PROOF: WE USE THE POLICY IMPROVEMENT THEOREM (CHAPTER 4). LET π' BE THE ϵ-GREEDY, AND π THE ϵ-SOFT POLICIES.

$$q_\pi(s, \pi'(s)) = \sum_a \pi'(a|s)\, q_\pi(s,a) =$$

$$\frac{\epsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \max_a q_\pi(s,a)$$

$$\geq \frac{\epsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|A(s)|}}{1-\epsilon}\, q_\pi(s,a)$$

(EQUALITY HOLDS WHEN $\pi(a)$ IS CHOSEN AS IN A, pg. 40)

COEFFICIENTS ≥ 0 AND SUM $=1$

$$= \frac{\epsilon}{|A(s)|} \sum_a q_\pi(s,a) - \frac{\epsilon}{|A(s)|} \sum_a q_\pi(s,a) + \sum_a \pi(a|s)\, q_\pi(s,a)$$

$$= v_\pi(s)$$

$$\Rightarrow \quad q_\pi(s, \pi'(s)) \overset{A}{\geq} v_\pi(s)$$

THEOREM 2: THE POLICY CALCULATED BY THE ALGORITHM CAN ONLY CONVERGE TO A POLICY THAT IS OPTIMAL AMONG ALL POLICIES THAT ARE ϵ-SOFT

PROOF: CONSIDER A NEW ENVIRONMENT THAT BEHAVES RANDOMLY; WITH PROB ϵ, IT CHOOSES A RANDOM ACTION UNIFORMLY. OTHERWISE, IT FOLLOWS SPECIFIED ACTION ACCORDING TO POLICY.

THEN OPTIMAL POLICY IN NEW ENVIRONMENT IS SAME AS BEST AMONG ϵ-SOFT POLICIES IN OLD ENVIRONMENT.

CONSIDER BELLMAN OPTIMALITY FOR NEW ENVIRONMENT:

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

$$= \max_a \left\{ (1-\epsilon) E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a, \text{NON RANDOM}\right] + \epsilon E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a, \text{RANDOM}\right]\right\}$$

$$= (1-\epsilon) \max_a E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a, \text{NON RANDOM}\right] + \frac{\epsilon}{|A(s)|} \sum_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \tilde{v}_*(s')\right] \Rightarrow$$

$$\left.\begin{aligned} \tilde{v}_*(s) = {} & (1-\epsilon) \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \tilde{v}_*(s')\right] \\ & + \frac{\epsilon}{|A(s)|} \sum_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \tilde{v}_*(s')\right] \end{aligned}\right\} \text{BELLMAN OPTIMALITY FOR NEW ENV.}$$

NOW ASSUME THAT ALGORITHM CONVERGES TO POLICY π. THEN:

$$v_\pi(s) = \sum_a \pi(a \mid s)\, q(s,a) =$$

$$\sum_a \frac{\epsilon}{|A(s)|} q(s,a) + (1-\epsilon) \max_a q(s,a)$$

$$= (1-\epsilon) \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right] + \frac{\epsilon}{|A(s)|} \sum_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s)\right]$$

SO $v_\pi(s)$ ALSO SATISFIES BELLMAN OPTIMALITY FOR NEW ENV. SO IT IS OPTIMAL AMONG ϵ-SOFT POLICIES FOR OLD ENV.

THEOREM 1: ALGORITHM SLOWLY IMPROVES

THEOREM 2: ALGORITHM STOPS WHEN OPTIMALITY IS ACHIEVED.

NB: PROOFS NOT STRICT

5.5 OFF-POLICY PREDICTION VIA IMPORTANCE SAMPLING

UNTIL NOW: SINGLE POLICY π USED BOTH FOR EXPLORING AND APPROACHING OPTIMAL $\Rightarrow$ CONFLICTING GOALS.

