MA3H7
Control Theory

Masoumeh Dashti and Andrew Stuart

January 3, 2015

Contents

Chapter 1

Introduction

1.1 Basic Concepts

We start by considering a system of ordinary differential equations (ODEs)

$$\begin{aligned} \dot{x} &= f(x,t), \quad t > 0 \\ x(0) &= x_0. \end{aligned}$$

Here $x_0 \in \mathbb{R}^n$ and $f : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ and the unknown solution is a function $x : [0,\infty) \to \mathbb{R}^n$. If f is independent of time t the equation is termed *autonomous*; otherwise it is *nonautonomous.*

We interpret these equations as a model describing the evolution of the *state* x of some system. The state is the minimum information required to characterize the system in the sense that, if $x(t_0)$ is known then $x(t)$ is uniquely determined for $t > t_0$.

Suppose now that we generalize the above setting to a situation where

$$\begin{aligned} \dot{x} &= f(x,t,u), \quad t > 0 \\ x(0) &= x_0 \end{aligned}$$

where $x_0 \in \mathbb{R}^n$ as before, and now $f : \mathbb{R}^n \times \mathbb{R}^+ \times U \to \mathbb{R}^n$ for some $U \subseteq \mathbb{R}^m$. We are interested in governing the evolution of the state x by means of the function u.

We call $\{u(t), t \geq 0\}$ a *control* and the corresponding state $\{x(t), t \geq 0\}$ is the *response*. Consider the set of *admissible controls*

$$\mathcal{U} = \Big\{ u : [0, \infty) \to U \subseteq \mathbb{R}^m \Big| u(\cdot) \,\text{is measurable} \Big\}. \tag{1.1}$$

The *control problem* is to determine a $u \in \mathcal{U}$ in order to induce a particular outcome for the response x. For example it might be desirable to choose u so that x ends up at a given target in a finite time, or asymptotically as time tends to infinity.

Two particular control problems will be of interest to us:

- An *open loop control* is a function $u \in \mathcal{U}$, chosen to depend only on the initial condition x_0, which ensures some particular control objective; we will concentrate on controlling the system to reach the origin in a finite time.
- A *closed loop control* is defined by choosing $u = c(x)$, for a function $c : \mathbb{R}^n \to U$, with the view of ensuring the stability of an equilibrium point $\overline{x}$; this is also referred to as a *feedback control*.

In addition to controlling the system, we will also be interested in *observing* the system. In particular it is frequently the case that the initial condition $x(0)$ is not known and that observations of the system are made to compensate for this fact. We will consider observation functions $y : [0, \infty) \to \mathbb{R}^p$ given by

$$y(t) = Dx(t), \quad t > 0$$

with $D \in \mathbb{R}^{p \times n}$. Typically we think of $p < n$ or, if $p = n$, then the case where D is not invertible. Thus it is not possible to determine the state of the system $x(t)$ at time t directly from $y(t)$ at time t. Instead we ask two related questions concerning determination of the system when x_0 is unknown, or incompletely known.

- The first, termed *observability*, concerns the question of whether, for some $T > 0$, we can determine x_0, given $\{y(t)\}_{t \in [0,T]}$ and $\{u(t)\}_{t \in [0,T]}$.
- The second, termed *filtering*, concerns the question of how to best estimiate $x(T)$ given $\{y(t)\}_{t \in [0,T]}$ and $\{u(t)\}_{t \in [0,T]}$ and, relatedly, whether our estimate approaches the true $x(T)$ as T grows.

Finally we will also be interested in *optimal control*: attempting to attain a given control objective whilst minimizing some measure of the cost such as the time taken or the L^2 norm of the control. For example we might replace the set of admissible controls $\mathcal{U}$ in (1.1) to the more restrictive set defined by the Hilbert space

$$\mathsf{U} = L^2\big((0,T);\mathbb{R}^m\big), \tag{1.2}$$

and then seek an admissible control which achieves the objective $x(T) = 0$ and minimizes the norm in U.

Throughout the notes we will also consider discrete-time analogues of the questions described above. We study the map

$$x_{k+1} = f(x_k, k, u_k), \quad k \in \mathbb{Z}^+.$$

The reader can readily generalize the concepts of open and closed loop controls, observability and filtering and optimal control to this situation – see Exercises 1-4 and 1-5.

1.2 Examples

Example 1.2.1. ROCKET

We start with an example which illustrates open loop control. Consider a vehicle, driven along a straight horizontal path powered by a rocket. Define:

- $s(t)$*: the distance of the vehicle from the origin at time* t*;*
- $\dot{s}(t)$ *the velocity of the vehicle at time* t*;*
- $u(t)$ *the force applied to the vehicle by the rocket;*
- m *the mass of the vehicle.*

Given the initial position and velocity of the vehicle, the control problem is to find out how to fire the engine to bring the vehicle to rest at the origin, in some specified time T*.*

Newton's second law gives us the equation

$$m\ddot{s} = u \tag{1.3}$$

and we may write this equation as a first order system as follows. First let

$$x = \begin{pmatrix} s \\ \dot{s} \end{pmatrix}.$$

Then

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

where

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ m^{-1} \end{pmatrix}.$$

Mathematically, the control problem is to choose u *so that* $x(T) = 0$*. We formulated the specific example as a first order system because it is more natural to develop a general theory in this form.*

We consider open loop control. First we consider the control problem with unrestricted controls, so that $U = \mathbb{R}$ *in the definition of* $\mathcal{U}$*. For this problem it is then possible to control the system to the origin in any finite positive time* $T > 0$*. To see this consider the function*

$$s(t) = \frac{t^3}{T^3}\big(v_0 T + 2s_0\big) - \frac{t^2}{T^2}\big(2v_0 T + 3s_0\big) + v_0 t + s_0.$$

This function satisfies

$$s(0) = s_0, \quad \dot{s}(0) = v_0, \quad s(T) = 0, \quad \dot{s}(T) = 0.$$

Substituting into the equation (1.3) *gives the control*

$$\begin{aligned} u(t) &= m\ddot{s}(t) \\ &= \frac{6mt}{T^3}\big(v_0 T + 2s_0\big) - \frac{2m}{T^2}\big(2v_0 T + 3s_0\big). \end{aligned} \tag{1.4}$$

This control will steer the system from the initial position/velocity of s_0/v_0 *into the origin, with zero position and velocity, in the finite time* $T > 0$*.*

However, in practice, the size of the engine, and maximum stress allowed on the vehicle, impose bounds on the control. For example we might wish to solve the control problem whilst insisting that u is chosen so that $-M \leq u(t) \leq M$ *for all* $t \geq 0$*, where* M *denotes the maximum absolute value of the control. Then the admissible controls* (1.3) *are defined by taking* $U = [-M, M]$ *in the definition of* $\mathcal{U}$*. Notice that, by choosing* T *large enough, we can ensure that the control given by* (1.4) *is admissible and so it is certainly possible to control this system provided that sufficiently long time-intervals are allowed (see Exercise 1-6).*

Example 1.2.2. INVERTED PENDULUM

We now illustrate a problem arising in closed loop control. Consider an inverted pendulum and define

- *length* ℓ;
- *mass* m;
- *gravity* g;
- *angle* $\phi(t)$ *to the upward vertical position;*
- *torque* $u(t)$ *applied at the pivot.*

We assume that all of the mass m *is concentrated at the end, and that friction is negligible. Given the initial position and velocity of the pendulum, the control problem is to choose the torque so as to bring the pendulum to the vertical position, asymptotically for large time. Note that, in the absence of a torque, the vertical configuration for the pendulum is unstable.*

By Newton's second law for angular momentum we have

$$m\ell^2\ddot{\phi} = mg\ell\sin(\phi) + u.$$

Choose units in which $m\ell^2 = mg\ell = 1$. *Then for small* ϕ *we have the approximate equations of motion*

$$\ddot{\phi} - \phi = u.$$

We work with this linear equation, and note that, if written as a first order system, it is a control problem as above with $n = 2$ *and* $m = 1$. *We may write it as the system*

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

where

$$x = \begin{pmatrix} \phi \\ \dot{\phi} \end{pmatrix}, \; A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Now consider a closed loop control given by

$$u(t) = -\alpha\phi(t) - \beta\dot{\phi}(t)$$

where $\alpha, \beta > 0$ are scalar parameters. We aim to choose these parameters to satisfy the control objective of stabilization. The equations of motion become

$$\ddot{\phi} + \beta\dot{\phi} + (\alpha - 1)\phi = 0.$$

The solution $\phi = \dot{\phi} = 0$ is stable if $\alpha > 1$ and unstable if $\alpha < 1$ since

$$\phi(t) = c^{+}\exp(\gamma^{+}t) + c^{-}\exp(\gamma^{-}t)$$

where

$$\gamma^{\pm} = \frac{1}{2}\Big(-\beta \pm \sqrt{\beta^2 - 4(\alpha - 1)}\Big).$$

Both the real parts of $\gamma^{\pm}$ are negative if $\alpha > 1$, leading to stability, whilst if $\alpha < 1$ then γ^{+} is positive, leading to instability.

Note that we may also write the closed loop control in matrix form. We have

$$u = Cx$$

where

$$C = \begin{pmatrix} -\alpha & -\beta \end{pmatrix}$$

and then

$$\dot{x} = (A + BC)x.$$

Written abstractly we see that the objective is to choose the design matrix C in the closed loop control in order to ensure that $A + BC$ has spectrum in the left-half plane.

Example 1.2.3. SIGNAL PROCESSING

Suppose we wish to find a signal $x^{\dagger}(t) : [0, \infty) \to \mathbb{R}^n$ which is known to satisfy the linear equation

$$\dot{x^{\dagger}} = Ax^{\dagger}, \quad x^{\dagger}(0) = x_0^{\dagger} \tag{1.5}$$

but that we do not know the initial condition $x_0^{\dagger} \in \mathbb{R}^n$. To compensate for this we are given observations $y \in \mathbb{R}^p$ of the system defined via

$$y(t) = Dx^{\dagger}(t), \tag{1.6}$$

with $D \in \mathbb{R}^{p \times n}$ for some $p < n$. In an attempt to determine $x^{\dagger}$ we consider the control system

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

with a closed loop control u chosen to equal $y - Dx$, where y is the data. This is an unrestricted control and so $U = \mathbb{R}^p$. We then obtain the equation

$$\dot{x} = Ax + B\bigl(y - Dx\bigr). \tag{1.7}$$

This equation combines our knowledge of the model, as it has elements of equation (1.5)*, as well as the data y given by* (1.6)*. It is natural to ask whether $e(t) := x(t) - x^\dagger(t)$ converges to* 0 *as t increases since then the true signal will be determined. Notice that $x^\dagger$ itself satisfies*

$$\dot{x^\dagger} = Ax^\dagger + B\bigl(y - Dx^\dagger\bigr)$$

since the last term is identically zero. Thus e satisfies

$$\dot{e} = \bigl(A - BD\bigr)e$$

and the convergence of e to zero can be studied via the spectral properties of the matrix $A - BD$; indeed this is a general case of the specific set-up considered at the end of Example 1.2.2, with $C = -D$.

Example 1.2.4. ROBOT

Consider the problem of specifying the velocity $u(t)$ of a robot, moving along a predetermined straight line with coordinate $x(t)$, in such a fashion that the robot ends at the origin after one time unit, given starting position $x(0) = x_0$. Here $n = m = 1$ and we consider unrestricted open loop controls, so that $U = \mathbb{R}$. The governing equation is simply

$$\dot{x} = u, \quad x(0) = x_0 \tag{1.8a}$$

$$x(1) = 0. \tag{1.8b}$$

There are an uncountably infinite number of solutions to this open loop control problem: choose any differentiable function $x(t)$ satisfying $x(0) = x_0$ and $x(1) = 0$ and then choose the control $u(t) = \dot{x}(t)$. However, what if we now impose the constraint that we wish to find the controller which minimizes the square integral of the control over the time interval $[0, 1]$*? We introduce the Hilbert space $\mathsf{U} = L^2\bigl((0,T);\mathbb{R}\bigr)$ and seek to minimize $J : \mathsf{U} \to \mathbb{R}^+$ given by*

$$J(u) := \frac{1}{2}\int_0^1 u(s)^2 ds,$$

subject to (1.8) *being satisfied. Since equation* (1.8) *implies that $u = \dot{x}$, this implies that we should minimize*

$$I(x) := \frac{1}{2}\int_0^1 |\dot{x}(s)|^2 ds \tag{1.9}$$

subject to $x(0) = x_0$ *and* $x(1) = 0$*. If we ask that* $\dot{x}$ *(and hence* u*) is square integrable on* $(0,1)$ *then this problem is solved by choosing* x *to satisfy the equation*

$$\ddot{x} = 0, \tag{1.10a}$$

$$x(0) = x_0 \qquad x(1) = 0. \tag{1.10b}$$

The solution is $x(t) = (1-t)x_0$*, corresponding to control* $u(t) = -x_0$*.*

1.3 Overview

In Example 1.2.1 we considered a linear control problem with $n = 2$ and $m = 1$ and two specific choices of admissible control set, determined by the choices $U = \mathbb{R}$ and $U = [-M, M]$. In both cases any initial state $x(0) = x_0$ can be steered to the origin in finite time, although the length of the time interval depends on M, the bound on the admissible controls, in the second case. It is not true, however, that all systems can be controlled in this way. Characterizing those x_0 which can be steered to the target is known as the *controllability problem* and is discussed in Chapter 3. That chapter is devoted to open loop control and we study controls from the admissible set $\mathcal{U}$ given by (1.1) and both unrestricted controls, with $U = \mathbb{R}^m$, and restricted controls with $U = [-1, 1]^m$.

In Example 1.2.2 a closed loop control was employed to stabilize an unstable equilibrium point. More general stability notions, and conditions for *stabilization*, are the subject of Chapter 4. Closed-loop controls are the main focus of this chapter.

Example 1.2.3 is prototypical of the problem of *filtering* which we introduce in Chapter 4, and then discuss in more detail in Chapter 5.

Choosing an optimal control based on some cost criterion, as in Example 1.2.4, is the subject of *optimal control theory* introduced in Chapter 5 in discrete time and then discussed in detail in Chapter 6. Here we work with admissible controls which are square-integrable functions from U given by (1.2).

Chapter 2 contains a range of background material in linear algebra, analysis, ordinary differential equations, stability theory, probability theory and the calculus of variations. We will not cover this material sequentially, but rather we will dip into it as needed as the course progresses sequentially through Chapters 3–6.

Exercises

Exercise 1-1. Consider the system

$$\ddot{x} + x = u$$

with $x(0) = x_0$ and $\dot{x}(0) = v_0$. Find an unrestricted control u which ensures that $x(T) = \dot{x}(T) = 0$.

Exercise 1-2. Consider the scalar equation

$$\dot{x} = ax + bu$$

where $a, b > 0$ are constants. Consider a closed loop controller in the form $u = cx$, for constant c. Under what conditions on the scalars b and c does $x(t) \to 0$ as $t \to \infty$.

Exercise 1-3. Consider I given by (1.9). Assume that $x \in C^2([0,1];\mathbb{R})$ and satisfies $x(0) = x_0$ and $x(1) = 0$. Let $h \in C^2([0,1];\mathbb{R})$ with $h(0) = h(1) = 0$. Let $\epsilon \neq 0$ and show that

$$I(x + \epsilon h) = I(x) - \epsilon \int_0^1 \ddot{x}(s)h(s)ds + \epsilon^2 I(h).$$

Deduce that, for all such functions h, $I(x + \epsilon h)$ is minimized as a function of x, for ϵ sufficiently small, by choosing $\ddot{x} = 0$ and $x(0) = x_0$, $x(1) = 0$.

Exercise 1-4. Consider the discrete-time system

$$x_{n+1} - x_n = u_n$$

with $x_0 = X$.

i) Show that, for any integer $N > 0$, there is an unrestricted control sequence $\{u_n\}_{n=0}^{N-1}$ such that $x_N = 0$.

ii) Find the unrestricted control which achieves $x_N = 0$ and minimizes $\sum_{n=0}^{n-1} u_n^2$.

iii) Assume that $X < 0$ and that the controls are restricted so that $u_n \in [0, 1]$. Show that the control objective $x_N = 0$ can only be achieved if $N \geq -X$.

Exercise 1-5. Generalize the concepts of open and closed loop controls, observability and filtering to the discrete-time situation.

Exercise 1-6. Consider $\mathcal{U}$ given by (1.1) with $m = 1$ and $U = [-M, M]$. Show that control (1.4) is admissible for T large enough.

Chapter 2

Background Material

This chapter gathers together a range of mathematical tools which will be used throughout. Section 2.1 describes some basic notational conventions, and then section 2.2 describes the norms and function spaces that we use throughout. Sections 2.3, 2.4 and 2.8 concern linear algebra, analysis and probability respectively. Sections 2.5, 2.6 and 2.7 concern differential equations overviewing linear equations, nonlinear equations and stability theory respectively. Section 2.9 concerns the calculus of variations.

2.1 Notation

We use $\mathbb{R}$ and $\mathbb{C}$ to denote the real and complex numbers respectively. Non-negative reals are denoted by $\mathbb{R}^+ := \{x \in \mathbb{R} | x \geq 0\}$. The positive integers are denoted by $\mathbb{N}$ and the set of all integers, including negative numbers and zero, is $\mathbb{Z} = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$. Non-negative integers are denoted by $\mathbb{Z}^+ = \{0, 1, 2, \cdots\}$.

2.2 Inner-Products and Norms

Throughout these notes we use the *Euclidean inner product* on $\mathbb{C}^r$ given by

$$\langle x, y\rangle = \sum_{j=1}^{n} \overline{x_j} y_j \qquad \forall x, y \in \mathbb{C}^r.$$

This induces the Euclidean norm

$$|x| = \sqrt{\Big(\sum_{j=1}^{r} |x_j|^2\Big)}, \qquad \forall x \in \mathbb{C}^r.$$

We also use $|\cdot|_\infty$ to denote the infinity norm on $\mathbb{C}^r$:

$$|x|_\infty = \max_{1\le j\le n} |x_j|.$$

These definitions are also used, with the natural simplifications, on $\mathbb{R}^r$.

For matrices $A \in \mathbb{R}^{p\times r}$ we will use the Frobenius norm given by

$$\|A\| = \sqrt{\big(\sum_{j=1,k=1}^{p,r} |a_{j,k}|^2\big)}.$$

We define the operator norm induced by the Euclidean norm, namely

$$|A| = \sup_{|x|=1} |Ax|.$$

Note that for operators norms $|AB| \le |A|\cdot|B|$ and hence $|A^k| \le |A|^k$.

We will also require infinite dimensional normed vector spaces from time to time. In particular we will consider Banach spaces $\Big(\mathsf{X}, \|\cdot\|\Big)$: complete normed vector spaces. Throughout we use $B(x;\epsilon)$ to denote a ball in X of radius ϵ centred at $x \in \mathsf{X}$:

$$B(x;\epsilon) := \{x' \in \mathsf{X} : \|x - x'\| < \epsilon\}.$$

The mostly widely used example of a Banach space in these notes is the space $C([0,T];\mathbb{R}^r)$ of continuous functions on $[0,T]$ taking values in $\mathbb{R}^r$.

On occasion it will be useful to use the Hilbert space structure which arises from adding an inner-product to our Banach space. Let $\Big(\mathsf{H}, \langle\cdot,\cdot\rangle, \|\cdot\|\Big)$

denote a Hilbert space. Thus $\|h\|^2 = \langle h, h\rangle$. If $\mathsf{H}_i, i = 1, 2$ are Hilbert spaces and $L : \mathsf{H}_1 \to \mathsf{H}_2$ is a linear operator then the *adjoint* of L is the operator $L^* : \mathsf{H}_2 \to \mathsf{H}_1$ defined by

$$\langle L^* a, b\rangle = \langle a, Lb\rangle$$

for all $a \in \mathsf{H}_2$ and $b \in \mathsf{H}_1$. The simplest example is the Euclidean space $\mathbb{R}^r$ found by restricting the complex Euclidean space above to reals. Another commonly occuring example of a Hilbert space arising in these notes is the space $L^2([0,T];\mathbb{R}^r)$ of square integrable functions on $[0,T]$ taking values in $\mathbb{R}^r$ with inner-product

$$\langle a, b\rangle = \int_0^T a^T(s)b(s)ds.$$

Other examples are the space $H^1([0,T];\mathbb{R}^r)$ with inner-product

$$\langle a, b\rangle_{H^1} = \langle a, b\rangle + \Big\langle \frac{da}{dt}, \frac{db}{dt}\Big\rangle$$

and the space $H_0^1([0,T];\mathbb{R}^r)$ with inner-product

$$\langle a, b\rangle_{H^1} = \Big\langle \frac{da}{dt}, \frac{db}{dt}\Big\rangle.$$

Finally we will sometimes use the Hilbert space $\mathsf{H} := L^2([0,T];\mathbb{R}^r)\times\mathbb{R}^p$ with inner-product defined as follows. Let $a_1, b_1 \in L^2([0,T];\mathbb{R}^r)$ and $a_2, b_2 \in \mathbb{R}^p$ so that $(a_1, a_2) \in \mathsf{H}$ and $(b_1, b_2) \in \mathsf{H}$. Then the inner-product is defined as

$$\langle (a_1, a_2), (b_1, b_2)\rangle = \langle a_1, b_1\rangle_{L^2} + \langle a_2, b_2\rangle_{\mathbb{R}^p}.$$

2.3 Linear Algebra

Definition 2.3.1. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive-definite (resp. negative-definite) if $\langle v, Av\rangle > 0$ (resp. < 0) for all $v \in \mathbb{R}^n\backslash\{0\}$. A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite (resp. negative semi-definite) if $\langle v, Av\rangle \geq 0$ (resp. ≤ 0) for all $v \in \mathbb{R}^n$.*

Thus whenever we refer to a matrix as being *definite* in one of these four ways we are implicitly stating that the matrix is symmetric.

Definition 2.3.2. *The* rank *of a matrix $G \in \mathbb{R}^{p\times q}$ is the number of linearly independent columns.*

Theorem 2.3.3. *The rank of G and of G^T are identical: the number of linearly indepenent rows and columns of a matrix $G \in \mathbb{R}^{p\times q}$ is the same.*

Let $g^{(j)} \in \mathbb{R}^p$ denote the j^{th} column of G; thus $1 \le j \le q$.

Theorem 2.3.4. *Consider a matrix $G \in \mathbb{R}^{p\times q}$. Then*

(i) $\operatorname{rank} G \le \min\{p, q\}$;

(ii) $\operatorname{rank} G < p$ *if and only if there is a non-zero vector* $y \in \mathbb{R}^p$ *which is orthogonal to every column of* G;

(iii) if $p \le q$ *then* $\operatorname{rank} G = p$ *if and only if there is a matrix* $H \in \mathbb{R}^{p\times p}$, *made of columns of* G*, with* $\det H \ne 0$*;*

(iv) if $\operatorname{rank} G = p$ *then there is a set of vectors* $z^{(j)} \in \mathbb{R}^p$ *such that* $\sum_{j=1}^{q} g^{(j)}(z^{(j)})^T = I$.

Proof. (i) The first item follows from the fact that the number of columns is equal to q, so that the number of linearly independent columns is less than or equal to q; and similarly the number of rows is equal to p, so that the number of linearly independent rows is less than or equal to p. Since the rank can be computed row-wise or column-wise, and the answer is the same, the result follows.

(ii) The second item follows from noting that $\operatorname{rank} G < p$ if and only if there is a linear combination of the p rows of G which returns the zero vector in $\mathbb{R}^q$:

$$\sum_{i=1}^{p} y_i g_i^{(j)} = 0, \quad 1 \le j \le q,$$

with $y = (y_1, \cdots, y_p)^T$ not identically zero. This is equivalent to the condition that $\langle y, g^{(j)}\rangle = 0$ for $1 \le j \le q$.

(iii) For the third item note that, since $p \le q$, $\operatorname{rank} G \le p$. Furthermore $\operatorname{rank} G = p$ if and only if there are p linearly independent columns $\{g^{(k_j)}\}_{j=1}^{p}$ with $k_j \in \{1, \cdots, q\}$. If $H = (g^{(k_1)}, \cdots, g^{(k_p)})$ then $\det H \ne 0$ if and only if the p columns are linearly independent. The result follows.

(iv) For the final item note that, if $\operatorname{rank} G = p$ then for every unit vector $e_i \in \mathbb{R}^p$ there is a set of real numbers $\{z_{i,j}\}_{j=1}^q$ such that

$$\sum_{j=1}^{q} z_{i,j} g^{(j)} = e_i, \quad 1 \leq i \leq p.$$

This can be written succintly as

$$\sum_{j=1}^{q} g^{(j)} \big(z^{(j)}\big)^T = I$$

where $z^{(j)} \in \mathbb{R}^p$ has i^{th} entry $z_{i,j}$. □

Definition 2.3.5. *Given a matrix $A \in \mathbb{R}^{r \times r}$, a vector $w \in \mathbb{C}^r$ is an* eigenvector *and $\lambda \in \mathbb{C}$ is an* eigenvalue *of A if*

$$Aw = \lambda w \quad \textit{and} \quad w \neq 0. \tag{2.1}$$

A vector $v \in \mathbb{C}^r$ is a generalized eigenvector *corresponding to λ, if there exists $\ell \in \mathbb{N}$ such that $(A - \lambda I)^\ell v = 0$ and $(A - \lambda I)^{\ell - 1} v \neq 0$. The case $\ell = 1$ corresponds to an eigenvector.*

Definition 2.3.6. *Given a matrix $A \in \mathbb{C}^{r \times r}$ we define the* characteristic polynomial *of A as*

$$\begin{aligned} p_A(z) &:= \det(zI - A) \\ &= \sum_{j=0}^{r} a_j z^j. \end{aligned}$$

Note that $a_r = 1$ by construction.

The eigenvalues of A are the roots of the polynomial p_A and, using this, the following may be proved:

Theorem 2.3.7. Cayley-Hamilton Theorem . *Let p_A be the characteristic polynomial of $A \in \mathbb{R}^{r \times r}$. Then $p_A(A) = 0$. From this it follows that A^r may be written as a linear combination of the set of matrices $\{I, A, A^2, \cdots, A^{r-1}\}$ as*

$$A^r = -\sum_{j=0}^{r-1} a_j A^j.$$

We recall some facts concerning the (real) **Jordan Canonical Form**. We need the following:

Theorem 2.3.8. *Let $\{\lambda_i\}_{i=1}^m$ be the eigenvalues of matrix $A \in \mathbb{R}^{n\times n}$, not counting multiplicities; thus $m \le n$. Then, corresponding to the eigenvalue λ_i, there are k_i generalized eigenvectors of A, for integer $k_i \in \{1, \dots, n\}$; furthermore $\sum_{i=1}^m k_i = n$. The n generalized eigenvectors $v^{(i)}$ can be chosen to be linearly independent and hence span $\mathbb{R}^n$. Consequently the matrix $V = (v^{(1)}, \cdots, v^{(n)})$ is invertible.*

Remarks 2.3.9. *On occasion (see Theorem 2.5.4) it will also be useful to express the eigenvalues $\{\lambda_i\}_{i=1}^n$ counting multiplicities. Corresponding to each eigenvalue λ_i there is then a generalized eigenvector $v^{(i)}$ and an integer ℓ_i such that $(A - \lambda_i I)^{\ell_i} v^{(i)} = 0$ and $(A - \lambda_i I)^{\ell_i - 1} v^{(i)} \neq 0$.*

For arbitrary $A \in \mathbb{R}^{n\times n}$ there exists nonsingular $Q \in \mathbb{R}^{n\times n}$ such that

$$Q^{-1}AQ = \tilde{A} := \begin{pmatrix} J_1 & 0 & \cdots & 0 & 0 \\ 0 & J_2 & \cdots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & J_{m-1} & 0 \\ 0 & 0 & \cdots & 0 & J_m \end{pmatrix}$$

for real square block matrices J_k whose dimensions sum to n. In the notation of Theorem 2.3.8, block J_i corresponds to eigenvalue λ_i of A and has dimension k_i.

If $k_i = 1$ and the eigenvalue is real then $J_i = \lambda_i$ whilst for $k_i > 1$ and real eigenvalue we have

$$J_i = \begin{pmatrix} \lambda_i & 0 & 0 & \cdots & 0 \\ 1 & \lambda_i & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \lambda_i & 0 \\ 0 & 0 & \cdots & 1 & \lambda_i \end{pmatrix}.$$

If $\lambda_i = \alpha_i + \iota\beta_i, \beta_i \neq 0$, $\iota^2 = -1$ and $\alpha_i, \beta_i \in \mathbb{R}$ then

$$J_i = \begin{pmatrix} R_i & 0 & 0 & \cdots & 0 \\ I_2 & R_i & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & R_i & 0 \\ 0 & 0 & \cdots & I_2 & R_i \end{pmatrix}$$

with

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad R_i = \begin{pmatrix} \alpha_i & -\beta_i \\ \beta_i & \alpha_i \end{pmatrix}.$$

We now describe a very useful, special, matrix:

Definition 2.3.10. *A matrix $A \in \mathbb{R}^{n \times n}$ is a* companion matrix *if it has the form*

$$A := \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & \cdots & -a_{n-1} \end{pmatrix} \tag{2.2}$$

for some vector $a = (a_0, \dots, a_{n-1})^T \in \mathbb{R}^n$.

Theorem 2.3.11. *Set $a_n = 1$. The characteristic polynomial of the companion matrix A given in Definition 2.3.10 is*

$$\begin{aligned} p_A(z) &:= \det(zI - A) \\ &= \sum_{j=0}^{n} a_j z^j. \end{aligned}$$

Theorem 2.3.12. *Let $A, B, P \in \mathbb{R}^{n \times n}$ with P invertible. Then:*

1. $\det(AB) = \det(A)\det(B)$*;*
2. $\det(P^{-1}) = \big(\det(P)\big)^{-1}$*;*
3. $\det(PAP^{-1}) = \det(A)$.

Proof of the preceding two theorems is left as Exercise 2-18.

2.4 Analysis

Let $\Big(\mathsf{X}, \|\cdot\|\Big)$ be a Banach space, that is a complete normed vector space.

Definition 2.4.1. *A sequence* $\{x_k\}_{k\in\mathbb{Z}^+}$ *in* X converges strongly *to* $x \in \mathsf{X}$ *if* $\|x - x_k\| \to 0$.

Definition 2.4.2. *We use the notation* $\partial\Omega$ *to denote the boundary of a set* $\Omega \subseteq \mathbb{R}^n$. *This is the set of points* $x \in \Omega$ *for which both* $B(x;\epsilon) \cap \Omega^c$ *and* $B(x;\epsilon) \cap \Omega$ *are non-empty for all* $\epsilon > 0$. *We use the notation* $\operatorname{Int}\Omega$ *to denote the interior of* Ω*: the set of points for which* $B(x;\epsilon) \subseteq \Omega$ *for some* $\epsilon > 0$.

Definition 2.4.3. *Let* X, Z *be Banach spaces and* $B \subset \mathsf{X}$ *an open set; then* $F : B \to \mathsf{Z}$ *is* Fréchet differentiable *at* $x_0 \in B$ *if there exists a bounded linear operator* $A : \mathsf{X} \to \mathsf{Z}$ *such that*

$$\lim_{\|h\|\to 0} \|h\|^{-1}\Big(\|F(x_0 + h) - F(x_0) - Ah\|\Big) = 0.$$

The operator A *is called the* Fréchet derivative *of* F *at* x_0.

Theorem 2.4.4. Implicit Function Theorem *Let* $F : (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \to F \in \mathbb{R}^n$ *be continuously differentiable and satisfy*

$$F(0,0) = 0, \quad \det D_y F(0,0) \neq 0.$$

Then there exists $\theta > 0$ *and a function* $x \in \mathbb{R}^n \mapsto \Gamma \in \mathbb{R}^n$ *such that* $\Gamma(0) = 0$, Γ *is continuous at* 0 *and* $F\big(x, \Gamma(x)\big) = 0$ *for any* $x \in B(0;\theta)$.

Proof. Define $D \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{n\times n}$ by $D := D_x F(0,0), E := D_y F(0,0)$ Then define $G : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$G(x,y) = E^{-1}\Big(F(x,y) - Dx\Big) - y.$$

Since F is continuously differentiable, G has the same property. We also have

$$G(0,0) = 0, \quad D_x G(0,0) = 0, \quad D_y G(0,0) = 0.$$

Therefore, we can find η small enough so that, for any $(x, y) \in \overline{B}(0;\eta) \times \overline{B}(0;\eta)$, we have

$$\begin{aligned} \|D_x G(x,y)\| &\leq \frac{1}{2}, \\ \|D_y G(x,y)\| &\leq \frac{1}{2}, \\ \|G(x,y)\| &\leq \frac{1}{2}\Big(\|x\| + \|y\|\Big). \end{aligned}$$

Now for each $x \in \overline{B}(0;\eta)$ define

$$H(x,\cdot) : y \mapsto -E^{-1}Dx - G(x,y)$$

and choose $\theta \in (0,\eta)$ small enough so that

$$\Big(\|E^{-1}D\| + \frac{1}{2}\Big)\theta < \frac{1}{2}\eta.$$

Then, for any $x \in B(0;\theta)$ and $y \in B(0;\eta)$,

$$\begin{aligned}\|H(x,y)\| &\leq \|E^{-1}D\|\|x\| + \frac{1}{2}\Big(\|x\| + \|y\|\Big)\\ &\leq \Big(\|E^{-1}D\| + \frac{1}{2}\Big)\|x\| + \frac{1}{2}\|y\|\\ &\leq \eta.\end{aligned}$$

From this it follows that $H(x,\cdot) : B(0;\eta) \mapsto B(0;\eta)$ for any $x \in B(0;\theta)$.

Having established that $H(x,\cdot)$ maps a closed subset into itself, it is natural to attempt to use the Contraction Mapping Theorem 2.4.6, which follows below. Thus we show that $H(x,\cdot)$ is a contraction on $B(0;\eta)$ – see Definition 2.4.5. We have

$$\begin{aligned}\|H(x,y) - H(x,z)\| &= \|G(x,y) - G(x,z)\|\\ &= \Big\|\int_0^1 \frac{d}{ds}G\big(x, sy + (1-s)z\big)ds\Big\|\\ &= \Big\|\int_0^1 D_yG\big(x, sy + (1-s)z\big)(y-z)ds\Big\|\\ &\leq \int_0^1 \|D_yG\big(x, sy + (1-s)z\big)\|ds\|y-z\|\\ &\leq \frac{1}{2}\|y-z\|.\end{aligned}$$

Hence H is a contraction on $B(0;\eta)$ and maps $B(0;\eta)$ into itself. Hence there is a unique fixed point $y^\star = \Gamma(x) \in B(0;\eta)$ of $H(x,\cdot)$, for all $x \in B(0;\theta)$. Since $H(0,0) = G(0,0) = 0$ we have $\Gamma(0) = 0$. To see that Γ is continuous we note that

$$\begin{aligned}\|\Gamma(x)\| &= \|H\big(x,\Gamma(x)\big)\|\\ &= \|E^{-1}Dx + G\big(x,\Gamma(x)\big)\|\\ &\leq \|E^{-1}D\|\|x\| + \frac{1}{2}\| + \frac{1}{2}\|\Gamma(x)\|\end{aligned}$$

and therefore, re-arranging,

$$\|\Gamma(x)\| \leq \Big(2\|E^{-1}D\| + 1\Big)\|x\|.$$

$$\|\Gamma(x) - \Gamma(0)\| \leq \|\Gamma(x)\| \leq \Big(2\|E^{-1}D\| + 1\Big)\|x\|.$$

Thus Γ is continuous. □

Definition 2.4.5. *Let $\Big(\mathsf{X}, \|\cdot\|\Big)$ be a Banach space and $B \subseteq \mathsf{X}$ a closed subset. A mapping $T : B \to B$ is a* contraction on B *if there exists $\lambda \in [0, 1)$ such that, for all $x_1, x_2 \in B$, $\|T(x_1) - T(x_2)\| \leq \lambda\|x_1 - x_2\|$.*

Theorem 2.4.6. Contraction Mapping Theorem *If T is a contraction on B then there is a unique solution of the equation $T(x) = x$ in B. Furthermore there is $C > 0$ such that the iteration $x_{k+1} = T(x_k)$ with $x_0 \in B$ satisfies $\|x_k - x\| \leq C\lambda^k\|x_0 - x\|$.*

Definition 2.4.7. *A subset Ω of a vector space is* symmetric *if $x \in \Omega$ implies $-x \in \Omega$.*

Definition 2.4.8. *A subset Ω of a vector space is* convex *if $x, y \in \Omega$ and $\lambda \in [0, 1]$ imply $\lambda x + (1 - \lambda)y \in \Omega$.*

We now discuss further structure which arises from the Hilbert space setting. Let $\Big(\mathsf{H}, \langle\cdot,\cdot\rangle, \|\cdot\|\Big)$ denote a Hilbert space.

