ELIGIBILITY TRACES

- CAN COMBINE WITH MOST TD-METHODS
- **FORWARD VIEW:** BRIDGE FROM TD TO MC METHODS. INTERMEDIATE SPECTRUM OF METHODS
- **BACKWARD VIEW:** TEMPORAL RECORD OF EVENT OCCURRENCE (STATE VISIT, ACTION TAKING), "MARKS THINGS FOR UNDERGOING LEARNING CHANGES"
  ONLY MARKEDS ARE ASSIGNED REWARD/PUNISHMENT

N-STEP TD

TD = ONE-STEP AHEAD
MC = ALL STEPS AHEAD TO END, ∞ → **N-STEP:** N STEPS AHEAD IE 2SA: $R_{t+1} + \gamma R_{t+2} + \gamma^2 V_t(S_{t+2})$ $\quad G_t^{t+n}(c) = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{N-1} R_H + \gamma^N C$

N-STEP BACKUP: $\Delta(S_t) = \alpha\left[G_t^{t+n}(V_t(S_{t+n})) - V_t(S_t)\right]$, **UPDATE:** $V_{t+1}(s) = V_t(s) + \Delta_t(s)$; ONLINE | OFFLINE $V_{t+1}(s) = V_t(s)$

$V_T(s) = V_{T-1}(s) + \sum_0^{T-1} \Delta_T(s)$

ERROR-REDUCTION PROPERTY: MAX N, LOWER ERROR → CONVERGENCE

- COOL, BUT INCONVENIENT TO IMPLEMENT BECAUSE WAITING.

FWD VIEW

- **IDEA:** LET'S AVERAGE DIFFERENT N-STEP RETURNS, OR AS LONG AS WEIGHT SUM TO 1. STILL ERROR REDUCTION ← **COMPLEX BACKUPS**
- **λ RETURN:** $L_T = (1-\lambda)\sum_N^{\infty} \lambda^{N-1} G_t^{t+N}(V_t(S_{t+n}))$, CONTAINS ALL N-STEPS BACKUPS, WITH EXP-DECAYING WEIGHTS. AFTER TERMINAL ALL RETS ARE G_t

  $\lambda = 1$ SAME AS CONSTANT α MC RETURN

  $\lambda = 0$ SAME AS TD(0) METHOD

  - CAN UPDATE ONLINE OR OFFLINE
  - FWD VIEW BECAUSE NEVER GO BACK TO PREV STATES

  ALGO: $\Delta_t(s_t) = \alpha[L_t - V_t(S_t)]$

BWD VIEW

- ELIGIBILITY TRACE FOR STATES $\begin{cases} E_T(s) = \gamma\lambda E_{T-1}(s) & s \neq S_T \\ E_T(s) = \gamma\lambda E_{T-1}(s) + 1 & s = S_T \end{cases}$ λ = TRACE-DECAY PARAM → ACCUMULATING TRACE

TD(λ) UPDATES:

$\begin{cases} \delta_T = R_{T+1} + \gamma V_t(S_{t+1}) - V_t(S_T) \\ \Delta V_t(s) = \alpha \delta_T E_T(s) \quad \text{FOR ALL } s \in S \end{cases}$

- UPDATES PROPORTIONAL TO RECENTLY VISITED STATES, ONLINE OR OFFLINE
- $\begin{cases} \lambda = 0 & \text{TD(0)} \\ \lambda = 1 & \text{CONSTANT } \alpha \text{ MC} \end{cases}$
- $\lambda = 1$, $\gamma = 1$ → NO DISCOUNT, NO DECAY EPISODIC MC
- BWD BECAUSE PROPAGATES OR UPDATES BACK IN TIME
- MORE GENERAL FORMULATION OF MC

REPLACEMENT TRACE: $E_T(S_T) = 1$ **DUTCH TRACE:** $E_T(S_T) = (1-\alpha)\gamma\lambda E_{T-1}(S_T) + 1$ → GENERALLY BEST FOR ON-LINE ALGO APPROXIMATING λ-RETURN.

TOTAL EXACT STEP-BY-STEP EQUIVALENCE OF λ BASED FWD AND BWD IMPLEMENTATIONS.

SARSA(λ)

- TRACES FOR S-A PAIRS, $Q_{T+1}(s,a) = Q_T(s,a) + \alpha\delta_T E_T(s,a)$, $\delta_T = R_{T+1} + \gamma Q_T(S_{T+1}, A_{T+1}) - Q(S_T, A_T)$
- POLICY IMPROVEMENT AS USUAL, EG ϵ-GREEDY
- ANY VARIANT

Q(λ)

- LOOKAHEADS ONLY UNTIL FIRST EXPLORATORY ACTION BECAUSE AFTER NO MORE RELATION TO GREEDY POLICY. N-STEP RETURNS ONLY UNTIL THEN.
- TRACES SET TO 0 WHEN EXPLORATORY NONGREEDY ACTION IS TAKEN. ALWAYS SET CUR (S,A) TRACE TO 1 AFTER ACTION, EVEN WHEN EXPLORATORY
- $Q_{T+1}(s,a) = Q_T(s,a) + \alpha\delta_T E_T(s,a)$ $\quad \delta_T = R_{T+1} + \gamma \max_{a'} Q_T(S_{T+1}, a') - Q_T(S_T, A_T)$
- MAY BE ONLY LITTLE FASTER THAN 1-STEP Q LEARNING
- BOOTSTRAPS EVEN WITH $\lambda = 1$ DUE TO TRACE CUTS AND USING VALUE ESTIMATES VS ACTUAL REWARDS, NOT OPTIMAL → **DECOUPLING NEEDED** W/ IMPORTANCE SAMPLING?

TRACES FOR ACTOR-CRITIC

- USE STATE TRACES FOR CRITIC, AND ACTION-STATE TRACES FOR ACTOR, SEPARATE SETS. USE SUITABLE ALGOS FOR THE UPDATES.

VARIABLE-λ

IDEA: $\lambda \to \lambda_T$, VARIES FROM STEP TO STEP. IE AS FCN OF STATE $\lambda_T = \lambda(S_T)$, DEPENDING ON CERTAINTY OF ESTIMATE.

IMPLEMENTATION REMARKS

- O(N) ON SEQUENTIAL PIPELINES BECAUSE TRACES ARE MOSTLY 0
- ONLY A FEW TIMES MORE TIME THAN NO TRACES.
- IF ANN → ONLY DOUBLE
- MAKE MORE SENSE IN ONLINE THAN IN OFFLINE SETTINGS