CHAPTER 6 TEMPORAL DIFFERENCE LEARNING

6.1 TD PREDICTION

NON STATIONARY EVERY-VISIT MONTE CARLO:

$$V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(S_t) \right]$$

→ α: STEP SIZE: IF $\alpha = 1$, THEN WE SET ESTIMATE EQUAL TO LAST MEASUREMENT

→ THE ABOVE FORMULA IS APPLIED WHEN EPISODE FINISHES

ONE-STEP TD, OR TD(0)

$$V(S_t) \leftarrow V(S_t) + \alpha \overbrace{\left[\underbrace{R_{t+1} + \gamma V(S_{t+1})} - \underbrace{V(S_t)} \right]}^{\text{CORRECTION}}$$

NOTE THE TEMPORAL DIFFERENCE BETWEEN THE TWO → BETTER ESTIMATE, USING ONE EXTRA DATA POINT, R_{t+1} — INITIAL ESTIMATE

(WE COULD USE MORE DATA POINTS, THIS IS THE TD(n) METHOD). PSEUDO CODE:

INPUT: POLICY π

PARAMETER: STEP SIZE $\alpha \in (0, 1]$

INITIALIZE $V(s)$ $\forall s \in \mathcal{S}^+$, BUT $V(\text{TERMINAL}) = 0$

LOOP FOR EACH EPISODE:

  INITIALIZE S

  LOOP FOR EACH STEP OF EPISODE:

    $A \leftarrow$ ACTION GIVEN BY π FOR S

    TAKE ACTION, OBSERVE R, S'

    $V(S) \leftarrow V(S) + \alpha \left[R + \gamma V(S') - V(S) \right]$

    $S \leftarrow S'$

  UNTIL S TERMINAL

LET US COMPARE 1) DP (DYNAMIC PROGRAMMING)
2) MC (MONTE CARLO)
3) TD (TEMPORAL DIFFERENCE)

$$v_\pi(s) \triangleq E_\pi\left[G_t \mid S_t = s \right]$$ ← MONTE CARLO ESTIMATES THIS

$$= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s \right]$$

$$= E_\pi\left[R_{t+1} + \gamma \underbrace{v_\pi(S_{t+1})} \mid S_t = s \right]$$

TD ESTIMATES BOTH

DP "ESTIMATES" THIS

IMPORTANT QUANTITY:

TD ERROR: $\delta_t \triangleq R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$

MONTE CARLO ERROR: $G_t - V(S_t)$.

THE TWO ARE CONNECTED. ASSUMING THE ESTIMATED VALUES DO NOT CHANGE,

$$G_t - V(S_t) = R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1})$$

$$= \delta_t + \gamma\left(G_{t+1} - V(S_{t+1})\right)$$

$$= \delta_t + \gamma \delta_{t+1} + \gamma^2 \left(G_{t+2} - V(S_{t+1})\right)$$

$$= \delta_t + \gamma \delta_{t+1} + \gamma^2 \delta_{t+2} + \dots + \gamma^{T-t}\left(\underset{0}{G_T} - \underset{0}{V(S_T)}\right)$$

$$= \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k$$

EXAMPLE 6.1. DRIVING HOME

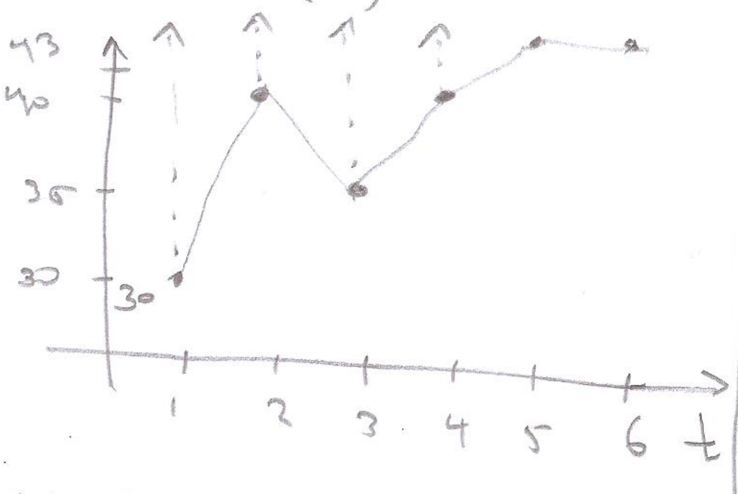
LET US CONSIDER ONE EPISODE OF DRIVING HOME, ACCORDING TO A SPECIAL POLICY (I.E. ROUTE TAKEN, ETC.)

