OPTIMIZATION AND CONTROL

Richard Weber, Lent 2002

Contents

ii

Schedules

I follow the outline in the Lecture Schedules. The first 6 lectures are devoted to **dynamic programming** in discrete-time and cover both finite and infinite-horizon problems; discounted-cost, positive, negative and average-cost programming; the time-homogeneous Markov case; stopping problems; value iteration and policy improvement.

The next 5 lectures are devoted to the **LQG model** (linear systems, quadratic costs) and draw out the important ideas of controllability and observability; the Ricatti equation; imperfect observation, certainly equivalence; the Kalman filter and linear least squares estimation.

The final 5 lectures are1 devoted to **continuous-time models** and include treatment of Pontryagin's maximum principle and the Hamiltonian; Markov decision processes on a countable state space; and controlled diffusion processes.

Each of the 16 lectures is designed to be a somewhat self-contained unit, e.g., there will be one lecture on 'Negative Programming', one on 'Controllability', etc. Examples and applications are important in this course, so there are one or more worked examples in each lecture.

Examples sheets

There are three examples sheets, corresponding to the thirds of the course. There are two or three questions for each lecture, some theoretical and some of a problem nature. Each question is marked to indicate the lecture with which it is associated.

Lecture Notes and Handouts

There are printed lecture notes for the course and other occasional handouts. There are sheets summarising notation and what you are expected to know for the exams.

The notes include a list of keywords and I will be drawing your attention to these as we go along. If you have a good grasp of the meaning of each of these keywords, then you will be well on your way to understanding the important concepts of the course.

WWW pages

Notes for the course, and other information are on the web at
`http://www.statslab.cam.ac.uk/~rrw1/oc/index.html`.

Books

The following books are recommended.

D. P. Bertsekas, *Dynamic Programming*, Prentice Hall, 1987.
D. P. Bertsekas, *Dynamic Programming and Optimal Control*,
Volumes I and II, Prentice Hall, 1995.
L. M. Hocking, *Optimal Control: An introduction to the theory and applications*, Oxford 1991.
S. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, 1983.
P. Whittle, *Optimization Over Time*. Volumes I and II, Wiley, 1982-83.

Ross's book is probably the easiest to read. However, it only covers Part I of the course. Whittle's book is good for Part II and Hocking's book is good for Part III. The recent book by Bertsekas is useful for all parts. Many other books address the topics of the course and a collection can be found in Sections 3B and 3D of the DPMMS library. Notation differs from book to book. My notation will be closest to that of Whittle's books and consistent throughout. For example, I will always denote a minimal cost function by $F(\cdot)$ (whereas, in the recommended books you will find F, V, ϕ, J and many others symbols used for this quantity.)

1 Dynamic Programming: The Optimality Equation

We introduce the idea of dynamic programming and the principle of optimality. We give notation for state-structured models, and introduce ideas of feedback, open-loop, and closed-loop controls, a Markov decision process, and the idea that it can be useful to model things in terms of time to go.

1.1 Control as optimization over time

Optimization is a key tool in modelling. Sometimes it is important to solve a problem optimally. Other times either a near-optimal solution is good enough, or the real problem does not have a single criterion by which a solution can be judged. However, even then optimization is useful as a way to test thinking. If the 'optimal' solution is ridiculous it may suggest ways in which both modelling and thinking can be refined.

Control theory is concerned with dynamic systems and their **optimization over time**. It accounts for the fact that a dynamic system may evolve stochastically and that key variables may be unknown or imperfectly observed (as we see, for instance, in the UK economy).

This contrasts with optimization models in the IB course (such as those for LP and network flow models); these static and nothing was random or hidden. It is these three new features: dynamic and stochastic evolution, and imperfect state observation, that give rise to new types of optimization problem and which require new ways of thinking.

We could spend an entire lecture discussing the importance of control theory and tracing its development through the windmill, steam governor, and so on. Such 'classic control theory' is largely concerned with the question of stability, and there is much of this theory which we ignore, e.g., Nyquist criterion and dynamic lags.

1.2 The principle of optimality

A key idea is that optimization over time can often be regarded as 'optimization in stages'. We trade off our desire to obtain the lowest possible cost at the present stage against the implication this would have for costs at future stages. The best action minimizes the sum of the cost incurred at the current stage and the least total cost that can be incurred from all subsequent stages, consequent on this decision. This is known as the Principle of Optimality.

Definition 1.1 (Principle of Optimality) *From any point on an optimal trajectory, the remaining trajectory is optimal for the corresponding problem initiated at that point.*

1.3 Example: the shortest path problem

Consider the 'stagecoach problem' in which a traveler wishes to minimize the length of a journey from town A to town J by first traveling to one of B, C or D and then onwards to one of E, F or G then onwards to one of H or I and the finally to J. Thus there are 4 'stages'. The arcs are marked with distances between towns.
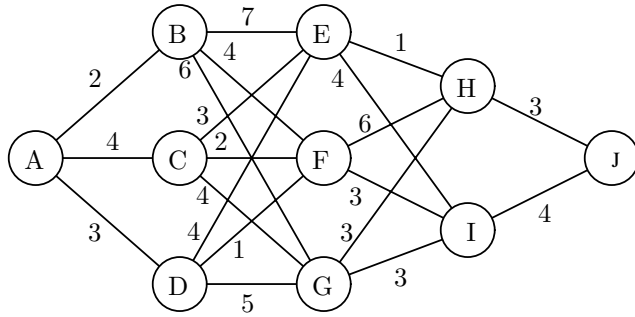
Road system for stagecoach problem

Solution. Let $F(X)$ be the minimal distance required to reach J from X. Then clearly, $F(J) = 0$, $F(H) = 3$ and $F(I) = 4$.

$$F(F) = \min[\, 6 + F(H), 3 + F(I)\,] = 7\,,$$

and so on. Recursively, we obtain $F(A) = 11$ and simultaneously an optimal route, i.e., A→D→F→I→J (although it is not unique). ∎

The study of dynamic programming dates from Richard Bellman, who wrote the first book on the subject (1957) and gave it its name. A very large number of problems can be treated this way.

1.4   The optimality equation

The optimality equation in the general case.   In **discrete-time** t takes integer values, say $t = 0, 1, \ldots$. Suppose u_t is a **control variable** whose value is to be chosen at time t. Let $U_{t-1} = (u_0, \ldots, u_{t-1})$ denote the partial sequence of controls (or decisions) taken over the first t stages. Suppose the cost up to the **time horizon** h is given by

$$\mathbf{C} = G(U_{h-1}) = G(u_0, u_1, \ldots, u_{h-1})\,.$$

Then the **principle of optimality** is expressed in the following theorem.

Theorem 1.2 (The principle of optimality) *Define the functions*

$$G(U_{t-1}, t) = \inf_{u_t, u_{t+1}, \ldots, u_{h-1}} G(U_{h-1})\,.$$

Then these obey the recursion

$$G(U_{t-1}, t) = \inf_{u_t} G(U_t, t+1) \quad t < h\,,$$

with terminal evaluation $G(U_{h-1}, h) = G(U_{h-1})$.

The proof is immediate from the definition of $G(U_{t-1}, t)$, i.e.,

$$G(U_{t-1}, t) = \inf_{u_t} \inf_{u_{t+1}, \ldots, u_{h-1}} G(u_0, \ldots, u_{t-1},\ u_t\ , u_{t+1}, \ldots, u_{h-1})\,.$$

2

The state structured case. The control variable u_t is chosen on the basis of knowing $U_{t-1} = (u_0, \ldots, u_{t-1})$, (which determines everything else). But a more economical representation of the past history is often sufficient. For example, we may not need to know the entire path that has been followed up to time t, but only the place to which it has taken us. The idea of a **state variable** $x \in \mathbb{R}^d$ is that its value at t, denoted x_t, is calculable from known quantities and obeys a **plant equation** (or law of motion)

$$x_{t+1} = a(x_t, u_t, t) \, .$$

Suppose we wish to minimize a cost function of the form

$$\mathbf{C} = \sum_{t=0}^{h-1} c(x_t, u_t, t) + \mathbf{C}_h(x_h) \, , \tag{1.1}$$

by choice of controls $\{u_0, \ldots, u_{h-1}\}$. Define the cost from time t onwards as,

$$\mathbf{C}_t = \sum_{\tau=t}^{h-1} c(x_\tau, u_\tau, \tau) + \mathbf{C}_h(x_h) \, , \tag{1.2}$$

and the minimal cost from time t onwards as an optimization over $\{u_t, \ldots, u_{h-1}\}$ conditional on $x_t = x$,

$$F(x, t) = \inf_{u_t, \ldots, u_{h-1}} \mathbf{C}_t \, .$$

Here $F(x, t)$ is the minimal future cost from time t onward, given that the state is x at time t. Then by an inductive proof, one can show as in Theorem 1.2 that

$$F(x, t) = \inf_u [c(x, u, t) + F(a(x, u, t), t+1)] \, , \quad t < h \, , \tag{1.3}$$

with terminal condition $F(x, h) = \mathbf{C}_h(x)$. Here x is a generic value of x_t. The minimizing u in (1.3) is the optimal control $u(x, t)$ and values of $x_0, \ldots, x_{t-1}$ are irrelevant. The **optimality equation** (1.3) is also called the **dynamic programming equation** (DP) or **Bellman equation**.

The DP equation defines an optimal control problem in what is called **feedback** or **closed loop** form, with $u_t = u(x_t, t)$. This is in contrast to the **open loop** formulation in which $\{u_0, \ldots, u_{h-1}\}$ are to be determined all at once at time 0. A **policy** (or strategy) is a rule for choosing the value of the control variable under all possible circumstances as a function of the perceived circumstances. To summarise:

(i) The optimal u_t is a function only of x_t and t, i.e, $u_t = u(x_t, t)$.

(ii) The DP equation expresses the optimal u_t is closed loop form. It is optimal whatever the past control policy may have been.

(iii) The DP equation is a backward recursion in time (from which we get the optimum at $h - 1$, then $h - 2$ and so on.) The later policy is decided first.

> '*Life must be lived forward and understood backwards.*' (Kierkegaard)

3

1.5 Markov decision processes

Consider now stochastic evolution. Let $X_t = (x_0, \ldots, x_t)$ and $U_t = (u_0, \ldots, u_t)$ denote the x and u histories at time t. As above, state structure is characterised by the fact that the evolution of the process is described by a state variable x, having value x_t at time t, with the following properties.

(a) *Markov dynamics:* (i.e., the stochastic version of the plant equation.)

$$P(x_{t+1} \mid X_t, U_t) = P(x_{t+1} \mid x_t, u_t).$$

(b) *Decomposable cost*, (i.e., cost given by (1.1)).

These assumptions define state structure. For the moment we also require.

(c) *Perfect state observation:* The current value of the state is observable. That is, x_t is known at the time at which u_t must be chosen. So, letting W_t denote the observed history at time t, we assume $W_t = (X_t, U_{t-1})$. Note that $\mathbf{C}$ is determined by W_h, so we might write $\mathbf{C} = \mathbf{C}(W_h)$.

These assumptions define what is known as a discrete-time **Markov decision process** (MDP). Many of our examples will be of this type. As above, the cost from time t onwards is given by (1.2). Denote the minimal expected cost from time t onwards by

$$F(W_t) = \inf_{\pi} E_{\pi}[\mathbf{C}_t \mid W_t],$$

where π denotes a policy, i.e., a rule for choosing the controls $u_0, \ldots, u_{h-1}$. We can assert the following theorem.

Theorem 1.3 *$F(W_t)$ is a function of x_t and t alone, say $F(x_t, t)$. It obeys the optimality equation*

$$F(x_t, t) = \inf_{u_t} \{c(x_t, u_t, t) + E[F(x_{t+1}, t+1) \mid x_t, u_t]\}, \quad t < h, \tag{1.4}$$

with terminal condition

$$F(x_h, h) = \mathbf{C}_h(x_h).$$

Moreover, the minimizing value of u_t in (1.4) (which is also only a function x_t and t) is optimal.

Proof. The value of $F(W_h)$ is $\mathbf{C}_h(x_h)$, so the asserted reduction of F is valid at time h. Assume it is valid at time $t + 1$. The DP equation is then

$$F(W_t) = \inf_{u_t}\{c(x_t, u_t, t) + E[F(x_{t+1}, t+1) \mid X_t, U_t]\}. \tag{1.5}$$

But, by assumption (a), the right-hand side of (1.5) reduces to the right-hand member of (1.4). All the assertions then follow. ∎

2   Some Examples of Dynamic Programming

We illustrate the method of dynamic programming and some useful 'tricks'.

2.1   Example: managing spending and savings

An investor receives annual income from a building society of x_t pounds in year t. He consumes u_t and adds $x_t - u_t$ to his capital, $0 \leq u_t \leq x_t$. The capital is invested at interest rate $\theta \times 100\%$, and so his income in year $t + 1$ increases to

$$x_{t+1} = a(x_t, u_t) = x_t + \theta(x_t - u_t).$$

He desires to maximize his total consumption over h years, $\mathbf{C} = \sum_{t=0}^{h-1} u_t$.

Solution. In the notation we have been using, $c(x_t, u_t, t) = u_t$, $\mathbf{C}_h(x_h) = 0$. This is a **time-homogeneous** model, in which neither costs nor dynamics depend on t. It is easiest to work in terms of '**time to go**', $s = h - t$. Let $F_s(x)$ denote the maximal reward obtainable, starting in state x and when there is time s to go. The dynamic programming equation is

$$F_s(x) = \max_{0 \leq u \leq x} [u + F_{s-1}(x + \theta(x - u))],$$

where $F_0(x) = 0$, (since no more can be obtained once time h is reached.) Here, x and u are generic values for x_s and u_s.

We can substitute backwards and soon guess the form of the solution. First,

$$F_1(x) = \max_{0 \leq u \leq x} [u + F_0(u + \theta(x - u))] = \max_{0 \leq u \leq x} [u + 0] = x.$$

Next,

$$F_2(x) = \max_{0 \leq u \leq x} [u + F_1(x + \theta(x - u))] = \max_{0 \leq u \leq x} [u + x + \theta(x - u)].$$

Since $u + x + \theta(x - u)$ linear in u, its maximum occurs at $u = 0$ or $u = x$, and so

$$F_2(x) = \max[(1 + \theta)x, 2x] = \max[1 + \theta, 2]x = \rho_2 x.$$

This motivates the guess $F_{s-1}(x) = \rho_{s-1} x$. Trying this, we find

$$F_s(x) = \max_{0 \leq u \leq x} [u + \rho_{s-1}(x + \theta(x - u))] = \max[(1 + \theta)\rho_{s-1}, 1 + \rho_{s-1}]x = \rho_s x.$$

Thus our guess is verified and $F_s(x) = \rho_s x$, where ρ_s obeys the recursion implicit in the above, and i.e., $\rho_s = \rho_{s-1} + \max[\theta\rho_{s-1}, 1]$. This gives

$$\rho_s = \begin{cases} s & s \leq s^* \\ (1 + \theta)^{s - s^*} s^* & s \geq s^* \end{cases},$$

where s^* is the least integer such that $s^* \geq 1/\theta$, i.e., $s^* = \lceil 1/\theta \rceil$. The optimal strategy is to invest the whole of the income in years $0, \ldots, h - s^* - 1$, (to build up capital) and then consume the whole of the income in years $h - s^*, \ldots, h - 1$. ∎

5

There are several things worth remembering from this example. (i) It is often useful to frame things in terms of time to go, s. (ii) Although the form of the dynamic programming equation can sometimes look messy, try working backwards from $F_0(x)$ (which is known). Often a pattern will emerge from which we can piece together a solution. (iii) When the dynamics are linear, the optimal control lies at an extreme point of the set of feasible controls. This form of policy, which either consumes nothing or consumes everything, is known as **bang-bang control**.

2.2 Example: exercising a stock option

The owner of a call option has the option to buy a share at fixed 'striking price' p. The option must be exercised by day h. If he exercises the option on day t and then immediately sells the share at the current price x_t, he can make a profit of $x_t - p$. Suppose the price sequence obeys the equation $x_{t+1} = x_t + \epsilon_t$, where the ϵ_t are i.i.d. random variables for which $E|\epsilon| < \infty$. The aim is to exercise the option optimally.

Let $F_s(x)$ be the value function (maximal expected profit) when the share price is x and there are s days to go. Show that (i) $F_s(x)$ is non-decreasing in s, (ii) $F_s(x) - x$ is non-increasing in x and (iii) $F_s(x)$ is continuous in x. Deduce that the optimal policy can be characterised as follows.

There exists a non-decreasing sequence $\{a_s\}$ such that an optimal policy is to exercise the option the first time that $x \geq a_s$, where x is the current price and s is the number of days to go before expiry of the option.

Solution. The state variable at time t is, strictly speaking, x_t plus a variable which indicates whether the option has been exercised or not. However, it is only the latter case which is of interest, so x is the effective state variable. Since dynamic programming makes its calculations backwards, from the termination point, it is often advantageous to write things in terms of the time to go, $s = h - t$. So if we let $F_s(x)$ be the value function (maximal expected profit) with s days to go then

$$F_0(x) = \max\{x - p, 0\},$$

and so the dynamic programming equation is

$$F_s(x) = \max\{x - p, E[F_{s-1}(x + \epsilon)]\}, \quad s = 1, 2, \ldots$$

Note that the expectation operator comes *outside*, not inside, $F_{s-1}(\cdot)$.

One can use induction to show (i), (ii) and (iii). For example, (i) is obvious, since increasing s means we have more time over which to exercise the option. However, for a formal proof

$$F_1(x) = \max\{x - p, E[F_0(x + \epsilon)]\} \geq \max\{x - p, 0\} = F_0(x).$$

Now suppose, inductively, that $F_{s-1} \geq F_{s-2}$. Then

$$F_s(x) = \max\{x - p, E[F_{s-1}(x + \epsilon)]\} \geq \{x - p, E[F_{s-2}(x + \epsilon)]\} = F_{s-1}(x),$$

6

whence F_s is non-decreasing in s. Similarly, an inductive proof of (ii) follows from

$$\underbrace{F_s(x) - x}_{} = \max\{-p, E[\underbrace{F_{s-1}(x + \epsilon) - (x + \epsilon)}_{}] + E(\epsilon)\},$$

since the left hand underbraced term inherits the non-increasing character of the right hand underbraced term. Thus the optimal policy can be characterized as stated. For from (ii), (iii) and the fact that $F_s(x) \geq x - p$ it follows that there exists an a_s such that $F_s(x)$ is greater that $x - p$ if $x < a_s$ and equals $x - p$ if $x \geq a_s$. It follows from (i) that a_s is non-decreasing in s. The constant a_s is the smallest x for which $F_s(x) = x - p$. ∎

2.3 Example: accepting the best offer

We are to interview h candidates for a job. At the end of each interview we must either hire or reject the candidate we have just seen, and may not change this decision later. Candidates are seen in random order and can be ranked against those seen previously. The aim is to maximize the probability of choosing the candidate of greatest rank.

