Optimization and Control

J.R. Norris

November 22, 2007

1 Controllable dynamical systems

Controllable dynamical systems may be considered both in discrete time, with parameter $n \in \mathbb{Z}^+ = \{0, 1, 2, \dots\}$, and in continuous time, with parameter $t \in \mathbb{R}^+ = [0, \infty)$. They may be deterministic or stochastic, that is to say random. They are the basic objects of interest in this course. We now present the four main types.

1.1 Discrete-time, deterministic

Let S and A be sets. A *discrete-time controllable dynamical system* with *state-space* S and *action-space*[1] A is a map $f : \mathbb{Z}^+ \times S \times A \to S$. The interpretation is that, if at time n, in state x, we choose action a, then we move to state $f(n, x, a)$ at time $n + 1$. When f has no dependence on its first argument, we call the system *time-homogeneous*. A *control* is a map $u : \mathbb{Z}^+ \to A$. Given a starting time and state (k, x) and a control u, the *controlled sequence* $(x_n)_{n \geqslant k}$ is defined by $x_k = x$ and the equation[2]

$$x_{n+1} = f(n, x_n, u_n), \quad n \geqslant k.$$

In the time-homogeneous case[3], we shall usually specify only a starting state x and take as understood the starting time $k = 0$.

1.2 Discrete-time, stochastic

Assume for now that S is countable. Write $\text{Prob}(S)$ for the set of probability measures on S. We identify each $p \in \text{Prob}(S)$ with the vector $(p_y : y \in S)$ given by $p_y = p(\{y\})$. A *discrete-time stochastic controllable dynamical system*[4] with *state-space* S and *action-space* A is a map $P : \mathbb{Z}^+ \times S \times A \to \text{Prob}(S)$. The interpretation is that, if at time n, in state

[1] In fact, since the actions available in each state are often different, it is convenient sometimes to specify for each state x an action-space A_x, which may depend on x. Then the product $S \times A$ is replaced everywhere by $\cup_{x \in S} \{x\} \times A_x$. This makes no difference to the theory, which we shall therefore explain in the simpler case, only reviving the notation A_x in certain examples.

[2] This is sometimes called the *plant equation*.

[3] We can always reduce to the time-homogeneous case as follows: define $\tilde{S} = \mathbb{Z}^+ \times S$ and, for $(n, x) \in \tilde{S}$, set $\tilde{f}((n, x), a) = (n + 1, f(n, x, a))$. If $(x_n)_{n \geqslant k}$ is the controlled sequence of f for starting time and state $\tilde{x} = (k, x)$ and control u, and if we set, for $n \geqslant 0$, $\tilde{u}_n = u_{k+n}$ and $\tilde{x}_n = (k + n, x_{k+n})$, then $(\tilde{x}_n)_{n \geqslant 0}$ is the controlled sequence of $\tilde{f}$ for starting state $\tilde{x}$ and control $\tilde{u}$.

[4] The term *Markov decision process* is also used, although P is not a process. However, we shall see that a choice of Markov control associates to P a Markov process.

x, we choose action a, then we move to y at time $n+1$ with probability $P(n,x,a)_y$. We write, for a function F on $\mathbb{Z}^+ \times S$,

$$PF(n,x,a) = \int_S F(n+1,y)P(n,x,a)(dy) = \sum_{y\in S} P(n,x,a)_y F(n+1,y).$$

Thus, also, $PF(n,x,a) = \mathbb{E}(F(n+1,Y))$, where Y is a random variable with distribution $P(n,x,a)$. Often, P will be *time-homogeneous*[5] and will be considered as a function $S\times A \to \text{Prob}(S)$. Then we shall write, for a function F on S,

$$PF(x,a) = \sum_{y\in S} P(x,a)_y F(y).$$

A *control* is a map $u : S^* \to A$, where

$$S^* = \{(n,x_k,x_{k+1},\dots,x_n) : k,n \in \mathbb{Z}^+, k \leqslant n, x_k,x_{k+1},\dots,x_n \in S\}.$$

Given a control u and a starting time and state (k,x), we specify the distribution of a random process $(X_n)_{n\geqslant k}$ by the requirement that for all $n \geqslant k$ and all $x_k,\dots,x_n \in S$,

$$\begin{aligned}\mathbb{P}(X_k = x_k, X_{k+1} = x_{k+1},\dots,X_n = x_n)\\ = \delta_{xx_k} P(k,x_k,u_k(x_k))_{x_{k+1}}\\ \times P(k+1,x_{k+1},u_{k+1}(x_k,x_{k+1}))_{x_{k+2}} \dots P(n-1,x_{n-1},u_{n-1}(x_k,\dots,x_{n-1}))_{x_n}.\end{aligned}$$

Thus, we determine the action that we take at each time n as a function u of n and of the history of the process up to that time. When we want to indicate the choice of control u and starting time and state (k,x), we shall write $\mathbb{P}^u_{(k,x)}$ in place of $\mathbb{P}$, and similarly $\mathbb{E}^u_{(k,x)}$ in place of $\mathbb{E}$. We take $k=0$ unless otherwise indicated and then write simply $\mathbb{P}^u_x$. We call $(X_n)_{n\geqslant k}$ the *controlled process*. A function $u : \mathbb{Z}^+ \times S \to A$ is called a *Markov control* and is identified with the control $(n,x_k,\dots,x_n) \mapsto u_n(x_n)$. A function $u : S \to A$ is called a *stationary Markov control*. In the *time-homogeneous* case, the controlled process determined by a stationary Markov control u is a (time-homogeneous) Markov chain on S, with transition matrix $P^u = (p^u_{xy} : x,y \in S)$ given by $p^u_{xy} = P(x,u(x))_y$. More generally, for any Markov control u, the controlled process $(X_n)_{n\geqslant 0}$ is a time-inhomogeneous Markov chain with time-dependent transition matrix $P^u(n) = (p^u_{xy}(n) : x,y \in S)$ given by $p^u_{xy}(n) = P(n,x,u_n(x))_y$.

Here is common way for a stochastic controllable dynamical system to arise: there is given a sequence of independent, identically distributed, random variables $(\varepsilon_n)_{n\geqslant 1}$, with values in a set E, say, and a function $G : \mathbb{Z}^+ \times S \times A \times E \to S$. We can then take $P(n,x,a)$ to be the distribution on S of the random variable $G(n,x,a,\varepsilon)$. Thus, for a function F on $\mathbb{Z}^+ \times S$, we have

$$PF(n,x,a) = \mathbb{E}(F(n+1,G(n,x,a,\varepsilon))). \tag{1}$$

Given a control u, this gives a ready-made way to realise the controlled process $(X_n)_{n\geqslant k}$, using the recursion[6]

$$X_{n+1} = G(n,X_n,U_n,\varepsilon_{n+1}), \quad U_n = u_n(X_k,\dots,X_n).$$

[5] A reduction to the time-homogeneous case can be made by a procedure analogous to that described in footnote 3. The details are left as an exercise.

[6] This is like the deterministic plant equation.

We shall call the pair $(G, (\varepsilon_n)_{n\geqslant 1})$ a *realised stochastic controllable dynamical system*. Every stochastic controllable dynamical system can be realised in this way; sometimes this is natural, at other times not. The notion of a realised stochastic controllable dynamical system provides a convenient way to generalize our discussion to the case where S is no longer countable. We shall consider in detail the case where $S = \mathbb{R}^n$, where the random variables ε_n are Gaussian, and where G is an affine function of x and ε.

1.3 Continuous-time, deterministic

Take now as state-space $S = \mathbb{R}^d$, for some $d \geqslant 1$. A *time-dependent vector field* on $\mathbb{R}^d$ is a map $b : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$. Given a starting point $x_0 \in \mathbb{R}^d$, we can attempt to define a continuous path $(x_t)_{t\geqslant 0}$ in $\mathbb{R}^d$, called the *flow* of b, by solving the differential equation $\dot{x}_t = b(t, x_t)$ for $t \geqslant 0$, with initial value x_0. In examples, we shall often calculate solutions explicitly. In Section 18 we shall show that continuity of b, or just piecewise continuity in time, together with the Lipschitz condition (4), guarantees the existence of a unique solution, even if we cannot calculate it explicitly. The Lipschitz condition is in turn implied by the existence and boundedness of the gradient $\nabla b = \partial b/\partial x$, which is usually easy to check.

A *continuous-time controllable dynamical system* with *action-space* A is given by a map $b : \mathbb{R}^+ \times \mathbb{R}^d \times A \to \mathbb{R}^d$. We interpret this as meaning that, if at time t, in state x, we choose action a, then we move at that moment with velocity $b(t, x, a)$. A *control* is a map $u : \mathbb{R}^+ \to A$. Given a control u, we obtain a vector field b^u by setting $b^u(t, x) = b(t, x, u_t)$. Then, given a starting time and place (s, x), the *controlled path* $(x_t)_{t\geqslant s}$ is defined by the differential equation $\dot{x}_t = b^u(t, x_t)$ for $t \geqslant s$, with initial value $x_s = x$. More generally, it is sometimes convenient to consider as a control a map $u : \mathbb{R}^+ \times \mathbb{R}^d \to A$. Then we set $b^u(t, x) = b(t, x, u(t, x))$ and solve the differential equation as before.

1.4 Continuous-time, stochastic

The most common continuous-time Markov processes fall into two types, jump processes and diffusions, each of which has a controllable counterpart. For simplicity, we give details only for the time-homogeneous case.

We shall consider jump processes only in the case where the state-space S is countable. In this context, Markov processes are called Markov chains[7]. A Markov chain is specified by a Q-matrix Q. Given a starting point $x_0 \in S$, there is an associated continuous-time Markov chain $(X_t)_{t\geqslant 0}$, starting from x_0, with generator matrix Q.

A *continuous-time jump-type stochastic controllable dynamical system* with *state-space* S and *action-space* A is given by a pair of maps $q : S \times A \to \mathbb{R}^+$ and $\pi : S \times A \to \mathrm{Prob}(S)$. We insist that $\pi(x, a)$ have no mass at x. If action a is chosen, then we jump from x at rate $q(x, a)$ to a new state, chosen with distribution $\pi(x, a)$. A *stationary Markov control* is a map $u : S \to A$, and serves to specify a Q-matrix Q^u, and hence a Markov chain, by

$$q^u_{xx} = -q(x, u(x)), \quad q^u_{xy} = q(x, u(x))\pi(x, u(x))_y, \quad y \neq x.$$

[7]These are one of the subjects of the Part II course Applied Probability.

The problem is to optimize the Markov chain over the controls. Note that this type of system has no deterministic analogue, as the only deterministic continuous-time time-homogeneous Markov process of jump type is a constant.

A diffusion process is a generalization of the differential equation $\dot{x}_t = b(x_t)$. Fix $m \geqslant 1$ and specify, in addition to the vector field b, called in this context the *drift*, m further vector fields $\sigma_1, \ldots, \sigma_m$ on S. Take m independent Brownian motions $B^1, \ldots, B^m$ and attempt to solve the stochastic differential equation[8]

$$dX_t = \sum_i \sigma_i(X_t) dB_t^i + b(X_t)dt.$$

The intuition behind this equation is that we move from x in an infinitesimal time δt by a normal random variable with mean $b(x)\delta t$ and with covariance matrix $\sum_i \sigma_i(x)\sigma_i(x)^T \delta t$. The solution $(X_t)_{t \geqslant 0}$, is a Markov process in S having continuous paths, which is known as a diffusion process.

A *continuous-time diffusive stochastic controllable dynamical system* with *state-space* S and *action-space* A is given by a family of maps $\sigma_i : S \times A \to \mathbb{R}^d, i = 1, \ldots, m$, and $b : S \times A \to \mathbb{R}^d$. We assume that these maps are all continuously differentiable on $\mathbb{R}^d$. If action a is chosen, then, intuitively, we move from x in an infinitesimal time δt by a normal random variable with mean $b(x, a)\delta t$ and with covariance matrix $\sum_i \sigma_i(x, a)\sigma_i(x, a)^T \delta t$. On specifying a *stationary Markov control* $u : S \to A$, we obtain the coefficients for a stochastic differential equation by $\sigma_i^u(x) = \sigma_i(x, u(x))$ and $b^u(x) = b(x, u(x))$, and hence, subject to some regularity conditions, we can define a diffusion process. Stochastic differential equations, diffusions, and the controllable systems and controls just introduced all have straightforward time-dependent generalizations.

[8]This discussion is intended only to sketch the outline of the theory, which is treated in the Part III course Stochastic Calculus and Applications. Provided that $\sigma_1, \ldots, \sigma_m$ and b are all differentiable, with bounded derivative, the equation has a unique maximal local solution, just as in the deterministic case.

2 The dynamic optimality equation

We introduce a *cost function*

$$c : \mathbb{Z}^+ \times S \times A \to \mathbb{R}.$$

We assume throughout that one of the following three conditions holds: either c is non-negative, or c is non-positive, or there is a convergent series of constants $\sum_n c_n < \infty$ such that $|c(n,.,.)| \leqslant c_n$ for all n. In the second case, we shall usually express everything in terms of the *reward function* $r = -c$. Given a controllable dynamical system f and a control u, we define the *total cost function* $V^u : \mathbb{Z}^+ \times S \to \mathbb{R}$ by

$$V^u(k,x) = \sum_{n=k}^{\infty} c(n, x_n, u_n),$$

where $(x_n)_{n \geqslant k}$ is given by $x_k = x$ and $x_{n+1} = f(n, x_n, u_n)$ for all $n \geqslant k$. On the other hand, given a stochastic controllable dynamical system P with control u, we define the *expected total cost function* $V^u : \mathbb{Z}^+ \times S \to \mathbb{R}$ by

$$V^u(k,x) = \mathbb{E}^u_{(k,x)} \sum_{n=k}^{\infty} c(n, X_n, U_n),$$

where $U_n = u_n(X_k, \ldots, X_n)$. In order to avoid the use of measure theory, we assume in the stochastic case, until Section 11, that the state-space S is countable. All the notions and results we present extend in a straightforward way to the case of a general measurable space $(S, \mathcal{S})$. Our assumptions on c are sufficient to ensure that the sums and expectations here are well-defined. The *infimal cost function* is defined by

$$V(k,x) = \inf_u V^u(k,x),$$

where the infimum is taken over all controls[9] [10]. A control u is *optimal* for (k,x) if $V^u(k,x) = V(k,x)$. The main problem considered in this course is the calculation of V and the identification of optimal controls u, when they exist, in this and some analogous

[9] Note that we have used a smaller class of controls in the deterministic case and a larger class in the stochastic case. Suppose we fix a starting time and state (k,x) and use a control u from the larger class in a deterministic controllable dynamical system, obtaining a controlled sequence $(x_n)_{n \geqslant k}$. Set $\tilde{u}_n = u_n(x_k, \ldots, x_n)$ for $n \geqslant k$, and define $\tilde{u}_n$ arbitrarily for $n \leqslant k-1$. Then $\tilde{u}$ belongs to the smaller class of controls and has the same controlled sequence starting from (k,x). Hence there is no loss in restricting the infimum to the smaller class, but we should be mindful that the infimum can no longer be approached simultaneously for all (k,x) by a single sequence of controls. The larger class is necessary in the stochastic case to specify an appropriate dependence of the control on the controlled process.

[10] In some applications we shall have costs of the more general form $c(n, X_n, U_n, X_{n+1})$. However, we can always reduce to the case under discussion using the formula

$$\mathbb{E}^u_{(k,x)}(c(k, n_k, U_n, X_{n+1}) | X_k, \ldots, X_n) = \bar{c}(n, X_n, U_n),$$

where

$$\bar{c}(n,x,a) = \sum_{y \in S} P(n,x,a)_y c(n,x,a,y).$$

contexts. In most practical cases, a simple search over all controls is infeasible, because there are too many possibilities. Instead, the main approach is based on the following result.

Proposition 2.1. *The infimal cost function satisfies the* dynamic optimality equation[11]

$$V(k,x) = \inf_a \{c(k,x,a) + V(k+1, f(k,x,a))\} \qquad (\textit{deterministic case}),$$
$$V(k,x) = \inf_a (c+PV)(k,x,a) \qquad (\textit{stochastic case}),$$

for all $k \geqslant 0$ *and* $x \in S$.

