Lecture 3

→ Prediction → You give it an MDP and a policy and it outputs V_π.

→ Control → You are given the MDP and output the optimal value function and policy

---

→ Policy Evaluation → To generate value functions given ~~you~~ the MDP and the policy.

problem → Evaluate a given policy π
Solution → We use bellman expectation equation for evaluation of the policy.

- We start with an arbitrary values of V_0
- We then complete 1 iteration of one step look ahead using the bellman equation
- We repeat this many times till we converge to the V_π.

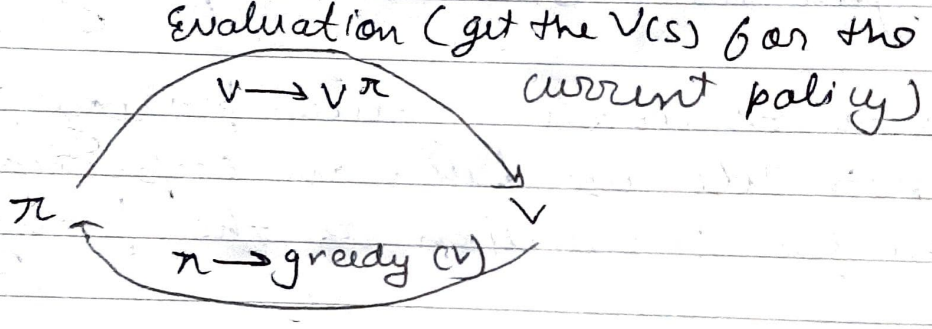
~~This~~ Sychronus Backup
→ At each iteration $K+1$
→ for all states $s \in S$
→ Update $V_{K+1}(s)$ from $V_K(s')$

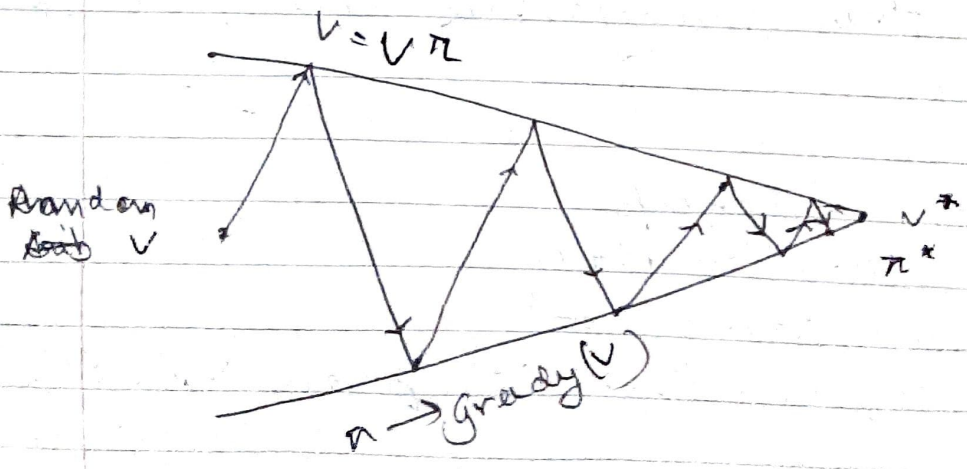
$$V^{K+1} = R^\pi + \gamma P^\pi V^K$$

→ Policy Iteration → make our policy better

two steps
- Evaluate the policy
- Improve the policy → $\pi' = \text{greedy}(V_\pi)$

Follow these iteratively, the process always converges to the optimal policy





policy Evaluation → Estimate V^π → Iterative policy evaluation

policy improvement → ~~Estimate~~ generate $\pi' \geq \pi$ greedy policy improvement

No matter where we start, we will always end up converging to V^* and π^*

→ Greedily → Look at the values of ~~pat~~ taking actions in a particular state and ~~go~~ take the action that ~~giv~~ has the maximum q^π value

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

This improves the value from any states s over one step.

$$q_\pi(s, \pi'(s)) = \max_a q_\pi(s,a) \geq q_\pi(s,\pi(s)) = V^\pi(s)$$

→ Another way of policy iteration →

Eg. Look at the bellman equation ones → update our value function → act greedly w.r.t that value function → and repeat

$$V^*(s) = \max_{a \in A} R_s^a + \gamma \sum_{s' \in S^+} P_{ss'}^a V^*(s')$$

apply this iteratively.

→ Value Iteration - Until now we had a value function for all states and we used that to generate a policy we then used the policy to generate value function

In this algorithm, we do not work with policy. Using bellman optimality equation we, in first step, update all the value functions starting from arbitrary values. We repeat the same using the new value functions until we converge to the optimal value function.

We don't know the optimal policy but we will know the optimal value function