Solution. Let W_t be the history of observations up to time t, i.e., after we have interviewed the tth candidate. All that matters are the value of t and whether the tth candidate is better than all her predecessors: let $x_t = 1$ if this is true and $x_t = 0$ if it is not. In the case $x_t = 1$, the probability she is the best of all h candidates is

$$P(\text{best of } h \mid \text{best of first } t) = \frac{P(\text{best of } h)}{P(\text{best of first } t)} = \frac{1/h}{1/t} = \frac{t}{h}.$$

Now the fact that the tth candidate is the best of the t candidates seen so far places no restriction on the relative ranks of the first $t - 1$ candidates; thus $x_t = 1$ and W_{t-1} are statistically independent and we have

$$P(x_t = 1 \mid W_{t-1}) = \frac{P(W_{t-1} \mid x_t = 1)}{P(W_{t-1})} P(x_t = 1) = P(x_t = 1) = \frac{1}{t}.$$

Let $F(0, t - 1)$ be the probability that under an optimal policy we select the best candidate, given that we have seen $t - 1$ candidates so far and the last one was *not* the best of those. Dynamic programming gives

$$F(0, t - 1) = \frac{t-1}{t}F(0, t) + \frac{1}{t}\max\left(\frac{t}{h}, F(0, t)\right) = \max\left(\frac{t-1}{t}F(0, t) + \frac{1}{h}, F(0, t)\right)$$

These imply $F(0, t - 1) \geq F(0, t)$ for all $t \leq h$. Therefore, since t/h and $F(0, t)$ are respectively increasing and non-increasing in t, it must be that for small t we have $F(0, t) > t/h$ and for large t we have $F(0, t) < t/h$. Let t_0 be the smallest t such that $F(0, t) \leq t/h$. Then

$$F(0, t - 1) = \begin{cases} F(0, t_0), & t < t_0, \\ \dfrac{t-1}{t}F(0, t) + \dfrac{1}{h}, & t \geq t_0. \end{cases}$$

7

Solving the second of these backwards from the point $t = h$, $F(0, h) = 0$, we obtain

$$\frac{F(0, t-1)}{t-1} = \frac{1}{h(t-1)} + \frac{F(0, t)}{t} = \cdots = \frac{1}{h(t-1)} + \frac{1}{ht} + \cdots + \frac{1}{h(h-1)},$$

whence

$$F(0, t-1) = \frac{t-1}{h} \sum_{\tau=t-1}^{h-1} \frac{1}{\tau}, \quad t \geq t_0.$$

Since we require $F(0, t_0) \leq t_0/h$, it must be that t_0 is the smallest integer satisfying

$$\sum_{\tau=t_0}^{h-1} \frac{1}{\tau} \leq 1.$$

For large h the sum on the left above is about $\log(h/t_0)$, so $\log(h/t_0) \approx 1$ and we find $t_0 \approx h/e$. The optimal policy is to interview $\approx h/e$ candidates, but without selecting any of these, and then select the first one thereafter that is the best of all those seen so far. The probability of success is $F(0, t_0) \sim t_0/h \sim 1/e = 0.3679$. It is surprising that the probability of success is so large for arbitrarily large h. ∎

There are a couple lessons in this example. (i) It is often useful to try to establish the fact that terms over which a maximum is being taken are monotone in opposite directions, as we did with t/h and $F(0, t)$. (ii) A typical approach is to first determine the form of the solution, then find the optimal cost (reward) function by backward recursion from the terminal point, where its value is known.

8

3 Dynamic Programming over the Infinite Horizon

We define the cases of discounted, negative and positive dynamic programming and establish the validity of the optimality equation for an infinite horizon problem.

3.1 Discounted costs

For a discount factor, $\beta \in (0, 1]$, the **discounted-cost criterion** is defined as

$$\mathbf{C} = \sum_{t=0}^{h-1} \beta^t c(x_t, u_t, t) + \beta^h \mathbf{C}_h(x_h) \,, \tag{3.1}$$

This simplifies things mathematically, particularly when we want to consider an infinite horizon. If costs are uniformly bounded, say $|c(x, u)| < B$, and discounting is strict ($\beta < 1$) then the infinite horizon cost is bounded by $B/(1 - \beta)$. In economic language, if there is an interest rate of $r\%$ per unit time, then a unit amount of money at time t is worth $\rho = 1 + r/100$ at time $t + 1$. Equivalently, a unit amount at time $t + 1$ has present value $\beta = 1/\rho$. The function, $F(x, t)$, which expresses the minimal present value at time t of expected-cost from time t up to h is

$$F(x, t) = \inf_{u_t, \ldots, u_{h-1}} E\left[\left. \sum_{\tau=t}^{h-1} \beta^{\tau-t} c(x_\tau, u_\tau, \tau) + \beta^{h-t} \mathbf{C}_h(x_h) \right| x_t = x \right]. \tag{3.2}$$

The DP equation is now

$$F(x, t) = \inf_u \left[c(x, u, t) + \beta E F(a(x, u, t), t + 1) \right], \quad t < h, \tag{3.3}$$

where $F(x, h) = \mathbf{C}_h(x)$.

3.2 Example: job scheduling

A collection of n jobs is to be processed in arbitrary order by a single machine. Job i has processing time p_i and when it completes a reward r_i is obtained. Find the order of processing that maximizes the sum of the discounted rewards.

Solution. Here we take 'time k' as the point at which the $n - k$ th job has just been completed and the state at time k as the collection of uncompleted jobs, say S_k. The dynamic programming equation is

$$F_k(S_k) = \max_{i \in S_k} [r_i \beta^{p_i} + \beta^{p_i} F_{k-1}(S_k - \{i\})].$$

Obviously $F_0(\emptyset) = 0$. Applying the method of dynamic programming we first find $F_1(\{i\}) = r_i \beta^{p_i}$. Then, working backwards, we find

$$F_2(\{i, j\}) \quad = \quad \max[r_i \beta^{p_i} + \beta^{p_i + p_j} r_j , \ r_j \beta^{p_j} + \beta^{p_j + p_i} r_i].$$

There will be 2^n equations to evaluate, but with perseverance we can determine $F_n(\{1, 2, \ldots, n\})$. However, there is a simpler way.

An interchange argument. Suppose that jobs are scheduled in the order $i_1, \ldots, i_k, i, j, i_{k+2}, \ldots, i_n$. Compare the reward of this schedule to one in which the order of jobs i and j are reversed: $i_1, \ldots, i_k, j, i, i_{k+2}, \ldots, i_n$. The rewards under the two schedules are respectively

$$R_1 + \beta^{T+p_i} r_i + \beta^{T+p_i+p_j} r_j + R_2 \quad \text{and} \quad R_1 + \beta^{T+p_j} r_j + \beta^{T+p_j+p_i} r_i + R_2,$$

where $T = p_{i_1} + \cdots + p_{i_k}$, and R_1 and R_2 are respectively the sum of the rewards due to the jobs coming before and after jobs i, j; these are the same under both schedules. The reward of the first schedule is greater if $r_i \beta^{p_i}/(1 - \beta^{p_i}) > r_j \beta^{p_j}/(1 - \beta^{p_j})$. Hence a schedule can be optimal only if the jobs are taken in decreasing order of the indices $r_i \beta^{p_i}/(1 - \beta^{p_i})$. This type of reasoning is known as an **interchange argument**. ■

There are a couple points to note. (i) An interchange argument can be useful for solving a decision problem about a system that evolves in stages. Although such problems can be solved by dynamic programming, an interchange argument – when it works – is usually easier. (ii) The decision points need not be equally spaced in time. Here they are the points at which a number of jobs have been completed.

3.3 The infinite-horizon case

In the finite-horizon case the cost function is obtained simply from (3.3) by the backward recursion from the terminal point. However, when the horizon is infinite there is no terminal point and so the validity of the optimality equation is no longer obvious.

Let us consider the time-homogeneous Markov case, in which costs and dynamics do not depend on t, i.e., $c(x, u, t) = c(x, u)$. Suppose also that there is no terminal cost, i.e., $\mathbf{C}_h(x) = 0$. Define the *s-horizon cost under policy* π as

$$F_s(\pi, x) = E_\pi \left[\sum_{t=0}^{s-1} \beta^t c(x_t, u_t) \,\middle|\, x_0 = x \right],$$

where E_π denotes expectation over the path of the process under policy π. If we take the infimum with respect to π we have the *infimal s-horizon cost*

$$F_s(x) = \inf_\pi F_s(\pi, x).$$

Clearly, this always exists and satisfies the optimality equation

$$F_s(x) = \inf_u \{ c(x, u) + \beta E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u] \}, \tag{3.4}$$

with terminal condition $F_0(x) = 0$.

The *infinite-horizon cost under policy* π is also quite naturally defined as

$$F(\pi, x) = \lim_{s \to \infty} F_s(\pi, x). \tag{3.5}$$

This limit need not exist, but it will do so under any of the following scenarios.

10

D (**discounted programming**): $0 < \beta < 1$, and $|c(x, u)| < B$ for all x, u.

N (**negative programming**): $0 < \beta \leq 1$ and $c(x, u) \geq 0$ for all x, u.

P (**positive programming**): $0 < \beta \leq 1$ and $c(x, u) \leq 0$ for all x, u.

Notice that the names 'negative' and 'positive' appear to be the wrong way around with respect to the sign of $c(x, u)$. However, the names make sense if we think of equivalent problems of maximizing rewards. Maximizing positive rewards (P) is the same thing as minimizing negative costs. Maximizing negative rewards (N) is the same thing as minimizing positive costs. In cases N and P we usually take $\beta = 1$.

The existence of the limit (possibly infinite) in (3.5) is assured in cases N and P by monotone convergence, and in case D because the total cost occurring after the sth step is bounded by $\beta^s B/(1 - \beta)$.

3.4   The optimality equation in the infinite-horizon case

The *infimal infinite-horizon cost* is defined as

$$F(x) = \inf_{\pi} F(\pi, x) = \inf_{\pi} \lim_{s \to \infty} F_s(\pi, x). \tag{3.6}$$

The following theorem justifies our writing an optimality equation.

Theorem 3.1 *Suppose D, N, or P holds. Then $F(x)$ satisfies the optimality equation*

$$F(x) = \inf_{u} \{c(x, u) + \beta E[F(x_1) \mid x_0 = x, u_0 = u)]\}. \tag{3.7}$$

Proof.   We first prove that '$\geq$' holds in (3.7). Suppose π is a policy, which chooses $u_0 = u$ when $x_0 = x$. Then

$$F_s(\pi, x) = c(x, u) + \beta E[F_{s-1}(\pi, x_1) \mid x_0 = x, u_0 = u]. \tag{3.8}$$

Either D, N or P is sufficient to allow us to takes limits on both sides of (3.8) and interchange the order of limit and expectation. In cases N and P this is because of monotone convergence. Infinity is allowed as a possible limiting value. We obtain

$$\begin{aligned}
F(\pi, x) &= c(x, u) + \beta E[F(\pi, x_1) \mid x_0 = x, u_0 = u] \\
&\geq c(x, u) + \beta E[F(x_1) \mid x_0 = x, u_0 = u] \\
&\geq \inf_{u} \{c(x, u) + \beta E[F(x_1) \mid x_0 = x, u_0 = u]\}.
\end{aligned}$$

Minimizing the left hand side over π gives '$\geq$'.

To prove '$\leq$', fix x and consider a policy π that having chosen u_0 and reached state x_1 then follows a policy π^1 which is suboptimal by less than ϵ from that point, i.e., $F(\pi^1, x_1) \leq F(x_1) + \epsilon$. Note that such a policy must exist, by definition of F, although π^1 will depend on x_1. We have

11

$$F(x) \leq F(\pi, x)$$
$$= c(x, u_0) + \beta E[F(\pi^1, x_1) \mid x_0 = x, u_0]$$
$$\leq c(x, u_0) + \beta E[F(x_1) + \epsilon \mid x_0 = x, u_0]$$
$$\leq c(x, u_0) + \beta E[F(x_1) \mid x_0 = x, u_0] + \beta \epsilon.$$

Minimizing the right hand side over u_0 and recalling ϵ is arbitrary gives '$\leq$'. ■

3.5 Example: selling an asset

A inventor has built a better mouse trap. Each day he is made an offer for it, which he may either accept or reject. Offers are independently and identically distributed with probability density function $g(x)$, $x \geq 0$. Each day there is a probability $1 - \beta$ that a competitor will invent a better mouse trap and make the first invention worthless. Find the policy that maximizes the inventor's expected return and express his maximal return as the unique root of an equation. Show that if $\beta > 1/2$, $g(x) = 2/x^3$, $x \geq 1$, then he should accept the first bid which is at least $\sqrt{\beta/(1-\beta)}$.

Solution. There are only two states, depending on whether a bid has been accepted or not. Let these be 0 and 1 respectively. The optimality equation is

$$F(1) = \int_{y=0}^{\infty} \max[y, \beta F(1)] \, g(y) \, dy$$

$$= \beta F(1) + \int_{y=0}^{\infty} \max[y - \beta F(1), 0] \, g(y) \, dy$$

$$= \beta F(1) + \int_{y=\beta F(1)}^{\infty} [y - \beta F(1)] \, g(y) \, dy$$

Hence

$$(1 - \beta)F(1) = \int_{y=\beta F(1)}^{\infty} [y - \beta F(1)] \, g(y) \, dy. \tag{3.9}$$

That this equation has a unique root, $F(1) = F^*$, follows from the fact that left and right hand sides are increasing and decreasing in $F(1)$ respectively. Thus he should sell his invention when he is first offered at least βF^*. His maximal reward is F^*.

Consider the case $g(y) = 2/y^3$, $y \geq 1$. The left hand side of (3.9) is less that the right hand side at $F(1) = 1$ provided $\beta > 1/2$. In this case the root is greater than 1 and we compute it as

$$(1 - \beta)F(1) = 2/\beta F(1) - \beta F(1)/[\beta F(1)]^2,$$

and thus $F^* = 1/\sqrt{\beta(1-\beta)}$ and $\beta F^* = \sqrt{\beta/(1-\beta)}$.

If $\beta \leq 1/2$ we should immediately accept the first offer. ■

Notice that discounting arises in this problem because at each stage there is a probability β that a 'catastrophe' will occur that brings things to a sudden end. This characterization of a manner in which discounting can arise is often quite useful.

12

4   Positive Programming

We address the special theory of maximizing positive rewards, (noting that there may be no optimal policy but that if a policy has a value function that satisfies the optimality equation then it is optimal), and the method of value iteration.

4.1   Example: possible lack of an optimal policy.

Positive programming concerns minimizing non-positive costs, $c(x, u) \leq 0$. The name originates from the equivalent problem of maximizing non-negative rewards, $r(x, u) \geq 0$, and for this section we present results in that setting. The following example shows that there may be no optimal policy.

Suppose the possible states are the non-negative integers and in state x we have a choice of either moving to state $x + 1$ and receiving no reward, or moving to state 0, obtaining reward $1 - 1/i$, and then remaining in state 0 thereafter and obtaining no further reward. The optimality equations is

$$F(x) = \max\{1 - 1/x, F(x + 1)\} \quad x > 0 \, .$$

Clearly $F(x) = 1$, $x > 0$, but the policy that chooses the maximizing action in the optimality equation always moves on to state $x+1$ and hence has zero reward. Clearly, there is no policy that actually achieves a reward of 1.

4.2   Characterization of the optimal policy

The following theorem provides a necessary and sufficient condition for a policy to be optimal: namely, its value function must satisfy the optimality equation. This theorem also holds for the case of strict discounting and bounded costs.

Theorem 4.1 *Suppose D or P holds and π is a policy whose value function $F(\pi, x)$ satisfies the optimality equation*

$$F(\pi, x) = \sup_u \{r(x, u) + \beta E[F(\pi, x) \mid x_0 = x, u_0 = u]\} \, .$$

Then π is optimal.

Proof.   Let π' be any policy and suppose it takes $u_t(x) = f_t(x)$. Since $F(\pi, x)$ satisfies the optimality equation,

$$F(\pi, x) \geq r(x, f_0(x)) + \beta E_{\pi'}[F(\pi, x_1) \mid x_0 = x, u_0 = f_0(x)] \, .$$

By repeated substitution of this into itself, we find

$$F(\pi, x) \geq E_{\pi'}\left[\sum_{t=0}^{s-1} \beta^t r(x_t, u_t) \,\middle|\, x_0 = x\right] + \beta^s E_{\pi'}[F(\pi, x_s) \mid x_0 = x] \, . \qquad (4.1)$$

13

In case P we can drop the final term on the right hand side of (4.1) (because it is non-negative) and then let $s \to \infty$; in case D we can let $s \to \infty$ directly, observing that this term tends to zero. Either way, we have $F(\pi, x) \geq F(\pi', x)$. ∎

4.3   Example: optimal gambling

A gambler has i pounds and wants to increase this to N. At each stage she can bet any fraction of her capital, say $j \leq i$. Either she wins, with probability p, and now has $i + j$ pounds, or she loses, with probability $q = 1 - p$, and has $i - j$ pounds. Let the state space be $\{0, 1, \ldots, N\}$. The game stops upon reaching state 0 or N. The only non-zero reward is 1, upon reaching state N. Suppose $p \geq 1/2$. Prove that the timid strategy, of always betting only 1 pound, maximizes the probability of the gambler attaining N pounds.

Solution. The optimality equation is

$$F(i) = \max_{j, j \leq i}\{pF(i + j) + qF(i - j)\}\,.$$

To show that the timid strategy is optimal we need to find its value function, say $G(i)$, and show that it is a solution to the optimality equation. We have $G(i) = pG(i + 1) + qG(i - 1)$, with $G(0) = 0$, $G(N) = 1$. This recurrence gives

$$G(i) = \begin{cases} \dfrac{1 - (q/p)^i}{1 - (q/p)^N} & p > 1/2, \\ \dfrac{i}{N} & p = 1/2. \end{cases}$$

If $p = 1/2$, then $G(i) = i/N$ clearly satisfies the optimality equation. If $p > 1/2$ we simply have to verify that

$$G(i) = \frac{1 - (q/p)^i}{1 - (q/p)^N} = \max_{j:j \leq i}\left\{ p\left[\frac{1 - (q/p)^{i+j}}{1 - (q/p)^N}\right] + q\left[\frac{1 - (q/p)^{i-j}}{1 - (q/p)^N}\right] \right\}\,.$$

It is a simple exercise to show that $j = 1$ maximizes the right hand side. ∎

4.4   Value iteration

The infimal cost function F can be approximated by **successive approximation** or **value iteration**. This is important and practical method of computing F. Let us define
$$F_\infty(x) = \lim_{s \to \infty} F_s(x) = \lim_{s \to \infty} \inf_\pi F_s(\pi, x)\,. \tag{4.2}$$

This exists (by monotone convergence under N or P, or by the fact that under D the cost incurred after time s is vanishingly small.)

Notice that (4.2) reverses the order of $\lim_{s \to \infty}$ and $\inf_\pi$ in (3.6). The following theorem states that we can interchange the order of these operations and that therefore

14

$F_s(x) \to F(x)$. However, in case N we need an additional assumption:

F (**finite actions**): There are only finitely many possible values of u in each state.

Theorem 4.2 *Suppose that D or P holds, or N and F hold. Then* $F_\infty(x) = F(x)$.

Proof. First we prove '$\leq$'. Given any $\bar\pi$,

$$F_\infty(x) = \lim_{s\to\infty} F_s(x) = \lim_{s\to\infty} \inf_\pi F_s(\pi, x) \leq \lim_{s\to\infty} F_s(\bar\pi, x) = F(\bar\pi, x).$$

Taking the infimum over $\bar\pi$ gives $F_\infty(x) \leq F(x)$.

Now we prove '$\geq$'. In the positive case, $c(x, u) \leq 0$, so $F_s(x) \geq F(x)$. Now let $s \to \infty$. In the discounted case, with $|c(x, u)| < B$, imagine subtracting $B > 0$ from every cost. This reduces the infinite-horizon cost under any policy by exactly $B/(1-\beta)$ and $F(x)$ and $F_\infty(x)$ also decrease by this amount. All costs are now negative, so the result we have just proved applies. [Aternatively, note that

$$F_s(x) - \beta^s B/(1 - \beta) \leq F(x) \leq F_s(x) + \beta^s B/(1 - \beta)$$

(can you see why?) and hence $\lim_{s\to\infty} F_s(x) = F(x)$.]

