

Chapter 5

STOCHASTIC CONTROL AND DYNAMIC PROGRAMMING

5.1 Formulation of the Stochastic Control Problem

Consider the nonlinear stochastic system in state space form

$$\begin{aligned}x_{k+1} &= f_k(x_k, u_k, w_k) \\ x(0) &= x_0\end{aligned}\tag{5.1}$$

for $k = 0, 1, \dots, N$, where $N < \infty$ in this chapter unless otherwise specified. We assume that $\{w_k, k = 0, \dots, N\}$ is an **independent sequence** of random vectors, with mean zero and covariance Q . The initial condition x_0 is assumed to be independent of w_k for all k , with mean m_0 and covariance Σ_0 . $\{u_k, k = 0, \dots, N\}$ is the control input sequence. We assume that for each k , the past history of the state x_k is available so that admissible control laws are of the form

$$u_k = \phi_k(x^k)$$

where $x^k = \{x_j, j = 0, \dots, k\}$ is the history of state trajectory, also denoted by \mathcal{X}_k . Such control laws are called closed-loop controls. Note that open-loop controls, in which u_k is a function of k only, is a special case of closed-loop controls. It is readily seen (see Exercises) that in general for stochastic systems, closed-loop control laws out-perform open-loop controls. We may therefore confine attention to closed-loop control laws of the form $\Phi = \{\phi_0, \phi_1, \dots, \phi_N\}$. Once the control law Φ is chosen, the basic underlying random processes $\{x_0, w_k, k = 0, \dots, N\}$ completely determine the process x_k and hence u_k through the closed-loop system equations

$$\begin{aligned}x_{k+1}^\Phi &= f_k(x_k^\Phi, \phi_k(\mathcal{X}_k^\Phi), w_k) \\ x(0) &= x_0 \\ u_k^\Phi &= \phi_k(\mathcal{X}_k^\Phi)\end{aligned}$$

where x_k^Φ denotes the state process that results when the control law Φ is used.

To compare the effectiveness of control, we construct a sequence of **real-valued** functions $L_k(x_k, u_k, w_k)$ which is to be interpreted as the cost incurred at stage k in state x_k using control u_k and with noise disturbance w_k . L_k is thus a function of random variables and its values are random. We define the cost of control by

$$J(\Phi) = E \sum_{k=0}^N L_k(x_k^\Phi, u_k^\Phi, w_k)$$

Once the control law Φ is chosen, $J(\Phi)$ can be evaluated. Different control laws can therefore be compared based on their respective costs.

Example 5.1.1:

Consider the linear stochastic system described by

$$x_{k+1} = x_k + u_k + w_k$$

with $Ex_0 = 0$, $Ex_0^2 = 1$, $EW_k = 0$, $EW_k^2 = 1$. Suppose $N = 2$, and the per stage costs are given by $L_k(x_k, u_k, w_k) = x_k^2$. Let $\Phi = \{\phi_0, \phi_1\}$ with $\phi_0(x) = -2x$, $\phi_1(x) = -3x$. The closed-loop system under this control policy satisfies

$$\begin{aligned} x_1^\Phi &= x_0 - 2x_0 + w_0 = -x_0 + w_0 \\ x_2^\Phi &= x_1^\Phi - 3x_1^\Phi + w_1 = -2(-x_0 + w_0) + w_1 = 2x_0 - 2w_0 + w_1 \end{aligned}$$

The cost criterion under the policy Φ is given by

$$\begin{aligned} J(\Phi) &= E[x_0^2 + (x_1^\Phi)^2 + (x_2^\Phi)^2] \\ &= Ex_0^2 + E(-x_0 + w_0)^2 + E(2x_0 - 2w_0 + w_1)^2 \\ &= 1 + 2 + 9 = 12 \end{aligned}$$

On the other hand, if we choose the policy $\Psi = \{\psi, \psi\}$, where $\psi(x) = -x$, the closed-loop system is given by

$$x_{k+1}^\Psi = w_k$$

Hence the cost criterion under the policy Ψ is given by

$$\begin{aligned} J(\Psi) &= E[x_0^2 + (x_1^\Psi)^2 + (x_2^\Psi)^2] \\ &= Ex_0^2 + Ew_0^2 + Ew_1^2 = 3 \end{aligned}$$

We see that for this example, the policy Ψ is superior to the policy Φ .

We can now formulate the stochastic optimal control problem as follows:

Stochastic Optimal Control Problem:

Find the control law Φ so that for the stochastic system (5.1), the cost $J(\Phi)$ incurred is minimized. The control law Φ which gives the smallest $J(\Phi)$ is called the optimal control law.

Let the optimal cost be defined as

$$J^* = \inf_{\Phi} J(\Phi)$$

The optimal control Φ^* is thus the policy satisfying

$$J(\Phi^*) = J^*$$

Since there are an uncountably infinite number of control laws to choose from, the above stochastic control problem might appear to be intractable. This fortunately turns out not to be the case. The rest of this chapter treats the dynamic programming method for solving the stochastic optimal control problem. Our treatment follows closely that given in Kumar and Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*.

5.2 Dynamic Programming

The main tool in stochastic control is the method of dynamic programming. This method enables us to obtain feedback control laws naturally, and converts the problem of searching for optimal policies into a sequential optimization problem. The basic idea is very simple yet powerful. We begin by defining a special class of policies.

Definition: A policy Φ is called Markov if each function ϕ_k is a function of x_k only, so that $u_k = \phi_k(x_k)$.

Note that if a Markov policy Φ is used, the corresponding state process will be a Markov process.

Let Φ be a fixed **Markov** policy. Define recursively the functions

$$\begin{aligned} V_N^\Phi(x) &= EL_N(x, \phi_N(x), w_N) \\ V_k^\Phi(x) &= EL_k(x, \phi_k(x), w_k) + EV_{k+1}^\Phi[f_k(x, \phi_k(x), w_k)] \end{aligned} \quad (5.2)$$

Since x is fixed, the expectation is with respect to w . We use the following notation

- (a) x_k^Φ is the state process generated when the Markov policy Φ is used.
- (b) $u_k^\Phi = \phi_k(x_k^\Phi)$ is the control input at time k when the Markov policy Φ is used.

Lemma 5.2.1 now shows that the functions $V_k^\Phi(x)$ represent the cost-to-go at time k when Φ is used.

Lemma 5.2.1 *Let Φ be a Markov policy. Then*

$$\begin{aligned} V_k^\Phi(x_k^\Phi) &= E\left[\sum_{j=k}^N L_j(x_j^\Phi, u_j^\Phi, w_j) \mid x_k^\Phi\right] \\ &= E\left[\sum_{j=k}^N L_j(x_j^\Phi, u_j^\Phi, w_j) \mid \mathcal{X}_k^\Phi\right] \end{aligned} \quad (5.3)$$

where the expectation is with respect to w_k .

Proof. For notational simplicity, we write L_j for $L_j(x_j^\Phi, u_j^\Phi, w_j)$ whenever there is no possibility of confusion. The proof is by backward induction, a procedure used most often in connection with dynamic programming. First note that Lemma 5.2.1 is true for $k = N$. Now assume, by induction, that it is true for $j = k + 1, \dots, N$. We have

$$\begin{aligned} E\left[\sum_{j=k}^N L_j \mid x_k^\Phi\right] &= E\left[L_k + \sum_{j=k+1}^N L_j \mid x_k^\Phi\right] \\ &= E[L_k \mid x_k^\Phi] + E\left\{E\left[\sum_{j=k+1}^N L_j \mid x_{k+1}^\Phi, x_k^\Phi\right] \mid x_k^\Phi\right\} \\ &= E[L_k \mid x_k^\Phi] + E\left\{E\left[\sum_{j=k+1}^N L_j \mid x_{k+1}^\Phi\right] \mid x_k^\Phi\right\} \text{ by the Markov nature of } x_k^\Phi \\ &= E[L_k \mid x_k^\Phi] + E[V_{k+1}^\Phi(x_{k+1}^\Phi) \mid x_k^\Phi] \\ &= E[L_k \mid x_k^\Phi] + E[V_{k+1}^\Phi(f_k(x_k^\Phi, u_k^\Phi, w_k)) \mid x_k^\Phi] \end{aligned} \quad (5.4)$$

It is readily be verified that the following property of conditional expectation holds: If z and w are two independent random variables,

$$E[h(z, w)|z] = E_w h(z, w) \quad (5.5)$$

where E_w denotes expectation with respect to the random variable w . Using (5.5) in (5.4) and noting that x_k^Φ and w_k are independent, the R.H.S. is seen to be $V_k^\Phi(x_k^\Phi)$. Hence the Lemma is also true for $j = k$. By induction, the Lemma is proved.

