Lecture Notes on Linear System Theory

John Lygeros* and Federico A. Ramponi†

*Automatic Control Laboratory, ETH Zurich
CH-8092, Zurich, Switzerland
lygeros@control.ee.ethz.ch

†Department of Information Engineering, University of Brescia
Via Branze 38, 25123, Brescia, Italy
federico.ramponi@unibs.it

January 3, 2015

Contents

Chapter 1

Introduction

1.1 Objectives of the course

This course has two main objectives. The first (and more obvious) is for students to learn something about linear systems. Most of the course will be devoted to linear time varying systems that evolve in continuous time $t \in \mathbb{R}_+$. These are dynamical systems whose evolution is defined through state space equations of the form

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),$$
$$y(t) = C(t)x(t) + D(t)u(t),$$

where $x(t) \in \mathbb{R}^n$ denotes the system state, $u(t) \in \mathbb{R}^m$ denotes the system inputs, $y(t) \in \mathbb{R}^p$ denotes the system outputs, $A(t) \in \mathbb{R}^{n\times n}$, $B(t) \in \mathbb{R}^{n\times m}$, $C(t) \in \mathbb{R}^{p\times n}$, and $D(t) \in \mathbb{R}^{p\times m}$ are matrices of appropriate dimensions, and where, as usual, $\dot{x}(t) = \frac{dx}{dt}(t)$ denotes the derivative of $x(t)$ with respect to time.

Time varying linear systems are useful in many application areas. They frequently arise as models of mechanical or electrical systems whose parameters (for example, the stiffness of a spring or the inductance of a coil) change in time. As we will see, time varying linear systems also arise when one linearizes a non-linear system around a trajectory. This is very common in practice. Faced with a nonlinear system one often uses the full nonlinear dynamics to design an optimal trajectory to guide the system from its initial state to a desired final state. However, ensuring that the system will actually track this trajectory in the presence of disturbances is not an easy task. One solution is to linearize the nonlinear system (i.e. approximate it by a linear system) around the optimal trajectory; the approximation is accurate as long as the nonlinear system does not drift too far away from the optimal trajectory. The result of the linearization is a time varying linear system, which can be controlled using the methods developed in this course. If the control design is done well, the state of the nonlinear system will always stay close to the optimal trajectory, hence ensuring that the linear approximation remains valid.

A special class of linear time varying systems are linear time invariant systems, usually referred to by the acronym LTI. LTI systems are described by state equations of the form

$$\dot{x}(t) = Ax(t) + Bu(t),$$
$$y(t) = Cx(t) + Du(t),$$

where the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ are constant for all times $t \in \mathbb{R}_+$. LTI systems are somewhat easier to deal with and will be treated in the course as a special case of the more general linear time varying systems.

The second and less obvious objective of the course is for students to experience something about doing automatic control research, in particular developing mathematical proofs and formal logical arguments. Linear systems are ideally suited for this task. There are two main reasons for this. The first is that almost all the derivations given in the class can be carried out in complete detail, down to the level of basic algebra. There are very few places where one has to invoke "higher powers", such as an obscure mathematical theorem whose proof is outside the scope of the course. One can generally continue the calculations until he/she is convinced that the claim is true. The second reason is that linear systems theory brings together two areas of mathematics, algebra and analysis. As we will soon see, the state space, $\mathbb{R}^n$, of the systems has both an algebraic structure (it is a vector space) and a topological structure (it is a normed space). The algebraic structure allows us to perform linear algebra operations, compute projections, eigenvalues, etc. The topological structure, on the other hand, forms the basis of analysis, the definition of derivatives, etc. The main point of linear systems theory is to exploit the algebraic structure to develop tractable "algorithms" that allow us to answer analysis questions which appear intractable by themselves.

For example, consider the time invariant linear system

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1.1}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times n}$, and $B \in \mathbb{R}^{n\times m}$. Given $x_0 \in \mathbb{R}^n$, $T > 0$ and a continuous function $u(\cdot) : [0, T] \to \mathbb{R}^m$ (known as the input trajectory) one can show (Chapter 3) that there exists a unique function $x(\cdot) : [0, T] \to \mathbb{R}^n$ such that

$$x(0) = x_0 \text{ and } \dot{x}(t) = Ax(t) + Bu(t), \quad \text{for all } t \in [0, T]. \tag{1.2}$$

This function is called the state trajectory (or simply the solution) of system (1.1) with initial condition x_0 under the input $u(\cdot)$. As we will see in Chapter 3, $u(\cdot)$ does not even need to be continuous for (1.2) to be true, provided one appropriately qualifies the statement "for all $t \in [0, T]$".

