| **EE223: Stochastic Systems: Estimation and Control.** | **SP'07** |
|---|---|
| Lecture 21 — April, 05 | |
| *Lecturer: Venkat Anantharam* | *scribe: Assane Gueye* |

In this lecture, we continue to explore the average cost DP problem started in lecture 20. We have already seen the main result of average cost DP and we have proven it. We will now derive a sufficient condition under which the hypothesis of the main result holds.

Let's first recall the model adopted for the average cost DP.

## 21.1 Average Cost Dynamic Programming

We consider a finite state controlled Markov chain with transition probability matrices

$$P(u) = [p_{ij}(u)], \quad i, j \in \mathfrak{X}, \text{ finite, and } u \in \mathfrak{U} \text{ finite} \tag{21.1}$$

Our aim is to minimize over all strategies $\Psi$ the average cost

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} g(X_k^\Psi, U_K^\Psi)$$

where $g(i,u)$ is the expected one step cost incurred when the system is in state $i$ and control $u$ is applied.

**Main Result:**
*Suppose that there exist $\lambda \in R$ and a vector $h = [h(1), \ldots, h(d)]^T$ ($d = |\mathcal{X}|$) such that*

$$\lambda + h(i) = \min_u \left[ g(i,u) + \sum_{j=1}^{d} p_{ij}(u) h(j) \right] \tag{21.2}$$

*holds for all $1 \le i \le d$.*
*Then $\lambda$ is the optimal expected cost and any $\mu : \mathfrak{X} \to \mathfrak{U}$ such that for all $i$ $\mu(i)$ minimizes in the $i$'th equation in (21.2) defines an optimal stationary Markov strategy.*

For a proof of this main result, refer to lecture 20 or to [1]. In the next section, we will study the optimality conditions for the average cost DP.

## 21.2 Optimality conditions: Blackwell Optimal Policies

The main result shows that the solutions of the average cost DP problem can be identified if we could find $\lambda$ (the optimal average cost) and $h$ (an associated vector of differential

costs) satisfying the average cost DP equation (average cost Bellman equation). However, it provides no assurance on the existence of $\lambda$ and $h$. In this section, we will study conditions under which the existence is guaranteed. For that we need to first define the notion of Blackwell optimality.

**Definition 21.1.** *Blackwell Optimal Strategy*
*A* Blackwell optimal strategy *is one given by a function $\mu : \mathfrak{X} \to \mathfrak{U}$ such that $\mu$ defines an optimal strategy in every $\alpha$-discounted problem (for the given Markov Chain and per stage cost $g(i,\cdot)$) for $\alpha \in (\bar{\alpha}, 1)$, where $0 < \bar{\alpha} < 1$ is some number.*

**Claim 21.2.** *A Blackwell optimal strategy always exists.*

A proof of this claim is given in [1], pg. 202.

Now we will derive a sufficient condition for the existence of the $\lambda$ and $h$ hypothesized in the main result. We will first recall, as an aside, some results about Markov chain.
Given a finite state MC with transition probability $P$, the following limit exists and is well defined.

$$P^* = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k \tag{21.3}$$

Write $P^* = \left[p^*_{ij}\right]$. The intuition is that $p^*_{ij}$ is the long-run proportion of time that the chain will be in state $j$ given that it started in state $i$. This intuition holds irrespective of the number of communicating classes: from a transient state the chain will eventually end up in one of the recurrent communicating classes, $p^*_{ij}$ will be non-zero only if $j$ lies in a recurrent communicating class, and for each $i$ the sum of $p^*_{ij}$ over all $j$ in a given recurrent communicating class will give the probability with which the chain started at $i$ eventually ends up in that recurrent communicating class. See any book on Markov chains for more details.
Furthermore, we have that

$$\left\| \frac{1}{D} \sum_{k=0}^{n-1+D} P^k - P^* \right\| < C\gamma^n, \quad \text{for } C < \infty \text{ and } \gamma < 1.$$

where $D$ is the l.c.m of the periods of the recurrent communicating classes.
This last inequality implies that

$$H = \sum_{k=0}^{\infty} (P^k - P^*)$$

is well defined. Its $(i,j)$ entry can be thought of as giving the expected excess over stationarity of the number of visits to state $j$, given that the chain started in state $i$.

