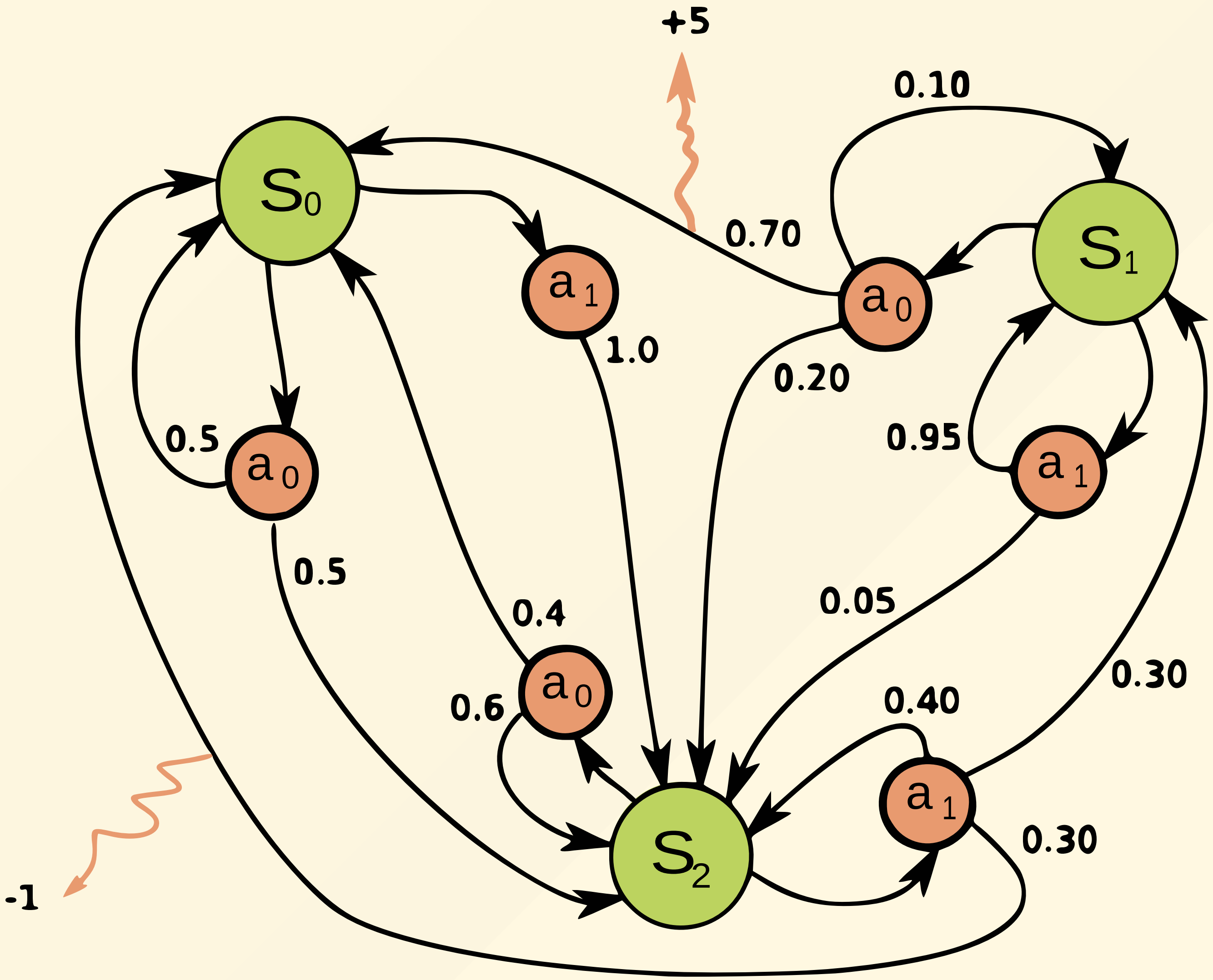
Definitions

1. Agent.
2. Environment.
3. State.
4. Observation.
5. Episode.

MDP

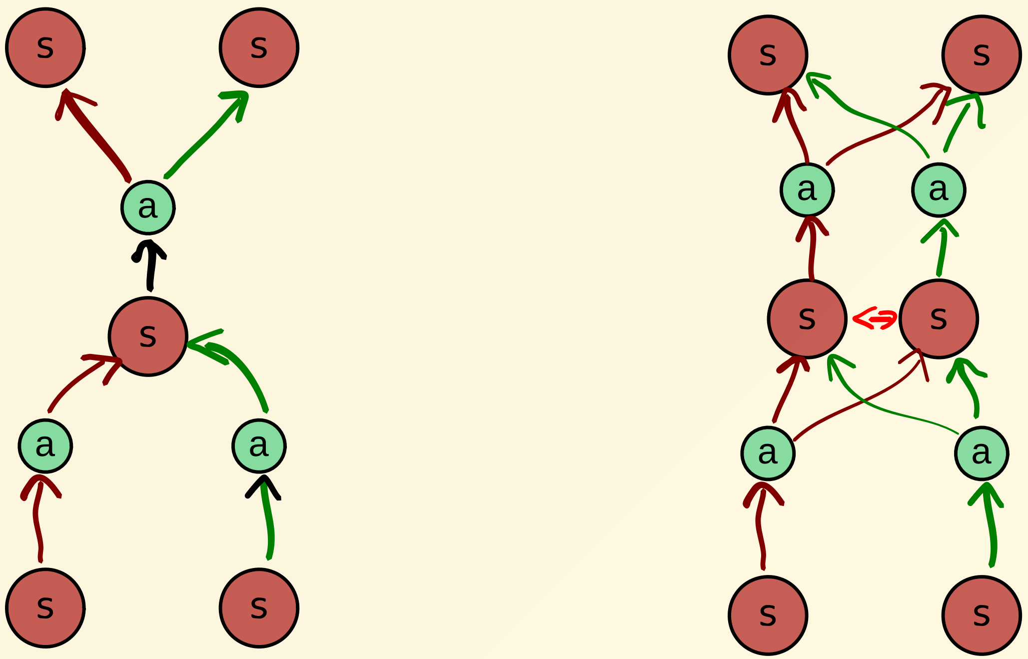
A MDP is a 4-tuple (S, A, P_a, R_a), where:

- $S :=$ is a set of states called the state space.
- $A :=$ is a set of actions called the action space.
- $A_s :=$ is a set of actions available from state $s \epsilon S$.
- $P_a(s, s') := \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a)$ is the probability that action a in state s at time t will lead to state s' at time $t + 1$.