System (1.1) is called controllable (Chapter 8) if and only if for all $x_0 \in \mathbb{R}^n$, for all $\hat{x} \in \mathbb{R}^n$, and for all $T > 0$, there exists $u(\cdot) : [0, T] \to \mathbb{R}^m$ such that the solution of system (1.1) with initial condition x_0 under the input $u(\cdot)$ is such that $x(T) = \hat{x}$. Controllability is clearly an interesting property for a system to have. If the system is controllable then we can guide it from any initial state to any final state by selecting an appropriate input. If not, there may be some desirable parts of the state space that we cannot reach from some initial states. Unfortunately, determining whether a system is controllable directly from the definition is impossible. This would require calculating all trajectories that start at all initial conditions. Except for trivial cases (like the linear system $\dot{x}(t) = u(t)$) this calculation is intractable, since the initial states, x_0, the times T of interest, and the possible input trajectories $u(\cdot) : [0, T] \to \mathbb{R}^m$ are all infinite. Fortunately, linear algebra can be used to answer the question without even computing a single solution (Chapter 8).

Theorem 1.1 *System (1.1) is controllable if and only if the matrix* $\begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n\times nm}$ *has rank* n.

The theorem shows how the seemingly intractable analysis question "is the system (1.1) controllable?" can be answered by a simple algebraic calculation of the rank of a matrix.

The treatment in these notes is inspired by [6] in terms of the level of mathematical rigour and at places the notation and conventions. Coverage and style of presentation of course differ substantially. There are many good reference books for linear systems theory, including [5, 1, 2, 9] and, primarily for linear time invariant systems, [11].

1.2 Proof methods

Most of the course will be devoted to proving theorems. The proof methods that we will encounter are just a set of tools, grounded in mathematical logic and widely accepted in the mathematical

community, that let us say that a proposition is true, *given* that others are true. A "Theorem" is indeed a logical statement that can be proven: This means that the truth of such statement can be established by applying our proof methods to other statements that we already accept as true, either because they have been proven before, or because we postulate so (for example the "axioms" of logic), or because we *assume* so in a certain context (for example, when we say "Let V be a vector space ..." we mean "Assume that the set V verifies the axioms of a vector space ...").
Theorems of minor importance, or theorems whose main point is to establish an intermediate step in the proof of another theorem, will be called "Lemmas", "Facts", or "Propositions"; An immediate consequence of a theorem that deserves to be highlighted separately is usually called a "Corollary". And a logical statement that we think may be true but cannot prove so is called a "Conjecture".

The logical statements we will most be interested in typically take the form

$$p \Rightarrow q$$

(p implies q). p is called the hypothesis and q the consequence.

Example (No smoke without fire) It is generally accepted that when there is smoke, there must be some a fire somewhere. This knowledge can be encoded by the logical implication

$$\begin{array}{ccccc} \text{If} & \text{there is smoke} & \text{then} & \text{there is a fire} \\ & p & \Rightarrow & q. \end{array}$$

This is a statement of the form $p \Rightarrow q$ with p the statement "there is smoke" and q the statement "there is a fire". ■

Hypotheses and consequences may typically depend on one or more free variables, that is, objects that in the formulation of hypotheses and consequences are left free to change.

Example (Greeks) Despite recent economic turbulence, it is generally accepted that Greek citizens are also Europeans. This knowledge can be encoded by the logical implication

$$\begin{array}{ccccc} \text{If} & X \text{ is a Greek} & \text{then} & X \text{ is a European} \\ & p(X) & \Rightarrow & q(X). \end{array}$$

A sentence like "X is a ..." is the verbal way of saying something belongs to a *set*; for example the above statement can also be written as

$$X \in \text{Greeks} \Rightarrow X \in \text{Europeans},$$

where "Greeks" and "Europeans" are supposed to be sets; the assertion that this implication is true for arbitrary X ($\forall X,\ X \in \text{Greeks} \Rightarrow X \in \text{Europeans}$) is equivalent to the set-theoretic statement of inclusion:

$$\text{Greeks} \subseteq \text{Europeans}.$$

You can visualize the implication and its set-theoretic interpretation in Figure 1.1. ■

There are several ways of proving that logical statements are true. The most obvious one is a direct proof: Start from p and establish a finite sequence of intermediate implications, p_1, p_2, ..., p_n leading to q

$$p \Rightarrow p_1 \Rightarrow p_2 \Rightarrow \ldots \Rightarrow p_n \Rightarrow q.$$

We illustrate this proof technique using a statement about the natural numbers.

