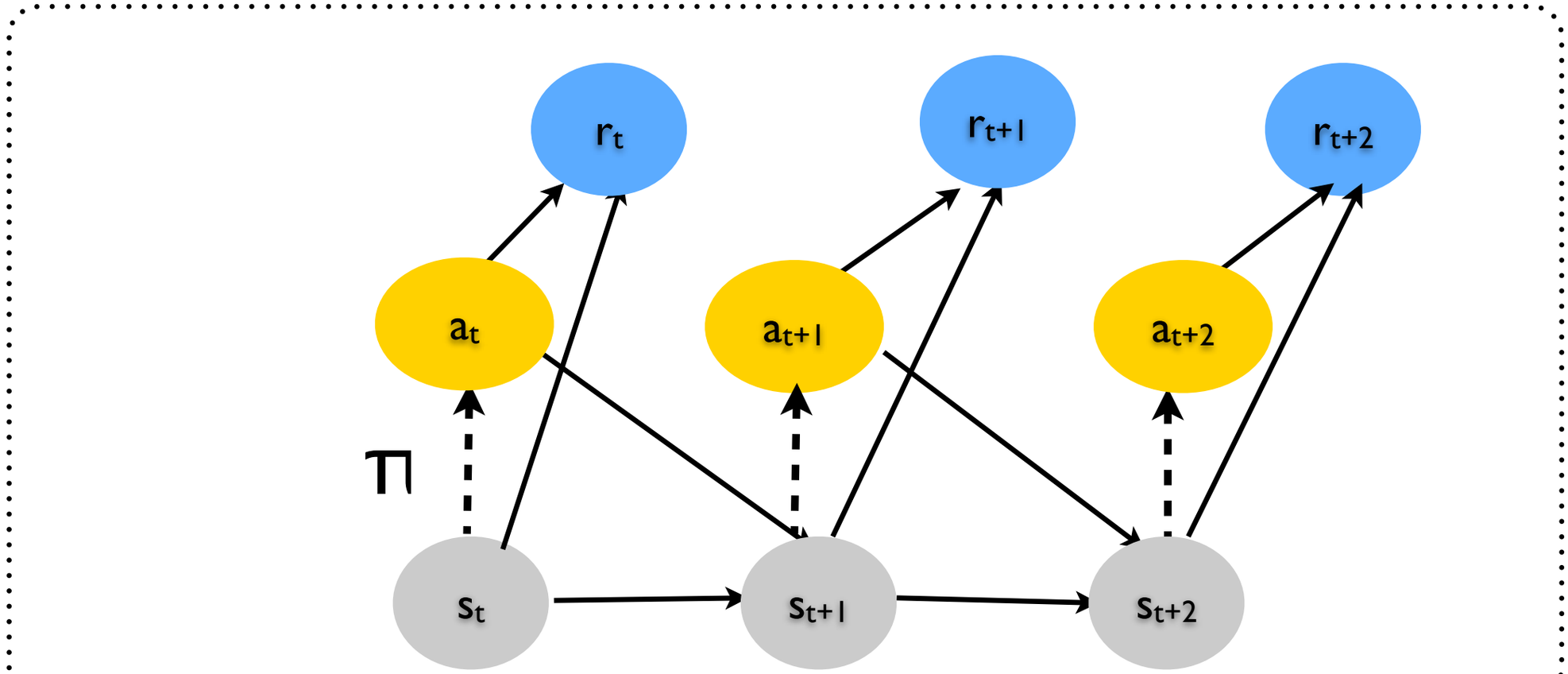
- MDP is a stochastic process that models the environment under different actions.
- It is defined on states, actions and rewards.
- Agent decides on the action, and receives a reward which depends on the action and states

Goal: what action should I take to maximize my chance to win a game?


r_t
r_{t+1}
r_{t+2}
a_t
a_{t+1}
a_{t+2}
π
s_t
s_{t+1}
s_{t+2}
x_t
x_{t+1}
x_{t+2}
MDP


r_t
r_{t+1}
r_{t+2}
a_t
a_{t+1}
a_{t+2}
π
s_t
s_{t+1}
s_{t+2}
x_t
x_{t+1}
x_{t+2}
MDP


r_t
r_{t+1}
r_{t+2}
a_t
a_{t+1}
a_{t+2}
π
s_t
s_{t+1}
s_{t+2}
x_t
x_{t+1}
x_{t+2}
MDP
HMM

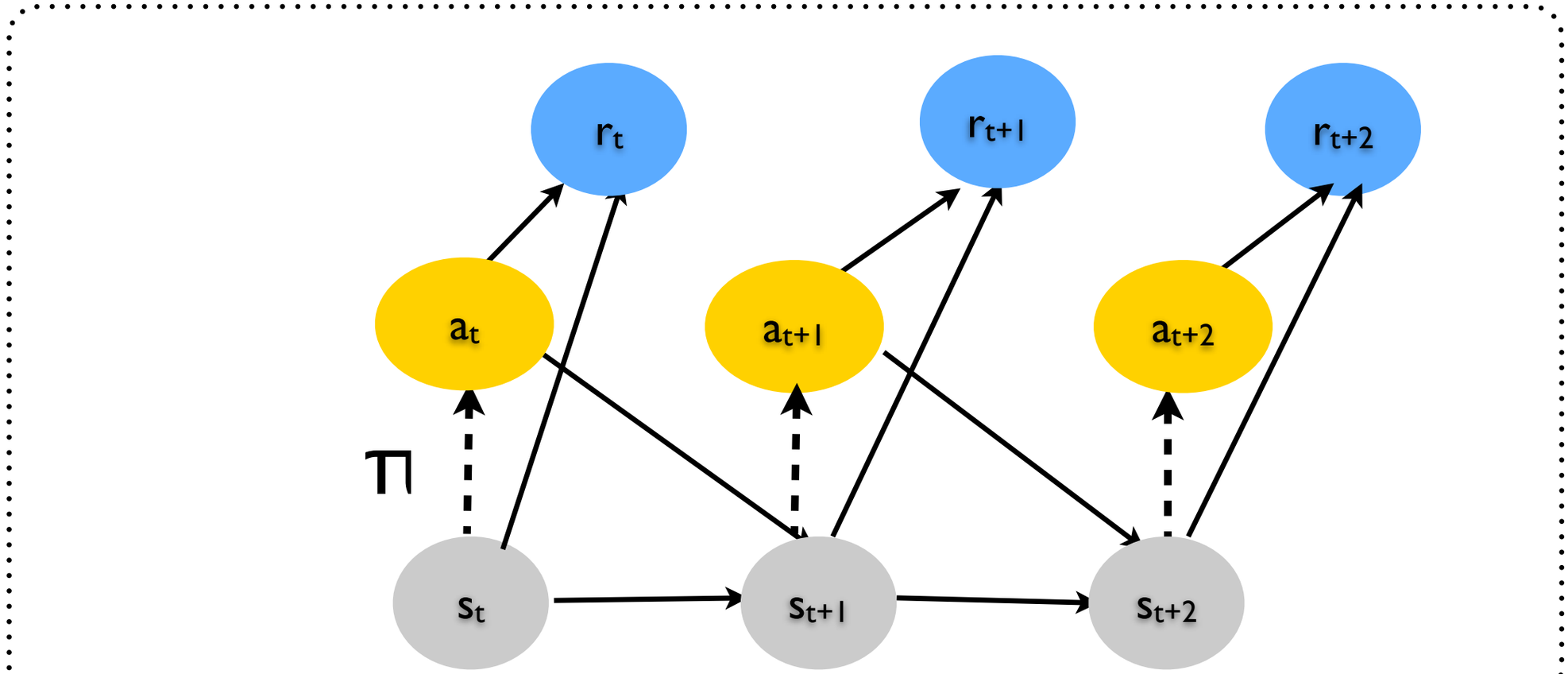


MDP

Transition probability $P(s_{t+1}|a_t, s_t)$

Reward probability $P(r_t|a_t, s_t)$

Policy $P(a_t|s_t)$ $\pi(a_t|s_t)$



MDP

Transition probability $P(s_{t+1}|a_t, s_t)$ P_{st+1st}

Reward probability $P(r_t|a_t, s_t)$

Policy $P(a_t|s_t)$ $\pi(a_t|s_t)$

Markov Decision Processes

- Set of states $S = \{s_1, s_2, ..., s_n\}$
- Set of actions $A = \{a_1, a_2, ..., a_n\}$
- Set of rewards $R = \{r_1, r_2, ..., r_n\}$
- Policy π gives an action for each state, $\pi: S \rightarrow A$

What are the Markov assumptions?

Markov Decision Processes

- Set of states $S = \{s_1, s_2, ..., s_n\}$
- Set of actions $A = \{a_1, a_2, ..., a_n\}$
- Set of rewards $R = \{r_1, r_2, ..., r_n\}$
- Policy π gives an action for each state, $\pi: S \rightarrow A$

What are the Markov assumptions?

$$P(r_t|s_t,a_t,s_{t-1},a_{t-1},..) = P(r_t|s_t,a_t)$$

$$P(s_{t+1}|s_t,a_t,s_{t-1},a_{t-1},..) = P(s_{t+1}|s_t,a_t)$$

Given a policy, π and a start state s, we can calculate the expectation of future rewards.

Value function, $V^{\pi}(s)$ is a measure for the expected discounted return:

$$V^{\pi}(s) = E\{r_0 + \gamma r_1 + \gamma^2 r_2 + ...|s_0 = s; \pi\}$$

$$= E\{\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s; \pi\}$$

$$= \sum_{t=0}^{\infty} \gamma^t E\{r_t | s_0 = s; \pi\}$$

Given a policy, π and a start state s, we can calculate the expectation of future rewards.

Value function, $V^\pi(s)$ is a measure for the expected discounted return:

$$V^\pi(s) = E\{r_0 + \gamma r_1 + \gamma^2 r_2 + ...|s_0 = s; \pi\}$$

$$= E\{\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s; \pi\}$$

$$= \sum_{t=0}^{\infty} \gamma^t E\{r_t | s_0 = s; \pi\}$$

Discounted by γ exponentially to the future

Given a policy, π and a start state s, we can calculate the expectation of future rewards.

Value function, $V^\pi(s)$ is a measure for the expected discounted return:

$$V^\pi(s) = E\{r_0 + \gamma r_1 + \gamma^2 r_2 + ...|s_0 = s; \pi\}$$

$$= E\{\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s; \pi\}$$

$$= \sum_{t=0}^{\infty} \gamma^t E\{r_t | s_0 = s; \pi\}$$

Discounted by γ exponentially to the future

$$V^\pi(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

Given a policy, π and a start state s, we can calculate the expectation of future rewards.

Value function, $V^\pi(s)$ is a measure for the expected discounted return:

$$V^\pi(s) = E\{r_0 + \gamma r_1 + \gamma^2 r_2 + ... | s_0 = s; \pi\}$$

$$= E\{\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s; \pi\}$$

$$= \sum_{t=0}^{\infty} \gamma^t E\{r_t | s_0 = s; \pi\}$$

Discounted by γ exponentially to the future

$$V^\pi(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

Does it converge?

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

But this for only one policy, π

How about the optimal value function, $V^*(s)$?

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

But this for only one policy, π

How about the optimal value function, $V^*(s)$?

Optimal value function

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

But this for only one policy, π

How about the optimal value function, $V^*(s)$?

Optimal value function

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

But this for only one policy, π

How about the optimal value function, $V^*(s)$?

Optimal value function

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

Optimal policy

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

But this for only one policy, π

How about the optimal value function, $V^*(s)$?

Optimal value function

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

Optimal policy

$$\pi^* \text{ optimal} \Leftrightarrow \forall_s : V^{\pi^*}(s) = V^*(s)$$

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E[r_t]$$

But this for only one policy, π

How about the optimal value function, $V^*(s)$?

Optimal value function

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

Optimal policy

$$\pi^* \text{ optimal} \Leftrightarrow \forall_s : V^{\pi^*}(s) = V^*(s)$$

Is there really an optimal policy for every MDP?

We can use the recursive property to compute the value function for each state

$$
\begin{aligned}
V^{\pi}(s) &= E\{r_0 + \gamma r_1 + \gamma^2 r_2 + ... | s_0 = s; \pi\} \\
&= E\{r_0 | s_0 = s; \pi\} + \gamma E\{r_1 + \gamma r_2 + \gamma^2 r_3 + ... | s_0 = s; \pi\} \\
&= R(\pi(s), s) + \gamma \sum_{s'} P(s' | \pi(s), s) E\{r_1 + \gamma r_2 + \gamma^2 r_3 + ... | s_1 = s'; \pi\} \\
&= R(\pi(s), s) + \gamma \sum_{s'} P(s' | \pi(s), s) V^{\pi}(s')
\end{aligned}
$$

For all possible transitions from state s

Recursive property holds for optimal value function V^*

Bellman's Equation

With actions

$$V^*(s) = \max_a [R(a,s) + \gamma \sum_{s'} P(s'|\pi(s), s) V^*(s')]$$

$$\pi^*(s) = \text{argmax}_a [R(a,s) + \gamma \sum_{s'} P(s'|\pi(s), s) V^*(s')]$$

No actions

$$V^*(s) = r_s + \gamma \sum_{s'} P(s'|\pi(s), s) V^*(s')$$

$$\pi^*(s) = r_s + \gamma \sum_{s'} P(s'|\pi(s), s) V^*(s')$$

Suppose we have n states:

$$V^*(s_1) = r_{s_1} + \gamma\left(p_{s_1 s_1}V^*(s_1) + p_{s_1 s_2}V^*(s_2) + \cdots + p_{s_1 s_n}V^*(s_n)\right)$$
$$V^*(s_2) = r_{s_1} + \gamma\left(p_{s_2 s_1}V^*(s_1) + p_{s_1 s_2}V^*(s_2) + \cdots + p_{s_n s_n}V^*(s_n)\right)$$
$$V^*(s_n) = r_{s_1} + \gamma\left(p_{s_n s_1}V^*(s_1) + p_{s_n s_2}V^*(s_2) + \cdots + p_{s_n s_n}V^*(s_n)\right)$$

It can be solved in n equations in closed form.

But we may not be able to do this every time, so we consult to value/ policy iteration to find the optimal values.