In the negative case,

$$\begin{aligned}
F_\infty(x) &= \lim_{s\to\infty} \min_u \{c(x, u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\} \\
&= \min_u \{c(x, u) + \lim_{s\to\infty} E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\} \\
&= \min_u \{c(x, u) + E[F_\infty(x_1) \mid x_0 = x, u_0 = u]\}, \qquad (4.3)
\end{aligned}$$

where the first equality follows because the minimum is over a finite number of terms and the second equality follows by Lebesgue monotone convergence (since $F_s(x)$ increases in s). Let π be the policy that chooses the minimizing action on the right hand side of (4.3). This implies, by substitution of (4.3) into itself, and using the fact that N implies $F_\infty \geq 0$,

$$\begin{aligned}
F_\infty(x) &= E_\pi\left[\sum_{t=0}^{s-1} c(x_t, u_t) + F_\infty(x_s) \;\middle|\; x_0 = x \right] \\
&\geq E_\pi\left[\sum_{t=0}^{s-1} c(x_t, u_t) \;\middle|\; x_0 = x \right].
\end{aligned}$$

Letting $s \to \infty$ gives $F_\infty(x) \geq F(\pi, x) \geq F(x)$. ∎

4.5   Example: pharmaceutical trials

A doctor has two drugs available to treat a disease. One is well-established drug and is known to work for a given patient with probability p, independently of its success for

15

other patients. The new drug is untested and has an unknown probability of success θ, which the doctor believes to be uniformly distributed over $[0, 1]$. He treats one patient per day and must choose which drug to use. Suppose he has observed s successes and f failures with the new drug. Let $F(s, f)$ be the maximal expected-discounted number of future patients who are successfully treated if he chooses between the drugs optimally from this point onwards. For example, if he uses only the established drug, the expected-discounted number of patients successfully treated is $p + \beta p + \beta^2 p + \cdots = p/(1 - \beta)$. The posterior distribution of θ is

$$f(\theta \mid s, f) = \frac{(s + f + 1)!}{s! f!} \theta^s (1 - \theta)^f, \quad 0 \leq \theta \leq 1,$$

and the posterior mean is $\bar{\theta}(s, f) = (s + 1)/(s + f + 2)$. The optimality equation is

$$F(s, f) = \max \left[\frac{p}{1 - \beta}, \frac{s + 1}{s + f + 2} \left(1 + \beta F(s + 1, f)\right) + \frac{f + 1}{s + f + 2} \beta F(s, f + 1) \right].$$

It is not possible to give a nice expression for F, but we can find an approximate numerical solution. If $s + f$ is very large, say 300, then $\bar{\theta}(s, f) = (s + 1)/(s + f + 2)$ is a good approximation to θ. Thus we can take $F(s, f) \approx (1 - \beta)^{-1} \max[p, \bar{\theta}(s, f)]$, $s + f = 300$ and work backwards. For $\beta = 0.95$, one obtains the following table.

f \ s	0	1	2	3	4	5
0	.7614	.8381	.8736	.8948	.9092	.9197
1	.5601	.6810	.7443	.7845	.8128	.8340
2	.4334	.5621	.6392	.6903	.7281	.7568
3	.3477	.4753	.5556	.6133	.6563	.6899
4	.2877	.4094	.4898	.5493	.5957	.6326

These numbers are the greatest values of p for which it is worth continuing with at least one more trial of the new drug. For example, with $s = 3$, $f = 3$ it is worth continuing with the new drug when $p = 0.6 < 0.6133$. At this point the probability that the new drug will successfully treat the next patient is 0.5 and so the doctor should actually prescribe the drug that is least likely to cure! This example shows the difference between a **myopic policy**, which aims to maximize immediate reward, and an optimal policy, which forgets immediate reward in order to gain information and possibly greater rewards later on. Notice that it is worth using the new drug at least once if $p < 0.7614$, even though at its first use the new drug will only be successful with probability 0.5.

16

5 Negative Programming

We address the special theory of minimizing positive costs, (noting that the action that extremizes the right hand side of the optimality equation gives an optimal policy), and stopping problems and their solution.

5.1 Stationary policies

A **Markov policy** is a policy that specifies the control at time t to be simply a function of the state and time. In the proof of Theorem 4.1 we used $u_t = f_t(x_t)$ to specify the control at time t. This is a convenient notation for a Markov policy, and we write $\pi = (f_0, f_1, \dots)$. If in addition the policy does not depend on time, it is said to be a **stationary Markov policy**, and we write $\pi = (f, f, \dots) = f^\infty$.

5.2 Characterization of the optimal policy

Negative programming concerns minimizing non-negative costs, $c(x, u) \geq 0$. The name originates from the equivalent problem of maximizing non-positive rewards, $r(x, u) \leq 0$.

The following theorem gives a necessary and sufficient condition for a stationary policy to be optimal: namely, it must choose the optimal u on the right hand side of the optimality equation.

Theorem 5.1 *Suppose D or N holds. Suppose $\pi = f^\infty$ is the stationary Markov policy such that*

$$
c(x, f(x)) + \beta E[F(x_1) \mid x_0 = x, u_0 = f(x)] \\
= \min_u [c(x, u) + \beta E[F(x_1) \mid x_0 = x, u_0 = u]].
$$

Then $F(\pi, x) = F(x)$, and π is optimal.

Proof. Suppose this policy is $\pi = f^\infty$. Then by substituting the optimality equation into itself and using the fact that π specifies the minimizing control at each stage,

$$
F(x) = E_\pi \left[\sum_{t=0}^{s-1} \beta^t c(x_t, u_t) \, \middle| \, x_0 = x \right] + \beta^s E_\pi \left[F(x_s) \mid x_0 = x \right]. \tag{5.1}
$$

In case N we can drop the final term on the right hand side of (5.1) (because it is non-negative) and then let $s \to \infty$; in case D we can let $s \to \infty$ directly, observing that this term tends to zero. Either way, we have $F(x) \geq F(\pi, x)$. ∎

A corollary is that an optimal policy always exists. Neither Theorem 5.1 or this corollary are true for positive programming (c.f., the example in Section 4.1).

5.3 Optimal stopping over a finite horizon

One way that the total-expected cost can be finite is if it is possible to enter a state from which no further costs are incurred. Suppose u has just two possible values: $u = 0$ (stop), and $u = 1$ (continue). Suppose there is a termination state, say 0, that is entered upon choosing the stopping action. Once this state is entered the system stays in that state and no further cost is incurred thereafter.

Suppose that stopping is mandatory, in that we must continue for no more that s steps. The finite-horizon dynamic programming equation is therefore

$$F_s(x) = \min\{k(x), c(x) + E[F_{s-1}(x_1 \mid x_0 = x, u_0 = 1)]\}, \qquad (5.2)$$

with $F_0(x) = k(x)$, $c(0) = 0$.

Consider the set of states in which it is at least as good to stop now as to continue one more step and then stop:

$$S = \{x \ : \ k(x) \leq c(x) + E[k(x_1) \mid x_0 = x, u_0 = 1)]\}.$$

Clearly, it cannot be optimal to stop if $x \notin S$, since in that case it would be strictly better to continue one more step and then stop. The following theorem characterises all finite-horizon optimal policies.

Theorem 5.2 *Suppose S is closed (so that once the state enters S it remains in S.) Then an optimal policy for all finite horizons is: stop if and only if $x \in S$.*

Proof. The proof is by induction. If the horizon is $s = 1$, then obviously it is optimal to stop only if $x \in S$. Suppose the theorem is true for a horizon of $s - 1$. As above, if $x \notin S$ then it is better to continue for more one step and stop rather than stop in state x. If $x \in S$, then the fact that S is closed implies $x_1 \in S$ and so $F_{s-1}(x_1) = k(x_1)$. But then (5.2) gives $F_s(x) = k(x)$. So we should stop if $s \in S$. ■

The optimal policy is known as a **one-step look-ahead rule** (OSLA).

5.4 Example: optimal parking

A driver is looking for a parking space on the way to his destination. Each parking space is free with probability p independently of whether other parking spaces are free or not. The driver cannot observe whether a parking space is free until he reaches it. If he parks s spaces from the destination, he incurs cost s, $s = 0, 1, \ldots$. If he passes the destination without having parked the cost is D. Show that an optimal policy is to park in the first free space that is no further than s^* from the destination, where s^* is the greatest integer s such that $(Dp + 1)q^s \geq 1$.

Solution. When the driver is s spaces from the destination it only matters whether the space is available ($x = 1$) or full ($x = 0$). The optimality equation gives

$$F_s(0) = qF_{s-1}(0) + pF_{s-1}(1),$$

$$F_s(1) = \min \begin{cases} s, & \text{(take available space)} \\ qF_{s-1}(0) + pF_{s-1}(1), & \text{(ignore available space)} \end{cases}$$

18

where $F_0(0) = D$, $F_0(1) = 0$.

Suppose the driver adopts a policy of taking the first free space that is s or closer. Let the cost under this policy be $k(s)$, where

$$k(s) = ps + qk(s-1),$$

with $k(0) = qD$. The solution general solution is of the form $k(s) = a + bs + cq^s$. So after substituting and using the boundary condition at $s = 0$, we have

$$k(s) = -\frac{q}{p} + s + \left(D + \frac{1}{p}\right)q^{s+1}, \quad s = 0, 1, \dots \ .$$

It is better to stop now (at a distance s from the destination) than to go on and take the first available space if s is in the stopping set

$$S = \{s : s \le k(s-1)\} = \{s : (Dp + 1)q^s \ge 1\}.$$

This set is closed (since s decreases) and so by Theorem 5.2 this stopping set describes the optimal policy. ∎

If the driver parks in the first available space past his destination and walk backs, then $D = 1 + qD$, so $D = 1/p$ and s^* is the greatest integer such that $2q^s \ge 1$.

5.5 Optimal stopping over the infinite horizon

Let us now consider the stopping problem over the infinite-horizon. As above, let $F_s(x)$ be the infimal cost given that we are required to stop by time s. Let $F(x)$ be the infimal cost when all that is required is that we stop eventually. Since less cost can be incurred if we are allowed more time in which to stop, we have

$$F_s(x) \ge F_{s+1}(x) \ge F(x).$$

Thus by monotone convergence $F_s(x)$ tends to a limit, say $F_\infty(x)$, and $F_\infty(x) \ge F(x)$.

Example: we can have $F_\infty > F$

Consider the problem of stopping a symmetric random walk on the integers, where $c(x) = 0$, $k(x) = \exp(-x)$. The policy of stopping immediately, π, has $F(\pi, x) = \exp(-x)$, and this satisfies the infinite-horizon optimality equation,

$$F(x) = \min\{\exp(-x), (1/2)F(x+1) + (1/2)F(x-1)\}.$$

However, π is not optimal. A symmetric random walk is recurrent, so we may wait until reaching as large an integer as we like before stopping; hence $F(x) = 0$. Inductively, one can see that $F_s(x) = \exp(-x)$. So $F_\infty(x) > F(x)$.

(Note: Theorem 4.2 says that $F_\infty = F$, but that is in a setting in which there is no terminal cost and for different definitions of F_s and F than we take here.)

19

Example: Theorem 4.1 is not true for negative programming

Consider the above example, but now suppose one is allowed never to stop. Since continuation costs are 0 the optimal policy for all finite horizons and the infinite horizon is never to stop. So $F(x) = 0$ and this satisfies the optimality equation above. However, $F(\pi, x) = \exp(-x)$ also satisfies the optimality equation and is the cost incurred by stopping immediately. Thus it is not true (as for positive programming) that a policy whose cost function satisfies the optimality equation is optimal.

The following lemma gives conditions under which the infimal finite-horizon cost does converge to the infimal infinite-horizon cost.

Lemma 5.3 *Suppose all costs are bounded as follows.*

$$(a)\ K = \sup_x k(x) < \infty \qquad (b)\ C = \inf_x c(x) > 0\,. \qquad (5.3)$$

Then $F_s(x) \to F(x)$ as $s \to \infty$.

Proof. (*starred*) Suppose π is an optimal policy for the infinite horizon problem and stops at the random time τ. Then its cost is at least $(s+1)CP(\tau > s)$. However, since it would be possible to stop at time 0 the cost is also no more than K, so

$$(s+1)CP(\tau > s) \leq F(x) \leq K\,.$$

In the s-horizon problem we could follow π, but stop at time s if $\tau > s$. This implies

$$F(x) \leq F_s(x) \leq F(x) + KP(\tau > s) \leq F(x) + \frac{K^2}{(s+1)C}\,.$$

By letting $s \to \infty$, we have $F_\infty(x) = F(x)$. ∎

Note that the problem posed here is identical to one in which we pay K at the start and receive a terminal reward $r(x) = K - k(x)$.

Theorem 5.4 *Suppose S is closed and (5.3) holds. Then an optimal policy for the infinite horizon is: stop if and only if $x \in S$.*

Proof. By Theorem 5.2 we have for all finite s,

$$F_s(x) \begin{array}{ll} = k(x) & x \in S\,, \\ < k(x) & x \notin S\,. \end{array}$$

Lemma 5.3 gives $F(x) = F_\infty(x)$. ∎

6 Average-cost Programming

We address the infinite-horizon average-cost case, the optimality equation for this case and the policy improvement algorithm.

6.1 Average-cost optimization

It can happen that the undiscounted expected total cost is infinite, but the accumulation of cost per unit time is finite. Suppose that for a stationary Markov policy π, the following limit exists:

$$\lambda(\pi, x) = \lim_{t \to \infty} \frac{1}{t} E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \,\middle|\, x_0 = x \right].$$

It is reasonable to expect that there is a well-defined notion of an optimal **average-cost** function, $\lambda(x) = \inf_\pi \lambda(\pi, x)$, and that under appropriate assumptions, $\lambda(x) = \lambda$ should not depend on x. Moreover, one would expect

$$F_s(x) = s\lambda + \phi(x) + \epsilon(s, x),$$

where $\epsilon(s, x) \to 0$ as $s \to \infty$. Here $\phi(x) + \epsilon(s, x)$ reflects a transient due to the initial state. Suppose that the state space and action space are finite. From the optimality equation for the finite horizon problem we have

$$F_s(x) = \min_u \{ c(x, u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u] \}. \tag{6.1}$$

So by substituting $F_s(x) \sim s\lambda + \phi(x)$ into (6.1), we obtain

$$s\lambda + \phi(x) \sim \min_u \{ c(x, u) + E[(s-1)\lambda + \phi(x_1) \mid x_0 = x, u_0 = u] \}$$

which suggests, what it is in fact, the average-cost optimality equation:

$$\lambda + \phi(x) = \min_u \{ c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u] \}. \tag{6.2}$$

Theorem 6.1 *Let λ denote the minimal average-cost. Suppose there exists a constant λ' and bounded function ϕ such that for all x and u,*

$$\lambda' + \phi(x) \leq c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u]. \tag{6.3}$$

Then $\lambda' \leq \lambda$. average-cost. This also holds when $\leq$ is replaced throughout by $\geq$.

Proof.   Suppose u is chosen by some policy π. By repeated substitution of (6.3) into itself we have

$$\phi(x) \leq -t\lambda' + E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \,\middle|\, x_0 = x \right] + E_\pi[\phi(x_t) \mid x_0 = x]$$

21

Divide this by t and let $t \to \infty$ to obtain

$$0 \leq -\lambda' + \lim_{t \to \infty} \frac{1}{t} E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \, \middle| \, x_0 = x \right],$$

where the final term on the right hand side is simply the average-cost under policy π. Minimizing the right hand side over π gives the result. $\blacksquare$

Theorem 6.2 *Suppose there exists a constant λ and bounded function ϕ satisfying (6.2). Then λ is the minimal average-cost and the optimal stationary policy is the one that chooses the optimizing u on the right hand side of (6.2).*

Proof. Equation (6.2) implies that (6.3) holds with equality when one takes π to be the stationary policy that chooses the optimizing u on the right hand side of (6.2). Thus π is optimal and λ is the minimal average-cost. $\blacksquare$

The average-cost optimal policy is found simply by looking for a bounded solution to (6.2). Notice that if ϕ is a solution of (6.2) then so is ϕ+(a constant), because the (a constant) will cancel from both sides of (6.2). Thus ϕ is undetermined up to an additive constant. In searching for a solution to (6.2) we can therefore pick any state, say $\bar{x}$, and arbitrarily take $\phi(\bar{x}) = 0$.

6.2 Example: admission control at a queue

Each day a consultant is presented with the opportunity to take on a new job. The jobs are independently distributed over n possible types and on a given day the offered type is i with probability a_i, $i = 1, \ldots, n$. Jobs of type i pay R_i upon completion. Once he has accepted a job he may accept no other job until that job is complete. The probability that a job of type i takes k days is $(1 - p_i)^{k-1} p_i$, $k = 1, 2, \ldots$. Which jobs should the consultant accept?

Solution. Let 0 and i denote the states in which he is free to accept a job, and in which he is engaged upon a job of type i, respectively. Then (6.2) is

$$\lambda + \phi(0) = \sum_{i=1}^{n} a_i \max[\phi(0), \phi(i)],$$

$$\lambda + \phi(i) = (1 - p_i)\phi(i) + p_i[R_i + \phi(0)], \quad i = 1, \ldots, n.$$

Taking $\phi(0) = 0$, these have solution $\phi(i) = R_i - \lambda/p_i$, and hence

$$\lambda = \sum_{i=1}^{n} a_i \max[0, R_i - \lambda/p_i].$$

The left hand side is increasing in λ and the right hand side is decreasing λ. Hence there is a root, say λ^*, and this is the maximal average-reward. The optimal policy takes the form: *accept only jobs for which $p_i R_i \geq \lambda^*$.* $\blacksquare$

22

6.3 Value iteration bounds

Value iteration in the average-cost case is based upon the idea that $F_s(x) - F_{s-1}(x)$ approximates the minimal average-cost for large s.

Theorem 6.3 *Define*

$$m_s = \min_x\{F_s(x) - F_{s-1}(x)\}, \qquad M_s = \max_x\{F_s(x) - F_{s-1}(x)\}. \qquad (6.4)$$

Then $m_s \leq \lambda \leq M_s$, where λ is the minimal average-cost.

Proof. (*starred*) Suppose that the first step of a s-horizon optimal policy follows Markov plan f. Then

$$F_s(x) = F_{s-1}(x) + [F_s(x) - F_{s-1}(x)] = c(x, f(x)) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = f(x)].$$

Hence

$$F_{s-1}(x) + m_s \leq c(x, u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u],$$

for all x, u. Applying Theorem 6.1 with $\phi = F_{s-1}$ and $\lambda' = m_s$, implies $m_s \leq \lambda$. The bound $\lambda \leq M_s$ is established in a similar way. ∎

This justifies the following **value iteration algorithm**. At termination the algorithm provides a stationary policy that is within $\epsilon \times 100\%$ of optimal.

(0) Set $F_0(x) = 0$, $s = 1$.
(1) Compute F_s from

$$F_s(x) = \min_u\{c(x, u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\}.$$

(2) Compute m_s and M_s from (6.4). Stop if $M_s - m_s \leq \epsilon m_s$. Otherwise set $s := s + 1$ and goto step (1).

6.4 Policy improvement

Policy improvement is an effective method of improving stationary policies.

Policy improvement in the average-cost case.