Definition 1.1 *A natural number $n \in \mathbb{N}$ is called <u>odd</u> if and only if there exists $k \in \mathbb{N}$ such that $n = 2k + 1$. It is called <u>even</u> if and only if there exists $k \in \mathbb{N}$ such that $n = 2k$.*

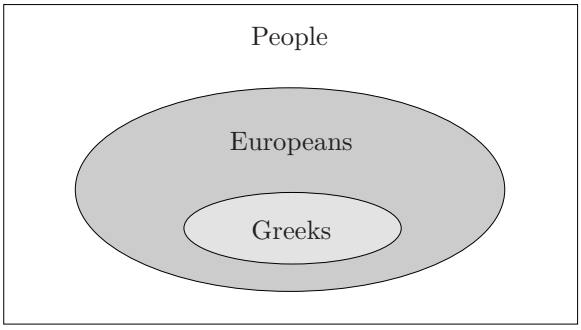

Figure 1.1: Set theoretic interpretation of logical implication.

One can indeed show that all natural numbers are either even or odd, and no natural number is both even and odd (Problem 1.1).

Theorem 1.2 *If n is odd then n^2 is odd.*

Proof:

$$\begin{aligned} n \text{ is odd} &\Leftrightarrow \exists k \in \mathbb{N} \ : \ n = 2k+1 \\ &\Rightarrow \exists k \in \mathbb{N} \ : \ n^2 = (2k+1)(2k+1) \\ &\Rightarrow \exists k \in \mathbb{N} \ : \ n^2 = 4k^2 + 4k + 1 \\ &\Rightarrow \exists k \in \mathbb{N} \ : \ n^2 = 2(2k^2 + 2k) + 1 \\ &\Rightarrow \exists l \in \mathbb{N} \ : \ n^2 = 2l + 1 \ \ (\text{namely, } l = 2k^2 + 2k \in \mathbb{N}) \\ &\Rightarrow n^2 \text{ is odd} \end{aligned}$$

■

This proof principle can also be exploited to perform *proof by induction*. Proof by induction concerns propositions, p_k, indexed by the natural numbers, $k \in \mathbb{N}$, and statements of the form

$$\forall k \in \mathbb{N}, \ \ p_k \text{ is true.}$$

One often proves such statements by showing that p_0 is true and then establishing and infinite sequence of implications

$$p_0 \Rightarrow p_1 \Rightarrow p_2 \Rightarrow \dots .$$

Clearly proving these implications one by one is impractical. It suffices, however, to establish that $p_k \Rightarrow p_{k+1}$ for all $k \in \mathbb{N}$, or in other words

$$[p_0 \wedge (p_k \Rightarrow p_{k+1}, \ \ \forall k \in \mathbb{N})] \Rightarrow [p_k, \ \ \forall k \in \mathbb{N}] \, .$$

We demonstrate this proof style with another statement about the natural numbers.

Definition 1.2 *The factorial, $n!$, of a natural number $n \in \mathbb{N}$ is the natural number $n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1$. By convention, if $n = 0$ we set $n! = 1$.*

Theorem 1.3 *For all $m, k \in \mathbb{N}$, $(m+k)! \geq m!k!$.*

Proof: It is easy to check that the statement holds for the special cases $m = k = 0$, $m = 0$ and $k = 1$, and $m = 1$ and $k = 0$. For the case $m = k = 1$, $(m+k)! = 2! \geq 1!1! = m!k!$.

Assume now that for some $m, k \in \mathbb{N}$, $(m+k)! \geq m!k!$ (we call this the induction hypothesis) and show that for $m, k+1$, $(m+k+1)! \geq m!(k+1)!$ (also $(m+1)!k!$ for the case $m+1, k$, by symmetry).

$$\begin{aligned}(m+k+1)! &= (m+k)!(m+k+1)\\ &\geq m!k!(m+k+1) \text{ (by the induction hypothesis)}\\ &\geq m!k!(k+1) \text{ (since } m \in \mathbb{N})\\ &= m!(k+1)!\end{aligned}$$

which completes the proof. ■

Even though there is no direct way to illustrate proof by induction using statements about Greeks and other nationalities, one could in principle use a similar line of reasoning to prove a statement like "once a Greek always a Greek" by arguing that children with at least one Greek parent are themselves Greek.