Proof. A simpler variant of the following argument may be used to prove the deterministic case. This is left as an exercise. Fix $k \in \mathbb{Z}^+$ and $x \in S$. Note that $V^u(k,x)$ depends on u only through $a = u_k(x)$ and through the control $\tilde{u}$, given for $n \geqslant k+1$ by $\tilde{u}_n(x_{k+1}, \ldots, x_n) = u_n(x, x_{k+1}, \ldots, x_n)$. By conditioning on X_{k+1}, we have

$$V^u(k,x) = c(k,x,a) + \sum_{y \in S} P(k,x,a)_y V^{\tilde{u}}(k+1,y).$$

Now, as we vary $\tilde{u}$ over all controls, we can approach the infimal value of $V^{\tilde{u}}(k+1,y)$ for all $y \in S$ simultaneously, so we obtain

$$\inf_{u_k(x)=a} V^u(k,x) = (c+PV)(k,x,a).$$

The optimality equation is now obtained on taking the infimum over $a \in A$. □

The idea of the proof is thus to condition on the first step and use the fact that the resulting constrained minimization is similar in form to the original. We used here the *Principle of Optimality*. In its most abstract form, this is just the fact that one can take an infimum over a set S given as a union $\cup_{a \in A} S_a$ by

$$\inf_{x \in S} f(x) = \inf_{a \in A} \inf_{x \in S_a} f(x).$$

Another, more concrete, instance is the fact any path of minimal length between two points must also minimize length between any two intermediate points on the path. Note that the proposition says nothing about uniqueness of solutions to the optimality equation. We shall look into this in a number of more specific contexts in the next few sections.

[11] Also called the *dynamic programming* or *Bellman* equation.

3 Finite-horizon dynamic optimization

We show how to optimize a controllable dynamical system over finitely many time steps. Fix a time horizon $n \in \mathbb{Z}^+$ and assume that

$$c(n, x, a) = C(x) \quad \text{and} \quad c(k, x, a) = 0, \quad k \geqslant n+1, \quad x \in S, \quad a \in A.$$

Thus the total cost function is given by

$$V^u(k, x) = \sum_{j=k}^{n-1} c(j, x_j, u_j) + C(x_n), \quad 0 \leqslant k \leqslant n,$$

in the deterministic case, and in the stochastic case by

$$V^u(k, x) = \mathbb{E}^u_{(k,x)} \left(\sum_{j=k}^{n-1} c(j, X_j, U_j) + C(X_n) \right), \quad 0 \leqslant k \leqslant n.$$

Note that $V(k, x) = 0$ for all $k \geqslant n+1$. Hence, the optimality equation can be written in the form

$$\begin{aligned} V(n, x) &= C(x), && x \in S, \\ V(k, x) &= \inf_a \{c(k, x, a) + V(k+1, f(k, x, a))\}, && 0 \leqslant k \leqslant n-1, \quad x \in S, \end{aligned}$$

in the deterministic case, and in the stochastic case by[12]

$$\begin{aligned} V(n, x) &= C(x), && x \in S, \\ V(k, x) &= \inf_a (c + PV)(k, x, a), && 0 \leqslant k \leqslant n-1, \quad x \in S. \end{aligned}$$

Both these equations have a unique solution, which moreover may be computed by a straightforward[13] backwards recursion from time n. Once we have computed V, an optimal control can be identified whenever we can find a minimizing action in the optimality equations for $0 \leqslant k \leqslant n-1$. The following easy result verifies this for the deterministic case.

[12] It is often convenient to write the equation in terms of the *time to go* $m = n - k$. Assume that P is time-homogeneous and set $V_m(x) = V(k, x)$ and $c_m(x, a) = c(k, x, a)$, then the optimality equations become $V_0(x) = C(x)$ and

$$V_{m+1}(x) = \inf_a (c_m + PV_m)(x, a), \quad 0 \leqslant m \leqslant n-1, \quad x \in S.$$

In particular, in the case where both P and c are time-homogeneous, if we define

$$V^u_n(x) = \mathbb{E}^u_x \sum_{j=0}^{n-1} c(X_j, U_j), \quad V_n(x) = \inf_u V^u_n(x),$$

then the functions V_n are given by $V_0(x) = 0$ and, for $n \geqslant 0$,

$$V_{n+1}(x) = \inf_a (c + PV_n)(x, a), \quad x \in S.$$

[13] Although straightforward in concept, the size of the state space may make this a demanding procedure in practice. It is worth remembering, as a possible alternative, the following *interchange argument*, when

Proposition 3.1. *Suppose we can find a control u, with controlled sequence $(x_0, \ldots, x_n)$ such that*

$$V(k, x_k) = c(k, x_k, u_k) + V(k+1, f(k, x_k, u_k)), \quad 0 \leqslant k \leqslant n-1.$$

Then u is optimal for $(0, x_0)$.

Proof. Fix a such a control u, and set

$$m_k = \sum_{j=0}^{k-1} c(j, x_j, u_j) + V(k, x_k), \quad 0 \leqslant k \leqslant n.$$

Then, for $0 \leqslant k \leqslant n-1$, since $x_{k+1} = f(k, x_k, u_k)$, we have

$$m_{k+1} - m_k = c(k, x_k, u_k) + V(k+1, x_{k+1}) - V(k, x_k) = 0.$$

Hence

$$V(0, x_0) = m_0 = m_n = \sum_{j=0}^{n-1} c(j, x_j, u_j) + C(x_n).$$

□

Example (Managing spending and saving). An investor holds a capital sum in a building society, which pays a fixed rate of interest $\theta \times 100\%$ on the sum held at each time $k = 0, 1, \ldots, n-1$. The investor can choose to reinvest a proportion a of the interest paid, which then itself attracts interest. No amounts invested can be withdrawn. How should the investor act to maximize total consumption by time $n-1$?

Take as state the present income $x \in \mathbb{R}^+$ and as action the proportion $a \in [0, 1]$ which is reinvested. The income next time is then

$$f(x, a) = x + \theta a x$$

and the reward this time is $r(x, a) = (1-a)x$. The optimality equation is given by

$$V(k, x) = \max_{0 \leqslant a \leqslant 1} \{(1-a)x + V(k+1, (1+\theta a)x)\}, \quad 0 \leqslant k \leqslant n-1,$$

---

seeking to optimize the order in which one performs a sequence of n tasks. Label the tasks $\{1, \ldots, n\}$ and write $c(\sigma)$ for the cost of performing the tasks in the order $\sigma = (\sigma_1, \ldots, \sigma_n)$. We examine the effect on $c(\sigma)$ of interchanging the order of two of the tasks. Suppose we can find a function f on $\{1, \ldots, n\}$ such that, for all σ and all $0 \leqslant i \leqslant n-1$,

$$c(\sigma') < c(\sigma) \quad \text{whenever} \quad f(\sigma_i) > f(\sigma_{i+1}),$$

where σ' is obtained from σ by interchanging the order of tasks σ_i and σ_{i+1}. Then the condition $f(\sigma_1) \leqslant \ldots \leqslant f(\sigma_n)$ is necessary for optimality of σ. This may be enough to reduce the number of possible optimal orders to 1. In any case, if we have also, for all σ and all $0 \leqslant i \leqslant n-1$,

$$c(\sigma') = c(\sigma) \quad \text{whenever} \quad f(\sigma_{i+1}) = f(\sigma_i),$$

then our optimality condition is also sufficient.

with $V(n,x) = 0$. Working back from time n, we see that $V(k,x) = c_{n-k}x$ for some constants $c_0, \ldots, c_n$, given by $c_0 = 0$ and

$$c_{m+1} = \max\{c_m + 1, (1+\theta)c_m\}, \quad 0 \leqslant m \leqslant n-1.$$

Hence

$$c_m = \begin{cases} m, & m \leqslant m^*, \\ m^*(1+\theta)^{m-m^*}, & m > m^*, \end{cases}$$

where $m^* = \lceil 1/\theta \rceil$. By Proposition 3.1, the optimal control is to reinvest everything before time $k^* = n - m^*$ and to consume everything from then on.

The optimality of a control in the stochastic case can verified using the following result.

Proposition 3.2. *Suppose we can find a Markov control u such that*

$$V(k,x) = (c + PV)(k, x, u_k(x)), \quad 0 \leqslant k \leqslant n-1, \quad x \in S.$$

Then u is optimal for all (k,x).

Proof. Fix such a Markov control u and write $(X_0, \ldots, X_n)$ for the associated Markov chain starting from $(0,x)$. Define

$$M_k = \sum_{j=0}^{k-1} c(j, X_j, U_j) + V(k, X_k), \quad 0 \leqslant k \leqslant n.$$

Then, for $0 \leqslant k \leqslant n-1$,

$$M_{k+1} - M_k = c(k, X_k, U_k) + V(k+1, X_{k+1}) - V(k, X_k),$$

so, for all $y \in S$,

$$\mathbb{E}^u(M_{k+1} - M_k | X_k = y) = (c + PV)(k, y, u_k(y)) - V(k,y) = 0.$$

Hence

$$V(0,x) = \mathbb{E}_x^u(M_0) = \mathbb{E}_x^u(M_n) = \mathbb{E}_x^u\left(\sum_{j=0}^{n-1} c(j, X_j, U_j) + C(X_n)\right).$$

The same argument works for all starting times k. □

Example (Exercising a stock option). You hold an option to buy a share at a fixed price p, which can be exercised at any time $k = 0, 1, \ldots, n-1$. The share price satisfies $Y_{k+1} = Y_k + \varepsilon_{k+1}$, where $(\varepsilon_k)_{k \geqslant 1}$ is a sequence of independent identically distributed random variables[14], with $\mathbb{E}(|\varepsilon|) < \infty$. How should you act to maximise your expected return?

Take as state the share price $x \in \mathbb{R}$, until we exercise the option, when we move to a terminal state ∂. Take as action space the set $\{0,1\}$, where $a = 1$ corresponds to exercising the option. The problem specifies a realised stochastic controllable dynamical system. We

[14]Thus we allow, unrealistically, the possibility that the price could be negative. This model might perhaps be used over a small time interval, with Y_0 large.

are working outside the countable framework here, but in the realised case, where PV is given straightforwardly by 1. The rewards and dynamics before termination are given by

$$r(x,a) = a(x-p), \quad G(x,a,\varepsilon) = \begin{cases} x+\varepsilon, & \text{if } a=0, \\ \partial, & \text{if } a=1, \end{cases}.$$

Hence the optimality equation is given by

$$V(k,x) = \max\{x-p, \mathbb{E}(V(k+1,x+\varepsilon))\}, \quad k=0,1,\ldots,n-1,$$

with $V(n,x)=0$. Note that $V(n-1,x)=(x-p)^+$. By backwards induction, we can show that $V(k,.)$ is convex for all k, and increases as k decreases. Set $p_k = \inf\{x \geqslant 0 : V(k,x) = x-p\}$. Then p_k increases as k decreases and the optimal control is to exercise the option as soon as $Y_k = p_k$.

4 Dynamic optimization for non-negative rewards

We show how to optimize a time-homogeneous stochastic controllable dynamical system with non-negative rewards over an infinite time-horizon[15].

Let P be a time-homogeneous stochastic controllable dynamical system with state-space S and action-space A. Suppose given a *reward function*

$$r : S \times A \to \mathbb{R}^+.$$

Given a control u, define the *expected total reward function*

$$V^u(x) = \mathbb{E}^u_x \sum_{n=0}^{\infty} r(X_n, U_n),$$

where, as usual, the notation signifies that $(X_n)_{n\geqslant 0}$ is the controlled process of u, starting from x, and where $U_n = u_n(X_0, \ldots, X_n)$. Define also the *optimal reward* or *value function*

$$V(x) = \sup_u V^u(x).$$

We are using notation inconsistent with Section 2 because we have defined V as the negative of the corresponding object in Section 2. The optimality equation transforms straightforwardly under this change of notation – one just replaces the infimum by a supremum.

Define for $n \geqslant 0$

$$V^u_n(x) = \mathbb{E}^u_x \sum_{k=0}^{n-1} r(X_k, U_k), \quad V_n(x) = \sup_u V^u_n(x).$$

By monotone convergence[16], since $r \geqslant 0$, $V^u_n(x) \uparrow V^u(x)$ as $n \to \infty$, for all x and u. So

$$V(x) = \sup_u \sup_n V^u_n(x) = \sup_n \sup_u V^u_n(x) = \sup_n V_n(x).$$

The functions V_n are finite-horizon optimal reward functions, which, taking advantage of time-homogeneity, can be computed iteratively using the optimality equation

$$V_{n+1}(x) = \sup_a (r + PV_n)(x, a),$$

so we have a way to compute V. This procedure is called *value iteration*.

Proposition 4.1. *The optimal reward function is the minimal non-negative solution of the* dynamic optimality equation

$$V(x) = \sup_a (r + PV)(x, a), \quad x \in S.$$

Hence, any control u, for which V^u also satisfies this equation, is optimal, for all starting states x.

[15] This is also called positive programming.

[16] This fundamental result of measure theory states that, for any sequence of measurable functions $0 \leqslant f_n \uparrow f$, and any measure μ, we have convergence of the corresponding integrals $\mu(f_n) \uparrow \mu(f)$. Here, it justifies the interchange of summation and expectation, for the expectation is a form of integral, and we just take f_n as the partial sum $\sum_{k=0}^{n-1} r(X_k, U_k)$.

Proof. We know that V satisfies the optimality equation by Proposition 2.1. Suppose F is another non-negative solution. Then $F \geqslant 0 = V_0$. Suppose inductively for $n \geqslant 0$ that $F \geqslant V_n$. Then

$$F(x) = \sup_a (r + PF)(x, a) \geqslant \sup_a (r + PV_n)(x, a) = V_{n+1}(x)$$

so the induction proceeds. Hence $F \geqslant \sup_n V_n = V$. □

Example (Possible lack of an optimal policy). Consider the controllable dynamical system $f(x, a) = a(x+1_{\{x \geqslant 1\}})$, with state-space $\mathbb{Z}^+$ and action-space $\{0, 1\}$. Take as reward function $r(x, a) = (1 - a)(1 - 1/x)$. Thus, in state $x \geqslant 1$, we can choose to jump up by 1, or to jump to 0, gaining a reward of $1 - 1/x$. Once at 0, no further reward is gained.

The optimality equations are given by $V(0) = 0$ and

$$V(x) = \max\{1 - 1/x, V(x + 1)\}, \quad x \geqslant 1.$$

It is straightforward to show that, for any $\lambda \in [1, \infty)$, the function V_λ, given by

$$V_\lambda(x) = \lambda 1_{\{x \geqslant 1\}},$$

is a solution of the optimality equations, and indeed that there are no other solutions. Then, by the proposition, we can identify the optimal reward function as the smallest of these functions, namely V_1. *However, there is no optimal control.* If we wait until we get to n, then we gain a reward $1 - 1/n$. But if we wait for ever, we get nothing. Note that waiting forever corresponds to the control $u(x) = 0$ for all x, which has the property that

$$V(x) = (r + PV)(x, u(x)), \quad x \in S.$$

So we see, contrary to the finite-horizon case, that this is not enough to guarantee optimality. We do have for this control that

$$V^u(x) = (r + PV^u)(x, u(x)), \quad x \in S.$$

However, V^u, which is the minimal non-negative solution of *this* equation, is identically zero.

Example (Optimal gambling). A gambler has one pound and wishes to increase it to N pounds. She can place bets on a sequence of favorable games, each independently having probability $p > 1/2$ of success, but her stake must be a whole number of pounds and may not exceed her current fortune. What strategy maximizes her chances of reaching her goal?

We take as state-space $S = \mathbb{Z}^+$. It is natural here to allow a state-dependent action-space[17] $A_x = \{0, 1, \ldots, x\}$. The optimality equations are given by

$$V(x) = \max_{a \in A_x} \{pV(x + a) + (1 - p)V(x - a)\}, \quad 1 \leqslant x \leqslant N - 1,$$

with $V(0) = 0$ and $V(x) = 1$ for all $x \geqslant N$. There is no systematic approach to solving these equations, so we guess that the timid strategy of betting one pound each time will

[17] See footnote 1.

be optimal. As motivation, we might compare the outcomes, firstly of betting two pounds once, and secondly, of successively betting one pound until we are either two up or two down. So, take $u(x) = 1$ for all x. Then, by a standard Markov chain argument,

$$V^u(x) = pV^u(x+1) + (1-p)V^u(x-1), \quad 1 \leqslant x \leqslant N-1,$$

with $V^u(0) = 1$ and $V^u(N) = 1$. These equations have *unique* solution

$$V^u(x) = (1-\lambda^x)/(1-\lambda^N),$$

where $\lambda = (1-p)/p \in (0,1)$. It now follows from the fact that V^u is concave that it satisfies the optimality equations too, so u is optimal.

5 Dynamic optimization for discounted costs

We show how to optimize a time-homogeneous stochastic controllable dynamical system with bounded costs, discounted[18] at rate $\beta \in (0,1)$.

