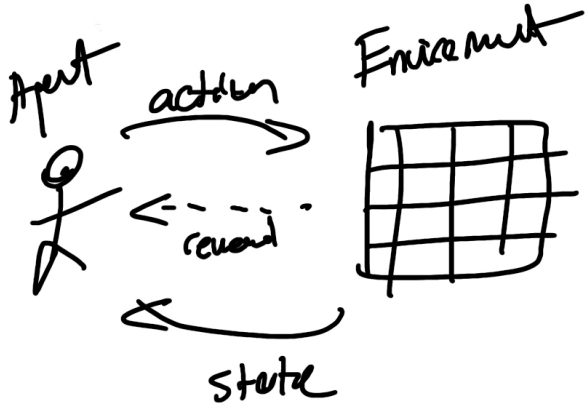
Reinforcement Learning Basics

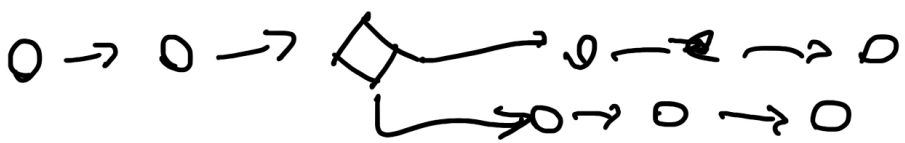


In the MDP world, the environment would be MDP, agent would be π

Behavior structures:

plan → fixed sequence of actions.
- by doing learning
- ↳ stochasticity

conditional plan → "if statements"



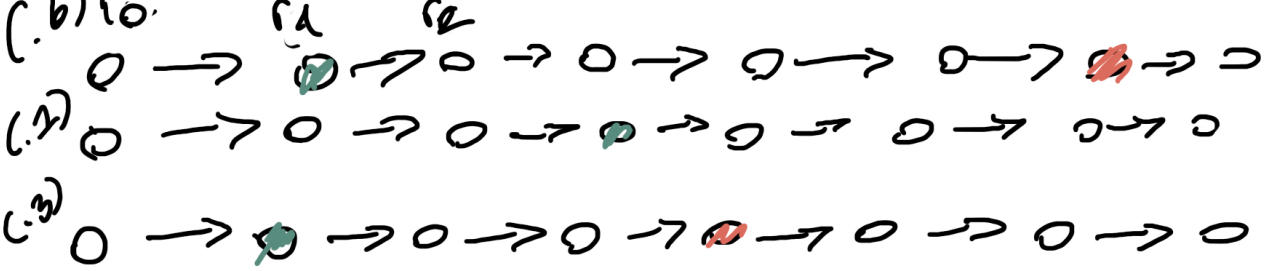
Stationary policy / universal plan
→ mapping from state to action
conditional but with the same if for every possible state

very large

evaluating a policy

Prob

(.6) r_0 r_1 r_2



state transitions to immediate rewards

Truncate according to horizon

Summarize sequences Return $\sum_{i=1}^{T} \gamma^i \cdot r_i$

Summarize over sequence

Real −0.2 Green +1

T: 5 γ: .8

Average expectation

Since horizon T: 5, we care abt 5.

$(.8 \cdot 1)$ for sequence 1. $(.8)$

$(.8^3 \cdot 1)$ for seq 2 → $(.512)$

$(.8^4 \cdot 1) + ((0.8)^4 \cdot -0.2))$ for seq 3 $(.0.718)$

Weighted average $(.8) \cdot (.6) + .512 \cdot .1 + .7808 \cdot 0.3$

→ $\boxed{.746624}$

evaluating a learner :

- value of returned policy
- computational complexity (time)
- sample complexity (time)
  ↳ how much data it needs.