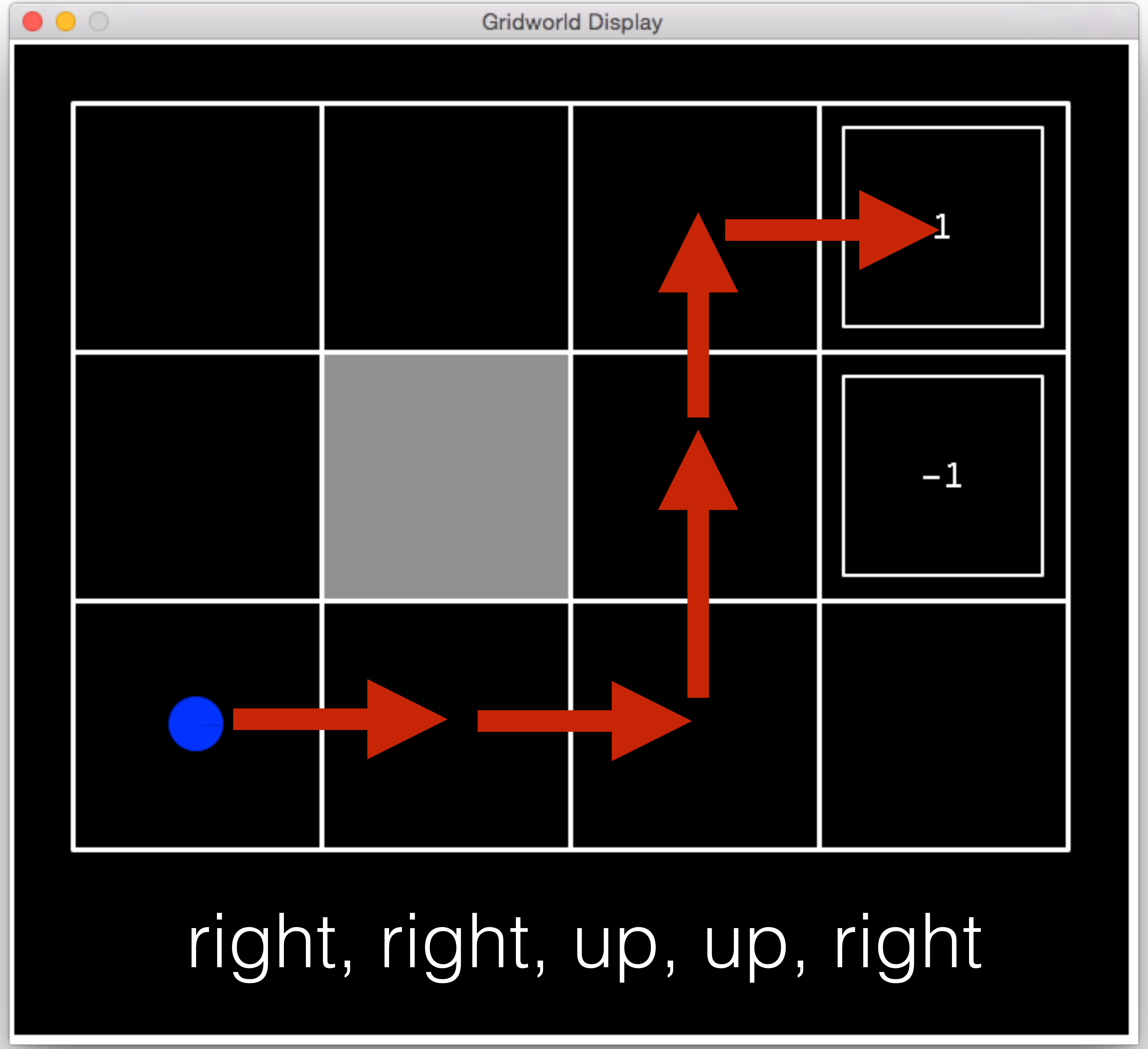
Markov Decision Processes

CS4804

Outline

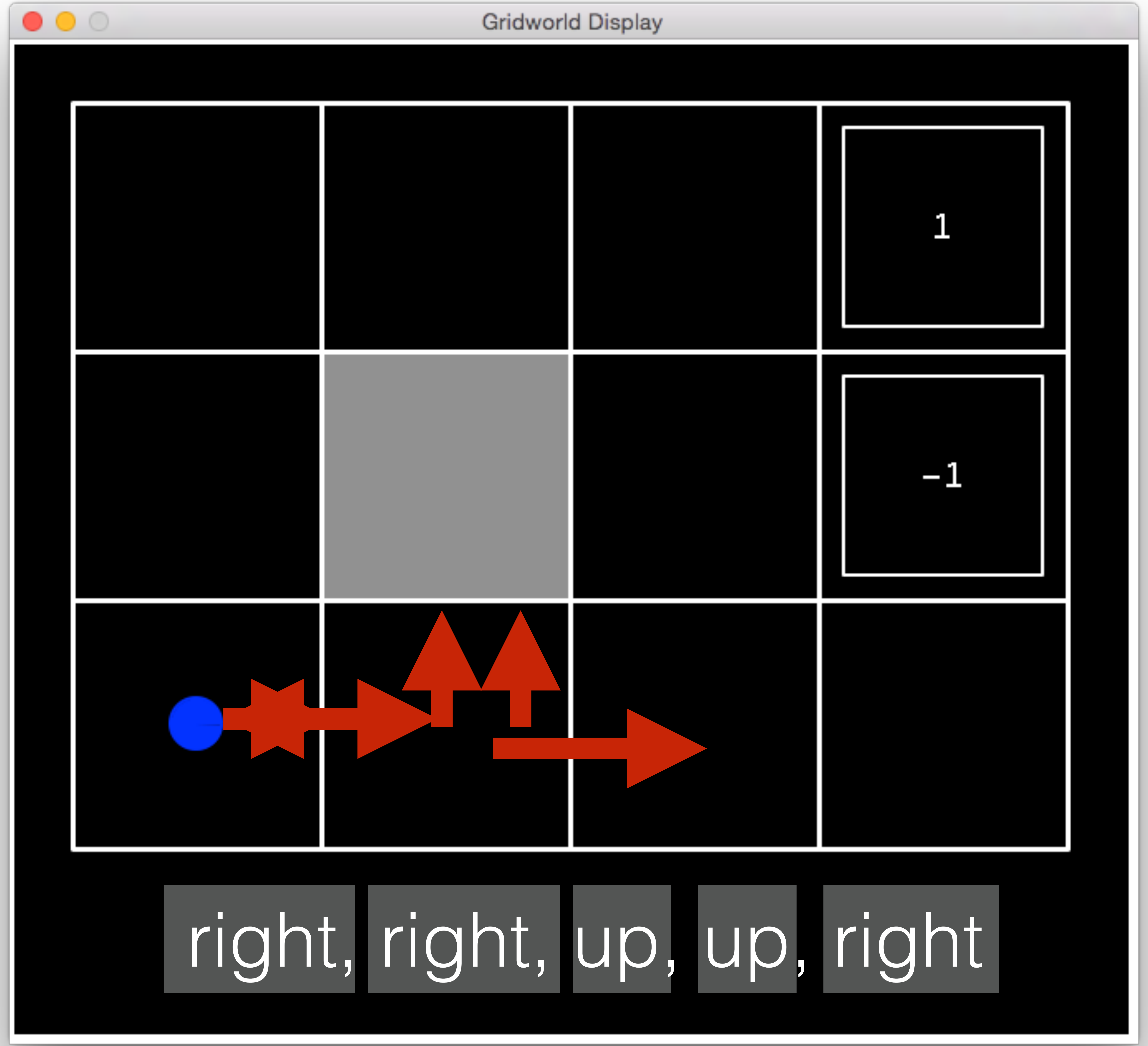
- Markov decision process: richer environment representation
- Reward functions
- Optimizing policies via value iteration