Let P be a time-homogeneous stochastic controllable dynamical system with state-space S and action-space A. Suppose given a cost function

$$c : S \times A \to \mathbb{R},$$

and suppose that $|c(x,a)| \leqslant C$ for all x, a, for some constant $C < \infty$. Given a control u, define the *expected discounted cost function*

$$V^u(x) = \mathbb{E}^u_x \sum_{n=0}^{\infty} \beta^n c(X_n, U_n).$$

Define also the *infimal discounted cost function*

$$V(x) = \inf_u V^u(x).$$

Our current set-up corresponds in the framework of Section 2, to the case of a time-dependent cost function $(n,x,a) \mapsto \beta^n c(x,a)$.

Define, for $n \geqslant 0$ and any control u,

$$V^u_n(x) = \mathbb{E}^u_x \sum_{k=0}^{n-1} \beta^k c(X_k, U_k), \quad V_n(x) = \inf_u V^u_n(x).$$

Note that

$$|V^u_n(x) - V^u(x)| \leqslant C \sum_{k=n}^{\infty} \beta^k = \frac{C\beta^n}{1-\beta},$$

so, taking the infimum over u, we have

$$|V_n(x) - V(x)| \leqslant \frac{C\beta^n}{1-\beta} \to 0, \quad \text{as } n \to \infty.$$

[18] Such a discounting of future costs is normal in financial models, and reflects the fact that money can be invested to earn interest. There is a second way in which a discounted problem may arise. Consider the set-up of Section 4, modified by the introduction of a *killing time* T, with $\mathbb{P}(T \geqslant n+1) = \beta^n$ for all $n \geqslant 0$, independent of the controlled process $(X_n)_{n \geqslant 0}$. The idea is that, at each time step, independently, there is a probability β that some external event will terminate the process, and that no further rewards will be received. Then consider the *expected total reward function* for control u given by

$$V^u(x) = \mathbb{E}^u_x \sum_{n=0}^{T-1} r(X_n, U_n) = \mathbb{E}^u_x \sum_{n=0}^{\infty} r(X_n, U_n) 1_{\{T \geqslant n+1\}}.$$

Now

$$\mathbb{E}^u_x (r(X_n, U_n) 1_{\{T \geqslant n+1\}} | X_n, U_n) = \beta^n r(X_n, U_n),$$

so our problem reduces to the optimization of the *expected discounted reward function*

$$V^u(x) = \mathbb{E}^u_x \sum_{n=0}^{\infty} \beta^n r(X_n, U_n).$$

Taking advantage of time-homogeneity, the finite-horizon cost functions V_n may be determined iteratively for $n \geqslant 0$ by $V_0(x) = 0$ and the optimality equations

$$V_{n+1}(x) = \inf_a (c + \beta P V_n)(x, a), \quad x \in S.$$

Hence, as in the case of non-negative rewards, we can compute V by *value iteration*.

Proposition 5.1. *The infimal discounted cost function is the unique bounded solution of the* dynamic optimality equation

$$V(x) = \inf_a (c + \beta P V)(x, a), \quad x \in S.$$

Moreover, any map $u : S \to A$ such that

$$V(x) = (c + \beta P V)(x, u(x)), \quad x \in S,$$

defines an optimal control, for every starting state x.

Proof. We know that V satisfies the optimality equation by Proposition 2.1, and

$$|V(x)| \leqslant C \sum_{n=0}^{\infty} \beta^n = \frac{C}{1-\beta} < \infty,$$

so V is bounded. Let now F be any bounded solution of the optimality equation and let u be any control. Consider the process

$$M_n = \sum_{k=0}^{n-1} \beta^k c(X_k, U_k) + \beta^n F(X_n), \quad n \geqslant 0.$$

Then

$$M_{n+1} - M_n = \beta^n c(X_n, U_n) + \beta^{n+1} F(X_{n+1}) - \beta^n F(X_n),$$

so, for all $y \in S$ and $a \in A$,

$$\mathbb{E}_x^u(M_{n+1} - M_n | X_n = y, U_n = a) = \beta^n c(y, a) + \beta^{n+1} P F(y, a) - \beta^n F(y) \geqslant 0$$

and so

$$F(x) = \mathbb{E}_x^u(M_0) \leqslant \mathbb{E}_x^u(M_n) = V_n^u(x) + \beta^n \mathbb{E}_x^u F(X_n).$$

On letting $n \to \infty$, using the boundedness of F, we obtain $F \leqslant V^u$. Since u was arbitrary, this implies that $F \leqslant V$.

In the special case where we can find a stationary Markov control $u : S \to A$ such that

$$F(x) = (c + \beta P F)(x, u(x)), \quad x \in S,$$

then, for all $y \in S$,

$$\mathbb{E}_x^u(M_{n+1} - M_n | X_n = y) = 0.$$

Hence

$$F(x) = \mathbb{E}_x^u(M_0) = \mathbb{E}_x^u(M_n) = V_n^u(x) + \beta^n \mathbb{E}_x^u F(X_n) \tag{2}$$

and so, letting $n \to \infty$, the final term vanishes and we find that $F = V^u$. In particular, in the case $F = V$, such a control u is optimal.

We do not know in general that there exists such a minimizing u but, given $\varepsilon > 0$, we can always choose $\tilde{u}$ such that

$$(c + \beta PF)(x, \tilde{u}(x)) \leqslant F(x) + \varepsilon, \quad x \in S,$$

which we can write in the form

$$F(x) = (\tilde{c} + \beta PF)(x, \tilde{u}(x)), \quad x \in S,$$

for a new cost function $\tilde{c} \geqslant c - \varepsilon$. The argument of the preceding paragraph, with $\tilde{c}$ in place of c and $\tilde{u}$ in place of u now shows that

$$F(x) = \mathbb{E}_x^u \sum_{k=0}^{\infty} \beta^k \tilde{c}(X_k, \tilde{u}(X_k)) \geqslant V^{\tilde{u}}(x) - \frac{\varepsilon}{1-\beta} \geqslant V(x) - \frac{\varepsilon}{1-\beta}.$$

Since $\varepsilon > 0$ was arbitrary, we conclude that $V = F$. □

6 Dynamic optimization for non-negative costs

We show how to optimize a time-homogeneous stochastic controllable dynamical system with non-negative costs over an infinite time-horizon[19].

Let P be a time-homogeneous stochastic controllable dynamical system with state-space S and action-space A. Suppose given a *cost function*

$$c : S \times A \to \mathbb{R}^+.$$

Given a control u, define, as above, the *expected total cost function* V^u and the *infimal cost function* V by

$$V^u(x) = \mathbb{E}^u_x \sum_{n=0}^{\infty} c(X_n, U_n), \quad V(x) = \inf_u V^u(x).$$

Recall from Section 4 that $V^u_n(x) \uparrow V^u(x)$ as $n \to \infty$, where

$$V^u_n(x) = \mathbb{E}^u_x \sum_{k=0}^{n-1} c(X_k, U_k).$$

Proposition 6.1. *Assume that A is finite. Then the infimal cost function is the minimal non-negative solution of the* dynamic optimality equation

$$V(x) = \min_a (c + PV)(x, a), \quad x \in S.$$

Moreover, any map $u : S \to A$ such that

$$V(x) = (c + PV)(x, u(x)), \quad x \in S,$$

defines an optimal control, for every starting state x.

Proof. We know by Proposition 2.1 that V is a solution of the optimality equation. Suppose that F is another non-negative solution. We use the finiteness of A to find a map $\tilde{u} : S \to A$ such that

$$F(x) = (c + PF)(x, \tilde{u}(x)), \quad x \in S.$$

The argument leading to equation (2) is valid when $\beta = 1$, so we have

$$F(x) = V^{\tilde{u}}_n(x) + \mathbb{E}^{\tilde{u}}_x F(X_n) \geqslant V^{\tilde{u}}_n(x).$$

On letting $n \to \infty$, we obtain $F \geqslant V^{\tilde{u}} \geqslant V$. Finally, when $F = V$ we can take $\tilde{u} = u$ to see that $V \geqslant V^u$, and hence that u defines an optimal control. □

The proposition allows us to see, in particular, that *value iteration* remains an effective way to approximate the infimal cost function in the current case. For let us set

$$V_n(x) = \inf_u V^u_n(x)$$

[19] This is also called negative programming – the problem can be recast in terms of non-positive rewards.

and note that $V_n(x) \uparrow V_\infty(x)$ as $n \to \infty$ for some function V_∞. Now $V_n^u \leqslant V^u$ for all n so, taking an infimum over controls we obtain $V_n \leqslant V$ and hence $V_\infty \leqslant V$. On the other hand we have the finite-horizon optimality equations

$$V_{n+1}(x) = \min_a (c + PV_n)(x, a), \quad x \in S,$$

and we can pass to the limit as $n \to \infty$ to see that V_∞ satisfies the optimality equation. But V is the minimal non-negative solution of this equation, so $V_\infty \geqslant V$, so $V_\infty = V$.

A second iterative approach to optimality is the method of *policy improvement*. We know that, for any given map $u : S \to A$, we have

$$V^u(x) = (c + PV^u)(x, u(x)), \quad x \in S.$$

If V^u does not satisfy the optimality equation, then we can find a strictly better control by choosing $\tilde{u} : S \to A$ such that

$$V^u(x) \geqslant (c + PV^u)(x, \tilde{u}(x)), \quad x \in S,$$

with strict inequality at some state x_0. Then, obviously, $V^u \geqslant V_0^{\tilde{u}} = 0$. Suppose inductively that $V^u \geqslant V_n^{\tilde{u}}$. Then

$$V^u(x) \geqslant (c + PV^u)(x, \tilde{u}(x)) \geqslant (c + PV_n^{\tilde{u}})(x, \tilde{u}(x)) = V_{n+1}^{\tilde{u}}(x), \quad x \in S,$$

so the induction proceeds and, letting $n \to \infty$, we obtain $V^u \geqslant V^{\tilde{u}}$, with strict inequality at x_0.

7 Optimal stopping

We show how optimal stopping problems for Markov chains can be treated as dynamic optimization problems.

Let $(X_n)_{n\geqslant 0}$ be a Markov chain on S, with transition matrix P. Suppose given two bounded functions

$$c : S \to \mathbb{R}, \quad f : S \to \mathbb{R},$$

respectively the *continuation cost* and the *stopping cost*. A random variable T, with values in $\mathbb{Z}^+ \cup \{\infty\}$, is a *stopping time* if, for all $n \in \mathbb{Z}^+$, the event $\{T = n\}$ depends only on $X_0, \ldots, X_n$. Define the *expected total cost function* V^T by

$$V^T(x) = \mathbb{E}_x \left(\sum_{k=0}^{T-1} c(X_k) + f(X_T) 1_{\{T<\infty\}} \right), \quad x \in S,$$

and define for $n \in \mathbb{Z}^+$ and $x \in S$,

$$V_n(x) = \inf_{T \leqslant n} V^T(x), \quad V_*(x) = \inf_{T<\infty} V^T(x), \quad V(x) = \inf_T V^T(x),$$

where the infima are taken over all stopping times T, first with the restriction $T \leqslant n$, then with $T < \infty$, and finally unrestricted. Where unbounded stopping times are involved, we assume that c and f are non-negative, so the sums and expectations are well defined. It is clear that $V_n(x) \geqslant V_{n+1}(x) \geqslant V_*(x) \geqslant V(x)$ for all n and x, as the infima are taken over progressively larger sets. The calculation of these functions and the determination, where possible, of minimizing stopping times are known as *optimal stopping problems*[20].

We translate these problems now into dynamic optimization problems, with state-space $S \cup \{\partial\}$ and action space $\{0, 1\}$. Action 0 will correspond to continuing, action 1 to stopping. On stopping, we go to ∂ and stay there. Define, for $x \in S$,

$$P(x, 0)_y = p_{xy}, \quad P(x, 1)_\partial = \delta_{y\partial},$$

and

$$c(x, a) = \begin{cases} c(x), & a = 0, \\ f(x), & a = 1. \end{cases}$$

Given a stopping time T, there exists for each $n \geqslant 0$ a set $B_n \subseteq S^{n+1}$ such that $\{T = n\} = \{(X_0, \ldots, X_n) \in B_n\}$. Define a control u by

$$u_n(x_0, \ldots, x_n) = \begin{cases} 1, & \text{if } (x_0, \ldots, x_n) \in B_n, \\ 0, & \text{otherwise.} \end{cases}$$

Note that we obtain all controls for starting time 0 in this way and that the controlled process is given by

$$\tilde{X}_n = \begin{cases} X_n, & n \leqslant T, \\ \partial, & n \geqslant T + 1. \end{cases}$$

[20] We limit our discussion to the time-homogeneous case. If there is a time dependence in the transition matrix or in the costs, a reduction to the time-homogeneous case can be achieved as in footnote 3, specifically, by considering the process $\tilde{X}_n = (k + n, X_{n+k})$.

Hence, V_n is the infimal cost function for the n-horizon problem, with final cost f, so satisfies $V_0(x) = f(x)$ and, for all $n \geqslant 0$,

$$V_{n+1}(x) = \min\{f(x), (c + PV_n)(x)\}, \quad x \in S.$$

Moreover, V is the infimal cost function for the infinite-horizon problem, so, if c and f are non-negative, then V is the minimal non-negative solution to

$$V(x) = \min\{f(x), (c + PV)(x)\}, \quad x \in S.$$

The V_* problem corresponds to a type of restriction on controls which we have not seen before. However the argument of Proposition 2.1 can be adapted to show that V_* also satisfies the optimality equation

$$V_*(x) = \min\{f(x), (c + PV_*)(x)\}, \quad x \in S.$$

Example. Consider a simple symmetric random walk on the integers with continuation cost $c(x) = 0$ and stopping cost $f(x) = 1 + e^{-x}$. Since f is convex, specifically since $f(x) \leqslant \frac{1}{2}f(x+1) + \frac{1}{2}f(x-1)$ for all x, a simple inductive argument[21] using the finite-horizon optimality equations shows that $V_n = f$ for all n. Since $(X_n)_{n\geqslant 0}$ is recurrent, the stopping time $T_n = \inf\{n \geqslant 0 : X_n = N\}$ is finite for all N, for every starting point x. So $V_*(x) \leqslant V^{T_n}(x) = 1 + e^{-N}$. Obviously, $V_*(x) \geqslant 1$, so $V_*(x) = 1$ for all x. Finally, $V = V^\infty = 0$. We note that $\inf_n V_n(x) > V_*(x) > V(x)$ for all x.

Proposition 7.1 (One step look ahead rule). *Suppose that $(X_n)_{n\geqslant 0}$ cannot escape from the set*

$$S_0 = \{x \in S : f(x) \leqslant (c + Pf)(x)\}.$$

Then, for all $n \geqslant 0$, the following stopping time is optimal for the n-horizon problem

$$T_n = \inf\{k \geqslant 0 : X_k \in S_0\} \wedge n.$$

Proof. The case $n = 0$ is trivially true. Suppose inductively that the claim holds for n. Then $V_n = f$ on S_0, so $PV_n = Pf$ on S_0 as we cannot escape. So, for $x \in S_0$,

$$V_{n+1}(x) = \min\{f(x), (c + PV_n)(x)\} = f(x)$$

and it is optimal to stop immediately. But, for $x \notin S_0$, it is better to wait, if we can. Hence the claim holds for $n + 1$ and the induction proceeds. □

[21] An alternative analysis of this example may be based on the *optional stopping theorem*, which is a fundamental result of martingale theory. This is introduced in the course Stochastic Financial Models and in the Part III course Advanced Probability. The random walk is a martingale, so, since f is convex, $(f(X_n))_{n\geqslant 0}$ is a submartingale. By optional stopping, $\mathbb{E}_x(f(X_T)) \geqslant \mathbb{E}_x(f(X_0)) = f(x)$ for all bounded stopping times T, so $V_n(x) = f(x)$ for all x. The fact that the conclusion of the optional stopping theorem does not extend to T_N is a well known sort of counterexample in martingale theory.

Example (Optimal parking). Suppose that you intend to park on the Backs, and wish to minimize the expected distance you will have to walk to Garrett Hostel Lane, and that a proportion p of the parking spaces are free. Assume that each parking space is free or occupied independently, that a queue of cars behind you take up immediately any space you pass by, and that no new spaces are vacated. Where should you park?

If you reach Garrett Hostel Lane without parking, then you should park in the next available space. This lies at a random distance (in spaces) D, with $\mathbb{P}(D = n) = (1-p)p^n$, for $n \geqslant 0$, so the expected distance to walk is $\mathbb{E}(D) = q/p$, where $q = 1 - p$. Here we have made the simplifying assumptions that Queen's Road is infinitely long and that there are no gaps between the spaces.

