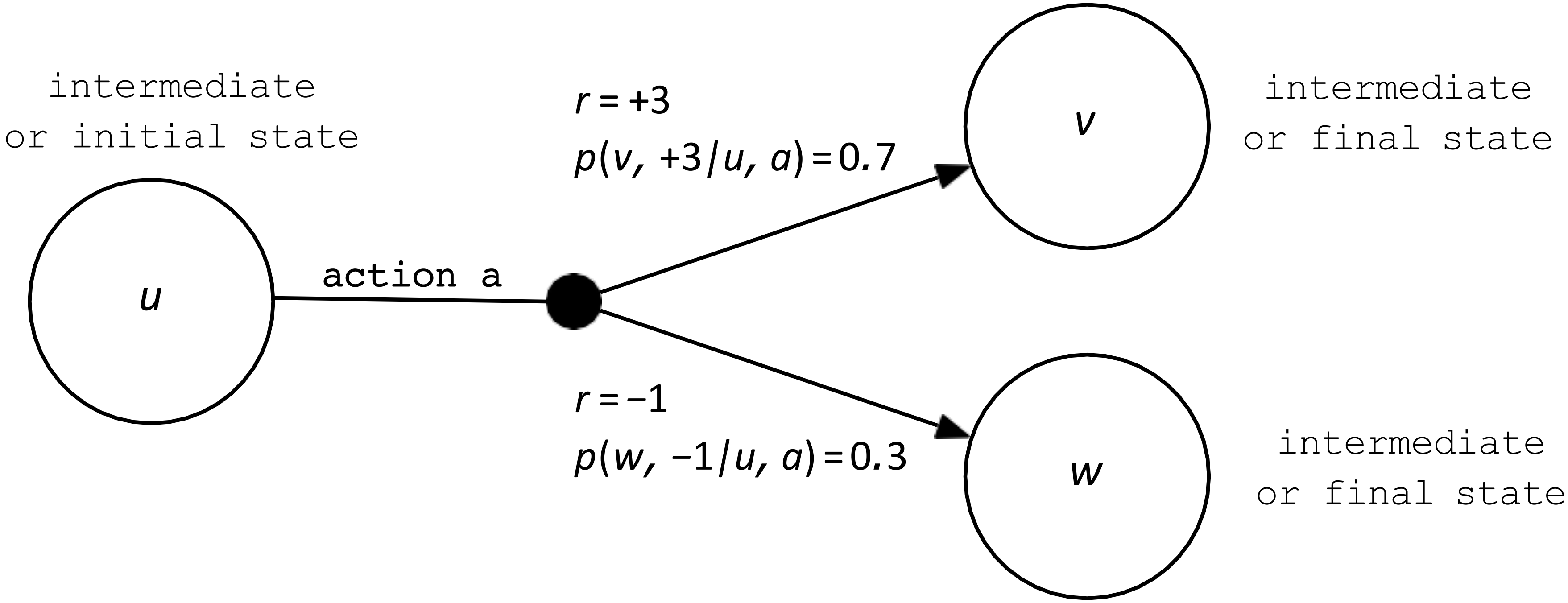
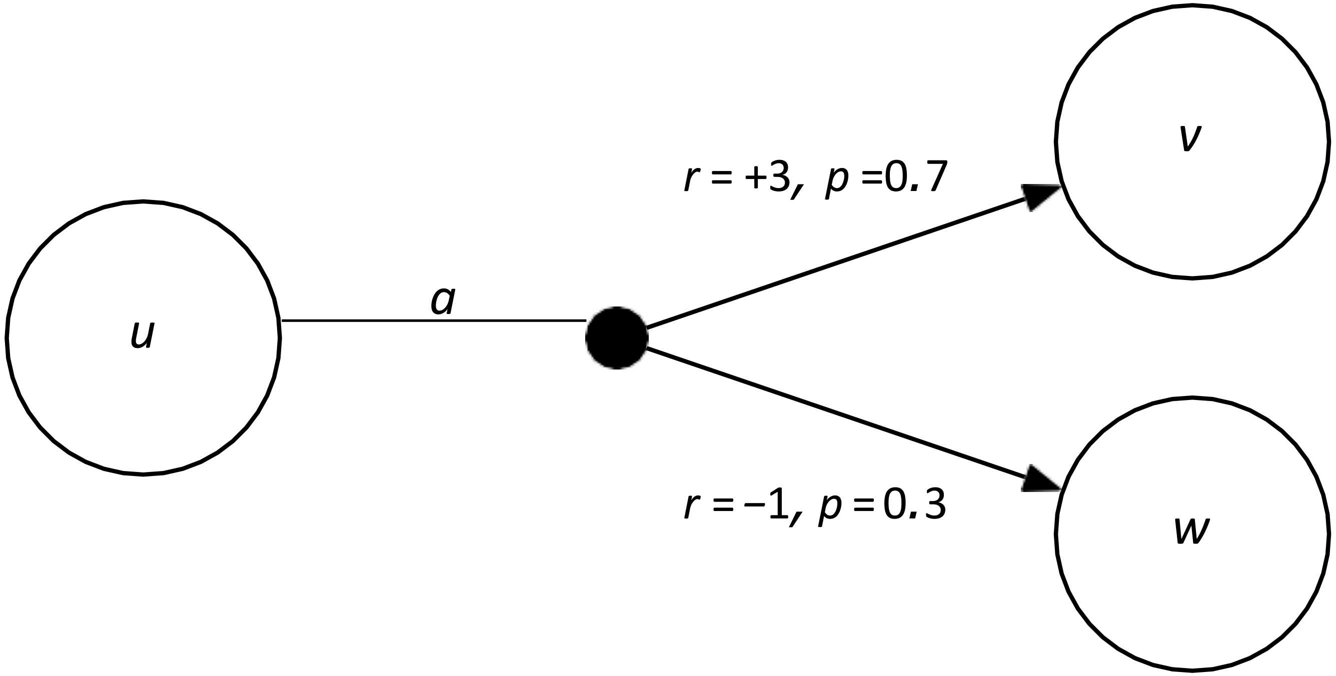
# MDP: Markov Decision Processes