Now define, for an arbitrary admissible policy Ψ , the cost-to-go at time k by

$$J_k^\Psi = E\left[\sum_{j=k}^N L_j(x_j^\Psi, u_j^\Psi, w_j) \mid \mathcal{X}_k^\Psi\right]$$

Then

$$J_0^\Psi = E\left[\sum_{j=0}^N L_j(x_j^\Psi, u_j^\Psi, w_j) \mid x_0\right]$$

and

$$EJ_0^\Psi = J(\Psi)$$

The next lemma defines a sequence of functions which form a lower bound to the cost-to-go.

Lemma 5.2.2 (*Comparison Principle*)

Let $V_k(x)$ be any function such that the following inequalities are satisfied for all x and u :

$$\begin{aligned} V_N(x) &\leq EL_N(x, u, w_N) \\ V_k(x) &\leq E_w L_k(x, u, w_k) + E_w V_{k+1}[f_k(x, u, w_k)] \end{aligned} \quad (5.6)$$

Let Ψ be any admissible policy. Then

$$V_k(x_k^\Psi) \leq J_k^\Psi \quad \text{for all } k \text{ w.p.1}$$

Proof. Again the proof is by backward induction. Lemma 5.2.2 is clearly true for $k = N$ by the definition of $V_N(x)$. Suppose it is true for $j = k + 1, \dots, N$. We need to show that it is true for $j = k$. By independence of w_k and x^k , (5.6) can be written as

$$\begin{aligned} V_k(x_k^\Psi) &\leq E\{L_k(x_k^\Psi, \psi_k(\mathcal{X}_k^\Psi), w_k) + V_{k+1}[f_k(x_k^\Psi, \psi_k(\mathcal{X}_k^\Psi), w_k)] \mid \mathcal{X}_k^\Psi\} \\ &\leq E\{L_k(x_k^\Psi, \psi_k(\mathcal{X}_k^\Psi), w_k) + J_{k+1}^\Psi \mid \mathcal{X}_k^\Psi\} \\ &= E\{L_k(x_k^\Psi, \psi_k(\mathcal{X}_k^\Psi), w_k) + E \sum_{j=k+1}^N [L_j(x_j^\Psi, \psi_j(\mathcal{X}_j^\Psi), w_j) \mid \mathcal{X}_{k+1}^\Psi] \mid \mathcal{X}_k^\Psi\} \\ &= E\left\{\sum_k^N L_j \mid \mathcal{X}_k^\Psi\right\} \\ &= J_k^\Psi \end{aligned}$$

Corollary 5.2.1 For any function $V_k(x)$ satisfying (5.6), $J^* \geq EV_0(x_0)$

The next result is the main optimality theorem of dynamic programming in the stochastic control context.

Theorem 5.1 Define the sequence of functions

$$\begin{aligned} V_N(x) &= \inf_u EL_N(x, u, w_N) \\ V_k(x) &= \inf_u \{E_w L_k(x, u, w_k) + E_w V_{k+1}[f_k(x, u, w_k)]\} \end{aligned} \quad (5.7)$$

(i) For any admissible policy Φ ,

$$V_k(x_k^\Phi) \leq J_k^\Phi$$

and

$$EV_0(x_0) \leq J(\Phi)$$

(ii) A Markov policy Φ^* is optimal if the infimum for (5.7) is achieved at Φ^* . Then

$$V_k(x_k^{\Phi^*}) = J_k^{\Phi^*} \quad w.p.1$$

and

$$EV_0(x_0) = J^* = J(\Phi^*)$$

(iii) A Markov policy Φ^* is optimal only if for each k , the infimum for (5.7) at each $x_k^{\Phi^*}$ is achieved by $\phi_k^*(x_k^{\Phi^*})$.

Proof. (i): V_k satisfies the Comparison Principle so that (i) obtains.

(ii): Let Φ be a Markovian policy which achieves the infimum. Then by Lemma 5.2.1 and (i)

$$V_k(x_k^\Phi) = J_k^\Phi \leq J_k^\Psi \quad \text{all } k \text{ and any admissible } \Psi$$

In particular, $J_0^\Phi = V_0(x_0) \Rightarrow \Phi$ is optimal by Corollary 5.2.1.

(iii): To prove (iii), we suppose Φ is Markovian and optimal. We prove by induction that Φ achieves the infimum. For $k = N$, (iii) is clearly true. For, if $\phi'_N \neq \phi_N$ achieves the infimum, we can define a Markov policy $\Phi' = (\phi_0, \dots, \phi_{N-1}, \phi'_N)$. Then since $EL_k = EL'_k$, $k \leq N-1$, we see that Φ not optimal.

Now suppose (iii) is true for $k+1$ and $J_{k+1}^\Phi = V_{k+1}(x_{k+1}^\Phi)$, but that it is not true for k . Then $\exists \phi'_k$ s.t.

$$\begin{aligned} &E_w L_k(x_k^\Phi, \phi_k(x_k^\Phi), w_k) + E_w V_{k+1}[f_k(x_k^\Phi, \phi_k(x_k^\Phi), w_k)] \\ &\geq E_w L_k(x_k^\Phi, \phi'_k(x_k^\Phi), w_k) + E_w V_{k+1}[f_k(x_k^\Phi, \phi'_k(x_k^\Phi), w_k)] \end{aligned} \quad (5.8)$$

Furthermore, strict inequality holds with positive probability so that expectation of L.H.S. of (5.8) $>$ expectation of R.H.S. Define

$$\Phi' = (\phi_0 \dots \phi_{k-1}, \phi'_k, \phi_{k+1} \dots \phi_N)$$

Then

$$EL_l = EL'_l \quad l \leq k-1$$

By the induction hypothesis, $\phi_{k+1} \dots \phi_N$ achieve the infimum. Since ϕ, ϕ' are both Markovian

$$\begin{aligned} EJ_{k+1}^\Phi(x_{k+1}^\Phi) &= EV_{k+1}(x_{k+1}^\Phi) \\ EJ_{k+1}^{\Phi'}(x_{k+1}^{\Phi'}) &= EV_{k+1}(x_{k+1}^{\Phi'}) \end{aligned}$$

We then have

$$\begin{aligned} J(\Phi) &= E \sum_0^{k-1} L_l + EL_k + EV_{k+1}(x_{k+1}^\Phi) \\ &> E \sum_0^{k-1} L'_l + EL'_k + EV_{k+1}(x_{k+1}^{\Phi'}) \\ &= J(\Phi') \end{aligned}$$

contradicting the optimality of Φ .

Based on Theorem 5.1, the solution to stochastic control problems can be obtained through the solution of the dynamic programming equation (5.7). It is to be solved recursively backwards, starting at $k = N$. For $k = N$ and each x , we have the corresponding optimal control $\phi_N^*(x)$. At every step $k < N$, we evaluate the R.H.S. of (5.7) for every possible value of x , and for each x , the optimal feedback law is given by

$$\phi_k^*(x) = \arg \min \{E_w L_k(x, u, w_k) + E_w V_{k+1}[f_k(x, u, w_k)]\}$$

Theorem 5.1 can be interpreted through the **Principle of Optimality** enunciated by Bellman:

Principle of Optimality

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Let us discuss how the Principle of Optimality determines the optimal control at time k . Suppose we are in state x at time k , and we take an arbitrary decision u . The Principle of Optimality states that if the resulting state is x_{k+1} , the remaining decisions must be optimal so that we must incur the optimal cost $V_{k+1}(x_{k+1})$. The optimal decision at time k must therefore be that u which minimizes the sum of the average cost at time k and the average value of $V_{k+1}(x_{k+1})$ over all possible transitions. This is precisely the content of the dynamic programming equation.

