- Lecture 16 Notes : Policy Gradient Algorithms

  - Generalized Policy Iteration - We don't need complete policy evaluation or greedy policy improvement.

    ↪ Relax the requirement for precise policy evaluation or improvement

    ↪ Idea : Do policy improvement with Gradient Ascent rather than doing an argmax

      ↪ Functional representation of policy function $\pi(s, a; \theta)$ and adjust θ little by little to improve VF

      ↪ We still have func. approx. of Action Value Function $Q(s, a; w)$

$$\pi(s, a; \theta) \quad : \quad \text{"Actor"}$$

$$Q(s, a; w) \quad : \quad \text{"Critic"}$$

      ↪ Critic Parameters w are optimized by loss-function minimization.

      ↪ Actor parameters θ are optimized w.r.t. Expected Returns max

      ↪ Major difference is in policy improvement. No longer just an argmax

- Advantages :

  - Finds best Stochastic Policy

    ↪ Relevant for partially observable MDPs

    ↪ Naturally Explores due to Stochastic policy representation

      ↪ Don't have to worry about explore/exploit

    ↪ Small changes in $\theta \Rightarrow$ Small changes in π

→ This avoids convergence issues seen in argmax-based algorithms

→ see disadvantages on slide 6

- <u>Theory</u> : Assume discrete-time, countable space, stationary MDPs

Notation: $P(s, a, s') = P^a_{s,s'}$; $R(s,a) = R^a_s$

$p_0 : \mathcal{N} \longrightarrow [0,1]$ = Initial State probability distribution

- Expected returns Objective

$$J(\theta) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \cdot R_{t+1} \right] \longleftarrow \text{Unconditional}$$

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{k=t}^{\infty} \gamma^{k-t} R_{k+1} \mid S_t = s \right]$$

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{k=t}^{\infty} \gamma^{k-t} R_{k+1} \mid S_t = s, \ A_t = a \right]$$

Advantage Function $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$

$$J(\theta) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \right] = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_\pi \left[R_{t+1} \right]$$

= ... See Slide 9 for details

... lots of algebra

$$= \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in A} \pi(s, a; \theta) \cdot R^a_s$$

Discounted aggregate State visitation measure (see slide 9)

$\leadsto$ You can get $\rho^{\pi}(s)$ through Sampling.

$$\nabla_\theta J(\theta) = \sum_{s \in N} \rho^{\pi}(s) \cdot \sum_{a \in A} \nabla_\theta \pi (s, a; \theta) \cdot Q^{\pi}(s, a)$$

Notice that this is unaffected by $\nabla \pi$

$\leadsto$ Continue from 40 minute mark, slide 10. Slides offer sufficient notes.
I just followed the discussion from here.

- Policy Gradient Proof was on a previous final