- $R_a(s, s')$ is the immediate reward recived after transition from s to s'.


+5
0.10
S_0
S_1
S_2
a_1
a_0
0.70
1.0
0.20
0.5
0.95
0.5
0.05
0.4
0.30
0.6
0.40
0.30
-1

Marcov property

$$\mathbb{P}(S_{t+1}|S_t, A_t) = \mathbb{P}(S_{t+1}|S_t, A_t, S_{t-1}, A_{t-1}, \dots)$$

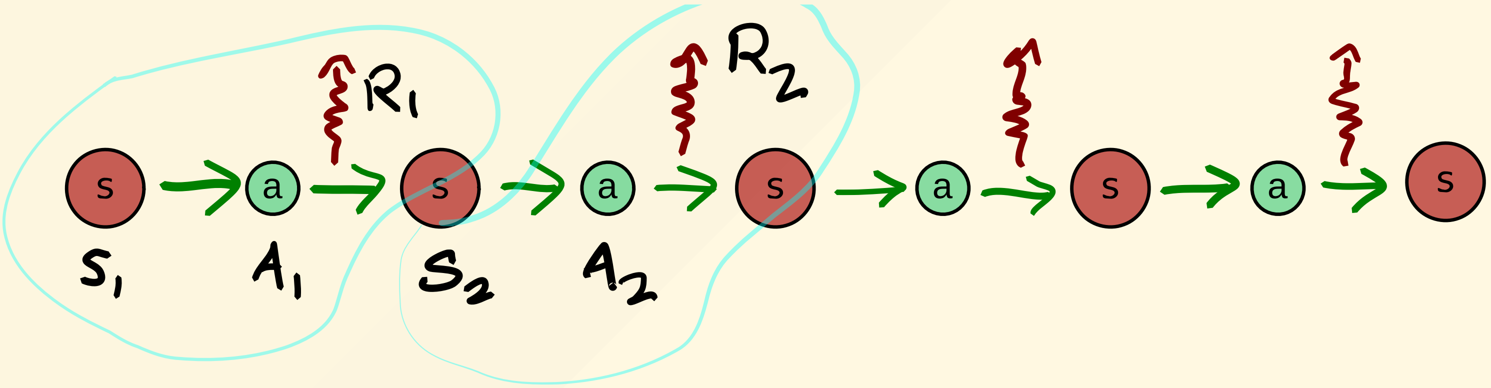


Episode

Is a sequence :

$$[(S_0, A_0, R_0), (S_1, A_1, R_1), ..., (S_T, A_T, R_T)]$$

its just one run.