Write V_n for the minimal expected distance starting from n spaces before Garrett Hostel Lane. Then $V_0 = q/p$ and, for $n \geqslant 1$, $V_n = qV_{n-1} + p\min\{n, V_{n-1}\}$. Set $n^* = \inf\{n \geqslant 0 : V_n < n\}$. For $n \leqslant n^*$, we have $V_n = qV_{n-1} + pn$, so $V_n = n + (2q^n - 1)q/p$. Hence $n^* = \inf\{n \geqslant 0 : 2q^n < 1\}$. For $n \geqslant n^*$, we have $V_n = V_{n^*}$. The optimal time to stop is thus the first free space no more than n^* spaces before the Lane. We leave as an exercise the to express this argument in terms of the general framework described above.

8 Dynamic optimization for long-run average costs

We show how to optimize the long-run average cost for a time-homogeneous stochastic controllable dynamical system with bounded instantaneous costs.

Let P be a time-homogeneous stochastic controllable dynamical system with state-space S and action-space A. Suppose given a bounded *cost function* $c : S \times A \to \mathbb{R}$. Define, as usual, for a control u,

$$V_n^u(x) = \mathbb{E}_x^u \sum_{k=0}^{n-1} c(X_k, U_k), \quad x \in S,$$

where $U_k = u_k(X_0, \ldots, X_k)$. A control u is *optimal*, starting from x, if the limit

$$\lambda = \lim_{n\to\infty} \frac{V_n^u(x)}{n}$$

exists and if, for all other controls $\tilde{u}$,

$$\liminf_{n\to\infty} \frac{V_n^{\tilde{u}}(x)}{n} \geqslant \lambda.$$

The limit λ is then the *minimal long-run average cost* starting from x.

Proposition 8.1. *Suppose there exists a constant λ and a bounded function θ on S such that*

$$\lambda + \theta(x) \leqslant (c + P\theta)(x, a), \quad x \in S, \quad a \in A.$$

Then, for all controls u, and all $x \in S$,

$$\liminf_{n\to\infty} \frac{V_n^u(x)}{n} \geqslant \lambda.$$

Proof. Fix u and set

$$M_n = \theta(X_n) + \sum_{k=0}^{n-1} c(X_k, U_k) - n\lambda.$$

Then

$$M_{n+1} - M_n = \theta(X_{n+1}) - \theta(X_n) + c(X_n, U_n) - \lambda,$$

so, for all $y \in S$ and $a \in A$,

$$\mathbb{E}_x^u(M_{n+1} - M_n | X_n = y, U_n = a) = P\theta(y, a) - \theta(y) + c(y, a) - \lambda \geqslant 0.$$

Hence

$$\theta(x) = \mathbb{E}_x^u(M_0) \leqslant \mathbb{E}_x^u(M_n) = \mathbb{E}_x^u(\theta(X_n)) - n\lambda + V_n^u(x)$$

and so

$$\frac{V_n^u(x)}{n} \geqslant \lambda + \frac{\theta(x)}{n} - \frac{\mathbb{E}_n^u(\theta(X_n))}{n}$$

and we conclude by letting $n \to \infty$. □

By a similar argument, which is left as an exercise, one can also prove the following result.

Proposition 8.2. *Suppose there exists a constant λ and a bounded function θ on S, and a map $u : S \to A$, such that*

$$\lambda + \theta(x) \geqslant (c + P\theta)(x, u(x)), \quad x \in S.$$

Then, for all $x \in S$,

$$\limsup_{n\to\infty} \frac{V_n^u(x)}{n} \leqslant \lambda.$$

By combining the above two results, we see that, if λ and θ satisfy the *dynamic optimality equation*

$$\lambda + \theta(x) = \inf_a (c + P\theta)(x, a), \quad x \in S,$$

and if the infimum is achieved at $u(x)$ for each $x \in S$, then λ is the minimal long-run average cost and u defines an optimal control, for all starting states x. Note that, since $P1 = 1$, we can add any constant to θ and still have a solution. So, we are free to impose the condition $\theta(x_0) = 0$ for any given $x_0 \in S$ when looking for solutions. The function θ can then be thought of as the (un-normalized) extra cost of starting at x rather than x_0.

Example (Consultant's job selection). Each day a consultant is either free or is occupied with some job, which may be of m different types $1, \ldots, m$. Whenever he is free, he is given the opportunity to take on a job for the next day. A job of type x is offered with probability π_x and the types of jobs offered on different days are independent. On any day when he works on a job of type x, he completes it with probability p_x, independently for each day, and on its completion he is paid R_x. Which jobs should he accept?

We take as state-space the set $\{0, 1, \ldots, m\}$, where 0 corresponds to the consultant being free and $1, \ldots, m$ correspond to his working on a job of that type. The optimality equations for this problem are given by

$$\begin{aligned}
\lambda + \theta(0) &= \sum_{x=1}^{m} \pi_x \max\{\theta(0), \theta(x)\}, \\
\lambda + \theta(x) &= (1 - p_x)\theta(x) + p_x(R_x + \theta(0)), \quad x = 1, \ldots, m.
\end{aligned}$$

Take $\theta(0) = 0$, then $\theta(x) = R_x - (\lambda/p_x)$ for $x = 1, \ldots, m$, so the optimal λ must solve $\lambda = G(\lambda)$, where

$$G(\lambda) = \sum_{x=1}^{m} \pi_x \max\{0, R_x - (\lambda/p_x)\}.$$

Since G is non-increasing, there is a unique solution λ. The optimal control is then to accept jobs of type x if and only if $p_x R_x \geqslant \lambda$.

The optimality equation can be written down simply by reflecting on the details of the problem. A check on the validity of this process is provided by seeing how this particular problem can be expressed in terms of the general theory. For this, we take for state 0 the action-space $A_0 = \{(\varepsilon_1, \ldots, \varepsilon_m) : \varepsilon_x \in \{0, 1\}\}$. Here the action $(\varepsilon_1, \ldots, \varepsilon_m)$ signifies that we accept a job of type x if and only if $\varepsilon_x = 1$. There is no choice to be made in states $1, \ldots, m$. We take, for $x = 1, \ldots, m$,

$$P(0, \varepsilon)_x = \pi_x \varepsilon_x, \quad P(0, \varepsilon)_0 = \sum_{x=1}^{m} \pi_x (1 - \varepsilon_x), \quad P(x)_0 = p_x, \quad P(x)_x = 1 - p_x,$$

and

$$r(0, \varepsilon) = 0, \quad r(x) = p_x R_x.$$

The reward function here gives the expected reward in state x, as in the discussion in footnote 10. We leave as an exercise to see that the general form of the optimality equations specializes to the particular equations claimed. The complicated form of action-space reflects the fact that, in this example, we in fact make our choice based on knowledge of the type of job offered, whereas, in the general theory, the action is chosen without such knowledge.

The following result provides a *value iteration* approach to long-run optimality. Recall that the finite-horizon optimality equations are $V_0(x) = 0$ and, for $k \geqslant 0$,

$$V_{k+1}(x) = \inf_a (c + PV_k)(x, a), \quad x \in S.$$

Set

$$\lambda_k^- = \inf_x \{V_{k+1}(x) - V_k(x)\}, \quad \lambda_k^+ = \sup_x \{V_{k+1}(x) - V_k(x)\}.$$

Proposition 8.3. *For all $k \geqslant 0$ and all controls u, we have*

$$\liminf_{n\to\infty} \frac{V_n^u(x)}{n} \geqslant \lambda_k^-.$$

Moreover, if there exists $u : S \to A$ such that

$$V_{k+1}(x) = (c + PV_k)(x, u(x)), \quad x \in S,$$

then

$$\limsup_{n\to\infty} \frac{V_n^u(x)}{n} \leqslant \lambda_k^+.$$

Proof. Note that

$$\lambda_k^- + V_k(x) \leqslant V_{k+1}(x) \leqslant (c + PV_k)(x, a), \quad x \in S, \quad a \in A,$$

and

$$\lambda_k^+ + V_k(x) \geqslant V_{k+1}(x) = (c + PV_k)(x, u(x)), \quad x \in S,$$

and apply the preceding two propositions with $\theta = V_k$. □

9 Full controllability of linear systems

We begin a detailed study of linear controllable dynamical systems by finding criteria for existence of controls to get from any given state to any other.

Consider the linear controllable dynamical system, with state-space $\mathbb{R}^d$ and action-space $\mathbb{R}^m$, given by

$$f(x,a) = Ax + Ba, \quad x \in \mathbb{R}^d, \quad a \in \mathbb{R}^m.$$

Here A is a $d \times d$ matrix and B is a $d \times m$ matrix. We say that f is *fully controllable in n steps*[22] if, for all $x_0, x \in \mathbb{R}^d$, there is a control $(u_0, \dots, u_{n-1})$ such that $x_n = x$. Here, $(x_0, \dots, x_n)$ is the controlled sequence, given by $x_{k+1} = f(x_k, u_k)$ for $0 \leqslant k \leqslant n-1$. We then seek to minimize the *energy* $\sum_{k=0}^{n-1} |u_k|^2$ over the set of such controls.

Proposition 9.1. *The system f is fully controllable in n steps if and only if* $\operatorname{rank}(M_n) = d$, *where M_n is the $d \times nm$ matrix $[A^{n-1}B, \dots, AB, B]$. Set $y = x - A^n x_0$ and $G_n = M_n M_n^T$. Then the minimal energy from x_0 to x in n steps is $y^T G_n^{-1} y$ and this is achieved uniquely by the control*

$$u_k^T = y^T G_n^{-1} A^{n-k-1} B, \quad 0 \leqslant k \leqslant n-1.$$

Proof. By induction on $n \geqslant 0$ we obtain

$$x_n = A^n x_0 + A^{n-1} B u_0 + \cdots + B u_{n-1} = A^n x_0 + M_n u, \quad u = \begin{pmatrix} u_0 \\ \vdots \\ u_{n-1} \end{pmatrix},$$

from which the first assertion is clear. Fix $x_0, x \in \mathbb{R}^d$ and a control u such that $M_n u = y$. Then, by Cauchy–Schwarz,

$$y^T G_n^{-1} y = y^T G_n^{-1} M_n u \leqslant (y^T G_n^{-1} M_n M_n^T G_n^{-1} y)^{1/2} |u|,$$

so $\sum_{k=0}^{n-1} |u_k|^2 = |u|^2 \geqslant y^T G_n^{-1} y$, with equality if and only if $u^T = y^T G_n^{-1} M_n$. □

Note that $\operatorname{rank}(M_n)$ is non-decreasing in n and, by Cayley–Hamilton[23], is constant for $n \geqslant d$.

Consider now the continuous-time linear controllable dynamical system

$$b(x,u) = Ax + Bu, \quad x \in \mathbb{R}^d, \quad u \in \mathbb{R}^m.$$

Given a starting point x_0, the controlled process for control $(u_t)_{t \geqslant 0}$ is given by the solution of $\dot{x}_t = b(x_t, u_t)$ for $t \geqslant 0$. We say that b is *fully controllable in time t* if, for all $x_0, x \in \mathbb{R}^d$, there exists a control $(u_s)_{0 \leqslant s \leqslant t}$ such that $x_t = x$. We then seek to minimize the *energy* $\int_0^t |u_s|^2 ds$ subject to $x_t = x$. Note that

$$\frac{d}{dt}(e^{-At} x_t) = e^{-At}(\dot{x}_t - A x_t) = e^{-At} B u_t,$$

[22] This notion is also called *controllability* in accounts where controllable dynamical systems are called something else.

[23] This standard result of linear algebra states that a matrix satisfies its own characteristic equation

so

$$x_t = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu_s ds.$$

Consider for $t \geqslant 0$ the $d \times d$ matrix

$$G(t) = \int_0^t e^{As}BB^T(e^{As})^T ds.$$

Lemma 9.2. *For all $t > 0$, $G(t)$ is invertible if and only if* $\operatorname{rank}(M_d) = d$.

Proof. If $\operatorname{rank}(M_d) \leqslant d-1$, then we can find $v \in \mathbb{R}^d \setminus \{0\}$ such that $v^T A^n B = 0$ for all $n \leqslant d-1$, and hence for all $n \geqslant 0$ by Cayley–Hamilton. Then $v^T e^{As}B = 0$ for all s and so $v^T G(t)v = 0$ for all $t \geqslant 0$. On the other hand, if $\operatorname{rank}(M_d) = d$, then, given $v \in \mathbb{R}^d$, there is a smallest $n \geqslant 0$ such that $v^T A^n B \neq 0$. Then $|v^T e^{As}B| \sim |v^T A^n B| s^n/n!$ as $s \downarrow 0$, so $v^T G(t)v > 0$ for all $t > 0$. □

Proposition 9.3. *The system b is fully controllable in time t if and only if $G(t)$ is invertible. The minimal energy for a control from x_0 to x in time t is $y^T G(t)^{-1} y$, where $y = x - e^{At}x_0$, and is achieved uniquely by the control*

$$u_s^T = y^T G(t)^{-1} e^{A(t-s)}B.$$

The proof is similar to the proof of the discrete-time result and is left as an exercise. As the invertibility of $G(t)$ does not depend on the value of $t > 0$, we speak from now of simply of *full controllability* in the case of continuous time linear systems.

Example (Broom balancing). You attempt to balance a broom upside-down by supporting the tip of the stick in your palm. Is this possible?

We can resolve the dynamics in components to reduce to a one-dimensional problem. Write u for the horizontal distance of the tip from a fixed point of reference, and write θ for angle made by the stick with the vertical. Suppose that all the mass resides in the head of the broom, at a distance L from the tip. Newton's Law gives, for the component perpendicular to the stick of the acceleration of the head

$$g \sin\theta = \ddot{u}\cos\theta + L\ddot{\theta}.$$

We investigate the linearized dynamics near the fixed point $\theta = 0$ and $u = 0$. Replace θ by $\varepsilon\theta$ and u by εu. Then

$$g\varepsilon\theta = \varepsilon\ddot{u} + L\varepsilon\ddot{\theta} + O(\varepsilon^2),$$

so, in terms of $x = u + L\theta$ the linearized system is $\ddot{x} = \alpha(x - u)$, where $\alpha = g/L$, that is,

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

Then $\operatorname{rank}[AB, B] = 2$, so the linearized system is fully controllable. This provides evidence that, when the broom is close to vertical, we can bring by a suitable choice of control from any initial condition to rest while vertical.

Example (Satellite in a planar orbit). The following equations of motion describe a satellite moving in a planar orbit with radial thrust u_r and tangential thrust u_θ:

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

For each $\rho > 0$, there is a solution with $\dot{\theta} = \omega = \sqrt{c/\rho^3}$. We linearize around this solution, setting $r = \rho + \varepsilon x$, $\dot{\theta} = \omega + \varepsilon z$, $u_r \varepsilon u$ and $u_\theta = \varepsilon v$. After some routine calculations, and introducing $y = \dot{x}$, we obtain the linear controllable dynamical system

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = A \begin{pmatrix} x \\ y \\ z \end{pmatrix} + B \begin{pmatrix} u \\ v \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 & 0 \\ 3\omega^2 & 0 & 2\omega\rho \\ 0 & -2\omega/\rho & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1/\rho \end{pmatrix}.$$

It is straightforward to check that $\operatorname{rank}[AB, B] = 3$, so the linear system is fully controllable. On the other hand, if the tangential thrust would fail, so $v = 0$, we would have to replace B by its first column B_1. We have

$$B_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad AB_1 = \begin{pmatrix} 1 \\ 0 \\ -2\omega/\rho \end{pmatrix}, \quad A^2 B_1 = \begin{pmatrix} 0 \\ -\omega^2 \\ 0 \end{pmatrix},$$

so $\operatorname{rank}[A^2 B_1, AB_1, B_1] = 2$ and the system is not fully controllable. In fact, it is the angular momentum which cannot be controlled, as

$$\frac{d}{dt}(r^2\dot{\theta}^2) = (2\omega\rho, 0, \rho^2)^T \begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix}, \quad (2\omega\rho, 0, \rho^2)^T A^n B = 0, \quad n \geqslant 0.$$

10 Linear systems with non-negative quadratic costs

The general theory of dynamic optimization for non-negative costs specializes in a computationally explicit way in the case of linear systems with quadratic costs.

