

Optimization and Control

Contents

Table of Contents	i
Schedules	iv
1 Dynamic Programming	1
1.1 Control as optimization over time	1
1.2 The principle of optimality	1
1.3 Example: the shortest path problem	1
1.4 The optimality equation	2
1.5 Markov decision processes	4
2 Examples of Dynamic Programming	5
2.1 Example: optimization of consumption	5
2.2 Example: exercising a stock option	6
2.3 Example: secretary problem	7
3 Dynamic Programming over the Infinite Horizon	9
3.1 Discounted costs	9
3.2 Example: job scheduling	9
3.3 The infinite-horizon case	10
3.4 The optimality equation in the infinite-horizon case	11
3.5 Example: selling an asset	12
4 Positive Programming	13
4.1 Example: possible lack of an optimal policy.	13
4.2 Characterization of the optimal policy	13
4.3 Example: optimal gambling	14
4.4 Value iteration	14
4.5 Example: pharmaceutical trials	15
5 Negative Programming	17
5.1 Example: a partially observed MDP	17
5.2 Stationary policies	18
5.3 Characterization of the optimal policy	18
5.4 Optimal stopping over a finite horizon	19
5.5 Example: optimal parking	20

6	Optimal Stopping Problems	21
6.1	Bruss's odds algorithm	21
6.2	Example: Stopping a random walk	22
6.3	Optimal stopping over the infinite horizon	22
6.4	Sequential Probability Ratio Test	24
6.5	Bandit processes	24
7	Bandit Processes and the Gittins Index	25
7.1	Multi-armed bandit problem	25
7.2	The two-armed bandit	26
7.3	Gittins index theorem	26
7.4	Playing golf with many balls	27
7.5	Example: Weitzman's problem	28
7.6	*Calculation of the Gittins index*	29
7.7	*Forward induction policies*	29
7.8	*Proof of the Gittins index theorem*	29
8	Average-cost Programming	31
8.1	Average-cost optimality equation	31
8.2	Example: admission control at a queue	32
8.3	Value iteration bounds	32
8.4	Policy improvement algorithm	33
9	Continuous-time Markov Decision Processes	35
9.1	Stochastic scheduling on parallel machines	35
9.2	Controlled Markov jump processes	37
9.3	Example: admission control at a queue	38
10	LQ Regulation	40
10.1	The LQ regulation problem	40
10.2	The Riccati recursion	42
10.3	White noise disturbances	42
10.4	LQ regulation in continuous-time	43
10.5	Linearization of nonlinear models	43
11	Controllability and Observability	44
11.1	Controllability and Observability	44
11.2	Controllability	44
11.3	Controllability in continuous-time	46
11.4	Example: broom balancing	46
11.5	Stabilizability	47
11.6	Example: pendulum	47
11.7	Example: satellite in a plane orbit	48

12 Observability and the LQG Model	49
12.1 Infinite horizon limits	49
12.2 Observability	49
12.3 Observability in continuous-time	51
12.4 Example: satellite in planar orbit	51
12.5 Imperfect state observation with noise	51
13 Kalman Filter and Certainty Equivalence	53
13.1 The Kalman filter	53
13.2 Certainty equivalence	54
13.3 The Hamilton-Jacobi-Bellman equation	55
13.4 Example: LQ regulation	56
13.5 Example: harvesting fish	57
14 Pontryagin’s Maximum Principle	60
14.1 Example: optimization of consumption	60
14.2 Heuristic derivation of Pontryagin’s maximum principle	61
14.3 Example: parking a rocket car	62
14.4 Adjoint variables as Lagrange multipliers	64
15 Using Pontryagin’s Maximum Principle	65
15.1 Transversality conditions	65
15.2 Example: use of transversality conditions	65
15.3 Example: insects as optimizers	66
15.4 Problems in which time appears explicitly	66
15.5 Example: monopolist	67
15.6 Example: neoclassical economic growth	68
16 Controlled Diffusion Processes	70
16.1 The dynamic programming equation	70
16.2 Diffusion processes and controlled diffusion processes	70
16.3 Example: noisy LQ regulation in continuous time	71
16.4 Example: passage to a stopping set	72

Schedules

Dynamic programming

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

LQG systems

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

Continuous-time models

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

Richard Weber, January 2013

1 Dynamic Programming

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

1.1 Control as optimization over time

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

Control theory is concerned with dynamic systems and their **optimization over time**. It accounts for the fact that a dynamic system may evolve stochastically and that key variables may be unknown or imperfectly observed.

The optimization models in the IB course (for linear programming and network flow models) were static and nothing was either random or hidden. In this course it is the additional features of dynamic and stochastic evolution, and imperfect state observation, that give rise to new types of optimization problem and which require new ways of thinking.

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

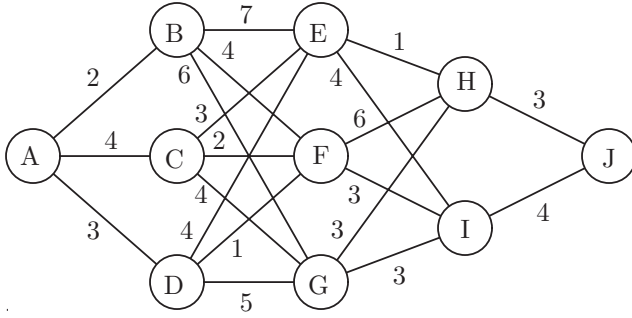
1.2 The principle of optimality

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

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

1.3 Example: the shortest path problem

Consider the 'stagecoach problem' in which a 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 $F(X)$ be the minimal distance required to reach J from X. Then clearly, $F(J) = 0$, $F(H) = 3$ and $F(I) = 4$.

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

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

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

1.4 The optimality equation

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

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

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

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

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

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.

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

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

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

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

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

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 . Then by an inductive proof, one can show as in Theorem 1.2 that

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

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**.

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

- (i) The optimal u_t is a function only of x_t and t , i.e. $u_t = u(x_t, t)$.
- (ii) The DP equation expresses the optimal u_t in closed-loop form. It is optimal whatever the past control policy may have been.
- (iii) The DP equation is a backward recursion in time (from which we get the optimum at $h - 1$, then $h - 2$ and so on.) The later policy is decided first.

‘Life must be lived forward and understood backwards.’ (Kierkegaard)

1.5 Markov decision processes

Consider now stochastic evolution. Let $X_t = (x_0, \dots, x_t)$ and $U_t = (u_0, \dots, u_t)$ denote the x and u histories at time t . As above, state structure is characterised by the fact that the evolution of the process is described by a state variable x , having value x_t at time t . The following assumptions define what is known as a discrete-time **Markov decision process** (MDP).

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

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

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

For the moment we also require the following:

(c) *Perfect state observation*: The current value of the state is observable. That is, x_t is known when choosing u_t . So, letting W_t denote the observed history at time t , we assume $W_t = (X_t, U_{t-1})$.

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

As in the previous section, the cost from time t onwards is given by (1.2). Denote the minimal expected cost from time t onwards by

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

where π denotes a policy, i.e. a rule for choosing the controls u_0, \dots, u_{h-1} .

The following theorem is then obvious.

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

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

with terminal condition

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

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

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

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

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

2 Examples of Dynamic Programming

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

2.1 Example: 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, $0 \leq u_t \leq x_t$. The capital is invested at interest rate $\theta \times 100\%$, and so his income in year $t + 1$ increases to

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

He desires to maximize total consumption over h years,

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

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

Solution. Since dynamic programming makes its calculations backwards, from the termination point, it is often advantageous to write things in terms of the ‘**time to go**’, $s = h - t$. Let $F_s(x)$ denote the maximal reward obtainable, starting in state x when there is time s to go. The dynamic programming equation is

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

where $F_0(x) = 0$, (since 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,

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

Next,

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

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

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

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

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

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

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

where s^* is the least integer such that $1 + s^* \leq (1 + \theta)s^* \iff s^* \geq 1/\theta$, i.e. $s^* = \lceil 1/\theta \rceil$. The optimal strategy is to invest the whole of the income in years $0, \dots, h - s^* - 1$, (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) Although the form of the dynamic programming equation can sometimes look messy, try working backwards from $F_0(x)$ (which is known). Often a pattern will emerge from which you can piece together a solution. (iii) When the dynamics are linear, the optimal control lies at an extreme point of the set of feasible controls. This form of policy, which either consumes nothing or consumes everything, is known as **bang-bang control**.

2.2 Example: exercising a stock option

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

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

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

Solution. The state variable at time t is, strictly speaking, x_t plus a variable which indicates whether the option has been exercised or not. However, it is only the latter case which is of interest, so x is the effective state variable. As above, we use time to go, $s = h - t$. So if we let $F_s(x)$ be the value function (maximal expected profit) with s days to go then

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

and so the dynamic programming equation is

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

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

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

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 than $x - p$ if $x < a_s$ and equals $x - p$ if $x \geq a_s$. It follows from (i) that a_s is non-decreasing in s . The constant a_s is the smallest x for which $F_s(x) = x - p$.