# Basic block: **state**, **action**, **model**, **reward**



# Basic block: **state**, **action**, **model**, **reward**



# Markov Decision Process: **MDP**

<u>Markov decision process data</u>

- A set of **states** $S$, a set of **actions** $A$ and a set of **rewards** $R$
- For each state $s \in S$ and action $a \in A$, a probability distribution $p(\cdot, \cdot | s, a)$ over $S \times R$
- A discount factor $\gamma \in [0, 1]$

<u>Distribution model</u>

The probability distribution $p$ is called **distribution model**, or simply model, of the MDP

<u>Focus on **finite MDP**</u>

From now on, assume that $S$, $A$ and $R$ are finite

# MDP: meaning of the model

Markov decision process data

- A set of **states** $S$, a set of **actions** $A$ and a set of **rewards** $R$
- For each state $s \in S$ and action $a \in A$, a probability distribution $p(\cdot, \cdot | s, a)$ over $S \times R$
- A discount factor $\gamma \in [0, 1]$

From distribution model to random variables $S_t$ and $R_t$

The probability distribution $p$ of the MDP gives the **next** state and reward:

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

# MDP: meaning of the model

Exercises

- Explain what $S_t$, $A_t$ and $R_t$ are
- Given $p$, give a formula for $\Pr(S_t = s' | S_{t-1} = s,\ A_{t-1} = a)$
- Given $p$, give a formula for $\mathbb{E}[R_t | S_{t-1} = s,\ A_{t-1} = a]$
- Given $p$, give a formula for $\mathbb{E}[R_t | S_{t-1} = s,\ A_{t-1} = a, S_t = s']$

# The M in MDP: **Markov** property

**Tabular** representation: transitions

An action $a \in A$ gives a **transition probability** from a state $s$ to a state $s'$ :

$$P_{ss'}^{a} \coloneqq p(s'|s, a =) = Pr(S_t = s'|S_{t-1} = s, A_{t-1} = a)$$

Thus, we have a **transition matrix** $P^a$ for each action $a$, and a corresponding underlying **Markov** stochastic process.

