In this chapter we consider a number of applications of discrete-time stochastic optimal control with perfect state information. These applications are special cases of the basic problem of Section 1.2 and can be addressed via the DP algorithm. In all these applications the stochastic nature of the disturbances is significant. For this reason, in contrast with the deterministic problems of the preceding two chapters, the use of closed-loop control is essential to achieve optimal performance.

4.1 LINEAR SYSTEMS AND QUADRATIC COST

In this section we consider the special case of a linear system

$$x_{k+1} = A_k x_k + B_k u_k + w_k, \qquad k = 0, 1, \ldots, N-1,$$

and the quadratic cost

$$\mathop{E}_{\substack{w_k \\ k=0,1,\ldots,N-1}} \left\{ x_N' Q_N x_N + \sum_{k=0}^{N-1} (x_k' Q_k x_k + u_k' R_k u_k) \right\}.$$

In these expressions, x_k and u_k are vectors of dimension n and m, respectively, and the matrices A_k, B_k, Q_k, R_k are given and have appropriate dimension. We assume that the matrices Q_k are positive semidefinite symmetric, and the matrices R_k are positive definite symmetric. The controls u_k are unconstrained. The disturbances w_k are independent random vectors with given probability distributions that do not depend on x_k and u_k. Furthermore, each w_k has zero mean and finite second moment.

The problem described above is a popular formulation of a regulation problem whereby we want to keep the state of the system close to the origin. Such problems are common in the theory of automatic control of a motion or a process. The quadratic cost function is often reasonable because it induces a high penalty for large deviations of the state from the origin but a relatively small penalty for small deviations. Also, the quadratic cost is frequently used, even when it is not entirely justified, because it leads to a nice analytical solution. A number of variations and generalizations have similar solutions. For example, the disturbances w_k could have nonzero means and the quadratic cost could have the form

$$E\left\{ (x_N - \overline{x}_N)' Q_N (x_N - \overline{x}_N) + \sum_{k=0}^{N-1} \left((x_k - \overline{x}_k)' Q_k (x_k - \overline{x}_k) + u_k' R_k u_k \right) \right\},$$

which expresses a desire to keep the state of the system close to a given trajectory $(\overline{x}_0, \overline{x}_1, \ldots, \overline{x}_N)$ rather than close to the origin. Another gener-

matrices, rather than being known. This case is considered at the end of this section.

Applying now the DP algorithm, we have

$$J_N(x_N) = x_N' Q_N x_N,$$

$$J_k(x_k) = \min_{u_k} E\left\{ x_k' Q_k x_k + u_k' R_k u_k + J_{k+1}(A_k x_k + B_k u_k + w_k) \right\}. \quad (4.1)$$

It turns out that the cost-to-go functions J_k are quadratic and as a result the optimal control law is a linear function of the state. These facts can be verified by straightforward induction. We write Eq. (4.1) for $k = N-1$,

$$\begin{aligned} J_{N-1}(x_{N-1}) = \min_{u_{N-1}} E\{ & x_{N-1}' Q_{N-1} x_{N-1} + u_{N-1}' R_{N-1} u_{N-1} \\ & + (A_{N-1} x_{N-1} + B_{N-1} u_{N-1} + w_{N-1})' Q_N \\ & \cdot (A_{N-1} x_{N-1} + B_{N-1} u_{N-1} + w_{N-1}) \}, \end{aligned}$$

and we expand the last quadratic form in the right-hand side. We then use the fact $E\{w_{N-1}\} = 0$ to eliminate the term $E\{w_{N-1}' Q_N (A_{N-1} x_{N-1} + B_{N-1} u_{N-1})\}$, and we obtain

$$\begin{aligned} J_{N-1}(x_{N-1}) = x_{N-1}' Q_{N-1} x_{N-1} + \min_{u_{N-1}} [& u_{N-1}' R_{N-1} u_{N-1} \\ & + u_{N-1}' B_{N-1}' Q_N B_{N-1} u_{N-1} + 2 x_{N-1}' A_{N-1}' Q_N B_{N-1} u_{N-1}] \\ & + x_{N-1}' A_{N-1}' Q_N A_{N-1} x_{N-1} + E\{ w_{N-1}' Q_N w_{N-1} \}. \end{aligned}$$

By differentiating with respect to u_{N-1} and by setting the derivative equal to zero, we obtain

$$(R_{N-1} + B_{N-1}' Q_N B_{N-1}) u_{N-1} = -B_{N-1}' Q_N A_{N-1} x_{N-1}.$$

The matrix multiplying u_{N-1} on the left is positive definite (and hence invertible), since R_{N-1} is positive definite and $B_{N-1}' Q_N B_{N-1}$ is positive semidefinite. As a result, the minimizing control vector is given by

$$u_{N-1}^* = -(R_{N-1} + B_{N-1}' Q_N B_{N-1})^{-1} B_{N-1}' Q_N A_{N-1} x_{N-1}.$$

By substitution into the expression for J_{N-1}, we have

$$J_{N-1}(x_{N-1}) = x_{N-1}' K_{N-1} x_{N-1} + E\{ w_{N-1}' Q_N w_{N-1} \},$$

where by straightforward calculation, the matrix K_{N-1} is verified to be

$$K_{N-1} = A_{N-1}' (Q_N - Q_N B_{N-1} (B_{N-1}' Q_N B_{N-1} + R_{N-1})^{-1} B_{N-1}' Q_N) A_{N-1}$$

The matrix K_{N-1} is clearly symmetric. It is also positive semidefinite. To see this, note that from the preceding calculation we have for $x \in \Re^n$

$$x'K_{N-1}x = \min_u \big[x'Q_{N-1}x + u'R_{N-1}u + (A_{N-1}x + B_{N-1}u)'Q_N(A_{N-1}x + B_{N-1}u)\big].$$

Since Q_{N-1}, R_{N-1}, and Q_N are positive semidefinite, the expression within brackets is nonnegative. Minimization over u preserves nonnegativity, so it follows that $x'K_{N-1}x \geq 0$ for all $x \in \Re^n$. Hence K_{N-1} is positive semidefinite.

Since J_{N-1} is a positive semidefinite quadratic function (plus an inconsequential constant term), we may proceed similarly and obtain from the DP equation (4.1) the optimal control law for stage $N-2$. As earlier, we show that J_{N-2} is a positive semidefinite quadratic function, and by proceeding sequentially, we obtain the optimal control law for every k. It has the form

$$\mu_k^*(x_k) = L_k x_k, \tag{4.2}$$

where the gain matrices L_k are given by the equation

$$L_k = -(B_k'K_{k+1}B_k + R_k)^{-1}B_k'K_{k+1}A_k,$$

and where the symmetric positive semidefinite matrices K_k are given recursively by the algorithm

$$K_N = Q_N, \tag{4.3}$$

$$K_k = A_k'\big(K_{k+1} - K_{k+1}B_k(B_k'K_{k+1}B_k + R_k)^{-1}B_k'K_{k+1}\big)A_k + Q_k. \tag{4.4}$$

Just like DP, this algorithm starts at the terminal time N and proceeds backwards. The optimal cost is given by

$$J_0(x_0) = x_0'K_0x_0 + \sum_{k=0}^{N-1} E\{w_k'K_{k+1}w_k\}.$$

The control law (4.2) is simple and attractive for implementation in engineering applications: the current state x_k is being fed back as input through the linear feedback gain matrix L_k as shown in Fig. 4.1.1. This accounts in part for the popularity of the linear-quadratic formulation. As we will see in Chapter 5, the linearity of the control law is still maintained even for problems where the state x_k is not completely observable (imperfect state information).

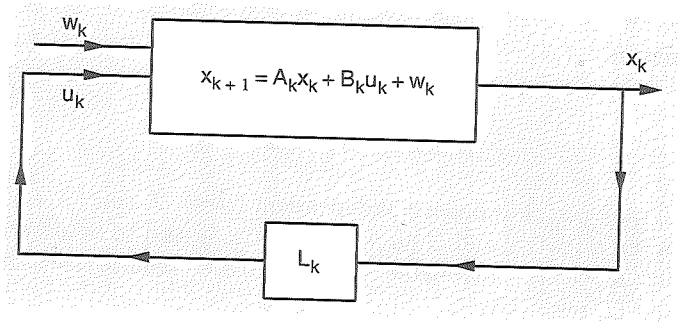

Figure 4.1.1 Linear feedback structure of the optimal controller for the linear-quadratic problem.

The Riccati Equation and Its Asymptotic Behavior

Equation (4.4) is called the *discrete-time Riccati equation*. It plays an important role in control theory. Its properties have been studied extensively and exhaustively. One interesting property of the Riccati equation is that if the matrices A_k, B_k, Q_k, R_k are constant and equal to A, B, Q, R, respectively, then the solution K_k converges as $k \to -\infty$ (under mild assumptions) to a steady-state solution K satisfying the *algebraic Riccati equation*

$$K = A'\big(K - KB(B'KB + R)^{-1}B'K\big)A + Q. \tag{4.5}$$

This property, to be proved shortly, indicates that for the system

$$x_{k+1} = Ax_k + Bu_k + w_k, \qquad k = 0, 1, \ldots, N-1,$$

and a large number of stages N, one can reasonably approximate the control law (4.2) by the control law $\{\mu^*, \mu^*, \ldots, \mu^*\}$, where

$$\mu^*(x) = Lx, \tag{4.6}$$

$$L = -(B'KB + R)^{-1}B'KA,$$

and K solves the algebraic Riccati equation (4.5). This control law is *stationary*; that is, it does not change over time.

We now turn to proving convergence of the sequence of matrices $\{K_k\}$ generated by the Riccati equation (4.4). We first introduce the notions of controllability and observability, which are very important in control

Definition 4.1.1: A pair (A, B), where A is an $n \times n$ matrix and B is an $n \times m$ matrix, is said to be *controllable* if the $n \times nm$ matrix

$$[B, AB, A^2B, \ldots, A^{n-1}B]$$

has full rank (i.e., has linearly independent rows). A pair (A, C), where A is an $n \times n$ matrix and C an $m \times n$ matrix, is said to be *observable* if the pair (A', C') is controllable, where A' and C' denote the transposes of A and C, respectively.