In the average-cost case a policy improvement algorithm can be based on the following observations. Suppose that for a policy $\pi = f^\infty$, we have that λ, ϕ is a solution to

$$\lambda + \phi(x) = c(x, f(x_0)) + E[\phi(x_1) \mid x_0 = x, u_0 = f(x_0)],$$

and suppose for some policy $\pi_1 = f_1^\infty$,

$$\lambda + \phi(x) \geq c(x, f_1(x_0)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_1(x_0)], \qquad (6.5)$$

with strict inequality for some x. Then following the lines of proof in Theorem 6.1

$$\lim_{t\to\infty} \frac{1}{t} E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \,\middle|\, x_0 = x \right] = \lambda \geq \lim_{t\to\infty} \frac{1}{t} E_{\pi_1} \left[\sum_{s=0}^{t-1} c(x_s, u_s) \,\middle|\, x_0 = x \right].$$

If there is no π_1 for which (6.5) holds then π satisfies (6.2) and is optimal. This justifies the following **policy improvement algorithm**

(0) Choose an arbitrary stationary policy π_0. Set $s = 1$.
(1) For a given stationary policy $\pi_{s-1} = f_{s-1}^\infty$ determine ϕ, λ to solve

$$\lambda + \phi(x) = c(x, f_{s-1}(x)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_{s-1}(x)].$$

This gives a set of linear equations, and so is intrinsically easier to solve than (6.2).
(2) Now determine the policy $\pi_s = f_s^\infty$ from

$$c(x, f_s(x)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_s(x)]$$
$$= \min_u \{ c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u] \},$$

taking $f_s(x) = f_{s-1}(x)$ whenever this is possible. By applications of Theorem 6.1, this yields a strict improvement whenever possible. If $\pi_s = \pi_{s-1}$ then the algorithm terminates and π_{s-1} is optimal. Otherwise, return to step (1) with $s := s + 1$.

   If both the action and state spaces are finite then there are only a finite number of possible stationary policies and so the policy improvement algorithm will find an optimal stationary policy in finitely many iterations. By contrast, the value iteration algorithm that can only obtain more and more accurate approximations of λ^*.

Policy improvement in the discounted-cost case.

In the case of strict discounting, the following theorem plays the role of Theorem 6.1. The proof is similar, by repeated substitution of (6.6) into itself.

Theorem 6.4 *Suppose there exists a bounded function G such that for all x and u,*

$$G(x) \leq c(x, u) + \beta E[G(x_1) \mid x_0 = x, u_0 = u]. \qquad (6.6)$$

Then $G \leq F$, where F is the minimal discounted-cost function. This also holds when $\leq$ is replaced throughout by $\geq$.

   The policy improvement algorithm is similar. E.g., step (1) becomes

(1) For a given stationary policy $\pi_{s-1} = f_{s-1}^\infty$ determine G to solve

$$G(x) = c(x, f_{s-1}(x)) + \beta E[G(x_1) \mid x_0 = x, u_0 = f_{s-1}(x)].$$

7 LQ Models

We present the LQ regulation model in discrete and continuous time, the Riccati equation, its validity in a problem with noise, and the idea of linearization.

7.1 The LQ regulation model

The elements needed to define a control optimization problem are specification of (i) the dynamics of the process, (ii) which quantities are observable at a given time, and (iii) an optimization criterion.

In the **LQG model** the plant equation and observation relations are linear, the cost is quadratic, and the noise is Gaussian (jointly normal). The LQG model is important because it has a complete theory and introduces some key concepts, such as controllability, observability and the certainty-equivalence principle.

Begin with a model in which the state x_t is fully observable and there is no noise. The plant equation of the time-homogeneous $[A, B, \ \cdot \ ]$ system has the linear form

$$x_t = Ax_{t-1} + Bu_{t-1}, \tag{7.1}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, A is $n \times n$ and B is $n \times m$. The cost function is

$$\mathbf{C} = \sum_{t=0}^{h-1} c(x_t, u_t) + \mathbf{C}_h(x_h), \tag{7.2}$$

with one-step and terminal costs

$$c(x, u) = x^\top Rx + u^\top Sx + x^\top S^\top u + u^\top Qu = \begin{bmatrix} x \\ u \end{bmatrix}^\top \begin{bmatrix} R & S^\top \\ S & Q \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}, \tag{7.3}$$

$$\mathbf{C}_h(x) = x^\top \Pi_h x. \tag{7.4}$$

All quadratic forms are non-negative definite, and Q is positive definite. There is no loss of generality in assuming that R, Q and Π_h are symmetric. This is a model for **regulation** of (x, u) to the point $(0, 0)$ (i.e., steering to a critical value).

To solve the optimality equation we shall need the following lemma.

Lemma 7.1 *Suppose x, u are vectors. Consider a quadratic form*

$$\begin{pmatrix} x \\ u \end{pmatrix}^\top \begin{pmatrix} \Pi_{xx} & \Pi_{xu} \\ \Pi_{ux} & \Pi_{uu} \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}.$$

Assume it is symmetric and $\Pi_{uu} > 0$, i.e., positive definite. Then the minimum with respect to u is achieved at

$$u = -\Pi_{uu}^{-1}\Pi_{ux}x,$$

and is equal to

$$x^\top \left[\Pi_{xx} - \Pi_{xu}\Pi_{uu}^{-1}\Pi_{ux} \right] x.$$

Proof. Suppose the quadratic form is minimized at u. Then

$$\begin{pmatrix} x \\ u+h \end{pmatrix}^\top \begin{pmatrix} \Pi_{xx} & \Pi_{xu} \\ \Pi_{ux} & \Pi_{uu} \end{pmatrix} \begin{pmatrix} x \\ u+h \end{pmatrix}$$

$$= x^\top \Pi_{xx} x + 2x^\top \Pi_{xu} u + \underbrace{2h^\top \Pi_{ux} x + 2h^\top \Pi_{uu} u} + u^\top \Pi_{uu} u + h^\top \Pi_{uu} h\,.$$

To be stationary at u, the underbraced linear term in $h^\top$ must be zero, so

$$u = -\Pi_{uu}^{-1} \Pi_{ux} x\,,$$

and the optimal value is $x^\top \left[\Pi_{xx} - \Pi_{xu} \Pi_{uu}^{-1} \Pi_{ux} \right] x.$ ∎

Theorem 7.2 *Assume the structure of (7.1)–(7.4). Then the value function has the quadratic form*

$$F(x,t) = x^\top \Pi_t x\,, \quad t < h\,, \tag{7.5}$$

and the optimal control has the linear form

$$u_t = K_t x_t\,, \quad t < h\,.$$

The time-dependent matrix Π_t satisfies the Riccati equation

$$\Pi_t = f\Pi_{t+1}\,, \quad t < h\,, \tag{7.6}$$

where f is an operator having the action

$$f\Pi = R + A^\top \Pi A - (S^\top + A^\top \Pi B)(Q + B^\top \Pi B)^{-1}(S + B^\top \Pi A)\,, \tag{7.7}$$

and Π_h has the value prescribed in (7.4). The $m \times n$ matrix K_t is given by

$$K_t = -(Q + B^\top \Pi_{t+1} B)^{-1}(S + B^\top \Pi_{t+1} A)\,, \quad t < h\,.$$

Proof. Assertion (7.5) is true at time h. Assume it is true at time $t+1$. Then

$$F(x,t) = \inf_u \left[c(x,u) + (Ax + Bu)^\top \Pi_{t+1}(Ax + Bu) \right]$$

$$= \inf_u \left[\begin{pmatrix} x \\ u \end{pmatrix}^\top \begin{pmatrix} R + A^\top \Pi_{t+1} A & S^\top + A^\top \Pi_{t+1} B \\ S + B^\top \Pi_{t+1} A & Q + B^\top \Pi_{t+1} B \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \right]$$

By Lemma 7.1 the minimum is achieved by $u = K_t x$, and the form of f comes from this also. ∎

7.2   The Riccati recursion

The backward recursion (7.6)–(7.7) is called the **Riccati equation**. Note that

(i) S can be normalized to zero by choosing a new control $u^* = u + Q^{-1}Sx$, and setting $A^* = A - BQ^{-1}S$, $R^* = R - S^\top Q^{-1}S$.

(ii) The optimally controlled process obeys $x_{t+1} = \Gamma_t x_t$. Here Γ_t is called the **gain matrix** and is given by

$$\Gamma_t = A + BK_t = A - B(Q + B^\top \Pi_{t+1}B)^{-1}(S + B^\top \Pi_{t+1}A).$$

(iii) An equivalent expression for the Riccati equation is

$$f\Pi = \inf_K \left[R + K^\top S + S^\top K + K^\top QK + (A + BK)^\top \Pi(A + BK) \right].$$

(iv) We might have carried out exactly the same analysis for a time-heterogeneous model, in which the matrices A, B, Q, R, S are replaced by A_t, B_t, Q_t, R_t, S_t.

7.3   Example: additive white noise

Suppose the plant equation (7.1) is replaced by

$$x_{t+1} = Ax_t + Bu_t + \epsilon_t,$$

where $\epsilon_t \in \mathbb{R}^n$ is vector **white noise**, defined by the properties $E\epsilon = 0$, $E\epsilon_t \epsilon_t^\top = N$ and $E\epsilon_t \epsilon_s^\top = 0$, $t \neq s$. The DP equation is then

$$F(x, t) = \inf_u \left[c(x, u) + E_\epsilon[(F(Ax + Bu + \epsilon, t + 1)] \right].$$

By definition $F(x, h) = x^\top \Pi_h x$. Try a solution $F(x, t) = x^\top \Pi_t x + \gamma_t$. This holds for $t = h$. Suppose it is true for $t + 1$, then

$$F(x, t) = \inf_u \left[c(x, u) + E(Ax + Bu + \epsilon)^\top \Pi_{t+1}(Ax + Bu + \epsilon) + \gamma_{t+1} \right]$$

$$= \inf_u \left[c(x, u) + E(Ax + Bu)^\top \Pi_{t+1}(Ax + Bu) \right]$$

$$+ 2E\left[\epsilon^\top (Ax + Bu) \right] + E\left[\epsilon^\top \Pi_{t+1}\epsilon \right] + \gamma_{t+1}$$

$$= \inf_u [\cdots] + 0 + \text{tr}(N\Pi_{t+1}) + \gamma_{t+1}.$$

Here we use the fact that

$$E\left[\epsilon^\top \Pi\epsilon \right] = E\left[\sum_{ij} \epsilon_i \Pi_{ij} \epsilon_j \right] = E\left[\sum_{ij} \epsilon_j \epsilon_i \Pi_{ij} \right] = \sum_{ij} N_{ji} \Pi_{ij} = \text{tr}(N\Pi).$$

Thus (i) Π_t follows the same Riccati equation as before, (ii) the optimal control is $u_t = K_t x_t$, and (iii)

$$F(x, t) = x^\top \Pi_t x + \gamma_t = x^\top \Pi_t x + \sum_{j=t+1}^{h} \text{tr}(N\Pi_j).$$

27

The final term can be viewed as the cost of correcting future noise. In the infinite horizon limit of $\Pi_t \to \Pi$ as $t \to \infty$, we incur an average cost per unit time of $\mathrm{tr}(N\Pi)$, and a transient cost of $x^\top \Pi x$ that is due to correcting the initial x.

7.4 Continuous-time LQ regulation

In continuous-time we take $\dot{x} = Ax + Bu$ and

$$\mathbf{C} = \int_0^h \begin{pmatrix} x \\ u \end{pmatrix}^\top \begin{pmatrix} R & S^\top \\ S & Q \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} dt + (x^\top \Pi x)_h \,.$$

We can obtain the continuous-time solution from the discrete time solution, moving forward in time in increments of Δ.

$$x_{t+1} \to x_{t+\Delta}, \quad A \to I + A\Delta, \quad B \to B\Delta, \quad R, S, Q \to R\Delta, S\Delta, Q\Delta \,.$$

Then as before, $F(x,t) = x^\top \Pi x$, where Π obeys the Riccati equation

$$\frac{\partial \Pi}{\partial t} + R + A^\top \Pi + \Pi A - (S^\top + \Pi B) Q^{-1} (S + B^\top \Pi) = 0 \,.$$

This is simpler than the discrete time version. The optimal control is

$$u(t) = K(t) x(t)$$

where

$$K(t) = -Q^{-1}(S + B^\top \Pi) \,.$$

The optimally controlled plant equation is $\dot{x} = \Gamma(t)x$, where $\Gamma(t) = A + BK = A - BQ^{-1}(S + B^\top \Pi)$.

7.5 Linear models as linearization

Linear models in continuous-time are important because they arise naturally via the linearization of nonlinear models. Consider the state-structured nonlinear model:

$$\dot{x} = a(x, u).$$

Suppose x, u are perturbed from an equilibrium $(\bar{x}, \bar{u})$ where $a(\bar{x}, \bar{u}) = 0$. Let $x' = x - \bar{x}$ and $u' = u - \bar{u}$ and immediately drop the primes. The linearized version is

$$\dot{x} = Ax + Bu$$

where

$$A = \left. \frac{\partial a}{\partial x} \right|_{(\bar{x}, \bar{u})}, \quad B = \left. \frac{\partial a}{\partial u} \right|_{(\bar{x}, \bar{u})} \,.$$

If $\bar{x}, \bar{u}$ is to be a stable equilibrium point then we must be able to choose a control that can stabilise the system in the neighbourhood of $(\bar{x}, \bar{u})$.

28

8 Controllability

We generalize LQ regulation to cases with disturbances and variable set points. We also define controllability and give conditions for controllability in discrete and continuous time.

8.1 Disturbances

Suppose the plant equation of the previous lecture, i.e.,

$$x_t = Ax_{t-1} + Bu_{t-1} \tag{8.1}$$

is replaced by

$$x_t = Ax_{t-1} + Bu_{t-1} + \alpha_t, \tag{8.2}$$

where $\{\alpha_0, \ldots, \alpha_{h-1}\}$ is a known sequence of **disturbances**. By an exactly similar (but algebraically daunting) inductive argument we can show that F has the form

$$F(x, t) = x^\top \Pi_t x - 2\sigma_t^\top x + \omega_t,$$

where Π_t satisfies exactly the same Riccati equation as before, (7.6)–(7.7), and

$$\sigma_t = \Gamma_t^\top (\sigma_{t+1} - \Pi_{t+1}\alpha_{t+1}), \quad t < h.$$

The optimal control is

$$u_t = K_t x_t + (Q + B^\top \Pi_{t+1} B)^{-1} B^\top (\sigma_{t+1} - \Pi_{t+1}\alpha_{t+1})$$

$$= K_t x_t - (Q + B^\top \Pi_{t+1} B)^{-1} B^\top$$

$$\times \left(\sum_{s=0}^{h-t-1} \Gamma_{t+1}^\top \Gamma_{t+2}^\top \cdots \Gamma_{t+s}^\top \Pi_{t+s+1}\alpha_{t+s+1} - \Gamma_{t+1}^\top \cdots \Gamma_{h-1}^\top \sigma_h \right), \quad t < h.$$

This consists of the same feedback term as before, $K_t x_t$, plus a feedforward term which, as a function of future disturbances, anticipates disturbances and acts intelligently. (In fact, $\sigma_h = 0$, but σ_h would be non-zero if the terminal cost had included a linear term in x_h.) We don't bother to work out the recursion for ω_t, since this does not affect the control.

8.2 Tracking

Suppose (8.2) holds and the cost is now

$$c(x, u) = \begin{bmatrix} x - \bar{x}_t \\ u - \bar{u}_t \end{bmatrix}^\top \begin{bmatrix} R & S^\top \\ S & Q \end{bmatrix} \begin{bmatrix} x - \bar{x}_t \\ u - \bar{u}_t \end{bmatrix}.$$

29

The problem is now one of **tracking** a sequence of values $(\bar{x}_t, \bar{u}_t)$, $t = 0, \ldots, h - 1$. We can solve this model by adopting new variables $x_t^* = x_t - \bar{x}_t$ and $u_t^* = u_t - \bar{u}_t$. The cost as a function of (x^*, u^*) takes standard form and the plant equation is now

$$x_t^* = Ax_{t-1}^* + Bu_{t-1}^* + \alpha_t^*,$$

and $\bar{\alpha}_t = \bar{x}_t - A\bar{x}_{t-1} - B\bar{u}_{t-1}$. The result in subsection 8.1 applies, except that α_t is replaced with $\alpha_t^* = \alpha_t - \bar{\alpha}_t$.

8.3 Controllability

Consider the $[A, B, \cdot]$ system with plant equation (8.1). The **controllability** question is: can we bring x to an arbitrary prescribed value by some u-sequence?

Definition 8.1 *The system is **r-controllable** if one can bring it from an arbitrary prescribed x_0 to an arbitrary prescribed x_r by some u-sequence $u_0, u_1, \ldots, u_{r-1}$. A system of dimension n is said to be **controllable** if it is r-controllable for some r*

Example. If B is square and non-singular then the system is 1-controllable, for

$$x_1 = Ax_0 + Bu_0 \quad \text{where} \quad u_0 = B^{-1}(x_1 - Ax_0).$$

Example. Consider the case, $(n = 2, m = 1)$,

$$x_t = \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} x_{t-1} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u_{t-1}.$$

This system is not 1-controllable. But

$$x_2 - A^2 x_0 = Bu_1 + ABu_0 = \begin{pmatrix} 1 & a_{11} \\ 0 & a_{21} \end{pmatrix} \begin{pmatrix} u_1 \\ u_0 \end{pmatrix}.$$

So it is 2-controllable if and only if $a_{21} \neq 0$.

More generally, by substituting (8.1) into itself, we see that we must find $u_0, u_1, \ldots, u_{r-1}$ to satisfy

$$\Delta = x_r - A^r x_0 = Bu_{r-1} + ABu_{r-2} + \cdots + A^{r-1}Bu_0, \tag{8.3}$$

for arbitrary Δ. In providing conditions for controllability we shall need to make use of the following theorem.

Theorem 8.2 (The Cayley-Hamilton theorem) *Any $n \times n$ matrix A satisfies its own characteristic equation. So that if*

$$\det(\lambda I - A) = \sum_{j=0}^{n} a_j \lambda^{n-j}$$

then

$$\sum_{j=0}^{n} a_j A^{n-j} = 0. \tag{8.4}$$

30

The implication is that $I, A, A^2, \ldots, A^{n-1}$ constitute a basis for A^r, $r = 0, 1, \ldots$. Proof. (*starred*) Define

$$\Phi(z) = \sum_{j=0}^{\infty} (Az)^j = (I - Az)^{-1} = \frac{\text{adj}(I - Az)}{\det(I - Az)}.$$

Then

$$\det(I - Az)\Phi(z) = \sum_{j=0}^{n} a_j z^j \Phi(z) = \text{adj}(I - Az),$$

which implies (8.4) since the coefficient of z^n must be zero. ∎

We are now in a position to characterise controllability.

