TL;DR FROM PREVIOUS LECTURE

- GPI ≡ ESTIMATE OF q_π ⟲ IMPROVEMENT OF π USING q_π ($\pi' = \varepsilon\text{-greedy}(q_\pi)$)

WHY SOLVE A MORE GENERAL PROBLEM?

IDEA: ① $\forall s \in S$, PARAMETRIZE THE POLICY: $\pi(a,s) = \pi(a,s,\theta) =: \pi(\theta)$

② CHOOSE A PERFORMANCE MEASURE $J(\theta)$

THE LARGER $J(\theta)$, THE BETTER $\pi(\theta)$

③ OPTIMIZE $J(\theta)$: $\theta^* := \underset{\theta}{\operatorname{argmax}}\, J(\theta)$

① SOFTMAX PARAMETRIZATION ≡ COMPUTE PREFERENCES $h(s,a,\theta)$ OF STATE-ACTION PAIRS (s,a) (FOR INSTANCE, USING A NN), AND DEFINE ACTION PROBABILITIES BY:

$$\pi(s,a,\theta) := \frac{e^{h(s,a,\theta)}}{\sum_{b \in A} e^{h(s,b,\theta)}}$$

② $J(\theta)$ = "EXPECTED TOTAL REWARD GIVEN BY $\pi(\theta)$"

= "TOTAL REWARD ALONG TRAJECTORIES τ WEIGHTED BY PROB(τ)"

- $\tau = s_0, a_0, r_1, s_1, a_1, \dots$
- $\text{PROB}(\tau) = \underbrace{\rho(s_0)}_{\text{STARTING DISTRIBUTION}} \cdot \pi(a_0, s_0, \theta) \cdot \underbrace{p(s_1, r_1, s_0, a_0)}_{\text{DISTRIBUTION MODEL}} \cdot \pi(a_1, s_1, \theta) \cdot \dots$
- "TOTAL REWARD ALONG τ" $\equiv \sum_{t=0}^{\infty} \gamma^t r_{t+1}$ = "DISCOUNTED RETURN" $=: R(\tau)$
- "EXPECTED TOTAL REWARD" $\equiv$

$$J(\theta) := \sum_{\tau} \text{PROB}(\tau) \cdot R(\tau) = \mathop{\mathbb{E}}_{\tau \sim \text{PROB}_{\rho,\theta}} [R(\tau)]$$

REMARK: IF $d_{s_0}^H(s)$ = "PROBABILITY OF REACHING s FROM s_0 IN H STEPS",

THEN $d_{s_0}(s) := \sum_{H=0}^{\infty} d_{s_0}^H(s)$ = "AVERAGE NUMBER OF VISITS TO s"

GIVES A DISTRIBUTION ON S: $\mu(s) = \dfrac{d_{s_0}(s)}{\sum_s d_{s_0}(s)}$

"ON-POLICY DISTRIBUTION"

$\eta(s)$ ≡ "PROBABILITY OF VISITING s STARTING FROM s_0"

THEOREM $J(\theta)$ CAN BE COMPUTED BY LOOKING ONE STEP FORWARD:

$$J(\theta) = \sum_{s,a} \mu(s) \cdot \pi(a,s,\theta) \cdot r(s,a)$$

"DISTRIBUTIONAL POINT OF VIEW"

$$= \sum_{s} \eta(s) \cdot \left(\sum_{a} \pi(a,s,\theta) \cdot q_{\pi_\theta}(s,a) \right)$$

EXERCISE: WRITE THIS WHEN STARTING FROM A FIXED s_0

(3) OPTIMIZE $J(\theta)$ ⇝ WE NEED DERIVATIVES $\nabla_\theta J = \nabla J$

"GRADIENT OF J WRT θ"

THIS LECTURE!