$$= E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \mid S_t = s\right]$$

$$\leq \dots \leq E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3}) \mid S_t = s\right]$$

$$\leq \dots E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s\right]$$

$$= v_{\pi'}(s).$$

QED.

THIS PROPERTY SUGGESTS THE FOLLOWING <u>POLICY IMPROVEMENT</u> WHICH IS GREEDY: ONCE WE HAVE EVALUATED A POLICY AND WE HAVE $v_\pi(s)$ AND $q_\pi(s,a)$, WE SET A NEW POLICY

$$\pi'(s) \triangleq \arg\max_a q_\pi(s,a)$$

$$= \arg\max_a E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \arg\max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right]$$

AFTER THE IMPROVEMENT, THERE ARE TWO POSSIBILITIES:

1) WE FIND A BETTER VALUE FUNCTION. THIS CAN ONLY HAPPEN FINITE NUMBER OF TIMES
2) WE DO NOT IMPROVE THE VALUE FUNCTION. THEN: $v_{\pi'} = v_\pi$ AND

$$v_{\pi'}(s) = \max_a E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right]$$

SO THE BELLMAN OPTIMALITY CONDITIONS HOLD!

SO AFTER FINITE NUMBER OF STEPS WE ARRIVE AT OPTIMAL POLICY.

SO FAR, WE DISCUSSED DETERMINISTIC POLICIES. IN FACT, POLICY IMPROVEMENT THEOREM HOLDS FOR ALL POLICIES. WHEN YOU TAKE THE ARGMAX, JUST ALLOCATE ALL PROBABILITY TO THE OPTIMAL ACTIONS ONLY

4.3 POLICY ITERATION

THE TWO PREVIOUS PARAGRAPHS CAN BE COMBINED

$$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} v_*$$

E: EVALUATION I: IMPROVEMENT.

THIS IS GUARANTEED TO CONVERGE!

$\boxed{\text{HW3: CALCULATE THE RESULTS OF EXAMPLE 4.2}}$

JACK'S CAR RENTAL

1) TWO LOCATIONS FOR CAR RENTAL WITH CAPACITY OF 20 CARS EACH
2) RETURNS AND REQUESTS ARE POISSON

$$\left(P(X = n) = e^{-\lambda} \frac{\lambda^n}{n!} \right)$$

3) RETURN RATES: 3, 2
4) REQUEST RATES: 3, 4
5) CAR RENTAL: \$10
6) CAR TRANSPORT: \$2
7) STATE: NUMBER OF CARS AT EACH LOCATION AT END OF DATE.

8) POLICY: NUMBER of CARS WE MOVE FROM (OVERNIGHT) LOCATION 1 TO LOCATION 2, MAXIMUM 5 BOTH WAYS

9) RETURNS BECOME AVAILABLE THE NEXT DAY

4.4. VALUE ITERATION

OBSERVATION: NO NEED TO FIND EXACT VALUE FUNCTION FOR INTERMEDIATE ITERATIONS. JUST MAKE ONE UPDATE.

THE TWO LOOPS ARE THEN COMBINED IN SIMPLE ITERATION:

$$v_{k+1}(s) \triangleq \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_k(s')\right]$$

SO NOW BELLMAN OPTIMALITY CONDITIONS BECOME AN ITERATIVE UPDATE!

VALUE ITERATION ALGORITHM

INPUT: THRESHOLD 1) $\theta > 0$, 2) INITIAL $V(s)$ $\forall s \in S^+$ (BUT $V(\text{TERMINAL}) = 0$)

LOOP:

$\Delta \leftarrow 0$

LOOP $\forall s \in S$:

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma V(s')\right]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

UNTIL $\Delta < \theta$

OUTPUT: $\pi(s) = \arg\max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma V(s')\right]$

OBSERVE: 1) WE USE SWEEPS, I.E. VALUE FUNCTIONS ARE MADE IMMEDIATELY AVAILABLE

2) POLICY IS CALCULATED EXPLICITLY ONLY AT THE END

3) IN EACH NEW POLICY EVALUATION, WE START WITH PREVIOUS ONE (WARM START?) (CURSE OF DIMENSIONALITY)

PROBLEM: STATE SPACE IS OFTEN SO HUGE, THAT THIS PLAIN VANILLA APPROACH ONLY WORKS IN SIMPLE PROBLEMS. IN REST OF COURSE TO TRY TO IMPROVE SPEED BY BEING SMARTER.