5.3 Inventory Control Example

The method of dynamic programming will now be illustrated with one of its standard application examples. A store needs to order inventory at the beginning of each day to fill the needs of customers. We assume that whatever stock ordered is delivered immediately. We assume, for simplicity, that the cost per unit stock order is 1 and the holding cost per unit item remaining unsold at the end of the day is also 1. Furthermore, there is a shortage cost per unit demand unfilled of 3. The stochastic control problem is: given the probability distribution for the random demand during the day, find the optimal planning policy for 2 days to minimize the expected cost, subject to a storage constraint of 2 items.

To analyze this problem let us introduce mathematical notation and make precise our assumptions.

Let x_k be the stock available at the beginning of the k th day, u_k the stock ordered at the beginning of the k th day, w_k the random demand during the k th day. The storage constraint of 2 units translate to the inequality $x_k + u_k \leq 2$. Since stock is nonnegative and integer-valued, we must also have $0 \leq x_k$, $0 \leq u_k$. The x_k process is then seen to satisfy the equation

$$x_{k+1} = \max(0, x_k + u_k - w_k) \tag{5.9}$$

Now let us assume that the probability distribution of w_k is the same for all k , given by

$$P(w_k = 0) = 0.1, \quad P(w_k = 1) = 0.7, \quad P(w_k = 2) = 0.2$$

Assume also that the initial stock $x_0 = 0$. The cost function is given by

$$L_k(x_k, u_k, w_k) = u_k + \max(0, x_k + u_k - w_k) + 3 \max(0, w_k - x_k - u_k) \quad (5.10)$$

$N = 1$ since we are planning for today and tomorrow. So the dynamic programming algorithm gives

$$V_k^* = \min_{0 \leq u_k \leq 2-x} E\{u_k + \max(0, x + u_k - w_k) + 3 \max(0, w_k - x - u_k) + V_{k+1}^*[\max(0, x + u_k - w_k)]\} \quad (5.11)$$

with $V_2^*(x) = 0$ for all x .

We now proceed backwards

$$V_1^*(x) = \min_{0 \leq u_1 \leq 2-x} E\{u_1 + \max(0, x + u_1 - w_1) + 3 \max(0, w_1 - x - u_1)\}$$

Now the values that x can take on are 0, 1, 2, and so is u_1 . Hence, using the probability distribution for w_1 , we get

$$V_1^*(0) = \min_{0 \leq u_1 \leq 2} \{u_1 + 0.1 \max(0, u_1) + 0.3 \max(0, -u_1) + 0.7 \max(0, u_1 - 1) + 2.1 \max(0, 1 - u_1) + 0.2 \max(0, u_1 - 2) + 0.6 \max(0, 2 - u_1)\} \quad (5.12)$$

For $u_1 = 0$, R.H.S. of (5.12) = $2.1 + 1.2 = 3.3$

For $u_1 = 1$, R.H.S. of (5.12) = $1 + 0.1 + 0.6 = 1.7$

For $u_1 = 2$, R.H.S. of (5.12) = $2 + 0.2 + 0.7 = 2.9$

Hence the minimizing u_1 for $x_1 = 0$ is 1 so that $\phi_1^*(0) = 1$, and $V_1^*(0) = 1.7$.

Similarly, for $x_1 = 1$, we obtain

$$\begin{aligned} V_1^*(1) &= \min_{0 \leq u_1 \leq 1} E\{u_1 + \max(0, 1 + u_1 - w_1) + 3 \max(0, w_1 - 1 - u_1)\} \\ &= 0.7 \text{ for the choice } u_1 = 0. \end{aligned}$$

Hence

$$\phi_1^*(1) = 0 \quad \text{and} \quad V_1^*(1) = 0.7$$

Finally, for $x_1 = 2$, we have

$$\begin{aligned} V_1^*(2) &= \min_{0 \leq u_1 \leq 0} E\{u_1 + \max(0, 2 + u_1 - w_1) + 3 \max(0, w_1 - 2 - u_1)\} \\ &= 0.9 \end{aligned}$$

In this case, no decision on u_1 is necessary since it is constrained to be 0. Hence $\phi_1^*(2) = 0$. Now to go back to $k = 0$, we apply (5.11) to get

$$\begin{aligned} V_0^*(x) &= \min_{0 \leq u_0 \leq 2-x} E\{u_0 + \max(0, x + u_0 - w_0) + 3 \max(0, w_0 - x - u_0) \\ &\quad + V_1^*[\max(0, x + u_0 - w_0)]\} \end{aligned} \quad (5.13)$$

Since the initial condition is taken to be $x = 0$, we need only compute $V_0^*(0)$. This gives

$$\begin{aligned} V_0^*(0) &= \min_{0 \leq u_0 \leq 2} E\{u_0 + \max(0, u_0 - w_0) + 3 \max(0, w_0 - u_0) \\ &\quad + V_1^*[\max(0, u_0 - w_0)]\} \\ &= \min_{0 \leq u_0 \leq 2} \{u_0 + 0.1 \max(0, u_0) + 0.3 \max(0, -u_0) \\ &\quad + 0.1 V_1^*[\max(0, u_0)] + 0.7 \max(0, u_0 - 1) + 2.1 \max(0, 1 - u_0) \\ &\quad + 0.7 V_1^*[\max(0, u_0 - 1)] + 0.2 \max(0, u_0 - 2) + 0.6 \max(0, 2 - u_0) \\ &\quad + 0.2 V_1^*[\max(0, u_0 - 2)]\} \end{aligned} \quad (5.14)$$

Using the values of $V_1^*(x)$ computed at the previous step, we find that for

$$\begin{aligned} u_0 = 0, & \quad \text{R.H.S. of (5.14) = 5.0} \\ u_0 = 1, & \quad \text{R.H.S. of (5.14) = 3.3} \\ u_0 = 2, & \quad \text{R.H.S. of (5.14) = 3.82} \end{aligned}$$

Hence, the minimizing u_0 is $u_0 = 1$ and

$$V_0^*(0) = 3.3 \text{ with } \phi_0^*(0) = 1 .$$

Had the initial state been 1, we would have

$$V_0^*(1) = 2.3 \text{ with } \phi_0^*(1) = 0 ;$$

and had x_0 been 2, we would have

$$V_0^*(2) = 1.82 \text{ with } \phi_0^*(2) = 0 .$$

The above calculations completely characterize the optimal policy Φ^* . Note that the optimal control policy is given as a look-up table, not as an analytical expression.

5.4 A Gambling Example

In general, dynamic programming equations cannot be solved analytically. One has to be content with generating a look-up table for the optimal policy through minimizing the right hand side of the dynamic programming equation. However, in some very special cases, it is possible to solve the dynamic programming equation. We give an illustrative example to show how this may be done.

A gambler enters a game whereby he may, at time k , stake any amount $u_k \geq 0$ that does not exceed his current fortune x_k (defined to be his initial capital plus his gain or minus his loss thus far). If he wins, he gets back his stake plus an additional amount equal to his stake so that his fortune will increase from x_k to $x_k + u_k$. If he loses, his fortune decreases to $x_k - u_k$. His probability of winning at each stake is p where $\frac{1}{2} < p < 1$, so that his probability of losing is $1 - p$. His objective is to maximize $E \log x_N$ where x_N is his fortune after N plays.