Theorem 8.3 *(i) The system $[A, B, \cdot\,]$ is r-controllable if and only if the matrix*

$$M_r = \begin{bmatrix} B & AB & A^2 B & \cdots & A^{r-1}B \end{bmatrix}$$

has rank n, or (ii) equivalently, if and only if the $n \times n$ matrix

$$M_r M_r^\top = \sum_{j=0}^{r-1} A^j (BB^\top)(A^\top)^j$$

is nonsingular (or, equivalently, positive definite.) (iii) If the system is r-controllable then it is s-controllable for $s \geq \min(n, r)$, and (iv) a control transferring x_0 to x_r with minimal cost $\sum_{t=0}^{r-1} u_t^\top u_t$ is

$$u_t = B^\top (A^\top)^{r-t-1}(M_r M_r^\top)^{-1}(x_r - A^r x_0), \quad t = 0, \ldots, r - 1.$$

Proof.   (i) The system (8.3) has a solution for arbitrary Δ if and only if M_r has rank n. (ii) $M_r M_r^\top$ is singular if and only if there exists w such that $M_r M_r^\top w = 0$, and

$$M_r M_r^\top w = 0 \iff w^\top M_r M_r^\top w = 0 \iff M_r^\top w = 0.$$

(iii) The rank of M_r is non-decreasing in r, so if it is r-controllable, then it is s-controllable for $s \geq r$. But the rank is constant for $r \geq n$ by the Cayley-Hamilton theorem. (iv) Consider the Lagrangian

$$\sum_{t=0}^{r-1} u_t^\top u_t + \lambda^\top (\Delta - \sum_{t=0}^{r-1} A^{r-t-1} B u_t),$$

giving

$$u_t = \tfrac{1}{2} B^\top (A^\top)^{r-t-1} \lambda.$$

Now we can determine λ from (8.3). ∎

31

8.4 Controllability in continuous-time

Theorem 8.4 *(i) The n dimensional system $[A, B, \;\cdot\;]$ is controllable if and only if the matrix M_n has rank n, or (ii) equivalently, if and only if*

$$G(t) = \int_0^t e^{As} B B^\top e^{A^\top s} \, ds,$$

is positive definite for all $t > 0$. (iii) If the system is controllable then a control that achieves the transfer from $x(0)$ to $x(t)$ with minimal control cost $\int_0^t u_s^\top u_s ds$ is

$$u(s) = B^\top e^{A^\top (t-s)} G(t)^{-1} (x(t) - e^{At} x(0)).$$

Note that there is now no notion of r-controllability. However, $G(t) \downarrow 0$ as $t \downarrow 0$, so the transfer becomes more difficult and costly as $t \downarrow 0$.

8.5 Example: broom balancing

Consider the problem of balancing a broom in an upright position on your hand. By Newton's laws, the system obeys $m\ddot{x} \cos\theta = m(\ddot{u} \cos\theta + L\ddot{\theta}) = mg \sin\theta$. For small θ we have $\theta \sim \sin\theta \sim (x - u)/L$, so with $\alpha = g/L$ the plant equation is

$$\ddot{x} = \alpha(x - u),$$

equivalently,

$$\frac{d}{dt} \begin{pmatrix} x \\ \dot{x} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \alpha & 0 \end{pmatrix} \begin{pmatrix} x \\ \dot{x} \end{pmatrix} + \begin{pmatrix} 0 \\ -\alpha \end{pmatrix} u \, .$$
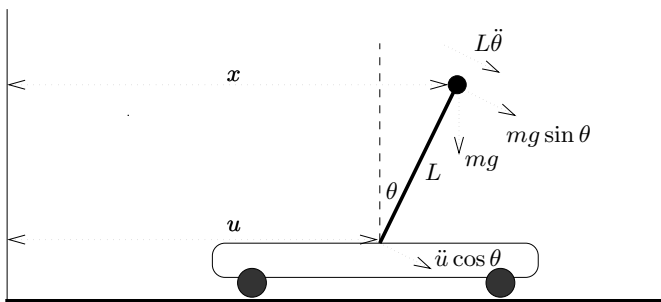

Figure 1: Force diagram for broom balancing

Since

$$\begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 0 & -\alpha \\ -\alpha & 0 \end{bmatrix},$$

the system is controllable if θ is initially small.

9 Infinite Horizon Limits

We presents one further example of controllability in continuous-time, the notion of stabilizability and the infinite horizon limit for the LQ regulation problem.

9.1 Example: satellite in a plane orbit

Consider a satellite of unit mass in a planar orbit and take polar coordinates (r, θ).

$$\ddot{r} = r\dot{\theta}^2 - \frac{c}{r^2} + u_r, \qquad \ddot{\theta} = -\frac{2\dot{r}\dot{\theta}}{r} + \frac{1}{r}u_\theta,$$

where u_r and u_θ are the radial and tangential components of thrust. If $u = 0$ then a possible orbit (such that $\dot{r} = \ddot{\theta} = 0$) is with $r = \rho$ and $\dot{\theta} = \omega = \sqrt{c/\rho^3}$.

Recall that one reason for taking an interest in linear models is that they tell us about controllability around an equilibrium point. Imagine there is a perturbing force. Take coordinates of perturbation

$$x_1 = r - \rho, \quad x_2 = \dot{r}, \quad x_3 = \theta - \omega t, \quad x_4 = \dot{\theta} - \omega.$$

Then, with $n = 4$, $m = 2$,

$$\dot{x} \sim \begin{pmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega\rho \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega/\rho & 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1/\rho \end{pmatrix} \begin{pmatrix} u_r \\ u_\theta \end{pmatrix} = Ax + Bu.$$

It is easy to check that $M_2 = \begin{bmatrix} B & AB \end{bmatrix}$ has rank 4 and that therefore the system is controllable.

But suppose $u_\theta = 0$ (tangential thrust fails). Then

$$B = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad M_4 = \begin{bmatrix} B & AB & A^2B & A^3B \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & -\omega^2 \\ 1 & 0 & -\omega^2 & 0 \\ 0 & 0 & -2\omega/\rho & 0 \\ 0 & -2\omega/\rho & 0 & 2\omega^3/\rho \end{bmatrix}.$$

Since $(2\omega\rho, 0, 0, \rho^2)M_4 = 0$, this is singular and has rank 3. The uncontrollable component is the angular momentum, $2\omega\rho\delta r + \rho^2\delta\dot{\theta} = \delta(r^2\dot{\theta})|_{r=\rho, \dot{\theta}=\omega}$.

On the other hand, if $u_r = 0$ then the system is controllable. We can change the radius by tangential breaking or thrust.

9.2 Stabilizability

Using the stationary control, $u_t = Kx_t$, we have $x_t = Ax_{t-1} + Bu_{t-1} = (A + BK)x_{t-1}$. So with $\Gamma = A + BK$, we have $x_t = \Gamma^t x_0$.

Definition 9.1 *We say that Γ is a **stability matrix** in the discrete-time sense if all eigenvalues of Γ lie strictly inside the unit disc in the complex plane, $|z| = 1$.*

*The $[A, B]$ system is said to **stabilizable** if there exists a K such that $A + BK$ is a stability matrix, and hence that $x_t \to 0$ as $t \to \infty$.*

Note that $u_t = Kx_t$ is linear and Markov. In seeking controls such that $x_t \to 0$ it is sufficient only to consider controls of this type since, as we see below, such controls arise as optimal controls for the infinite-horizon LQ regulation problem.

9.3 Stability in continuous-time

In continuous time the plant equation is $\dot{x} = Ax + Bu$. So if we use a control $u = Kx$ and set $\Gamma = A + BK$,

$$\dot{x} = \Gamma x, \quad x = e^{\Gamma t}x_0, \quad \text{where } e^{\Gamma t} = \sum_{j=0}^{\infty}(\Gamma t)^j/j!$$

We say Γ is a stability matrix in the continuous-time sense if all eigenvalues of Γ have negative real part, and hence that $x_t \to 0$ as $t \to \infty$.

9.4 Example: pendulum

Consider a pendulum of length L, unit mass bob and angle θ to the vertical. Suppose we wish to stabilise θ to zero by application of a force u. Then

$$\ddot{\theta} = -gL\sin\theta + u.$$

We change the state variable to $x = (\theta, \dot{\theta})$ and write

$$\frac{d}{dt}\begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} \dot{\theta} \\ -gL\sin\theta + u \end{pmatrix}$$

$$\sim \begin{pmatrix} \dot{\theta} \\ -gL\theta \end{pmatrix} + \begin{pmatrix} 0 \\ u \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 1 \\ -gL & 0 \end{pmatrix}\begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}u.$$

Suppose we try to stabilise with a control $u = -K\theta = -Kx_1$. Then

$$A + BK = \begin{pmatrix} 0 & 1 \\ -gL - K & 0 \end{pmatrix}$$

and this has eigenvalues $\pm\sqrt{-gL - K}$. So either $-gL - K > 0$ and one eigenvalue has a positive real part, in which case there is in fact instability, or $-gL - K < 0$ and eigenvalues are purely imaginary, which means we will in general have oscillations. So successful regulation must be a function of $\dot{\theta}$ as well, (and this would come out of solution to the LQ regulation problem.)

9.5 Infinite-horizon LQ regulation

Consider the time-homogeneous case and write the finite-horizon cost in terms of time to go s. The terminal cost, when $s = 0$, is denoted $F_0(x) = x^\top \Pi_0 x$. In all that follows we take $S = 0$, without loss of generality.

Lemma 9.2 *Suppose $\Pi_0 = 0$, $R \geq 0$, $Q \geq 0$ and $[A, B, \cdot]$ is controllable or stabilizable. Then $\{\Pi_s\}$ has a finite limit Π.*

Proof. Costs are non-negative, so $F_s(x)$ is non-decreasing in s. Now $F_s(x) = x^\top \Pi_s x$. Thus $x^\top \Pi_s x$ is non-decreasing in s for every x. To show that $x^\top \Pi_s x$ is bounded we use one of two arguments.

If the system is controllable then $x^\top \Pi_s x$ is bounded because there is a policy which, for any $x_0 = x$, will bring the state to zero in at most n steps and at finite cost and can then hold it at zero with zero cost thereafter.

If the system is stabilizable then there is a K such that $\Gamma = A + BK$ is a stability matrix and using $u_t = Kx_t$, we have

$$F_s(x) \leq x^\top \left[\sum_{t=0}^\infty (\Gamma^\top)^t (R + K^\top Q K) \Gamma^t \right] x < \infty.$$

Hence in either case we have an upper bound and so $x^\top \Pi_s x$ tends to a limit for every x. By considering $x = e_j$, the vector with a unit in the jth place and zeros elsewhere, we conclude that the jth element on the diagonal of Π_s converges. Then taking $x = e_j + e_k$ it follows that the off diagonal elements of Π_s also converge. ∎

Both value iteration and policy improvement are effective ways to compute the solution to an infinite-horizon LQ regulation problem. Policy improvement goes along the lines developed in Lecture 6.

The following theorem establishes the efficacy of value iteration. It is similar to Theorem 4.2 which established the same fact for D, N and P programming. The LQ regulation problem is a negative programming problem, however we cannot apply Theorem 4.2, because in general the terminal cost of $x^\top \Pi_0 x$ is not zero.

Theorem 9.3 *Suppose that $R > 0$, $Q > 0$ and the system $[A, B, \cdot]$ is controllable. Then (i) The equilibrium Riccati equation*

$$\Pi = f\Pi \tag{9.1}$$

has a unique non-negative definite solution Π. (ii) For any finite non-negative definite Π_0 the sequence $\{\Pi_s\}$ converges to Π. (iii) The gain matrix Γ corresponding to Π is a stability matrix.

Proof. (*starred*) Define Π as the limit of the sequence $f^{(s)}0$. By the previous lemma we know that this limit exists and that it satisfies (9.1).

Consider $u_t = Kx_t$ and $x_{t+1} = (A + BK)x_t = \Gamma x_t = \Gamma^t x_0$, for arbitrary x_0, where $K = -(Q + B^\top \Pi B)^{-1} B^\top \Pi A$ and $\Gamma = A + BK$. We can write (9.1) as

$$\Pi = R + K^\top Q K + \Gamma^\top \Pi \Gamma. \tag{9.2}$$

and hence

$$x_t^\top \Pi x_t = x_t^\top (R + K^\top Q K)x_t + x_{t+1}^\top \Pi x_{t+1} \geq x_{t+1}^\top \Pi x_{t+1}.$$

Thus $x_t^\top \Pi x_t$ decreases and, being bounded below by zero, it tends to a limit. Thus $x_t^\top (R + K^\top Q K)x_t$ tends to 0. Since $R + K^\top Q K$ is positive definite this implies $x_t \to 0$, which implies (iii). Hence for arbitrary finite non-negative definite Π_0,

$$\Pi_s = f^{(s)} \Pi_0 \geq f^{(s)} 0 \to \Pi. \tag{9.3}$$

However, if we choose the fixed policy $u_t = Kx_t$ then it follows that

$$\Pi_s \leq f^{(s)} 0 + (\Gamma^\top)^s \Pi_0 \Gamma^s \to \Pi. \tag{9.4}$$

Thus (9.3) and (9.4) imply (ii).

Finally, if non-negative definite $\tilde{\Pi}$ also satisfies (9.1) then $\tilde{\Pi} = f^{(s)} \tilde{\Pi} \to \Pi$, whence (i) follows. ■

9.6   The $[A, B, C]$ system

The notion of controllability rested on the assumption that the initial value of the state was known. If, however, one must rely upon imperfect observations, then the question arises whether the value of state (either in the past or in the present) can be determined from these observations. The discrete-time system $[A, B, C]$ is defined by the plant equation and observation relation

$$x_t = Ax_{t-1} + Bu_{t-1},$$
$$y_t = Cx_{t-1}.$$

Here $y \in \mathbb{R}^r$ is observed, but x is not. We suppose C is $r \times n$. The observability question is whether or not we can infer x_0 from the observations $y_1, y_2, \ldots$. The notion of observability stands in dual relation to that of controllability; a duality that indeed persists throughout the subject.

10   Observability

We present observability and the LQG model with process and observation noise.

10.1   Observability

The previous lecture introduced the $[A, B, C]$ system,

$$x_t = Ax_{t-1} + Bu_{t-1}, \tag{10.1}$$

$$y_t = Cx_{t-1}. \tag{10.2}$$

$y \in \mathbb{R}^r$ is observed, but x is not. C is $r \times n$. The **observability** question is: can we infer x at a prescribed time by subsequent y values?

Definition 10.1 *A system is said to be **r-observable** if x_0 can be inferred from knowledge of the subsequent observations $y_1, \ldots, y_r$ and relevant control values $u_0, \ldots, u_{r-2}$ for any initial x_0. A system of dimension n is said to be **observable** if it is r-observable for some r.*

From (10.1) and (10.2) we can determine y_t in terms of x_0 and subsequent controls:

$$x_t = A^t x_0 + \sum_{s=0}^{t-1} A^s B u_{t-s-1},$$

$$y_t = Cx_{t-1} = C \left[A^{t-1} x_0 + \sum_{s=0}^{t-2} A^s B u_{t-s-2} \right].$$

Thus, if we define the 'reduced observation'

$$\tilde{y}_t = y_t - C \left[\sum_{s=0}^{t-2} A^s B u_{t-s-2} \right],$$

then x_0 is to be determined from the system of equations

$$\tilde{y}_t = CA^{t-1} x_0, \quad 1 \le t \le r. \tag{10.3}$$

By hypothesis, these equations are mutually consistent, and so have a solution; the question is whether this solution is unique. This is the reverse of the situation for controllability, when the question was whether the equation for u had a solution at all, unique or not. Note that an implication of the system definition is that the property of observability depends only on the matrices A and C; not upon B at all.

Theorem 10.2 *(i) The system $[A, \; \cdot \; , C]$ is r-observable if and only if the matrix*

$$N_r = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{r-1} \end{bmatrix}$$

has rank n, or (ii) equivalently, if and only if the $n \times n$ matrix

$$N_r^\top N_r = \sum_{j=0}^{r-1} (A^\top)^j C^\top C A^j$$

is nonsingular. (iii) If the system is r-observable then it is s-observable for $s \geq \min(n, r)$, and (iv) the determination of x_0 can be expressed

$$x_0 = (N_r^\top N_r)^{-1} \sum_{j=1}^{r} (A^\top)^{j-1} C^\top \tilde{y}_j. \tag{10.4}$$

Proof. If the system has a solution for x_0 (which is so by hypothesis) then this solution must is unique if and only if the matrix N_r has rank n, whence assertion (i). Assertion (iii) follows from (i). The equivalence of conditions (i) and (ii) can be verified directly as in the case of controllability.

If we define the deviation $\eta_t = \tilde{y}_t - CA^{t-1} x_0$ then the equation amounts to $\eta_t = 0$, $1 \leq t \leq r$. If these equations were not consistent we could still define a 'least-squares' solution to them by minimizing any positive-definite quadratic form in these deviations with respect to x_0. In particular, we could minimize $\sum_{t=0}^{r-1} \eta_t^\top \eta_t$. This minimization gives (10.4). If equations (10.3) indeed have a solution (i.e., are mutually consistent, as we suppose) and this is unique then expression (10.4) must equal this solution; the actual value of x_0. The criterion for uniqueness of the least-squares solution is that $N_r^\top N_r$ should be nonsingular, which is also condition (ii). ∎

Note that we have again found it helpful to bring in an optimization criterion in proving (iv); this time, not so much to construct one definite solution out of many, but rather to construct a 'best-fit' solution where an exact solution might not have existed. This approach lies close to the statistical approach necessary when observations are corrupted by noise.

10.2   Observability in continuous-time

Theorem 10.3 *(i) The n-dimensional continuous-time system $[A, \, \cdot \, , C]$ is observable if and only if the matrix N_n has rank n, or (ii) equivalently, if and only if*

$$H(t) = \int_0^t e^{A^\top s} C^\top C e^{As} \, ds$$

is positive definite for all $t > 0$. (iii) If the system is observable then the determination of $x(0)$ can be written

$$x(0) = H(t)^{-1} \int_0^t e^{A^\top s} C^\top \tilde{y}(s) \, ds,$$

where

$$\tilde{y}(t) = y(t) - \int_0^t CA^{t-s} Bu(s) \, ds.$$

38

10.3   Examples

Example. Observation of population

Consider two populations whose sizes are changing according to the equations

$$\dot{x}_1 = \lambda_1 x_1, \qquad \dot{x}_2 = \lambda_2 x_2.$$

Suppose we observe $x_1 + x_2$, so

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \quad C = (1 \quad 1), \quad N_2 = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix}.$$

and so the individual populations are observable if $\lambda_1 \neq \lambda_2$.

Example. Radioactive decay

Suppose $j = 1, 2, 3$ and decay is from 1 to 2 at rate α, from 1 to 3 at rate β and from 2 to 3 at rate γ. We observe only the accumulation in state 3. Then $\dot{x} = Ax$, where

$$A = \begin{pmatrix} -\alpha - \beta & 0 & 0 \\ \alpha & -\gamma & 0 \\ \beta & \gamma & 0 \end{pmatrix}, \quad C = (0 \quad 0 \quad 1),$$

$$N_3 = \begin{pmatrix} C \\ CA \\ CA^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ \beta & \gamma & 0 \\ \alpha\gamma - \beta(\alpha + \beta) & -\gamma^2 & 0 \end{pmatrix}.$$

The determinant equals $\gamma(\alpha + \beta)(\gamma - \beta)$, so we require $\gamma > 0$, $\alpha + \beta > 0$, $\beta \neq \gamma$ for a nonzero determinant and observability.

Example. Satellite

Recall the linearised equation $\dot{x} = Ax$, for perturbations of the orbit of a satellite, (here taking $\rho = 1$), where

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} r - \rho \\ \dot{r} \\ \theta - \omega t \\ \dot{\theta} - \omega \end{pmatrix} \quad A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega & 0 & 0 \end{pmatrix}.$$

By taking $C = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$ we see that we can see that system is observable on the basis of angle measurements alone but, not observable for $\tilde{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$, i.e., on the basis of radius movements alone.

$$N_4 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega & 0 & 0 \\ -6\omega^3 & 0 & 0 & -4\omega^2 \end{bmatrix} \quad \tilde{N}_4 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega \\ 0 & -\omega^2 & 0 & 0 \\ -3\omega^4 & 0 & 0 & -2\omega^3 \end{bmatrix}$$

10.4 Imperfect state observation with noise

The full LQG model, whose description has been deferred until now, assumes linear dynamics, quadratic costs and Gaussian noise. Imperfect observation is the most important point. The model is

$$x_t = Ax_{t-1} + Bu_{t-1} + \epsilon_t, \tag{10.5}$$
$$y_t = Cx_{t-1} + \eta_t, \tag{10.6}$$

where ϵ_t is process noise, y_t is the observation at time t and η_t is the observation noise. The state observations are degraded in that we observe only Cx_{t-1}. Assume

$$\mathrm{cov}\begin{pmatrix} \epsilon \\ \eta \end{pmatrix} = E \begin{pmatrix} \epsilon \\ \eta \end{pmatrix}\begin{pmatrix} \epsilon \\ \eta \end{pmatrix}^\top = \begin{pmatrix} N & L \\ L^\top & M \end{pmatrix}$$

and that $x_0 \sim N(\hat{x}_0, V_0)$. Let $W_t = (Y_t, U_{t-1}) = (y_1, \ldots, y_t; u_0, \ldots, u_{t-1})$ denote the observed history up to time t. Of course we assume that t, A, B, C, N, L, M, $\hat{x}_0$ and V_0 are also known; W_t denotes what might be different if the process were rerun. In the next lecture we turn to the question of estimating x from y.

