Optimization and Control

Contents

Table of Contents i

Schedules iv

1 Dynamic Programming 1
- 1.1 Control as optimization over time . . . 1
- 1.2 The principle of optimality . . . 1
- 1.3 Example: the shortest path problem . . . 1
- 1.4 The optimality equation . . . 2
- 1.5 Markov decision processes . . . 4

2 Examples of Dynamic Programming 5
- 2.1 Example: managing spending and savings . . . 5
- 2.2 Example: exercising a stock option . . . 6
- 2.3 Example: secretary problem . . . 7

3 Dynamic Programming over the Infinite Horizon 9
- 3.1 Discounted costs . . . 9
- 3.2 Example: job scheduling . . . 9
- 3.3 The infinite-horizon case . . . 10
- 3.4 The optimality equation in the infinite-horizon case . . . 11
- 3.5 Example: selling an asset . . . 12

4 Positive Programming 13
- 4.1 Example: possible lack of an optimal policy. . . . 13
- 4.2 Characterization of the optimal policy . . . 13
- 4.3 Example: optimal gambling . . . 14
- 4.4 Value iteration . . . 14
- 4.5 Example: pharmaceutical trials . . . 15

5 Negative Programming 17
- 5.1 Stationary policies . . . 17
- 5.2 Characterization of the optimal policy . . . 17
- 5.3 Optimal stopping over a finite horizon . . . 18
- 5.4 Example: optimal parking . . . 18
- 5.5 Optimal stopping over the infinite horizon . . . 19

6 Bandit Processes and the Gittins Index **21**
6.1 Multi-armed bandit problem 21
6.2 Gittins index theorem 22
6.3 Calibration 23
6.4 Proof of the Gittins index theorem 23
6.5 Calculation of the Gittins index 24
6.6 Forward induction policies 24

7 Average-cost Programming **25**
7.1 Average-cost optimality equation 25
7.2 Example: admission control at a queue 26
7.3 Value iteration bounds 26
7.4 Policy improvement algorithm 27

8 LQ Regulation **29**
8.1 The LQ regulation problem 29
8.2 The Riccati recursion 31
8.3 White noise disturbances 31
8.4 LQ regulation in continuous-time 32

9 Controllability **33**
9.1 Controllability 33
9.2 Controllability in continuous-time 34
9.3 Linearization of nonlinear models 35
9.4 Example: broom balancing 35
9.5 Example: satellite in a plane orbit 36

10 Stabilizability and Observability **37**
10.1 Stabilizability 37
10.2 Example: pendulum 37
10.3 Infinite-horizon LQ regulation 38
10.4 The $[A, B, C]$ system 38
10.5 Observability in continuous-time 40

11 Kalman Filter and Certainty Equivalence **41**
11.1 Example: observation of population 41
11.2 Example: satellite in planar orbit 41
11.3 Imperfect state observation with noise 41
11.4 The Kalman filter 42
11.5 Certainty equivalence 44

12 Dynamic Programming in Continuous Time **45**
12.1 Example: parking a rocket car 45
12.2 The Hamilton-Jacobi-Bellman equation 45
12.3 Example: LQ regulation 46
12.4 Example: harvesting fish 47

13 Pontryagin's Maximum Principle **49**
13.1 Heuristic derivation 49
13.2 Example: parking a rocket car (continued) 50
13.3 Transversality conditions 51
13.4 Adjoint variables as Lagrange multipliers 52

14 Applications of the Maximum Principle **53**
14.1 Example: a moon lander 53
14.2 Example: use of transversality conditions 54
14.3 Problems in which time appears 54
14.4 Example: monopolist 55
14.5 Example: insects as optimizers 56

15 Controlled Markov Jump Processes **57**
15.1 The dynamic programming equation 57
15.2 The case of a discrete state space 57
15.3 Uniformization in the infinite horizon case 58
15.4 Example: admission control at a queue 59

16 Controlled Diffusion Processes **61**
16.1 Diffusion processes and controlled diffusion processes 61
16.2 Example: noisy LQ regulation in continuous time 62
16.3 Example: a noisy second order system 62
16.4 Example: passage to a stopping set 63

Index **65**

Schedules

Dynamic programming
The principle of optimality. The dynamic programming equation for finite-horizon problems. Interchange arguments. Markov decision processes in discrete time. Innite-horizon problems: positive, negative and discounted cases. Value interation. Policy improvment algorithm. Stopping problems. Average-cost programming. [6]

LQG systems
Linear dynamics, quadratic costs, Gaussian noise. The Riccati recursion. Controllability. Stabilizability. Innite-horizon LQ regulation. Observability. Imperfect state observation and the Kalman lter. Certainty equivalence control. [5]

Continuous-time models
The optimality equation in continuous time. Pontryagins maximum principle. Heuristic proof and connection with Lagrangian methods. Transversality conditions. Optimality equations for Markov jump processes and diffusion processes. [5]

Richard Weber, January 2012

1 Dynamic Programming

Dynamic programming and the principle of optimality. Notation for state-structured models. Feedback, open-loop, and closed-loop controls. Markov decision processes.

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 when an optimal solution is not required it can be useful to test one's thinking by following an optimization approach. 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.

The optimization models in the IB course (for linear programming and network flow models) were static and nothing was either random or hidden. In this course it is the additional features of 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.

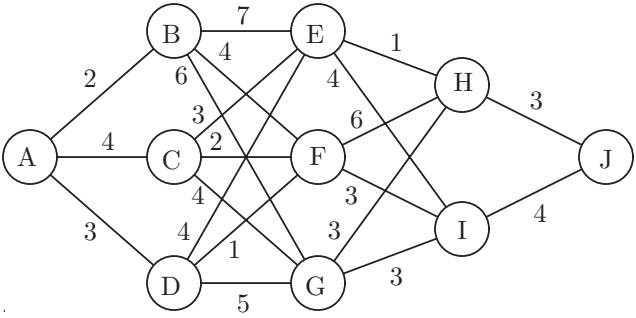
1.2 The principle of optimality

A key idea in this course is that optimization over time can often be seen as 'optimization in stages'. We trade off our desire to obtain the least 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 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 $F(\mathrm{X})$ be the minimal distance required to reach J from X. Then clearly, $F(\mathrm{J}) = 0$, $F(\mathrm{H}) = 3$ and $F(\mathrm{I}) = 4$.

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

and so on. Recursively, we obtain $F(\mathrm{A}) = 11$ 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.

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

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, can be found 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 **separable 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 in 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)

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. The following assumptions define what is known as a discrete-time **Markov decision process** (MDP).

(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) *Separable (or decomposible) cost function*, (i.e. cost given by (1.1)).

For the moment we also require the following:

(c) *Perfect state observation:* The current value of the state is observable. That is, x_t is known when choosing u_t. 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 $\mathbf{C} = \mathbf{C}(W_h)$.

As in the previous section, 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 $u_0, \ldots, u_{h-1}$.

The following theorem is then obvious.

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, a 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 Examples of Dynamic Programming

Examples of dynamic programming problems and some useful tricks to solve them. The idea that it can be useful to model things in terms of time to go.

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

$$x_{t+1} = a(x_t, u_t) = x_t + \theta(x_t - u_t). \tag{2.1}$$

He desires to maximize his total consumption over h years,

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

The plant equation (2.1) specifies a **Markov decision process** (MDP). When we add to this the aim of maximizing the performance measure $\mathbf{C}$ we have what is called a **Markov decision problem**. For both we use the abbreviation MDP. In the notation we have been using, $c(x_t, u_t, t) = u_t$, $\mathbf{C}_h(x_h) = 0$. This is termed a **time-homogeneous** model because neither costs nor dynamics depend on t.

Solution. 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**', $s = h - t$. Let $F_s(x)$ denote the maximal reward obtainable, starting in state x 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

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

where s^* is the least integer such that $1+s^* \le (1+\theta)s^* \iff s^* \ge 1/\theta$, 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 $h-s^*, \ldots, h-1$.

There are several things worth learning 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 you 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 she exercises the option on day t and then immediately sells the share at the current price x_t, she 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 $x \ge a_s$, 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. As above, we use 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, $F_{s-1}(\cdot)$.

It easy to show (i), (ii) and (iii) by induction on s. 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

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

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: secretary problem

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 t th candidate. All that matters are the value of t and whether the t th 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 $F(t - 1)$ be the probability that under an optimal policy we select the best candidate, given that we have passed over the first $t - 1$ candidates. Dynamic programming gives

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

The first term deals with what happens when the tth candidate is not the best so far; we should certainly pass over her. The second term deals with what happens when she is the best so far. Now we have a choice: either accept her (and she will turn out to be best with probability t/h), or pass over her.

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

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

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

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

whence

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

Since we require $F(t_0) \leq t_0/h$, 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 $t_0 \approx h/e$. Thus the optimal policy is to interview $\approx h/e$ candidates, but without selecting any of these, and then select the first candidate thereafter who is the best of all those seen so far. The probability of success is $F(0) = F(t_0) \sim t_0/h \sim 1/e = 0.3679$. It is surprising that the probability of success is so large for arbitrarily large h.

There are a couple things to learn from 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 $F(t)$. (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.

3 Dynamic Programming over the Infinite Horizon

Cases of discounted, negative and positive dynamic programming. Validity of the optimality equation over the infinite horizon.

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 finance, 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

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

where E_π denotes expectation over the future path of the process under policy π. The DP equation is now

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

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-to-go k' as the point at which the $n - k$ th job has just been completed and there remains a set of k 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 jobs are processed in the order $i_1, \ldots, i_k, i, j, i_{k+3}, \ldots, i_n$. Compare the reward that is obtained if the order of jobs i and j is reversed: $i_1, \ldots, i_k, j, i, i_{k+3}, \ldots, i_n$. 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 times at which jobs complete.

3.3 The infinite-horizon case

In the finite-horizon case the value 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.

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

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

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

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

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 (e.g. if $\beta = 1$, $x_{t+1} = -x_t$ and $c(x,u) = x$), but it will do so under any of the following three scenarios.