The stochastic control problem is characterized by the state equation

$$x_{k+1} = x_k + u_k w_k$$

where $P(w_k = 1) = p$, $P(w_k = -1) = 1 - p$. Since there are no per stage costs, we can write down the dynamic programming equation

$$V_k(x) = \max_u E[V_{k+1}(x + u_k w_k)]$$

with terminal condition

$$V_N(x) = \log x$$

Since it is not obvious what is the form of the function $V_k(x)$, we do one step of dynamic programming computation starting from the known terminal condition at time N .

$$\begin{aligned} V_{N-1}(x) &= \max_u E \log(x + u w_{N-1}) \\ &= \max_u \{p \log(x + u) + (1 - p) \log(x - u)\} \end{aligned}$$

Differentiating, we get

$$\frac{p}{x+u} - \frac{1-p}{x-u} = 0$$

Simplifying, we get

$$u_{N-1} = (2p-1)x_{N-1}$$

It is straightforward to verify that this is the maximizing value of u_{N-1} . Upon substituting into the right hand side of $V_{N-1}(x)$, we obtain

$$\begin{aligned} V_{N-1}(x) &= p \log 2px + (1-p) \log 2(1-p)x \\ &= p \log 2p + p \log x + (1-p) \log 2(1-p) + (1-p) \log x \\ &= \log x + p \log 2p + (1-p) \log 2(1-p) \end{aligned}$$

We see that the function $\log x + \alpha_k$ fits the form of $V_{N-1}(x)$ as well as $V_N(x)$. This suggests that we try the following guess for the optimal value function

$$V_k(x) = \log x + \alpha_k$$

Putting into the dynamic programming equation, we find

$$\begin{aligned} \log x + \alpha_k &= \max_u E\{\log(x + uw_{k-1}) + \alpha_{k+1}\} \\ &= \max_u \{p \log(x+u) + (1-p) \log(x-u) + \alpha_{k+1}\} \end{aligned}$$

Noting that the maximization is the same as that for time $N-1$, we have again the optimizing u_k given by

$$u_k = (2p-1)x_k$$

Substituting, we obtain

$$\begin{aligned} \log x + \alpha_k &= p \log(2px) + (1-p) \log 2(1-p)x + \alpha_{k+1} \\ &= p \log 2p + p \log x + (1-p) \log 2(1-p) + (1-p) \log x + \alpha_{k+1} \\ &= \log x + \alpha_{k+1} + p \log 2p + (1-p) \log 2(1-p) \end{aligned}$$

We see that the trial solution indeed solves the dynamic programming equation if we set the sequence α_k to be given by the equation

$$\begin{aligned} \alpha_k &= \alpha_{k+1} + p \log 2p + (1-p) \log 2(1-p) \\ &= \alpha_{k+1} + \log 2 + p \log p + (1-p) \log(1-p) \end{aligned}$$

with terminal condition $\alpha_N = 0$. This completely determines the optimal policy for this gambling problem.

5.5 The Curse of Dimensionality

In principle, dynamic programming enables us to solve general discrete time stochastic control problems. However, unless we are lucky enough to be able to solve the dynamic programming equation analytically, we would need to search for the optimal value of u for each x . If we examine the computational effort involved, we quickly see that in practice, there are difficulties in applying the dynamic programming algorithm. To get a feeling about the numbers involved, suppose the state space is finite and contains N_x elements. Similarly, let the total number of elements in the control set be N_u and let the planning horizon be N stages. Then at every stage, we need to evaluate V^* at N_x values of the state. If we look at the right hand

side of (5.7), we see that for each x , we have to evaluate the value of $E_w L_k(x, u, w_k) + E_w V_{k+1}^*[f_k(x, u, w_k)]$ for N_u values of u . So the number of function evaluations per stage is of the order of $N_x N_u$. For N stages then, the total number of function evaluations would be $N_x N_u N$. Often the state is a continuous variable. Discretization of the state space is used to produce a finite approximating set. For good accuracy, N_x is often large. Thus, with any planning horizon N greater than 10, as is common, we shall be burdened with a significant computational problem. Although this rough analysis does not take into account much more efficient computational methods associated with dynamic programming, it does give an indication to the rapid increase in the computational difficulties. This computational difficulty associated with the method of dynamic programming is often called the curse of dimensionality, and has effectively prevented it from being applied to many practical problems.

For the theoretically inclined, there are interesting technical problems associated with the dynamic programming equation. Two such mathematical problems are the following:

- (1) We have to show the minimization in (5.7) can be carried out at every stage. Typical assumptions which enable us to do that are the following:
 - (a) Assume that the control set is finite. Then the minimization of the right hand side at every stage is easily determined by simply searching over the control set.
 - (b) Assume that the control set is compact (for Euclidean space, this is the same as closed and bounded) and show, from other assumptions connected with the problem, that the R.H.S. is continuous in u so that the minimum exists.
- (2) The quantities appearing in (5.7) makes probabilistic sense, i.e., they are all valid random variables. Such measure-theoretic questions can be avoided if the underlying stochastic process is a Markov chain with countable state space.

Of course, all these problems disappear if we can actually solve the dynamic programming equation explicitly. Such cases are rare and are often of limited scope and interest, as in the gambling example. There is however one important class of stochastic control problems which have broad applicability and for which we have a simple solution. This is the linear regulator problem which we shall treat next.

5.6 The Stochastic Linear Regulator Problem

The system process is given by the equation

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Gw_k \\ x_{k_0} &= x_0 \end{aligned} \tag{5.15}$$

where w_k is an independent sequence of random vectors with $Ew_k = 0$ and $Ew_k w_k^T = Q$, and $Ex_0 = m$, $\text{cov}(x_0) = \Sigma_0$, x_0 independent of w_k . The cost criterion is given by

$$J = E \left\{ x_N^T M x_N + \sum_{k=k_0}^{N-1} \|Dx_k + Fu_k\|^2 \right\} \tag{5.16}$$

where $M \geq 0$ and $F^T F > 0$. The control set is the entire \mathbb{R}^m space, hence the control values are unconstrained. The form of the cost is motivated by the desire to regulate the state of the system x_k to zero at time N without making any large excursions in its trajectory, and at the same time, not spending too much control effort.

The dynamic programming equation for this problem can be written down immediately.

$$V_k(x) = \min_u \{ \|Dx + Fu\|^2 + E\{V_{k+1}[Ax + Bu + Gw_k]\} \} \quad (5.17)$$

with terminal condition $V_N(x) = x^T M x$.

The great simplicity of this problem lies in the fact that we can actually solve the dynamic programming equation (5.17) analytically. To this end, we first note 2 preliminary results.

(i) For any random vector x with mean m and covariance Σ , and any $S \geq 0$, we have

$$\begin{aligned} E(x^T S x) &= E\{(x - m)^T S (x - m)\} + E m^T S x + E x^T S m - m^T S m \\ &= \text{tr } S \Sigma + m^T S m \end{aligned} \quad (5.18)$$

(ii) For $R_1 > 0$, any R_2 , and R_3 symmetric,

$$\begin{aligned} g(u) &= u^T R_1 u + u^T R_2 x + x^T R_2^T u + x^T R_3 x \\ &= (u + R_1^{-1} R_2 x)^T R_1 (u + R_1^{-1} R_2 x) + x^T (R_3 - R_2^T R_1^{-1} R_2) x \end{aligned}$$

Hence for each x , the value of u which minimizes $g(u)$ is given by

$$u = -R_1^{-1} R_2 x$$

with the resulting value of $g(u)$ given by

$$g(u) = x^T (R_3 - R_2^T R_1^{-1} R_2) x$$

Now noting the form of the cost and the terminal condition, we try a solution for $V_k(x)$ in the form

$$V_k(x) = x^T S_k x + q_k \quad (5.19)$$

Applying (5.17), we see immediately that $S_N = M$ and $q_N = 0$

$$\begin{aligned} E\{V_{k+1}[Ax + Bu_k + Gw_k]\} &= (Ax + Bu_k)^T S_{k+1} [Ax + Bu_k] \\ &\quad + \text{tr } S_{k+1} G Q G^T + q_{k+1} \end{aligned} \quad (5.20)$$

so that (5.17) becomes

$$\begin{aligned} x^T S_k x + q_k &= \min_{u_k} \{ \|Dx + Fu_k\|^2 + (Ax + Bu_k)^T S_{k+1} (Ax + Bu_k) \\ &\quad + \text{tr } S_{k+1} G Q G^T + q_{k+1} \} \end{aligned} \quad (5.21)$$