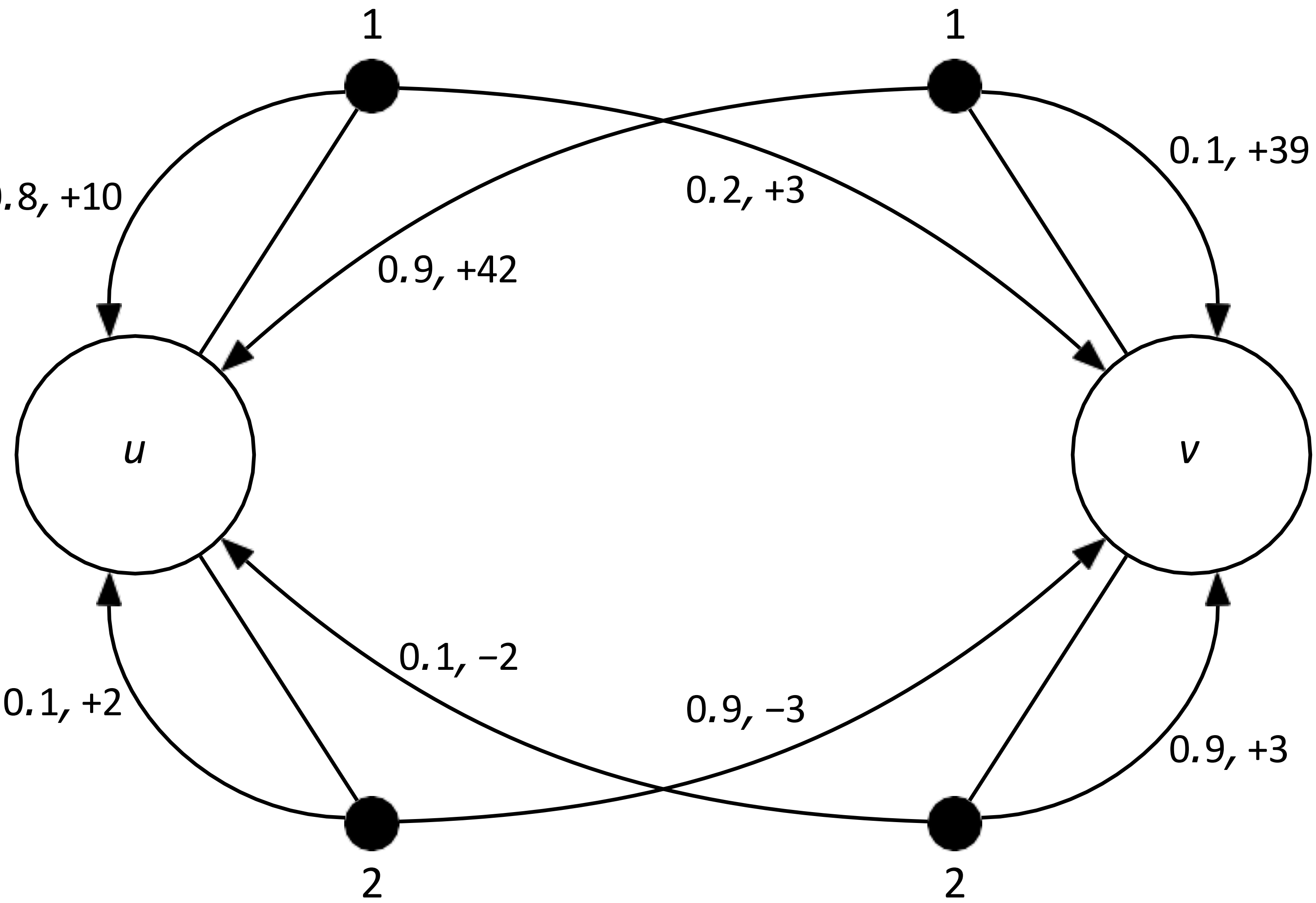
**Tabular** representation: rewards

An action $a \in A$ gives an **average reward** for any state $s$:

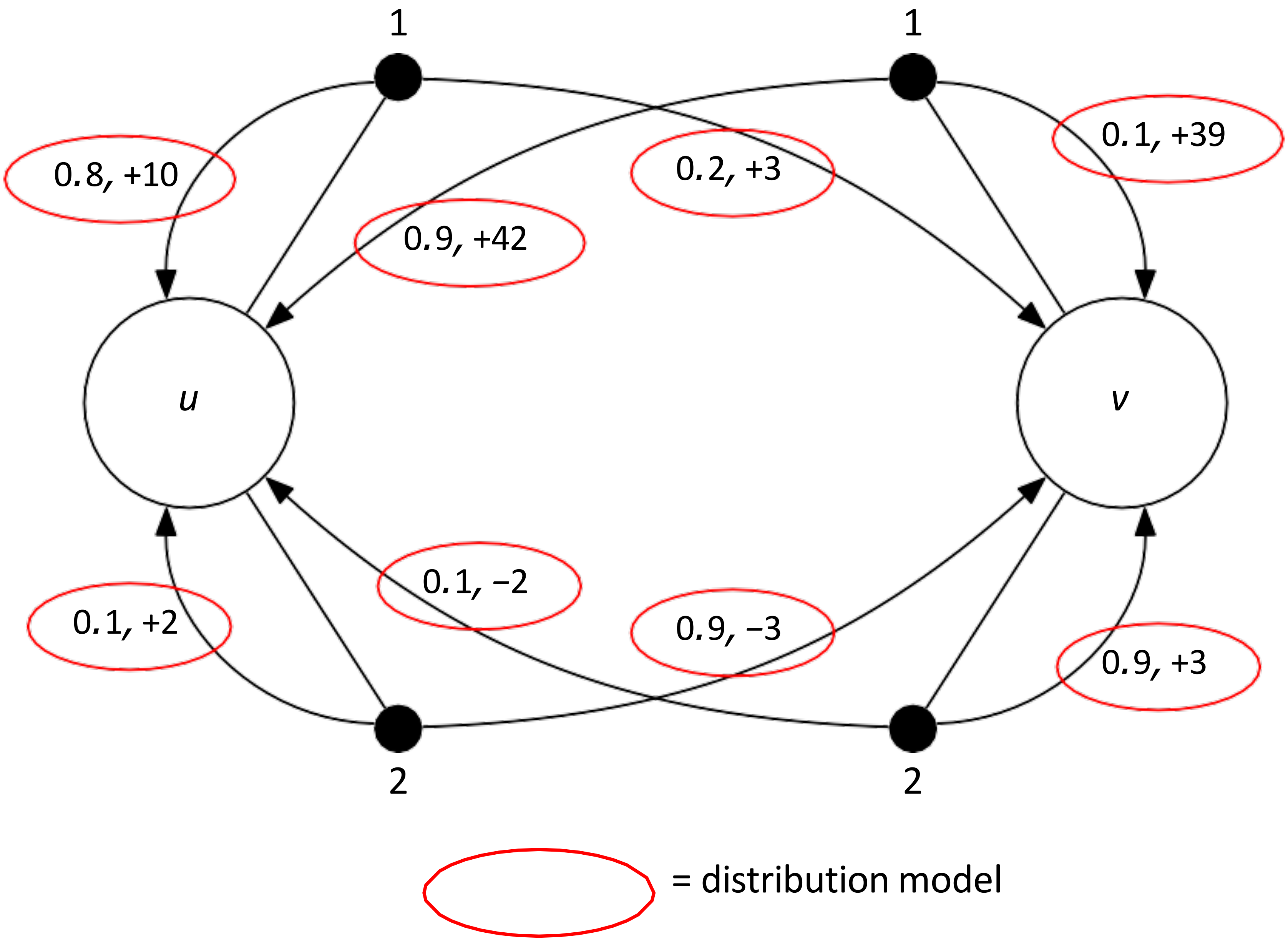
$$R_s^a = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a]$$

Thus, we have an **average reward vector** $R^a$ for any action $a$.