collect reward



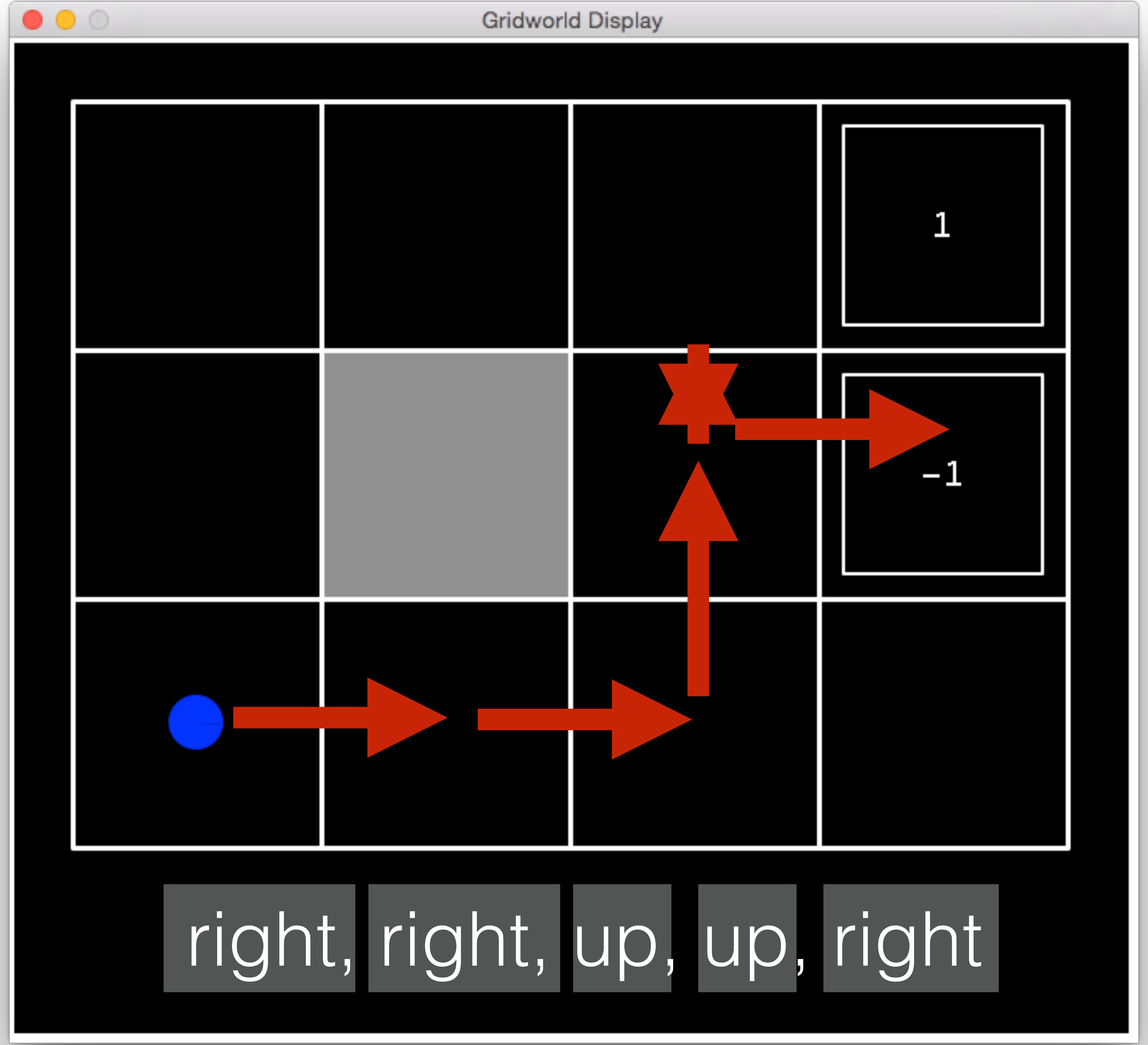
collect reward

stochastic transitions



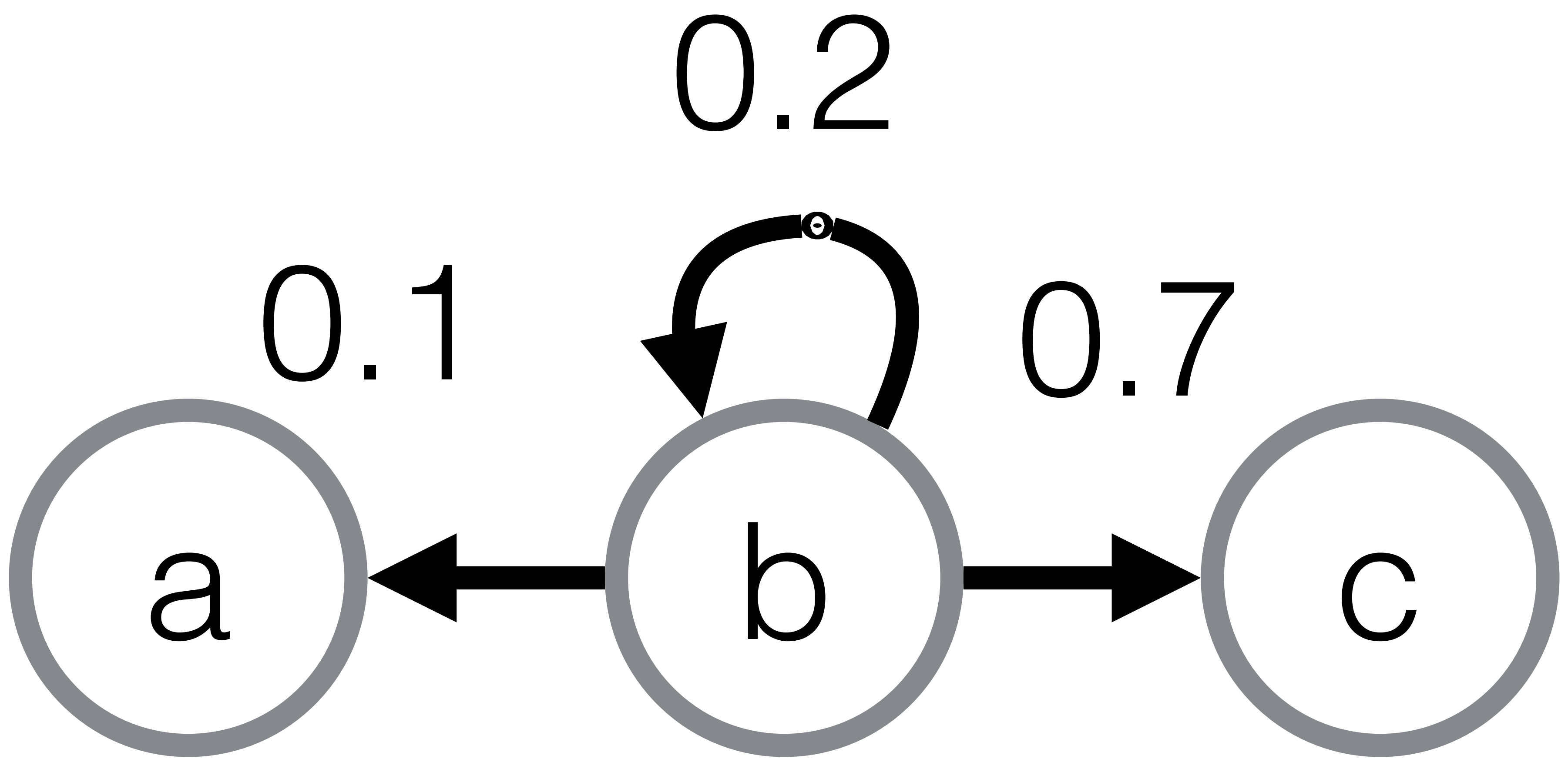
collect reward

stochastic transitions



Actions and Transitions

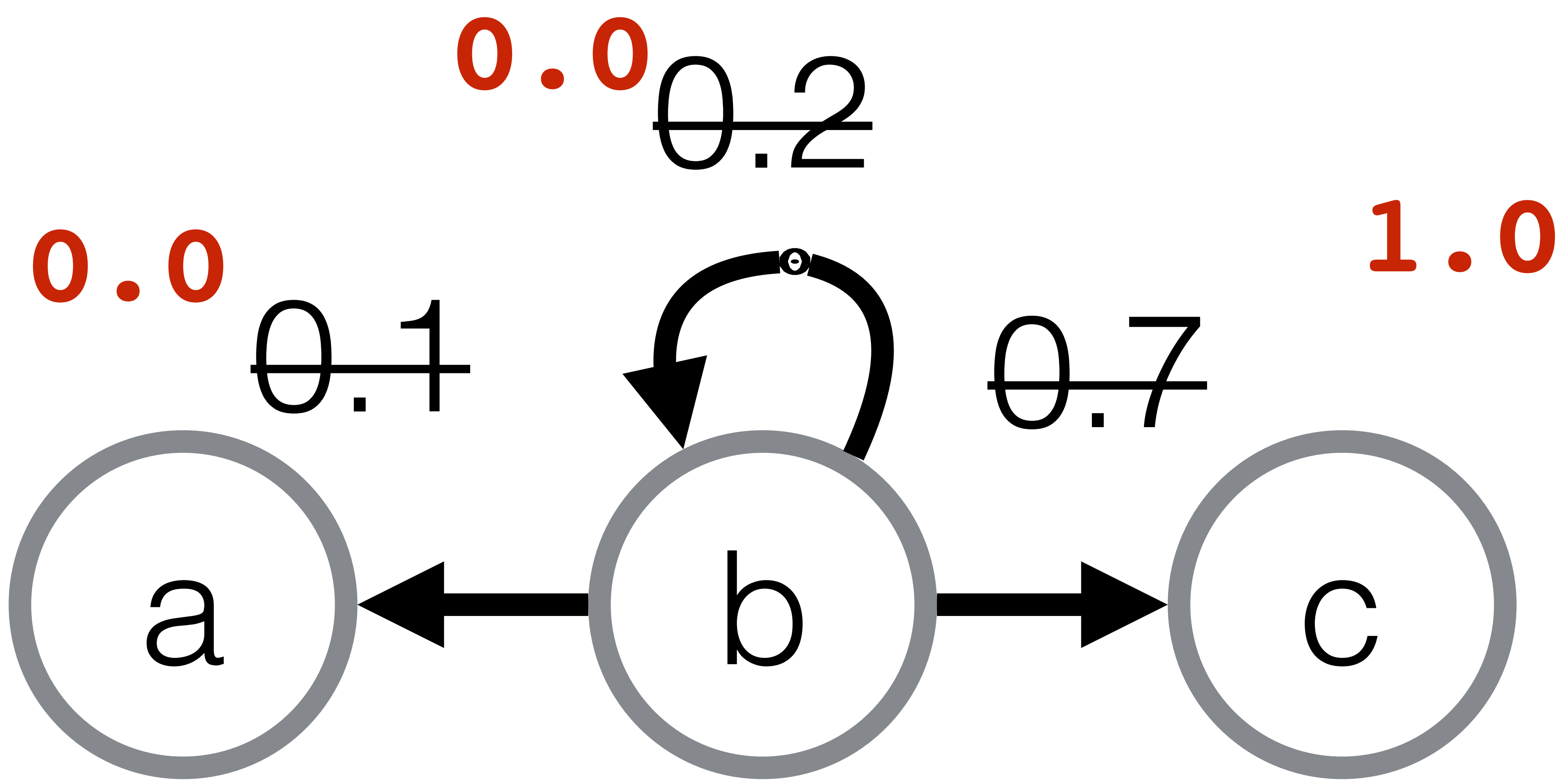
- Pr(**s'** | **s**, **a**)
  - Probability we **transition** to **s'** if we choose **action** **a** in state **s**