CHAPTER 5: MONTE CARLO

5.1 MONTE CARLO PREDICTION

DEFINITION OF MONTE CARLO: WE SIMULATE ENTIRE EPISODES AND USE THEM TO APPROXIMATE $v_\pi(s)$, $q_\pi(s,a)$, $v_*(s)$, $q_*(s,a)$.

for now, focus on APPROXIMATING $v_\pi(s)$ AND $q_\pi(s,a)$

NOTE THAT $v_\pi(s) = E_\pi(G_t | S_t = s)$, $q_\pi(s,a) = E_\pi[G_t | S_t = s, A_t = a]$

THIS SUGGESTS THE FOLLOWING ALGORITHM (pg. 92)

INPUT: π, POLICY TO EVALUATE

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

$Returns(s) \leftarrow$ EMPTY LIST, $\forall s \in S$

LOOP FOREVER:

GENERATE EPISODE FOLLOWING π:

$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, A_3, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

LOOP FOR EACH STEP, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

UNLESS S_t APPEARS IN $S_0, S_1, \ldots, S_{t-1}$:

APPEND G TO RETURNS(S_t)

$V(S_t) \leftarrow$ AVERAGE(RETURNS(S_t))

COMMENTS: 1) THIS IS FIRST-VISIT MC, WHICH GUARANTEES THAT MEASUREMENTS ARE INDEPENDENT $\Longrightarrow$ SLLN

2) THERE IS ALSO EVERY-VISIT VARIANT, WHICH ALSO CONVERGES

3) SIMPLE MODIFICATION ALLOWS ALSO ESTIMATION OF $q_\pi(s,a)$

4) SPEED OF CONVERGENCE IS SPECIFIED BY CENTRAL LIMIT THEOREM:

$x_1, x_2, x_3 \ldots$ IID WITH μ, σ. THEN

$$\frac{\sum_{i=1}^{N} x_i - N\mu}{\sqrt{N} \cdot \sigma} \sim N(0,1), \quad \text{SO} \quad \sum_{i=1}^{N} x_i \text{ DEVIATES}$$

BY A MULTIPLE OF $\sqrt{N}$

COMPARISON WITH DP

1) NO NEED TO COMPUTE PROBABILITY DISTRIBUTIONS
2) NO BOOTSTRAPPING, IE USING OTHER VALUES OF $v_{\pi}(s')$
 SO WE CAN FOCUS RESOURCES ON SPECIFIC STATES AND STATE-ACTION PAIRS
3) WE NEED TO MAKE SURE THAT SIMULATION VISITS ALL
 STATE-ACTION PAIRS, EVEN THOSE INCOMPATIBLE WITH CURRENT POLICY

5.3 MONTE-CARLO CONTROL

APPLY CLASSIC POLICY ITERATION WITH MONTE CARLO:

$$\pi_0 \xrightarrow{E} q_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} q_{\pi_1} \rightarrow \dots \xrightarrow{I} \pi_* \xrightarrow{E} q_*$$

TWO PROBLEMS WE WILL FACE:

① WE NEED TO ~~HAVE~~ DATA POINTS FOR ALL (s,a) STATE ACTION PAIRS. SOLUTION: SO WE DO EXPLORING STARTS, I.E. WE PICK A RANDOM (s,a) PAIR TO START

② WE NEED TO MAKE INFINITE ITERATIONS FOR CONVERGENCE OF POLICY ESTIMATION. OBVIOUSLY WE CANNOT DO THIS, SO WE DO FINITE ITERATIONS, OR JUST ONE

WE END UP WITH FOLLOWING ALGORITHM, THAT HAS NOT BEEN SHOWN TO CONVERGE YET!