STATE	ELAPSED TIME (REWARDS)	PREDICTED TIME TO GO	PREDICTED TOTAL TIME
1) LEAVING OFFICE FRIDAY AT 6	0	30	30
2) REACH CAR, RAINING*	5	35	40
3) EXITING HIGHWAY	20	15	35
4) 2 NDARY ROAD BEHIND TRUCK	30	10	40
5) ENTERING HOME STREET	40	3	43
6) ARRIVE HOME	43	0	43

(* THERE IS ANOTHER STATE WHERE YOU VISIT CAR BUT IT DOES NOT RAIN)

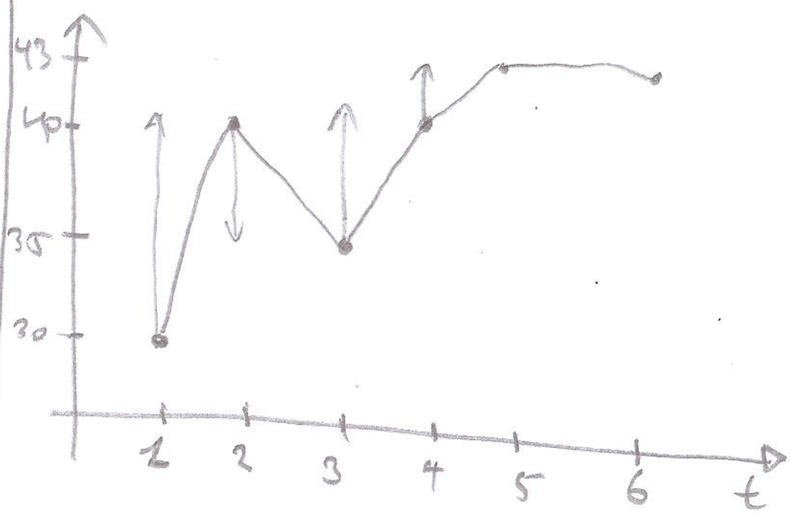
MONTE CARLO WITH $\alpha = 1$

PREDICTIONS $V(S_t)$



AT THE END, THESE ALL BECOME EQUAL TO 43

TD (0) WITH $\alpha = 1$

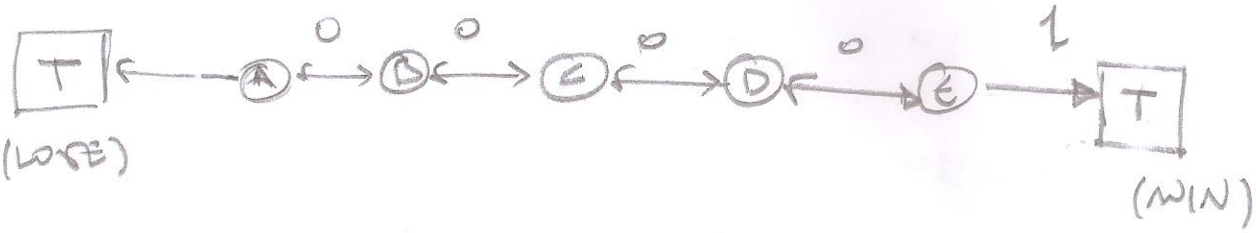


AFTER EACH ITERATION, THESE BECOME EQUAL TO SUBSEQUENT PREDICTION

6.2 ADVANTAGES OF TD PREDICTION METHODS

1) W.R.T DP, NO NEED FOR MODEL
2) W.R.T. MC, NO NEED TO WAIT FOR EPISODE TO END
3) THEY CONVERGE (I.E. $V_t(s) \to v_\pi(s)$)
    1) IN THE MEAN, IF α FIXED AND <u>SUFFICIENTLY</u> SMALL
    2) WITH PROBABILITY 1, IF $\alpha \to 0$ IN USUAL MANNER ((2.7))
4) TD IS USUALLY FASTER.

