Lecture 13

* Policy evaluation of finite state systems:

WLOG assume $S = [n]$ and fix a stationary policy μ.
Let us define the following vectors and matrices:

$$J := [J(1) \dots J(n)]^T, \quad TJ := [(TJ)(1) \dots (TJ)(n)]^T$$

$$T_\mu J := [(T_\mu J)(1) \dots (T_\mu J)(n)]^T$$

$$P_\mu := \begin{bmatrix} P(1,\mu(1),1) & \dots & P(1,\mu(1),n) \\ \vdots & & \\ P(n,\mu(n),1) & \dots & P(n,\mu(n),n) \end{bmatrix}$$

and $\quad g_\mu := [g(1,\mu(1)), \dots, g(n,\mu(n))]^T$

So now we can write

$$T_\mu J = g_\mu + \alpha P_\mu J$$

and by cor. 2 J_μ is the solution of the equation

$$J = g_\mu + \alpha P_\mu J$$

or, $\quad J_\mu = (I - \alpha P_\mu)^{-1} g_\mu$

Note that $I - \alpha P_\mu$ is always invertible as P_μ is stochastic matrix.

* Policy Iteration:

Algorithm

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

Step 2: (Policy evaluation) Given the stationary policy μ^k, Compute J_{μ^k} as the solution of

$$(I - \alpha P_{\mu^k}) J = g_{\mu^k}$$

Step 3: (Policy improvement) Obtain a new stationary policy μ^{k+1} satisfying

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$$

If $J_{\mu^k} = T J_{\mu^k}$ then stop else return to step 2 and repeat.

* The algorithm is based only the following proposition:

Proposition 6: Let μ and $\bar{\mu}$ be stationary policies such that $T_{\bar{\mu}} J_\mu = T J_\mu$ or, equivalently for $i = 1, \ldots, n$,

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

Then we have

$$J_{\bar{\mu}}(i) \leq J_\mu(i), \quad i = 1, \ldots, n$$

Furthermore if μ is not optimal then the inequality is strict for at least one i.

Proof: Since $J_\mu = T_\mu J_\mu$ and, by hypothesis $T_{\bar{\mu}} J_\mu = T J_\mu$. We have for every i,

$$\begin{aligned} J_\mu(i) &= g(i, \mu(i)) + \alpha \sum_{j=1}^{n} p(i, \mu(i), j)\, J_\mu(j) \\ &\geq g(i, \bar{\mu}(i)) + \alpha \sum_{j=1}^{n} p(i, \bar{\mu}(i), j)\, J_\mu(j) \\ &= (T_{\bar{\mu}} J_\mu)(i) \end{aligned}$$

Applying this repeatedly we get

$$J_\mu \geq T_{\bar{\mu}} J_\mu \geq \cdots \geq T_{\bar{\mu}}^k J_\mu \geq \cdots \geq \lim_{N \to \infty} T_{\bar{\mu}}^N J_\mu = J_{\bar{\mu}}$$

If $J_\mu = J_{\bar{\mu}}$ then

$$T_{\bar{\mu}} J_{\bar{\mu}} = T J_{\bar{\mu}}$$

$$T_{\bar{\mu}} J_{\bar{\mu}} = T J_{\bar{\mu}}$$

by Prop 5, $J_{\bar{\mu}} = J^*$. Thus μ must be optimal. ($\Rightarrow \cancel{\Leftarrow}$) $\boxtimes$