Definition 2.4.9. *A sequence $\{h_k\}_{k\in\mathbb{Z}^+}$ in H* converges weakly *to $h \in \mathsf{H}$ if, for all $\ell \in \mathsf{H}$,*

$$\langle h_k, \ell\rangle \to \langle h, \ell\rangle.$$

We write $h_k \rightharpoonup h$.

Recall (Definition 2.4.1) that a sequence $\{h_k\}_{k\in\mathbb{Z}^+}$ in H *converges strongly* to $h \in \mathsf{H}$ if $\|h - h_k\| \to 0$. We then write $h_k \to h$. Strong convergence implies weak convergence, but the converse is not true.

Theorem 2.4.10. *Every bounded sequence $\{h_k\}_{k\in\mathbb{Z}^+}$ in H contains a weakly convergent subsequence in H.*

Theorem 2.4.11. *Let $C \subset \mathsf{H}$ be closed, bounded and convex. Every sequence $\{h_k\}_{k\in\mathbb{Z}^+}$ in C contains a weakly convergent subsequence with limit in C.*

Theorem 2.4.12. *If $h_k \rightharpoonup h$ then*

$$\liminf_{k\to\infty} \|h_k\|^2 \geq \|h\|.$$

We say that the function $x \mapsto \|x\|^2$ is weakly lower semicontinuous*.*

Proof.

$$\begin{aligned}\|h_k\|^2 &= \|h + (h_k - h)\|^2 \\ &= \|h\|^2 + 2\langle h, h_k - h\rangle + \|h_k - h\|^2 \\ &\geq \|h\|^2 + \langle h, h_k - h\rangle.\end{aligned}$$

By weak convergence we have $\langle h, h_k - h\rangle \to 0$ as $h_k \to h$ and the desired result follows. □

2.5 Linear ODEs

Consider the system

$$\begin{aligned}\dot{x} &= Ax + a, \quad \text{for almost all } t \in [0,T] &\quad (2.3a)\\ x(0) &= x_0, &\quad (2.3b)\end{aligned}$$

with $A(t) \in \mathbb{R}^{n\times n}$, $a(t) \in \mathbb{R}^n$ for any $t > 0$ and with solution $x(t) \in \mathbb{R}^n$, any $t > 0$.

A key role in what follows is played by the matrix equation

$$\begin{aligned}\dot{S} &= AS, \quad \text{for almost all } t \in [0,T] &\quad (2.4a)\\ S(0) &= I; &\quad (2.4b)\end{aligned}$$

The matrix-valued function $S : \mathbb{R}^+ \to \mathbb{R}^{n\times n}$ is called the *fundamental solution.*

We define a solution of (2.3) to be a solution of the integral equation

$$x(t) = x_0 + \int_0^t A(s)x(s)ds + \int_0^t a(s)ds. \quad (2.5)$$

Likewise we define a solution of (2.4) to be a solution of the integral equation

$$S(t) = I + \int_0^t A(s)S(s)ds. \quad (2.6)$$

We start by studying the nonautonomous fundamental solution, where A may depend on time t, and then specialize to the autonomous case, where A is constant.

Remarks 2.5.1. *We work on a fixed interval $[0,T]$ but, under the assumptions made here, T may be chosen arbitrarily in $\mathbb{R}^+$.*

2.5.1 Nonautonomous Fundamental Solution

In the general nonautonomous setting we have the following result.

Theorem 2.5.2. *In the system (2.3) assume that the entries of $A(\cdot), a(\cdot)$ are locally integrable. Then:*

1. *there exists exactly one function $x : [0,T] \to \mathbb{R}^n$, $T < \infty$, with absolutely continuous elements solving (2.3);*
2. *there exists exactly one function $S : [0,T] \to \mathbb{R}^{n\times n}$, $T < \infty$, with absolutely continuous elements solving (2.4);*
3. *the matrix $S(t)$ is invertible for any $t \in [0,T]$ and the unique solution of (2.3) is written as*

$$x(t) = S(t)x_0 + \int_0^t S(t)S^{-1}(s)a(s)ds, \quad t \in [0,T] \tag{2.7}$$

and satisfies

$$x(t) = S(t)S(t_0)^{-1}x(t_0) + \int_{t_0}^t S(t)S^{-1}(s)a(s)ds, \quad t \in [0,T]. \tag{2.8}$$

Proof. 1). Solution of equation (2.3) is defined to be solution of the integral equation (2.5). Since $A(s)$ is locally integrable $\exists T_1 > 0$ such that

$$\int_0^{T_1} |A(s)|ds = \alpha \in (0,1).$$

In fact we may choose a sequence $T_j \to \infty$ such that

$$\int_{T_j}^{T_{j+1}} |A(s)|ds = \alpha.$$

To see that the sequence accumulates at infinity note that, if the sequence were to accumulate at $T^\star < \infty$ then, with $T_0 = 0$,

$$\int_0^{T^\star} |A(s)|ds = \sum_{j=0}^{\infty} \int_{T_j}^{T_{j+1}} |A(s)|ds = \sum_{j=0}^{\infty} \alpha = \infty$$

contradicting local integrability.

For $t \in [0, T_1]$ define $\mathcal{L}$ by

$$\big(\mathcal{L}x\big)(t) = x_0 + \int_0^t a(s)ds + \int_0^t A(s)x(s)ds.$$

The operator $\mathcal{L}$ defines a continuous transformation from $\mathsf{X} := C([0, T_1]; \mathbb{R}^n)$ into itself. Recall that X is a Banach space when equipped with the norm

$$\|x\| := \sup_{0 \le t \le T_1} |x(t)|,$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^n$, and also the induced matrix norm on $\mathbb{R}^{n \times n}$. We now show that $\mathcal{L}$ is a contraction on X because, for any $x, y \in \mathsf{X}$, we have

$$\begin{aligned}
\|\mathcal{L}x - \mathcal{L}y\| &= \sup_{0 \le t \le T_1} \big|(\mathcal{L}x)(t) - (\mathcal{L}y)(t)\big| \\
&\le \Big(\int_0^{T_1} |A(s)|ds\Big) \sup_{0 \le t \le T_1} |x(t) - y(t)| \\
&= \alpha \|x - y\|
\end{aligned}$$

which establishes that $\mathcal{L}$ is a contraction on X since $\alpha < 1$. Hence $x = \mathcal{L}x$ has a unique solution in X by the contraction mapping principle (Theorem 2.4.6). But the equation $x = \mathcal{L}x$ is of course simply the integral equation equivalent to (2.3) and the existence of a unique solution on $[0, T_1]$ follows. Consider T_1 as the initial time and arguing similarly on the interval $[T_1, T_2]$ we extend existence and uniqueness to $[0, T_2]$. We proceed inductively on $[T_j, T_{j+1}]$ to get the result on any interval $[0, T]$ with $T < \infty$.

2). Having obtained existence and uniqueness of solutions of (2.3), we define $S(t)$ as follows. Let $x_0^{(i)} = [0, \cdots, 0, 1, 0, \cdots, 0]^T$, choose $a = 0$ and let $x^{(i)}(t)$ be the resulting solution of (2.3). Then define

$$S(t) = \Big(x^{(1)}(t), x^{(2)}(t), \cdots, x^{(n)}(t)\Big).$$

By construction this is the the solution to (2.4).

3). Assume for contradiction that $S(t)$ is not invertible for all $t > 0$ and let $T_0 \in [0, T]$ be the first time at which S fails to be invertible, noting that T_0 is strictly positive. This is because a matrix S is invertible if and only if $\det S \neq 0$ and because $S(0) = I$ we have $\det S(0) = 1$. By continuity of the solution $S(t)$ in t, and continuity of the function $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$, we deduce that $\det S(t) > 0$ for some interval $t \in [0, T_0)$.

For $t \in [0, T_0)$ we have

$$\begin{aligned} 0 &= \frac{d}{dt}\Big(S(t)S^{-1}(t)\Big) \\ &= \Big(\frac{dS(t)}{dt}\Big)S^{-1}(t) + S(t)\frac{d}{dt}\Big(S^{-1}(t)\Big) \\ &= A(t) + S(t)\frac{d}{dt}\Big(S^{-1}(t)\Big). \end{aligned}$$

Rearranging gives the identity

$$\frac{d}{dt}\Big(S^{-1}(t)\Big) = -S^{-1}(t)A(t), \quad t \in [0, T_0).$$

Now let $y(t)$ solve the equation

$$\begin{aligned} &\dot{y}^T = -y^T A, \quad \text{for almost all } t \in [0, T] \\ &y(0) = y_0. \end{aligned}$$

By an argument similar to that used in the proof of 1.) it follows that this equation has a unique solution and, hence, so too does the matrix equation

$$\begin{aligned} &\dot{\Phi} = -\Phi A, \quad \text{for almost all } t \in [0, T] \\ &\Phi(0) = I. \end{aligned}$$

Clearly, then, by uniqueness, we have $S^{-1}(t) = \Phi(t)$. But $\lim_{t\to T_0^-} \det \Phi(t)$ is finite and so $\det S(T_0) \neq 0$.

Having existence and uniqueness of $x(t)$, together with invertibility of $S(t)$, establishing the identities (2.7) and (2.8) are the remaining steps in the proof. From (2.7) we see that $x(0) = x_0$ so that the initial condition is satisfied. From (2.8) we have

$$S^{-1}(t)x(t) = S^{-1}(t_0)x(t_0) + \int_{t_0}^{t} S^{-1}(s)a(s)ds$$

and differentiating gives

$$\Big(\frac{d}{dt}S^{-1}(t)\Big)x(t) + S^{-1}(t)\frac{d}{dt}x(t) = S^{-1}(t)a(t)$$

and using the expression above for the derivative of $S^{-1}(t)$ we obtain

$$-S^{-1}(t)A(t)x(t) + S^{-1}(t)\frac{d}{dt}x(t) = S^{-1}(t)a(t).$$

Rearranging and multiplying through by $S(t)$ shows that

$$\frac{d}{dt}x(t) = A(t)x(t) + a(t).$$

Thus the solution of (2.8) is also the solution of (2.3). The proof is complete.

□

2.5.2 Autonomous Fundamental Solution

We now develop some theory relevant to the case of autonomous linear systems: more precisely, we study the case where $A \in \mathbb{R}^{n\times n}$ is a constant matrix independent of time. Consider the following definition of the matrix $\exp(At)$:

$$S(t) = \exp(At) := \sum_{k=0}^{\infty} \frac{1}{k!}A^k t^k. \tag{2.9}$$

We first show that this sum is well-defined, and we then show that it coincides with the preceding definition of $S(t)$ from Theorem 2.5.2, when specialized to the autonomous case. Writing $e^{At} = \exp(At)$ we have the following:

Theorem 2.5.3. *Let $A \in \mathbb{R}^{n\times n}$ be a constant matrix. Then the matrix series* (2.9) *for the matrix exponential e^{At} converges and satisfies*

- $e^{tA}e^{sA} = e^{(t+s)A}$;
- $\big(e^{tA}\big)^{-1} = e^{-tA}$;
- $\frac{d}{dt}\big(e^{tA}\big) = Ae^{tA}$;
- $Ae^{tA} = e^{tA}A$.

Thus $S(t) = e^{At}$ coincides with $S(t)$ appearing in Theorem 2.5.2 and solving (2.4)*. Furthermore the unique solution of the equation* (2.3) *can, in this case, be written as*

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}a(s)ds. \tag{2.10}$$

Proof. To see that the infinite sum is well-defined, note that for the matrix norm $|\cdot|$ induced by the Euclidean norm we have

$$\begin{aligned}|S(t)| &\leq \sum_{k=0}^{\infty} \frac{1}{k!}|A^k|t^k \\ &\leq \sum_{k=0}^{\infty} \frac{1}{k!}|A|^k t^k \\ &= \exp(|A|t) \\ &< \infty.\end{aligned}$$

In fact, the same argument shows that, for all $t \in [0, T]$, $|S(t)| \leq \exp(|A|T)$ and hence, by the Weierstrass M-test, that the series is uniformly convergent on $[0, T]$. See Exercise 2-4 for the remaining points in the proof. □

Recall that the Cayley-Hamilton Theorem 2.3.7 implies that A^n may be written as a linear combination of the set of matrices $\{I, A, A^2, \cdots, A^{n-1}\}$. Iterating on this idea, using the expression for e^{At} as a power series in A, demonstrates the following theorem. Recall the vectors $\{v^{(j)}\}_{j=1}^n$ denoting the generalized eigenvectors of A. By Theorem 2.3.8 these vectors form a basis for $\mathbb{R}^n$ and, without loss of generality, we assume that the normalization $|v^{(j)}| = 1$ is chosen. Furthermore, corresponding to every generalized eigenvector is an eigenvalue λ_i (counting multiplicities) and Remark 2.3.9 tells us that there is an integer $\ell_i \leq n$ such that

$$\begin{aligned}(A - \lambda_i I)^r v^{(i)} &\neq 0, \quad r < \ell_i \\ (A - \lambda_i I)^r v^{(i)} &= 0, \quad r = \ell_i.\end{aligned}$$

Note that, in constrast to Theorem 2.3.8, in the following theorem the eigenvalues λ_i are enumerated according to their multiplicities.

Theorem 2.5.4. *Let $A \in \mathbb{R}^{n\times n}$. Then, for any $x_0 \in \mathbb{R}^n$, there is $\alpha = (\alpha_1, \cdots, \alpha_n)^T \in \mathbb{R}^n$ and $\ell = (\ell_1, \cdots, \ell_n)^T \in \{1, \cdots, n\}^n$ such that*

$$e^{At}x_0 = \sum_{i=1}^{n} \alpha_i \, e^{At} v^{(i)} = \sum_{i=1}^{n} \alpha_i \Big(\sum_{j=0}^{\ell_i - 1} (A - \lambda_i I)^j \frac{t^j}{j!} \Big) e^{\lambda_i t} v^{(i)}. \tag{2.11}$$

Proof. For $v^{(i)}$ a generalized eigenvector of A we have a corresponding eigenvalue $\lambda = \lambda_i$ and integer $\ell = \ell_i$ as defined preceding the theorem statement. Furthermore, for $v = v^{(i)}$ and $\lambda = \lambda_i$,

$$\begin{aligned} e^{At}v &= e^{(A-\lambda I)t} e^{\lambda t} v \\ &= \Big(I + (A - \lambda I)t + \cdots + (A - \lambda I)^{\ell - 1} \frac{t^{\ell-1}}{(\ell-1)!} \Big) e^{\lambda t} v \\ &= \sum_{j=0}^{\ell-1} (A - \lambda I)^j \frac{t^j}{j!} e^{\lambda t} v. \end{aligned}$$

Now let $\{v^{(i)}\}_{i=1}^n$ denote the generalized eigenvectors of A, and note that this set spans $\mathbb{R}^n$. By Theorem 2.3.8, for any $x_0 \in \mathbb{R}^n$ we may write

$$x_0 = \sum_{i=1}^{n} \alpha_i v^{(i)} = V\alpha$$

where $\alpha = (\alpha_1, \cdots, \alpha^n)^T$. Thus (2.11) follows. □

In the following corollary we enumerate the eigenvalues according to multiplicities and use the notation introduced in Remark 2.3.9.

Corollary 2.5.5. *Let $A \in \mathbb{R}^{n\times n}$ and define*

$$\overline{\lambda} := \max\{\mathrm{Re}\,\lambda : \lambda \, \text{is an eigenvalue of} \, A\}.$$

Then:

(i) if $\overline{\lambda} < 0$ then there is constant $C > 0$ such that

$$|e^{At}| \le C \exp\big(\frac{1}{2}\overline{\lambda} t\big);$$

(ii) if $\overline{\lambda} \leq 0$ and the integer $\ell_i = 1$ for every generalized eigenvector associated with an eigenvalue with zero real part, then there is constant $C > 0$ such that

$$|e^{At}| \leq C;$$

(iii) if $\overline{\lambda} > 0$ then there is constant $C > 0$ such that

$$|e^{At}| \leq C \exp\bigl(2\overline{\lambda}t\bigr).$$

Proof. In this proof we use the same notation as above, with m_i denoting the multiplicity of eigenvalue λ_i with corresponding generalized eigenvector $v^{(i)}$.

(i) Recall that (2.11) gives

$$e^{At}x_0 = \sum_{i=1}^{n} \alpha_i \Bigl(\sum_{j=0}^{\ell_i - 1} (A - \lambda_i I)^j \frac{t^j}{j!}\Bigr) e^{\lambda_i t} v^{(i)}.$$

By the assumptions of the theorem there is constant $K > 0$ such that

$$\max_{1 \leq i \leq n, 0 \leq j \leq n-1} \Bigl| e^{\lambda_i t} \frac{t^j}{j!} \Bigr| \leq K e^{\frac{1}{2}\overline{\lambda}t}.$$

Define

$$M = \max_{1 \leq i \leq n, 0 \leq j \leq n-1} |(A - \lambda_i I)^j|.$$

Then

$$\begin{aligned} |e^{At}x_0| &\leq n^2 \Bigl(\max_{1 \leq i \leq n, 0 \leq j \leq n-1} |(A - \lambda_i I)^j|\Bigr) \Bigl(\max_{1 \leq i \leq n, 0 \leq j \leq n-1} e^{\lambda_i t} \frac{t^j}{j!}\Bigr) \|\alpha\|_\infty \\ &\leq n^2 M K e^{\frac{1}{2}\overline{\lambda}t} \|\alpha\|_\infty. \end{aligned}$$

Now note that, by Theorem 2.3.8, there is invertible V, made from the columns $\{v^{(i)}\}_{i=1}^n$, such that $x_0 = V\alpha$. Hence

$$\|\alpha\|_\infty \leq |\alpha| \leq |V^{-1}||x_0|.$$

Combining this and the previous inequality gives

$$|e^{At}x_0| \leq cMKe^{\frac{1}{2}\lambda t} \times |V^{-1}||x_0|$$

and the desired result follows.

(ii) Without loss we assume that, for index $1 \leq i \leq r$ we have $\ell_i = 1$, whilst for index $r+1 \leq i \leq n$ we have generalized eigenvectors ($\ell_i > 1$). The representation (2.11) then gives

$$e^{At}x_0 = \sum_{i=1}^{r} \alpha_i e^{\lambda_i t} v_i + \sum_{i=r+1}^{n} \alpha_i \Big(\sum_{j=0}^{\ell_i - 1} (A - \lambda_i I)^j \frac{t^j}{j!} \Big) e^{\lambda_i t} v_i. \tag{2.12}$$

For $1 \leq i \leq r$ we have $\mathrm{Re}\lambda_i \leq 0$ and so we have

$$|e^{\lambda_i t}| \leq 1.$$

For $r+1 \leq i \leq n$ we have $\mathrm{Re}\lambda_i < 0$ and hence there is $C > 0$ and $\lambda < 0$ such that

$$\max_{r+1 \leq i \leq n, 1 \leq j \leq n} \quad e^{\lambda_i t} \frac{t^j}{j!} \leq C e^{\frac{1}{2}\lambda t};$$

Therefore

$$\begin{aligned} \sup_{t \geq 0} |e^{At}x_0| &\leq \sup_{t \geq 0} \|\alpha\|_\infty \big(r + n^2 C M e^{\frac{1}{2}\lambda t} \big) \\ &\leq \|\alpha\|_\infty \big(r + n^2 CM \big) \end{aligned}$$

with M as defined in the proof of **(i)**. Note that

$$x_0 = \sum_{i=1}^{n} \alpha_i v_i = V\alpha$$

where, by Theorem 2.3.8, $V \in \mathbb{R}^{n \times n}$ is invertible. Thus

$$\|\alpha\|_\infty \leq |\alpha| \leq |V^{-1}||x_0|.$$

Hence we may choose δ so that, if $|x_0| < \delta$, then $\big(r + n^2 CM \big)\|\alpha\|_\infty < \epsilon$ and the desired bound follows.

The proof of **(iii)** is similar to that of **(i)** and is omitted. □

Consider equation (2.3) in the case where A is constant in time and $a \in L^2\big((0,T);\mathbb{R}^n\big)$; notice that this choice of a is locally integrable so that Theorem 2.5.2 shows the existence of a solution

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)} a(s) ds. \tag{2.13}$$

Define the operator $\mathcal{A} : H^1\big((0,T);\mathbb{R}^n\big) \to L^2\big((0,T);\mathbb{R}^n\big) \times \mathbb{R}^n$ by

$$\mathcal{A}(x) = \begin{pmatrix} \frac{dx}{dt} - Ax \\ x(0) \end{pmatrix}. \tag{2.14}$$

We may show the following:

Theorem 2.5.6. *The operator $\mathcal{A}$ has bounded inverse $\mathcal{A}^{-1} : L^2\big((0,T);\mathbb{R}^n\big) \times \mathbb{R}^n \to H^1\big((0,T);\mathbb{R}^n\big)$.*

Proof. Consider the equation $\mathcal{A}x(t) = (a, x_0)^T$. Using (2.13) and the fact that

$$\frac{dx}{dt} = Ax + a, \quad x(0) = x_0,$$

we have the existence of a constant C, independent of $(a, x_0)^T$, such that

$$\begin{aligned} |x(t)| &\le C\big(|x_0| + \int_0^t |a(s)|ds\big) \\ \Big|\frac{dx}{dt}(t)\Big| &\le |A||x(t)| + |a(t)|. \end{aligned}$$

Since $a \in L^2$ it follows that

$$\sup_{0 \le t \le T} |x(t)| \le C\Big(|x_0| + \sqrt{T}\|a\|_{L^2}\Big)$$

and hence that

$$\Big|\frac{dx}{dt}(t)\Big|^2 \le 2\|A\|^2 C^2 \Big(|x_0| + \sqrt{T}\|a\|_{L^2}\Big)^2 + 2|a(t)|^2.$$

Integrating these two estimates shows that there is constant K, independent of $(a, x_0)^T$, such that

$$\|x\|_{H^1}^2 = \int_0^T \Big(|x(t)|^2 + \Big|\frac{dx}{dt}(t)\Big|^2\Big)dt \le K\Big(|x_0| + \|a\|_{L^2}\Big)^2$$

as required. □

2.6 Nonlinear ODEs

We briefly review existence, uniqueness and continuity of solutions for the nonautnonomous ODE

$$\dot{x} = g(x,t) \tag{2.15a}$$

$$x(0) = x_0 \tag{2.15b}$$

with $x \in \mathbb{R}^n$, $g : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$.

Theorem 2.6.1. *Assume that, for each $x \in \mathbb{R}^n$, the function $g(x,\cdot) : [0,T] \to \mathbb{R}^n$ is measurable. If, for an integrable non-negative function $c : [0,T] \to [0,\infty)$, we have, for all $x, y \in \mathbb{R}^n$ and $t \in [0,T]$,*

$$|g(x,t)| \le c(t)\big(1 + |x|\big),$$
$$|g(x,t) - g(y,t)| \le c(t)|x-y|,$$

then, for all $x_0 \in \mathbb{R}^n$, equation (2.15) has exactly one solution $x(\cdot) = \varphi(\cdot, x_0)$ on $[0,T]$.

Theorem 2.6.2. *Consider equation (2.15) and assume that:*

- *the conditions of Theorem 2.6.1 hold;*
- *$D_x g(\cdot, t)$ is continuous;*
- *$g(\cdot,\cdot)$ and $D_x g(\cdot,\cdot)$ are bounded on bounded subsets of $[0,T] \times \mathbb{R}^n$.*

Then the mapping $\xi \in \mathbb{R}^n \mapsto \varphi(\cdot, \xi) \in C([0,T];\mathbb{R}^n)$ is differentiable at any $\xi \in \mathbb{R}^n$. The function $t \mapsto X(t) := D_\xi \varphi(t, x_0)$ is absolutely continuous and satisfies

$$\dot{X} = D_x g\big(\varphi(t,x_0), t\big) X$$
$$X(0) = I.$$

Corollary 2.6.3. *Consider equation (2.15) in the case where g depends on a parameter $\gamma \in \mathbb{R}^\ell$ so that $g = g(x,\gamma,t)$ with $(x,\gamma) \in \mathbb{R}^n \times \mathbb{R}^\ell$ and $t \in [0,T]$. Assume that g satisfies the assumptions of Theorem 2.6.2 with $x \in \mathbb{R}^n$ replaced by $(x,\gamma) \in \mathbb{R}^n \times \mathbb{R}^\ell$ and denote the solution of (2.15) by $\varphi(t,\gamma,\xi)$. Then for arbitrary $t \in [0,T], \gamma_0 \in \mathbb{R}^\ell$ and $x_0 \in \mathbb{R}^n$ the mapping $\gamma \in \mathbb{R}^\ell \mapsto \varphi(\cdot,\gamma,\xi)$ is differentiable at γ_0 and $G(t) := D_\gamma \varphi(t, \gamma_0, x_0)$ satisfies*

$$\dot{G} = D_x g\big(\varphi(t,\gamma_0,x_0),t\big) G + D_\gamma g\big(\varphi(t,\gamma_0,x_0),t\big)$$
$$G(0) = 0.$$

Proof. We write equation (2.15) as

$$\begin{aligned}
\dot{x} &= g(x, \gamma, t)\\
\dot{\gamma} &= 0\\
x(0) &= x_0\\
\gamma(0) &= \gamma.
\end{aligned}$$

Define

$$z = \begin{pmatrix} x \\ \gamma \end{pmatrix}, z_0 = \begin{pmatrix} x_0 \\ \gamma \end{pmatrix}, h(z,t) = \begin{pmatrix} g(x,\gamma,t) \\ 0 \end{pmatrix}.$$

Then we have

$$\begin{aligned}
\dot{z} &= g(z, t)\\
z(0) &= z_0.
\end{aligned}$$

Applying Theorem 2.6.2 to this system gives the desired result. □

2.7 Stability of ODEs

Consider the nonautonomous equation (2.15) and assume that $g(0,t) \equiv 0$ for all $t \geq 0$ and that $g : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ satisfies the conditions of Theorem 2.6.2.

Definition 2.7.1. *Consider the system* (2.15). *The origin is* stable *if the solution $\varphi(t, x_0)$ satisfies the following: for any $\epsilon > 0 \, \exists \, \delta > 0$ such that*

$$|x_0| < \delta \Longrightarrow \sup_{t \geq 0} |x(t)| < \epsilon.$$

The origin is asymptotically stable *if it is stable and $\exists r > 0$ such that*

$$|x_0| < r \Longrightarrow \lim_{t \to \infty} |x(t)| = 0.$$

The origin is globally asymptotically stable *if, for all $x_0 \in \mathbb{R}^n$,*

$$\lim_{t \to \infty} |x(t)| = 0.$$

The origin is exponentially stable *if it is stable and $\exists M, \omega, r > 0$ such that*

$$|x_0| < r \Longrightarrow |x(t)| \leq M e^{-\omega t} |x_0| \; \forall t \geq 0.$$

Remarks 2.7.2. *Although these definitions are stated regarding (asymptotic) stability of the origin, they are easily translated to the definitions regarding the stability of any point $\overline{x}$ for which $g(\overline{x}, t) \equiv 0$ so that $\varphi(t, \overline{x}) = \overline{x}$ for all $t \geq 0$.*

2.7.1 Linear Systems

Consider now the linear system

$$\dot{x} = Ax, \quad t > 0 \tag{2.16a}$$

$$x(0) = x_0. \tag{2.16b}$$

Theorem 2.7.3. *Consider the system* (2.16) *The origin is asymptotically stable if and only if, for all eigenvalues λ of A,* $\operatorname{Re}\lambda < 0$.

Proof. **Only If.** Suppose that for an eigenvalue λ of A we have $\operatorname{Re}\lambda \geq 0$ and let z be the corresponding eigenvector. Note that $Az = \lambda z$ implies that $e^{At}z = e^{\lambda t}z$. Let $z = x + \mathbf{i}y$ and $\lambda = \alpha + \mathbf{i}\omega$ with $x, y \in \mathbb{R}^n$ and $\alpha, \omega \in \mathbb{R}$. Suppose first that $x \neq 0$. Since z is the eigenvector corresponding to λ, $\bar{z} = x - \mathbf{i}y$ is the eigenvector corresponding to[1] $\bar{\lambda} = \alpha - \mathbf{i}\omega$ and we can write

$$\begin{aligned} 2e^{At}x &= e^{At}(z + \bar{z}) = e^{\alpha t}(e^{\mathbf{i}\omega t}z + e^{-\mathbf{i}\omega t}\bar{z}) \\ &= 2e^{\alpha t}\big(x\cos\omega t - y\sin\omega t\big) \end{aligned}$$

which does not converge to zero as $t \to \infty$ if $\alpha \geq 0$. If $x = 0$, then instead we consider

$$\begin{aligned} -2e^{At}y &= \mathbf{i}e^{At}(z - \bar{z}) \\ &= -e^{\alpha t}\big(y\exp(\mathbf{i}\omega t) + y\exp(-\mathbf{i}\omega t)\big) \end{aligned}$$

and again this does not converge to zero as $t \to \infty$ if $\alpha \geq 0$.

If This follows directly from Corollary 2.5.5(i). □

Remarks 2.7.4. *The proof reveals that, for the linear system* (2.16)*, if all eigenvalues have negative real parts then the origin is in fact* exponentially *stable.*

Recall Theorem 2.3.8 and recall that ℓ_i is the number of generalized eigenvectors corresponding to eigenvalue λ_i of A. If $\ell_i = 1$ we simply have an eigenvector and no other generalized eigenvectors.

Theorem 2.7.5. *Consider the system* (2.16)*. The origin is stable if, for all eigenvalues λ_i of A,* $\operatorname{Re}\lambda_i \leq 0$ *and, any eigenvalue with* $\operatorname{Re}\lambda_i = 0$ *has no generalized eigenvectors associated with it, apart from the eigenvector itself: $\ell_i = 1$. Conversely, if the origin is stable then* $\operatorname{Re}\lambda_i \leq 0$ *for all i.*

[1] complex conjugate $\bar{\lambda}$ not to be confused with $\overline{\lambda}$ defined in Corollary 2.5.5.

Proof. The first part of the theorem is a consequence of Corollary 2.5.5(ii). For the second part, assume that there is i such that $\operatorname{Re}\lambda_i > 0$ and that the corresponding eigenvector is $v^{(i)}$. For simplicity we assume that both the eigenvalue and eigenvector are real. Then with $x(0) = v^{(i)}$ we have $e^{At}x(0) = e^{\lambda_i t}x(0) \to \infty$ as $t \to \infty$ proving instability. (The case of complex eigenvalue is similar). □

Example 2.7.6. *Consider the matrix*

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

This has a single eigenvalue of zero, with eigenvector $(1, 0)$ *and generalized eigenvector* $(0, 1)$*. Neither of the preceding two theorems apply and indeed the system* (2.16) *is unstable. To see this note that* $\ddot{x_1} = 0$ *so that* $x_1(t)$ *grows linearly with* t*.*

2.7.2 Lyapunov Functions

In this section we consider the autonomous system

$$\dot{x} = g(x) \tag{2.17a}$$

$$x(0) = x_0 \tag{2.17b}$$

with $x \in \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n$.

Definition 2.7.7. *Let* $G \subset \mathbb{R}^n$ *be a neighbourhood of* 0 *and* $V : G \to \mathbb{R}$ *a continuously differentiable function. Then* V *is a* Lyapunov function *if:*

- $V(x) > 0$ *for any* $x \in G\backslash\{0\}$ *and* $V(0) = 0$;
- $\dot{V}(x) := \langle DV(x), g(x)\rangle \leq 0 \quad \forall x \in G.$

If, in addition,

$$\dot{V}(x) = \langle DV(x), g(x)\rangle < 0 \quad \forall x \in G\backslash\{0\},$$

then V *is a* strict Lyapunov function.

Notice that $\dot{V}\big(x(t)\big) = \frac{d}{dt}\Big(V\big(x(t)\big)\Big)$ so that a (strict) Lyapunov function is a non-negative function which is zero only at the origin and is non-increasing (descreasing) along trajectories. Intuitively this links it to stability and the remainder of this section substantiates this intuition.

Lemma 2.7.8. *Let $B(0;R)$ be a ball of radius R centred at the origin in $\mathbb{R}^n$. Assume that the function $W : \overline{B(0;R)} \to \mathbb{R}$ is continuous with $W(0) = 0$ and $W(x) > 0$ for any $x \in \overline{B(0;R)}\backslash\{0\}$. Then there are increasing functions $\alpha : \mathbb{R} \to \mathbb{R}, \beta : \mathbb{R} \to \mathbb{R}$ with $\alpha(0) = \beta(0) = 0$ and $\alpha(r) > 0, \beta(r) > 0$ for $r > 0$ such that*

$$\alpha(\|x\|) \leq W(x) \leq \beta(\|x\|).$$

Proof. For any $x \in \overline{B(0;R)}$ with $\|x\| = r \leq R$ we have

$$\min_{r\leq\|y\|\leq R} W(y) \leq W(x) \leq \max_{\|y\|\leq r} W(y)$$

and both $\alpha(r) := \min_{r\leq\|y\|\leq R} W(y)$ and $\beta(r) = \max_{\|y\|\leq r} W(y)$ are increasing functions of r. Thus the lemma is proved. □

Theorem 2.7.9. *Consider the ODE* (2.17)*. Then:*

- *(i) the origin is stable if a Lyapunov function exists on a neighbourhood G of the origin;*
- *(ii) the origin is asymptotically stable if a strict Lyapunov function exists on a neighbourhood G of the origin.*

Proof. (i) We first observe that

$$\frac{d}{dt}\Big(V\big(x(t)\big)\Big) = \langle DV\big(x(t)\big), \frac{dx}{dt}(t)\rangle = \langle DV\big(x(t)\big), g(x(t)\big) \leq 0$$

so that V is nonincreasing along trajectories. Now choose R to ensure $B(0;R) \subset G$ and $\alpha(\cdot), \beta(\cdot)$ as in Lemma 2.7.8 so that

$$\alpha(\|x\|) \leq V(x) \leq \beta(\|x\|) \quad \forall x \in B(0;R).$$

For arbitrary $\epsilon \in (0, R)$ choose δ such that $\beta(\delta) < \alpha(\epsilon)$. Let $\|x_0\| < \delta$. Then we have, using the fact that V is nonincreasing along trajectories to obtain the second inequality,

$$\begin{aligned}\alpha\big(\|\varphi(t,x_0)\|\big) &\le V\big(\varphi(t,x_0)\big)\\ &\le V(x_0)\\ &\le \beta\big(\|x_0\|\big)\\ &\le \beta(\delta)\\ &\le \alpha(\epsilon).\end{aligned}$$

Thus $\alpha\big(\|\varphi(t,x_0)\|\big) < \alpha(\epsilon)$ and since α is increasing this implies that $\|\varphi(t,x_0)\| < \epsilon$.

