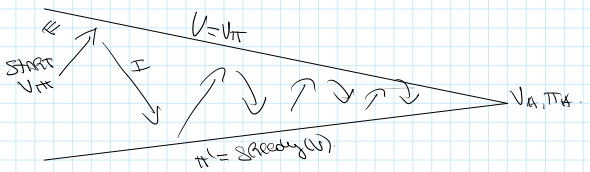
David Silver RL Lecture 5 - Model Free Control

Tuesday, March 16, 2021 | 3:45 PM

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

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

- 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 estimates for $Q_\pi(s,a)\ \forall a \neq \pi(s)$.

⇒ POLICY IMPROVEMENT won't select BEST ACTIONS

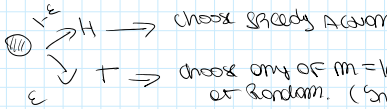
ε-Greedy EXPLORATION.

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

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

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

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 class thm, 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:

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

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

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

Because $\sum_a \frac{\pi(a|s) - \varepsilon/m}{1-\varepsilon} = 1$

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

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

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

$= \sum_a q_\pi(s,a) \lfloor \pi(a|s) \rfloor = q_\pi(s, \pi(s)) = v_\pi(s)$ ∎

MONTE-CARLO POLICY ITERATION / MONTE-CARLO CONTROL.

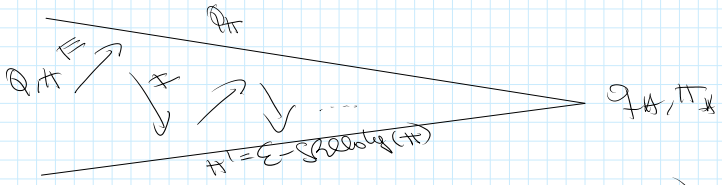
POLICY EVALUATION: MONTE-CARLO POLICY EVALUATION FOR Q.

POLICY IMPROVEMENT: ε-GREEDY POLICY IMPROVEMENT.

How can we make this more efficient?

- IMPROVE your policy AFTER EVERY EPISODE.

The theorem does not 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 GPI EVENTUALLY BECOMES GREEDY W/ RESPECT TO Q.

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

Balances EXPLORATION vs. EXPLOITATION.

THEOREM: GLIE MONTE-CARLO CONTROL CONVERGES TO q_*

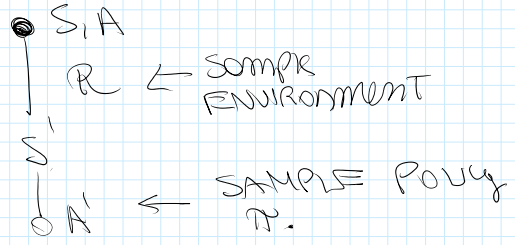
↳ FIRST ALGORITHM TO FULLY SOLVE AN MDP W/ UNKNOWN DYNAMICS.

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 $\Rightarrow$ 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 $\pi^* = \underset{a}{\text{argmax}}\, Q_{\pi^*}(s,a)$. therefore, $Q(s,a)$ EVALUATION USING TD-Learning would look somewhat like this:

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

PROBLEM: Can we follow π^* ? depends on Q_{π^*}.

↳ has been invalidated randomly.

so maybe use OFF-policy. Follow any exploratory policy that visits all (s,a) forever.