D (**discounted programming**):	$0 < \beta < 1$,	and $\lvert c(x,u)\rvert < B$	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 $c(x,u)$. The names actually come from equivalent problems of maximizing rewards, like $r(x,u)$ $(= -c(x,u))$. 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 the optimality equation (i.e. (3.7)).

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

$$
\begin{aligned}
F(x) &\le F(\pi, x) \\
&= c(x, u_0) + \beta E[F(\pi^1, x_1) \mid x_0 = x, u_0] \\
&\le c(x, u_0) + \beta E[F(x_1) + \epsilon \mid x_0 = x, u_0] \\
&\le c(x, u_0) + \beta E[F(x_1) \mid x_0 = x, u_0] + \beta\epsilon.
\end{aligned}
$$

Minimizing the right hand side over u_0 and recalling that ϵ is arbitrary gives '$\le$'. □

3.5 Example: selling an asset

A spectulator owns a rare collection of tulip bulbs and each day has one opportunity to sell it, which he may either accept or reject. The potential sale prices are independently and identically distributed with probability density function $g(x)$, $x \ge 0$. Each day there is a probability $1-\beta$ that the market for tulip bulbs will collapse, making his bulb collection completely worthless. Find the policy that maximizes his expected return and express it as the unique root of an equation. Show that if $\beta > 1/2$, $g(x) = 2/x^3$, $x \ge 1$, then he should sell the first time the sale price is at least $\sqrt{\beta/(1-\beta)}$.

Solution. There are only two states, depending on whether he has sold the collection or not. Let these be 0 and 1, respectively. The optimality equation is

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

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 when he can get at least βF^*. His maximal reward is F^*.

Consider the case $g(y) = 2/y^3$, $y \ge 1$. 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 $\beta \le 1/2$ he should sell at any price.

Notice that discounting arises in this problem because at each stage there is a probability $1-\beta$ that a 'catastrophe' will occur that brings things to a sudden end. This characterization of the way that discounting can arise is often quite useful.

4 Positive Programming

Special theory for maximizing positive rewards. We see that there can be no optimal policy. However, if a given policy has a value function that satisfies the optimality equation then that policy is optimal. Value iteration algorithm.

4.1 Example: possible lack of an optimal policy.

Positive programming is about maximizing non-negative rewards, $r(x, u) \geq 0$, or minimizing non-positive costs, $c(x, u) \leq 0$. The following example shows that there may be no optimal policy.

Example 4.1. 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 $1 - 1/x$, 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 4.2. *Suppose D or P holds and π is a policy whose value function $F(\pi, x)$ satisfies the optimality equation*

$$F(\pi, x) = \sup_u \{r(x, u) + \beta E[F(\pi, x_1) \mid x_0 = x, u_0 = u]\}.$$

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

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

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 whole number of pounds not exceeding 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, say π, is optimal we need to find its value function, say $G(i) = F(\pi, x)$, and then 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

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

Let W_j be the expression inside $\{\ \}$ on the right hand side. It is simple calculation to show that $W_{j+1} < W_j$ for all $j \geq 1$. Hence $j = 1$ maximizes the right hand side.

4.4 Value iteration

An important and practical method of computing F is **successive approximation** or **value iteration**. Starting with $F_0(x) = 0$, we can successively calculate, for $s = 1, 2, \ldots,$

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

So $F_s(x)$ is the infimal cost over s steps. Now let

$$F_\infty(x) = \lim_{s\to\infty} F_s(x) = \lim_{s\to\infty} \inf_\pi F_s(\pi, x). \tag{4.2}$$

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 $F_s(x) \to F(x)$. 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 4.3. *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. [Alternatively, 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,

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

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 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

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

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

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

Notice that after the first time that the doctor decides is not optimal to use the new drug it cannot be optimal for him to return to using it later, since his indformation about that drug cannot have changed while not using it.

It is not possible to give a closed-form 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 (the known success probability of the well-established drug) for which it is worth continuing with at least one more trial of the new drug. For example, suppose $p = 0.6$ and 6 trials with the new drug have given $s = f = 3$. Then since $p = 0.6 < 0.6133$ we should treat the next patient with the new drug. At this point the probability that the new drug will successfully treat the next patient is 0.5 and so the doctor will actually be treating that patient with the drug that is least likely to cure!

Here we see a tension going on between desires for **exploitation** and **exploration**. A **myopic policy** seeks only to maximize immediate reward. However, an optimal policy takes account of the possibility of gaining information that could lead to greater rewards being obtained 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. Of course as the discount factor β tends to 0 the optimal policy will looks more and more like the myopic policy.

5 Negative Programming

The special theory of minimizing positive costs. We see that action that extremizes the right hand side of the optimality equation is an optimal policy. 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 4.2 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 can write $\pi = (f_0, f_1, \dots)$ to denote such a policy. If in addition the policy does not depend on time and is non-randomizing in its choice of action then it is said to be a **deterministic stationary Markov policy**, and we write $\pi = (f, f, \dots) = f^\infty$.

5.2 Characterization of the optimal policy

Negative programming is about maximizing non-positive rewards, $r(x, u) \leq 0$, or minimizing non-negative costs, $c(x, u) \geq 0$. 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. Note that in this theorem we are requiring that the infimum over u is attained as a minimum over u.

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

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

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

(i.e. $u = f(x)$ is the value of u which minimizes the r.h.s. of the DP equation.)

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,

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

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 if assumption F holds then an optimal policy exists. Neither Theorem 5.1 or this corollary are true for positive programming (see Example 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. We let $c(x, 0) = k(x)$ (stopping cost) and $c(x, 1) = c(x)$ (continuation cost).

Suppose that $F_s(x)$ denotes the minimum total cost when we are constrained to stop within the next s steps. This gives a finite-horizon problem with dynamic programming equation

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

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. If S is closed then the following theorem gives us the form of the optimal policies for all finite-horizons.

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 rule).

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

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

where $F_0(0) = D$, $F_0(1) = 0$. It is easy to see that $qF_{s-1}(0) + pF_{s-1}(1)$ is decreasing in s. So the stopping condition of $qF_{s-1}(0) + pF_{s-1}(1) \geq s$ will be satisfied for small s.

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 general solution is of the form $k(s) = -q/p + s + cq^s$. So after substituting and using the boundary condition at $s = 0$, we have

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

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

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

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

We might let D be the expected distance that that the driver must walk if he takes the first available space at the destination or further down the road. In this case, $D = 1 + qD$, so $D = 1/p$ and s^* is the greatest integer such that $2q^s \geq 1$.

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 the sth step. 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

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

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

Example 5.3 (showing that $F_\infty > F$ is possible). 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 $F_\infty(x) > F(x)$.

Remark. In Theorem 4.3 we had $F_\infty = F$, but for that theorem $k(x) = 0$. In the above example, $k(x) > 0$ and $F_s(x)$ is the infimal cost over the set of policies that must stop within s steps.

Example 5.4 (showing Theorem 4.2 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 5.5. *Suppose all costs are bounded as follows.*

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

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

Proof. 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

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

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

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

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 5.6. *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 5.5 gives $F(x) = F_\infty(x)$. □

6 Bandit Processes and the Gittins Index

The multi-armed bandit problem. Bandit processes. Gittins index theorem.

6.1 Multi-armed bandit problem

A **multi-armed bandit** is a slot-machine with multiple arms. The arms differ in the distributions of rewards that they pay out when pulled. An important special case is when arm i is a so-called **Bernoulli bandit**, with parameter p_i. We have already met this as the drug-testing model in §4.5. Such an arm pays £1 with probability p_i, and £0 with probability $1-p_i$; this happens independently each time the arm is pulled. If there are n such arms, and a gambler knows the true values of $p_1, \ldots, p_n$, then obviously he maximizes his expected reward by always pulling the arm of maximum p_i. However, if he does not know these values, then he must choose each successive arm on the basis of the information he has obtained by playing, updated in a Bayesian manner on the basis of observing the rewards he has obtained on previous pulls. The aim in the **multi-armed bandit problem** (MABP) is to maximize the expected total discounted reward.

More generally, we consider a problem of controlling the evolution of n independent reward-producing Markov processes decision processes. The action space of each process contains just two controls, which cause the process to be either 'continued' or 'frozen'. At each instant (in discrete time) exactly one of these so-called **bandit processes** is *continued* (and reward from it obtained), while all the other bandit processes are *frozen*. The continued process can change state; but frozen processes do not change state. Reward is accrued only from the bandit process that is continued. This creates what is termed a **simple family of alternative bandit processes** (SFAB). The word 'simple' means that all the n bandit processes are available at all times.

Let $x(t) = (x_1(t), \ldots, x_n(t))$ be the states of the n bandits. Let i_t denote the bandit process that is continued at time t under some policy π. In the language of Markov decision problems, we wish to find the value function:

$$F(x) = \sup_{\pi} E\left[\sum_{t=0}^{\infty} r_{i_t}(x_{i_t}(t))\beta^t \,\middle|\, x(0) = x \right],$$

where the supremum is taken over all policies π that are realizable (or non-anticipatory), in the sense that i_t depends only on the problem data and $x(t)$, not on any information which only becomes known only after time t.

Setup in this way, we have an infinite-horizon discounted-reward Markov decision problem. It therefore has a deterministic stationary Markov optimal policy. Its dynamic programming is

$$F(x) = \max_{i:i\in\{1,\ldots,n\}} \left\{ r_i(x) + \beta \sum_{y\in E_i} P_i(x_i, y) F(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) \right\}. \quad (6.1)$$

6.2 Gittins index theorem

Remarkably, the problem posed by a SFAB (or a MABP) can be solved by an **index policy**. That is, we can compute a number (called an index), separately for each bandit process, such that the optimal policy is always to continue the bandit process having the currently greatest index. We can illustrate the idea of an index policy with the example from job scheduling that we previous met in §3.2.