2.3 Example: secretary problem

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

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

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

Now the fact that the t th 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 t th 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

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

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

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

whence

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

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

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

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

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

3 Dynamic Programming over the Infinite Horizon

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

3.1 Discounted costs

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

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

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

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

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

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

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

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

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

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

3.3 The infinite-horizon case

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

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

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

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

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

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

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

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

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

Notice that the names ‘negative’ and ‘positive’ appear to be the wrong way around with respect to the sign of $c(x, u)$. 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 s th 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

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

The following theorem justifies our writing the optimality equation (i.e. (3.7)).

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

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

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

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

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

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

Minimizing the left hand side over π gives ‘ \geq ’.

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

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

Minimizing the right hand side over u_0 and recalling that ϵ is arbitrary gives ‘ \leq ’. \square

3.5 Example: selling an asset

A speculator owns a rare collection of tulip bulbs and each day has an opportunity to sell it, 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

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

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 he 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 than the right hand side at $F(1) = 1$ provided $\beta > 1/2$. In this case the root is greater than 1 and we compute it as

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

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

If $\beta \leq 1/2$ 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 the way that discounting can arise is often quite useful.

4 Positive Programming

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

4.1 Example: possible lack of an optimal policy.

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

Example 4.1. Suppose the possible states are the non-negative integers and in state x we have a choice of either moving to state $x + 1$ and receiving no reward, or moving to state 0, obtaining reward $1 - 1/x$, and then remaining in state 0 thereafter and obtaining no further reward. The optimality equations is

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

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

4.2 Characterization of the optimal policy

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

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

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

Then π is optimal.

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

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

By repeated substitution of this into itself, we find

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

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 $s \rightarrow \infty$; in case D we can let $s \rightarrow \infty$ directly, observing that this term tends to zero. Either way, we have $F(\pi, x) \geq F(\pi', x)$. \square

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 $\{0, 1, \dots, N\}$. 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

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

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

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

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

4.4 Value iteration

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

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

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

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

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

Notice that (4.2) reverses the order of $\lim_{s \rightarrow \infty}$ and \inf_{π} in (3.6). The following theorem states that we can interchange the order of these operations and that therefore $F_s(x) \rightarrow F(x)$. However, in case N we need an additional assumption:

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

Theorem 4.3. *Suppose that D or P holds, or N and F hold. Then $F_\infty(x) = F(x)$.*

Proof. First we prove ‘ \leq ’. Given any $\bar{\pi}$,

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

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

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

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

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

In the negative case,

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

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

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

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

4.5 Example: pharmaceutical trials

A doctor has two drugs available to treat a disease. One is well-established drug and is known to work for a given patient with probability p , independently of its success for other patients. The new drug is untested and has an unknown probability of success θ , which the doctor believes to be uniformly distributed over $[0, 1]$. He treats one patient per day and must choose which drug to use. Suppose he has observed s successes and f failures with the new drug. Let $F(s, f)$ be the maximal expected-discounted number of

future patients who are successfully treated if he chooses between the drugs optimally from this point onwards. For example, if he uses only the established drug, the expected-discounted number of patients successfully treated is $p + \beta p + \beta^2 p + \dots = p/(1 - \beta)$. 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

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

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

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

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

These numbers are the greatest values of p (the known success probability of the well-established drug) for which it is worth continuing with at least one more trial of the new drug. For example, suppose $p = 0.6$ and 6 trials with the new drug have given $s = f = 3$. Then since $p = 0.6 < 0.6133$ we should treat the next patient with the new drug. At this point the probability that the new drug will successfully treat the next patient is 0.5 and so the doctor will actually be treating that patient with the drug that is least likely to cure!

Here we see a tension going on between desires for **exploitation** and **exploration**. A **myopic policy** seeks only to maximize immediate reward. However, an optimal policy takes account of the possibility of gaining information that could lead to greater rewards being obtained later on. Notice that it is worth using the new drug at least once if $p < 0.7614$, even though at its first use the new drug will only be successful with probability 0.5. Of course as the discount factor β tends to 0 the optimal policy will look more and more like the myopic policy.

5 Negative Programming

The special theory of minimizing nonnegative costs. We see that action that extremizes the right hand side of the optimality equation is optimal. Stopping problems and OSLA rule.

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 a **partially observable Markov decision process** (POMDP). The decision-maker cannot directly observe the underlying state, but he must maintain a probability distribution over the set of possible states, based on his observations, and the underlying MDP. This distribution is updated by 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

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

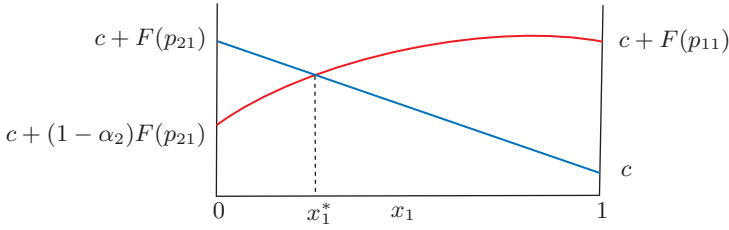
The arguments of $F_{s-1}(\cdot)$ are the posterior probabilities that the object in location 1, given that we have search location 1 (or 2) and not found it.

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 .

Since $F_s \rightarrow F$ 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 $s \rightarrow \infty$ 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 the answer is of this form for all values of the 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

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

or $F_{t+1} = \mathcal{L}(f)F_{t+1}$, 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, $r(x, u) \leq 0$, or minimizing non-negative costs, $c(x, u) \geq 0$. The following theorem gives a necessary and sufficient condition for a stationary policy to be optimal: namely, it must choose the optimal u on the right hand side of the optimality equation. Note that in this theorem we are requiring that the infimum over u is attained as a minimum over u (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*

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

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

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

Proof. The proof is really the same as the final part of proving Theorem 4.3. By substituting the optimality equation into itself and using the fact that π specifies the minimizing control at each stage,

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

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 $s \rightarrow \infty$; in case D we can let $s \rightarrow \infty$ directly, observing that this term tends to zero. Either way, we have $F(x) \geq F(\pi, x)$. \square

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, that is entered upon choosing the stopping action. Once this state is entered the system stays in that state and no further cost is incurred thereafter. We let $c(x, 0) = k(x)$ (stopping cost) and $c(x, 1) = c(x)$ (continuation cost). 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. The dynamic programming equation is

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

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:

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

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 $s \in S$. \square

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 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 first 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 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 and their solution.

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 k th 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 s th event to the last event (the n th)

$$R_s = r_s + \dots + r_n$$

and finds the greatest 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

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

Clearly the stopping set is closed, so the OSLA-rule is optimal. The probability of stopping on the last 1 is $(q_{s^*} \dots q_n)(r_{s^*} + \dots + r_n)$ and (by solving a little optimization problem) 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

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

We can also solve a 'groups' version of the secretary problem. Suppose we see h groups of candidates, of sizes n_1, \dots, n_h . We wish to stop with the group that contains the best of all the candidates. Then $p_1 = 1$, $p_2 = n_2/(n_1 + n_2), \dots, p_h = n_h/(n_1 + \dots + n_h)$. The odds algorithm tells us to stop if group i contains the best candidate so far and $i \geq s^*$, where s^* is the greatest integer s such that

$$\frac{n_s}{\sum_{i=1}^{s-1} n_i} + \frac{n_{s+1}}{\sum_{i=1}^s n_i} + \dots + \frac{n_h}{\sum_{i=1}^{h-1} n_i} \geq 1.$$

6.2 Example: Stopping a random walk

Suppose that x_t follows a random walk on $\{0, \dots, N\}$. At any time t we may stop the walk and take a positive reward $r(x_t)$. In states 0 and N we must stop. The aim is to maximize $Er(x_T)$.

Solution. The dynamic programming equation is

$$\begin{aligned} F(0) &= r(0), \quad F(N) = r(N) \\ F(x) &= \max \left\{ r(x), \frac{1}{2}F(x-1) + \frac{1}{2}F(x+1) \right\}, \quad 0 < x < N. \end{aligned}$$

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

We say F is 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 $F_0 = 0$, we have $G \geq F_0$. So we can prove by induction that

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

Theorem 4.3 tells us that $F_s(x) \rightarrow F(x)$ as $s \rightarrow \infty$. Hence $F \leq G$.

A OSLA rule is not optimal here. The optimal rule is to stop iff $F(x) = r(x)$.

6.3 Optimal stopping over the infinite horizon

Consider now a general stopping problem over the infinite-horizon with $k(x), c(x)$ as previously, and with the aim of minimizing total expected cost. Let $F_s(x)$ be the infimal cost given that we are required to stop by the s th step. Let $F(x)$ be the infimal cost when all that is required is that we stop eventually. Since less cost can be incurred if we are allowed more time in which to stop, we have

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

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

Example 6.1. Consider the problem of stopping a symmetric random walk on the integers, where $c(x) = 0$, $k(x) = \exp(-x)$. The policy of stopping immediately, say π , has $F(\pi, x) = \exp(-x)$, and since e^{-x} is a convex function this satisfies the infinite-horizon optimality equation,

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

However, π is not optimal. The random walk is recurrent, so we may wait until reaching as large an integer as we like before stopping; hence $F(x) = 0$. Thus we see two things:

- (i) It is possible that $F_\infty > F$. This is because $F_s(x) = e^{-x}$, but $F(x) = 0$.
- (ii) Theorem 4.2 is not true for negative programming. Policy π has $F(\pi, x) = e^{-x}$ and this satisfies the optimality equation. Yet π is not optimal.