Sometimes direct proof $p \Rightarrow q$ is difficult. In this case we try to find other statements that are logically equivalent to $p \Rightarrow q$ and prove these instead. An example of such a statement is $\neg q \Rightarrow \neg p$, or in logic notation

$$(\neg q \Rightarrow \neg p) \Leftrightarrow (p \Rightarrow q).$$

Example (Greeks (cont.)) The statement that all the Greeks are Europeans is also equivalent to

$$\begin{array}{ccccc}\text{If} & X \text{ is not a European} & \text{then} & X \text{ is not a Greek}\\ & \neg q(X) & \Rightarrow & \neg p(X).\end{array}$$

In turn, this is equivalent to the set theoretic statement

$$\text{non-Europeans} \quad \subseteq \quad \text{non-Greeks}.$$

If we stipulate a priori that all the possible X we may consider in our discourse belong to some "big" set (i.e., "People"), in fact this is also equivalent to

$$\text{People} \setminus \text{Europeans} = \text{Europeans}^c \quad \subseteq \quad \text{Greeks}^c = \text{People} \setminus \text{Greeks}.$$

where $\setminus$ denotes the difference of two sets and the superscript c denotes the set-complement with respect to "People".

Exercise 1.1 Visualize this set theoretic interpretation by a picture similar to Figure 1.1.

■

A proof where we show that $p \Rightarrow q$ is true by showing $\neg q \Rightarrow \neg p$ is true is known as a *proof by contraposition*. We illustrate this proof technique by another statement about the natural numbers.

Theorem 1.4 *If n^2 is odd then n is odd.*

Proof: Let $p =$ "n^2 is odd", $q =$ "n is odd". Assume n is even ($\neg q$) and show that n^2 is even ($\neg p$).

$$\begin{aligned}n \text{ is even} &\Leftrightarrow \exists k \in \mathbb{N} \ : \ n = 2k\\ &\Rightarrow \exists k \in \mathbb{N} \ : \ n^2 = 4k^2\\ &\Rightarrow \exists l \in \mathbb{N} \ : \ n = 2l \ \ (\text{namely, } l = 2k^2 \in \mathbb{N})\\ &\Rightarrow n^2 \text{ is even.}\end{aligned}$$

■

Another common method that can be used to indirectly prove that $p \Rightarrow q$ is to suppose that p is true, to suppose that q is false, and to apply other proof methods to derive a *contradiction*. A contradiction is a proposition of the form $r \wedge \neg r$ (like "There is smoke and there is no smoke", or "n is even and n is odd"); all such statements are postulated to be *false* by virtue of their mere structure, and irrespective of the proposition r. If, by assuming p is true and q is false we are able to reach a false assertion, we must admit that if p is true the consequence q cannot be false, in other words that p implies q. This method is known as *proof by contradiction*.

Example (Greeks and Chinese) Suppose the following implications: for all X,

$$\begin{aligned} X \text{ is a Greek} &\Rightarrow X \text{ is a European} \\ X \text{ is a Chinese} &\Rightarrow X \text{ is an Asian} \\ X \text{ is an Asian} &\Rightarrow X \text{ is not a European} \end{aligned}$$

We show by contradiction that every Greek is not a Chinese, more formally

$$\begin{array}{ccccc} \text{If} & X \text{ is a Greek} & \text{then} & X \text{ is not a Chinese} \\ & p(X) & \Rightarrow & q(X) \end{array}$$

Indeed, suppose $p(X)$ and *the converse* of $q(X)$, that is, X is a Chinese. By direct deduction,

$$\begin{array}{c} X \text{ is a Greek} \wedge X \text{ is a Chinese} \\ \Downarrow \\ X \text{ is a European} \wedge X \text{ is an Asian} \\ \Downarrow \\ X \text{ is a European} \wedge X \text{ is not a European} \end{array}$$

Since the conclusion is a contradiction for all X, we must admit that $p(X) \wedge \neg q(X)$ is false or, which is the same, that $p(X) \Rightarrow q(X)$. The set-theoretic interpretation is as follows: By postulate,

$$\text{Europeans} \ \cap \ \text{non-Europeans} \ = \ \emptyset$$

On the other hand, by deduction,

$$(\text{Greeks} \ \cap \ \text{Chinese}) \ \subseteq \ (\text{Europeans} \ \cap \ \text{non-Europeans})$$

It follows that Greeks $\cap$ Chinese is also equal to the empty set. *Therefore* (here is the point of the above proof), Greeks $\subseteq$ non-Chinese.