(ii) By assumption $-\dot{V}(x) > 0$ for any $x \in G\backslash\{0\}$. Applying the result of Lemma 2.7.8 to the positive function $-\dot{V}$ on $B(0;R)\backslash\{0\}$ we have that there exists γ increasing and such that

$$-\dot{V}(x) \ge \gamma\big(\|x\|\big) \quad \forall x \in B(0;R). \tag{2.18}$$

Choose δ as in part (i) so that $\|\varphi(t,x_0)\| < \epsilon$. To show asymptotic stability we need to show that, for any $\hat{\epsilon} \in (0,\epsilon)$, $\exists \tau = \tau(\hat{\epsilon}) > 0$ such that $\|\varphi(t,x_0)\| < \hat{\epsilon}$ for any $t \ge \tau$. Choose $\hat{\delta}$ so that $\beta(\hat{\delta}) < \alpha(\hat{\epsilon})$. We need to show the existence of $\hat{t} \ge 0$ such that $\varphi(\hat{t},x_0) \in B(0;\hat{\delta})$ because then an argument similar to that used in (i) shows that $\varphi(t,x_0) \in B(0;\hat{\epsilon})$ for all $t > \hat{t}$. Suppose, for contradiction, that such a $\hat{t}$ does not exist, so that $\|\varphi(t,x_0)\| > \hat{\delta}$ for all $t \ge 0$. By (2.18) we have

$$\begin{aligned}V\big(\varphi(\tau,x_0)\big) - V(x_0) &= \int_0^\tau \dot{V}(s)ds\\ &\le -\int_0^\tau \gamma\Big(\|\varphi(s,x_0)\|\Big)ds.\end{aligned}$$

Therefore

$$\begin{aligned}V\big(\varphi(\tau,x_0)\big) &\le V(x_0) - \int_0^\tau \gamma\Big(\|\varphi(s,x_0)\|\Big)ds\\ &\le \beta\big(\|x_0\|\big) - \int_0^\tau \gamma(\hat{\delta})ds\\ &\le \beta(\delta) - \tau\gamma(\hat{\delta}).\end{aligned}$$

Now choose τ so that $\tau\gamma(\hat{\delta}) > \beta(\delta)$ and, since $V \ge 0$, the desired contradiction follows. □

Definition 2.7.10. *The domain of attraction of the equilibrium point* 0 *of* (2.17) *is*

$$\mathcal{A} = \{x \in \mathbb{R}^n : \varphi(t,x) \to 0 \text{ as } t \to \infty\}.$$

Corollary 2.7.11. *Let V be a strict Lyapunov function on domain $B(0;R) \subset \mathbb{R}^n$, for* (2.17). *Define $G_\rho = \{x \in B(0;R) : V(x) < \rho\}$, for $\rho < \infty$, and suppose that $\overline{G_\rho}$ is compact and that $G_\rho \subset \operatorname{Int} B(0;R)$. Then $G_\rho \subset \mathcal{A}$.*

Proof. Let $x_0 \in G_\rho$. We have

$$\alpha\big(\|\varphi(t,x_0)\|\big) \leq V\big(\varphi(t,x_0)\big) \leq V(x_0) < \rho.$$

Choose $r < R$ so that $\alpha(r) = \rho$ noting that then $\varphi(t,x_0) \in B(0;R)$ for any $t \geq 0$. With γ defined as in the proof of Theorem 2.7.9. Then we have

$$\begin{aligned} V\big(\varphi(t,x_0)\big) - V(x_0) &\leq -\int_0^t \dot{V}(s)ds \\ &= -\int_0^t \gamma\big(\varphi(s,x_0)\big)ds. \end{aligned}$$

To show asymptotic stability we need to show that, for any $\delta > 0$ there is $\tau = \tau(\delta) > 0$ such that $\varphi(t,x_0) \in B(0;\delta)$. Again we suppose for contradiction that such a τ does not exist. But

$$\begin{aligned} V\big(\varphi(t,x_0)\big) &\leq V(x_0) - \int_0^t \gamma(\delta)ds \\ &< \rho - \gamma(\delta)t. \end{aligned}$$

Choosing $t > \rho/\gamma(\delta)$ gives a contradiction. □

Remarks 2.7.12. *To understand the compactness requirement in the preceding corollary, consider the case*

$$V(x) = \frac{x_1^2}{1+x_1^2} + x_2^2.$$

The region $\{x \in \mathbb{R}^2 : V(x) \leq \rho\}$ is compact for $\rho < 1$, but not for $\rho > 1$. The figure shows that an initial state can diverge.

The preceding corollary implies the following:

Theorem 2.7.13. *Suppose that a strict Lyapunov function exists for* (2.17) *on $\mathbb{R}^n$ and that $V(x) \to \infty$ as $|x| \to \infty$ (i.e. V is radially unbounded). Then $x = 0$ is globally asymptotically stable.*

Example 2.7.14. *Consider the nonlinear model of a damped pendulum (with no control) so that, for $\theta \in (-\pi/2, \pi/2)$,*

$$\ddot{\theta} + \dot{\theta} + \sin(\theta) = 0.$$

Writing $x_1 = \theta$ and $x_2 = \dot{\theta}$ we obtain the system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\sin x_1 - x_2. \end{aligned}$$

Consider $V(x) = \big(1 - \cos(x_1)\big) + \frac{1}{2}x_2^2$. Then

$$\begin{aligned} \dot{V}(x) &= \dot{x}_1 \sin(x_1) + \dot{x}_2 x_2 \\ &= x_2 \sin(x_1) + x_2\big(-\sin(x_1) - x_2\big) \\ &= -x_2^2 \\ &\le 0. \end{aligned}$$

Thus the origin is stable, by Theorem 2.7.9(i). The damping term $-x_2^2$ suggests asymptotic stability, but does not prove it.

Example 2.7.15. *Consider the equation*

$$\dot{x} = -x(1 - x)$$

and set $V(x) = \frac{1}{2}x^2$. Then

$$\dot{V}(x) = x\dot{x} = -x^2(1 - x)$$

and we see that $\dot{V}(x) < 0$ for $|x| \in (0,1)$. Taking $G = (-1,1)$ we deduce, using Theorem 2.7.9(ii), that the origin is asymptotically stable and taking $G_{\frac{1}{2}} = \{x \in G : V(x) < \frac{1}{2}\}$ we deduce from Corollary 2.7.11 that $(-1,1) \subset \mathcal{A}$.

Remarks 2.7.16. *By techniques similar to those in the proof of Theorem 2.7.9 we can show the following (see Exercise 2-12). Let $V : G \to \mathbb{R}$ be a continuously differentiable function with $V(x) > 0$ for $x \in G \backslash \{0\}$ and $V(0) = 0$. If $\dot{V}(x) > 0$ for any $x \in G \backslash \{0\}$ then $x = 0$ is unstable.*

2.7.3 Lyapunov Function for Linear Systems

We now consider the linear version of (2.17), obtaining the equation

$$\dot{x} = Ax \tag{2.19a}$$

$$x(0) = x_0 \tag{2.19b}$$

with $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$.

In order to try and identify a Lyapunov function, set

$$V(x) = \langle x, Px\rangle \tag{2.20}$$

with $P \in \mathbb{R}^{n\times n}$ being symmetric positive definite. Then $V(x) > 0$ for $x \in \mathbb{R}^n\backslash\{0\}$ and

$$\begin{aligned}\langle DV(x), Ax\rangle &= \langle Ax, Px\rangle + \langle x, PAx\rangle \\ &= \langle x, (A^TP + PA)x\rangle.\end{aligned}$$

(Note that we can also write this as

$$\langle DV(x), Ax\rangle = 2\langle x, PAx\rangle$$

but that the matrix PA is not then symmetric.) If we define

$$Q = A^TP + PA \tag{2.21}$$

then, to use Lyapunov stability theory, we would like to know when there is positive-definite symmetric P satisfying this equation, for some negative definite matrix Q.

Theorem 2.7.17. *Let $Q \in \mathbb{R}^{n\times n}$ be a negative definite matrix. Then there exists positive definite $P \in \mathbb{R}^{n\times n}$ solving the* Lyapunov equation (2.21) *if and only if the origin is asymptotically stable for* (2.19).

Proof. Fix negative-definite Q and, to prove the "only if" part assume that there is a positive-definite solution $P \in \mathbb{R}^{n\times n}$ of (2.21). Then $V(x)$ given by (2.20) is a strict Lyapunov function for (2.19) and, by Theorem 2.7.9, $x = 0$ is asymptotically stable.

Now fix negative-definite Q and assume that $x = 0$ is asymptotically stable, so that all eigenvalues of A have negative real part, by Theorem 2.7.3. Define

$$P = -\int_0^\infty e^{A^T t} Q e^{At} dt. \tag{2.22}$$

The matrix P is well-defined by this expression because, by Corollary 2.5.5, noting that A^T has the same eigenvalues as A,

$$|e^{A^T t} Q e^{At}| \leq c|Q| \exp(\overline{\lambda} t) \tag{2.23}$$

where, recall, $\overline{\lambda} := \max\{\operatorname{Re}\lambda : \lambda \,\text{is an eigenvalue of}\, A\} < 0$. Note that P is positive-definite since e^{At} is invertible and $-Q$ is positive-definite so that, for some $\alpha > 0$

$$\begin{aligned}
\langle v, Pv \rangle &= -\int_0^\infty \langle v, e^{A^T t} Q e^{At} v \rangle dt \\
&= -\int_0^\infty \langle e^{At} v, Q e^{At} v \rangle dt \\
&\geq \alpha \int_0^\infty |e^{At} v|^2 dt \\
&> 0.
\end{aligned}$$

Furthermore, we have, using (2.23) in the last line,

$$\begin{aligned}
A^T P + PA &= -\int_0^\infty \Big(A^T e^{tA^T} Q e^{tA} + e^{tA^T} Q e^{tA} A \Big) dt \\
&= -\int_0^\infty \frac{d}{dt} \Big(e^{tA^T} Q e^{tA} \Big) dt \\
&= Q - \lim_{t \to \infty} e^{tA^T} Q e^{tA} \\
&= Q.
\end{aligned}$$

□

Corollary 2.7.18. *For any $Q \in \mathbb{R}^{n \times n}$ there is a unique solution P of equation* (2.21) *if all eigenvalues of A have negative real part.*

Proof. See Exercise 2-8. Note Q is not required to be negative-definite, hence P will not be positive definite in general. □

Example 2.7.19. *Consider the linear system*

$$\ddot{\xi} + \dot{\xi} + \xi = 0.$$

Let $x_1 = \xi$ *and* $x_2 = \dot{\xi}$. *We have*

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 - x_2. \end{aligned}$$

which is in the form of (2.19) *with*

$$A = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}.$$

Let $Q = -I$. *Then the Lyapunov equation* (2.21) *becomes*

$$\begin{pmatrix} p_1 & p_2 \\ p_2 & p_3 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} p_1 & p_2 \\ p_2 & p_3 \end{pmatrix} = - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Multiplying the matrices gives

$$\begin{pmatrix} -p_2 & p_1 - p_2 \\ -p_3 & p_2 - p_3 \end{pmatrix} + \begin{pmatrix} -p_2 & -p_3 \\ p_1 - p_2 & p_2 - p_3 \end{pmatrix} = - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

From the symmetry this gives the three independent equations

$$\begin{aligned} 2p_2 &= 1 \\ p_1 - p_2 - p_3 &= 0 \\ 2p_2 - 2p_3 &= -1 \end{aligned}$$

with solution $p_1 = 3/2, p_2 = \frac{1}{2}, p_3 = 1$.

Thus we deduce that

$$V(x) = \frac{3}{2}x_1^2 + x_1x_2 + x_2^2$$

is a Lyapunov function. To see that it is positive away from the origin note that

$$V(x) = \frac{5}{4}x_1^2 + \big(\frac{1}{2}x_1 + x_2\big)^2$$

showing that $V(x) > 0$ *for* $x \neq 0$. *Note also that*

$$\begin{aligned} \dot{V}(x) &= 3x_1\dot{x}_1 + x_1\dot{x}_2 + \dot{x}_1x_2 + 2x_2\dot{x}_2 \\ &= 3x_1x_2 - x_1\big(x_1 + x_2\big) + x_2^2 - 2x_2\big(x_1 + x_2\big) \\ &= -x_1^2 - x_2^2 \end{aligned}$$

so that $V\big(x(t)\big)$ *is deccreasing along trajectories, away from the origin. Of course the fact that* $\dot{V}(x) = -|x|^2$ *is a consequence of our choice* $Q = -I$.

2.7.4 Linearization

Consider the nonlinear system (2.17) and assume that $g(\overline{x}) = 0$.

Theorem 2.7.20. *Assume that g is differentiable at $\overline{x}$. Then the system (2.17) is asymptotically stable at $x = \overline{x}$ if all eigenvalues of the matrix $A = Dg(\overline{x})$ have strictly negative real parts.*

Proof. Without loss of generality we may take $\overline{x} = 0$ (by shifting the origin). Define $h(x) = g(x) - Ax$, noting that $h : \mathbb{R}^n \to \mathbb{R}^n$. By differentiability of g at 0 we have

$$\lim_{\|x\|\to 0} \frac{\|h(x)\|}{\|x\|} = 0. \tag{2.24}$$

Since the linear system (2.19) with $A = Dg(0)$ is asymptotically stable there exists a positive-definite matrix $P > 0$ such that

$$PA + A^T P = -I. \tag{2.25}$$

Define $V(x) = \langle x, Px\rangle$. Then

$$\begin{aligned} \dot{V}(x) &= \langle g(x), Px\rangle + \langle x, Pg(x)\rangle \\ &= \langle Ax + h(x), Px\rangle + \langle x, P\big(Ax + h(x)\big)\rangle \\ &= \langle x, \big(A^T P + PA\big)x\rangle + 2\langle x, Ph(x)\rangle. \end{aligned}$$

Thus, by (2.25),

$$\dot{V}(x) \le -\|x\|^2 + 2\|x\|\|P\|\|h(x)\|.$$

By (2.24) we deduce that, for any $\epsilon > 0$ there is $\delta > 0$ such that $\|x\| < \delta$ implies $\|h(x)\| < \epsilon\|x\|$. In the set $G = \{\|x\| < \delta\}$ we have

$$\begin{aligned} \dot{V}(x) &< -\|x\|^2 + 2\epsilon\|P\|\|x\|^2 \\ &= -\big(1 - 2\epsilon\|P\|\big)\|x\|^2 \end{aligned}$$

so that by choosing $\epsilon < (2\|P\|)^{-1}$ we get $\dot{V}(x) < 0$ and the result follows by Theorem 2.7.9. □

Remarks 2.7.21. *If (2.17) is exponentially stable at 0 then all eigenvlaues of $A = Dg(0)$ have negative real part. This may be shown by substituting $y(t) = e^{-\eta t}x(t)$ for $\eta \in (0, \omega)$ in (2.17).*

Example 2.7.22. *Consider the nolinear system*

$$\ddot{\xi} + \dot{\xi} - \xi + \xi^2 = 0.$$

Let $x_1 = \xi$ *and* $x_2 = \dot{\xi}$. *We then have*

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= x_1 - x_2 - x_1^2.\end{aligned}$$

This is in the form of (2.17) *with*

$$g(x) = \begin{pmatrix} x_2 \\ x_1 - x_2 - x_1^2 \end{pmatrix}.$$

There are two equilibrium points: $x = (0,0)$ *and* $x = (1,0)$. *The derivative of* g *is*

$$Dg(x) = \begin{pmatrix} 0 & 1 \\ 1 - 2x_1 & -1 \end{pmatrix}.$$

At the equilibrium point $x = (0,0)$ *we have*

$$Dg(x) = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$$

with eigenvalues $\lambda^{\pm} = \frac{1}{2}\big(-1 \pm \sqrt{5}\big)$ *and Theorem 2.7.20 does not apply. On the other hand, at the equilibrium point* $x = (1,0)$ *we have*

$$Dg(x) = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}$$

with eigenvalues $\lambda^{\pm} = \frac{1}{2}\big(-1 \pm i\sqrt{5}\big)$ *and Theorem 2.7.20 gives asymptotic stability.*

2.8 Probability

In this section we use lowercase letters to denote both the random variable *and* the argument of the probability density functions (pdfs) for the random variable. The same letter with a † superscript represents a particular realization of the relevant random variable.

Throughout we consider random variables x on $\mathbb{R}^n$ and we assume that there is a *probability density function* (pdf) $\rho_x : \mathbb{R}^n \to \mathbb{R}^+$ such that, for any Borel set[1] $A \subseteq \mathbb{R}^n$,

$$\mathbb{P}(x \in A) = \int_A \rho_x(x)dx.$$

In words this formula expresses the probability that the random variable x lies in the set A. The pdf enables us to calculate this probability for arbitrary sets A. Necessarily we have

$$\int_{\mathbb{R}^n} \rho_x(x)dx = 1,$$

stating that, with probability one, x lies somewhere in $\mathbb{R}^n$.

For any function $f : \mathbb{R}^n \to \mathbb{R}^r$ we define the *expectation* of f to be

$$\mathbb{E}\big(f(x)\big) = \int_{\mathbb{R}^n} f(x)\rho_x(x)dx.$$

Two important special cases of this are as follows. We define the *mean* $\overline{x} \in \mathbb{R}^n$ by

$$\overline{x} = \mathbb{E}x = \int_{\mathbb{R}^n} x\rho_x(x)dx$$

and the *covariance* $C_x \in \mathbb{R}^{n\times n}$ by

$$C_x = \mathbb{E}(x-\overline{x})(x-\overline{x})^T = \int_{\mathbb{R}^n} (x-\overline{x})(x-\overline{x})^T \rho_x(x)dx.$$

Particularly important for us are the following random variables:

Definition 2.8.1. *Let $m \in \mathbb{R}^n$ and let $C \in \mathbb{R}^{n\times n}$ be a symmetric positive-definite matrix. Then $x \in \mathbb{R}^n$ is a* Gaussian random variable *with mean m and covariance matrix C if*

$$\rho_x(x) = \frac{1}{\sqrt{(2\pi)^n \det C}} \exp\Big(-\frac{1}{2}\langle (x-m), C^{-1}(x-m)\rangle\Big).$$

We write $x \sim N(m, C)$.

Remarks 2.8.2.

- *The matrix $L = C^{-1}$ is known as the* precision matrix *and is well-defined under our assumptions. However it is useful to allow for the case where C is not invertible.*

[1] A set which can be formed through countable union, countable intersection and complementation of open sets.

- *The degenerate case where $C = 0$ is known as a* Dirac mass *and corresponds to a random variable which takes the value m with probability one.*

- *The definition of Gaussian can be extended to positive semi-definite C by means of the* characteristic function. *Heuristically the measure in this case may be viewed as an independent product of Diracs and Gaussians with invertible covariance, in an appropriate coordinate system.*

- *A random variable on $\mathbb{R}^n$ is Gaussian if and only if the pdf may be written as*
$$\rho_x(x) \propto \exp\bigl(-J(x; m, C)\bigr)$$
where
$$J(x; m, C) = \frac{1}{2}\bigl|C^{-\frac{1}{2}}(x - m)\bigr|^2$$
for some vector $m \in \mathbb{R}^n$ and positive-definite matrix $C \in \mathbb{R}^{n \times n}$. Furthermore, the mean is then the minimizer over x of $J(x; m, C)$.

We now consider a joint random variable $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ with pdf $\rho_{x,y} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^+$. We define the *marginal* of x to be the random variable on $\mathbb{R}^n$ with pdf
$$\rho_x(x) = \int_{\mathbb{R}^m} \rho_{x,y}(x, y) dy.$$
Similarly we may define the marginal of y with pdf ρ_y given by
$$\rho_y(y) = \int_{\mathbb{R}^n} \rho_{x,y}(x, y) dx.$$

Of particular significance for us is the *conditional random variable* x given a single realization of the random variable y; we denote this realization by $y^\dagger$. This conditional random variable is a random variable on $\mathbb{R}^n$ with pdf $\rho_{x|y} : \mathbb{R}^n \to \mathbb{R}^+$ given by
$$\rho_{x|y}(x; y^\dagger) = \frac{\rho_{x,y}(x, y^\dagger)}{\rho_y(y^\dagger)}.$$
We denote this random variable by $x|y = y^\dagger$. Where it causes no confusion to do so we simply write $x|y$ as the relevant random variable and denote its density by $\rho_{x|y}(x; y)$. We then have
$$\rho_{x|y}(x; y) = \frac{\rho_{x,y}(x, y)}{\rho_y(y)}. \tag{2.26}$$

Similarly we may define the conditional random variable y given $x = x^\dagger$ with pdf $\rho_{y|x} : \mathbb{R}^m \to \mathbb{R}^+$ given by

$$\rho_{y|x}(y; x^\dagger) = \frac{\rho_{x,y}(x^\dagger, y)}{\rho_x(x^\dagger)}; \tag{2.27}$$

we denote this random variable by $y|x = x^\dagger$. Again, where it causes no confusion, we simply write the random variable as $y|x$ and denote its density by $\rho_{y|x}(y; x)$. Then we have

$$\rho_{y|x}(y; x) = \frac{\rho_{x,y}(x, y)}{\rho_x(x)}. \tag{2.28}$$

From these definitions it is easy to deduce:

Theorem 2.8.3. Bayes' Theorem *Assume that for $y \in \mathbb{R}^m$ the marginal density is non-zero: $\rho_y(y) \neq 0$. The pdf of the conditional random variable $x|y$ may be computed from the conditional random variable $y|x$ by means of the formula*

$$\rho_{x|y}(x; y) = \frac{\rho_{y|x}(y; x)\rho_x(x)}{\rho_y(y)}.$$

Proof. This follows directly from the fact that

$$\rho_{x,y}(x, y) = \rho_{y|x}(y; x)\rho_x(x)$$

and

$$\rho_{x,y}(x, y) = \rho_{x|y}(x; y)\rho_y(y).$$

□

Remarks 2.8.4. *We sometimes write Bayes' Theorem succinctly as*

$$\mathbb{P}(x|y) \propto \mathbb{P}(y|x)\mathbb{P}(x).$$

Also we will use this theorem when everything is conditioned on a third random variable z and it is then written succinctly as

$$\mathbb{P}(x|y, z) \propto \mathbb{P}(y|x, z)\mathbb{P}(x|z).$$

Example 2.8.5. *Bayes' Theorem is particularly useful for us when considering Gaussian distributions. In particular we will encounter the situation where a joint random variable $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ is specified as follows: we assume that the marginal distribution of x is a Gaussian $N(m_0, C_0)$; and that the conditional distribution of $y|x$ is a Gaussian $N(Hx, \Gamma)$. For simplicity we will assume that C_0 and Γ are positive-definite in this scenario.*

Theorem 2.8.6. *Consider the joint random variable (x, y) as specified in Example 2.8.5, with C_0 and Γ positive-definite. Then:*

- *(i) the joint random variable (x, y) is itself Gaussian with positive-definite covariance Σ and mean a given by the formulae*
$$\Sigma^{-1}a = r, \quad r = \begin{pmatrix} C_0^{-1}m_0 \\ 0 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} C_0^{-1} + H^T\Gamma^{-1}H & -H^T\Gamma^{-1} \\ -\Gamma^{-1}H & \Gamma^{-1} \end{pmatrix};$$
- *(ii) the conditional random variable $x|y$ is also Gaussian and has positive-definite covariance C' and mean m' given by the formulae*
$$(C')^{-1} = C_0^{-1} + H^T\Gamma^{-1}H$$
and
$$(C')^{-1}m' = C_0^{-1}m_0 + H^T\Gamma^{-1}y.$$

Proof. Rearranging (2.28) we deduce that
$$\rho_{x,y}(x, y) = \rho_{y|x}(y; x)\rho_x(x).$$
Thus we see that
$$\rho_{x,y}(x, y) \propto \exp\Big(-J(x, y)\Big) \tag{2.29}$$
where
$$J(x, y) = \frac{1}{2}|\Gamma^{-\frac{1}{2}}(y - Hx)|^2 + \frac{1}{2}|C_0^{-\frac{1}{2}}(x - m_0)|^2.$$
We aim to write that as a quadratic form in $z = (x^T, y^T)^T$ in order to identify the relevant mean and covariance. Expanding the quadratic forms gives
$$\begin{aligned} J(x, y) =& \frac{1}{2}\langle x, (C_0^{-1} + H^T\Gamma^{-1}H)x\rangle + \frac{1}{2}\langle y, \Gamma^{-1}y\rangle - \frac{1}{2}\langle x, H^T\Gamma^{-1}y\rangle - \frac{1}{2}\langle y, \Gamma^{-1}Hx\rangle \\ & - \langle x, C_0^{-1}m_0\rangle + \frac{1}{2}\langle m_0, C_0^{-1}m_0\rangle \\ =& \frac{1}{2}\langle z, \Sigma^{-1}z\rangle - \langle x, C_0^{-1}m_0\rangle + \frac{1}{2}\langle m_0, C_0^{-1}m_0\rangle \\ =& \frac{1}{2}\langle (z - a), \Sigma^{-1}(z - a)\rangle + \frac{1}{2}\langle m_0, C_0^{-1}m_0\rangle - \frac{1}{2}\langle a, \Sigma^{-1}a\rangle. \end{aligned}$$

Here r, Σ are as defined in the theorem statement. It can be shown (see Exercise 2-14) that Σ^{-1} is positive-definite and symmetric since Γ and C_0 are positive-definite and symmetric. Hence the preceding equation for a is well-defined. Furthermore, we deduce that

$$\rho_{x,y}(x,y) \propto \exp\Bigl(-\frac{1}{2}\bigl\langle (z-a), \Sigma^{-1}(z-a)\bigr\rangle\Bigr),$$

with constant of proportionality independent of both x and y, from which it follows that (x, y) is Gaussian $N(a, \Sigma)$.

A similar approach shows that $x|y$ is also Gaussian. From (2.29) and (2.26) we deduce that the required density for $\rho_{x|y}(x, y)$ is also proportional to $\exp\bigl(-J(x, y)\bigr)$, with constant of proportionality depending on y, but not on x. Thus we now want to write $J(x, y)$ as a quadratic form in x alone in order to determine the relevant Gaussian distribution. For this we note that

$$\begin{aligned} J(x,y) =& \frac{1}{2}\langle x, (C_0^{-1} + H^T\Gamma^{-1}H)x\rangle - \langle x, H^T\Gamma^{-1}y + C_0^{-1}m_0\rangle \\ &+ \frac{1}{2}\langle y, \Gamma^{-1}y\rangle + \frac{1}{2}\langle m_0, C_0^{-1}m_0\rangle \\ =& \frac{1}{2}\langle x, (C')^{-1}x\rangle - \langle x, H^T\Gamma^{-1}y + C_0^{-1}m_0\rangle + \frac{1}{2}\langle m_0, C_0^{-1}m_0\rangle + \frac{1}{2}\langle y, \Gamma^{-1}y\rangle \\ =& \frac{1}{2}\langle (x-m'), (C')^{-1}(x-m')\rangle + \frac{1}{2}\langle m_0, C_0^{-1}m_0\rangle + \frac{1}{2}\langle y, \Gamma^{-1}y\rangle - \frac{1}{2}\langle m', (C')^{-1}m'\rangle \end{aligned}$$

where C' and m' are as defined in the theorem statement. It may be shown (Exercise 2-14) that C' is positive-definite and symmetric since Γ and C_0 are positive-definite and symmetric, and hence that C' and m' are well-defined. By means of Theorem 2.8.3 we deduce that

$$\rho_{x|y}(x;y) \propto \exp\Bigl(-\frac{1}{2}\langle (x-m'), (C')^{-1}(x-m')\rangle\Bigr),$$

with constant of proportionality independent of x, from which it follows that $x|y$ is Gaussian $N(m', C')$. □

Theorem 2.8.7. *If $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ is Gaussian then the marginal distribution of x is also Gaussian. Furthermore, if the joint random variable (x, y) has precision matrix L with the block form*

$$L = \begin{pmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{pmatrix}$$

then the matrices $L_{xx} \in \mathbb{R}^{n\times n}$ and $L_{yy} \in \mathbb{R}^{m\times m}$ are both symmetric and positive-definite, and $L_{yx} = L_{xy}^T \in \mathbb{R}^{m\times n}$. Furthermore, the mean of the marginal distribution is $\mathbb{E}x$ and the precision matrix is $L_{xx} - L_{xy}L_{yy}^{-1}L_{yx}$.

Proof. The fact that the matrices L_{xx}, L_{yy} are both symmetric and positive-definite is the subject of Exercise 2-15. $L_{yx} = L_{xy}^T$ since L symmetric.

We first consider the case where (x, y) has mean zero. Thus the pdf has the form, for $C = L^{-1}$,

$$\rho_{x,y}(x, y) = \frac{1}{\sqrt{(2\pi)^{n+m}\det C}} \exp\bigl(-J(x, y)\bigr)$$

where, for $L_{yy}h = L_{yx}x$ we have

$$\begin{aligned} J(x, y) &= \frac{1}{2}\langle x, L_{xx}x\rangle + \langle L_{yx}x, y\rangle + \frac{1}{2}\langle y, L_{yy}y\rangle \\ &= \frac{1}{2}\langle x, L_{xx}x\rangle + \frac{1}{2}\langle (y + h), L_{yy}(y + h)\rangle - \frac{1}{2}\langle h, L_{yy}h\rangle. \end{aligned}$$

From this we see that

$$\begin{aligned} \int_{\mathbb{R}^m} \rho_{x,y}(x, y)dy &= \int_{\mathbb{R}^m} \frac{1}{\sqrt{(2\pi)^{n+m}\det C}} \exp\bigl(-J(x, y)\bigr)dy \\ &= \sqrt{\frac{(2\pi)^m \det L_{yy}^{-1}}{(2\pi)^{n+m}\det C}} \exp\bigl(-\frac{1}{2}\langle x, L_{xx}x\rangle + \frac{1}{2}\langle h, L_{yy}h\rangle\bigr) I \end{aligned}$$

where

$$\begin{aligned} I &= \int_{\mathbb{R}^m} \frac{1}{\sqrt{(2\pi)^m \det L_{yy}^{-1}}} \exp\bigl(-\frac{1}{2}\langle (y + h), L_{yy}(y + h)\rangle\bigr)dy \\ &= 1. \end{aligned}$$

From this it follows that the marginal distribution of x is a mean zero Gaussian with precision operator $L_{xx} - L_{xy}L_{yy}^{-1}L_{yx}$.

If (x, y) has mean (a, b) then a similar calculation shows that

$$\int_{\mathbb{R}^m} \rho_{x,y}(x, y)dy = \sqrt{\frac{(2\pi)^m \det L_{yy}^{-1}}{(2\pi)^{n+m}\det C}} \exp\bigl(-\frac{1}{2}\langle x', L_{xx}x'\rangle + \frac{1}{2}\langle h', L_{yy}h'\rangle\bigr)$$

where $x' = x - a$ and $L_{yy}h' = L_{yx}(x - a)$. From this it follows that the marginal distribution of x is a Gaussian with mean a and precision operator $L_{xx} - L_{xy}L_{yy}^{-1}L_{yx}$. □

Remarks 2.8.8. *We prove Theorem 2.8.7 in the case of positive-definite covariance. However it extends to the positive semi-definite case.*

Theorem 2.8.9. *Let $z \in \mathbb{R}^p$ be a Gaussian random variable with mean $m_z \in \mathbb{R}^p$ and covariance $C_z \in \mathbb{R}^{p\times p}$. Let $w = a + Az$ for some $a \in \mathbb{R}^q$ and $A \in \mathbb{R}^{q\times p}$. Then $w \in \mathbb{R}^q$ is a Gaussian random variable with mean $m_w = a + Am_z$ and covariance $C_w = AC_zA^T$.*

Proof. The fact that an affine linear transformation preserves Gaussianity may be checked directly by using change of variable formula for pdf's. It is then clear that, since expectation is a linear operation,

$$\begin{aligned} m_w &= \mathbb{E}w \\ &= \mathbb{E}(a + Az) \\ &= a + A\mathbb{E}(z) \\ &= a + Am_z. \end{aligned}$$

Similarly

$$\begin{aligned} C_w &= \mathbb{E}\big((w - m_w)(w - m_w)^T\big) \\ &= \mathbb{E}\big(A(z - m_z)(z - m_z)^T A^T\big) \\ &= A\mathbb{E}\big((z - m_z)(z - m_z)^T\big)A^T \\ &= AC_zA^T. \end{aligned}$$

□

Two random variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are said to be *independent* if the pdf of the random variable $z = (x, y) \in \mathbb{R}^{n+m}$ factorizes as the product of the two pdf's of x and y:

$$\rho_{x,y}(x, y) = \rho_x(x)\rho_y(y). \tag{2.30}$$

More generally a set of K independent random variables x_k will have pdf which factors as the product of the K densities for each x_k.

A sequence of random variables $\{x_i\}_{i\in\mathbb{N}}$ is said to be *independent, identically distributed*, i.i.d. for short, if x_i and x_j have the same probability distribution for all i and j and are independent for $i \neq j$.

Two random variables $x, y \in \mathbb{R}^n$ are *uncorrelated* if $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$.

Theorem 2.8.10. *Two Gaussian random variables $x, y \in \mathbb{R}^n$ are independent if and only if they are* uncorrelated.

The following result concerning block matrix inversion, with application to conditioned Gaussians, is very useful.

Lemma 2.8.11. *Consider a positive-definite matrix C with the block form*

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^T & C_{22} \end{pmatrix}.$$

Then C_{22} is positive-definite symmetric and the Schur complement *S defined by $S = C_{11} - C_{12}C_{22}^{-1}C_{12}^T$ is positive-definite symmetric. Furthermore*

$$C^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}C_{12}C_{22}^{-1} \\ -C_{22}^{-1}C_{12}^T S^{-1} & C_{22}^{-1} + C_{22}^{-1}C_{12}^T S^{-1}C_{12}C_{22}^{-1} \end{pmatrix}.$$

Now let (x, y) be jointly Gaussian with distribution $N(m, C)$ and $m = (m_1^T, m_2^T)^T$. Then the conditional distribution of $x|y$ is Gaussian with mean m' and covariance matrix C' given by

$$m' = m_1 + C_{12}C_{22}^{-1}(y - m_2),$$
$$C' = C_{11} - C_{12}C_{22}^{-1}C_{12}^T.$$

Proof. To see that C_{22} is positive-definite let $\xi = (0^T, \xi_2^T)^T$ and note that, for $\xi_2 \neq 0$,

$$0 < \langle \xi, C\xi \rangle = \langle \xi_2, C_{22}\xi_2 \rangle.$$

To see that S is positive-definite assume for contradiction that

$$\langle \xi_1, C_{11}\xi_1 - C_{12}C_{22}^{-1}C_{12}^T\xi_1 \rangle \leq 0$$

for some non-zero ξ_1. Now let

$$\xi = \big(\xi_1^T, (-C_{22}^{-1}C_{12}^T\xi_1)^T\big)^T.$$

Then

$$\langle \xi, C\xi \rangle = \langle \xi_1, C_{11}\xi_1 - C_{12}C_{22}^{-1}C_{12}^T\xi_1 \rangle \leq 0,$$

a contradiction. Thus C^{-1} as written is well-defined. Multiplication establishes that it is indeed the inverse of C.