Single machine scheduling. Recall the example in §3.2 in which n jobs are to be processed successively on one machine. Job i has a known processing times t_i, assumed to be a positive integer. On completion of job i a positive reward r_i is obtained. If job 1 is processed first, and job 2 is processed immediately after it, then the sum of discounted rewards obtained from these two jobs is $r_1\beta^{t_1} + r_2\beta^{t_1+t_2}$. If the processing order of the jobs is interchanged, we obtain $r_2\beta^{t_2} + r_1\beta^{t_1+t_2}$. A little algebra shows that the first ordering has the greater reward if $r_1\beta^{t_1}/(1-\beta^{t_1}) > r_2\beta^{t_2}/(1-\beta^{t_2})$. Using this idea, it is not hard to see that the total discounted reward obtained from the n jobs is maximized by processing them in decreasing order of indices, computed as $G_i = r_i\beta^{t_i}(1-\beta)/(1-\beta^{t_i})$. Note that $G_i \to r_i/t_i$ as $\beta \to 1$.

The appropriate index for the MABP is in the same spirit, but more complicated. The key result for the MABP is the following, named after its originator, John Gittins.

Theorem 6.1 (Gittins Index Theorem)**.** *The problem posed by a SFABP, as setup above, is solved by always continuing the process having the greatest* **Gittins index***, which is defined for bandit process i as*

$$G_i(x_i) = \sup_{\tau>0} \frac{E\left[\sum_{t=0}^{\tau-1} \beta^t r_i(x_i(t)) \,\middle|\, x_i(0) = x_i\right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \,\middle|\, x_i(0) = x_i\right]}, \tag{6.2}$$

where τ is a stopping time constrained to take a value in the set $\{1, 2, \dots\}$.

By a **stopping time** τ we mean a time that can be recognized when it occurs. In fact, it can be shown that τ attains the supremum when $\tau = \min\{t : G_i(x_i(t)) \le G_i(x_i(0)), \tau > 0\}$, that is, τ is the first time at which the process reaches a state in which the Gittins index is no greater than it was initially.

Examining (6.2), we see that the Gittins index is the maximal possible quotient of a numerator that is 'expected total discounted *reward* over τ steps', and denominator that is 'expected total discounted *time* over τ steps', where τ is at least 1 step. Notice that the Gittins index can be computed for all states of B_i as a function only of the data $r_i(\cdot)$ and $P_i(\,\cdot\,,\cdot\,)$. That is, it can be computed without knowing anything about the other bandit processes.

In the job scheduling example just considered, the optimal stopping time on the right hand side of (6.2) is $\tau = t_i$, the numerator is $r_i\beta^{t_i}$ and the denominator is $1 + \beta + \cdots + \beta^{t_i-1} = (1-\beta^{t_i})/(1-\beta)$. Thus, $G_i = r_i\beta^{t_i}(1-\beta)/(1-\beta^{t_i})$.

The Index Theorem above is due to Gittins and Jones, who had obtained it by 1970, and presented it in 1972. The solution of the MABP impressed many experts

as surprising and beautiful. Peter Whittle describes a colleague of high repute, asking another colleague '*What would you say if you were told that the multi-armed bandit problem had been solved*?' The reply was '*Sir, the multi-armed bandit problem is not of such a nature that it can be solved*'.

6.3 Calibration

An alternative characterization of $G_i(x_i)$ is

$$G_i(x_i) = \sup\left\{\lambda : \frac{\lambda}{1-\beta} \le \sup_{\tau>0} E\left[\sum_{t=0}^{\tau-1} \beta^t r_i(x_i(t)) + \beta^\tau \frac{\lambda}{1-\beta} \,\Big|\, x_i(0) = x_i\right]\right\}. \quad (6.3)$$

That is, we consider a simple family of two bandit processes: bandit process B_i and a **calibrating bandit process**, say Λ, which pays out a known reward λ at each step it is continued. The Gittins index of B_i is the value of λ for which we are indifferent as to which of B_i and Λ to continue initially. Notice that once we decide to switch from continuing B_i to continuing Λ, at time τ, then information about B_i does not change and so it must be optimal to stick with continuing Λ ever after.

6.4 Proof of the Gittins index theorem

Various proofs have been given of the index theorem, all of which are useful in developing insight about this remarkable result. The following one is due to Weber (1992).

Proof of Theorem 6.1. We start by considering a problem in which only bandit process B_i is available. Let us define the **fair charge**, $\gamma_i(x_i)$, as the maximum amount that a gambler would be willing to pay per step in order to be permitted to continue B_i for at least one more step, and being free to stop continuing it whenever he likes thereafter. This is

$$\gamma_i(x_i) = \sup\left\{\lambda : 0 \le \sup_{\tau>0} E\left[\sum_{t=0}^{\tau-1} \beta^t \Big(r_i(x_i(t)) - \lambda\Big) \,\Big|\, x_i(0) = x_i\right]\right\}. \quad (6.4)$$

Notice that the right hand sides of (6.3) and (6.4) are equivalent and so $\gamma_i(x_i) = G_i(x_i)$. Notice also that the time τ will be the first time that $G_i(x_i(\tau)) < G_i(x_i(0))$.

We next define the **prevailing charge** for B_i at time t as $g_i(t) = \min_{s\le t} \gamma_i(x_i(s))$. So $g_i(t)$ actually depends on $x_i(0), \ldots, x_i(t)$ (which we omit from its argument for convenience). Note that $g_i(t)$ is a nonincreasing function of t and its value depends only on the states through which bandit i evolves. The proof of the Index Theorem is completed by verifying the following facts, each of which is almost obvious.

(i) Suppose that in the problem with n available bandit processes, $B_1, \ldots, B_n$, the gambler not only collects rewards, but also must pay the prevailing charge of the bandit that he chooses to continue. Then he cannot do better than just break even (i.e. have expected reward 0).

This is because he could only make a strictly positive profit (in expected value) if this happens for at least one bandit. Yet the prevailing charge has been defined in such a way that he can only just break even.

(ii) He maximizes the expected discounted sum of the prevailing charges that he pays by always continuing the bandit with the greatest prevailing charge.

This is because he will thereby interleave the n nonincreasing sequences of prevailing charges into one nonincreasing sequence of prevailing charges. This way of interleaving them maximizes their discounted sum.

(iii) Using this strategy he also just breaks even; so this strategy, (of always continuing the bandit with the greatest $g_i(x_i)$), must maximize the expected discounted sum of the rewards that he can obtain from this SFAB. □

6.5 Calculation of the Gittins index

How can we compute the Gittins index value for each possible state of a bandit process B_i? The input is the data of $r_i(\cdot)$ and $P_i(\cdot,\cdot)$. If the state space of B_i is finite, say $E_i = \{1,\ldots,k_i\}$, then the Gittins indices can be computed in an iterative fashion. First we find the state of greatest index, say 1 such that $1 = \arg\max_j r_i(j)$. Having found this state we can next find the state of second-greatest index. If this is state j, then $G_i(j)$ is computed in (6.2) by taking τ to be the first time that the state is not 1. This means that the second-best state is the state j which maximizes

$$\frac{E[r_i(j) + \beta r_i(1) + \cdots + \beta^{\tau-1} r_i(1)]}{E[1 + \beta + \cdots + \beta^{\tau-1}]},$$

where τ is the time at which, having started at $x_i(0) = j$, we have $x_i(\tau) \neq 1$. One can continue in this manner, successively finding states and their Gittins indices, in decreasing order of their indices. If B_i moves on a finite state space of size k_i then its Gittins indices (one for each of the k_i states) can be computed in time $O(k_i^3)$.

If the state space of a bandit process is infinite, as in the case of the Bernoulli bandit, there may be no finite calculation by which to determine the Gittins indices for all states. In this circumstance, we can approximate the Gittins index using something like the value iteration algorithm. Essentially, one solves a problem of maximizing right hand side of (6.2), subject to $\tau \leq N$, where N is large.

6.6 Forward induction policies

If we put $\tau = 1$ on the right hand side of (6.2) then it evaluates to $Er_i(x_i(t))$. If we use this as an index for choosing between projects, we have a **myopic policy** or **one-step-look-ahead policy**. The Gittins index policy generalizes the idea of a one-step-look-ahead policy, since it looks-ahead by some optimal time τ, so as to maximize, on the right hand side of (6.2), a measure of the rate at which reward can be accrued. This defines a so-called **forward induction policy**.

7 Average-cost Programming

The average-cost optimality equation. Policy improvement algorithm.

7.1 Average-cost optimality equation

Suppose that for a stationary Markov policy π, the following limit exists:

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

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

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

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

$$F_s(x) = \min_u \{c(x, u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\}. \tag{7.1}$$

So by substituting $F_s(x) \sim s\lambda + \phi(x)$ into (7.1), 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 that the average-cost optimality equation should be:

$$\lambda + \phi(x) = \min_u \{c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u]\}. \tag{7.2}$$

Theorem 7.1. *Suppose there exists a constant λ and bounded function ϕ satisfying (7.2). Let π be the policy which in each state x chooses u to minimize the right hand side. Then λ is the minimal average-cost and π is the optimal stationary policy.*

Proof. Suppose u is chosen by some policy π'. By repeated substitution of (7.2) into itself we have

$$\phi(x) \le -t\lambda + E_{\pi'} \left[\sum_{\tau=0}^{t-1} c(x_\tau, u_\tau) \,\middle|\, x_0 = x \right] + E_{\pi'}[\phi(x_t) \mid x_0 = x].$$

with equality if $\pi' = \pi$. Divide this by t and let $t \to \infty$. Boundedness of ϕ ensures that $(1/t)E_{\pi'}[\phi(x_t) \mid x_0 = x] \to 0$. So we obtain

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

with equality if $\pi' = \pi$. □

So an average-cost optimal policy can be found by looking for a bounded solution to (7.2). Notice that if ϕ is a solution of (7.2) then so is ϕ+(a constant), because the (a constant) will cancel from both sides of (7.2). Thus ϕ is undetermined up to an additive constant. In searching for a solution to (7.2) we can therefore pick any state, say $\bar{x}$, and arbitrarily take $\phi(\bar{x}) = 0$. The function ϕ is called the **relative value function**.