<u>EXAMPLE 6.2</u> RANDOM WALK (GAMBLER'S RUIN)



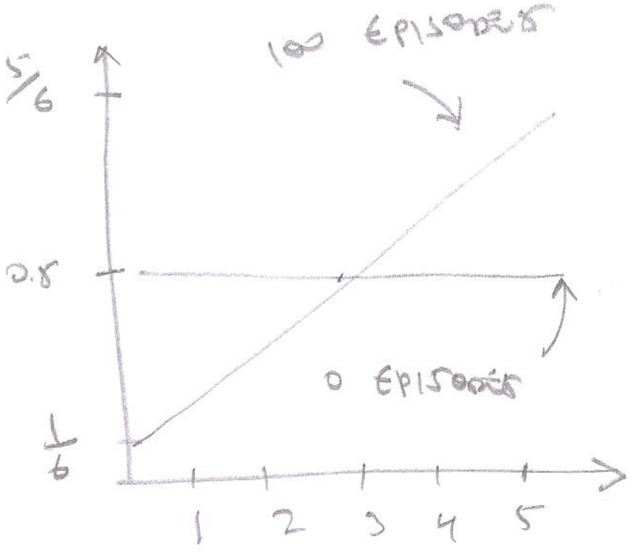
WHAT ARE PROBABILITIES P_A, P_B, P_C, P_D, P_E THAT STARTING AT RESPECTIVE STATE WE WIN?
OBSERVE THAT

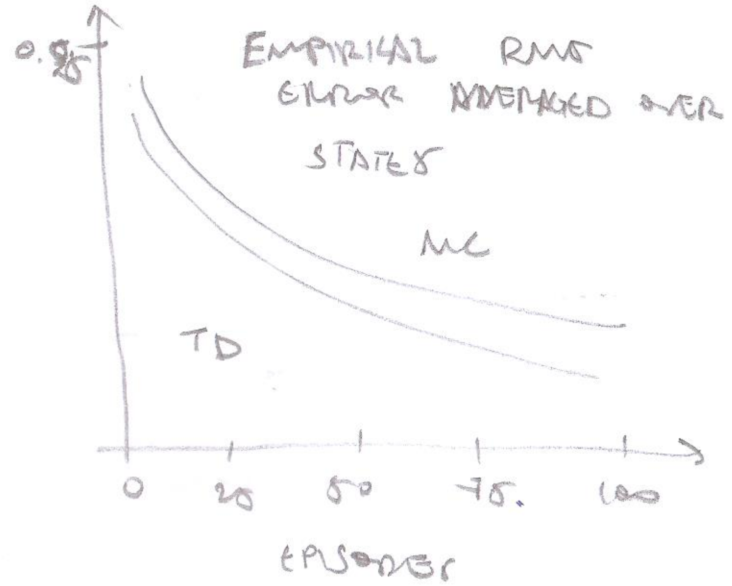
$$\left.\begin{aligned}
P_A &= \tfrac{1}{2}\cdot 0 + \tfrac{1}{2} P_B \\
P_B &= \tfrac{1}{2} P_A + \tfrac{1}{2} P_C \\
P_C &= \tfrac{1}{2} P_B + \tfrac{1}{2} P_D \\
P_D &= \tfrac{1}{2} P_E + \tfrac{1}{2} P_C
\end{aligned}\right\} \Rightarrow \quad P_A = \frac{1}{6},\ P_B = \frac{2}{6},\ P_C = \frac{3}{6},\ P_D = \frac{4}{6},\ P_E = \frac{5}{6}$$

$$P_E = \tfrac{1}{2}\cdot 1 + \tfrac{1}{2} P_D$$

EASY TO GENERALIZE FOR BIASED WALKS AND n STATES.