Now we turn to the conditioning result. Define

$$C^{-1} := L = \begin{pmatrix} L_{11} & L_{12} \\ L_{12}^T & L_{22} \end{pmatrix}.$$

We first prove the result for $m = 0$. Note that L_{11} is positive-definite symmetric since L is, by a similar argument to that showing that C_{22} is positive. The pdf for (x, y) is proportional to $\exp\bigl(-J(x, y)\bigr)$ where

$$\begin{aligned}
J(x,y) =& \frac{1}{2}\langle x, L_{11}x\rangle + \langle x, L_{12}y\rangle + \frac{1}{2}\langle y, L_{22}y\rangle, \\
=& \frac{1}{2}\langle (x + L_{11}^{-1}L_{12}y), L_{11}(x + L_{11}^{-1}L_{12}y)\rangle \\
&+ \frac{1}{2}\langle y, L_{22}y\rangle - \frac{1}{2}\langle L_{12}y, L_{11}^{-1}L_{12}y\rangle.
\end{aligned}$$

It follows that $x|y$ has mean $-L_{11}^{-1}L_{12}y$ and covariance $C' = L_{11}^{-1}$. Shifting $x \mapsto x - m_1$ and $y \mapsto y - m_2$ gives the desired result for the mean and covariance expressed in terms of L. To express the covariance in terms of C we use the first conclusion of the lemma to deduce that $L_{11} = S^{-1}$so that

$$C' = L_{11}^{-1} = S = C_{11} - C_{12}C_{22}^{-1}C_{12}^T.$$

For the mean note that, also by the first part of the lemma,

$$L_{11}^{-1}L_{12} = -SS^{-1}C_{12}C_{22}^{-1}$$

as required. □

We now apply this Lemma to Example 2.8.5.

Example 2.8.12. *Consider the jointly Gaussian random variable (x, y) constructed in Example 2.8.5, with C_0 and Γ positive-definite. Thus $x \sim N(m_0, C_0)$ and $y|x \sim N(Hx, \Gamma)$. Theorem 2.8.6 shows that, provided C_0 and Γ are positive-definite, $x|y$ is Gaussian with mean m' and covariance C' given by*

$$(C')^{-1} = C_0^{-1} + H^T\Gamma^{-1}H \tag{2.31a}$$

$$(C')^{-1}m' = C_0^{-1}m_0 + H^T\Gamma^{-1}y. \tag{2.31b}$$

Here we will show that, provided Γ is invertible, the covariance and mean of $x|y$ are given by

$$C' = C_0 - C_0H^T(\Gamma + HC_0H^T)^{-1}HC_0 \tag{2.32a}$$

$$m' = m_0 + C_0H^T(\Gamma + HC_0H^T)^{-1}(y - Hm_0). \tag{2.32b}$$

The statement $y|x \sim N(Hx, \Gamma)$ means that $y = Hx + \xi$ where $\xi \sim N(0, \Gamma)$ is independent of x. To see this note that $\Gamma + HC_0H^T$ is invertible because, for any $v \neq 0$,

$$\langle v, (\Gamma + HC_0H^T)v\rangle \geq \langle v, \Gamma v\rangle > 0.$$

The mean of (x, y) is $(m_0^T, (Hm_0)^T)^T$. The covariance of (x, y) is found as follows:

$$\begin{aligned}
\mathbb{E}(x - m_0)(x - m_0)^T &= C_0,\\
\mathbb{E}(x - m_0)(y - Hm_0)^T &= \mathbb{E}(x - m_0)(Hx - Hm_0)^T + \mathbb{E}(x - m_0)\xi^T\\
&= \mathbb{E}(x - m_0)(x - m_0)^T H^T\\
&= C_0H^T\\
\mathbb{E}(y - Hm_0)(y - Hm_0)^T &= \mathbb{E}\big(y - Hx + H(x - m_0)\big)\big(y - Hx + H(x - m_0)\big)^T\\
&= \mathbb{E}\big(\xi + H(x - m_0)\big)\big(\xi + H(x - m_0)\big)^T\\
&= \mathbb{E}\xi\xi^T + \mathbb{E}H(x - m_0)(x - m_0)^T H^T\\
&= \mathbb{E}\xi\xi^T + HC_0H^T\\
&= \Gamma + HC_0H^T.
\end{aligned}$$

The desired result now follows from careful application of Lemma 2.8.11. □

2.9 Calculus of Variations

Let U be a Hilbert space and let $\mathsf{U}_{\text{ad}} \subseteq \mathsf{U}$ denote a closed convex subset in U. Define

$$J(u) = \frac{\beta}{2}\|u\|^2. \tag{2.33}$$

In all cases in this section the norm and inner-product will be in the natural space where the variable(s) in question live and so we will not denote this space explicitly in the inner-products and norm.

Theorem 2.9.1. *Let $\beta > 0$ and assume that there exists $u^\star \in \mathsf{U}_{\text{ad}}$ such that $J(u^\star) < \infty$. Then there exists unique $\overline{u} \in \mathsf{U}_{\text{ad}}$ such that $J(\overline{u}) \leq J(u)$ for all $u \in \mathsf{U}_{\text{ad}}$.*

This theorem is a special case of a more general result which we now state and prove. Let X, Y and Z be Hilbert spaces and $\mathcal{A} \in L(\mathsf{X}, \mathsf{Z}), \mathcal{B} \in L(\mathsf{U}, \mathsf{Z})$

and $D \in L(\mathsf{X}, \mathsf{Y})$. Let $\mathsf{X}_{\mathrm{ad}} \subseteq \mathsf{X}$ denote a closed convex subset in X, choose $g \in \mathsf{Z}$ and define

$$\mathsf{F}_{\mathrm{ad}} = \big\{(x, u) \in \mathsf{X}_{\mathrm{ad}} \times \mathsf{U}_{\mathrm{ad}} | \mathcal{A}x + \mathcal{B}u = g\big\}.$$

Define, for fixed $y \in \mathsf{Y}$,

$$J(x, u) = \frac{\alpha}{2}\|y - Dx\|^2 + \frac{\beta}{2}\|u\|^2. \tag{2.34}$$

Remarks 2.9.2. *For Theorem 2.9.1 (resp. 2.9.3) the existence of a feasible point $u^\star$ (resp. $(x^\star, u^\star)$) follows automatically provided that U_{ad} (resp. F_{ad}) is non-empty. However we formulate the statement and proofs in a slightly more general way which is easily adapted to more complicated choices of objective functional J.*

Theorem 2.9.3. *Let $\beta > 0$, $\alpha \geq 0$ and assume that there exists $(x^\star, u^\star) \in \mathsf{F}_{\mathrm{ad}}$ such that $J(x^\star, u^\star) < \infty$. If $\mathcal{A}$ has bounded inverse then there exists unique $(\overline{x}, \overline{u}) \in \mathsf{F}_{\mathrm{ad}}$ such that $J(\overline{x}, \overline{u}) \leq J(x, u)$ for all $(x, u) \in \mathsf{F}_{\mathrm{ad}}$.*

Proof. Since $J \geq 0$ and since there is a feasible point $(x^\star, u^\star) \in \mathsf{F}_{\mathrm{ad}}$ with $J(x^\star, u^\star) < \infty$ it follows that

$$\overline{J} := \inf_{(x,u) \in \mathsf{F}_{\mathrm{ad}}} J(x, u)$$

exists and is finite. Let $\{x_k, u_k\}$ be a minimizing sequence in F_{ad} so that

$$\lim_{k \to \infty} J(x_k, u_k) = \overline{J}.$$

Now, for any $\delta > 0$ we may choose the minimizing sequence so that

$$\frac{\beta}{2}\|u_k\|^2 \leq J(x_k, u_k) \leq \overline{J} + \delta.$$

Thus $\{u_k\}$ is bounded in U and hence $\{x_k\}$ is bounded in X, because $\mathcal{A}$ has bounded inverse. This implies, along a (relabelled) subsequence, the existence of a weak limit in $\mathsf{X} \times \mathsf{U}$:

$$\{x_k, u_k\} \rightharpoonup (\overline{x}, \overline{u}).$$

We now show that the limit $(\overline{x}, \overline{u}) \in \mathsf{F}_{\mathrm{ad}}$. To this end we note that

$$\{x_k, u_k\} \in M := \mathsf{F}_{\mathrm{ad}} \bigcap \overline{B}_{\mathsf{X} \times \mathsf{U}}(r)$$

where $\overline{B}_{\mathsf{X}\times\mathsf{U}}(r)$ is the closed centred ball of radius r in $\mathsf{X}\times\mathsf{U}$. Since M is bounded, closed and convex it is weakly sequentially compact and we deduce from Theorem 2.4.11 that $(\overline{x},\overline{u})\in M\subset\mathsf{F}_{\mathrm{ad}}$.

Finally observe that J is weakly lower semicontinuous because the square of any Hilbert space norm has this property, and $\beta>0$. Thus

$$\overline{J}=\lim_{k\to\infty}J(x_k,u_k)\geq J(\overline{x},\overline{u})\geq\overline{J}.$$

It follows that $J(\overline{x},\overline{u})=\overline{J}$ and hence that $(\overline{x},\overline{u})\in\mathsf{F}_{\mathrm{ad}}$ attains the infimum of J in F_{ad}. Finaly, since $\beta>0$ we have that $u\mapsto J\big(\mathcal{A}^{-1}(g-Bu),u\big)$ is strictly convex and hence there cannot exist more than one minimizer. □

Characterizing the minimizer $(\overline{x},\overline{u})$ may be achieved by working with Lagrange multipliers. If we define $L:\mathsf{X}\times\mathsf{U}\times\mathsf{Z}\to\mathbb{R}$ by

$$L(x,u,p)=J(x,u)+\langle p,\mathcal{A}x+\mathcal{B}u-g\rangle$$

then the constrained minimizer of $J(x,u)$ over

$$\mathsf{F}_{\mathrm{ad}}=\big\{(x,u)\in\mathsf{X}\times\mathsf{U}|\mathcal{A}x+\mathcal{B}u=g\big\}$$

will be found by making $L(x,u,p)$ stationary with respect to (x,u,p). We now extend this to consider the case of minimizing $J(x,u)$ over

$$\mathsf{F}_{\mathrm{ad}}=\big\{(x,u)\in\mathsf{X}_{\mathrm{ad}}\times\mathsf{U}|\mathcal{A}x+\mathcal{B}u=g\big\}$$

where X_{ad} comprises a finite set of linear constraints. Let $f\in\mathbb{R}^r$ and define

$$\mathsf{X}_{\mathrm{ad}}=\{x\in\mathsf{X}|\ell(x)=f\}\tag{2.35}$$

for some bounded linear functional $\ell:\mathsf{X}\to\mathbb{R}^r$. We now define the Lagrangian $\tilde{L}:\mathsf{X}\times\mathsf{U}\times\mathsf{Z}\times\mathbb{R}^r\to\mathbb{R}$ by

$$\tilde{L}(x,u,p,\rho)=L(x,u,p)+\langle\rho,\ell(x)-f\rangle.$$

The following theorem is proved by considering the stationary points of this Lagrangian.

Theorem 2.9.4. *Consider the setting of Theorem 2.9.3 in the case where* $\mathsf{U}_{\mathrm{ad}}=\mathsf{U}$ *and* X_{ad} *given by* (2.35). *Then there is a Lagrange multiplier*

$(\overline{p}, \overline{\rho}) \in \mathsf{Z} \times \mathbb{R}^r$ *such that the minimizer* $(\overline{x}, \overline{u}) \in \mathsf{F}_{\mathrm{ad}}$ *of* J *satisfies the following equations:*

$$\langle \mathcal{A}\overline{x} + \mathcal{B}\overline{u} - g, \delta p \rangle = 0 \quad \forall \delta p \in \mathsf{Z} \tag{2.36a}$$

$$\langle D_x J(\overline{x}, \overline{u}) + \mathcal{A}^* \overline{p} + \ell^*(\overline{\rho}), \delta x \rangle = 0 \quad \forall \delta x \in \mathsf{X} \tag{2.36b}$$

$$\langle D_u J(\overline{x}, \overline{u}) + \mathcal{B}^* \overline{p}, \delta u \rangle = 0 \quad \forall \delta u \in \mathsf{U} \tag{2.36c}$$

$$\langle \ell(\overline{x}) - f, \delta \rho \rangle = 0 \quad \forall \delta \rho \in \mathbb{R}^r. \tag{2.36d}$$

Proof. The constrained minimizer of $J(x, u)$ is found by making $\tilde{L}(x, u, p, \rho)$ stationary with respect to (x, u, p, ρ). □

Example 2.9.5. *We return to the* ROBOT *Example 1.2.4. We wish to minimize*

$$J(u) := \frac{1}{2} \int_0^1 u(s)^2 ds \tag{2.37}$$

subject to

$$\frac{dx}{dt} = u, \quad x(0) = x_0 \tag{2.38}$$

and

$$x(1) = 0. \tag{2.39}$$

To employ the preceding theory from the calculus of variations we note that $\alpha = 0$ *and* $\beta = r = 1$ *and we define*

$$\begin{aligned} \mathsf{U} &= L^2\big((0,1); \mathbb{R}\big), & \mathsf{U}_{\mathrm{ad}} &= \mathsf{U} \\ \mathsf{X} &= H^1\big((0,1); \mathbb{R}\big), & \mathsf{X}_{\mathrm{ad}} &= \big\{x \in \mathsf{X} | x(1) = 0\big\} \\ \mathsf{Z} &= L^2\big((0,1); \mathbb{R}\big) \times \mathbb{R}. \end{aligned}$$

Thus $\ell(x) = x(1)$ *and* $f = 0$.

We define $\mathcal{A}, \mathcal{B}$ *and* g *by*

$$\mathcal{A}x = \begin{pmatrix} \frac{dx}{dt} \\ x(0) \end{pmatrix}, \quad \mathcal{B}u = \begin{pmatrix} -u \\ 0 \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ x_0 \end{pmatrix}.$$

The equation $\mathcal{A}x + \mathcal{B}u = g$ *is then simply an abstract form of* (2.38). *Because*

$$x(t) = x_0 + \int_0^t u(s) ds$$

and

$$\frac{dx}{dt}(t) = u(t)$$

we have

$$\begin{aligned}
\|x\|_{H^1}^2 &= \int_0^1 \Big(|\frac{dx}{dt}|^2 + |x|^2 \Big) dt \\
&\leq \int_0^1 \Big(|u(t)|^2 + 2|x_0|^2 + 2\big|\int_0^t u(s)ds\big|^2 \Big) dt \\
&\leq \int_0^1 |u(t)|^2 dt + 2|x_0|^2 + 2\int_0^1 |u(s)|^2 ds \\
&= 3\|u\|_{L^2}^2 + 2|x_0|^2 \\
&= 3\|\mathcal{B}u\|^2 + 2\|g\|^2 \\
&\leq 3\|\mathcal{B}u - g\|^2.
\end{aligned}$$

The last line holds because of the zero structure of $\mathcal{B}u$ and g. From the preceding we deduce that $\mathcal{A}^{-1}$ is bounded, because $x = \mathcal{A}^{-1}(g - \mathcal{B}u)$. Thus Theorem 2.9.3 implies the existence of a minimizer $(\overline{x}, \overline{u})$.

We now apply Theorem 2.9.4 to determine defining equations for $(\overline{x}, \overline{u})$, together with the Lagrange multipliers $(\overline{p}, \overline{\rho})$. From (2.36a) *and* (2.36d) *we obtain* (2.38) *and* (2.39) *with (x, u) replaced by $(\overline{x}, \overline{u})$:*

$$\frac{d\overline{x}}{dt} = \overline{u}, \quad \overline{x}(0) = x_0 \tag{2.40}$$

and

$$\overline{x}(1) = 0. \tag{2.41}$$

To calculate (2.36b) *we must find the adjoints of $\mathcal{A}$ and ℓ. To this end, let $\overline{p} = (\overline{\lambda}, q) \in \mathsf{Z}$ and we* assume *that $\overline{\lambda}$ is differentiable – we discuss this assumption at the end of the example. We note that under this smoothness assumption we have*

$$\begin{aligned}
\langle \overline{p}, \mathcal{A}\delta x \rangle &= \int_0^1 \overline{\lambda} \frac{d\delta x}{dt} dt + q\delta x(0) \\
&= \overline{\lambda}(1)\delta x(1) - \overline{\lambda}(0)\delta x(0) - \int_0^1 \frac{d\overline{\lambda}}{dt} \delta x dt + q\delta x(0) \\
&:= \langle \mathcal{A}^* \overline{p}, \delta x \rangle.
\end{aligned}$$

Also

$$\langle \overline{\rho}, \ell(\delta x)\rangle = \overline{\rho}\delta x(1) := \langle \ell^*(\overline{\rho}), \delta x\rangle.$$

Thus, since J *is independent of* x*,* (2.36b) *gives*

$$\langle \mathcal{A}^*\overline{p} + \ell^*(\overline{\rho}), \delta x\rangle = 0 \quad \forall \delta x \in \mathsf{X}$$

and hence that

$$\big(\overline{\lambda}(1) + \overline{\rho}\big)\delta x(1) + \big(q - \overline{\lambda}(0)\big)\delta x(0) - \int_0^1 \frac{d\overline{\lambda}}{dt}\delta x\, dt = 0, \quad \forall \delta x \in X.$$

This identity holds if we choose the Lagrange mutipliers $\overline{\lambda}, q$ *and* $\overline{\rho}$ *to satisfy*

$$\frac{d\overline{\lambda}}{dt} = 0 \tag{2.42a}$$

$$\overline{\lambda}(1) = -\overline{\rho}, \tag{2.42b}$$

$$q = \overline{\lambda}(0). \tag{2.42c}$$

Finally we note that $D_u J(\overline{x}, \overline{u}) = \overline{u}$ *and that*

$$\langle \overline{p}, \mathcal{B}\delta u\rangle = \int_0^1 (-\overline{\lambda}\delta u)dt := \langle \mathcal{B}^*\overline{p}, \delta u\rangle$$

so that (2.36c) *implies that*

$$\int_0^1 (\overline{u} - \overline{\lambda})\delta u\, dt = 0 \quad \forall \delta u \in \mathsf{U}$$

so that

$$\overline{u} = \overline{\lambda}. \tag{2.43}$$

We note that equations (2.40)*,* (2.41)*,* (2.42a) *and* (2.43) *imply that*

$$\frac{d^2\overline{x}}{dt^2} = 0, \quad \overline{x}(0) = x_0, \quad \overline{x}(1) = 0.$$

Therefore $\overline{x}(t) = (1-t)x_0$ *and a candidate for the optimal control is* $\overline{u}(t) = -x_0$.

In Chapter 6 we will actually show that this procedure produces the optimal control. The preceding analysis uses Theorem 2.9.4 which simply exhibits necessary *conditions to be satisfied by the optimal control. Without proving uniqueness of a solution to the identities given by these necessary conditions we are unable to deduce that the candidate solution we exhibit is actually the optimal control. In this regard, note that we assumed that* $\overline{\lambda}$ *was differentiable and performed an integration by parts; we have not ruled out the possibility of other solutions to the variational equations of Theorem 2.9.4 with less regularity and therefore we do not yet know that we have exhibited the optimal choice. Chapter 6 will contain an analysis which exhibits the optimality of the candidate control explicitly.*

We will also be interested in the case where $\mathsf{X}_{\text{ad}} = \mathsf{X}$. The following theorem may be proved by use of the Lagrangian L defined preceding Theorem 2.9.4.

Theorem 2.9.6. *Consider the setting of Theorem 2.9.3 in the case where* $\mathsf{U}_{\text{ad}} = \mathsf{U}$ *and* $\mathsf{X}_{\text{ad}} = \mathsf{X}$*. Then there is a Lagrange multiplier* $\overline{p} \in \mathsf{Z}$ *such that the minimizer* $(\overline{x}, \overline{u}) \in \mathsf{F}_{\text{ad}}$ *of* J *satisfies the following equations:*

$$\begin{aligned} \langle \mathcal{A}\overline{x} + \mathcal{B}\overline{u} - g, \delta p\rangle &= 0 \quad \forall \delta p \in \mathsf{Z} \\ \langle D_x J(\overline{x}, \overline{u}) + \mathcal{A}^*\overline{p}, \delta x\rangle &= 0 \quad \forall \delta x \in \mathsf{X} \\ \langle D_u J(\overline{x}, \overline{u}) + \mathcal{B}^*\overline{p}, \delta u\rangle &= 0 \quad \forall \delta u \in \mathsf{U}. \end{aligned}$$

Exercises

Exercise 2-1. Assume that A is diagonalizable: there is a invertible transformation X such that $A = X\Lambda X^{-1}$ and Λ is diagonal. Prove the Cayley-Hamilton Theorem 2.3.7.

Exercise 2-2. If $x(t)$ satisfies $\dot{x}(t) = A\,x(t)$ with constant matrix $A \in \mathbb{R}^{2\times 2}$,

$$x(t) = \begin{pmatrix} \mathrm{e}^{-t} \\ -2\mathrm{e}^{-t} \end{pmatrix} \quad \text{when} \quad x(0) = \begin{pmatrix} 1 \\ -2 \end{pmatrix},$$

and

$$x(t) = \begin{pmatrix} \mathrm{e}^{-2t} \\ -\mathrm{e}^{-2t} \end{pmatrix} \quad \text{when} \quad x(0) = \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

find the general solution to $\dot{x} = Ax$, and matrix A.

Exercise 2-3. Calculate e^{At} for

$$a)\ A = \begin{pmatrix} -2 & 1 \\ -1 & -4 \end{pmatrix}, \qquad b)\ A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

Exercise 2-4. Prove Theorem 2.5.3.

Exercise 2-5. Let A and B be square and constant matrices. Show that the solution to the matrix differential equation

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = A\,S(t) + S(t)B, \quad S(0) = C,$$

is $S(t) = \exp(At)C\exp(Bt)$.

Exercise 2-6. Find an explicit formula for the solution of the differential equation $\dot{x} = Ax + bu$, $x(0) = x_0$ where

$$A = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad u(t) = \sin t.$$

Exercise 2-7. Consider the system

$$\begin{aligned} \dot{x}(t) &= f(x, \gamma, t) \\ x(0) &= x_0. \end{aligned}$$

Let $x \in \mathbb{R}^n$, parameter $\gamma \in \mathbb{R}^l$, $t \in [0, T]$ and assume that $f(x, \gamma, t)$ satisfies the assumptions of Theorem 2.6.2 with $\mathbb{R}^n$ replaced by $\mathbb{R}^n \times \mathbb{R}^l$. Use the result of Theorem 2.6.2 to show that for any $t \in [0, T]$, $\gamma_0 \in \mathbb{R}^l$ and $x_0 \in \mathbb{R}^n$ the function $\gamma \mapsto \varphi(t, \gamma, x_0)$ is Fréchet differentiable at γ_0 and

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\gamma \varphi(t, \gamma_0, x_0) = D_x f(\varphi(t, \gamma_0, x_0), \gamma_0, t) D_\gamma \varphi(t, \gamma_0, x_0) + D_\gamma f(\varphi(t, \gamma_0, x_0), \gamma_0, t).$$

Exercise 2-8. Prove Corollary 2.7.18

Exercise 2-9. Consider the nonlinear map

$$x_{k+1} = f(x_k)$$

with fixed point at 0 so that $f(0) = 0$. Define *stability, asymptotic stability, global asymptotic stability* and *exponential stability* of this fixed point, using Definition 2.7.1 for the continuous time case as a guide.

Exercise 2-10. Use the Liapunov function method to show the stability of the origin in the following systems

i) $\begin{cases} \dot{x}_1 = -x_1^3 - 2x_2^2 \\ \dot{x}_2 = x_1 x_2 - x_2^3 \end{cases},$

ii) $\begin{cases} \dot{x}_1 = x_2(1-x_1) \\ \dot{x}_2 = -x_1(1-x_2) \end{cases},$
use $V(x) = -x_1 - \log(1-x_1) - x_2 - \log(1-x_2)$.

What is the domain of attraction of the origin for the system in (i)?

Exercise 2-11.

i) A gradient system is given by

$$\dot{x}(t) = -DW(x(t)) = -\left(\frac{\partial W(x)}{\partial x_1}, \dots, \frac{\partial W(x)}{\partial x_n}\right)$$

with $W : G \to \mathbb{R}$, ($G \subset \mathbb{R}^n$) a twice continuously differentiable function. Show that if $\bar{x}$ is a minimum of W which is also an isolated critical point (so that there exists $\delta > 0$ such that for any $x \in B_\delta(\bar{x}) \setminus \{\bar{x}\}$, $W(x) > W(\bar{x})$ and $DW(\bar{x})$ is invertible) then $\bar{x}$ is an asymptotically stable equilibrium point of the above gradient system.

ii) Find the asymptotically stable equilibria of

$$\dot{x} = -DV(x)$$

with $V : \mathbb{R}^2 \to \mathbb{R}$ given by $V(x_1, x_2) = x_1^2(x_1 - 1)^2 + x_2^2$.

Exercise 2-12. Consider the system $\dot{x} = f(x)$, with $f : \mathbb{R}^n \to \mathbb{R}^n$ continuously differentiable and $f(0) = 0$. Let $\mathcal{G}$ be an open neighborhood of the origin in $\mathbb{R}^n$ and $V : \mathcal{G} \to \mathbb{R}$ a continuously differentiable function with $V(0) = 0$. Assume that $V(x) > 0$ in $\mathcal{G} \setminus \{0\}$. Show that

i) if $\dot{V}(x) = \langle f(x), DV(x) \rangle$ is non-negative in $\mathcal{G}$ then the origin of the above system is not asymptotically stable,

ii) if $\dot{V}(x) > 0$ in $\mathcal{G} \setminus \{0\}$ then the origin is unstable.

Exercise 2-13. Verify that for the Gaussian distribution given in Definition 2.8.1 the mean and covariance are indeed given by m and C.

Exercise 2-14. Prove that the matrices Σ, C' and $(C')^{-1}$ defined in proving Theorem 2.8.6 are positive-definite and symmetric.

Exercise 2-15. Prove that the matrices L_{xx} and L_{yy} defined in proving Theorem 2.8.7 are positive-definite and symmetric.

Exercise 2-16. Consider a joint random variable $(X, Y) \in \mathbb{R}^2$ specified by the distributions $X \sim N(0, \sigma^2)$ and $Y|X \sim N(X, \gamma^2)$. Find the mean and covariance of the random variables (X, Y) and $X|Y$ and find the marginal distributions of X and of Y.

Exercise 2-17. Prove that the two expressions for the conditional mean m' and covariance C' given in Theorem 2.8.6 and Example 2.8.12 agree.

Exercise 2-18. Prove Theorems 2.3.11 and 2.3.12.

Exercise 2-19. Show that if $\{\eta_j\}$ is a sequence of real, independent and identically distributed, mean zero Gaussian random variables, then

$$\mathbb{E}[(\sum_{j=1}^{N} \eta_j)^2] = \sum_{j=1}^{N} \mathbb{E}[\eta_j^2].$$

Exercise 2-20. Show that if $x_1 \sim N(m_1, C_1)$, $x_2 \sim N(m_2, C_2)$ are independent Gaussian random variables, then $(x_1, x_2) \sim N(m, C)$, where

$$m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \; C = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}.$$

Chapter 3

Controllability

This chapter is primarily concerned with open-loop control. Section 3.1 concerns discrete-time linear systems with constant matrices. In section 3.2 we set-up the continuous time problem, in general. Section 3.3 then specifies to continuous time linear systems, considering constant matrices initially, then the general case, and finally returning to constant design matrices and restricted controls. In section 3.4 nonlinear problems are considered.

3.1 Discrete-Time Linear Systems

Although most of our analysis will concern continuous-time, possibly nonlinear, systems, we start by considering discrete-time linear systems as this introduces key ideas in a straightforward setting. We study systems with the form

$$x_{k+1} = Ax_k + Bu_k, \tag{3.1}$$

where $x_k \in \mathbb{R}^n$ is the state of the system at discrete time $k \in \mathbb{Z}^+$ and $u_k \in \mathbb{R}^m$ is the control to be applied to the system at discrete time $k \in \mathbb{Z}^+$. The matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are fixed, independent of k.

Definition 3.1.1. *For constant matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, the* controllability matrix *is*

$$G = G(A,B) := \Big(B, AB, A^2B, \dots, A^{n-1}B\Big) \in \mathbb{R}^{n\times nm}.$$

Recall (Definition 2.3.2) that the rank of a matrix is the number of linearly independent rows or columns. For G it follows from Theorem 2.3.4(i) that $\operatorname{rank} G \leq n$ since $G \in \mathbb{R}^{n\times nm}$.

Theorem 3.1.2. - *(i) The controllability matrix has* $\operatorname{rank} G < n$ *if and only if there is a vector $y \in \mathbb{R}^n \backslash \{0\}$ such that*

$$y^T A^k B = 0, \quad \text{for } k = 0, \dots, n-1.$$

- *(ii) If the controllability matrix has* $\operatorname{rank} G = n$ *then, for $I \in \mathbb{R}^{n\times n}$ the identity, there are matrices $K_j \in \mathbb{R}^{m\times n}$ such that*

$$BK_1 + ABK_2 + \cdots + A^{n-1}BK_n = I. \tag{3.2}$$

Proof. The first item follows from Theorem 2.3.4(ii) whilst the second follows from Theorem 2.3.4(iv). □

To be concrete we concentrate on the problem of controlling the system so that it reaches the origin in a finite number of steps. This suggests the following:

Definition 3.1.3. *The* controllable set $\mathcal{C}$ *is the set of initial conditions x_0 for* (3.1) *for which there is an integer $\ell \in \mathbb{Z}^+$ and a control sequence $\{u_j\}_{j=0}^{\ell-1} \in \mathbb{R}^{m\ell}$ such that $x_\ell = 0$. If $\mathcal{C} = \mathbb{R}^n$ then the system is said to be* controllable*. In the case of linear control problems, we will also say that (A, B) is controllable in this case.*

Theorem 3.1.4. *Assume that A is invertible. Then $\mathcal{C} = \mathbb{R}^n$ if and only if* $\operatorname{rank} G = n$.

Proof. It is straighforward to prove from (3.1) that

$$x_k = A^k x_0 + \sum_{j=0}^{k-1} A^{k-j-1} B u_j. \tag{3.3}$$

We first prove that $\mathcal{C} = \mathbb{R}^n$ implies $\operatorname{rank} G = n$. Assume for purposes of contradiction that $\operatorname{rank} G < n$ but that $\mathcal{C} = \mathbb{R}^n$. We deduce from Theorem 3.1.2(i) that there is a non-zero vector $y \in \mathbb{R}^n$ such that

$$y^T A^k B = 0, \quad \text{for } k = 0, \dots, n-1. \tag{3.4}$$

Since A is invertible the characteristic polynomial does not have zero as a root: $p_A(0) \neq 0$. Thus, in the Cayley-Hamilton Theorem 2.3.7, $a_0 \neq 0$ and it follows that, for $b_j = -a_j/a_0$,

$$I = \sum_{j=1}^{n} b_j A^j \tag{3.5}$$

so that

$$A^{-1} = \sum_{j=1}^{n} b_j A^{j-1}.$$

From this and (3.4) it follows that

$$y^T A^{-1} B = 0$$

so that (3.4) can be extended to hold for $k = -1$. Then, from (3.5), we have

$$A^{-2} = \sum_{j=1}^{n} b_j A^{j-2}.$$

From this and (3.4), now extended to $k = -1$, it follows that

$$y^T A^{-2} B = 0.$$

Proceeding by induction we see that

$$y^T A^k B = 0, \quad \text{for } k \leq n-1. \tag{3.6}$$

Since the problem is assumed controllable we deduce from (3.3) that, for some $\ell \in \mathbb{Z}^+$ and some control sequence $\{u_j\}_{j=0}^{\ell-1}$,

$$0 = A^\ell x_0 + \sum_{j=0}^{\ell-1} A^{\ell-j-1} B u_j.$$

Inverting A^ℓ gives

$$0 = x_0 + \sum_{j=0}^{\ell-1} A^{-j-1} B u_j.$$

From this and (3.6) it follows that

$$\langle y, x_0 \rangle = 0.$$

But since x_0 is arbitrary this implies that

$$\langle y, x \rangle = 0, \quad \forall x \in \mathbb{R}^n.$$

Hence $y = 0$ a contradiction. Hence $\mathcal{C} = \mathbb{R}^n$ implies $\operatorname{rank} G = n$.

We now prove that $\operatorname{rank} G = n$ implies that $\mathcal{C} = \mathbb{R}^n$. From Theorem 3.1.2(ii) we have matrices $K_j \in \mathbb{R}^{m \times n}$ such that

$$BK_1 + ABK_2 + \cdots + A^{n-1}BK_n = I.$$

Thus

$$Bv_0 + ABv_1 + \cdots + A^{n-1}Bv_{n-1} = -A^n x_0$$

where $v_{j-1} = -K_j A^n x_0$. Now set $u_j = v_{n-j-1}$. Then, by (3.3),

$$\begin{aligned}
x_n &= A^n x_0 + \sum_{j=0}^{n-1} A^{n-j-1} B u_j \\
&= A^n x_0 + \sum_{j=0}^{n-1} A^{n-j-1} B v_{n-j-1} \\
&= A^n x_0 + \sum_{j=0}^{n-1} A^j B v_j \\
&= A^n x_0 - A^n x_0 \\
&= 0.
\end{aligned}$$

Thus we have established that any starting point can be controlled to the origin in n steps, provided that $\operatorname{rank} G = n$. Thus $\operatorname{rank} G = n$ implies that $\mathcal{C} = \mathbb{R}^n$ and the proof is complete. □

3.2 Continuous-Time Systems: Setup

The remainder of the chapter concerns continuous time systems. In this section we describe the notation, together with various definitions, that we will use throughout. Consider the ODE

$$\dot{x} = f(x, u), \quad t > 0 \tag{3.7a}$$

$$x(0) = x_0 \tag{3.7b}$$

where $x_0 \in \mathbb{R}^n$, $u : [0, \infty) \to U \subset \mathbb{R}^m$ is the control and $x : [0, \infty) \to \mathbb{R}^n$ is the response. The right-hand side satisfies $f : \mathbb{R}^n \times U \to \mathbb{R}^n$.

Throughout this chapter we consider open-loop control, and assume that our control objective is to drive the system state to the origin, as in the discrete setting. Recall the set of admissible controls

$$\mathcal{U} = \Big\{ u : [0, \infty) \to U \subseteq \mathbb{R}^m \Big| u(\cdot) \,\text{is measurable} \Big\}.$$

We mainly consider unrestricted controls, for which $U = \mathbb{R}^m$, and restricted controls defined via the choice $U = [-1, 1]^m$. In these cases we have, respectively,

$$\mathcal{U} = \Big\{ u : [0, \infty) \to \mathbb{R}^m \Big| u(\cdot) \,\text{is measurable} \Big\}$$

and

$$\mathcal{U} = \Big\{ u : [0, \infty) \to [-1, 1]^m \Big| u(\cdot) \,\text{is measurable} \Big\}.$$

We also prove some general results in which we assume structural properties on U, applying to both $U = \mathbb{R}^m$ and $U = [-1, 1]^m$, but also to other choices of U.