Outline

1. MDP overview
2. Value Iteration
3. Policy Iteration
4. Previous Exam Questions

Value iteration

Iterative algorithm, start with $V^0(s_i)$
This could be initialized to 0

$$V^1(s_1) = r_{s_1}$$

$$V^2(s_1) = r_{s_1} + \gamma(\sum_k p_{s_1 s_k} V^1(s_k))$$

$$V^{t+1}(s_1) = r_{s_1} + \gamma(\sum_k p_{s_1 s_k} V^t(s_k))$$

$$|V_{s_i}^{t+1} - V_{s_i}^t|_{t \to \infty} < \epsilon$$

Outline

1. MDP overview
2. Value Iteration
3. Policy Iteration
4. Previous Exam Questions

POLICY ITERATION

- Initialization:

  Randomly choose π_0 set t = 0
- Policy evaluation:

  For each s_i, compute $V^*(s_i)$
- Policy update:

$$\pi_t(s_i) = \max_a r_i + \gamma(\sum_j P(s_j|a, s_i)V^*(s_j)$$

- If not converged, t = t+ l

Policy iteration vs Value Iteration

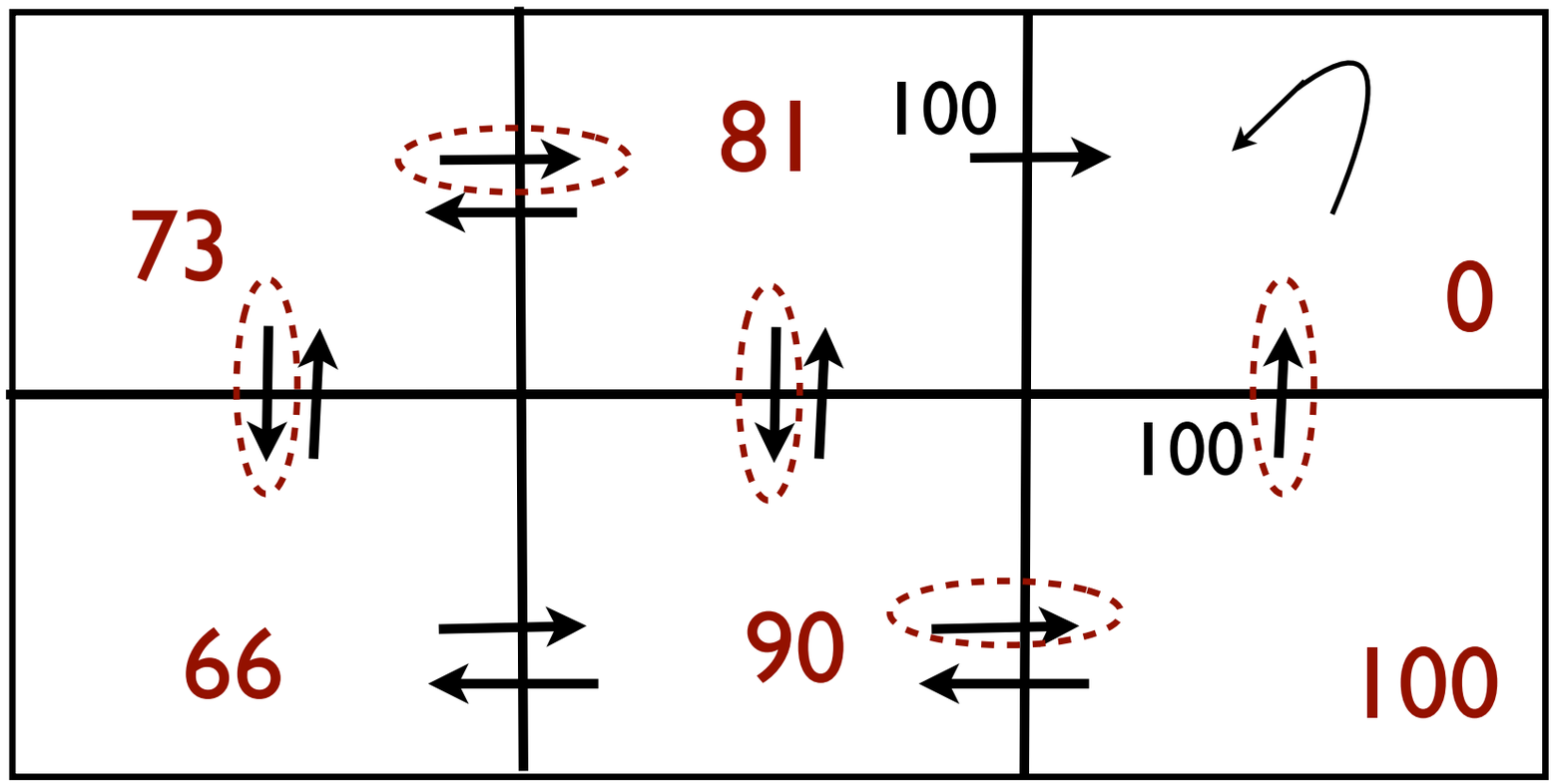
- Policy iteration is good when the initial policy guess is right
- If we have too many actions value iteration may be slow
- Otherwise value iteration is safer

Markov Decision Processes

1. MDP overview
2. Value Iteration
3. Policy Iteration
4. Previous Exam Questions

Warm-up

Suppose policy π is circled. Suppose γ is 0.9. What are the V^{π} (s) values? Immediate rewards are written next to transitions, transitions with no immediate reward has 0 value.

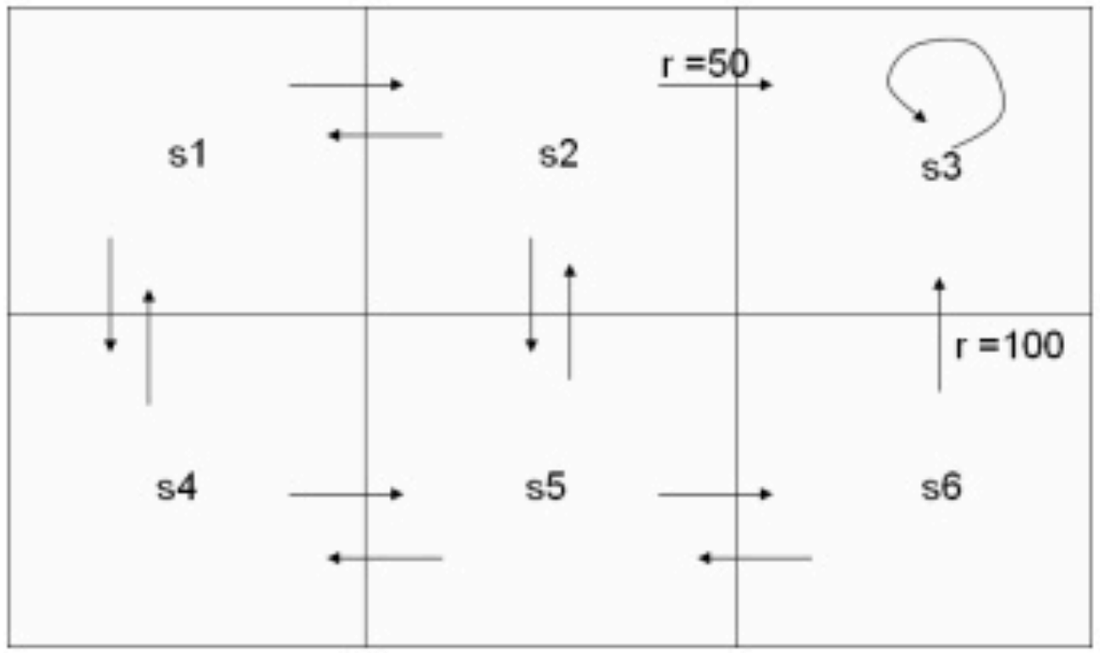


10-701 Final Exam, Fall 2006

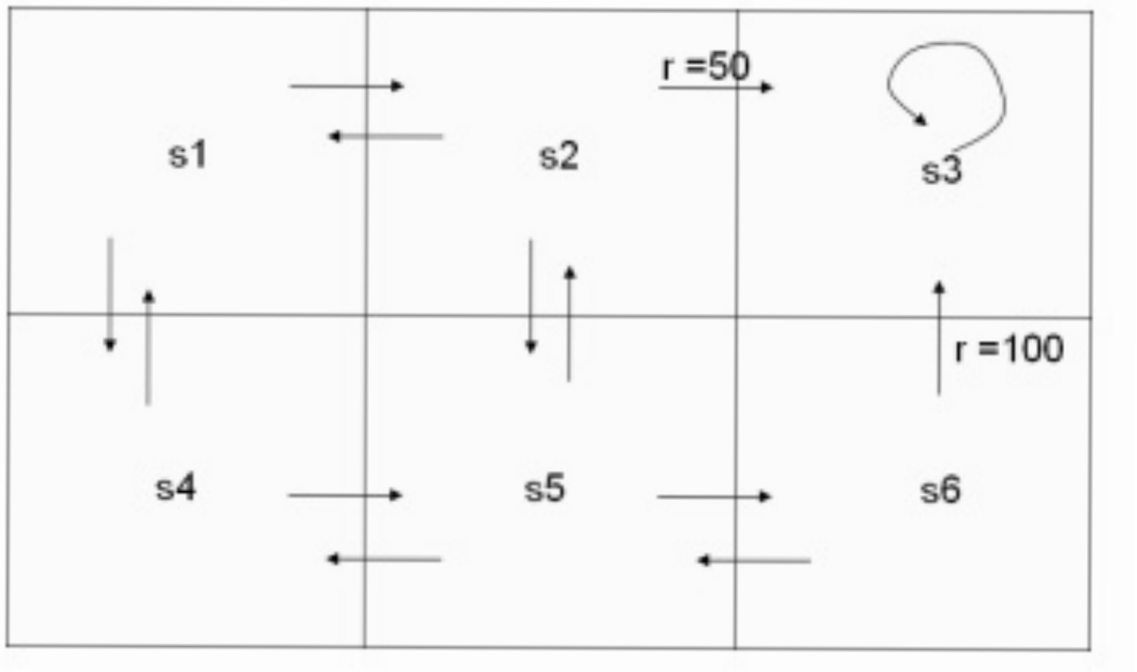
9 MDPs and Reinforcement Learning [16pts]

Part A. [10pts]

Consider the following deterministic Markov Decision Process (MDP), describing a simple robot grid world. Notice the values of the *immediate rewards* are written next to transitions. Transitions with no value have an immediate reward of 0. **Assume the discount factor** $\gamma = 0.8$.