a = right

Actions and Transitions

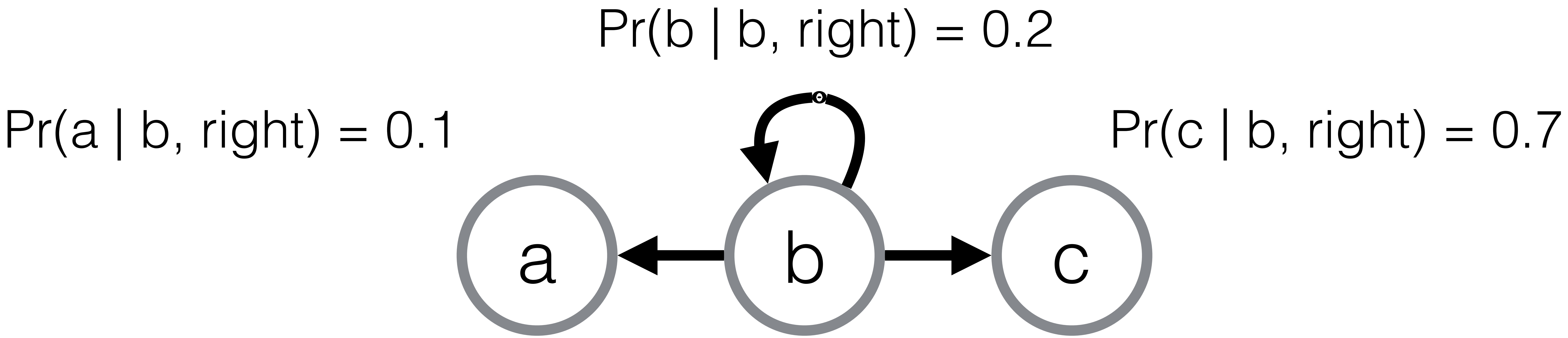
- Pr(**s'** | **s**, **a**)
  - Probability we **transition** to **s'** if we choose **action** **a** in state **s**



a = right

Actions and Transitions

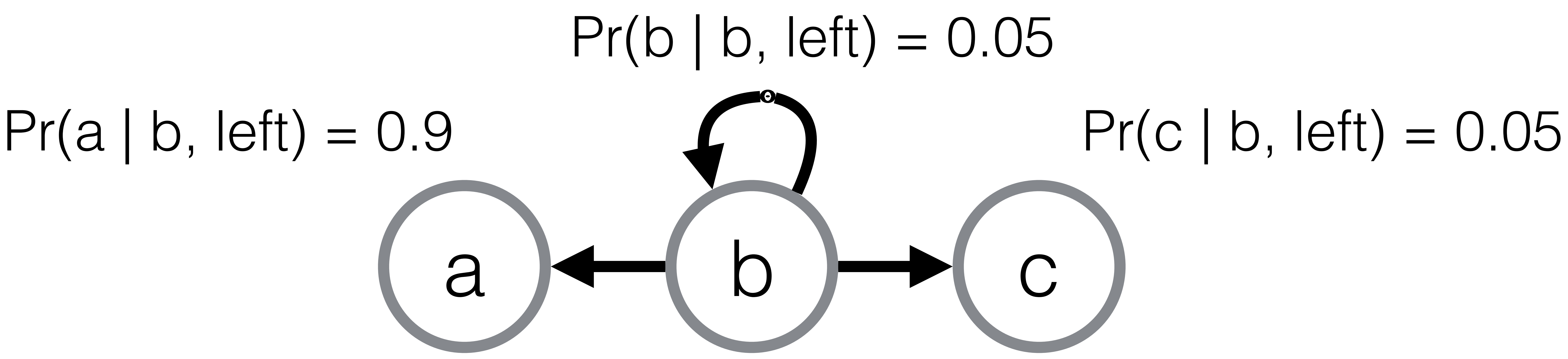
- Pr(**s'** | **s**, **a**)
  - Probability we **transition** to **s'** if we choose **action** **a** in state **s**



a = right

Actions and Transitions

- Pr(**s’** | **s**, **a**)
  - Probability we **transition** to **s’** if we choose **action** **a** in state **s**



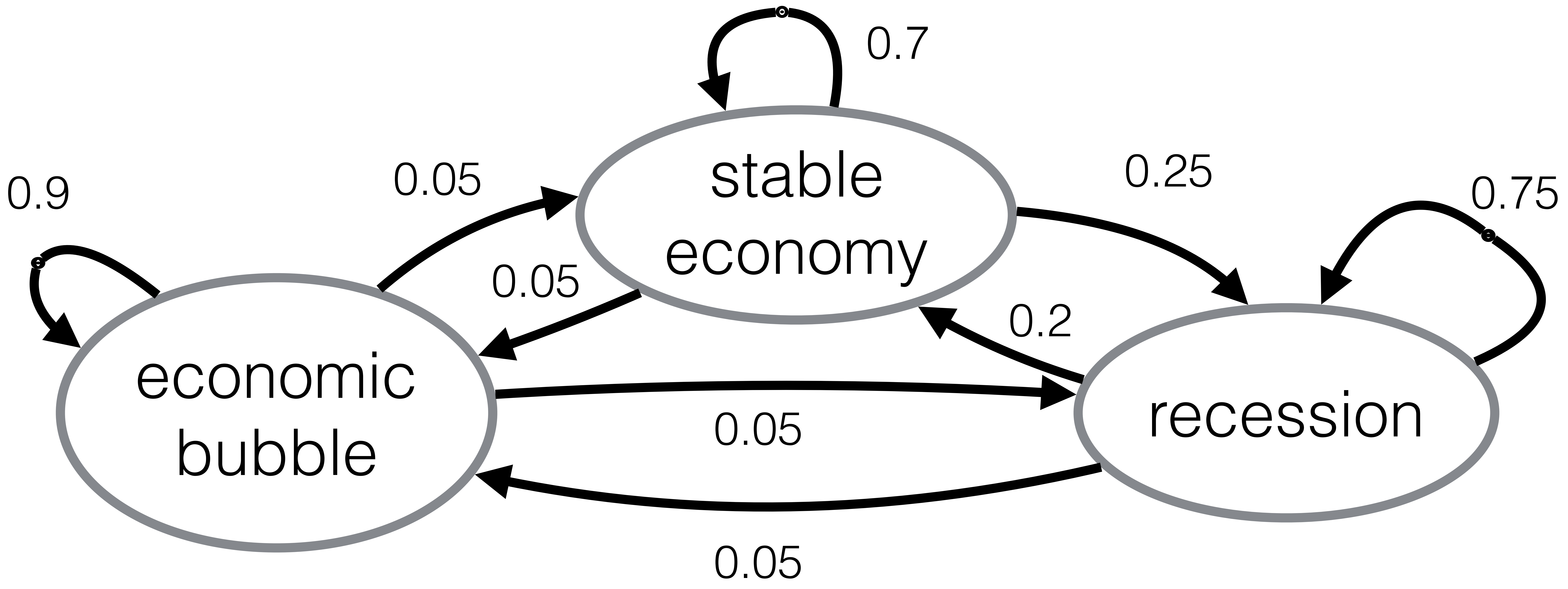
a = left

Preview: Markov Models

Markov Decision Process: Pr(**s’** | **s**, **a**)