Definition 3.2.1. *The* fixed time t controllable set *is*

$$\mathcal{C}(t) = \Big\{ x_0 \in \mathbb{R}^n \Big| \exists u \in \mathcal{U} : x(t) = 0 \Big\}.$$

Then

$$\mathcal{C} = \bigcup_{t>0} \mathcal{C}(t)$$

is the controllable set. *If $\mathcal{C} = \mathbb{R}^n$ then the system is said to be* controllable. *We will also say that (A, B) is controllable in this case.*

Remarks 3.2.2. *Real systems may be subjected to disturbances, not accounted for in our model equation* (3.7)*, which act to push the system away from the target state. In such situations it is desirable to be able to steer all nearby states to the target. Hence it is important that $\mathcal{C}$ contains a neighbourhood of the origin. The ideal situation would be, of course, that $\mathcal{C} = \mathbb{R}^n$. For both the linear and nonlinear systems, and for unrestricted and restricted controls, we will study the structure of the set $\mathcal{C}$ in what follows.*

Definition 3.2.3. *The* reachable set *at time t, $K(t, x_0)$, is defined as follows:*

$$K(t, x_0) = \Big\{ x_1 \Big| \exists u \in \mathcal{U} : x(0) = x_0, x(t) = x_1 \Big\}.$$

Let $x_1 \in K(t, x_0)$. Then we may chose u so that $x(t)$ solves (3.7) with $x(0) = x_0$ and $x(t) = x_1$. Define $z(t) = x(t_1 - t)$ and $\tilde{u} = u(t_1 - t)$. It follows that z solving the time-reversed system

$$\dot{z} = -f(z, \tilde{u}), \quad t > 0 \tag{3.8a}$$

$$z(0) = x_1 \tag{3.8b}$$

will satisfy $z(t_1) = x_0$. If we define K^- to be the reachable set for the time-reversed system then we have proved the following theorem which will enable us to trasfer results from controllable sets to reachable sets, and vice-versa.

Theorem 3.2.4. *For the ODE* (3.7) $\mathcal{C}(t_1) = K^-(t_1, 0)$.

3.3 Linear Systems

In this section we study controllability problems for the linear system

$$\dot{x} = Ax + Bu, \quad t > 0 \tag{3.9a}$$

$$x(0) = x_0. \tag{3.9b}$$

We assume:

Assumption 3.3.1. The control $u \in \mathcal{U}$ is locally integrable, the matrix-valued functions A, B satisfy $A \in C(\mathbb{R}^+, \mathbb{R}^{n\times n})$ and $B \in C(\mathbb{R}^+, \mathbb{R}^{n\times m})$.

Before reading this material it is useful to recap the basic results concerning linear ODEs which are contained in section 2.5. By Theorem 2.5.2 we see that, under Assumption 3.3.1, there is a unique $x \in C([0,T]; \mathbb{R}^n)$ solving (3.9) for any $T > 0$ (see Remark 2.5.1). If U is bounded then it is sufficient to assume that u is measurable, and to drop local integrability. In this situation $u \in L^\infty(\mathbb{R}^+; \mathbb{R}^m)$ and therefore locally integrability of u follows.

We will repeatedly use the fact that the solution of (3.9) is given, from (2.7), by the formula

$$x(t) = S(t)x_0 + \int_0^t S(t)S^{-1}(s)B(s)u(s)ds, \quad t \in [0, T]. \tag{3.10}$$

From this it follows that

$$S(t)^{-1}x(t) = x_0 + \int_0^t S^{-1}(s)B(s)u(s)ds, \quad t \in [0, T]. \tag{3.11}$$

Note that $x_0 \in \mathcal{C}(t)$ thus implies that there is $u \in \mathcal{U}$ such that

$$x_0 = -\int_0^t S^{-1}(s)B(s)u(s)ds. \tag{3.12}$$

In the setting where A and B are constant, we hence obtain

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)ds; \tag{3.13}$$

and

$$e^{-tA}x(t) = x_0 + \int_0^t e^{-sA}Bu(s)ds. \tag{3.14}$$

In this case $x_0 \in \mathcal{C}(t)$ hence implies that there is $u \in \mathcal{U}$ such that

$$x_0 = -\int_0^t e^{-sA}Bu(s)ds. \tag{3.15}$$

The controllability matrix is defined as in Definition 3.1.1:

$$G = G(A, B) := \Big(B, AB, A^2B, \dots, A^{n-1}B\Big).$$

3.3.1 Unrestricted Controls

The following theorem is analogous to Theorem 3.1.4; note, however, that it is not necessary to assume that A is invertible in the continuous time setting.

Theorem 3.3.2. *Consider the linear system* (3.9) *with constant A and B, and unrestricted controls. Then $\mathcal{C} = \mathbb{R}^n$ if and only if* $\operatorname{rank} G = n$.

Proof. We first prove that $\mathcal{C} = \mathbb{R}^n$ implies that rank$G = n$. Suppose that $\mathcal{C} = \mathbb{R}^n$ and assume, for contradiction, rank$G < n$. It follows from Theorem 3.1.2(i) that there is vector $y \in \mathbb{R}^n \backslash \{0\}$ such that

$$y^T A^k B = 0, \quad \text{for } k = 0, \dots, n-1.$$

From the Cayley-Hamilton Theorem 2.3.7 it follows that

$$A^n = -a_{n-1}A^{n-1} - \dots - a_1 A - a_0 I.$$

From this it follows that

$$y^T A^n B = -a_{n-1} y^T A^{n-1} B - \cdots - a_1 y^T AB - a_0 y^T B = 0.$$

Similarly

$$y^T A^{n+1} B = -a_{n-1} y^T A^n B - \cdots - a_1 y^T A^2 B - a_0 y^T AB = 0.$$

Proceeding by induction we see that

$$y^T A^k B = 0, \quad \text{for } k \in \mathbb{Z}^+. \tag{3.16}$$

Since, by (2.9),

$$e^{-As} = \sum_{k=0}^{\infty} \frac{(-1)^k A^k}{k!} s^k$$

the identities (3.16) imply that

$$y^T e^{-As} B = 0, \quad \text{for all} \quad s \in \mathbb{R}. \tag{3.17}$$

Since $\mathcal{C} = \mathbb{R}^n$, equation (3.15) shows that for any $x_0 \in \mathbb{R}^n$, there is a time $t = t(x_0)$ and a control $u : [0, t) \to \mathbb{R}^m$, such that

$$x_0 = -\int_0^t e^{-sA} B u(s) ds.$$

Hence $y^T x_0 = 0$. But x_0 is arbitrary, and hence $y = 0$ contradicting the assumption that rank$G < n$.

Now assume that rank$G = n$. Lemma 3.3.3 which follows shows that $\mathcal{C}(t) = \mathbb{R}^n$ for any $t > 0$ and hence that $\mathcal{C} = \mathbb{R}^n$. This completes the proof. □

Lemma 3.3.3. *Consider the linear system* (3.9) *with constant A and B, and unrestricted controls. Then* rank$G = n$ *implies that $\mathcal{C}(t) = \mathbb{R}^n$ for any $t > 0$.*

Proof. By Theorem 3.1.2(ii) (see equation (3.2)) the condition rank$G = n$ implies that, for $I \in \mathbb{R}^{n \times n}$ the identity, there are matrices $K_j \in \mathbb{R}^{m \times n}$ such that

$$BK_1 + ABK_2 + \cdots + A^{n-1} BK_n = I.$$

or

$$\sum_{j=1}^{n} A^{j-1} BK_j = I. \tag{3.18}$$

Now fix any $t > 0$. Let φ be any function in $C^{n-1}([0,t];\mathbb{R})$ satisfying

$$\frac{d^j\varphi}{dt^j}(0) = \frac{d^j\varphi}{dt^j}(t) = 0, \quad j = 0, 1, 2, \ldots, n-1$$

and

$$\int_0^t \varphi(s)ds = 1.$$

A polynomial of sufficiently high degree can be used, for example. Now set

$$\psi(s) = -e^{As}x_0\varphi(s), \ s \in [0,t]$$

and note that, by construction, ψ satisfies the same homogeneous boundary conditions as φ:

$$\frac{d^j\psi}{dt^j}(0) = \frac{d^j\psi}{dt^j}(t) = 0, \quad j = 0, 1, 2, \ldots, n-1.$$

Now choose the control

$$\begin{aligned} u(s) &= K_1\psi(s) + K_2\frac{d\psi}{ds}(s) + \cdots + K_n\frac{d^{n-1}\psi}{ds^{n-1}}(s), \\ &= \sum_{j=1}^{n} K_j \frac{d^{j-1}\psi}{ds^{j-1}}(s), \quad s \in [0,t]. \end{aligned} \tag{3.19}$$

We will show that this control implies that $x(t) = 0$.

Note that, using the boundary conditions on ψ and Theorem 2.5.3, integration by parts shows that

$$\int_0^t e^{A(t-s)}BK_j\frac{d^{j-1}\psi}{ds^{j-1}}(s)ds = \int_0^t e^{A(t-s)}ABK_j\frac{d^{j-2}\psi}{ds^{j-2}}(s)ds.$$

Further integrations by parts demonstrate that, for $j = 1, \ldots, n$,

$$\int_0^t e^{A(t-s)}BK_j\frac{d^{j-1}\psi}{ds^{j-1}}(s)ds = \int_0^t e^{A(t-s)}A^{j-1}BK_j\psi(s)ds.$$

Hence, summing this identity over $j = 1, \ldots, n$ and using (3.18) and (3.19) on the right and left hand sides respectively, we obtain

$$\int_0^t e^{A(t-s)}Bu(s)ds = \int_0^t e^{A(t-s)}\psi(s)ds.$$

Thus, using (3.13),

$$\begin{aligned}
x(t) &= e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)ds \\
&= e^{At}x_0 + \int_0^t e^{A(t-s)}\psi(s)ds \\
&= e^{At}x_0 - \int_0^t e^{A(t-s)}e^{As}x_0\varphi(s)ds \\
&= e^{At}x_0 - e^{At}x_0\int_0^t \varphi(s)ds \\
&= 0.
\end{aligned}$$

Since x_0 is arbitrary we have shown that $\mathcal{C}(t) = \mathbb{R}^n$ for any $t > 0$ and hence that $\mathcal{C} = \mathbb{R}^n$. □

Corollary 3.3.4. *Consider the fixed matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. For any $y \in \mathbb{R}^n$ define $w(t) = y^T e^{-At}B$. Then $\operatorname{rank} G = n$ if and only if $w(\cdot) \neq 0$ (w is not the zero function) for every $y \neq 0$.*

Proof. **If**. Assume that $\operatorname{rank} G < n$. Then the proof of Theorem 3.3.2 demonstrates that there exists non-zero y for which (3.17) holds:

$$y^T e^{-As}B = 0, \quad \text{for all} \quad s \in \mathbb{R}. \tag{3.20}$$

For **only if** assume that there exists non-zero y such that (3.20) holds. Differentiating $k = 0, \dots, n-1$ times and setting $s = 0$ gives

$$y^T A^k B = 0, \quad \text{for } k = 0, \dots, n-1.$$

Thus, by Theorem 3.1.2(i) we see that $\operatorname{rank} G < n$. □

Given two matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$ define $y = y(A, B) \in \mathbb{R}^k$, with $k = n^2 + mn$, to be the vector made up of the entries of the two matrices. Now consider the two linear systems

$$\begin{aligned}
\dot{x} &= A_1 x + B_1 u, \\
\dot{x} &= A_2 x + B_2 u
\end{aligned}$$

with $A_i \in \mathbb{R}^{n\times n}, B_i \in \mathbb{R}^{n\times m}$. Define the *distance* between the two systems to be the Euclidean norm of the vector $y(A_1, B_1) - y(A_2, B_2)$ and note that

$$|y(A_1, B_1) - y(A_2, B_2)|^2 = \|A_1 - A_2\|^2 + \|B_1 - B_2\|^2,$$

with $\|\cdot\|$ the Frobenius norm and $|\cdot|$ the Euclidean norm. This distance function then makes the set of all linear control systems

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

into a metric space $\mathcal{M}$.

Theorem 3.3.5. *Consider the linear system* (3.9) *with constant A and B, and unrestricted controls. The set of all controllable linear autonomous systems is open and dense in the metric space $\mathcal{M}$.*

Proof. **Openness** We need to show that if, for the system

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

we have $\operatorname{rank} G(A,B) = n$ then $\operatorname{rank} G(\tilde{A},\tilde{B}) = n$ provided

$$\|A - \tilde{A}\|^2 + \|B - \tilde{B}\|^2 \leq \epsilon^2$$

for ϵ sufficiently small.

Recall from Theorem 2.3.4(iii) that $\operatorname{rank} G(A,B) = n$ if and only if $\exists H \in \mathbb{R}^{n\times n}$, a submatrix formed from the columns of $G(A,B)$, with $\det H \neq 0$. The determinant is a continuous function of its entries and hence, for ϵ small enough we have $\det \tilde{H} \neq 0$ if $\tilde{H}$ is the submatrix of $G(\tilde{A},\tilde{B})$ chosen using the same columns as used to construct H. Thus $\operatorname{rank} G(\tilde{A},\tilde{B}) = n$.

Density We need to show that, if $\operatorname{rank} G(A,B) < n$ then, for any $\epsilon > 0$, $\exists \tilde{A} \in \mathbb{R}^{n\times n}$ and $\tilde{B} \in \mathbb{R}^{n\times m}$ such that

$$\|A - \tilde{A}\|^2 + \|B - \tilde{B}\|^2 \leq \epsilon^2$$

holds and $\operatorname{rank} G(\tilde{A},\tilde{B}) = n$.

Consider any $n \times n$ submatrix H made from columns of $G(A,B)$. We can think of $\det H$ as a polynomial of the $k = n^2 + mn$ entries of A and B:

$$\det H = \phi\big(y(A,B)\big).$$

Then, since $\operatorname{rank} G(A,B) < n$ we have $\phi\big(y(A,B)\big) = 0$. But a nontrivial polynomial ϕ cannot vanish identically on any $k-$dimensional ball centred at $y(A,B)$, because otherwise all partial derivatives are zero and all coefficients vanish. Therefore $\exists \xi \in \mathbb{R}^k \neq 0$ such that $\phi\big(y(A,B) + \lambda\xi\big) \neq 0$ for all λ sufficiently small. From this it follows that $\exists \tilde{A}, \tilde{B}$ arbitrarily close to A, B such that $\operatorname{rank} G(\tilde{A},\tilde{B}) = n$. □

Remarks 3.3.6. *The above theorem implies that, if a physical process is modelled by the linear system* (3.9)*, and the parameters are only approximately known then establishing controllability is still useful: if a given instance of* (3.9) *is controllable then so is any nearby system.*

3.3.2 Symmetric and Convex Controls

We continue our study of the linear control system given by (3.9), now returning to the general setting where A and B may depend on time t. We recall that the control u is contained in the set $\mathcal{U}$ given by (1.1), and the set U which defines it. Recall also Definitions 2.4.7 and 2.4.8 concerning symmetric and convex sets respectively.

Theorem 3.3.7. *Consider the linear system* (3.9) *under Assumptions 3.3.1 and assume that U in definition* (1.1) *of $\mathcal{U}$ is convex (resp. symmetric). Then, for the linear system* (3.9)*, the controllable set $\mathcal{C}$ is convex (resp. symmetric).*

Proof. Note first that U convex (resp. symmetric) implies that $\mathcal{U}$ is convex (resp. symmetric). First we prove the stated result concerning symmetry. Let $x_0 \in \mathcal{C}(t)$ for some $t > 0$. It suffices to show that $-x_0 \in \mathcal{C}(t)$. By (3.12) we have

$$x_0 = -\int_0^t S^{-1}(s)B(s)u(s)ds \tag{3.21}$$

and hence

$$-x_0 = -\int_0^t S^{-1}(s)B(s)\big(-u(s)\big)ds.$$

With initial condition $-x_0$ and choosing the control $-u(t)$ in (3.10) gives

$$\begin{aligned} x(t) &= -S(t)x_0 - \int_0^t S(t)S^{-1}(s)B(s)u(s)ds \\ &= -S(t)\Big(x_0 + \int_0^t S^{-1}(s)B(s)u(s)ds\Big) \\ &= 0 \end{aligned}$$

where the last line follows from (3.21). This proves symmetry.

Now we consider convexity. Assume that $x_0, \hat{x}_0 \in \mathcal{C}$. Then there exist $t, \hat{t}$ such that $x_0 \in \mathcal{C}(t)$ and $\hat{x}_0 \in \mathcal{C}(\hat{t})$. Without loss of generality we assume that

$t \leq \hat{t}$. Then, from (3.12), for some $u, \hat{u} \in \mathcal{U}$,

$$x_0 = -\int_0^t S^{-1}(s)B(s)u(s)ds$$
$$\hat{x}_0 = -\int_0^{\hat{t}} S^{-1}(s)B(s)\hat{u}(s)ds.$$

Now define

$$\tilde{u}(s) = u(s), \quad 0 \leq s \leq t,$$
$$\tilde{u}(s) = 0, \quad t < s \leq \hat{t}.$$

Clearly

$$x_0 = -\int_0^{\hat{t}} S^{-1}(s)B(s)\tilde{u}(s)ds$$

and therefore $x_0 \in \mathcal{C}(\hat{t})$. For $\lambda \in [0,1]$ we have

$$\lambda x_0 + (1-\lambda)\hat{x}_0 = -\int_0^{\hat{t}} S^{-1}(s)B(s)\big(\lambda\tilde{u}(s) + (1-\lambda)\hat{u}(s)\big)ds. \tag{3.22}$$

Now $\mathcal{U}$ is convex becuase U is convex and hence $\big(\lambda\tilde{u}(s) + (1-\lambda)\hat{u}(s)\big) \in \mathcal{U}$. If $x(0) = \lambda x_0 + (1-\lambda)\hat{x}_0$ then, by (3.10) and (3.22)

$$x(\hat{t}) = S(\hat{t})\big(\lambda x_0 + (1-\lambda)\hat{x}_0\big) + \int_0^{\hat{t}} S(\hat{t})S^{-1}(s)B(s)\big(\lambda\tilde{u}(s) + (1-\lambda)\hat{u}(s)\big)ds$$
$$= 0.$$

Thus $\lambda x_0 + (1-\lambda)\hat{x}_0 \in \mathcal{C}(\hat{t}) \subseteq \mathcal{C}$ and the result is proved. □

Remarks 3.3.8.

- *Since both $U = \mathbb{R}^m$ and $U = [-1,1]^m$ are convex and symmetric, the preceding theorem holds for both these cases.*
- *The arguments in the proof may be readily adapted to show that if U is convex (resp. symmetric), then $\mathcal{C}(t)$ is convex (resp. symmetric) for any $t \in \mathbb{R}^+$.*

We now prove a related result concerning convexity which will be useful to us in the study of optimal control in Chapter 6. Let $\mathsf{U} = L^2\big([0,T];\mathbb{R}^m\big)$ and let

$$\mathsf{U}_{\text{ad}} = \{u \in \mathsf{U} : x(T) = 0\}.$$

Lemma 3.3.9. U_{ad} *is convex.*

Proof. A function $u \in \mathsf{U}_{\mathrm{ad}}$ if and only if

$$0 = e^{AT}x_0 + \int_0^T e^{A(T-s)}Bu(s)ds.$$

If $u_i \in \mathsf{U}_{\mathrm{ad}}$ for $i = 1, 2$ then

$$0 = e^{AT}x_0 + \int_0^T e^{A(T-s)}Bu_i(s)ds.$$

Multiplying this equation by λ for $i = 1$ and by $(1 - \lambda)$ for $i = 2$ and adding gives

$$0 = e^{AT}x_0 + \int_0^T e^{A(T-s)}B\big(\lambda u_1(s) + (1 - \lambda)u_2(s)\big)ds.$$

Thus $\lambda u_1 + (1 - \lambda)u_2 \in \mathsf{U}_{\mathrm{ad}}$. □

Example 3.3.10. *Consider the two dimensional system*

$$\begin{aligned}\dot{x}_1 &= 0,\\ \dot{x}_2 &= u\end{aligned}$$

with $x_i(t) \in \mathbb{R}$ *and* $u : \mathbb{R}^+ \to [-1, 1]$. *This system is of the form* (3.9) *with*

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

It is clear that, since $x_1(t) = x_1(0)$ *for all* $t > 0$, *the controllable set is simply*

$$\mathcal{C} = \big\{(x_1, x_2)|x_1 = 0\big\}.$$

This set is symmetric and convex. However, note that $\mathcal{C}$ *does not contain a neighbourhood of the origin.*

3.3.3 Restricted Controls

We now proceed to study restricted controls. Motivated by the example at the end of the previous section, it is natural to ask what conditions are required on the matrices A and B to ensure that, when the controls are restricted, $\mathcal{C}$ contains a neighbourhood of the origin. This is the goal of the next theorem which generalizes Theorem 3.3.2 to restricted controls. Recall (Definition 2.4.2) that the interior and boundary of a set $\Omega \subseteq \mathbb{R}^n$.

Theorem 3.3.11. *Consider the linear autonomous system* (3.9) *with A and B constant matrices. For restricted controls we have*

$$\operatorname{rank} G = n \quad \text{if and only if} \quad 0 \in \operatorname{Int} \mathcal{C}.$$

Proof. Suppose rank$G < n$. Then the linear span of the columns of G has dimension less than or equal to $n-1$. It follows from Theorem 3.1.2(i) that there is vector $y \in \mathbb{R}^n \backslash \{0\}$ such that

$$y^T A^k B = 0, \quad \text{for } k = 0, \ldots, n-1.$$

Following the proof of Theorem 3.3.2 we deduce, by the Cayley-Hamilton theorem, that

$$y^T A^k B = 0, \quad \text{for } k \in \mathbb{Z}^+$$

and hence that $y^T e^{-As} B = 0$. Now assume that $x_0 \in \mathcal{C}(t)$. Then, by (3.14), there exists $u \in \mathcal{U}$ such that

$$x_0 = -\int_0^t e^{-As} B u(s) ds$$

so that

$$y^T x_0 = -\int_0^t y^T e^{-As} B u(s) ds = 0.$$

Thus $\mathcal{C}(t)$ lies in the hyperplane orthogonal to $y \neq 0$ for all $t \geq 0$. Hence $\mathcal{C} = \bigcup_{t>0} \mathcal{C}(t)$ lies in the same hyperplane, from which it follows that $\operatorname{Int} \mathcal{C} = \emptyset$, and in particular that $0 \notin \operatorname{Int} \mathcal{C}$.

To show the converse, suppose that $0 \notin \operatorname{Int} \mathcal{C}$ so that $0 \notin \operatorname{Int} \mathcal{C}(t)$ for all $t > 0$. Note that $0 \in \mathcal{C}(t)$ for all $t > 0$ as can be seen by choosing u to be identically zero. We deduce that $0 \in \partial \mathcal{C}(t)$. Since $\mathcal{C}(t)$ is convex (see Remarks 3.3.8) we deduce the existence of a hyperplane through 0 such that $\mathcal{C}(t)$ lies on one side of the hyperplane. More precisely (see Lemma 3.3.15) there exists a non-zero vector $y \in \mathbb{R}^n$ such that $y^T x_0 \leq 0$ for all $x_0 \in \mathcal{C}(t)$. Pick any $u \in \mathcal{U}$ and any $t > 0$ and define x_0 by

$$x_0 = -\int_0^t e^{-As} B u(s) ds.$$

Then (3.14) shows that $x_0 \in \mathcal{C}(t)$ and therefore

$$y^T x_0 = -\int_0^t y^T e^{-As} B u(s) ds \leq 0.$$

Hence

$$\int_0^t y^T e^{-As} Bu(s) ds \geq 0 \quad \forall u \in \mathcal{U}.$$

By Lemma 3.3.14 below this implies that

$$y^T e^{-As} B = 0 \quad \forall s \geq 0. \tag{3.23}$$

Setting $s = 0$ gives $y^T B = 0$. Differentiating the identity (3.23) with respect to s and setting $s = 0$ gives $y^T AB = 0$ and differentiating k times and setting $s = 0$ gives

$$y^T A^k B = 0, \quad k = 1, 2, \ldots, n-1.$$

By Theorem 3.1.2(i) this establishes that rankG $< n$ as required. □

Corollary 3.3.12. *Consider the linear autonomous system* (3.9) *with A and B constant matrices. For unrestricted or restricted controls we have the following: if* $\operatorname{rank} G < n$ *then there exists a hyperplane in* $\mathbb{R}^n$ *which contains* $\mathcal{C}(t)$ *for any* $t > 0$.

Corollary 3.3.13. *Consider the linear autonomous system* (3.9) *with A and B constant matrices with unrestricted or restricted controls. For any $y \in \mathbb{R}^n$ define $w(t) = y^T e^{-At} B$. Then* $\operatorname{rank} G = n$ *if and only if $w(\cdot) \neq 0$ (w is not the zero function) for every $y \neq 0$.*

Proof. We have already proved this result in the case of unrestricted controls as Corollary 3.3.4; the case of unrestricted controls is similar and is left as Exercise 3-15. □

Lemma 3.3.14. *Let $t > 0$ and $y \in \mathbb{R}^n$. If for any $u \in \mathcal{U}$ we have*

$$\int_0^t y^T e^{-As} Bu(s) ds \geq 0 \tag{3.24}$$

then $y^T e^{-As} B = 0$.

Proof. Let $v : \mathbb{R}^+ \to \mathbb{R}^m$ be defined by $v^T(s) = y^T e^{-As} B$. Since $u \in \mathcal{U}$ implies $-u \in \mathcal{U}$ we can replace u by $-u$ in (3.24) which changes the sign of the identity. Hence

$$\int_0^t y^T e^{-As} Bu(s) ds = \int_0^t v^T(s) u(s) ds = 0$$

for any $u \in \mathcal{U}$. Now suppose that for some $s_0 \in [0, t]$ we have $v(s_0) \neq 0$. By continuity there is an interval I containing s_0 on which $v(s) \neq 0$ for any $s \in I$.

Define $u \in \mathcal{U}$ by

$$\begin{aligned} u(s) &= 0, \quad s \neq I, \\ u(s) &= \lambda v(s)/|v(s)|^2, \quad s \in I \end{aligned}$$

where $\lambda > 0$ is chosen so that $|u(s)|_\infty \leq 1$ for all s, so that $u \in \mathcal{U}$. Then

$$\int_0^t v^T(s) u(s) ds = \int_I \lambda ds > 0.$$

This is a contradiction and the desired result follows. □

Recall Definition 2.4.2 of the boundary of a set.

Lemma 3.3.15. *Let $\Omega \subset \mathbb{R}^n$ be a convex set. Then z is in $\partial\Omega$ if and only if there exists a vector $y \in \mathbb{R}^n$ such that $y^T(x - z) < 0$ for any $x \in \Omega$.*

Proof. First suppose that $z \in \partial\Omega$. Since Ω is convex there is a hyperplane $\mathcal{H}$ through z such that Ω is entirely on one side of $\mathcal{H}$. Choose $y \perp \mathcal{H}$ pointing in a direction away from Ω. Appeal to the figure to see that $y^T(x - z) < 0$.

Now we prove the converse. Suppose that $z \in \operatorname{Int}\Omega$. Then for any $y \in \mathbb{R}^n$ there is $\epsilon > 0$ such that $z + \epsilon y \in \operatorname{Int}\Omega$. For $x = z + \epsilon y$ we have

$$y^T(z - x) = y^T(-\epsilon y) = -\epsilon\|y\|^2 < 0$$

and the proof is complete. □

Example 3.3.16. *Consider the $1-$dimensional system*

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

Since $u(t) \in [-1, 1]$ for all $t \geq 0$ we deduce that $\dot{x}(t) \geq 0$ if $x(t) \geq 1$ and $\dot{x}(t) \leq 0$ if $x(t) \leq 1$. Thus $\mathcal{C} \subset (-1, 1)$. For any $x_0 \in \mathcal{C}(t)$ we have

$$x_0 = -\int_0^t e^{-s} u(s) ds$$

for some $u \in \mathcal{U}$. Thus

$$\begin{aligned} x_0 &\leq \int_0^t e^{-s}|u(s)|ds \\ &= \int_0^t e^{-s}ds \\ &= 1 - e^{-t} < 1. \end{aligned}$$

Now let $|x_0| \leq 1 - e^{-t}$ and define the control u as follows. For $s \in [0, t^*]$ we set

$$\begin{aligned} u(s) &= 1, \quad x_0 < 0 \\ u(s) &= -1, \quad x_0 > 0; \end{aligned}$$

and for $s \in (t^*, t]$ we set $u = 0$. Here t^* is chosen so that $x(t^*) = 0$ which requires, in the case $x_0 > 0$, that

$$0 = e^{t^*}x_0 - \int_0^{t^*} e^{(t^*-s)}ds$$

so that t^* is the unique solution of the equation

$$\int_0^{t^*} e^{-s}ds = x_0$$

given by $1 - e^{-t^*} = x_0$ (note that $0 \leq x_0 < 1 - e^{-t}$ so the equation indeed has a unique solution $t^* \in [0, t]$.) A similar argument holds when $x_0 < 0$.

Then we see that

$$\mathcal{C}(t) = [-1 + e^{-t}, 1 - e^{-t}]$$

so that

$$\mathcal{C} = \bigcup_{t>0} \mathcal{C}(t) = (-1, 1).$$

The preceding example demonstrates a situation, with $n = 1$, where $0 \in \operatorname{Int} \mathcal{C}$ but $\mathcal{C} \neq \mathbb{R}$. We now give conditions which ensure global controllability, i.e. $\mathcal{C} = \mathbb{R}^n$.

Theorem 3.3.17. *Let $\operatorname{rank} G = n$ for the linear system (3.9) with $A \in \mathbb{R}^{n\times n}$ constant. Consider a control $u \in \mathcal{U}$ with $U = [-1, 1]^m$. If for every eigenvalue of A we have $\operatorname{Re} \lambda < 0$ then $\mathcal{C} = \mathbb{R}^n$.*

Proof. Let $B_0(\delta)$ denote the ball of radius δ centred at the origin in $\mathbb{R}^n$. By Theorem 3.3.11 we have $0 \in \operatorname{Int} \mathcal{C}$ and hence, for small enough δ, $B_0(\delta) \subset \mathcal{C}$.

Now for any given $x_0 \in \mathbb{R}^n$ set $u(t) = 0$ for $t \in [0, t^*]$, where t^* will be determined in what follows. The dynamics on $[0, t^*]$ is given by

$$\begin{aligned} x(t) &= e^{At} x_0 \\ &= \sum_{i=1}^{n} \alpha_i \Big(\sum_{j=0}^{k_i - 1} (A - \lambda_i I)^j \frac{t^j}{j} \Big) e^{\lambda_i t} v^{(i)} \end{aligned}$$

as we showed in (2.11). Here the λ_i are the eigenvalues of A and all have negative real parts, whilst the $v^{(i)}$ are the generalized eigenvectors of A. It is immediate that there is $t^* = t^*(\delta) > 0$ such that $\|x(t^*)\| \leq \delta$.

Let $x^* = x(t^*)$. Since $x^* \in B_0(\delta) \subset \mathcal{C}$, there exists $u^* \in \mathcal{U}$ which brings the system (3.9) started at $x_0 = x^*$ to the origin in finite time T. For the system started at x_0, define

$$u(t) = \left\{ \begin{array}{cc} 0 & t \in [0, t^*] \\ u^*(t - t^*) & t \in (t^*, t^* + T]. \end{array} \right\}.$$

Then $x(t^* + T) = 0$ by construction and we have proved the desired result. □

Example 3.3.18. ROCKET *The matrix A here has a zero eigenvalue of mutiplicity two and so the preceding theorem does not apply. We show later that the assumptions can be weakened to deal with some situations of this type.*

Theorem 3.3.19. *Consider the linear system* (3.9) *with constant A and B, and restricted controls. The set of systems for which the controllable set contains an open neighbourhood of the origin is open and dense in the metric space $\mathcal{M}$.*

Proof. The proof is similar to that for Theorem 3.3.5, combined with Theorem 3.3.11. Indeed, by Theorem 3.3.11, $\operatorname{rank} G(A, B) = n$ implies $0 \in \text{Int } \mathcal{C}$. The proof of Theorem 3.3.5 shows that the set of matrices $\{A, B\}$ with $\operatorname{rank} G(A, B) = n$ is open and dense in the space of pairs of metrices of $n \times n, n \times m$ size. Hence the linear control systems for which $0 \in \text{Int } \mathcal{C}$ are dense in the metric space of all linear control systems (with the same norm as above). □

For the following it is useful to recall the real Jordan canonical form discussed at the end of section 2.3.

Theorem 3.3.20. *Consider the linear system* (3.9) *with* $A \in \mathbb{R}^{n\times n}$ *constant. Consider a control* $u \in \mathcal{U}$ *with* $U = [-1,1]^m$*. Then* $\mathcal{C} = \mathbb{R}^n$ *if and only if* $\operatorname{rank} G = n$ *and* $\operatorname{Re}\lambda \leq 0$ *for every eigenvalue* λ *of* A*.*

Proof. **If** Suppose that $\operatorname{rank} G = n$ and $\operatorname{Re}\lambda \leq 0$ for every eigenvalue λ of A. Assume for contradiction that $\mathcal{C} \neq \mathbb{R}^n$. By convexity of $\mathcal{C}$ (Theorem 3.3.7) we have from Lemma 3.3.15 that $\exists z \in \mathbb{R}^n$, on the boundary of $\mathcal{C}$, and non-zero $y \in \mathbb{R}^n$ such that $y^T(x_0 - z) \leq 0$ for any $x_0 \in \mathcal{C}$. Hence if $\mathcal{C} \neq \mathbb{R}^n$ there exists $y \in \mathbb{R}^n$ and $\alpha = y^T z \in \mathbb{R}$ such that $y^T x_0 \leq \alpha$ for all $x_0 \in \mathcal{C}$. We will establish a contradiction to this, by showing that for any nonzero $y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ we can find $x_0 \in \mathcal{C}$ such that $y^T x_0 > \alpha$.

Let nonzero $y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ be given. Then $x_0 \in \mathcal{C}$ iff there exists $t \in \mathbb{R}^+$ and $u \in \mathcal{U}$ such that

$$x_0 = -\int_0^t e^{-As} Bu(s)ds.$$

Thus we want to show that $\exists u \in \mathcal{U}$ such that

$$-\int_0^t y^T e^{-As} Bu(s)ds > \alpha.$$

Let $v(s) = (y^T e^{-As} B)^T \in \mathbb{R}^m$. Since $\operatorname{rank} G = n$, we know by Corollary 3.3.4 that $v(\cdot) \neq 0$ on $[0,t]$. Now define

$$u_0(s) = \left\{ \begin{array}{ll} -\frac{v(s)}{|v(s)|} & \text{when } v(s) \neq 0, \\ 0, & \text{when } v(s) = 0 \end{array} \right\}$$

and $u(s) = \lambda u_0(s)$ with $\lambda > 0$ chosen to ensure that $|u|_\infty \leq 1$ so that $u \in \mathcal{U}$. Then

$$-\int_0^t v^T(s)u(s)ds = \lambda \int_0^t |v(s)|ds.$$

We will show that

$$\lim_{t\to\infty} \int_0^t |v(s)|ds = \infty$$

so that, for some $t > 0$,

$$-\int_0^t v^T(s)u(s)ds = \lambda \int_0^t |v(s)|ds > \alpha.$$

We argue this step by contradiction too. Assume that

$$\int_0^\infty |v(s)|ds < \infty \tag{3.25}$$

and let

$$\phi(t) := \int_t^\infty v(s)ds$$

noting that the convergence of the integral in (3.25) implies immediately that

$$\phi(t) \to 0 \text{ as } t \to \infty. \tag{3.26}$$

Let p be the characteristic polynomial of A. Since $p(A) = 0$ we have

$$\begin{aligned} p\big(-\frac{d}{dt}\big)v^T(t) &= p\big(-\frac{d}{dt}\big)\big(y^T e^{-tA}B\big) \\ &= y^T\Big(p\big(-\frac{d}{dt}\big)e^{-tA}\Big)B \\ &= y^T p\big(A\big)e^{-tA}B \\ &= 0. \end{aligned}$$

Since $-\frac{d}{dt}\phi = v$ we deduce that $\phi(t)$ satisfies

$$-\frac{d}{dt}p\big(-\frac{d}{dt}\big)\phi = 0.$$

Hence each component of ϕ is a linear combination of terms of the form $a(t)e^{\mu t}$ where μ solves $\mu p(\mu) = 0$. After imposing an ordering on the μ we obtain

$$\begin{aligned} \mu_j = -\lambda_j, \quad j = 1, \ldots, n \\ \mu_{j+1} = 0. \end{aligned}$$

Thus the k^{th} component of $\phi(t)$ may be written as

$$k^{(k)}(t) = \sum_{j=1}^{n} a_j^{(k)}(t)e^{-\lambda_j t} + a_{n+1}^{(k)}(t)$$

and the $a_j^{(k)}$ are polynomials in t. Since $-\lambda_j \geq 0$ this contradicts (3.26) and the **If** part of the proof is concluded.