7.2 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 (7.2) is

$$\begin{aligned}\lambda + \phi(0) &= \sum_{i=1}^{n} a_i \max[\phi(0), \phi(i)],\\ \lambda + \phi(i) &= (1-p_i)\phi(i) + p_i[R_i + \phi(0)], \quad i = 1, \ldots, n.\end{aligned}$$

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* $p_i R_i \geq \lambda^*$.

7.3 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. For the rest of this lecture we suppose the state space is finite.

Theorem 7.2. *Define*

$$m_s = \min_x \{F_s(x) - F_{s-1}(x)\}, \qquad M_s = \max_x \{F_s(x) - F_{s-1}(x)\}. \tag{7.3}$$

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

Proof. Suppose the optimal policy over a time horizon of s steps starts by taking $u_0 = f_s(x_0)$. If we were to use this as a stationary Markov policy, say $\pi_s = f_s^\infty$, we would obtain some average-cost, say λ_s. Suppose $\pi = f^\infty$ is the average-cost optimal policy over the infinite horizon, taking $u = f(x)$, with average-cost λ ($\leq \lambda_s$). Then

$$\begin{aligned} F_{s-1}(x) + [F_s(x) - F_{s-1}(x)] = F_s(x) &= c(x, f_s(x)) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = f_s(x)] \\ &\leq c(x, f(x)) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = f(x)] \end{aligned}$$

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

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

We substitute this into itself $t-1$ times to get

$$F_{s-1}(x) \leq -m_s t + E_\pi \left[\sum_{\tau=0}^{t-1} c(x_\tau, u_\tau) \,\middle|\, x_0 = x \right] + E_\pi[F_{s-1}(x_t) \mid x_0 = x].$$

Divide by t and let $t \to \infty$ to get $m_s \leq \lambda$. A bound $\lambda_s \leq M_s$ is found similarly using

$$\begin{aligned} F_{s-1}(x) &\geq -M_s + c(x,u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = f_s(x)]. \\ &\geq -M_s t + E_{\pi_s} \left[\sum_{\tau=0}^{t-1} c(x_\tau, u_\tau) \,\middle|\, x_0 = x \right] + E_{\pi_s}[F_{s-1}(x_t) \mid x_0 = x]. \end{aligned}$$

□

This justifies use of a **value iteration algorithm** in which we calculate F_s until $M_s - m_s \leq \epsilon m_s$. At that point the stationary policy f_s^∞ achieves an average-cost that is within $\epsilon \times 100\%$ of optimal.

7.4 Policy improvement algorithm

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

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

and there exists a policy $\pi_1 = f_1^\infty$ such that

$$\lambda + \phi(x) \geq c(x, f_1(x_0)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_1(x_0)], \tag{7.4}$$

for all x, and with strict inequality for some x (and thus $f_1 \neq f$). Then following the lines of proof in Theorems 7.1 and 7.2 (repeatedly substituting (7.4) into itself),

$$\lambda \geq \lim_{t\to\infty} \frac{1}{t} E_{\pi_1} \left[\sum_{\tau=0}^{t-1} c(x_\tau, u_\tau) \,\middle|\, x_0 = x \right]. \tag{7.5}$$

So π_1 is at least as good as π. If there is no π_1 then π satisfies (7.2) and so π is optimal. This justifies the following **policy improvement algorithm**

(0) Choose an arbitrary stationary policy π_0. Set $s = 1$.

(1) For 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 (7.2). The average-cost of π_{s-1} is λ.

(2) Now determine the policy $\pi_s = f_s^{\infty}$ from

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

taking $f_s(x) = f_{s-1}(x)$ whenever this is possible. If $\pi_s = \pi_{s-1}$ then we have a solution to (7.2) and so π_{s-1} is optimal. Otherwise π_s is a new policy. By the calculation in (7.5) this has an average-cost no more than λ, so π_s is at least as good as π_{s-1}, We now 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 must find an optimal stationary policy in finitely many iterations. By contrast, the value iteration algorithm only obtains increasingly accurate approximations of λ^*.

Example 7.3. Consider again the example of §7.2. Let us start with a policy π_0 which accept only jobs of type 1. The average-cost of this policy can be found by solving

$$\lambda + \phi(0) = a_1\phi(1) + \sum_{i=2}^{n} a_i\phi(0),$$

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

The solution is $\lambda = a_1 p_1 R_1/(a_1 + p_1)$, $\phi(0) = 0$, $\phi(1) = p_1 R_1/(a_1 + p_1)$, and $\phi(i) = R_i - \lambda/p_i$, $i \geq 2$. The first use of step (1) of the policy improvement algorithm will create a new policy π_1, which improves on π_0, by accepting jobs for which $\phi(i) = \max\{\phi(0), \phi(i)\}$, i.e. for which $\phi(i) = R_i - \lambda/p_i > 0 = \phi(0)$.

If there are no such i then π_0 is optimal. So we may conclude that π_0 is optimal if and only if $p_i R_i \leq a_1 p_1 R_1/(a_1 + p_1)$ for all $i \geq 2$.

Policy improvement in the discounted-cost case.

In the case of strict discounting the policy improvement algorithm is similar:

(0) Choose an arbitrary stationary policy π_0. Set $s = 1$.

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

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

(2) Now determine the policy $\pi_s = f_s^{\infty}$ from

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

taking $f_s(x) = f_{s-1}(x)$ whenever this is possible. Stop if $f_s = f_{s-1}$. Otherwise return to step (1) with $s := s + 1$.

8 LQ Regulation

LQ regulation model in discrete and continuous time. Riccati equation, and its validity in the model with additive white noise.

8.1 The LQ regulation problem

As we have seen, the elements of 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 illuminates 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

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

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

$$\mathbf{C} = \sum_{t=0}^{h-1} c(x_t, u_t) + \mathbf{C}_h(x_h), \tag{8.2}$$

with one-step and terminal costs

$$c(x, u) = x^\top Rx + u^\top Sx + x^\top S^\top u + u^\top Qu = \begin{pmatrix} x \\ u \end{pmatrix}^\top \begin{pmatrix} R & S^\top \\ S & Q \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}, \tag{8.3}$$

$$\mathbf{C}_h(x) = x^\top \Pi_h x. \tag{8.4}$$

All quadratic forms are non-negative definite ($\succeq 0$), and Q is positive definite ($\succ 0$). 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 8.1. *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 8.2. *Assume the structure of* (8.1)–(8.4)*. Then the value function has the quadratic form*

$$F(x,t) = x^\top \Pi_t x, \quad t \le h, \tag{8.5}$$

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*

$$\Pi_t = f\Pi_{t+1}, \quad t < h, \tag{8.6}$$

where Π_h *has the value given in* (8.4)*, and* f *is an operator having the action*

$$f\Pi = R + A^\top \Pi A - (S^\top + A^\top \Pi B)(Q + B^\top \Pi B)^{-1}(S + B^\top \Pi A). \tag{8.7}$$

The $m \times n$ *matrix* K_t *is given by*

$$K_t = -(Q + B^\top \Pi_{t+1} B)^{-1}(S + B^\top \Pi_{t+1} A), \quad t < h. \tag{8.8}$$

Proof. Assertion (8.5) is true at time h. Assume it is true at time $t+1$. Then

$$\begin{aligned} F(x,t) &= \inf_u \left[c(x,u) + (Ax + Bu)^\top \Pi_{t+1}(Ax + Bu)\right] \\ &= \inf_u \left[\begin{pmatrix} x \\ u \end{pmatrix}^\top \begin{pmatrix} R + A^\top \Pi_{t+1} A & S^\top + A^\top \Pi_{t+1} B \\ S + B^\top \Pi_{t+1} A & Q + B^\top \Pi_{t+1} B \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \right]. \end{aligned}$$

Lemma 8.1 shows the minimizer is $u = K_t x$, and gives the form of f. □

8.2 The Riccati recursion

The backward recursion (8.6)–(8.7) is called the **Riccati equation**.

(i) Since the optimal control is linear in the state, say $u = Kx$, an equivalent expression for the Riccati equation is

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

(ii) The optimally controlled process obeys $x_{t+1} = \Gamma_t x_t$. We call Γ_t the **gain matrix** and it 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) S can be normalized to zero by choosing a new control $u^* = u + Q^{-1}Sx$, and setting $A^* = A - BQ^{-1}S$, $R^* = R - S^\top Q^{-1} S$.

(iv) Similar results are true 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.

(v) It is also possible to analyse models in which

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

for a known sequence of disturbances $\{\alpha_t\}$, or in which the cost function is

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

so that the aim is to track a sequence of values $(\bar{x}_t, \bar{u}_t)$, $t = 0, \ldots, h-1$.

8.3 White noise disturbances

Suppose the plant equation (8.1) is now

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

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 dynamic programming equation is then

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

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

$$\begin{aligned} F(x, t) &= \inf_u \left[c(x, u) + E(Ax + Bu + \epsilon)^\top \Pi_{t+1}(Ax + Bu + \epsilon) + \gamma_{t+1} \right] \\ &= \inf_u \left[c(x, u) + E(Ax + Bu)^\top \Pi_{t+1}(Ax + Bu) \right] \\ &\qquad + 2E\left[\epsilon^\top (Ax + Bu)\right] + E\left[\epsilon^\top \Pi_{t+1} \epsilon\right] + \gamma_{t+1} \\ &= \inf_u [\cdots] + 0 + \mathrm{tr}(N\Pi_{t+1}) + \gamma_{t+1}, \end{aligned}$$

where $\mathrm{tr}(A)$ means the trace of matrix A. Here we use the fact that

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

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

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

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.

8.4 LQ regulation in continuous-time

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

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

We can obtain the continuous-time solution from the discrete time solution by moving forward in time in increments of Δ. Make the following replacements.