# Example

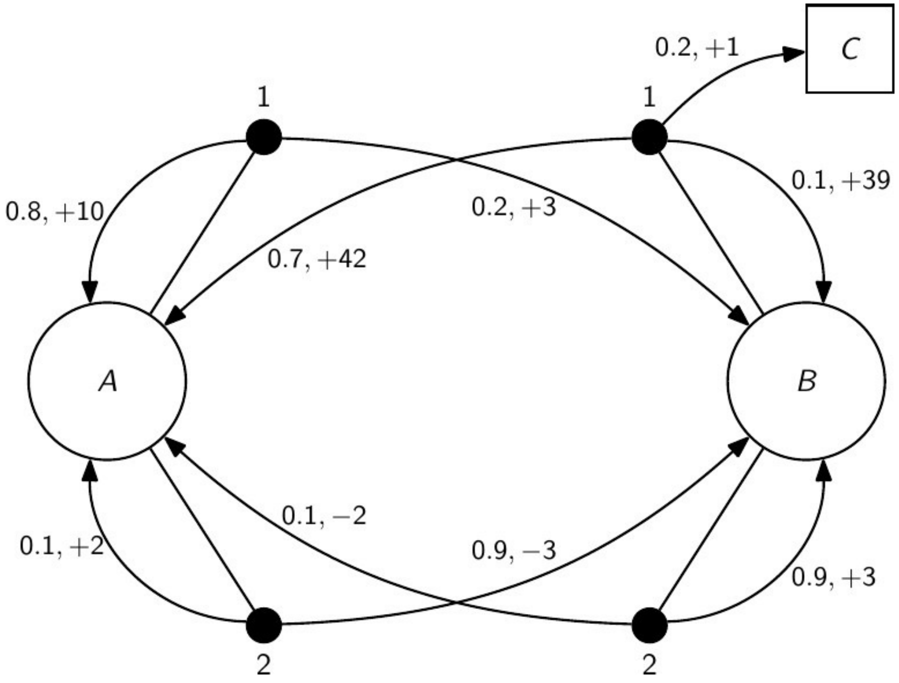


# Example



1
1
0.8, +10
0.2, +3
0.1, +39
0.9, +42
u
v
0.1, −2
0.1, +2
0.9, −3
0.9, +3
2
2
= distribution model

# Episodic MDP




- If there is a special **terminal state** reachable from every state, the MDP is **episodic**
- Otherwise, the MDP is **continuing**
- **Episode**: any sample $S_0, A_0, R_1, S_1, \ldots$ terminating in the final state

Exercise