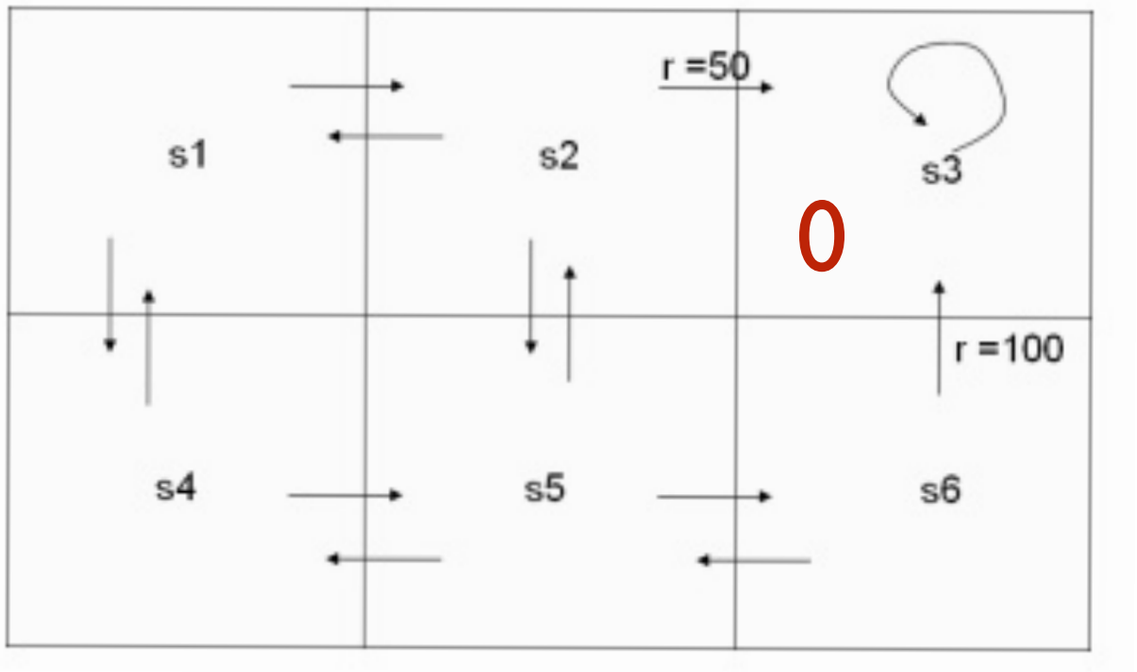


A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



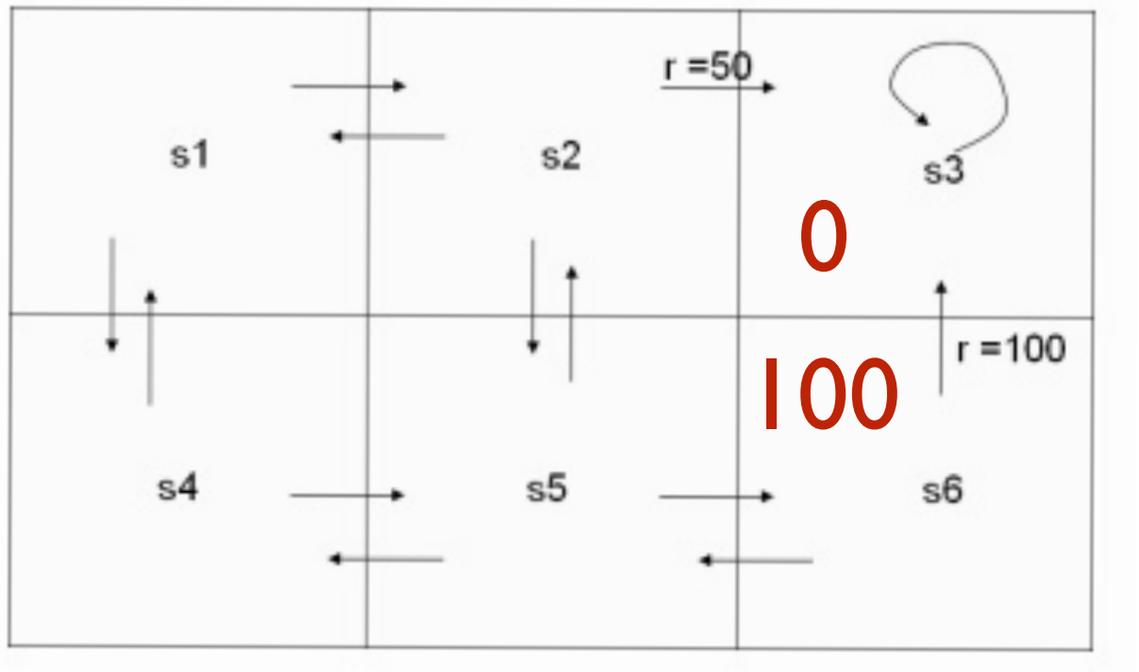
$\gamma = 0.8$

A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



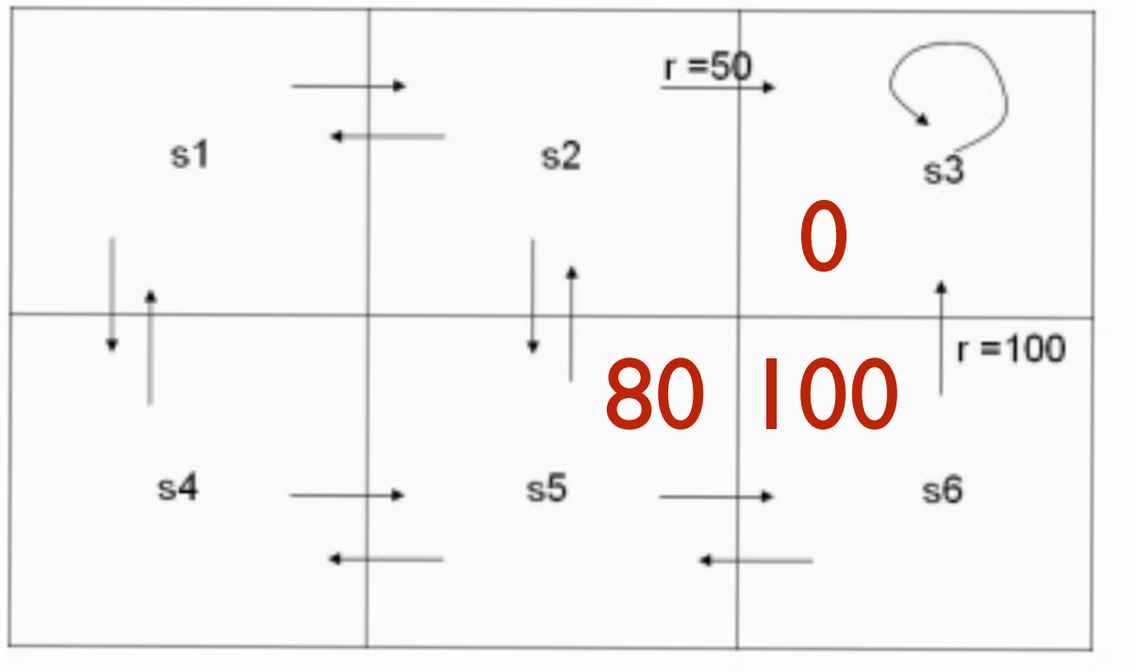
$\gamma = 0.8$

A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



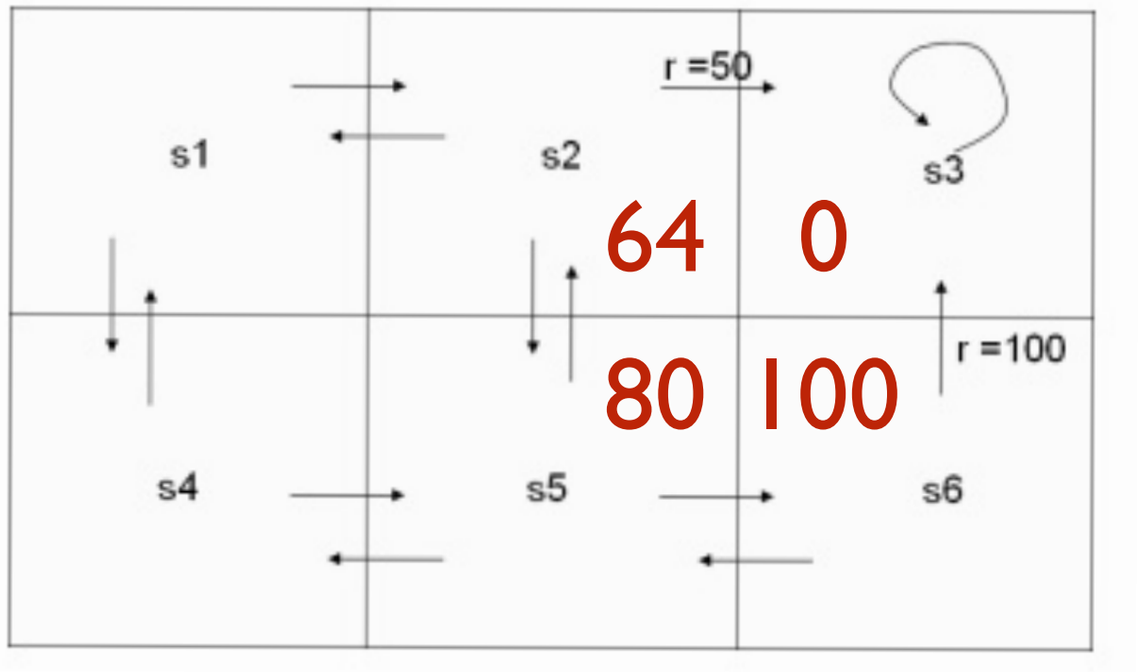
$\gamma = 0.8$

A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



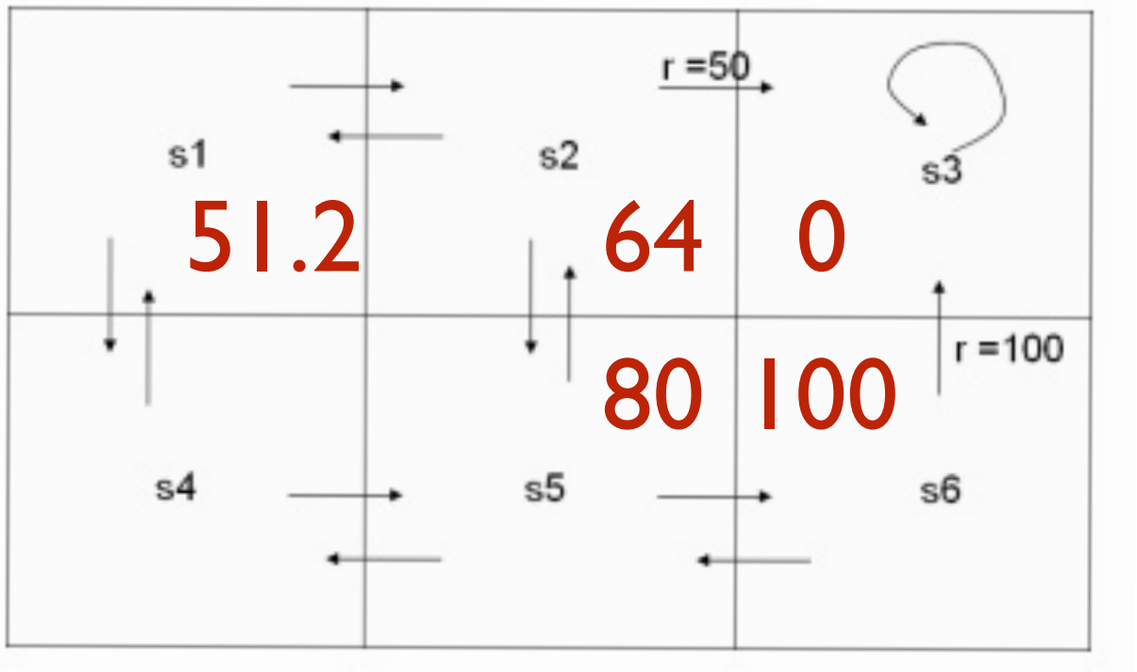
$\gamma = 0.8$

A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



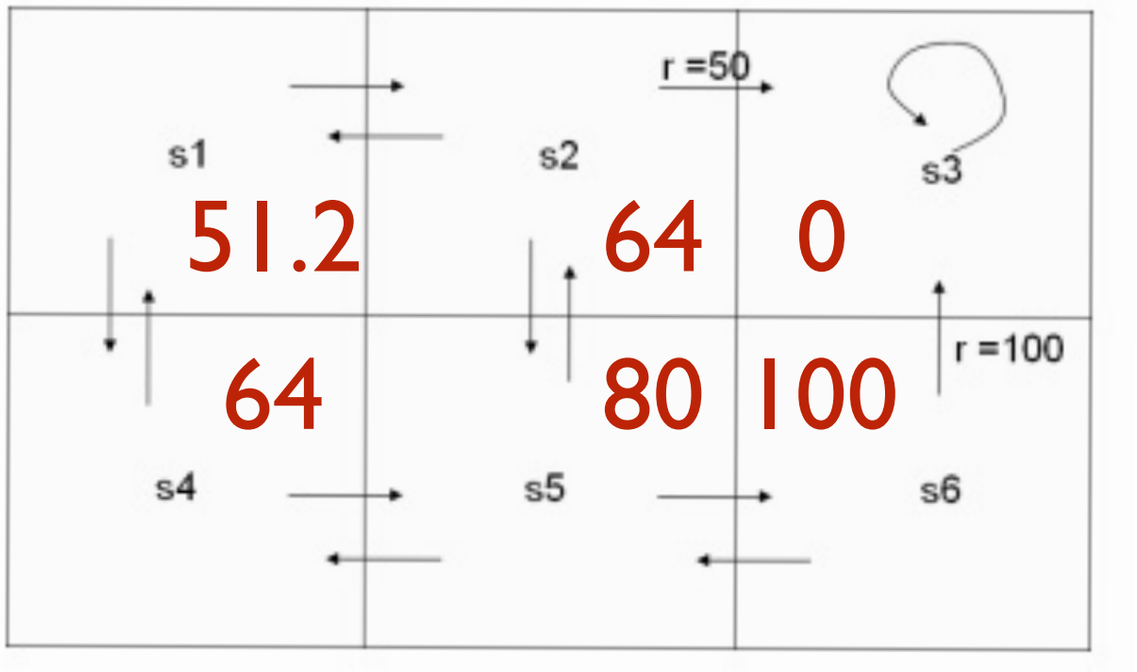
$\gamma = 0.8$

A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



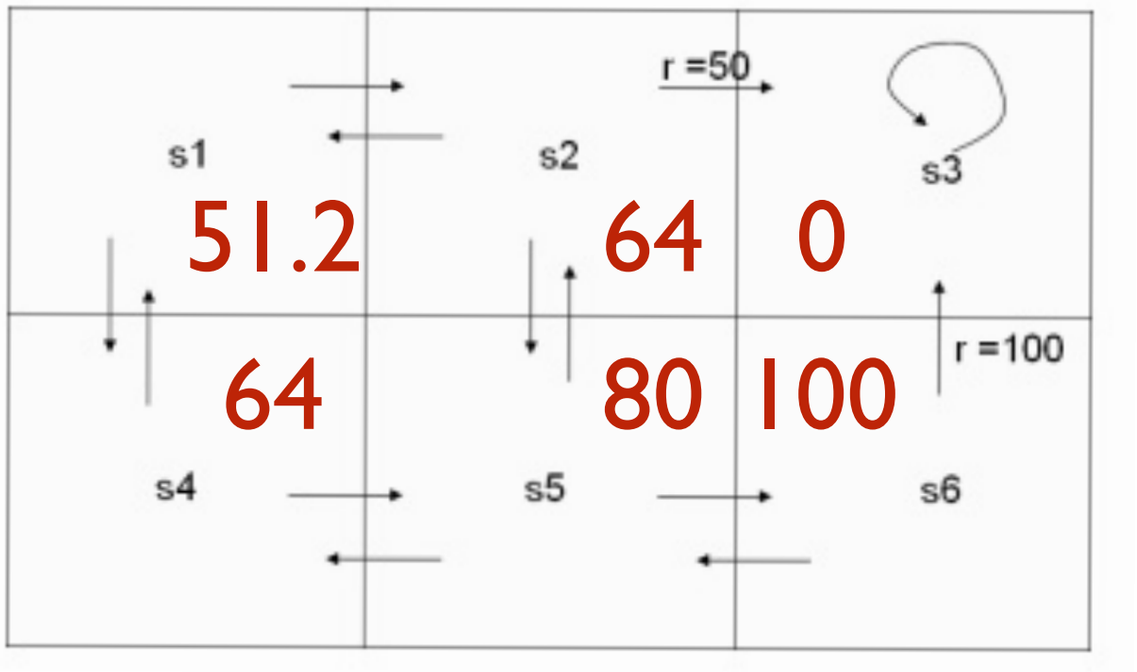
$\gamma = 0.8$

A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



$\gamma = 0.8$

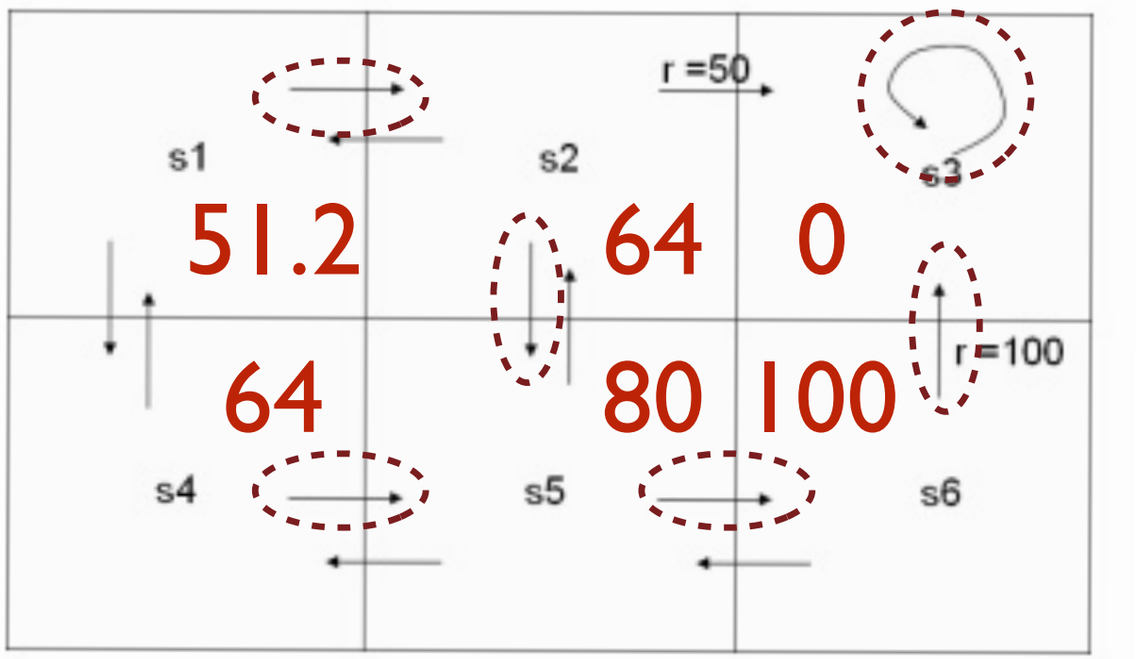
A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



$\gamma = 0.8$

B. Mark the state-action transition arrows that correspond to one optimal pol
If there is a tie, always choose the state with the smallest index.

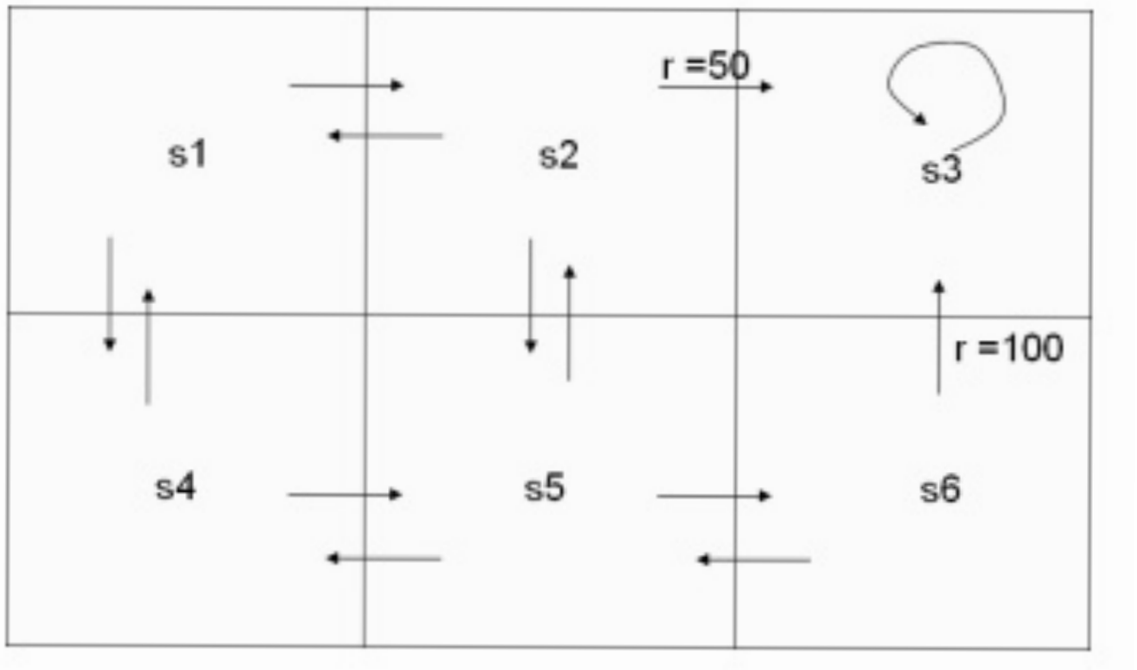
A. For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram



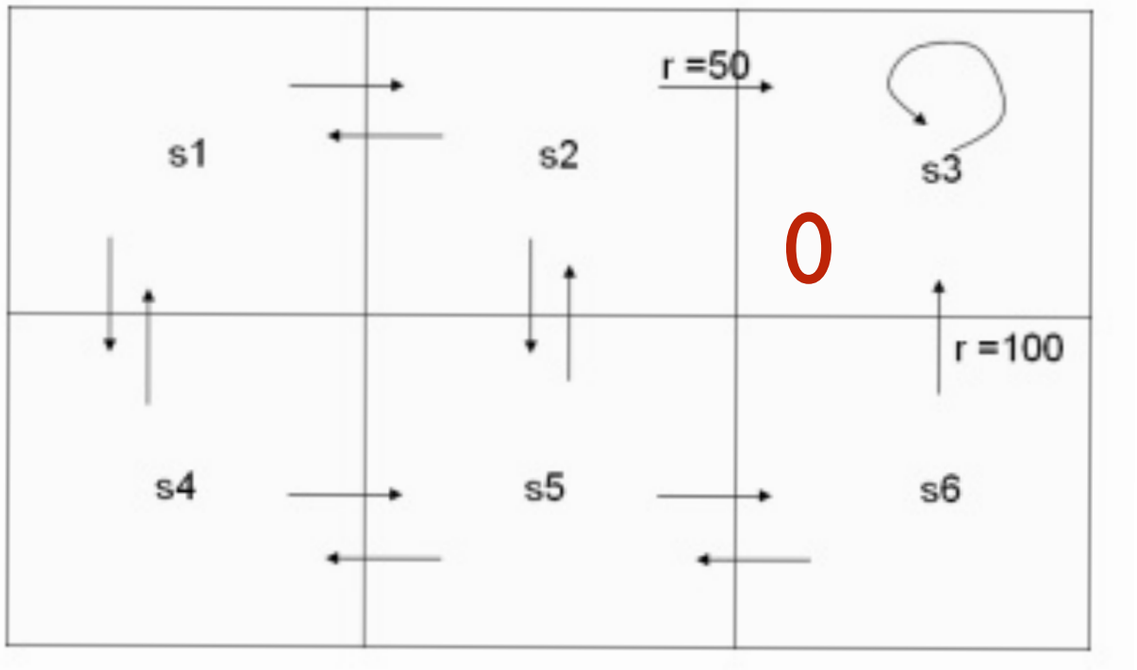
$\gamma = 0.8$

B. Mark the state-action transition arrows that correspond to one optimal pol
If there is a tie, always choose the state with the smallest index.

