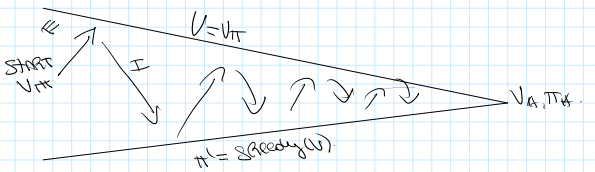
LECTURE 5 - MODEL FREE CONTROL

GOAL: Find V_* / π_* W/O knowledge of MDP dynamics

Recap: Generalised Policy Iteration (GPI)



- We want to make this FRAMEWORK MODEL-FREE.

PROBLEM #1:

- We CAN'T DO model-FREE policy EVALUATION

For V_π, because ACTING greedy would require knowledge of the MDP

$$\pi' = \text{greedy}(V_\pi) = \underset{a}{\text{Argmax}}\, Q_\pi(s,a) = \underset{a}{\text{argmax}}\, R_s^a + \sum_{s'} P(s'|s,a)\, V_\pi(s')$$

- Instead, we can DO POLICY EVALUATION on $q_\pi(s,a)$!

↳ GREEDY POLICY IMPROVEMENT OVER q IS MODEL-FREE

PROBLEM #2:

- Because we ARE DOING Q-POLICY EVALUATION WITH SAMPLING ⇒ IF our policy π is DETERMINISTIC we won't have samples FOR $Q_\pi(s,a)$ $\forall a \neq \pi(s)$.

⇒ POLICY IMPROVEMENT won't select BEST ACTIONS

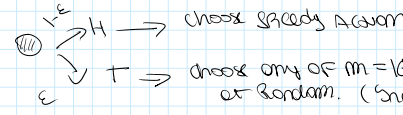
ε-Greedy EXPLORATION.

- IDEA: Modify Policy Improvement step to make sure we don't get stuck.

$$\pi'(a|s) = \begin{cases} \frac{\varepsilon}{m} + 1 - \varepsilon & \text{if } a = \underset{a^*}{\text{Argmax}}\, Q_\pi(s,a) \\ \frac{\varepsilon}{m} & \text{otherwise} \end{cases}$$

$$(m-1)\frac{\varepsilon}{m} + \frac{\varepsilon}{m} + 1 - \varepsilon = \varepsilon - \frac{\varepsilon}{m} + \frac{\varepsilon}{m} - \varepsilon + 1 = 1 \quad \blacksquare$$

Intuition: We FLIP A COIN, H comes up w/ prob $1-\varepsilon$ and T w/ prob ε.



Is ε-greedy a policy IMPROVEMENT?

According to last Lecture theorem, it is enough to verify $Q_\pi(s, \pi'(s)) \geq V_\pi(s)$ $\forall s \in S$.

Because ⇒ $V_{\pi'}(s) \geq V_\pi(s)$ $\forall s \in S$.

proof:

$$Q_\pi(s,\pi'(s)) = \sum_{a \in A} \pi'(a|s)\, q_\pi(s,a) =$$

$$\frac{\varepsilon}{m} \sum_{a \in A} q_\pi(s,a) + (1-\varepsilon) \max_a q_\pi(s,a) \leq$$

$$\frac{\varepsilon}{m} \sum_{a \in A} q_\pi(s,a) + (1-\varepsilon) \sum_{a \in A} \frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon}\, q_\pi(s,a)$$

Because $\sum_a \frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon} = 1$ Because $\max_a q_\pi(s,a) \leq \max_a q_\pi(s,a) \left[\sum_a \frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon}\right]$

$$= \max_a q_\pi(s,a) \left[\frac{1-\varepsilon}{1-\varepsilon}\right].$$

$$= \sum_a q_\pi(s,a) \left[\pi(a|s)\right] = q_\pi(s, \pi(s))$$

$$= V_\pi(s) \quad \blacksquare$$

The theorem does not apply in this case Because we only have an ε-greedy

MONTE-CARLO POLICY ITERATION / MONTE-CARLO CONTROL.

POLICY EVALUATION: MONTE-CARLO POLICY EVALUATION FOR Q.

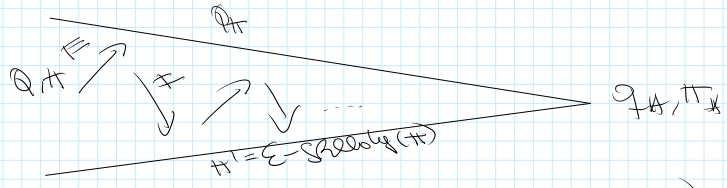
POLICY EVALUATION: MONTE-CARLO POLICY EVALUATION FOR Q.

POLICY IMPROVEMENT: ε-GREEDY POLICY IMPROVEMENT.

