TD - LEARNING

- NO ENV MODEL, LEARN ON ESTIMATES W/O WAITING FOR FINAL STATE. 'MC + DP' → BOOTSTRAP(AND SAMPLING) | TD AND MC ARE **SAMPLE BACKUPS** (VS FULL BACKUPS) THEY LOOK AT SAMPLE SUCCESSOR.
- **TD(0)**: $V(S_t) \leftarrow V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right]$, UPDATES AT NEXT TIME-STEP
- ERROR PROPORTIONAL TO CHANGE OVER TIME OF PREDICTIONS.
- PROVED TO CONVERGE W/ STATIC SMALL AND STOCHASTICALLY DECREASING α
- **ADVANTAGES**
  - NO MODEL
  - NATURALLY ONLINE, FULLY INCREMENTAL
  - LESS PRONE TO ISSUES DUE TO DISCOUNTING | TRUNCATING
  - EMPIRICALLY FASTER CONVERGENCE THAN MC
  - TD APPROXIMATES MLE WITH N MEMORY, VS N^2 OF CLASSIC FORM!
- **ON BATCH UPDATES**
  - WHEN UPDATES ONLY AFTER EVERY NEW BATCH.
  - TD(0) CONVERGES DETERMINISTICALLY TO MLE. CERTAINTY-EQUIVALENCE ESTIMATE
  - α-MC CONVERGE TO MINIMIZED MSE ON TRAINING DATA
  - **NONBATCH**: NO CONVERGENCE FOR EITHER, BUT STEPS IN THAT DIRECTION.

SARSA

- ON-POLICY TD CONTROL • GENERAL GPI FRAMEWORK • S-A PAIR TO S-A PAIR LEARNING. WE ESTIMATE $Q_\pi(S,A)$ FOR π ON ALL (S,A)
- $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma Q(S',A') - Q(S,A)\right]$ • $Q_\pi \rightleftarrows \pi$ • THOUGHT TO CONVERGE IF (S,A) VISITED ∞ TIMES AND π CONVERGES IN THE LIMIT
- π CAN BE ε-GREEDY, ε-SOFT, ETC...

Q - LEARNING

- OFF-POLICY TD CONTROL • $Q(S,A): Q(S,A) + \alpha\left[R_{t+1} + \gamma \max_a Q(S',a) - Q(S,A)\right]$
- Q DIRECTLY APPROXES q_* • CONVERGENCE: ALL PAIRS CONTINUE TO BE UPDATED • MAXIMIZES OVER ALL POSSIBLE ACTIONS a AT NEXT STATE
- IN ONLINE LEARNING SARSA PREFERRABLE DUE TO CONSIDERING ACTION SELECTION • ε DECREASING OVER TIME → ON CONVERGENCE

X AFTERSTATE VALUE FUNCTION: MODEL STATE AFTER THE ACTION. USEFUL IN IE GAMES, QUEUES. BETTER BECAUSE REDUCE SEARCH SPACES. GPI ON FOR THESE.

ACTOR - CRITIC METHODS

- EXTENSION OF REINFORCEMENT COMPARISON TO TD AND FULL RL PROBLEM: ON-POLICY • MINIMAL COMPUTATION TO SELECT ACTIONS; CAN LEARN EXPLICITLY STOCHASTIC ~~PARTS~~ POLICIES
- **ACTOR**: SEPARATE STRUCTURE REPRESENTING POLICY INDEPENDENT OF VALUE FUNCTION, IMPROVES POLICY • GENERALIZED POLICY ITERATION
- **CRITIC**: ESTIMATED VALUE FUNCTION → CRITIQUE IS TD ERROR, DRIVES LEARNING IN BOTH A AND C; EVALUATES CUR POLICY
  - STATE-VALUE USUALLY

$$-\ \delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \xrightarrow{C} P(S,A) \xleftarrow{A} P(S,A) + \beta\delta_t,\quad P(S,A) \leftarrow P(S,A) + \beta\delta_t\left[1 - \pi_t(S,A)\right] \text{ ALTERNATIVELY}$$

R - LEARNING

- OFF-POLICY CONTROL • NO DISCOUNTED EXPERIENCE, NO EPISODES • **MAXIMUM REWARD PER TIMESTEP**
- VALUE FCN FOR π DEFINED RELATIVE TO AVG EXPECTED REWARD PER TIMESTEP UNDER π • ASSUMES PROCESS ERGODIC AND Q^π INDEPENDENT OF STARTING STATE
- $Q^\pi = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E_\pi\{R_t\}$ • SHORT-TERM TRANSIENT DETERMINES VALUE STATE $\tilde{V}^\pi(s) = \sum_{t=1}^{\infty} E_\pi\{R_{t+1} - Q^\pi \mid s_t = s\}$, $\tilde{Q}^\pi(s,a) = \sum E_\pi\{R_{t+1} - Q^\pi \mid s_t = s, a_t = a\}$

→ **ACTOR**: BEHAVIOR POLICY, MAY BE ~~DETERMINISTIC~~ STOCHASTIC, ACTS ON GREEDY WRT CRITIC, OR GRADIENT ASCENT ON SOME FIXED PARAMETRIC POLICY

CRITIC: TARGET POLICY, ~~GREEDY~~ IMPROVES, MAY BE DETERMINISTIC. ε-GREEDY WRT CURRENT A-V FUNCTION ~~IS BEHAVIOR POLICY~~. SARSA/Q/... FOR LEARNING

DOUBLE - Q LEARNING

- VANILLA Q OVERESTIMATES VALUES BECAUSE USES SAME VALUES TO SELECT AND EVALUATE ACTIONS
$$\begin{cases} \theta_{t+1} = \theta_t + \alpha\left(Y_t^Q - Q(S_t,A_t,\theta_t)\right) \nabla_{\theta_t} Q(S_t,A_t,\theta_t) \\ Y_t^Q = R_{t+1} + \gamma \max_a Q(S_{t+1},a,\theta_t) \end{cases}$$
- IDEA: DECOUPLE SELECTION FROM EVALUATION → HAVE 2 VALUE FUNCTIONS θ, θ'
  ASSIGN EXPERIENCE RANDOMLY. FOR EACH UPDATE USE ONE FOR GREEDY POLICY, OTHER FOR VALUE

$$Y^Q = R_{t+1} + \gamma Q\left(S_{t+1}, \underbrace{\operatorname{argmax}_a Q(S_{t+1},a,\theta_t)}_{\text{SELECTION}}, \theta_t\right) \rightarrow Y^{DQ} = R_{t+1} + \gamma Q\left(S_{t+1}, \underbrace{\operatorname{argmax}_a Q(S_{t+1},a,\theta_t)}_{\text{SELECTION}}, \underbrace{\theta'_t}_{\text{EVALUATION}}\right) \quad \bullet\ \theta \text{ IS CURRENT VALUE}$$

- IN DQN → **DOUBLE DQN** WE HAVE TARGET NET ON CENTRAL NODE → USE THAT AS θ'_t! NO NEED TO INTRODUCE OTHER NETWORKS FOR EVAL OF GREEDY POLICY