The optimal feedback law is given by, according to Theorem 5.1, the minimizing value of the R.H.S. of (5.21). We find, using preliminary result (ii), that

$$u_k = \phi_k^*(x_k) = -(F^T F + B^T S_{k+1} B)^{-1} (B^T S_{k+1} A + F^T D) x_k \quad (5.22)$$

This is then the optimal policy. On substituting (5.22) into (5.21) and grouping the quadratic terms together, we see that S_k must satisfy

$$\begin{aligned} S_k &= A^T S_{k+1} A + D^T D - (A^T S_{k+1} B + D^T F) (F^T F + B^T S_{k+1} B)^{-1} (B^T S_{k+1} A + F^T D) \\ S_N &= M \end{aligned} \quad (5.23)$$

q_k must satisfy

$$\begin{aligned} q_k &= q_{k+1} + \text{tr } S_{k+1} G Q Q^T \\ q_N &= 0 \end{aligned} \quad (5.24)$$

(5.24) can be solved explicitly for q_k , to give

$$q_k = \sum_{j=k}^{N-1} \text{tr } S_{j+1} G Q Q^T \quad (5.25)$$

The optimal cost is given by

$$\begin{aligned} EV_{k_0}(x_0) &= E x_0^T S_{k_0} x_0 + q_{k_0} \\ &= m_0^T S_{k_0} m_0 + \text{tr } S_{k_0} \Sigma_0 + \sum_{j=k_0}^{N-1} \text{tr } S_{j+1} G Q Q^T \end{aligned} \quad (5.26)$$

There are several things to notice about the solution of the linear regulator problem.

- (5.23) may be recognized as a discrete time Riccati difference equation. It is identical in form to the Riccati difference equation which features so prominently in the Kalman filter equations. We can put them into one-to-one correspondence by the following table:

Regulator	Filter
$k \leq N$	$k \geq k_0$
A	A^T
B	C^T
$D^T D$	$G Q Q^T$
$F^T F$	$H R H^T$
$D^T F$	$G T H^T$
$D^T [I - F(F^T F)^{-1} F^T] D$	$G [Q - T H^T (H R H^T)^{-1} H T^T] G^T$

This is an illustration of the intimate relation within linear-quadratic control and linear filtering, and is also referred to as the duality between filtering and control.

- The optimal feedback law is the same one as the linear regulator problem for deterministic systems, i.e., for the case where $w_k = 0$ and x_0 fixed. On the one hand, this says that the linear feedback law is optimal even in the face of additive disturbances, a clearly desirable engineering property. On the other hand, it also says that the naive control scheme of setting all disturbances to its mean values and solving the resulting deterministic control problem is in fact optimal. So for this problem, the stochastic aspects do not really play an important role. This is due to the very special nature of the linear regulator problem.
- The manner in which the stochastic aspects enter is basically through the modification of the optimal cost. If the problem were deterministic, then the optimal cost in (5.26) would contain only the term $m_0^T S_{k_0} m_0$. The random nature of the initial state x_0 contributes the additional term $\text{tr } S_{k_0} \Sigma_0$, and

the random nature of the disturbance w_k contributes the term $\sum_{j=k_0}^{N-1} \text{tr } S_{j+1} G Q Q^T$.

5.7 Asymptotic Properties of the Linear Regulator

The asymptotic properties of the linear regulator again centre on those of the Riccati difference equation. The asymptotic behaviour of the Riccati equation has already been studied in the filtering context. We can summarize the results as follows:

Let

$$\begin{aligned}\hat{A} &= A - B(F^T F)^{-1}F^T D \\ \hat{D} &= (I - F(F^T F)^{-1}F^T)^{\frac{1}{2}} D\end{aligned}$$

(or any \hat{D} satisfying $\hat{D}^T \hat{D} = D^T(I - F(F^T F)^{-1}F^T)D$). If (A, B) is stabilizable and (\hat{D}, \hat{A}) detectable, then there exists a unique solution, in the class of positive semidefinite matrices, to the algebraic Riccati equation

$$S = A^T S A + D^T D - (A^T S B + D^T F)(F^T F + B^T S B)^{-1}(B^T S A + F^T D). \quad (5.27)$$

Moreover, the closed-loop system matrix $A - B(F^T F + B^T S B)^{-1}(B^T S A + F^T D)$ is stable. For any $M \geq 0$, S_k , the solution of (5.23) $\xrightarrow[k \rightarrow -\infty]{} S$.

If we consider the stationary version of the feedback law (5.22), i.e.

$$\phi(x_k) = -(F^T F + B^T S B)^{-1}(B^T S A + F^T D)x_k \quad (5.28)$$

Where S is the unique positive semidefinite solution of (5.27), the resulting closed-loop system is given by

$$x_{k+1} = (A - B(F^T F + B^T S B)^{-1}(B^T S A + F^T D))x_k + Gw_k \quad (5.29)$$

If we denote the covariance of x_k by Σ_k , then by stability of (5.29), $\Sigma_k \xrightarrow[k \rightarrow \infty]{} \Sigma$. This means that the second moments of x_k are finite in the infinite interval and second moment stability obtains. In particular, if x_0 is Gaussian and w_k is a white Gaussian sequence, the closed-loop system (5.29) will also generate a Gaussian process. It converges to a stationary Gaussian process as $k \rightarrow \infty$. Note that because of the noise input, x_k will not go to zero as $k \rightarrow \infty$.

5.8 Stochastic Control of Linear Systems with Partial Observations

In Section 5.5, we considered the linear regulator problem when the entire state x_k is observed. In this section, we assume that x_k is not directly observable. Our system is given by

$$x_{k+1} = Ax_k + Bu_k + Gw_k \quad (5.30)$$

$$y_k = Cx_k + Hv_k \quad (5.31)$$

$$x_{k_0} = x_0$$

where we assume w_k and v_k to be independent *Gaussian* random sequences with $Ew_k = Ev_k = 0$, $Ew_k w_j^T = Q\delta_{kj}$, $Ev_k v_j^T = R\delta_{kj}$ with $R > 0$, $HRH^T > 0$, and $Ew_k v_j^T = T\delta_{kj}$. Furthermore, x_0 is assumed to be a Gaussian random vector with mean m_0 and covariance P_0 , independent of w_k and v_k .

The control problem is to minimize

$$J = E \left\{ x_N^T M x_N + \sum_{k=k_0}^{N-1} \|Dx_k + Fu_k\|^2 \right\}$$

The crucial distinction between the present problem and that in Section 5.5 is that the control law cannot be made a function of x_k . It can only be allowed to depend on the past observations. It is thus very important to specify the admissible laws.

Let $\mathcal{Y}_k = \sigma\{y(s), k_0 \leq s \leq k\}$, the sigma field generated by $\{y(s), k_0 \leq s \leq k\}$. This represents the information contained in the observations so that we have a causal control policy with a one-step delay in the information feedback. We take the admissible control laws to be $\Phi = \{\phi_{k_0}, \dots, \phi_{N-1}\}$ where ϕ_k is a (Borel) function of \mathcal{Y}_{k-1} . The interpretation is that u_k depends on y_s , $k_0 \leq s \leq k-1$. Once \mathcal{Y}_{k-1} is known, the value u_k is also determined.