Remark. In Theorem 4.3 we had $F_\infty = F$, but 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). However, Example 6.1 $k(x) > 0$ 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.*

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

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

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

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

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

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

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

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

Theorem 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. By Theorem 5.3 we have for all finite s ,

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

Lemma 6.2 gives $F(x) = F_\infty(x)$. □

6.4 Sequential Probability Ratio Test

A statistician wishes to decide between two hypotheses, $H_0 : f = f_0$ and $H_1 : f = f_1$ on the basis of i.i.d. observations drawn from a distribution with density f . *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 $x = (x_1, \dots, x_n)$. The posterior probabilities are in the likelihood ratio

$$\ell_n(x) = \frac{f_1(x_1) \cdots f_1(x_n) p_1}{f_0(x_1) \cdots f_0(x_n) p_0}.$$

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

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

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

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

This is a similar problem to the one we solved about searching for a hidden 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 an observation 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, $a_0^* \leq a_1^*$: *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 Bandit processes

A **bandit process** is a special type of MDP 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 $u = 1$ we obtain a reward $r(x_t)$ and the state changes, to x_{t+1} , according to the Markov dynamics $P(x_{t+1} | x_t, u_t = 1)$.

A **simple family of alternative bandit processes** (SFABP) is a collection of n such bandit processes. At each time $t = 0, 1, \dots$ we must select exactly one bandit to receive the continuation action, while all others are frozen.

This provides 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?
- Which of n oil fields should we explore next?

7 Bandit Processes and the Gittins Index

The multi-armed bandit problem. Gittins index theorem.

7.1 Multi-armed bandit problem

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

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

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:

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

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

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

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

7.2 The two-armed bandit

Consider first a problem with one arm B_1 which always pays λ , and a second arm of general type. The optimality equation, when B_2 is in its state x , is

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

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 at some later step, τ . Notice that once we switch to B_1 we will never wish switch back to B_2 because things remain the same as when we first switched away from B_2 .

We are to choose the **stopping time** τ so as to optimally switch from continuing B_2 to continuing B_1 . 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.

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

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

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

$G(x)$ 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.

It can be easily shown that $\tau = \min\{t : G_i(x_i(t)) \leq G_i(x_i(0)), \tau > 0\}$, 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.

7.3 Gittins index theorem

Remarkably, the problem posed by a SFABP (or a MABP) can be solved by using these Gittins indices.

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 had obtained it by 1970, and presented it in 1972. The solution of the MABP impressed many experts as surprising and beautiful. Peter Whittle describes a colleague of high repute, asking another colleague ‘*What would you say if you were told that the multi-armed bandit problem had been solved?*’ The reply was ‘*Sir, the multi-armed bandit problem is not of such a nature that it can be solved*’.

Single machine scheduling. Recall the example in §3.2 in which n jobs are to be processed successively on one machine. Job i has a known processing times t_i , assumed to be a positive integer. On completion of job i a positive reward r_i is obtained. We used an interchange argument to show that the discounted sum of rewards is maximized by processing jobs in decreasing order of the index $r_i\beta^{t_i}/(1 - \beta^{t_i})$.

Now we do this using Gittins index. 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 $1 + \beta + \dots + \beta^{t_i-1} = (1 - \beta^{t_i})/(1 - \beta)$. Thus, $G_i = r_i\beta^{t_i}(1 - \beta)/(1 - \beta^{t_i})$. Note that $G_i \rightarrow r_i/t_i$ as $\beta \rightarrow 1$.

7.4 Playing golf with many balls

There have been many proofs of the index theorem, all of which are useful in developing insight. The following proof is perhaps the simplest (due to Weber (1992)). It is fun to explain it in a special context (invented by Dimitri, Tetali and Winkler, 2003).

A golfer is playing with n balls. They are at positions x_1, \dots, x_n . If he plays ball i it will next land at location y with probability $P(x_i, y)$. He wishes to minimize the expected number of shots taken to get one ball in the hole (location 0).

To represent this as a SFABP we set rewards and costs all 0, except that a reward R is obtained by continuing a bandit that is in state 0. So if some bandit (ball) reaches state 0, say with the golfer’s t th shot, he will continue to play it there, obtaining reward $(\beta^t + \beta^{t+1} + \dots)R$. Suppose the golfer pays a ‘green fee’ of $R/(1 - \beta)$. Then he will be trying to maximize

$$-\frac{R}{1 - \beta} + (\beta^t + \beta^{t+1} + \dots)R = -(1 + \beta + \dots + \beta^{t-1})R$$

which tends to $-tR$ as $\beta \rightarrow 1$, so he’ll want to minimize the expected number of shots taken to sink a ball. We claim this problem can be solved by a (Gittins) index policy.

Proof. Suppose the golfer is playing with just one ball, which is in location x_i . The golfer faces a cost of 1 for each shot he takes until the ball is sunk. So to motivate him to play on, we offer him a prize $g(x_i)$ which he receives if he plays at least one more shot and eventually sinks the ball. He is allowed to quit subsequently if the ball lands in a bad place and the offered prize is no longer sufficiently motivating. But if this ever happens, we will increase the offered prize, so that it again becomes just advantageous

for him to keep playing. This defines an nondecreasing sequence of offered prizes for ball i . Notice that it is defined independently of the other balls.

Now he plays with n balls. For each of them we define an offered prize, as above. It is a function of its location, just as if he were playing only with that ball.

The key idea is that with these offered prizes the golfer can keep playing until a ball is sunk, and he will just break even. He is guaranteed to collect the least possible prize when a ball is finally sunk if he follows the policy of always playing the ball for which the least prize is on offer. But the prizes were invented to make the game is ‘just fair’, and so this policy must also minimize the expected number of shots required to sink a ball in the hole. The prize $g(x_i)$ is of course the Gittins index for location x_i . \square

7.5 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 can be modelled as 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 . We wish to maximize the expected value of

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

where we open boxes i_1, \dots, i_τ and then take the best prize thereafter. In the limit as $\beta \rightarrow 1$ this objective corresponds to that of Weitzman’s problem, namely,

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

and so we can find the solution using the Gittins index theorem.

The Gittins index of an opened box is $r(x_i)$. The index of an unopened box i is the solution to

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

or, by setting $g_i = G/(1 - \beta)$, and letting $\beta \rightarrow 1$, 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 $g_i = -c_i + (1 - p_i)g_i + p_i r_i \implies g_i = r_i - c_i/p_i$.

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.*

7.6 *Calculation of the Gittins index*

How can we compute the Gittins index value for each possible state of a bandit process B_i ? The input is the data of $r_i(\cdot)$ and $P_i(\cdot, \cdot)$. If the state space of B_i is finite, say $E_i = \{1, \dots, k_i\}$, then the Gittins indices can be computed in an iterative fashion. First we find the state of greatest index, say 1 such that $1 = \arg \max_j r_i(j)$. Having found this state we can next find the state of second-greatest index. If this is state j , then $G_i(j)$ is computed in (7.3) by taking τ to be the first time that the state is not 1. This means that the second-best state is the state j which maximizes

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

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

If the state space of a bandit process is infinite, as in the case of the Bernoulli bandit, there may be no finite calculation by which to determine the Gittins indices for all states. In this circumstance, we can approximate the Gittins index using something like the value iteration algorithm. Essentially, one solves a problem of maximizing right hand side of (7.3), subject to $\tau \leq N$, where N is large.

7.7 *Forward induction policies*

If we put $\tau = 1$ on the right hand side of (7.3) then it evaluates to $E r_i(x_i(t))$. If we use this as an index for choosing between projects, we have a **myopic policy** or **one-step-look-ahead policy**. The Gittins index policy generalizes the idea of a one-step-look-ahead policy, since it looks-ahead by some optimal time τ , so as to maximize, on the right hand side of (7.3), a measure of the rate at which reward can be accrued. This defines a so-called **forward induction policy**.

7.8 *Proof of the Gittins index theorem*

The proof in the §7.4 has all the key ideas. For completeness we give a fuller proof of the Gittins index theorem.

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 a gambler would be willing to pay per step in order to be permitted to continue B_i for at least one more step, and being free to stop continuing it whenever he likes thereafter. This is

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

Notice that (7.2) and (7.4) are equivalent definitions 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 $x_i(0), \dots, x_i(t)$ (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 happen 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. □

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:

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

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

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

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

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

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:

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

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.*

Proof. Suppose u is chosen by some policy π' . By repeated substitution of (8.2) into itself we have

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

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

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

with equality if $\pi' = \pi$. □

So an average-cost optimal policy can be found by looking for a bounded solution to (8.2). Notice that if ϕ is a solution of (8.2) then so is $\phi + (\text{a constant})$, 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 $\phi(\bar{x}) = 0$. The function ϕ is called the **relative value function**.

8.2 Example: admission control at a queue

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

Solution. Let 0 and i denote the states in which he is free to accept a job, and in which he is engaged upon a job of type i , respectively. Then (8.2) is

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

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

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

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

8.3 Value iteration bounds

Value iteration in the average-cost case is based upon the idea that $F_s(x) - F_{s-1}(x)$ approximates the minimal average-cost for large s . For the rest of this lecture we suppose the state space is finite.

