2.997 Decision-Making in Large-Scale Systems February 18
MIT, Spring 2004 Handout #7

Lecture Note 5

1 Relationship between Discounted and Average-Cost Problems

In this lecture, we will show that optimal policies for discounted-cost problems with large enough discount factor are also optimal for average-cost problems. The analysis will also show that, if the optimal average cost is the same for all initial states, then the average-cost Bellman's equation has a solution.

Note that the optimal average cost λ^* is independent of the initial state. Recall that

$$J_u(x) = \limsup_{N\to\infty} \frac{1}{N} \mathrm{E}\left[\sum_{t=0}^{N-1} g_u(x_t)|x_0 = x\right]$$

or, equivalently,

$$J_u = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} P_u^t g_u.$$

We also let $J_{u,\alpha}$ denote the discounted cost-to-go function associated with policy u when the discount factor is α, i.e.,

$$J_{u,\alpha} = \sum_{t=0}^{\infty} \alpha^t P_u^t g_u = (I - \alpha P_u)^{-1} g_u.$$

The following theorem formalizes the relationship between the discounted cost-to-go function and average cost.

Theorem 1 *For every stationary policy u, there is h_u such that*

$$J_{u,\alpha} = \frac{1}{1-\alpha} J_u + h_u + O(|1-\alpha|). \tag{1}$$

Theorem 1 follows easily from the following proposition.

Proposition 1 *For all stationary policies u, we have*

$$(I - \alpha P_u)^{-1} = \frac{1}{1-\alpha} P_u^* + H_u + O(|1-\alpha|)^1, \tag{2}$$

where

$$P_u^* = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} P_u^t, \tag{3}$$

$$H_u = (I - P_u + P_u^*)^{-1} - P_u^*, \tag{4}$$

$$P_u P_u^* = P_u^* P_u = P_u^* P_u^* = P_u^*, \tag{5}$$

$$P_u^* H_u = 0, \tag{6}$$

$$P_u^* + H_u = I + P_u H_u. \tag{7}$$

[1] $O(|1-\alpha|)$ is a function satisfying $\lim_{\alpha\to 1} O(|1-\alpha|) = 0$.

Proof: Let $M_\alpha = (1-\alpha)(I-\alpha P_u)^{-1}$. Then, since

$$|M_\alpha(x,y)| = (1-\alpha)\left|\sum_{t=0}^{\infty} \alpha^t P_u^t(x,y)\right| \le (1-\alpha)\left|\sum_{t=0}^{\infty} \alpha^t \cdot 1\right| = 1,$$

$M_\alpha(x,y)$ is in the form of

$$M_\alpha(x,y) = \frac{p(\alpha)}{q(\alpha)},$$

where $p(\cdot)$ and $q(\cdot)$ are polynomials such that $q(1) \neq 1$. We conclude that the limit $\lim_{\alpha\to 1} M_\alpha$ exists. Let $P_u^* = \lim_{\alpha\to 1} M_\alpha$. We can do Taylor's expansion of M_α around $\alpha = 1$, so that

$$M_\alpha = P_u^* + (1-\alpha)H_u + O((1-\alpha)^2)$$

where $H_u = -\frac{dM_\alpha}{d\alpha}$. Therefore

$$\boxed{(I-\alpha P_u)^{-1} = \tfrac{1}{1-\alpha}P_u^* + H_u + O(|1-\alpha|)}$$

for some P_u^* and H_u.

Next, observe that

$$(1-\alpha)(I-\alpha P_u)(I-\alpha P_u)^{-1} = (1-\alpha)I$$

for all α. Taking the limit as $\alpha \to 1$ yields

$$(I - P_u)P_u^* = 0,$$

so that $P_u^* = P_u P_u^*$. We can use the same reasoning to conclude that $P_u^* = P_u^* P_u$. We also have

$$(I-\alpha P_u)P_u^* = (1-\alpha)P_u^*,$$

hence for every α we have

$$P_u^* = (1-\alpha)(I-\alpha P_u)^{-1}P_u^*,$$

and taking the limit as $\alpha \to 1$ yields $P_u^* P_u^* = P_u^*$.

We now show that, for every $t \ge 1$, $P_u^t - P_u^* = (P_u - P_u^*)^t$. For $t = 1$, it is trivial. Suppose that the result holds up to $n-1$, i.e., $P_u^{n-1} - P_u^* = (P_u - P_u^*)^{n-1}$. Then $(P_u - P_u^*)(P_u - P_u^*)^{n-1} = (P_u - P_u^*)(P_u^{n-1} - P_u^*) = P_u^n - P_u P_u^* - P_u^* P_u^{n-1} + P_u^* P_u^* = P_u^n - P_u^* - P_u^* P_u^{n-2} + P_u^* = P_u^n - P_u^*$. By induction, we have $P_u^t - P_u^* = (P_u - P_u^*)^t$.