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

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 slightly 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

$$\Gamma(t) = A + BK = A - BQ^{-1}(S + B^\top \Pi).$$

9 Controllability

Controllability in discrete and continuous time. Linearization of nonlinear models.

9.1 Controllability

Consider the $[A, B, \cdot]$ system with plant equation $x_{t+1} = Ax_t + Bu_t$. The **controllability** question is: can we bring x to an arbitrary prescribed value by some u-sequence?

Definition 9.1. 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 **controllable** if it is r-controllable for some r

Example 9.2. 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 9.3. 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 the plant equation into itself, we see that we must find $u_0, u_1, \ldots, u_{r-1}$ to satisfy

$$\Delta = x_r - A^r x_0 = Bu_{r-1} + ABu_{r-2} + \cdots + A^{r-1}Bu_0, \tag{9.1}$$

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

Theorem 9.4. (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

$$\sum_{j=0}^{n} a_j A^{n-j} = 0. \tag{9.2}$$

The implication is that $I, A, A^2, \ldots, A^{n-1}$ contains basis for A^r, $r = 0, 1, \ldots$. We are now in a position to characterise controllability.

Theorem 9.5. *(i) The system $[A, B, \cdot]$ is r-controllable if and only if the matrix*

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

has rank n, (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 (9.1) has a solution for arbitrary Δ if and only if M_r has rank n.

(ii) That is, if and only if there does not exist nonzero w such that $w^\top M_r = 0$. Note that

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

(iii) The rank of M_r is non-decreasing in r, so if the system is r-controllable, it is $(r+1)$-controllable. By the Cayley-Hamilton theorem, the rank is constant for $r \geq n$.

(iv) Consider the Lagrangian

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

giving

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

Now we can determine λ from (9.1). □

9.2 Controllability in continuous-time

Theorem 9.6. *(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*

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

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 $t \downarrow 0$.

9.3 Linearization of nonlinear models

Linear models 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 bring the system back to $(\bar{x}, \bar{u})$ from any nearby starting point.

9.4 Example: broom balancing

Consider the problem of balancing a broom in an upright position on your hand. By Newton's laws, the system obeys $m(\ddot{u}\cos\theta + L\ddot{\theta}) = mg\sin\theta$.

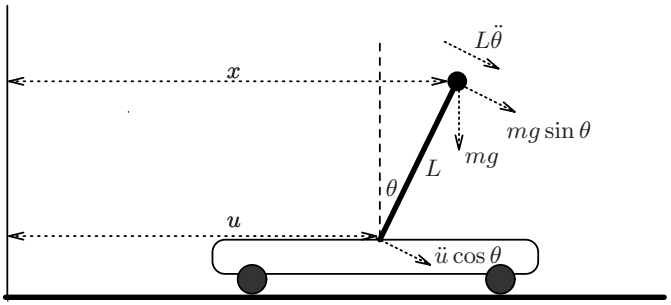


Figure 1: Force diagram for broom balancing

For small θ we have $\cos\theta \sim 1$ and $\theta \sim \sin\theta = (x - u)/L$. So with $\alpha = g/L$

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

equivalently,

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

Since

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

the system is controllable if θ is initially small.

9.5 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 $u_r = u_\theta = 0$ then there is a possible equilibrium in which the orbit is a circle of radius $r = \rho$, $\dot{\theta} = \omega = \sqrt{c/\rho^3}$ and $\dot{r} = \ddot{\theta} = 0$.

Consider a perturbation of this orbit and measure the deviations from the orbit by

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

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.

Suppose $u_r = 0$ (radial thrust fails). Then

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1/\rho \end{bmatrix} \quad M_4 = \begin{bmatrix} B & AB & A^2B & A^3B \end{bmatrix} = \begin{bmatrix} 0 & 0 & 2\omega & 0 \\ 0 & 2\omega & 0 & -2\omega^3 \\ 0 & 1/\rho & 0 & -4\omega^2/\rho \\ 1/\rho & 0 & -4\omega^2/\rho & 0 \end{bmatrix}.$$

which is of rank 4, so the system is still controllable. We can change the radius by tangential braking or thrust.

But if $u_\theta = 0$ (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 only rank 3. In fact, the uncontrollable component is the angular momentum, $2\omega\rho\delta r + \rho^2\delta\dot{\theta} = \delta(r^2\dot{\theta})|_{r=\rho,\dot{\theta}=\omega}$.

10 Stabilizability and Observability

Stabilizability. LQ regulation problem over the infinite horizon. Observability.

10.1 Stabilizability

Suppose we apply the stationary closed-loop control $u = Kx$ so that $\dot{x} = Ax + Bu = (A + BK)x$. So with $\Gamma = A + BK$, we have

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

Similarly, in discrete-time, we have can take the stationary control, $u_t = Kx_t$, so that $x_t = Ax_{t-1} + Bu_{t-1} = (A + BK)x_{t-1}$. Now $x_t = \Gamma^t x_0$.

We are interested in choosing Γ so that $x_t \to 0$ and $t \to \infty$.

Definition 10.1.

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

Γ is a **stability matrix** in the discrete-time sense if all its eigenvalues of lie strictly inside the unit disc in the complex plane, $|z| = 1$, and hence $x_t \to 0$ as $t \to \infty$.

The $[A, B]$ system is said to **stabilizable** if there exists a K such that $A + BK$ is a stability matrix.

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

10.2 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

$$\ddot{\theta} = -(g/L)\sin\theta + u.$$

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

$$\frac{d}{dt}\begin{pmatrix}\theta\\ \dot{\theta}\end{pmatrix} = \begin{pmatrix}\dot{\theta}\\ -(g/L)\sin\theta + u\end{pmatrix} \sim \begin{pmatrix}0 & 1\\ -g/L & 0\end{pmatrix}\begin{pmatrix}\theta\\ \dot{\theta}\end{pmatrix} + \begin{pmatrix}0\\ 1\end{pmatrix} u.$$

Suppose we try to stabilise with a control that is a linear function of only θ (not $\dot{\theta}$), so $u = Kx = (-\kappa, 0)x = -\kappa\theta$. Then

$$\Gamma = A + BK = \begin{pmatrix}0 & 1\\ -g/L & 0\end{pmatrix} + \begin{pmatrix}0\\ 1\end{pmatrix}\begin{pmatrix}-\kappa & 0\end{pmatrix} = \begin{pmatrix}0 & 1\\ -g/L - \kappa & 0\end{pmatrix}.$$

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

10.3 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 10.2. *Suppose $\Pi_0=0$, $R\succeq 0$, $Q\succeq 0$ 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

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

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 7.

10.4 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

$$x_t = Ax_{t-1} + Bu_{t-1}, \tag{10.1}$$
$$y_t = Cx_{t-1}. \tag{10.2}$$

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

Definition 10.3. A system is said to be **r-observable** if x_0 can be inferred from knowledge of the observations $y_1, \dots, y_r$ and relevant control values $u_0, \dots, u_{r-2}$ for any initial x_0. An n-dimensional system is **observable** if it is r-observable for some r.

The notion of observability stands in dual relation to that of controllability; a duality that indeed persists throughout the subject.

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 10.4. *(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.

10.5 Observability in continuous-time

Theorem 10.5. *(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

$$\tilde{y}(t) = y(t) - \int_0^t C e^{A(t-s)} B u(s)\, ds.$$

11 Kalman Filter and Certainty Equivalence

More about observability. Least squares estimation and the LQG model. The Kalman filter. Certainty equivalence.

11.1 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

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

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

11.2 Example: satellite in planar orbit

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

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

By taking $C = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$ we see that the 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.

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

11.3 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

$$x_t = Ax_{t-1} + Bu_{t-1} + \epsilon_t, \tag{11.1}$$
$$y_t = Cx_{t-1} + \eta_t, \tag{11.2}$$

where ϵ_t is process noise. The state observations are degraded in that we observe only the p-vector $y_t = Cx_{t-1} + \eta_t$, where η_t is observation noise. Typically $p < n$. In this $[A, B, C]$ system A is $n \times n$, B is $n \times m$, and C is $p \times n$. Assume

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

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.

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

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

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

$$E(x \mid y) = V_{xy}V_{yy}^{-1}y, \tag{11.3}$$

and

$$cov(x \mid y) = V_{xx} - V_{xy}V_{yy}^{-1}V_{yx}. \tag{11.4}$$

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 (11.4) □

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.

11.4 The Kalman filter

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* (11.1)–(11.2)*. 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*

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

where the time-dependent matrix V_t satisfies a Riccati equation

$$V_t = gV_{t-1}, \quad t < h,$$

where V_0 is given, and g is the operator having the action

$$gV = N + AVA^\top - (L + AVC^\top)(M + CVC^\top)^{-1}(L^\top + CVA^\top). \tag{11.6}$$

The $p \times m$ matrix H_t is given by

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

Compare this to the very similar statement of Theorem 8.2. Notice that (11.6) computes V_t forward in time ($V_t = gV_{t-1}$), whereas (8.7) computes Π_t backward in time ($\Pi_t = f\Pi_{t+1}$).

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

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

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**

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

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

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

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). □

11.5 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)*

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

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 8.2.