Consider the linear controllable dynamical system

$$f(x,a) = Ax + Ba, \quad x \in \mathbb{R}^d, \quad a \in \mathbb{R}^m,$$

with non-negative quadratic cost function

$$c(x,a) = x^T R x + x^T S^T a + a^T S x + a^T Q a,$$

where R is a $d \times d$ symmetric matrix, S is am $m \times d$ matrix and Q is an $m \times m$ symmetric matrix. We assume throughout that Q is positive-definite. We begin with some calculations regarding partial minimization of quadratic forms. Note that

$$\inf_a c(x,a) = c(x,Kx) = x^T(R - S^T Q^{-1} S)x,$$

where $K = -Q^{-1}S$. Thus the requirement that c be non-negative imposes the constraint that $R - S^T Q^{-1} S$ is non-negative definite. For a non-negative definite matrix Π, we can write

$$c(x,a) + f(x,a)^T \Pi f(x,a) = \tilde{c}(x,a) = x^T \tilde{R} x + x^T \tilde{S}^T a + a^T \tilde{S} x + a^T \tilde{Q} a,$$

where $\tilde{R} = R + A^T \Pi A$, $\tilde{S} = S + B^T \Pi A$ and $\tilde{Q} = Q + B^T \Pi B$. Since $B^T \Pi B$ is non-negative definite, $\tilde{Q}$ is positive-definite. Hence

$$\inf_a \{c(x,a) + f(x,a)^T \Pi f(x,a)\} = \tilde{c}(x, K(\Pi)x) = x^T r(\Pi) x, \tag{3}$$

where

$$K(\Pi) = -\tilde{Q}^{-1}\tilde{S}, \quad r(\Pi) = \tilde{R} - \tilde{S}^T \tilde{Q}^{-1} \tilde{S}.$$

Since the left-hand side of equation (3) is non-negative, $r(\Pi)$ must be non-negative definite. Fix now a non-negative definite matrix Π_0 and consider the n-horizon problem with final cost $c(x) = x^T \Pi_0 x$. Define, as usual, for $n \geqslant 0$,

$$V_n^u(x) = \sum_{k=0}^{n-1} c(x_k, u_k) + c(x_n), \quad V_n(x) = \inf_u V_n^u(x),$$

where $x_0 = x$ and $x_{k+1} = Ax_k + Bu_k, k \geqslant 0$. Then (see footnote 12) $V_0 = c$ and

$$V_{n+1}(x) = \inf_a \{c(x,a) + V_n(Ax + Ba)\}, \quad n \geqslant 0.$$

Hence we obtain the following result by using equation (3) and an induction on $n \geqslant 0$.

Proposition 10.1. *Define* $(\Pi_n)_{n \geqslant 0}$ *by the* Riccati recursion

$$\Pi_{n+1} = r(\Pi_n), \quad n \geqslant 0.$$

Then,

$$V_n(x) = x^T \Pi_n x$$

and the optimal sequence $(x_0, \ldots, x_n)$ *is given by*

$$x_k = \Gamma_{n-k} \ldots \Gamma_{n-1} x_0, \quad k = 0, 1, \ldots, n,$$

where $\Gamma_n = A + BK(\Pi_n)$ *is the* gain matrix.

We turn now to the infinite-horizon case. Define, as usual,

$$V^u(x) = \sum_{k=0}^{\infty} c(x_k, u_k), \quad V(x) = \inf_u V^u(x).$$

Note that, if f is fully controllable, we can choose u so that $x_k = 0$ and $u_k = 0$ for all $k \geqslant d$, so $V(x) < \infty$ for all $x \in \mathbb{R}^d$.

A matrix A is a (discrete-time) *stability matrix* if $A^n \to 0$ as $n \to \infty$. We call f *stabilizable* if $A + BK$ is a stability matrix for some K. We use the matrix norm $|A| = \sup\{|Ax| : |x| = 1\}$, for which $|Ax| \leqslant |A||x|$ for all $x \in \mathbb{R}^d$, $|A| = |A^T|$ and $|AB| \leqslant |A||B|$. Then A is a stability matrix if and only if $|A|^n \leqslant C\alpha^n$ for all $n \geqslant 0$, for some constants $C < \infty$ and $\alpha \in [0, 1)$.

Example. Suppose

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Then $f(x, a) = Ax + Ba$ is stabilized by $K = \begin{pmatrix} -2 & 0 \end{pmatrix}$, but f is not fully controllable.

Note that, if f is stabilized by K, and we set $u_n = Kx_n$, then $x_n = \Gamma^n x_0$, where $\Gamma = A + BK$. Choose $C < \infty$ and $\alpha < 1$ such that $|\Gamma^n| \leqslant C\alpha^n$ for all $n \geqslant 0$. Then, for all $x \in \mathbb{R}^d$,

$$V(x) \leqslant V^u(x) = x^T \sum_{n=0}^{\infty} (\Gamma^n)^T Q_K \Gamma^n x \leqslant C^2 |Q_K||x|^2/(1 - \alpha^2) < \infty,$$

where

$$Q_K = \begin{pmatrix} I \\ K \end{pmatrix}^T \begin{pmatrix} R & S^T \\ S & Q \end{pmatrix} \begin{pmatrix} I \\ K \end{pmatrix}.$$

Proposition 10.2. *Assume that f is fully controllable or stabilizable. Then the infimal cost function is given by*

$$V(x) = x^T \Pi x, \quad x \in \mathbb{R}^d,$$

where Π is the minimal non-negative definite solution to the equilibrium Riccati equation

$$\Pi = r(\Pi),$$

and, for $K = K(\Pi)$, $u(x) = Kx$ defines an optimal control. Moreover, if Q_K is positive-definite, in particular, if c is positive-definite, then $\Gamma = A + BK$ is a stability matrix, Π is the only non-negative definite solution to $\Pi = r(\Pi)$, and, for any non-negative definite matrix Π_0, if we define $\Pi_{n+1} = r(\Pi_n)$, for $n \geqslant 0$, then $\Pi_n \to \Pi$ as $n \to \infty$.

Proof. By Proposition 2.1,

$$V(x) = \inf_a \{c(x, a) + V(Ax + Ba)\}, \quad x \in \mathbb{R}^d.$$

Take $\Pi_0 = 0$ in the preceding proposition to obtain for the infimal cost function of the n-horizon problem with no final cost,

$$x^T \Pi_n x = V_n(x) \uparrow V_\infty(x) \leqslant V(x), \quad x \in \mathbb{R}^d.$$

Since f is fully controllable or stabilizable, $V(x) < \infty$ for all $x \in \mathbb{R}^d$. Hence[24] there is a non-negative definite matrix Π such that $V_\infty(x) = x^T \Pi x$ for all x. Since r is continuous, we can let $n \to \infty$ in $\Pi_{n+1} = r(\Pi_n)$ to obtain $\Pi = r(\Pi)$. Then

$$V_\infty(x) = \min_a \{c(x,a) + V_\infty(Ax + Ba)\}, \quad x \in \mathbb{R}^d,$$

with minimum at $a = u(x) = K(\Pi)x$. Then $V_\infty \geqslant V^u \geqslant V$ by the argument of Proposition 6.1, so $V(x) = x^T \Pi x$ and u is optimal. For $\Gamma = A + BK$, we have

$$\sum_{n=0}^{\infty} (\Gamma^n)^T Q_K \Gamma^n = \Pi < \infty,$$

so, if Q_K is positive-definite, then Γ is a stability matrix.

Consider the n-horizon problem with final cost $x^T \tilde{\Pi}_0 x$, where $\tilde{\Pi}_0$ is any non-negative definite matrix. The infimal cost function is $\tilde{V}_n(x) = x^T \tilde{\Pi}_n x$, where $\tilde{\Pi}_{n+1} = r(\tilde{\Pi}_n)$ for $n \geqslant 0$. Then

$$V_n(x) \leqslant \tilde{V}_n(x) \leqslant V_n^u(x) + x^T (\Gamma^n)^T \tilde{\Pi}_0 \Gamma^n x.$$

If $r(\tilde{\Pi}_0) = \tilde{\Pi}_0$, then we obtain $\Pi \leqslant \tilde{\Pi}_0$, so Π is the minimal non-negative solution. In the case where Q_K is positive-definite, for general $\tilde{\Pi}_0$, as $n \to \infty$, the final term tends to 0, so we obtain

$$x^T \Pi x \leqslant \lim_{n \to \infty} x^T \tilde{\Pi}_n x \leqslant x^T \Pi x, \quad x \in \mathbb{R}^d,$$

so $\tilde{\Pi}_n \to \Pi$. In particular Π is the only solution to $r(\Pi) = \Pi$. □

[24] Write $e_1, \ldots, e_d$ for the standard basis in $\mathbb{R}^d$, then $V_n(e_i \pm e_j)$ converges to a finite limit for all i, j, and so, by polarization, does $(\Pi_n)_{ij} = e_i^T \Pi_n e_j$. Denote the limit by Π_{ij}. Then $\Pi = (\Pi_{ij})$ is symmetric and $x^T \Pi_n x \to x^T \Pi x$ for all $x \in \mathbb{R}^d$.

11 Certainty-equivalent control

We show that the addition of noise to a linear system with quadratic costs does not change the optimal control, as a function of state.

Consider the realised stochastic controllable dynamical system $(G, (\varepsilon_n)_{n\geqslant 1})$, where

$$G(x, a, \varepsilon) = Ax + Ba + \varepsilon, \quad x \in \mathbb{R}^d, \quad a \in \mathbb{R}^m,$$

and where $(\varepsilon_n)_{n\geqslant 1}$ are independent $\mathbb{R}^d$-valued random variables, with mean $\mathbb{E}(\varepsilon) = 0$ and variance $\mathbb{E}(\varepsilon\varepsilon^T) = N$. Thus the controlled process, for a given starting point $X_0 = x$, is given by

$$X_{n+1} = AX_n + BU_n + \varepsilon_{n+1}$$

where $U_n = u_n(X_0, \ldots, X_n)$ is the control. We study the n-horizon problem with non-negative quadratic instantaneous costs $c(x, a)$ and final cost $c(x)$, as in the preceding section. Thus

$$c(x, a) = x^T Rx + x^T S^T a + a^T Sx + a^T Qa, \quad c(x) = x^T \Pi_0 x.$$

Set

$$V_n^u(x) = \mathbb{E}_x^u \left(\sum_{k=0}^{n-1} c(X_k, U_k) + c(X_n) \right), \quad V_n(x) = \inf_u V_n^u(x).$$

Suppose inductively that

$$V_n(x) = x^T \Pi_n x + \gamma_n.$$

This is true for $n = 0$ if we take $\gamma_0 = 0$. By a straightforward generalization[25] of Proposition 2.1, V_{n+1} is given by the optimality equation

$$V_{n+1}(x) = \inf_a \{c(x, a) + \mathbb{E}(V_n(Ax + Ba + \varepsilon))\}.$$

We have

$$\begin{aligned} \mathbb{E}(V_n(Ax + Ba + \varepsilon)) &= \mathbb{E}((Ax + Ba + \varepsilon)^T \Pi_n (Ax + Ba + \varepsilon)) + \gamma_n \\ &= (Ax + Ba)^T \Pi_n (Ax + Ba) + \mathbb{E}(\varepsilon^T \Pi_n \varepsilon) + \gamma_n \end{aligned}$$

and we showed in the preceding section that

$$\inf_a \{c(x, a) + (Ax + Ba)^T \Pi_n (Ax + Ba)\} = x^T r(\Pi_n) x,$$

with minimizing action $a = K(\Pi_n)$. Also

$$\mathbb{E}(\varepsilon^T \Pi_n \varepsilon) = \sum_{i,j} \mathbb{E}(\varepsilon_i (\Pi_n)_{ij} \varepsilon_j) = \sum_{i,j} \mathbb{E}(N_{ij} (\Pi_n)_{ij}) = \text{trace}(N\Pi_n).$$

So $V_{n+1}(x) = x^T \Pi_{n+1} x + \gamma_{n+1}$, where $\Pi_{n+1} = r(\Pi_n)$ and $\gamma_{n+1} = \gamma_n + \text{trace}(N\Pi_n)$. By induction, we have proved the following result.

[25] We have moved out of the setting of a countable state space used in Section 2. For a function F on $S \times A$, instead of writing PF as a sum, we can use the formula $PF(x, a) = \mathbb{E}(F(G(x, a, \varepsilon)))$.

Proposition 11.1. *For the linear system*

$$X_{n+1} = AX_n + BU_n + \varepsilon_{n+1},$$

with independent perturbations $(\varepsilon_n)_{n\geqslant 1}$*, having mean* 0 *and variance* N*, and with non-negative quadratic costs as above, the infimal cost function is given by*

$$V_n(x) = x^T \Pi_n x + \gamma_n$$

and the n*-horizon optimal control is* $U_k = K(\Pi_{n-1-k})X_k$*.*

This is *certainty-equivalent control* as the optimal control is the same as for $\varepsilon = 0$.

12 LQG systems and the Kalman filter

We introduce the LQG model and show how to reduce it to a stochastic controllable dynamical system using the Kalman filter. The LQG system is the system of equations

$$\begin{aligned} X_{n+1} &= AX_n + BU_n + \varepsilon_{n+1}, \\ Y_{n+1} &= CX_n + \eta_{n+1}, \qquad\qquad n \geqslant 0. \end{aligned}$$

Here A, B and C are given matrices and the random variables $X_0, \binom{\varepsilon_1}{\eta_1}, \binom{\varepsilon_2}{\eta_2}, \dots$ are independent Gaussians, X_0 having mean x and variance Σ_0 and, for $n \geqslant 1$, ε_n and η_n having mean 0 and

$$\operatorname{var}(\varepsilon_n) = N, \quad \operatorname{cov}(\varepsilon_n, \eta_n) = L, \quad \operatorname{var}(\eta_n) = M.$$

The *state* X_n takes values in $\mathbb{R}^d$, the *observation* Y_n takes values in $\mathbb{R}^p$ and the control values U_n are in $\mathbb{R}^m$. We complete the model by specifying a *control*, which is a function $u : (\mathbb{R}^p)^* \to \mathbb{R}^m$, and setting $U_n = u_n(Y_1, \dots, Y_n)$. We emphasise that what is different now is that we no longer observe the state, but have to estimate the state value on the basis of the observations. Set

$$V_n^u(x, \Sigma_0) = \mathbb{E}^u_{(x,\Sigma_0)} \left(\sum_{k=0}^{n-1} c(X_k, U_k) + c(X_n) \right), \quad V_n(x, \Sigma_0) = \inf_u V_n^u(x, \Sigma_0).$$

Lemma 12.1. *Let X and Y be jointly Gaussian, with mean 0 and with*

$$\operatorname{var}(X) = U, \quad \operatorname{cov}(X, Y) = W, \quad \operatorname{var}(Y) = V,$$

with V invertible. Set $\hat{X} = WV^{-1}Y$ and $Z = X - \hat{X}$. Then Z is independent of Y with

$$\operatorname{var}(Z) = U - WV^{-1}W^T.$$

Proof. Note that Y and Z are jointly Gaussian, so zero covariance will imply independence. We compute

$$\operatorname{cov}(Z, Y) = \operatorname{cov}(X, Y) - WV^{-1} \operatorname{var}(Y) = W - WV^{-1}V = 0$$

and

$$\operatorname{var}(Z) = \operatorname{cov}(Z, X) = \operatorname{var}(X) - WV^{-1} \operatorname{cov}(Y, X) = U - WV^{-1}W^T.$$

□

We now obtain a recursive scheme, called the *Kalman filter*, which determines for $n \geqslant 1$ the mean and variance of the conditional distribution of X_n, given the observations $Y_1, \dots, Y_n$. Suppose inductively that *we can write $X_n = \hat{X}_n + \Delta_n$, where $\hat{X}_n$ is a function of $Y_1, \dots, Y_n$, and where Δ_n is independent of $Y_1, \dots, Y_n$, with distribution $N(0, \Sigma_n)$.* This is true for $n = 0$, with $\hat{X}_0 = x$. We have

$$\begin{aligned} X_{n+1} &= A\hat{X}_n + BU_n + \xi_{n+1}, & \xi_{n+1} &= \varepsilon_{n+1} + A\Delta_n, \\ Y_{n+1} &= C\hat{X}_n + \zeta_{n+1}, & \zeta_{n+1} &= \eta_{n+1} + C\Delta_n. \end{aligned}$$

Note that the *innovations* ξ_{n+1} and ζ_{n+1} are zero-mean Gaussians and are independent of $Y_1, \ldots, Y_n$, with

$$\operatorname{var}(\xi_{n+1}) = \tilde{N} = N + A\Sigma_n A^T, \quad \operatorname{var}(\zeta_{n+1}) = \tilde{M} = M + C\Sigma_n C^T,$$

$$\operatorname{cov}(\xi_{n+1}, \zeta_{n+1}) = \tilde{L} = L + A\Sigma_n C^T.$$

Set

$$H_{n+1} = H(\Sigma_n) = \tilde{L}\tilde{M}^{-1}, \quad \Sigma_{n+1} = \sigma(\Sigma_n) = \tilde{N} - \tilde{L}\tilde{M}^{-1}\tilde{L}^T.$$

By the lemma, $\xi_{n+1} = \hat{\varepsilon}_{n+1} + \Delta_{n+1}$, where

$$\hat{\varepsilon}_{n+1} = H_{n+1}\zeta_{n+1} = H_{n+1}(Y_{n+1} - C\hat{X}_n)$$

and where Δ_{n+1} is independent of ζ_{n+1}, and hence of $Y_1, \ldots, Y_{n+1}$, with distribution $N(0, \Sigma_{n+1})$. Note that

$$\operatorname{var}(\hat{\varepsilon}_{n+1}) = H_{n+1} \operatorname{var}(\zeta_{n+1}) H_{n+1}^T = \tilde{L}\tilde{M}^{-1}\tilde{L}^T = \tilde{N} - \Sigma_{n+1} = N + A\Sigma_n A^T - \Sigma_{n+1}.$$

Now $X_{n+1} = \hat{X}_{n+1} + \Delta_{n+1}$, where

$$\hat{X}_{n+1} = A\hat{X}_n + BU_n + \hat{\varepsilon}_{n+1},$$

which is a function of $Y_1, \ldots, Y_{n+1}$, as required. This establishes the induction.

