IMPORTANCE SAMPLING RATIO FOR THE STATE-ACTION VALUES:

$$P_r\left\{S_{t+1}, A_{t+1}, S_{t+2}, A_{t+2}, \ldots, S_{T-1}, A_{T-1}, S_T \mid S_t, A_t, A_{t+1:T-1} \sim \pi\right\}$$

$$= \left[\prod_{k=t}^{T-2} P(S_{k+1} \mid S_k, A_k)\, \pi(A_{k+1} \mid S_{k+1})\right] P(S_T \mid S_{T-1}, A_{T-1})$$

WE DEFINE THE PRODUCT OF THE TWO FOR THE TWO POLICIES:

$$\rho_{t:T-1} = \prod_{k=t}^{T-2} \frac{\pi(A_{k+1} \mid S_{k+1})}{b(A_{k+1} \mid S_{k+1})}$$

NEXT, WE SHOW THAT $E_b\left[\rho_{t:T-1} G_t \mid S_t = s, A_t = a\right]$

$$= q_\pi(s, a)$$

INDEED:

$$q_\pi(s,a) = E_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

$$= \sum_{\text{ALL TRAJECTORIES}} P[\text{TRAJ} \mid \pi, S_t = s, A_t = a]\, E_\pi[G_t \mid S_t = s, A_t = a, \text{TRAJ}]$$

TRAJECTORY CONDITIONING

$$= \sum_{\text{ALL TRAJ}} \rho_{t:T-1}\, P[\text{TRAJ} \mid b, S_t = s, A_t = a]\, E_b[G_t \mid S_t = s, A_t = a, \text{TRAJ}]$$

NUMBER

$$= \sum_{\text{ALL TRAJ}} P[\text{TRAJ} \mid b, S_t = s, A_t = a]\, E_b[\rho_{t:T-1} G_t \mid S_t = s, A_t = a, \text{TRAJ}]$$

$$= E_b[\rho_{t:T-1} G_t \mid S_t = s, A_t = a)$$