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 $c(x,u) = x^\top Rx + 2u^\top Sx + u^\top Qu$, we have

$$\begin{aligned}
F(W_{t-1}) &= \min_{u_{t-1}} E\left[c(x_{t-1}, u_{t-1}) + \hat{x}_t \Pi_t \hat{x}_t + \cdots \mid W_{t-1}, u_{t-1}\right] \\
&= \min_{u_{t-1}} E\Bigg[c(\hat{x}_{t-1} - \Delta_{t-1}, u_{t-1}) \\
&\qquad + (A\hat{x}_{t-1} + Bu_{t-1} + H_t\zeta_t)^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1} + H_t\zeta_t) \\
&\qquad + \cdots \,\Big|\, W_{t-1}, u_{t-1}\Bigg] \qquad (11.9) \\
&= \min_{u_{t-1}} \left[c(\hat{x}_{t-1}, u_{t-1}) + (A\hat{x}_{t-1} + Bu_{t-1})^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1})\right] + \cdots ,
\end{aligned}$$

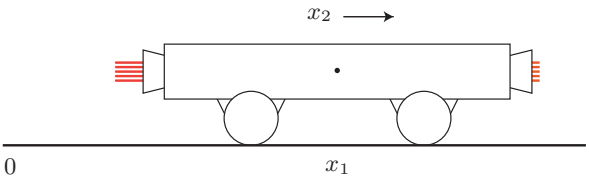
where we use the fact that, conditional on W_{t-1}, u_{t-1}, the quantities Δ_{t-1} and ζ_t have zero means and are policy independent. So when we evalute (11.9) the expectations of all terms which are linear in these quantities are zero, like $E[u_{t-1}^\top S\Delta_{t-1}]$, and the expectations of all terms which are quadratic in these quantities, like $E[\Delta_{t-1}^\top R\Delta_{t-1}]$, are policy independent (and so may be included as part of $+\cdots$). □

12 Dynamic Programming in Continuous Time

The HJB equation for dynamic programming in continuous time.

12.1 Example: parking a rocket car

A car has rocket engines at both ends. Its initial position and velocity are $x_1(0)$ and $x_2(0)$.



By firing the rockets (causing acceleration of u in the forward or reverse direction) we wish to park the car in minimum time, i.e. minimize T such that $x_1(T) = x_2(T) = 0$. The dynamics are $\dot{x}_1 = x_2$ and $\dot{x}_2 = u$, where u is constrained by $|u| \leq 1$.

Let $F(x)$ be minimum time that is required to park the car. Then

$$F(x_1, x_2) = \min_{-1 \leq u \leq 1} \Big\{ \delta + F((x_1, x_2) + (x_2, u)\delta) \Big\}.$$

By making a Taylor expansion and then letting $\delta \to 0$ we find

$$0 = \min_{-1 \leq u \leq 1} \left\{ 1 + \frac{\partial F}{\partial x_1} x_2 + \frac{\partial F}{\partial x_2} u \right\} \tag{12.1}$$

with the boundary condition $F(0, 0) = 0$. This is a difficult equation to solve, but we can at least see that the optimal control will be a **bang-bang control** with $u = -\operatorname{sign}(\frac{\partial F}{\partial x_2})$ and so F satisfies

$$0 = 1 + \frac{\partial F}{\partial x_1} x_2 - \left| \frac{\partial F}{\partial x_2} \right|.$$

12.2 The Hamilton-Jacobi-Bellman equation

We generalize the above. In continuous time the plant equation is,

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

Consider a discounted cost of

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

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 of order δ in the Taylor series expansion we obtain,

$$\inf_u \left[c(x,u,t) - \alpha F + \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} a(x,u,t) \right] = 0, \quad t < h, \tag{12.2}$$

with $F(x,h) = \mathbf{C}(x,h)$. In the undiscounted case, we simply put $\alpha = 0$. Notice that in (12.1) we have $\alpha = 0$ and the term of $\frac{\partial F}{\partial t}$ disappears because $h = \infty$.

Equation (12.2) is called the **Hamilton-Jacobi-Bellman equation** (HJB). Its heuristic derivation we have given above is justified by the following theorem.

Theorem 12.1. *Suppose a policy π, using a control u, has a value function F which satisfies the HJB equation* (12.2) *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

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

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

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

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

12.3 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 $2x^\top \Pi A x = x^\top \Pi A x + x^\top A^\top \Pi x$. This must hold for all x. We have a solution to the HJB equation if $\Pi(t)$ satisfies the Riccati differential equation

$$R + \Pi A + A^\top \Pi - \Pi B Q^{-1} B^\top \Pi + \frac{d\Pi}{dt}$$

with a given boundary value for $\Pi(h)$.

12.4 Example: harvesting fish

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 $\int_0^T ue^{-\alpha t}\,dt$, subject to $0 \le u \le u_{\max}$, where $u_{\max}$ is the greatest possible fishing rate.

Solution. The DP equation (with discounting) is

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

Since u occurs linearly with the maximization we again have a bang-bang optimal control, of the form

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

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

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

Let us make a guess that $F(x)$ is concave, and then deduce that

$$u = \begin{bmatrix} 0 \\ \text{undetermined, but effectively } a(\bar{x}) \\ u_{\max} \end{bmatrix} \text{ for } x \begin{bmatrix} < \\ = \\ > \end{bmatrix} \bar{x}. \tag{12.4}$$

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 will switch 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

$$F(\bar{x}) = \int_0^\infty e^{-\alpha t} a(\bar{x}) dt = a(\bar{x})/\alpha. \tag{12.5}$$

So from (12.3) and (12.5) we have

$$F_x(x) = \frac{\alpha F(x) - u(x)}{a(x) - u(x)} \to 1 \text{ as } x \nearrow \bar{x} \text{ or } x \searrow \bar{x}. \tag{12.6}$$

For F to be concave, F_{xx} must be negative if it exists. So we must have

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

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

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

where the last line follows because (12.6) 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 $a'(\bar{x}) = \alpha$. Choosing this as our $\bar{x}$, we can verify that $F(x)$ is concave, as we conjectured from the start.

We now have the complete solution. The control in (12.4) has a value function F which satisfies the HJB equation.

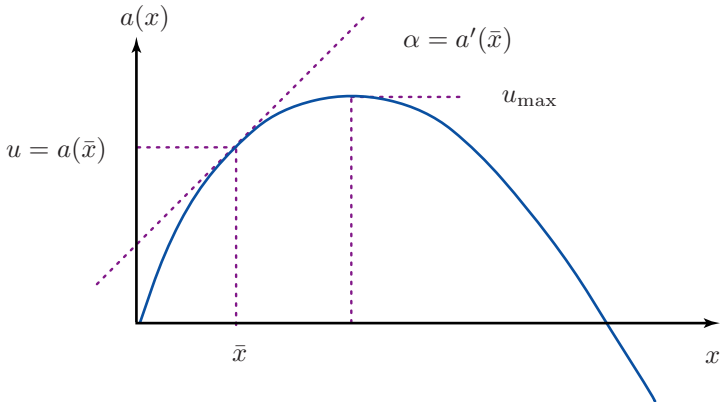


Figure 2: Growth rate $a(x)$ subject to environment pressures

Notice that we sacrifice long term yield for immediate return. If the initial population is greater than $\bar{x}$ then the optimal policy is to fish at rate $u_{\max}$ until we reach $\bar{x}$ and then fish at rate $u = a(\bar{x})$. As $\alpha \nearrow a'(0)$, $\bar{x} \searrow 0$. If $\alpha \geq a'(0)$ then it is optimal to wipe out the entire fish stock.

13 Pontryagin's Maximum Principle

Pontryagin's maximum principle. Transversality conditions. Rocket car example.

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 dynamic programming approach is a calculation for a general initial value of the state. Thus, when PMP is useful, it finds an open-loop prescription of the optimal control, whereas dynamic programming is useful for finding a closed-loop prescription. PMP can be used as both a computational and analytic technique (and in the second case can solve the problem for general initial value.)

We begin by considering a time-homogenous formulation, with plant equation $\dot{x} = a(x, u)$ and instantaneous cost $c(x, u)$. The trajectory is to be controlled until it reaches some stopping set S, where there is a terminal cost $K(x)$. As in (12.2) the value function $F(x)$ obeys the a dynamic programming equation (without discounting)

$$\inf_{u \in \mathcal{U}} \left[c(x, u) + \frac{\partial F}{\partial x} a(x, u) \right] = 0, \quad x \notin S, \tag{13.1}$$

with terminal condition

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

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 as the state moves along the optimal trajectory. The proof that F_x exists in the required sense is actually a tricky technical matter. We 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*

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

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

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. Assuming u is the optimal control we have from (13.1) in incremental form as

$$F(x,t) = c(x,u)\delta + F(x + a(x,u)\delta, t+\delta) + o(\delta).$$

Now use the chain rule to differentiate with respect to x_i and this yields

$$\begin{aligned}
&\frac{d}{dx_i}F(x,t) = \delta\frac{d}{dx_i}c(x,u) + \sum_j \frac{\partial}{\partial x_j}F(x + a(x,u)\delta, t+\delta)\frac{d}{dx_i}(x_j + a_j(x,u)\delta) \\
\implies \quad &-\lambda_i(t) = \delta\frac{dc}{dx_i} - \lambda_i(t+\delta) - \delta\sum_j \lambda_j(t+\delta)\frac{da_j}{dx_i} + o(\delta) \\
\implies \quad &\frac{d}{dt}\lambda_i(t) = \frac{dc}{dx_i} - \sum_j \lambda_j(t)\frac{da_j}{dx_i}
\end{aligned}$$

which is (13.6). □

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 $\lambda(t)$.

13.2 Example: parking a rocket car (continued)

In the example of §12.1, $\dot{x}_1 = x_2$, $\dot{x}_2 = u$, $|u| \le 1$, and we sought to minimize

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

where T is the first time at which $x = (0,0)$. 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 $\dot{\lambda}_i = -\partial H/\partial x_i$, so

$$\dot{\lambda}_1 = 0, \qquad \dot{\lambda}_2 = -\lambda_1. \tag{13.8}$$

Suppose that at termination $\lambda_1 = \alpha$, $\lambda_2 = \beta$. Then in terms of time to go we can compute

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

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.

From (13.1) we see that the maximized value of H must be 0. So at termination (when $x_2 = 0$), we conclude that we must have $|\beta| = 1$. 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** $x_1 = \pm x_2^2/2$ (shown dotted in Figure 3).

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,

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

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.