One may show that if the pair (A, B) is controllable, then for any initial state x_0, there exists a sequence of control vectors $u_0, u_1, \ldots, u_{n-1}$ that force the state x_n of the system

$$x_{k+1} = Ax_k + Bu_k$$

to be equal to zero at time n. Indeed, by successively applying the above equation for $k = n-1, n-2, \ldots, 0$, we obtain

$$x_n = A^n x_0 + Bu_{n-1} + ABu_{n-2} + \cdots + A^{n-1}Bu_0$$

or equivalently

$$x_n - A^n x_0 = (B, AB, \ldots, A^{n-1}B) \begin{pmatrix} u_{n-1} \\ u_{n-2} \\ \vdots \\ u_0 \end{pmatrix}. \tag{4.7}$$

If (A, B) is controllable, the matrix $(B, AB, \ldots, A^{n-1}B)$ has full rank and as a result the right-hand side of Eq. (4.7) can be made equal to any vector in $\Re^n$ by appropriate selection of $(u_0, u_1, \ldots, u_{n-1})$. In particular, one can choose $(u_0, u_1, \ldots, u_{n-1})$ so that the right-hand side of Eq. (4.7) is equal to $-A^n x_0$, which implies $x_n = 0$. This property explains the name "controllable pair" and in fact is often used to define controllability.

The notion of observability has an analogous interpretation in the context of estimation problems; that is, given measurements $z_0, z_1, \ldots, z_{n-1}$ of the form $z_k = Cx_k$, it is possible to infer the initial state x_0 of the system $x_{k+1} = Ax_k$, in view of the relation

$$\begin{pmatrix} z_{n-1} \\ \vdots \\ z_1 \\ z_0 \end{pmatrix} = \begin{pmatrix} CA^{n-1} \\ \vdots \\ CA \\ C \end{pmatrix} x_0.$$

Alternatively, it can be seen that observability is equivalent to the property

The notion of stability is of paramount importance in control theory. In the context of our problem it is important that the stationary control law (4.6) results in a stable closed-loop system; that is, in the absence of input disturbance, the state of the system

$$x_{k+1} = (A + BL)x_k, \qquad k = 0, 1, \ldots,$$

tends to zero as $k \to \infty$. Since $x_k = (A + BL)^k x_0$, it follows that the closed-loop system is stable if and only if $(A + BL)^k \to 0$, or equivalently (see Appendix A), if and only if the eigenvalues of the matrix $(A + BL)$ are strictly within the unit circle.

The following proposition shows that for a stationary controllable system and constant matrices Q and R, the solution of the Riccati equation (4.4) converges to a positive definite symmetric matrix K for an arbitrary positive semidefinite symmetric initial matrix. In addition, the proposition shows that the corresponding closed-loop system is stable. The proposition also requires an observability assumption, namely, that Q can be written as $C'C$, where the pair (A, C) is observable. Note that if r is the rank of Q, there exists an $r \times n$ matrix C of rank r such that $Q = C'C$ (see Appendix A). The implication of the observability assumption is that in the absence of control, if the state cost per stage $x_k'Qx_k$ tends to zero or equivalently $Cx_k \to 0$, then also $x_k \to 0$.

To simplify notation, we reverse the time indexing of the Riccati equation. Thus, P_k in the following proposition corresponds to K_{N-k} in Eq. (4.4). A graphical proof of the proposition for the case of a scalar system is given in Fig. 4.1.2.

Proposition 4.4.1: Let A be an $n \times n$ matrix, B be an $n \times m$ matrix, Q be an $n \times n$ positive semidefinite symmetric matrix, and R be an $m \times m$ positive definite symmetric matrix. Consider the discrete-time Riccati equation

$$P_{k+1} = A'\big(P_k - P_k B(B'P_k B + R)^{-1}B'P_k\big)A + Q, \qquad k = 0, 1, \ldots, \tag{4.8}$$

where the initial matrix P_0 is an arbitrary positive semidefinite symmetric matrix. Assume that the pair (A, B) is controllable. Assume also that Q may be written as $C'C$, where the pair (A, C) is observable. Then:

(a) There exists a positive definite symmetric matrix P such that for every positive semidefinite symmetric initial matrix P_0 we have

$$\lim_{k \to \infty} P_k = P.$$

Furthermore, P is the unique solution of the algebraic matrix equation

$$P = A'\big(P - PB(B'PB + R)^{-1}B'P\big)A + Q \tag{4.9}$$

within the class of positive semidefinite symmetric matrices.

(b) The corresponding closed-loop system is stable; that is, the eigenvalues of the matrix

$$D = A + BL, \tag{4.10}$$

where

$$L = -(B'PB + R)^{-1}B'PA, \tag{4.11}$$

are strictly within the unit circle.

Proof: The proof proceeds in several steps. First we show convergence of the sequence generated by Eq. (4.8) when the initial matrix P_0 is equal to zero. Next we show that the corresponding matrix D of Eq. (4.10) satisfies $D^k \to 0$. Then we show the convergence of the sequence generated by Eq. (4.8) when P_0 is any positive semidefinite symmetric matrix, and finally we show uniqueness of the solution of Eq. (4.9).

Initial Matrix $P_0 = 0$. Consider the optimal control problem of finding $u_0, u_1, \ldots, u_{k-1}$ that minimize

$$\sum_{i=0}^{k-1}(x_i'Qx_i + u_i'Ru_i)$$

subject to

$$x_{i+1} = Ax_i + Bu_i, \qquad i = 0, 1, \ldots, k-1,$$

where x_0 is given. The optimal value of this problem, according to the theory of this section, is $x_0'P_k(0)x_0$,

where $P_k(0)$ is given by the Riccati equation (4.8) with $P_0 = 0$. For any control sequence $(u_0, u_1, \ldots, u_k)$ we have

$$\sum_{i=0}^{k-1}(x_i'Qx_i + u_i'Ru_i) \le \sum_{i=0}^{k}(x_i'Qx_i + u_i'Ru_i)$$

and hence

$$x_0'P_k(0)x_0 = \min_{u_i \atop i=0,\ldots,k-1} \sum_{i=0}^{k-1}(x_i'Qx_i + u_i'Ru_i)$$

$$\le \min_{u_i \atop i=0,\ldots,k} \sum_{i=0}^{k}(x_i'Qx_i + u_i'Ru_i)$$

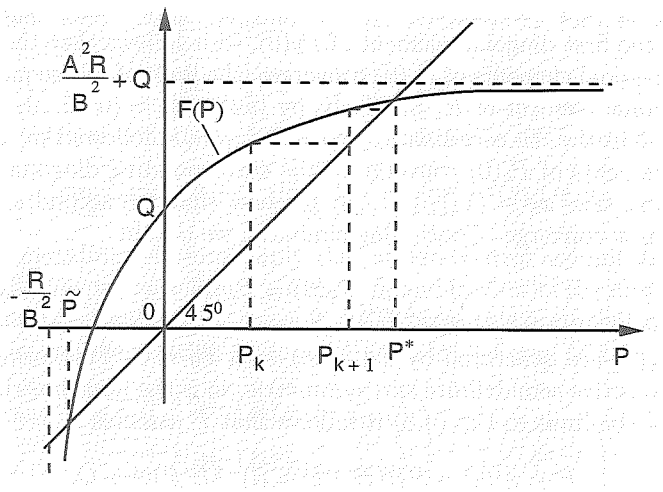

Figure 4.1.2 Graphical proof of Prop. 4.4.1 for the case of a scalar stationary system (one-dimensional state and control), assuming that $A \neq 0$, $B \neq 0$, $Q > 0$, and $R > 0$. The Riccati equation (4.8) is given by

$$P_{k+1} = A^2\left(P_k - \frac{B^2P_k^2}{B^2P_k + R}\right) + Q,$$

which can be equivalently written as

$$P_{k+1} = F(P_k),$$

where the function F is given by

$$F(P) = \frac{A^2RP}{B^2P + R} + Q.$$

Because F is concave and monotonically increasing in the interval $(-R/B^2, \infty)$, as shown in the figure, the equation $P = F(P)$ has one positive solution P^* and one negative solution $\tilde{P}$. The Riccati iteration $P_{k+1} = F(P_k)$ converges to P^* starting anywhere in the interval $(\tilde{P}, \infty)$ as shown in the figure.

where both minimizations are subject to the system equation constraint $x_{i+1} = Ax_i + Bu_i$. Furthermore, for a fixed x_0 and for every k, $x_0'P_k(0)x_0$ is bounded from above by the cost corresponding to a control sequence that forces x_0 to the origin in n steps and applies zero control after that. Such a sequence exists by the controllability assumption. Thus the sequence $\{x_0'P_k(0)x_0\}$ is nondecreasing with respect to k and bounded from above, and therefore converges to some real number for every $x_0 \in \Re^n$. It follows that the sequence $\{P_k(0)\}$ converges to some matrix P in the sense that each of the sequences of the elements of $P_k(0)$ converges to the corresponding

ing elements of P. To see this, take $x_0 = (1, 0, \ldots, 0)$. Then $x_0' P_k(0) x_0$ is equal to the first diagonal element of $P_k(0)$, so it follows that the sequence of first diagonal elements of $P_k(0)$ converges; the limit of this sequence is the first diagonal element of P. Similarly, by taking $x_0 = (0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in the ith coordinate, for $i = 2, \ldots, n$, it follows that all the diagonal elements of $P_k(0)$ converge to the corresponding diagonal elements of P. Next take $x_0 = (1, 1, 0, \ldots, 0)$ to show that the second elements of the first row converge. Continuing similarly, we obtain

$$\lim_{k \to \infty} P_k(0) = P,$$

where $P_k(0)$ are generated by Eq. (4.8) with $P_0 = 0$. Furthermore, since $P_k(0)$ is positive semidefinite and symmetric, so is the limit matrix P. Now by taking the limit in Eq. (4.8) it follows that P satisfies

$$P = A'\big(P - PB(B'PB + R)^{-1}B'P\big)A + Q.$$

In addition, by direct calculation we can verify the following useful equality

$$P = D'PD + Q + L'RL, \tag{4.12}$$

where D and L are given by Eqs. (4.10) and (4.11). An alternative way to derive this equality is to observe that from the DP algorithm corresponding to a finite horizon N we have for all states x_{N-k}

$$x_{N-k}' P_{k+1}(0) x_{N-k} = x_{N-k}' Q x_{N-k} + \mu_{N-k}^*(x_{N-k})' R \mu_{N-k}^*(x_{N-k}) + x_{N-k+1}' P_k(0) x_{N-k+1}.$$

By using the optimal controller expression $\mu_{N-k}^*(x_{N-k}) = L_{N-k} x_{N-k}$ and the closed-loop system equation $x_{N-k+1} = (A + BL_{N-k})x_{N-k}$, we thus obtain

$$P_{k+1}(0) = Q + L_{N-k}' R L_{N-k} + (A + BL_{N-k})' P_k(0)(A + BL_{N-k}). \tag{4.13}$$

Equation (4.12) then follows by taking the limit as $k \to \infty$ in Eq. (4.13).

