Chapter 5. Blackwell Optimality

Let $\{X_n\}$ be a MCP with state space S and feasible control set $U(x)$ at state $x \in S$. Consider the infinite horizon problem

$$v(x) \doteq \inf_{\{u_n\}} J(x; \{u_n\}) \doteq \inf_{\{u_n\}} E_x \left[\sum_{j=0}^{\infty} \beta^j c(X_j; u_j) \right].$$

Here the discount factor $\beta \in (0, 1)$. To distinguish among discount factors, the above optimization problem is referred to as the β-discount problem, and a policy is said to be β-optimal if it is optimal for the β-discount problem. We have the following result.

Theorem 1 *Assume that S is a finite space, and for each $x \in S$, $U(x)$ is also a finite space. Then there exists a $\bar{\beta} \in (0, 1)$ and a stationary control policy $u^* : x \mapsto U(x)$ such that $\{u^*\}$ is β-optimal for all $\beta \in (\bar{\beta}, 1)$.*

Such policy $\{u^*\}$ is often said to be *Blackwell optimal*. As we will show later on, Blackwell optimality is closely related to the long run average cost problems.

Proof. We denote by $v_\beta(x)$ the value function for the β-discount problem. Similarly for $J_\beta(x; \{u_n\})$. For each β-discount problem, there exists an optimal control policy that is stationary, thanks to the results in Chapter 4. Since the state space S and the feasible control sets are all finite space, there are a finite number of stationary policies in total. Therefore, there must exist a stationary policy, say u^*, and a sequence of $\beta_n \uparrow 1$, such that $\{u^*\}$ is optimal for the β_n-discount problem.

We will prove the theorem by contradiction. Suppose the claim in the theorem is not true. Then there exists a sequence $\alpha_n \uparrow 1$ such that $\{u^*\}$ is not optimal for the α_n-discount problem. It follows that for each α_n, there exists a state $x_n \in S$, such that

$$J_{\alpha_n}(x_n; \{u^*\}) > v_{\alpha_n}(x_n).$$

Since the state space is finite, one can assume that $x_n \equiv \bar{x} \in S$ (use a subsequence if necessary). Similarly, choosing a subsequence if necessary, one can further assume that there exists a stationary policy $\{\bar{u}\}$ that is optimal for all α_n-discount problems.

For conclusion, one can find two sequences $\beta_n \uparrow 1$ and $\alpha_n \uparrow 1$, and a state $\bar{x} \in S$ such that

$$J_{\alpha_n}(\bar{x}; \{u^*\}) > J_{\alpha_n}(\bar{x}; \{\bar{u}\}), \qquad J_{\beta_n}(\bar{x}; \{u^*\}) \leq J_{\beta_n}(\bar{x}; \{\bar{u}\}).$$

Now think of the above quantities as functions of the discount factor. That is, let

$$f(\beta) \doteq J_\beta(\bar{x}; \{u^*\}), \qquad g(\beta) \doteq J_\beta(\bar{x}; \{\bar{u}\}).$$

Then we have

$$f(\alpha_n) - g(\alpha_n) > 0, \qquad f(\beta_n) - g(\beta_n) \leq 0,$$

which implies that the equation $f - g = 0$ has infinitely many roots over interval $(0, 1)$.

We show that both f and g are continuous, rational functions of $\beta \in (0, 1)$. If this is the case, we have obtained a contradiction, since for any rational functions there exist at most finitely many roots. The proof is the same for f and g, so we will only show for f. By definition,

$$f(\beta) = E_{\bar{x}}\left[\sum_{j=0}^{\infty} \beta^j c(X_j, \bar{u}(X_j))\right] = \sum_{x \in S}\left[c(x; \bar{u}(x)) \sum_{j=0}^{\infty} \beta^j P^j(\bar{x}; x)\right],$$

where P^j is the j-step transition probability matrix under the stationary control policy $\bar{u}$. Let $P = P^1$ be the probability transition probability matrix. Then

$$\sum_{j=0}^{\infty} \beta^j P^j = I + \beta P + (\beta P)^2 + \cdots = (I - \beta P)^{-1}.$$

This is a matrix with each component a rational function of β, which is automatically continuous. We complete the proof. ■

0.1 Example: Service control for a queuing network

The Blackwell optimality is not exclusively for finite state space S and finite control choices. The example in this section is a special case from [2]. The premise of the problem is about the service control of a single station with multiclass job arrivals.