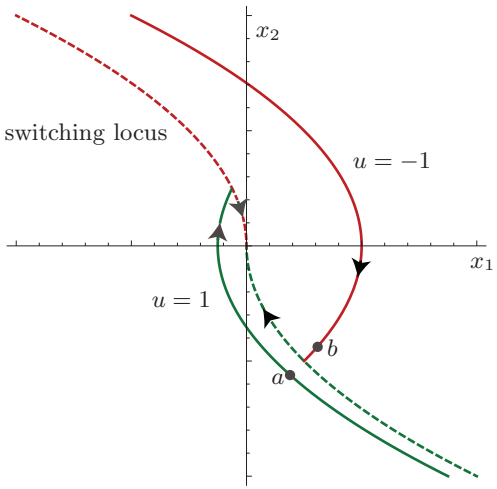


Figure 3: Optimal trajectories for parking a rocket car

Notice that the path is sensitive to the initial conditions. Consider the two points, a and b, just either side of the switching locus. From a we take $u = 1$ then $u = -1$. From b we first take $u = -1$, then $u = 1$.

13.3 Transversality conditions

In the above analysis we appealed to (13.1) to see that H must be maximized to 0. We can make this a generally valid assertion, and also say some things about the terminal value of $\lambda(T)$ (the so-called **transversality conditions**.)

Theorem 13.2. *(i) $H = 0$ on the optimal path. (ii) The terminal condition*

$$(\lambda + K_x)^\top \sigma = 0 \tag{13.9}$$

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). To see (ii), suppose that x is a point at which the optimal trajectory first enters S. Then $x \in S$ and so $F(x) = K(x)$. Suppose $x + \epsilon\sigma + o(\epsilon) \in S$. Then

$$\begin{aligned} 0 &= F(x + \epsilon\sigma + o(\epsilon)) - K(x + \epsilon\sigma + o(\epsilon)) \\ &= F(x) - K(x) + (F_x(x) - K_x(x))^\top \sigma\epsilon + o(\epsilon) \end{aligned}$$

Together with $F(x) = K(x)$ this gives $(F_x - K_x)^\top \sigma = 0$. Since $\lambda = -F_x$ we get $(\lambda + K_x)^\top \sigma = 0$. □

13.4 Adjoint variables as Lagrange multipliers

We may also think of $\lambda(t)$ as a Lagrange multiplier for the constraint $\dot{x} = a(x, u)$ (at time t). Consider the Lagrangian

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

This is to be maximized over (x, u, λ) paths having the property that $x(t)$ first enters the set S at time T. We integrate $\lambda^\top \dot{x}$ by parts to obtain

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

We now think about varying both $x(t)$ and $u(t)$, but without regard to the constraint $\dot{x} = a(x, u)$. The quantitiy within the integral must be stationary with respect to $x = x(t)$ and hence $\dot{\lambda} + \lambda^\top a_x - c_x = 0 \implies \dot{\lambda} = -H_x$, i.e. (13.6).

The Lagrangian must also be stationary with respect to small variations in $x(T)$ that are in a direction σ such that $x(T) + \epsilon\sigma$ is in the stopping set (or within $o(\epsilon)$ of it), and this gives $(K_x(x(T)) + \lambda(T))^\top \sigma = 0$, i.e. (13.9).

It is good to have this alternative viewpoint, but it is informal and less easy to rigorise than the 'proofs' of (13.6) in §13.1, and (13.9) in §13.3

14 Applications of the Maximum Principle

Examples of typical arguments for synthesizing a solution to an optimal control problem by use of Pontryagin's maximum principle. Problems in which time t appears explicitly.

14.1 Example: a moon lander

Consider a moon lander whose state variables are

$$\begin{aligned} x &= \text{its height above the moon's surface} \\ v &= \text{its velocity (in the direction away from the moon)} \\ m &= \text{its mass} \end{aligned}$$

We can control a thrust force $u \in [0, 1]$ by burning rocket fuel. This slows the descent and also reduces the lander's mass. The dynamics are

$$\begin{aligned} \dot{x} &= v \\ m\dot{v} &= -mg + u \\ \dot{m} &= -ku \end{aligned}$$

where g is the gravitational constant of the moon, and k rate at which fuel mass is consumed to create unit thrust.

Suppose we wish to maximize the mass that we safely land on the moon (including unspent fuel). So we take a terminal cost of $K(x, v, m) = -m(T)$, where T is the time of a soft landing, which we specified as $x(T) = v(T) = 0$. Taking adjoint variables p, q, r, the Hamiltonian is

$$H = pv + q(-g + u/m) - rku = pv - qg + u(q/m - kr).$$

We know that u must be chosen in $[0, 1]$ to maximize H. Let $\Delta = q/m - kr$. So optimal $u = 0$ or $u = 1$ as $\Delta < 0$ or $\Delta > 0$, respectively. Thus u is a bang-bang control.

From PMP we have $\dot{p} = 0$, $\dot{q} = -p$ and $\dot{r} = uq/m^2$. From these we find $d\Delta/dt = -p/m$. Since p is a constant we may conclude that Δ is either increasing or decreasing in time. The only form of solution that could provide a soft landing is the one for which Δ is increasing; initially the lander is allowed to free fall ($u = 0$), until some time t^*, and then it is deaccelerated under maximum thrust ($u = 1$) so a soft landing is achieved. Once we know the solution must be of this form then it is not hard to work out what the value of t^* must be (which will depend on the initial values x, v, m). Of course for some initial conditions a soft landing is not possible. There is also a constraint $m \geq m_0$, where m_0 is the non-fuel mass of the lander.

Similarly, we might address the question of how to return the lander to its mother ship in the least time, or how to gain maximum height upon leaving the moon.

14.2 Example: use of transversality conditions

Suppose $\dot{x}_1 = x_2$, $\dot{x}_2 = u$, $x(0) = (0, 0)$, u is unconstrained, and cost to be minimized is

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

Here $K(x) = -x_1(1)$. 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

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

The terminal x is unconstrained so in the transversality condition of $(\lambda + K_x)^\top \sigma = 0$, σ is arbitrary and so we also have

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

Thus the solution must be $\lambda_1(t) = 1$ and $\lambda_2(t) = 1 - t$. The optimal control is $u(t) = 1 - t$.

Note that there is often more than one way to set up a control problem. In this problem, we might have taken $K = 0$, but included a cost of $-\int_0^1 x_2\, dt = -x_1(1) + x_1(0)$.

14.3 Problems in which time appears

Thus far, $c(\cdot)$, $a(\cdot)$ and $K(\cdot)$ have been function of (x, u), but not t. Sometimes we wish to solve problems in t appears, such as when $\dot{x} = a(x, u, t)$. We can cope with this generalization by the simple mechanism of introducing a new variable that equates to time. Let $x_0 = t$, with $\dot{x}_0 = a_0 = 1$.

Having been augmented by this variable, the Hamiltonian gains a term and becomes

$$\tilde{H} = \lambda_0 a_0 + H = \lambda_0 a_0 + \sum_{i=1}^{n} \lambda_i a_i - c$$

where $\lambda_0 = -F_t$ and $a_0 = 1$. Theorem 13.2 says that $\tilde{H}$ must be maximized to 0. Equvalently, on the optimal trajectory,

$$H(x, u, \lambda) = \sum_{i=1}^{n} \lambda_i a_i - c \text{ must be maximized to } -\lambda_0.$$

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

$$\dot{\lambda}_0 = -H_t = c_t - \lambda a_t, \tag{14.1}$$

and transversality condition

$$(\lambda + K_x)^\top \sigma + (\lambda_0 + K_t)\tau = 0, \tag{14.2}$$

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 time-dependacy and terminal conditions on $x(T)$ and/or T. For example, we can draw the following inferences.

(i) If K is time-independent (so $K_t = 0$) and the terminal time T is unconstrained (so τ is arbitrary) then the transversality condition implies $\lambda_0(T) = 0$. Since H is always maximized to $-\lambda_0(t)$ it must be maximized to 0 at T.

(ii) If a, c are only functions of (x, u) then $\dot\lambda_0 = c_t - \lambda^\top a_t = 0$, and so $\lambda_0(t)$ is constant on the optimal trajectory. Since H is always maximized to $-\lambda_0(t)$ it must be maximized to a constant on the optimal trajectory.

(iii) If both (i) and (ii) are true then H is maximized to 0 along the entire optimal trajectory. We had this in the problem of parking in minimal time, §13.2.

14.4 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, $\int_0^T e^{-\alpha t} up(u)\, dt$, (where T is unconstrained.)

Solution. Notice that t appears in the cost function. The plant equation is $\dot x = -u$ and the Hamiltonian is

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

Note that $K = 0$. 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 $\lambda_0(T) = -K_t|_{t=T} = 0$. Therefore, since we require H to be 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 the time at which all wine has been sold, and so

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

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

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

14.5 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. In this problem the Hamiltonian is

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

and $K(w,q) = -q$. The adjoint equations and transversality conditions give

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

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

$$u = \begin{array}{c} 0 \\ 1 \end{array} \quad \text{if} \quad \Delta(t) := \lambda_1 a - c \begin{array}{c} < \\ > \end{array} 0.$$

At $\Delta(T) = -c$, we must have $u(T) = 0$. If t is a little less than T, λ_1 is small and $u = 0$ so the equation for λ_1 is

$$\dot{\lambda}_1 = \lambda_1 b - c. \tag{14.3}$$

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 $\Delta(t) = \lambda_1 a - c \ge 0$. 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 (14.3) from t to T, to give $\lambda_1(t) = (c/b)(1 - e^{-(T-t)b})$ and so the switch occurs where $\lambda_1 a - c = 0$, i.e. $(a/b)(1 - e^{-(T-t)b}) = 1$, 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.

15 Controlled Markov Jump Processes

Control problems in a continuous-time stochastic setting. Markov jump processes 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

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

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**

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

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

$$E\big[\phi(x(t+\delta t)) \mid x(t)=j, u(t)=u\big]$$
$$= \sum_{k \neq j} q_{jk}(u)\phi(k)\delta t + \left(1 - \sum_{k \neq j} q_{jk}(u)\delta t\right)\phi(j) + o(\delta t),$$

whence it follows that

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

and the DP equation becomes

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

This is the optimality equation for a controlled **Markov jump process**, which evolves in time like this:

- The state is j. We choose some control, say u.
- After a time that is exponentialy distributed with parameter $q_j(u) = \sum_{k \neq j} q_{jk}(u)$, (i.e. having mean $1/q_j(u)$), the state jumps.
- The jump is to state k ($\neq j$) with probability $q_{jk}(u)/q_j(u)$.

