RL POLICIES

BIOE 498/598 PJ

Review

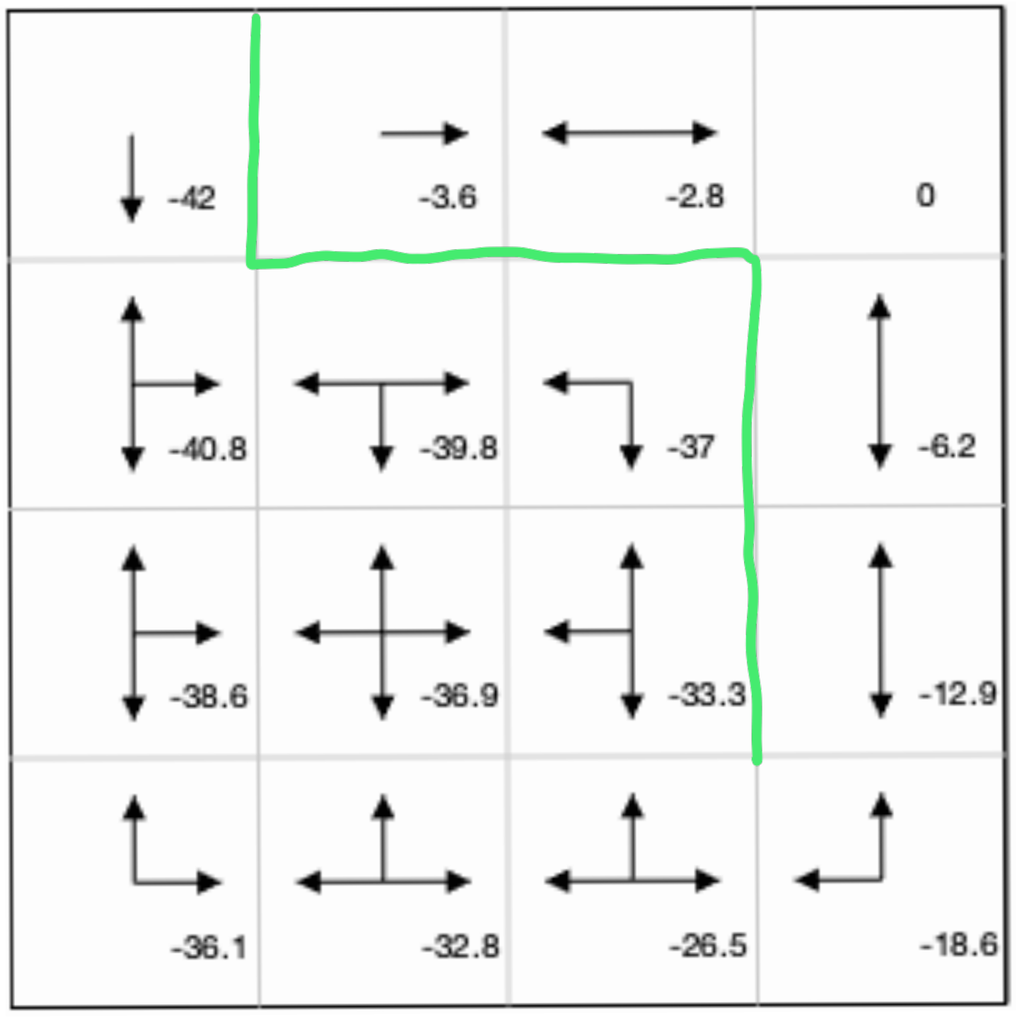
RL uses MDP framework

$$s_0 \xrightarrow{a,r} s_1 \xrightarrow{a,r} s_2 \cdots s_T$$

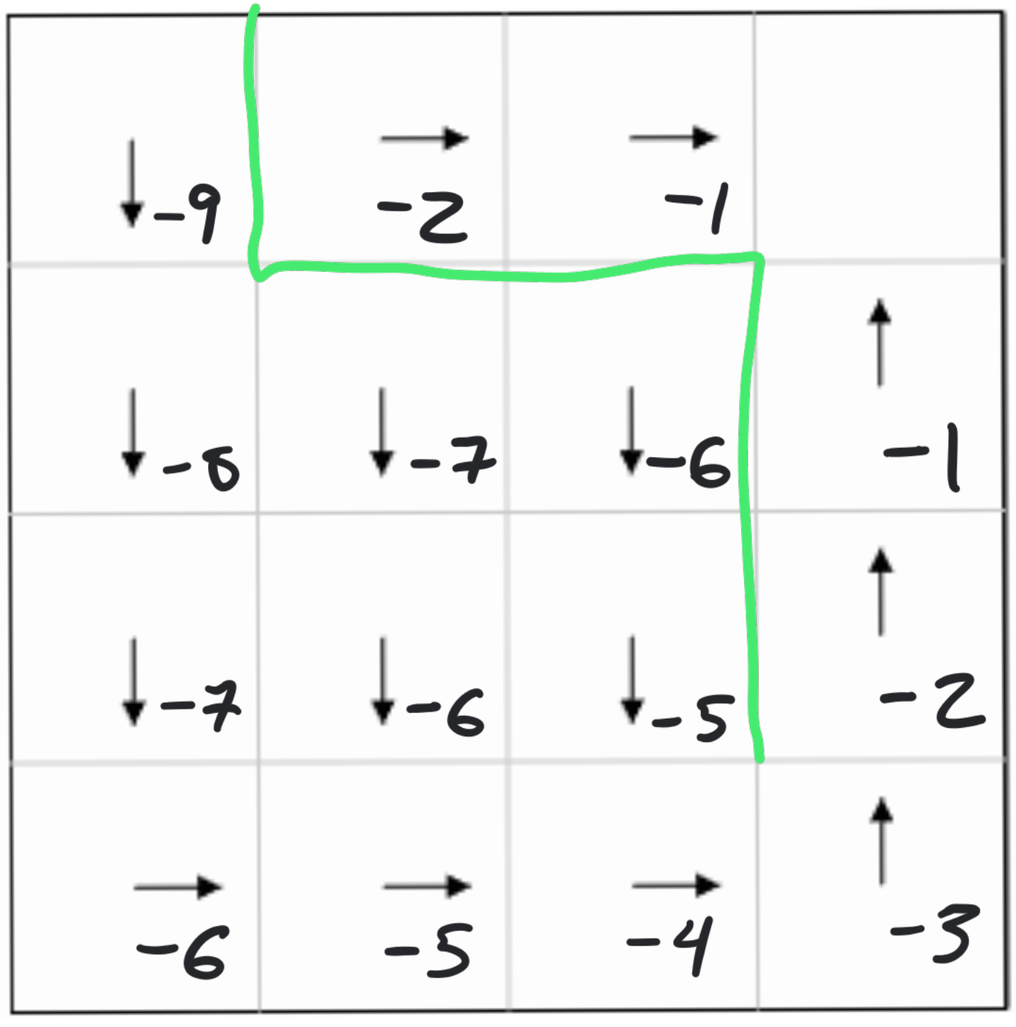
$$\max_a \left\{ r_1 + \sum_{i=2}^{T} r_i \right\} = \max_a \left\{ r_1 + V(s') \right\}$$

Policy π is a function that choose actions for every state.