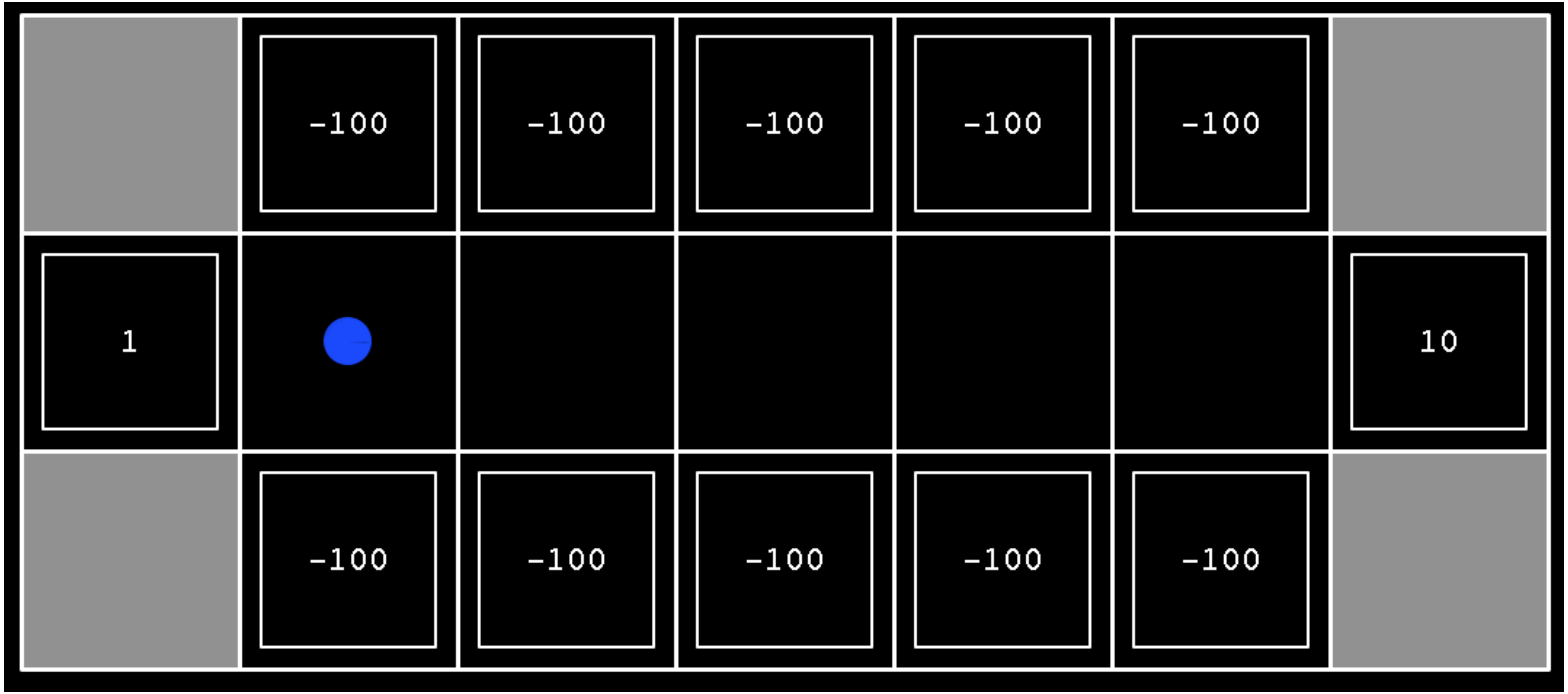
Markov Process Pr(**s’** | **s**)

Preview: Markov Models

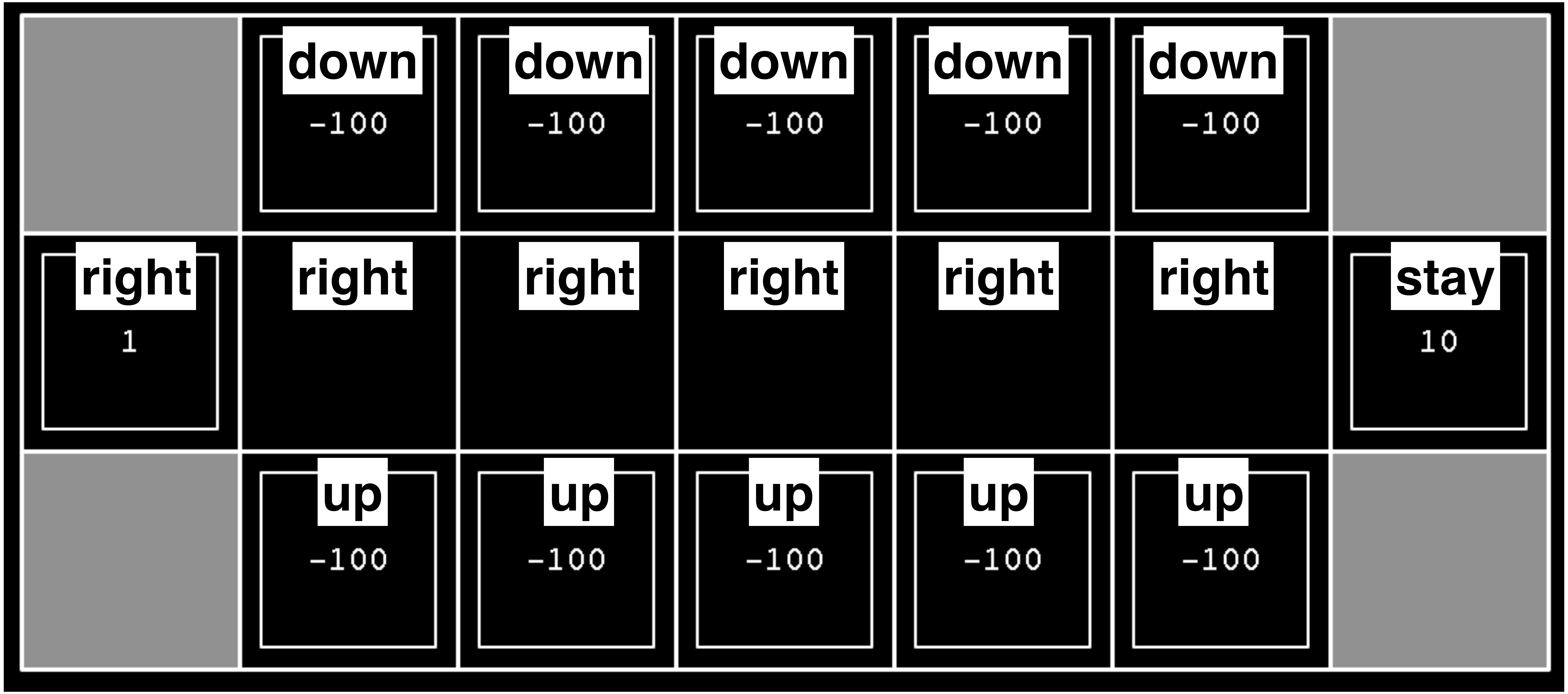


Markov Process Pr(**s'** | **s**)

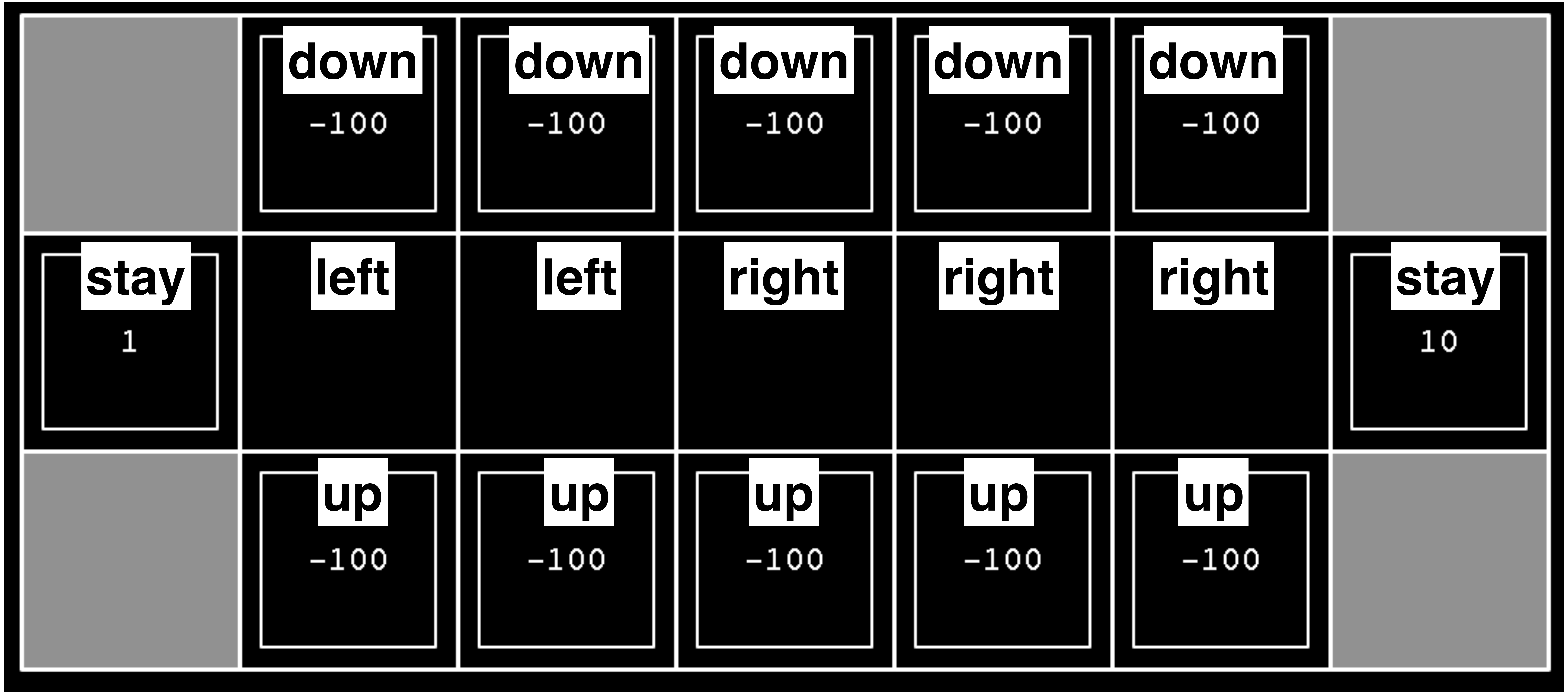
Reward function R(s)