Theorem 8.2. *Define*

$$m_s = \min_x \{F_s(x) - F_{s-1}(x)\}, \quad M_s = \max_x \{F_s(x) - F_{s-1}(x)\}. \quad (8.3)$$

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

Proof. Suppose $\pi = f^\infty$ is the average-cost optimal policy over the infinite horizon, taking $u = f(x)$, with average-cost λ . Then

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

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

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

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

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

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

8.4 Policy improvement algorithm

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

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

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

$$\lambda + \phi(x) \geq c(x, f_1(x_0)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_1(x_0)], \quad (8.4)$$

for all x , and with strict inequality for some x (and thus $f_1 \neq f$). Then following the lines of proof in Theorems 8.1 and 8.2 (repeatedly substituting (8.4) into itself),

$$\lambda \geq \lim_{t \rightarrow \infty} \frac{1}{t} E_{\pi_1} \left[\sum_{\tau=0}^{t-1} c(x_\tau, u_\tau) \mid x_0 = x \right]. \quad (8.5)$$

So π_1 is at least as good as π . If there is no 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 (8.5) this has an average-cost no more than λ , so π_s is at least as good as π_{s-1} . We now return to step (1) with $s := s + 1$.

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

Example 8.3. 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 finished in expected time $1/\lambda_1 + \dots + 1/\lambda_n$. 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 $\lambda_1 < \dots < \lambda_n$. We would like to prove that for all $i, j \in S \subseteq \{1, \dots, n\}$,

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

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

Stochastic sequential assignment problem with arrivals

Let π be LLR. 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 $i \geq 3$,

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

If for some $3 \leq i < j$ we were to have $\delta_i(\pi, S) > \delta_j(\pi, S)$ then this would require that either $\delta_i(S^1) > \delta_j(S^1)$ or $\delta_i(S^2) > \delta_j(S^2)$. But our induction hypothesis for (9.1) rules these out.

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

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

By comparing (9.2) and (9.3) we see that we could only have $\delta_i(S) < \delta_1(S)$ for $i \geq 3$ if at least one of $\delta_i(S^1) < \delta_3(S^1)$ or $\delta_i(S^2) < \delta_1(S^2)$ is true. These are ruled out by our induction hypothesis. Similarly, we cannot have $\delta_i(S) < \delta_2(S)$ for $i \geq 3$.

This completes a step of 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

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

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 wears two stockings each day, 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 $0, 1, \dots$. There is a constant service rate μ (meaning that the service times of customers are distributed $\mathcal{E}(\mu)$, and we may control the arrival rate u between 0 and a maximum value M . Let $c(x, u) = ax - Ru$. There is a **holding cost** a per unit time for each customer in the queue and receiving a reward R at the point that each new customer is admitted (and therefore incurring reward at rate Ru when the arrival rate is u). Suppose there is discounting at rate α . The problem is one of choosing $0 \leq u_t \leq M$ to minimize

$$E \left[\int_0^\infty (ax_t - Ru_t) e^{-\alpha t} dt \right].$$

Let us take $B = \alpha + M + \mu$, and without loss of generality assume $B = 1$.

After uniformization the discounted-cost optimality equation will look like

$$F(0) = \inf_{u \in [0, M]} \{-Ru + \beta[uF(1) + (M - u)F(j - 1)]\}$$

$$F(x) = \inf_{u \in [0, M]} \{ax - Ru + \beta[uF(x + 1) + \mu F(x - 1) + (M - u)F(x)]\}, \quad x \geq 1.$$

So we can see that the optimal control is bang-bang, taking $u = 0$ or $u = M$ as the coefficient of u , namely $-R + F(x + 1) - F(x)$, is positive or negative. One can set up a value iteration form of this, i.e.

$$F_{k+1}(0) = \inf_{u \in [0, M]} \{-Ru + \beta[uF_k(1) + (M - u)F_k(j - 1)]\}$$

$$F_{k+1}(x) = \inf_{u \in [0, M]} \{ax - Ru + \beta[uF_k(x + 1) + \mu F_k(x - 1) + (M - u)F_k(x)]\}, \quad x \geq 1$$

and then prove by induction that $F_k(x)$ is concave in x . This means that there exists a **threshold rule** such that the optimal policy will be of the form:

$$u = \begin{cases} 0 & \text{as } x \geq x^* \\ M & \text{as } x < x^* \end{cases}.$$

Time-average cost optimality. The optimality equation is

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

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

Thus u should be chosen to be 0 or M as $-R + \phi(x+1) - \phi(x)$ is positive or negative.

Let us consider what happens under the policy that take $u = M$ for all x . The relative costs for this policy, say f , are given by

$$f(x) + \gamma = ax - R\lambda + Mf(x+1) + \mu f(x-1), \quad x \geq 1.$$

The solution to the homogeneous part of this recursion is of the form $f(x) = d_1 1^x + d_2 (\mu/M)^x$. Assuming $M < \mu$ and we desire a solution for f that does not grow exponentially, we take $d_2 = 0$ and so the solution is effectively the solution to the inhomogeneous part, i.e.

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

Applying the idea of policy improvement, we conclude that a better policy is to take $u = 0$ (i.e. don't admit a customer) if $-R + f(x+1) - f(x) > 0$, i.e. if

$$\frac{(x+1)a}{\mu - M} - R > 0.$$

Further policy improvement is needed to reach the optimal policy. However, this policy already exhibits an interesting property: it rejects customers for smaller queue length x than does a policy which rejects a customer if and only if

$$\frac{(x+1)a}{\mu} - R > 0.$$

This second policy is optimal if one is purely concerned with whether or not an individual customer that joins when there are x customers in front of him will show a profit on the basis of the difference between the reward R and his expected holding cost $(x+1)a/\mu$. This example exhibits the difference between **individual optimality** (which is myopic) and **social optimality**. The socially optimal policy is more reluctant to admit customers because it anticipates that more customers are on the way; thus it feels less badly about forgoing the profit on a customer that presents himself now, recognizing that admitting such a customer can cause customers who are admitted after him to suffer greater delay.

10 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.

10.1 The LQ regulation problem

As we have seen, the elements of a control optimization problem are specification of (i) the dynamics of the process, (ii) which quantities are observable at a given time, and (iii) an optimization criterion.

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

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

$$x_t = Ax_{t-1} + Bu_{t-1}, \quad (10.1)$$

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

$$\mathbf{C} = \sum_{t=0}^{h-1} c(x_t, u_t) + \mathbf{C}_h(x_h), \quad (10.2)$$

with one-step and terminal costs

$$c(x, u) = x^\top Rx + u^\top Su + x^\top S^\top u + u^\top Qu = \begin{pmatrix} x \\ u \end{pmatrix}^\top \begin{pmatrix} R & S^\top \\ S & Q \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}, \quad (10.3)$$

$$\mathbf{C}_h(x) = x^\top \Pi_h x. \quad (10.4)$$

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

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

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

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

and is equal to

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

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

$$\begin{aligned} & \begin{pmatrix} x \\ u+h \end{pmatrix}^\top \begin{pmatrix} \Pi_{xx} & \Pi_{xu} \\ \Pi_{ux} & \Pi_{uu} \end{pmatrix} \begin{pmatrix} x \\ u+h \end{pmatrix} \\ &= x^\top \Pi_{xx} x + 2x^\top \Pi_{xu} u + \underbrace{2h^\top \Pi_{ux} x + 2h^\top \Pi_{uu} u}_{\text{linear term}} + u^\top \Pi_{uu} u + h^\top \Pi_{uu} h. \end{aligned}$$

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

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

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

Theorem 10.2. *Assume the structure of (10.1)–(10.4). Then the value function has the quadratic form*

$$F(x, t) = x^\top \Pi_t x, \quad t \leq h, \quad (10.5)$$

and the optimal control has the linear form

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

The time-dependent matrix Π_t satisfies the Riccati equation

$$\Pi_t = f \Pi_{t+1}, \quad t < h, \quad (10.6)$$

where Π_h has the value given in (10.4), and f is an operator having the action

$$f\Pi = R + A^\top \Pi A - (S^\top + A^\top \Pi B)(Q + B^\top \Pi B)^{-1}(S + B^\top \Pi A). \quad (10.7)$$

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

$$K_t = -(Q + B^\top \Pi_{t+1} B)^{-1}(S + B^\top \Pi_{t+1} A), \quad t < h. \quad (10.8)$$

Proof. Assertion (10.5) is true at time h . Assume it is true at time $t+1$. Then

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

Lemma 10.1 shows the minimizer is $u = K_t x$, and gives the form of f . \square

10.2 The Riccati recursion

The backward recursion (10.6)–(10.7) is called the **Riccati equation**.

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

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

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

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

(iii) S can be normalized to zero by choosing a new control $u^* = u + Q^{-1}Sx$, and setting $A^* = A - BQ^{-1}S$, $R^* = R - S^\top Q^{-1}S$. So $A^*x + Bu^* = Ax + Bu$ and $c(x, u) = x^\top Rx + u^{*\top}Qu^*$.

