The sequence $\{v_k\} \longrightarrow v_\pi$ as $k \to \infty$ under the same conditions that guarantee the existence of v_π.

1. either $\gamma < 1$ ($\because \gamma \in [0, 1)$)

2. episodic task

This iterative algorithm is called "iterative policy evaluation"

expected updates:- In DP, updates are based on an expectation over all possible next states, rather than on a sample next state.

When you use the updated values in the subsequent iteration — in place algorithm. ex - $S = \{s_1, s_2, s_3\}$

→ see the pseudo code for policy evaluation from TB.

iteration 0, $V(s_1) \to 0.5$
$V(s_2) \to -1$
$V(s_3) \to 2$

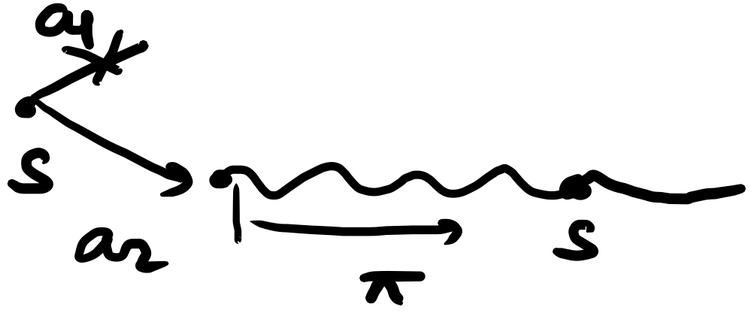
iteration 1, $V(s_1) \leftarrow$ Update
$V(s_2)$

Policy improvement :- deterministic vs. Stochastic policy :-

$\pi(a|s)$
$a \in A(s) = \{a_1, a_2, a_3\}$

Let us say we have V_π for some arbitrary det. policy π. If we change the policy to

deterministically choose $a \neq \pi(s)$ - whether we should do it or not?



$$q_\pi(s,a) = E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$\uparrow$ (a)

$\because \quad q_\pi(s,a) = E\left[G_t \mid S_t = s, A_t = a\right]$

(see earlier lectures for proof of (a))

↳ Choose a & thenfollow π, leading to value of s being $q_\pi(s,a)$

Q. $q_\pi(s,a) \geq V_\pi(s)$

Suppose the Q. has a true answer, then certainly a

policy where you select a whomever s' is encountered, the new policy would in fact be a better one overall. This is what one would expect.