Policy Iteration

Start with a randomly chosen initial policy π_0

Iterate until there is no change in utilities:

1. Policy evaluation, given a policy π_i, calculate the utility $U_i(s)$ of every state s using policy π_i by solving the system of equations:

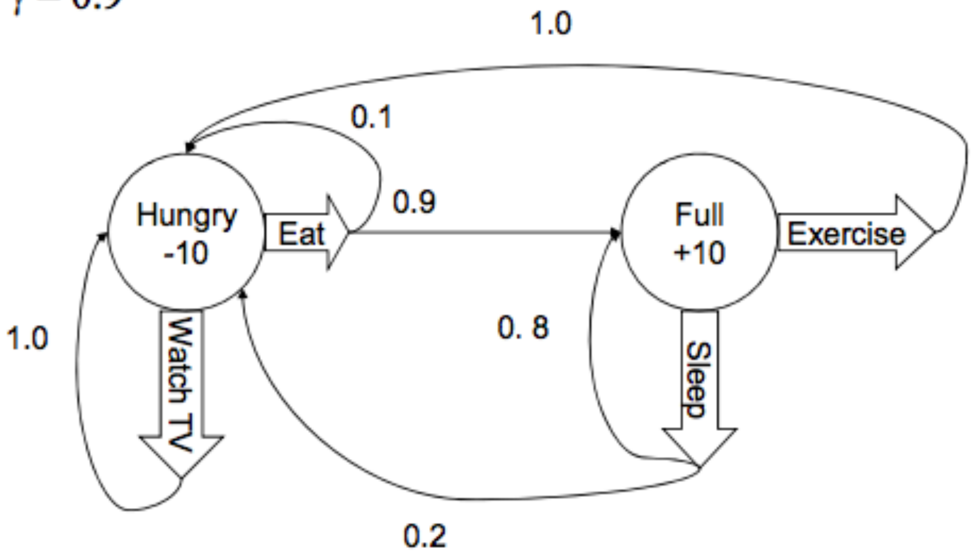
$$U_i(s) = R(s) + \gamma \sum_{s'} P(s' \mid s, \pi_i(s)) U_i(s')$$

2. Policy improvement: calculate the new policy π_i using one-step look-ahead based on U_i (s):

$$\pi^*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{s'} P(s' \mid s, a) U(s')$$

Policy Iteration Example

Do one iteration of policy iteration on the MDP below. Assume an initial policy of π_1(Hungry) = Eat and π_1(Full) = Sleep. Let $\gamma = 0.9$



Policy Iteration Example

Policy Evaluation Phase

Use initial policy for Hungry: π_1(Hungry) = Eat

U_1(Hungry) = -10 + (0.9)[(0.1)U_1(Hungry)+(0.9)U_1(Full)]

$\Rightarrow$$U_1$(Hungry) = -10 + (0.09)$U_1$(Hungry)+(0.81)$U_1$(Full)

$\Rightarrow$(0.91)U_1(Hungry)-(0.81)U_1(Full) = -10

Use initial policy for Full: π_1(Full) = Sleep.

U_1(Full) = 10 + (0.9)[(0.8)U_1(Full) + (0.2)U_1(Hungry)]

$\Rightarrow$$U_1$(Full) = 10 + (0.72)$U_1$(Full) + (0.18)$U_1$(Hungry)]

$\Rightarrow$(0.28)U_1(Full) - (0.18)U_1(Hungry) = 10

Policy Iteration Example

$(0.91)U_1(Hungry)-(0.81)U_1(Full) = -10$ ....(Equation 1)
$(0.28)U_1(Full) - (0.18)U_1(Hungry)=10$ ...(Equation 2)

Solve for $U_1(Hungry)$ and $U_1(Full)$

From Equation 1:

$(0.91)U_1(Hungry) = -10+(0.81)U_1(Full)$

$\Rightarrow U_1(Hungry) = (-10/0.91)+(0.81/0.91)U_1(Full)$

$\Rightarrow U_1(Hungry)=-10.9+(0.89)U_1(Full)$

Policy Iteration Example

$(0.91)U_1(\text{Hungry})-(0.81)U_1(\text{Full}) = -10$ ....(Equation 1)
$(0.28)U_1(\text{Full}) - (0.18)U_1(\text{Hungry})=10$ ...(Equation 2)

Solve for $U_1(\text{Hungry})$ and $U_1(\text{Full})$

Substitute $U_1(\text{Hungry})=-10.9+(0.89)U_1(\text{Full})$ into Equation 2

$(0.28)U_1(\text{Full}) - (0.18)[-10.9+(0.89)U_1(\text{Full})]=10$

$\Rightarrow (0.28)U_1(\text{Full}) + 1.96-(0.16)U_1(\text{Full})=10$

$\Rightarrow (0.12)U_1(\text{Full})=8.04$

$\Rightarrow U_1(\text{Full})=67$

$\Rightarrow U_1(\text{Hungry})=-10.9+(0.89)(67)=-10.9+59.63=48.7$

Policy Iteration Example

- π_2(Hungry) =Eat
- π_2(Full) = Sleep

Policy Iteration Example

$\pi_2(\text{Hungry})$

$$= \underset{\{\text{Eat},\text{WatchTV}\}}{\text{argmax}} \begin{cases} \text{T(Hungry, Eat, Full)U}_1\text{(Full)} + \\ \quad \text{T(Hungry, Eat, Hungry)U}_1\text{(Hungry)} & \text{[Eat]} \\ \text{T(Hungry, WatchTV, Hungry)U}_1\text{(Hungry)} & \text{[WatchTV]} \end{cases}$$

$$= \underset{\{\text{Eat},\text{WatchTV}\}}{\text{argmax}} \begin{cases} (0.9)\text{U1(Full)} + (0.1)\text{U1(Hungry)} & \text{[Eat]} \\ (1.0)\text{U1(Hungry)} & \text{[WatchTV]} \end{cases}$$

$$= \underset{\{\text{Eat},\text{WatchTV}\}}{\text{argmax}} \begin{cases} (0.9)(67) + (0.1)(48.7) & \text{[Eat]} \\ (1.0)(48.7) & \text{[WatchTV]} \end{cases}$$

$$= \underset{\{\text{Eat},\text{WatchTV}\}}{\text{argmax}} \begin{cases} 65.2 & \text{[Eat]} \\ 48.7 & \text{[Watch]} \end{cases}$$

$= \text{Eat}$

Policy Iteration Example

$$\pi_2(\text{Full})$$

$$= \underset{\{\text{Exercise,Sleep}\}}{\text{argmax}} \begin{Bmatrix} \text{T(Full, Exercise, Hungry)U}_1\text{(Hungry)} & \text{[Exercise]} \\ \text{T(Full, Sleep, Full)U}_1\text{(Full)} + & \\ \quad \text{T(Full, Sleep, Hungry)U}_1\text{(Hungry)} & \text{[Sleep]} \end{Bmatrix}$$

$$= \underset{\{\text{Exercise,Sleep}\}}{\text{argmax}} \begin{Bmatrix} (1.0)\text{U}_1\text{(Hungry)} & \text{[Exercise]} \\ (0.8)\text{U}_1\text{(Full)} + (0.2)\text{U}_1\text{(Hungry)} & \text{[Sleep]} \end{Bmatrix}$$

$$= \underset{\{\text{Exercise,Sleep}\}}{\text{argmax}} \begin{Bmatrix} (1.0)(48.7) & \text{[Exercise]} \\ (0.8)(67) + (0.2)(48.7) & \text{[Sleep]} \end{Bmatrix}$$

$$= \underset{\{\text{Exercise,Sleep}\}}{\text{argmax}} \begin{Bmatrix} 48.7 & \text{[Exercise]} \\ 63.34 & \text{[Sleep]} \end{Bmatrix}$$

$$= \text{Sleep}$$