(iv) Similar results are true if $x_{t+1} = A_t x_t + B_t u_t + \alpha_t$, where $\{\alpha_t\}$ is a known sequence of disturbances, and the aim is to track a sequence of values (\bar{x}_t, \bar{u}_t) , $t = 0, \dots, h-1$, so the cost is

$$c(x, u, t) = \begin{pmatrix} x - \bar{x}_t \\ u - \bar{u}_t \end{pmatrix}^\top \begin{pmatrix} R_t & S_t^\top \\ S_t & Q_t \end{pmatrix} \begin{pmatrix} x - \bar{x}_t \\ u - \bar{u}_t \end{pmatrix}.$$

10.3 White noise disturbances

Suppose the plant equation (10.1) 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 $E\epsilon = 0$, $E\epsilon_t \epsilon_s^\top = N$ and $E\epsilon_t \epsilon_s^\top = 0$, $t \neq s$. The dynamic programming equation is then

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

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

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

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

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

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

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

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

10.4 LQ regulation in continuous-time

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

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

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

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

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

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

This is slightly simpler than the discrete time version. The optimal control is $u(t) = K(t)x(t)$, where $K(t) = -Q^{-1}(S + B^\top \Pi)$.

The optimally controlled plant equation is $\dot{x} = \Gamma(t)x$, where

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

10.5 Linearization of nonlinear models

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

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

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

$$\dot{x}' = \dot{x} = a(\bar{x} + x', \bar{u} + u') = Ax' + Bu$$

where

$$A_{ij} = \left. \frac{\partial a_i}{\partial x_j} \right|_{(\bar{x}, \bar{u})}, \quad B_{ij} = \left. \frac{\partial a_i}{\partial u_j} \right|_{(\bar{x}, \bar{u})}.$$

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.

11 Controllability and Observability

Controllability in discrete and continuous time. Stabilizability.

11.1 Controllability and Observability

The discrete-time system $[A, B, \cdot]$ is defined by the plant equation

$$x_t = Ax_{t-1} + Bu_{t-1}, \quad (11.1)$$

The **controllability** question is: can we bring x to an arbitrary prescribed value by some u -sequence?

The discrete-time system $[A, B, C]$ is defined by (11.1) and observation relation

$$y_t = Cx_{t-1}. \quad (11.2)$$

$y_t \in \mathbb{R}^p$ is observed, but x_t is not. C is $p \times n$. The **observability** question is: can we infer x_0 from subsequent y values?

Definition 11.1. The $[A, B, \cdot]$ system is **r-controllable** if one can bring it from an arbitrary prescribed x_0 to an arbitrary prescribed x_r by some u -sequence u_0, u_1, \dots, u_{r-1} . A system of dimension n is **controllable** if it is r -controllable for some r

Definition 11.2. The $[A, B, C]$ system is said to be **r-observable** if x_0 can be inferred from knowledge of the observations y_1, \dots, y_r and relevant control values u_0, \dots, u_{r-2} for any initial x_0 . An n -dimensional system is **observable** if r -observable for some r .

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

11.2 Controllability

Example 11.3. Consider the case, $(n = 2, m = 1)$,

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

This system is not 1-controllable. But

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

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

In general, by substituting the plant equation (11.1) into itself, we see that we must find u_0, u_1, \dots, u_{r-1} to satisfy

$$\Delta = x_r - A^r x_0 = Bu_{r-1} + ABu_{r-2} + \dots + A^{r-1}Bu_0, \quad (11.3)$$

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

Theorem 11.4. (The Cayley-Hamilton theorem) Any $n \times n$ matrix A satisfies its own characteristic equation. So that if

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

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

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

Theorem 11.5. (i) The system $[A, B, \cdot]$ is r -controllable iff the matrix

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

has rank n , (ii) equivalently, iff the $n \times n$ matrix

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

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

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

Proof. (i) The system (11.3) has a solution for arbitrary Δ iff M_r has rank n .

(ii) That is, iff there does not exist nonzero w such that $w^\top M_r = 0$. Now

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

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

(iv) Consider the Lagrangian

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

giving

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

Now we can determine λ from (11.3). □

11.3 Controllability in continuous-time

Theorem 11.6. (i) The n dimensional system $[A, B, \cdot]$ is controllable iff the matrix M_n has rank n , or (ii) equivalently, iff

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

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

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

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

11.4 Example: broom balancing

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

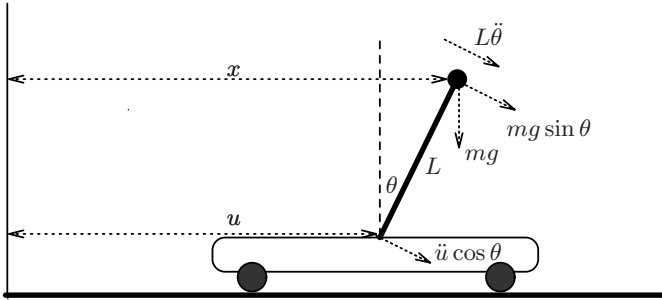


Figure 1: Force diagram for broom balancing

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

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

equivalently,

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

Since

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

the system is controllable if θ is initially small.

11.5 Stabilizability

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

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

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

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

Definition 11.7.

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

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

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

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

11.6 Example: pendulum

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

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

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

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

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

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

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

11.7 Example: satellite in a plane orbit

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

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

where u_r and u_θ are the radial and tangential components of thrust. If $u_r = u_\theta = 0$ then there is a possible equilibrium in which the orbit is a circle of radius $r = \rho$, $\dot{\theta} = \omega = \sqrt{c/\rho^3}$ and $\dot{r} = \dot{\theta} = 0$.

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

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

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

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

It is easy to check that $M_2 = [B \ AB]$ has rank 4 and that therefore the system is controllable.

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

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

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

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

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

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, $2\omega\rho\delta r + \rho^2\delta\dot{\theta} = \delta(r^2\dot{\theta})|_{r=\rho, \dot{\theta}=\omega}$.

12 Observability and the LQG Model

LQ regulation problem over the infinite horizon. More on observability. Least squares estimation and the LQG model.

12.1 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 12.1. *Suppose $\Pi_0 = 0$, $R \succeq 0$, $Q \succeq 0$ and $[A, B, \cdot]$ is controllable or stabilizable. Then $\{\Pi_s\}$ has a finite limit Π .*

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

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

If the system is stabilizable then there is a K such that $\Gamma = A + BK$ is a stability matrix. Using $u_t = Kx_t$, we have $x_t = \Gamma^t x$ and $u_t = K\Gamma^t x$, so

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

Hence in either case we have an upper bound and so $x^\top \Pi_s x$ tends to a limit for every x . By considering $x = e_j$, the vector with a unit in the j th place and zeros elsewhere, we conclude that the j th 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. \square

Both value iteration and policy improvement are effective ways to compute the solution to an infinite-horizon LQ regulation problem. Policy improvement goes along the lines developed in Lecture 8.

12.2 Observability

From (11.1) and (11.2) we can determine y_t in terms of x_0 and subsequent controls:

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

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

$$\tilde{y}_t = CA^{t-1}x_0, \quad 1 \leq t \leq r. \quad (12.1)$$

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

Theorem 12.2. (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

$$x_0 = (N_r^\top N_r)^{-1} \sum_{j=1}^r (A^\top)^{j-1} C^\top \tilde{y}_j. \quad (12.2)$$

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

If we define the deviation $\eta_t = \tilde{y}_t - CA^{t-1}x_0$ then the equation amounts to $\eta_t = 0$, $1 \leq t \leq r$. If these equations were not consistent we could still define a ‘least-squares’ solution to them by minimizing any positive-definite quadratic form in these deviations with respect to x_0 . In particular, we could minimize $\sum_{t=0}^{r-1} \eta_t^\top \eta_t$. This minimization gives (12.2). If equations (12.1) indeed have a solution (i.e. are mutually consistent, as we suppose) and this is unique then expression (12.2) 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). \square

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. This approach lies close to the statistical approach necessary when observations are corrupted by noise.

12.3 Observability in continuous-time

Theorem 12.3. (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

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

12.4 Example: satellite in planar orbit

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

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

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

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

12.5 Imperfect state observation with noise

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

portant point. The model is

$$x_t = Ax_{t-1} + Bu_{t-1} + \epsilon_t, \quad (12.3)$$

$$y_t = Cx_{t-1} + \eta_t, \quad (12.4)$$

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

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

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

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

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

$$E(x | y) = V_{xy}V_{yy}^{-1}y, \quad (12.5)$$

and

$$\text{cov}(x | y) = V_{xx} - V_{xy}V_{yy}^{-1}V_{yx}. \quad (12.6)$$

Proof. Both y and $x - V_{xy}V_{yy}^{-1}y$ are linear functions of x and y and therefore they are Gaussian. From $E[(x - V_{xy}V_{yy}^{-1}y)y^\top] = 0$ 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 (12.6) \square

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.

13 Kalman Filter and Certainty Equivalence

The Kalman filter. Certainty equivalence. The HJB equation for dynamic programming in continuous time.