Note that

$$\begin{aligned}\mathbb{E}(c(X_k, U_k)) &= \mathbb{E}(c(\hat{X}_k + \Delta_k, U_k)) \\ &= \mathbb{E}(\Delta_k^T R \Delta_k) + \mathbb{E}(c(\hat{X}_k, U_k)) = \operatorname{trace}(R\Sigma_k) + \mathbb{E}(c(\hat{X}_k, U_k))\end{aligned}$$

and, similarly,

$$\mathbb{E}(c(X_n)) = \operatorname{trace}(\Pi_0 \Sigma_n) + \mathbb{E}(c(\hat{X}_n)).$$

Hence

$$V_n(x, \Sigma_0) = \hat{V}_n(x, \Sigma_0) + \sum_{k=0}^{n-1} \operatorname{trace}(R\Sigma_k) + \operatorname{trace}(\Pi_0 \Sigma_n),$$

where $\hat{V}_n$ is the infimal cost function of the stochastic controllable dynamical system

$$\hat{X}_{n+1} = A\hat{X}_n + BU_n + \hat{\varepsilon}_{n+1}, \quad \Sigma_{n+1} = \sigma(\Sigma_n),$$

where $(\hat{\varepsilon}_n)_{n \geqslant 1}$ are independent, and $\hat{\varepsilon}_{n+1}$ has distribution $N(0, \hat{N}(\Sigma_n))$, with

$$\hat{N}(\Sigma) = N + A\Sigma A^T - \sigma(\Sigma).$$

This system can be treated by a small variation of the method in the preceding section. In particular, certainty-equivalence holds: the optimal control for the n-horizon problem is $U_k = K(\Pi_{n-1-k})\hat{X}_k$. The product form of this control is remarkable as, given A, B and n, $K(\Pi_{n-1-k})$ depends only on the cost functions, whilst the controllable dynamical system for $\hat{X}_k$ is independent of the costs. This is called the *separation principle*.

Example. We investigate from first principles one of the simplest control problems with noisy observation. We shall follow the same lines as in the general theory and use similar notation. The system has scalar state and observations and is given by

$$X_{n+1} = X_n + U_n, \quad Y_{n+1} = X_{n+1} + \eta_{n+1}, \quad n \geqslant 0,$$

where the random variable X_0 and η_n, $n \geqslant 1$ are independent, with $X_0 \sim N(x, v)$ and $\eta_n \sim N(0, 1)$, for all n, and where $U_n = u_n(Y_1, \ldots, Y_n)$. We fix a time-horizon n and aim to choose u to minimize

$$V_n^u(x, v) = \mathbb{E}_{(x,v)}^u \left(\sum_{k=0}^{n-1} U_k^2 + DX_n^2 \right).$$

Consider first the control problem for $x_k = \mathbb{E}(X_k)$: we seek to minimize $\sum_{k=0}^{n-1} u_k^2 + Dx_n^2$ subject to $x_{k+1} = x_k + u_k$ and $x_0 = x$. The minimum is $Dx^2/(1 + Dn)$, achieved when $u_k = -Dx_k/(1 + D(n - k))$.

Next, we calculate the Kalman filter. We determine recursively for $n \geqslant 0$ a function $\hat{X}_n$ of $Y_1, \ldots, Y_n$ such that $X_n = \hat{X}_n + \Delta_n$, with Δ_n independent of $Y_1, \ldots, Y_n$. Write $v_n = \mathrm{var}(\Delta_n)$. For $n = 0$ we can take $\hat{X}_0 = x$ and $v_0 = v$. At the nth step, we write

$$\begin{aligned} X_{n+1} &= \hat{X}_n + U_n + \xi_{n+1}, & \xi_{n+1} &= \Delta_n, \\ Y_{n+1} &= \hat{X}_n + U_n + \zeta_{n+1}, & \zeta_{n+1} &= \Delta_n + \eta_{n+1}, \end{aligned}$$

where the innovations ξ_{n+1} and ζ_{n+1} are independent of $Y_1, \ldots, Y_n$. We aim to split

$$\xi_{n+1} = H_{n+1}\zeta_{n+1} + \Delta_{n+1},$$

where Δ_{n+1} is independent of ζ_{n+1} and hence of $Y_1, \ldots, Y_{n+1}$. On taking variances in this equation, we obtain

$$v_n = H_{n+1}^2(v_n + 1) + v_{n+1}.$$

On the other hand, by taking the covariance with ζ_{n+1}, we have

$$v_n = H_{n+1}(v_n + 1).$$

These equations imply that $H_{n+1} = v_{n+1}$ and determine a recursion $v_{n+1}^{-1} = 1 + v_n^{-1}$, so $v_n^{-1} = n + v_0^{-1}$, and so $v_n = v/(1 + vn)$.

Now $\hat{X}_{n+1} = \hat{X}_n + U_n + \hat{\varepsilon}_{n+1}$, where $\hat{\varepsilon}_{n+1} = H_{n+1}\zeta_{n+1}$, so

$$\mathrm{var}(\hat{\varepsilon}_{n+1}) = s_{n+1} = H_{n+1}^2(1 + v_n) = \left(\frac{v}{1 + (n+1)v} \right)^2 \left(1 + \frac{v}{1 + nv} \right).$$

By certainty-equivalence, the optimal control for the n-horizon problem is given by $U_k = -D\hat{X}_k/(1 + D(n - k))$, so

$$\hat{X}_{k+1} = \frac{1 + D(n - k - 1)}{1 + D(n - k)} \hat{X}_k + \hat{\varepsilon}_{n+1}.$$

On taking variances, we obtain the recursion

$$\operatorname{var}(\hat{X}_{k+1}) = \left(\frac{1 + D(n-k-1)}{1 + D(n-k)}\right)^2 \operatorname{var}(\hat{X}_k) + s_{n+1}.$$

Finally, the minimal expected cost is

$$\mathbb{E}^u_{(x,v)}\left(\sum_{k=0}^{n-1} U_k^2 + DX_n^2\right) = \sum_{k=0}^{n-1} \frac{D^2}{(1 + D(n-k))^2} \operatorname{var}(\hat{X}_k) + D \operatorname{var}(\hat{X}_n) + D \operatorname{var}(\Delta_n).$$

13 Observability

We introduce the notion of observability for deterministic linear systems.

Consider the system

$$x_{n+1} = Ax_n, \quad y_{n+1} = Cx_n, \quad n \geqslant 0.$$

Here the state x takes values in $\mathbb{R}^d$, the observation y takes values in $\mathbb{R}^p$, and A and C are matrices of appropriate dimensions. We say that the system is *observable in n-steps* if $y_1, \ldots, y_n$ determine uniquely the initial state x_0, for all $x_0 \in \mathbb{R}^d$. It is *observable* if it is observable in n-steps for some $n \geqslant 1$. Note that

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = N_n x_0, \quad N_n = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix},$$

so the system is observable in n-steps if and only if $\operatorname{rank}(N_n) = d$. Hence, by the Cayley–Hamilton theorem, the system is observable if and only if $\operatorname{rank}(N_d) = d$.

The continuous-time system

$$\dot{x}_t = Ax_t, \quad \dot{y}_t = Cx_t, \quad y_0 = 0, \quad t \geqslant 0$$

is *observable in time t* if $(y_s)_{0 \leqslant s \leqslant t}$ determines uniquely the initial state x_0, for all $x_0 \in \mathbb{R}^d$. Since

$$\left(\frac{d}{dt}\right)^n \Big|_{t=0} y_t = CA^{n-1}x_0, \quad n \geqslant 1,$$

it is clear that, for any $t > 0$, the condition $\operatorname{rank}(N_d) = d$ is sufficient for observability in time t. On the other hand, if $\operatorname{rank}(N_d) \leqslant d-1$, then there exists $x_0 \in \mathbb{R}^d \setminus \{0\}$ such that $CA^n x_0 = 0$ for $n = 0, 1, \ldots, d-1$, and hence for all n by Cayley–Hamilton. Hence

$$y_t = \int_0^t Ce^{sA}x_0 ds = 0$$

for all $t \geqslant 0$ and we cannot distinguish x_0 from 0. The condition $\operatorname{rank}(N_d) = d$ is thus equivalent to observability (in any time $t > 0$).

Example (The sum of two populations). Suppose $\dot{x}_t = \lambda x_t$, $\dot{z}_t = \mu z_t$ and we observe $y_t = x_t + z_t$. Can we determine x_0 and z_0?. In this case

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

So we can determine x_0 and z_0, provided $\lambda \neq \mu$. Even though we are provided with the extra information $x_0 + z_0$, if $\lambda = \mu$, we will have $y_t = (x_0 + z_0)e^{\lambda t}$, so we can never recover x_0 alone.

Example (Radioactive decay). Suppose atoms of element 1 can decay in two ways, to atoms of element 2 at rate α and to atoms of element 3 at rate β. Suppose that atoms of element 2 also decay to atoms of element 3, at rate γ. We observe the number of atoms of element 3. Can we determine the initial numbers of atoms of elements 1 and 2?

Here we have the system

$$\dot{x}_t^1 = -(\alpha+\beta)x_t^1, \quad x_t^2 = \alpha x_t^1 - \gamma x_t^2, \quad \dot{x}_t^3 = \beta x_t^1 + \gamma x_t^2,$$

so

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

Note that $\det N_3 = \gamma(\alpha+\beta)(\gamma-\beta)$, so the system is observable if $\gamma > 0, \alpha+\beta > 0$ and $\beta \neq \gamma$. It is easy to see that it is not so otherwise.

14 The LQG model in equilibrium

We show that full controllability of the $(A, B, .)$-system, together with observability of the $(A, ., C)$-system, is sufficient for the existence of an equilibrium control in the LQG model. We also discuss the optimal such control.

In Section 9 we showed that the $(A, B, .)$-system is fully controllable if and only if $\text{rank}(M_d) = d$. Also, by Proposition 10.2, this condition implies that the $(A, B, .)$-system is stabilizable, that is, there exists a matrix K such that $|A + BK| < 1$.

In the previous section we saw that the $(A, ., C)$-system is observable if and only if $\text{rank}(N_d) = d$. Now

$$N_d^T = \begin{pmatrix} C^T & C^T A^T & \dots & C^T (A^T)^{d-1} \end{pmatrix},$$

so, by comparing with the form of M_d, we deduce that observability implies the existence of a matrix H such that $|A - HC| = |A^T - C^T H^T| < 1$. We call this last condition *asymptotic observability*.

In the remainder of this section, we consider the LQG model

$$\begin{aligned} X_{n+1} &= AX_n + BU_n + \varepsilon_{n+1}, \\ Y_{n+1} &= CX_n + \eta_{n+1}, \qquad\qquad n \geqslant 0, \end{aligned}$$

as in Section 12, and we assume stability and asymptotic observability, that is, the existence of K and H such that $|A + BK| < 1$ and $|A - HC| < 1$. We assume also that both the instantaneous costs and the noise are non-degenerate, that is to say, the matrices

$$\begin{pmatrix} R & S \\ S^T & Q \end{pmatrix}, \quad \begin{pmatrix} N & L^T \\ L & M \end{pmatrix}$$

are both positive-definite.

Theorem 14.1. *Under the above assumptions, the equations $\Pi = r(\Pi)$ and $\Sigma = \sigma(\Sigma)$ both have unique solutions in the set of non-negative-definite matrices. Set $H = H(\Sigma)$ and $K = K(\Pi)$. Define recursively $\hat{X}_0 = 0$ and*

$$U_n = K\hat{X}_n, \quad \hat{X}_{n+1} = A\hat{X}_n + BU_n + H(Y_{n+1} - C\hat{X}_n), \quad n \geqslant 0.$$

Then the long-run average expected cost is given by

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E} \sum_{k=0}^{n-1} c(X_k, U_k) = \text{trace}(R\Sigma) + \text{trace}(\hat{N}\Pi),$$

where $\hat{N} = N + A\Sigma A^T - \Sigma$. Moreover our choice of H and K minimizes the long-run average expected cost.

Outline proof. We showed existence and uniqueness of Π in Proposition 10.2. The existence and uniqueness of Σ can be deduced by comparing the forms of the equations $\Pi = r(\Pi)$ and $\Sigma = \sigma(\Sigma)$. From Section 12, the minimal long-run average expected cost from Δ is $\text{trace}(R\Sigma)$. From Section 11, the minimal long-run average expected cost of the controllable dynamical system for $\hat{X}$ is $\text{trace}(\hat{N}\Pi)$, using control K. □

15 The Hamilton–Jacobi–Bellman equation

We begin a study of deterministic continuous-time controllable dynamical systems with a heuristic derivation of the Hamilton–Jacobi–Bellman equation. Then we prove that any suitably well-behaved solution of this equation must coincide with the infimal cost function and that the minimizing action gives an optimal control.

Recall from Subsection 1.3 that a continuous-time controllable dynamical system is a map

$$b : \mathbb{R}^+ \times \mathbb{R}^d \times A \to \mathbb{R}^d.$$

We now assume that the action-space A is a subset of $\mathbb{R}^p$ for some p, in examples A is often simply an interval in $\mathbb{R}$. We assume also that b is continuous, and is differentiable in x with bounded derivative. A control is a map $u : \mathbb{R}^+ \to A$. Given a control u and a starting time and state (s, x), we define[26] the controlled path $(x_t)_{t \geqslant s}$ as the solution of the differential equation

$$\dot{x}_t = b(t, x_t, u_t), \quad t \geqslant s, \quad x_s = x.$$

We shall consider two types of optimization problem. In the first type, we fix a *stopping set* $D \subseteq \mathbb{R}^d$ and a *time-horizon* $T < \infty$ and specify continuous and bounded *cost functions*[27]

$$c : [0, T) \times \mathbb{R}^d \times A \to \mathbb{R}, \quad C : \{T\} \times D \to \mathbb{R}.$$

We say that a control u is *feasible*, starting from (s, x), if, for the associated controlled path starting from (s, x), we have $x_T \in D$. If there is no such control, then we say (s, x) is *infeasible*. In the second type of problem, we also fix a stopping set $D \subseteq \mathbb{R}^d$, which is the boundary of some open set $S \subseteq \mathbb{R}^d$, but the time of arrival in D is *unconstrained*. We specify continuous and bounded cost functions

$$c : \mathbb{R}^+ \times S \times A \to \mathbb{R}, \quad C : \mathbb{R}^+ \times D \to \mathbb{R}.$$

We say that a control u is *feasible*, starting from (s, x), if $\tau < \infty$, where

$$\tau = \inf\{t \geqslant 0 : x_t \in D\}.$$

In order to give a unified treatment of the two cases, we shall, in the first case, set $\tau = T$ and write $\tilde{S} = ([0, T) \times \mathbb{R}^d)$ and $\tilde{D} = \{T\} \times D$. In the the second case, we write $\tilde{S} = \mathbb{R}^+ \times S$ and $\tilde{D} = \mathbb{R}^+ \times D$.