Consider a single server system whose customers are of two separate classes. Customers of class i arrive according to independent Poisson processes with arrival rate λ_i $(i = 1, 2)$. Service times for customers are iid with common distribution F, irregardless of the class. The mean service time is denoted by μ. It is assumed to be strictly positive. The moment generating function for the service is

$$\varphi(\beta) \doteq \int_0^\infty e^{-\beta t} dF(t),$$

which is assumed to be finite for all β.

As the decision maker, one can choose to idle ("0"), serve class 1 customer ("1"), or serve class 2 customer ("2"). If the decision is idleness, then this decision cannot be reversed until there is a new arrival to the system. If the decision is to serve any customer, then the system will keep serving this customer without interruption until the customer is served and leaves the system.

A reward of $R_i > 0$ is received upon completion of each class i service ($i = 1, 2$). The reward is discounted at interest rate β, so that the reward R_k received at time t has a present value $e^{-\beta t} R_k$. The objective is to maximize the total discounted reward. A policy is said to be *Blackwell optimal* if there exists a $\bar{\beta} > 0$ such that it is β-optimal for all $\beta \in (0, \bar{\beta})$.

We consider two specific stationary policies.

Definition 1 The *full priority rule* is the policy which always serves class 1 customer whenever one is present, and otherwise serves a class 2 customer, and choosing idleness only when the system is empty. The *restricted pripority rule* is the policy which always serves class 1 customer whenever one is present, and otherwises chooses idleness, thus never serving class 2 customer.

Without loss of generality, we assume $R_1 > R_2$. We have the following result.

Proposition 1 *Assume the system is stable, that is $\lambda_1 \mu + \lambda_2 \mu < 1$. Then for any fixed $\beta > 0$, either the full priority rule or the restricted priority rule is optimal. Furthermore, the full priority rule is Blackwell optimal.*

This optimization, though continuous time, can be set up into the discrete time framework. The decision point for the decision maker are those epochs at which either a service is completed or a customer arrives to find the server idle. The state at each decision points are $X = (X^1, X^2)$ where X^i is the number of class i customers. The random noise ξ, indeed is the real time between two epochs. Its distribution depends on the control, but nonetheless can be specified.

More precisely, let $X_n = (X_n^1, X_n^2)$ be the state at the n-th epoch, and ξ_{n+1} be the time between n-th epoch and $(n+1)$-th epoch. The reward for this period (after discounting back to the n-th epoch)

$$e^{-\beta \xi_{n+1}} c(X_n, u_n; \xi_{n+1}) = e^{-\beta \xi_{n+1}} \left[R_1 1_{\{u_n = 1\}} + R_2 1_{\{u_n = 2\}} \right].$$

Another slight difference from the settings in Chapter 4 is that the discounting is random in the sense that $\{\xi_n\}$ is a sequence of random variables. But this makes little difference in the analysis.

Denote by $Q(x', dt|x, u)$ as the conditional probability of $(X_{n+1}, \xi_{n+1}) = (x', dt)$ given $(X_n, u_n) = (x, u)$. Let

$$\Lambda \doteq \lambda_1 + \lambda_2.$$

It is not difficult to verify that, for example, with $\delta_1 \doteq (1, 0)$ and $\delta_2 \doteq (0, 1)$,

$$Q(x', dt|x, 0) = \begin{cases} \lambda_1 e^{-\Lambda t}\, dt & ; \quad \text{if } x' = x + \delta_1 \\ \lambda_2 e^{-\Lambda t}\, dt & ; \quad \text{if } x' = x + \delta_2 \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

and the identification of Q for $u = \{1, 2\}$ is left as an exercise.