We consider the issues of state estimation and optimal control and shall show:

(i) $\hat{x}_t$ can be calculated recursively from the Kalman filter (a linear operator):

$$\hat{x}_t = A\hat{x}_{t-1} + Bu_{t-1} + H_t(y_t - C\hat{x}_{t-1}),$$

which is like the plant equation except that the noise is given by an **innovation process**, $\tilde{y}_t = y_t - C\hat{x}_y$, rather than the white noise.

(ii) If there is full information (i.e., $y_t = x_t$) and the optimal control is $u_t = K_t x_t$, then without full information the optimal control is $u_t = K_t \hat{x}_t$, where $\hat{x}_t$ is the best linear least squares estimate of x_t based on the information (Y_t, U_{t-1}) at time t.

Many of the ideas we encounter in this analysis are unrelated to the special state structure and are therefore worth noting as general observations about control with imperfect information.

11 Kalman Filtering and Certainty Equivalence

We presents the important concepts of the Kalman filter, certainty-equivalence and the separation principle.

11.1 Preliminaries

Lemma 11.1 *Suppose x and y are jointly normal with zero means and covariance matrix*

$$\operatorname{cov}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix}.$$

Then the distribution of x conditional on y is Gaussian, with

$$E(x \mid y) = V_{xy}V_{yy}^{-1}y, \tag{11.1}$$

and

$$\operatorname{cov}(x \mid y) = V_{xx} - V_{xy}V_{yy}^{-1}V_{yx}. \tag{11.2}$$

Proof. Both y and $x - V_{xy}V_{yy}^{-1}y$ are linear functions of x and y and therefore they are Gaussian. From $E\left[(x - V_{xy}V_{yy}^{-1}y)y^\top\right] = 0$ it follows that they are uncorrelated and this implies they are independent. Hence the distribution of $x - V_{xy}V_{yy}^{-1}y$ conditional on y is identical with its unconditional distribution, and this is Gaussian with zero mean and the covariance matrix given by (11.2) ■

The estimate of x in terms of y defined as $\hat{x} = Hy = V_{xy}V_{yy}^{-1}y$ is known as the **linear least squares estimate** of x in terms of y. Even without the assumption that x and y are jointly normal, this linear function of y has a smaller covariance matrix than any other unbiased estimate for x that is a linear function of y. In the Gaussian case, it is also the maximum likelihood estimator.

11.2 The Kalman filter

Let us make the LQG and state-structure assumptions of Section 10.4.

$$x_t = Ax_{t-1} + Bu_{t-1} + \epsilon_t, \tag{11.3}$$
$$y_t = Cx_{t-1} + \eta_t, \tag{11.4}$$

Notice that both x_t and y_t can be written as a linear functions of the unknown noise and the known values of $u_0, \ldots, u_{t-1}$. Thus the distribution of x_t conditional on $W_t = (Y_t, U_{t-1})$ must be normal, with some mean $\hat{x}_t$ and covariance matrix V_t. The following theorem describes recursive updating relations for these two quantities.

Theorem 11.2 (The Kalman filter) *Suppose that conditional on W_0, the initial state x_0 is distributed $N(\hat{x}_0, V_0)$ and the state and observations obey the recursions of*

the LQG model (11.3)–(11.4). Then conditional on W_t, the current state is distributed $N(\hat{x}_t, V_t)$. The conditional mean and variance obey the updating recursions

$$\hat{x}_t = A\hat{x}_{t-1} + Bu_{t-1} + H_t(y_t - C\hat{x}_{t-1}), \qquad (11.5)$$

$$V_t = N + AV_{t-1}A^\top - (L + AV_{t-1}C^\top)(M + CV_{t-1}C^\top)^{-1}(L^\top + CV_{t-1}A^\top), \quad (11.6)$$

where

$$H_t = (L + AV_{t-1}C^\top)(M + CV_{t-1}C^\top)^{-1}. \qquad (11.7)$$

Proof. The proof is by induction on t. Consider the moment when u_{t-1} has been determined but y_t has not yet observed. The distribution of (x_t, y_t) conditional on (W_{t-1}, u_{t-1}) is jointly normal with means

$$E(x_t \mid W_{t-1}, u_{t-1}) = A\hat{x}_{t-1} + Bu_{t-1},$$
$$E(y_t \mid W_{t-1}, u_{t-1}) = C\hat{x}_{t-1}.$$

Let $\Delta_{t-1} = \hat{x}_{t-1} - x_{t-1}$, which by an inductive hypothesis is $N(0, V_{t-1})$. Consider the **innovations**

$$\xi_t = x_t - E(x_t \mid W_{t-1}, u_{t-1}) = x_t - (A\hat{x}_{t-1} + Bu_{t-1}) = \epsilon_t - A\Delta_{t-1},$$
$$\zeta_t = y_t - E(y_t \mid W_{t-1}, u_{t-1}) = y_t - C\hat{x}_{t-1} = \eta_t - C\Delta_{t-1}.$$

Conditional on (W_{t-1}, u_{t-1}), these quantities are normally distributed with zero means and covariance matrix

$$\mathrm{cov}\begin{bmatrix} \epsilon_t - A\Delta_{t-1} \\ \eta_t - C\Delta_{t-1} \end{bmatrix} = \begin{bmatrix} N + AV_{t-1}A^\top & L + AV_{t-1}C^\top \\ L^\top + CV_{t-1}A^\top & M + CV_{t-1}C^\top \end{bmatrix} = \begin{bmatrix} V_{\xi\xi} & V_{\xi\zeta} \\ V_{\zeta\xi} & V_{\zeta\zeta} \end{bmatrix}.$$

Thus it follows from Lemma 11.1 that the distribution of ξ_t conditional on knowing $(W_{t-1}, u_{t-1}, \zeta_t)$, (which is equivalent to knowing W_t), is normal with mean $V_{\xi\zeta}V_{\zeta\zeta}^{-1}\zeta_t$ and covariance matrix $V_{\xi\xi} - V_{\xi\zeta}V_{\zeta\zeta}^{-1}V_{\zeta\xi}$. These give (11.5)–(11.7). ∎

11.3 Certainty equivalence

We say that a quantity a is *policy-independent* if $E_\pi(a \mid W_0)$ is independent of π.

Theorem 11.3 *Suppose LQG model assumptions hold. Then (i)*

$$F(W_t) = \hat{x}_t^\top \Pi_t \hat{x}_t + \cdots \qquad (11.8)$$

where $\hat{x}_t$ is the linear least squares estimate of x_t whose evolution is determined by the Kalman filter in Theorem 11.2 and '$+\cdots$' indicates terms that are policy independent; (ii) the optimal control is given by

$$u_t = K_t\hat{x}_t,$$

where Π_t and K_t are the same matrices as in the full information case of Theorem 7.2.

It is important to grasp the remarkable fact that (ii) asserts: *the optimal control u_t is exactly the same as it would be if all unknowns were known and took values equal to their linear least square estimates (equivalently, their conditional means) based upon observations up to time t.* This is the idea known as **certainty equivalence**. As we have seen in the previous section, the distribution of the estimation error $\hat{x}_t - x_t$ does not depend on U_{t-1}. The fact that the problems of optimal estimation and optimal control can be decoupled in this way is known as the **separation principle**.

Proof. The proof is by backward induction. Suppose (11.8) holds at t. Recall that

$$\hat{x}_t = A\hat{x}_{t-1} + Bu_{t-1} + H_t\zeta_t\,, \qquad \Delta_{t-1} = \hat{x}_{t-1} - x_{t-1}.$$

Then with a quadratic cost of the form $c(x, u) = x^\top R x + 2u^\top S x + u^\top Q u$, we have

$$F(W_{t-1}) = \min_{u_{t-1}} E\left[c(x_{t-1}, u_{t-1}) + \hat{x}_t \Pi_t \hat{x}_t + \cdots \mid W_{t-1}, u_{t-1}\right]$$

$$= \min_{u_{t-1}} E\left[c(\hat{x}_{t-1} - \Delta_{t-1}, u_{t-1})\right.$$

$$\left. + (A\hat{x}_{t-1} + Bu_{t-1} + H_t\zeta_t)^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1} + H_t\zeta_t) \mid W_{t-1}, u_{t-1}\right]$$

$$= \min_{u_{t-1}} \left[c(\hat{x}_{t-1}, u_{t-1}) + (A\hat{x}_{t-1} + Bu_{t-1})^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1})\right] + \cdots \,,$$

where we use the fact that conditional on W_{t-1}, u_{t-1}, both Δ_{t-1} and ζ_t have zero means and are policy independent. This ensures that when we expand the quadratics in powers of Δ_{t-1} and $H_t\zeta_t$ the expected value of the linear terms in these quantities are zero and the expected value of the quadratic terms (represented by $+\cdots$) are policy independent. ∎

11.4 Example: inertialess rocket with noisy position sensing

Consider the scalar case of controlling the position of a rocket by inertialess control of its velocity but in the presence of imperfect position sensing.

$$x_t = x_{t-1} + u_{t-1}, \qquad y_t = x_t + \eta_t\,,$$

where η_t is white noise with variance 1. Suppose it is desired to minimize

$$E\left[\sum_{t=0}^{h-1} u_t^2 + D x_h^2\right].$$

Notice that the observational relation differs from the usual model of $y_t = C x_{t-1} + \eta_t$. To derive a Kalman filter formulae for this variation we argue inductively from scratch. Suppose $\hat{x}_{t-1} - x_{t-1} \sim N(0, V_{t-1})$. Consider a linear estimate of x_t,

$$\hat{x}_t = \hat{x}_{t-1} + u_{t-1} + H_t(y_t - \hat{x}_{t-1} - u_{t-1})\,.$$

(The relevant innovation process is now $\tilde{y}_t = y_t - \hat{x}_{t-1} - u_{t-1}$.) Subtracting the plant equation and substituting for x_t and y_t gives

$$\Delta_t = \Delta_{t-1} + H_t(\eta_t - \Delta_{t-1}).$$

The variance of Δ_t is therefore

$$\text{var}\,\Delta_t = V_{t-1} - 2H_t V_{t-1} + H_t^2(1 + V_{t-1}).$$

Minimizing this with respect to H_t gives $H_t = V_{t-1}(1 + V_{t-1})^{-1}$, so the variance in the least squares estimate of x_t obeys the recursion,

$$V_t = V_{t-1} - V_{t-1}^2(1 + V_{t-1})^{-1} = V_{t-1}/(1 + V_{t-1}).$$

Hence

$$V_t^{-1} = V_{t-1}^{-1} + 1 = \cdots = V_0^{-1} + t.$$

If there is complete lack of information at the start, then $V_0^{-1} = 0$, $V_t = 1/t$ and

$$\hat{x}_t = \hat{x}_{t-1} + u_{t-1} + \frac{V_{t-1}(y_t - \hat{x}_{t-1} - u_{t-1})}{1 + V_{t-1}} = \frac{(t-1)(\hat{x}_{t-1} + u_{t-1}) + y_t}{t}.$$

As far at the optimal control is concerned, suppose an inductive hypothesis that $F(W_t) = \hat{x}_t^2 \Pi_t + \cdots$, where '$\cdots$' denotes policy independent terms. Then

$$F(W_{t-1}) = \inf_u \left\{ u^2 + E[\hat{x}_{t-1} + u + H_t(y_t - \hat{x}_{t-1} - u)]^2 \Pi_t + \cdots \right\}$$
$$= \inf_u \left\{ u^2 + (\hat{x}_{t-1} + u)^2 \Pi_t + E[H_t(\eta_t - \Delta_{t-1})]^2 \Pi_t + \cdots \right\}.$$

Minimizing over u we obtain the usual Riccati recursion of

$$\Pi_{t-1} = \Pi_t - \Pi_t^2/(1 + \Pi_t) = \Pi_t/(1 + \Pi_t).$$

Hence $\Pi_t = D/(1 + D(h-t))$ and the optimal control is the certainty equivalence control $u_t = -D\hat{x}_t/(1 + D(h-t))$. This is the same control as in the deterministic case, but with x_t replaced by $\hat{x}_t$.

12 Dynamic Programming in Continuous Time

We consider deterministic dynamic programming in continuous time.

12.1 The optimality equation

In continuous time the plant equation is,

$$\dot{x} = a(x, u, t).$$

Let us consider a discounted cost of

$$\mathbf{C} = \int_0^T e^{-\alpha t} c(x, u, t)\, dt + e^{-\alpha T} \mathbf{C}(x(T), T).$$

The discount factor over δ is $e^{-\alpha\delta} = 1 - \alpha\delta + o(\delta)$. So the optimality equation is,

$$F(x, t) = \inf_u \left[c(x, u, t)\delta + e^{-\alpha\delta} F(x + a(x, u, t)\delta, t + \delta) + o(\delta) \right].$$

By considering the term that multiplies δ in the Taylor series expansion we obtain,

$$\inf_u \left[c(x, u, t) - \alpha F + \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} a(x, u, t) \right] = 0, \quad t < T, \tag{12.1}$$

with $F(x, T) = \mathbf{C}(x, T)$. In the undiscounted case, we simply put $\alpha = 0$.

The following theorem states that if we can find a policy whose value function satisfies the DP equation then that policy is optimal.

Theorem 12.1 *Suppose a policy π, using a control u, has a value function F which satisfies the DP equation (12.1) for all values of x and t. Then π is optimal.*

Proof.   Consider any other policy, using control v, say. Then along the trajectory defined by $\dot{x} = a(x, v, t)$ we have

$$-\frac{d}{dt} e^{-\alpha t} F(x, t) = e^{-\alpha t} \left[c(x, v, t) - \left(c(x, v, t) - \alpha F + \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} a(x, v, t) \right) \right]$$

$$\leq e^{-\alpha t} c(x, v, t).$$

Integrating this inequality along the v path, from $x(0)$ to $x(T)$, gives

$$F(x(0), 0) - e^{-\alpha T} \mathbf{C}(x(T), T) \leq \int_{t=0}^{T} e^{-\alpha t} c(x, v, t)\, dt.$$

Thus the v path incurs a cost of at least $F(x(0), 0)$, and hence π is optimal. ∎

12.2 Example: LQ regulation

The undiscounted continuous time DP equation for the LQ regulation problem is

$$0 = \inf_u \left[x^\top R x + u^\top Q u + F_t + F_x^\top (Ax + Bu) \right] .$$

Suppose we try a solution of the form $F(x, t) = x^\top \Pi(t) x$, where $\Pi(t)$ is a symmetric matrix. Then $F_x = 2\Pi(t)x$ and the optimizing u is $u = -\frac{1}{2} Q^{-1} B^\top F_x = -Q^{-1} B^\top \Pi(t) x$. Therefore the DP equation is satisfied with this u if

$$0 = x^\top \left[R + \Pi A + A^\top \Pi - \Pi B Q^{-1} B^\top \Pi + \frac{d\Pi}{dt} \right] x ,$$

where we use the fact that $2x^\top \Pi A x = x^\top \Pi A x + x^\top A^\top \Pi x$. Hence we have a solution to the DP equation if $\Pi(t)$ satisfies the Riccati differential equation of Section 7.4.

12.3 Example: estate planning

A man is considering his lifetime plan of investment and expenditure. He has an initial level of savings $x(0)$ and no other income other than that which he obtains from investment at a fixed interest rate. His total capital is therefore governed by the equation

$$\dot{x}(t) = \beta x(t) - u(t) ,$$

where $\beta > 0$ and u is his rate of expenditure. He wishes to maximize

$$\int_0^T e^{-\alpha t} \sqrt{u(t)} \, dt,$$

for a given T. Find his optimal policy.

Solution. The optimality equation is

$$0 = \sup_u \left[\sqrt{u} - \alpha F + \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} (\beta x - u) \right] .$$

Suppose we try a solution of the form $F(x, t) = f(t)\sqrt{x}$. For this to work we need

$$0 = \sup_u \left[\sqrt{u} - \alpha f \sqrt{x} + f' \sqrt{x} + \frac{f}{2\sqrt{x}} (\beta x - u) \right] .$$

By $d[\ ]/du = 0$, the optimizing u is $u = x/f^2$ and the optimized value is

$$(\sqrt{x}/f) \left[\frac{1}{2} - (\alpha - \tfrac{1}{2}\beta) f^2 + f f' \right] . \tag{12.2}$$

Therefore we shall have a solution if we choose f to make the bracketed term in (12.2) equal to 0. We have the boundary condition $F(x, T) = 0$, which imposes $f(T) = 0$. Thus we find

$$f(t)^2 = \frac{1 - e^{-(2\alpha - \beta)(T-t)}}{2\alpha - \beta} .$$

46

By Theorem 12.1 we have established the form of optimal control, which in closed loop form is $u = x/f^2$. ∎

12.4 Example: harvesting

A fish population of size x obeys the plant equation,

$$\dot{x} = a(x, u) = \begin{cases} a(x) - u & x > 0, \\ a(x) & x = 0. \end{cases}$$

The function $a(x)$ reflects the facts that the population can grow when it is small, but is subject to environmental limitations when it is large. It is desired to maximize the discounted total harvest $\int_0^T u e^{-\alpha t}\, dt$.

Solution. The DP equation (with discounting) is

$$\sup_u \left[u - \alpha F + \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x}[a(x) - u] \right] = 0, \quad t < T.$$

Hence u occurs linearly with the maximization and so we have a bang-bang optimal control of the form

$$u = \begin{bmatrix} 0 \\ \text{indeterminate} \\ u_{\max} \end{bmatrix} \text{ for } F_x \begin{bmatrix} > \\ = \\ < \end{bmatrix} 1,$$

where $u_{\max}$ is the largest practicable fishing rate.