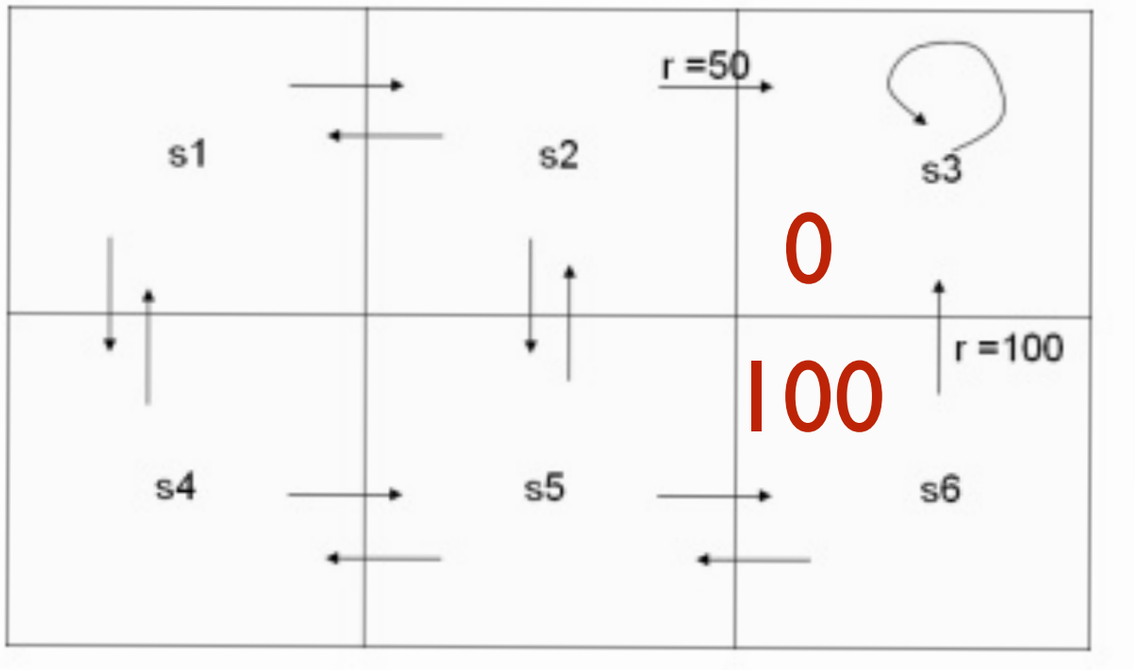
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



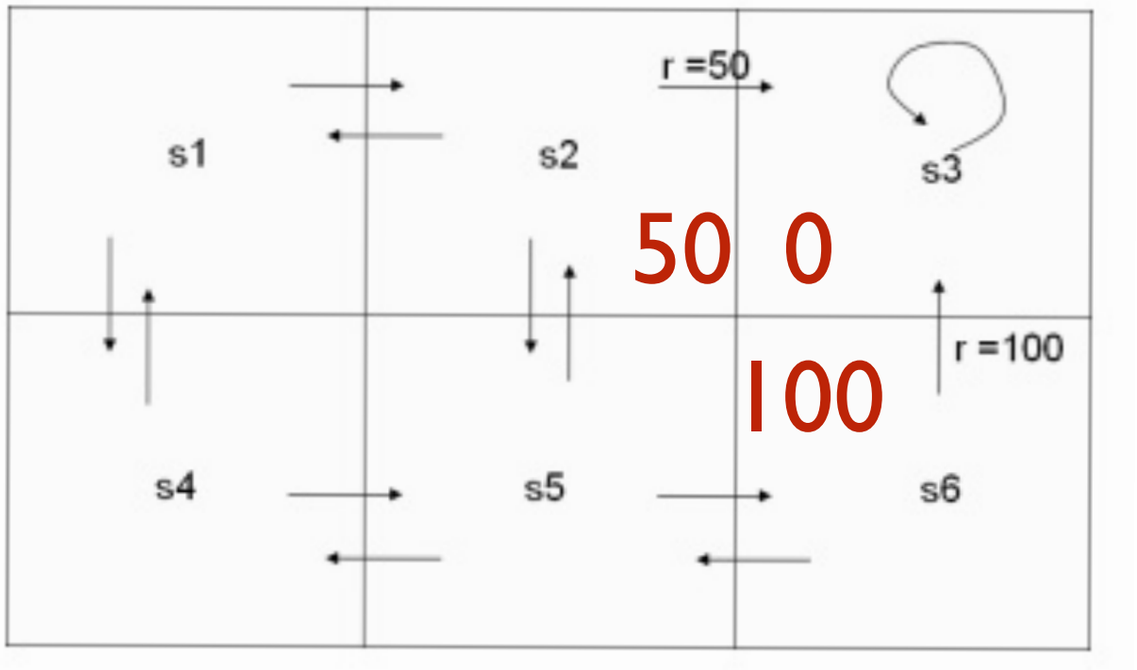
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



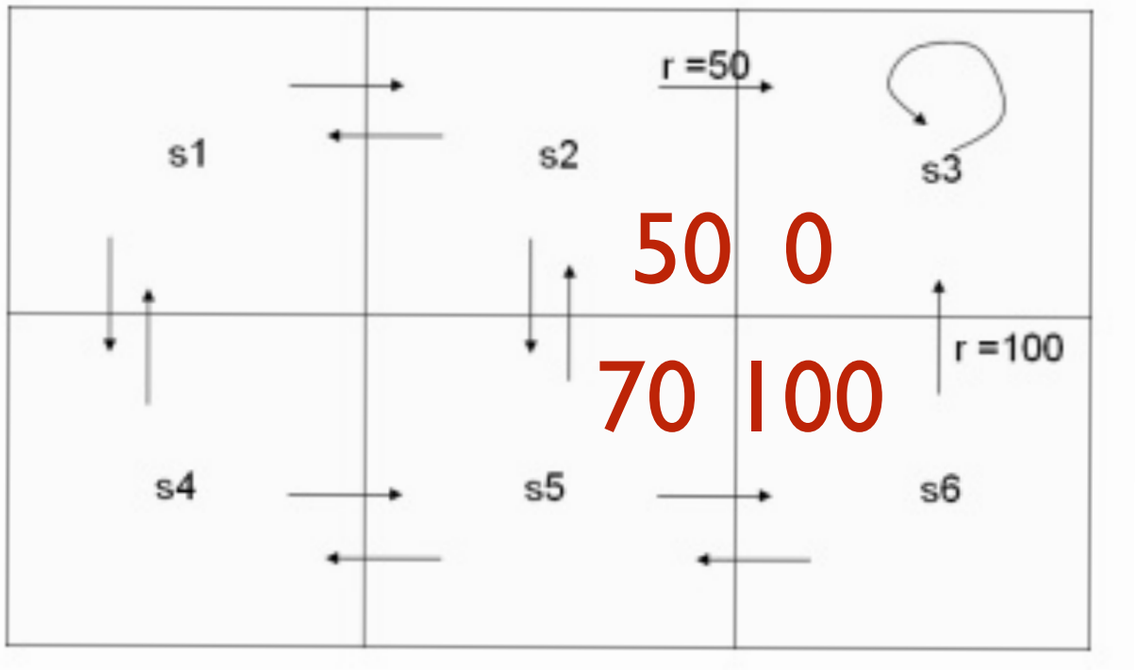
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



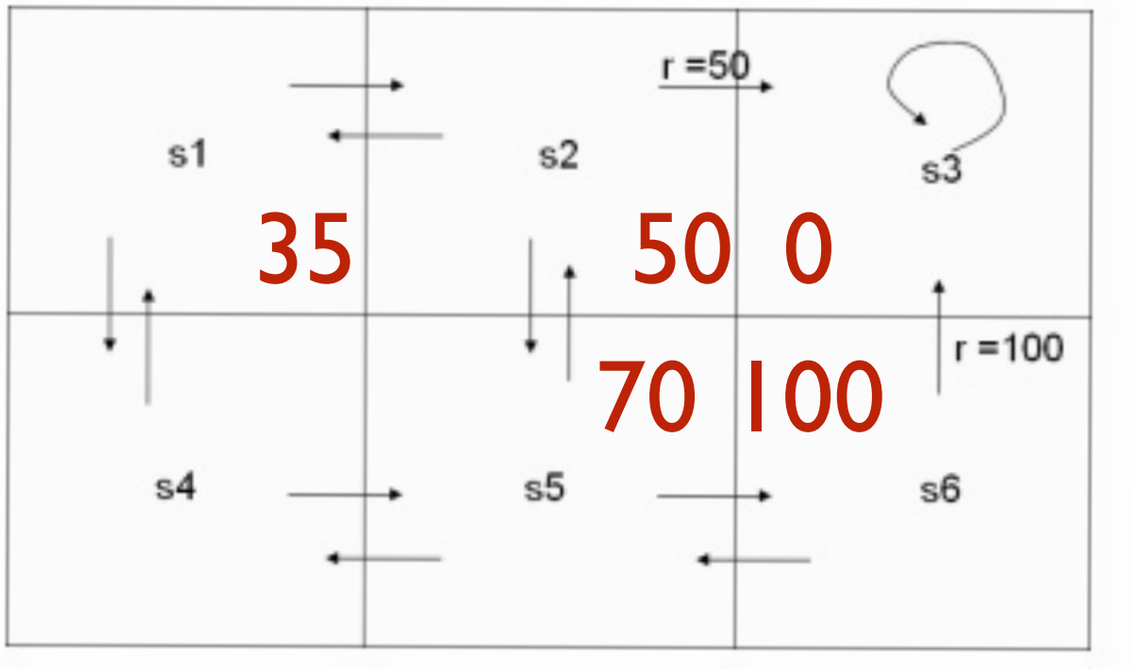
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



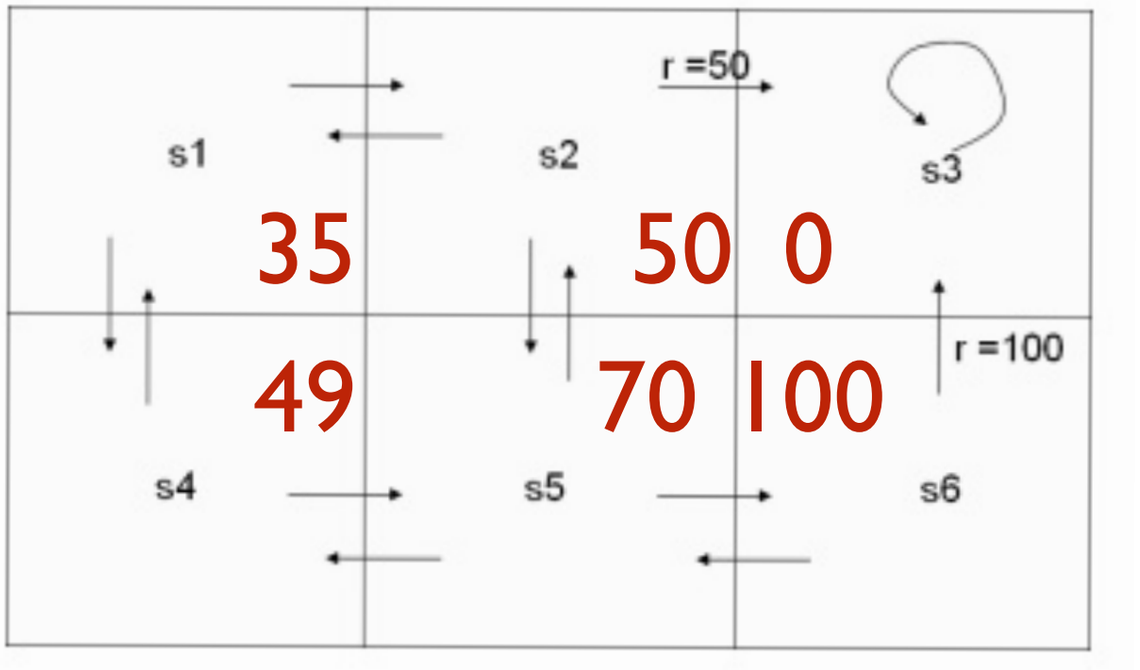
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



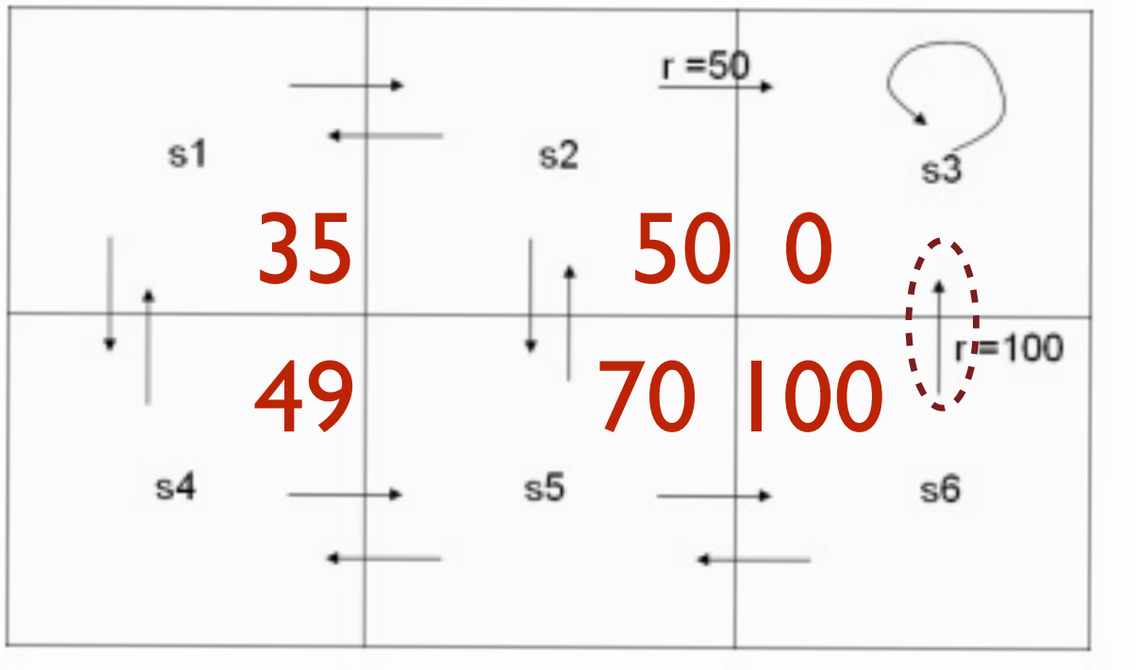
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