By re-writing the expression for $H$, one can get

$$\begin{aligned} H &= \sum_{k=0}^{\infty}(P^k - P^*) \\ &= (I - P^*) + \sum_{k=1}^{\infty}(P^k - P^*) \\ &= (I - P^*) + P\sum_{k=0}^{\infty}(P^k - P^*) \qquad (21.4) \\ &= I - P^* + PH \qquad (21.5) \end{aligned}$$

where (21.4) comes from the fact that $PP^* = P^*P = P^*P^* = P^*$. For more details about this fact, refer to [1] page 195. From (equation 21.5), we see that $H$ satisfies

$$P^* + H = I + PH \qquad (21.6)$$

Notice that equality (21.6) (a matrix equation) has the same shape as equation (21.2).

Now, notice that for any stochastic matrix $P$ we have

$$\begin{aligned} (I - \alpha P)^{-1} &= \sum_{k=0}^{\infty}\alpha^k P^k \\ &= \left(\sum_{k=0}^{\infty}\alpha^k\right)P^* + \sum_{k=0}^{\infty}\alpha^k(P^k - P^*) \\ &= (1-\alpha)^{-1}P^* + H + \sum_{k=0}^{\infty}\alpha^k(P^k - P^*) - \sum_{k=0}^{\infty}(P^k - P^*) \\ &= (1-\alpha)^{-1}P^* + H + \sum_{k=0}^{\infty}(\alpha^k - 1)(P^k - P^*) \\ &= (1-\alpha)^{-1}P^* + H + \Gamma_\alpha\ , \qquad (21.7) \end{aligned}$$

where $P^*$ is defined in terms of $P$ as in equation (21.3). Note that $\Gamma_\alpha$, as defined (21.7), tends to zero as $\alpha \to 1$.

The $\alpha$-discounted overall expected cost for the stationary strategy $\mu$ is given by:

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

where $J_{\alpha,\mu} = [J_{\alpha,\mu}(1), \ldots, J_{\alpha,\mu}(d)]^T$, $g_\mu = [g(1,\mu(1)), \ldots, g(d,\mu(d))]^T$, and $P_\mu$ has entries $p_{ij}(\mu(i))$.

Further, for a given strategy $\mu : \mathfrak{X} \to \mathfrak{U}$, if we let $J_\mu$ denote the associated long term average cost i.e.

$$J_\mu = P^*_\mu g_\mu \Leftrightarrow J_\mu(i) = \sum_{k=1}^{d} p_{i,j}(\mu(i)) g_\mu(j) \ , \text{ for all } i,$$

and let $P^*_\mu$ be defined in terms of $P_\mu$ as in equation (21.3), then we have

$$\begin{aligned} J_{\alpha,\mu} &= (I - \alpha P_\mu)^{-1} g_\mu \\ &= \left((1-\alpha)^{-1} P^*_\mu + H_\mu + \Gamma_{\alpha,\mu}\right) g_\mu \\ &= (1-\alpha)^{-1} P^*_\mu g_\mu + H_\mu g_\mu + \Gamma_{\alpha,\mu} g_\mu \\ &= (1-\alpha)^{-1} J_\mu + h_\mu + o(1-\alpha) \ . \end{aligned} \tag{21.8}$$

Here $H_\mu$ and $\Gamma_{\alpha,\mu}$ are associated to $P_\mu$ as in equation21.7. Also, $h_\mu = H_\mu g_\mu$. Finally, $o(1-\alpha)$ is a function that tends to zero as $\alpha \to 1$, and we may take this to happen uniformly over all $\mu$ because there are only finitely many possibilities for $\mu$.
Equality (21.8) holds for every strategy $\mu$ and $\alpha \in (0,1)$. In particular, for a Blackwell optimal strategy $\mu^*$, we have

$$J_{\alpha,\mu^*} = (1-\alpha)^{-1} J_{\mu^*} + h_{\mu^*} + o(1-\alpha) \tag{21.9}$$