13.1 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} .

$$\begin{aligned} x_t &= A^t x_0 + A^{t-1} B u_0 + \dots + B u_{t-1} + A^{t-1} \epsilon_0 + \dots + A \epsilon_{t-1} + \epsilon_t \\ y_t &= C \left(A^{t-1} x_0 + A^{t-2} B u_0 + \dots + B u_{t-2} + A^{t-2} \epsilon_0 + \dots + A \epsilon_{t-2} + \epsilon_{t-1} \right) + \eta_t \end{aligned}$$

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 13.1. (The Kalman filter) *Suppose that conditional on W_0 , the initial state x_0 is distributed $N(\hat{x}_0, V_0)$ and the state and observations obey the recursions of the LQG model (12.3)–(12.4). Then conditional on W_t , the current state is distributed $N(\hat{x}_t, V_t)$. The conditional mean and variance obey the updating recursions*

$$\hat{x}_t = A\hat{x}_{t-1} + B u_{t-1} + H_t(y_t - C\hat{x}_{t-1}), \quad (13.1)$$

where the time-dependent matrix V_t satisfies a Riccati equation

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

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

$$gV = N + A V A^\top - (L + A V C^\top)(M + C V C^\top)^{-1}(L^\top + C V A^\top). \quad (13.2)$$

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

$$H_t = (L + A V_{t-1} C^\top)(M + C V_{t-1} C^\top)^{-1}. \quad (13.3)$$

Compare this to the very similar statement of Theorem 10.2. Notice that (13.2) computes V_t forward in time ($V_t = g V_{t-1}$), whereas (10.7) computes Π_t backward in time ($\Pi_t = f \Pi_{t+1}$).

Proof. The proof is by induction on t . Consider the moment when u_{t-1} has been 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

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

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

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

Thus it follows from Lemma 12.4 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 (13.1)–(13.3). \square

13.2 Certainty equivalence

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

Theorem 13.2. *Suppose LQG model assumptions hold. Then (i) the value function is of the form*

$$F(W_t) = \hat{x}_t^\top \Pi_t \hat{x}_t + \dots \quad (13.4)$$

where \hat{x}_t is the linear least squares estimate of x_t whose evolution is determined by the Kalman filter in Theorem 13.1 and ‘ $+\dots$ ’ 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 10.2.

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

Proof. The proof is by backward induction. Suppose (13.4) holds at t . Recall that

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

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

$$\begin{aligned}
F(W_{t-1}) &= \min_{u_{t-1}} E [c(x_{t-1}, u_{t-1}) + \hat{x}_t^\top \Pi_t \hat{x}_t + \cdots \mid W_{t-1}, u_{t-1}] \\
&= \min_{u_{t-1}} E \left[c(\hat{x}_{t-1} - \Delta_{t-1}, u_{t-1}) \right. \\
&\quad + (A\hat{x}_{t-1} + Bu_{t-1} + H_t \zeta_t)^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1} + H_t \zeta_t) \\
&\quad \left. + \cdots \mid W_{t-1}, u_{t-1} \right] \\
&= \min_{u_{t-1}} [c(\hat{x}_{t-1}, u_{t-1}) + (A\hat{x}_{t-1} + Bu_{t-1})^\top \Pi_t (A\hat{x}_{t-1} + Bu_{t-1})] + \cdots,
\end{aligned} \tag{13.5}$$

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 evaluate (13.5) 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

13.3 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,

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

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

$$\inf_u \left[c(x, u, t) - \alpha F + \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} a(x, u, t) \right] = 0, \quad t < h, \tag{13.6}$$

with $F(x, h) = \mathbf{C}(x, h)$. In the undiscounted case, we simply put $\alpha = 0$. Notice that in (14.8) we have $\alpha = 0$ and the term of $\frac{\partial F}{\partial t}$ disappears because $h = \infty$.

Equation (13.6) 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 13.3. Suppose a policy π , using a control u , has a value function F which satisfies the HJB equation (13.6) for all values of x and t . Then π is optimal.

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

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

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

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

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

13.4 Example: LQ regulation

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

$$0 = \inf_u [x^\top R x + u^\top Q u + F_t + F_x^\top (A x + B u)].$$

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

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

where we use the fact that $2x^\top \Pi A x = x^\top \Pi A x + x^\top A^\top \Pi x$. This must hold for all x . So we have a solution to the HJB equation if $\Pi(t)$ satisfies the Riccati differential equation

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

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

13.5 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 $0 \leq u \leq u_{\max}$, where u_{\max} is the greatest possible fishing rate.

Solution. The DP equation (with discounting) is

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

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

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

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

$$\sup_u \left[u - \alpha F + \frac{\partial F}{\partial x} [a(x) - u] \right] = 0. \quad (13.7)$$

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

$$u = \begin{bmatrix} 0 \\ \text{undetermined, but effectively } a(\bar{x}) \\ u_{\max} \end{bmatrix} \text{ for } x \begin{bmatrix} < \\ = \\ > \end{bmatrix} \bar{x}. \quad (13.8)$$

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

$$F(\bar{x}) = \int_0^\infty e^{-\alpha t} a(\bar{x}) dt = a(\bar{x}) / \alpha. \quad (13.9)$$

So from (13.7) and (13.9) we have

$$F_x(x) = \frac{\alpha F(x) - u(x)}{a(x) - u(x)} \rightarrow 1 \text{ as } x \nearrow \bar{x} \text{ or } x \searrow \bar{x}. \quad (13.10)$$

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 (13.10) 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 (13.8) has a value function F which satisfies the HJB equation.

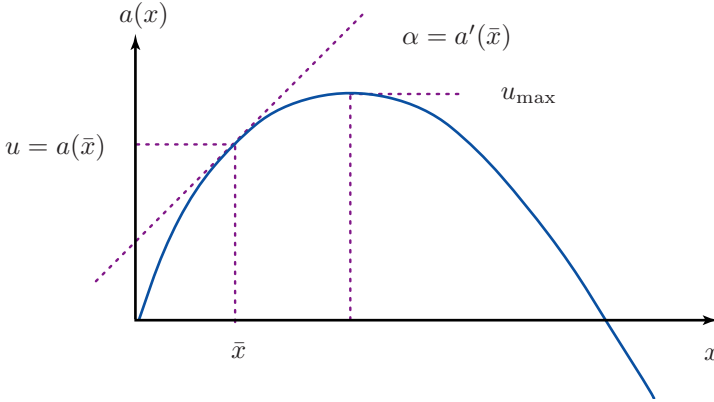


Figure 2: Growth rate $a(x)$ subject to environment pressures

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

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

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

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

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

14 Pontryagin's Maximum Principle

Pontryagin's maximum principle. Optimization of consumption. Parking a rocket car. Adjoint variables as Lagrange multipliers.

14.1 Example: optimization of consumption

Suppose that given $x(0)$, κ and T , all positive, we wish to choose $u(t)$ to maximize

$$\int_0^T \log u(t) dt + \kappa \log x(T), \quad \text{subject to } \dot{x}(t) = ax(t) - u(t), \quad 0 \leq t \leq T.$$

Solution. Try using a Lagrange multiplier $\lambda(t)$ for the constraint $\dot{x}(t) = ax(t) - u(t)$. The Lagrangian is

$$L = \kappa \log x(T) + \int_0^T [\log u - \lambda(\dot{x} - (ax - u))] dt$$

Now use integration by parts, and define $H(x, u, \lambda) = \log u + \lambda(ax - u)$.

$$\begin{aligned} L &= \kappa \log x(T) - \lambda(t)x(t) \Big|_0^T + \int_0^T [\log u + \dot{\lambda}x + \lambda(ax - u)] dt \\ &= [\kappa \log x(T) - \lambda(T)x(T)] + \lambda(0)x(0) + \int_0^T [\dot{\lambda}x + H(x, u, \lambda)] dt. \end{aligned}$$

To make L stationary with respect to both $x(t)$ and $u(t)$, at every point within the integrand, we need

$$\begin{aligned} \dot{\lambda} + \frac{\partial}{\partial x} H(x, u, \lambda) &= 0 \implies \dot{\lambda} = -a\lambda \\ \frac{\partial}{\partial u} H(x, u, \lambda) &= 0 \implies u = 1/\lambda, \end{aligned}$$

and so $\lambda(t) = \lambda(0)e^{-at}$, $u(t) = \lambda(0)^{-1}e^{at}$ and $\dot{x}(t) = ax(t) - \lambda(0)^{-1}e^{at}$.

If the value of $x(T)$ is prescribed (and $< e^{aT}x(0)$ so u need not be negative), then we can solve this differential equation for x , choosing $\lambda(0)$ so that $x(t)$ takes prescribed values $x(0)$ and $x(T)$ at $t = 0$ and T . We get (after some algebra)

$$u(t) = \left(\frac{x(0) - x(T)e^{-aT}}{(T-t)x(0) - x(T)e^{-aT}} \right) x(t).$$

If the value of $x(T)$ is free, then stationarity of L with respect to $x(T)$ requires $\kappa/x(T) - \lambda(T) = 0$ which (after some algebra) implies $\lambda(0) = (\kappa + T)/x(0)$ and

$$u(t) = \frac{1}{T + \kappa} x(0) e^{at} = \frac{1}{(T-t) + \kappa} x(t).$$

If $a > 1/(\kappa + T)$ the trajectory is one in which $x(t)$ is initially increasing and then decreasing; otherwise $x(t)$ is decreasing. The optimal 'inheritance' left at T is

$$x(T) = \frac{\kappa}{\kappa + T} x(0) e^{aT}.$$

14.2 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, $x(0)$. In comparison, the dynamic programming approach is a calculation for a general initial value of the state. Thus, when PMP is useful, it finds an open-loop prescription of the optimal control, whereas dynamic programming is useful for finding a closed-loop prescription. PMP can be used as both a computational and analytic technique (and in the second case can solve the problem for general initial value.)