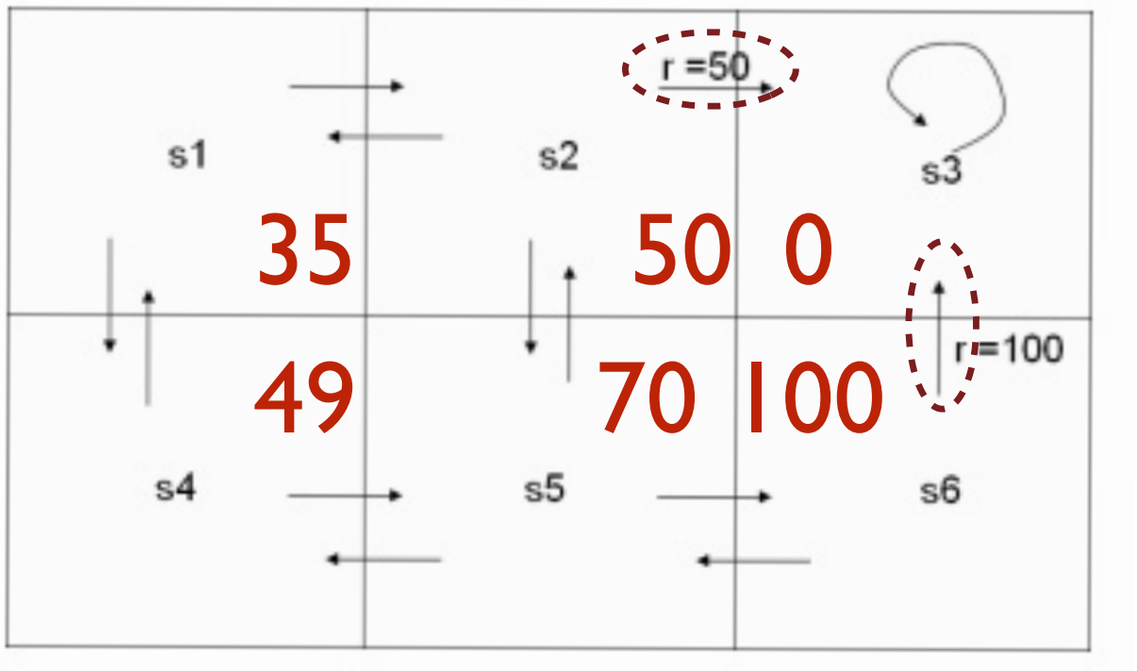
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



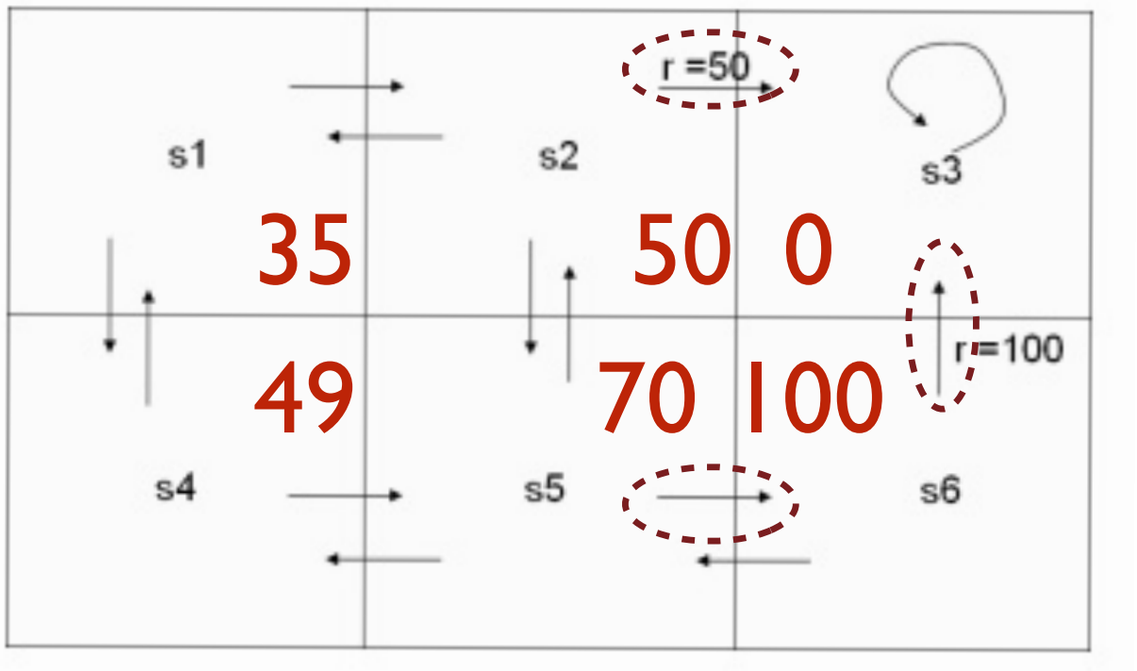
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



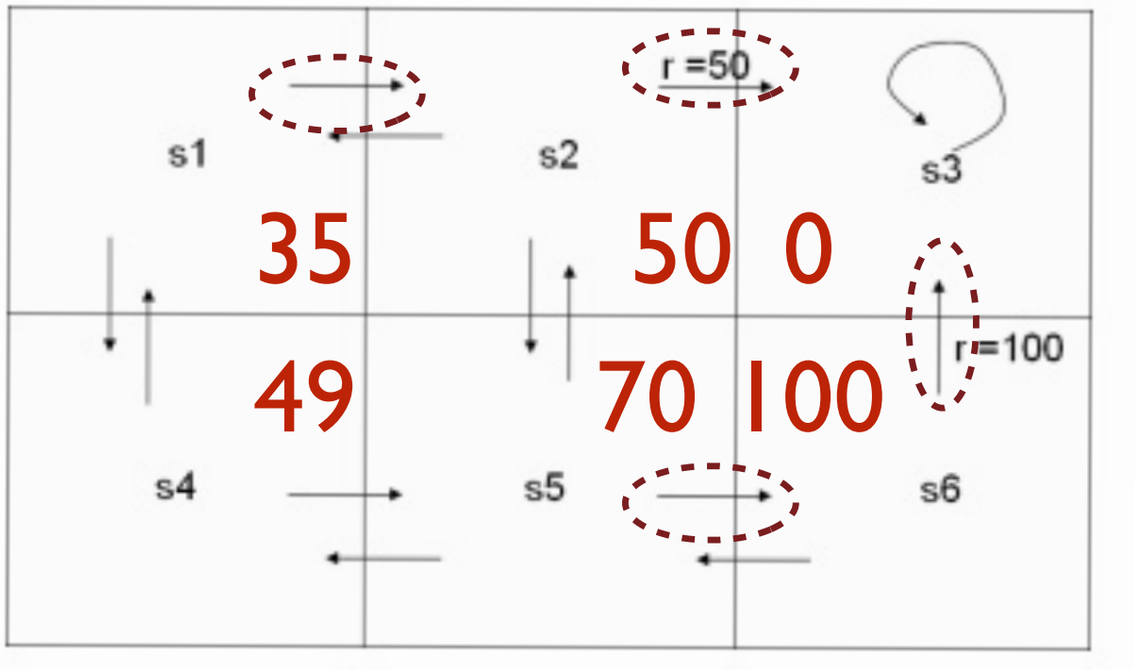
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



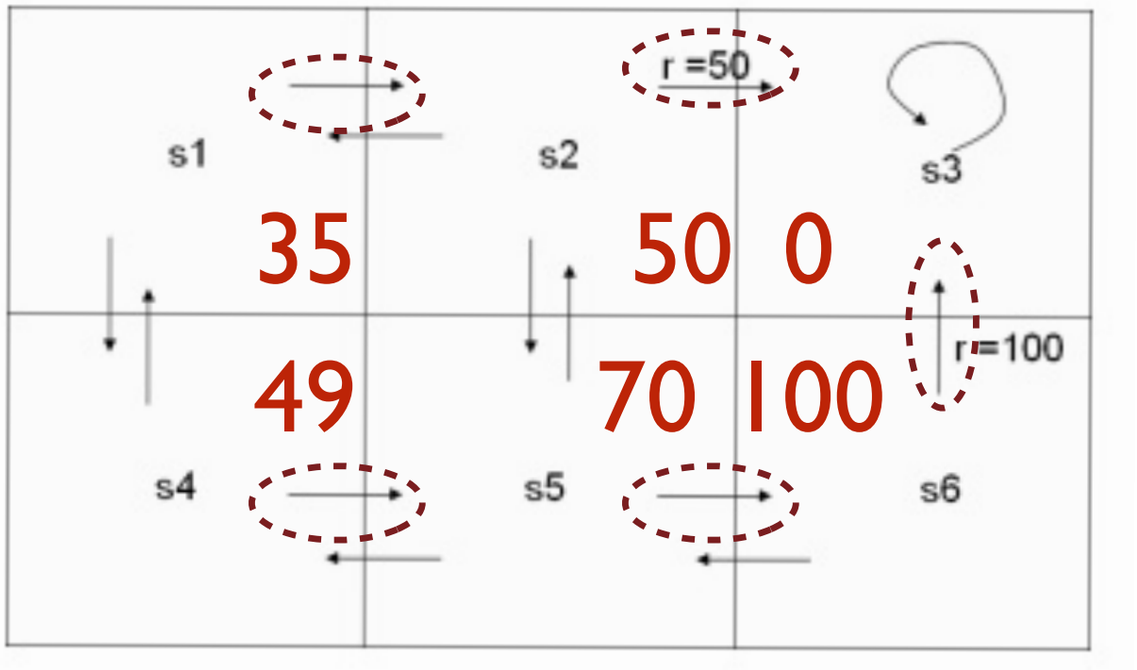
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



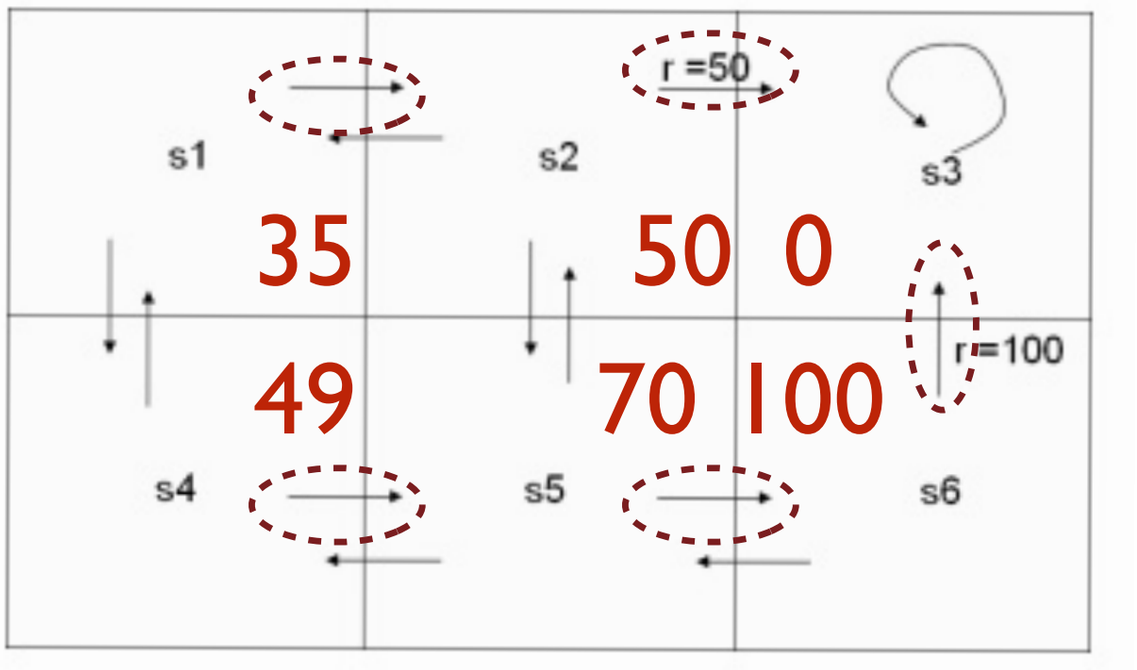
C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.



C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.

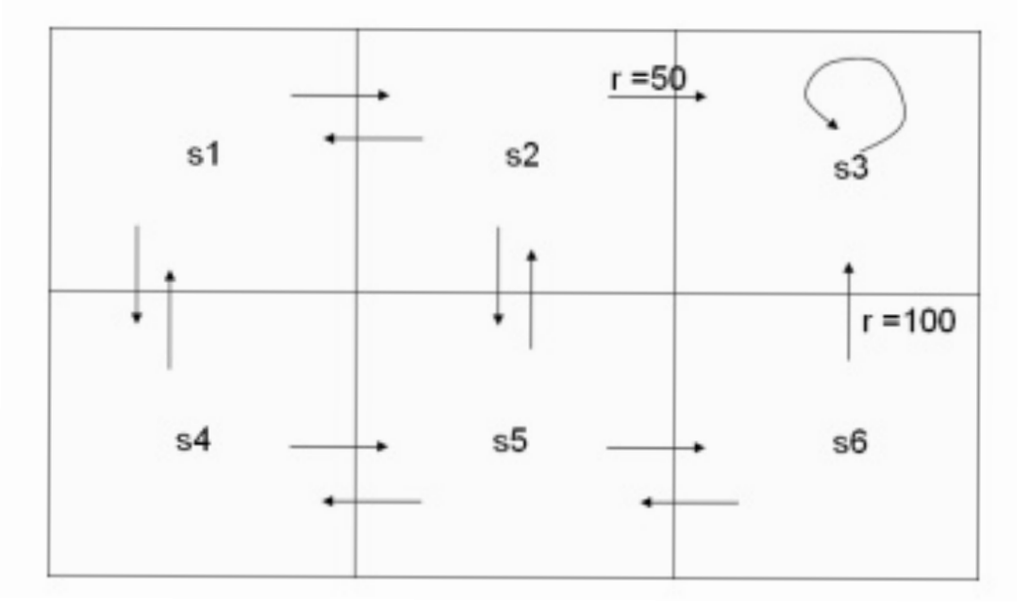


C. Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which ¼(s) values (i.e., which policy actions) change.

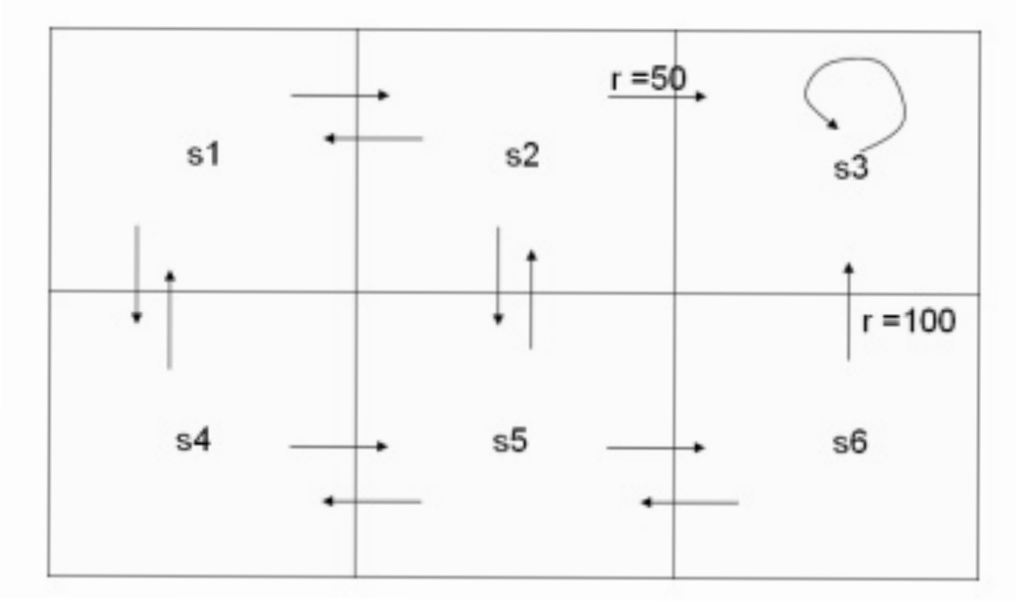


New value for γ: 0.7
Changed policy actions: S2 -> S3

A4. How many complete loops (iterations) of value iteration are sufficient to guarantee finding the optimal policy for this MDP? Assume that values are initialized to zero, and that states are considered in an arbitrary order on each iteration.



A4. How many complete loops (iterations) of value iteration are sufficient to guarantee finding the optimal policy for this MDP? Assume that values are initialized to zero, and that states are considered in an arbitrary order on each iteration.