EXPLAINS ADVANTAGE OF CASINO





SO, WHY IS IT FASTER? IT USES FRESHER INFO.
ALSO, CONSIDER NEXT EXAMPLE:

EXAMPLE 6.4

YOU SIMULATE 8 EPISODES:

1) A, 0, B, 0 2) B, 1, 3) B, 1, 4) B, 1

STATES REWARDS 5) B, 1 6) B, 1 7) B, 1 8) B, 0

(AFTER B, WE ENTER TERMINAL STATE)

WHAT IS YOUR ESTIMATE OF $v_\pi(A)$ AND $v_\pi(B)$?
ASSUME $\gamma = 1$

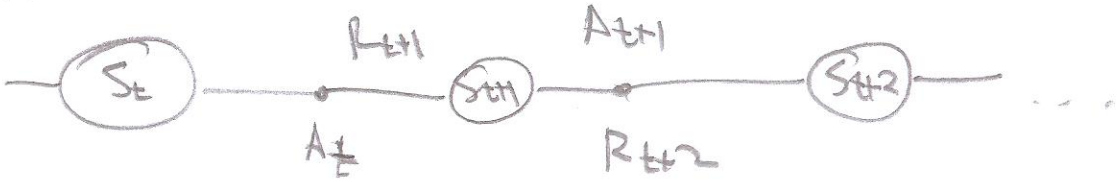
→ IF YOU SAY $v_\pi(B) = 0{,}75$, $v_\pi(A) = 0.75$ THEN YOU SAY WHAT TD(0) ACHIEVES, IF WE TRAIN THE ALGORITHM REPEATEDLY ON ABOVE 8 EPISODES

→ IF YOU SAY $v_\pi(B) = 0{,}75$, $v_\pi(A) = 0$ THEN YOU SAY WHAT MC ACHIEVES, IF WE TRAIN THE ALGORITHM REPEATEDLY ON ABOVE 8 EPISODES

IF MODEL IS MARKOVIAN, FIRST RESULT IS BETTER

6.4. SARSA: ON-POLICY TD CONTROL

CONSIDER A SEQUENCE OF STATE-ACTION PAIRS IN AN EPISODE:



THE SAME UPDATE FOR $V(S_t)$ NOW HOLDS FOR $Q(S_t, A_t)$:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \right]$$

(IF S_{t+1} = TERMINAL, $Q(S_{t+1}, A_{t+1}) = 0$)

UPDATE USES $S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}$, AND SO IS CALLED SARSA

SARSA (ON-POLICY TD. CONTROL) FOR ESTIMATING $Q \approx q_*$

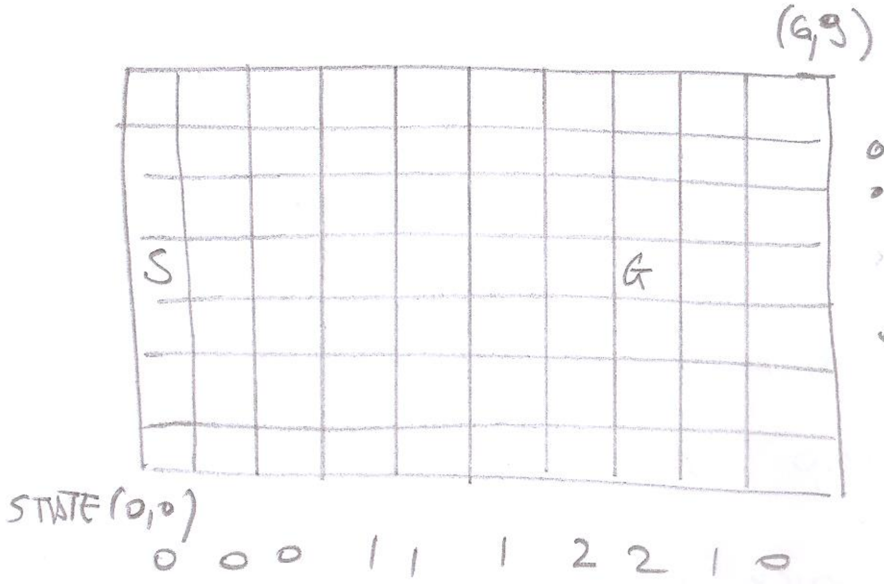
ALGORITHM PARAMETERS: STEP SIZE $\alpha \in (0, 1]$, SMALL $\epsilon > 0$

INITIALIZE $Q(s, a)$ $\forall s \in \mathcal{S}^+$, $a \in \mathcal{A}(s)$, ARBITRARILY BUT $Q(\text{TERMINAL}, \cdot) = 0$