Return

A given episode $[(S_0, A_0, R_0), (S_1, A_1, R_1), \ldots, (S_T, A_T, R_T)]$ of MDP and a given $\gamma \epsilon [0, 1]$.

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

or

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

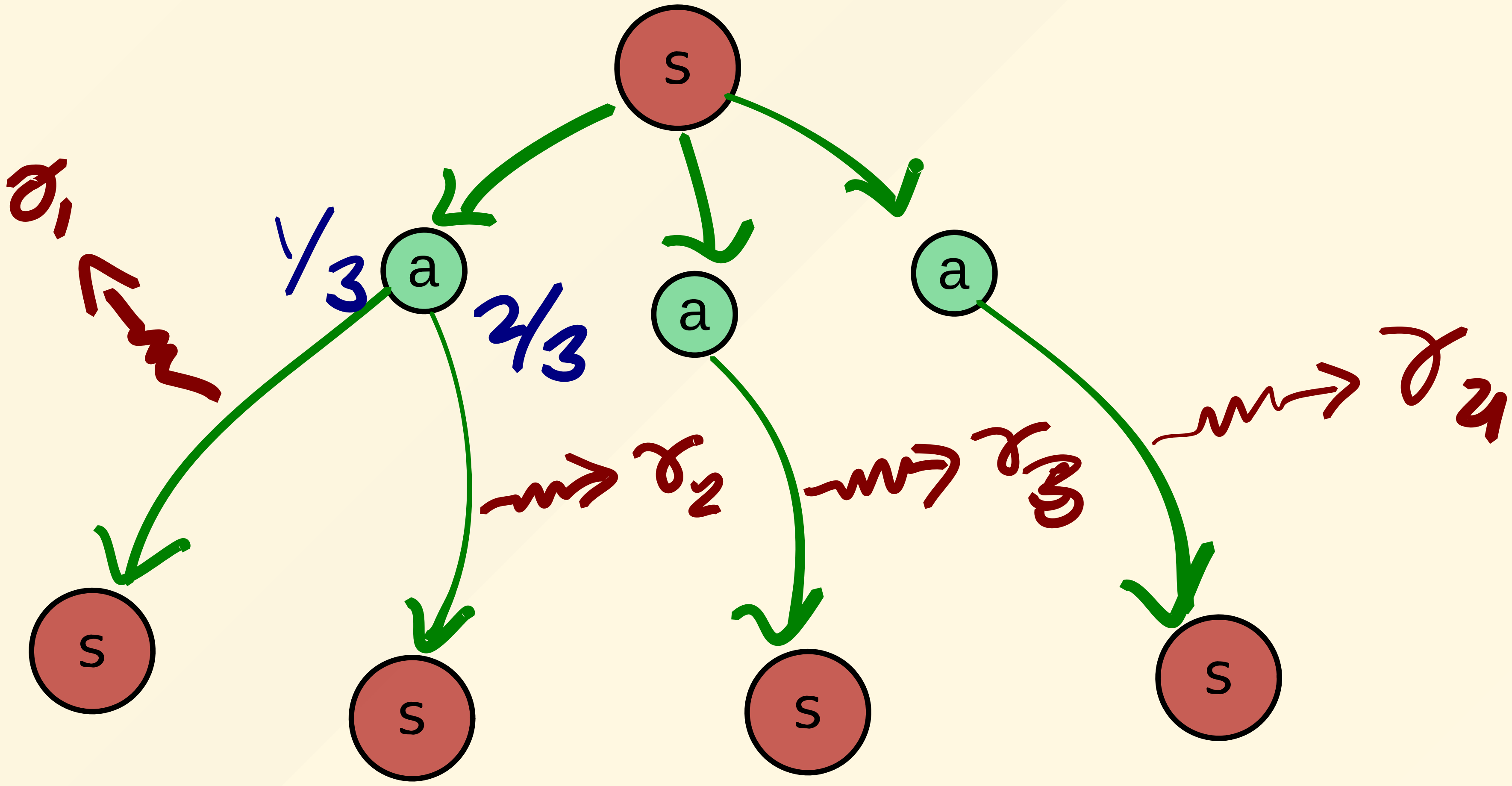
Reward function.