The key to the solution of the problem is that under the linear-Gaussian assumptions, the estimation and control can be separated from each other. Introduce the system

$$\begin{aligned}\bar{x}_{k+1} &= A\bar{x}_k + Gw_k \\ \bar{x}_{k_0} &= x_0\end{aligned}\tag{5.32}$$

$$\bar{y}_k = C\bar{x}_k + Hv_k\tag{5.33}$$

Lemma 5.8.1 *For any admissible policy Φ , $\bar{\mathcal{Y}}_k = \mathcal{Y}_k$, $k = k_0, \dots, N-1$. In other words, $\bar{\mathcal{Y}}_k$ contains the same amount of information as \mathcal{Y}_k .*

Proof: Let $\tilde{x}_k = x_k - \bar{x}_k$. Then

$$\begin{aligned}\tilde{x}_{k+1} &= A\tilde{x}_k + Bu_k \\ \tilde{x}_{k_0} &= 0\end{aligned}\tag{5.34}$$

We claim that \tilde{x}_k depends only on \mathcal{Y}_{k-2} . This is clearly true for \tilde{x}_{k_0+1} because $\tilde{x}_{k_0+1} = A\tilde{x}_{k_0} + Bu_{k_0} = Bu_{k_0}$, which is assumed to be dependent on \mathcal{Y}_{k_0-1} (i.e. no observed information). Suppose, by induction, that \tilde{x}_k depends only on \mathcal{Y}_{k-2} . Then since $\tilde{x}_{k+1} = A\tilde{x}_k + Bu_k$, the R.H.S. depends only on \mathcal{Y}_{k-1} , and the claim follows.

Now $y_{k_0} = \bar{y}_{k_0}$. Assume by induction that $\mathcal{Y}_j = \bar{\mathcal{Y}}_j$, $j \leq k-1$. Then

$$\begin{aligned}y_k &= Cx_k + Hv_k = C\bar{x}_k + Hv_k + C\tilde{x}_k \\ &= \bar{y}_k + C\tilde{x}_k\end{aligned}\tag{5.35}$$

Using the previous claim, the R.H.S. of (5.35) depends only on $\bar{\mathcal{Y}}_k$. Hence $\mathcal{Y}_k \subset \bar{\mathcal{Y}}_k$. But from (5.35), we also see that $\bar{\mathcal{Y}}_k \subset \mathcal{Y}_k$ so that $\bar{\mathcal{Y}}_k = \mathcal{Y}_k$.

Remark: If we allow u_k to depend on \mathcal{Y}_k , Lemma 5.8.1 still holds, with virtually no change in the proof. In this case, there is no delay in the information available for control.

We may now split the system into two parts

$$x_k = \bar{x}_k + \tilde{x}_k\tag{5.36}$$

using (5.32) and (5.34). Furthermore, the estimate

$$\begin{aligned}\hat{x}_{k+1|k} &= E\{x_{k+1}|\mathcal{Y}_k\} = E\{\bar{x}_{k+1} + \tilde{x}_{k+1}|\mathcal{Y}_k\} \\ &= E\{\bar{x}_{k+1}|\mathcal{Y}_k\} + \tilde{x}_{k+1}\end{aligned}\tag{5.37}$$

But $E\{\bar{x}_{k+1}|\mathcal{Y}_k\} = E\{\bar{x}_{k+1}|\bar{\mathcal{Y}}_k\}$ corresponds to the optimal conditional mean estimate in the Kalman filtering problem. So (5.37) becomes

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k-1} + K_k(\bar{y}_k - C\hat{x}_{k|k-1}) + A\tilde{x}_k + Bu_k\tag{5.38}$$

where K_k is the Kalman filter gain. But using (5.37), we have

$$\begin{aligned}\hat{x}_{k+1|k} &= A\hat{x}_{k|k-1} + Bu_k + K_k(y_k - C\tilde{x}_k - C\hat{x}_{k|k-1}) \\ &= A\hat{x}_{k|k-1} + Bu_k + K_k(y_k - C\hat{x}_{k|k-1})\end{aligned}\quad (5.39)$$

If we compare (5.39) to the standard Kalman filter, we see that the additional term Bu_k in the state equation appears in the same additive manner in the estimation equation (5.39). This is a consequence of our assumption about admissible laws.

The next step in the development is the simplification of the cost. Consider the term

$$\begin{aligned}E\{x_k^T D^T D x_k | \mathcal{Y}_{k-1}\} &= E\{(x_k - \hat{x}_{k|k-1})^T D^T D (x_k - \hat{x}_{k|k-1}) | \mathcal{Y}_{k-1}\} + \hat{x}_{k|k-1}^T D^T D \hat{x}_{k|k-1} \\ &= \text{tr } D^T D P_{k|k-1} + \hat{x}_{k|k-1}^T D^T D \hat{x}_{k|k-1}\end{aligned}$$

Hence

$$E(x_k^T D^T D x_k) = E(E\{x_k^T D^T D x_k | \mathcal{Y}_{k-1}\}) = \text{tr } D^T D P_{k|k-1} + E(\hat{x}_{k|k-1}^T D^T D \hat{x}_{k|k-1}) \quad (5.40)$$

Similarly, noting that u_k is known given \mathcal{Y}_{k-1} ,

$$E(x_k^T D^T F u_k) = E(E\{x_k^T D^T F u_k | \mathcal{Y}_{k-1}\}) = E(\hat{x}_{k|k-1}^T D^T F u_k) \quad (5.41)$$

Note that the 1st term on the R.H.S. of (5.40) is independent of u_k . Using (5.40) and (5.41), we obtain the following expression for the cost

$$\begin{aligned}J &= E\{\hat{x}_{N|N-1}^T M \hat{x}_{N|N-1} + \sum_{k=k_0}^{N-1} [\|D\hat{x}_{k|k-1} + F u_k\|^2]\} \\ &\quad + \text{terms independent of control}\end{aligned}\quad (5.42)$$

Now (5.39) may be written as

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k-1} + Bu_k + K_k \nu_k \quad (5.43)$$

where $\nu_k = y_k - C\hat{x}_{k|k-1} = \bar{y}_k - C\hat{x}_{k|k-1}$ is the innovations process. According to the results in Section 3.2, ν_k is also a Gaussian white noise process, and in the form of $\bar{y}_k - C\hat{x}_{k|k-1}$, can be seen to be independent of u_k . We have now reduced the problem from one with partial observations to one with complete observations in that $\hat{x}_{k+1|k}$ is the state of the system, known at time $k+1$ from (5.39), with cost criterion

$$\hat{J} = E\{\hat{x}_{N|N-1}^T M \hat{x}_{N|N-1} + \sum_{k=k_0}^{N-1} \|D\hat{x}_{k|k-1} + F u_k\|^2\}$$

since the terms in (5.42) which are independent of the control will not affect the choice of the control law. The results of Section 5.6 are now directly applicable and we obtain

$$\begin{aligned}u_k &= -(F^T F + B^T S_{k+1} B)^{-1} (B^T S_{k+1} A + F^T D) \hat{x}_{k|k-1} \\ &= \phi_k(\mathcal{Y}_{k-1})\end{aligned}\quad (5.44)$$

since $\hat{x}_{k|k-1}$ depends only on \mathcal{Y}_{k-1} .

The result obtained in (5.44) characterizing the optimal control in the partially observed linear regulator problem is usually known as the **Separation Theorem**. The name comes from the fact that the feedback law

$$\phi_k(x) = -(F^T F + B^T S_{k+1} B)^{-1} (B^T S_{k+1} A + F^T D) x$$

is precisely the optimal control law for the deterministic linear regulator problem with quadratic cost. The Separation Theorem says then that if we have additive Gaussian white noise in the system, the optimal feedback law should be applied to the best estimate of the state of the system. This separates the task of designing the optimal stochastic control into 2 parts: that of designing the optimal deterministic feedback law, and that of designing the optimal estimator. This constitutes one of the most important results in system theory.