Stability of the Closed-Loop System. Consider the system

$$x_{k+1} = (A + BL)x_k = Dx_k \tag{4.14}$$

for an arbitrary initial state x_0. We will show that $x_k \to 0$ as $k \to \infty$. We have for all k, by using Eq. (4.12),

$$x_{k+1}' P x_{k+1} - x_k' P x_k = x_k'(D'PD - P)x_k = -x_k'(Q + L'RL)x_k.$$

Hence

$$x_{k+1}' P x_{k+1} = x_0' P x_0 - \sum_{i=0}^{k} x_i'(Q + L'RL)x_i. \tag{4.15}$$

The left-hand side of this equation is bounded below by zero, so it follows that

$$\lim_{k \to \infty} x_k'(Q + L'RL)x_k = 0.$$

Since R is positive definite and Q may be written as $C'C$, we obtain

$$\lim_{k \to \infty} Cx_k = 0, \qquad \lim_{k \to \infty} Lx_k = \lim_{k \to \infty} \mu^*(x_k) = 0. \tag{4.16}$$

The preceding relations imply that as the control asymptotically becomes negligible, we have $\lim_{k \to \infty} Cx_k = 0$, and in view of the observability assumption, this implies that $x_k \to 0$. To express this argument more precisely, let us use the relation $x_{k+1} = (A + BL)x_k$ [cf. Eq. (4.14)], to write

$$\begin{pmatrix} C\left(x_{k+n-1} - \sum_{i=1}^{n-1} A^{i-1} B L x_{k+n-i-1}\right) \\ C\left(x_{k+n-2} - \sum_{i=1}^{n-2} A^{i-1} B L x_{k+n-i-2}\right) \\ \vdots \\ C(x_{k+1} - BLx_k) \\ Cx_k \end{pmatrix} = \begin{pmatrix} CA^{n-1} \\ CA^{n-2} \\ \vdots \\ CA \\ C \end{pmatrix} x_k. \tag{4.17}$$

Since $Lx_k \to 0$ by Eq. (4.16), the left-hand side tends to zero and hence the right-hand side tends to zero also. By the observability assumption, however, the matrix multiplying x_k on the right side of (4.17) has full rank. It follows that $x_k \to 0$.

Positive Definiteness of P. Assume the contrary, i.e., there exists some $x_0 \neq 0$ such that $x_0' P x_0 = 0$. Since P is positive semidefinite, from Eq. (4.15) we obtain

$$x_k'(Q + L'RL)x_k = 0, \qquad k = 0, 1, \ldots$$

Since $x_k \to 0$, we obtain $x_k' Q x_k = x_k' C'C x_k = 0$ and $x_k' L'RL x_k = 0$, or

$$Cx_k = 0, \qquad Lx_k = 0, \qquad k = 0, 1, \ldots$$

Thus all the controls $\mu^*(x_k) = Lx_k$ of the closed-loop system are zero while we have $Cx_k = 0$ for all k. Based on the observability assumption, we will show that this implies $x_0 = 0$, thereby reaching a contradiction. Indeed, consider Eq. (4.17) for $k = 0$. By the preceding equalities, the left-hand side is zero and hence

$$0 = \begin{pmatrix} CA^{n-1} \\ \vdots \\ CA \\ C \end{pmatrix} x_0.$$

Since the matrix multiplying x_0 above has full rank by the observability assumption, we obtain $x_0 = 0$, which contradicts the hypothesis $x_0 \neq 0$

Arbitrary Initial Matrix P_0. Next we show that the sequence of matrices $\{P_k(P_0)\}$, defined by Eq. (4.8) when the starting matrix is an arbitrary positive semidefinite symmetric matrix P_0, converges to $P = \lim_{k\to\infty} P_k(0)$. Indeed, the optimal cost of the problem of minimizing

$$x_k' P_0 x_k + \sum_{i=0}^{k-1} (x_i' Q x_i + u_i' R u_i) \tag{4.18}$$

subject to the system equation $x_{i+1} = Ax_i + Bu_i$ is equal to $x_0' P_k(P_0) x_0$. Hence we have for every $x_0 \in \Re^n$

$$x_0' P_k(0) x_0 \leq x_0' P_k(P_0) x_0.$$

Consider now the cost (4.18) corresponding to the controller $\mu(x_k) = u_k = Lx_k$, where L is defined by Eq. (4.11). This cost is

$$x_0' \left(D^{k\prime} P_0 D^k + \sum_{i=0}^{k-1} D^{i\prime} (Q + L'RL) D^i \right) x_0$$

and is greater or equal to $x_0' P_k(P_0) x_0$, which is the optimal value of the cost (4.18). Hence we have for all k and $x \in \Re^n$

$$x' P_k(0) x \leq x' P_k(P_0) x \leq x' \left(D^{k\prime} P_0 D^k + \sum_{i=0}^{k-1} D^{i\prime} (Q + L'RL) D^i \right) x.$$

We have proved that

$$\lim_{k\to\infty} P_k(0) = P,$$

and we also have, using the fact $\lim_{k\to\infty} D^{k\prime} P_0 D^k = 0$, and the relation $Q + L'RL = P - D'PD$ [cf. Eq. (4.12)],

$$\begin{aligned} \lim_{k\to\infty} \left\{ D^{k\prime} P_0 D_k + \sum_{i=0}^{k-1} D^{i\prime} (Q + L'RL) D^i \right\} & \\ = \lim_{k\to\infty} \left\{ \sum_{i=0}^{k-1} D^{i\prime} (Q + L'RL) D^i \right\} & \\ = \lim_{k\to\infty} \left\{ \sum_{i=0}^{k-1} D^{i\prime} (P - D'PD) D^i \right\} & \\ = P. & \end{aligned} \tag{4.19}$$

Combining the preceding three equations, we obtain

for an arbitrary positive semidefinite symmetric initial matrix P_0.

Uniqueness of Solution. If $\tilde{P}$ is another positive semidefinite symmetric solution of the algebraic Riccati equation (4.9), we have $P_k(\tilde{P}) = \tilde{P}$ for all $k = 0, 1, \ldots$ From the convergence result just proved, we then obtain

$$\lim_{k\to\infty} P_k(\tilde{P}) = P,$$

implying that $\tilde{P} = P$. **Q.E.D.**

The assumptions of the preceding proposition can be relaxed somewhat. Suppose that, instead of controllability of the pair (A, B), we assume that the system is *stabilizable* in the sense that there exists an $m \times n$ feedback gain matrix G such that the closed-loop system $x_{k+1} = (A + BG)x_k$ is stable. Then the proof of convergence of $P_k(0)$ to some positive semidefinite P given previously carries through. [We use the stationary control law $\mu(x) = Gx$ for which the closed-loop system is stable to ensure that $x_0' P_k(0) x_0$ is bounded.] Suppose that, instead of observability of the pair (A, C), the system is assumed *detectable* in the sense that A is such that if $u_k \to 0$ and $Cx_k \to 0$ then it follows that $x_k \to 0$. (This essentially means that instability of the system can be detected by looking at the measurement sequence $\{z_k\}$ with $z_k = Cx_k$.) Then Eq. (4.16) implies that $x_k \to 0$ and that the system $x_{k+1} = (A + BL)x_k$ is stable. The other parts of the proof of the proposition follow similarly, with the exception of positive definiteness of P, which cannot be guaranteed anymore. (As an example, take $A = 0$, $B = 0$, $C = 0$, $R > 0$. Then both the stabilizability and the detectability assumptions are satisfied, but $P = 0$.)

To summarize, if the controllability and observability assumptions of the proposition are replaced by the preceding stabilizability and detectability assumptions, the conclusions of the proposition hold with the exception of positive definiteness of the limit matrix P, which can now only be guaranteed to be positive semidefinite.

Random System Matrices

We consider now the case where $\{A_0, B_0\}, \ldots, \{A_{N-1}, B_{N-1}\}$ are not known but rather are independent random matrices that are also independent of $w_0, w_1, \ldots, w_{N-1}$. Their probability distributions are given, and they are assumed to have finite second moments. This problem falls again within the framework of the basic problem by considering as disturbance at each time k the triplet (A_k, B_k, w_k). The DP algorithm is written as

$$J_N(x_N) = x_N' Q_N x_N,$$

$$J_k(x_k) = \min_{\quad} \quad E \quad \{x_k' Q_k x_k + u_k' R_k u_k + J_{k+1}(A_k x_k + B_k u_k + w_k)\}$$

Calculations very similar to those for the case where A_k, B_k are not random show that the optimal control law has the form

$$\mu_k^*(x_k) = L_k x_k,$$

where the gain matrices L_k are given by

$$L_k = -\big(R_k + E\{B_k' K_{k+1} B_k\}\big)^{-1} E\{B_k' K_{k+1} A_k\},$$

and where the matrices K_k are given by the recursive equation

$$K_N = Q_N,$$

$$K_k = E\{A_k' K_{k+1} A_k\} - E\{A_k' K_{k+1} B_k\}\big(R_k + E\{B_k' K_{k+1} B_k\}\big)^{-1} E\{B_k' K_{k+1} A_k\} + Q_k. \tag{4.20}$$

In the case of a stationary system and constant matrices Q_k and R_k it is not necessarily true that the above equation converges to a steady-state solution. This is demonstrated in Fig. 4.1.3 for a scalar system, where it is shown that if the expression

$$T = E\{A^2\}E\{B^2\} - \big(E\{A\}\big)^2\big(E\{B\}\big)^2$$

exceeds a certain threshold, the matrices K_k diverge to ∞ starting from any nonnegative initial condition. A possible interpretation is that if there is a lot of uncertainty about the system, as quantified by T, optimization over a long horizon is meaningless. This phenomenon has been called the *uncertainty threshold principle*; see Athans, Ku, and Gershwin [AGK77], and Ku and Athans [KuA77].

