RL Week - 3

L1 Bellman Equations

- Value of state → expected long term reward starting from that state, depends on the agents policy.
- Bellman Eq^n for policy π

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

$$= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma E_\pi[G_{t+1} \mid S_{t+1} = s'] \right]$$

$$= \sum_a \pi(a|s) \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right] \quad \forall s \in S$$

→ linear eq^n in $|S|$ variables
→ unique sol^n exists
→ for the first reward, we calculate the probability, but from next step onwards, take the expected value into account.

→ we assume things markov & stationary

Eg: Ac^{ns} → N, S, W, E . deterministic . If would-take agent off the grid . No movement but reward = 0. Other ac^{ns} produce reward = 0, except spl. ac^{ns}.

State value func^{n} for equiprobable random policy. $\gamma = 0.9$

L2 Bellman Optimality Equa^{ns}

- for finite MDPs, policies can be partially ordered.

$\pi \geq \pi'$ iff. $v_\pi(s) \geq v_{\pi'}(s) \quad \forall s \in S$

- There is always atleast 1 (or many) policies that is better than or equal to all the others. This will be our optimal policy. That is π_*. ($\pi_* \geq \pi \ \forall \pi$)
- Optimal policies share the same optimal state-value func^{n}.

$$v_*(s) = \max_\pi v_\pi(s) \quad \forall s \in S$$

$$q_*(s,a) = \max_\pi q_\pi(s,a) \quad \forall s \in S, a \in A$$

- there is 1 deterministic policy (atleast 1) out of all policies.

Eg. Grid world setting

$$M = \langle S, A, p, \gamma \rangle$$

- many policies, but at least 1 optimal policy.
- Bellman Optimality eq^n for v_*

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s, a)$$

$$= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a]$$

$$= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma \underbrace{G_{t+1}}_{\gamma v_*(S_{t+1})} \mid S_t = s, A, a)$$

$$v_*(s) = \max_a \sum_{s', a} p(s', a \mid s, a)[a + \gamma v_*(s')]$$

- The value of a state under an optimal policy must equal the expected return for the best ac^n from that state.

$$q_*(s, a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a\right]$$

$$= \sum_{s',r} p(s', r | s, a)\left[r + \gamma \max_{a'} q_*(s', a') \right]$$

- Any policy that is greedy w.r.t v_* is an optimal policy. ∴, given v_*, 1 step look ahead search produces the long term optimal actions.

grid world v^*

22.0 22.4 →
19.8 — ·

- given q_*, the agent does not have to do a 1-step ahead search.

$$\pi_*(s) = \arg\max_{a \in A(s)} q_*(s, a)$$

L3 DP Policy Iteration

- Stochastic DP is the solution method of choice for MDPs (greedy, conquer & divide rule)
  1. req. comp. knowledge of system dynamics (transition matrix & reward)
  2. computationally expensive

evaluation	v_π	q_π
control	v_*	q_*

3. curse of dimensionality
4. guaranteed to converge

- Given a policy π, compute the state value func^n v_π.

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \left[r + \gamma v_\pi(s') \right]$$

↳ we iteratively solve it. You will have $|S|$ sized of variables. $|S|$ linear equations. There will be a single unique soln^n. So, these are simultaneous linear equations.

$$v_0 = 0 \quad \forall s$$

$$v_{k+1}(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \left[r + \gamma v_k(s') \right]$$

similarly, calculate v_{k+2} using v_{k+1}. This iteration solves the problem gives approximate results.

Input π

$\theta > 0 \rightarrow$ Algo param

$V(s)$ initialize, $V(\text{terminal}) = 0$

$\Delta \leftarrow 0$

(updates in place)

Loop $s \in S$

$v \leftarrow V(s)$

$V(s) \leftarrow \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) [r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < 0$

L4 Policy Improvement

- Given a π, we can now find v_π.
- For a state s, is it possible to choose ac^n $a \neq \pi(s)$

Value of a in state s —

$$q_\pi(s, a) = E[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = a]$$

$$= \sum_{s', r} p(s', r | s, a)[r + \gamma v_\pi(s')]$$

- So, it is better to switch to ac^n a for state s iff. $q_\pi(s, a) > v_\pi(s)$

- Now we do this for all states, that gives new policy π' (greedy wrt v_π)

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

$$= \arg\max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right]$$

$\because, v_{\pi'} \geq v_\pi \Rightarrow$ Policy Improvement $\forall s \in S$

- What if $v_{\pi'} = v_\pi$, then

$$v_{\pi'}(s) = \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_{\pi'}(s')\right]$$

↳ Bellman optimality eqn again

$\therefore, v_{\pi'} = v_*$ & both π & π' are the optimal policies.

$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \pi_2 \cdots$

$E \rightarrow$ evaluate, $I \rightarrow$ improve
↳ greedification

Policy Iteration