	S1	S2	S3	S4	S5	S6
T=0	0	0	0	0	0	0
T=1	0	50	0	0	0	100
T=2	40	50	0	0	80	100
T=3	40	64	0	64	80	100
T=4	51.2	64	0	64	80	100

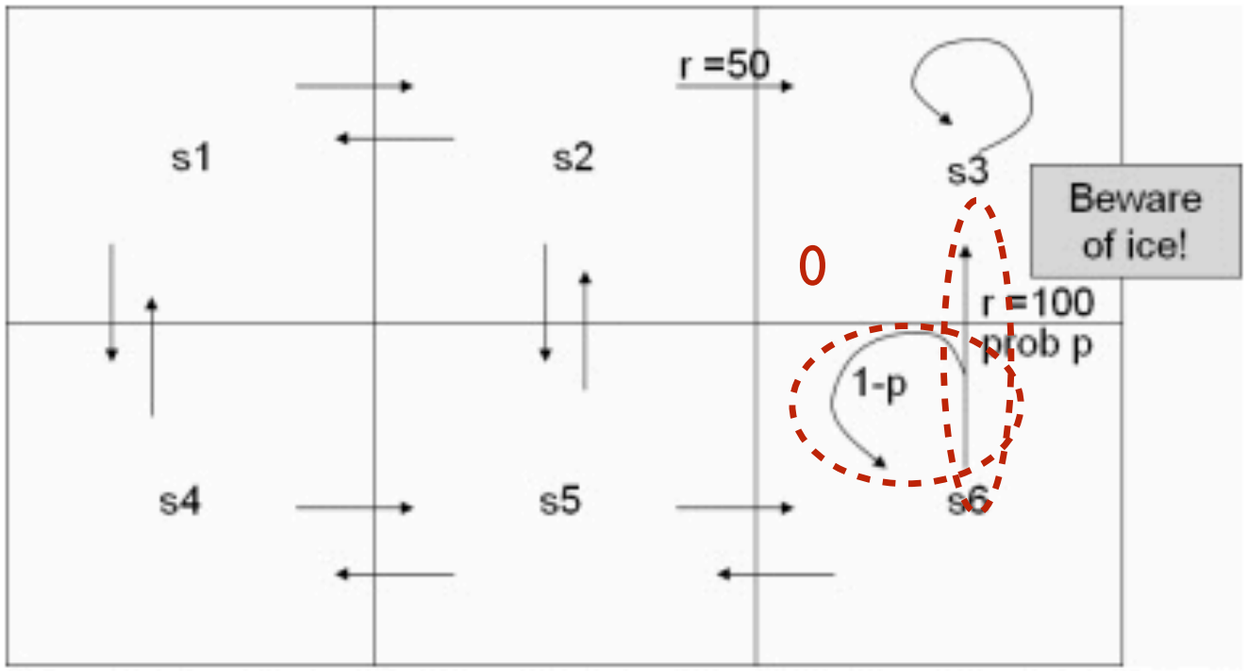
A5. Is it possible to change the immediate reward function so that V^* changes but the optimal policy π^* remains unchanged? If yes, give such a change, and describe the resulting change to V. Otherwise, explain in at most 2 sentences why this is impossible.

A5. Is it possible to change the immediate reward function so that V^* changes but the optimal policy π^* remains unchanged? If yes, give such a change, and describe the resulting change to V. Otherwise, explain in at most 2 sentences why this is impossible.

Modify each reward equally, V^* will change but the optimal policy will remain the same

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



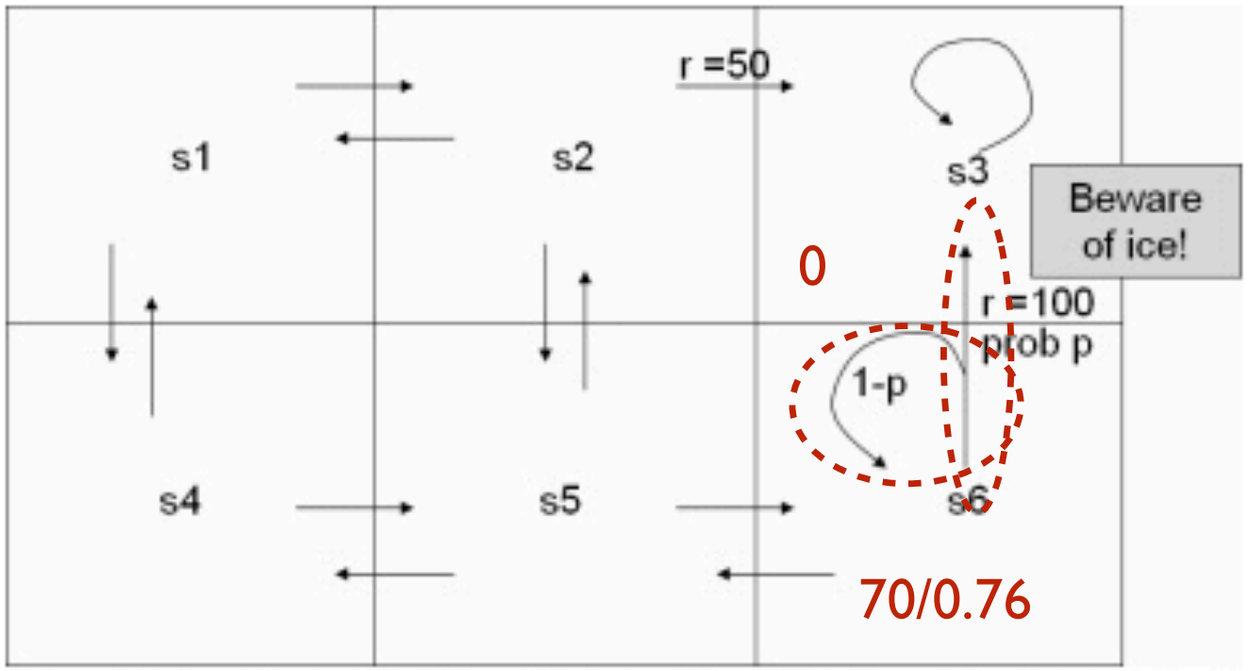
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



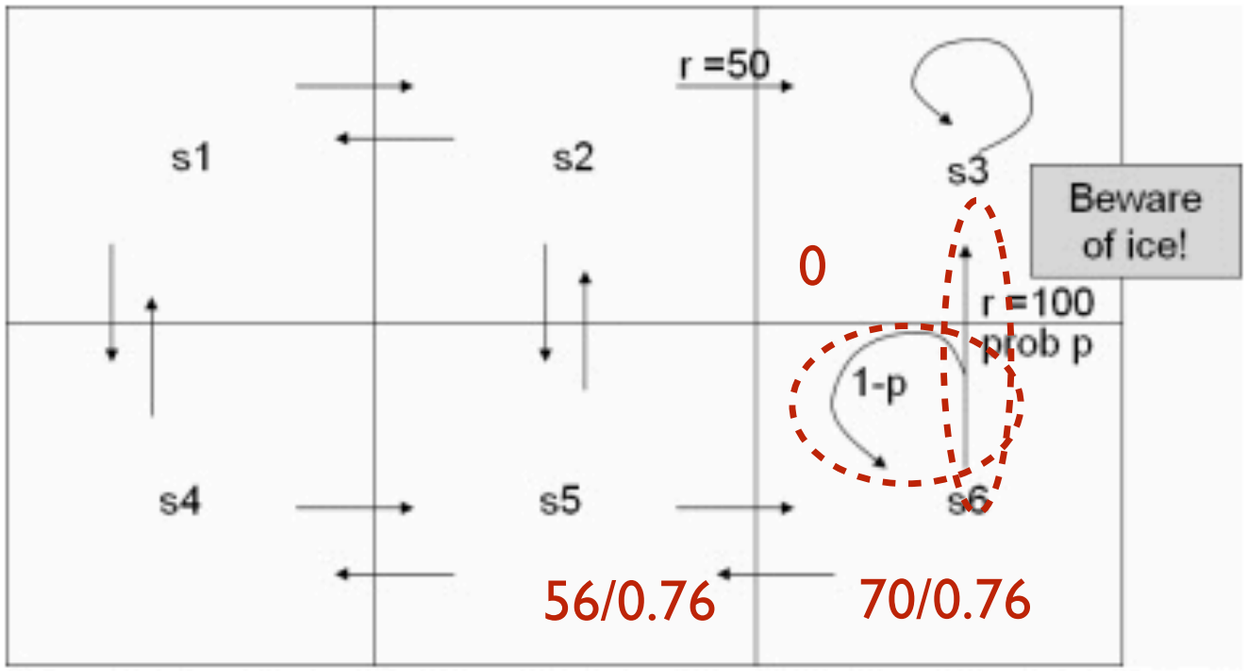
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



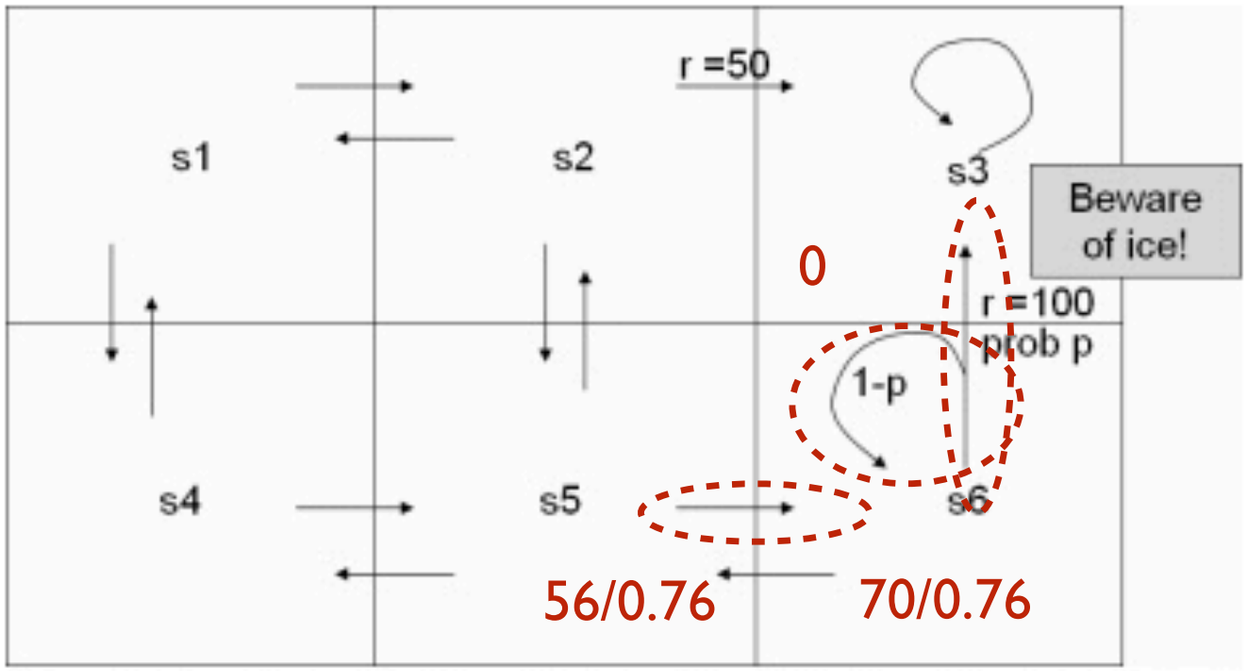
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



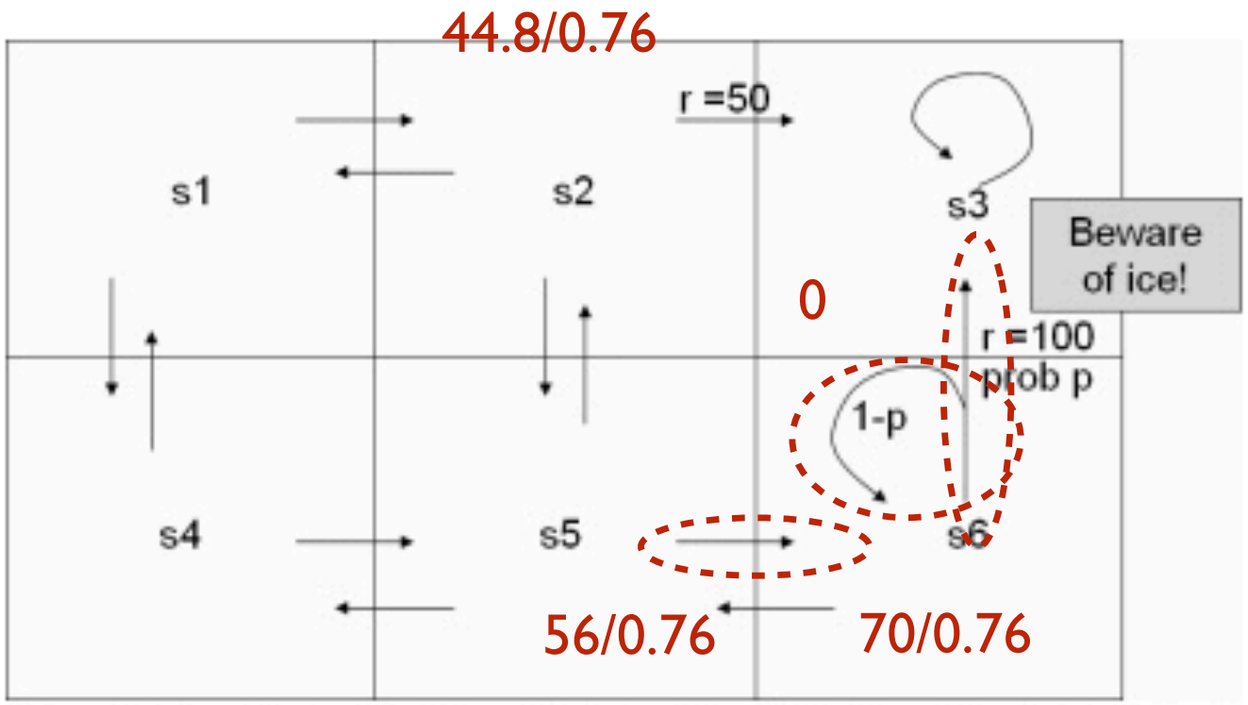
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s)r(s', a, s) + \gamma P(s'|\pi(s), s)V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



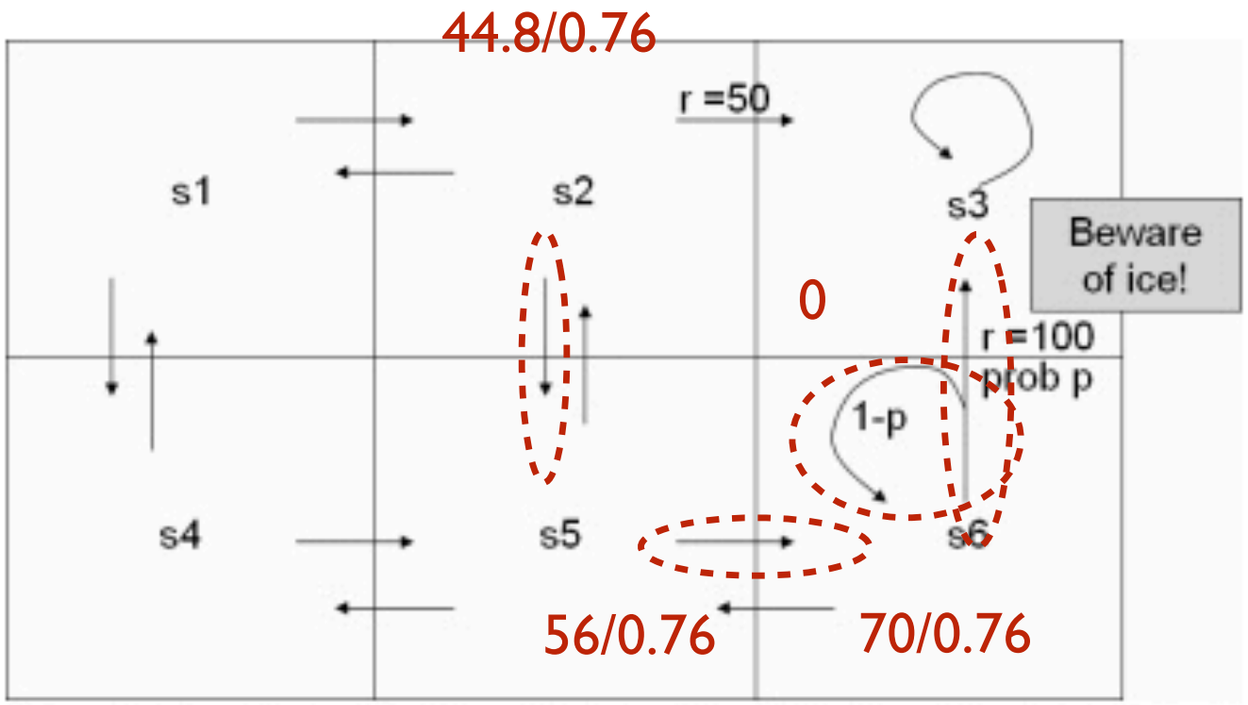
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