15.3 Uniformization in the infinite horizon case

We now explain how (in the infinite horizon case) the continuous-time DP equation (15.2) may be rewritten to look like a discrete-time DP equation. After doing this, all the ideas of Lectures 1–7 can be applied. In the case of an infinite horizon we have $F_t(j,t) = 0$, and for discounting, as in (12.2), we must include a term of $-\alpha F(j)$. So the optimality equation becomes

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

Suppose we choose B large enough so that for all j and u,

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

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

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

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

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

where

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

This makes the dynamic programming equation look exactly like a case of discounted dynamic programming in discrete time. All the results we have for in lectures 1–7 can now be used (e.g., value iteration, OSLA rules, etc.) This trick is called **uniformization**. Its effect is to make the times between jumps i.i.d. (exponentially distributed).

- The state is j. We choose a control, u.
- After a time that is exponentialy distributed with parameter $B+\alpha$ (i.e. having mean $1/(B+\alpha)$) the state jumps.
- The jump is to
  - k ($\neq j$) with probability $q_{jk}(u)/(B+\alpha)$,
  - back to j, with probability $(B - q_j(u))/(B+\alpha)$, or
  - to a terminating state, with probability $\alpha/(B+\alpha)$.

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,

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

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

$$\phi(j) + \bar{\gamma} = \inf_u \left[\bar{c}(j,u) + \sum_k p_{jk}(u)\phi(k) \right], \tag{15.5}$$

which has exactly the same form as the discrete-time average-cost optimality equation of Lecture 7. The theorems and techniques of that lecture can now be applied.

15.4 Example: admission control at a queue

Consider a queue in which the number of customers waiting may be $0, 1, \ldots$, 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

$$\begin{aligned}\phi(0) + \gamma &= \inf_u[-Ru + u\phi(1) + (\mu + \lambda - u)\phi(0)],\\ &= \inf_u[u\{-R + \phi(1) - \phi(0)\} + (\mu + \lambda)\phi(0)],\end{aligned}$$

$$\begin{aligned}\phi(x) + \gamma &= \inf_u[ax - Ru + u\phi(x+1) + \mu\phi(x-1) + (\lambda - u)\phi(x)],\\ &= \inf_u[ax + u\{-R + \phi(x+1) - \phi(x)\} + \mu\phi(x-1) + \lambda\phi(x)], \quad x > 0.\end{aligned}$$

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 are 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

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**. It can be understood as a $\delta \to 0$ limit of a symmetric random walk in which steps $\pm\sqrt{\delta}$ are made at times $\delta, 2\delta, \ldots$.) The specification is internally consistent because, for example,

$$B(t) = B(t_1) + [B(t) - B(t_1)]$$

and for $0 \le t_1 \le t$ 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

$$\begin{aligned}\Lambda(u)\phi(x) &= \lim_{\delta t \downarrow 0} E\left[\left.\frac{\phi(x(t+\delta t)) - \phi(x)}{\delta t}\right| x(t) = x, u(t) = u\right] \\ &= \phi_x a + \tfrac{1}{2}\phi_{xx} g^2 \\ &= \phi_x a + \tfrac{1}{2} N \phi_{xx},\end{aligned}$$

where $N(x,u) = g(x,u)^2$. So this **controlled diffusion process** has DP equation

$$\inf_u \left[c + F_t + F_x a + \tfrac{1}{2} N F_{xx}\right] = 0, \tag{16.1}$$

and in the vector case

$$\inf_u \left[c + F_t + F_x^\top a + \tfrac{1}{2}\mathrm{tr}(N F_{xx})\right] = 0.$$

16.2 Example: noisy LQ regulation 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 12.3,

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

and also, as in Section 8.3,

$$\frac{d\gamma}{dt} + \mathrm{tr}(N\Pi(t)) = 0, \quad \text{giving} \quad \gamma(t) = \int_t^T \mathrm{tr}(N\Pi(\tau))\, d\tau.$$

16.3 Example: a noisy second order system

Consider a special case of LQ regulation:

$$\underset{u}{\text{minimize}}\, E\left[x(T)^2 + \int_0^T u(t)^2\, dt\right]$$

where for $0 \le t \le T$,

$$\dot{x}(t) = y(t) \quad \text{and} \quad \dot{y}(t) = u(t) + \epsilon(t),$$

$u(t)$ is the control variable, and $\epsilon(t)$ is **Gaussian white noise**,

Note that if we define $z(t) = x(t) + (T-t)y(t)$ then

$$\dot{z} = \dot{x} - y + (T-t)\dot{y} = (T-t)u + (T-t)\epsilon(t)$$

where $z(T) = x(T)$. Hence the problem can be posed only in terms of scalars u and z.

Recalling what we know about LQ models, let us conjecture that the optimality equation is of a form

$$V(z,t) = z^2 P_t + \gamma_t. \tag{16.2}$$

We could use (16.1). But let us argue from scratch. For (16.2) to work we will need

$$\begin{aligned} z^2 P_t + \gamma_t &= \min_u \left\{ u^2\delta + E\left[(z+\dot{z}\delta)^2 P_{t+\delta} + \gamma_{t+\delta}\right]\right\} \\ &= \min_u \left\{ u^2\delta + \left[z^2 + 2(T-t)zu\delta + (T-t)^2\delta\right] P_{t+\delta} + \gamma_{t+\delta}\right\} + o(\delta) \end{aligned}$$

The optimizing u is

$$u = -(T-t)P_{t+\delta} z.$$

Substituting this and letting $\delta \to 0$ we have

$$-z^2\dot{P}_t - \dot{\gamma}_t = -z^2(T-t)^2 P_t^2 + (T-t)^2 P_t.$$

Thus

$$-\dot{\gamma}_t = (T-t)^2 P_t$$

and

$$\dot{P}_t = (T-t)^2 P_t^2.$$

Using the boundary condition $P_T = 1$, we find that the solution to the above differential equation is

$$P_t = \left(1 + \tfrac{1}{3}(T-t)^3\right)^{-1},$$

and the optimal control is

$$u(t) = -(T-t)\left(1 + \tfrac{1}{3}(T-t)^3\right)^{-1} z(t).$$

16.4 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 $\frac{1}{2}\int_0^T (L + Qu^2)dt$, 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

$$\inf_u \left[\frac{L + Qu^2}{2} + u \frac{\partial F}{\partial x} \right] = 0, \quad 0 < x < 1, \tag{16.3}$$

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 (16.3) 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.

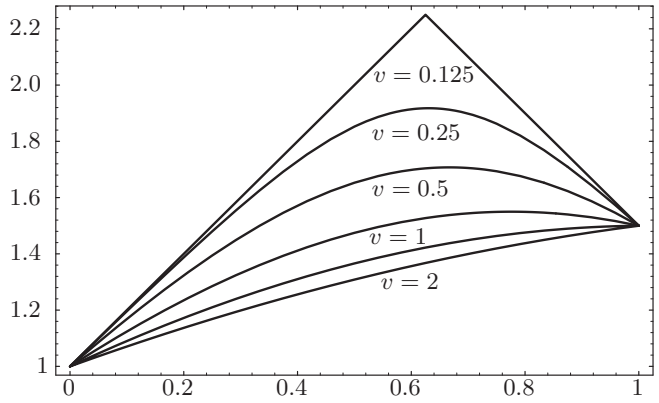


Figure 4: $F(x)$ against x for the passage to a stopping set

The figure 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.

Index

adjoint variable, 49
average-cost, 25

bandit process, 21
bang-bang control, 6, 45, 47, 53
Bellman equation, 3
Bernoulli bandit, 21
Brownian motion, 61

calibrating bandit process, 23
certainty equivalence, 44
chattering, 47
closed-loop, 3, 49
control theory, 1
control variable, 2
controllability, 33
controllable, 33
controlled diffusion process, 61, 62

decomposable cost, 4
deterministic stationary Markov policy, 17, 21
diffusion process, 61
discount factor, 9
discounted programming, 11
discounted-cost criterion, 9
discrete-time, 2
dynamic programming equation, 3

exploration and exploitation, 16

fair charge, 23
feedback, 3
finite actions, 14
forward induction policy, 24

gain matrix, 31
Gaussian white noise, 62
Gittins index, 22

Hamilton-Jacobi-Bellman equation, 46
Hamiltonian, 49

index policy, 22
individual optimality, 60
infinitesimal generator, 57
innovations, 43
interchange argument, 10

linear least squares estimate, 42
LQG model, 29

Markov decision problem, 5
Markov decision process, 4, 5
Markov dynamics, 4
Markov jump process, 58
Markov policy, 17
multi-armed bandit, 21
multi-armed bandit problem, 21
myopic policy, 16, 24

negative programming, 11

observability, 39
observable, 39
one-step look-ahead rule, 18, 24
open-loop, 3, 49
optimality equation, 3
optimization over time, 1

perfect state observation, 4
plant equation, 3
policy, 3
policy improvement algorithm, 27
Pontryagin's maximum principle, 49
positive programming, 11
power, 63
prevailing charge, 23
principle of optimality, 1
principle of optimality, 2

r-controllable, 33
r-observable, 39
regulation, 29
relative value function, 26

Riccati equation, 30, 31, 43

secretary problem, 7
separable cost function, 3
separation principle, 44
simple family of alternative bandit processes, 21
social optimality, 60
stability matrix, 37
stabilizable, 37
state variable, 3
stochastic differential equation, 61
stopping problem, 18
stopping time, 22
successive approximation, 14
switching locus, 51

time horizon, 2
time to go, 5
time-homogeneous, 5, 10
transition intensity, 57
transversality conditions, 51

uniformization, 59

value function, 6
value iteration, 14
value iteration algorithm, 27
value iteration bounds, 26

white noise, 31
Wiener process, 61