On Certainty Equivalence

We close this section by making an observation about the simplifications that arise when the cost is quadratic. Consider the minimization over u of

$$\underset{w}{E}\big\{(ax + bu + w)^2\big\},$$

where a and b are given scalars, x is known, and w is a random variable. The optimum is attained for

$$u^* = -\left(\frac{a}{b}\right)x - \left(\frac{1}{b}\right)E\{w\}.$$

Thus u^* depends on the probability distribution of w only through the

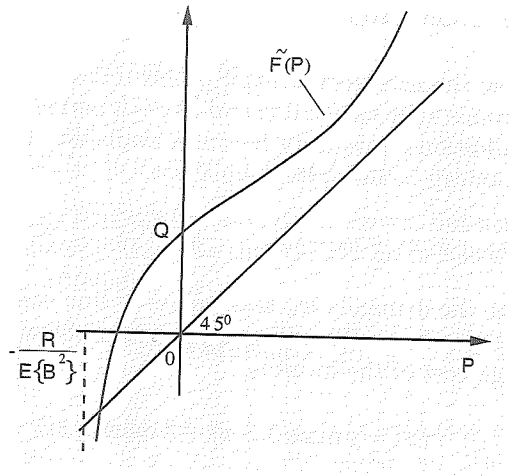

Figure 4.1.3 Graphical illustration of the asymptotic behavior of the generalized Riccati equation (4.20) in the case of a scalar stationary system (one-dimensional state and control). Using P_k in place of K_{N-k}, this equation is written as

$$P_{k+1} = \tilde{F}(P_k),$$

where the function $\tilde{F}$ is given by

$$\tilde{F}(P) = \frac{E\{A^2\}RP}{E\{B^2\}P + R} + Q + \frac{TP^2}{E\{B^2\}P + R},$$

$$T = E\{A^2\}E\{B^2\} - \big(E\{A\}\big)^2\big(E\{B\}\big)^2.$$

If $T = 0$, as in the case where A and B are not random, the Riccati equation becomes identical with the one of Fig. 4.1.2 and converges to a steady-state. Convergence also occurs when T has a small positive value. However, as illustrated in the figure, for T large enough, the graph of the function $\tilde{F}$ and the 45-degree line that passes through the origin do not intersect at a positive value of P, and the Riccati equation diverges to infinity.

for the corresponding deterministic problem where w is replaced by $E\{w\}$. This property is called the *certainty equivalence principle* and appears in various forms in many (but not all) stochastic control problems involving linear systems and quadratic cost. For the first problem of this section, where A_k, B_k are known, certainty equivalence holds because the optimal control law (4.2) is the same as the one that would be obtained from the corresponding deterministic problem where w_k is not random but rather is known and is equal to zero (its expected value). However, for the problem where A_k, B_k are random, the certainty equivalence principle does not hold, since if one replaces A_k, B_k with their expected values in Eq. (4.20)

3

Undiscounted Problems

Contents

3.1. Unbounded Costs per Stage . . . . . p. 124
3.2. Linear Systems and Quadratic Cost . . . . . p. 140
3.3. Inventory Control . . . . . p. 142
3.4. Optimal Stopping . . . . . p. 145
3.5. Optimal Gambling Strategies . . . . . p. 150
3.6. Nonstationary and Periodic Problems . . . . . p. 157
3.7. Notes, Sources, and Exercises . . . . . p. 162

In this chapter we consider total cost infinite horizon problems where we allow costs per stage that are unbounded above or below. Also, the discount factor α does not have to be less than one. The complications resulting are substantial, and the analysis required is considerably more sophisticated than the one given thus far. We also consider applications of the theory to important classes of problems. The exercise section touches on several related topics.

3.1 UNBOUNDED COSTS PER STATE

In this section we consider the total cost infinite horizon problem of Section 1.1 under one of the following two assumptions.

Assumption P: (Positivity) The cost per stage g satisfies

$$0 \leq g(x,u,w), \qquad \text{for all } (x,u,w) \in S \times C \times D. \tag{3.1}$$

Assumption N: (Negativity) The cost per stage g satisfies

$$g(x,u,w) \leq 0, \qquad \text{for all } (x,u,w) \in S \times C \times D. \tag{3.2}$$

Somewhat paradoxically, problems corresponding to Assumption P are sometimes referred to in the research literature as *negative DP problems*. This choice of name is due to historical reasons. It was introduced in an early paper by Strauch [Str66], where the problem of maximizing the infinite sum of negative rewards per stage was considered. Similarly, problems corresponding to Assumption N are sometimes referred to as *positive DP problems* (Blackwell [Bla65], Strauch [Str66]). Assumption N arises in problems where there is a nonnegative reward per stage and the total expected reward is to be *maximized*.

Note that when $\alpha < 1$ and g is either bounded above or below, we may add a suitable scalar to g in order to satisfy Eq. (3.1) or Eq. (3.2), respectively. An optimal policy will not be affected by this change since, because of the discount factor, the addition of a constant r to g merely adds $(1-\alpha)^{-1}r$ to the cost of every policy.

One complication arising from unbounded costs per stage is that, for

$\pi = \{\mu_0, \mu_1, \ldots\}$, the cost $J_\pi(x_0)$ may be ∞ (in the case of Assumption P) or $-\infty$ (in the case of Assumption N). Here is an example:

Example 3.1.1

Consider the scalar system

$$x_{k+1} = \beta x_k + u_k, \qquad k = 0, 1, \ldots,$$

where $x_k \in \Re$ and $u_k \in \Re$, for all k, and β is a positive scalar. The control constraint is $|u_k| \leq 1$, and the cost is

$$J_\pi(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k |x_k|.$$

Consider the policy $\tilde{\pi} = \{\tilde{\mu}, \tilde{\mu}, \ldots\}$, where $\tilde{\mu}(x) = 0$ for all $x \in \Re$. Then

$$J_{\tilde{\pi}}(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k \beta^k |x_0|,$$

and hence

$$J_{\tilde{\pi}}(x_0) = \begin{cases} 0 & \text{if } x_0 = 0 \\ \infty & \text{if } x_0 \neq 0 \end{cases} \qquad \text{if} \qquad \alpha\beta \geq 1,$$

while

$$J_{\tilde{\pi}}(x_0) = \frac{|x_0|}{1-\alpha\beta} \qquad \text{if} \qquad \alpha\beta < 1.$$

Note a peculiarity here: if $\beta > 1$ the state x_k diverges to ∞ or to $-\infty$, but if the discount factor is sufficiently small ($\alpha < 1/\beta$), the cost $J_{\tilde{\pi}}(x_0)$ is finite.

It is also possible to verify that when $\beta > 1$ and $\alpha\beta \geq 1$ the optimal cost $J^*(x_0)$ is equal to ∞ for $|x_0| \geq 1/(\beta-1)$ and is finite for $|x_0| < 1/(\beta-1)$. What happens here is that when $\beta > 1$ the system is unstable, and in view of the restriction $|u_k| \leq 1$ on the control, it may not be possible to force the state near zero once it has reached sufficiently large magnitude.

The preceding example shows that there is not much that can be done about the possibility of the cost function being infinite for some policies. To cope with this situation, we conduct our analysis with the notational understanding that the costs $J_\pi(x_0)$ and $J^*(x_0)$ may be ∞ (or $-\infty$) under Assumption P (or N, respectively) for some initial states x_0 and policies π. In other words, we consider $J_\pi(\cdot)$ and $J^*(\cdot)$ to be extended real-valued functions. In fact, the entire subsequent analysis is valid even if the cost $g(x,u,w)$ is ∞ or $-\infty$ for some (x,u,w), as long as Assumption P or Assumption N holds.

The line of analysis of this section is fundamentally different from the one of the discounted problem of Section 1.2. For the latter problem,

this section, the tails of the cost sequences may not be small, and for this reason, the control is much more focused on affecting the long-term behavior of the state. For example, let $\alpha = 1$, and assume that the stage cost at all states is nonzero except for a cost-free and absorbing termination state. Then, a primary task of control under Assumption P (or Assumption N) is roughly to bring the state of the system to the termination state or to a region where the cost per stage is nearly zero as *quickly* as possible (as *late* as possible, respectively). Note the difference in control objective between Assumptions P and N. It accounts for some strikingly different results under the two assumptions.

Main Results – Bellman's Equation

We now present results that characterize the optimal cost function J^*, as well as optimal stationary policies. We also give conditions under which value iteration converges to the optimal cost function J^*. In the proofs we will often need to interchange expectation and limit in various relations. This interchange is valid under the assumptions of the following theorem.