5.9 Stability of the closed-loop System

Equation (5.30) together with the control law (5.44) give rise to the closed-loop system

$$x_{k+1} = Ax_k - B(F^T F + B^T S_{k+1} B)^{-1} (B^T S_{k+1} A + F^T D) \hat{x}_{k|k-1} + Gw_k \quad (5.45)$$

Let $e_{k|k-1} = x_k - \hat{x}_{k|k-1}$. Then $e_{k|k-1}$ satisfies the equation

$$\begin{aligned} e_{k+1|k} &= Ae_{k|k-1} - (AP_{k|k-1}C^T + GTH^T)(CP_{k|k-1}C^T + HRH^T)^{-1}Ce_{k|k-1} \\ &\quad - (AP_{k|k-1}C^T + GTH^T)(CP_{k|k-1}C^T + HRH^T)^{-1}Hv_k + Gw_k \end{aligned} \quad (5.46)$$

Let

$$\begin{aligned} L_k &= (F^T F + B^T S_{k+1} B)^{-1} (B^T S_{k+1} A + F^T D) \\ K_k &= (AP_{k|k-1}C^T + GTH^T)(CP_{k|k-1}C^T + HRH^T)^{-1} \end{aligned}$$

(5.45) and (5.46) may be combined to give the following system

$$\begin{bmatrix} x_{k+1} \\ e_{k+1|k} \end{bmatrix} = \begin{bmatrix} A - BL_k & BL_k \\ 0 & A - K_k C \end{bmatrix} \begin{bmatrix} x_k \\ e_{k|k-1} \end{bmatrix} + \begin{bmatrix} Gw_k \\ Gw_k - K_k H v_k \end{bmatrix} \quad (5.47)$$

If the algebraic Riccati equations associated with S_k and $P_{k|k-1}$ have unique stabilizing solutions, then we may consider the stationary control law given by

$$u_k = -(F^T F + B^T S B)^{-1} (B^T S A + F^T D) \hat{x}_{k|k-1}$$

where $\hat{x}_{k|k-1}$ is generated by the stationary filter given by (3.11). Let

$$\begin{aligned} L &= (F^T F + B^T S B)^{-1} (B^T S A + F^T D) \\ K &= (APC^T + GTH^T)(CPC^T + HRH^T)^{-1} \end{aligned}$$

The closed-loop system then takes the form

$$\begin{bmatrix} x_{k+1} \\ e_{k+1|k}^s \end{bmatrix} = \begin{bmatrix} A - BL & BL \\ 0 & A - KC \end{bmatrix} \begin{bmatrix} x_k \\ e_{k|k-1}^s \end{bmatrix} + \begin{bmatrix} Gw_k \\ Gw_k - KH v_k \end{bmatrix} \quad (5.48)$$

This is again a system of the form

$$\xi_{k+1} = \hat{A}\xi_k + \eta_k$$

and the stability of ξ_k , in the sense of boundedness of its covariance, is governed by the stability of \hat{A} . But the block triangular nature of \hat{A} shows that the stability of \hat{A} is determined by the stability of $A - BL$ and that of $A - KC$. Using our previous results concerning asymptotic behaviour of the Kalman filter and the linear regulator, we can immediately state the following result.

Theorem 5.2 *If the pairs (A, B) and (\check{A}, \check{G}) are stabilizable, and the pairs (C, A) and (\hat{D}, \hat{A}) are detectable, then the stationary control law*

$$u_k = -(F^T F + B^T S B)^{-1} (B^T S A + F^T D) \hat{x}_{k|k-1} \quad (5.49)$$

where S is given by the unique positive semidefinite solution of the algebraic Riccati equation (5.27) and $\hat{x}_{k|k-1}$ is given by the stationary filter (3.11), gives rise to a stable closed-loop system.

In connection with stationary control laws we may consider infinite time control problems. Note that we cannot in general formulate the cost criterion associated with an infinite time control problem as

$$E \sum_{k=0}^{\infty} \|Dx_k + Fu_k\|^2,$$

since the noise terms will make the above cost infinite no matter what the control law is. This may be seen from the optimal cost for the finite time problem which contains the term $\sum_{k=k_0}^{N-1} \text{tr } S_{k+1} G Q G^T$. If as $N \rightarrow \infty$, $S_k \rightarrow S$, the infinite sum will become unbounded. One way of formulating a meaningful infinite time problem is to take the average cost per unit time criterion

$$J_r = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} E \|Dx_k + Fu_k\|^2 \quad (5.50)$$

It can be shown that if the conditions of Theorem 5.2 hold, the control law (5.49) is in fact optimal for the cost (5.50). See, for example, Kushner, *Introduction to Stochastic Control* and the exercises.

5.10 Exercises

1. This problem illustrates the fact that in stochastic control, closed-loop control generally out-performs open-loop control. Consider the linear stochastic system

$$x_{k+1} = x_k + u_k + w_k$$

with cost criterion

$$J(\Phi) = E \sum_{k=0}^N x_k^2$$

where $N \geq 1$, $Ex_0 = 0$, $Ex_0^2 = 1$, $EW_k = 0$, $EW_k^2 = 1$, and w_k is an independent sequence, also independent of x_0 .

- (a) Let u_k be any deterministic sequence (corresponding to open-loop control). Determine the cost criterion in terms of N and u_k .
- (b) Let u_k be given by the closed-loop control law $u_k = -x_k$. Determine the cost criterion associated with this policy and show that it is strictly less than the cost criterion determined in (a), regardless of the open-loop control sequence used in (a).
2. Let x_k denote the price of a given stock on the k th day and suppose that

$$x_{k+1} = x_k + w_k$$

$$x_0 = 10$$

where w_k forms an independent, identically distributed sequence with probability distribution $P(w_k = 0) = 0.1$, $P(w_k = 1) = 0.4$, $P(w_k = -1) = 0.5$. You have the option to buy one share of the stock at a fixed price, say 9. You have 3 days in which to exercise the option ($k = 0, 1, 2$). If you exercise the option, and the stock price is x , your profit is $\max(x - 9, 0)$. Formulate this as a stochastic control problem and find the optimal policy to maximize your expected profit.

3. Consider the following gambling problem. On each play of a certain game, a gambler has a probability p of winning, with $0 < p < 1/2$. He begins with an initial amount of M dollars. On each play he may bet any amount up to his entire fortune. If he bets u dollars and wins, he gains u dollars, while if he loses he loses the u dollars he has bet. Let x_k be his fortune at time k . Then we readily see that x_k satisfies the following equation

$$x_{k+1} = x_k + u_k w_k$$

where u_k satisfies $0 \leq u_k \leq x_k$, and w_k is an independent sequence with $P(w_k = 1) = p$ and $P(w_k = -1) = 1 - p$. The total number of plays is fixed to be N and the gambler would like to construct an optimal policy to maximize Ex_N^2 where x_N is the fortune he has at time N .

- (a) Formulate the problem as a stochastic control problem and obtain the dynamic programming equation which characterizes the optimal reward.
- (b) Characterize the optimal policy in terms of the parameter p .
(Hint: Guess the form of the optimal reward $V_k(x)$. Be careful about the maximization.)

4. An employer has N applicants for an advertised position. Each applicant has an independent nonnegative score which obeys a common probability distribution known to the employer. The actual score is found by interviewing the applicant. An applicant is either appointed or rejected after the interview. Once rejected, the applicant is lost. The position must be filled by the employer. The problem is to find the optimal appointment policy which maximizes the expected score of the candidate appointed.

We formulate the problem as a dynamic programming problem. Let the score associated with the k th candidate be w_k with density function $p(w)$. w_k is an independent identically distributed sequence by assumption. Let x_k be the state of the process, which is either the score of the k th candidate, or if an appointment has already been made, the distinguished state F . The two control values at time k are 1 for appoint or 2 for reject. We can therefore write the state equation as

$$x_{k+1} = f(x_k, u_k, w_{k+1})$$

where

$$\begin{aligned} f(x_k, u_k, w_{k+1}) &= F && \text{if } x_k = F \text{ or } u_k = 1 \\ &= w_{k+1} && \text{if } u_k = 2 \end{aligned}$$

- (i) Determine the per stage “reward” $L(x_k, u_k)$ as a function of x_k, u_k .
 - (ii) Obtain the dynamic programming equation for this optimization problem. Be sure to include the starting (terminal) condition for the optimal cost.
 - (iii) Show that for $k \leq N - 1$, the optimal control is to appoint the k th candidate if $x_k > \alpha_k$ and reject if $x_k < \alpha_k$ while both appointment and rejection are optimal if $x_k = \alpha_k$. Characterize α_k . (Hint: Set $\alpha_k = EV_{k+1}(w_{k+1})$ and obtain a difference equation for α_k .)
 - (iv) Suppose $p(w) = 1$, $0 \leq w \leq 1$, and $N = 4$. Determine the α_k sequence and hence the optimal policy.
5. This problem treats the optimal control of a simple partially observed scalar linear system with quadratic criterion.

