EE 223: Stochastic Systems: Estimation and Control	Spring 2007
Lecture 22 — April 10	
Lecturer: Venkat Anantharam	*Scribe: Lei Shi*

22.1 Average Cost DP

$$P(u) = [p_{ij}(u)] \qquad i,j \in \mathcal{X} \qquad u \in \mathcal{U} \qquad |\mathcal{X}| = d \qquad \mathcal{U} \text{ finite}, \tag{22.1}$$

Aim: Minimize over Ψ (causal control strategy):

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} E[g(X_k^\Psi, U_k^\Psi)] \tag{22.2}$$

We already know that (under some conditions) optimal strategies exist in the class of stationary Markov strategies. More specifically, consider

$$\lambda + h(i) = \min_u [g(i,u) + \sum p_{ij}(u)h(j)] \tag{22.3}$$

(to be solved for λ and $(h(1), ..., h(d))$). If a solution exists, then λ is the optimal average cost and any minimizing $\mu : \mathcal{X} \mapsto \mathcal{U}$ gives an optimal strategy.

Assuming the existence of optimal strategies within the class of stationary Markov strategies, one can use a linear programming approach to find an optimal strategy using the "ergodic control" idea. (One of the early papers in this area is that of Manne, "Linear programming and sequential decisions", *Management Science*, Vol. 6, pp. 259 -267, 1960. It is also developed in the book of Derman, *Finite State Markovian Decision Processes*, Academic Press, 1970. It has been popularized in recent years by Vivek Borkar. For instance, see Chapter 11 of *Handbook of Markov Decision Processes*, edited by Eugene Feinberg and Adam Shwartz – this chapter is written by Borkar.)

Assume $P_\pi = [\sum_u p_{ij}\pi(i,u)]$ is irreducible for all stationary Markov randomized strategies (defined by $\pi(i,u), i \in \mathcal{X}, u \in \mathcal{U}$, where $\sum_u \pi(i,u) = 1$ for all i.)

There is a unique invariant probability distribution $\eta_\pi = [\eta_\pi(1), ..., \eta_\pi(d)]$, s.t. $\eta_\pi P_\pi = \eta_\pi$. Let $\nu_\pi(i,u) = \eta_\pi(i) \cdot \pi(i,u)$. This ν_π, a probability distribution on $\mathcal{X} \times \mathcal{U}$, is called the stationary occupation measure corresponding to π.)

Fact: the set of all ν_π as π ranges over all stationary randomized Markov strategies is a closed convex polytope whose extreme points (vertices) are the stationary occupation measures associated to deterministic stationary Markov strategies.

Consequence: we can consider the linear program

$$\min_{\nu_\pi} \sum_{i,u} g(i,u)\nu_\pi(i,u) \tag{22.4}$$

The solution will occur at an extreme point. This gives an optimal stationary (deterministic) Markov strategy. Of course, the optimum strategy need not be unique in general.

We show that the set of stationary occupation measures associated to stationary randomized Markov strategies is convex. The fact that it is closed comes from a simple continuity argument. Let $\pi^{(0)}$ and $\pi^{(1)}$ define two randomized stationary Markov strategies, let $\lambda \in (0,1)$, and let $\pi^{(\lambda)} = \lambda\pi^{(1)} + (1-\lambda)\pi^{(0)}$. Let $\eta_{\pi^{(0)}}, \eta_{\pi^{(1)}}$ be invariant for $\pi^{(0)}, \pi^{(1)}$ respectively.

$$\nu_{\pi^{(0)}}(i,u) = \eta_{\pi^{(0)}}(i)\pi^{(0)}(i,u) \text{ and } \nu_{\pi^{(1)}}(i,u) = \eta_{\pi^{(1)}}(i)\pi^{(1)}(i,u) \tag{22.5}$$

We'd like to show that $\lambda\nu_{\pi^{(1)}} + (1-\lambda)\nu_{\pi^{(0)}}$ is also a stationary occupation measure associated to a stationary randomized Markov strategy.

Let $\widetilde{\eta} = \lambda\eta_{\pi^{(1)}} + (1-\lambda)\eta_{\pi^{(0)}}$, and look at the randomized stationary Markov strategy defined by

$$\widetilde{\pi}(i,u) = \frac{\lambda\eta_{\pi^{(1)}}(i)}{\widetilde{\eta}(i)}\pi^{(1)}(i,u) + \frac{(1-\lambda)\eta_{\pi^{(0)}}(i)}{\widetilde{\eta}(i)}\pi^{(0)}(i,u) \tag{22.6}$$

Note that this strategy need not be the one associated to $\pi^{(\lambda)}$. We can check that $\sum_u \widetilde{\pi}(i,u) = 1$ for all i to confirm that this indeed defines a stationary randomized Markov strategy.

A straightforward calculation shows that $\widetilde{\eta}$ is the invariant distribution of $\widetilde{\pi}$. Also

$$\widetilde{\eta}(i)\widetilde{\pi}(i,u) = \lambda\nu_{\pi^{(1)}}(i,u) + (1-\lambda)\nu_{\pi^{(0)}}(i,u)\ . \tag{22.7}$$

Thus we have verified that the stationary occupation measure corresponding to $\tilde{\pi}$ is $\lambda\nu_{\pi^{(1)}} + (1-\lambda)\nu_{\pi^{(0)}}$.