Only If Here we must assume that either $\operatorname{rank} G < n$ or that $\operatorname{Re}\lambda > 0$ for some eigenvalue λ of A and we consider the two cases separately. The first assumption, $\operatorname{rank} G < n$, implies by Corollary 3.3.12 that $\mathcal{C}(t)$ is contained in a hyperplane in $\mathbb{R}^n$ for all $t > 0$ and it follows that $\mathcal{C}$ is contained in the same hyperplane. Hence $\mathcal{C} \neq \mathbb{R}^n$ and we are done. Thus it remains to consider the situation where $\operatorname{Re} \lambda_1 > 0$ for some eigenvalue λ_1 of A. Recall the real Jordan form $A = Q^{-1} J Q$ and set $x = Qz$. Then (3.9) gives

$$\begin{aligned} \dot{z}(t) &= Q^{-1} A Q z + Q^{-1} B u \\ &= \tilde{A} z + w \end{aligned}$$

Note that, since $|u(t)|_\infty \leq 1$ for all t, it follows that there is constant $K > 0$ such that $|w|_\infty \leq K$ for all t.

Without loss of generality we may assume that λ_1 appears in J_1 and consider the case of real and complex λ_1 seperately. In the real case we find that the first component of z satisfies the equation

$$\dot{z}^{(1)} = \lambda_1 z(1) + w^{(1)}.$$

Now chose $z^{(1)}(0) > \frac{K}{\lambda_1}$. Then $z^{(1)}(t)$ is always increasing and the origin cannot be reached. Hence $\mathcal{C} \neq \mathbb{R}^n$.

Now consider the complex case where $\lambda_1 = \alpha + i\beta$ for some $\alpha > 0$. The first two equations for z take the form

$$\begin{aligned} \dot{z}^{(1)} &= \alpha z^{(1)} - \beta z^{(2)} + w^{(1)} \\ \dot{z}^{(2)} &= \beta z^{(1)} + \alpha z^{(2)} + w^{(2)}. \end{aligned}$$

Let $v = (z^{(1)}, z^{(2)})^T$ and $\xi = (w^{(1)}, w^{(2)})^T$. Taking an inner-product of the equations with v we find that

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} |v|^2 &= \alpha |v|^2 + v^T \xi \\ &\geq |v| \Big(\alpha |v| - |\xi| \Big) \end{aligned}$$

Note that $|\xi| \leq \sqrt{2} K$. From this it follows that, if $|v(0)| > 2\sqrt{K}/\alpha$ then $|v(t)|$ is increasing and the origin cannot be reached. Hence $\mathcal{C} \neq \mathbb{R}^n$. This completes the proof. □

3.4 Nonlinear Systems

In this section we study the controllable set for the nonlinear system (3.7):

$$\begin{aligned}\dot{x} &= f(x,u)\\ x(0) &= x_0\end{aligned}$$

with $x \in \mathbb{R}^n$, $f : \mathbb{R}^n \times [-1,1]^m \to \mathbb{R}^n$ and $u \in \mathcal{U}$. As in the linear problems studied previously, we assume that our target is $0 \in \mathbb{R}^n$. We denote the solution, and its dependence on initial condition x_0 and control u, by $x(t) = \varphi(t, x_0, u)$.

Assumption 3.4.1. The function $f : \mathbb{R}^n \times [-1,1]^m \to \mathbb{R}^n$ is continuous and there is $c > 0$ such that, for all $x, y \in \mathbb{R}^n$ and $u \in [-1,1]^m$,

$$\begin{aligned}|f(x,u)| &\leq c\big(1 + |x| + |u|\big),\\ |f(x,u) - f(y,u)| &\leq c|x-y|,\end{aligned}$$

The following is an immediate corollary of Theorem 2.6.1:

Corollary 3.4.2. *In equation* (3.7) *suppose that f satisfies Assumptions 3.4.1. Then for any given integrable function $u : [0,T] \to [-1,1]^m$ there exists exactly one solution of equation* (3.7)*.*

Recall the reachable set from Definition 3.2.3 and the time-reversed system given by (3.8):

$$\begin{aligned}\dot{z} &= -f(z, \tilde{u}), \quad t > 0\\ z(0) &= x_1\end{aligned}$$

We denote the solution, and its dependence on initial condition x_0 and control u, by $x(t) = \varphi^-(t, x_1, u)$. Let $K^-(t) = K^-(t,0)$ the reachable set from $x_1 = 0$. The reachable set $K^-(t)$ for (3.8) is the same as the controllable set $\mathcal{C}(t)$ for (3.7). Therefore, the statement that $B_\delta(0) \subset \mathcal{C}(t)$ for (3.7) is equivalent to saying that $B_\delta(0) \subset K^-(t)$ for (3.8). We use this fact to prove the following.

Theorem 3.4.3. *Consider the system* (3.7) *with $f(0,0) = 0$ and $f : \mathbb{R}^n \times [-1,1]^m \to \mathbb{R}^n$ continuously differentiable and satisfying Assumptions 3.4.1. Let $A := D_x f(0,0)$ and $B := D_u f(0,0)$. If $\operatorname{rank} G(A,B) = n$ then $0 \in \operatorname{Int} \mathcal{C}$.*

Proof. We will show that $\exists\delta$ sufficiently small and T sufficiently large so that $B_\delta(0) \subset K^-(T)$ for the time-reversed system (3.8).

Consider first the linearized time-reversed system

$$\begin{aligned} \dot{z} &= -Az - Bu, \\ z(0) &= 0. \end{aligned}$$

Since $\operatorname{rank} G(A,B) = n$ we deduce from Theorem 3.3.11 that there exists $\delta > 0$ such that $0 \in \mathbb{R}^n$ can be steered to any point in $B_\delta(0)$.

Let $e_1, \ldots, e_n$ be canonical basis vectors in $\mathbb{R}^n$ and choose $\alpha_i \in (0,\delta), \quad i = 1, \ldots, n$. Then there exist $(u_i, T_i) \in \mathcal{U} \times \mathbb{R}^+$ such that

$$\alpha_i e_i = -\int_0^{T_i} e^{-A(1-s)} B u_i(s) ds, \quad i = 1, \ldots, n.$$

Set $T = \max_{1 \le j \le n} T_j$ and assume, without loss of generality, that $T = T_1$. Extend the $u_i(s)$, currently defined on $(0, T_i)$ to $(0, T)$ by setting $u_i \equiv 0$ for $t \in (T_i, T)$. We then have

$$\alpha_i e_i = -\int_0^{T} e^{-A(1-s)} B u_i(s) ds, \quad i = 1, \ldots, n. \tag{3.27}$$

Now set $\gamma = (\gamma_1, \ldots, \gamma_n)$ and construct the control

$$u(t,\gamma) = \sum_{i=1}^n \gamma_i u_i(t).$$

We will show that, by appropriate choice of the γ, we can steer the solution $\varphi^-\big(t, 0, u(t,\gamma)\big)$ of (3.8) with $\tilde{u} = u(t,\gamma)$ and $x_1 = 0$ to an arbitrary point in $B_\delta(0)$. Note that $\varphi^-\big(t, 0, u(t,\gamma)\big)$ satisfies

$$\varphi^-\big(t, 0, u(t,\gamma)\big) = -\int_0^t f\Big(\varphi^-\big(s, 0, u(s,\gamma)\big), u(s,\gamma)\Big)\, ds. \tag{3.28}$$

Let

$$v_i(t) = \frac{\partial \varphi^-}{\partial \gamma_i}\big(t, 0, u(t,\gamma)\big)\Big|_{\gamma=0}$$

so that

$$V(t) = \big(v_1(t)\ v_2(t)\ \ldots\ v_n(t)\big) = \frac{\partial \varphi^-}{\partial \gamma}\big(t, 0, u(t,\gamma)\big)\Big|_{\gamma=0}.$$

This derivative exists by Exercise 2-7.

By differentiating (3.28) with respect to each γ_i we find

$$\begin{aligned} V(t) &= -\int_0^t \frac{\partial}{\partial \gamma} f\Big(\varphi^-\big(s,0,u(s,\gamma)\big), u(s,\gamma)\Big)\Big|_{\gamma=0} ds \\ &= -\int_0^t D_x f\big(\varphi^-(s,0,0),0\big)V(s) + D_u f\big(\varphi^-(s,0,0),0\big)\big(u_1(s) \ \dots \ u_n(s)\big)\, ds \\ &= -\int_0^t AV(s) + B\big(u_1(s) \ \dots \ u_n(s)\big)\, ds \end{aligned}$$

Hence

$$v_i(t) = -\int_0^t A_f v_i(s) + B_f u_i(s)\, ds$$

which implies that $v_i(T) = \alpha_i e_i$ and therefore

$$V(T) = \big(v_1(T) \ \dots \ v_n(T)\big) = \big(\alpha_1 e_1 \ \dots \ \alpha_n e_n\big).$$

Since $\alpha_i > 0$ for $i = 1, \dots, n$ and e_i are linearly independent, $V(T)$ is invertible. Now consider the map $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ to be defined as

$$F(x,\gamma) = x - \varphi^-\big(T, 0, \sum_{i=1}^n \gamma^i u_i\big).$$

The function F is continuously differentiable, with $F(0,0) = 0$ and $D_\gamma F(0,0) = -V(T)$ invertible. Hence by the Implicit Function Theorem (Theorem 2.4.4) there exists $\theta > 0$ and $\Gamma(x) = \big(\Gamma^1(x), \cdots, \Gamma^n(x)\big)$ continuous at zero with $\Gamma(0) = 0$ such that $F\big(x, \Gamma(x)\big) = 0$. Therefore

$$\varphi^-\big(t, 0, \sum_{i=1}^n \Gamma^i(x) u_i(t)\big) = 0$$

for any $x \in B_\theta(0)$. Thus for each $x \in B_\theta(0)$ the control $\sum_{i=1}^n \Gamma^i(x) u_i(t)$ steers the system from 0 at time $t = 0$ to x at time $t = T$.

Thus we have proved that $B_\theta(0) \subset K^-(T)$ for (3.8) and hence that $B_\theta(0) \subset \mathcal{C}(T)$ for (3.7). □

Remarks 3.4.4.

- *By Exercise 3-7 the above theorem implies that the controllable set $\mathcal{C}$ is open.*
- *Another tool used to discuss such examples is the concept of Lie bracket, based on ideas from differential geometry.*

- $\operatorname{rank} G(A,B) < n$ *does not imply* $0 \notin \operatorname{Int}\mathcal{C}$.

Example 3.4.5. *Consider the system*

$$\dot{x}_1 = x_1 + \sin x_2 + x_1 e^{x_2} \tag{3.29a}$$

$$\dot{x}_2 = x_2^2 + u \tag{3.29b}$$

where $u \in \mathcal{U}$. *We have*

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

Thus

$$\operatorname{rank}\Big(B, AB\Big) = \operatorname{rank} I = 2$$

and so the system (3.29) *is locally controllable by Theorem 3.4.3.*

Now replace u *by* u^3 *in* (3.29) *to obtain*

$$\dot{x}_1 = x_1 + \sin x_2 + x_1 e^{x_2} \tag{3.30a}$$

$$\dot{x}_2 = x_2^2 + u^3 \tag{3.30b}$$

where $u \in \mathcal{U}$. *Theorem 3.4.3 does not apply in this case because* $B = 0$. *But the system is still locally controllable because, if* $u^\star$ *steers* (3.29) *from* $x_0 \in B_\delta(0)$ *to* 0 *then* $(u^\star)^{\frac{1}{3}}$ *steers* (3.30) *from* x_0 *to* 0.

Exercises

Exercise 3-1. Consider the discrete-time control system (3.1) with $n = 2, m = 1$ and $B = (1,0)^T$ and unrestricted controls. Under what condition on $A \in \mathbb{R}^{2\times 2}$ is the system controllable?

Exercise 3-2. Consider a discrete-time control system of the form

$$z_{n+k} + \sum_{j=0}^{k-1} \alpha_j z_{n+j} = u_n.$$

Here $z_n, u_n \in \mathbb{R}$ and the control sequence $\{u_n\}_{n\geq 0}$ is unrestricted. By introducing the vector

$$x_n := (z_n, \cdots, z_{n+k-1})^T$$

write the system in the form (3.1) for matrices A, B and appropriate dimensions m, n which you should specify. Is the system controllable?

Exercise 3-3. Consider the control system

$$\dot{x} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} x + u\, b$$

with unestricted control $u : [0, \infty) \to \mathbb{R}$ measurable. Consider choosing b to be equal to each of the three unit vectors $e_i \in \mathbb{R}^3$. For which i is $\mathcal{C} = \mathbb{R}^3$? If $\mathcal{C} \neq \mathbb{R}^3$ then characterize $\mathcal{C}$.

Exercise 3-4. Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ and assume that rank $B = p$. Prove that
rank $[B,\ AB, \ldots, A^{n-1}B] = n$ implies that rank $[B,\ AB, \ldots, A^{n-p}B] = n$. Using this result, restate Theorem 3.3.2.

Exercise 3-5. Consider the linear system

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0 \in \mathbb{R}^n$$

with $x(t) \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $u : [0, \infty) \to \mathbb{R}^m$ measurable. Prove Theorem 3.3.2 by using Theorem 3.3.11 and comparing the controllable set for $u(t) \in \Omega$ with the controllable set for $u(t) \in k\Omega$, $k > 1$, where $k\Omega = \{kv | v \in \Omega\}$.)

Exercise 3-6. Let A, B and u be as in the previous exercise. Show that the system

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} = Ay + Bu, \qquad y(0) \in \mathbb{R}^n, \quad \dot{y}(0) \in \mathbb{R}^n$$

is controllable in $\mathbb{R}^{2n}$ if and only if the pair (A, B) is controllable. (Zabczyk 1995)

Exercise 3-7. Consider the system

$$\dot{x} = f(x, u), \qquad x(0) = x_0 \in \mathbb{R}^n$$

with $x(t) \in \mathbb{R}^n$, $f : \mathbb{R}^n \times [-1,1]^m \to \mathbb{R}^n$, and the control $u : [0,\infty) \to [-1,1]^m$ measurable. Assume that $f(x,u)$ is continuously differentiable on $\mathbb{R}^n \times [-1,1]^m$ and $f(0,0) = 0$. Show that the controllable set $\mathcal{C}$ is open if and only if $0 \in \operatorname{Int} \mathcal{C}$.

Exercise 3-8. *Kalman decomposition.* Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Assume that $\operatorname{rank} G(A,B) = l$. Show that there exists a nonsigular matrix P such that

$$PAP^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \qquad PB = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},$$

with $A_{11} \in \mathbb{R}^{l\times l}$, $A_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$ and $B_1 \in \mathbb{R}^{l\times m}$. Prove that $\operatorname{rank} G(A_{11}, B_1) = l$.

Exercise 3-9. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 \, \mathrm{e}^{x_1} - x_1 - u_1 \\ \dot{x}_2 &= x_1 \, \mathrm{e}^{x_2} - x_2 + u_2, \end{aligned}$$

with $u : [0,\infty) \to [-1,1]^2$ measurable. Show that $\mathcal{C} \neq \mathbb{R}^2$ and $0 \in \operatorname{Int} \mathcal{C}$.

Exercise 3-10. Let $g(\xi_1, \ldots, \xi_{n+1})$, with $\xi_1, \ldots, \xi_{n+1} \in \mathbb{R}$, be continuously differentiable. Assume that $g(0,\ldots,0) = 0$, $\frac{\partial g}{\partial \xi_{n+1}}(0,\ldots,0) \neq 0$. Show that the system

$$\frac{\mathrm{d}^n z}{\mathrm{d}t^n} = g\left(\frac{\mathrm{d}^{n-1} z}{\mathrm{d}t^{n-1}}, \ldots, z(t), u(t)\right)$$

considered as a system in $\mathbb{R}^n$ is locally controllable at $0 \in \mathbb{R}^n$ (Zabczyk 1992).

Exercise 3-11. Consider the system $\dot{x} = f(x,u)$ with $x(0) = x_0 \in \mathbb{R}^n$, $f : \mathbb{R}^n \times [-1,1]^m \to \mathbb{R}^n$ continuously differentiable and $f(0,0) = 0$. Let $A = f_x(0,0)$ and $B = f_u(0,0)$ and assume that $\operatorname{rank} G(A,B) = n$. Show that if the solution $\varphi(t,x_0) \equiv 0$ of the free system $\dot{x} = f(x,0)$ is globally asymptotically stable then the controllable set $\mathcal{C} = \mathbb{R}^n$.

Exercise 3-12. Consider the system

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= -\sin x_1 + u\end{aligned}$$

with unrestricted control u.Take $V(x) = 2x_1^2 + x_2^2 + 2x_1x_2$ and use the Lyapunov function method to show that the origin is asymptotically stable if the control u is chosen appropriately.

Exercise 3-13. Consider the system

$$\dot{x} = Ax + Bu + c(t), \quad x(0) = x_0 \in \mathbb{R}^n$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $u : [0,\infty) \to [-1,1]^m$ measurable and $c : [0,\infty) \to \mathbb{R}^n$ a given continuous function. Show that the reachable set $K(t,x_0)$ is closed and convex.

Exercise 3-14. Consider the system

$$\dot{x} = f(x,u), \quad x(0) = x_0 \in \mathbb{R}^n$$

with $u(t) \in [-1,1]^m$. Let $\tau_1 > 0$ be fixed and show that if

$$|x(t)^T f(x(t),u(t))| \leq c_1\|x(t)\|^2 + c_2,$$

with c_1 and c_2 constant and for $t \in [0,\tau_1]$, then $\|x(t)\|$ is uniformly bounded on $[0,\tau_1]$.

Exercise 3-15. Prove Corollary 3.3.13

Exercise 3-16. By studying the proof of Theorem 3.1.4 show that, for the discrete time control system (3.1), $\operatorname{rank} G = n$ implies that $\mathcal{C} = \mathbb{R}^n$. Is it necessary to assume that A is invertible?

Chapter 4

Stability and Stabilization

The previous chapter was concerned entirely with open loop control where, for given initial condition, the objective is to find a control which ensures that the state is driven to the origin in finite time. We note that the controllability matrix $G(A, B)$ plays a big role in this theory. This chapter is concerned with closed loop control, where the objective is to choose a relationship between control u and state x which ensures asymptotic stability of the origin. Although this differs from open loop control we will see that, once again, the controllability matrix $G(A, B)$ plays a big role in the theory.

In section 4.1 we introduce stabilizability in the context of discrete time linear systems. The remainder of the chapter concerns the continuous time setting. We are interested in closed-loop (or feedback) control for the system (3.7):

$$\begin{aligned} \dot{x} &= f(x, u), \\ x(0) &= x_0. \end{aligned}$$

Thus we are interested in finding a map $u \mapsto c(x)$ which stabilizes this system in the neighbourhood of a particular state of interest. We will concentrate on linear maps c. In section 4.2 we study an important link between asymptotic stability and controllability. Section 4.3 generalizes stabilizability to continuous time linear systems. In section 4.4 we introduce an important link between stabilizability of linear systems and controllability of systems with scalar controls. Section 4.5 concerns nonlinear systems and stabilizability.

4.1 Discrete-Time Linear Systems

As in the previous chapter, we introduce ideas via discrete time problems. We study the linear control system (3.1) which has the form

$$x_{k+1} = Ax_k + Bu_k \tag{4.1}$$

for constant matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Our aim is to determine feedback (closed-loop) controls which stabilize the origin for this iteration. Thus we introduce $K \in \mathbb{R}^{m\times n}$ and consider closed-loop linear controls of the form $u_k = Kx_k$. Our aim is to choose the matrix K to achieve desirable objectives in (4.1). With this linear feedback control (4.1) becomes

$$x_{k+1} = (A + BK)x_k. \tag{4.2}$$

We will be particularly interested in the control objective of making the origin stable. (See Exercise 2-9).

Definition 4.1.1. *The system* (4.1) *is* stabilizable *if there exists a matrix* $K \in \mathbb{R}^{m\times n}$ *such that the origin for the linear system* (4.2) *is asymptotically stable.*

Example 4.1.2. *As an example of why this might be a useful concept consider a dynamical system of the form*

$$x^\dagger_{k+1} = Ax^\dagger_k, \tag{4.3}$$

where $x^\dagger \in \mathbb{R}^n$ *and* $A \in \mathbb{R}^{n\times n}$*. Imagine that we do not know the initial starting point of the system, but that we do observe a sequence* $\{y_k\}_{k\in\mathbb{Z}^+}$ *in* $\mathbb{R}^m$ *where*

$$y_k = Dx^\dagger_k \tag{4.4}$$

for some $D \in \mathbb{R}^{m\times n}$*. Think of the case where* D *is not invertible, for example if* $m < n$*. To learn the sequence* $\{x^\dagger_k\}_{k\in\mathbb{N}}$ *we combine the model* (4.3) *with the observed data and consider the control system* (4.1) *with*

$$u_k = y_k - Dx_k.$$

This leads to the closed-loop control

$$x_{k+1} = Ax_k + B(y_k - Dx_k). \tag{4.5}$$

It is natural to ask whether this recovers the true signal for large enough k*.*

Since $y_k = Dx_k^\dagger$ we have from (4.3) *that*

$$x_{k+1}^\dagger = Ax_k^\dagger + B(y_k - Dx_k^\dagger). \tag{4.6}$$

Defining $e_k = x_k - x_k^\dagger$ and subtracting (4.6) *from* (4.5) *we see that*

$$e_{k+1} = (A - BD)e_k. \tag{4.7}$$

We wish to determine when the sequence $e_k \to 0$ as $k \to \infty$. The next theorem addresses this question.

Theorem 4.1.3. *Consider Example 4.1.2. Assume that the control system* (4.1) *is stabilizable. Then there is a choice of observation matrix $D \in \mathbb{R}^{m\times n}$ such that, for any $x_0 \in \mathbb{R}^n$, we have that $|x_k - x_k^\dagger| \to 0$ as $k \to \infty$.*

Proof. Simply take $D = -K$ in the argument preceding the theorem statement. □

4.2 Controllability and Stability for Linear Systems

As a warm-up we consider the stability of the linear system (2.19):

$$\dot{x} = Ax, \tag{4.8a}$$

$$x(0) = x_0. \tag{4.8b}$$

Although no control appears directly in this stability question we will ascertain conditions which relate it to to the controllability of the system

$$\dot{z} = A^T z + D^T u. \tag{4.9}$$

Here Corollary 3.3.4 will play a big role, as it will in subsequent sections.

Theorem 4.2.1. *Let $A \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{m\times n}$ and suppose that $\operatorname{rank} G(A^T, D^T) = n$; thus* (4.9) *is controllable. Then equation* (4.8) *is asymptotically stable at* 0 *if and only if there exists positive-definite P such that*

$$PA + A^T P = -D^T D. \tag{4.10}$$

Proof. Assume first that A is asymptotically stable. Set $Q = -D^T D$ and note that this is negative semi-definite. By applying the same methods as used to prove Theorem 2.7.17 we deduce that

$$P = \int_0^\infty e^{A^T t} D^T D e^{At} dt$$

solves the desired equation. It remains to establish that it is positive-definite; this does not follow from the proof used in Theorem 2.7.17 since $Q = -D^T D$ is not strictly negative-definite. Instead we use controllability of (4.9) to establish positive-definiteness of P. Since $\operatorname{rank} G(A^T, D^T) = n$ Corollary 3.3.4 shows that $v^T e^{A^T t} D^T$ is not identically zero for any $v \neq 0$. From this it follows that P is positive-definite:

$$\langle v, Pv \rangle = \int_0^\infty |De^{At} v|^2 dt > 0.$$

To prove the converse we suppose that there exists positive definite P solving (4.10). Let $V(x) = \langle x, Px \rangle$ and note that

$$\begin{aligned}\dot{V}(x) = \langle DV(x), Ax \rangle &= \langle x, (A^T P + PA)x \rangle \\ &= -\langle x, D^T Dx \rangle \\ &= -\langle Dx, Dx \rangle \\ &= -|Dx|^2 \\ &\leq 0\end{aligned}$$

proving stability by Theorem 2.7.9(i). It follows that $\operatorname{Re} \lambda \leq 0$ for any eigenvalue λ of A by Theorem 2.7.5. To show asymptotic stability it suffices to show that $\operatorname{Re} \lambda < 0$ for all such λ, by Theorem 2.7.3. For contradiction assume that there is an eigenvalue/eigenvector pair (λ_1, v_1) with $\operatorname{Re} \lambda_1 = 0$. We have

$$\begin{aligned}v_1^* e^{A^T t} P e^{At} v_1 &= (e^{\lambda_1 t} v_1)^* P e^{\lambda_1 t} v_1 \\ &= v_1^* e^{-\lambda_1 t} P e^{\lambda_1 t} v_1 \\ &= v_1^* P v_1\end{aligned}$$

demonstrating that

$$v_1^* \Big(e^{A^T t} P e^{At} - P \Big) v_1 = 0. \tag{4.11}$$

Now observe that

$$\begin{aligned}\frac{d}{dt}\Big(e^{A^Tt}Pe^{At}\Big) &= e^{A^Tt}\Big(A^TP + PA\Big)e^{At}\\ &= -e^{A^Tt}D^TDe^{At}.\end{aligned}$$

Integrating between 0 and t gives the identity

$$e^{A^Tt}Pe^{At} - P = -\int_0^t e^{A^Ts}D^TDe^{As}ds$$

so that

$$\begin{aligned}v_1^*\Big(e^{A^Tt}Pe^{At} - P\Big)v_1 &= -\int_0^t \big(De^{As}v_1\big)^*\big(De^{As}v_1\big)ds\\ &= -\int_0^t |De^{As}v_1|^2 ds \neq 0,\end{aligned}$$

again using Corollary 3.3.4, contradicting (4.11). □

4.3 Stabilizability of Linear Systems

Recall Example 1.2.2 which concerns stabilization of the INVERTED PENDULUM in a vertical position. That example motivates the contents of this section. We generalize to consider the linear system

$$\dot{x} = Ax + Bu, \tag{4.12a}$$

$$x(0) = x_0 \in \mathbb{R}^n \tag{4.12b}$$

with $u \in \mathcal{U}$ an unrestricted control. Here $x \in \mathbb{R}^n, u \in \mathbb{R}^m$ and $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$. Recall that if the system (4.12) is controllable we say, briefly, (A, B) is controllable (see Definition 3.2.1).

We study feedback controls with $u = Kx$, for $K \in \mathbb{R}^{m\times n}$, giving rise to the resulting linear system

$$\dot{x} = (A + BK)x, \tag{4.13a}$$

$$x(0) = x_0 \in \mathbb{R}^n. \tag{4.13b}$$

Definition 4.3.1. *The system* (4.12) *is* stabilizable *if there exists a matrix* $K \in \mathbb{R}^{m\times n}$ *such that the linear system* (4.13) *is asymptotically stable.*

Theorem 4.3.2. *If (A, B) is controllable then it is stabilizable.*

We do not prove this theorem at this stage, but we note

Example 4.3.3. *We revisit Example 1.2.3. The previous theorem shows that, if (A, B) is controllable then there exists an observation matrix $D \in \mathbb{R}^{p\times n}$ such that the equation for e is asymptotically stable. As a consequence, the control system (1.7) has solution which satisfies $|x(t) - x^\dagger(t)| \to 0$ as $t \to \infty$. Thus the true signal can be recovered from the observation, asymptotically for large time, provided the system is controllable and an appropriate observation matrix is chosen.*

4.4 Linear Control Problems With Scalar Control

A key role in control theory is played by control problems of the form

$$\frac{d^n z}{dt^n} + a_{n-1}\frac{d^{n-1}z}{dt^{n-1}} + \cdots + a_1\frac{dz}{dt} + a_0 z = u.$$

Here the objective is to use the *scalar* control u to control the vector $y \in \mathbb{R}^n$ given by

$$y = \begin{pmatrix} z \\ \frac{dz}{dt} \\ \vdots \\ \frac{d^{n-2}z}{dt^{n-2}} \\ \frac{d^{n-1}z}{dt^{n-1}} \end{pmatrix}$$

from a given initial configuration to a desired state, which we take as 0. Many systems of this form arise in applications: the ROCKET Example 1.2.1 and the INVERTED PENDULUM EXAMPLE 1.2.2 both have this form with $n = 2$, the equation being Newton's second law and the controller being an applied force.

In addition to this applied motivation we will show that systems of this form play a useful role in the understanding of the more general linear control problem. To this end it is useful to observe that the system can be written in the following canonical form. Define the companion matrix (see Definition

2.3.10) $\tilde{A}$ and vector $\tilde{b}$ by

$$\tilde{A} := \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & \cdots & -a_{n-1} \end{pmatrix}, \ \tilde{b} := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{4.14}$$

Then

$$\dot{y} = \tilde{A}y + \tilde{b}u, \tag{4.15a}$$
$$y(0) = y_0 \in \mathbb{R}^n. \tag{4.15b}$$

Now we consider the general control system

$$\dot{x} = Ax + bu, \tag{4.16a}$$
$$x(0) = x_0 \in \mathbb{R}^n, \tag{4.16b}$$

for $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^{n \times 1}$. Thus the control is scalar. Our first theorem shows that, if controllable, then this system can be converted, via the transformation $y = Px$, to a canonical control problem of the form (4.15). The proof of the theorem specifies the transformation matrix P.

Theorem 4.4.1. *Suppose that the system* (4.16) *is controllable. Then* $\exists$ *nonsingular* $P \in \mathbb{R}^{n \times n}$ *such that* $y = Px$ *solves* (4.15) *with* $\tilde{A} = PAP^{-1}$ *and* $\tilde{b} = Pb$. *Furthermore the defining vector* $a = (a_0, \cdots, a_{n-1}) \in \mathbb{R}^n$ *is formed from the coefficients of the charateristic polynomial of* A*:*

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

where $a_n = 1$.

Proof. Once P is constructed with the desired matrix properties, application of it to (4.16) gives

$$P\dot{x} = PAx + Pbu = PAP^{-1}y + Pbu = \tilde{A}y + \tilde{b}u$$

and hence that

$$\dot{y} = \tilde{A}y + \tilde{b}u$$

as required. Thus it remains to construct P with the desired matrix properties. Recall that, by Theorem 3.3.2,

$$G = G(A, b) = \begin{pmatrix} b, & Ab, & \cdots, & A^{n-1}b \end{pmatrix} \in \mathbb{R}^{n\times n}$$

satisfies $\operatorname{rank} G = n$, and is hence invertible. Now define the vectors $v^{(j)}$ by the condition that

$$(G^{-1})^T = \big(v^{(1)}, \cdots, v^{(n)}\big)$$

so that $v^{(j)}$ is the j^{th} row of G^{-1}. Because $G^{-1}G = I$, we deduce that

$$\langle v^{(j)}, A^{k-1}b\rangle = \delta_{jk}. \tag{4.17}$$

Now define P by the identity

$$P^T = \big(v^{(n)}, A^T v^{(n)}, \cdots, (A^T)^{n-1}v^{(n)}\big).$$

We need to show that P is nonsingular and satisfies $PAP^{-1} = \tilde{A}$, $Pb = \tilde{b}$. To show that $\det P \neq 0$ suppose for contradiction that $\exists \gamma = (\gamma_1, \cdots, \gamma_n) \in \mathbb{R}^n \setminus \{0\}$ such that

$$\sum_{j=1}^{n} \gamma_j (v^{(n)})^T A^{j-1} = 0.$$

Applying the identity above to b on the right and using (4.17) we deduce that $\gamma_n = 0$. Applying the identity to Ab on the right and again using (4.17) then gives $\gamma_{n-1} = 0$. Proceeding inductively we deduce that $\gamma \equiv 0$, giving the desired contradiction.

By the Cayley-Hamilton Theorem 2.3.7 we have

$$PA = \begin{pmatrix} (v^{(n)})^T A \\ (v^{(n)})^T A^2 \\ \vdots \\ (v^{(n)})^T A^n \end{pmatrix} = \begin{pmatrix} (v^{(n)})^T A \\ (v^{(n)})^T A^2 \\ \vdots \\ -\sum_{j=0}^{n-1} a_j (v^{(n)})^T A^j \end{pmatrix}$$
$$= \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & \cdots & -a_{n-1} \end{pmatrix} \begin{pmatrix} (v^{(n)})^T \\ (v^{(n)})^T A \\ \vdots \\ (v^{(n)})^T A^{n-2} \\ (v^{(n)})^T A^{n-1} \end{pmatrix} = \tilde{A}P.$$

Finally observe that, by (4.17),

$$Pb = \begin{pmatrix} (v^{(n)})^T b \\ (v^{(n)})^T Ab \\ \vdots \\ (v^{(n)})^T A^{n-2} b \\ (v^{(n)})^T A^{n-1} b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

and the proof is complete. □

This result is interesting in its own right, for problems driven by a scalar control. But the next lemma shows that there is a direct link to general linear control problems, a link which we then exploit in Theorem 4.4.3.

Lemma 4.4.2. *Suppose that (A, B) is controllable. Then $\exists F \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$ such that $(A + BF, b)$ is controllable.*

Proof. Since the original system (4.12) is controllable $\exists u^{(0)} \in \mathbb{R}^m$ such that $Bu^{(0)} \neq 0$ because otherwise $B = 0$. We first show that $\exists \{u^{(j)}\}_{j=0}^{n-1}$, all in $\mathbb{R}^m$, such that the vectors $\{d^{(j)}\}_{j=1}^{n}$ now constructed are linearly independent. We set

$$\begin{aligned} d^{(1)} &= Bu^{(0)} \\ d^{(2)} &= Ad^{(1)} + Bu^{(1)} \\ \vdots &= \vdots \\ d^{(n)} &= Ad^{(n-1)} + Bu^{(n-1)}. \end{aligned}$$

For contradiction, suppose not: $\{d^{(1)}, \cdots, d^{(k)}\}$ are linearly independent for some $k \in \{1, \cdots n-1\}$ but for any $u \in \mathbb{R}^m$, $d^{(k+1)} = Ad^{(k)} + Bu \in E = \text{span}\,\{d^{(1)}, \cdots, d^{(k)}\}$.
Take $u = 0$. Then we deduce that $Ad^{(k)} \in E$. But since $Ad^{(k)} + Bu \in E$ for any $u \in \mathbb{R}^m$ we conclude that $Bu \in E$ for any $u \in \mathbb{R}^m$. But $Ad^{(j)} = d^{(j+1)} - Bu^{(j)}$ for $j = 1, \ldots, k-1$. Since $d^{(j+1)} \in E$ by assumption and $Bu \in E$ for any u, we deduce that $Ad^{(j)} \in E$ for $j = 1, \cdots, k-1$. We already know that $Ad^{(k)} \in E$ and so $Ad^{(j)} \in E$ for $j = 1, \cdots, k$. Hence we have $AE \subset E$.

