Policy Evaluation

Input: $\pi(a|s)$

Output: $V_\pi(s)$ or $Q_\pi(s,a)$

How to find $V(s)$ using DP (same technique can be applied to $Q(s,a)$)

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s'}\sum_r p(s',r|s,a)\{r + \gamma V_\pi(s')\}$$

Everything is known except for the V's.

Finding $V(s)$ iteratively, first initialize $V_0(s)=0$ or random for all states (0 for terminals)

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

k's are not timesteps in the environment but are in our code

Repeating this again and again $V_\pi(s) = V_\infty(s)$

when do we stop? We know that we approach the answer as $k \to \infty$ but we can't wait for an infinite amount of time

$$\Delta = \max|V_{k+1}(s) - V_k(s)|$$

You get to pick the threshold Δ for your own desire of accuracy

The idea of obtaining the best policy is called policy improvement.

Given a policy, how can I find a better policy?

Assume we're given some π and we've found $V_\pi(s)$ and $Q_\pi(s,a)$

Suppose we take an action not prescribed by the policy for state s

This is what $Q_\pi(s,a)$ tells us! Expected future return for doing 'a' in 's' following π thereafter.

if $Q_\pi(s,a) > V_\pi(s,a)$, then our return for the episode is better than if we had just followed π the whole time.

Making this small change will improve our expected return.

How do we pick the best action 'a' to take? Just look at all the values, pick the one that gives us the max.

$$a^* = \underset{a}{\mathrm{argmax}}\, Q_\pi(s,a)$$

What if we perform this other action (a^*) every time we visit state 's'?

then we have been a new policy, $\pi'(s) \neq \pi(s)$

The RHS of Bellman applies only for π, not π'

Policy Improvement Theorem

$$\left.\begin{array}{l}\text{if } Q_\pi(s,\pi'(s)) \geq V_\pi(s) \\ \text{Then } V_{\pi'}(s) \geq V_\pi(s) \text{ for all } s \in S\end{array}\right\}$$

if we have strict inequality in the first statement we also have in the second.

We discussed changing the action for a single time. What if we perform this process to all states? After this process we will be improving our policy for all states

If we reach a point where $\pi'(s) = \pi(s)$ then it must be true that $V_{\pi'}(s) = V_{\pi}(s)$ Then from the Bell man Optimality equation, if we reach this point and the equation is satisfied then we found the optimal policy

What if we keep on doing policy improvement over and over again? This is the concept of policy Iteration

$$\left.\begin{aligned} V_{\pi_0} &= PE(\pi_0)\\ \pi_1 &= PI(V_{\pi_0})\\ V_{\pi_1} &= PE(\pi_1)\\ \pi_2 &= PI(V_{\pi_1})\\ &\vdots \end{aligned}\right\}$$

We do this process until our policy stops improving → This is called policy iteration

This loop can go on forever, so we can quit when the policy is stable or when the value is stable

Remember, optimal values are unique but optimal policies are not.