Dynamic Programming and Optimal Control

Richard Weber

Autumn 2014

Contents

1 Dynamic Programming

Dynamic programming and the principle of optimality. Notation for state-structured models. An example, with a bang-bang optimal control.

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 a near-optimal solution is adequate. Many real life problems do 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 explore a problem by following an optimization approach. If the 'optimal' solution is ridiculous then that 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 origins of 'control theory' can be traced to the wind vane used to face a windmill's rotor into the wind and the centrifugal governor invented by Jame Watt. Such 'classic control theory' is largely concerned with the question of stability, and much of this is outside this course, e.g., Nyquist criterion and dynamic lags. However, control theory is not merely concerned with the control of mechanisms. but it can be used to study a multitude of dynamical systems, in biology, telecommunications, manufacturing, heath services, finance, economics, etc.

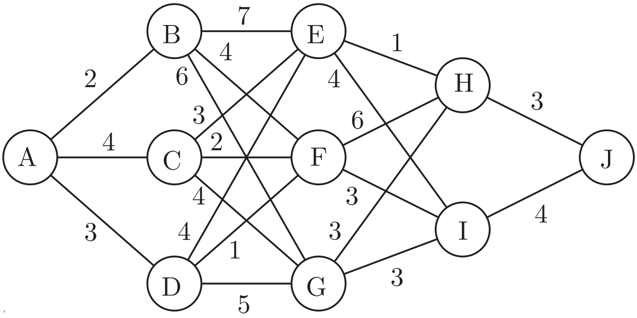
1.2 The principle of optimality

A key idea is that optimization over time can often be seen as 'optimization in stages'. We trade off cost incurred at the present stage against the implication this has for the least total cost that can be incurred from all future stages. The best action minimizes the sum of these two costs. 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 traveller wishes to minimize the length of a journey from town A to town J by first travelling 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).

Dynamic programming dates from Richard Bellman, who in 1957 wrote the first book on the subject and gave it its name.

1.4 The optimality equation

The optimality equation in the general case. In a **discrete-time** model, t takes integer values, $t = 0, 1, \ldots$. 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

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

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 $U_{t-1} = (u_0, \dots, u_{t-1})$, (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 $\{u_0, \dots, u_{h-1}\}$. 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 $\{u_t, \dots, u_{h-1}\}$ conditional on $x_t = x$,

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

Here $F(x,t)$ is the minimal future cost from time t onward, given that the state is x at time t. 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 $x_0, \dots, x_{t-1}$ are irrelevant.

The **optimality equation** (1.3) is also called the **dynamic programming equation** (DP) or **Bellman equation**.

1.5 Example: optimization of consumption

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

$$x_{t+1} = a(x_t, u_t) = x_t + \theta(x_t - u_t). \tag{1.4}$$

He desires to maximize total consumption over h years,

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

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

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

where $F_0(x) = 0$, (since nothing more can be consumed 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,

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

Next,

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

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

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

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 $0, \dots, h - s^* - 1$, (to build up capital) and then consume the whole of the income in years $h - s^*, \dots, h - 1$.

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) The dynamic programming equation my look messy. But try working backwards from $F_0(x)$, which is known. A pattern may emerge from which you can guess the general solution. You can then prove it correct by induction.

(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 Markov Decision Problems

Feedback, open-loop, and closed-loop controls. Markov decision processes and problems. Examples and some useful tricks. Secretary problem.

2.1 Features of the state-structured case

In the state-structured case the DP equation provides the optimal control in what is called **feedback** or **closed-loop** form, with $u_t = u(x_t, t)$. This contrasts with **open-loop** formulation in which $\{u_0, \ldots, u_{h-1}\}$ are to be chosen all at once at time 0.

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)

2.2 Markov decision processes

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. Assumption (a), below, defines what is known as a discrete-time **Markov decision process**. When we add assumption (b) and seek to minimize $\mathbf{C}$ then we have what is called a **Markov decision problem**. For both we use the abbreviation MDP.

(a) *Markov dynamics.* The stochastic version of the plant equation is

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

(b) *Separable (or decomposable) cost function.* Cost $\mathbf{C}$ is given by (1.1).

For the moment we also require the following:

(c) *Perfect state observation.* The current value of the state is always observable. That is, x_t is known when choosing u_t. So known fully at time t is $W_t = (X_t, U_{t-1})$.

Note that $\mathbf{C}$ is determined by W_h, so we might write $\mathbf{C} = \mathbf{C}(W_h)$.

As previously, 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 $u_0, \ldots, u_{h-1}$.

In general, 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.

The following theorem is then obvious.

Theorem 2.1. *$F(W_t)$ is a function of x_t and t alone, say $F(x_t, t)$. It obeys the optimality equation*

$$F(x_t, t) = \inf_{u_t} \left\{ c(x_t, u_t, t) + E[F(x_{t+1}, t+1) \mid x_t, u_t] \right\}, \quad t < h, \tag{2.1}$$

with terminal condition

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

Moreover, a minimizing value of u_t in (2.1) *(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

$$F(W_t) = \inf_{u_t} \{ c(x_t, u_t, t) + E[F(x_{t+1}, t+1) \mid X_t, U_t] \}. \tag{2.2}$$

But, by assumption (a), the right-hand side of (2.2) reduces to the right-hand member of (2.1). All the assertions then follow. □

2.3 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 $x \geq a_s$, where x is the current price and s is the number of days to go before expiry of the option.

Solution. The state at time t is, strictly speaking, x_t plus a variable to indicate 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 previously, we use time to go, $s = h - t$. So letting $F_s(x)$ 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 $F_s(x) = x - p$.

2.4 Example: secretary problem

Suppose we are to interview h candidates for a secretarial job. After seeing each candidate we must either hire or permanently reject her. Candidates are seen in random order and can be ranked against those seen previously. The aim is to maximize the probability of choosing the best candidate.

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

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

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

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

Let $F(t - 1)$ 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

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

Now t_0 is the smallest integer satisfying $F(t_0) \leq t_0/h$, or equilvalently

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

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 for 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

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

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 $F_n(\{1,2,\dots,n\})$. However, there is a simpler way.

An interchange argument

Suppose jobs are processed in the order $i_1, \dots, i_k, i, j, i_{k+3}, \dots, i_n$. Compare the reward that is obtained if the order of jobs i and j is reversed: $i_1, \dots, i_k, j, i, i_{k+3}, \dots, i_n$. 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**. The optimal policy we have obtained is an example of an **index policy**.

Note these points. (i) An interchange argument can be useful when a system evolves in stages. Although one might use 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 $F_0(x) = 0$.

Sometimes a nice way to write (4.2) is as $F_s = \mathcal{L} F_{s-1}$ where $\mathcal{L}$ is the operator with action

$$\mathcal{L}\phi(x) = \inf_u \{ c(x,u) + \beta E[\phi(x_1) \mid x_0 = x, u_0 = u] \}.$$

This operator transforms a scalar function of the state x to another scalar function of x. Note that $\mathcal{L}$ is a **monotone operator**, in the sense that if $\phi_1 \le \phi_2$ then $\mathcal{L}\phi_1 \le \mathcal{L}\phi_2$.

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**):	$0 < \beta \le 1$,	and $c(x,u) \ge 0$	for all x, u.
P (**positive programming**):	$0 < \beta \le 1$,	and $c(x,u) \le 0$	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 the intuitively obvious optimality equation (i.e. (3.7)). The theorem is obvious, but its proof is not.

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 '$\ge$' 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

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

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

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

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

3.5 Example: selling an asset

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

Solution. There are only two states, depending on whether she 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 she should sell when she can get at least βF^*. Her maximal reward is F^*.

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

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

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

If $\beta \le 1/2$ she 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 a way in which discounting can arise is often quite useful.

What if past offers remain open?

Suppose we change the problem so that past offers remain open. The state is now the best of the offers received so far, and the optimality equation is

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

However, the solution is exactly the same as before: sell at the first time an offer exceeds βF^*. Intuitively, this makes sense. If it is optimal to reject an offer then there is no reason we should wish to retrieve it later.

To prove this fact mathematically, we could consider the policy π which takes the same actions as does the policy that is optimal when offers do not remain open. Its value function is

$$F(x) = \begin{cases} x, & x \ge \beta F^* \\ \beta F^*, & x \le \beta F^*. \end{cases}$$

We can then carry through a calculation to show that this functions satisfies the optimality equation, and hence, by Theorem 4.2, it is optimal.

4 Positive Programming

Special theory for maximizing positive rewards. There may 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. Clinical trials.

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 $0, 1, 2, \ldots$ and in state x we may either move to state $x+1$ and receive no reward, or move to state 0, obtain reward $1 - 1/x$, and then remain in state 0 thereafter, 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 in case D

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

$$F_s(x) = \inf_u \{c(x, u) + \beta E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\}. \tag{4.2}$$

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

$$F_\infty(x) = \lim_{s \to \infty} F_s(x) = \lim_{s \to \infty} \inf_\pi F_s(\pi, x). \tag{4.3}$$

This limit 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, given any $\bar{\pi}$,

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

Taking the infimum over $\bar{\pi}$ gives

$$F_\infty(x) \leq F(x). \tag{4.4}$$

A nice way to write (4.2) is as $F_s = \mathcal{L}F_{s-1}$ where $\mathcal{L}$ is the operator with action

$$\mathcal{L}\phi(x) = \inf_u \{c(x,u) + \beta E[\phi(x_1) \mid x_0 = x, u_0 = u]\}.$$

This operator transforms a scalar function of the state x to another scalar function of x. Note that $\mathcal{L}$ is a **monotone operator**, in the sense that if $\phi_1 \leq \phi_2$ then $\mathcal{L}\phi_1 \leq \mathcal{L}\phi_2$.

The following theorem states that $\mathcal{L}^s(0) = F_s(x) \to F(x)$ as $s \to \infty$.

Theorem 4.3. *Suppose that D holds. Then* $\lim_{s\to\infty} F_s(x) = F(x)$.

Proof. Firstly, the optimal policy is certainly no more costly than a policy that minimizes the expected cost over the first s steps and then behaves arbitrarily thereafter, incurring an expected cost no more than $\beta^s B/(1-\beta)$. So

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

Secondly, under any policy π,

$$\begin{aligned} F(\pi, x) &= E_\pi(\text{cost over first } s \text{ steps}) + E_\pi(\text{cost over remaining steps}) \\ &\geq F_s(x) - \beta^s B/(1-\beta). \end{aligned}$$

It follows that $F_s(x) \to F(x)$, uniformly in x. □

To prove the theorem for the N and P cases, we need to add to (4.4) proofs that $F_\infty(x) \geq F(x)$. This is not examinable, but is provided for completeness in §4.6.

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

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

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

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

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 can approximate F using value iteration, finding $F \approx \mathcal{L}^n(0)$ for large n. An alternative, is the following.

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 then 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 be successful!

Here we see a tension between **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.

The above is an example of a **two-armed bandit problem** and a foretaste for Lecture 7 in which we will learn about the **multi-armed bandit problem** and how to optimally conduct trials amongst several alternative drugs.

4.6 *Value iteration in cases N and P*

This subsection, and others with headings enclosed as *...*, is not lectured or examinable, but is included for completeness.

For case N we need an additional assumption:

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

Theorem 4.4. *Suppose that P holds, or N and F hold. Then* $\lim_{s\to\infty} F_s(x) = F(x)$.

Proof. We have (4.4), so must prove '$\geq$'.
In case P, $c(x,u) \leq 0$, so $F_s(x) \geq F(x)$. Letting $s \to \infty$ proves the result.
In case N,

$$\begin{aligned} F_\infty(x) &= \lim_{s\to\infty} \min_u \{c(x,u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\} \\ &= \min_u \{c(x,u) + \lim_{s\to\infty} E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\} \\ &= \min_u \{c(x,u) + E[F_\infty(x_1) \mid x_0 = x, u_0 = u]\}, \end{aligned} \tag{4.5}$$

where the first equality is because the minimum is over a finite number of terms and the second equality is by Lebesgue monotone convergence, noting that $F_s(x)$ increases in s. Let π be the policy that chooses the minimizing action on the right hand side of (4.5). Then by substitution of (4.5) into itself, and the fact that N implies $F_\infty \geq 0$,

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

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

Remark. It is interesting that a D case can always be reduced to a P or N case. To see this, recall that in the D case, $|c(x,u)| < B$. Imagine subtracting $B > 0$ from every cost. This reduces the infinite-horizon cost under any policy by exactly $B/(1-\beta)$, and any optimal policy is unchanged. All costs are now negative, so we now have a P case. Similarly, adding B to every cost reduces a D case to an N case.

Thus Theorem 4.4 implies Theorem 4.3.

5 Negative Programming

Special theory for minimizing positive costs. The action that extremizes the right hand side of the optimality equation is optimal. Stopping problems and their solution.

5.1 Example: a partially observed MDP

Example 5.1. A hidden object moves between two location according to a Markov chain with probability transition matrix $P = (p_{ij})$. A search in location i costs c_i, and if the object is there it is found with probability α_i. The aim is to minimize the expected cost of finding the object.

This is example of what is called a **partially observable Markov decision process** (POMDP). In a POMDP the decision-maker cannot directly observe the underlying state. Instead, he must maintain a probability distribution over the set of possible states, based on his observations, and the underlying MDP. This distribution is updated using the usual Bayesian calculations.

Solution. Let x_i be the probability that the object is in location i (where $x_1 + x_2 = 1$). Value iteration of the dynamic programming equation is via

$$F_s(x_1) = \min\Bigg\{ c_1 + (1-\alpha_1 x_1) F_{s-1}\left(\frac{(1-\alpha_1)x_1 p_{11} + x_2 p_{21}}{1-\alpha_1 x_1}\right), \\ c_2 + (1-\alpha_2 x_2) F_{s-1}\left(\frac{(1-\alpha_2)x_2 p_{21} + x_1 p_{11}}{1-\alpha_2 x_2}\right)\Bigg\}.$$

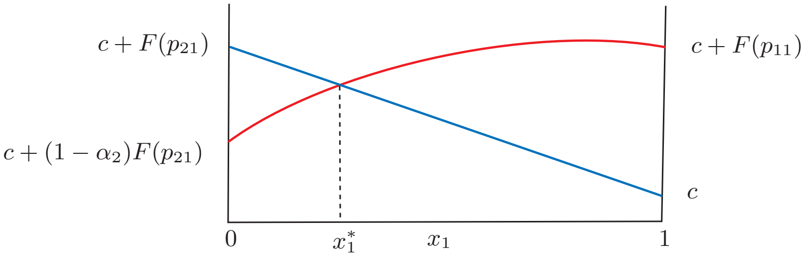
The arguments of $F_{s-1}(\cdot)$ are the posterior probabilities that the object in location 1, given that we have search location 1 (or 2) and not found it.

Now $F_0(x_1) = 0$, $F_1(x_1) = \min\{c_1, c_2\}$, $F_2(x)$ is the minimum of two linear functions of x_1. If F_{s-1} is the minimum of some collection of linear functions of x_1 it follows that the same can be said of F_s. Thus, by induction, F_s is a concave function of x_1.

By application of our theorem that $F_s \to F$ in the N and F case, we can deduce that the infinite horizon return function, F, is also a concave function. Notice that in the optimality equation for F (obtained by letting $s \to \infty$ in the equation above), the left hand term within the $\min\{\cdot,\cdot\}$ varies from $c_1 + F(p_{21})$ to $c_1 + (1-\alpha_1)F(p_{11})$ as x_1 goes from 0 to 1. The right hand term varies from $c_2 + (1-\alpha_2)F(p_{21})$ to $c_2 + F(p_{11})$ as x_1 goes from 0 to 1.