The same contraction mapping argument asserts that the value function, say $v(x)$, is the unique bounded function satisfying the DPE

$$v(x) = \max_{u \in U(x)} \left[R(x; u) + \sum_{x'} \int_0^\infty e^{-\beta t} v(x') Q(x', dt|x, u) \right]$$

with

$$R(x; u) \doteq \begin{cases} 0 & ; \quad \text{if } u = 0 \\ R_1 \cdot \varphi(\beta) & ; \quad \text{if } u = 1 \\ R_2 \cdot \varphi(\beta) & ; \quad \text{if } u = 2 \end{cases}$$

Fix an arbitrary $\beta > 0$, let $\alpha \in (0, 1)$ be the unique positive solution satisfying the equation

$$\alpha = \varphi \left[\beta + \lambda_1 (1 - \alpha) \right].$$

Abusing notation a bit, let $\varphi = \varphi(\beta)$. Define

$$\begin{aligned} C_1 &\doteq \varphi (1 - \varphi)^{-1} R_1 \\ C_2 &\doteq (1 - \alpha)^{-1} \left[R_2 \varphi + (\varphi - \alpha) C_1 \right] \\ W &\doteq \lambda_1 (1 - \alpha) C_1 [\beta + \lambda_1 (1 - \alpha)]^{-1}. \end{aligned}$$

We have the following lemma.

Lemma 1 *For any fixed β, we have $C_1 \geq C_2$ and $C_1 > W$. If $C_1 \geq W \geq C_2$, then the restricted priority policy is β-optimal. If $C_1 \geq C_2 \geq W$, then the full priority policy is β-optimal.*

We will leave the proof of the lemma to the end.

Proof of Proposition 1. Thanks to Lemma 1, it suffices to study the behavior of $C_2 - W$ as $\beta \to 0$. However, it is not difficult to show that as $\beta \to 0$,

$$\lim_{\beta \downarrow 0} \beta[C_2 - W] = R_2(1 - \lambda_1\mu)\mu^{-1} > 0,$$

observing that (check!)

$$\lim_{\beta \downarrow 0} \frac{1-\alpha}{\beta} = \frac{\mu}{1-\lambda_1\mu},$$

and

$$\lim_{\beta \downarrow 0} \frac{1-\varphi}{\beta} = \mu.$$

We complete the proof. ■

Proof of Lemma 1. The inequalities $C_1 \geq C_2$ and $C_1 > W$ are trivial. For $C_1 \geq W \geq C_2$, we claim that the value function is

$$v(x) = C_1 \left(1 - \alpha^{x^1} \frac{\beta}{\beta + \lambda_1(1-\alpha)}\right), \quad x = (x^1, x^2).$$

As for $C_1 \geq C_2 \geq W$, let $\bar{\alpha}$ be the unique solution to the equation

$$\bar{\alpha} = \varphi[\beta + \Lambda(1 - \bar{\alpha})].$$

Then the value function is

$$\begin{aligned} v(x) &= C_1 + (C_2 - C_1)\alpha^{x^1} \\ &\quad + [\beta + \Lambda(1-\bar{\alpha})]^{-1} \left[\lambda_1(1-\alpha)(C_1 - C_2) - \beta C_2\right] \bar{\alpha}^{x^1 + x^2} \end{aligned}$$

The proof is just direct verification. We omit the details. A more elegant proof can be found in [1]. ■

References

[1] J.M. Harrison. Dynamic scheduling of a multiclass queue: discount optimality. *Operations Research*, 23:270–282, 1975.

[2] J.M. Harrison. Dynamic scheduling of a two-class queue: small interest rates. *SIAM J. Appl. Math.*, 31:51–61, 1976.