How can we make this more efficient?

- IMPROVE your POLICY AFTER EVERY EPISODE.

Apply in this case Because we only have an Estimate of Q_π. And then Assumes Q_π has been computed EXACTLY! ↳ we might not have MONOTONIC IMPROVEMENT.



GLIE (GREEDY IN LIMIT W/ INFINITE EXPLORATION)

COME UP W/ A SCHEDULE FOR EXPLORATION. ST

(1) EVERY STATE and ACTION IS VISITED ∞ TIMES
(2) the POLICY we OBTAIN IN LIMIT EVENTUALLY BECOMES GREEDY W/ RESPECT TO Q.

Balances EXPLORATION vs - EXPLOITATION.

ε-GREEDY IS GLIE IF $\varepsilon \to 0$ ($\varepsilon = 1/k$).

THEOREM: GLIE MONTE-CARLO CONTROL CONVERGES to q_*

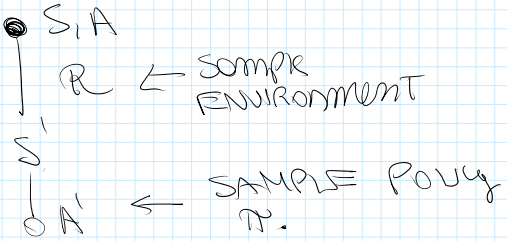
↳ FIRST ALGORITHM to FULLY SOLVE an MDP W/ UNKNOWN DYNAMICS.

TEMPORAL DIFFERENCE LEARNING CONTROL

2 ways → ON-POLICY.
↳ OFF-POLICY.

SARSA / ON-POLICY TD CONTROL.

POLICY EVALUATION STEP: EVALUATE $Q^\pi(S,A)$ USING TD LEARNING.



$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(R_{t+1} + \gamma Q(S',A') - Q(S,A)\right).$$

POLICY IMPROVEMENT: ε-GREEDY POLICY IMPROVEMENT.

↳ CAN BE DONE AFTER EVERY STEP.

OFF POLICY LEARNING.

- GOAL: EVALUATE $\pi(S|A)$ to COMPUTE V_π or $q_\pi(S,A)$ WHILE FOLLOWING / SAMPLING FROM $\mu(S|A)$.

- π: TARGET POLICY - μ: BEHAVIOURAL POLICY

Q-LEARNING (EVALUATE q_π)

- Goal: OFF-POLICY Learning FOR $q_\pi(S,A)$.
- Next Action is chosen using behaviour policy $A_{t+1} \sim \mu(S)$
- But we consider Alternative Successor Action $A' \sim \pi(S)$ sampled from target policy.
- Update $Q(S,A)$ towards value of A'.

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma Q(S',A') - Q(S,A)\right]$$

Special case: $\mu = \varepsilon\text{-greedy}(Q)$ ← BEHAVIOURAL POLICY

$\pi = \text{greedy}(Q)$ ← TARGET POLICY

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_{A'} Q(S',A') - Q(S,A)\right]$$

- In this case, Q directly APPROXIMATES Q^* independent of the policy we FOLLOW ⇒ OFF-POLICY (Follow any but target IS π^*).
- FOR convergence we Require μ to keep on VISITING (S,A) (ε-greedy is Enough).

NOTE: WE COULD USE A completely Random μ, And the Algorithm would still converge to q^*.

THEOREM: Q-Learning Converges to q^*.

QUESTION: IS IT TRUE THIS UPDATE DIRECTLY EVALUATES Q^* USING TD-Learning BECAUSE $\pi^* = \text{greedy}(Q^*)$?

CAN WE THINK OF THE ALG. AS SAMPLING FROM Q^* the Bellman Optimality EQUATION??

π^* satisfies $\pi^* = \text{Argmax}\ Q_{\pi^*}(S,A)$. therefore $Q(S,A)$ EVALUATION USING TD-Learning would look somewhat like this:

$$\begin{aligned} Q(S,A) &= Q(S,A) + \alpha\left[R + \gamma Q(S', \pi_A(S')) - Q(S,A)\right] \\ &= Q(S,A) + \alpha\left[R + \gamma Q(S', \text{Argmax}\ Q_{\pi^*}(S,a')) - Q(S,A)\right] \\ &= Q(S,A) + \alpha\left[R + \gamma Q(S', \underset{a'}{\text{argmax}}\ Q(S,a')) - Q(S,A)\right] \\ &= Q(S,A) + \alpha\left[R + \gamma Q(S', \max_{a'} Q(S,a')) - Q(S,A)\right] \end{aligned}$$

PROBLEM: Can we Follow π^*? Depends on Q_{π^*}. It has been "Solved" Randomly.

So Maybe USE OFF-POLICY. Follow ANY Exploratory policy that VISITS ALL (S,A) FOREVER.