LOOP FOR EACH EPISODE:
- INITIALIZE S
- CHOOSE A FROM S USING Q (E.G. ϵ-GREEDY)
- LOOP FOR EACH STEP OF EPISODE:
  - TAKE ACTION A, OBSERVE R, S'
  - CHOOSE A' FROM S' USING Q (E.G. ϵ-GREEDY)
  - $Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma Q(S', A') - Q(S, A)]$
  - $S \leftarrow S'$; $A \rightarrow A'$
- UNTIL S TERMINAL

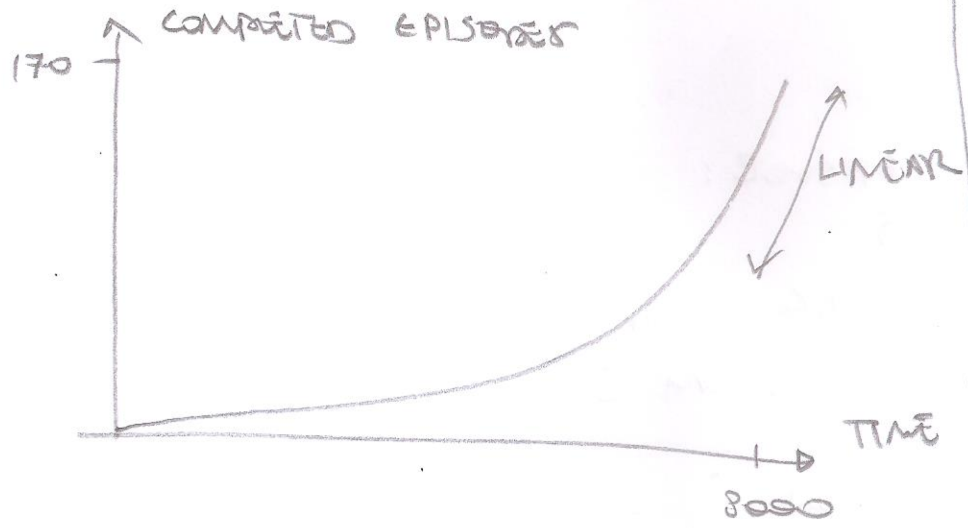
(57)

EXAMPLE 6.5 WINDY GRIDWORLD



- ALL REWARDS ARE -1
- EPISODIC TASKS: WE FINISH AT G
- WE FINISH AT S
- $\epsilon = 0.1$, $\alpha = 0.5$
- INITIAL $Q(s,a)$ $\forall s,a$

OBSERVE: TRIVIAL TO SOLVE WITH DIJKSTRA'S ALGORITHM, IF WE HAD MODEL.



HOMEWORK #6

REPRODUCE THIS FIGURE

6.5 Q-LEARNING: OFF POLICY TD-CONTROL

CONSIDER ALTERNATIVE UPDATE RULE:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

OBSERVE THAT:

$$(3.20) \quad q_*(s,a) = \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right]$$

THEREFORE, Q-LEARNING TRIES TO DISCOVER OPTIMAL POLICY. THEREFORE IT IS AN OFF-POLICY ALGORITHM, BECAUSE THERE ARE TWO POLICIES:

1) TARGET POLICY: THE OPTIMAL ONE
2) BEHAVIOR POLICY: ϵ-GREEDY, OR ANY OTHER THAT DOES SOME EXPLORING

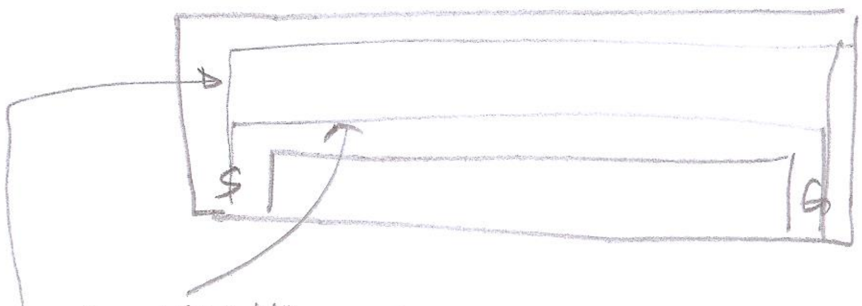
Q-LEARNING CAN ALSO BE SHOWN TO CONVERGE.