Now note that

$$\begin{aligned}
H_u &= \lim_{\alpha\to 1} \frac{M_\alpha - P_u^*}{1-\alpha} \\
&= \lim_{\alpha\to 1}\left[(I-\alpha P_u)^{-1} - \frac{P_u^*}{1-\alpha}\right] \\
&= \lim_{\alpha\to 1}\left[\sum_{t=0}^{\infty} \alpha^t (P_u^t - P_u^*)\right] \\
&= \lim_{\alpha\to 1}\left[I - P_u^* + \sum_{t=1}^{\infty} \alpha^t (P_u - P_u^*)^t\right] \\
&= \lim_{\alpha\to 1}\left[\sum_{t=0}^{\infty} \alpha^t (P_u - P_u^*)^t - P_u^*\right] \\
&= (I - P_u + P_u^*)^{-1} - P_u^*.
\end{aligned}$$

Hence $\boxed{H_u = (I - P_u + P_u^*)^{-1} - P_u^*.}$

We now show $P_u^* H_u = 0$. Observe

$$\begin{aligned} P_u^* H_u &= P_u^* \left[(I - P_u + P_u^*)^{-1} - P_u^* \right] \\ &= \sum_{t=0}^{\infty} P_u^* (P_u - P_u^*)^t - P_u^* \\ &= P_u^* - P_u^* = 0. \end{aligned}$$

Therefore, $\boxed{P_u^* H_u = 0.}$

Observe $(I - P_u + P_u^*)H_u = I - (I - P_u + P_u^*)P_u^* = I - P_u^*$. Since $P_u^* H_u = 0$, we have $\boxed{P_u^* + H_u = I + P_u H_u.}$

By multiplying P_u^k to $P_u^* + H_u = I + P_u H_u$, we have

$$P_u^k P_u^* + P_u^k H_u = P_u^k + P_u^{k+1} H_u, \quad \forall\, k$$

Summing from $k = 0$ to $k = N - 1$, we have

$$N P_u^* + \sum_{k=0}^{N-1} P_u^k H_u = \sum_{k=0}^{N-1} P_u^k + \sum_{k=1}^{N} P_u^k H_u,$$

or, equivalently,

$$N P_u^* = \sum_{k=0}^{N-1} P_u^k + (P_u^N - I) H_u.$$

Dividing both sides by N and letting $N \to \infty$, then we have

$$\boxed{\lim_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{k=0}^{N-1} P_u^k = P_u^*.}$$

□

Since $P_u^* = P_u^* P_u$ and P_u itself is a stochastic matrix, the rows of P_u^* are of special meanings. Let π_u denote a row of P_u. Then $\pi_u = \pi_u P_u$ and $\pi_u(x) = \sum_y \pi_u(y) P_u(y, x)$. Then $P_u(x_1 = x | x_0 \sim \pi_u) = \sum_y \pi_u(y) P_u(y, x)$. We can conclude that any row in matrix P_u^* is a stationary distribution for the Markov chain under the policy u. However, does this observation mean that all rows in P_u^* are identical?

Theorem 2

$$J_{u,\alpha} = \frac{J_u}{1-\alpha} + h_u + O(|1-\alpha|)$$

Proof:

$$\begin{aligned} J_{u,\alpha} &= (I - \alpha P_u)^{-1} g_u \\ &= \left(\frac{P_u^*}{1-\alpha} + H_u + O(|1-\alpha|) \right) g_u \\ &= \frac{P_u^* g_u}{1-\alpha} + H_u g_u + O(|1-\alpha|) \\ &= \frac{1}{1-\alpha} \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} P_u^t g_u + \underbrace{h_u}_{=H_u g_u} + O(|1-\alpha|) \\ &= \frac{J_u}{1-\alpha} + \underbrace{h_u}_{=H_u g_u} + O(|1-\alpha|). \end{aligned}$$

□

2 Blackwell Optimality

In this section, we will show that policies that are optimal for the discounted-cost criterion with large enough discount factors are also optimal for the average-cost criterion. Indeed, we can actually strengthen the notion of average-cost optimality and establish the existence of policies that are optimal for all large enough discount factors.

Definition 1 (Blackwell Optimality) *A stationary policy u^* is called Blackwell optimal if $\exists \bar{\alpha} \in (0,1)$ such that u^* is optimal $\forall\, \alpha \in [\bar{\alpha}, 1)$.*

Theorem 3 *There exists a stationary Blackwell optimal policy and it is also optimal for the average-cost problem among all stationary policies.*

Proof: Since there are only finitely many policies, we must have for each state x a policy μ_x such that $J_{u_x,\alpha}(x) \leq J_{u,\alpha}(x)$ for all large enough α. If we take the policy μ^* to be given by $\mu^*(x) = \mu_x(x)$, then μ^* must satisfy Bellman's equation

$$J_{u^*,\alpha} = \min_u \{g_u + \alpha P_u J_{u^*,\alpha}\}$$

for all large enough α, and we conclude that μ^* is Blackwell optimal.