Policy π(s)


down
-100
down
-100
down
-100
down
-100
down
-100
right
1
right
right
right
right
right
stay
10
up
-100
up
-100
up
-100
up
-100
up
-100

Policy $\pi(s)$


down
down
down
down
down
-100
-100
-100
-100
-100
stay
left
left
right
right
right
stay
1
10
up
up
up
up
up
-100
-100
-100
-100
-100

How Good is a Policy?

$$U([s_0, s_1, \ldots, s_T]) = \sum_{t=0}^{T} R(s_t)$$

$$U([s_0, s_1, \ldots, s_T]) = \sum_{t=0}^{T} \gamma^t R(s_t) \qquad \gamma \in (0, 1]$$

How Good is a Policy?

$$U([s_0, s_1, \ldots, s_T]) = \sum_{t=0}^{\cancel{T}\,\infty} \gamma^t R(s_t) \qquad \gamma \in (0, 1]$$

$$U^\pi(s) = \mathrm{E}_{\Pr([s_0, s_1, \ldots] | s_0 = s, \pi)} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t) \right]$$

$$\pi_s^* = \arg\max_\pi U^\pi(s)$$

$$U([s_0, s_1, \ldots, s_T]) = \sum_{t=0}^{\cancel{T}\ \infty} \gamma^t R(s_t) \qquad \gamma \in (0, 1]$$

$$U^\pi(s) = \mathrm{E}_{\Pr([s_0, s_1, \ldots] | s_0 = s, \pi)} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t) \right]$$

$$\pi_s^* = \arg\max_\pi U^\pi(s) = \pi_{s'}^*, \text{ for any } s'$$

$$U(s) = U^{\pi^*}(s)$$

$$\pi^*(s) = \arg\max_{a \in A(s)} \sum_{s'} P(s' | s, a) U(s')$$

$$\pi^*(s) = \arg\max_{a \in A(s)} \sum_{s'} P(s'|s,a)U(s')$$

U(*s*') = expected utility given optimal play from *s*'

$$U(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a)U(s')$$

Bellman equation

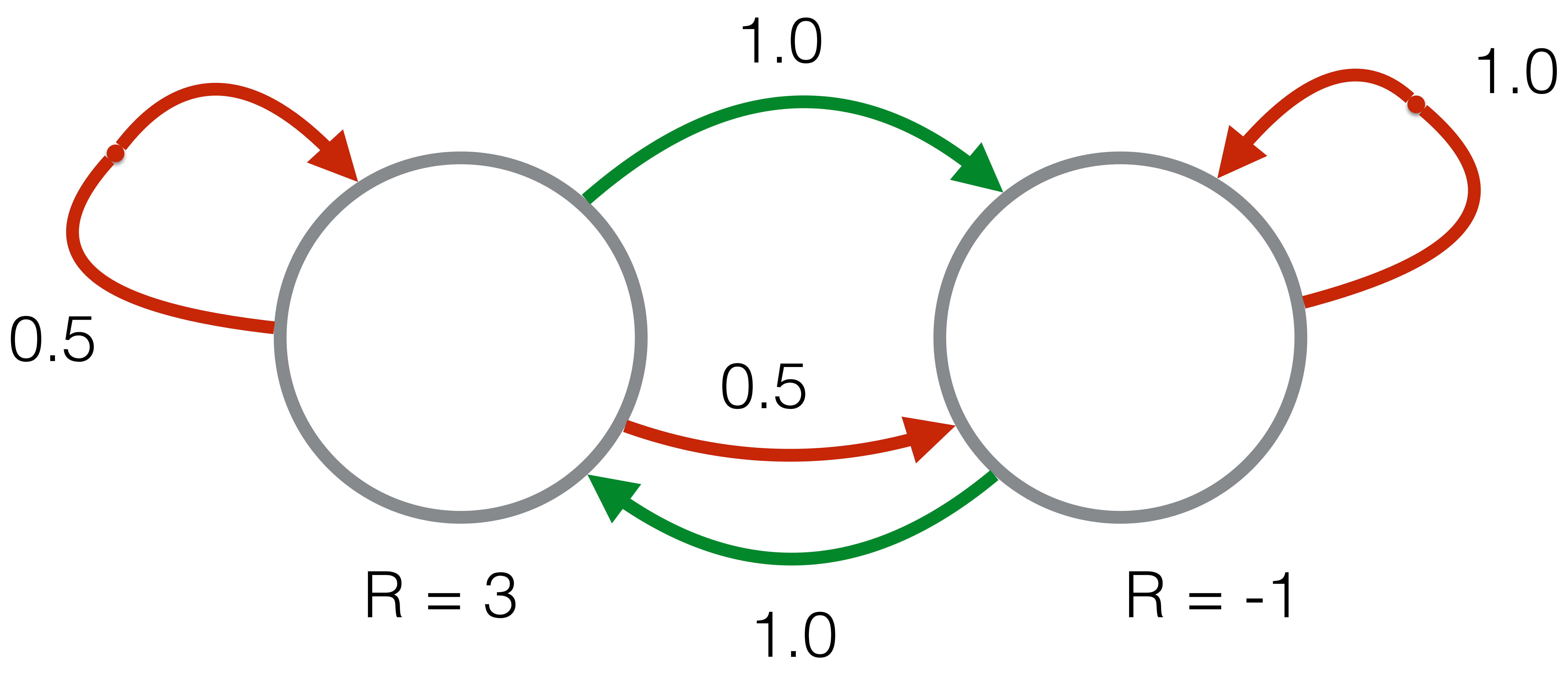
Value Iteration

$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$U(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U(s')$$

