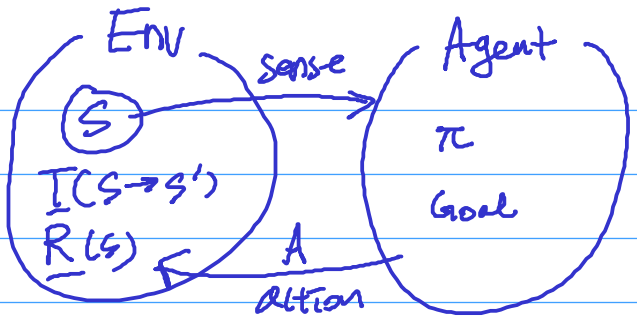
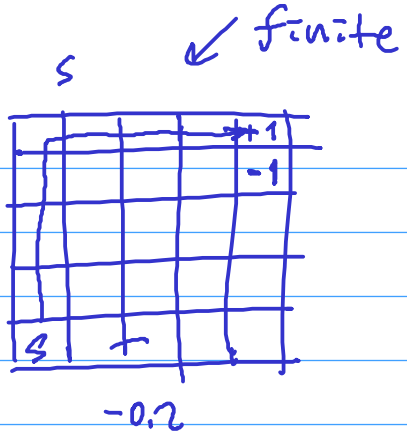




Markovian

MDP

T, R

$V(s) \to E[G_t | s]$

$Q(s,a) \to E[G_t | s, a]$

Bellman Eqn $V(s) \leftrightarrow V(s')$

$Q(s,a) \leftrightarrow Q(s', a')$

DP

- Value Iteration
- policy Iter

$\pi(a|s)$

Value Iter

$V(s) = 0$

for i

$$V(s) \leftarrow \max_a \left(R(s) + \gamma \sum_{s'} P(s'|s,a) V(s') \right)$$

π^* ↑ argmax( )

policy Iter

$V = 0$, $\pi \to$ rand

evaluate π, V

$$V_\pi = R(s) + \gamma \sum_{s'} P(s'|s,\pi) V(s')$$

$V(a)$

T, R

RL: T not known, R not known →

- model-based: $\hat{T}, \hat{R}$
- model-free: $V \to$ Passive RL; $Q \to$ active RL

Passive RL

MC: → Ignores Bellman → $V(s)$ → $\text{Avg}(R_i) \to$

TD: Moving average V → · Consistent, · converge faster

one episode

$s_1 \to s_2 \to s_3 \to s_4 \cdots \to s_T$

$$V^{k+1}(s) \leftarrow V^k(s) + \alpha [R_k + \gamma V_k(s') - V_k(s)]$$

π fixed

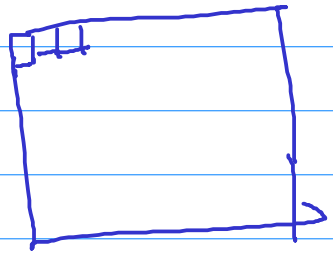
π → improve Q-function

on-policy : one π → ε → random, $1-\varepsilon$ → learned policy } explore vs exploit → ε-greedy
(MC, TD) (-) behaves suboptimal.

off-policy : two π → target policy → optimal policy
(MC, TD) behavior policy → ε-greedy (random)

on-policy TD control : Sarsa
off-policy TD control : Q-learning

State, Tables
V(s)
Q(s,a)



Approximation RL

function approximator → "regression"