Now let u^* be Blackwell optimal. Also suppose that $\bar{u}$ is optimal for the average-cost problem. Then

$$\frac{J_{u^*}}{1-\alpha} + h_{u^*} + O(|1-\alpha|) \leq \frac{J_{\bar{u}}}{1-\alpha} + h_{\bar{u}} + O(|1-\alpha|),\ \forall \alpha \geq \bar{\alpha}.$$

Taking the limit as $\alpha \to 1$, we conclude that

$$J_{u^*} \leq J_{\bar{u}},$$

and u^* must be optimal for the average-cost problem. □

Remark 1 *It is actually possible to establish average-cost optimality of Blackwell optimal policies among the set of all policies, not only stationary ones.*

Remark 2 *An algorithm for computing Blackwell optimal policies involves lexicographic optimization of J_u, h_u and higher-order terms in the Taylor expansion of $J_{u,\alpha}$.*

Theorem 3 implies that if the optimal average cost is the same regardless of the initial state, then the average-cost Bellman's equation has a solution. Combined with Theorem 1 of the previous lecture, it follows that this is a necessary and sufficient condition for existence of Bellman's equation solution.

Corollary 1 *If $J_{u^*} = \lambda^* e$, then $\lambda e + h = Th$ has a solution (λ^*, h_{u^*}) with u^* which is Blackwell optimal.*

Proof: We have, for all large enough α,

$$\begin{aligned}
J_{u^*,\alpha} &= \min_u \{g_u + \alpha P_u J_{u^*,\alpha}\} \\
\frac{J_{u^*}}{1-\alpha} + h_{u^*} + O((1-\alpha)^2) &= \min_u \left\{ g_u + \alpha P_u \left(\frac{J_{u^*}}{1-\alpha} + h_{u^*} + O((1-\alpha)^2) \right) \right\} \\
\frac{\lambda^* e}{1-\alpha} + h_{u^*} + O((1-\alpha)^2) &= \min_u \left\{ g_u + \alpha P_u \left(\frac{\lambda^* e}{1-\alpha} + h_{u^*} + O((1-\alpha)^2) \right) \right\} \\
\lambda^* + h_{u^*} + O((1-\alpha)^2) &= \min_u \left\{ g_u + \alpha P_u \left(h_{u^*} + O((1-\alpha)^2) \right) \right\}.
\end{aligned}$$

Taking the limit as $\alpha \to 1$, we get

$$\lambda^* e + h_{u^*} = \min_u \{g_u + P_u h_{u^*}\} = T h_{u^*}.$$

□

In the average-cost setting, existence of a solution to Bellman's equation actually depends on the structure of transition probabilities in the system. Some sufficient conditions for the optimal average cost to be the same regardless of the initial state are given below.

Definition 2 *We say that two states x, y* communicate *under policy u if there are $k, \bar{k} \in \{1, 2, \dots\}$ such that $P_u^k(x,y) > 0$, $P_u^{\bar{k}}(y,x) > 0$.*

Definition 3 *We say that a state x is* recurrent *under policy u if, conditioned on the fact that it is visited at least once, it is visited infinitely many times.*

Definition 4 *We say that a state x is* transient *under policy u if it is only visited finitely many times, regardless of the initial condition of the system.*

Definition 5 *We say that a policy u is* unichain *if all of its recurrent states communicate.*

We state without proof the following theorem.

Theorem 4 *Either of the following conditions is sufficient for the optimal average cost to be the same regardless of the initial state:*

1. *There exists a unichain optimal policy.*
2. *For every pair of states x and y, there is a policy u such that x and y communicate.*

3 Value Iteration

We want to compute

$$\min_u \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} P_u^t g_u$$

One way to obtain this value is to calculate a finite but very large N to approximate the limit and speculate that such an limit is accurate. Hence we consider

$$T^k J = \min_u \mathrm{E}\left[\sum_{t=0}^{k-1} g_u(x_t) + J_0(x_k)\right]$$

Recall $J^*(x,T) \cong \lambda^* T + h^*(x)$. Choose some state x and $\bar{x}$, we have

$$J^*(x,T) - J^*(\bar{x},T) = h^*(x) - h^*(\bar{x})$$

Then

$$h_k(x) = J^*(x,k) - \delta^k, \qquad \text{for some } \delta^1, \delta^2, \ldots$$

Note that, since $(\lambda^*, h^* + ke)$ is a solution to Bellman's equation for all k whenever (λ^*, h^*) is a solution, we can choose the value of a single state arbitrarily. Letting $h^*(\bar{x}) = 0$, we have the following commonly used version of value iteration;

$$h_{k+1}(x) = (Th_k)(x) - (Th_k)(\bar{x}) \tag{8}$$

Theorem 5 *Let h_k be given by (8). Then if $h_k \to \bar{h}$, we have $\lambda^* = (T\bar{h})(\bar{x})$ and $h^* = \bar{h}$, $\lambda^* e + h^* = Th^*$.*

Note that there must exist a solution to the average-cost Bellman's equation for value iteration to converge. However, it can be shown that existence of a solution is not a sufficient condition.