Given MDP we define reward function. $s, s' \epsilon S, a \epsilon A$

$$r(s) = \mathbb{E}_{a,s'}[R_{t+1}|S_t = s]$$

$$r(s, a) = \mathbb{E}_{s'}[R_{t+1}|S_t = s, A_t = a]$$

$$r(s, a, s') = \mathbb{E}[R_{t+1}|S_t = s, A_t = a, S_{t+1} = s']$$



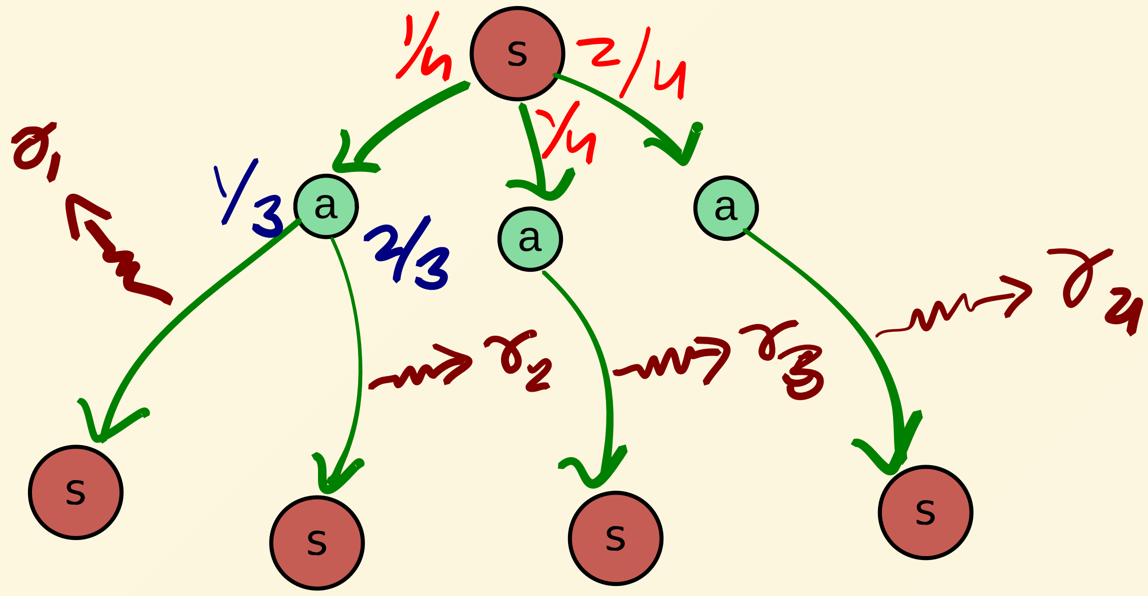
s
a
a
a
s
s
s
s
1/3
2/3
r_1
r_2
r_3
r_4

Policy

Given a MPD we define

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$$

- A policy fully defines the behavior of an agent.



Now we acn talk about:

$$r(s) = \mathbb{E}_{a,s'}[R_{t+1} | S_t = s]$$

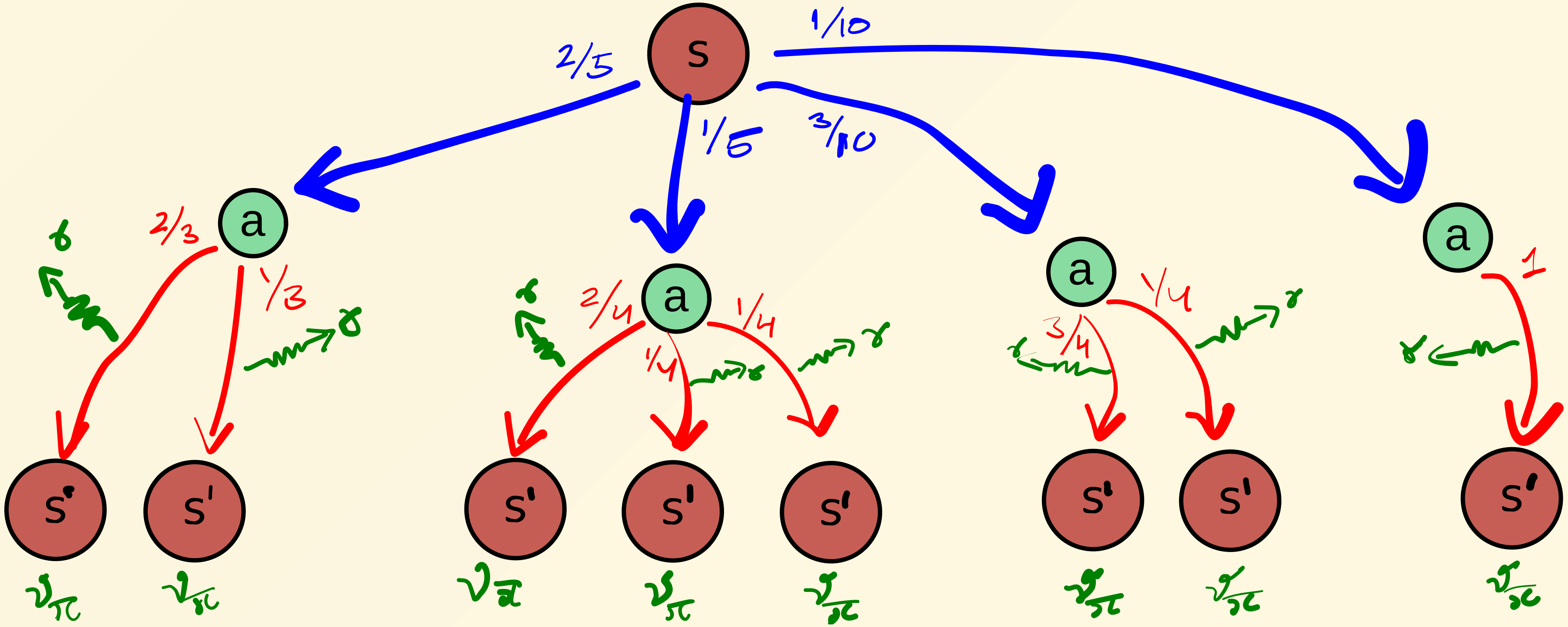
State-value function V

Given a MDP and a policy π on it we define

- $$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] \;\; \forall s \epsilon S$$
- $$V_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] \;\; \forall s \epsilon S$$
- $$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')], \quad \forall s \epsilon S$$