Suppose $F(x, t) \to F(x)$ as $T \to \infty$, and $\partial F/\partial t \to 0$. Then

$$\sup_u \left[u - \alpha F + \frac{\partial F}{\partial x}[a(x) - u] \right] = 0. \tag{12.3}$$

If $F(x)$ is concave,

$$u = \begin{bmatrix} 0 \\ \text{indeterminate, but effectively } a(\bar{x}) \\ u_{\max} \end{bmatrix} \text{ for } x \begin{bmatrix} < \\ = \\ > \end{bmatrix} \bar{x}.$$

Clearly, $\bar{x}$ is the operating point. We suppose

$$\dot{x} = \begin{cases} a(x) > 0, & x < \bar{x} \\ a(x) - u_{\max} < 0, & x > \bar{x}. \end{cases}$$

We say that there is **chattering** about the point $\bar{x}$, in the sense that u switches between its maximum and minimum values either side of $\bar{x}$, effectively taking the value $a(\bar{x})$ at $\bar{x}$. To determine $\bar{x}$ we note that

$$F(\bar{x}) = \int_0^\infty e^{-\alpha t} a(\bar{x}) dt = a(\bar{x})/\alpha. \tag{12.4}$$

47

So from (12.3) and (12.4) we have

$$F_x(x) = \frac{\alpha F(x) - u(x)}{a(x) - u(x)} \to 1 \text{ as } x \nearrow \bar{x} \text{ or } x \searrow \bar{x}. \tag{12.5}$$

Assuming F is concave, F_{xx} is negative if it exists. So

$$F_{xx} = \frac{\alpha F_x}{a(x) - u} - \left(\frac{\alpha F - u}{a(x) - u}\right)\left(\frac{a'(x)}{a(x) - u}\right)$$

$$= \left(\frac{\alpha F - u}{a(x) - u}\right)\left(\frac{\alpha - a'(x)}{a(x) - u}\right)$$

$$\simeq \frac{\alpha - a'(x)}{a(x) - u(x)}$$

where the last line follows because (12.5) holds in a neighbourhood of $\bar{x}$. It is required that F_{xx} be negative. But the denominator changes sign at $\bar{x}$, so the numerator must do so also, and therefore we must have $a'(\bar{x}) = \alpha$; this determines $\bar{x}$.
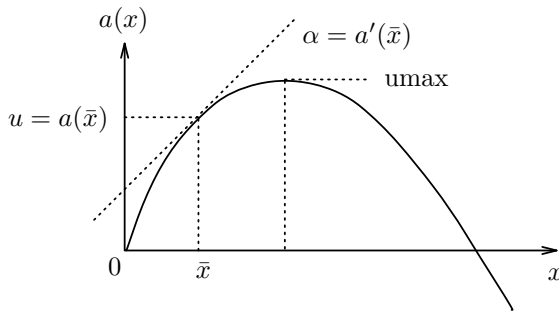

Figure 2: Growth rate $a(x)$ subject to environment pressures

There is a sacrifice of long term yield for immediate return if $\alpha > 0$. If the initial population is greater than $\bar{x}$ then the optimal policy is to overfish at $u_{\max}$ until we reach the new $\bar{x}$ and then fish at rate $u = a(\bar{x})$. As $\alpha \nearrow a'(0)$, $\bar{x} \searrow 0$, i.e., for sufficiently large α is becomes optimal to wipe out the fish population. ■

48

13 Pontryagin's Maximum Principle

We explain Pontryagin's maximum principle, derive it and give examples of its use.

13.1 Heuristic derivation

Pontryagin's maximum principle (PMP) states *a necessary condition that must hold on an optimal trajectory.* It is a calculation for a *fixed* initial value of the state, $x(0)$. In comparison, the DP approach is a calculation for a general initial value of the state. PMP can be used as both a computational and analytic technique (and in the second case can solve the problem for general initial value.)

Consider first a time-invariant formulation, with plant equation $\dot{x} = a(x, u)$, instantaneous cost $c(x, u)$, stopping set $\mathcal{S}$ and terminal cost $\mathbf{K}(x)$. The value function $F(x)$ obeys the DP equation (without discounting)

$$\inf_u \left[c(x, u) + \frac{\partial F}{\partial x} a(x, u) \right] = 0 \,, \tag{13.1}$$

outside $\mathcal{S}$, with terminal condition

$$F(x) = \mathbf{K}(x) \,, \quad x \in \mathcal{S} \,. \tag{13.2}$$

Define the **adjoint variable**

$$\lambda = -F_x \tag{13.3}$$

This is a column n-vector, and is to be regarded as a function of time on the path. The proof that F_x exists in the required sense is actually a tricky technical matter. Also define the **Hamiltonian**

$$H(x, u, \lambda) = \lambda^\top a(x, u) - c(x, u) \,, \tag{13.4}$$

a scalar, defined at each point of the path as a function of the current x, u and λ.)

Theorem 13.1 (PMP) *Suppose $u(t)$ and $x(t)$ represent the optimal control and state trajectory. Then there exists an adjoint trajectory $\lambda(t)$ such that together $u(t)$, $x(t)$ and $\lambda(t)$ satisfy*

$$\dot{x} = H_\lambda, \qquad [= a(x, u)] \tag{13.5}$$

$$\dot{\lambda} = -H_x, \qquad [= -\lambda^\top a_x + c_x] \tag{13.6}$$

and for all t, $0 \le t \le T$, and all feasible controls v,

$$H(x(t), v, \lambda(t)) \le H(x(t), u(t), \lambda(t)) \,, \tag{13.7}$$

i.e., the optimal control $u(t)$ is the value of v maximizing $H((x(t), v, \lambda(t))$.

'Proof.' Our heuristic proof is based upon the DP equation; this is the most direct and enlightening way to derive conclusions that may be expected to hold in general.

Assertion (13.5) is immediate, and (13.7) follows from the fact that the minimizing value of u in (13.1) is optimal. We can write (13.1) in incremental form as

$$F(x) = \inf_u [c(x, u)\delta + F(x + a(x, u)\delta)] + o(\delta).$$

Using the chain rule to differentiate with respect to x_i yields

$$-\lambda_i(t) = c_{x_i}\delta - (1 + a_{x_i}\delta)^\top \lambda_i(t + \delta) + o(\delta)$$

whence (13.6) follows. ∎

Notice that (13.5) and (13.6) each give n equations. Condition (13.7) gives a further m equations (since it requires stationarity with respect to variation of the m components of u.) So in principle these equations, if nonsingular, are sufficient to determine the $2n + m$ functions $u(t)$, $x(t)$ and $\lambda(t)$.

One can make other assertions, including specification of end-conditions (the so-called **transversality conditions**.)

Theorem 13.2 *(i) $H = 0$ on the optimal path. (ii) The sole initial condition is specification of the initial x. The terminal condition*

$$(\lambda + \mathbf{K}_x)^\top \sigma = 0 \tag{13.8}$$

holds at the terminal x for all σ such that $x + \epsilon\sigma$ is within $o(\epsilon)$ of the termination point of a possible optimal trajectory for all sufficiently small positive ϵ.

'Proof.' Assertion (i) follows from (13.1), and the first assertion of (ii) is evident. We have the terminal condition (13.2), from whence it follows that $(F_x - \mathbf{K}_x)\sigma = 0$ for all x, σ such that x and $x + \epsilon\sigma$ lie in $\mathcal{S}$ for all small enough positive ϵ. However, we are only interested in points where an optimal trajectory makes it first entry to $\mathcal{S}$ and at these points (13.3) holds. Thus we must have (13.8). ∎

13.2 Example: skateboarding

Consider the problem of accelerating a skateboard in such a way as to maximize the total distance traveled in a given time, minus the effort expended. The system is

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}u, \tag{13.9}$$

where $x_2 = \dot{x}_1$, $x_1(0) = x_2(0) = 0$, and

$$\mathbf{C} = \tfrac{1}{2}\int_0^T u(t)^2\, dt - x_1(T).$$

50

Here $K(x) = -x_1(T)$ and the Hamiltonian is

$$H(x, u, \lambda) = \lambda_1 x_2 + \lambda_2 u - \tfrac{1}{2} u^2,$$

which is maximized at $u(t) = \lambda_2(t)$. Now $\dot{\lambda}_i = -\partial H/\partial x_i$ gives

$$\dot{\lambda}_1 = 0, \qquad \dot{\lambda}_2 = -\lambda_1. \tag{13.10}$$

In the terminal condition, $(\lambda + \mathbf{K}_x)^\top \sigma = 0$, σ is arbitrary and so we also have

$$\lambda_1(T) - 1 = 0, \qquad \lambda_2(T) = 0.$$

Thus the solution must be $\lambda_1(t) = 1$ and $\lambda_2(t) = T - t$. Hence the optimal applied force is $u(t) = T - t$, which decreases linearly with time and reaches zero at T.

13.3 Example: Bush problem

The problem is now to bring the skateboard to rest at given position in minimal time. This is a problem for which the classic calculus of variations fails. It is one of the earliest applications of the maximum principle. The plant equation is (13.9) as before, but now we impose the constraint $|u| \leq 1$. The problem is to reach $x = (0,0)$ in minimal time, so

$$\mathbf{C} = \int_0^T 1 \, dt \, .$$

The Hamiltonian is

$$H = \lambda_1 x_2 + \lambda_2 u - 1 \, ,$$

which is maximized by $u = \text{sign}(\lambda_2)$. The adjoint variables satisfy (13.10) as above, so if at termination $\lambda_1 = \alpha$, $\lambda_2 = \beta$, then in terms of time to go,

$$\lambda_1 = \alpha, \qquad \lambda_2 = \beta + \alpha s,$$

and these reveal the form of the solution: there is at most one change of sign of λ_2 on the optimal path; u is maximal in one direction and then possibly maximal in the other.

Appealing to the fact that $H = 0$ at termination (when $x_2 = 0$), we conclude that $|\beta| = 1$. We now consider the case $\beta = 1$. The case $\beta = -1$ is similar.

If $\beta = 1$, $\alpha \geq 0$ then $\lambda_2 = 1 + \alpha s \geq 0$ for all $s \geq 0$ and

$$u = 1, \qquad x_2 = -s, \qquad x_1 = s^2/2.$$

In this case the optimal trajectory lies on the parabola $x_1 = x_2^2/2$, $x_1 \geq 0, x_2 \leq 0$. This is half of the **switching locus** $x_1 = \pm x_2^2/2$.

If $\beta = 1$, $\alpha < 0$ then $u = -1$ or $u = 1$ as the time to go is greater or less than $s_0 = 1/|\alpha|$. In this case,

$$
\begin{aligned}
u &= -1, & x_2 &= (s - 2s_0), & x_1 &= 2s_0 s - \tfrac{1}{2}s^2 - s_0^2, & s &\geq s_0, \\
u &= 1, & x_2 &= -s, & x_1 &= \tfrac{1}{2}s^2, & s &\leq s_0.
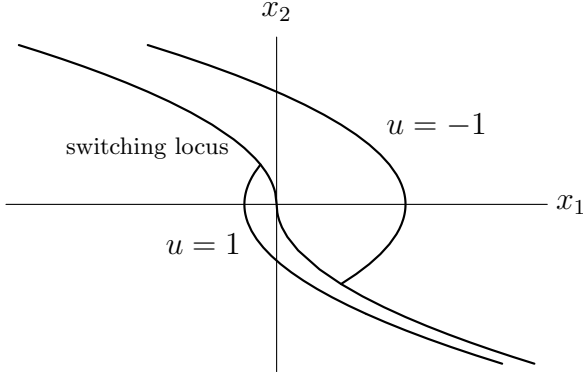\end{aligned}
$$

Figure 3: Optimal trajectories for the Bush problem

The control rule expressed as a function of s is open-loop, but in terms of (x_1, x_2) and the switching locus, it is closed-loop.

Notice that the path is sensitive to the initial conditions, in that the optimal path is very different for two points just either side of the switching locus.

13.4 Connection with Lagrangian multipliers

An alternative way to understand the maximum principle is to think of λ as a Lagrangian multiplier associated with the constraint $\dot{x} = a(x, u)$. Consider the Lagrangian form

$$L = \int_0^T \left[-c - \lambda^\top (\dot{x} - a) \right] dt - \mathbf{K}(x(T)),$$

to be maximized with respect to the (x, u, λ) path. Here $x(t)$ first enters a stopping set $\mathcal{S}$ at $t = T$. We integrate $\lambda^\top \dot{x}$ by parts to obtain

$$L = -\lambda(T)^\top x(T) + \lambda(0)^\top x(0) + \int_0^T \left[\dot{\lambda}^\top x + \lambda^\top a - c \right] dt - \mathbf{K}(x(T)).$$

The integrand must be stationary with respect to $x(t)$ and hence $\dot{\lambda} = -H_x$. The expression must also be stationary with respect to $\epsilon > 0$, $x(T) + \epsilon\sigma \in \mathcal{S}$ and hence $(\lambda(T) + \mathbf{K}_x(x(T)))^\top \sigma = 0$, i.e., (13.6) and (13.8). It is good to have this alternative view, but the treatment is less immediate and less easy to rigorise.

14 Applications of the Maximum Principle

We discuss the terminal conditions of the maximum principle and further examples of its use. The arguments are typical of those used to synthesise a solution to an optimal control problem by use of the maximum principle.

14.1 Problems with terminal conditions

Suppose a, c, $\mathcal{S}$ and $\mathbf{K}$ are alle t-dependent. The DP equation for $F(x, t)$ is now be

$$\inf_u [c + F_t + F_x a] = F_t - \sup_u [\lambda^\top a - c] = 0 , \tag{14.1}$$

outside a stopping set $\mathcal{S}$, with $F(x, t) = \mathbf{K}(x, t)$ for (x, t) in $\mathcal{S}$. However, we can reduce this to a formally time-invariant case by augmenting the state variable x by the variable t. We then have the augmented variables

$$x \to \begin{bmatrix} x \\ t \end{bmatrix} \quad a \to \begin{bmatrix} a \\ 1 \end{bmatrix} \quad \lambda \to \begin{bmatrix} \lambda \\ \lambda_0 \end{bmatrix} .$$

We keep the same definition (13.4) as before, that $H = \lambda^\top a - c$, and take $\lambda_0 = -F_t$. It now follows from (14.1) that on the optimal trajectory

$$H(x, u, \lambda) \text{ is maximized to } -\lambda_0 .$$

Theorem 13.1 still holds, as can be verified. However, to (13.6) we can now add

$$\dot{\lambda}_0 = -H_t = c_t - \lambda a_t . \tag{14.2}$$

and transversality condition

$$(\lambda + \mathbf{K}_x)^\top \sigma + (\lambda_0 + \mathbf{K}_t) \tau = 0 , \tag{14.3}$$

which must hold at the termination point (x, t) if $(x + \epsilon \sigma, t + \epsilon \tau)$ is within $o(\epsilon)$ of the termination point of an optimal trajectory for all small enough positive ϵ. We can now understand what to do with various types of terminal condition.

If the stopping rule specifies only a **fixed terminal time** T then τ must be zero and σ is unconstrained, so that (14.3) becomes $\lambda(T) = -\mathbf{K}_x$. The skateboard problem in Section 13.2 is such an example.

If there is a **free terminal time** then τ is unconstrained and so (14.3) gives $-\lambda_0(T) = \mathbf{K}_T$. An example of this case appears in Section 14.2 below.

If the system is time-homogeneous, in that a and c are independent of t, but the terminal cost $\mathbf{K}(x, T)$ depends on T, then (14.2) implies that λ_0 is constant and so the maximized value of H is constant on the optimal orbit. The Bush problem in Section 13.3 can be treated this way. We take $\mathbf{K}(x, T) = T$ and deduce from the transversality condition that since τ is unconstrained, $\lambda_0 = -\mathbf{K}_T = -1$. Thus $H = \lambda_1 x_2 + \lambda_2 u$ is maximized to 1 at all points of the optimal trajectory.

14.2 Example: monopolist

Miss Prout holds the entire remaining stock of Cambridge elderberry wine for the vintage year 1959. If she releases it at rate u (in continuous time) she realises a unit price $p(u) = (1 - u/2)$, for $0 \le u \le 2$ and $p(u) = 0$ for $u \ge 2$. She holds an amount x at time 0 and wishes to release it in a way that maximizes her total discounted return, $\int_0^T e^{-\alpha t} u p(u)\, dt$, (where T is unconstrained.)

Solution. The plant equation is $\dot{x} = -u$ and the Hamiltonian is

$$H(x, u, \lambda) = e^{-\alpha t} u p(u) - \lambda u = e^{-\alpha t} u(1 - u/2) - \lambda u\,.$$

Note that $\mathbf{K} = 0$. Maximizing with respect to u and using $\dot{\lambda} = -H_x$ gives

$$u = 1 - \lambda e^{\alpha t}\,, \qquad \dot{\lambda} = 0\,, \qquad t \ge 0\,,$$

so λ is constant. The terminal time is unconstrained so the transversality condition gives $\lambda_0(T) = -\mathbf{K}_T = 0$. Therefore, since H is maximized to $-\lambda_0(T) = 0$ at T, we have $u(T) = 0$ and hence

$$\lambda = e^{-\alpha T}\,, \qquad u = 1 - e^{-\alpha(T-t)}\,, \quad t \le T\,,$$

where T is then determined as the solution to

$$x = \int_0^T u\, dt = T - \left(1 - e^{-\alpha T}\right)/\alpha\,.$$

Thus u is implicitly a function of x, through T. The optimal value function is

$$F(x) = \int_0^T (u - u^2/2) e^{-\alpha t}\, dt = \frac{1}{2} \int_0^T \left(e^{\alpha t} - e^{\alpha t - 2\alpha T}\right) dt = \frac{\left(1 - e^{\alpha T}\right)^2}{2\alpha}\,.$$

■

14.3 Example: insects as optimizers

A colony of insects consists of workers and queens, of numbers $w(t)$ and $q(t)$ at time t. If a time-dependent proportion $u(t)$ of the colony's effort is put into producing workers, ($0 \le u(t) \le 1$, then w, q obey the equations

$$\dot{w} = auw - bw\,, \quad \dot{q} = c(1 - u)w\,,$$

where a, b, c are constants, with $a > b$. The function u is to be chosen to maximize the number of queens at the end of the season. Show that the optimal policy is to produce only workers up to some moment, and produce only queens thereafter.

Solution. The Hamiltonian is

$$H = \lambda_1(auw - bw) + \lambda_2 c(1 - u)w\,.$$

54

The adjoint equations and transversality conditions (with $\mathbf{K} = -q$) give

$$
\begin{aligned}
-\dot{\lambda}_0 &= H_t = 0 \\
-\dot{\lambda}_1 &= H_w = \lambda_1(au - b) + \lambda_2 c(1 - u)\,, \\
-\dot{\lambda}_2 &= H_q = 0
\end{aligned}
\qquad
\begin{aligned}
\lambda_1(T) &= -K_w = 0 \\
\lambda_2(T) &= -K_q = 1
\end{aligned}\,,
$$

and hence $\lambda_0(t)$ is constant and $\lambda_2(t) = 1$ for all t. Therefore H is maximized by

$$
u = \begin{matrix} 0 \\ 1 \end{matrix} \quad \text{as } \lambda_1 a - c \begin{matrix} \leq \\ \geq \end{matrix} 0\,.
$$

At T, this implies $u(T) = 0$. If t is a little less than T, λ_1 is small and $u = 0$ so the equation for λ_1 is

$$
\dot{\lambda}_1 = \lambda_1 b - c. \tag{14.4}
$$

As long as λ_1 is small, $\dot{\lambda}_1 < 0$. Therefore as the *remaining time s* increases, $\lambda_1(s)$ increases, until such point that $\lambda_1 a - c \geq 0$. The optimal control becomes $u = 1$ and then $\dot{\lambda}_1 = -\lambda_1(a - b) < 0$, which implies that $\lambda_1(s)$ continues to increase as s increases, right back to the start. So there is no further switch in u.

The point at which the single switch occurs is found by integrating (14.4) backwards from $s = 0$, to give $\lambda_1(s) = (c/b)(1 - e^{sb})$ and so the switch occurs where $\lambda_1 a - c = 0$, i.e., $(a/b)(1 - e^{sb}) = 1$, or

$$
t_{\text{switch}} = T + (1/b)\log(1 - b/a).
$$

Experimental evidence suggests that social insects do closely follow this policy and adopt a switch time that is nearly optimal for their natural environment. ∎

14.4   Example: rocket thrust optimization

Regard a rocket as a point mass with position x, velocity v and mass m. Mass is changed only by expansion of matter in the jet. Suppose the jet has vector velocity k relative to the rocket and the rocket is subject to external force f. Then the condition of momentum conservation yields

$$
(m - \delta m)(v + \delta v) + (v - k)\delta m - mv = f \delta t\,,
$$

and this gives the so-called 'rocket equation',

$$
m\dot{v} = k\dot{m} + f\,.
$$

Suppose the jet speed $|k| = 1/b$ is fixed, but the direction and the rate of expulsion of mass can be varied. Then the control is the thrust vector $u = k\dot{m}$, subject to $|u| \leq 1$, say. Find the control that maximizes the height that the rocket reaches.