NOW: TWO POLICIES: TARGET POLICY π, BEHAVIOR POLICY b $\Rightarrow$ OFF-POLICY METHODS

ADVANTAGE: MORE GENERAL

DISADVANTAGE: SLOWER

FIRST, WE FOCUS ON PREDICTION OF π USING b

REQUIREMENT: COVERAGE $\pi(a|s) > 0 \Rightarrow b(a|s)$

INDEED, WE NEED TO ENCOUNTER WHAT WE MEASURE

WE DEFINE IMPORTANCE-SAMPLING RATIO $\rho_{t:T-1}$ AS FOLLOWS

PROBABILITY OF STATE-ACTION TRAJECTORY

$$\Pr\{A_t, S_{t+1}, A_{t+1}, \dots, A_{T-1}, S_T \mid S_t, A_{t:T-1} \sim \pi\}$$

$$= \pi(A_t|S_t)\, p(S_{t+1}|S_t, A_t)\, \pi(A_{t+1}|S_{t+1}) \dots p(S_T|S_{T-1}, A_{T-1})$$

$$= \prod_{k=t}^{T-1} \pi(A_k|S_k)\, p(S_{k+1}|S_k, A_k).$$

LIKEWISE FOR POLICY b.

HOW MUCH MORE PROBABLE TO SEE THIS TRAJECTORY WHEN USING π INSTEAD OF b?

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

SO THIS RATIO ONLY DEPENDS ON THE POLICIES, NOT THE ENVIRONMENT

OBSERVE THAT $\rho_{t:T-1}$ IS ALSO A RANDOM VARIABLE BUT DEPENDS ONLY ON TRAJECTORY

ITS USEFULNESS COMES FROM THE FOLLOWING

$$v_\pi(s) = E_\pi\left[G_t \mid S_t = s\right]$$

$$\underset{\substack{\text{(NESTED}\\\text{EXPECTATION)}}}{=} \sum_{\substack{\text{ALL TRAJECTORIES TRAJ}\\\text{LEADING TO T}}} P[\text{TRAJ} \mid \pi]\, E\left[G_t \mid S_t = s, \text{TRAJ}\right]$$

$$= \sum_{\substack{\text{ALL TRAJECTORIES TRAJ}\\\text{LEADING TO T}}} \rho_{t:T-1}\, P[\text{TRAJ} \mid b]\, E\left[G_t \mid S_t = s, \text{TRAJ}\right] =$$

$$\sum_{\substack{\text{ALL TRAJECTORIES TRAJ}\\\text{LEADING TO T}}} P[\text{TRAJ} \mid b]\, E\left[\rho_{t:T-1}\, G_t \mid S_t = s, \text{TRAJ}\right]$$

$$\underset{\substack{\text{(NESTED}\\\text{EXPECTATION)}}}{=} E_b\left[\rho_{t:T-1}\, G_t \mid S_t = s\right] \Rightarrow$$

$$\boxed{v_\pi(s) = E_b\left[\rho_{t:T-1}\, G_t \mid S_t = s\right]}$$

(NOTE: WE ALSO HAVE

$$v_\pi(s) = E_\pi\left[G_t \mid S_t = s\right]$$
$$v_b(s) = E_b\left[G_t \mid S_t = s\right]$$

)

WE INTRODUCE NOTATION:

→ LET t BE THE TIME INDEX OF CONSECUTIVE EPISODES FOLLOWING POLICY b. (SO IF AT $t=100$ WE END EPISODE, AT TIME 101 WE START NEW EPISODE

→ LET $\mathcal{T}(s)$ THE SET OF TIMES WE VISIT STATE s (EITHER FOR FIRST TIME, OR ALL TIMES, ACCORDING TO THE VARIANT USED)

→ LET $T(t)$ THE FIRST TIME OF TERMINATION FOLLOWING t.

THEN, WE CAN ESTIMATE $v_\pi(s)$ AS FOLLOWS:

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|} \quad (A)$$

EXPRESSES HOW MANY MORE OR FEWER TIMES WE WOULD HAVE SEEN THAT STATE UNDER π

THIS IS ORDINARY IMPORTANCE SAMPLING

ANOTHER FORMULA IS FREQUENTLY USED, AND IS ACTUALLY PREFERABLE:

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}} \quad (B)$$

WEIGHTED IMPORTANCE SAMPLING.