Consider the special case of $\alpha_1 = 1$ and $c_1 = c_2 = c$. Then the left hand term is the linear function $c + (1-x_1)F(p_{21})$. This means we have the picture below, where the blue and red curves corresponds to the left and right hand terms, and intersect exactly once since the red curve is concave.

Thus the optimal policy can be characterized as "*search location 1 iff the probability that the object is in location 1 exceeds a threshold x_1^**".



The value of x_1^* depends on the parameters, α_i and p_{ij}. It is believed that the answer is of this form for any parameters, but this is still an unproved conjecture.

5.2 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 $\pi = (f, f, \dots) = f^\infty$.

For such a policy we might write

$$F_t(\pi, x) = c(x, f(x)) + E[F_{t+1}(\pi, x_1) \mid x_0 = x, u_t = f(x)]$$

or $F_t = \mathcal{L}(f)F_{t+1}$, where $\mathcal{L}(f)$ is the operator having action

$$\mathcal{L}(f)\phi(x) = c(x, f(x)) + E[\phi(x_1) \mid x_0 = x, u_0 = f(x)].$$

5.3 Characterization of the optimal policy

Negative programming is about maximizing non-positive rewards, $r(x,u) \le 0$, or minimizing non-negative costs, $c(x,u) \ge 0$. 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 (as would be the case if we make the finite actions assumptions, F).

Theorem 5.2. *Suppose D or N holds. Suppose $\pi = f^\infty$ is the stationary Markov policy such that*

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

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. 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 5.2 or this corollary are true for positive programming (see Example 4.1).

5.4 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. It is entered upon choosing the stopping action, and once entered the system stays in that state and no further cost is incurred thereafter. Let $c(x, 0) = k(x)$ (stopping cost) and $c(x, 1) = c(x)$ (continuation cost). This defines a **stopping problem**.

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 optimality 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 5.3. *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).

5.5 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, $s = 0, 1, \dots$. If he passes the destination without having parked then 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

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

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

Now we solve the problem using the idea of a OSLA rule. It is better to stop now (at a distance s from the destination) than to go on and take the next available space if s is in the stopping set

$$S = \{s : s \leq k(s-1)\}$$

where $k(s-1)$ is the expected cost if we take the first available space that is $s-1$ or closer. Now

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

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

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

So

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

This set is closed (since s decreases) and so by Theorem 5.3 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 $2q^s \geq 1$.

6 Optimal Stopping Problems

More on stopping problems. Bruss's odds algorithm. Sequential probability ratio test. Prospecting.

6.1 Bruss's odds algorithm

A doctor, using a special treatment, codes 1 for a successful treatment, 0 otherwise. He treats a sequence of n patients and wants to minimize any suffering, while achieving a success with every patient for whom that is possible. Stopping on the last 1 would achieve this objective, so he wishes to maximize the probability of this.

Solution. Suppose X_k is the code of the kth patient. Assume $X_1, \dots, X_n$ are independent with $p_k = P(X_k = 1)$. Let $q_k = 1 - p_k$ and $r_k = p_k/q_k$. **Bruss's odds algorithm** sums the odds from the sth event to the last event (the nth)

$$R_s = r_s + \cdots + r_n$$

and finds the greatest integer s, say s^*, for which $R_s \geq 1$. We claim that by stopping the first time that code 1 occurs amongst patients $\{s^*, s^* + 1, \dots, n\}$, the doctor maximizes probability of stopping on the last patient who can be successfully treated.

To prove this claim we just check optimality of a OSLA-rule. The stopping set is

$$\begin{aligned} S &= \{i : q_{i+1} \cdots q_n > (p_{i+1} q_{i+2} q_{i+3} \cdots q_n) + (q_{i+1} p_{i+2} q_{i+3} \cdots q_n) \\ &\qquad\qquad + \cdots + (q_{i+1} q_{i+2} q_{i+3} \cdots p_n)\} \\ &= \{i : 1 > r_{i+1} + r_{i+2} + \cdots + r_n\} \\ &= \{s^*, s^* + 1, \dots, n\}. \end{aligned}$$

Clearly the stopping set is closed, so the OSLA-rule is optimal. The probability of stopping on the last 1 is $(q_{s^*} \cdots q_n)(r_{s^*} + \cdots + r_n)$ and, by solving a little optimization problem, we can see that this is always $\geq 1/e = 0.368$, provided $R_1 \geq 1$.

We can use the odds algorithm to re-solve the secretary problem. Code 1 when a candidate is better than all who have been seen previously. Our aim is to stop on the last candidate coded 1. We proved previously that $X_1, \dots, X_h$ are independent and $P(X_t = 1) = 1/t$. So $r_i = (1/t)/(1 - 1/t) = 1/(t-1)$. The algorithm tells us to ignore the first $s^* - 1$ candidates and the hire the first who is better than all we have seen previously, where s^* is the greatest integer s for which

$$\frac{1}{s-1} + \frac{1}{s} + \cdots + \frac{1}{h-1} \geq 1 \qquad \left(\equiv \text{ the least } s \text{ for which } \frac{1}{s} + \cdots + \frac{1}{h-1} \leq 1\right).$$

6.2 Example: stopping a random walk

Suppose the state space is $\{0, \dots, N\}$. In state x_t we may stop and take positive reward $r(x_t)$, or we may continue, in which case x_{t+1} is obtained by a step of a symmetric random walk. However, in states 0 and N we must stop. We wish to maximize $Er(x_T)$.

Solution. This is an example in which a OSLA rule is not optimal. The dynamic programming equation is

$$F(x) = \max\left\{r(x), \tfrac{1}{2}F(x-1) + \tfrac{1}{2}F(x+1)\right\}, \quad 0 < x < N,$$

with $F(0) = r(0)$, $F(N) = r(N)$. We see that

(i) $F(x) \geq \frac{1}{2}F(x-1) + \frac{1}{2}F(x+1)$, so $F(x)$ is concave.

(ii) Also $F(x) \geq r(x)$.

A function with properties (i) and (ii) is called a **concave majorant** of r. In fact, F can be characterized as the smallest concave majorant of r. For suppose that G is any other concave majorant of r. Starting with $F_0(x) = 0$, we have $G \geq F_0$. So we can prove by induction that

$$\begin{aligned}
F_s(x) &= \max\left\{r(x), \tfrac{1}{2}F_{s-1}(x-1) + \tfrac{1}{2}F_{s-1}(x-1)\right\} \\
&\leq \max\left\{r(x), \tfrac{1}{2}G(x-1) + \tfrac{1}{2}G(x+1)\right\} \\
&\leq \max\left\{r(x), G(x)\right\} \\
&= G(x).
\end{aligned}$$

Theorem 4.3 for case P tells us that $F_s(x) \to F(x)$ as $s \to \infty$. Hence $F \leq G$.

The optimal rule is to stop iff $F(x) = r(x)$.

6.3 Optimal stopping over the infinite horizon

Consider now a stopping problem over the infinite-horizon with $k(x)$ and $c(x)$ defined as previously. Let $F_s(x)$ be the infimal expected cost given that we are required to stop by the sth step. Let $F(x)$ be the infimal expected 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 6.1. Consider the problem of stopping a symmetric random walk on the integers, where $c(x) = 0$, $k(x) = \exp(-x)$. Inductively, we find that $F_s(x) = \exp(-x)$. This is because e^{-x} is a convex function. However, since the random walk is recurrent, we may wait until reaching as large an integer as we like before stopping; hence $F(x) = 0$. Thus $F_s(x) \not\to F(x)$. We see two things:

(i) It is possible that $F_\infty > F$.

(ii) Theorem 4.2 does not hold for negative programming. The policy of stopping immediately, say π, has $F(\pi, x) = e^{-x}$, and this satisfies the optimality equation

$$F(x) = \max\left\{e^{-x}, \tfrac{1}{2}F(x-1) + \tfrac{1}{2}F(x+1)\right\}.$$

But π is not optimal.

Remark. The above example does not contradict Theorem 4.3, which said $F_\infty = F$, because for that theorem we assumed $F_0(x) = k(x) = 0$ and $F_s(x)$ was the infimal cost possible over s steps, and thus $F_s \leq F_{s+1}$ (in the N case). Example 6.1 differs because $k(x) > 0$ and $F_s(x)$ is the infimal cost amongst the set of policies that are required to stop within s steps. Now $F_s(x) \geq F_{s+1}(x)$.

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

Lemma 6.2. *Suppose all costs are bounded as follows.*

$$(a)\ K = \sup_x k(x) < \infty \qquad (b)\ C = \inf_x c(x) > 0. \tag{6.1}$$

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. Clearly $(s+1)CP(\tau > s) < K$, otherwise it would be optimal to stop immediately. In the s-horizon problem we could follow π, but stop at time s if $\tau > s$. This implies

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

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

Theorem 6.3. *Suppose S is closed and* (6.1) *holds. Then an optimal policy for the infinite horizon is: stop if and only if $x \in S$.*

Proof. As usual, it is not optimal to stop if $x \notin S$. If $x \in S$, then by Theorem 5.3,

$$F_s(x) = k(x), \quad x \in S.$$

Lemma 6.2 gives $F(x) = \lim_{s\to\infty} F_s(x) = k(x)$, and so it is optimal to stop. □

6.4 Example: sequential probability ratio test

From i.i.d. observations drawn from a distribution with density f, a statistician wishes to decide between two hypotheses, $H_0 : f = f_0$ and $H_1 : f = f_1$ *Ex ante* he believes the probability that H_i is true is p_i, where $p_0 + p_1 = 1$. Suppose that he has the sample $x = (x_1, \ldots, x_n)$. The posterior probabilities are in the likelihood ratio

$$\ell_n = \frac{P(f = f_1 \mid x_1, \ldots, x_n)}{P(f = f_0 \mid x_1, \ldots, x_n)} = \frac{f_1(x_1)\cdots f_1(x_n)}{f_0(x_1)\cdots f_0(x_n)}\frac{p_1}{p_0} = \frac{f_1(x_n)}{f_0(x_n)}\ell_{n-1}.$$

Suppose it costs γ to make an observation. Stopping and declaring H_i true results in a cost c_i if wrong. This leads to the optimality equation for minimizing expected cost

$$F(\ell) = \min\Bigg\{ c_0\frac{\ell}{1+\ell}, c_1\frac{1}{1+\ell},$$
$$\gamma + \frac{\ell}{1+\ell}\int F(\ell f_1(y)/f_0(y))f_1(y)dy + \frac{1}{1+\ell}\int F(\ell f_1(y)/f_0(y))f_0(y)dy\Bigg\}$$

Taking $H(\ell) = (1+\ell)F(\ell)$, the optimality equation can be rewritten as

$$H(\ell) = \min\left\{ c_0\ell, c_1, (1+\ell)\gamma + \int H(\ell f_1(y)/f_0(y))f_0(y)dy \right\}.$$

We have a very similar problem to that of searching for a moving object. The state is ℓ_n. We can stop (in two ways) or continue by paying for another observation, in which case the state makes a random jump to $\ell_{n+1} = \ell_n f_1(x)/f_0(x)$, where x is a sample from f_0. We can show that $H(\cdot)$ is concave in ℓ, and that therefore the optimal policy can be described by two numbers, $a_0^* < a_1*$: *If* $\ell_n \leq a_0^*$*, stop and declare* H_0 *true; If* $\ell_n \geq a_1^*$*, stop and declare* H_1 *true; otherwise take another observation.*

6.5 Example: prospecting

We are considering mining in location i where the return will be R_i per day. We do not know R_i, but believe it is distributed $U[0,i]$. The first day of mining incurs a prospecting cost of c_i, after which we will know R_i. What is the greatest daily g that we would be prepared to pay to mine in location i? Call this G_i. Assume we may abandon mining whenever we like.

$$G_i = \sup\left[g : 0 \leq -c_i - g + E[R_i] + \frac{\beta}{1-\beta}E\max\{0, R_i - g\} \right]$$

For $\beta = 0.9$, $i = 1$, and $c_1 = 1$ this gives $G_1 = 0.5232$.

Now suppose that there is also a second location we might prospect, $i = 2$. We think its reward, R_2, is *ex ante* distributed $U[0,2]$. For $c_2 = 3$ this gives $G_2 = 0.8705$.

Suppose the true cost of mining in either location is $g = 0.5$ per day. Since $G_2 > G_1 > g$ we might conjecture the following is optimal.

- Prospect location 2 and learn R_2.

  If $R_2 > G_1 = 0.5232$ stop and mine there ever after.

- Otherwise

  - Prospect location 1. Now having learned both of R_1, R_2, we mine in the best location if $\max\{R_1, R_2\} > g = 0.5$.
  - Otherwise abandon mining.

This is a conjecture. That it is optimal follows from the Gittins index theorem.

Notice that also,

$$G_i = \sup\left[g : \frac{g}{1-\beta} \leq -c_i + E[R_i] + \frac{\beta}{1-\beta}E\max\{g, R_i\} \right].$$

This provides another interpretation of G_i, as the greatest daily return of an existing mine for which we would be willing to prospect in the new mine i, switching back to the old mine if R_i turns out to be less than G_i.

7 Bandit Processes and the Gittins Index

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

7.1 Index policies

Recall the **single machine scheduling** example in §3.2 in which n jobs are to be processed successively on one machine. Job i has a known processing time t_i, assumed to be a positive integer. On completion of job i a positive reward r_i is obtained. We used an interchange argument to show that total discounted reward obtained from the n jobs is maximized by the **index policy** of always processing the uncompleted job of greatest index, computed as $r_i\beta^{t_i}(1-\beta)/(1-\beta^{t_i})$.

Notice that if we were to interrupt the processing a job before it is finished, so as to carry out some processing on a different job, this would be against the advice of the index policy. For the index, $r_i\beta^{t_i}(1-\beta)/(1-\beta^{t_i})$, increases as t_i decreases.

7.2 Bandit processes and the multi-armed bandit problem

A **bandit process** is a special type of Markov decision process in which there are just two possible actions: $u = 0$ (freeze) or $u = 1$ (continue). The control $u = 0$ produces no reward and the state does not change (hence the term 'freeze'). Under $u = 1$ there is reward $r(x_t)$ and the state changes, to x_{t+1}, according to the Markov dynamics $P(x_{t+1} \mid x_t, u_t = 1)$.

A **simple family of alternative bandit processes** (SFABP) is a collection of n such bandit processes.

Given a SFABP, the **multi-armed bandit problem** (MABP) is to maximize the expected total discounted reward obtained over an infinite number of steps. At each step, $t = 0, 1, \dots$, exactly one of the bandit processes is to be continued. The others are frozen.

Let $x(t) = (x_1(t), \dots, x_n(t))$ 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.

This provide a very rich modelling framework. With it we can model questions like:

- Which of n drugs should we give to the next patient?
- Which of n jobs should we work on next?
- When of n oil fields should we explore next?