MONTE CARLO ES (EXPLORING STARTS) FOR ESTIMATING π_*

INITIALIZE

$\pi(s) \in \mathcal{A}(s)$ ARBITRARILY, $\forall s \in \mathcal{S}$ (SO POLICY IS DETERMINISTIC)

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

RETURNS(s,a) $\leftarrow$ EMPTY LIST, $\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

LOOP FOREVER

CHOOSE $S_0 \in \mathcal{S}, A_0 \in \mathcal{A}(S_0)$ SO THAT ALL PAIRS HAVE PROB >0

GENERATE AN EPISODE FROM S_0, A_0 THAT FOLLOWS π:

$S_0, A_0, R_1, \dots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

LOOP FOR EACH STEP OF EPISODE $t = T-1, T-2, \dots, 0$

- $G \leftarrow \gamma G + R_{t+1}$
- UNLESS THE PAIR S_t, A_t APPEARS IN $S_0, A_0, \dots S_{t-1}, A_{t-1}$
  - APPEND G TO RETURNS(S_t, A_t)
  - $Q(S_t, A_t) \leftarrow$ AVERAGE(RETURNS(S_t, A_t))
  - $\pi(S_t) \leftarrow \operatorname{ARGMAX}_a Q(S_t, a)$

OBSERVE:

1) THE LOOP CANNOT CONVERGE TO SUBOPTIMAL POLICY, BECAUSE BELLMAN OPTIMALITY EQUATIONS ONLY WORK FOR OPTIMAL POLICY
2) BUT WE DON'T KNOW IF IT CONVERGES TO OPTIMAL POLICY (WE FULLY EXPECT IT DOES)

5.4 MONTE CARLO CONTROL WITHOUT EXPLORING STARTS

DEFINITION: AN ϵ-SOFT POLICY HAS: $\pi(a|s) \geq \frac{\epsilon}{|\mathcal{A}(s)|}$

DEFINITION: AN ϵ-GREEDY POLICY MAXIMIZES $q(s,a)$ OVER a FOR A FRACTION OF TIME $1-\epsilon$ AND FOR REST OF TIME IS RANDOM

NEW ALGORITHM: INSTEAD OF $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$, WE NOW HAVE:

$$A^* \leftarrow \arg\max_a Q(S_t, a)$$

$\forall a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1-\epsilon+\frac{\epsilon}{|\mathcal{A}(S_t)|} & \text{if } a = A^* \\ \frac{\epsilon}{|\mathcal{A}(S_t)|}, & \text{if } a \neq A^* \end{cases}$$

$\boxed{\text{HW \#4: APPLY BOTH ALGORITHMS FOR BLACK JACK, EXAMPLE 4.3}}$

BLACK JACK: (DETAILED DESCRIPTION IN BOOK)

ONE EPISODE:

1) DEALER SHOWS ONE CARD (A-10)
2) WE DRAW CARDS WITH AIM TO REACH 21 BUT NOT EXCEED IT.
3) WHEN WE STOP, DEALER TRIES TO REACH 21
4) WHOEVER STOPS CLOSER WINS.

OTHER RULES: FACE CARDS COUNT AS 10

ACES COUNT AS 1 OR 11 WHICHEVER IS BETTER

5) INFINITE CARDS IN STACK

SOLUTION:

STATE IS (MY CARDS, DEALER'S OPEN CARD, I HAVE ACE)

12-21 ; A-10 ; 0,1

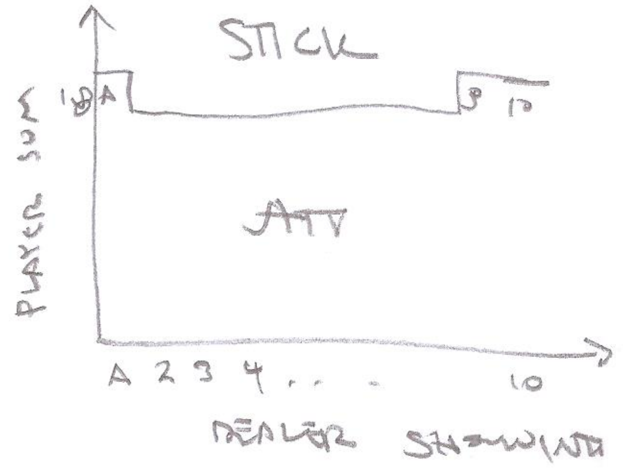
ACE COUNTS AS 11.