We begin by considering a time-homogeneous formulation, with plant equation $\dot{x} = a(x, u)$ and instantaneous cost $c(x, u)$. The trajectory is to be controlled until it reaches some stopping set S , where there is a terminal cost $K(x)$. As in (13.6) the value function $F(x)$ obeys the a dynamic programming equation (without discounting)

$$\inf_{u \in \mathcal{U}} \left[c(x, u) + \frac{\partial F}{\partial x} a(x, u) \right] = 0, \quad x \notin S, \quad (14.1)$$

with terminal condition

$$F(x) = K(x), \quad x \in S. \quad (14.2)$$

Define the **adjoint variable**

$$\lambda = -F_x. \quad (14.3)$$

This is a column n -vector, and is to be regarded as a function of time as the state moves along the optimal trajectory. The proof that F_x exists in the required sense is actually a tricky technical matter. We also define the **Hamiltonian**

$$H(x, u, \lambda) = \lambda^\top a(x, u) - c(x, u), \quad (14.4)$$

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

Theorem 14.1. (PMP) *Suppose $u(t)$ and $x(t)$ represent the optimal control and state trajectory. Then there exists an adjoint trajectory $\lambda(t)$ such that together $u(t)$, $x(t)$ and $\lambda(t)$ satisfy*

$$\dot{x} = H_\lambda, \quad [= a(x, u)] \quad (14.5)$$

$$\dot{\lambda} = -H_x, \quad [= -\lambda^\top a_x + c_x] \quad (14.6)$$

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

$$H(x(t), v, \lambda(t)) \leq H(x(t), u(t), \lambda(t)), \quad (14.7)$$

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

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

Assertion (14.5) is immediate, and (14.7) follows from the fact that the minimizing value of u in (14.1) is optimal. Assuming u is the optimal control we have from (14.1) 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

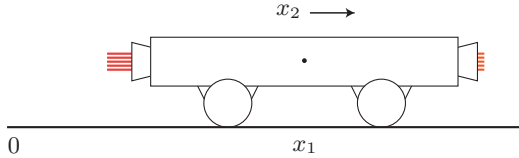
$$\begin{aligned} \frac{d}{dx_i} F(x, t) &= \delta \frac{d}{dx_i} c(x, u) + \sum_j \frac{\partial}{\partial x_j} F(x + a(x, u)\delta, t + \delta) \frac{d}{dx_i} (x_j + a_j(x, u)\delta) \\ \implies -\lambda_i(t) &= \delta \frac{dc}{dx_i} - \lambda_i(t + \delta) - \delta \sum_j \lambda_j(t + \delta) \frac{da_j}{dx_i}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

which is (14.6). □

Notice that (14.5) and (14.6) each give n equations. Condition (14.7) 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 $\lambda(t)$.

14.3 Example: parking a rocket car

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



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

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

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

$$0 = \min_{-1 \leq u \leq 1} \left\{ 1 + \frac{\partial F}{\partial x_1} x_2 + \frac{\partial F}{\partial x_2} u \right\} \quad (14.8)$$

with boundary condition $F(0,0) = 0$. We can see that the optimal control will be a **bang-bang control** with $u = -\text{sign}(\frac{\partial F}{\partial x_2})$ 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 $u = \text{sign}(\lambda_2)$. The adjoint variables satisfy $\dot{\lambda}_i = -\partial H / \partial x_i$, so

$$\dot{\lambda}_1 = 0, \quad \dot{\lambda}_2 = -\lambda_1. \quad (14.9)$$

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

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

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

From (14.1) or (14.8) 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

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

In this case the optimal trajectory lies on the parabola $x_1 = x_2^2/2$, $x_1 \geq 0$, $x_2 \leq 0$. This is half of the **switching locus** $x_1 = \pm x_2^2/2$ (shown dotted in Figure 3).

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

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

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

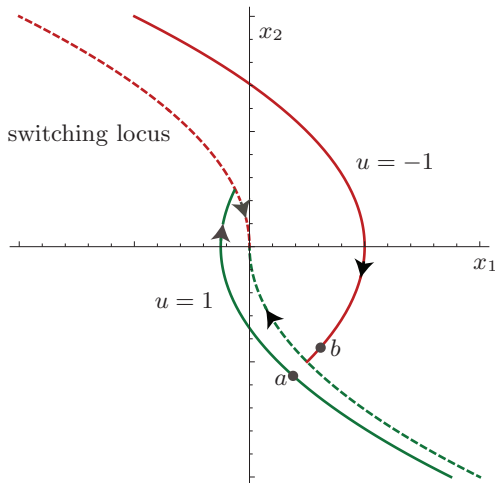


Figure 3: Optimal trajectories for parking a rocket car. Notice that the trajectories starting from two nearby points, a and b , are qualitatively different. From a we take first $u = 1$ then $u = -1$. From b we first take $u = -1$, then $u = 1$.

14.4 Adjoint variables as Lagrange multipliers

We have already seen in §14.1 that it is possible to think of $\lambda(t)$ as a Lagrange multiplier for the constraint $\dot{x} = a(x, u)$ (at time t). Consider the Lagrangian

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

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

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

We now think about varying both $x(t)$ and $u(t)$, but without regard to the constraint $\dot{x} = a(x, u)$. The quantity within the integral must be stationary with respect to $x = x(t)$ and hence $\dot{\lambda} + \lambda^\top a_x - c_x = 0 \implies \dot{\lambda} = -H_x$, i.e. (14.6).

If $x(T)$ is unconstrained then the Lagrangian must also be stationary with respect to small variations in $x(T)$ that are in a direction σ such that $x(T) + \epsilon\sigma$ is in the stopping set (or within $o(\epsilon)$ of it), and this gives $(K_x(x(T)) + \lambda(T))^\top \sigma = 0$, i.e. the so-called **transversality conditions**, which we will say more about in (15.1).

It is good to have this alternative viewpoint, but it is informal and less easy to rigourise than the ‘proofs’ of in §14.2, and §15.1

15 Using Pontryagin's Maximum Principle

Transversality conditions. Examples with Pontryagin's maximum principle.

15.1 Transversality conditions

In (14.1) we see that H must be maximized to 0. We can make this a generally valid assertion, and also say some things about the terminal value of $\lambda(T)$ (the so-called **transversality conditions**.)

Theorem 15.1. (i) $H = 0$ on the optimal path. (ii) The terminal condition

$$(\lambda + K_x)^\top \sigma = 0 \quad (15.1)$$

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

'Proof.' Assertion (i) follows from (14.1). To see (ii), suppose that x is a point at which the optimal trajectory first enters S . Then $x \in S$ and so $F(x) = K(x)$. Suppose $x + \epsilon\sigma + o(\epsilon) \in S$. Then

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

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

15.2 Example: use of transversality conditions

Suppose $\dot{x}_1 = x_2$, $\dot{x}_2 = u$, $x(0) = (0, 0)$, u is unconstrained, and we wish to minimize

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

Here $K(x) = -x_1(1)$. The Hamiltonian is

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

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

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

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

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

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

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

15.3 Example: insects as optimizers

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

$$H = \lambda_1(auw - bw) + \lambda_2c(1 - u)w$$

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

$$\begin{aligned} -\dot{\lambda}_1 &= H_w = \lambda_1(au - b) + \lambda_2c(1 - u) & \lambda_1(T) &= -K_w = 0 \\ -\dot{\lambda}_2 &= H_q = 0 & \lambda_2(T) &= -K_q = 1 \end{aligned} ,$$

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

$$u = \begin{cases} 0 & \text{if } \Delta(t) := \lambda_1 a - c < 0 \\ 1 & \text{if } \Delta(t) := \lambda_1 a - c \geq 0 \end{cases}$$

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

$$\dot{\lambda}_1 = \lambda_1 b - c. \tag{15.2}$$

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 $\Delta(t) = \lambda_1 a - c \geq 0$. The optimal control becomes $u = 1$ and then $\dot{\lambda}_1 = -\lambda_1(a - b) < 0$, which implies that $\lambda_1(s)$ continues to increase as s increases, right back to the start. So there is no further switch in u .

The point at which the single switch occurs is found by integrating (15.2) from t to T , to give $\lambda_1(t) = (c/b)(1 - e^{-(T-t)b})$ and so the switch occurs where $\lambda_1 a - c = 0$, i.e. $(a/b)(1 - e^{-(T-t)b}) = 1$, or

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

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

15.4 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 15.1 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 14.1 still holds. However, to (14.6) we can now add

$$\dot{\lambda}_0 = -H_t = c_t - \lambda a_t, \quad (15.3)$$

and transversality condition

$$(\lambda + K_x)^\top \sigma + (\lambda_0 + K_t) \tau = 0, \quad (15.4)$$

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

We can now understand what to do with various types of time-dependancy and terminal conditions on $x(T)$ and/or T . For example, we can draw the following inferences.

- (i) If K is time-independent (so $K_t = 0$) and the terminal time T is unconstrained (so τ is arbitrary) then the transversality condition implies $\lambda_0(T) = 0$. Since H is always maximized to $-\lambda_0(t)$ it must be maximized to 0 at T .
- (ii) If a, c are only functions of (x, u) then $\dot{\lambda}_0 = c_t - \lambda^\top a_t = 0$, and so $\lambda_0(t)$ is constant on the optimal trajectory. Since H is always maximized to $-\lambda_0(t)$ it must be maximized to a constant on the optimal trajectory.
- (iii) If both (i) and (ii) are true then H is maximized to 0 along the entire optimal trajectory. We had this in the problem of parking in minimal time, §14.3.