$$a = \pi(s)$$



V(s) under random moves

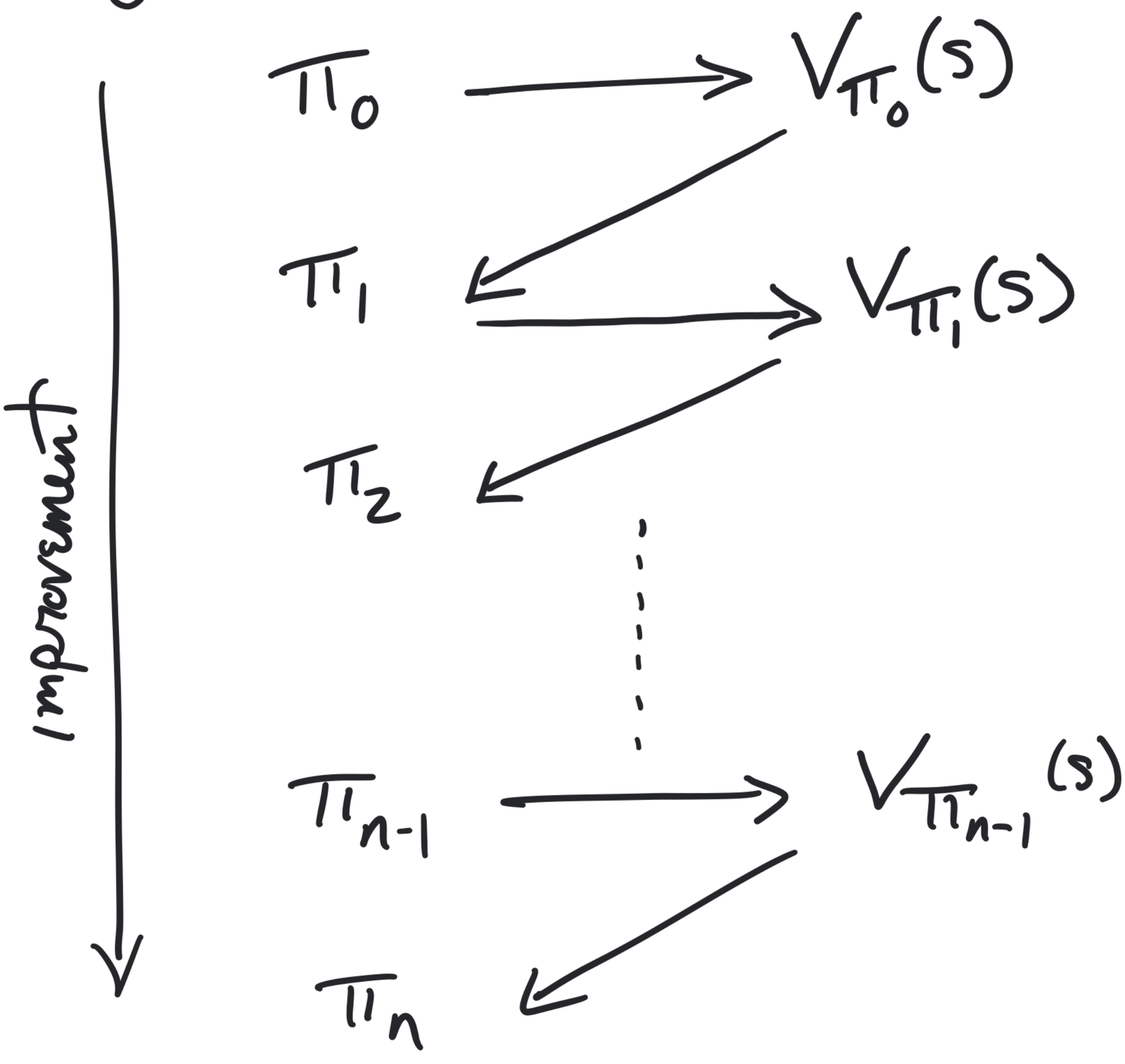


V(s) new policy

Value functions depend on the policy!