We have an infinite-horizon discounted-reward Markov decision problem. It has a deterministic stationary Markov optimal policy. The optimality equation is

$$F(x) = \max_{i:i\in\{1,\dots,n\}} \left\{ r_i(x) + \beta \sum_{y\in E_i} P_i(x_i, y) F(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n) \right\}. \qquad (7.1)$$

7.3 The two-armed bandit

Consider a MABP with just two bandits. Bandit B_1 always pays λ, and bandit B_2 is of general type. The optimality equation, when B_2 is in its state x, is

$$\begin{aligned} F(x) &= \max\left\{ \frac{\lambda}{1-\beta}, r(x) + \beta \sum_y P(x,y)F(y) \right\} \\ &= \max\left\{ \frac{\lambda}{1-\beta}, \sup_{\tau>0} E\left[\sum_{t=0}^{\tau-1} \beta^t r(x(t)) + \beta^\tau \frac{\lambda}{1-\beta} \,\middle|\, x(0) = x \right] \right\}. \end{aligned}$$

The left hand choice within $\max\{\cdot,\cdot\}$ corresponds to continuing B_1. The right hand choice corresponds to continuing B_2 for at least one step and then switching to B_1 a some later step, τ. Notice that once we switch to B_1 we will never wish switch back to B_2 because information remains the same as when we first switched from B_2 to B_1.

We are to choose the **stopping time** τ optimally. Because the two terms within the $\max\{\cdot,\cdot\}$ are both increasing in λ, and are linear and convex, respectively, there is a unique λ, say λ^*, for which they are equal.

$$\lambda^* = \sup\left\{ \lambda : \frac{\lambda}{1-\beta} \le \sup_{\tau>0} E\left[\sum_{t=0}^{\tau-1} \beta^t r(x(t)) + \beta^\tau \frac{\lambda}{1-\beta} \,\middle|\, x(0) = x \right] \right\}. \qquad (7.2)$$

Of course this λ depends on $x(0)$. We denote its value as $G(x)$. After a little algebra we have the definition

$$G(x) = \sup_{\tau>0} \frac{E\left[\sum_{t=0}^{\tau-1} \beta^t r(x(t) \,\middle|\, x(0) = x\right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \,\middle|\, x(0) = x\right]}. \qquad (7.3)$$

$G(x)$ is called the **Gittins index** (of state x), named after its originator, John Gittins. The definition above is by a **calibration**, the idea being that we find a B_1 paying a constant reward λ, such that we are indifferent as to which bandit to continue next.

In can be easily shown that $\tau = \min\{t : G_i(x_i(t)) \le G_i(x_i(0)), \tau > 0\}$, that is, τ is the first time B_2 is in a state where its Gittins index is no greater than it was initially.

In (7.3) we see that the Gittins index is the maximal possible quotient of 'expected total discounted *reward* over τ steps', divided by 'expected total discounted *time* over τ steps', where τ is at least 1 step. 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.

7.4 Gittins index theorem

Remarkably, the problem posed by a SFABP (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.

Theorem 7.1 (Gittins Index Theorem). *The problem posed by a SFABP, as setup above, is solved by always continuing the process having the greatest* **Gittins index***.*

The Index Theorem is due to Gittins and Jones, who obtained it in 1970, and presented it in 1972. The solution is 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*'.

The optimal **stopping time** τ in (7.3) is $\tau = \min\{t : G_i(x_i(t)) \leq G_i(x_i(0)), \tau > 0\}$, that is, τ is the first time at which the process reaches a state whose Gittins index is no greater than Gittins index at $x_i(0)$.

In the single machine scheduling example of §7.1, the optimal stopping time on the right hand side of (7.3) 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, $G_i = r_i\beta^{t_i}(1-\beta)/(1-\beta^{t_i})$. Note that $G_i \to r_i/t_i$ as $\beta \to 1$.

7.5 *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 7.1. 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 an agent would be willing to pay per step if he must continue B_i for one more step, and then stop whenever he likes thereafter. This is

$$\gamma_i(x_i) = \sup\left\{\lambda : 0 \leq \sup_{\tau>0} E\left[\sum_{t=0}^{\tau-1}\beta^t\Big(r_i(x_i(t)) - \lambda\Big)\,\Big|\, x_i(0) = x_i\right]\right\}. \tag{7.4}$$

Notice that (7.2) and (7.4) 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 $g_i(t) = \min_{s\leq t}\gamma_i(x_i(s))$. 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, $B_1, \dots, B_n$, the agent not only collects rewards, but also pays the prevailing charge of whatever bandit that he chooses to continue at each step. Then he cannot do better than just break even (i.e. expected value of rewards minus prevailing charges is 0).

This is because he could only make a strictly positive profit (in expected value) if this were to 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) If he always continues the bandit of greatest prevailing charge then he will interleave the n nonincreasing sequences of prevailing charges into a single nonincreasing sequence of prevailing charges and so maximize 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 also maximize the expected discounted sum of the rewards can be obtained from this SFABP. □

7.6 Example: Weitzman's problem

'Pandora' has n boxes, each of which contains an unknown prize. *Ex ante* the prize in box i has a value with probability distribution function F_i. She can learn the value of the prize by opening box i, which costs her c_i to do. At any stage she may stop and take as her reward the maximum of the prizes she has found. She wishes to maximize the expected value of the prize she takes, minus the costs of opening boxes.

Solution. This problem is similar to 'prospecting' problem in §6.5. It can be modelled in terms of a SFABP. Box i is associated with a bandit process B_i, which starts in state 0. The first time it is continued there is a cost c_i, and the state becomes x_i, chosen by the distribution F_i. At all subsequent times that it is continued the reward is $r(x_i) = (1-\beta)x_i$, and the state remains x_i. Suppose we wish to maximize the expected value of

$$-\sum_{t=1}^{\tau} \beta^{t-1} c_{i_t} + \max\{r(x_{i_1}), \dots, r(x_{i_\tau})\} \sum_{t=\tau}^{\infty} \beta^t$$
$$= -\sum_{t=1}^{\tau} \beta^{t-1} c_{i_t} + \beta^\tau \max\{x_{i_1}, \dots, x_{i_\tau}\}.$$

The Gittins index of an opened box is $r(x_i)/(1-\beta) = x_i$. The index of an unopened box i is the solution to

$$\frac{G_i}{1-\beta} = -c_i + \frac{\beta}{1-\beta} E \max\{r(x_i), G_i\}.$$

Pandora's optimal strategy is thus: *Open boxes in decreasing order of G_i until first reaching a point that a revealed prize is greater than all G_i of unopened boxes.*

The undiscounted case In the limit as $\beta \to 1$ this objective corresponds to that of Weitzman's problem, namely,

$$-\sum_{t=1}^{\tau} c_{i_t} + \max\{x_{i_1}, \dots, x_{i_\tau}\}.$$

By setting $g_i = G/(1-\beta)$, and letting $\beta \to 1$, we get an index that is the solution of $g_i = -c_i + E \max\{x_i, g_i\}$.

For example, if F_i is a two point distribution with $x_i = 0$ or $x_i = r_i$, with probabilities $1 - p_i$ and p_i, then $g_i = -c_i + (1 - p_i)g_i + p_i r_i \implies g_i = r_i - c_i/p_i$.

8 Average-cost Programming

The average-cost optimality equation. Policy improvement algorithm.

8.1 Average-cost optimality equation

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

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

Plausibly, there is a well-defined optimal **average-cost**, $\lambda(x) = \inf_\pi \lambda(\pi, x)$, and we expect $\lambda(x) = \lambda$ should not depend on x. 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 that is due to the initial state. Suppose that in each state the action space is finite. From the optimality equation for the finite horizon problem we have

$$F_s(x) = \min_u \{c(x, u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\}. \tag{8.1}$$

So by substituting $F_s(x) \sim s\lambda + \phi(x)$ into (8.1), 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:

$$\lambda + \phi(x) = \min_u \{c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u]\}. \tag{8.2}$$

Theorem 8.1. *Suppose there exists a constant λ and bounded function ϕ satisfying (8.2). 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.*

The proof follows by application of the following two lemmas.

Lemma 8.2. *Suppose the exists a constant λ and bounded function ϕ such that*

$$\lambda + \phi(x) \le c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u] \quad \textit{for all } x, u. \tag{8.3}$$

Then $\lambda \le \inf_\pi \lambda(\pi, x)$.

Proof. Let π be any policy. By repeated substitution of (8.3) into itself,

$$\phi(x) \le -t\lambda + E_\pi \left[\sum_{\tau=0}^{t-1} c(x_\tau, u_\tau) \,\middle|\, x_0 = x \right] + E_\pi[\phi(x_t) \mid x_0 = x]. \tag{8.4}$$

Divide by t, let $t \to \infty$, and then take the infimum over π. □

Lemma 8.3. *Suppose the exists a constant λ and bounded function ϕ such that for each x there exists some $u = f(x)$ such that*

$$\lambda + \phi(x) \geq c(x,u) + E[\phi(x_1) \mid x_0 = x, u_0 = f(x)]. \tag{8.5}$$

Let $\pi = f^\infty$. Then $\lambda \geq \lambda(\pi, x) \geq \inf_\pi \lambda(\pi, x)$.

Proof. Repeated substitution of (8.5) into itself gives (8.4) but with the inequality reversed. Divide by t and let $t \to \infty$. This gives $\lambda \geq \lambda(\pi, x) \geq \inf_\pi \lambda(\pi, x)$. □

So an optimal average-cost policy can be found by looking for a bounded solution to (8.2). Notice that if ϕ is a solution of (8.2) then so is ϕ+(a constant), because the (a constant) will cancel from both sides of (8.2). Thus ϕ is undetermined up to an additive constant. In searching for a solution to (8.2) we can therefore pick any state, say $\bar{x}$, and arbitrarily take $\phi(\bar{x}) = 0$. We can do this in whatever way is most convenient. The function ϕ is called the **relative value function**.

8.2 Example: admission control at a queue

Each day a consultant is has 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, $i = 1, \ldots, n$. A job of type i pays R_i upon completion. Once he has accepted a job he may accept no other job until the job is complete. The probability a job of type i takes k days is $(1 - p_i)^{k-1} p_i$, $k = 1, 2, \ldots$. Which jobs should he 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 (8.2) 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 increases in λ and the right hand side decreases in λ. Equality holds for some λ^*, which is the maximal average-reward. The optimal policy is: *accept only jobs for which $p_i R_i \geq \lambda^*$.*

8.3 Value iteration bounds

For the rest of this lecture we suppose the state space is finite and there are only finitely many actions in each state.

Theorem 8.4. *Define*

$$m_s = \min_x \{F_s(x) - F_{s-1}(x)\}, \qquad M_s = \max_x \{F_s(x) - F_{s-1}(x)\}. \tag{8.6}$$

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

Proof. For any x, u,

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

Now apply Lemma 8.2 with $\phi = F_{s-1}$, $\lambda = m_s$.

Similarly, for each x there is a $u = f_s(x)$, such that

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

Now apply Lemma 8.3 with $\phi = F_{s-1}$, $\lambda = M_s$. □

This justifies a **value iteration algorithm**: Calculate F_s until $M_s - m_s \le \epsilon m_s$. At this point the stationary policy f_s^∞ has average-cost that is within $\epsilon \times 100\%$ of optimal.

8.4 Policy improvement algorithm

In the average-cost case a **policy improvement algorithm** is be based on the following observations. Suppose that for a policy $\pi_0 = f^\infty$, we have that λ, ϕ solve

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

Then λ is the average-cost of policy π.

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

$$\lambda + \phi(x) \ge c(x, f_1(x_0)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_1(x_0)], \tag{8.7}$$

for all x, and with strict inequality for some x (and thus $f_1 \ne f$). Then just as in the proof of Lemma 8.3, we find

$$\lambda \ge \lim_{t\to\infty} \frac{1}{t} E_{\pi_1}\left[\sum_{\tau=0}^{t-1} c(x_\tau, u_\tau) \,\middle|\, x_0 = x\right]. \tag{8.8}$$

In fact, if the Markov chain induced by π_1 is irreducible the inequality above will be strict and π_1 will be strictly better than π. If there is no such π_1 then π satisfies (8.2) 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 (8.2). 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 (8.2) and so π_{s-1} is optimal. Otherwise π_s is a new policy. By the calculation in (8.8) 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 state and action 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 the minimal average cost.

Example 8.5. Consider again the example of §8.2. 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

$$\lambda + \phi(0) = a_1\phi(1) + \sum_{i=2}^n a_i\phi(0),$$
$$\lambda + \phi(i) = (1 - p_i)\phi(i) + p_i[R_i + \phi(0)], \quad i = 1, \dots, n.$$

The solution is $\lambda = a_1p_1R_1/(a_1 + p_1)$, $\phi(0) = 0$, $\phi(1) = p_1R_1/(a_1 + p_1)$, 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 $p_iR_i \leq a_1p_1R_1/(a_1 + p_1)$ 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 $s := s + 1$.

9 Continuous-time Markov Decision Processes

Control problems in a continuous-time stochastic setting. Markov jump processes when the state space is discrete. Uniformization.

9.1 *Stochastic scheduling on parallel machines*

A collection of n jobs is to be processed on a single machine. They have processing times $X_1, \dots, X_n$, which are *ex ante* distributed as independent exponential random variables, $X_i \sim \mathcal{E}(\lambda_i)$ and $EX_i = 1/\lambda_i$, where $\lambda_1, \dots, \lambda_n$ are known.

If jobs are processed in order $1, 2, \dots, n$, they finish in expected time $1/\lambda_1 + \cdots + 1/\lambda_n$. So the order of processing does not matter.

But now suppose there are m $(2 \leq m < n)$ identical machines working in parallel. Let C_i be the **completion time** of job i.

- $\max_i C_i$ is called the **makespan** (the time when all jobs are complete).
- $\sum_i C_i$ is called the **flow time** (sum of completion times).

Suppose we wish to minimize the expected makespan. We can find the optimal order of processing by stochastic dynamic programming. But now we are in continuous time, $t \geq 0$. So we need the important facts:

(i) $\min(X_i, X_j) \sim \mathcal{E}(\lambda_i + \lambda_j)$; (ii) $P(X_i < X_j \mid \min(X_i, X_j) = t) = \lambda_i/(\lambda_i + \lambda_j)$.

Suppose $m = 2$. The optimality equations are

$$\begin{aligned} F(\{i\}) &= \frac{1}{\lambda_i} \\ F(\{i,j\}) &= \frac{1}{\lambda_i + \lambda_j}[1 + \lambda_i F(\{j\}) + \lambda_j F(\{i\})] \\ F(S) &= \min_{i,j \in S} \frac{1}{\lambda_i + \lambda_j}[1 + \lambda_i F(S^i) + \lambda_j F(S^j)], \end{aligned}$$

where S is a set of uncompleted jobs, and we use the abbreviated notation $S^i = S \setminus \{i\}$.

It is helpful to rewrite the optimality equation. Let $\Lambda = \sum_i \lambda_i$. Then

$$\begin{aligned} F(S) &= \min_{i,j \in S} \frac{1}{\Lambda}\left[1 + \lambda_i F(S^i) + \lambda_j F(S^j) + \sum_{k \neq i,j} \lambda_k F(S)\right] \\ &= \min_{\substack{u_i \in [0,1], i \in S, \\ \sum_i u_i \leq 2}} \frac{1}{\Lambda}\left[1 + \Lambda F(S) + \sum_i u_i \lambda_i (F(S^i) - F(S))\right] \end{aligned}$$

This is helpful, because in all equations there is now the same divisor, Λ. An event occurs after a time that is exponentially distributed with parameter Λ, but with probability λ_k/Λ this is a 'dummy event' if $k \neq i, j$. This trick is known as **uniformization**. Having set this up we might also then say let $\Lambda = 1$.

We see that it is optimal to start by processing the two jobs in S for which $\delta_i(S) := \lambda_i(F(S^i) - F(S))$ is least.

The policy of always processing the m jobs of smallest [largest] λ_i is called the Lowest [Highest] Hazard Rate first policy, and denoted LHR [HHR].

Theorem 9.1.
(a) Expected makespan is minimized by LHR.