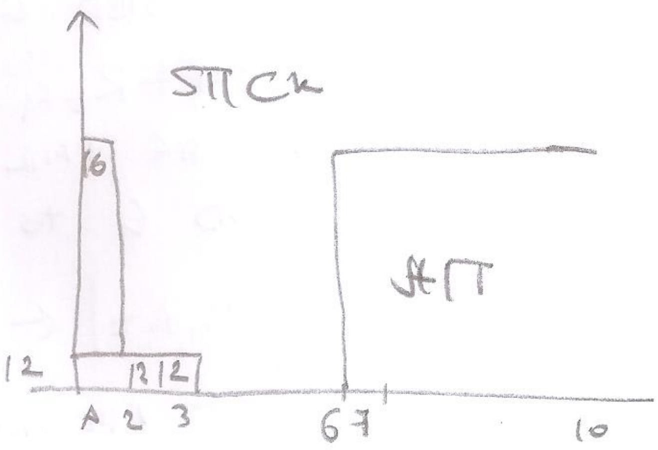
$\Rightarrow 2 \cdot 10 \cdot 10 = 200$

NO NEED TO CONSIDER OTHER STATES, WHEN THERE IS NOTHING TO DECIDE

ACTION IN EACH STATE IS STICK (0) OR HIT (1)

OPTIMAL POLICY (pg. 100)





OBSERVATIONS:

1) $\gamma = 1$
2) EACH STATE VISITED AT MOST ONCE
3) DEALER'S OPEN HAND REMAINS FIXED, SO IN SOME SENSE WE HAVE 10 INDEPENDENT PROBLEMS
4) INTUITION: IF WE HAVE A USEABLE ACE, WE HAVE ADVANTAGE, BECAUSE WE CAN EXCEED 21 ONCE
5) WE NEED TO BE MORE AGGRESSIVE WHEN THE DEALER HAS A GOOD CARD
6) NO NEED TO COUNT CARDS, IN THIS SETTING

RESULTING ALGORITHM:

ALGORITHM PARAMETER: $\epsilon > 0$

INITIALIZE:

$\pi \leftarrow$ ARBITRARY SOFT POLICY

$Q(s, a) \in \mathbb{R}$ ARBITRARILY FOR ALL $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

RETURNS$(s, a) \leftarrow$ EMPTY LIST, FOR ALL $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

REPEAT FOREVER (FOR EACH EPISODE)

- GENERATE AN EPISODE FOLLOWING π: $S_0, A_0, R_1, \ldots S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$
- LOOP FOR EACH STEP OF EPISODE, $t = T-1, T-2, \ldots, 0$:
  - $G \leftarrow \gamma G + R_{t+1}$
  - UNLESS THE PAIR (S_t, A_t) APPEARS IN $S_0, A_0, \ldots S_{t-1}, A_{t-1}$:
    - APPEND G TO RETURNS(S_t, A_t)
    - $Q(S_t, A_t) \leftarrow$ AVERAGE(RETURNS(S_t, A_t))
    - $A^* \leftarrow \text{ARGMAX}_a\, Q(S_t, a)$ (TIES BROKEN ARBITRARILY)
    - FOR ALL $a \in \mathcal{A}(S_t)$:
      - $$\pi(a \mid S_t) \leftarrow \begin{cases} 1 - \epsilon + \dfrac{\epsilon}{|\mathcal{A}(S_t)|}, & a = A^* \\[2ex] \dfrac{\epsilon}{|\mathcal{A}(S_t)|}, & a \neq A^* \end{cases}$$

        A

SO, DOES THIS ALGORITHM REACH OPTIMAL POLICY? LET'S EXPLORE

THEOREM 1: ANY ϵ-GREEDY POLICY π' W.R.T. q_π IS BETTER THAN ANY OTHER ϵ-SOFT POLICY π

(WHICH MEANS THE ALGORITHM TRIES TO CHANGE TO SOMETHING (π') BETTER THAN π