See inside This equation

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

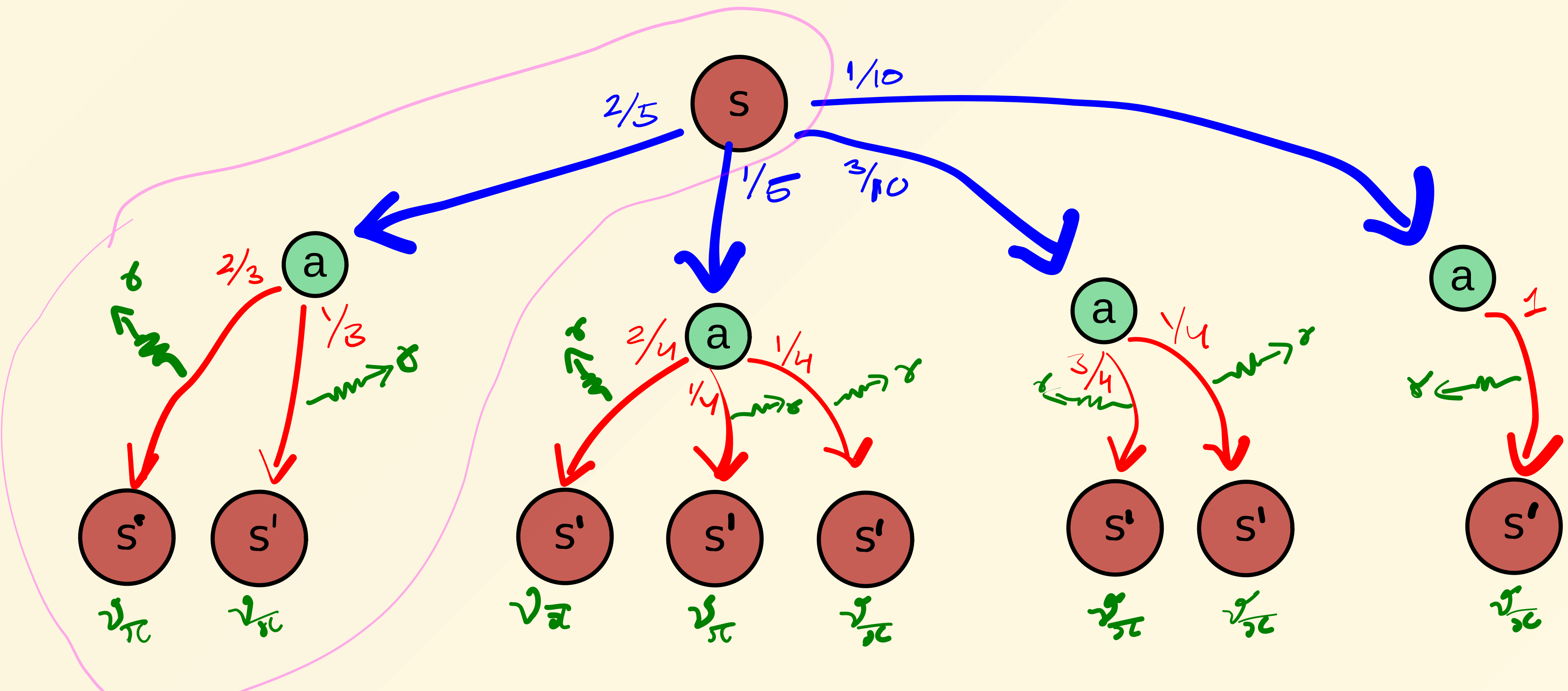


Action-value function Q

- $$q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$
- $$q_\pi(s,a) = \mathbb{E}_\pi[R_t + \gamma G_{t+1} | S_t = s, A_t = a]$$
- $$q_\pi(s,a) = \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')], \forall s \epsilon S, \forall a \epsilon A$$

See inside This equation

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s')], \forall s \epsilon S, \forall a \epsilon A$$



Action-advantage function A

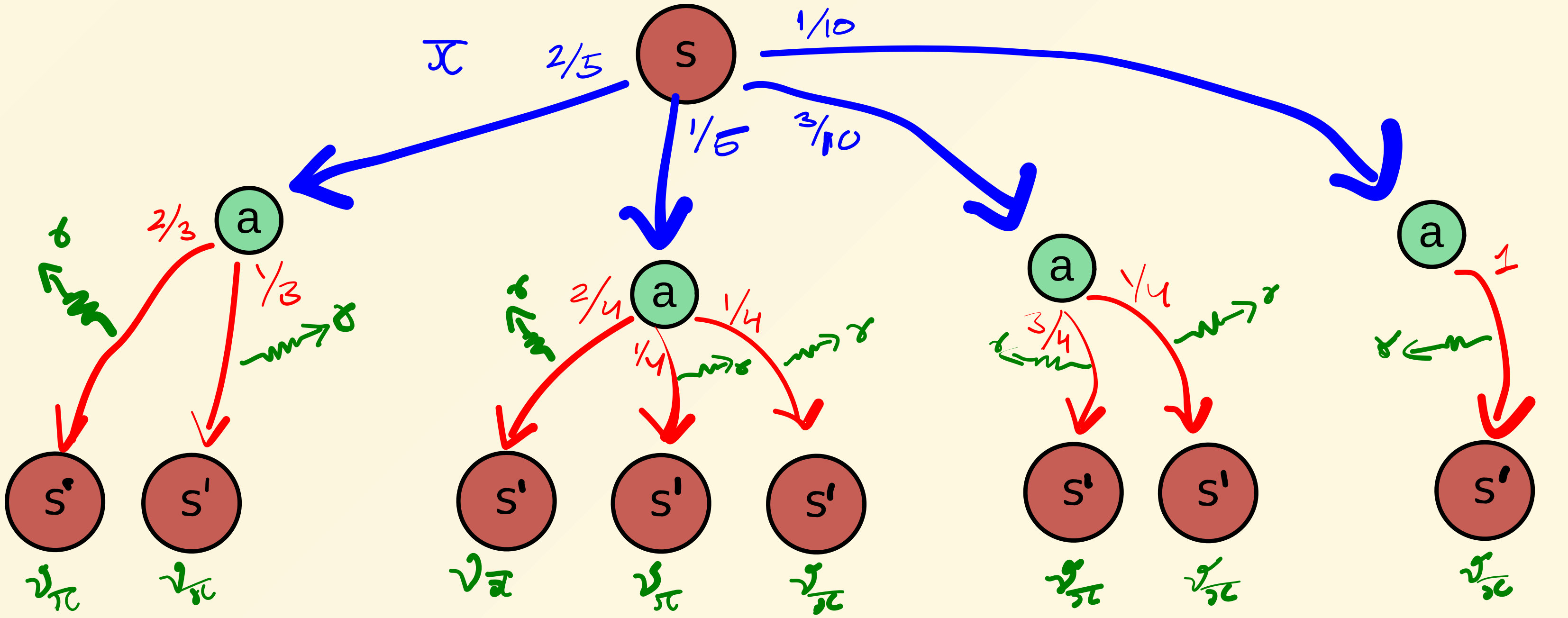
$$a_\pi(s, a) = q_\pi(s, a) - v_\pi(s)$$