(b) Expected flow time is minimized by HHR.

(c) $E[C_{(n-m+1)}]$ (expected time there is first an idle machine) is minimized by LHR.

Proof. (*starred*) We prove only (a), and for ease assume $m = 2$ and $\lambda_1 < \cdots < \lambda_n$. We would like to prove that for all $i, j \in S \subseteq \{1, \dots, n\}$,

$$i < j \iff \delta_i(S) < \delta_j(S) \quad \text{(except possibly if both } i \text{ and } j \text{ are the jobs that would be processed by the optimal policy).} \tag{9.1}$$

Truth of (9.1) would imply that jobs should be started in the order $1, 2, \dots, n$.

Let π be LHR. Take an induction hypothesis that (9.1) is true and that $F(S) = F(\pi, S)$ when S is a strict subset of $\{1, \dots, n\}$. Now consider $S = \{1, \dots, n\}$. We examine $F(\pi, S)$, and $\delta_i(\pi, S)$, under π. Let S^k denote $S \setminus \{k\}$. For $i \geq 3$,

$$\begin{aligned} F(\pi, S) &= \frac{1}{\lambda_1 + \lambda_2}[1 + \lambda_1 F(S^1) + \lambda_2 F(S^2)] \\ F(\pi, S^i) &= \frac{1}{\lambda_1 + \lambda_2}[1 + \lambda_1 F(S^{1i}) + \lambda_2 F(S^{2i})] \\ \implies \delta_i(\pi, S) &= \frac{1}{\lambda_1 + \lambda_2}[\lambda_1 \delta_i(S^1) + \lambda_2 \delta_i(S^2)], \quad i \geq 3. \end{aligned} \tag{9.2}$$

Suppose $3 \leq i < j$. The inductive hypotheses that $\delta_i(S^1) \leq \delta_j(S^1)$ and $\delta_i(S^2) \leq \delta_j(S^2)$ imply $\delta_i(\pi, S) \leq \delta_j(\pi, S)$.

Similarly, we can compute $\delta_1(\pi, S)$.

$$\begin{aligned} F(\pi, S) &= \frac{1}{\lambda_1 + \lambda_2 + \lambda_3}[1 + \lambda_1 F(S^1) + \lambda_2 F(S^2) + \lambda_3 F(\pi, S)] \\ F(\pi, S^1) &= \frac{1}{\lambda_1 + \lambda_2 + \lambda_3}[1 + \lambda_1 F(S^1) + \lambda_2 F(S^{12}) + \lambda_3 F(S^{13})] \\ \implies \delta_1(\pi, S) &= \frac{1}{\lambda_1 + \lambda_2 + \lambda_3}[\lambda_2 \delta_1(S^2) + \lambda_3 \delta_1(\pi, S) + \lambda_1 \delta_3(S^1)] \\ &= \frac{1}{\lambda_1 + \lambda_2}[\lambda_1 \delta_3(S^1) + \lambda_2 \delta_1(S^2)]. \end{aligned} \tag{9.3}$$

Using (9.2), (9.3) and using our inductive hypothesis, we deduce $\delta_1(\pi, S) \leq \delta_i(\pi, S)$. A similar calculation may be done for $\delta_2(\pi, S)$.

This completes a step of an inductive proof by showing that (9.1) is true for S, and that $F(S) = F(\pi, S)$. We only need to check the base of the induction. This is provided by the simple calculation

$$\begin{aligned}\delta_1(\{1,2\}) = \lambda_1(F(\{2\}) - F(\{1,2\})) &= \lambda_1\left[\frac{1}{\lambda_2} - \frac{1}{\lambda_1+\lambda_2}\left(1 + \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1}\right)\right]\\ &= -\frac{\lambda_2}{\lambda_1+\lambda_2} \leq \delta_2(\{1,2\}). \qquad \square\end{aligned}$$

The proof of (b) is very similar, except that the inequality in (9.1) should be reversed. The base of the induction comes from $\delta_1(\{1,2\}) = -1$.

The proof of (c) is also similar. The base of the induction is provided by $\delta_1(\{1,2\}) = \lambda_1(0 - 1/(\lambda_1 + \lambda_2))$. Since we are seeking to maximize $EC_{(n-m+1)}$ we should process jobs for which δ_i is greatest, i.e., least λ_i. The problem in (c) is known as the **Lady's nylon stocking problem**. We think of a lady (having $m = 2$ legs) who starts with n stockings, wears two at a time, each of which may fail, and she wishes to maximize the expected time until she has only one good stocking left to wear.

9.2 Controlled Markov jump processes

The above example illustrates the idea of a controlled **Markov jump process**. It evolves in continuous time, and in a discrete state space. In general:

- The state is i. We choose some control, say u ($u \in A(i)$, a set of available controls).
- After a time that is exponentially distributed with parameter $q_i(u) = \sum_{j\neq i} q_{ij}(u)$, (i.e. having mean $1/q_i(u)$), the state jumps.
- Until the jump occurs cost accrues at rate $c(i,u)$.
- The jump is to state j ($\neq i$) with probability $q_{ij}(u)/q_i(u)$.

The infinite-horizon optimality equation is

$$F(i) = \min_{u\in A(i)}\left\{\frac{1}{q_i(u)}\left[c(i,u) + \sum_j q_{ij}(u)F(j)\right]\right\}.$$

Suppose $q_i(u) \leq B$ for all i, u and use the **uniformization** trick,

$$F(i) = \min_{u\in A(i)}\left\{\frac{1}{B}\left[c(i,u) + (B - q_i(u))F(i) + \sum_j q_{ij}(u)F(j)\right]\right\}.$$

We now have something that looks exactly like a discrete-time optimality equation

$$F(i) = \min_{u\in A(i)}\left\{\bar{c}(i,u) + \sum_j p_{ij}(u)F(j)\right\}$$

where $\bar{c}(i,u) = c(i,u)/B$, $p_{ij}(u) = q_{ij}(u)/B$, $j \neq i$, and $p_{ii}(u) = 1 - q_i(u)/B$.

This is great! It means we can use all the methods and theorems that we have developed previously for solving discrete-time dynamic programming problems.

We can also introduce discounting by imagining that there is an 'exponential clock' of rate α which takes the state to a place where no further cost or reward is obtained. This leads to an optimality equation of the form

$$F(i) = \min_u \left\{ \bar{c}(i,u) + \beta \sum_j p_{ij}(u) F(j) \right\},$$

where $\beta = B/(B+\alpha)$, $\bar{c}(i,u) = c(i,u)/(B+\alpha)$, and $p_{ij}(u)$ is as above.

9.3 Example: admission control at a queue

The number of customers waiting in a queue is $0, 1, \ldots, N$. There is a constant service rate μ (meaning that the service times of customers are distributed as i.i.d. exponential random variables with mean $1/\mu$, and we may control the arrival rate u to any value in $[m, M]$. Let $c(x,u) = ax - Ru$. This comes from a **holding cost** a per unit time for each customer in the system (queueing or being served) and reward R is obtained as each new customer is admitted (and therefore incurring reward at rate Ru when the arrival rate is u). No customers are admitted if the queue size is N.

Time-average cost optimality. We use the uniformization trick. Arrivals are at rate M, but this is sum of actual arrivals at rate u, and dummy (or ficticious) arrivals at rate $M - u$. Service completions are happening at rate μ, but these are dummy service completions if $x = 0$. Assume $M + \mu = 1$ so that some event takes place after a time that is distributed $\mathcal{E}(1)$.

Let γ denote the minimal average-cost. The optimality equation is

$$\begin{aligned}
\phi(x) + \gamma &= \inf_{u \in [m,M]} \Big\{ ax - Ru + u\phi(x+1) + \mu\phi(x-1) + (M-u)\phi(x) \Big\}, \\
&= \inf_{u \in [m,M]} \Big\{ ax + u[-R + \phi(x+1) - \phi(x)] + \mu\phi(x-1) + M\phi(x) \Big\}, \quad 1 \leq x < N,
\end{aligned}$$

$$\begin{aligned}
\phi(0) + \gamma &= \inf_{u \in [m,M]} \Big\{ -Ru + u\phi(1) + (\mu + M - u)\phi(0) \Big\}, \\
&= \inf_{u \in [m,M]} \Big\{ u[-R + \phi(1) - \phi(0)] + (\mu + M)\phi(0) \Big\},
\end{aligned}$$

$$\phi(N) + \gamma = aN + M\phi(N) + \mu\phi(N-1).$$

Thus u should be chosen to be m or M as $-R + \phi(x+1) - \phi(x)$ is positive or negative.

Let us consider what happens under the policy that takes $u = M$ for all x. The relative costs for this policy, say $\phi = f$, and average cost γ' are given by

$$f(0) + \gamma' = -RM + Mf(1) + \mu f(0), \tag{9.4}$$

$$f(x) + \gamma' = ax - RM + Mf(x+1) + \mu f(x-1), \quad 1 \leq x < N \tag{9.5}$$

$$f(N) + \gamma' = aN + Mf(N) + \mu f(N-1). \tag{9.6}$$

The general solution to the homogeneous part of the recursion in (9.5) is

$$f(x) = d_1 1^x + d_2(\mu/M)^x$$

and a particular solution is $f(x) = Ax^2 + Bx$, where

$$A = \frac{1}{2(\mu - M)}, \quad B = \frac{a}{2(\mu - M)^2} + \frac{\gamma' + RM}{M - \mu}.$$

We can now solve for γ' and d_2 so that (9.4) and (9.6) are also satisfied. The solution is not pretty, but if we assume $\mu > M$ and take the limit $N \to \infty$ the solution becomes

$$f(x) = \frac{ax(x+1)}{2(\mu - M)}, \qquad \gamma' = \frac{aM}{\mu - M} - MR.$$

Applying the idea of policy improvement, we conclude that a better policy is to take $u = m$ (i.e. slow arrivals) if $-R + f(x+1) - f(x) > 0$, i.e. if

$$R < \frac{(x+1)a}{\mu - M}.$$

Further iterations of policy improvement would be needed to reach the optimal policy. At this point the problem becomes one to be solved numerically, not in algebra! However, this first step of policy improvement already exhibits an interesting property: it uses $u = m$ at a smaller queue size than would a myopic policy, which might choose to use $u = m$ when the net benefit obtained from the next customer is negative, i.e.

$$R < \frac{(x+1)a}{\mu}.$$

The right hand side is the expected cost this customer will incur while waiting. This example exhibits the difference between **individual optimality** (which is myopic) and **social optimality**. The socially optimal policy is more reluctant to admit a customer because, it anticipates further customers are on the way; it takes account of the fact that if it admits a customer then the customers who are admitted after him will suffer delay. As expected, the policies are nearly the same if the arrival rate M is small.

Of course we might expect that policy improvement will eventually terminate with a policy of the form: *use $u = m$ iff $x \geq x^*$.*

10 Restless Bandits

10.1 Examples

Again, we start with a family of n alternative Markov decision processes. Given their states at time t, say $x_1(t), \dots, x_n(t)$, we are to choose actions $u_1(t), \dots, u_n(t)$ to apply at time t. As in the multi-armed bandit problem, we suppose that there are just two available actions, so $u_i(t) \in \{0, 1\}$. We generalize the SFABP set up in two ways.

Our first generalization is to require that at each time t exactly m ($< n$) of the bandits be given the 'active' action $u_i = 1$.

Our second generalization is that the 'passive' action $u = 0$ no longer freezes a bandit; instead, the state evolves, but differently from its continuation under $u = 1$.

Example 10.1. Suppose the state of a bandit is a measure of its vigour. The active and passive actions correspond to notions of work and rest. Performance is positively related to vigour (or lack of fatigue), which increases with rest and decreases with work. For example, suppose that x takes values in $\{1, \dots, k\}$. If the active action $u = 1$ is applied to a bandit in state x, then there accrues an immediate reward of $r(x)$, increasing in x, but vigour decreases to $\max\{x - 1, 1\}$. The passive action $u = 0$ produces no reward, but vigour increases to $\min\{x + 1, k\}$.

Example 10.2. The active and passive actions might correspond to notions of 'observation' and 'no observation'. Suppose that each bandit is in one of two conditions: 0 and 1, associated with being 'bad' or 'good', respectively. It moves back and forth between these two conditions independently of any actions applied by the decision-maker, according to a 2-state Markov chain. Each bandit is now a POMDP. So far as the decision-maker is concerned the state of the bandit is the probability that it is in good condition. Under the action $u = 1$ the condition is observed, and if this is found to be i then $x(t + 1) = p_{i1}$. Moreover, if the condition is good then a reward is obtained. Under the action $u = 0$ the underlying condition of the process is not observed, and so, in a Bayesian manner, $x(t + 1) = x(t)p_{11} + (1 - x(t))p_{01}$. No reward is obtained.

Example 10.3. The active and passive actions correspond to running the process at different speeds. For example, suppose for $0 < \epsilon < 1$,

$$P(j|i, 0) = \begin{cases} \epsilon P(j|i, 1), & i \neq j \\ (1 - \epsilon) + \epsilon P(i|i, 1), & i = j \end{cases}$$

Thus a bandit which is operated continuously with $u = 1$ has the same stationary distribution as one that is operated continuously with $u = 0$. But the process moves faster when $u = 1$.

10.2 *Whittle index policy*

Let $\Omega = \{(u_1,\dots,u_n) : u_i \in \{0,1\}$ for all i, and $\sum_i u_i(t) = m\}$. The optimality equation is

$$F(x) = \max_{u\in\Omega}\left\{\sum_i r(x_i,u_i) + \beta \sum_{y_1,\dots,y_n} F(y_1,\dots,y_n)\prod_i P(y_i \,|\, x_i,u_i)\right\}.$$

Let us focus upon average reward (i.e. the limit as $\beta \to 1$). This is attractive because performance does not depend on the initial state. Assuming that the n bandits are statistically equivalent it is plausible that, under an optimal policy, bandit i will be given the action $u_i = 1$ for precisely a fraction m/n of the time. This motivates interest in an upper bound on the maximal average reward that can be obtained by considering a single bandit and asking how it should be controlled if we wish to maximize the average reward obtained from that bandit, subject to a relaxed constraint that $u_i = 1$ is employed for a fraction of exactly m/n of the time.

So consider a stationary Markov policy for operating a single restless bandit. Let z_x^u be the proportion of time that the bandit is in state x and that under this policy the action u is taken. An upper bound for our problem can be found from a linear program in variables $\{z_x^u : x \in E,\ u \in \{0,1\}\}$:

$$\text{maximize} \sum_{x,u} r(x,u) z_x^u \tag{10.1}$$

subject to

$$\sum_{x,u} z_x^u = 1 \tag{10.2}$$

$$\sum_x z_x^0 \geq 1 - m/n \tag{10.3}$$

$$\sum_u z_x^u = \sum_{y,u} z_y^u P(x \,|\, y,u), \text{ for all } x \tag{10.4}$$

$$z_x^u \geq 0, \text{ for all } x,u. \tag{10.5}$$

Here (10.4) are equations that determine the stationary probabilities. Notice that we have put an inequality in (10.3). Let us justify this by making the assumption that action $u = 1$ (which we call the **active action** is in some sense better than $u = 0$ (which we call the **passive action**. So if constraint (10.3) did not exist then we would wish to take $u = 1$ in all states. At optimality (10.3) will hold with equality.