Bellman equation

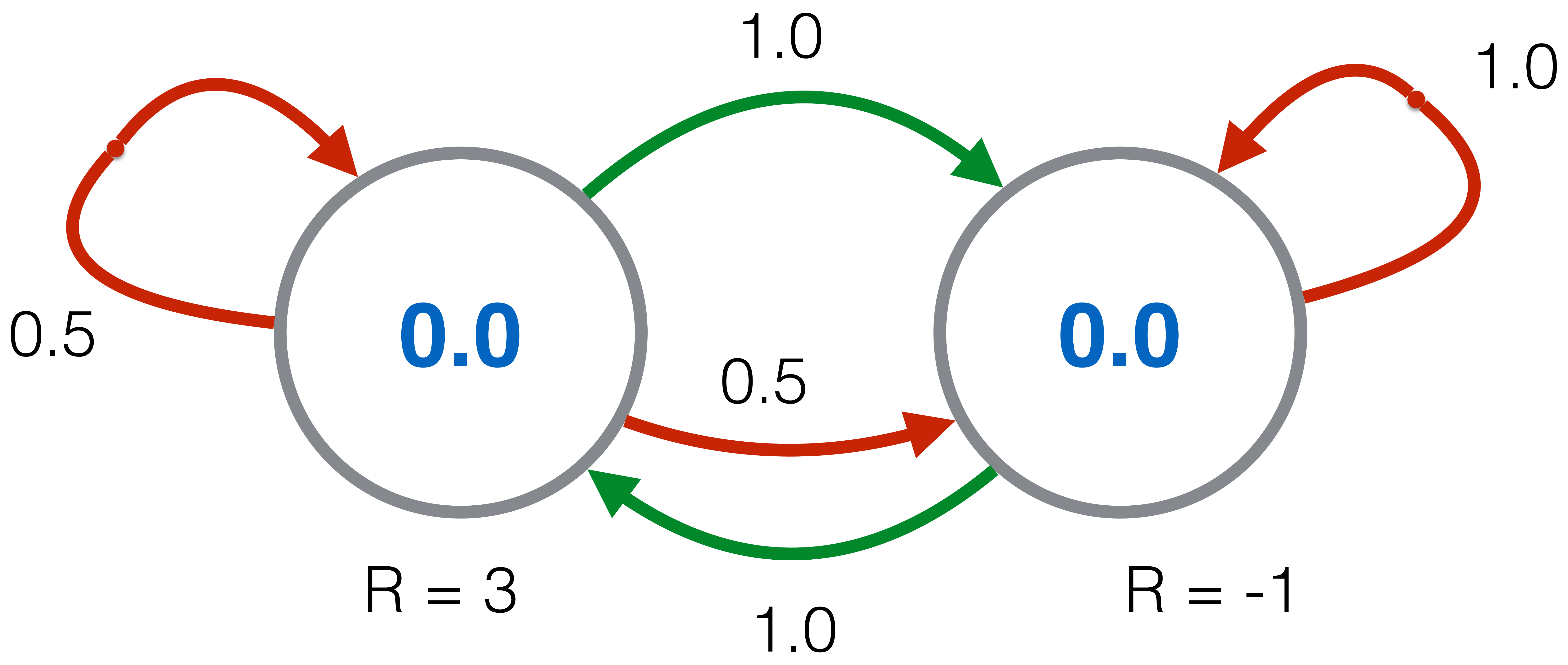
Value Iteration Example



Value Iteration Example

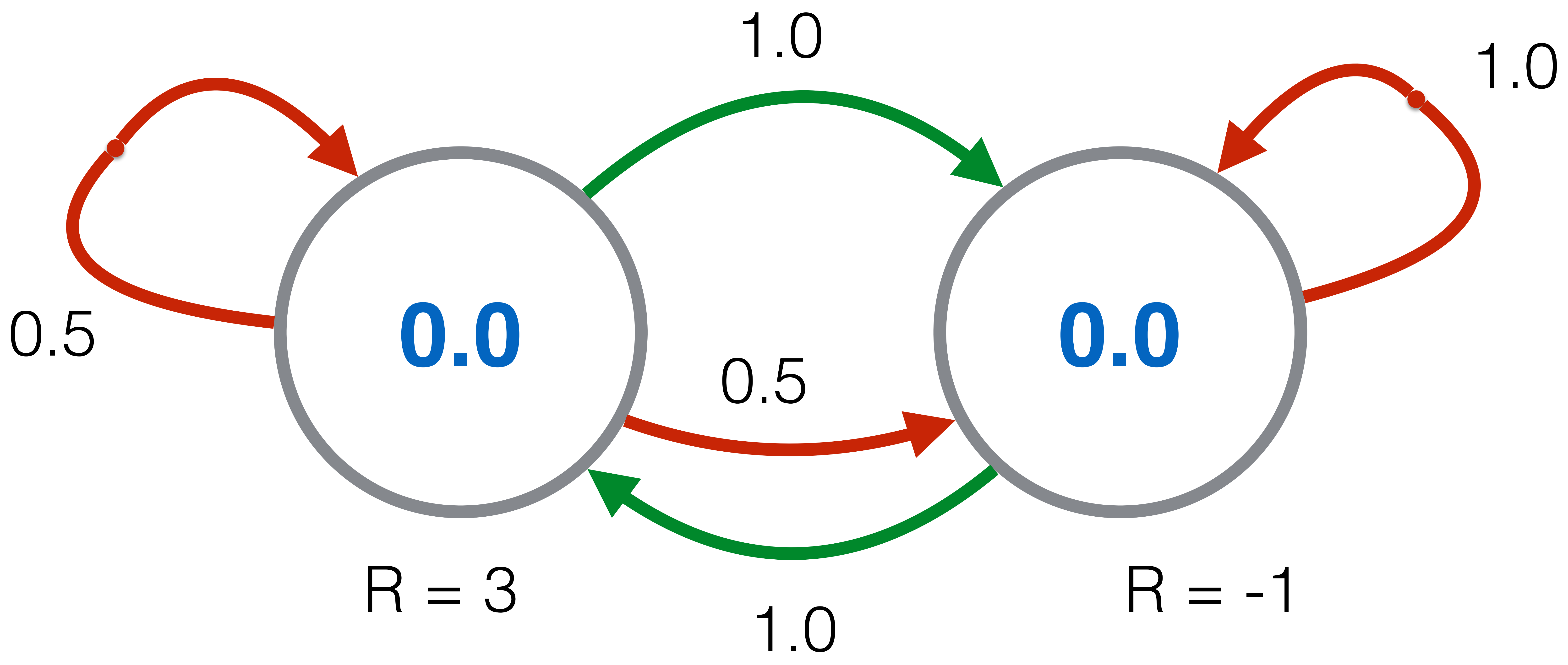
$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$