Monotone Convergence Theorem: Let $P = (p_1, p_2, \ldots)$ be a probability distribution over $S = \{1, 2, \ldots\}$. Let $\{h_N\}$ be a sequence of extended real-valued functions on S such that for all $i \in S$ and $N = 1, 2, \ldots,$

$$0 \leq h_N(i) \leq h_{N+1}(i).$$

Let $h : S \mapsto [0, \infty]$ be the limit function

$$h(i) = \lim_{N \to \infty} h_N(i).$$

Then

$$\lim_{N \to \infty} \sum_{i=1}^{\infty} p_i h_N(i) = \sum_{i=1}^{\infty} p_i \lim_{N \to \infty} h_N(i) = \sum_{i=1}^{\infty} p_i h(i).$$

Proof: We have

$$\sum_{i=1}^{\infty} p_i h_N(i) \leq \sum_{i=1}^{\infty} p_i h(i).$$

By taking the limit, we obtain

$$\lim \sum_{i=1}^{\infty} p_i h_N(i) \leq \sum_{i=1}^{\infty} p_i h(i),$$

so there remains to prove the reverse inequality. For every integer $M \geq 1$, we have

$$\lim_{N \to \infty} \sum_{i=1}^{\infty} p_i h_N(i) \geq \lim_{N \to \infty} \sum_{i=1}^{M} p_i h_N(i) = \sum_{i=1}^{M} p_i h(i),$$

and by taking the limit as $M \to \infty$ the reverse inequality follows. **Q.E.D.**

Similar to all the infinite horizon problems considered so far, the optimal cost function satisfies Bellman's equation.

Proposition 3.1.1: (Bellman's Equation) Under either Assumption P or N the optimal cost function J^* satisfies

$$J^*(x) = \min_{u \in U(x)} \mathop{E}_{w} \{ g(x, u, w) + \alpha J^*\big(f(x, u, w)\big) \}, \qquad x \in S$$

or, equivalently,

$$J^* = TJ^*.$$

Proof: For any admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$, consider the cost $J_\pi(x)$ corresponding to π when the initial state is x. We have

$$J_\pi(x) = \mathop{E}_{w} \{ g\big(x, \mu_0(x), w\big) + V_\pi\big(f(x, \mu_0(x), w)\big) \}, \tag{3.3}$$

where, for all $x_1 \in S$,

$$V_\pi(x_1) = \lim_{N \to \infty} \mathop{E}_{\substack{w_k \\ k=1,2,\ldots}} \left\{ \sum_{k=1}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}.$$

Thus, $V_\pi(x_1)$ is the cost from stage 1 to infinity using π when the initial state is x_1. We clearly have

$$V_\pi(x_1) \geq \alpha J^*(x_1), \qquad \text{for all } x_1 \in S.$$

Hence, from Eq. (3.3),

$$\begin{aligned} J_\pi(x) &\geq \mathop{E}_{w} \{ g\big(x, \mu_0(x), w\big) + \alpha J^*\big(f(x, \mu_0(x), w)\big) \} \\ &\geq \min_{u \in U(x)} \mathop{E}_{w} \{ g(x, u, w) + \alpha J^*\big(f(x, u, w)\big) \}. \end{aligned}$$

Taking the minimum over all admissible policies, we obtain

$$\begin{aligned} \min_{\pi} J_\pi(x) &= J^*(x) \\ &\geq \min_{u \in U(x)} \mathop{E}_{w} \{ g(x, u, w) + \alpha J^*\big(f(x, u, w)\big) \} \end{aligned}$$

Thus there remains to prove that the reverse inequality also holds. We prove this separately for Assumption N and for Assumption P.

Assume P. The following proof of $J^* \leq TJ^*$ under this assumption would be considerably simplified if we knew that there exists a μ such that $T_\mu J^* = TJ^*$. Since in general such a μ need not exist, we introduce a positive sequence $\{\epsilon_k\}$, and we choose an admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$ such that

$$(T_{\mu_k} J^*)(x) \leq (TJ^*)(x) + \epsilon_k, \qquad x \in S, \quad k = 0, 1, \ldots$$

Such a choice is possible because under P, we have $0 \leq J^*(x)$ for all x. By using the inequality $TJ^* \leq J^*$ shown earlier, we obtain

$$(T_{\mu_k} J^*)(x) \leq J^*(x) + \epsilon_k, \qquad x \in S, \quad k = 0, 1, \ldots$$

Applying $T_{\mu_{k-1}}$ to both sides of this relation, we have

$$\begin{aligned}(T_{\mu_{k-1}} T_{\mu_k} J^*)(x) &\leq (T_{\mu_{k-1}} J^*)(x) + \alpha\epsilon_k \\ &\leq (TJ^*)(x) + \epsilon_{k-1} + \alpha\epsilon_k \\ &\leq J^*(x) + \epsilon_{k-1} + \alpha\epsilon_k.\end{aligned}$$

Continuing this process, we obtain

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J^*)(x) \leq (TJ^*)(x) + \sum_{i=0}^{k} \alpha^i \epsilon_i.$$

By taking the limit as $k \to \infty$ and noting that

$$J^*(x) \leq J_\pi(x) = \lim_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J_0)(x) \leq \lim_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J^*)(x),$$

where J_0 is the zero function, it follows that

$$J^*(x) \leq J_\pi(x) \leq (TJ^*)(x) + \sum_{i=0}^{\infty} \alpha^i \epsilon_i, \qquad x \in S.$$

Since the sequence $\{\epsilon_k\}$ is arbitrary, we can take $\sum_{i=0}^{\infty} \alpha^i \epsilon_i$ as small as desired, and we obtain $J^*(x) \leq (TJ^*)(x)$ for all $x \in S$. Combining this with the inequality $J^*(x) \geq (TJ^*)(x)$ shown earlier, the result follows (under Assumption P).

Assume N and let J_N be the optimal cost function for the corresponding N-stage problem

$$J_N(x_0) = \min_\pi E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\}.$$

We first show that

$$J^*(x) = \lim_{N\to\infty} J_N(x), \qquad x \in S. \tag{3.4}$$

Indeed, in view of Assumption N, we have $J^* \leq J_N$ for all N, so

$$J^*(x) \leq \lim_{N\to\infty} J_N(x), \qquad x \in S. \tag{3.5}$$

Also, for all $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \geq J_N(x_0),$$

and by taking the limit as $N \to \infty$,

$$J_\pi(x) \geq \lim_{N\to\infty} J_N(x), \qquad x \in S.$$

Taking the minimum over π, we obtain $J^*(x) \geq \lim_{N\to\infty} J_N(x)$, and combining this relation with Eq. (3.5), we obtain Eq. (3.4).

For every admissible μ, we have

$$T_\mu J_N \geq J_{N+1},$$

and by taking the limit as $N \to \infty$, and using the monotone convergence theorem and Eq. (3.4), we obtain

$$T_\mu J^* \geq J^*.$$

Taking the minimum over μ, we obtain $TJ^* \geq J^*$, which combined with the inequality $J^* \geq TJ^*$ shown earlier, proves the result under Assumption N. **Q.E.D.**

Similar to Cor. 1.2.2.1, we have:

Corollary 3.1.1.1: Let μ be a stationary policy. Then under Assumption P or N, we have

$$J_\mu(x) = \mathop{E}_w\big\{g\big(x, \mu(x), w\big) + \alpha J_\mu\big(f(x, \mu(x), w)\big)\big\}, \qquad x \in S$$

or, equivalently,

$$J_\mu = T_\mu J_\mu. \tag{3.6}$$

Contrary to discounted problems with bounded cost per stage, the optimal cost function J^* under Assumption P or N need not be the unique solution of Bellman's equation. Consider the following

Example 3.1.2

Let $S = [0, \infty)$ (or $S = (-\infty, 0]$) and

$$g(x, u, w) = 0, \qquad f(x, u, w) = \frac{x}{\alpha}.$$

Then for every β, the function J given by $J(x) = \beta x$ for all $x \in S$, is a solution of Bellman's equation, so T has an infinite number of fixed points. Note, however, that there is a unique fixed point within the class of bounded functions, the zero function $J_0(x) \equiv 0$, which is the optimal cost function for this problem. More generally, it can be shown by using the following Prop. 3.1.2 that if $\alpha < 1$ and there exists a bounded function that is a fixed point of T, then that function must be equal to the optimal cost function J^* (see Exercise 3.5). When $\alpha = 1$, Bellman's equation may have an infinity of solutions even within the class of bounded functions. This is because if $\alpha = 1$ and $J(\cdot)$ is any solution, then for any scalar r, $J(\cdot) + r$ is also a solution.

The optimal cost function J^*, however, has the property that it is the smallest (under Assumption P) or largest (under Assumption N) fixed point of T in the sense described in the following proposition.

Proposition 3.1.2:

(a) Under Assumption P, if $\tilde{J} : S \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \geq T\tilde{J}$ and either $\tilde{J}$ is bounded below and $\alpha < 1$, or $\tilde{J} \geq 0$, then $\tilde{J} \geq J^*$.

(b) Under Assumption N, if $\tilde{J} : S \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \leq T\tilde{J}$ and either $\tilde{J}$ is bounded above and $\alpha < 1$, or $\tilde{J} \leq 0$, then $\tilde{J} \leq J^*$.

Proof: (a) Under Assumption P, let r be a scalar such that $\tilde{J}(x) + r \geq 0$ for all $x \in S$ and if $\alpha \geq 1$ let $r = 0$. For any sequence $\{\epsilon_k\}$ with $\epsilon_k > 0$, let $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \ldots\}$ be an admissible policy such that, for every $x \in S$ and k,

$$\mathop{E}_{w}\{g\big(x, \mu_k(x), w\big) + \alpha \tilde{J}\big(f\big(x, \mu_k(x), w\big)\big)\} \leq (T\tilde{J})(x) + \epsilon_k. \tag{3.7}$$

Such a policy exists since $(T\tilde{J})(x) > -\infty$ for all $x \in S$. We have for any initial state $x_0 \in S$,

$$\begin{aligned} J^*(x_0) &= \min_{\pi} \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\leq \min_{\pi} \liminf_{N\to\infty} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\leq \liminf_{N\to\infty} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\}. \end{aligned}$$

Using Eq. (3.7) and the assumption $\tilde{J} \geq T\tilde{J}$, we obtain

$$\begin{aligned} &E\left\{\alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}(x_k), w_k\big)\right\} \\ &= E\left\{\alpha^N \tilde{J}\big(f\big(x_{N-1}, \tilde{\mu}_{N-1}(x_{N-1}), w_{N-1}\big)\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\} \\ &\leq E\left\{\alpha^{N-1} \tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} + \alpha^{N-1}\epsilon_{N-1} \\ &\leq E\left\{\alpha^{N-2} \tilde{J}(x_{N-2}) + \sum_{k=0}^{N-3} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\} + \alpha^{N-2}\epsilon_{N-2} \\ &\quad + \alpha^{N-1}\epsilon_{N-1} \\ &\quad \vdots \\ &\leq \tilde{J}(x_0) + \sum_{k=0}^{N-1} \alpha^k \epsilon_k. \end{aligned}$$

Combining these inequalities, we obtain

$$J^*(x_0) \leq \tilde{J}(x_0) + \lim_{N\to\infty} \left(\alpha^N r + \sum_{k=0}^{N-1} \alpha^k \epsilon_k\right).$$

Since $\{\epsilon_k\}$ is an arbitrary positive sequence, we may select $\{\epsilon_k\}$ so that $\lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k \epsilon_k$ is arbitrarily close to zero, and the result follows.