The optimal value of the **dual LP** problem is equal to g, where this can be found from the average reward dynamic programming equation

$$\phi(x) + g = \max_{u\in\{0,1\}}\left\{r(x,u) + \lambda(1-u) + \sum_y \phi(y)P(y \,|\, x,u(x))\right\}. \tag{10.6}$$

Here λ and $\phi(x)$ are the Lagrange multipliers for constraints (10.3) and (10.4), respectively. The multiplier λ is positive and may be interpreted as a *subsidy* for taking the passive action. It is interesting to see how (10.6) can be obtained from (10.1)–(10.4). However, we might have simply taken as our starting point a problem of maximizing average reward when there is a subsidy for taking the passive action.

In general, the solution of (10.6) partitions the state space E into three sets, E_0, E_1 and E_{01}, where, respectively the optimal action is $u = 0$, $u = 1$, or some randomization between both $u = 0$ and $u = 1$. Let us avoid uninteresting pathologies by supposing that the state space is finite, and that every pure policy gives rise to a Markov chain with one recurrent class. Then the set E_{01}, where there is randomization, need never contain more than 1 state, a fact that is known for general Markov decision processes with constraints.

It is reasonable to expect that as the subsidy λ increases in (10.6) the set of states E_0 (in which $u = 0$ is optimal) should increase monotonically. This need not happen in general. However, if it does then we say the bandit is **indexable**. Whittle defines as an index the least value of the subsidy λ such that $u = 0$ is optimal. We call this the **Whittle index**, denoting it $W(\cdot)$, where $W(x) = \inf\{\lambda : x \in E_0(\lambda)\}$. It can be used to define a heuristic policy (the **Whittle index policy**) in which, at each instant, one engages m bandits with the greatest indices, i.e. those that are the last to leave the set E_1 as the subsidy for the passive action increases. The Whittle index extends the Gittins optimal policy for classical bandits; Whittle indices can be computed separately for each bandit; they are the same as the Gittins index in the case that $u = 0$ is a freezing action, so that $P(j|i, 0) = \delta_{ij}$.

10.3 *Whittle indexability*

The discussion so far begs two questions: (i) under what assumptions is a restless bandit indexable, and (ii) how good is the Whittle index policy? Might it be optimal, or very nearly optimal as n becomes large?

Interestingly, there are special classes of restless bandit for which one can prove indexability. Bandits of the type in Example 10.2 are indexable. The dual-speed bandits in Example 10.3 are indexable. A restless bandit is also indexable if the passive action transition probabilities, $P(j \mid i, 0)$, depend only on j (the destination state).

10.4 *Fluid models of large stochastic systems*

It is often interesting to think about problems in some 'large N' limit. Consider, for example, N identical independently running single server queues, of type $M/M/1$, each with its own Poisson arrival stream of rate λ and server of rate μ. The probability that a given queue has i customers is $\pi_i = \rho^i(1-\rho)$ where $\rho = \lambda/\mu$. The queues are running independently and so we would expect the number of them that have i customers to be $N\pi_i$. Suppose we start off with $Nx_i(0)$ of the queues having i customers, where $\sum_i x_i(0) = 1$. Since N is large, transitions will be happening very fast, and so using the law of large numbers we expect to see

$$\begin{aligned}
\frac{d}{dt}x_0(t) &= \mu x_1(t) - \lambda x_0(t) \\
\frac{d}{dt}x_i(t) &= \lambda x_{i-1}(t) + \mu x_{i+1}(t) - (\lambda+\mu)x_i(t).
\end{aligned}$$

We have replaced our stochastic system by a deterministic fluid approximation. (There are theorems which talk about the convergence when $N \to \infty$.) These differential equation will produce a trajectory $x_i(t) \to \pi_i$ as $t \to \infty$.

The same things happens even if we link the behaviour of the queues. Suppose we have total processing effort $N\mu$. Rather than place μ per queue, as above, we now decide to allocate 2μ to all queues having more than the median number of customers. Suppose $\sum_{i=1}^{j-1} x_i(t) + \alpha x_j(t) = 0.5$, so queues with $\leq j$ customers are not served, those with $\geq j+1$ are served, and some service $(1-\alpha)x_j(t)\mu$ effort is wasted. Now the fluid approximation is

$$\begin{aligned}
\frac{d}{dt}x_0(t) &= \mu x_1(t) - \lambda x_0(t) \\
\frac{d}{dt}x_i(t) &= \lambda x_{i-1}(t) - \lambda x_i(t), \quad 0 < i \leq j-1 \\
\frac{d}{dt}x_j(t) &= \lambda x_{j-1}(t) + 2\mu x_{j+1}(t) - \lambda x_j(t) \\
\frac{d}{dt}x_k(t) &= \lambda x_{k-1}(t) + 2\mu x_{k+1}(t) - (\lambda+2\mu)x_k(t), \quad k \geq j+1.
\end{aligned}$$

The appropriate set of differential equations will of course change depending upon which j is the median queue size. There will still be convergence to some equilibrium point (which we might hope will have a smaller average queue size.)

10.5 *Asymptotic optimality*

We now turn to the question of optimality or near optimality of the Whittle index policy. Taking $m = \alpha n$, let $R_W^{(n)}(\alpha)$, $R_{\text{opt}}^{(n)}(\alpha)$ and $r(\alpha)$ denote, respectively, the average reward that is obtained from n restless bandits under the Whittle index policy, under an optimal policy, and from a single bandit under the relaxed policy (that the bandit receive the action $u = 1$ for a fraction α of the time). Then

$$R_W^{(n)}(\alpha) \leq R_{\text{opt}}^{(n)}(\alpha) \leq nr(\alpha).$$

It is plausible that the Whittle index policy should be **asymptotically optimal** as n becomes large, in the sense that $r(\alpha) - R_W^{(n)}(\alpha)/n \to 0$ as $n \to \infty$. This is true if certain differential equations have an asymptotically stable equilibrium point (i.e. a point to which they converge from any starting state). These are the differential equations which describe a fluid approximation to the stochastic process of the bandit states evolving under the Whittle index policy.

Suppose bandits move on a state space of size k and let $z_i(t)$ be the proportion of the bandits in state i. The 'fluid approximation' for large n is given by piecewise linear differential equations, of the form:

$$\frac{dz}{dt} = A(z)x + b(z),$$

where $A(z)$ and $b(z)$ are constants within k polyhedral regions which partition the positive orthant of $\mathbb{R}^k$. For example for $k = 2$,

$$\frac{dz_i}{dt} = \sum_j q_{ji}(z)z_j - \sum_j q_{ij}(z)z_i$$

$$\frac{dz_1}{dt} = \begin{cases} -(q_{12}^0 + q_{21}^0)z_1 + (q_{12}^0 - q_{12}^1)\rho + q_{21}^0\,, & z_1 \geq \rho \\ -(q_{12}^1 + q_{21}^1)z_1 - (q_{21}^0 - q_{21}^1)\rho + q_{21}^0\,, & z_1 \leq \rho \end{cases}$$

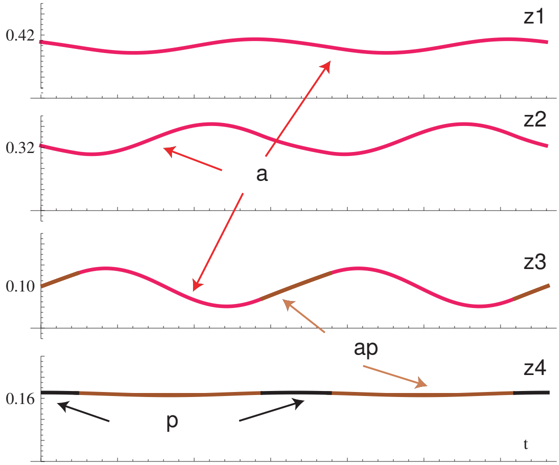
It turns out that there are examples of dual-speed restless bandits (needing $k > 3$) in which the differential equations have an asymptotically stable equilibrium cycle (rather than a stable equilibrium point), and this can lead to suboptimality of the Whittle index policy. However, in examples, the suboptimality was never found to be more than about one part in 10^4.

Theorem 10.4. *If bandits are indexable, and the fluid model has an asymptotically stable equilibrium point, then the Whittle index heuristic is asymptotically optimal, — in the sense that the reward per bandit tends to the reward that is obtained under the relaxed policy.*

Here is an example where the Whittle index does not quite provide asymptotically optimal performance.

$$\left(q_{ij}^0\right) = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 2 & -2 & 0 & 0 \\ 0 & 56 & -\frac{113}{2} & \frac{1}{2} \\ 1 & 1 & \frac{1}{2} & -\frac{5}{2} \end{pmatrix}, \quad \left(q_{ij}^1\right) = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 2 & -2 & 0 & 0 \\ 0 & \frac{7}{25} & -\frac{113}{400} & \frac{1}{400} \\ 1 & 1 & \frac{1}{2} & -\frac{5}{2} \end{pmatrix}$$

$$r^0 = (0, 1, 10, 10)\,, \quad r^1 = (10, 10, 10, 0)\,, \quad \rho = 0.835$$



Equilibrium point is $(\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4) = (0.409, 0.327, 0.100, 0.164)$. $\bar{z}_1 + \bar{z}_2 + \bar{z}_3 = 0.836$. The equilibrium is a cycle. Relaxed policy obtains average reward 10 per bandit. Heuristic obtains only 9.9993 per bandit.

It is tempting to try to generalize the idea of a Whittle index for restless bandits to problems with a discounted reward criterion, starting with the appropriate functional equation in place of and adding a subsidy for use of the passive action. However, there is no asymptotic optimality result for this case that is analogous to the result of for the average reward case. The use of discounted versions of Whittle indices can actually end up recommending the worst of all priority policies, and a payoff that is very far from the optimum. This is because the identity of the optimal priority policy can critically depend on the starting state of the n restless bandits, whereas the ordering of Whittle indices is calculated without reference to the starting state.

11 Sequential Assignment and Allocation Problems

Having met the Secretary problem, Multi-armed bandit problem, etc., we now turn to some other very well-known and interesting problems that I have personally enjoyed.

11.1 Sequential stochastic assignment problem

Derman, Lieberman and Ross (1974) defined the following **sequential stochastic assignment problem** (SSAP). It has been applied in many contexts, including kidney transplantation, aviation security, buying decisions in supply chains, and real estate.

There are n workers available to perform n jobs. First job 1 appears, followed by job 2, etc. Associated with the jth job is a random variable X_j which takes the value x_j. $X_1, \dots, X_n$ are i.i.d. random variables with distribution function G. If a 'perfect' worker is assigned to the value x_j job, a reward x_j is obtained. However, none of the workers is perfect, and whenever the ith worker is assigned to any type x_j job, the (expected) reward is given by $p_i x_j$, where $0 < p_i < 1$ is a known constant. After a worker is assigned, he is unavailable for future assignments. The problem is to assign the n workers to the n jobs so as to maximize the total expected reward. A policy is a rule for assigning workers to jobs. Let random variable i_j be the worker (identified by number) assigned to the jth arriving job. The expected reward to be maximized is

$$\sum_{j=1}^{n} E[p_{i_j} X_j].$$

The optimal policy is given by the following theorem. The surprise is that the thresholds $\alpha_{i,n}$ *are independent of the p's.*

Theorem 11.1. *For each $n > 1$, there exist numbers $a_{0,n} > a_{1,n} > a_{2,n} > \cdots > a_{n,n} = 0$ such that whenever there are n stages to go and $p_1 > \cdots > p_n$ then the optimal choice in the initial stage is to use p_i if the random variable X_1 is contained in the interval $(a_{i-1,n}, a_{i,n}]$. The $a_{i,n}$ depend on G but are independent of the p_i.*

Furthermore $a_{i,n}$ is the expected value, in an $n-1$ stage problem, of the quantity to which the ith largest p is assigned (assuming an optimal policy is followed), and

$$F(p_1, \dots, p_{n-1}) = \sum_{i=1}^{n-1} p_i a_{i,n}, \quad p_1 > \cdots > p_{n-1}.$$

Proof. The proof is by induction. Assuming it is true for $n-1$, we have for n,

$$\begin{aligned} F(p_1 \dots, p_n \mid x) &= \max_i \{ x p_k + F(p_1, \dots, p_{k-1}, p_{k+1}, \dots, p_n) \} \\ &= \max_k \left\{ \sum_{j=1}^{k-1} p_j a_{j,n} + p_k x + \sum_{j=k+1}^{n} p_j a_{j-1,n} \right\}. \end{aligned}$$

We use the **Hardy-Littlewood rearrangement inequality**, which says that if $a < A$ and $b < B$ then $ab + AB > aB + bA$. That is, it is best to match smallest with smallest, largest with largest.

Suppose $X_1 = x_1$ and $\{a_{1,n}, \dots, a_{i-1,n}, x_1, a_{i,n}, \dots, a_{n-1,n}\}$ is a decreasing sequence. Then the optimum matching of these numbers against the decreasing numbers $\{p_1, p_2, \dots, p_n\}$, is to form the sum of the products obtained by matching, for each j, the jth largest of $\{a_{1,n}, \dots, a_{i-1,n}, x_1, a_{i,n}, \dots, a_{n-1,n}\}$ with jth largest of $\{p_1, p_2, \dots, p_n\}$, which means that x_1 should be matched with p_i. Notice that $a_{k,n+1}$ is the expected value of the X that gets matched to p_k. This value does not depend on the values of the p_is. □

We could implement the optimal strategy by offering each job to the workers in decreasing order of their p values. Worker i will accept the job if workers $1, \dots, i-1$ reject it, and then $X \geq a_{i,n}$, since the jobs is worth is as much to him as he would expect to get if he forgoes it and then faces a $n-1$ stage problem (where his expected match is $a_{i,n}$). This is nice. We can obtain the socially optimal allocation by presenting the workers with a problem that they each solve from an individually optimal viewpoint.

11.2 Sequential allocation problems

Groundwater Management Burt (1965). Each day water is to be pumped from an aquifer, and replenished by a random amount of rainfall R_t. The aim is to maximize an expected sum of utilities $\sum a(y_t)$ minus pumping cost $\sum c(x_t, y_t)$.

$$F(x, s) = \max_{y \in [0,x]} \{a(y) - c(x, y) + \beta E F(x - y + R_s, s - 1)\}$$

$$F(x, 0) = 0. \qquad x \text{ is level of water in an aquifer.}$$

Here $s = h - t$ is time-to-go.

Investment problem Derman, Lieberman and Ross (1975). With probability p_t there is an opportunities to invest part of ones capital at time t. The aim is maximize the expected sum of $\sum a(y_t)$.

$$F(x, s) = q_s F(x, s - 1) + p_s \max_{y \in [0,x]} \{a(y) + F(x - y, s - 1)\}$$

$$F(x, 0) = 0. \qquad x \text{ is remaining capital of dollars.}$$

General fighter problem With probability p_t there is an opportunity for a fighter to shoot down an enemy plane. If m missiles are used then the enemy plane is destroyed with probability $a(m)$ and the fighter survives the dogfight with probability $c(m)$. The aim is to maximize the expected number of enemy planes destroyed.

$$F(n, s) = q_s F(n, s - 1) + p_s \max_{m \in \{1,\dots,n\}} \{a(m) + c(m) F(n - m, s - 1)\}$$

$$F(n, 0) = 0. \quad n \text{ is remaining stock of missiles.}$$

Bomber problem Klinger and Brown (1968). With probability p_t a bomber must defend itself against an attack and wishes to maximize the probability of reaching its final target.

$$\begin{aligned}P(n,s) &= P(\text{survive to for } s \text{ further distance})\\ &= q_s P(n,s-1) + p_s \max_{m\in\{1,\dots,n\}} c(m)P(n-m,s-1)\end{aligned}$$

$P(n,0)=1$. n is remaining stock of missiles. Typically, $c(m)=1-\theta^m$.