From this we deduce that $A^\ell Bu \in E$ for any $u \in \mathbb{R}^m$ and integer $\ell \geq 0$. We now show, using controllability, that $E = \mathbb{R}^n$ and hence that $k = n$. Suppose, for contradiction, that E is not equal to $\mathbb{R}^n$. Then $\exists x \in \mathbb{R}^n \backslash \{0\}$ such that $x \perp A^\ell B u$ for any $u \in \mathbb{R}^m$ and any integer $\ell \geq 0$. Therefore $x^T A^\ell B = 0$ for any integer $\ell \geq 0$. But this contradicts controllability by Theorems 3.3.2 and 3.1.2.

Now define b and F as follows: choose arbitrary $u^{(n)} \in \mathbb{R}^m$ and set

$$\begin{aligned} b = d^{(1)} &= Bu^{(0)} \\ Fd^{(j)} &= u^{(j)}, \quad j = 1, \dots, n. \end{aligned}$$

The latter is a valid definition of $F \in \mathbb{R}^{m \times n}$ as it defines its action on a set of n linearly independent vectors in $\mathbb{R}^n$. We then have, for $j = 1, \dots, n-1$,

$$\begin{aligned} d^{(j+1)} &= Ad^{(j)} + Bu^{(j)} \\ &= (A + BF)d^{(j)} \\ &= (A + BF)^j d^{(1)} \\ &= (A + BF)^j b. \end{aligned}$$

Thus

$$\left(\begin{array}{cccc} b & (A+BF)b & \cdots & (A+BF)^{n-1}b \end{array} \right) = \left(\begin{array}{cccc} d^{(1)} & d^{(2)} & \cdots & d^{(n)} \end{array} \right)$$

so that $\mathrm{rank} G(A + BF, b) = n$ as required. □

Before reading the following theorem it is useful to recall Definition 4.3.1 and Theorem 4.3.2. In particular note that Theorem 4.3.2 exhibits a particular choice of stabilizing matrix K which makes the origin stable for (4.13). The following theorem shows that, in fact, there exist an uncountable set of stabilizing matrices $K \in \mathbb{R}^{m \times n}$ for a given controllable system since we may ensure that $A + BK$ has any characteristic polynomial, and hence any set of eigenvalues, that we choose.

Theorem 4.4.3. *Suppose (A, B) is controllable and let $\beta = (\beta_0, \cdots, \beta_{n-1})^T \in \mathbb{R}^n$ be arbitrary. Then there exists $K \in \mathbb{R}^{m \times n}$ such that the characteristic polynomial of $A + BK$ is determined by β as follows:*

$$\det(\lambda I - A - BK) = \lambda^n + \beta_{n-1}\lambda^{n-1} + \cdots + \beta_0.$$

Proof. By Lemma 4.4.2 $\exists F \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^n$ such that $(A+BF, b)$ is controllable. Then by Theorem 4.4.1 $\exists P$ with $\det P \neq 0$ and such that

$$\tilde{A} = P(A+BF)P^{-1} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & \cdots & -a_{n-1} \end{pmatrix}$$

and

$$\tilde{b} = Pb = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

Recall that, in this construction, the vector $a = (a_0, \cdots, a_{n-1})^T$ is the vector defining the characteristic polynomial of $A+BF$; by Theorem 2.3.11 it is also the vector defining the characteristic polynomial of $\tilde{A}$. For any $g = (g_1, \cdots, g_n)^T \in \mathbb{R}^n$, the $n \times n$ matrix $\tilde{b}g^T$ has its last row equal to g^T and is zero everywhere else. Hence we have

$$\tilde{A} + \tilde{b}g^T = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 + g_1 & -a_1 + g_2 & \cdots & \cdots & -a_{n-1} + g_n \end{pmatrix}$$

with (again using Theorem 2.3.11) the characteristic polynomial

$$\det(\lambda I - \tilde{A} - \tilde{b}g^T) = \lambda^n + (a_{n-1} - g_n)\lambda^{n-1} + \cdots + (a_0 - g_1). \tag{4.18}$$

On the other hand we have

$$\begin{aligned} \det(\lambda I - \tilde{A} - \tilde{b}g^T) &= \det(\lambda PP^{-1} - P(A+BF)P^{-1} - Pbg^T PP^{-1}) \\ &= \det\Big(P(\lambda I - A - BF - bg^T P)P^{-1}\Big) \\ &= \det\Big(\lambda I - A - BF - bg^T P\Big). \end{aligned} \tag{4.19}$$

By the proof of Lemma 4.4.2, $b = Bu^{(0)}$. Choose $K = F + u^{(0)} g^T P$. Equations (4.18) and (4.19) imply that

$$\det\Big(\lambda I - A - BK\Big) = \lambda^n + (a_{n-1} - g_n)\lambda^{n-1} + \cdots + (a_0 - g_1).$$

Now choose $g_i = a_{i-1} - \beta_{i-1}$. The result follows. □

4.5 Stabilizability of Nonlinear Systems

We return to the nonlinear control problem (3.7):

$$\begin{aligned} \dot{x} &= f(x,u), \quad t > 0 \\ x(0) &= x_0. \end{aligned}$$

We consider unrestricted controls $\mathcal{U}$. We assume that $f(\overline{x}, \overline{u}) = 0$ for some $\overline{x} \in \mathbb{R}^n, \overline{u} \in \mathbb{R}^m$. Our aim is to determine controls which stabilize the origin.

Theorem 4.5.1. *Let Assumptions 3.4.1 hold, Let f be continuously differentiable and set $A = D_x f(\overline{x}, \overline{u})$ and $B = D_u f(\overline{x}, \overline{u})$. Assume that the linear system* (4.12) *is controllable with this choice of A, B. Then $\exists F \in \mathbb{R}^{m \times n}$ such that the closed-loop system*

$$\dot{x} = f\big(x, \overline{u} + F(x - \overline{x})\big) \tag{4.20}$$

is asymptotically stable at $x = \overline{x}$.

Proof. Since the linear system (4.12) is controllable, we deduce from Theorem 4.4.3 that $\exists F \in \mathbb{R}^{m \times n}$ such that all eigenvalues of $A_c := A + BF$ have negative real parts. If we can show that A_c is the linearization of the right-hand side of (4.20) then the result follows by Theorem 2.7.20.

To establish the required linearization result we proceed as follows. Define

$$f_c(x) = f\big(x, \overline{u} + F(x - \overline{x})\big).$$

By the chain rule we have

$$D_x f_c(x) = D_x f\big(x, \overline{u} + F(x - \overline{x})\big) + D_u f\big(x, \overline{u} + F(x - \overline{x})\big) F$$

and setting $x = \overline{x}$ gives

$$\begin{aligned} D_x f_c(\overline{x}) &= D_x f(\overline{x}, \overline{u}) + D_u f(\overline{x}, \overline{u}) F \\ &= A + BF \end{aligned}$$

as required.

□

Exercises

Exercise 4-1. Consider the control system (4.1) with $n = 2$ and

$$A = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Note that if $B = 0$ the system is not stable. If $B = (1, 0)^T$ and $B = (0, 1)^T$ then determine whether the system is stabilizable. Give an explanation for your findings.

Exercise 4-2. Consider the discrete system (4.3) for a signal $\{x_k^\dagger\}_{k \in \mathbb{Z}^+}$, with $n = 2$ and A as in Exercise 4-1. Assume that observations are made in the form (4.4) with $m = 1$ and $D = (\frac{1}{2}, \frac{1}{2})\mathbb{R}^{1 \times 2}$. Design a control of the form (4.1), that is find a matrix $B \in \mathbb{R}^{2 \times 1}$ such that $\lim_{k \to \infty} |x_k - x_k^\dagger| = 0$.

Exercise 4-3. Consider system (2.19) with

$$A = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}.$$

By studying the related control problem (4.9) for an appropriate choice of D, show that (2.19) is asymptotically stable.

Exercise 4-4. Consider the control problem (4.12) in the case where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

and $B = (\frac{1}{2}, \frac{1}{2})^T$. Exhibit an explicit choice of K for which the feedback control $u = Kx$ gives rise to a stable system, using the proof of Theorem 4.3.2.

Exercise 4-5. By use of Theorem 4.4.3 exhibit an uncountable family of feedback controls $u = Kx$ which stabilizes the system from Exercise 4-4.

Exercise 4-6. Recall Exercise 3-3. For the choices of $b = e_i$ which make the system controllable, exhibit the transformation (4.14) which renders the system in canonical form (4.15).

Exercise 4-7. Consider the system $\dot{x} = Ax + Bu$ with

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Find a feedback matrix $F \in \mathbb{R}^{2\times 3}$ such that the characteristic polynomial of $A + BF$ is $(\lambda + 1)^3$. Use the following steps:

i) Find vectors $u_0, u_1, u_2 \in \mathbb{R}^2$ such that the vectors $e_1 = Bu_0$, $e_2 = Ae_1 + Bu_1$ and $e_3 = Ae_2 + Bu_2$ are linearly independent.

ii) Find $\tilde{F}$ such that $\tilde{F}e_1 = u_1$ and $\tilde{F}e_2 = u_2$. Check that the pair $(A + B\tilde{F}, Bu_0)$ is controllable.

iii) Find a nonsigular matrix $P \in \mathbb{R}^{3\times 3}$ such that the pair $(P(A+B\tilde{F})P^{-1}, PBu_0)$ is in control canonical form.

iv) Find a vector $g \in \mathbb{R}^3$ such that the matrix $A + B\tilde{F} + Bu_0 g^T P$ has the required characteristic polynomial.

Exercise 4-8. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1 - x_1 x_2^2 + x_2 \\ \dot{x}_2 &= u + x_2 \end{aligned}$$

Find a matrix $F \in \mathbb{R}^{1\times 2}$ such that the closed loop control $u = Fx$ makes the origin asymptotically stable.

Exercise 4-9. For the system

$$\begin{aligned}\dot{x}_1 &= x_1 x_2 + x_2\\ \dot{x}_2 &= u\end{aligned}$$

Find a matrix $F \in \mathbb{R}^{1\times 2}$ so that the control $u = Fx$ makes the origin asymptotically stable. Show that the origin is not globally asymptotically controllable (Sontag 1998).

Chapter 5

Observing and Filtering

In this chapter we consider questions relating to the combination of observed data and mathematical model. The objective is to use the observations to compensate for uncertainty in the initial condition for the model, or uncertainty in the model itself. We study *observability*, which concerns determination of the initial condition from the observed data, and *filtering* which concerns estimation of the current state of the system, given data up to that time. We demonstrate links between this theory and the theories of controllability, developed in previous chapters, and detectability which we introduce here.

5.1 Discrete Time: Observability and Duality

Discrete time systems will play a central role in this chapter. Consider the system

$$x_{k+1} = Ax_k + Bu_k, \tag{5.1a}$$

$$y_k = Dx_k \tag{5.1b}$$

where $x_k \in \mathbb{R}^n, u_k \in \mathbb{R}^m, y_k \in \mathbb{R}^p$ and $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}$ and $D \in \mathbb{R}^{p \times n}$.

Observability Question: Let $\mathsf{J} = \{0, \ldots, J-1\}$. Given the observation $\{y_k\}_{k \in \mathsf{J}}$ and the input control function $\{u_k\}_{k \in \mathsf{J}}$ can we find $\{x_k\}_{k \in \mathsf{J}}$?

Recall from (3.3) that the general solution to this problem may be written as

$$x_k = A^k x_0 + \sum_{j=0}^{k-1} A^{k-j-1} B u_j.$$

Using this we find that

$$\begin{aligned} y_k &= D x_k \\ &= D A^k x_0 + D \sum_{j=0}^{k-1} A^{k-j-1} B u_j \end{aligned}$$

so that

$$D A^k x_0 = y_k - D \sum_{j=0}^{k-1} A^{k-j-1} B u_j.$$

The observability question is hence equivalent to asking if x_0 is uniquely determined by this identity, given $\{y_j\}_{j \in \mathsf{J}}$ and $\{u_j\}_{j \in \mathsf{J}}$. We make this a precise definition.

Definition 5.1.1. *The system* (5.1) *is* observable *on* J *if, given* $\{y_j\}_{j \in \mathsf{J}}$ *and* $\{u_j\}_{j \in \mathsf{J}}$*, the initial condition* x_0 *is determined uniquely.*

Example 5.1.2. *Consider the system* $x_{k+1} = x_k + u_k$ *where* $x_k, u_k \in \mathbb{R}^2$ *and let* $y_k = \langle e, x_k \rangle$ *where* $e = (1, 0)^T$*. It is clear that the observation sequence* $\{y_k\}$ *contains no information about the second component of* $\{x_k\}$ *and hence that the system is not observable. It is, however, controllable: in the abstract notation of* (3.1) *we have* $m = n = 2$ *and* $A = B = I$*. Hence the controllability matrix has rank* 2 *and Theorem 3.1.4 shows that the system is controllable.*

The previous example shows that controllability and observability are not the same. However, they are linked through a duality principle which we now explain.

Theorem 5.1.3. *Suppose* A *is invertible. The system* (5.1) *is observable on* J *with* $J = n$ *if and only if the corresponding* dual system

$$\begin{aligned} z_{k+1} &= A^T z_k + D^T v_k, && (5.2a) \\ w_k &= B^T z_k, && (5.2b) \end{aligned}$$

with unrestricted control v_k*, is controllable.*

Proof. **If** Suppose that (5.1) is not observable on J with $J = n$. Then for given $\{y_j\}_{j\in\mathsf{J}}$ and $\{u_j\}_{j\in\mathsf{J}}$ there exists sequences x, η with $x \neq \eta$ such that

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ \eta_{k+1} &= A\eta_k + Bu_k, \end{aligned}$$

and

$$y_k = Dx_k = D\eta_k, \quad k = 0, \ldots, n-1. \tag{5.3}$$

Let $\xi_k = x_k - \eta_k$ and $\xi_0 = x_0 - \eta_0 \neq 0$. Then

$$\xi_{k+1} = A\xi_k,$$

and so

$$\xi_k = A^k\xi_0.$$

By (5.3) we deduce that

$$\xi_0^T(A^T)^k D^T = 0, \quad k = 0, \ldots, n-1$$

for some $\xi_0 \neq 0$. By Theorem 3.1.2(i) we deduce that $\operatorname{rank} G(A^T, D^T) < n$ and so (5.2) is not controllable. Thus controllability of (5.2) implies observability of (5.1) on J with $J = n$.

Only if. Suppose now that (5.2) is not controllable. Then

$$\operatorname{rank} G(A^T, D^T) = \operatorname{rank}\Big(D^T, A^T D^T, \ldots, (A^T)^{n-1}D^T\Big) < n.$$

By Theorem 3.1.2(i) we deduce that $\exists\xi \in \mathbb{R}^n\backslash\{0\}$, orthogonal to all columns of $G(A^T, D^T)$ and so $\xi^T(A^T)^k D^T = 0$ for $k = 0, \ldots, n-1$. Hence

$$DA^k\xi = 0, \quad k = 0, \ldots, n-1.$$

Now consider the system

$$x_{k+1} = Ax_k + Bu_k, \quad x_0 = \lambda\xi,$$

for any $\lambda \in \mathbb{R}$ and let $u_k \equiv 0$. Then

$$y_k = Dx_k = DA^k x_0 = \lambda DA^k\xi = 0$$

for all $k \in \mathsf{J}$ with $J = n$. Since this is true for all λ it demonstrates that the system (5.1) is not observable on J with $J = n$. Hence observability of (5.1) on J with $J = n$ implies controllability of (5.2). $\square$

Example 5.1.4. *We revisit Example 5.1.2. The dual system* (5.2) *has matrix* $A = I$ *and* $D = (1, 0)$. *Clearly* $(A^T)^n D^T = D^T$ *for all* $n \in \mathbb{N}$ *and hence* $\operatorname{rank} G(A^T, D^T) = 1 < 2$. *Hence the dual system is not controllable and the original system is not observable on* J *with* $J = n$.

Remarks 5.1.5. *Note that observability of* (5.1) *is, by Theorem 5.1.3, determined entirely by the properties of the matrix pair* (A, D). *For this reason we refer simply to the pair* (A, D) *being observable.*

5.2 Continuous Time: Observability and Duality

Consider the system

$$\dot{x} = Ax + Bu, \quad x(0) = x_0 \tag{5.4a}$$

$$y = Dx \tag{5.4b}$$

where $x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m, y(t) \in \mathbb{R}^p$ and $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}$ and $D \in \mathbb{R}^{p \times n}$.

Observability Question: Given the observation $y(\cdot)$ and the input control function $u(\cdot)$, can we find $\{x(t)\}_{t>0}$?

Using (2.10) we have

$$\begin{aligned} y(t) &= Dx(t) \\ &= De^{At}x_0 + D\int_0^t e^{A(t-s)}Bu(s)ds \end{aligned}$$

so that

$$De^{At}x_0 = y(t) - D\int_0^t e^{A(t-s)}Bu(s)ds. \tag{5.5}$$

The observability question is hence equivalent to asking if there exists a unique x_0 determined by (5.5), given $y(\cdot)$ and $u(\cdot)$.

Definition 5.2.1. *The system* (5.4) *is* observable *if, given* $y(\cdot)$ *and* $u(\cdot)$ *on any time interval* $[0, t)$, *with* $t > 0$, *the initial condition* x_0 *is determined uniquely.*

Example 5.2.2. *Consider the system*

$$\begin{aligned}\dot{x}_1 &= x_1 + u_1,\\ \dot{x}_2 &= x_2 + u_2,\\ y &= x_1.\end{aligned}$$

Since x_1 and x_2 are independent it is clear that the observation y contains no information about x_2. Hence the system is not observable. It is, however, controllable: in the abstract notation of (3.9) *we have $m = n = 2$ and $A = B = I$. Hence the controllability matrix has rank* 2 *and Theorem 3.3.2 shows that the system is controllable.*

It is clear from equation (5.5) that the system is observable when $p = n$ and D is invertible. However if $p < n$ then observability turns out to be equivalent to controllability for a related dual problem. The situation is very similar to that arising in discrete time in section 5.1.

Theorem 5.2.3. *The system* (5.4) *is observable if and only if the corresponding* dual system

$$\dot{z} = A^T z + D^T v, \quad z(0) = z_0 \tag{5.6a}$$

$$w = B^T z \tag{5.6b}$$

with unrestricted control v is controllable.

Proof. **If** Suppose that (5.4) is not observable. Then for fixed $y(\cdot)$ and $u(\cdot)$ there exists $x_0 \neq \eta_0$ such that

$$\begin{aligned}\dot{x} &= Ax + Bu, \quad x(0) = x_0\\ \dot{\eta} &= A\eta + Bu, \quad \eta(0) = \eta_0\end{aligned}$$

and

$$y(t) = Dx(t) = D\eta(t). \tag{5.7}$$

Let $\xi = x - \eta$ and $\xi_0 = x_0 - \eta_0$. Then ξ satisfies

$$\dot{\xi} = A\xi, \quad \xi(0) = \xi_0$$

and so

$$\xi(t) = e^{At}\xi_0.$$

Hence, by (5.7),

$$D\xi(t) = De^{At}\xi_0 = 0.$$

Setting $t = 0$ we get $D\xi_0 = 0$. Differentiating k times and setting $t = 0$ gives $DA^k\xi_0 = 0$ so that $\xi_0^T(A^T)^kD^T = 0$ and therefore

$$\xi_0^T(A^T)^kD^T = 0, \quad k = 0, \ldots, n-1.$$

By Theorem 3.1.2(i) we deduce that $\operatorname{rank} G(A^T, D^T) < n$ and so (5.6) is not controllable. Thus controllability of (5.6) implies observability of (5.4).

Only if Suppose now that (5.6) is not controllable. Then

$$\operatorname{rank} G(A^T, D^T) = \operatorname{rank}\Big(D^T, A^TD^T, \ldots, (A^T)^{n-1}D^T\Big) < n.$$

By Theorem 3.1.2(i) we deduce that $\exists\xi \in \mathbb{R}^n\backslash\{0\}$, orthogonal to all columns of $G(A^T, D^T)$ and so $\xi^T(A^T)^kD^T = 0$ for $k = 0, \ldots, n-1$. Hence

$$DA^k\xi = 0, \quad k = 0, \ldots, n-1.$$

Consider

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = \lambda\xi,$$

for any $\lambda \in \mathbb{R}$ and and let $u(\cdot) = 0$. We will show that $y(t) = Dx(t) = 0$ for all $t \geq 0$; since this is true for all λ it demonstrates that the system (5.4) is not observable. By the Cayley-Hamilton Theorem 2.3.7

$$A^n = -a_{n-1}A^{n-1} - \cdots - a_0I$$

for $a_i \in \mathbb{R}$ coefficients of the characteristic polynomial and $i = 0, \cdots, n-1$. Thus $DA^n\xi = 0$ and, similarly,

$$DA^{n+1}\xi = DA\big(-a_{n-1}A^{n-1} - \cdots - a_0I\big)\xi = 0.$$

Proceeding inductively we deduce that $DA^k\xi = 0$ for all $k \in \mathbb{Z}^+$. Since

$$x(t) = \lambda e^{tA}\xi = \lambda\sum_{k=0}^{\infty}\frac{t^kA^k}{k!}\xi$$

we have

$$Dx(t) = \lambda\sum_{k=0}^{\infty}\frac{t^kDA^k}{k!}\xi = 0$$

for every $\lambda \in \mathbb{R}$ and so (5.4) is not observable (alternatively, one can apply Corollary 3.3.4 to get the same result). Hence observability of (5.4) implies controllability of (5.6). □

Example 5.2.4. *We revisit Example 5.2.2. The dual system* (5.6) *has matrix* $A = I$ *and* $D = (1, 0)$. *Clearly* $(A^T)^n D^T = D^T$ *for all* $n \in \mathbb{N}$ *and hence* $\operatorname{rank} G(A^T, D^T) = 1 < 2$. *Hence the system is not observable.*

Remarks 5.2.5.

- *Note that observability of* (5.4) *is, by Theorem 5.2.3, determined entirely by the properties of the matrix pair* (A, D). *For this reason we refer simply to the pair* (A, D) *being observable.*
- *Consider the question of determining the asymptotic stability of the origin for* (2.19) *and note that this is equivalent to determining asymptotic stability of the equation*

$$\dot{x} = A^T x.$$

Recall the system (5.4). *Together Theorems 5.2.3 and Theorem 4.2.1 state that if* (5.4) *is observable then the equation* (2.19) *for* x *is asymptotically stable if and only if there is a positive-definite* P *satisfying*

$$PA + A^T P = -D^T D.$$

Consider the linear system

$$\dot{x} = (A + LD)x, \tag{5.8a}$$

$$x(0) = x_0 \in \mathbb{R}^n. \tag{5.8b}$$

Definition 5.2.6. *The system* (5.4) *is* detectable *if there exists matrix* $L \in \mathbb{R}^{n \times p}$ *such that the linear system* (5.8) *is asymptotically stable.*

Theorem 5.2.7. *If* (A, D) *is observable then it is detectable.*

Proof. By Theorem 5.2.3, the matrix pair (A, D) is observable if and only if the matrix pair (A^T, D^T) is controllable. By Theorem 4.3.2 we deduce that if (A, D) is observable then (A^T, D^T) is stabilizable. There exists $K \in \mathbb{R}^{p \times n}$ such that

$$\dot{x} = (A^T + D^T K)x,$$

$$x(0) = x_0 \in \mathbb{R}^n.$$

is asymptotically stable. Hence the system (5.8) is asymptotically stable with $L = K^T$. □

Example 5.2.8. *Recall Example 1.2.3 concerning* SIGNAL PROCESSING. *Theorem 5.2.7 demonstrates that, provided* (A, D) *is observable, there is a choice of control matrix* B *such that* $|x(t) - x^\dagger(t)| \to 0$ *as* $t \to \infty$.

5.3 Discrete Time: Kalman Filter

Observability is concerned with determining the *initial state* of a dynamical system, from observations at later times. Another natural question is to determine the *current state*, from observations up to that time. This is an idea that we introduced in Example 4.1.2 of section 4.1 where we showed how an *ad hoc* combination of model and observed data could be used to create a control system which converges to the true signal, if the control system is stabilizable and an appropriate observation is made. In the notation that we will use in this section that *ad hoc* control rule system can be written as

$$m_{k+1} = Am_k + B(y_k - Dm_k).$$

In this section we derive a similar control rule of the form

$$m_{k+1} = Am_k + K_{k+1}(y_{k+1} - DAm_k), \tag{5.9}$$

by using ideas from probability to design the control. We study a problem analogous to that in section 5.1, in the presence of observational noise. Under specific assumptions on the form of the noise we show how to derive the control rule (5.9).

To this end we consider the system

$$x_{k+1} = Ax_k, \quad k \in \mathbb{Z}^+ \tag{5.10a}$$

$$y_{k+1} = Dx_{k+1} + \eta_{k+1}, \quad k \in \mathbb{Z}^+ \tag{5.10b}$$

where $x \in \mathbb{R}^n, y \in \mathbb{R}^p, \eta \in \mathbb{R}^p$ and $A \in \mathbb{R}^{n \times n}, D \in \mathbb{R}^{p \times n}$. We assume that we observe $\{y_k\}_{k \in \mathbb{N}}$ and that the η_k represent noise that enters the observation of Dx_k. As in section 5.1 and Example 4.1.2 we assume that x_0 is not known, and that the objective is to determine the state of the system from the observations.

The presence of noise in the observations introduces novel aspects into this problem; in particular, it allows us to exploit the structure of this noise to make rational decisions about how to estimate the state. We will work in the setting where it is assumed that the η_k are an i.i.d. sequence of random variables with distribution $\eta_1 \sim N(0, \Gamma)$. We will also assume that, although x_0 is not known precisely, it is drawn from a distribution $N(m_0, C_0)$ and is independent of the noise sequence $\{\eta_k\}$. These statistical assumptions, together with the linearity of the problem, will enable us to work entirely in a Gaussian framework.

In this setting the system states $\{x_j\}_{j\geq 0}$ and the observations $\{y_j\}_{j\geq 1}$ are all random variables. We let $Y_k = \{y_j\}_{j=1}^{k} \in \mathbb{R}^{kp}$ and $X_k = \{x_j\}_{j=0}^{k} \in \mathbb{R}^{(k+1)n}$. Our objective is to find the distribution of the random variable $x_k|Y_k$. We will first show that it is Gaussian with distribution $N(m_k, C_k)$ and we will then determine equations for the mean m_k and covariance C_k. This will constitute a control system in which the update $m_k \mapsto m_{k+1}$ provides the optimal combination of model and data required to estimate the state.

Lemma 5.3.1. *The random variables (X_k, Y_k), $X_k|Y_k$ and $x_k|Y_k$ are all Gaussian.*

Proof. We have that x_0 is Gaussian and, by (5.10a), X_k is a linear transformation of x_0. Hence, by Theorem 2.8.9, X_k is Gaussian. The random variable $Y_k|X_k$ is also Gaussian; in fact it is the independent product of the Gaussians $N(Dx_j, \Gamma)$ for $j = 1, \dots, k$ (see Exercise 2-20). By Theorem 2.8.6 with $n \mapsto n(k+1)$ and $m \mapsto pk$ we deduce that (X_k, Y_k) and $X_k|Y_k$ are both Gaussian. Finally note that $x_k|Y_k$ is simply the marginal distribution of the random variable $X_k|Y_k$ with respect to x_k and is itself Gaussian by virtue of Theorem 2.8.7. □

Two theorems now follow in which we use Example 2.8.12, and formulae (2.31) and (2.32) respectively, to derive the Kalman filter. In the first theorem we assume that A is invertible and that C_0 and Γ are positive definite. The more general case, in which we assume only that Γ is positive-definite, is addressed afterwards, in the second theorem. Under this invertibility assumption on A we have the following.

Theorem 5.3.2. *Assume that $x_0 \sim N(m_0, C_0)$ and that the observational noise sequence $\{\eta_k\}_{k\in\mathbb{N}}$ is independent of x_0 and is i.i.d. with $\eta_1 \sim N(0, \Gamma)$. Assume further that A is invertible and that C_0 and Γ are positive definite. Then $x_k|Y_k \sim N(m_k, C_k)$ where C_k is positive definite and*

$$C_{k+1}^{-1} = (AC_kA^T)^{-1} + D^T\Gamma^{-1}D \tag{5.11a}$$

$$C_{k+1}^{-1}m_{k+1} = (AC_kA^T)^{-1}Am_k + D^T\Gamma^{-1}y_{k+1}. \tag{5.11b}$$

Proof. Assume, for the purposes of induction, that $x_k|Y_k \sim N(m_k, C_k)$, with C_k positive-definite, interpreting $x_0|Y_0$ as simply x_0 (since Y_0 is not defined). We note that the matrix C_{k+1} is well-defined by (5.11a). This is because that formula gives, for $x \in \mathbb{R}^n \backslash \{0\}$,

$$\begin{aligned}\langle x, C_{k+1}^{-1} x\rangle &= \langle x, (A^{-1})^T C_k^{-1} A^{-1} x\rangle + \langle x, D^T \Gamma^{-1} D x\rangle \\ &= \langle A^{-1} x, C_k^{-1} A^{-1} x\rangle + \langle D x, \Gamma^{-1} D x\rangle \\ &\geq \langle A^{-1} x, C_k^{-1} A^{-1} x\rangle \\ &> 0.\end{aligned}$$

In the penultimate line we have used the fact that Γ and hence Γ^{-1} is positive definite, but note that D may not be invertible, hence the inequality is *not strict*; and in the last line we have used that C_k, and hence C_k^{-1}, is positive-definite, and A is invertible so that the inequality is *strict*. Thus C_{k+1}^{-1} is positive-definite symmetric and hence so is C_{k+1}.

The derivation proceeds in two steps. In the first (i) we use the linear map $x_{k+1} = A x_k$ to find $x_{k+1}|Y_k$ from $x_k|Y_k$; in the second (ii) we use Bayes' Theorem to find $x_{k+1}|Y_{k+1}$ from $x_{k+1}|Y_k$. For (i) note that by Theorem 2.8.9 and the inductive hypothesis we know that $x_{k+1}|Y_k \sim N(A m_k, A C_k A^T)$. For (ii) note that Bayes' Theorem in the form of Remarks 2.8.4 gives

$$\begin{aligned}\mathbb{P}(x_{k+1}|Y_{k+1}) &= \mathbb{P}(x_{k+1}|\{Y_k, y_{k+1}\}) \\ &\propto \mathbb{P}(y_{k+1}|\{Y_k, x_{k+1}\}) \mathbb{P}(x_{k+1}|Y_k) \\ &= \mathbb{P}(y_{k+1}|x_{k+1}) \mathbb{P}(x_{k+1}|Y_k).\end{aligned}$$

The last step uses the fact that $y_{k+1}|\{Y_k, x_{k+1}\} = y_{k+1}|x_{k+1}$ because of the independence of the η_k.

Now $y_{k+1}|x_{k+1} \sim N(D x_{k+1}, \Gamma)$ and $x_{k+1}|Y_k \sim N(A m_k, A C_k A^T)$. Thus Bayes' theorem shows that $x_{k+1}|Y_{k+1}$ is a Gaussian random variable with pdf $\rho(x) \propto \exp\bigl(-J(x)\bigr)$ where

$$J(x) = \frac{1}{2}\Bigl|\Gamma^{-\frac{1}{2}}\bigl(y_{k+1} - D x\bigr)\Bigr|^2 + \frac{1}{2}\Bigl|(A C_k A^T)^{-\frac{1}{2}}\bigl(x - A m_k\bigr)\Bigr|^2. \qquad (5.12)$$

We can complete the square and write

$$J(x) = \frac{1}{2}\Bigl|C_{k+1}^{-\frac{1}{2}}(x - m_{k+1})^2\Bigr|^2.$$

Details are not repated as the derivation is simply an application of Theorem 2.8.6(ii) with $x_{k+1} \mapsto x$, $y_{k+1} \mapsto y$, $m_0 = A m_k, C_0 = A C_k A^T, H = D$ and Γ as is. The desired formulae are also displayed in Example 2.8.12, formulae (2.31). □

The form (5.11) for the Kalman filter, which works with inverse of the covariance operator, is the simplest form in which to derive the filter, and is also useful for analysis. However, for the purposes of implementation, the following form is useful. It is also valid without the restrictive assumptions on invertibility of A and C_0 that we introduced for the preceding derivation.

We define the *predicted mean* $\hat{m}_{k+1}$ and *predicted covariance* $\widehat{C}_{k+1}$ by

$$\hat{m}_{k+1} = Am_k \tag{5.13a}$$

$$\widehat{C}_{k+1} = AC_kA^T. \tag{5.13b}$$

We define the *innovation* by

$$d_{k+1} = y_{k+1} - D\hat{m}_{k+1}, \tag{5.14}$$

the *innovation covariance* by

$$S_{k+1} = D\widehat{C}_{k+1}D^T + \Gamma \tag{5.15}$$

and the *Kalman gain* by

$$K_{k+1} = \widehat{C}_{k+1}D^TS_{k+1}^{-1}. \tag{5.16}$$

Theorem 5.3.3. *Assume that $x_0 \sim N(m_0, C_0)$ and that the observational noise sequence $\{\eta_k\}_{k\in\mathbb{N}}$ is independent of x_0 and is i.i.d. with $\eta_1 \sim N(0, \Gamma)$. Assume further that Γ is positive definite. Then $x_k|Y_k \sim N(m_k, C_k)$ where*

$$C_{k+1} = (I - K_{k+1}D)\widehat{C}_{k+1} \tag{5.17a}$$

$$m_{k+1} = \hat{m}_{k+1} + K_{k+1}d_{k+1}. \tag{5.17b}$$

Proof. The derivation proceeds using Bayes' Theorem, as in the proof of Theorem 5.3.2, leading to formula (5.12) for the negative logarithm of the pdf. From that expression, applying the formulae (2.32) for conditioned Gaussians from Example 2.8.12, we find that

$$C_{k+1} = \widehat{C}_{k+1} - \widehat{C}_{k+1}D^TS_{k+1}^{-1}D\widehat{C}_{k+1}$$
$$m_{k+1} = \hat{m}_{k+1} + \widehat{C}_{k+1}D^TS_{k+1}^{-1}d_{k+1}.$$

Substituting the definition (5.16) of the Kalman gain matrix gives the desired form of the Kalman update equations. □

Remarks 5.3.4. *Note that the update formula (5.17b) may be written as*

$$m_{k+1} = Am_k + K_{k+1}(y_{k+1} - DAm_k).$$

As such it is structurally very similar to the ad hoc *control rule* (4.5) *that we wrote down in Example 4.1.2 in order to estimate the state of the system by combining the observations and model. In the notation of this section that* ad hoc *control rule may be written as*

$$m_{k+1} = Am_k + B(y_k - Dm_k).$$

However the Kalman update formula has two important differences: (i) the fixed control matrix B has been replaced by a time-dependent *control matrix K_{k+1}: the Kalman gain; (ii) the difference between data and model is now evaluated at time $k+1$ and not at time k as the innovation is $y_{k+1} - DAm_k$ rather than $y_k - Dm_k$. The first difference is particularly striking: the additional Gaussian structure imposed on the noise enables us to* design *a choice of* time-dependent *control matrix from first principles.*

5.4 Discrete Time: Kalman Smoother

In *Kalman filtering* we seek the distribution of $x_k|Y_k$ so that our estimate of the solution x_k at time k depends only on data upto time k. The *Kalman smoother* seeks the distribution of $X_J|Y_J$ so that the estimate of the solution x_k at any time $k \in \{0, \cdots, J\}$ depends on all the data $\{y_j\}_{j=1,\cdots,J}$.