Next we argue that every extreme point of this convex set of stationary occupation measures corresponds to a deterministic stationary Markov strategy. Suppose π is a randomized strategy for which, for some $\lambda \in (0,1)$, we have

$$\pi(1,u) = \lambda\pi^{(1)}(1,u) + (1-\lambda)\pi^{(0)}(1,u) \tag{22.8}$$

for all u, where $\pi^{(1)} \neq \pi^{(0)}$, and further, $\pi(i,u) \equiv \pi^{(1)}(i,u) \equiv \pi^{(0)}(i,u)$ for all $i \neq 1$ for all $u \in \mathcal{U}$. We need to argue that ν_π cannot be an extreme point of the convex set of stationary occupation measures that we are discussing.

Let $\eta_\pi, \eta_{\pi^{(1)}}$ and $\eta_{\pi^{(0)}}$ be the respective invariant measures. We will define γ via

$$\frac{\gamma\eta_{\pi^{(1)}}(1)}{\gamma\eta_{\pi^{(1)}}(1) + (1-\gamma)\eta_{\pi^{(0)}}(1)} = \lambda \tag{22.9}$$

Let $\widetilde{\eta} = \gamma\eta_{\pi^{(1)}} + (1-\gamma)\eta_{\pi^{(0)}}$. It is straightforward to check that $\widetilde{\eta}$ is the invariant distribution associated to π, i.e.$\eta_\pi = \widetilde{\eta}$. Note that $\widetilde{\eta}$ need not equal $\lambda\eta_{\pi^{(1)}} + (1-\lambda)\eta_{\pi^{(0)}}$ in general.

We thus have

$$\begin{aligned}
\nu_\pi(i,u) &= \widetilde{\eta}(i)\pi(i,u) \\
&= \lambda\widetilde{\eta}(i)\pi^{(1)}(i,u) + (1-\lambda)\widetilde{\eta}(i)\pi^{(0)}(i,u) \\
&= \gamma\eta_{\pi^{(1)}}(i)\pi^{(1)}(i,u) + (1-\gamma)\eta_{\eta^{(0)}}(i)\pi^{(0)}(i,u) \\
&= \gamma\nu_{\pi^{(1)}}(i,u) + (1-\gamma)\nu_{\eta^{(0)}}(i,u)\ .
\end{aligned}$$

We have thus managed to express ν_π as a convex combination of $\nu_{\pi^{(1)}}$ and $\nu_{\pi^{(0)}}$. If $\nu_{\pi^{(1)}} \neq \nu_{\pi^{(0)}}$ this would imply that ν_π is not an extreme point of the convex set of stationary occupation measures associated to stationary randomized Markov strategies. This is because, by definition, no extreme point of a closed convex set can be expressed as a nontrivial convex combination of two distinct elements in the set.

To complete the characterization of the extreme points, we verify that $\nu_{\pi^{(1)}} \neq \nu_{\pi^{(0)}}$. Indeed, if we had $\nu_{\pi^{(1)}} = \nu_{\pi^{(0)}}$, then for all i and all u we would have $\eta_{\pi^{(1)}}(i)\pi^{(1)}(i,u) = \eta_{\pi^{(0)}}(i)\pi^{(0)}(i,u)$. Since $\pi^{(1)}(i,u) = \pi^{(0)}(i,u)$ for all $i \neq 1$ and all u, this implies that $\eta_{\pi^{(1)}}(i) = \eta_{\pi^{(0)}}(i)$ for all $i \neq 1$. But then we would have $\eta_{\pi^{(1)}}(1) = \eta_{\pi^{(0)}}(1)$. This would then imply that $\pi^{(1)}(1,u) = \pi^{(0)}(1,u)$ for all u. But we started out by assuming that this was not true. Thus it must be the case that $\nu_{\pi^{(1)}} \neq \nu_{\pi^{(0)}}$.

We now recall linear programming duality

Table 22.1. LP duality

$\min c^T x$	$\max \pi^T b$
$a_i^T x = b_i, i \in M$	π_i unrestricted, $i \in M$)
$a_i^T x \geq b_i, i \in \bar{M}$	$\pi_i \geq 0, i \in \bar{M}$
$x_j \geq 0, j \in N$	$\pi^T A_j \leq C_j, j \in N$
x_j unrestricted for $j \in \bar{N}$	$\pi^T A_j = C_j, j \in \bar{N}$

Here A_j are the columns of the matrix whose rows are a_i

The natural LP for the average cost DP problem comes from:
$\lambda + h(i) \leq g(i,u) + \sum_j p_{ij}(u)h(j)$ for all u, and for all i

This can be written as the problem of maximizing λ, (the variables are $\lambda, h(1), \cdots, h(d)$), where $\lambda, h(1), \cdots, h(d)$ are all unrestricted, subject to the constraints:

$$\begin{bmatrix} \lambda & h(1) & \cdots & h(d) \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ \cdots & -p_{ij}(u) & \cdots \\ \cdots & 1-p_{ii}(u) & \cdots \end{bmatrix} \leq \begin{bmatrix} \cdots & g(i,u) & \cdots \end{bmatrix}$$

The dual LP is to minimize

$$\begin{bmatrix} \cdots & g(i,u) & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \nu(i,u) \\ \vdots \end{bmatrix}$$

subject to

$$\begin{bmatrix} 1 & \cdots & 1 \\ \cdots & -p_{ij}(u) & \cdots \\ \cdots & 1-p_{ii}(u) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \vdots \\ \nu(i,u) \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where $\nu(i,u) \geq 0$, for all (i,u).

Thus the linear program defined by the ergodic control approach to the long term average cost DP problem is the LP dual of the natural LP for the problem that comes from the corresponding average cost Bellman equation.