The bomber problem can also be posed in continuous time, as

$$P(n,t)=e^{-t}+\int_0^t \max_{m\in\{1,\dots,n\}} c(m)P(n-m,s)e^{-(t-s)}\,ds.$$

Intuitively obvious properties of an optimal policy are (for the bomber problem, and other problems similarly)

$$\begin{array}{lcl}\textbf{(A)} & m(n,s) & \searrow \text{ as } s\nearrow\\ \textbf{(B)} & m(n,s) & \nearrow \text{ as } n\nearrow\\ \textbf{(C)} & n-m(n,s) & \nearrow \text{ as } n\nearrow\end{array}$$

Properties like these are sometimes quite easy to prove. Sometimes this is by a value iteration approaching, proving that the value function has appropriate concavity properties. Or sometimes an interhange arguments helps. Consider **(C)** for the Bomber Problem. We shall assume that $\log c(m)$ is concave in m.

Proof of **(C)** *of the Bomber Problem.* Let

$$m(n,s)=\arg\max_{m\in\{1,\dots,n\}} c(m)P(n-m,s-1).$$

We wish to show that $n-m(n,s-1)$ is nondecreasing in n. Suppose this were not the case. So perhaps $m=m(n,s-1)$ but $m'=m(n+1,s-1)$ with $n-m>n+1-m'$, i.e., $m'>m+1$. Consider the product of the survival probabilities

$$c(m)P(n-m,s-1)\times c(m')P(n+1-m',s-1) \tag{11.1}$$

Let $\bar{m}=m'-1$ and $\bar{m}'=m+1$. This different choice of amounts to fire in state $(n,s-1)$ and $(n+1,s-1)$ would have a product of survival probabilities

$$\begin{aligned}&c(\bar{m}))P(n-\bar{m},s-1)\times c(\bar{m}')P(n+1-\bar{m}',s-1)\\ &=c(m'-1)P(n+1-m',s-1)\times c(m+1)P(n-m,s-1)\end{aligned} \tag{11.2}$$

$$\begin{aligned}&(11.2)-(11.1)\\ &\quad=\Big[c(m+1)c(m'-1)-c(m)c(m')\Big]P(n-m,s-1)P(n+1-m',s-1)\\ &\quad\geq 0,\end{aligned}$$

since $\log c(m)$ is concave means that $m' > m+1 \implies \dfrac{c(m+1)}{c(m)} > \dfrac{c(m')}{c(m'-1)}$.

Hence at least one of our original m and m' must not have been optimal. □

However **(B)** for the Bomber problem remains an unproven conjecture, as also is **(A)** for the General fighter problem. It has been shown there are **(B)** is not true for the Bomber problem if $c(m)$ is an arbitrary concave function. However, opinion is very divided about whether or not **(B)** might be true for the special concave function $c(m) = 1 - \theta^m$. Very extensive computation have turned up no counterexample.

11.3 *SSAP with arrivals*

Suppose workers and jobs arrive according to Poisson processes with respective rates γ and λ. The workers are identical with $p_i = 1$ (the so-called house-selling case). Rewards are exponentially discounted with rate α. Job values are i.i.d., say $U[0,1]$.

Suppose that an arriving job is offered to the workers in inverse order of their arrival times; so the worker that arrived most recently has first right of refusal for jobs, and workers try to maximize their own expected job values. The individually optimal (IO) policy is the Nash equilibrium of a noncooperative game; that is, if all workers follow the IO policy and each worker is trying to maximize its own expected job value, then no worker will have incentive to deviate from the IO policy. Righter (2011) has shown that the IO policy is unique and has proved the following.

Theorem 11.2. *The IO policy is socially optimal (maximizing total discounted return). Thus the socially optimal policy is a threshold policy.*

A worker who is ith to be offered a job of value x should accept it iff $x \geq t_i$, where $t_1 > t_2 \cdots$ and t_i is the expected discounted job value that is allocated to worker i under the IO policy. Use uniformization, so $\gamma + \lambda + \alpha = 1$. This mean we can now think of γ and λ as probabilities, and of our problem as taking place in discrete time. The thresholds are given by thinking about worker i being indifferent between accepting and rejecting a job:

$$\begin{aligned} t_i = {} & \underbrace{\gamma\, t_{i+1}}_{\text{new worker arrives}} \\ & + \lambda\Big[\underbrace{t_i P(X < t_i)}_{\substack{\text{new job assigned to} \\ \text{worker behind } i}} + \underbrace{E[X\, 1_{\{t_i \leq X < t_{i-1}\}}]}_{\substack{\text{new job assigned to} \\ \text{worker } i}} + \underbrace{t_{i-1} P(X \geq t_{i-1})}_{\substack{\text{new job assigned to} \\ \text{worker ahead of } i}}\Big]. \end{aligned}$$

Proof. We show that following the IO policy, starting with the first decision, is better (for the sum of the worker's obtained values) than following an arbitrary policy for the first decision and then switching to the IO policy thereafter (call the latter policy π). Essentially we are showing that the policy improvement algorithm cannot improve IO.

Suppose at the first decision, $t_{i-1} > x \geq t_i$ but the job is assigned to no worker by π, and the IO policy is used thereafter. Workers $1, 2, \ldots, i-1$ and all future arriving workers will have the same expected job value (EJV) under IO and π. Worker i would have had x under IO, but will only get t_i under π. Workers $i+1,\dots,n$ will also have greater EJVs from time 1 onward once i has been assigned a job, so for them also IO is better than π. So the job should be assigned.

What if there are n workers present and $x < t_n$? IO rejects the job. If a policy π assigns the job to a worker, which we may take to be worker n, then all other workers have the same EVJ under IO and π, but worker n is taking x, whereas under π his EJV would be t_n, which is greater. So π cannot be optimal. □

11.4 *SSAP with a postponement option*

Consider now a SSAP with m perfect workers ($p_i = 1$), and n ($> m$) jobs to be presented sequentially, with i.i.d. values $X_1,\dots,X_n \sim U[0,1]$, and discounting. We no longer demand that a job must be assigned or rejected upon first seeing its value. It may be held in a queue, for possible assignment later. For a state in which there are $s = n-t$ jobs still to see, m workers still unassigned, and a queue holding jobs of values $x_1 > \cdots > x_m$ (some of these can be 0) the optimality equations are

$$F_{s,m}(x_1,\dots,x_m) = \int_0^1 G_{s,m}(T(x,x_1,\dots,x_m))\,dx$$

$$G_{s,m}(x_1,\dots,x_m) = \max_{i\in\{0,\dots,m\}} \left\{ \sum_{j=1}^{i} x_j + \beta F_{s,m-i}(x_{i+1},\dots,x_m) \right\},$$

where $T(x,x_1,\dots,x_m)$ is the vector formed from $\{x,x_1,\dots,x_m\}$ by discarding the smallest element and then rearranging the rest in decreasing order. We have written the optimality equation in two parts. The first equation is about receiving the next job. The second is about assigning jobs that are presently in the postponement queue. Feng and Hartman (2012) have proved the following.

Theorem 11.3. *An optimal policy is to assign the available job of greatest value, x, iff $x \geq \alpha_{m,s}$ where the threshold $\alpha_{m,s}$ depends on the numbers of unassigned workers, m, and jobs left to see, s, but not on the values of the other jobs in the queue.*

The final part of this statement is rather surprising. One might have expected that it would be optimal to assign the jobs of greatest value if and only if it is greater than some threshold, but where this is a complicated function of m, s and the values of the other jobs that are in the queue and are available to be assigned later.

The proof is very complicated. It would be nice to find an simpler proof, perhaps by thinking about re-casting the problem into one about individual optimality, as Righter did with the arrivals case of the SSAP.

11.5 *Stochastic knapsack and bin packing problems*

Similar to the problems we have addressed thus far is the stochastic **knapsack problem**. It has been applied within the field of revenue management. The capacity of the knapsack is a given amount of resource that can be used to fulfill customer demand over a certain time frame. Some examples include: rooms in a hotel aimed at weekend tourists, or seats on an airplane that must be sold before departure. Items of arrive with random sizes X_i and values R_i. We wish to maximize the expected total value of the items we can fit into a knapsack of size c.

For example, Coffman, Flatto, Weber (1987) have studied a stochastic **bin packing problem**. Items of sizes $X_1, X_2, \dots, X_n$ are encountered sequentially. These are i.i.d. with distribution function G. We wish to maximize the expected number of items that we can pack into a bin of size c. Each item must be accepted or rejected at the time it is encountered.

Let x by the space that is left in the bin. The dynamic programming equation is

$$
\begin{aligned}
&F_s(x) = (1 - G(x))F_{s-1}(x) + \int_0^x \max\{F_{s-1}(x), 1 + F_{s-1}(x-y)\} dG(y) \\
&F_0(x) = 0.
\end{aligned}
$$

The optimal rule is to accept the sth to last item iff doing so would leave remaining space in the bin of at least $z_{s-1,x}$, where $F_{s-1}(z_{s-1,x}) + 1 = F_{s-1}(x)$.

12 LQ Regulation

Models with linear dynamics and quadratic costs in discrete and continuous time. Riccati equation, and its validity with additive white noise. Linearization of nonlinear models.

12.1 The LQ regulation problem

A control problem is specified by the dynamics of the process, which quantities are observable at a given time, and an optimization criterion.

In the **LQG model** the dynamical and observational equations 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.

To begin, suppose 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

$$x_t = Ax_{t-1} + Bu_{t-1}, \tag{12.1}$$

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

$$\mathbf{C} = \sum_{t=0}^{h-1} c(x_t, u_t) + \mathbf{C}_h(x_h), \tag{12.2}$$

with one-step and terminal costs

$$c(x,u) = x^\top Rx + u^\top Sx + x^\top S^\top u + u^\top Qu = \begin{pmatrix} x \\ u \end{pmatrix}^\top \begin{pmatrix} R & S^\top \\ S & Q \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}, \tag{12.3}$$

$$\mathbf{C}_h(x) = x^\top \Pi_h x. \tag{12.4}$$

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

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

which is symmetric, with $\Pi_{uu} > 0$, 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. Consider the identity, obtained by 'completing the square',

$$\begin{aligned}&\begin{pmatrix}x\\u\end{pmatrix}^\top\begin{pmatrix}\Pi_{xx} & \Pi_{xu}\\ \Pi_{ux} & \Pi_{uu}\end{pmatrix}\begin{pmatrix}x\\u\end{pmatrix}\\ &\quad=\left(u+\Pi_{uu}^{-1}\Pi_{ux}x\right)^\top\Pi_{uu}\left(u+\Pi_{uu}^{-1}\Pi_{ux}x\right)+x^\top\left(\Pi_{xx}-\Pi_{xu}\Pi_{uu}^{-1}\Pi_{ux}\right)x. \qquad (12.5)\end{aligned}$$

□

Theorem 12.2. *Assuming* (12.1)–(12.4)*, the value function has the quadratic form*

$$F(x,t)=x^\top\Pi_t x, \quad t\le h, \qquad (12.6)$$

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*

$$\Pi_t=f\Pi_{t+1}, \quad t<h, \qquad (12.7)$$

where Π_h *has the value given in* (12.4)*, and* f *is an operator having the action*

$$f\Pi=R+A^\top\Pi A-(S^\top+A^\top\Pi B)(Q+B^\top\Pi B)^{-1}(S+B^\top\Pi A). \qquad (12.8)$$

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

$$K_t=-(Q+B^\top\Pi_{t+1}B)^{-1}(S+B^\top\Pi_{t+1}A), \quad t<h. \qquad (12.9)$$

Proof. Assertion (12.6) 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 12.1 shows the minimizer is $u=K_t x$, and gives the form of f. □

The backward recursion (12.7)–(12.8) is called the **Riccati equation**.

12.2 White noise disturbances

Suppose the plant equation (12.1) 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 $E\epsilon_t\epsilon_s^\top=0$, $t\ne s$. The dynamic programming equation is then

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

with $F(x,h) = x^\top \Pi_h x$. 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

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

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

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

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

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

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 $\operatorname{tr}(N\Pi)$, and a transient cost of $x^\top \Pi x$ that is due to correcting the initial x.

12.3 Example: control of an inertial system

Consider a system, with state $(x_t, v_t) \in \mathbb{R}^2$, being position and velocity,

$$\begin{pmatrix} x_{t+1} \\ v_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ v_t \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u_t + \begin{pmatrix} 0 \\ \epsilon_t \end{pmatrix},$$

with $\{u_t\}$ being controls making changes in velocity, and $\{\epsilon_t\}$ being independent disturbances, with means 0 and variances N. This is as §10.3 with $n = 2$, $m = 1$.

Suppose we wish to minimize the expected value of

$$\textstyle\sum_{t=0}^{h-1} u_t^2 + \Pi_0 x_h^2, \quad \text{which equals } \sum_{t=0}^{h-1} u_t^2 + \Pi_0 z_h^2,$$

when re-write the problem in terms of the scalar variable $z_t = x_t + (h-t)v_t$. This is the expected value of x_h if no further control are applied. In terms of $s = h - t$,

$$z_{s-1} = z_s + (s-1)u_t + (s-1)\epsilon_t.$$

Try $F_{s-1}(z) = z^2\Pi_{s-1} + \gamma_{s-1}$, which is true at $s = 1$, since $F_0(z) = z^2\Pi_0$. Then

$$\begin{aligned}
F_s(z) &= \inf_u \left[u^2 + EF_{s-1}(z + (s-1)u + (s-1)\epsilon)\right] \\
&= \inf_u \left[u^2 + E\left[z + (s-1)u + (s-1)\epsilon\right]^2 \Pi_{s-1} + \gamma_{s-1}\right] \\
&= \inf_u \left[u^2 + \left[(z + (s-1)u)^2 + (s-1)^2 N\right] \Pi_{s-1} + \gamma_{s-1}\right]. \\
&= \inf_u \Bigg[(1 + (s-1)^2\Pi_{s-1}) \left(u + \frac{(s-1)\Pi_{s-1}z}{1 + (s-1)^2\Pi_{s-1}}\right)^2 \\
&- \frac{(s-1)^2\Pi_{s-1}^2 z^2}{1 + (s-1)^2\Pi_{s-1}} + \Pi_{s-1}z^2 + (s-1)^2 N\Pi_{s-1} + \gamma_{s-1}\Bigg] \\
&= -\frac{(s-1)^2\Pi_{s-1}^2 z^2}{1 + (s-1)^2\Pi_{s-1}} + \Pi_{s-1}z^2 + (s-1)^2 N\Pi_{s-1} + \gamma_{s-1} \\
&= \frac{\Pi_{s-1}z^2}{1 + (s-1)^2\Pi_{s-1}} + (s-1)^2 N\Pi_{s-1} + \gamma_{s-1}
\end{aligned}$$

So we obtain the Riccati equation

$$\Pi_s = \frac{\Pi_{s-1}}{1 + (s-1)^2\Pi_{s-1}},$$

and optimal control

$$u_t = -\frac{(s-1)\Pi_{s-1}z_t}{1 + (s-1)^2\Pi_{s-1}} = -(s-1)\Pi_s(x_t + sv_t).$$

By taking the reciprocal of the Riccati equation for Π_s, we have

$$\Pi_s^{-1} = \Pi_{s-1}^{-1} + (s-1)^2 = \cdots = \Pi_0^{-1} + \sum_{i=1}^{s-1} i^2 = \Pi_0^{-1} + \tfrac{1}{6}s(s-1)(2s-1).$$

We also see that

$$\gamma_s = (s-1)^2 N\Pi_{s-1} + \gamma_{s-1}.$$

13 Controllability

Controllability in discrete and continuous time.