- Write an episode, and compute its probability of happening. Hint: tricky question.

# The D in MDP: **decisions**

Where are the decisions?

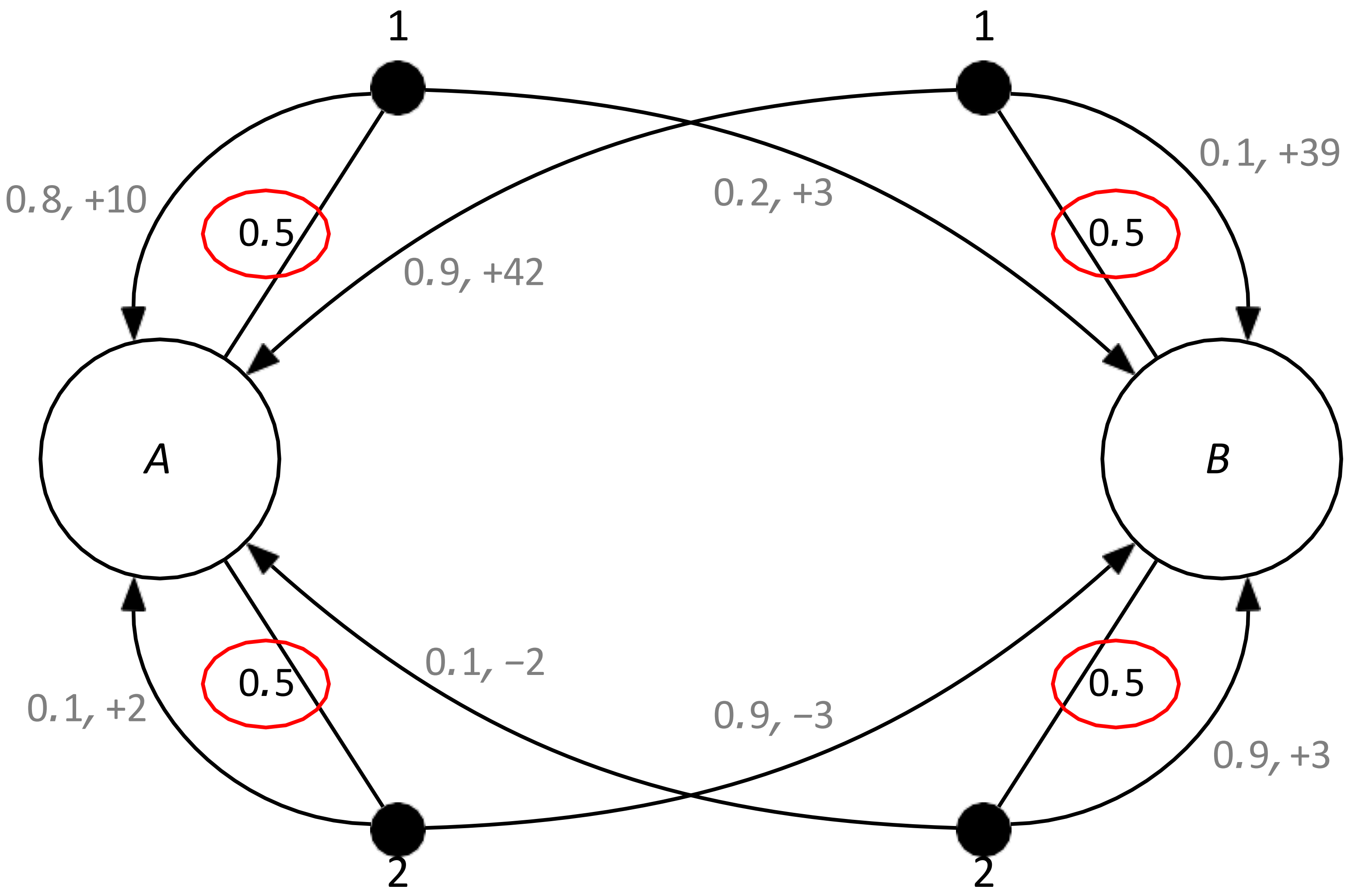
- In any state $s$, **the agent must choose** between available actions $a$
- When choosing $a$ from $s$, the environment answers $s'$ with probability $P_{ss'}^{a}$ . Environment decision.
- The agent behaviour is given by probabilities $\pi(a|s)$: "how likely I'm going to choose $a$ from $s$?". Agent decision.

