Lecture 10

* Policy Iteration:

- Value iteration / DP iteration is not guaranteed to terminate finite time. Only the limiting iterate is the optimal value.
- Though we can derive a number of iteration dependent error bound.

* Policy iteration is one alternative of value iteration which ends in finite time.

The algorithm is as follows:

1. Start with a stationary policy, μ^0. $k \leftarrow 0$
2. do:
3. evaluate J_{μ^k} as the solution of the system of linear equations:
$$J(i) = g(i, \mu^k(i)) + \sum_{j=1}^{n} p(i, \mu^k(i), j)\, J(j), \quad i \in [n]$$
4. $$\mu^{k+1}(i) \in \arg\min_{u \in \mathcal{U}(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, J_{\mu^k}(j) \right\}, \quad i \in [n]$$
5. $k \leftarrow k+1$
6. while $J_{\mu^k} \neq J_{\mu^{k-1}}$

Proposition 2: Under assumption 1, the policy iteration algorithm for the SSPP generates an improving sequence of policies, i.e, $J_{\mu^{k+1}}(i) \leq J_{\mu^k}(i)$ $\forall\, i$ and k and terminates with an optimal policy.

Proof: For any k, consider the sequence generated by the recursion

$$J_{N+1}(i) = g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} p(i, \mu^{k+1}(i), j)\, J_N(j), \quad i \in [n]$$

$$J_{N+1}(i) = g(i, \mu^{k+1}(i)) + \sum_{j=1} p(i, \mu^{k+1}(i), j)\, J_N(j), \quad i \in [n]$$

where $N = 0, 1, \ldots$, and

$$J_0(i) = J_{\mu^k}(i), \quad i = 1, \ldots, n$$

So, from the algo,

$$\begin{aligned} J_0(i) &= g(i, \mu^k(i)) + \sum_{j=1}^{n} p(i, \mu^k(i), j) \cdot J_0(j) \\ &\geq g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} p(i, \mu^{k+1}(i), j)\, J_0(j) \\ &= J_1(i) \qquad \forall\, i \in [n] \end{aligned}$$

By using the above inequality we obtain

$$\begin{aligned} J_1(i) &= g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} p(i, \mu^{k+1}(i), j)\, J_0(j) \\ &\geq g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} p(i, \mu^{k+1}(i), j)\, J_1(j) \\ &= J_2(i) \qquad \forall\, i \in [n] \end{aligned}$$

and by continuing like this we get

$$J_0(i) \geq J_1(i) \geq \cdots \geq J_N(i) \geq J_{N+1}(i) \geq \cdots, \quad \forall\, i \in [n]$$

Since by Prop. 1, $\lim_{N \to \infty} J_N(i) = J_{\mu^{k+1}}(i) \;\; \forall\, i \in [n]$,

$$J_{\mu^{k+1}}(i) \leq J_{\mu^k}(i) \qquad \forall\, i \in [n]$$

Now if $J_{\mu^k}(i) = J_{\mu^{k+1}}(i) \;\; \forall\, i \in [n]$ then

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

Thus J_{μ_k} solves Bellman's equation and by prop 1, μ^k is optimal. ▨

* Infinite Horizon Discounted Cost problem

- In this section we are interested to minimize

$$J_\pi(x_0) = \lim_{N \to \infty} \mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), x_{k+1}) \right]$$

$$J_\pi(x_0) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), x_{k+1})\right]$$

where $\alpha \in (0,1)$.

- We assume that g is absolutely bounded.

* We can have the results we obtained in the previous section for this setting as well. The proof technique is exactly the same (and somewhat simpler). So we will leave the proof of the following proposition for Homework.

<u>Proposition 1</u>: The following hold for the discounted cost problem:

a. The value iteration algorithm

$$J_{k+1}(i) = \min_{u\in \mathcal{U}(i)} \left\{ g(i,u) + \alpha \sum_{j=1}^{n} p(i,u,j)\, J_k(j) \right\}$$

Converges to the optimal costs $J^*(i)$, $i=1,\dots,n$, starting from arbitrary initial conditions $J_0(1),\dots,J_0(n)$.

b. The optimal costs $J^*(1),\dots,J^*(n)$ of the discounted problem satisfy Bellman's equation,

$$J^*(i) = \min_{u\in \mathcal{U}(i)} \left\{ g(i,u) + \alpha \sum_{j=1}^{n} p(i,u,j)\, J^*(i) \right\}$$

and in fact they are the unique solution of this equation.

c. For any stationary policy μ, the costs $J_\mu(1),\dots,J_\mu(2)$ are the unique solution of the equation

$$J_\mu(i) = g(i,\mu(i)) + \alpha \sum_{j=1}^{n} p(i,\mu(i),j)\, J_\mu(j), \quad i \in [n]$$

Furthermore, given any initial condition $J_0(1),\dots,J_0(n)$, the Sequence $J_k(i)$ generated by DP iteration

$$J_{k+1}(i) = g(i,\mu(i)) + \alpha \sum_{j=1}^{n} p(i,\mu(i),j)\, J_k(j), \quad i\in[n]$$

Converges to the cost $J_\mu(i)$ for every $i \in [n]$.

d. A stationary policy μ is optimal if and only if for every

d. A stationary policy μ is optimal if and only if for every state i, $\mu(i)$ attains the minimum in Bellman's equation.

e. The policy iteration algorithm given by

$$\mu^{k+1}(i) \in \underset{u \in U(i)}{\operatorname{argmin}} \left\{ g(i,u) + \alpha \sum_{j=1}^{n} p(i,u,j) \, J_{\mu^k}(j) \right\}, \quad i \in [n]$$

generates an improving sequence of policies and terminates with an optimal policy.

Proof: HW.

- We will take another route to prove the above results in the next few lectures.