CHAPTER 7 - n-STEP BOOTSTRAPPING

7.1 n-STEP TD PREDICTION

BASIC IDEA: TD METHODS LOOK ONE STEP AHEAD, MC LOOKS ALL THE WAY TO THE END, LET'S BRIDGE THE GAP.

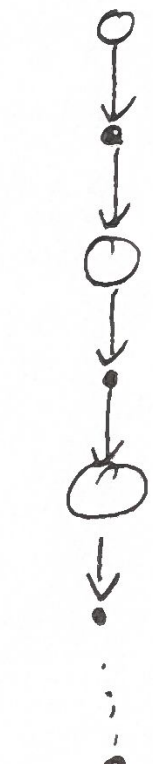
IN TERMS OF BACKUP DIAGRAMS:

1-STEP TD

2-STEP TD

n-STEP TD

∞-STEP TD / MONTE CARLO



IN TERMS OF THE CORRECTIONS WE USE (CALLED TARGET)

MONTE CARLO: THE TARGET IS THE RETURN

$$G_t \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T$$

ONE STEP UPDATES:

$$G_{t:t+1} \triangleq R_{t+1} + \gamma V_t(S_{t+1})$$

($V_t : \mathcal{S} \to \mathbb{R}$ IS THE ESTIMATE OF v_π AT TIME t)

NEW NOTATION →

↑ REWARD

↑ ESTIMATE OF VALUE

TWO-STEP UPDATES:

$$G_{t:t+2} \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 V_{t+1}(S_{t+2})$$

n-STEP UPDATE:

$$G_{t:t+n} \overset{\Delta}{=} \begin{cases} R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^{n} V_{t+n-1}(S_{t+n}), & 0 \le t < T-n \\ G_t, & \text{if } t+n \ge T \end{cases}$$

OBSERVING THAT TO UPDATE THE VALUE OF A STATE, WE NEED TO WAIT UNTIL TIME $t+n$. THEREFORE, THE UPDATE IS:

$$V_{t+n}(S_t) \overset{\Delta}{=} V_{t+n-1}(S_t) + \alpha \left[G_{t:t+n} - V_{t+n-1}(S_t) \right],$$

ALSO, $V_{t+n}(s) = V_{t+n-1}(s) \quad \forall s \neq S_t$.

n-STEP TD FOR ESTIMATING $V \simeq v_\pi$

INPUT: POLICY π ALGORITHM PARAMETERS: $\begin{cases} \text{STEP SIZE } \alpha \in (0,1], \\ n \in \mathbb{N}^* \end{cases}$

INITIALIZE $V(s)$ ARBITRARILY, $\forall s \in \mathcal{S}$

ALL STORES AND ACCESS OPERATIONS (FOR S_t, R_t) CAN TAKE INDEX $\bmod(n+1)$

LOOP FOR EACH EPISODE:

  INITIALIZE AND STORE $S_0 \neq$ TERMINAL

  $T \leftarrow \infty$

  LOOP FOR $t = 0, 1, \dots$

    IF $t < T$, THEN:

      TAKE ACTION ACCORDING TO $\pi(\cdot|S_t)$

      OBSERVE AND STORE NEXT REWARD R_{t+1} AND STATE S_{t+1}

      IF S_{t+1} TERMINAL, $T \leftarrow t+1$

    $\tau \leftarrow t-n+1$ (WE WILL UPDATE ESTIMATE FOR STATE S_τ)

    IF $\tau \ge 0$:

      $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$

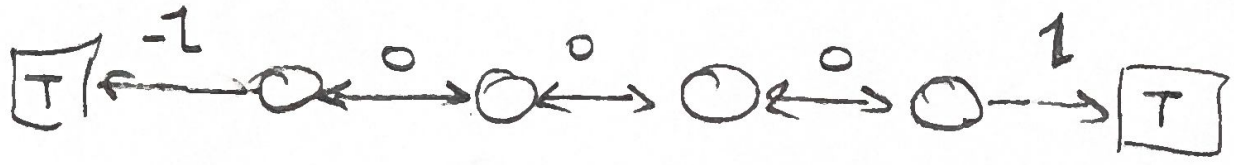
      IF $\tau+n < T$, THEN $G \leftarrow G + \gamma^n V(S_{\tau+n})$