13.1 Controllability

The discrete-time system $[A, B, \cdot\,]$, with dynamical equation

$$x_t = Ax_{t-1} + Bu_{t-1}, \tag{13.1}$$

is said to be **r-controllable** if from any x_0 it can be brought to any x_r by some sequence of controls $u_0, u_1, \ldots, u_{r-1}$. It is **controllable** if it is r-controllable for some r

Example 13.1. Consider the case, ($n = 2$, $m = 1$),

$$x_1 - Ax_0 = Bu_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} u_0.$$

This system is not 1-controllable. But

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

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

In general, by substituting the plant equation (13.1) into itself, we see that we must find $u_0, u_1, \ldots, u_{r-1}$ to satisfy

$$\Delta = x_r - A^r x_0 = Bu_{r-1} + ABu_{r-2} + \cdots + A^{r-1}Bu_0, \tag{13.2}$$

for arbitrary Δ. In providing conditions for controllability we use the following theorem.

Theorem 13.2. (The Cayley-Hamilton theorem) *Any $n \times n$ matrix A satisfies its own characteristic equation. So $\sum_{j=0}^{n} a_j A^{n-j} = 0$, where*

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

The implication is that $I, A, A^2, \ldots, A^{n-1}$ contains a basis for A^r, $r = 0, 1, \ldots$. This implies the following theorem.

Theorem 13.3. *Suppose A is $n \times n$. The system $[A, B, \cdot\,]$ is r-controllable iff the matrix*

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

has rank n. It is controllable iff M_n has rank n.

If the system is controllable then a control transferring x_0 to x_r with minimal cost $\sum_{t=0}^{r-1} u_t^\top u_t$ is

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

13.2 Controllability 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 $(= \Pi(t))$ obeys the Riccati equation

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

We find $u(t) = K(t)x(t)$, where $K(t) = -Q^{-1}(S + B^\top \Pi)$, and $\dot{x} = \Gamma(t)x$. These are slightly simpler than in discrete time.

Theorem 13.4. *The n dimensional system $[A, B, \cdot\,]$ is controllable iff the matrix M_n has rank n.*

Note that there is now no notion of r-controllability. However, 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

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

$G(t) \downarrow 0$ as $t \downarrow 0$, so the transfer becomes more difficult and costly as $t \downarrow 0$.

13.3 Linearization of nonlinear models

Linear models are important because they arise naturally via the linearization of non-linear models. Consider a continuous time 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 $u' = u - \bar{u}$. The linearized version is

$$\dot{x}' = \dot{x} = a(\bar{x} + x', \bar{u} + u') = Ax' + Bu, \quad \text{where } A_{ij} = \left.\frac{\partial a_i}{\partial x_j}\right|_{(\bar{x},\bar{u})}, \; B_{ij} = \left.\frac{\partial a_i}{\partial u_j}\right|_{(\bar{x},\bar{u})}.$$

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.

13.4 Example: broom balancing

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

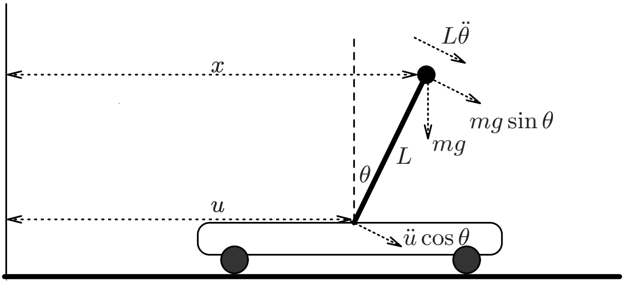


Figure 1: Force diagram for broom balancing

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

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

Since

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

the system is controllable if θ is initially small.

Suppose we try to control two brooms at equilibrium simultaneously. Then

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \beta & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ -\alpha \\ 0 \\ -\beta \end{pmatrix} \quad M_4 = \begin{pmatrix} 0 & -\alpha & 0 & -\alpha^2 \\ -\alpha & 0 & -\alpha^2 & 0 \\ 0 & -\beta & 0 & -\beta^2 \\ -\beta & 0 & -\beta^2 & 0 \end{pmatrix}$$

where $\alpha = g/L_1$, $\beta = g/L_2$.

So M_4 is of rank 4 only if α and β are different. It is not possible to simultaneously balance two brooms whose lengths are the same.

13.5 Infinite horizon limits

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 13.5. *Suppose* $\Pi_0 = 0$*,* $R \succeq 0$*,* $Q \succeq 0$ *and* $[A, B, \cdot\,]$ *is controllable. 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.

Hence $x^\top \Pi_s x$ is increasing and has a finite upper bounds, and so 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. □

13.6 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

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

The matrix Γ is called the **gain matrix**.

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 $x_t = \Gamma^t x_0$.

Γ is called a **stability matrix** if $x_t \to 0$ as $t \to \infty$.

In the continuous-time this happens iff all eigenvalues have negative real part.

In the discrete-time time it happens if all eigenvalues of lie strictly inside the unit disc in the complex plane, $|z| = 1$.

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

We can see that Lemma 13.5 holds under the condition that the system is stabilizble. Using $u_t = Kx_t$, we have $x_t = \Gamma^t x$ and $u_t = K\Gamma^t x$, so

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

So $F_s(x)$ is increasing, bounded above, and so tends to a limit.

14 Observability

LQ regulation problem over the infinite horizon. Observability.

14.1 Observability

The discrete-time system $[A, B, C]$ has (13.1), plus the observation equation

$$y_t = Cx_{t-1}. \tag{14.1}$$

The value of $y_t \in \mathbb{R}^p$ is observed, but x_t is not. C is $p \times n$.

This 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 x_0. A system is **observable** if r-observable for some r.

From (13.1) and (14.1) we can determine y_t in terms of x_0 and subsequent controls:

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

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

$$\tilde{y}_t = CA^{t-1}x_0, \quad 1 \le t \le r. \tag{14.2}$$

By hypothesis, these equations are mutually consistent, and so have a solution; the question is whether this solution is unique.

Theorem 14.1. *(i) The system $[A, \cdot, C]$ is r-observable iff the matrix*

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

has rank n, or (ii) equivalently, iff 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*

$$x_0 = (N_r^\top N_r)^{-1} \sum_{j=1}^{r} (A^\top)^{j-1} C^\top \tilde{y}_j. \tag{14.3}$$

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

If we define the deviation $\eta_t = \tilde{y}_t - CA^{t-1}x_0$ then the equations amount 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 (14.3). If equations (14.2) indeed have a solution (i.e. are mutually consistent, as we suppose) and this is unique then expression (14.3) 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). □

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 to construct a 'best-fit' solution where an exact solution might not have existed.

14.2 Observability in continuous-time

Theorem 14.2. *(i) The n-dimensional continuous-time system* $[A, \cdot, C]$ *is observable iff the matrix* N_n *has rank n, or (ii) equivalently, iff*

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

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

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

14.3 *Example: satellite in a plane orbit*

A satellite of unit mass in a planar orbit has polar coordinates (r, θ) obeying

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

where u_r and u_θ are the radial and tangential components thrust. If $u_r = u_\theta = 0$ then there is an equilibrium orbit as 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, after some algebra,

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

Controllability. It is easy to check that $M_2 = \begin{bmatrix} B & AB \end{bmatrix}$ has rank 4 and so 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, 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, $2\omega\rho\delta r + \rho^2\delta\dot\theta = \delta(r^2\dot\theta)|_{r=\rho,\dot\theta=\omega}$.

Observability. 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.

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

15 Kalman Filter and Certainty Equivalence

The Kalman filter. Certainty equivalence. Separation principle

15.1 Imperfect state observation with noise

The full LQG model assumes linear dynamics, quadratic costs and Gaussian noise. Imperfect observation is the most important point. The model is

$$x_t = Ax_{t-1} + Bu_{t-1} + \epsilon_t, \tag{15.1}$$

$$y_t = Cx_{t-1} + \eta_t, \tag{15.2}$$

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 Gaussian white noise with

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

and that $x_0 \sim N(\hat{x}_0, V_0)$. Let $W_t = (Y_t, U_{t-1}) = (y_1, \dots, y_t; u_0, \dots, u_{t-1})$ 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 15.1. *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

$$E(x \mid y) = V_{xy}V_{yy}^{-1}y, \tag{15.3}$$

and

$$cov(x \mid y) = V_{xx} - V_{xy}V_{yy}^{-1}V_{yx}. \tag{15.4}$$

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 (15.4) □

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.

15.2 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 $u_0, \dots, u_{t-1}$.

$$\begin{aligned} x_t &= A^t x_0 + A^{t-1}Bu_0 + \cdots + Bu_{t-1} + A^{t-1}\epsilon_0 + \cdots + A\epsilon_{t-1} + \epsilon_t \\ y_t &= C\Big(A^{t-1}x_0 + A^{t-2}Bu_0 + \cdots + Bu_{t-2} + A^{t-2}\epsilon_0 + \cdots + A\epsilon_{t-2} + \epsilon_{t-1}\Big) + \eta_t \end{aligned}$$

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. Notice that V_t is policy independent (does not depend on $u_0, \dots, u_{t-1}$).

The following theorem describes recursive updating relations for $\hat{x}_t$ and V_t.

Theorem 15.2. *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* (15.1)–(15.2)*. 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*

$$\hat{x}_t = A\hat{x}_{t-1} + Bu_{t-1} + H_t(y_t - C\hat{x}_{t-1}), \tag{15.5}$$

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

$$gV = N + AVA^\top - (L + AVC^\top)(M + CVC^\top)^{-1}(L^\top + CVA^\top). \tag{15.6}$$

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

$$H_t = (L + AV_{t-1}C^\top)(M + CV_{t-1}C^\top)^{-1}. \tag{15.7}$$

The updating of $\hat{x}_t$ in (15.5) is known as the **Kalman filter**. The estimate of x_t is a combination of a prediction, $A\hat{x}_{t-1} + Bu_{t-1}$, and observed error in predicting y_t.

Compare (15.6) to the similar Riccati equation in Theorem 12.2. Notice that (15.6) computes V_t forward in time ($V_t = gV_{t-1}$), whereas (12.8) 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 chosen 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 15.1 that the distribution of ξ_t conditional on knowing $(W_{t-1}, u_{t-1}, \zeta_t)$, (which is equivalent to knowing $W_t = (Y_t, U_{t-1})$), 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 (15.5)–(15.7). □

15.3 Certainty equivalence

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

Theorem 15.3. *Suppose LQG model assumptions hold. Then (i) the value function is of the form*

$$F(W_t) = \hat{x}_t^\top \Pi_t \hat{x}_t + \cdots \qquad (15.8)$$

where $\hat{x}_t$ is the linear least squares estimate of x_t whose evolution is determined by the Kalman filter in Theorem 15.2 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 12.2.

Proof. The proof is by backward induction. Suppose (15.8) holds at t. Recall that

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

Using the fact that $c(x,u)$ is a quadratic cost,

$$\begin{aligned}F(W_{t-1}) &= \min_{u_{t-1}} E\left[c(x_{t-1}, u_{t-1}) + \hat{x}_t \Pi_t \hat{x}_t + \cdots \mid W_{t-1}, u_{t-1}\right]\\ &= \min_{u_{t-1}} E\Bigg[c(\hat{x}_{t-1} - \Delta_{t-1}, u_{t-1})\\ &\qquad + (A\hat{x}_{t-1} + Bu_{t-1} + H_t\zeta_t)^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1} + H_t\zeta_t)\\ &\qquad + \cdots \,\Bigg|\, W_{t-1}, u_{t-1}\Bigg] \qquad (15.9)\\ &= \min_{u_{t-1}} \left[c(\hat{x}_{t-1}, u_{t-1}) + (A\hat{x}_{t-1} + Bu_{t-1})^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1})\right] + \cdots ,\end{aligned}$$

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 (15.9) the expectations of all terms which are linear in these quantities are zero, like $E[\hat{x}_{t-1}^\top R\Delta_{t-1}]$, 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 $+\cdots$). $\square$

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**.

15.4 *Risk-sensitive LEQG*

Suppose we wish to minimize

$$\gamma_\pi(\theta) = -\theta^{-1}\log\left[E_\pi\left(e^{-\theta C}\right)\right] \sim E_\pi C - \tfrac{1}{2}\theta\,\mathrm{var}_\pi(C) + \cdots$$

The variance of C enters as a first order term in θ. When θ is positive, zero or negative we are correspondingly risk-seeking, risk-neutral or risk-averse.

The LQG model with cost function $\gamma_\pi(C)$, where C is of the usual quadratic form, is quite naturally labelled LEQG (EQ meaning 'exponential of a quadratic').

At time t, when we are about to choose u_t. Certain things are known. i.e. $u_0, \ldots, u_{t-1}$ and $y_1, \ldots, y_t$. Other things are unknown, such as $x_0, \ldots, x_h$, $y_{t+1}, \ldots, y_{h-1}$, $\hat{x}_{t+1}, \ldots, \hat{x}_{h-1}$. Suppose, by an inductive hypothesis, we know that controls at times $s = t+1, \ldots, h-1$ will be certain linear functions of the estimated state $\hat{x}_s$. Then, conditional on known information all unknowns are jointly Gaussian. Assume $\theta > 0$. We can compute

$$\begin{aligned}F(W_t, t) &= -(1/\theta)\log\sup_{u_t} E\left[e^{-\theta C_t(\square)} \,\Big|\, W_t\right]\\ &= -(1/\theta)\log\sup_{u_t}\int e^{-\theta C_t(\square) - D(\square)}\,d\square\end{aligned}$$

where this is to be understood as integrating out all Gaussian unknowns against their joint density function, $\exp(-D(\square))$, where D is a quadratic in these variables. A key fact about integrating out Gaussian variables is that

$$\int e^{-\theta C_t(\square) - D(\square)}\,d\square \propto e^{-\inf_\square[\theta C_t(\square) + D(\square)]}$$

where the proportionality constant is policy independent and infimum on the right hand side is achieved at $\square = \hat{\square}$. For $\theta = 0$ this means least squares estimates. Thus we see a risk-sensitive certainty equivalence and separation principles in operation. We should first determine the minimizing $\hat{\square}$, and then choose u_t to minimize $S_t = C_t(\hat{\square}) + \theta^{-1}D(\hat{\square})$.

16 Dynamic Programming in Continuous Time

The HJB equation for dynamic programming in continuous time.

16.1 Example: LQ regulation in continuous time

Suppose $\dot{x} = u$, $0 \le t \le T$. The cost is to be minimized is $\int_0^T u^2 dt + Dx(T)^2$.