- The advantage function describes how much better it is to take action a instead of following policy π.
- It can be nagitive.

Bellman optimality equations

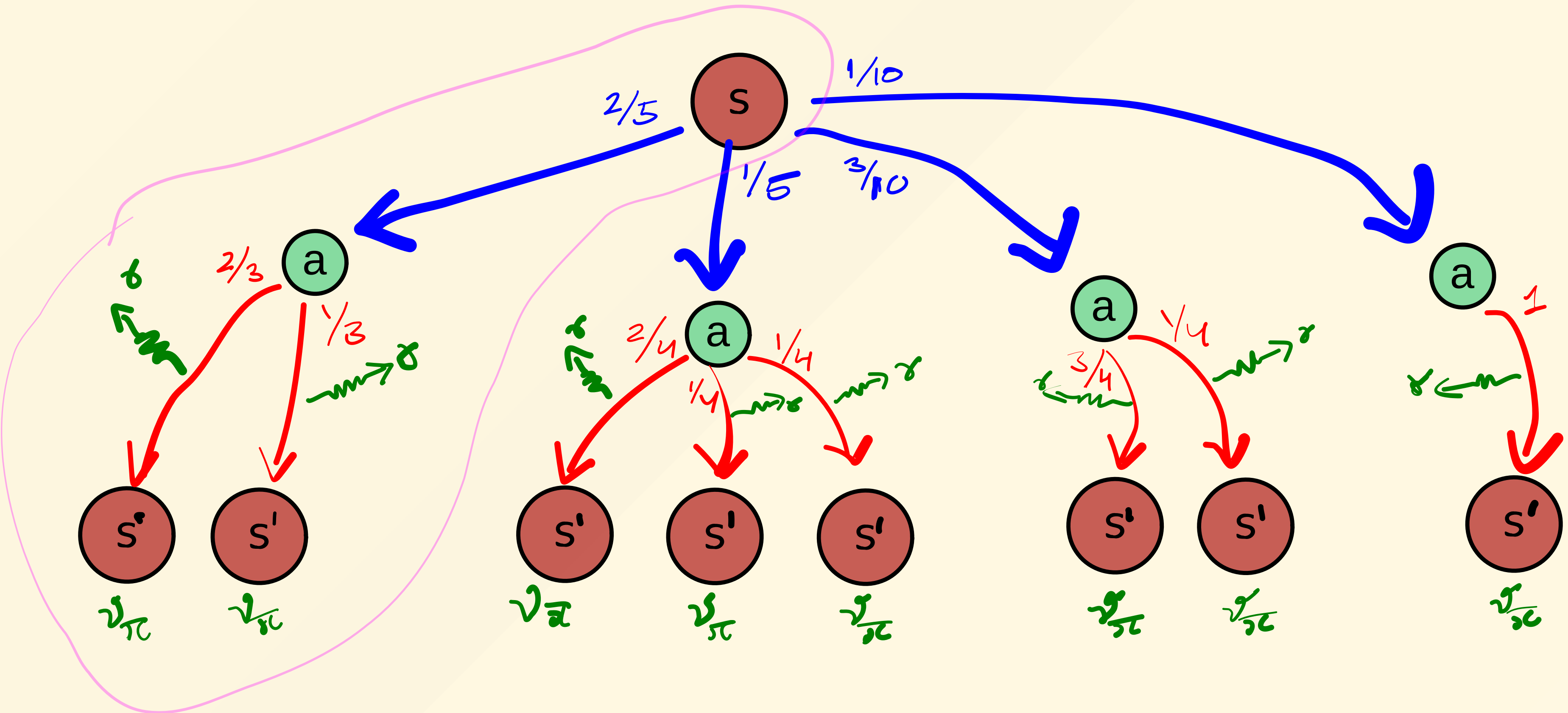
optimal state-value function

$$v_*(s) = \max_{\pi}[v_\pi(s)] \quad \forall s \epsilon S$$



Optimal action-value function.

$$q_*(s,a) = \max_{\pi}[q_\pi(s,a)], \forall s \epsilon S, \forall a \epsilon A$$



Policy-evaluation

