REINFORCEMENT LEARNING

POLICY: STATE → ACTION MAPPING

REWARD FUNCTION: AGENT STATE (ENV. STATE, ACTION) → REWARD MAPPING, SHORT TERM

VALUE FUNCTION: FOR A STATE, AMOUNT OF REWARD EXPECTED TO ACCUMULATE OVER FUTURE, LONG TERM PREDICTION

MODEL: MIMICS ENVIRONMENT BEHAVIOR (STATE, ACTION) → NEXT STATE, REWARD PREDICTION

SIMPLE TD RULE: $V(s) \leftarrow V(s) + \alpha\left[V(s') - V(s)\right]$, α STEPSIZE IS PROGRESSIVELY ADJUSTED TO AUGMENT $V(s)$ AFTER GREEDY MOVE.

- EVALUATIVE VS INSTRUCTIVE FEEDBACK → NO WAY OF DIRECTLY KNOWING 'OPTIMAL ACTION' AND DO INFERENCE ABOUT CORRECT ACTIONS
- EXPLORATION VS EXPLOITATION, HOW TO BALANCE.

ACTION-VALUE METHODS:

SAMPLE-AVERAGE: $\hat{Q}_t(a) = \frac{\sum_{T}^{n} R_{Ka}}{K_a}$ SAMPLE AVG OF REWARDS WHEN ACTION SELECTED → $Q_T(a^*) = \max_a \hat{Q}_t(a)$

ε-GREEDY: $Q_T(a^*) = (1-\varepsilon)\max_a \hat{Q}_t(a) + \varepsilon[\text{ANY } a]$ (UNIFORM), ε-GREEDY EXPLORES.

- MORE VARIANCE IN REWARDS (THEY ARE R.V) → MORE EXPLORATION IS GOOD
- HETEROSKEDASTIC REWARDS → EXPLORATION PAYS OFF.

SOFTMAX SELECTION: UNIFORM SELECTION FOR NONGREEDY MOVES IS BAD, $P(a|\cdot) = \frac{e^{Q_T(a)/\tau}}{\sum_N e^{Q_T(i)/\tau}}$ • τ TEMPERATURE. $\tau \to 0$ FULLY GREEDY; $\tau \to \infty$ FULLY EQUIPROBABLE

BINARY BANDIT TASK: 2 POSSIBLE ACTIONS, EACH HAS CLEAR SUCCESS/FAILURE PROBABILITY. NOT SUM TO 1, INFER IF FAILURE → OTHER WAS CORRECT

- LRP ALGO: GRADUALLY INCREASES PROB OF ACTION BEING BEST, ALWAYS UPDATES
- LRI ALGO: LRP BUT ONLY UPDATES WHEN SUCCESS

FOR NONSTATIONARY PROBLEMS: DO NOT WEIGHT SAME. DO WEIGHTED/EXPONENTIAL MOVING AVG $Q_n = (1-\alpha)Q_0 + \sum_1^n \alpha(1-\alpha)^{n-1} R_i$. SUM OF WEIGHTS IS 1

- α MAY BE VARIED FROM STEP TO STEP

HOW TO SET INITIAL ESTIMATES $Q_0(a)$?

THE MORE OPTIMISTIC → THE MORE WE ARE FAVORING EXPLORATION IN THE BEGINNING

REINFORCEMENT COMPARISON:

- ASSESS REWARD BY COMPARISON W/ REFERENCE REWARD, IE AVG OF PREVIOUSLY RECEIVED REWARDS, SOMETIMES MORE EFFECTIVE THAN ACTION-VALUE. PRECURSOR TO ACTOR-CRITIC
- HAVE PREFERENCE VALUES FOR ACTIONS $\pi_t(a) = P\{a_t = a\} = \frac{e^{P_t(a)}}{\sum_n e^{P_t(i)}}$

REFERENCE UPDATE: $\bar{R}_{T+1} = \bar{R}_T + \alpha\left[R_t - \bar{R}_T\right]$ PREFERENCE UPDATES $P_{T+1}(a) = P_T(a_T) + \beta\left[R_t - \bar{R}_T\right]$

PURSUIT:

HAVE BOTH A-V ESTIMATES AND ACTION PREFERENCES • PREFERENCE 'PURSUES' GREEDY A-V ESTIMATED ACTION

- a^*_{T+1} GREEDY ACTION,

$$\begin{cases} \pi_{T+1}(a^*_{T+1}) = \pi_T(a^*_{T+1}) + \beta\left[1 - \pi_T(a^*_{T+1})\right] & \text{GREEDY ACTION UPDATED INCREMENTS} \\ \pi_{T+1}(a) = \pi_T(a) + \beta\left[0 - \pi_T(a)\right] & a \neq a^*_{T+1} \text{ OTHER ACTIONS DECREMENTED.} \end{cases}$$

- ACTION-VALUES $Q_{T+1}(a)$ UPDATED IE VIA SAMPLE AVG

=

SSOCIATIVE TASK: IE MULTIPLE N-ARMED BANDIT TASK. TASK CHANGES RANDOMLY AT EACH PLAY. WE COULD TREAT IT AS A SINGLE NONSTATIONARY TASK BUT NOT VERY WELL. IF WE ARE GIVEN INFO ON WHICH TASK AT EACH PLAY, ASSOCIATIVE TASK. SEARCH FOR BEST ACTIONS AS WELL AS ASSOCIATING THEM WITH SITUATION/TASK