Definition

A **policy** $\pi$ is a probability distribution over actions given states:

$$\pi(a|s) := \Pr(A_t = a|S_t = s)$$

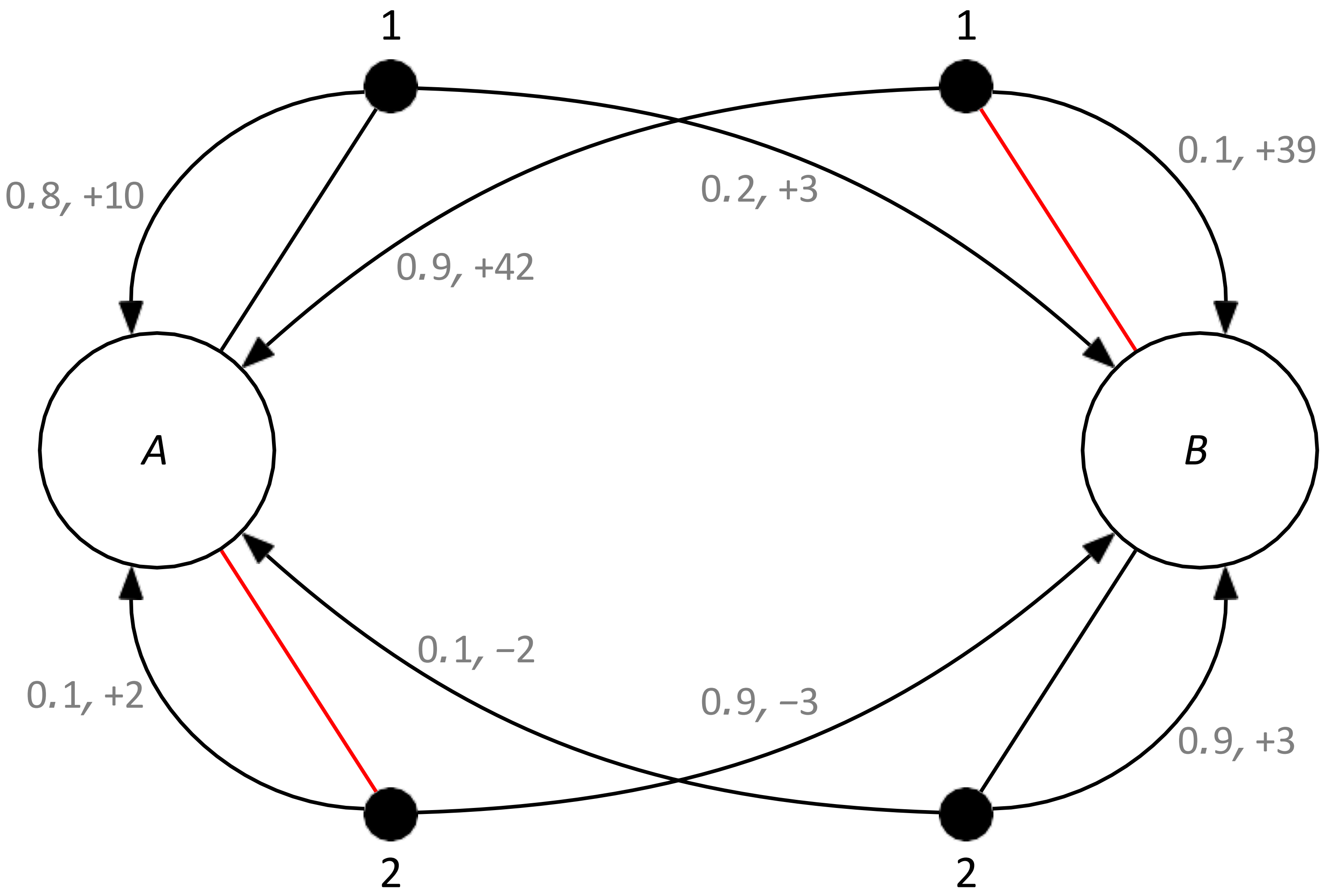
# Example: uniform stochastic policy




What can we do?

At every step, we choose the action according to the probability.

# Example: deterministic policy




What can we do?

At every step, we choose the given action.

# Tabular representation

$S$ and $A$ are finite

A policy can be represented by a table: every line in the table corresponds to a state.

Stochastic policy

| | |
|---|---|
| A | [0.5,0.5] |
| B | [0.5,0.5] |

Deterministic policy

| | |
|---|---|
| A | 2 |
| B | 1 |

# The **return**: towards the goal

Definition

- **Total return** of an episode ending at time $T$: the value of the random variable $G_t := R_{t+1} + R_{t+2} + \cdots + R_T$ for the episode
- If the MDP is continuing, we need a **discount factor**:

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

  Why?

- Transforming the *terminal* state in *absorbing* with reward 0, we can use a **unified notation** for episodic and continuing MDP:

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

- In episodic tasks we can use $\gamma = 1$, in continuing tasks we must use $\gamma < 1$

# The **return**: towards the goal

Why the discount

- The discount factor measures how much do we care about rewards far in the future
- A reward $r$ after $k + 1$ time-steps is worth "only" $\gamma^k r$ : we say **myopic evaluation** if $\gamma \sim 0$, **far-sighted evaluation** if $\gamma \sim 1$
- Convenience: avoids infinite returns in cyclic MDP
- We shouldn't trust our model too much: uncertainty about the future may not be fully represented
- If the reward is financial, immediate rewards may earn more interest than delayed rewards
- Animal and human behaviour shows preference for immediate reward

# How much are states and actions worth?

Remark

The total return $G_t$ at time $t$ is a random variable:

$$G_t \coloneqq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ \dots = \sum_{k=0}^{+\infty} \gamma^k R_{t+k+1}$$

Thus, it makes sense to compute its expected value.

Definition: **state-value function**

The **state-value function** $v_\pi(s)$ for a MDP is the return we can expect to accumulate starting from state $s$, **following the policy** $\pi$:

$$v_\pi(s) := \mathbb{E}_\pi[G_t | S_t = s]$$

Exercise

Is the above definition/notation correct?