Theorem 5.4.1. *The random variable $X_J|Y_J$ is Gaussian. If the (Gaussian) marginal distribution $x_0|Y_J$ is denoted $N(m', C')$ then the (Gaussian) marginal distributions $x_k|Y_J = N\big(A^k m', A^k C'(A^k)^T\big)$.*

Proof. The fact that $X_J|Y_J$ is Gaussian follows from Lemma 5.3.1. Consequently all the marginals $x_k|Y_J$ are Gaussian by Theorem 2.8.7. Since $x_k = A^k x_0$ we see that $x_k|Y_J$ is simply the image of $x_0|Y_J$ under the linear map A^k. Thus the desired result concerning the expression for the distributions of $x_k|Y_J$ in terms of $x_0|Y_J$ follows from Theorem 2.8.9 with $a = 0$ and $A \to A^k$. □

Theorem 5.4.2. *The Gaussian random variable* $x_0|Y_J \sim N(m', C')$ *where*

$$(C')^{-1} = C_0^{-1} + \sum_{j=1}^{J} (A^T)^j D^T \Gamma^{-1} D A^j$$

$$(C')^{-1} m' = C_0^{-1} m_0 + \sum_{j=1}^{J} (A^T)^j D^T \Gamma^{-1} y_j.$$

Proof. We have $x_0 \sim N(m_0, C_0)$. By Remarks 2.8.2, the density of this random variable is thus proportional to

$$\exp\bigl(-\frac{1}{2}|C_0^{-\frac{1}{2}}(x_0 - m_0)|^2\bigr). \tag{5.18}$$

Since $x_j = A^j x_0$ it follows that $y_j|x_0 \sim N(DA^j x_0, \Gamma)$ with density proportional to

$$\exp\bigl(-\frac{1}{2}|\Gamma^{-\frac{1}{2}}(y_j - DA^j x_0)|^2\bigr).$$

Furthermore, the random variables $y_j|x_0$ are independent, because the η_j are i.i.d. Thus, by (2.30), the random variable $Y_J|x_0$ has pdf which is the product of the pdfs for $y_j|x_0$ and $j = 1, \cdots, J$ and is hence proportional to

$$\exp\bigl(-\sum_{j=1}^{J} \frac{1}{2}|\Gamma^{-\frac{1}{2}}(y_j - DA^j x_0)|^2\bigr). \tag{5.19}$$

Combining (5.18) and (5.19) according to the Bayes' Theorem 2.8.3 (with $x \to x_0$ and $y \to Y_J$) we see that $x_0|Y_J$ has pdf proportional to $\exp\bigl(-J(x)\bigr)$ where

$$\begin{aligned} J(x) &= \frac{1}{2}|C_0^{-\frac{1}{2}}(x - m_0)|^2 + \frac{1}{2}\sum_{j=1}^{J} |\Gamma^{-\frac{1}{2}}(y_j - DA^j x)|^2 \\ &= \frac{1}{2}\langle (x - m_0), \mathbb{C}_0^{-1}(x - m_0)\rangle + \frac{1}{2}\sum_{j=1}^{J}\langle (y_j - DA^j x), \Gamma^{-1}(y_j - DA^j x)\rangle. \end{aligned}$$

Again we have used x as the argument of the pdf for $x_0|Y_J$, for economy of notation. Writing

$$J(x) = \frac{1}{2}\langle (x - m'), (C')^{-1}(x - m')\rangle + k$$

where k is independent of x we find, by matching the quadratic and linear terms in x respectively, the covariance and mean as given in the theorem statement. □

Theorem 5.4.3. *The means m'_k of $x_k|Y_J$, $k = 0, \cdots, J$, satisfy the following equations for $j = 1, \cdots, J$:*

$$\begin{aligned} m'_j &= Am'_{j-1}, & m'_0 &= m_0 + C_0 A^T \lambda_1 \\ \lambda_j &= A^T \lambda_{j+1} + D^T \Gamma^{-1}(y_j - Dm'_j), & \lambda_{J+1} &= 0. \end{aligned}$$

Proof. The proof of Theorem 5.4.2 shows that m'_0 is the minimizer of

$$J(x) := \frac{1}{2}|C_0^{-\frac{1}{2}}(x - m_0)|^2 + \frac{1}{2}\sum_{j=1}^{J}|\Gamma^{-\frac{1}{2}}(y_j - DA^j x)|^2.$$

Putting $x_j = A^j x_0$ and $X = \big((x_0)^T, \cdots, (x_J)^T\big)^T$ we see that m'_0 may be identified as the first component of the minimizer $M = \big((m'_0)^T, \cdots, (m'_J)^T\big)^T$ of

$$I(X) = \frac{1}{2}|C_0^{-\frac{1}{2}}(x_0 - m_0)|^2 + \frac{1}{2}\sum_{j=1}^{J}|\Gamma^{-\frac{1}{2}}(y_j - Dx_j)|^2$$

with respect to X satisfying the linear constraints

$$x_{j+1} = Ax_j, \quad j = 0, \cdots, J-1.$$

Under these constraints the minimizer will satisfy

$$m'_{j+1} = Am'_j$$

as required.

By using Lagrange multipliers we deduce that the minimizer of this problem coincides with critical points of the functional

$$L(X, \lambda) = I(X) + \sum_{j=1}^{J}\langle \lambda_j, x_j - Ax_{j-1}\rangle$$

where $\lambda = \big(\lambda_1^T, \cdots, \lambda_J^T\big)^T$ and, for notational convenience in what follows below, we define $\lambda_{J+1} = 0$.

Taking the derivative of L with respect to λ_j for $j = 1, \cdots, J$ and with respect to x_0, and setting $x_j = m'_j$ gives

$$m'_j = Am'_{j-1}, \quad m'_0 = m_0 + C_0 A^T \lambda_1.$$

Taking the derivative of L with respect to x_j for $j = 1, \cdots, J$, using $\lambda_{J+1} = 0$ and setting $x_j = m'_j$ gives

$$\lambda_j = A^T \lambda_{j+1} + D^T \Gamma^{-1}(y_j - Dm'_j), \quad \lambda_{J+1} = 0.$$

□

Remarks 5.4.4. *Recall that, because x_k is just the image of x_0 under the linear transformation, $x \mapsto A^k x$ it follows from Theorem 5.4.1 that $x_k|Y_J$ is Gaussian with mean $m'_j = A^j m'$ and covariance $C'_j = A^j C'(A^T)^j$ with m', C' as given in Theorem 5.4.2. It is the case, of course, that m'_0 given in Theorem 5.4.3 must be equal to m' from Theorem 5.4.2. We verify this by means of a direct calculation. From the iteration for the λ_j in Theorem 5.4.3 we have that*

$$\lambda_1 = \sum_{j=1}^{J} (A^T)^{j-1} D^T \Gamma^{-1} \big(y_j - DA^j m'_0\big)$$

so that

$$A^T \lambda_1 = \sum_{j=1}^{J} (A^T)^{j} D^T \Gamma^{-1} \big(y_j - DA^j m'_0\big).$$

Using the identity $m'_0 = m_0 + C_0 A^T \lambda_1$ we find that

$$(C_0)^{-1} m'_0 = C_0^{-1} m_0 + \sum_{j=1}^{J} (A^T)^{j} D^T \Gamma^{-1} \big(y_j - DA^j m'_0\big).$$

Collecting terms involving m'_0, and using the definition of C' from Theorem 5.4.2, we find that

$$(C')^{-1} m'_0 = C_0^{-1} m_0 + \sum_{j=1}^{J} (A^T)^{j} D^T \Gamma^{-1} y_j.$$

This shows that $m'_0 = m'$ where m' is as given in Theorem 5.4.2.

Exercises

Exercise 5-1. Is the sytem (5.1) with

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

observable on J with $J = 3$?

Exercise 5-2. Derive the equations for the mean m_k and variance c_k of the Kalman filter when applied to the system

$$x_{k+1} = ax_k, \quad x_0 \sim N(m_0, c_0) \tag{5.20}$$

$$y_k = x_k + \eta_k \tag{5.21}$$

where $a \neq 0$, $x_k, y_k, \eta_k \in \mathbb{R}$ and $\eta_1 \sim N(0, \sigma^2)$. Show that

$$c_k^{-1} = a^{-2k} c_0^{-1} + \Big(\sum_{j=0}^{k-1} a^{-2j}\Big)\sigma^{-2}.$$

Hence deduce that:

- if $|a| > 1$ then $c_k \to \sigma^2(|a|^2 - 1)/|a|^2$ as $k \to \infty$;
- if $|a| < 1$ then $a^{2k} c_k^{-1} \to c_0^{-1} + \frac{\sigma^{-2}}{a^{-2}-1}$;
- if $|a| = 1$ then $k^{-1} c_k^{-1} \to \sigma^{-2}$.

Hence prove that, for all values of a and all c_0, there is $K \in \mathbb{N}$ such that $c_k < \sigma^2$ for all $k > K$ and that, furthermore, if $|a| \leq 1$, $c_k \to 0$ as $k \to \infty$.

Exercise 5-3. Consider the equation for the mean update in the Kalman filter in the setting of Exercise 5-2. Now assume that the data y_k is derived from a deterministic true signal $x_k^\dagger$ generated as follows:

$$x_{k+1}^\dagger = a x_k^\dagger,$$

$$y_k = x_k^\dagger + \eta_k.$$

Show that the error e_k between the mean m_k and the true signal $x_k^\dagger$ satisfies the equation

$$c_{k+1}^{-1}e_{k+1} = a^{-1}c_k^{-1}e_k + \sigma^{-2}\eta_{k+1}.$$

Show that, if $|a| \neq 1$, then there is $K \in \mathbb{N}$ such that

$$|e_{k+1}| \leq \lambda|e_k| + |\eta_{k+1}|, \quad k \geq K.$$

where $\lambda = |a|$ if $|a| < 1$ and $\lambda = |a|^{-1}$ if $|a| > 1$.

Using the fact that η_k is an i.i.d. sequence of mean zero Gaussians with variance σ^2, prove that

$$\limsup_{k\to\infty} \mathbb{E}|e_k| \leq \sigma(1 - |\lambda|)^{-1}.$$

Exercise 5-4. Use Exercise 5-3 to show that $d_k := a^k c_k^{-1}e_k$ satisfies

$$d_{k+1} = d_k + \sigma^{-2}a^{k+1}\eta_{k+1}$$

and hence that

$$\mathbb{E}|d_{k+1}|^2 = \mathbb{E}|d_k|^2 + \sigma^{-2}a^{2(k+1)}.$$

Deduce the following:

- if $|a| > 1$ then $\mathbb{E}|e_k|^2 \to \sigma^2(|a|^2 - 1)/|a|^2$ as $k \to \infty$;
- if $|a| \leq 1$ then $\mathbb{E}|e_k|^2 \to 0$ as $k \to \infty$.

Exercise 5-5. Is the sytem (5.4) with

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$

observable? What about the sytem (5.4) with

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}?$$

and sytem (5.4) with

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad D = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}?$$

Exercise 5-6. Consider the linear system

$$x_{k+1} = Ax_k + Bu_k, \quad x_0 = x.$$

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$. Assume that we wish to find the control which ensures $x_K = 0$ and minimizes

$$J(u) = \frac{1}{2} \sum_{k=0}^{K-1} |u_k|^2.$$

Use Lagrange mutipliers to show that the optimal solution is given by control $\{u_j\}$ and state $\{x_j\}$ satisfying

$$\begin{aligned} u_j &= B^T \lambda_j, \quad j = 0, \ldots, K-1, \\ x_{j+1} &= Ax_j + Bu_j, \quad j = 0, \ldots, K-1, \\ \lambda_{j-1} &= A^T \lambda_j, \quad j = 1, \ldots, K-1 \end{aligned}$$

and $x_0 = x, x_K = 0$.

Chapter 6

Optimal Control

This chapter uses the calculus of variations to study two problems related to optimal control in the context of unconstrained linear control problems. Ideas from section 2.9, concerning the calculus of variations, play a central role. The chapter contains two sections. Section 6.1 concerns the optimal choice of control which achieves the objective $x(T) = 0$, for some fixed $T > 0$, with minimal $L^2\big([0,T];\mathbb{R}^m\big)$ norm. Section 6.2 concerns the optimal choice of control which, for some fixed $T > 0$, minimizes the sum of the square of the $L^2\big([0,T];\mathbb{R}^m\big)$ norm of the control and the square of the $L^2\big([0,T];\mathbb{R}^p\big)$ norm of the data/model mismatch $y - Dx$; thus we attempt to find control which leads to output which matches the data, whilst simultaneously ensuring that the control is not too big. A discrete analogue of the continuous time analysis of section 6.1 may be found in Exercise 5-6 and the reader is advised to solve this problem before reading section 6.1. Section 5.4 contains a discrete analysis similar to the continuous time analysis of section 6.2. However in the discrete setting we attempted to find the optimal fit to the data by varying the *initial condition*; in 6.2 we match the data by choosing an appropriate control. Exercise 6-2 studies the problem of choosing the initial condition to best match the data.

6.1 Minimum Energy Controls

In Chapter 3 we consider the unrestricted control problem

$$\dot{x} = Ax + Bu, \quad t > 0 \tag{6.1a}$$

$$x(0) = x_0. \tag{6.1b}$$

In Lemma 3.3.3 we showed that, provided the controllability matrix $G(A, B)$ has full rank, then all initial conditions can be controlled to the origin in finite time $T > 0$. But we did not exhibit any selection mechanism for choosing the best control. In this section we seek to do so by finding the control which, for given fixed $T > 0$, drives the solution of (6.1) to satisfy $x(T) = 0$ and which minimizes

$$J(u) = \frac{1}{2}\|u\|^2 \tag{6.2}$$

where the norm is in $\mathsf{U}_m = L^2\big([0, T]; \mathbb{R}^m\big)$.

We aim to minimize J over all controls which achieve the objective $x(T) = 0$. First we view this end point condition as a constraint on u and define

$$\mathsf{U}_{\mathrm{ad}} = \{u \in \mathsf{U}_m | x(T) = 0\}.$$

Lemma 3.3.9 shows that the set U_{ad} is convex. Using this fact we have the following:

Theorem 6.1.1. *Assume that (A, B) is controllable. Fix any $T > 0$. Then there is a unique minimizer $\overline{u} \in \mathsf{U}_{\mathrm{ad}}$ of J such that $J(\overline{u}) \leq J(u)$ for all $u \in \mathsf{U}_{\mathrm{ad}}$.*

Proof. We use Theorem 2.9.1 which requires that U_{ad} is closed and convex. We know that U_{ad} is convex and hence now establish that it is closed. If $\{u_k\}$ is a sequence of controls in U_{ad} with limit u in U_m then (3.13) shows that

$$\int_0^T e^{A(T-s)} B u_k(s) ds = -e^{AT} x_0.$$

Define the linear operator $\mathcal{L} : \mathsf{U}_m \to \mathbb{R}^n$ by

$$\big(\mathcal{L}\phi\big)(s) := \int_0^T e^{A(T-s)} B \phi(s) ds.$$

The operator $\mathcal{L}$ is bounded from U_m into $\mathbb{R}^n$ since $e^{A(T-s)}B$ is bounded in $L^\infty\big([0,T];\mathbb{R}^{n\times m}\big)$. Hence we deduce that

$$\int_0^T e^{A(T-s)}Bu(s)ds = \lim_{k\to\infty}\int_0^T e^{A(T-s)}Bu_k(s)ds = -e^{AT}x_0.$$

Thus $u \in \mathsf{U}_{\mathrm{ad}}$ and so the set U_{ad} is closed as well as convex. A feasible point in U_{ad} exists since the problem is controllable, by Lemma 3.3.3. Theorem 2.9.1 gives the desired result. □

Now we would like to characterize the minimizing control $\overline{u}$, and the resulting minimizing state $\overline{x}$. To this end, we employ Theorems 2.9.3 and 2.9.4. Let $X = H^1\big([0,T];\mathbb{R}^n\big)$, $\mathsf{U}_n = L^2\big([0,T];\mathbb{R}^n\big)$ and $\mathsf{Z} = \mathsf{U}_n \times \mathbb{R}^n$. Recall $\mathcal{A} : \mathsf{X} \to \mathsf{Z}$ defined by (2.14) and that Theorem 2.5.6 shows that this operator has bounded inverse. We also define $\mathcal{B} : \mathsf{U}_m \to \mathsf{Z}$ by

$$\mathcal{B}u = \begin{pmatrix} -Bu \\ 0 \end{pmatrix}. \tag{6.3}$$

Then we have the problem of minimizing J given by (6.2) subject to the constraints

$$\mathcal{A}x + \mathcal{B}u = g := \begin{pmatrix} 0 \\ x_0 \end{pmatrix} \tag{6.4}$$

and

$$\ell(x) := x(T) = 0. \tag{6.5}$$

We define

$$\mathsf{X}_{\mathrm{ad}} = \{x \in \mathsf{X} | x(T) = 0\} \tag{6.6}$$

and

$$\mathsf{F}_{\mathrm{ad}} = \big\{(x,u) \in \mathsf{X}_{\mathrm{ad}} \times \mathsf{U}_m | \mathcal{A}x + \mathcal{B}u = g\big\}.$$

Note that we have now put the constraint $x(T)$ on the set X, rather than directly on U_m. Note also that X_{ad} is a closed and convex set in X and that U_m is convex (see Exercise 6-1). We have the following result.

Theorem 6.1.2. *Let the assumptions of Theorem 6.1.1 hold. The problem of minimizing J given by (6.2) over $(x,u) \in \mathsf{F}_{\mathrm{ad}}$ has a unique solution.*

Furthermore if control $\overline{u}$ and state $\overline{x}$ are given via solution of the boundary value problem

$$\frac{d\overline{x}}{dt} = A\overline{x} + B\overline{u}, \quad \overline{x}(0) = x_0 \tag{6.7a}$$

$$\frac{d\overline{\lambda}}{dt} = -A^T\overline{\lambda}, \quad \overline{x}(T) = 0, \tag{6.7b}$$

and the identity $\overline{u} = B^T\overline{\lambda}$ then this pair solves the variational equations (2.36) *for appropriate choice of the Lagrange multipliers $\overline{p}, \overline{\rho}$.*

Proof. A feasible point exists in F_{ad} since the problem is controllable, by Lemma 3.3.3. The existence of a unique solution follows from Theorem 2.9.3 since X_{ad} and U_m are closed and convex and since $\mathcal{A}^{-1}$ is bounded. To characterize the solution we apply Theorem 2.9.4, showing that $(\overline{x}, \overline{u})$ satisfies the equations (2.36) which we repeat here for convenience:

$$\begin{aligned}
\langle \mathcal{A}\overline{x} + \mathcal{B}\overline{u} - g, \delta p\rangle &= 0 \quad \forall \delta p \in \mathsf{Z}\\
\langle D_x J(\overline{x}, \overline{u}) + \mathcal{A}^*\overline{p} + \ell^*(\overline{\rho}), \delta x\rangle &= 0 \quad \forall \delta x \in \mathsf{X}\\
\langle D_u J(\overline{x}, \overline{u}) + \mathcal{B}^*\overline{p}, \delta u\rangle &= 0 \quad \forall \delta u \in \mathsf{U}\\
\langle \ell(\overline{x}) - f, \delta\rho\rangle &= 0 \quad \forall \delta\rho \in \mathbb{R}^r.
\end{aligned}$$

The differential equation for $\overline{x}$ follows from (2.36a). We write $\overline{p} = (\overline{\lambda}, q) \in \mathsf{Z}$ and assume that $\overline{\lambda}$ is differentiable. Recall that here $\ell(x) = x(T)$ and that $f = 0$ so that $\ell(x) = f$ imposes the constraint that creates X_{ad}. Since J is independent of x, equation (2.36b) gives

$$\begin{aligned}
0 &= \langle \mathcal{A}^*\overline{p} + \ell^*(\overline{\rho}), \delta x\rangle\\
&= \langle \overline{p}, \mathcal{A}\delta x\rangle + \langle \overline{\rho}, \ell(\delta x)\rangle\\
&= \int_0^T \Big\langle \overline{\lambda}(t), \frac{d}{dt}\delta x(t) - A\delta x(t)\Big\rangle dt + \langle q, \delta x(0)\rangle + \langle \overline{\rho}, \delta x(T)\rangle\\
&= -\int_0^T \Big\langle \frac{d}{dt}\overline{\lambda}(t) + A^T\overline{\lambda}(t), \delta x(t)\Big\rangle dt + \langle \overline{\rho} + \overline{\lambda}(T), \delta x(T)\rangle\\
&\qquad + \langle q - \overline{\lambda}(0), \delta x(0)\rangle.
\end{aligned}$$

From this we deduce that if the Lagrange multipliers are related by $q = \overline{\lambda}(0)$ and $\overline{\rho} = -\overline{\lambda}(T)$ then we obtain the following weak form of the equation for $\overline{\lambda}(t)$:

$$\int_0^T \Big\langle \frac{d}{dt}\overline{\lambda}(t) + A^T\overline{\lambda}(t), \delta x(t)\Big\rangle dt = 0 \quad \forall \delta x \in X.$$

This is satisfied if $\overline{\lambda}$ solves the strong form of the equation given in the theorem statement.

Note that

$$\begin{aligned}\langle \mathcal{B}^* \overline{p}, \delta u\rangle &= \langle \overline{p}, \mathcal{B}\delta u\rangle \\ &= -\langle \overline{\lambda}, B\delta u\rangle \\ &= -\langle B^T \overline{\lambda}, \delta u\rangle.\end{aligned}$$

Thus

$$\langle \mathcal{B}^* \overline{p}, \delta u\rangle = -\langle B^T \overline{\lambda}, \delta u\rangle. \tag{6.8}$$

Hence equation (2.36c) is equivalent to

$$\langle \overline{u} - B^T \overline{\lambda}, \delta u\rangle = 0, \quad \forall \delta u \in U_m$$

which is satisfied under the assumptions of the theorem. Finally note that (2.36d) is satisfied when we impose the constraint that $\overline{x}(T) = 0$. □

Using equations (6.7) we can characterize the solution of this optimal control problem rather explicitly. We define

$$Q(T) = \int_0^T e^{As} B B^T e^{A^T s} ds$$

and note Corollary 3.3.4 shows that this matrix is invertible.

Theorem 6.1.3. *Under the assumptions of Theorem 6.1.1, the optimal control $\overline{u}$ is given by*

$$\overline{u}(t) = -B^T e^{A^T(T-t)} Q(T)^{-1} e^{AT} x_0.$$

Proof. We note, from (6.7b) that $\overline{\lambda}(t) = e^{-A^T t}\overline{\lambda}(0)$ so that $\overline{u}(t) = B^T e^{-A^T t}\overline{\lambda}(0)$. Thus the optimal state solves

$$\frac{d\overline{x}}{dt} = A\overline{x} + BB^T\big(e^{-A^T t}\overline{\lambda}(0)\big), \quad \overline{x}(0) = x_0.$$

Hence

$$\overline{x}(t) = e^{At} x_0 + \Big(\int_0^t e^{A(t-s)} BB^T e^{-A^T s} ds\Big)\overline{\lambda}(0).$$

Since $x(T) = 0$ we obtain

$$\begin{aligned}
\overline{\lambda}(0) &= -\Big(\int_0^T e^{-As} BB^T e^{-A^T s}\, ds\Big)^{-1} x_0 \\
&= -\Big(e^{-AT} \int_0^T e^{A(T-s)} BB^T e^{A^T(T-s)} ds\, e^{-A^T T}\Big)^{-1} x_0 \\
&= -\Big(e^{-AT} \int_0^T e^{As} BB^T e^{A^T s} ds\, e^{-A^T T}\Big)^{-1} x_0 \\
&= -e^{A^T T} Q(T)^{-1} e^{AT} x_0.
\end{aligned}$$

Combining with the expression

$$\overline{u}(t) = B^T e^{-A^T t} \overline{\lambda}(0)$$

gives the desired expression for the optimal control.

It remains to show that it is indeed optimal. To this end we note that any control $u \in \mathsf{U}_{\text{ad}}$ satisfies

$$0 = e^{AT} x_0 + \int_0^T e^{A(T-t)} Bu(t) dt.$$

Using this identity we find that

$$\begin{aligned}
\int_0^T \langle u(t), \overline{u}(t) \rangle dt &= -\int_0^T \Big\langle u(t), B^T e^{A^T(T-t)} Q(T)^{-1} e^{AT} x_0 \Big\rangle dt \\
&= -\int_0^T \Big\langle e^{A(T-t)} Bu(t), Q(T)^{-1} e^{AT} x_0 \Big\rangle dt \\
&= -\Big\langle \int_0^T e^{A(T-t)} Bu(t) dt, Q(T)^{-1} e^{AT} x_0 \Big\rangle \\
&= \Big\langle e^{AT} x_0, Q(T)^{-1} e^{AT} x_0 \Big\rangle.
\end{aligned}$$

This expression is independent of the specific choice of control $u \in \mathsf{U}_{\text{ad}}$ and so we deduce that, in particular,

$$\int_0^T \langle \overline{u}(t), \overline{u}(t) \rangle dt = \Big\langle e^{AT} x_0, Q(T)^{-1} e^{AT} x_0 \Big\rangle.$$

and, subtracting, we obtain

$$\int_0^T \langle u(t) - \overline{u}(t), \overline{u}(t) \rangle dt = 0.$$

From this it follows that

$$\begin{aligned}J(u) &= J(\overline{u} + u - \overline{u})\\ &= J(\overline{u}) + J(u - \overline{u})\\ &\geq J(\overline{u}).\end{aligned}$$

This establishes the desired optimality.

□

Remarks 6.1.4. *It is a consequence of Theorem 6.1.2 that the function $\overline{u}$ given in the previous theorem does indeed control the system* (3.9) *to the origin at time T. However it is instructive to establish this explicitly. Note that*

$$\begin{aligned}\overline{x}(T) &= e^{AT}x_0 + \int_0^T e^{A(T-s)}B\overline{u}(s)ds\\ &= e^{AT}x_0 - \Big(\int_0^T e^{A(T-s)}BB^Te^{A^T(T-s)}ds\Big)Q(T)^{-1}e^{AT}x_0\\ &= e^{AT}x_0 - \Big(\int_0^T e^{As}BB^Te^{A^Ts}ds\Big)Q(T)^{-1}e^{AT}x_0\\ &= e^{AT}x_0 - e^{AT}x_0\\ &= 0.\end{aligned}$$

6.2 Matching Data

Consider the system

$$\dot{x} = Ax + Bu, \quad x(0) = x_0 \tag{6.9a}$$
$$y = Dx + \eta \tag{6.9b}$$

where $x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m, y(t) \in \mathbb{R}^p$, $\eta(t) \in \mathbb{R}^p$ and $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$ and $D \in \mathbb{R}^{p\times n}$. We no longer impose the control objective that $x(T) = 0$.

The data y is a noisy observation of Dx and η is an unknown noise. We aim to choose the control u so that the solution x to (6.9) best matches the observed data $y \in \mathsf{U}_p = L^2\big([0,T];\mathbb{R}^p\big)$ and, as in the previous section, is not large in the L^2 sense. Specifically we minimize, for some $\alpha > 0$,

$$J(x,u) = \frac{\alpha}{2}\|y - Dx\|^2 + \frac{1}{2}\|u\|^2, \tag{6.10}$$

where the norms are in U_p and U_m respectively, subject to the constraint that (6.1) holds.

Now we would like to characterize the minimizing control $\overline{u}$, and the resulting minimizing state $\overline{x}$. To this end, we employ Theorems 2.9.3 and 2.9.4. We again set $X = H^1\big([0,T];\mathbb{R}^n\big)$, $\mathsf{U}_n = L^2\big([0,T];\mathbb{R}^n\big)$ and $\mathsf{Z} = \mathsf{U}_n \times \mathbb{R}^n$. Recall $\mathcal{A} : \mathsf{X} \to \mathsf{Z}$ defined by (2.14) and that Theorem 2.5.6 shows that this operator has bounded inverse. We also define $\mathcal{B}$ as in (6.3). Then we have the problem of minimizing J given by (6.10) over

$$\mathsf{F}_{\mathrm{ad}} = \big\{(x,u) \in \mathsf{X} \times \mathsf{U}_m | \mathcal{A}x + \mathcal{B}u = g\big\}.$$

Recall the optimality equations from Theorem 2.9.6:

$$\begin{aligned} \langle \mathcal{A}\overline{x} + \mathcal{B}\overline{u} - g, \delta p\rangle &= 0 \quad \forall \delta p \in \mathsf{Z} && (6.11\mathrm{a})\\ \langle D_x J(\overline{x},\overline{u}) + \mathcal{A}^*\overline{p}, \delta x\rangle &= 0 \quad \forall \delta x \in \mathsf{X} && (6.11\mathrm{b})\\ \langle D_u J(\overline{x},\overline{u}) + \mathcal{B}^*\overline{p}, \delta u\rangle &= 0 \quad \forall \delta u \in \mathsf{U}. && (6.11\mathrm{c}) \end{aligned}$$

We have the following result.

Theorem 6.2.1. *The problem of minimizing J given by* (6.10) *over* F_{ad} *has a unique solution. Furthermore if control* $\overline{u}$ *and state* $\overline{x}$ *are given via solution of the boundary value problem*

$$\begin{aligned} \frac{d\overline{x}}{dt} &= A\overline{x} + B\overline{u}, \quad \overline{x}(0) = x_0\\ \frac{d\overline{\lambda}}{dt} &= -A^T\overline{\lambda} + \alpha D^T\big(D\overline{x} - y), \quad \overline{\lambda}(T) = 0, \end{aligned}$$

and the identity $\overline{u} = B^T\overline{\lambda}$ *then this pair solves the variational equations* (6.11) *for appropriate choice of the Lagrange multiplier* $\overline{p}$.

Proof. A feasible point is identified by taking $u = 0$ and $x(t) = e^{At}x_0$. The existence of a unique solution then follows from Theorem 2.9.3, since $\mathcal{A}^{-1}$ is bounded and X and U_m are convex (see Exercise 6-1) and of course closed by the properties of Hilbert space. To charaterize the solution we apply Theorem 2.9.4, and equations (6.11). The differential equation for $\overline{x}$ follows from (6.11a). We write $\overline{p} = (\overline{\lambda}, q) \in \mathsf{Z}$.

To write down (6.11b) we note that

$$\begin{aligned}\langle \mathcal{A}^*\overline{p}, \delta x\rangle &= \langle \overline{p}, \mathcal{A}\delta x\rangle \\ &= \int_0^T \Big\langle \overline{\lambda}(t), \frac{d}{dt}\delta x(t) - A\delta x(t)\Big\rangle dt + \langle q, \delta x(0)\rangle \\ &= -\int_0^T \Big\langle \frac{d}{dt}\overline{\lambda}(t) + A^T\overline{\lambda}(t), \delta x(t)\Big\rangle dt + \langle \overline{\lambda}(T), \delta x(T)\rangle \\ &\qquad + \langle q - \overline{\lambda}(0), \delta x(0)\rangle.\end{aligned}$$

Furthermore $D_x J(x,u) = \alpha D^T\big(Dx - y\big)$. Thus equation (6.11b) gives

$$\begin{aligned}\int_0^T \Big\langle \frac{d}{dt}\overline{\lambda}(t) + A^T\overline{\lambda}(t) - \alpha D^T\big(D\overline{x}(t) - y(t)\big), \delta x(t)\Big\rangle dt \\ - \langle \overline{\lambda}(T), \delta x(T)\rangle - \langle q - \overline{\lambda}(0), \delta x(0)\rangle = 0.\end{aligned}$$

From this we deduce that if the Lagrange multipliers are related by $q = \overline{\lambda}(0)$ then we obtain the following weak form of the equation for $\overline{\lambda}(t)$:

$$\int_0^T \Big\langle \frac{d}{dt}\overline{\lambda}(t) + A^T\overline{\lambda}(t) - \alpha D^T\big(D\overline{x}(t) - y(t)\big), \delta x(t)\Big\rangle dt - \langle \overline{\lambda}(T), \delta x(T)\rangle = 0 \quad \forall \delta x \in X.$$

This is satisfied by the solution of the strong form of the equation given in the theorem statement. Finally we note that (6.8) shows that equation (6.11c) gives us

$$\langle \overline{u} - B^T\overline{\lambda}, \delta u\rangle = 0, \quad \forall \delta u \in U;$$

this equation is satisfied if $\overline{u} = B^T\overline{\lambda}$. □

Note that we have not demonstrated that the solution exhibited does indeed minimize J subject to the desired constraint. We have simply shown that it satisfies the necessary variational equations given by Theorem 2.9.4.

6.3 Exercises

Exercise 6-1. Show that X_{ad} given by (6.6) is a closed convex subset of X. Show that X and U_m are convex.

Exercise 6-2. Consider the linear dynamical system in $\mathbb{R}^n$ given by

$$\frac{dx}{dt} = Ax \tag{6.12}$$

for some fixed matrix $A \in \mathbb{R}^{n\times n}$. Assume that we only have a noisy observation m_0 of the initial condition $x(0) = x_0$ so that we do not know it exactly. To compensate we are given data $y \in L^2\big([0,T];\mathbb{R}^p\big)$ which is supposed to be a noisy measurement of Dx for some $D \in \mathbb{R}^{p\times n}$. Define $I : \mathbb{R}^n \to \mathbb{R}^+$ given by

$$I(x_0) = \frac{\alpha}{2}\|y - De^{A\cdot}x_0\|^2 + \frac{1}{2}|x_0 - m_0|^2$$

and note that minimizing $I(x_0)$ over all x_0 determines an initial condition which attempts to keep the solution x close to the measurement y and keeps the initial condition close to the observed initial condition m_0. By differentiating I with respect to x_0 find this optimal initial condition.

Exercise 6-3. Consider the setting of Exercise 6-2. Use Lagrange mutipliers to find the initial condition and function x which minimizes

$$J(x_0, x) = \frac{\alpha}{2}\|y - Dx\|^2 + \frac{1}{2}|x_0 - m_0|^2,$$

for $\alpha > 0$, subject to the constraint that (6.12) holds. You may assume that any terms appearing in the resulting variational equations have sufficient regularity to enable an integration by parts. Demonstrate that the result agrees with the result obtained in the preceding exercise. Explain why this is so.

Exercise 6-4. Consider the linear control problem (6.1). Use Lagrange mutipliers to find the control u which minimizes

$$J(x_0, x, u) = \frac{\alpha}{2}|x(T)|^2 + \frac{1}{2}\|u\|^2$$

for $\alpha > 0$, subject to the constraint that (6.1) relates x and u. You may assume that any terms appearing in the resulting variational equations have sufficient regularity to enable an integration by parts.

Chapter 7

References

For a concise mathematical introduction to control theory see [7]. This book covers all of the topics in our course, with the exception of the Kalman filter. It is the best single reference for this course, apart from these notes. A more geometric viewpoint on the subject is contained in the book [5]. A nice engineering treatment of the subject may be found in [1], but beware that some of the proofs in this book are incomplete. The subject of optimal control is covered in [4] and [6], and in the on-line lecture notes [2]. The Kalman filter is studied in [3]; the Wikipedia entry is also a good starting point for this topic:

http : //en.wikipedia.org/wiki/Kalman_filter

Acknowledgements The authors are grateful to Sergios Agapiou, Kui Lin, Daniel Sanz, Abhishek Shukla and the previous students from this course for feedback which has improved the notes.

Bibliography

[1] S. Barnett. *Introduction to mathematical control theory.* Oxford, Clarendon Press, 1975.

[2] L.C. Evans. *An introduction to mathematical optimal control theory.* 2005. `http://math.berkeley.edu/ evans/control.course.pdf`.

[3] A.C. Harvey. *Forecasting, structural time series models and the Kalman filter.* Cambridge Univ Pr, 1990.

[4] J. Macki and A. Strauss. *Introduction to optimal control theory.* Springer, 1982.

[5] E.D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Verlag, 1998.

[6] J.L. Speyer and D.H. Jacobson. *Primer on optimal control theory*, volume 20. Society for Industrial & Applied, 2010.

[7] J. Zabczyk. *Mathematical control theory: an introduction.* Springer, 1992.