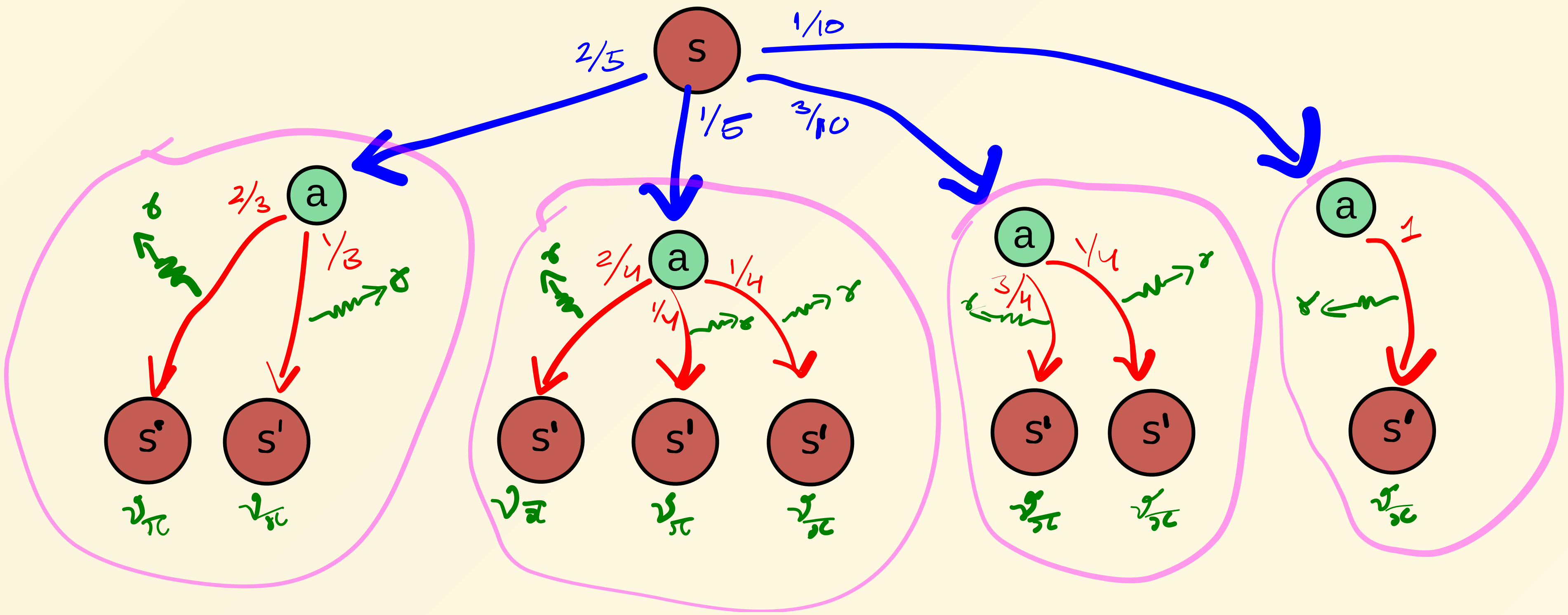
Given a policy we evaluate value function by iteration.

$$v_{k+1}(s) = \sum_{a} \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_k(s')]$$

when $k \to \infty, v_k \to v_\pi$

Policy-improvement equation

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



Value-iteration equation

$$v_{k+1}(s) = \max_a \sum_{s',r} p(s', r | s, a)[r + \gamma v_k(s')]$$

Summary

- MDP
- Markove property
- Episode

THANKS