Lecture 16

* Policy Iteration :

Algorithm

Step 1 : (Initialization) Guess an initial policy μ^0

Step 2 : (Policy evaluation) Given the stationary policy μ^k; Compute corresponding average and differential costs λ^k and $h^k(i)$ satisfying

$$\lambda^k + h^k(i) = g(i, \mu^k(i)) + \sum_{j=1}^{n} P(i, \mu^k(i), j)\, h^k(j) \quad \forall\, i \in [n]$$

$$h^k(n) = 0$$

Step 3 : (Policy improvement) Obtain a new stationary policy μ^{k+1} satisfying, where $\forall$ i, $\mu^{k+1}(i)$ is such that

$$g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} P(i, \mu^{k+1}(i), j)\, h^k(j)$$
$$= \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} P(i,u,j)\, h^k(j) \right\}$$

If $\lambda^{k+1} = \lambda^k$ and $h^{k+1}(i) = h^k(i)\ \forall\, i$, then stop else return to step 2 and repeat.

* The algorithm is based only the following proposition :

Proposition 2 : Under assumption 1, in the policy iteration algo, for each k we either have

$$\lambda^{k+1} < \lambda^k$$

or else we have

$$\lambda^{k+1} = \lambda^k, \quad h^{k+1}(i) \leq h^k(i), \quad i \in [n]$$

Furthermore, the algorthm terminates and the policies μ^k and μ^{k+1} obtained upon termination are optimal.

Proof: To simplify notation, denote $\mu^k = \mu$, $\mu^{k+1} = \bar{\mu}$, $\lambda^k = \lambda$, $\lambda^{k+1} = \bar{\lambda}$, $h^k = h$, $h^{k+1} = \bar{h}$. Define for $N = 1, 2, \ldots$

$$h_N(i) = g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\, h_{N-1}(j), \quad i \in [n]$$

where $h_0(i) = h(i)$

Note that $h_N(i)$ is the N-stage cost of policy $\bar{\mu}$ starting from i when the termination cost function is h. Thus we have

$$\bar{\lambda} = J_{\bar{\mu}}(i) = \lim_{N \to \infty} \frac{1}{N} h_N(i) \quad \forall\, i \in [n]$$

Since the contribution of the terminal cost function vanishes as $N \to \infty$. By definition of $\bar{\mu}$ and by Prop 1(c), we have $\forall\, i$

$$\begin{aligned} h_1(i) &= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\, h_0(j) \\ &\leq g(i, \mu(i)) + \sum_{j=1}^{n} P(i, \mu(i), j)\, h_0(j) \\ &= \lambda + h_0(i) \end{aligned}$$

From the above equation we also obtain

$$\begin{aligned} h_2(i) &= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\, h_1(j) \\ &\leq g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\,(\lambda + h_0(j)) \\ &\leq \lambda + g(i, \mu(i)) + \sum_{j=1}^{n} P(i, \mu(i), j)\, h_0(i) \\ &= 2\lambda + h_0(i) \end{aligned}$$

and proceeding in this way, $\forall\, i$ and N we have

$$\frac{1}{N} h_N(i) \leq \lambda + \frac{1}{N} h_0(i)$$

and by taking limit as $N \to \infty$, we get $\bar{\lambda} \leq \lambda$.

If $\bar{\lambda} = \lambda$ then μ^{k+1} is a policy improvement step for assoc. SSPP with cost per stage

$$g(i, u) - \lambda$$

Furthermore, $h(i)$ and $\bar{h}(i)$ are the optimal costs starting from

Furthermore, $h(i)$ and $\bar{h}(i)$ are the optimal costs starting from i and corresponding to μ and $\bar{\mu}$, respectively, is associated SSPP. Thus by a previous proposition

$$\bar{h}(i) \leq h(i) \quad \forall\, i.$$

Let us now show that when the algorithm terminates with $\bar{\lambda} = \lambda$ and $\bar{h}(i) = h(i)$ $\forall\, i \in [n]$, μ and $\bar{\mu}$ are optimal.

We have then $\forall\, i$

$$\begin{aligned} \lambda + h(i) = \bar{\lambda} + \bar{h}(i) &= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} p(i, \bar{\mu}(i), j)\, \bar{h}(j) \\ &= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} p(i, \bar{\mu}(i), j)\, h(j) \\ &= \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, h(j) \right\} \end{aligned}$$

Thus λ and h satisfy Bellman's equation. By prop 1, λ must be equal to optimal average cost.

Furthermore, $\bar{\mu}(i)$ attains minimum of r.h.s. of the Bellman's equation $\forall i$. So $\bar{\mu}$ is optimal. ▨