Also, by applying equality (21.6) to $g_\mu$ we get

$$J_\mu + h_\mu = g_\mu + P_\mu h_\mu \text{ for every strategy } \mu. \tag{21.10}$$

Now consider the Blackwell strategy $\mu^*$. For all $\alpha \in (\bar{\alpha}, 1)$ and for every strategy $\mu$ we have

$$g_{\mu^*} + \alpha P_{\mu^*} J_{\alpha,\mu^*} \leq g_\mu + \alpha P_\mu J_{\alpha,\mu^*}$$

Combining this inequality with (21.9), we obtain (after arranging the terms)

$$\begin{aligned} &0 \leq g_\mu - g_{\mu^*} + \alpha \left(P_\mu - P_{\mu^*}\right) J_{\alpha,\mu^*} \\ \Leftrightarrow\ &0 \leq g_\mu - g_{\mu^*} + \alpha \left(P_\mu - P_{\mu^*}\right) \left((1-\alpha)^{-1} J_{\mu^*} + h_{\mu^*} + o(1-\alpha)\right) \end{aligned} \tag{21.11}$$

Multiplying the last inequality by $(1-\alpha)$ and letting $\alpha \to 1$, we get

$$0 \leq P_\mu J_{\mu^*} - P_{\mu^*} J_{\mu^*} \Leftrightarrow P_{\mu^*} J_{\mu^*} \leq P_\mu J_{\mu^*} \tag{21.12}$$

Further, if $\mu$ is such that $P_{\mu^*} J_{\mu^*} = P_\mu J_{\mu^*}$, then the inequality 21.11 can be written as

$$0 \leq g_\mu - g_{\mu^*} + \alpha \left(P_\mu - P_{\mu^*}\right) \left(h_{\mu^*} + o(1-\alpha)\right)$$

Again, by taking the limit as $\alpha \to 1$, we obtain, using equation (21.10),

$$J_{\mu^*} + h_{\mu^*} = g_{\mu^*} + P_{\mu^*} h_{\mu^*} \leq g_\mu + P_\mu h_{\mu^*} \tag{21.13}$$

As a consequence of (21.13), we obtain the following sufficient condition for the existence of $\lambda$ and $h$ in the main result.

**Theorem 21.3.** *If the optimal average cost is the same for all initial conditions, then there exist $\lambda$ and $h$ as in the main result.*

**Corollary 21.4.** *If the chain $P_\mu$ is irreducible for every strategy $\mu$, then there exist $\lambda$ and $h$ as in the main result.*

**Proof:** (*of the theorem*) If $\mu^*$ is Blackwell optimal (we can always find such $\mu^*$), then we have $J_{\mu^*}(i) = \lambda$ for all $i$, because the optimal long term average cost is the same for all initial conditions. Letting $J_{\mu^*} = \lambda e$ with $e = [1, 1, \ldots, 1]^T$, we have for all $\mu$ that $P_\mu J_{\mu^*} = \lambda e = P_{\mu^*} J_{\mu^*}$ (because the sum of the entries of each row of $P_\mu$ is equal to 1).
From equation (21.13) we have $J_{\mu^*} + h_{\mu^*} \leq g_\mu + P_\mu h_{\mu^*}$ for all $\mu$, with equality when $\mu = \mu^*$.

We conclude that if the optimal average cost is the same for all initial conditions, then

$$\lambda + h(i) = \min_u \left( g(i,u) + \sum_{j=1}^{d} p_{ij}(u) h(j) \right)$$

where

$$\lambda = J_{\mu^*}(1) \text{ and } h(i) = h_{\mu^*}(i)$$

with $\mu^*$ being Blackwell optimal. □

The corollary follows immediately, since if the conditions of the corollary hold then the optimal average cost must be the same for all initial conditions.

# Bibliography

[1] Dimitri P. Bertsekas, *Dynamic Programming and Optimal Control, Volume 2, Second Edition* (Athena Scientific)

[2] Dimitri P. Bertsekas, *Dynamic Programming and Optimal Control, Volume 1, Third Edition* (Athena Scientific)