(b) Under Assumption N, let r be a scalar such that $\tilde{J}(x) + r \leq 0$ for all $x \in S$, and if $\alpha \geq 1$, let $r = 0$. We have for every initial state $x_0 \in S$,

$$\begin{aligned} J^*(x_0) &= \min_{\pi} \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\geq \min_{\pi} \limsup_{N\to\infty} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\geq \limsup_{N\to\infty} \min_{\pi} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\}, \end{aligned} \tag{3.8}$$

where the last inequality follows from the fact that for any sequence $\{h_N(\xi)\}$ of functions of a parameter ξ we have

$$\min_{\xi} \limsup_{N\to\infty} h_N(\xi) \geq \limsup_{N\to\infty} \min_{\xi} h_N(\xi).$$

This inequality follows by writing

$$h_N(\xi) \geq \min_{\xi} h_N(\xi)$$

and by subsequently taking the lim sup of both sides and the minimum over ξ of the left-hand side.

Now we have, by using the assumption $\tilde{J} \leq T\tilde{J}$,

$$\begin{aligned}
&\min_{\pi} E\left\{\alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\qquad = \min_{\pi} E\left\{\alpha^{N-1} \min_{u_{N-1}\in U(x_{N-1})} E_{w_{N-1}}\big\{g(x_{N-1}, u_{N-1}, w_{N-1}) \right. \right. \\
&\qquad\qquad \left. \left. + \alpha\tilde{J}\big(f(x_{N-1}, u_{N-1}, w_{N-1})\big)\big\} + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\qquad \geq \min_{\pi} E\left\{\alpha^{N-1}\tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\qquad\qquad \vdots \\
&\qquad \geq \tilde{J}(x_0).
\end{aligned}$$

Using this relation in Eq. (3.8), we obtain

$$J^*(x_0) \geq \tilde{J}(x_0) + \lim_{N\to\infty} \alpha^N r = \tilde{J}(x_0).$$

Q.E.D.

As before, we have the following corollary:

Corollary 3.1.2.1: Let μ be an admissible stationary policy.

(a) Under Assumption P, if $\tilde{J} : S \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \geq T_\mu \tilde{J}$ and either $\tilde{J}$ is bounded below and $\alpha < 1$, or $\tilde{J} \geq 0$, then $\tilde{J} \geq J_\mu$.

(b) Under Assumption N, if $\tilde{J} : S \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \leq T_\mu \tilde{J}$ and either $\tilde{J}$ is bounded above and $\alpha < 1$, or $\tilde{J} \leq 0$, then $\tilde{J} \leq J_\mu$.

Conditions for Optimality of a Stationary Policy

Under Assumption P, we have the same optimality condition as for dis-

Proposition 3.1.3: (Necessary and Sufficient Condition for Optimality under P) Let Assumption P hold. A stationary policy μ is optimal if and only if

$$TJ^* = T_\mu J^*.$$

Proof: If $TJ^* = T_\mu J^*$, Bellman's equation ($J^* = TJ^*$) implies that $J^* = T_\mu J^*$. From Cor. 3.1.2.1(a) we then obtain $J^* \geq J_\mu$, showing that μ is optimal. Conversely, if $J^* = J_\mu$, we have using Cor. 3.1.1.1, $TJ^* = J^* = J_\mu = T_\mu J_\mu = T_\mu J^*$. Q.E.D.

Note that when $U(x)$ is a finite set for every $x \in S$, the above proposition implies the existence of an optimal stationary policy under Assumption P. This need not be true under Assumption N (see the subsequent Example 3.4.4).

Unfortunately, the sufficiency part of the above proposition need not be true under Assumption N; i.e., we may have $TJ^* = T_\mu J^*$ while μ is not optimal. This is illustrated in the following example.

Example 3.1.3

Let $S = C = (-\infty, 0]$, $U(x) = C$ for all $x \in S$, and

$$g(x, u, w) = f(x, u, w) = u,$$

for all $(x, u, w) \in S \times C \times D$. Then $J^*(x) = -\infty$ for all $x \in S$, and every stationary policy μ satisfies the condition of the preceding proposition. On the other hand, when $\mu(x) = 0$ for all $x \in S$, we have $J_\mu(x) = 0$ for all $x \in S$, and hence μ is not optimal.

Under Assumption N, we have a different characterization of an optimal stationary policy.

Proposition 3.1.4: (Necessary and Sufficient Condition for Optimality under N) Let Assumption N hold. A stationary policy μ is optimal if and only if

$$TJ_\mu = T_\mu J_\mu. \tag{3.9}$$

Proof: If $TJ_\mu = T_\mu J_\mu$, then from Cor. 3.1.1.1 we have $J_\mu = T_\mu J_\mu$, so that

implies that μ is optimal. Conversely, if $J_\mu = J^*$, then $T_\mu J_\mu = J_\mu = J^* = TJ^* = TJ_\mu$. Q.E.D.

The interpretation of the preceding optimality condition is that persistently using μ is optimal if and only if this performs at least as well as using any $\overline{\mu}$ at the first stage and using μ thereafter. Under Assumption P this condition is not sufficient to guarantee optimality of the stationary policy μ, as the following example shows.

Example 3.1.4

Let $S = (-\infty, \infty)$, $U(x) = (0, 1]$ for all $x \in S$,

$$g(x, u, w) = |x|, \qquad f(x, u, w) = \alpha^{-1}ux,$$

for all $(x, u, w) \in S \times C \times D$. Let $\mu(x) = 1$ for all $x \in S$. Then $J_\mu(x) = \infty$ if $x \neq 0$ and $J_\mu(0) = 0$. Furthermore, we have $J_\mu = T_\mu J_\mu = TJ_\mu$, as the reader can easily verify. It can also be verified that $J^*(x) = |x|$, and hence the stationary policy μ is not optimal.

The Value Iteration Method

We now turn to the question whether the DP algorithm converges to the optimal cost function J^*. Let J_0 be the zero function on S,

$$J_0(x) = 0, \qquad x \in S.$$

Then under Assumption P, we have

$$J_0 \leq TJ_0 \leq T^2J_0 \leq \cdots \leq T^kJ_0 \leq \cdots,$$

while under Assumption N, we have

$$J_0 \geq TJ_0 \geq T^2J_0 \geq \cdots \geq T^kJ_0 \geq \cdots$$

In either case the limit function

$$J_\infty(x) = \lim_{k\to\infty} (T^kJ_0)(x), \qquad x \in S,$$

is well defined, provided we allow the possibility that J_∞ can take the value ∞ (under Assumption P) or $-\infty$ (under Assumption N). The question is whether the value iteration method is valid in the sense

This question is, of course, of computational interest, but it is also of analytical interest since, if we know that $J^* = \lim_{k\to\infty} T^kJ_0$, we can infer properties of the unknown function J^* from properties of the k-stage optimal cost functions T^kJ_0, which are defined in a concrete algorithmic manner.

We will show that $J_\infty = J^*$ under Assumption N. It turns out, however, that under Assumption P, we may have $J_\infty \neq J^*$ (see Exercise 3.1). We will later provide easily verifiable conditions guaranteeing that $J_\infty = J^*$ under Assumption P. We have the following proposition.

Proposition 3.1.5:

(a) Let Assumption P hold and assume that

$$J_\infty(x) = (TJ_\infty)(x), \qquad x \in S.$$

Then if $J : S \mapsto \Re$ is any bounded function and $\alpha < 1$, or otherwise if $J_0 \leq J \leq J^*$, we have

$$\lim_{k\to\infty} (T^kJ)(x) = J^*(x), \qquad x \in S. \tag{3.10}$$

(b) Let Assumption N hold. Then if $J : S \mapsto \Re$ is any bounded function and $\alpha < 1$, or otherwise if $J^* \leq J \leq J_0$, we have

$$\lim_{k\to\infty} (T^kJ)(x) = J^*(x), \qquad x \in S.$$

Proof: (a) Since under Assumption P, we have

$$J_0 \leq TJ_0 \leq \cdots \leq T^kJ_0 \leq \cdots \leq J^*,$$

it follows that $\lim_{k\to\infty} T^kJ_0 = J_\infty \leq J^*$. Since J_∞ is also a fixed point of T by assumption, we obtain from Prop. 3.1.2(a) that $J^* \leq J_\infty$. It follows that

$$J_\infty = J^*,$$

and hence Eq. (3.10) is proved for the case $J = J_0$.

For the case where $\alpha < 1$ and J is bounded, let r be a scalar such that

$$J_0 - re \leq J \leq J_0 + re.$$

Applying T^k to this relation, we obtain

Since $T^k J_0$ converges to J^*, as shown earlier, this relation implies that $T^k J$ converges also to J^*.

In the case where $J_0 \leq J \leq J^*$, we have by applying T^k

$$T^k J_0 \leq T^k J \leq J^*, \qquad k = 0, 1, \ldots$$

Since $T^k J_0$ converges to J^*, so does $T^k J$.