Solution. The plant equation (in $\mathbb{R}^3$) is

$$
\begin{aligned}
\dot{x} &= v \\
m\dot{v} &= u + f \\
\dot{m} &= -b|u|\,.
\end{aligned}
$$

We take dual variables p, q, r corresponding to x, v, m. Then

$$H = p^\top v + \frac{q^\top (u + f)}{m} - rb|u| - c,$$

(where if the costs are purely terminal $c = 0$), and u must maximize

$$\frac{q^\top u}{m} - br|u|.$$

The optimal u is in the direction of q so $u = |u|q/|q|$ and $|u|$ maximizes

$$|u| \left(\frac{|q|}{m} - br \right).$$

Thus we have that the optimal thrust should be

$$\begin{matrix} \text{maximal} \\ \text{intermediate} \\ \text{null} \end{matrix} \quad \text{as} \quad \left(\frac{|q|}{m} - rb \right) \quad \begin{matrix} > \\ = \\ < \end{matrix} \quad 0.$$

The control is bang/bang and p, q, r are determined from the dual equations.

If the rocket is launched vertically then $f = -mg$ and the dual equations give $\dot p = 0$, $\dot q = -p$ and $\dot r = qu/m^2 > 0$. Suppose we want to maximize the height that the rocket attains. Let m_0 be the mass of the rocket structure, so that the maximum height has been reached if $m = m_0$ and $v \le 0$. Since $\mathbf{K} = -x$ at termination, the transversality conditions give $p(T) = 1$, $q(T) = 0$. Thus $p(s) = 1$, $q(s) = s$, and $|u|$ must maximize $|u| (s/m - br)$. One can check that $(d/ds)(s/m - rb) > 0$, and hence we should use full thrust from launch up to some time, and thereafter coast to maximum height on zero thrust. ∎

56

15 Controlled Markov Jump Processes

We conclude with models for controlled optimization problems in a continuous time stochastic setting. This lecture is about controlled Markov jump processes, which are relevant when the state space is discrete.

15.1 The dynamic programming equation

The DP equation in incremental form is

$$F(x,t) = \inf_u \{c(x,u)\delta t + E[F(x(t+\delta t), t+\delta t) \mid x(t) = x, u(t) = u)]\}.$$

If appropriate limits exist then this can be written in the limit $\delta t \downarrow 0$ as

$$\inf_u [c(x,u) + F_t(x,t) + \Lambda(u)F(x,t)] = 0.$$

Here $\Lambda(u)$ is the operator defined by

$$\Lambda(u)\phi(x) = \lim_{\delta t \downarrow 0}\left[\frac{E[\phi(x(t+\delta t)) \mid x(t) = x, u(t) = u] - \phi(x)}{\delta t}\right] \qquad (15.1)$$

or

$$\Lambda(u)\phi(x) = \lim_{\delta t \downarrow 0} E\left[\frac{\phi(x(t+\delta t)) - \phi(x)}{\delta t} \,\middle|\, x(t) = x, u(t) = u\right]$$

the conditional expectation of the 'rate of change' of $\phi(x)$ along the path. The operator Λ converts a scalar function of state, $\phi(x)$, to another such function, $\Lambda\phi(x)$. However, its action depends upon the control u, so we write it as $\Lambda(u)$. It is called the **infinitesimal generator** of the controlled Markov process. Equation (15.1) is equivalent to

$$E[\phi(x(t+\delta t) \mid x(t) = x, u(t) = u] = \phi(x) + \Lambda(u)\phi(x)\delta t + o(\delta t).$$

This equation takes radically different forms depending upon whether the state space is discrete or continuous. Both are important, and we examine their forms in turn, beginning with a discrete state space.

15.2 The case of a discrete state space

Suppose that x can take only values in a discrete set, labelled by an integer j, say, and that the **transition intensity**

$$\lambda_{jk}(u) = \lim_{\delta t \downarrow 0}\frac{1}{\delta t}P(x(t+\delta t) = k \mid x(t) = j, u(t) = u)$$

is defined for all u and $j \neq k$. Then

$$E\big[\phi(x(t+\delta t)) \mid x(t) = j, u(t) = u\big]$$

$$= \sum_{k \neq j} \lambda_{jk}(u)\phi(k)\delta t + \left(1 - \sum_{k \neq j}\lambda_{jk}(u)\delta t\right)\phi(j) + o(\delta t),$$

whence it follows that

$$\Lambda(u)\phi(j) = \sum_k \lambda_{jk}(u)[\phi(k) - \phi(j)]$$

and the DP equation becomes

$$\inf_u \left[c(j,u) + F_t(j,t) + \sum_k \lambda_{jk}(u)[F(k,t) - F(j,t)] \right] = 0. \qquad (15.2)$$

This is the optimality equation for a **Markov jump process**.

15.3   Uniformization in the infinite horizon case

In this section we explain how (in the infinite horizon case) the continuous time DP equation (15.2) can be rewritten to look like a discrete time DP equation. Once this is done then all the ideas of Lectures 1–6 can be applied. In the discounted cost case (15.2) undergoes the usual modification to

$$\inf_u \left[c(j,u) - \alpha F(j,t) + F_t(j,t) + \sum_k \lambda_{jk}(u)[F(k,t) - F(j,t)] \right] = 0.$$

In the infinite horizon case, everything becomes independent of time and we have

$$\inf_u \left[c(j,u) - \alpha F(j) + \sum_k \lambda_{jk}(u)[F(k) - F(j)] \right] = 0. \qquad (15.3)$$

Suppose we can choose a B large enough that it is possible to define

$$\lambda_{jj}(u) = B - \sum_{k \neq j} \lambda_{jk}(u) \geq 0,$$

for all j and u. By adding $(B + \alpha)F(j)$ to both sides of (15.3), the DP equation can be written

$$(B + \alpha)F(j) = \inf_u \left[c(j,u) + \sum_k \lambda_{jk}(u)F(k) \right],$$

Finally, dividing by $B + \alpha$, this can be written as

$$F(j) = \inf_u \left[\bar{c}(j,u) + \beta \sum_k p_{jk}(u)F(k) \right], \qquad (15.4)$$

where

$$\bar{c}(j,u) = \frac{c(j,u)}{B + \alpha}, \quad \beta = \frac{B}{B + \alpha}, \quad p_{jk}(u) = \frac{\lambda_{jk}(u)}{B} \quad \text{and} \quad \sum_k p_{jk}(u) = 1.$$

58

This makes the dynamic programming equation look like a case of discounted dynamic programming in discrete time, or of negative programming if $\alpha = 0$. All the results we have for those cases can now be used (e.g., value iteration, OSLA rules, etc.) The trick of using a large B to make the reduction from a continuous to a discrete time formulation is called **uniformization**.

In the undiscounted case we could try a solution to (15.2) of the form $F(j, t) = -\gamma t + \phi(j)$. Substituting this in (15.2), we see that this gives a solution provided,

$$0 = \inf_u \left[c(j, u) - \gamma + \sum_k \lambda_{jk}(u)[\phi(k) - \phi(j)] \right] .$$

By adding $B\phi(j)$ to both sides of the above, then dividing by B, setting $\bar\gamma = \gamma/B$, and making the other substitutions above (but with $\alpha = 0$), this is equivalent to

$$\phi(j) + \bar\gamma = \inf_u \left[\bar c(j, u) + \sum_k p_{jk}(u)\phi(k) \right] , \qquad (15.5)$$

which has the same form as the discrete-time average-cost optimality equation of Lecture 6. The theorems and techniques of that lecture can now be applied.

15.4 Example: admission control at a queue

Consider a queue of varying size $0, 1, \dots$, with constant service rate μ and arrival rate u, where u is controllable between 0 and a maximum value λ. Let $c(x, u) = ax - Ru$. This corresponds to paying a cost a per unit time for each customer in the queue and receiving a reward R at the point that each new customer is admitted (and therefore incurring reward at rate Ru when the arrival rate is u). Let us take $B = \lambda + \mu$, and without loss of generality assume $B = 1$. The average cost optimality equation from (15.5) is

$$\phi(0) + \gamma = \inf_u [-Ru + u\phi(1) + (\mu + \lambda - u)\phi(0)] ,$$
$$= \inf_u [u\{-R + \phi(1) - \phi(0)\} + (\mu + \lambda)\phi(0)] ,$$

$$\phi(x) + \gamma = \inf_u [ax - Ru + u\phi(x + 1) + \mu\phi(x - 1) + (\lambda - u)\phi(x)] ,$$
$$= \inf_u [ax + u\{-R + \phi(x + 1) - \phi(x)\} + \mu\phi(x - 1) + \lambda\phi(x)] , \quad x > 0 .$$

Thus u should be chosen to be 0 or 1 as $-R + \phi(x + 1) - \phi(x)$ is positive or negative.

Let us consider what happens under the policy that take $u = \lambda$ for all x. The relative costs for this policy, say f, are given by

$$f(x) + \gamma = ax - R\lambda + \lambda f(x + 1) + \mu f(x - 1) , \quad x > 0 .$$

The solution to the homogeneous part of this recursion is of the form $f(x) = d_1 1^x + d_2(\mu/\lambda)^x$. Assuming $\lambda < \mu$ and we desire a solution for f that does not grow exponentially, we take $d_2 = 0$ and so the solution is effectively the solution to the inhomogeneous part, i.e.,

$$f(x) = \frac{ax(x+1)}{2(\mu - \lambda)}, \qquad \gamma = \frac{a\lambda}{\mu - \lambda} - \lambda R,$$

Applying the idea of policy improvement, we conclude that a better policy is to take $u = 0$ (i.e., don't admit a customer) if $-R + f(x+1) - f(x) > 0$, i.e., if

$$\frac{(x+1)a}{\mu - \lambda} - R > 0.$$

Further policy improvement would probably be needed to reach the optimal policy. However, this policy already exhibits an interesting property: it rejects customers for smaller queue length x than does a policy which rejects a customer if and only if

$$\frac{(x+1)a}{\mu} - R > 0.$$

This second policy is optimal if one is purely concerned with whether or not an individual customer that joins when there are x customers in front of him will show a profit on the basis of the difference between the reward R and his expected holding cost $(x+1)a/\mu$. This example exhibits the difference between **individual optimality** (which is myopic) and **social optimality**. The socially optimal policy is more reluctant to admit customers because it anticipates that more customers on the way; thus it feels less badly about forgoing the profit on a customer that presents himself now, recognizing that admitting such a customer can cause customers who are admitted after him to suffer greater delay. As expected, the policies are nearly the same if the arrival rate λ is small.

16 Controlled Diffusion Processes

We give a brief introduction to controlled continuous time stochastic models with a continuous state space, i.e., controlled diffusion processes.

16.1 Diffusion processes and controlled diffusion processes

The **Wiener process** $\{B(t)\}$, is a scalar process for which $B(0) = 0$, the increments in B over disjoint time intervals are statistically independent and $B(t)$ is normally distributed with zero mean and variance t. ('B' stands for **Brownian motion**.) This specification is internally consistent because, for example,

$$B(t) = B(t_1) + [B(t) - B(t_1)]$$

and for $0 \leq t_1 \leq t$ the two terms on the right-hand side are independent normal variables of zero mean and with variance t_1 and $t - t_1$ respectively.

If δB is the increment of B in a time interval of length δt then

$$E(\delta B) = 0, \quad E[(\delta B)^2] = \delta t, \quad E[(\delta B)^j] = o(\delta t), \quad \text{for } j > 2,$$

where the expectation is one conditional on the past of the process. Note that since

$$E[(\delta B/\delta t)^2] = O\left[(\delta t)^{-1}\right] \to \infty,$$

the formal derivative $\epsilon = dB/dt$ (continuous-time 'white noise') does not exist in a mean-square sense, but expectations such as

$$E\left[\left\{\int \alpha(t)\epsilon(t)dt\right\}^2\right] = E\left[\left\{\int \alpha(t)dB(t)\right\}^2\right] = \int \alpha(t)^2 dt$$

make sense if the integral is convergent.

Now consider a **stochastic differential equation**

$$\delta x = a(x, u)\delta t + g(x, u)\delta B,$$

which we shall write formally as

$$\dot{x} = a(x, u) + g(x, u)\epsilon.$$

This, as a Markov process, has an infinitesimal generator with action

$$\Lambda(u)\phi(x) = \lim_{\delta t \downarrow 0} E\left[\frac{\phi(x(t + \delta t)) - \phi(x)}{\delta t}\middle| x(t) = x, u(t) = u\right]$$
$$= \phi_x a + \tfrac{1}{2}\phi_{xx}g^2$$
$$= \phi_x a + \tfrac{1}{2}N\phi_{xx},$$

where $N(x, u) = g(x, u)^2$. The DP equation is thus

$$\inf_u \left[c + F_t + F_x a + \tfrac{1}{2} N F_{xx} \right] = 0.$$

In the vector case this becomes the **controlled diffusion process**

$$\inf_u \left[c + F_t + F_x a + \tfrac{1}{2} \mathrm{tr}(N F_{xx}) \right] = 0.$$

16.2   Example: LQG in continuous time

The dynamic programming equation is

$$\inf_u \left[x^\top R x + u^\top Q u + F_t + F_x^\top (Ax + Bu) + \tfrac{1}{2} \mathrm{tr}(N F_{xx}) \right] = 0.$$

In analogy with the discrete and deterministic continuous cases that we have considered previously, we try a solution of the form,

$$F(x, t) = x^\top \Pi(t) x + \gamma(t).$$

This leads to the same Riccati equation as in Section 12.2,

$$0 = x^\top \left[R + \Pi A + A^\top \Pi - \Pi B Q^{-1} B^\top \Pi + \frac{d\Pi}{dt} \right] x,$$

and also, as in Section 7.3,

$$\frac{d\gamma}{dt} + \mathrm{tr}(N\Pi(t)) = 0, \quad \text{giving} \quad \gamma(t) = \int_t^T \mathrm{tr}(N\Pi(\tau))\, d\tau.$$

16.3   Example: passage to a stopping set

Consider a problem of movement on the unit interval $0 \le x \le 1$ in continuous time, $\dot{x} = u + \epsilon$, where ϵ is white noise of **power** v. The process terminates at time T when x reaches one end or the other of the the interval. The cost is made up of an integral term $\tfrac{1}{2}\int_0^T (L + Qu^2)dt$, penalising both control and time spent, and a terminal cost which takes the value C_0 or C_1 according as termination takes place at 0 or 1.

Show that in the deterministic case $v = 0$ one should head straight for one of the termination points at a constant rate and that the value function $F(x)$ has a piecewise linear form, with possibily a discontinuity at one of the boundary points if that boundary point is the optimal target from no interior point of the interval.

Show, in the stochastic case, that the dynamic programming equation with the control value optimized out can be linearised by a transformation $F(x) = \alpha \log \phi(x)$ for a suitable constant α, and hence solve the problem.

Solution. In the deterministic case the optimality equation is

$$\inf_u \left[\frac{L + Qu^2}{2} + u \frac{\partial F}{\partial x} \right] = 0, \quad 0 < x < 1, \tag{16.1}$$

with boundary conditions $F(0) = C_0$, $F(1) = C_1$. If one goes (from x) for $x = 0$ at speed w one incurs a cost of $C_0 + (x/2w)(L + Qw^2)$ with a minimum over w value of $C_0 + x\sqrt{LQ}$. Indeed (16.1) is solved by

$$F(x) = \min\left[C_0 + x\sqrt{LQ}, C_1 + (1-x)\sqrt{LQ}\right].$$

The minimizing option determines the target and the optimal w is $\sqrt{L/Q}$.

In the stochastic case

$$\inf_u\left[\frac{L + Qu^2}{2} + u\frac{\partial F}{\partial x} + \frac{v}{2}\frac{\partial^2 F}{\partial x^2}\right] = 0.$$

So $u = -Q^{-1}F_x$ and

$$L - Q^{-1}\left(\frac{\partial F}{\partial x}\right)^2 + v\frac{\partial^2 F}{\partial x^2} = 0.$$

Make the transform $F(x) = -Qv\log\phi(x)$ so $\phi(x) = e^{-F(x)/Qv}$. Then

$$Qv^2\frac{\partial^2\phi}{\partial x^2} - L\phi = 0,$$

with solution

$$\phi(x) = k_1\exp\left(\frac{x}{v}\sqrt{L/Q}\right) + k_2\exp\left(-\frac{x}{v}\sqrt{L/Q}\right).$$

We choose the constants k_1, k_2 to meet the two boundary conditions on F. The figure
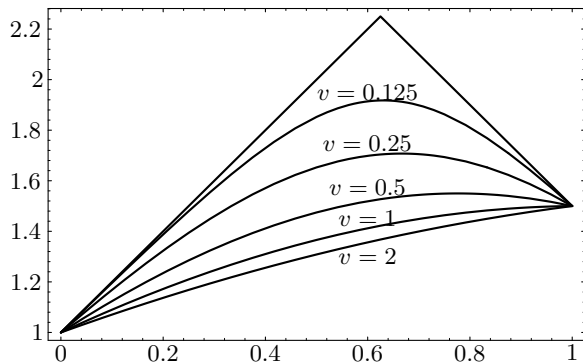
$F(x)$ against x for the passage to a stopping set

shows the solution for $L = 1$, $Q = 4$, $C_0 = 1$, $C_1 = 1.5$ and $v = 0.125, 0.25, 0.5, 1, 2$ and the deterministic solution. Notice that noise actually reduces cost by lessening the time until absorption at one or the other of the endpoints. ∎

63

16.4   Addendum on PMP: control subject to constraints

The necessary condition that the optimal u maximizes $H(x, u, \lambda)$ holds even when u is constrained to lie in some given set. The following examples illustrates this idea.

Example: princess and monster

Consider a princess who is jogging with speed r in the counterclockwise direction around a circular running track of radius r, and so has a position whose horizontal and vertical components at time t are $(r \cos t, r \sin t)$, $t \geq 0$. A monster who is initially located at the centre of the circle can move with velocity u, where $|u| \leq 1$. (This is in contrast to the set up on Examples Sheet 3, where we supposed a different constraint, $|u_i| \leq 1$.) The monster wishes to catch the princess in minimal time.

Solution. Let us take horizontal and vertical coordinates, (x, y). The Hamiltonian is

$$H(x, u, \lambda) = \lambda_1(u_1 + r \sin t) + \lambda_2(u_2 - r \cos t) - 1 \,.$$

This must be maximized over u_1, u_2 with $u_1^2 + u_2^2 \leq 1$. Since $\dot{\lambda}_1 = -H_x = 0$, $\dot{\lambda}_2 = -H_y = 0$ we see that λ_1 and λ_2 are constants, and hence maximizing $\lambda_1 u_1 + \lambda_2 u_2$ subject to the constraint gives

$$u_1 = \frac{\lambda_1}{\sqrt{\lambda_1^2 + \lambda_2^2}} \,, \qquad u_2 = \frac{\lambda_2}{\sqrt{\lambda_1^2 + \lambda_2^2}} \,.$$

The key point is that u_1 and u_2 are constants, and thus the monster should follow a straight line from the center of the circle to its rim, choosing the direction so as to intercept the princess in minimal time (which is r.)

Of course this form of optimal policy is completely intuitive and could be established by ad hoc arguments rather than by appealing to PMP. ∎

Index

65