Value Estimation

Model of the Environment

- Where does an action take us?

$$\delta(s, a) \mapsto s'$$

- How much reward do we receive?

$$r(s, a) \mapsto \Re$$

The values of different states are interrelated
Bellman's Equation:

$$V^\pi(s) = r(s, a) + \gamma \cdot V^\pi(\delta(s, a)) \qquad \text{where } a = \pi(s)$$

Can we solve *Bellman's equation*?

$$V^\pi(s) = r(s,a) + \gamma \cdot V^\pi(\delta(s,a)) \qquad \text{where } a = \pi(s)$$

- Direct solution (linear equation system)
- Iteratively (*value iteration*)

$$V^\pi_{k+1}(s) \leftarrow r(s,\pi(s)) + \gamma \cdot V^\pi_k(\delta(s,\pi(s)))$$

How do we find an optimal policy $\pi^\star$?
Easy if the optimal value function $V^\star$ was known:

$$\pi^\star(s) = \arg\max_a \left(r(s,a) + \gamma \cdot V^\star(\delta(s,a))\right)$$

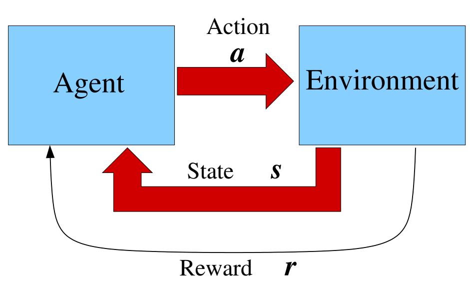
Optimal version of Bellman's equation

$$V^\star(s) = \max_a \left(r(s,a) + \gamma \cdot V^\star(\delta(s,a))\right)$$

Hard to solve

Policy iteration:

Interleaved calculation of policy and values



Normal scenario: *Unknown environment*
$r(s,a)$ and $\delta(s,a)$ are not known

V^π must be estimated from experience

Monte-Carlo Method

- Start at a random s
- Follow π, store the rewards and s_t
- When the goal is reached, update $V^{\pi}(s)$-estimation for all visited states with the future reward we actually received

Very slow convergence