(b) It was shown earlier [cf. Eq. (3.4)] that under Assumption N, we have

$$J_\infty(x) = \lim_{k \to \infty} (T^k J_0)(x) = J^*(x).$$

The proof from this point is identical to that for part (a). **Q.E.D.**

We now derive conditions guaranteeing that $J_\infty = TJ_\infty$ holds under Assumption P, which by Prop. 3.1.5 implies that $J_\infty = J^*$. We prove two propositions. The first admits an easy proof but requires a finiteness assumption on the control constraint set. The second is harder to prove but requires a weaker compactness assumption.

Proposition 3.1.6: Let Assumption P hold and assume that the control constraint set is finite for every $x \in S$. Then

$$J_\infty = TJ_\infty = J^*.$$

Proof: As shown in the proof of Prop. 3.1.5(a), we have for all k, $T^k J_0 \leq J_\infty \leq J^*$. Applying T to this relation, we obtain

$$\begin{aligned}(T^{k+1} J_0)(x) &= \min_{u \in U(x)} \mathop{E}_{w} \big\{ g(x,u,w) + \alpha (T^k J_0)\big(f(x,u,w)\big) \big\} \\ &\leq (TJ_\infty)(x),\end{aligned} \tag{3.11}$$

and by taking the limit as $k \to \infty$, it follows that

$$J_\infty \leq TJ_\infty.$$

Suppose that there existed a state $\tilde{x} \in S$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \tag{3.12}$$

Let u_k minimize in Eq. (3.11) when $x = \tilde{x}$. Since $U(\tilde{x})$ is finite, there must exist some $\tilde{u} \in U(\tilde{x})$ such that $u_k = \tilde{u}$ for all k in some infinite subset K of the positive integers. By Eq. (3.11) we have for all $k \in K$

$$\begin{aligned}(T^{k+1} J_0)(\tilde{x}) &= \mathop{E}_{w} \big\{ g(\tilde{x},\tilde{u},w) + \alpha (T^k J_0)\big(f(\tilde{x},\tilde{u},w)\big) \big\} \\ &\leq (TJ_\infty)(\tilde{x}).\end{aligned}$$

Taking the limit as $k \to \infty$, $k \in K$, we obtain

$$\begin{aligned}J_\infty(\tilde{x}) &= \mathop{E}_{w} \big\{ g(\tilde{x},\tilde{u},w) + \alpha J_\infty\big(f(\tilde{x},\tilde{u},w)\big) \big\} \\ &\geq (TJ_\infty)(\tilde{x}) \\ &= \min_{u \in U(\tilde{x})} \mathop{E}_{w} \big\{ g(\tilde{x},u,w) + \alpha J_\infty\big(f(\tilde{x},u,w)\big) \big\}.\end{aligned}$$

This contradicts Eq. (3.12), so we have $J_\infty(\tilde{x}) = (TJ_\infty)(\tilde{x})$. **Q.E.D.**

The following proposition strengthens Prop. 3.1.6 in that it requires a compactness rather than a finiteness assumption. We recall (see Appendix A of Vol. I) that a subset X of the n-dimensional Euclidean space $\Re^n$ is said to be *compact* if every sequence $\{x_k\}$ with $x_k \in X$ contains a subsequence $\{x_k\}_{k \in K}$ that converges to a point $x \in X$. Equivalently, X is compact if and only if it is closed and bounded. The empty set is (trivially) considered compact. Given any collection of compact sets, their intersection is a compact set (possibly empty). Given a sequence of nonempty compact sets $X_1, X_2 \ldots, X_k, \ldots$ such that

$$X_1 \supset X_2 \supset \cdots \supset X_k \supset X_{k+1} \supset \cdots$$

their intersection $\cap_{k=1}^\infty X_k$ is both nonempty and compact. In view of this fact, it follows that if $f : \Re^n \mapsto [-\infty, \infty]$ is a function such that the set

$$F_\lambda = \big\{ x \in \Re^n \mid f(x) \leq \lambda \big\} \tag{3.13}$$

is compact for every $\lambda \in R$, then there exists a vector x^* minimizing f; i.e., there exists an $x^* \in \Re^n$ such that

$$f(x^*) = \min_{x \in \Re^n} f(x).$$

To see this, take a sequence $\{\lambda_k\}$ such that $\lambda_k \to \min_{x \in \Re^n} f(x)$ and $\lambda_k \geq \lambda_{k+1}$ for all k. If $\min_{x \in \Re^n} f(x) < \infty$, such a sequence exists and the sets

$$F_{\lambda_k} = \{ x \in \Re^n \mid f(x) \leq \lambda_k \}$$

are nonempty and compact. Furthermore, $F_{\lambda_k} \supset F_{\lambda_{k+1}}$ for all k, and hence the intersection $\cap_{k=1}^\infty F_{\lambda_k}$ is also nonempty and compact. Let x^* be any vector in $\cap_{k=1}^\infty F_{\lambda_k}$. Then

$$f(x^*) \leq \lambda_k, \qquad k = 1, 2, \ldots,$$

and taking the limit as $k \to \infty$, we obtain $f(x^*) \leq \min_{x \in \Re^n} f(x)$, proving

that the set F_λ of Eq. (3.13) is compact for all λ is when f is continuous and $f(x) \to \infty$ as $\|x\| \to \infty$.

Proposition 3.1.7: Let Assumption P hold, and assume that the sets

$$U_k(x,\lambda) = \left\{ u \in U(x) \,\middle|\, \underset{w}{E}\{g(x,u,w) + \alpha(T^k J_0)\big(f(x,u,w)\big)\} \le \lambda \right\} \tag{3.14}$$

are compact subsets of a Euclidean space for every $x \in S$, $\lambda \in \Re$, and for all k greater than some integer $\overline{k}$. Then

$$J_\infty = TJ_\infty = J^*. \tag{3.15}$$

Furthermore, there exists a stationary optimal policy.

Proof: As in Prop. 3.1.6, we have $J_\infty \le TJ_\infty$. Suppose that there existed a state $\tilde{x} \in S$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \tag{3.16}$$

Clearly, we must have $J_\infty(\tilde{x}) < \infty$. For every $k \ge \overline{k}$, consider the sets

$$U_k\big(\tilde{x}, J_\infty(\tilde{x})\big) = \left\{ u \in U(\tilde{x}) \,\middle|\, \underset{w}{E}\{g(\tilde{x},u,w) + \alpha(T^k J_0)\big(f(\tilde{x},u,w)\big)\} \le J_\infty(\tilde{x}) \right\}.$$

Let also u_k be a point attaining the minimum in

$$(T^{k+1}J_0)(\tilde{x}) = \min_{u \in U(\tilde{x})} \underset{w}{E}\{g(\tilde{x},u,w) + \alpha(T^k J_0)\big(f(\tilde{x},u,w)\big)\};$$

i.e., u_k is such that

$$(T^{k+1}J_0)(\tilde{x}) = \underset{w}{E}\{g(\tilde{x},u_k,w) + \alpha(T^k J_0)\big(f(\tilde{x},u_k,w)\big)\}.$$

Such minimizing points u_k exist by our compactness assumption. For every $k \ge \overline{k}$, consider the sequence $\{u_i\}_{i=k}^\infty$. Since $T^k J_0 \le T^{k+1} J_0 \le \cdots \le J_\infty$, it follows that

$$\underset{w}{E}\{g(\tilde{x},u_i,w) + \alpha(T^k J_0)\big(f(\tilde{x},u_i,w)\big)\} \le \underset{w}{E}\{g(\tilde{x},u_i,w) + \alpha(T^i J_0)\big(f(\tilde{x},u_i,w)\big)\}$$

Therefore $\{u_i\}_{i=k}^\infty \subset U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$, and since $U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$ is compact, all the limit points of $\{u_i\}_{i=k}^\infty$ belong to $U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$ and at least one such limit point exists. Hence the same is true of the limit points of the whole sequence $\{u_i\}_{i=\overline{k}}^\infty$. It follows that if $\tilde{u}$ is a limit point of $\{u_i\}_{i=\overline{k}}^\infty$ then

$$\tilde{u} \in \cap_{k=\overline{k}}^\infty U_k\big(\tilde{x}, J_\infty(\tilde{x})\big).$$

This implies by Eq. (3.14) that for all $k \ge \overline{k}$

$$J_\infty(\tilde{x}) \ge \underset{w}{E}\{g(\tilde{x},\tilde{u},w) + \alpha(T^k J_0)\big(f(\tilde{x},\tilde{u},w)\big)\} \ge (T^{k+1}J_0)(\tilde{x}).$$

Taking the limit as $k \to \infty$, we obtain

$$J_\infty(\tilde{x}) = \underset{w}{E}\{g(\tilde{x},\tilde{u},w) + \alpha J_\infty\big(f(\tilde{x},\tilde{u},w)\big)\}.$$

Since the right-hand side is greater than or equal to $(TJ_\infty)(\tilde{x})$, Eq. (3.16) is contradicted. Hence $J_\infty = TJ_\infty$ and Eq. (3.15) is proved in view of Prop. 3.1.5(a).