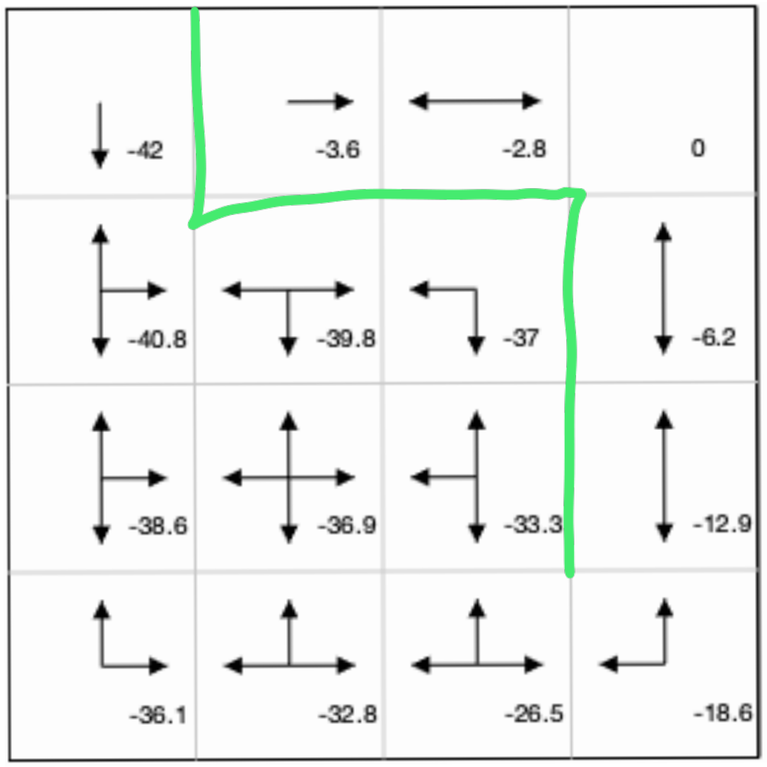
$V_\pi(s)$

Policy Improvement



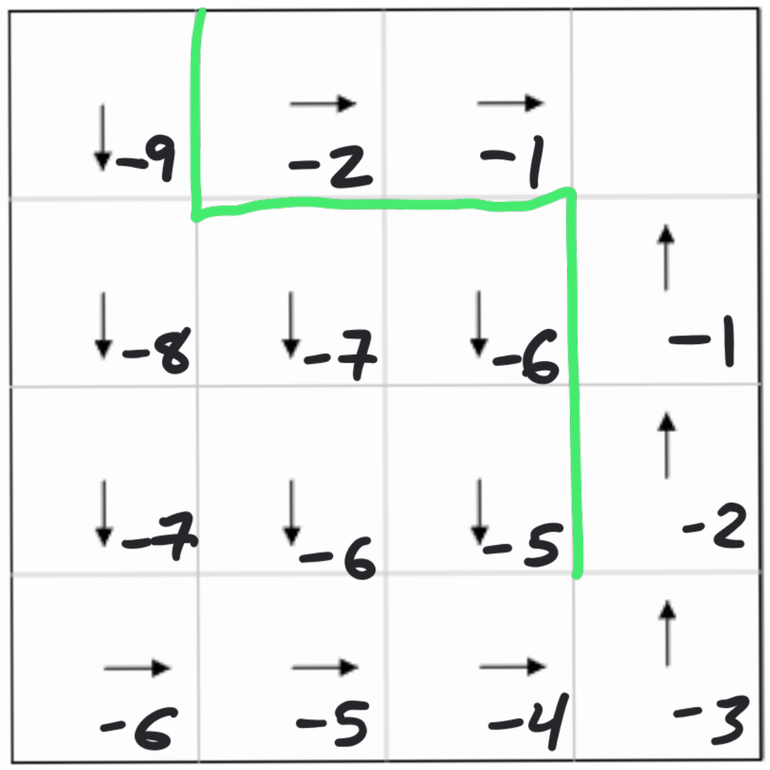
π_0 = Random

$V_{\pi_0}(s)$



π_1

$V_{\pi_1}(s)$



GRIDWORLD with A Bad policy

π_0 is biased toward the bottom left.

$\pi_0 \rightarrow V_{\pi_0}(s)$

$V_{\pi_0}(s) \rightarrow \pi_1$

$\pi_1 \rightarrow V_{\pi_1}(s)$

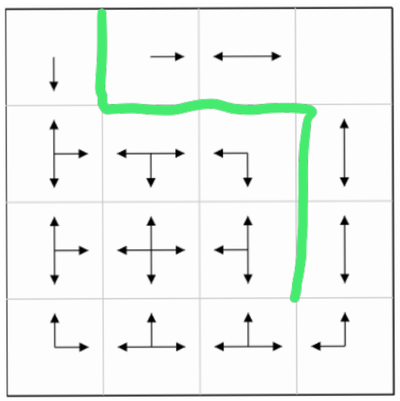
$V_{\pi_1}(s) \rightarrow \pi_2$

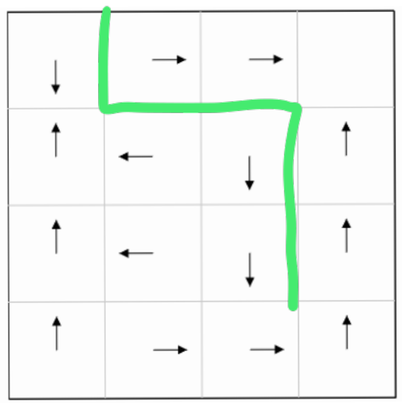