      $V(S_\tau) \leftarrow V(S_\tau) + \alpha \left[G - V(S_\tau) \right]$

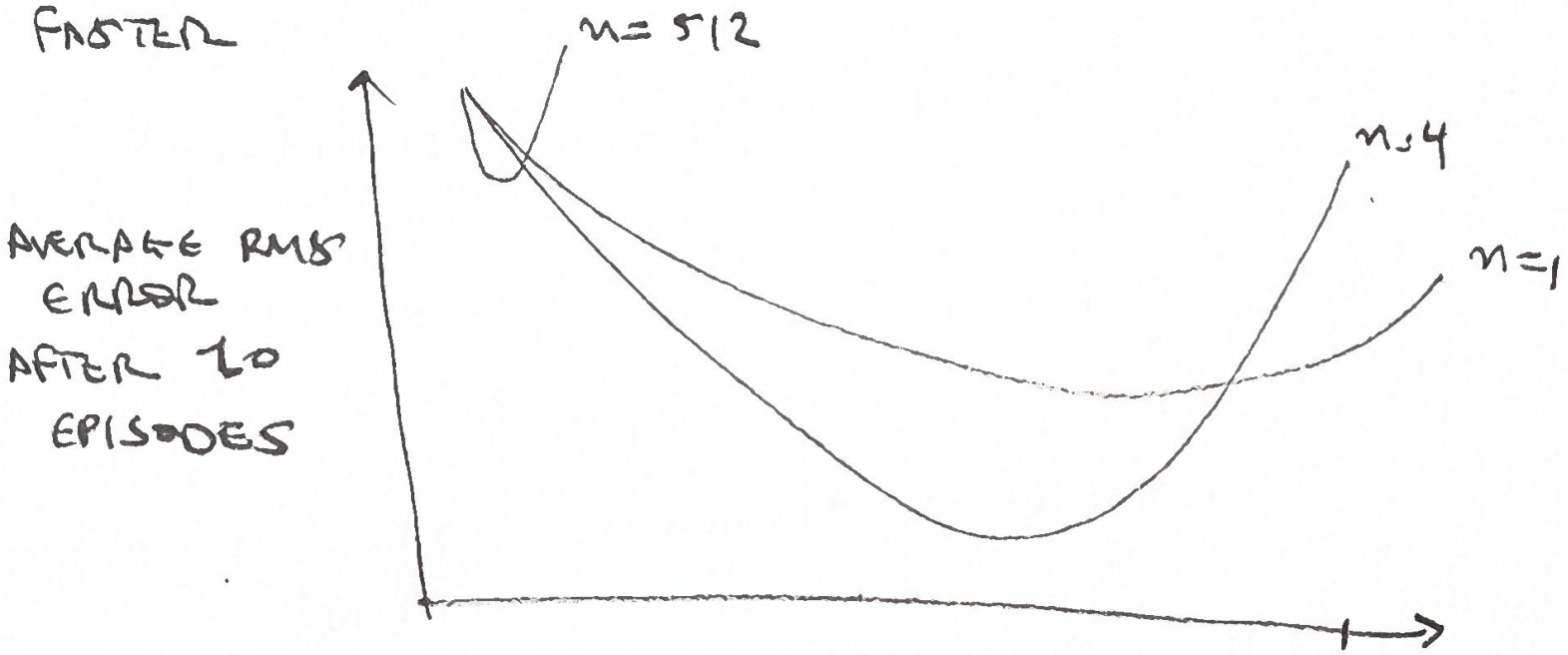
  UNTIL $\tau = T-1$

THIS ALGORITHM CAN BE SHOWN TO CONVERGE.