The *total cost* for a feasible control u, starting from $(s, x) \in \tilde{S}$, is defined by

$$V^u(s, x) = \int_s^\tau c(t, x_t, u_t)dt + C(\tau, x_\tau).$$

The *infimal cost function* V is defined by

$$V(s, x) = \inf_u V^u(s, x),$$

[26] The basic theory of existence and uniqueness for solutions of differential equations is reviewed, and its application in this setting is explained, in Section 18.

[27] As usual, any problem of maximizing rewards can be treated as a problem of minimizing negative costs, so we do not discuss the theory for this sort of problem separately.

where the infimum is taken over all continuous feasible controls starting from (s, x), and $V(s, x) = \infty$ if there are no such controls.

Suppose we start from $(t, x) \in \tilde{S}$ and choose action a until a short time later $t + \delta$, then switching to an optimal control. Comparing this control with the optimal control from (t, x), we obtain, up to terms which are small compared to δ,

$$V(t, x) \leqslant c(t, x, a)\delta + V(t + \delta, x + b(t, x, a)\delta)$$

On the other hand, by optimizing the right-hand side over a we might expect to get arbitrarily close to $V(t, x)$. We expand to first order

$$V(t + \delta, x + b(t, x, a)\delta) = V(t, x) + \dot{V}(t, x)\delta + \nabla V(t, x)b(t, x, a)\delta + O(\delta^2).$$

On substituting this in the inequality, rearranging, dividing by δ and letting $\delta \to 0$, we obtain

$$\inf_a \{c(t, x, a) + \dot{V}(t, x) + \nabla V(t, x)b(t, x, a)\} = 0, \quad (t, x) \in \tilde{S}.$$

This is called the *Hamilton–Jacobi–Bellman equation*. It is the optimality equation for continuous-time systems. The final cost C provides a boundary condition $V = C$ on $\tilde{D}$.

Proposition 15.1. *Suppose that there exists a function $F : \tilde{S} \cup \tilde{D} \to \mathbb{R}$, differentiable with continuous derivative, and that, for a given starting point $(s, x) \in \tilde{S}$, there exists a continuous feasible control u^* such that*

$$c(t, x, a) + \dot{F}(t, x) + \nabla F(t, x)b(t, x, a) \geqslant 0$$

for all $(t, x) \in \tilde{S}$ and $a \in A$, with equality when $t \in [s, \tau^*)$ and $(x, a) = (x_t^*, u_t^*)$. Suppose also that $F = C$ on $\tilde{D}$. Then $F(s, x) = V(s, x)$ and u^* defines an optimal control starting from (s, x).

Proof. It will suffice to consider the case $s = 0$. Fix any continuous feasible control $u : \mathbb{R}^+ \to A$ and set

$$m_t = \int_0^t c(s, x_s, u_s)ds + F(t, x_t), \quad 0 \leqslant t \leqslant \tau.$$

Then m is continuous on $[0, \tau]$ and differentiable on $[0, \tau)$, with

$$\dot{m}_t = c(t, x_t, u_t) + \dot{F}(t, x_t) + \nabla F(t, x_t)b(t, x_t, u_t) \geqslant 0,$$

and with equality if $u = u^*$. Therefore

$$F(0, x) = m_0 \leqslant m_\tau = \int_0^\tau c(s, x_s, u_s)ds + C(\tau, x_\tau) = V^u(0, x),$$

with equality if $u = u^*$. □

The proposition sets up a possible way to calculate the infimal cost function and to find an optimal control. One tries to solve the Hamilton–Jacobi–Bellman equation

$$\inf_a \{c(t, x, a) + \dot{V}(t, x) + \nabla V(t, x)b(t, x, a)\} = 0, \quad (t, x) \in \tilde{S},$$

and to identify, for each $(t,x) \in \tilde{S}$ a minimizing action $u(t,x)$. Then, given a starting point $(s,x) \in \tilde{S}$, we attempt to solve the differential equation $\dot{x}_t = b^u(t,x_t)$, where $b^u(t,x) = b(t,x,u(t,x))$ and check that $\tau < \infty$ and $x_\tau \in D$. The control $u_t^* = u(t,x_t)$ then has $(x_t)_{s \leqslant t \leqslant \tau}$ as its controlled process starting from (s,x), so u^* has the minimizing property required by the proposition. In this case, we say that the function u *defines a feasible control for starting point* (s,x). It is often the case that the minimizing function $u(t,x)$ depends *discontinuously* but *piecewise continuously* on (t,x), and so do the associated controls. It is not hard to extend the proposition to this case, though we will not give details. In practice, the main hope to solve the HJB equation is to guess its shape as a function of x, to find the minimizing action $u(t,x)$ explicitly, and thereby to reduce the problem to a differential equation in t. These steps are illustrated in the next two examples.

Example (Linear system with quadratic costs). Consider the linear system with state-space $\mathbb{R}^d$ and action-space $\mathbb{R}^p$ given by $b(x,a) = Ax + Ba$, where A and B are matrices of appropriate dimensions. Take as cost function the non-negative quadratic function $c(x,a) = x^T R x + a^T Q a$, which we shall assume to vanish only if $x = 0$ and $a = 0$. Suppose the final cost is also quadratic and non-negative, thus $C(x) = x^T \Pi(T) x$, for some matrix $\Pi(T)$.

As in the discrete-time case, let us try in the HJB equation a solution of the form $V(t,x) = x^T \Pi(t) x$, for some non-negative definite matrices $\Pi(t)$. We have

$$\begin{aligned}&\inf_a \{c(x,a) + \dot{V}(t,x) + \nabla V(t,x) b(x,a)\} \\ &\quad = \inf_a \{x^T(R + \Pi A + A^T\Pi + \dot{\Pi})x + x^T \Pi B a + a^T B^T \Pi x + a^T Q a\} = x^T(\tilde{R} - \tilde{S}^T Q^{-1} \tilde{S})x\end{aligned}$$

at $a = -Q^{-1}\tilde{S}x$, where $\tilde{R} = R + \Pi A + A^T \Pi + \dot{\Pi}$ and $\tilde{S} = B^t \Pi$. (See Section 10.) Hence V is a solution if and only if $(\Pi(t))_{0 \leqslant t \leqslant T}$ satisfies the continuous-time *Riccati equation*

$$\dot{\Pi} + R + \Pi A + A^T \Pi - \Pi B Q^{-1} B^T \Pi = 0.$$

Example (Managing investment income). The following may be considered as a model for optimizing utility from investment income over a prescribed lifetime T. We seek to maximize

$$\int_0^T e^{-\alpha s} \sqrt{u_s} ds$$

subject to $\dot{x}_t = \beta x_t - u_t$ and $x_t \geqslant 0$ for all $0 \leqslant t \leqslant T$. Thus α is the personal discount rate, β is the rate of interest, and $\sqrt{u}$ is the utility gained from income at rate u.

The optimality equation is

$$\sup_a \{e^{-\alpha t}\sqrt{a} + \dot{V}(t,x) + (\beta x - a) V'(t,x)\} = 0, \quad 0 \leqslant t \leqslant T,$$

with boundary condition $V(T,x) = 0$. By scaling, the maximal reward function must have the form

$$V(t,x) = e^{-\alpha t}\sqrt{v(t)x}$$

for some function v. By substitution in the optimality equation we obtain $\dot{v} - (2\alpha - \beta)v + 1 = 0$ with maximizing action $a = x/v(t)$. Hence

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

and the optimal control is $u_t = x_t/v(t)$.

A short-cut is available for this example using the Cauchy-Schwarz inequality. We have the constraint

$$0 = e^{-\beta T} x_T = x_0 - \int_0^T e^{-\beta s} u_s ds.$$

By Cauchy-Schwarz,

$$\int_0^T e^{-\alpha s} \sqrt{u_s} ds = \int_0^T (e^{-\beta s} u_s)^{1/2} (e^{-(2\alpha-\beta)s})^{1/2} ds \leqslant \sqrt{x_0} \left(\int_0^T e^{-(2\alpha-\beta)s} ds \right)^{1/2},$$

which confirms our calculation of $V(0, x_0)$.

16 Pontryagin's maximum principle

This is a powerful method for the computation of optimal controls, which has the crucial advantage that it does not require prior evaluation of the infimal cost function. We describe the method and illustrate its use in three examples. We also give two derivations of the principle, one in a special case under impractically strong conditions, and the other, at a heuristic level only, as an analogue of the method of Lagrange multipliers for constrained optimization.

We continue with the set-up of the preceding section but assume from now on that b, c and C are differentiable in t and x with continuous derivatives, and that the stopping set D is a hyperplane, thus $D = \{y\} + \Sigma$ for some $y \in \mathbb{R}^d$ and some vector subspace Σ of $\mathbb{R}^d$. Define for $\lambda \in \mathbb{R}^d$ the *Hamiltonian*

$$H(t, x, u, \lambda) = \lambda^T b(t, x, u) - c(t, x, u).$$

Pontryagin's maximum principle states that, if $(x_t, u_t)_{t \leqslant \tau}$ is optimal, then there exist *adjoint paths* $(\lambda_t)_{t \leqslant \tau}$ in $\mathbb{R}^d$ and $(\mu_t)_{t \leqslant \tau}$ in $\mathbb{R}$ with the following properties: for all $t \leqslant \tau$,

(i) $H(t, x_t, u, \lambda_t) + \mu_t$ has maximum value 0, achieved at $u = u_t$,

(ii) $\dot{\lambda}_t^T = -\lambda_t^T \nabla b(t, x_t, u_t) + \nabla c(t, x_t, u_t)$,

(iii) $\dot{\mu}_t = -\lambda_t^T \dot{b}(t, x_t, u_t) + \dot{c}(t, x_t, u_t)$,

(iv) $\dot{x}_t = b(t, x_t, u_t)$.

Moreover the following *transversality conditions* hold[28]:

(v) $(\lambda_\tau^T + \nabla C(\tau, x_\tau))\sigma = 0$ for all $\sigma \in \Sigma$,

and, in the time-unconstrained case,

(vi) $\mu_\tau + \dot{C}(\tau, x_\tau) = 0$.

Note that, in the time-unconstrained case, if b, c and C are time-independent, then $\mu_t = 0$ for all t.

The Hamiltonian serves as a way of remembering the first four statements, which could be expressed alternatively as

$$\text{(i) } 0 = \partial H/\partial u, \quad \text{(ii) } \dot{\lambda} = -\partial H/\partial x, \quad \text{(iii) } \dot{\mu} = -\partial H/\partial t, \quad \text{(iv) } \dot{x} = \partial H/\partial \lambda.$$

Beware that the reformulation of (i) is not always correct, for example in cases where the set of actions is an interval and where the maximum is achieved at an endpoint.

Example (Bringing a particle to rest in minimal time). Suppose we can apply a force to a particle, moving on a line, which imparts to it an acceleration a with $|a| \leqslant 1$ in the chosen units. For a given initial position q_0 and velocity p_0, how can we bring the particle to rest at the origin in the shortest time?

[28] Subject to the avoidance of certain pathological behaviour.

Take state $x = (q, p)$ and adjoint variable $\lambda = (\alpha, \beta)$. The problem is time-independent, so there is no need to consider μ. We have $\dot{q}_t = p_t$ and $\dot{p}_t = u_t$, with $|u_t| \leqslant 1$. We seek to minimize $\tau = \inf\{t \geqslant 0 : q_t = p_t = 0\} = \int_0^\tau 1dt$. So take $c = 1, C = 0$ and $D = \{(0, 0)\}$. The Hamiltonian is

$$H = \alpha p + \beta u - 1$$

so $u_t^* = \text{sgn}(\beta_t)$ and $\alpha_t p_t + |\beta_t| = 1$. The adjoint equations are

$$\dot{\alpha}_t = -\partial H/\partial q = 0, \quad \dot{\beta}_t = -\partial H/\partial p = -\alpha_t.$$

So α is a constant and $\beta_t = \beta_\tau + \alpha s$, where $s = \tau - t$ is the time-to-go. Since $p_\tau = 0$, we must have $\beta_\tau = \pm 1$. There remains the problem of determining the values of α and β_τ as a function of (q_0, p_0). We do this backwards.

Suppose $\beta_\tau = 1$ and $\alpha \geqslant 0$, then $\beta_t \geqslant 0$ for all $t \leqslant \tau$, so $u_t = 1$, $p_t = -s$ and $q_t = s^2/2 = p_t^2/2$. On the other hand, if $\beta_\tau = 1$ and $\alpha < 0$, then the preceding calculation applies only for $s \leqslant s_0 = 1/|\alpha|$; once $s > s_0$, we have $\beta_t < 0$, so $u_t = -1$, and integrating the equations of motion back from s_0, we get $p_t = s - 2s_0$ and $q_t = 2s_0 s - s^2/2 - s_0^2$. Similar calculations apply for $\beta_\tau = -1$.

Thus we find there is a *switching locus* given by $q = -sgn(p)p^2/2$. Each initial state (q_0, p_0) above the locus lies on a unique parabola $q = -p^2/2 + c$, with $c > 0$. The optimal control is initially to take $a = -1$, thereby moving round the parabola to hit the switching locus. On hitting the locus, the acceleration acceleration changes sign, bringing the particle to rest at the origin by moving along the locus.

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, then she realises a unit price $p(u) = 1 - u/2$ for $0 \leqslant u \leqslant 2$ and $p(u) = 0$ for $u \geqslant 2$. She holds amount x at time 0. What is her maximal total discounted return

$$\int_0^\infty e^{-\alpha t} u_t p(u_t) dt$$

and how should she achieve it?

The current stock evolves by $\dot{x}_t = -u_t$. Set $\tau = \inf\{t \geqslant 0 : x_t = 0\}$. Note that the rewards from any two controls which agree on $[0, n]$ can differ by at most $\int_n^\infty e^{-\alpha t} dt = e^{-\alpha n}/\alpha$ so it will suffice to find an optimal control among those for which $\tau < \infty$. So let us restrict now to such controls. We take $A = [0, \infty)$, $c = -e^{-\alpha t} up(u)$, $C = 0$ and $D = \{0\}$. The Hamiltonian is

$$H = -\lambda u + e^{-\alpha t} up(u),$$

which is maximized to a positive value at $u = 1 - \lambda e^{\alpha t}$, provided this is positive, and to 0 at 0 otherwise. The adjoint equation $\dot{\lambda}_t = -\partial H/\partial x = 0$ shows that λ is a constant, and the transversality condition $\mu_\tau = \dot{C} = 0$ shows that H is maximized to 0 at τ, so $u_\tau = 0$, and so $\lambda_\tau = e^{-\alpha\tau}$. Now

$$x = \int_0^\tau u_t dt = \int_0^\tau (1 - e^{-\alpha(\tau - t)}) dt = \tau - (1 - e^{-\alpha\tau})/\alpha.$$

This equation is satisfied by a unique $\tau \in (0, \infty)$, though we cannot solve it explicitly, and then the optimal control is $u_t = 1 - e^{-\alpha(\tau - t)}$. Finally, the maximal reward is

$$V(x) = \int_0^\infty e^{-\alpha t} u_t p(u_t) dt = \frac{(1 - e^{-\alpha\tau})^2}{2\alpha}.$$

Example (Insect optimization). A colony of insects consists of workers and queens, numbering w_t and q_t at time t. If a proportion u_t of the workers' effort at time t is devoted to producing more workers, then the numbers evolve according to the differential equations

$$\dot{w}_t = a u_t w_t - b w_t, \quad \dot{q}_t = (1 - u_t) w_t,$$

where a, b are positive constants, with $a > b$. How should the workers behave to maximize the number of queens produced by the end of the season?

Write T for the length of the season and take as state the number of workers. Then $c = -(1 - u)w, C = 0$ and $D = \mathbb{R}$. The Hamiltonian is

$$H = \lambda(au - b)w + (1 - u)w = \begin{cases} (1 - \lambda b)w, & \text{if } u = 0, \\ (\lambda a - \lambda b)w, & \text{if } u = 1. \end{cases}$$

So $u_t = 0$ if $\lambda_t a < 1$ and $u_t = 1$ if $\lambda_t a \geqslant 1$. The adjoint equations are

$$\dot{\lambda}_t = -\partial H / \partial w = \begin{cases} \lambda_t b - 1, & \text{if } u_t = 0, \\ -\lambda_t (a - b), & \text{if } u_t = 1. \end{cases}$$

Hence, for small time-to-go $s = T - t$, we have $u_t = 0$, so $\lambda_t = (1 - e^{-bs})/b$. We switch to $u_t = 1$ when $a(1 - e^{-bs})/b = 1$, that is, at

$$s_0 = \frac{1}{b} \log \left(\frac{a}{a - b} \right).$$

There is only one switch because $\dot{\lambda}_t$ is always negative. Hence, regardless of the length of the season, the workers should produce only more workers until there is s_0 time to go, when they should all switch to making queens.