To show that there exists an optimal stationary policy, observe that Eq. (3.15) and the last relation imply that $\tilde{u}$ attains the minimum in

$$J^*(\tilde{x}) = \min_{u \in U(\tilde{x})} \underset{w}{E}\{g(\tilde{x},u,w) + \alpha J^*\big(f(\tilde{x},u,w)\big)\}$$

for a state $\tilde{x} \in S$ with $J^*(\tilde{x}) < \infty$. For states $\tilde{x} \in S$ such that $J^*(\tilde{x}) = \infty$, every $u \in U(\tilde{x})$ attains the preceding minimum. Hence by Prop. 3.1.3(a) an optimal stationary policy exists. **Q.E.D.**

The reader may verify by inspection of the preceding proof that if $\mu_k(\tilde{x})$, $k = 0, 1, \ldots$, attains the minimum in the relation

$$(T^{k+1}J_0)(\tilde{x}) = \min_{u \in U(x)} \underset{w}{E}\{g(\tilde{x},u,w) + \alpha(T^k J_0)\big(f(\tilde{x},u,w)\big)\},$$

then if $\mu^*(\tilde{x})$ is a limit point of $\{\mu_k(\tilde{x})\}$, for every $\tilde{x} \in S$, the stationary policy μ^* is optimal. Furthermore, $\{\mu_k(\tilde{x})\}$ has at least one limit point for every $\tilde{x} \in S$ for which $J^*(\tilde{x}) < \infty$. Thus *the value iteration method under the assumptions of either Prop. 3.1.6 or Prop. 3.1.7 yields in the limit not only the optimal cost function J^* but also an optimal stationary policy.*

Other Computational Methods

Unfortunately, policy iteration is not a valid procedure under either P or N in the absence of further conditions. If μ and $\overline{\mu}$ are stationary policies such that $T_{\overline{\mu}} J_\mu = TJ_\mu$, then it can be shown that under Assumption P we have

To see this, note that $T_{\overline{\mu}}J_\mu = TJ_\mu \leq T_\mu J_\mu = J_\mu$ from which we obtain $\lim_{N\to\infty} T^N_{\overline{\mu}} J_\mu \leq J_\mu$. Since $J_{\overline{\mu}} = \lim_{N\to\infty} T^N_{\overline{\mu}} J_0$ and $J_0 \leq J_\mu$, we obtain $J_{\overline{\mu}} \leq J_\mu$. However, $J_{\overline{\mu}} \leq J_\mu$ by itself is not sufficient to guarantee the validity of policy iteration. For example, it is not clear that strict inequality holds in Eq. (3.17) for at least one state $x \in S$ when μ is not optimal. The difficulty here is that the equality $J_\mu = TJ_\mu$ does not imply that μ is optimal, and additional conditions are needed to guarantee the validity of policy iteration. However, for special cases such conditions can be verified (see for example Section 3.2 and Exercise 3.16).

It is possible to devise a computational method based on mathematical programming when S, C, and D are finite sets by making use of Prop. 3.1.2. Under N and $\alpha = 1$, the corresponding (linear) program is (compare with Section 1.3.4)

$$\text{maximize} \quad \sum_{i=1}^{n} \lambda_i$$

$$\text{subject to} \quad \lambda_i \leq g(i,u) + \sum_{j=1}^{n} p_{ij}(u)\lambda_j, \qquad i = 1,2,\ldots,n, \quad u \in U(i).$$

When $\alpha = 1$ and Assumption P holds, the corresponding program takes the form

$$\text{minimize} \quad \sum_{i=1}^{n} \lambda_i$$

$$\text{subject to} \quad \lambda_i \geq \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u)\lambda_j \right], \qquad i = 1,\ldots,n,$$

but unfortunately this program is not linear or even convex.

3.2 LINEAR SYSTEMS AND QUADRATIC COST

Consider the case of the linear system

$$x_{k+1} = Ax_k + Bu_k + w_k, \qquad k = 0,1,\ldots,$$

where $x_k \in \Re^n$, $u_k \in \Re^m$ for all k, and the matrices A, B are known. As in Sections 4.1 and 5.2 of Vol. I, we assume that the random disturbances w_k are independent with zero mean and finite second moments. The cost function is quadratic and has the form

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{w_k} \left\{ \sum_{k=0}^{N-1} \alpha^k \big(x_k' Q x_k + \mu_k(x_k)' R \mu_k(x_k) \big) \right\},$$

where $\alpha \in (0,1)$, Q is a positive semidefinite symmetric $n \times n$ matrix, and R is a positive definite symmetric $m \times m$ matrix. Clearly, Assumption P of Section 3.1 holds.

Our approach will be to use the DP algorithm to obtain the functions TJ_0, $T^2J_0, \ldots$, as well as the pointwise limit function $J_\infty = \lim_{k\to\infty} T^k J_0$. Subsequently, we show that J_∞ satisfies $J_\infty = TJ_\infty$ and hence, by Prop. 3.1.5(a), $J_\infty = J^*$. The optimal policy is then obtained from the optimal cost function J^* by minimizing in Bellman's equation (cf. Prop. 3.1.3).

As in Section 4.1 of Vol. I, we have

$$J_0(x) = 0, \qquad x \in \Re^n,$$

$$(TJ_0)(x) = \min_u [x'Qx + u'Ru] = x'Qx, \qquad x \in \Re^n,$$

$$\begin{aligned}(T^2J_0)(x) &= \min_u E\{x'Qx + u'Ru + \alpha(Ax+Bu+w)'Q(Ax+Bu+w)\} \\ &= x'K_1x + \alpha E\{w'Qw\}, \qquad x \in \Re^n,\end{aligned}$$

$$(T^{k+1}J_0)(x) = x'K_kx + \sum_{m=0}^{k-1} \alpha^{k-m} E\{w'K_m w\}, \qquad x \in \Re^n, \quad k = 1,2,\ldots,$$

where the matrices $K_0, K_1, K_2, \ldots$ are given recursively by

$$K_0 = Q,$$

$$K_{k+1} = A'\big(\alpha K_k - \alpha^2 K_k B(\alpha B'K_kB + R)^{-1}B'K_k\big)A + Q, \qquad k = 0,1,\ldots$$

By defining $\tilde{R} = R/\alpha$ and $\tilde{A} = \sqrt{\alpha}A$, the preceding equation may be written as

$$K_{k+1} = \tilde{A}'\big(K_k - K_kB(B'K_kB + \tilde{R})^{-1}B'K_k\big)\tilde{A} + Q,$$

and is of the form considered in Section 4.1 of Vol. I. By using the result shown there, we have that the generated matrix sequence $\{K_k\}$ converges to a positive definite symmetric matrix K,

$$K_k \to K,$$

provided the pairs $(\tilde{A}, B)$ and $(\tilde{A}, C)$, where $Q = C'C$, are controllable and observable, respectively. Since $\tilde{A} = \sqrt{\alpha}A$, controllability and observability of (A,B) or (A,C) are clearly equivalent to controllability and observability of $(\tilde{A},B)$ or $(\tilde{A},C)$, respectively. The matrix K is the unique solution of the equation

Because $K_k \to K$, it can also be seen that the limit

$$c = \lim_{k\to\infty} \sum_{m=0}^{k-1} \alpha^{k-m} E\{w' K_m w\}$$

is well defined, and in fact

$$c = \frac{\alpha}{1-\alpha} E\{w' K w\}. \tag{3.19}$$

Thus, in conclusion, if the pairs (A, B) and (A, C) are controllable and observable, respectively, the limit of the functions $T^k J_0$ is given by

$$J_\infty(x) = \lim_{k\to\infty} (T^k J_0)(x) = x' K x + c. \tag{3.20}$$

Using Eqs. (3.18)-(3.20), it can be verified by straightforward calculation that for all $x \in S$

$$J_\infty(x) = (TJ_\infty)(x) = \min_u \left[x' Q x + u' R u + \alpha E\{J_\infty(Ax + Bu + w)\}\right] \tag{3.21}$$

and hence, by Prop. 3.1.5(a), $J_\infty = J^*$. Another way to prove that $J_\infty = TJ_\infty$ is to show that the assumption of Prop. 3.1.7, is satisfied; i.e., the sets

$$U_k(x, \lambda) = \left\{u \mid E\{x' Q x + u' R u + \alpha (T^k J_0)(Ax + Bu + w)\} \le \lambda\right\}$$

are compact for all k and scalars λ. This can be verified using the fact that $T^k J_0$ is a positive semidefinite quadratic function and R is positive definite. The optimal stationary policy μ^*, obtained by minimization in Eq. (3.21), has the form

$$\mu^*(x) = -\alpha (\alpha B' K B + R)^{-1} B' K A x, \qquad x \in \Re^n.$$

This policy is attractive for practical implementation since it is linear and stationary. A number of generalized versions of the problem of this section, including the case of imperfect state information, are treated in the exercises. Interestingly, the problem can be solved by policy iteration (see Exercise 3.16), even though, as discussed in Section 3.1, policy iteration is not valid in general under Assumption P.

3.3 INVENTORY CONTROL

Let us consider a discounted, infinite horizon version of the inventory control problem of Section 4.2 in Vol. I. Inventory stock evolves according to the equation

We assume that the successive demands w_k are independent and bounded, and have identical probability distributions. We also assume for simplicity that there is no fixed cost. The case of a nonzero fixed cost can be treated similarly. The cost function is

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots,N-1}} \left\{ \sum_{k=0}^{N-1} \alpha^k \big(c\mu_k(x_k) + H\big(x_k + \mu(x_k) - w_k\big)\big) \right\},$$

where

$$H(y) = p \max(0, -y) + h \max(0, y).$$

The DP algorithm is given by

$$J_0(x) = 0,$$

$$(T^{k+1} J_0)(x) = \min_{0\le u} E\{cu + H(x + u - w) + \alpha (T^k J_0)(x + u - w)\}.$$

We first show that the optimal cost is finite for all initial states:

$$J^*(x_0) = \min_\pi J_\pi(x_0) < \infty, \qquad \text{for all } x_0 \in S.$$

Indeed, consider the policy $\tilde\pi = \{\tilde\mu, \tilde\mu, \ldots\}$, where $\tilde\mu$ is defined by

$$\tilde\mu(x) = \begin{cases} 0 & \text{if } x \ge 0, \\ -x & \text{if } x < 0. \end{cases}$$

Since w_k is nonnegative and bounded, it follows that the inventory stock x_k when the policy $\tilde\pi$ is used satisfies

$$-w_{k-1} \le x_k \le \max(0, x_0), \qquad k = 1, 2, \ldots,$$

and is bounded. Hence $\tilde\mu(x_k)$ is also bounded. It follows that the cost per stage incurred when $\tilde\pi$ is used is bounded, and in view of the presence of the discount factor we have

$$J_{\tilde\pi}(x_0) < \infty, \qquad x_0 \in S.$$

Since $J^* \le J_{\tilde\pi}$, the finiteness of the optimal cost follows.

Next we observe that, under the assumption $c < p$, the functions $T^k J_0$ are real-valued and convex. Indeed, we have

$$J_0 \le TJ_0 \le \cdots \le T^k J_0 \le \cdots \le J^*,$$

which implies that $T^k J_0$ is real-valued. Convexity follows by induction as