$\pi_2 \rightarrow V_{\pi_2}(s)$

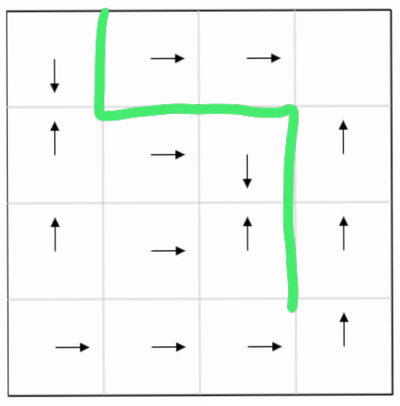
$V_{\pi_2}(s) \rightarrow \pi_3$

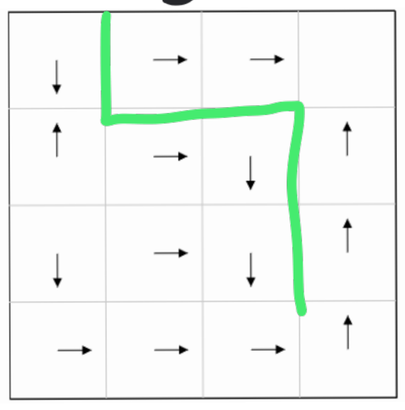
$\pi_3 \rightarrow V_{\pi_3}(s)$

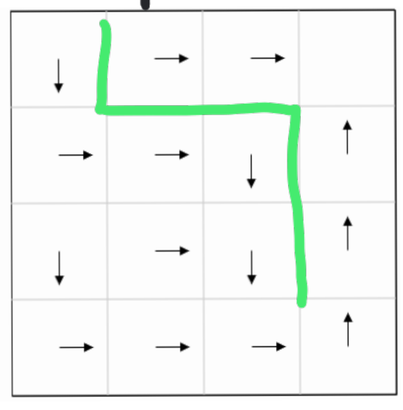
$V_{\pi_3}(s) \rightarrow \pi_4$











Policy improvement can be incremental

$\pi_i \xrightarrow{\text{partially}} V_{\pi_i}(s)$

$V_{\pi_i}(s) \rightarrow \pi_{i+1}$

General policy iteration