Method 1. By dynamic programming, for small δ,

$$F(x,t) = \inf_u \Big[u^2\delta + F(x+u\delta, t+\delta) \Big]$$

with $F(x,T) = Dx^2$. This gives

$$0 = \inf_u \Big[u^2 + uF_x(x,t) + F_t(x,t) \Big].$$

So $u = -(1/2)F_x(x,t)$ and hence $(1/4)F_x^2 = F_t$. Can we guess a solution to this? Perhaps by analogy with our known discrete time solution $F(x,t) = \Pi(t)x^2$. In fact,

$$F(x,t) = \frac{Dx^2}{1+(T-t)D}, \quad \text{and so } u(0) = -\tfrac{1}{2}F_x = -\frac{D}{1+TD}x(0).$$

Method 2. Suppose we use a Lagrange multiplier $\lambda(t)$ for the constraint $\dot{x} = u$ at time t, and then consider maximization of the Lagrangian

$$L = -Dx(T)^2 + \int_0^T \Big[-u^2 - \lambda(\dot{x} - u) \Big] \, dt$$

which using integration by parts gives

$$= -Dx(T)^2 - \lambda(T)x(T) + \lambda(0)x(0) + \int_0^T \Big[-u^2 + \dot{\lambda}x + \lambda u) \Big] dt.$$

Stationarity with respect to small changes in $x(t)$, $u(t)$ and $x(T)$ requires $\dot{\lambda} = 0$, $u = (1/2)\lambda$ and $2Dx(T) + \lambda(T) = 0$, respectively. Hence u is constant,

$$x(T) = x(0) + uT = x(0) + (1/2)\lambda T = x(0) - TDx(T).$$

From this we get $x(T) = x(0)/(1+TD)$ and $u(t) = -Dx(0)/(1+TD)$.

16.2 The Hamilton-Jacobi-Bellman equation

In continuous time the plant equation is,

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

Consider a discounted cost of

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

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

$$F(x,t) = \inf_u \left[c(x,u,t)\delta + (1-\alpha\delta)F(x + a(x,u,t)\delta, t+\delta) + o(\delta)\right].$$

By considering the term of order δ in the Taylor series expansion we obtain,

$$\inf_u \left[c(x,u,t) - \alpha F + \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} a(x,u,t) \right] = 0, \quad t < h, \tag{16.1}$$

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

Equation (16.1) is called the **Hamilton-Jacobi-Bellman equation** (HJB). Its heuristic derivation we have given above is justified by the following theorem. It can be viewed as the equivalent, in continuous time, of the backwards induction that we use in discrete time to verify that a policy is optimal because it satisfies the the dynamic programming equation.

Theorem 16.1. *Suppose a policy π, using a control u, has a value function F which satisfies the HJB equation* (16.1) *for all values of x and t. Then π is optimal.*

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

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

The inequality is because the term round brackets is non-negative. Integrating this inequality along the v path, from $x(0)$ to $x(h)$, gives

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

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

16.3 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 $\int_0^T u e^{-\alpha t}\, dt$, subject to $0 \le u \le u_{\max}$.

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 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

$$\sup_u \left[u - \alpha F + \frac{\partial F}{\partial x}[a(x) - u] \right] = 0. \tag{16.2}$$

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

$$u = \begin{bmatrix} 0 \\ \text{undetermined, but effectively } a(\bar{x}) \\ u_{\max} \end{bmatrix} \text{ for } x \begin{bmatrix} < \\ = \\ > \end{bmatrix} \bar{x}. \tag{16.3}$$

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

$$F(\bar{x}) = \int_0^\infty e^{-\alpha t} a(\bar{x}) dt = a(\bar{x})/\alpha. \tag{16.4}$$

So from (16.2) and (16.4) we have

$$F_x(x) = \frac{\alpha F(x) - u(x)}{a(x) - u(x)} \to 1 \text{ as } x \nearrow \bar{x} \text{ or } x \searrow \bar{x}. \tag{16.5}$$

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

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

where the last line follows because (16.5) holds in a neighbourhood of $\bar{x}$. It is required that F_{xx} be negative. But the denominator changes sign at $\bar{x}$, so the numerator must do so also, and therefore we must have $a'(\bar{x}) = \alpha$. We now have the complete solution. The control in (16.3) has a value function F which satisfies the HJB equation.

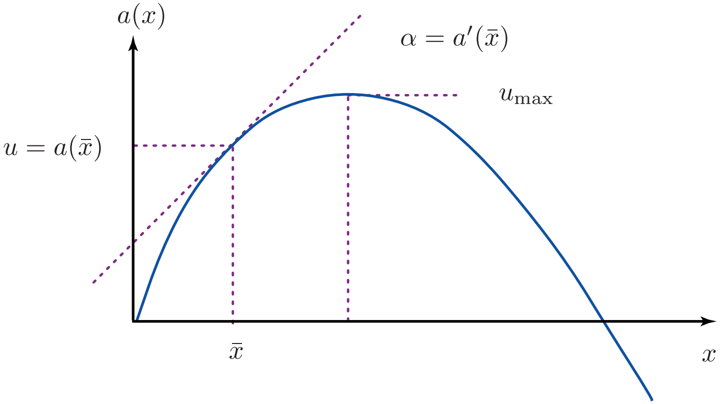


Figure 2: Growth rate $a(x)$ subject to environment pressures

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.

Finally, it would be good to verify that $F(x)$ is concave, as we conjectured from the start. The argument is as follows. Suppose $x > \bar{x}$. Then

$$\begin{aligned} F(x) &= \int_0^T u_{\max} e^{-\alpha t} dt + \int_T^\infty a(\bar{x}) e^{-\alpha t} dt \\ &= a(\bar{x})/\alpha + (u_{\max} - a(\bar{x}))\big(1 - e^{-\alpha T}\big)/\alpha \end{aligned}$$

where $T = T(x)$ is the time taken for the fish population to decline from x to $\bar{x}$, when $\dot{x} = a(x) - u_{\max}$. Now

$$\begin{aligned} T(x) = \delta + T(x + (a(x) - u_{\max})\delta) &\implies 0 = 1 + (a(x) - u_{\max})T'(x) \\ &\implies T'(x) = 1/(u_{\max} - a(x)) \end{aligned}$$

So $F''(x)$ has the same sign as that of

$$\frac{d^2}{dx^2}\big(1 - e^{-\alpha T}\big) = -\frac{\alpha e^{-\alpha T}(\alpha - a'(x))}{(u_{\max} - a(x))^2},$$

which is negative, as required, since $\alpha = a'(\bar{x}) \geq a'(x)$, when $x > \bar{x}$. The case $x < \bar{x}$ is similar.

17 Pontryagin's Maximum Principle

Pontryagin's maximum principle. Transversality conditions. Parking a rocket car.

17.1 Heuristic derivation of Pontryagin's maximum principle

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, $x(0)$. Thus, when PMP is useful, it finds an open-loop prescription of the optimal control. 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 problem with plant equation $\dot{x} = a(x, u)$ and instantaneous cost $c(x, u)$, both independent of t. The trajectory is to be controlled until it reaches some stopping set S, where there is a terminal cost $K(x)$. As in (16.1) the value function $F(x)$ obeys the dynamic programming equation (without discounting)

$$\inf_{u \in \mathcal{U}} \left[c(x, u) + \frac{\partial F}{\partial x} a(x, u) \right] = 0, \quad x \notin S, \tag{17.1}$$

with terminal condition

$$F(x) = K(x), \quad x \in S. \tag{17.2}$$

Define the **adjoint variable**

$$\lambda = -F_x. \tag{17.3}$$

This is column n-vector is 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**

$$H(x, u, \lambda) = \lambda^\top a(x, u) - c(x, u), \tag{17.4}$$

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

Theorem 17.1. (PMP) *Suppose $u(t)$ and $x(t)$ represent the optimal control and state trajectory. Then there exists an adjoint trajectory $\lambda(t)$ such that*

$$\dot{x} = \quad H_\lambda, \qquad [= a(x, u)\,] \tag{17.5}$$

$$\dot{\lambda} = -H_x, \qquad [= -\lambda^\top a_x + c_x\,] \tag{17.6}$$

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

$$H(x(t), v, \lambda(t)) \le H(x(t), u(t), \lambda(t)) = 0, \tag{17.7}$$

Moreover, if $x(T)$ is unconstrained then at $x = x(T)$ we must have

$$(\lambda(T) + K_x)^\top \sigma = 0 \tag{17.8}$$

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.' 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 (17.5) is immediate, and (17.7) follows from the fact that the minimizing value of u in (17.1) is optimal. Assuming u is the optimal control we have from (17.1) 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

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

which is (17.6).

Now 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 $(\lambda + K_x)^\top\sigma = 0$. □

Notice that (17.5) and (17.6) each give n equations. Condition (17.7) 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 $\lambda(t)$.

Requirements of (17.8) are known as **transversality conditions**.

17.2 Example: parking a rocket car

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

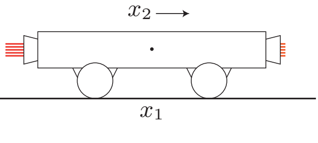


Figure 3: Optimal trajectories for parking problem

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 rocket car. Then

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

By making a Taylor expansion and then letting $\delta \to 0$ we find the HJB equation:

$$0 = \min_{-1 \leq u \leq 1} \left\{ 1 + \frac{\partial F}{\partial x_1} x_2 + \frac{\partial F}{\partial x_2} u \right\} \tag{17.9}$$

with boundary condition $F(0,0) = 0$. We can 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

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

Now let us tackle the same problem using PMP. We wish to minimize

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

where T is the first time at which $x = (0,0)$. For dynamics if $\dot{x}_1 = x_2$, $\dot{x}_2 = u$, $|u| \leq 1$, the Hamiltonian is

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

which is maximized by $u = \operatorname{sign}(\lambda_2)$. The adjoint variables satisfy $\dot{\lambda}_i = -\partial H / \partial x_i$, so

$$\dot{\lambda}_1 = 0, \qquad \dot{\lambda}_2 = -\lambda_1. \tag{17.10}$$

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

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

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 (17.1) or (17.9) 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 4).

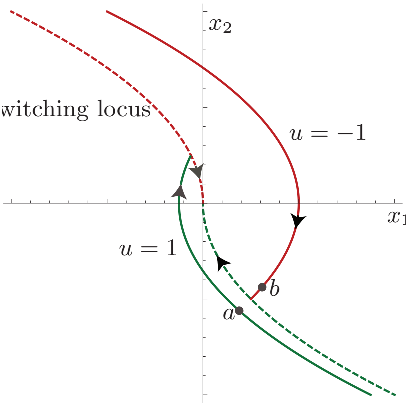


Figure 4: Optimal trajectories for parking a rocket car. Notice that the trajectories starting from two nearby points, a and b, are qualitatively different. The car starts to the right of the origin and moving towards it. From b one starts by further accelerating the movement towards the orgin. From a one starts by braking movement towards the origin.

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.

18 Using Pontryagin's Maximum Principle

Problems with explicit time. Examples with Pontryagin's maximum principle.

18.1 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

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

and hennce $\lambda_2(t) = 1$ for all t. Since H is maximized by u,

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

Since $\Delta(T) = -c$, 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

$$\dot{\lambda}_1 = \lambda_1 b - c. \tag{18.1}$$

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 (18.1) 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.

18.2 Problems in which time appears explicitly

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 17.1 says that $\tilde{H}$ must be maximized to 0. Equivalently, 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 17.1 still holds. However, to (17.6) we can now add

$$\dot{\lambda}_0 = -H_t = c_t - \lambda a_t, \tag{18.2}$$

and transversality condition

$$(\lambda + K_x)^\top \sigma + (\lambda_0 + K_t)\tau = 0, \tag{18.3}$$

which must hold at the termination point (x, t) if $(x + \epsilon\sigma, t + \epsilon\tau)$ is within $o(\epsilon)$ of the termination point of an optimal trajectory for all small enough positive ϵ.

18.3 Example: monopolist

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

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

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

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

$$x(0) = \int_0^T u\,dt = T - \left(1 - e^{-\alpha T}\right)/\alpha.$$

Thus $u(0) = 1 - e^{-\alpha T}$ is implicitly a function of $x(0)$, through T.

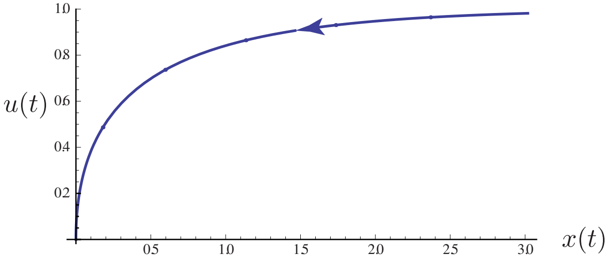


Figure 5: Trajectories of $x(t), u(t)$, for $\alpha = 1$.

The optimal value function is

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

18.4 Example: neoclassical economic growth

Suppose x is the existing capital per worker and u is consumption of capital per worker. The plant equation is

$$\dot{x} = f(x) - \gamma x - u, \tag{18.4}$$

where $f(x)$ is production per worker (which depends on capital available to the worker), and $-\gamma x$ represents depreciation of capital. We wish to choose u to maximize

$$\int_{t=0}^T e^{-\alpha t} g(u) dt,$$

where $g(u)$ measures utility and T is prescribed.

Solution. This is really the same as the fish harvesting example in §16.3, with $a(x) = f(x) - \gamma x$. So let us take

$$\dot{x} = a(x) - u. \tag{18.5}$$

It is convenient to take

$$H = e^{-\alpha t}\left[g(u) + \lambda(a(x) - u)\right]$$

so including a discount factor in the definition of u, corresponding to expression of F in terms of present values. Here λ is a scalar. Then $g'(u) = \lambda$ (assuming the maximum is at a stationary point), and

$$\frac{d}{dt}\left(e^{-\alpha t}\lambda\right) = -H_x = -e^{-\alpha t}\lambda a'(x) \tag{18.6}$$

or

$$\dot{\lambda}(t) = (\alpha - a'(x))\lambda(t). \tag{18.7}$$

From $g'(u) = \lambda$ we have $g''(u)\dot{u} = \dot{\lambda}$ and hence from (18.7) we obtain

$$\dot{u} = \frac{1}{\sigma(u)}[a'(x) - \alpha], \tag{18.8}$$

where

$$\sigma(u) = -\frac{g''(u)}{g'(u)}$$

is the elasticity of marginal utility. Assuming g is strictly increasing and concave we have $\sigma > 0$. So (x, u) are determined by (18.5) and (18.8). An equilibrium solution at $\bar{x}, \bar{u}$ is determined by

$$\bar{u} = a(\bar{x}) \quad a'(\bar{x}) = \alpha,$$

These give the balanced growth path; interestingly, it is independent of g.

This provides an example of so-called **turnpike theory**. For sufficiently large T the optimal trajectory will move from the initial $x(0)$ to within an arbitrary neighbourhood of the balanced growth path (the turnpike) and stay there for all but an arbitrarily small fraction of the time. As the terminal time becomes imminent the trajectory leaves the neighbourhood of the turnpike and heads for the terminal point $x(T) = 0$.

18.5 *Diffusion processes*

How might we introduce noise in a continuous-time plant equation? In the example of §16.1 we might try to write $\dot{x} = u + v\epsilon$, where v is a constant and ϵ is noise. But how should we understand ϵ? A sensible guess (based on what we know about sums of i.i.d. random variables and the Central Limit Theorem) is that $B(t) = \int_0^t \epsilon(s)\,ds$ should be distributed as Gaussian with mean 0 and variance t. The random process, $B(t)$, which fits the bill, is called **Brownian motion**. But much must be made precise.

Just as we previously derived the HJB equation before, we now find

$$\begin{aligned} F(x,t) &= \inf_u \Big[u^2\delta + E[F(x + u\delta + vB(\delta), t + \delta)]\Big] \\ \implies 0 &= \inf_u \Big[u^2 + uF_x + v^2F_{xx} + F_t\Big] \implies 0 = -(1/2)F_x^2 + v^2F_{xx} + F_t, \end{aligned}$$

where $F(x, T) = Dx^2$. The solution to this p.d.e. is (unsurprisingly)

$$F(x,t) = \frac{Dx^2}{1 + (T-t)D} + 2v^2\log(1 + (T-t)D).$$

Index