$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$

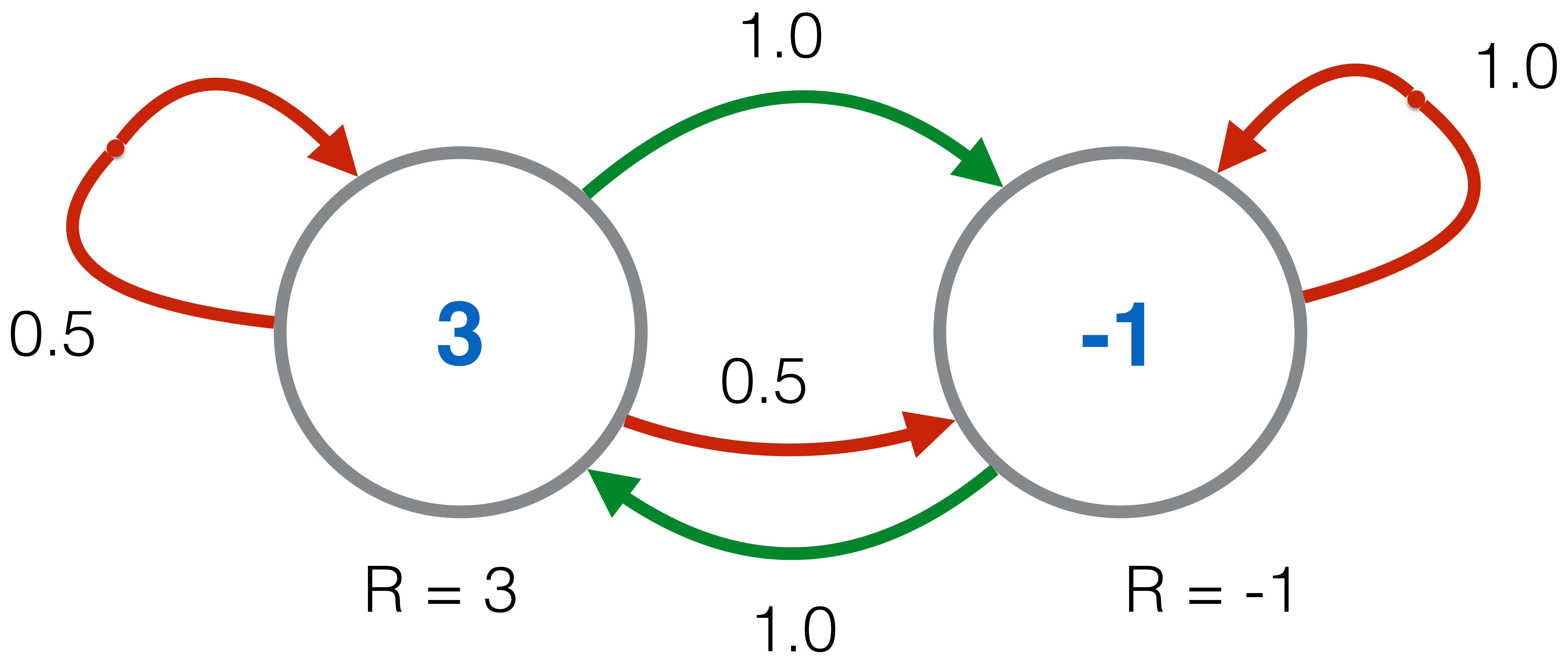


3 + 0.5 max{ 1.0 * 0.0, 0.5 * 0.0 + 0.5 * 0.0 } = 3

-1 + 0.5 max{ 1.0 * 0.0, 1.0 * 0.0 } = -1

$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$

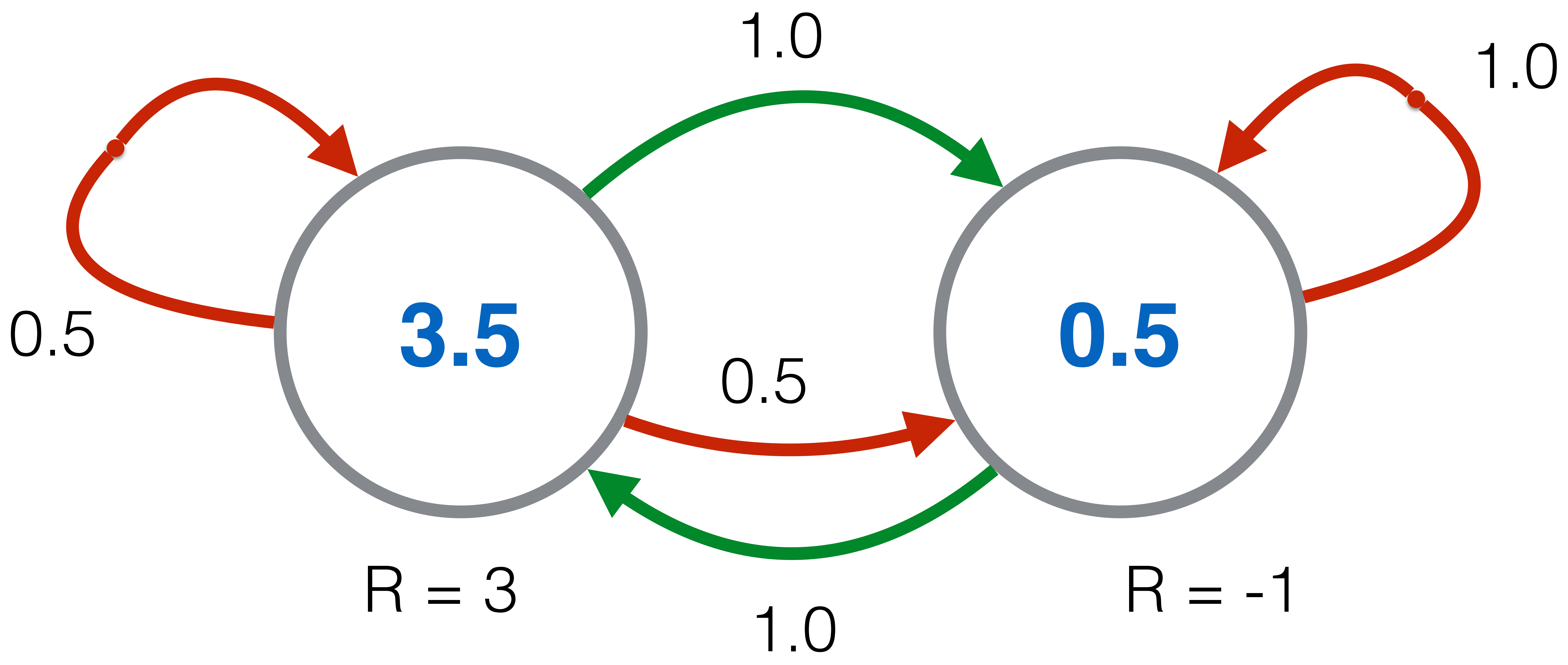


3 + 0.5 max{ 1.0 * (-1), 0.5 * 3 + 0.5 * (-1) } = 3 + 0.5 max{ -1, 1} = 3.5

-1 + 0.5 max{ 1.0 * 3, 1.0 * (-1) } = 0.5

$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$

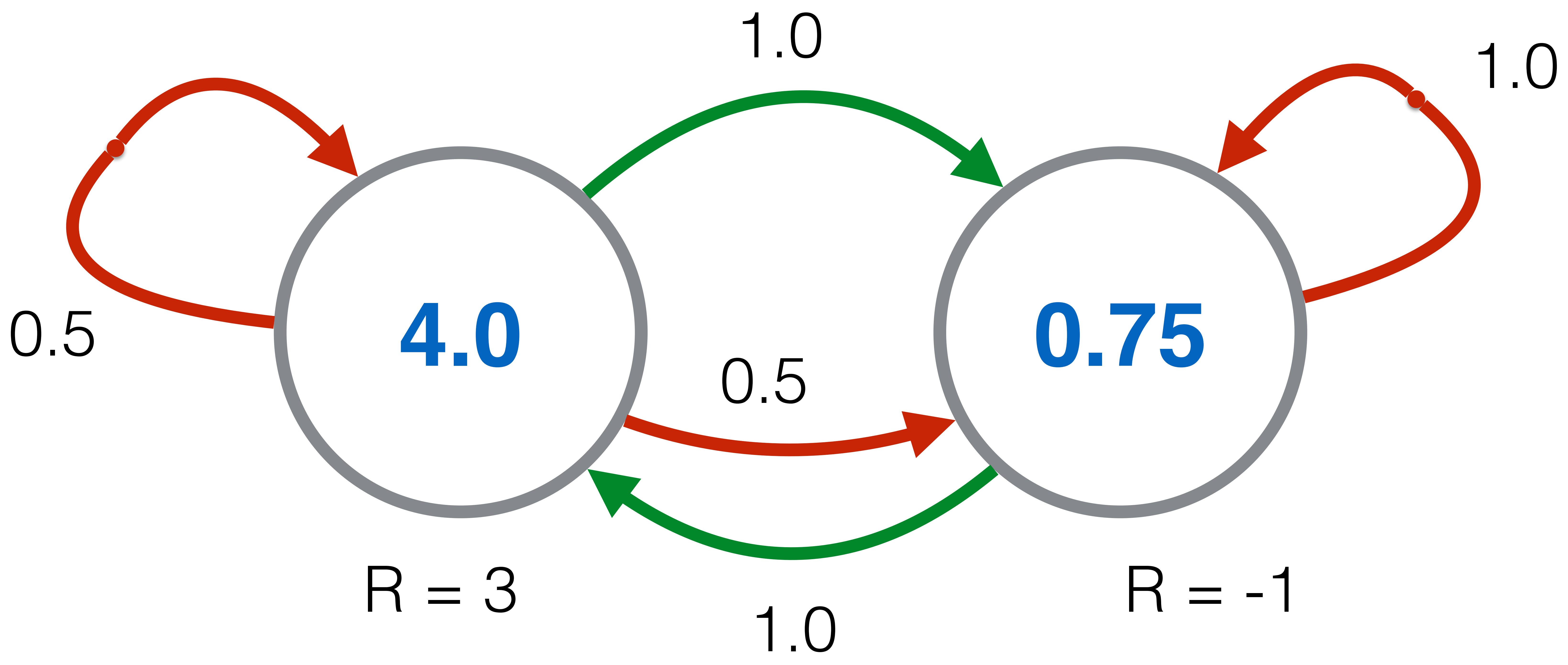


3 + 0.5 max{ 1.0 * 0.5, 0.5 * 3.5 + 0.5 * 0.5 } = 3 + 0.5 max{ 0.5, 2} = 4

-1 + 0.5 max{ 1.0 * 3.5, 1.0 * 0.5 } = 0.75

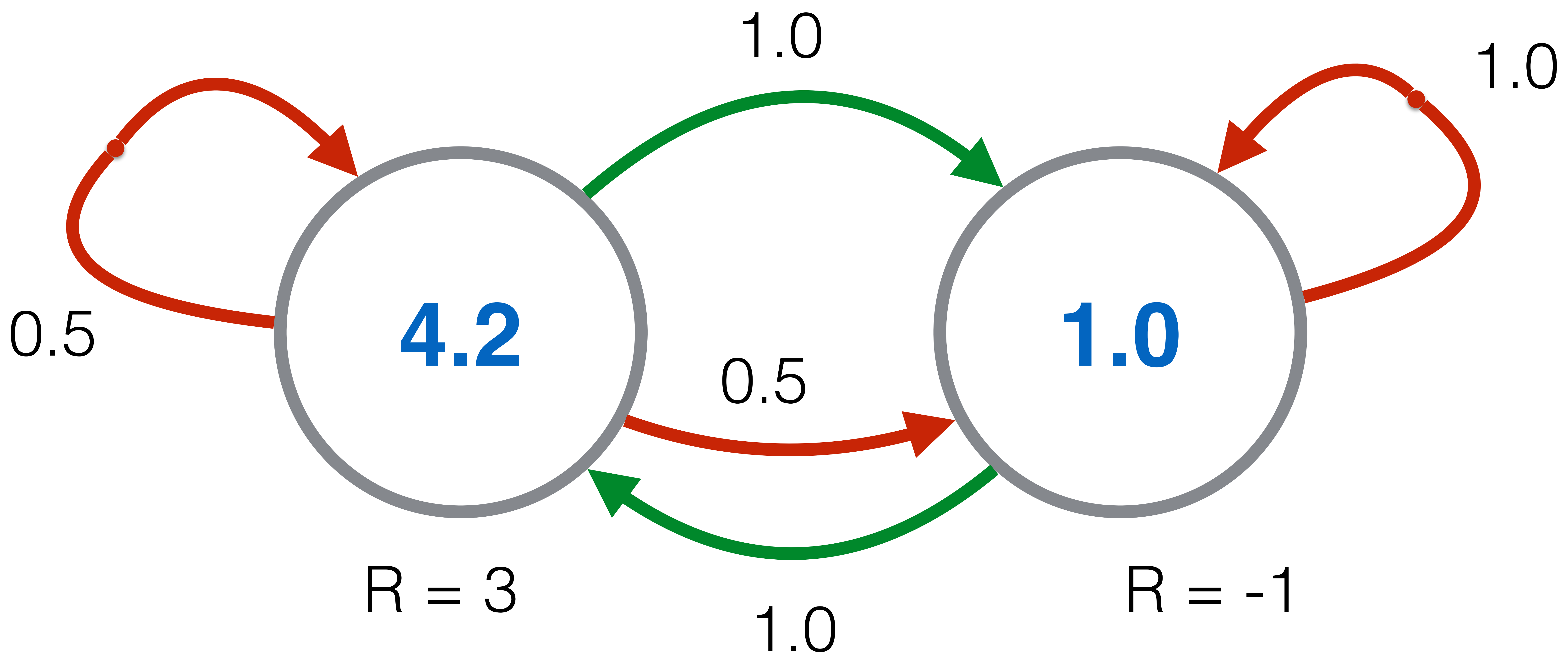
$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$