15.5 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 $0 \leq u \leq 2$ and $p(u) = 0$ for $u \geq 2$. 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

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

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

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

where T is then the time at which all wine has been sold, and so

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

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

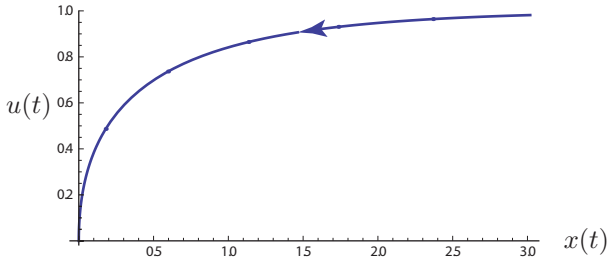


Figure 4: Trajectories of $x(t), u(t)$, for $\alpha = 1$.

The optimal value function is

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

15.6 Example: neoclassical economic growth

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

$$\dot{x} = f(x) - \gamma x - u, \tag{15.5}$$

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 §13.5, with $a(x) = f(x) - \gamma x$. So let us take

$$\dot{x} = a(x) - u. \quad (15.6)$$

It is convenient to take

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

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

$$\frac{d}{dt} (e^{-\alpha t} \lambda) = -H_x = -e^{-\alpha t} \lambda a'(x) \quad (15.7)$$

or

$$\dot{\lambda}(t) = (\alpha - a'(x))\lambda(t). \quad (15.8)$$

From $g'(u) = \lambda$ we have $g''(u)\dot{u} = \dot{\lambda}$ and hence from (15.8) we obtain

$$\dot{u} = \frac{1}{\sigma(u)} [a'(x) - \alpha], \quad (15.9)$$

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 (15.6) and (15.9). 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 $x(T) = 0$.

16 Controlled Diffusion Processes

Control problems in a continuous-time, continuous state space, stochastic setting.

16.1 The dynamic programming equation

The DP equation in incremental form is

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

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

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

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

$$\Lambda(u)\phi(x) = \lim_{\delta t \downarrow 0} \left[\frac{E[\phi(x(t + \delta t)) \mid x(t) = x, u(t) = u] - \phi(x)}{\delta t} \right] \quad (16.1)$$

or

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

the conditional expectation of the ‘rate of change’ of $\phi(x)$ along the path. The operator Λ converts a scalar function of state, $\phi(x)$, to another such function, $\Lambda\phi(x)$. However, its action depends upon the control u , so we write it as $\Lambda(u)$. It is called the **infinitesimal generator** of the controlled Markov process. Equation (16.1) is equivalent to

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

This equation takes radically different forms depending upon whether the state space is discrete or continuous. Both are important.

If the state space is discrete we have the Markov jump process of Lecture 9. In this case $\Lambda(u)\phi(i) = \sum_j q_{ij}(u)[\phi(j) - \phi(i)]$. Now we turn to the case of continuous state space.

16.2 Diffusion processes and controlled diffusion processes

The **Wiener process** $\{B(t)\}$, is a scalar process for which $B(0) = 0$, the increments in B over disjoint time intervals are statistically independent and $B(t)$ is normally distributed with zero mean and variance t . (‘ B ’ stands for **Brownian motion**. It can be understood as a $\delta \rightarrow 0$ limit of a symmetric random walk in which steps $\pm\sqrt{\delta}$ are made at times $\delta, 2\delta, \dots$) The specification is internally consistent because, for example,

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

and for $0 \leq t_1 \leq t$ the two terms on the right-hand side are independent normal variables of zero mean and with variance t_1 and $t - t_1$ respectively.

If δB is the increment of B in a time interval of length δt then

$$E(\delta B) = 0, \quad E[(\delta B)^2] = \delta t, \quad E[(\delta B)^j] = o(\delta t), \quad \text{for } j > 2,$$

where the expectation is one conditional on the past of the process. Note that since

$$E[(\delta B/\delta t)^2] = O[(\delta t)^{-1}] \rightarrow \infty,$$

the formal derivative $\epsilon = dB/dt$ (continuous-time ‘white noise’) does not exist in a mean-square sense, but expectations such as

$$E \left[\left\{ \int \alpha(t) \epsilon(t) dt \right\}^2 \right] = E \left[\left\{ \int \alpha(t) dB(t) \right\}^2 \right] = \int \alpha(t)^2 dt$$

make sense if the integral is convergent.

Now consider a **stochastic differential equation**

$$\delta x = a(x, u) \delta t + g(x, u) \delta B,$$

which we shall write formally as

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

This, as a Markov process, has an infinitesimal generator with action

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

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

$$\inf_u \left[c + F_t + F_x a + \frac{1}{2} N F_{xx} \right] = 0, \tag{16.2}$$

and in the vector case

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

16.3 Example: noisy LQ regulation in continuous time

The dynamic programming equation is

$$\inf_u \left[x^\top R x + u^\top Q u + F_t + F_x^\top (A x + B u) + \frac{1}{2} \text{tr}(N F_{xx}) \right] = 0.$$

In analogy with the discrete and deterministic continuous cases that we have considered previously, we try a solution of the form,

$$F(x, t) = x^\top \Pi(t) x + \gamma(t).$$

This leads to the same Riccati equation as in Section 13.4,

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

and also, as in Section 10.3,

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

16.4 Example: passage to a stopping set

Consider a problem of movement on the unit interval $0 \leq x \leq 1$ in continuous time, $\dot{x} = u + \epsilon$, where ϵ is white noise of power v . The process terminates at time T when x reaches one end or the other of the the interval. The cost is made up of an integral term $\frac{1}{2} \int_0^T (L + Qu^2) dt$, penalising both control and time spent, and a terminal cost which takes the value C_0 or C_1 according as termination takes place at 0 or 1.

Show that in the deterministic case $v = 0$ one should head straight for one of the termination points at a constant rate and that the value function $F(x)$ has a piecewise linear form, with possibly a discontinuity at one of the boundary points if that boundary point is the optimal target from no interior point of the interval.

Show, in the stochastic case, that the dynamic programming equation with the control value optimized out can be linearised by a transformation $F(x) = \alpha \log \phi(x)$ for a suitable constant α , and hence solve the problem.

Solution. In the deterministic case the optimality equation is

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

with boundary conditions $F(0) = C_0$, $F(1) = C_1$. If one goes (from x) for $x = 0$ at speed w one incurs a cost of $C_0 + (x/2w)(L + Qw^2)$ with a minimum over w value of $C_0 + x\sqrt{LQ}$. Indeed (16.3) is solved by

$$F(x) = \min \left[C_0 + x\sqrt{LQ}, C_1 + (1 - x)\sqrt{LQ} \right].$$

The minimizing option determines the target and the optimal w is $\sqrt{L/Q}$.

In the stochastic case

$$\inf_u \left[\frac{L + Qu^2}{2} + u \frac{\partial F}{\partial x} + \frac{v}{2} \frac{\partial^2 F}{\partial x^2} \right] = 0.$$

So $u = -Q^{-1}F_x$ and

$$L - Q^{-1} \left(\frac{\partial F}{\partial x} \right)^2 + v \frac{\partial^2 F}{\partial x^2} = 0.$$

Make the transform $F(x) = -Qv \log \phi(x)$ so $\phi(x) = e^{-F(x)/Qv}$. Then the above equation simplifies to

$$Qv^2 \frac{\partial^2 \phi}{\partial x^2} - L\phi = 0,$$

with solution

$$\phi(x) = k_1 \exp\left(\frac{x}{v} \sqrt{L/Q}\right) + k_2 \exp\left(-\frac{x}{v} \sqrt{L/Q}\right).$$

We choose the constants k_1, k_2 to meet the two boundary conditions on F .

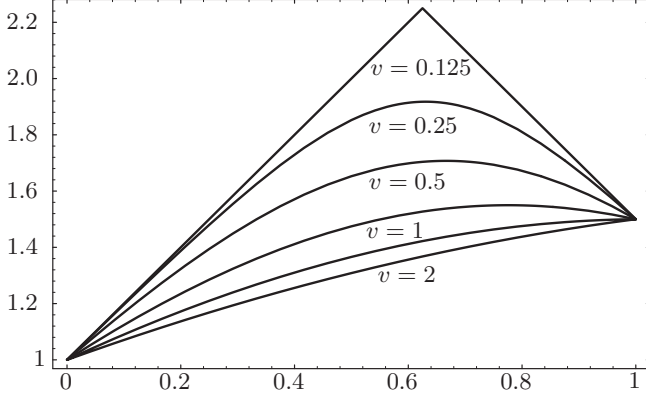


Figure 5: $F(x)$ against x for the passage to a stopping set

The figure shows the solution for $L = 1$, $Q = 4$, $C_0 = 1$, $C_1 = 1.5$ and $v = 0.125, 0.25, 0.5, 1, 2$ and the deterministic solution.

Notice that for these parameter choices the presence of noise actually reduces cost. This is because we are nearly indifferent as to which endpoint we hit, and L is small relative to Q . So it will be good to keep u small and let the noise do most of the work in bringing the state to endpoint.