BOTH (A) AND (B) CONVERGE TO $v_\pi(s)$, AS $t \to \infty$

THE CONVERGENCE OF (A) FOLLOWS BY THE SLLN

FOR THE CONVERGENCE OF (B) WE NOTE:

$$\frac{\dfrac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|}}{\dfrac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}{|\mathcal{T}(s)|}} \to \frac{v_\pi(s)}{1}$$

(CONSIDER A MODIFIED ENVIRONMENT WHERE ALL REWARDS ARE ZERO, EXCEPT THE REWARD MOVING TO A TERMINAL STATE, WHICH IS 1)

PROS AND CONS :

A IS UNBIASED, I.E. ITS MEAN IS $v_\pi(s)$

A HAS LARGE VARIANCE, BECAUSE THE RATIOS MIGHT BE LARGE

B IS BIASED, I.E. ITS MEAN $\neq v_\pi(s)$

B HAS MUCH SMALLER VARIANCE,

OVERALL, B IS PREFERABLE

FOR EXAMPLE, COMPARE A, B WHEN $|\mathcal{T}(s)| = 1$ (FOR FIRST-VISIT CASE) AND THERE IS A SINGLE RETURN G_t

A $V(s) = \rho_{t:T(t)-1}\, G_t$ WHICH IS UNBIASED, BUT HAS LARGE VARIANCE

SAY 10

B $V(s) = G_t$ WHICH IS BIASED, BUT HAS SMALLER VARIANCE

(NOTE: EVERY-VISIT ESTIMATES ARE BOTH BIASED)

5.6 INCREMENTAL IMPLEMENTATION

WE DEVELOP ALGORITHM THAT SIMULATES ACCORDING TO POLICY b AND COMPUTES

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}\, G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

WE HAVE SEQUENCE of

$$G_1, G_2, \ldots, G_{n-1}, \ldots$$

$$W_1, W_2, \ldots, W_{n-1}, \ldots \qquad (W_i = P_{t_i} : T(t_i)) - 1$$

WE NEED TO FORM ESTIMATE

$$V_n \stackrel{\Delta}{=} \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \quad n \geq 2, \quad V_1 = \text{ARBITRARY}$$

WE ALSO DEFINE $\quad C_n = \sum_{k=1}^{n} W_k, \quad C_0 = 0$

WE NOTE THAT

$$V_{n+1} = \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k} = \frac{\sum_{k=1}^{n-1} W_{k-1} G_{k-1}}{C_n} + \frac{W_n G_n}{C_n}$$

$$= \frac{\sum_{k=1}^{n-1} W_k G_k}{C_n} + \frac{W_n V_n}{C_n} + \frac{W_n}{C_n}\left[G_n - V_n\right]$$

$$= \underbrace{\frac{\sum_{k=1}^{n-1} W_k G_k}{C_{n-1}}}_{V_n} \cdot \frac{C_{n-1}}{C_n} + \frac{W_n V_n}{C_n} + \frac{W_n}{C_n}\left[G_n - V_n\right]$$

$$= V_n\left[\frac{C_{n-1} + W_n}{C_n}\right] + \frac{W_n}{C_n}\left[G_n - V_n\right] \Rightarrow$$

$$V_{n+1} = V_n + \frac{W_n}{C_n}\left[G_n - V_n\right], \quad n \geq 1$$

THIS LEADS TO FOLLOWING ALGORITHM (pg. 110)

OFF-POLICY MC PREDICTION FOR ESTIMATING $Q \simeq q_\pi$

INPUT: TARGET POLICY π

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

$Q(s,a) \in \mathbb{R}$ (ARBITRARILY)

$C(s,a) \leftarrow 0$

LOOP FOREVER (FOR EACH EPISODE)

$b \leftarrow$ ANY POLICY WITH COVERAGE OF π

GENERATE AN EPISODE FOLLOWING $b: S_0, A_0, R_1, \ldots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$W \leftarrow 1$

LOOP FOR EACH STEP OF EPISODE, $t = T-1, T-2, \ldots, 0$ WHILE $W \neq 0$

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

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \dfrac{W}{C(S_t, A_t)}\left[G - Q(S_t, A_t)\right]$

$W \leftarrow W \dfrac{\pi(A_t | S_t)}{b(A_t | S_t)}$ (*)

THIS CAN BE EASILY MODIFIED TO BECOME A CONTROL ALGORITHM, AS WE SHOW NEXT

(*) SHOULD PROBABLY BE UP!