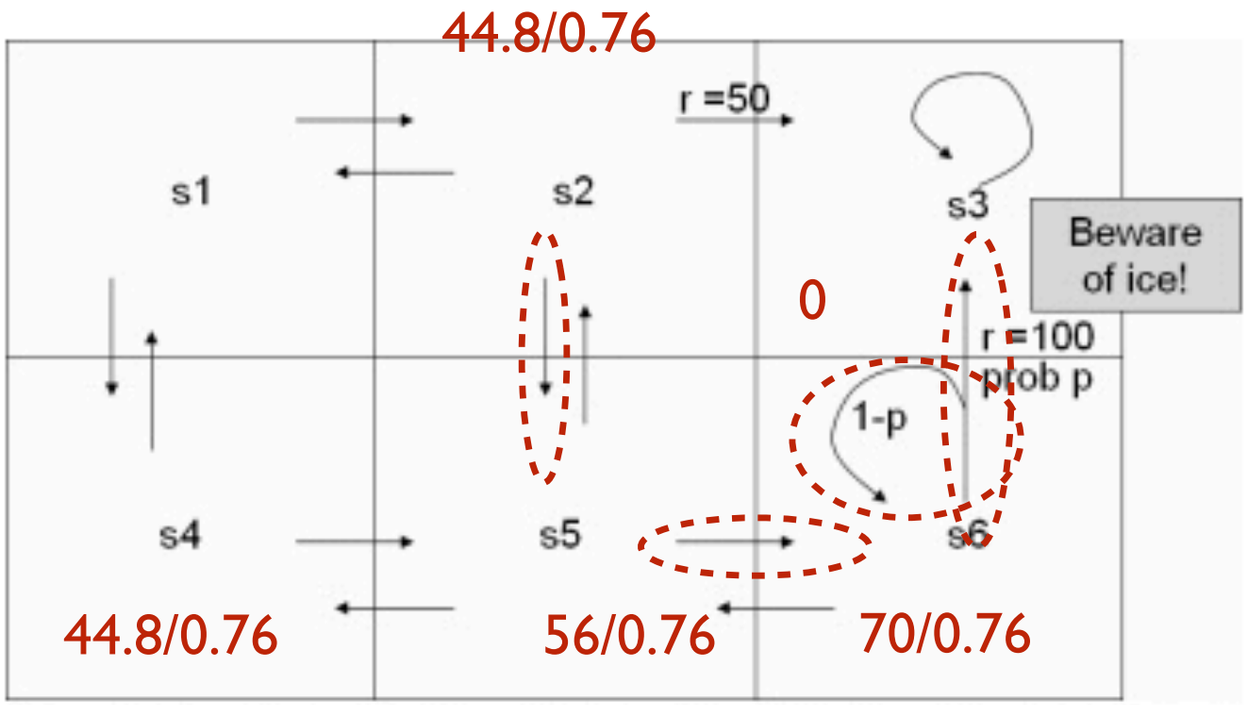
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



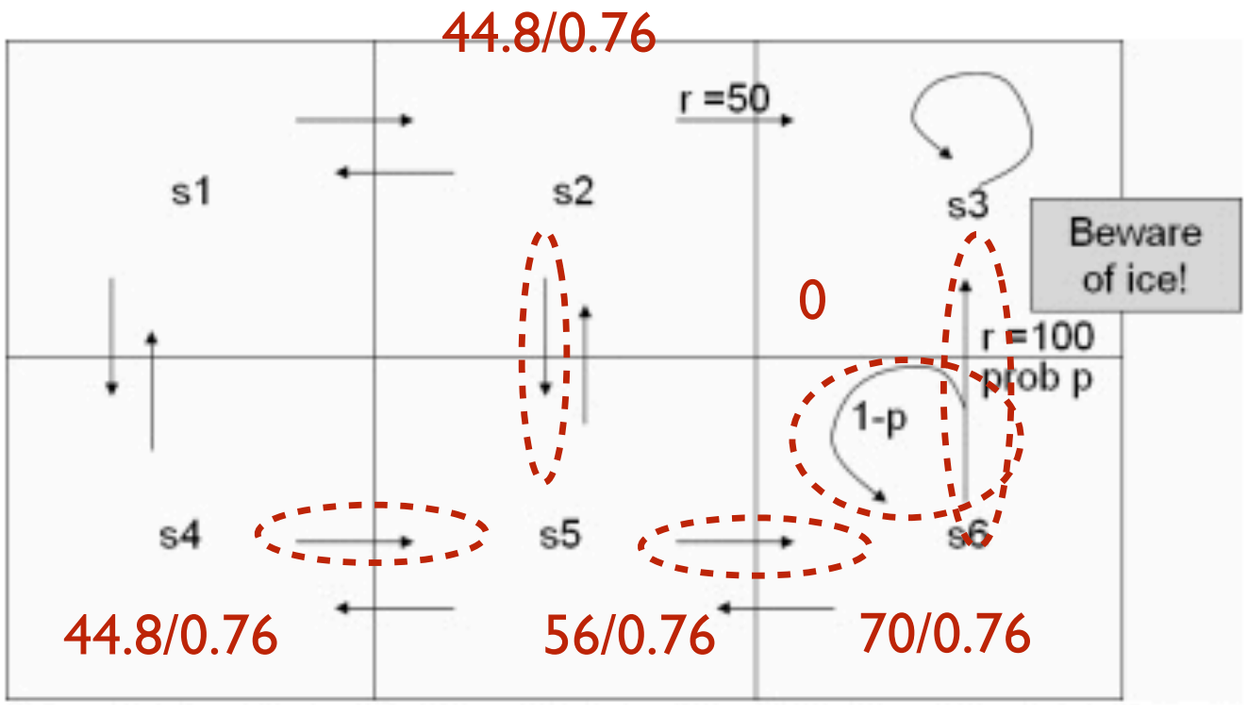
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



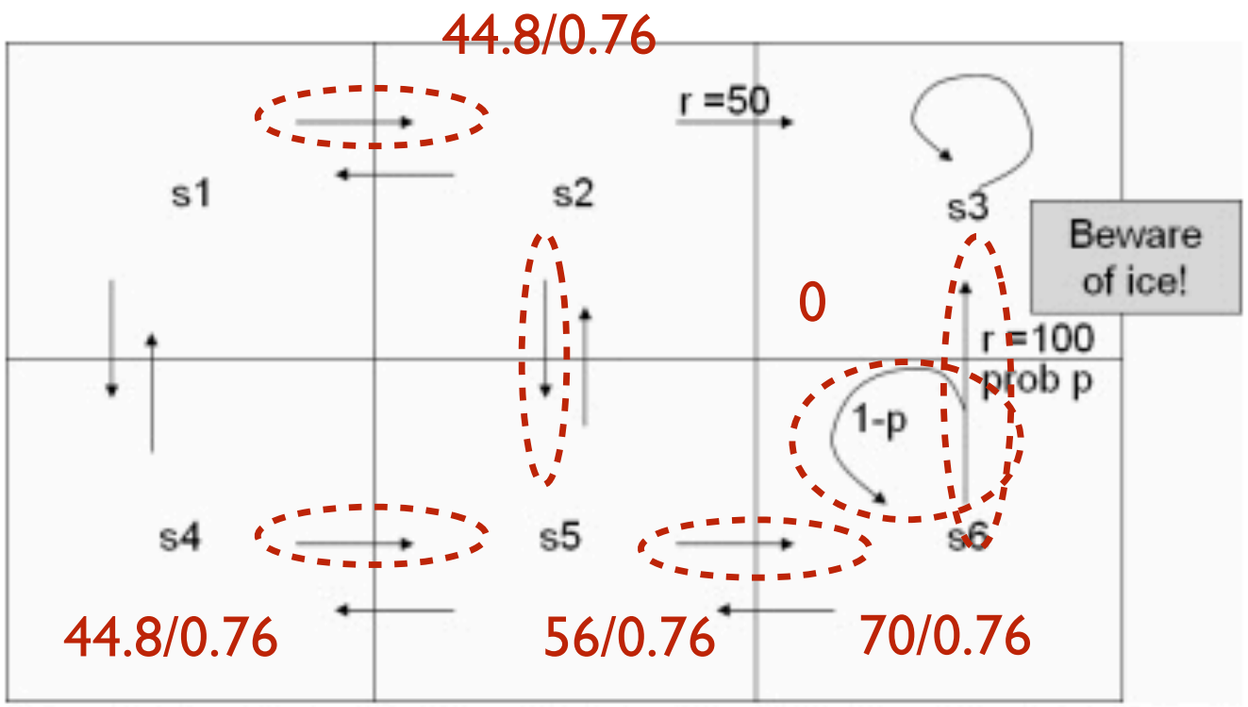
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



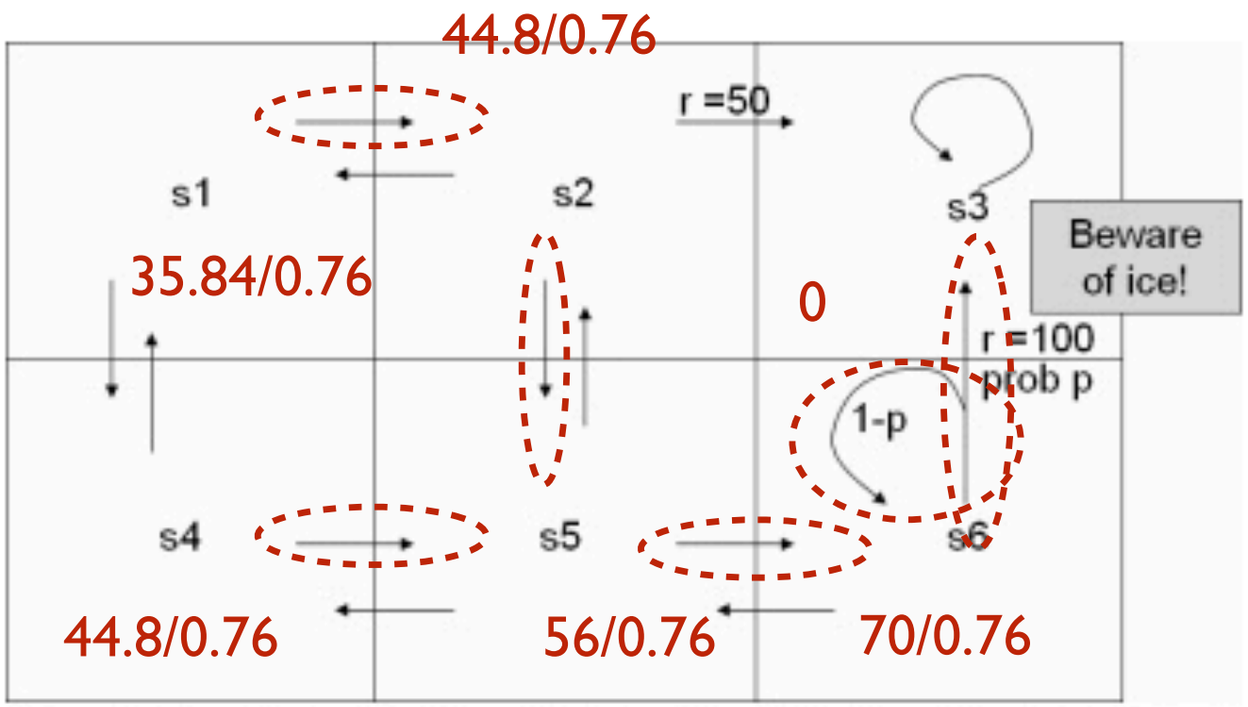
B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$

Part B. (6pts)

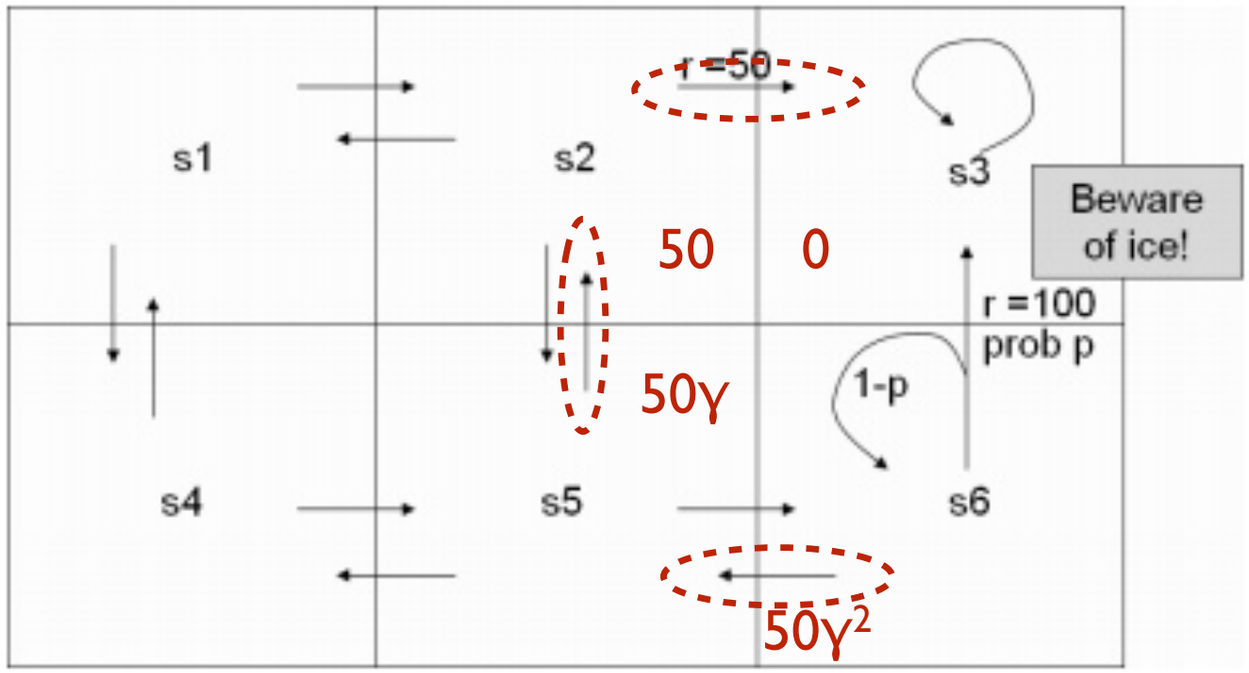
It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1-p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor $\gamma = 0.8$.**



B1. Assume p = 0.7. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V^*(s) = \sum_{s'} P(s'|\pi(s), s) r(s', a, s) + \gamma P(s'|\pi(s), s) V^*(s')]$$

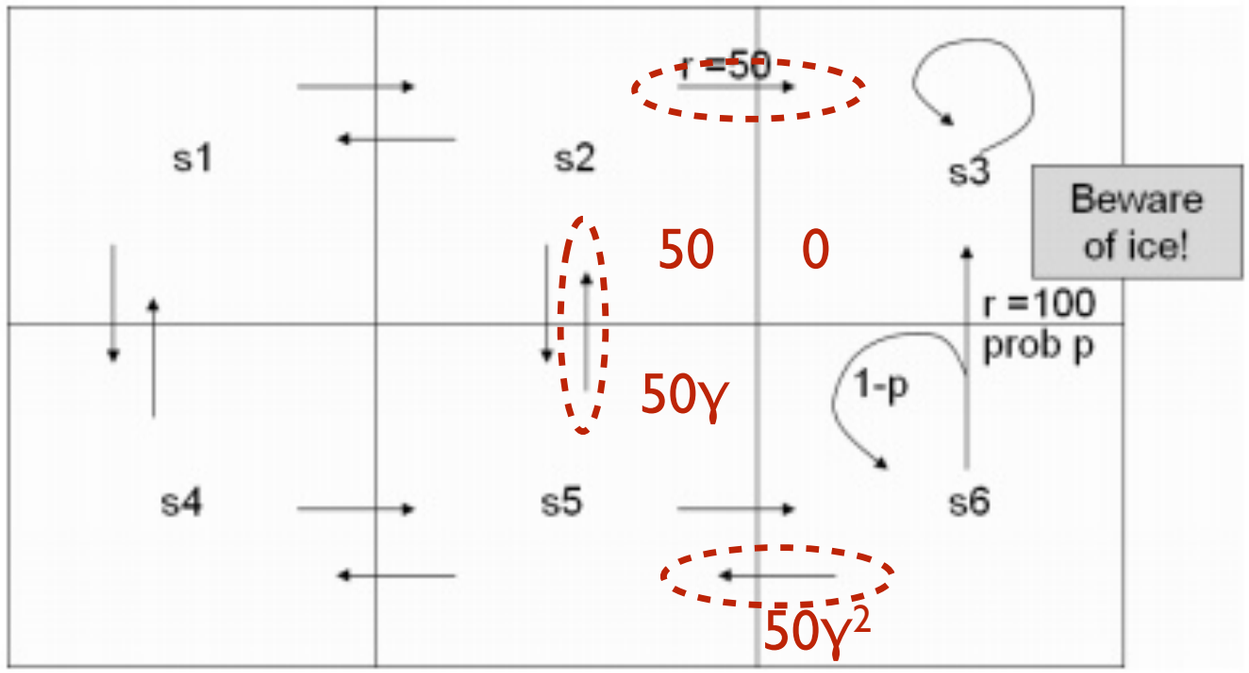
$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = 70 + 0.24V_6^*$$
$$V_6^* = \frac{70}{0.76}$$



B2. (2pts) How bad does the ice have to get before the robot will prefer to completely avoid it? Answer this question by giving a value for p below which the optimal policy chooses actions that completely avoid the ice, even choosing the action “go west" over “go north" when the robot is in state s6

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = \frac{100p}{1-\gamma(1-p)}$$

Also...