# How much are states and actions worth?

Total return

$$G_t := R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ \ldots = \sum_{k=0}^{+\infty} \gamma^k R_{t+k+1}$$

State-value function

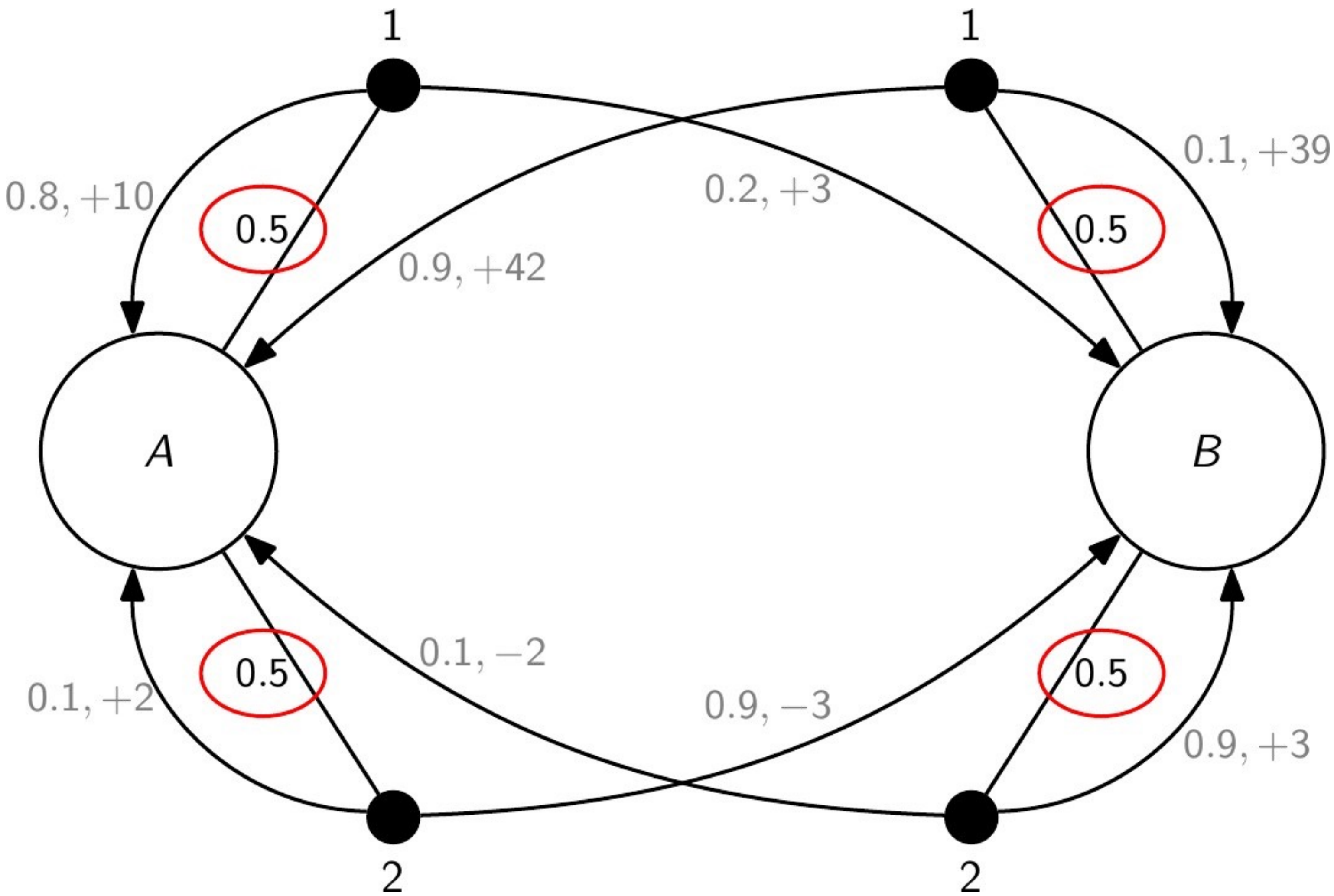
$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$

Definition: **action-value function**

The **action-value function** $q_\pi(s,a)$ for a MDP is the return we can expect to accumulate starting from a state $s$, choosing action $a$, and then **following the policy** $\pi$:

$$q_\pi(s,a) := \mathrm{E}_\pi[G_t | S_t = s, A_t = a]$$

# Example




Exercise

Compute $q_\pi(A, 1)$, $q_\pi(A, 2)$, $q_\pi(B, 1)$ and $q_\pi(B, 2)$ for the uniform policy $\pi$.