- Stochastic policy:

$$\pi(a \mid s) = P[\text{Action} = a \mid \text{state} = s]$$

Value Function

The value function predicts the (discounted) future reward in a state given a policy

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \ \gamma^\infty R_{t+\infty} \mid S_t = s]$$

A Markov decision process (S, A, P, R, γ) is a Markov reward process (S, P, R, γ) with associated finite set of actions A. It consists of

- a finite set of states S
- a finite set of actions A
- a reward function
  - $R_s^a = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
- a discount factor $0 \leq \gamma \in \mathbb{R} \leq 1$
- a stochastic matrix P describing state transition
  - $P_{s,s'}^a = P[S_{t+1} \mid S_t = s, A_t = a]$

Reminder: Bellman Equation (Expectation)

Both value functions (for expectation) can recursively be decomposed in the same way, into

- immediate reward and
- discounted future reward

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

$$q_\pi(s, a) = \mathbb{E}[R_{t+1} + \gamma q_\pi(s_{t+1}, a_{t+1}) \mid S_t = s, A_t = a]$$