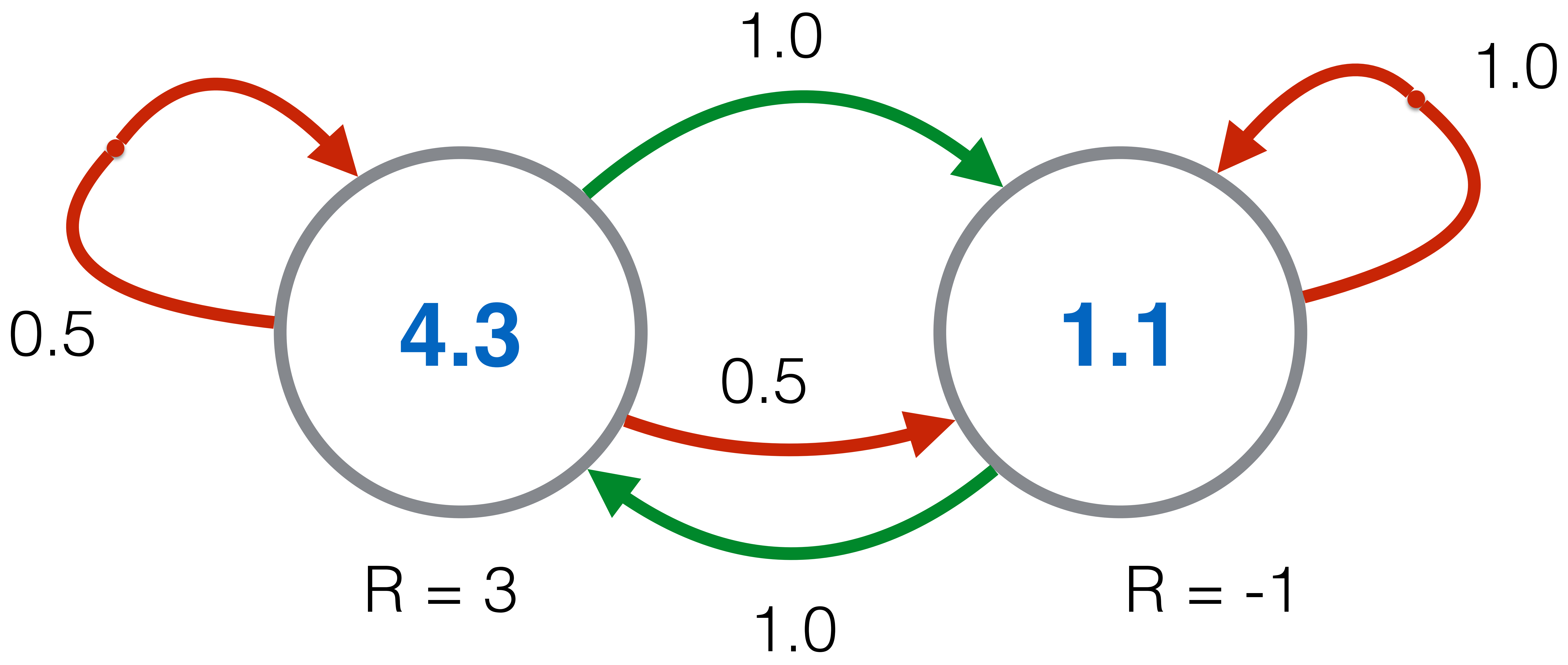
$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$



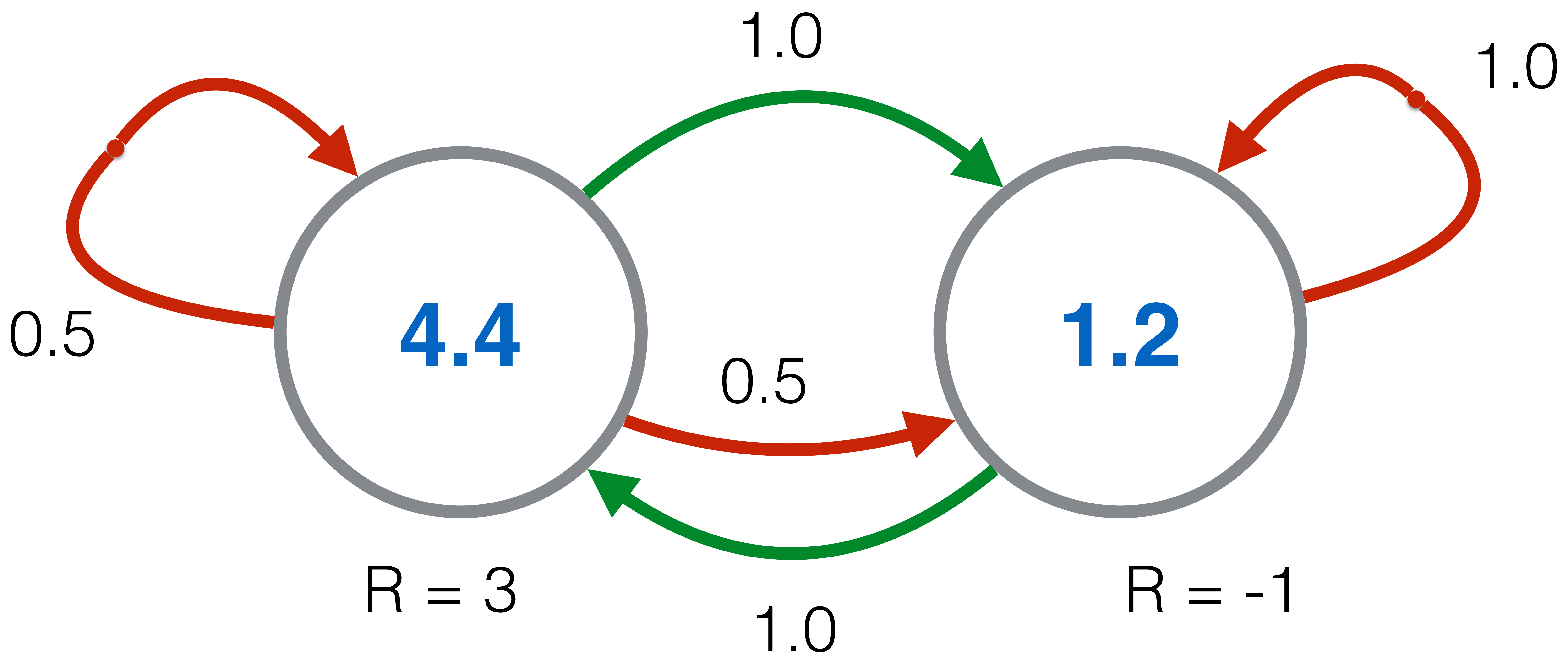
$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$



$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

$$\gamma = 0.5$$

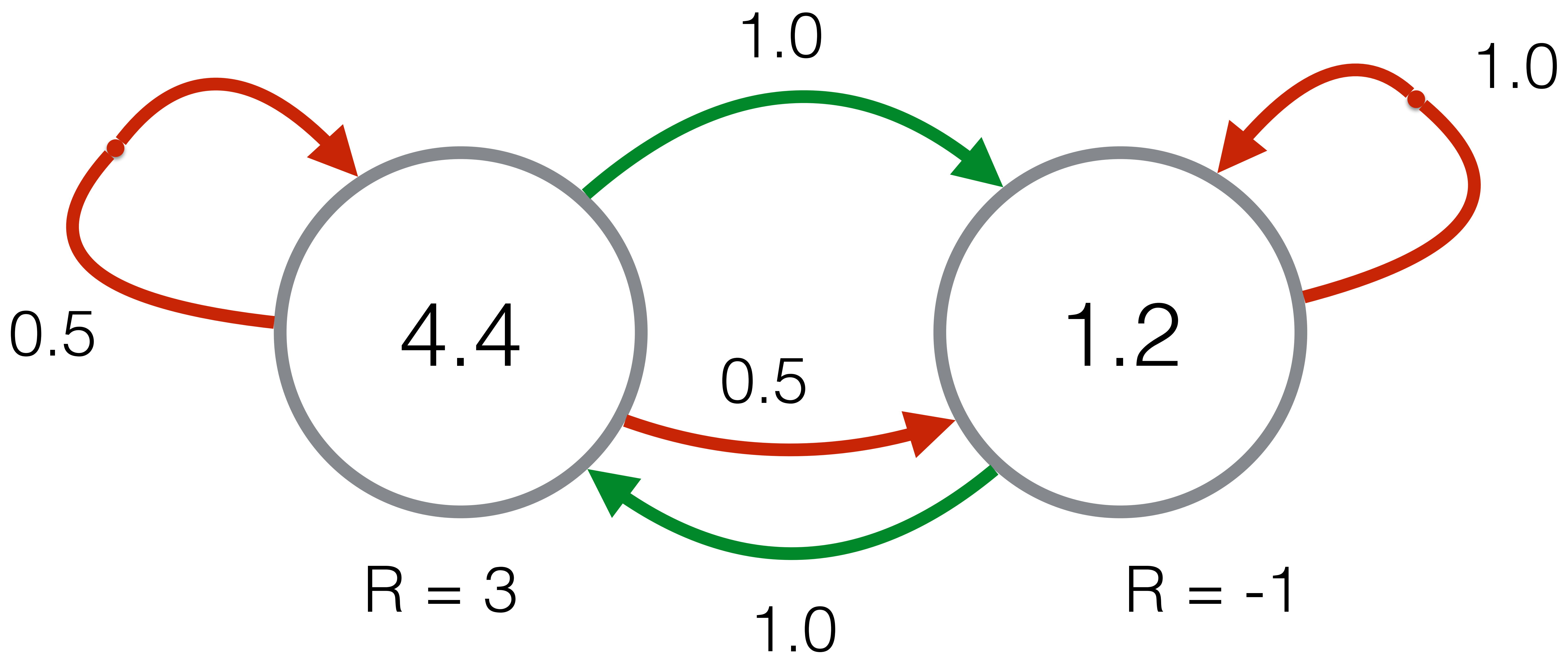


3 + 0.5 max { 1.0 * **1.2**, 0.5 * **4.4** + 0.5 * **1.2** }

3 + 0.5 max { 1.2, 2.2 + 0.6 }

3 + 0.5 max { 1.2, 2.8 }

3 + 0.5 * 2.8 = 4.4



$$\pi^*(s) = \arg\max_{a \in A(s)} \sum_{s'} P(s'|s,a)U(s')$$

Summary

- Markov decision processes
  - actions have probabilistic state transitions
- Discounted reward function
- Optimal policy maximizes expected reward
- Value iteration
- Chapter 17 to end of 17.2