A heuristic derivation of Pontryagin's maximum principle can be made by analogy with the method of Lagrange multipliers for constrained optimization problems. Recall that to maximize $f(x)$ subject to a d-dimensional constraint $g(x) = b$, one introduces the *Lagrangian*

$$L(x, \lambda) = f(x) - \lambda^T (g(x) - b),$$

where $\lambda \in \mathbb{R}^d$. For each λ, we seek $x(\lambda)$ to maximize $L(x, \lambda)$ and then seek λ so that $g(x(\lambda)) = b$. Then $x(\lambda)$ is the desired maximizer. Now, suppose we wish to maximize

$$-\int_0^T c(x_t, u_t) dt - C(x_T)$$

subject to $\dot{x}_t = b(x_t, u_t)$. We might try to maximize for each path $(\lambda_t)_{t \leqslant T}$

$$\begin{aligned} L(x, \lambda) &= \int_0^T \{-c(x_t, u_t) - \lambda_t^T(\dot{x}_t - b(x_t, u_t))\}dt - C(x_T) \\ &= -\lambda_T^T x_T + \lambda_0^T x_0 + \int_0^T \{\dot{\lambda}_t^T x_t + \lambda_t^T b(x_t, u_t) - c(x_t, u_t)\}dt - C(x_T). \end{aligned}$$

Then to maximize over x we might set

$$0 = \partial L/\partial x_t = \dot{\lambda}_t^T + \lambda_t^T \nabla b(x_t, u_t) - \nabla c(x_t, u_t),$$

which is the adjoint equation, and, in permitted directions,

$$0 = \partial L/\partial x_T = -\lambda_T^T - \nabla C(x_T),$$

which is the transversality condition.

The following result establishes the validity of Pontryagin's maximum principle, subject to the existence of a twice continuously differentiable solution to the Hamilton-Jacobi-Bellman equation, with well-behaved minimizing actions. These hypotheses are unnecessarily strong and are too strong for many applications. A proof of the principle under weaker hypotheses lies beyond the scope of this course. We assume that the action space A is an open subset in $\mathbb{R}^p$ and that b and the cost functions c and C are continuously differentiable.

Proposition 16.1. *Suppose that there exists a function $F : \tilde{S} \cup \tilde{D} \to \mathbb{R}$, twice differentiable with continuous derivatives, and a function $u : \tilde{S} \to A$ such that*

$$c(t, x, a) + \dot{F}(t, x) + \nabla F(t, x) b(t, x, a) \geqslant 0$$

for all $a \in A$, with equality when $a = u(t, x)$, for all $(t, x) \in \tilde{S}$. Suppose also that $F = C$ on $\tilde{D}$. Fix a starting point $(0, x)$ and assume that u defines a continuous feasible control and controlled path $(u_t, x_t)_{t \leqslant \tau}$ starting from $(0, x)$. Set $\mu_t = -\dot{F}(t, x_t)$ and $\lambda_t^T = -\nabla F(t, x_t)$, then

$$\begin{aligned} \dot{\lambda}_t^T &= -\lambda_t^T \nabla b(t, x_t, u_t) + \nabla c(t, x_t, u_t), \\ \dot{\mu}_t &= -\lambda_t^T \dot{b}(t, x_t, u_t) + \dot{c}(t, x_t, u_t), \end{aligned}$$

and, for any $\sigma \in \Sigma$, we have

$$(\lambda_\tau^T + \nabla C)(\tau, x_\tau))\sigma = 0,$$

and, in the time-unconstrained case,

$$\mu_\tau + \dot{C}(\tau, x_\tau) = 0.$$

Proof. Define, for $(t, x) \in \tilde{S}$ and $a \in A$,

$$J(t, x, a) = c(t, x, a) + \dot{F}(t, x) + \nabla F(t, x) b(t, x, a).$$

Then $J(t,x,a) \geqslant 0$ and $J(t,x,u(t,x)) = 0$ so, since A is open, we have

$$(\partial J/\partial a)(t,x,u(t,x)) = 0,$$

and hence

$$0 = (\partial/\partial x)J(t,x,u(t,x)) = \nabla J(t,x,u(t,x)), \quad 0 = (\partial/\partial t)J(t,x,u(t,x)) = \dot{J}(t,x,u(t,x)).$$

Write $a = u(t,x)$, then

$$0 = \nabla J(t,x,a) = \nabla c(t,x,a) + \nabla F(t,x)\nabla b(t,x,a) + \{\nabla \dot{F}(t,x) + \nabla^2 F(t,x)b(t,x,a)\}$$

and

$$0 = \dot{J}(t,x,a) = \dot{c}(t,x,a) + \nabla F(t,x)\dot{b}(t,x,a) + \{\nabla \dot{F}(t,x)b(t,x,a) + \ddot{F}(t,x)\}.$$

Hence

$$\begin{aligned}\dot{\lambda}_t^T &= -\nabla \dot{F}(t,x_t) - \nabla^2 F(t,x_t)b(t,x_t,u_t) \\ &= \nabla c(t,x_t,u_t) + \nabla F(t,x_t)\nabla b(t,x_t,u_t) = \nabla c(t,x_t,u_t) - \lambda_t^T \nabla b(t,x_t,u_t)\end{aligned}$$

and

$$\begin{aligned}\dot{\mu}_t &= -\ddot{F}(t,x_t) - \nabla \dot{F}(t,x_t)b(t,x_t,u_t) \\ &= \dot{c}(t,x,u_t) + \nabla F(t,x_t)\dot{b}(t,x_t,u_t) = \dot{c}(t,x,u_t) - \lambda_t^T \dot{b}(t,x_t,u_t).\end{aligned}$$

On differentiating the equality $F = C$ at (τ, x_τ) in the direction σ , we obtain

$$(\lambda_\tau^T + \nabla C)(\tau,x_\tau))\sigma = 0,$$

and, in the time-unconstrained case, we can differentiate at (τ, x_τ) in t to obtain

$$\mu_\tau + \dot{C}(\tau,x_\tau) = 0.$$

□

17 Continuous-time stochastic systems

The discussion in this section will not be rigorous. A stochastic controllable dynamical system of jump type is given by a function

$$q : \mathbb{R}^+ \times \{(x, y) \in S \times S : x \neq y\} \times A \to \mathbb{R}^+.$$

We assume that the state-space S is countable. We write $q_{xy}(t, a) = q(t, x, y, a)$. For x, y distinct, $q_{xy}(t, a)$ gives the rate of jumping from x to y when at time t we choose action a. It is convenient to write

$$q_{xx}(t, a) = -\sum_{y \neq x} q_{xy}(t, a).$$

We consider Markov controls $u : \mathbb{R}^+ \times S \to A$ and set

$$q^u_{xy}(t) = q_{xy}(t, u(t, x)).$$

Then the controlled process $(X_t)_{t \geqslant s}$ for control u, starting from (s, x_s) satisfies $X_s = x_s$ and, for all $t \geqslant s$ and $x \in S$, conditional on $X_t = x$, as $\delta \downarrow 0$,

$$X_{t+\delta} = \begin{cases} x, & \text{with probability } 1 + q^u_{xx}(t)\delta + o(\delta), \\ y, & \text{with probability } q^u_{xy}(t)\delta + o(\delta), \text{ for all } y \neq x. \end{cases}$$

We consider the same sorts of control problem as in Section 15, where now we take an expectation in defining the cost functions

$$V^u(s, x) = \mathbb{E}^u_{(s,x)} \left(\int_s^\tau c(X_t, U_t)dt + C(X_\tau) \right), \quad V(s, x) = \inf_u V^u(s, x).$$

We now give a derivation of the optimality equation for V. Suppose we start at (t, x) and choose action a until time $t + \delta$, then switch to an optimal control. On comparing the resulting expected total cost with that of an optimal control from the outset, we obtain

$$V(t, x) \leqslant c(x, a)\delta + \mathbb{E}(V(t + \delta, X_{t+\delta})|X_t = x).$$

Now expand to first order in δ

$$\begin{aligned} \mathbb{E}(V(t + \delta, X_{t+\delta})|X_t = x) &= V(t + \delta, x)(1 + q_{xx}(t, a)\delta) + \sum_{y \neq x} V(t + \delta, y)q_{xy}(t, a)\delta + o(\delta) \\ &= V(t, x) + \dot{V}(t, x)\delta + \sum_{y \in S} q_{xy}(t, a)V(t, y)\delta + o(\delta). \end{aligned}$$

So

$$0 \leqslant \{c(x, a) + \dot{V}(t, x) + QV(t, x, a)\}\delta + o(\delta),$$

with equality if a is chosen optimally, where

$$QV(t, x, a) = \sum_{y \in S} q_{xy}(t, a)V(t, y).$$

Thus we obtain the optimality equation

$$\inf_a \{c(x,a) + \dot{V}(t,x) + QV(t,x,a)\} = 0$$

and we expect to find the optimal control as the minimizing action a.

Now we shall give an analogous discussion in the case of a diffusive stochastic controllable dynamical system. We specify two functions

$$\sigma, b : \mathbb{R}^+ \times \mathbb{R} \times A \to \mathbb{R}.$$

The function σ^2 is the *diffusivity* and determines the size of the stochastic fluctuations or *noise* in the dynamics. The function b is the *drift* and determines the average velocity. Given a choice of Markov control $u : \mathbb{R}^+ \times \mathbb{R} \to A$, set $\sigma^u(t,x) = \sigma(t,x,u(t,x))$ and $b^u(t,x) = b(t,x,u(t,x))$. The corresponding dynamics can be described infinitesimally[29], conditional on $X_t = x$, by

$$X_{t+\delta} = x + \sigma^u(t,x)\Delta + b^u(t,x)\delta + o(\delta),$$

as $\delta \to 0$, where $\mathbb{E}(\Delta) = 0$ and $\mathbb{E}(\Delta^2) = \delta$. We define cost functions V^u and V exactly as in the jump case.

Let us now derive the optimality equation for V. Suppose we start at (t,x) and choose action a until time $t+\delta$, then switch to an optimal control. On comparing the resulting expected total cost with that of an optimal control from the outset, we obtain

$$V(t,x) \leqslant c(x,a)\delta + \mathbb{E}\left(V(t+\delta, x + \sigma^u(t,x)\Delta + b^u(t,x)\delta + o(\delta))\right).$$

We expand to fisrt order in δ

$$\begin{aligned}
&\mathbb{E}\left(V(t+\delta, x + \sigma^u(t,x)\Delta + b^u(t,x)\delta + o(\delta))\right) \\
&\quad = V(t,x) + \dot{V}(t,x)\delta + V'(t,x)(b(t,x,a)\delta + \sigma(t,x,a)\Delta) + \frac{1}{2}V''(t,x)\sigma(t,x,a)^2\Delta^2 + o(\delta).
\end{aligned}$$

So

$$0 \leqslant \{c(x,a) + \dot{V}(t,x) + LV(t,x,a)\}\delta + o(\delta)$$

with equality if a is chosen optimally, where

$$LV(t,x,a) = \frac{1}{2}\sigma(t,x,a)^2 V''(t,x) + b(t,x,a)V'(t,x).$$

Thus the optimality equation is

$$\inf_a \{c(x,a) + \dot{V}(t,x) + LV(t,x,a)\} = 0$$

and we expect to find the optimal control as the minimizing action a

[29] A rigorous formulation rests on the theory of stochastic integration. The infinitesimal formula given is replaced by the stochastic integral equation

$$X_t = x + \int_0^t \sigma^u(s, X_s) dB_s + \int_0^t b^u(s, X_s) ds,$$

where $(B_t)_{t \geqslant 0}$ is a Brownian motion.

Example (Escape to the boundary). Consider the diffusive controllable dynamical system in $[-1,1]$, with constant diffusivity $\sigma^2 = 1$ and with drift $b(t,x,u) = u$. Suppose we wish to minimize

$$V^u(x) = \frac{1}{2}\mathbb{E}_x^u\left(\tau + \int_0^\tau U_s^2 ds\right),$$

where $\tau = \inf\{t \geqslant 0 : |X_t| = 1\}$, $x \in [-1,1]$, and $U_s = u(X_s)$. The optimality equation is

$$\inf_u \left\{\frac{1+u^2}{2} + uV'(x) + \frac{1}{2}\sigma^2 V''(x)\right\} = 0.$$

The left hand side is minimized to

$$\frac{1}{2}(\sigma^2 V''(x) - V'(x)^2 + 1)$$

by taking $u = -V'(x)$. We can solve the differential equation with boundary conditions $V(-1) = V(1) = 0$ to obtain

$$V'(x) = -\tanh \lambda x,$$

where $\lambda = 1/\sigma^2$. It follows easily now that $V(x)$ is increasing in λ, and takes the limiting values $V(x) = 0$ as $\lambda \to 0$ and $V(x) = 1 - |x|$ as $\lambda \to \infty$. This fits with the intuitively reasonable idea that noise makes it easier to escape to the boundary.

18 Existence and uniqueness of solutions for differential equations

The possibility defining a controlled path $(x_t)_{0\leqslant t\leqslant T}$ using a differential equation is assured by the following result, at least in the case where b and u are continuous. The result does not form an examinable part of the course.

Proposition 18.1. *Let $b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ be continuous and suppose that, for some $K < \infty$, for all $0 \leqslant t \leqslant T$,*

$$|b(t,x) - b(t,y)| \leqslant K|x-y|, \quad x, y \in \mathbb{R}^d. \tag{4}$$

Then, for all $x_0 \in \mathbb{R}^d$, there is a unique differentiable function $(x_t)_{0\leqslant t\leqslant T}$ such that

$$\dot{x}_t = b(t, x_t), \quad 0 \leqslant t \leqslant T.$$

Proof. Note that, by continuity, $C = \sup_{t\leqslant T} |b(t,x_0)| < \infty$. Set $x_t(0) = x_0$ for all $t \geqslant 0$ and define recursively for $n \geqslant 0$

$$x_t(n+1) = x_0 + \int_0^t b(s, x_s(n))ds, \quad 0 \leqslant t \leqslant T. \tag{5}$$

Set $f_n(t) = \sup_{s\leqslant t} |x_s(n) - x_s(n-1)|$. Then

$$f_1(t) \leqslant \int_0^t |b(s,x_0)|ds \leqslant Ct, \quad t \leqslant T,$$

and, for $n \geqslant 1$,

$$f_{n+1}(t) = \sup_{s\leqslant t} \left| \int_0^t \{b(s, x_s(n)) - b(s, x_s(n-1))\}ds \right| \leqslant K \int_0^t f_n(s)ds.$$

Then, by induction $f_n(t) \leqslant CK^{n-1}t^n/n!$. Hence, $\sum_n f_n(T) < \infty$, so the functions $x(n)$ converge uniformly on $[0,T]$ to a continuous limit x. We can let $n \to \infty$ in (5) to obtain

$$x_t = x_0 + \int_0^t b(s, x_s)ds, \quad 0 \leqslant t \leqslant T.$$

Since the integrand here is continuous, we deduce that x is differentiable in t and satisfies $\dot{x}_t = b(t, x_t)$ for all t. Finally, if $(y_t)_{t\leqslant T}$ also has this property, we can define $f(t) = \sup_{s\leqslant t} |x_s - y_s|$. Then f is bounded, say by $B < \infty$, so, arguing as above,

$$f(t) \leqslant K \int_0^t f(s)ds \leqslant BK^{n-1}t^n/n!$$

for all n and t. Hence $x_t = y_t$ for all t. □

We remark that the assumption of continuity in t is unnecessarily strong, especially if the differential equation is interpreted in its integrated form. In particular, in several examples we shall wnat to consider the case where b has a discontinuity in t at some time. The (uniform in t) *Lipschitz condition* 4 is implied by a uniform bound on the gradient of b (in x). If it can be seen *a priori* that any solution stays in a given convex subset U of $\mathbb{R}^d$, then it is only necessary to establish such a bound on U.

We can apply the proposition to the control problem provided that we assume b is continuous on $[0,T] \times \mathbb{R}^d \times A$ and satisfies, for some $K < \infty$, for all $0 \leqslant t \leqslant T$ and $a \in A$,

$$|b(t,x,a) - b(t,y,a)| \leqslant K|x-y|, \quad x, y \in \mathbb{R}^d.$$

Then, for any continuous control $u : [0,T] \to A$, the differential equation

$$\dot{x}_t = b^u(t, x_t), \quad 0 \leqslant t \leqslant T,$$

has a unique solution, where $b^u(t,x) = b(t,x,u_t)$.