Q-LEARNING (OFF-POLICY TD CONTROL) FOR ESTIMATING π^*

ALGORITHM PARAMETERS: STEP SIZE $\alpha \in (0,1]$, SMALL $\epsilon > 0$
INITIALIZE $Q(s,a)$, FOR ALL $s \in \mathcal{S}^+$, $a \in \mathcal{A}(s)$ ARBITRARILY EXCEPT THAT $Q(\text{TERMINAL}, \cdot) = 0$

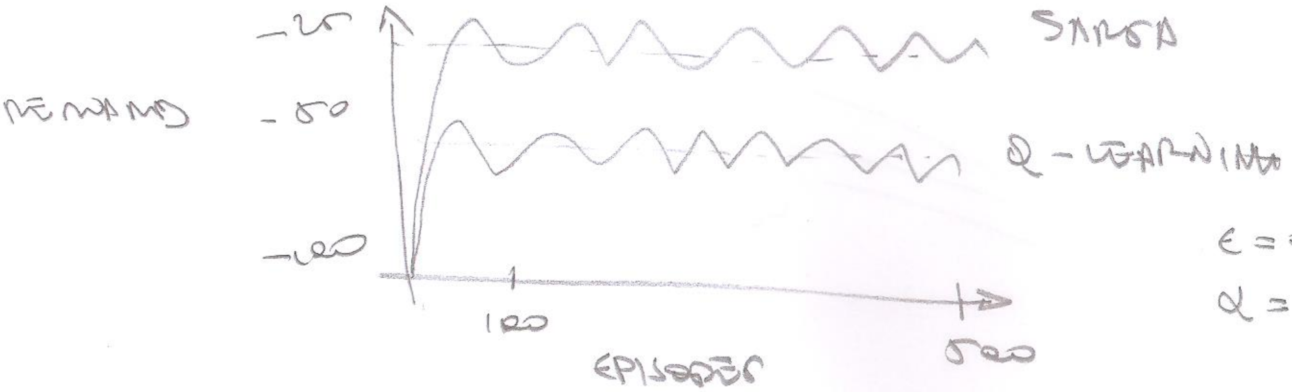
LOOP FOR EACH EPISODE:
- INITIALIZE S
- LOOP FOR EACH STEP OF EPISODE:
  - CHOOSE A FROM S USING Q AND ϵ-GREEDY POLICY (OR ANOTHER)
  - TAKE ACTION A, OBSERVE R, S'
  - $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
  - $S \leftarrow S'$
- UNTIL S TERMINAL

Q-LEARNING CAN PERFORM WORSE THAN SARSA.
CONSIDER THE CLIFF WALKING EXPERIMENT



Q LEARNING DISCOVERS THE OPTIMAL ROUTE AND USES IT, AND ONCE IN A WHILE IT FALLS IN THE CLIFF

BUT SARSA LEARNS TO STAY AWAY FROM THE CLIFF!



$\epsilon = 0.1$
α = CONSTANT

6.6 EXPECTED SARSA

WE MODIFY THE UPDATE RULE OF Q-LEARNING TO TAKE AVERAGE, INSTEAD OF MAXIMUM:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma E_\pi \big[Q(S_{t+1}, A_{t+1}) \,|\, S_{t+1} \big] - Q(S_t, A_t) \Big]$$