SO, WHY IS THIS BETTER? CONSIDER EXAMPLE 6.2 (RANDOM WALK) WITH 19 STATES. THIS IS NOW EXAMPLE 7.1



WITH 1-STEP TD, AFTER THE FIRST EPISODE, ONLY ONE STATE CHANGES VALUE ESTIMATE. WITH n-STEP, n STATES WILL CHANGE THEIR VALUE ESTIMATE. SO THIS IS FASTER



7.2 n-STEP SARSA

NOW WE MOVE FROM PREDICTION TO CONTROL. WE NEED TO FOCUS ON ACTION VALUES. WE DEFINE: (for $n \geq 1$)

$$G_{t:t+n} \triangleq \begin{cases} R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^{n} Q_{t+n-1}(S_{t+n}, A_{t+n}) & 0 \leq t < T-n \\ G_t = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{T-t-1} R_T & t+n \geq T \end{cases}$$

SO THE UPDATE BECOMES

$$Q_{t+n}(S_t, A_t) \triangleq Q_{t+n-1}(S_t, A_t) + \alpha \left[G_{t:t+n} - Q_{t+n-1}(S_t, A_t) \right]$$

AND

$$Q_{t+n}(s, a) = Q_{t+n-1}(s, a) \quad \forall\, (s, a) \neq (S_t, A_t)$$

THIS IS n-STEP SARSA

n-STEP SARSA FOR ESTIMATING $Q \simeq q^*$ (or q_π)

INITIALIZE $Q(s,a)$ ARBITRARILY, FOR ALL $s \in S, a \in A$

INITIALIZE π TO BE ϵ-GREEDY WITH RESPECT TO Q, (or A FIXED POLICY π)

ALGORITHM PARAMETERS: $\alpha \in (0,1]$, SMALL $\epsilon > 0$, $n \in \mathbb{N}^*$

(ALL STORE AND ACCESS OPERATIONS (FOR S_t, R_t, A_t) CAN TAKE THEIR INDEX mod $n+1$)

LOOP FOR EACH EPISODE

- INITIALIZE AND STORE $S_0 \neq$ TERMINAL
- SELECT AND STORE AN ACTION $A_0 \sim \pi(\cdot|S_0)$
- $T \leftarrow \infty$
- LOOP FOR $t = 0, 1, 2, \ldots$
  - IF $t < T$, THEN:
    - TAKE ACTION A_t
    - OBSERVE AND STORE THE NEXT REWARD AS R_{t+1} AND THE NEXT STATE AS S_{t+1}
    - IF S_{t+1} TERMINAL, $T \leftarrow t+1$
    - ELSE: SELECT AND STORE AN ACTION $A_{t+1} \sim \pi(\cdot|S_{t+1})$
  - $\tau \leftarrow t - n + 1$ (# τ IS THE TIME WHOSE ESTIMATE IS UPDATED)
  - IF $\tau \geq 0$:
    - $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$
    - IF $\tau + n < T$, THEN $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
    - $Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha [G - Q(S_\tau, A_\tau)]$
    - IF π IS BEING LEARNED, THEN ENSURE IT IS ϵ-GREEDY WRT Q
- UNTIL $\tau = T - 1$

BACKUP DIAGRAMS

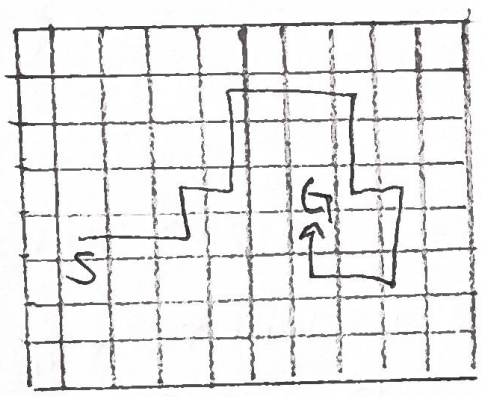
1-STEP SARSA (SARSA(0))

2-STEP SARSA

n-STEP SARSA

∞-STEP SARSA (MONTE CARLO)

SO WHY IS IT BETTER? FIGURE 7.4



CONSIDER NEW EPISODE IN GRID WORLD. ASSUME ALL $q=0$ INITIALLY.

