III. Policy Iteration & Value Iteration Algos

So far, we have ***offline*** algos. We are also interested in online algos. Next we show the Bellman eqns provide fixed-point eqns for online learning

- - - - - - - -

Consider discounted optimal control formulation

$$\text{minimize } \sum_{k=0}^{\infty} \gamma^k \cdot c(x_k, u_k) \qquad \gamma \in [0,1]$$

$$\text{subject to : } x_{k+1} = f(x_k, u_k)$$

Define $V_\pi(x)$ as the value fcn corresponding to policy π ← may not be optimal.

Note: $V_\pi(x_k) = \sum_{\tau=k}^{\infty} \gamma^{\tau-k} \cdot c(x_\tau, u_\tau)$

① Policy Eval

② Policy Improv.

$$= c(x_k, u_k) + \gamma \cdot \underbrace{\sum_{\tau=k+1}^{\infty} \gamma^{\tau-(k+1)} c(x_\tau, u_\tau)}_{= \gamma \cdot V_\pi(x_{k+1})}$$

$$V_\pi(x_k) = c(x_k, u_k) + \gamma \cdot V_\pi(x_{k+1}) \leftarrow$$

all where $u_k = \pi(x_k)$

Observation & Question: This eqn is implicit in $V_\pi(\cdot)$ and suggests iterative scheme ...

$$V_\pi^{j+1}(x_k) = c(x_k, u_k) + \gamma \cdot V_\pi^{j}(x_{k+1}) \quad ; \quad \text{Start } V_\pi^0(x_k) = 0 \ \forall x_k$$

$j = 0, 1, \ldots$

Q: Does V_π^j converge as $j \to \infty$? A: YES!

Algo 1 (Iterative Policy Evaluation)

To compute the value fcn corresponding to some arbitrary policy π:

For $j = 0, 1, \ldots$

$$V_\pi^{j+1}(x_k) = c(x_k, u_k) + \gamma \cdot V_\pi^j(x_{k+1}) \quad \forall\, x_k \in X$$

where $u_k = \pi(x_k)$

$$V_\pi^0(x_k) = 0 \quad \forall\, x_k \in X$$

Sutton & Barto refer $V_\pi^j(x_k)$ as $j \to \infty$ as a "full backup"

2) Policy Improvement. To improve a given policy, an intuitive idea uses Bellman's Principle of Opt'l. Eqn:

$$\pi^{NEW} = \arg\min_{\pi(\cdot)} \left\{ c(x_k, \pi(x_k)) + \gamma \cdot \underbrace{V_{\pi^{OLD}}(x_{k+1})}_{\text{from policy eval}} \right\}$$

$$\text{where } x_{k+1} = f(x_k, \pi(x_k))$$

Bertsekas [1996] has prove π^{NEW} is improved wrt. π^{OLD} in the sense $V_{\pi^{NEW}}(x_k) \leq V_{\pi^{OLD}}(x_k)$

$$\forall x_k \in X$$

SUMMARY

Model-based mean need to know $f(\circ,\circ)$

Policy Evalution

Given an arbitrary policy π

Find V_π

For $j = 0, 1, \cdots$

$$V_\pi^{j+1}(x_k) = c(x_k, u_k) + \gamma \cdot V_\pi^j(x_{k+1})$$

$$V_\pi^0(x_k) = 0 \ \forall x_k \in X$$

Where $u_k = \pi(x_k)$, $x_{k+1} = f(x_k, u_k)$

Policy Improvement

Given $V_{\pi^{OLD}}$ for some arbitrary policy π^{OLD}, find improved policy π^{NEW}

$$\pi^{NEW} = \arg\min_{\pi(\cdot)} \left\{ c(x_k, \pi(x_k)) + \gamma \cdot V_{\pi^{OLD}}(x_{k+1}) \right\}$$

where $x_{k+1} = f(x_k, \pi(x_k)) \ \forall x_k \in X$