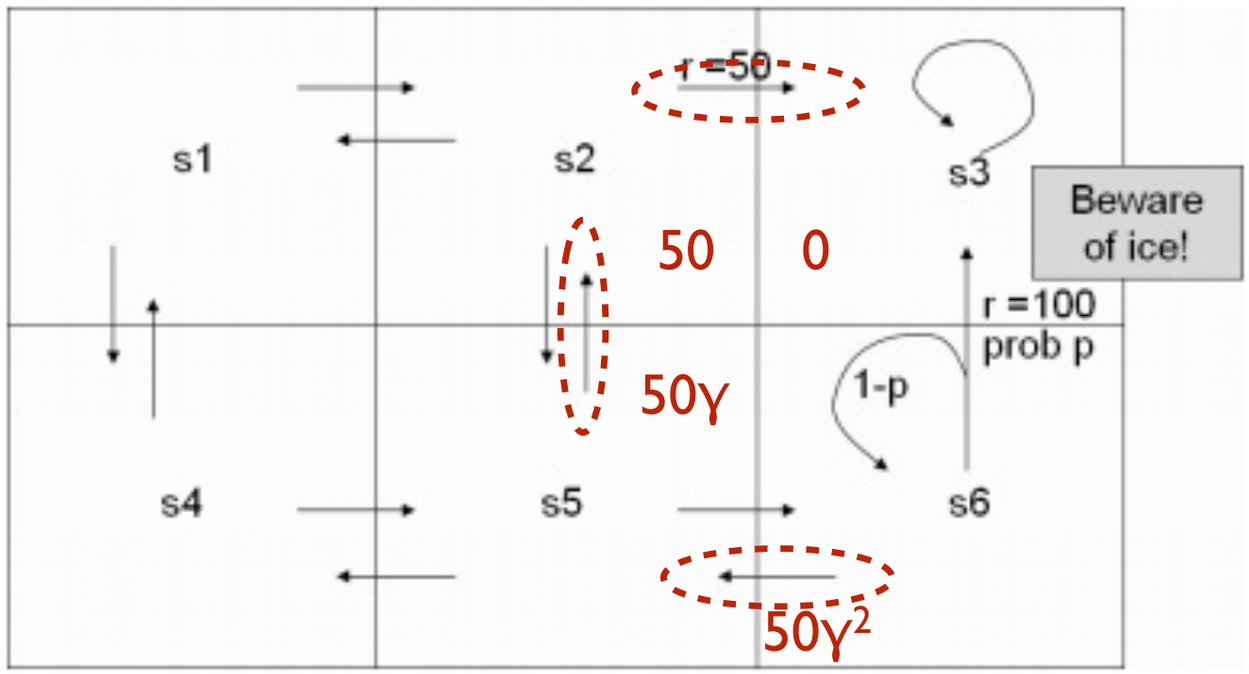


B2. (2pts) How bad does the ice have to get before the robot will prefer to completely avoid it? Answer this question by giving a value for p below which the optimal policy chooses actions that completely avoid the ice, even choosing the action “go west" over “go north" when the robot is in state s6

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = \frac{100p}{1-\gamma(1-p)}$$

Also...

$$50\gamma^2 = V_6^*$$



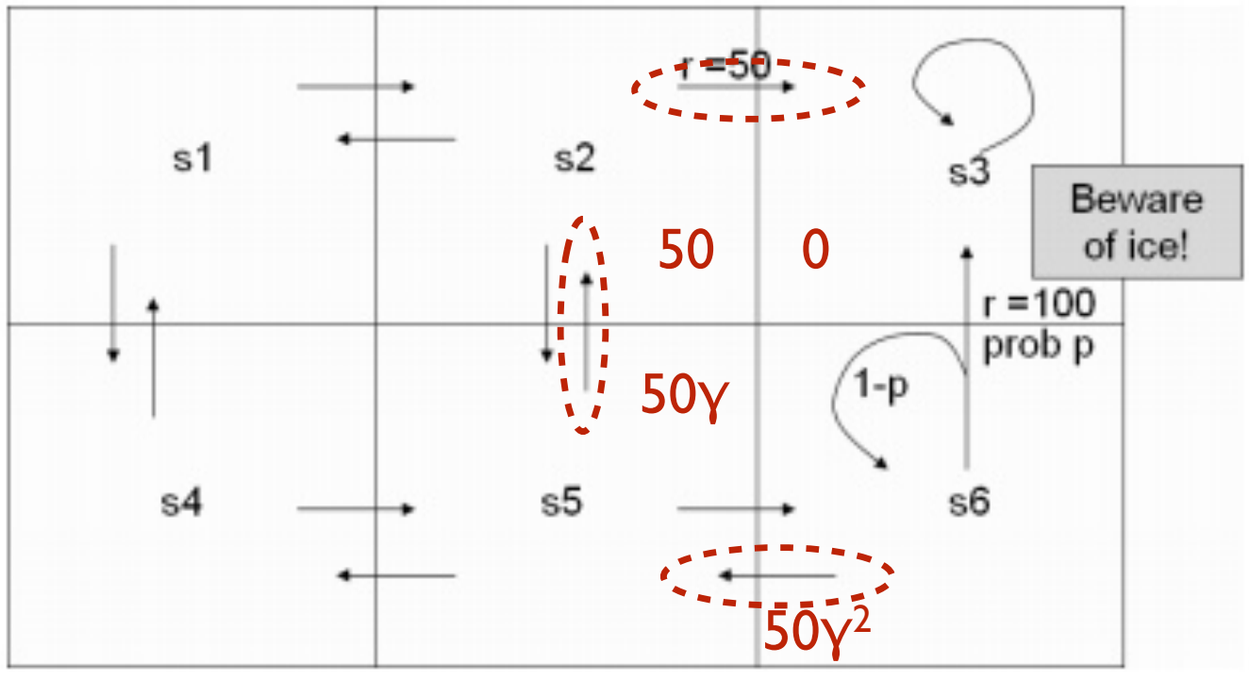
B2. (2pts) How bad does the ice have to get before the robot will prefer to completely avoid it? Answer this question by giving a value for p below which the optimal policy chooses actions that completely avoid the ice, even choosing the action “go west" over “go north" when the robot is in state s6

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = \frac{100p}{1-\gamma(1-p)}$$

Also...

$$50\gamma^2 = V_6^*$$

$$50\gamma^2 = \frac{100p}{1-\gamma(1-p)}$$



B2. (2pts) How bad does the ice have to get before the robot will prefer to completely avoid it? Answer this question by giving a value for p below which the optimal policy chooses actions that completely avoid the ice, even choosing the action “go west" over “go north" when the robot is in state s6

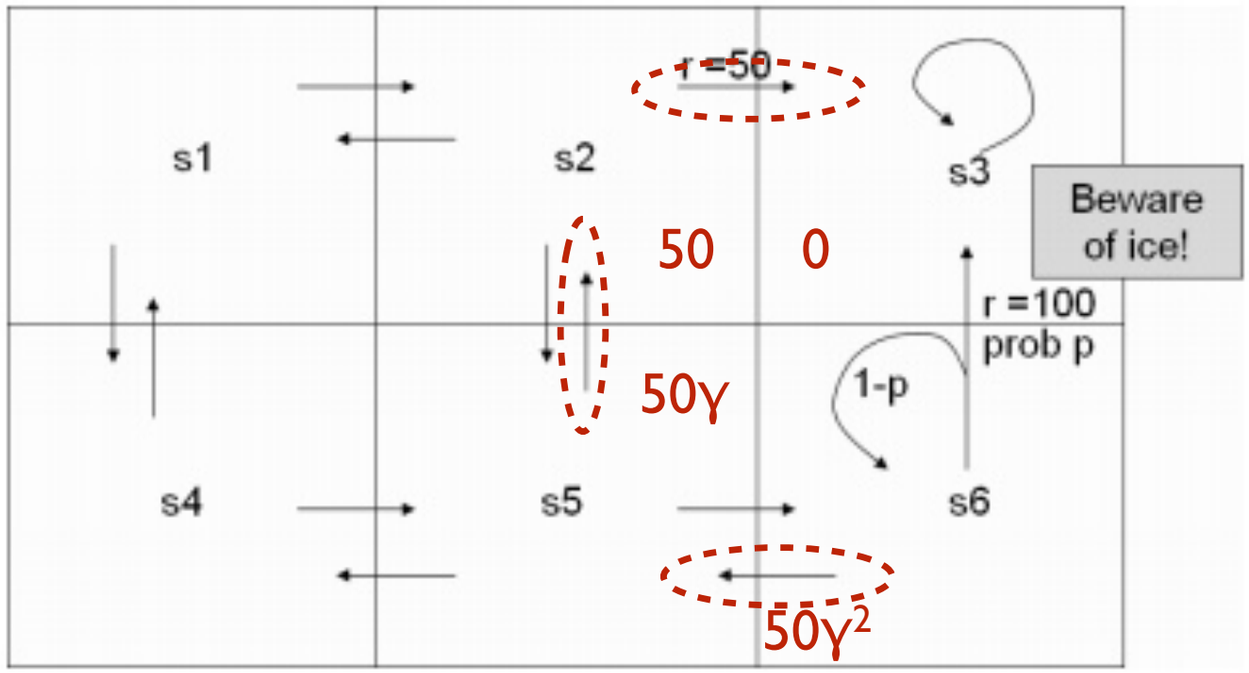
$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = \frac{100p}{1-\gamma(1-p)}$$

Also...

$$50\gamma^2 = V_6^*$$

$\gamma = 0.8$

$$50\gamma^2 = \frac{100p}{1-\gamma(1-p)}$$



B2. (2pts) How bad does the ice have to get before the robot will prefer to completely avoid it? Answer this question by giving a value for p below which the optimal policy chooses actions that completely avoid the ice, even choosing the action “go west" over “go north" when the robot is in state s6

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = \frac{100p}{1-\gamma(1-p)}$$

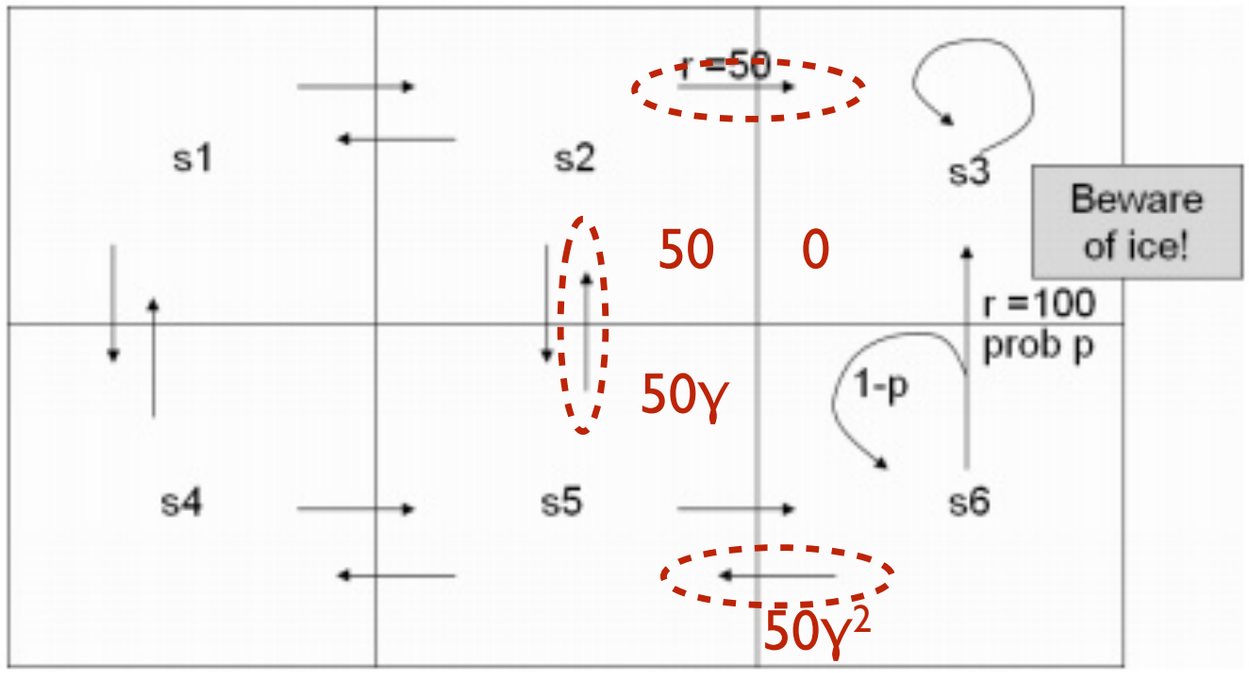
Also...

$$50\gamma^2 = V_6^*$$

$\gamma = 0.8$

$$50\gamma^2 = \frac{100p}{1-\gamma(1-p)}$$

Solve the eq.



B2. (2pts) How bad does the ice have to get before the robot will prefer to completely avoid it? Answer this question by giving a value for p below which the optimal policy chooses actions that completely avoid the ice, even choosing the action “go west" over “go north" when the robot is in state s6

$$V_6^* = 100p + \gamma(1-p)V_6^*$$
$$V_6^* = \frac{100p}{1-\gamma(1-p)}$$

Also... $50\gamma^2 = V_6^*$

$\gamma = 0.8$

$$50\gamma^2 = \frac{100p}{1-\gamma(1-p)}$$

Solve the eq.

p = 8/93

Questions?

References

- A good reading is : “Reinforcement Learning: A Survey (1996)” Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore. Journal of Artificial Intelligence Research, 4, 237-285. Here is a link where you can download the paper: http://www.autonlab.org/autonweb/14686/version/3/part/5/data/kaelbling-reinforcement.pdf?branch=main&language=en

- Marc Toussaint lecture notes (retrieved from http://userpage.fu-berlin.de/mtoussai/notes/markov-decision-processes.pdf)