1) 1-STEP SARSA STRENGTHENS THE LAST ACTION

2) 10-STEP SARSA STRENGTHENS THE LAST 10 ACTIONS

MODIFICATION FOR EXPECTED SARSA

$$G_{t:t+n} \doteq \begin{cases} R_{t+1} + \dots + \gamma^{n-1} R_{t+n} + \gamma^{n} \bar{V}_{t+n-1}(S_{t+n}), & t+n < T \\ G_t, & t+n \geq T \end{cases}$$

WHERE $\bar{V}_t(s) \doteq \sum_a \pi(a|s)\, q_t(s,a)$ ← EXPECTED APPROXIMATE VALUE

7.3 n-STEP OFF-POLICY LEARNING

RECALL OFF-POLICY METHODS: TWO POLICIES
- b: BEHAVIOR POLICY
- π: TARGET π FOR WHICH WE LEARN (I.E. GREEDY WITH RESPECT TO CURRENT Q)

TO MODIFY THE EVALUATION OF THE STATE VALUE ESTIMATE, WE USE, IN THE CASE OF n-STEP TD:

$$V_{t+n}(S_t) \triangleq V_{t+n-1}(S_t) + \alpha \rho_{t:t+n-1}\left[G_{t:t+n} - V_{t+n-1}(S_t)\right]$$

WHERE

$$\rho_{t:h} \triangleq \prod_{k=t}^{\min(h,T-1)} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

(THIS LEARNS π)

FOR EXAMPLE:
1) IF ANY OF A_k CANNOT BE TAKEN BY π, THEN $\rho_{t:h}=0$ AND WE DO NOT LEARN
2) IF A COMBINATION OF ACTIONS IS MUCH MORE COMMON IN π, THEN $\rho \gg 1$ AND WE LEARN A LOT ABOUT SOMETHING WE WOULD SEE WITH π MUCH MORE OFTEN
3) IF $\pi = b$, THEN THE COEFFICIENT $\rho = 1$ ALWAYS

IN THE CASE OF n-STEP SARSA, WE HAVE

$$Q_{t+n}(S_t, A_t) \triangleq Q_{t+n-1}(S_t, A_t) + \alpha \rho_{t+1:t+n}\left[G_{t:t+n} - Q_{t+n-1}(S_t, A_t)\right]$$

↑ WINDOW IS NOW DIFFERENT BECAUSE WE HAVE TAKEN ACTION.

OFF-POLICY n-STEP SARSA FOR ESTIMATING $Q \approx q_*$ OR q_π

INPUT: ARBITRARY BEHAVIOR POLICY b SUCH THAT $b(a|s) > 0$, $\forall a \in \mathcal{A}, s \in \mathcal{S}$

INITIALIZE $Q(s,a)$ ARBITRARILY $\forall s \in \mathcal{S}, a \in \mathcal{A}$

INITIALIZE π TO BE GREEDY WITH RESPECT TO Q, OR AS A FIXED GIVEN POLICY

ALGORITHM PARAMETERS: STEP SIZE $\alpha \in (0,1]$, $n \in \mathbb{N}^*$

ALL STORE / ACCESS OPERATIONS (FOR S_t, R_t, A_t) CAN TAKE THEIR INDEX MOD $n+1$

LOOP FOR EACH EPISODE:

  INITIALIZE AND STORE $S_0 \neq$ TERMINAL

  SELECT AND STORE AN ACTION $A_0 \sim b(\cdot|S_0)$

  $T \leftarrow \infty$

  LOOP FOR $t = 0, 1, 2, \ldots$

    IF $t < T$ THEN:

      TAKE ACTION A_t

      OBSERVE AND STORE THE NEXT REWARD AS R_{t+1} AND THE NEXT STATE AS S_{t+1}

      IF S_{t+1} IS TERMINAL, THEN

        $T \leftarrow t+1$

      ELSE

        SELECT AND STORE AN ACTION $A_{t+1} \sim b(\cdot|S_{t+1})$

    $\tau \leftarrow t - n + 1$ (τ IS THE TIME WHOSE ESTIMATE IS BEING UPDATED)

    IF $\tau \geq 0$:

      $\rho \leftarrow \prod_{i=\tau+1}^{\min(\tau+n, T-1)} \frac{\pi(A_i|S_i)}{b(A_i|S_i)}$

      $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$

      IF $\tau + n < T$ THEN: $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$

      $Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha \rho [G - Q(S_\tau, A_\tau)]$

      IF π IS BEING LEARNED, THEN ENSURE THAT $\pi(\cdot|S_\tau)$ IS GREEDY WRT. Q

  UNTIL $\tau = T - 1$