- (a) Let

$$x_{k+1} = x_k + u_k$$

$$y_k = x_k + v_k$$

$$J = E \left\{ qx_N^2 + \sum_{k=0}^{N-1} f^2 u_k^2 \right\}, \quad q > 0$$

$Ex_0 = m_0$, $cov(x_0) = p_0 > 0$, $Ev_k = 0$, $Ev_k v_j = r\delta_{kj}$, $r > 0$. Admissible controls u_k are functions of y_τ , $0 \leq \tau \leq k - 1$. Find the optimal control law explicitly in terms of the given parameters. You’ll have to solve two Riccati difference equations.

- (b) Let $X_k = E\hat{x}_{k|k-1}^2$. Determine the difference equation satisfied by X_k . Express the control effort Eu_k^2 in terms of X_k .
 - (c) Let $N = 4$, $q = 10$, $f = 1$, $m_0 = 1$, $p_0 = 1$, $r = 1$. Find sequentially Eu_k^2 for $k = 0, 1, 2, 3$.
6. (i) Infinite time problems can also be solved directly using dynamic programming. Consider the system

$$x_{k+1} = Ax_k + Bu_k + Gw_k \tag{ex6.1}$$

where the state x_k is perfectly observed. The cost criterion to be minimized is

$$J_\rho = E \sum_{k=0}^{\infty} \rho^k \|Dx_k + Fu_k\|^2$$

Show that if there exists a function $V(x)$ such that $\rho^k EV(x_k) \xrightarrow{k \rightarrow \infty} 0$ and that $V(x)$ satisfies the dynamic programming equation

$$V(x) = \min_u \{ \|Dx + Fu\|^2 + \rho EV(Ax + Bu + Gw_k) \} \quad (ex6.2)$$

then the optimal control law is given by

$$u_k = \arg \min \{ \|Dx_k + Fu_k\|^2 + \rho EV(Ax_k + Bu_k + Gw_k) \} \quad (ex6.3)$$

Determine the function $V(x)$ and the control law u_k explicitly, making appropriate assumptions about properties of solutions to an algebraic Riccati equation.

(ii) Similar results can be obtained for the average cost per unit time problem

$$J_{av} = \lim_{N \rightarrow \infty} \frac{1}{N} E \sum_{k=0}^{N-1} \|Dx_k + Fu_k\|^2$$

Show that if there exist a real number λ and a function $W(x)$ such that $\frac{1}{N} EW(x_N) \xrightarrow{N \rightarrow \infty} 0$ and that

$$\lambda + W(x) = \min_u [\|Dx + Fu\|^2 + EW(Ax + Bu + Gw_k)] \quad (ex6.4)$$

then the control which minimizes the R.H.S. of (ex6.4) is the optimal control. Determine the function $W(x)$ explicitly and the optimal control law. Finally, show that λ is the optimal cost.

(Hint: Consider the identity $E \left\{ \sum_{j=1}^N W(x_j) - E[W(x_j) | x_{j-1}, u_{j-1}] \right\} = 0$.

Show that $E[W(x_j) | x_{j-1}, u_{j-1}] \geq \lambda + W(x_{j-1}) - \|Dx_{j-1} + Fu_{j-1}\|^2$ and substitute this into the identity.)

7. A singular quadratic control problem is one in which there is no penalty on the control. This problem shows how sometimes a singular control problem can be transformed into a nonsingular one. Suppose the scalar transfer function

$$H(z) = \frac{b_1 z^{n-1} + \dots + b_n}{z^n + a_1 z^{n-1} + \dots + a_n} \quad b_1 \neq 0$$

is realized by a state space representation of the form

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

$$y_k = cx_k$$

so that $c(zI - A)^{-1}b = H(z)$. Without loss of generality, we may take (c, A) to be in observable canonical form

$$A = \begin{bmatrix} 0 & & & -a_n \\ 1 & & & \\ \vdots & \ddots & & \vdots \\ & & 1 & -a_1 \end{bmatrix} \quad c = [0 \dots 0 \quad 1]$$

Then

$$b = \begin{bmatrix} b_n \\ \vdots \\ b_1 \end{bmatrix}$$

Suppose the control problem is to minimize

$$J = \sum_{k=0}^{\infty} y_k^2$$

This is then a singular control problem.

- (a) Show that J is minimized if and only if $J_1 = \sum_{k=0}^{\infty} y_{k+1}^2$ is minimized.
 (b) Express J_1 in the form of

$$J_1 = \sum_{k=0}^{\infty} \|Dx_k + Fu_k\|^2 \quad \text{with} \quad F^T F > 0$$

What are D and F ?

- (c) Put $v_k = u_k + (F^T F)^{-1} F^T D x_k$ and express the system equations in terms of v_k , i.e., find \hat{A} and \hat{b} so that

$$x_{k+1} = \hat{A}x_k + \hat{b}v_k$$

Express J_1 also in terms of v_k , i.e. find \hat{D} and \hat{F} so that $J_1 = \sum_{k=0}^{\infty} \|\hat{D}x_k + \hat{F}v_k\|^2$, $\hat{F}^T \hat{F} > 0$.

- (d) Give conditions in terms of the original system matrices under which (\hat{A}, \hat{b}) is stabilizable and (\hat{D}, \hat{A}) is detectable. Determine the optimal control (which is also stabilizing) in this case.
 (e) Determine the necessary and sufficient conditions for detectability of (\hat{D}, \hat{A}) using the original transfer function $H(z)$.
8. We discussed the solution to the LQG problem when there is a one-step delay in the information available for control. Assume that $E(w_k v_k^T) = 0$, i.e. $T = 0$, but the admissible control laws are of the form $u_k = \phi_k(y^k)$. Imitate the derivation of Section 5.8 to show that the optimal control law in this case for the finite time problem is

$$u_k = -(F^T F + B^T S_{k+1} B)^{-1} (B^T S_{k+1} A + F^T D) \hat{x}_{k|k}$$

9. (i) For the control algebraic Riccati equation (CARE), assume $F = 0$. (CARE) now reads, assuming that the indicated inverse exists,

$$S = A^T S A + D^T D - A^T S B (B^T S B)^{-1} B^T S A$$

Assume that B is a $n \times 1$ column vector and D is a $1 \times n$ row vector, $DB \neq 0$. Verify that $D^T D$ is a solution of CARE. Give appropriate structural conditions under which this solution is the unique positive semidefinite solution which stabilizes the closed-loop system.

(Hint: Refer to problem 7.)

- (ii) Consider the system

$$y_k + a y_{k-1} = u_{k-1} + b u_{k-2} + e_k + c e_{k-1}$$

A state space representation of this system is

$$x_{k+1} = \begin{bmatrix} 0 & 0 \\ 1 & -a \end{bmatrix} x_k + \begin{bmatrix} b \\ 1 \end{bmatrix} u_k + \begin{bmatrix} c \\ 1 \end{bmatrix} e_{k+1}$$

$$y_k = [0 \quad 1]x_k$$

Find the time-invariant control law using LQG theory which minimizes $\lim_{N \rightarrow \infty} \frac{1}{N} E \sum_{j=0}^{N-1} y_j^2$ where u_k is allowed to be a function of y^k . Check all the structural assumptions needed (stabilizability, detectability, etc.) and solve as many equations explicitly as you can.

(Hint: Use the result obtained in (i).)