OR

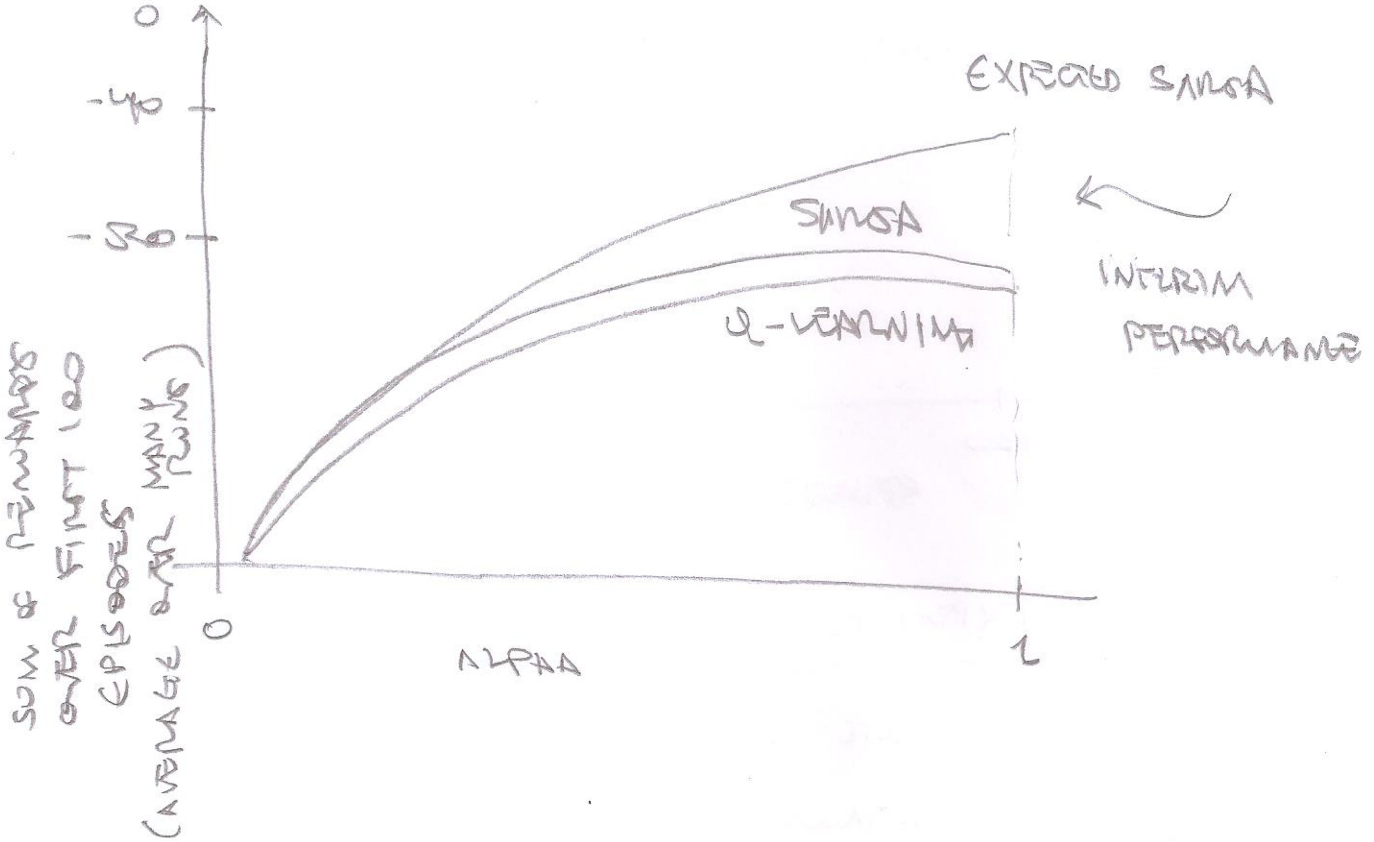
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1})\, Q(S_{t+1}, a) - Q(S_t, A_t) \Big]$$

(OBSERVE:
$$q_\pi(s, a) = \sum_{s', r} p(s', r | s, a) \Big[r + \gamma \sum_{a'} \pi(a'|s')\, q(s', a') \Big]$$
)

SO THIS SCHEME MOVES DETERMINISTICALLY IN THE EXPECTED DIRECTION TOWARDS WHICH SARSA MOVES, AND SO IT CALLED <u>EXPECTED</u> SARSA

BECAUSE IT HAS LESS VARIANCE THAN SARSA, IT GENERALLY PERFORMS BETTER THAN SARSA

CLIFF WALKING PERFORMANCE



$\boxed{\text{HOMEWORK \#7}}$

REPRODUCE THIS FIGURE, ONLY FOR THE INTERIM CURVES.

BONUS IF YOU CAN EXPLAIN WHY THE ASYMPTOTIC PERFORMANCE FOR SARSA IS BELOW THE INTERIM PERFORMANCE

USE ϵ-GREEDY WITH $\epsilon = 0.1$