5.7 OFF-POLICY MONTE CARLO CONTROL

INITIALIZE $\forall s \in S, a \in A(s)$:

$Q(s,a) \in \mathbb{R}$ (ARBITRARILY)

$C(s,a) \leftarrow 0$ (SUM OF WEIGHTS)

$\pi(s) \leftarrow \text{ARGMAX}_a\, Q(s,a)$ (TIES BROKEN CONSISTENTLY)

LOOP FOREVER (FOR EACH EPISODE)

$b \leftarrow$ ANY SOFT POLICY

GENERATE AN EPISODE USING b: $S_0, A_0, R_1, \ldots S_{T-1}, A_{T-1}$

$G \leftarrow 0$

$W \leftarrow 1$

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

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

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \dfrac{W}{C(S_t, A_t)}\left[G - Q(S_t, A_t)\right]$

$\pi(S_t) \leftarrow \text{ARGMAX}_a\, Q(S_t, a)$ (TIES BROKEN CONSISTENTLY)

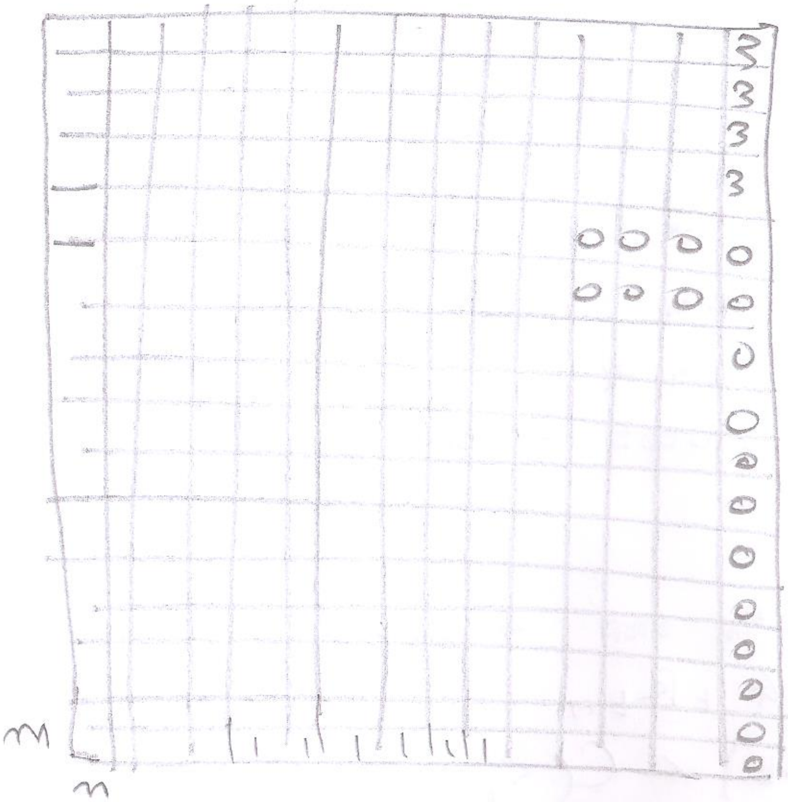
IF $A_t \neq \pi(S_t)$ THEN EXIT INNER LOOP (AND GO TO NEXT EPISODE)

* $W \leftarrow W \dfrac{1}{b(A_t | S_t)}$

* PROBABLY PLACED WRONG IN THE BOOK

ONE PROBLEM: WE NEED TO CUT SHORT THE PARSING EPISODE WHEN WE CHANGE THE POLICY

$\boxed{\text{HW \#5: SOLVE THE RACE TRACK EXERCISE, 5.12}}$



0 : NO GO
1 : START
2 : ROAD
3 : END

→ STATE: LOCATION AND VELOCITY

→ VELOCITY: VALID LOCATIONS UP/DOWN AND LEFT/RIGHT
(MAX +5) ⇒ 11 × 11 CASES

→ ACTION: CHANGE VELOCITY COMPONENTS +1, -1, 0

→ REWARD: -1 PER STEP NOT FINISHING

→ IF YOU HIT BOUNDARY, YOU GO BACK AT START WITH ZERO VELOCITY

→ WITH PROB 0.1, VELOCITY INCREMENTS ARE ZERO
(OIL ON THE TRACK!)

→ USE OFF-POLICY METHOD OF PREVIOUS PAGE

COMMENT: MOST EFFICIENT WAY TO SOLVE THIS IS WRITE IT AS A <u>ROUTING</u> PROBLEM, ON A GRAPH, WHERE THERE IS A GRAPH NODE FOR EACH STATE

BUT OUR METHOD GENERALIZES BETTER