Exercise 1.2 Visualize this set theoretic interpretation by a picture similar to Figure 1.1.

■

We will illustrate this fundamental proof technique with another statement, about rational numbers.

Definition 1.3 *The real number $x \in \mathbb{R}$ is called rational if and only if there exist integers $n, m \in \mathbb{Z}$ with $m \neq 0$ such that $x = n/m$.*

Theorem 1.5 *(Pythagoras) $\sqrt{2}$ is not rational.*

Proof: (Euclid) Assume, for the sake of contradiction, that $\sqrt{2}$ is rational. Then there exist $n, m \in \mathbb{Z}$ with $m \neq 0$ such that $\sqrt{2} = n/m$. Since $\sqrt{2} > 0$, without loss of generality we can take $n, m \in \mathbb{N}$; if they happen to be both negative multiply both by -1 and replace them by the resulting numbers.

Without loss of generality, we can further assume that m and n have no common divisor; if they do, divide both by their common divisors until there are no common divisors left and replace m and n by the resulting numbers. Now

$$\begin{aligned}\sqrt{2} = \frac{n}{m} \Rightarrow 2 = \frac{n^2}{m^2} \\ \Rightarrow n^2 = 2m^2 \\ \Rightarrow n^2 \text{ is even} \\ \Rightarrow n \text{ is even (Theorem 1.4 and Problem 1.1)} \\ \Rightarrow \exists k \in N \ : \ n = 2k \\ \Rightarrow \exists k \in N \ : \ 2m^2 = n^2 = 4k^2 \\ \Rightarrow \exists k \in N \ : \ m^2 = 2k^2 \\ \Rightarrow m^2 \text{ is even} \\ \Rightarrow m \text{ is even (Theorem 1.4 and Problem 1.1).}\end{aligned}$$

Therefore, n and m are both even and, according to Definition 1.1, 2 divides both. This contradicts the fact that n and m have no common divisor. Therefore $\sqrt{2}$ cannot be rational. ■

Exercise 1.3 What is the statement p in Theorem 1.5? What is the statement q? What is the statement r in the logical contradiction $r \wedge \neg r$ reached at the end of the proof?

Two statements are equivalent if one implies the other and vice versa,

$$(p \Leftrightarrow q) \text{ is the same as } (p \Rightarrow q) \wedge (q \Rightarrow p)$$

Usually showing that two statements are equivalent is done in two steps: Show that $p \Rightarrow q$ and then show that $q \Rightarrow p$. For example, consider the following statement about the natural numbers.

Theorem 1.6 *n^2 is odd if and only if n is odd.*

Proof: n is odd implies that n^2 is odd (by Theorem 1.2) and n^2 is odd implies that n is odd (by Theorem 1.4). Therefore the two statements are equivalent. ■

This is argument is related to the canonical way of proving that two sets are equal, by proving two set inclusions $A \subseteq B$ and $B \subseteq A$. To prove these inclusions one proves two implications:

$$\begin{aligned}X \in A \Rightarrow X \in B \\ X \in B \Rightarrow X \in A\end{aligned}$$

or, in other words, $X \in A \Leftrightarrow X \in B$.

Finally, let us close this brief discussion on proof techniques with a subtle caveat: If p is a false statement then any implication of the form $p \Rightarrow q$ is true, irrespective of what q is.

Example (Maximal natural number) Here is a proof that there is no number larger than 1.

Theorem 1.7 *Let $N \in \mathbb{N}$ be the largest natural number. Then $N = 1$.*

Proof: Assume, for the sake of contradiction, that $N > 1$. Then N^2 is also a natural number and $N^2 > N$. This contradicts the fact that N is the largest natural number. Therefore we must have $N = 1$. ■

Obviously the "theorem" in this example is saying something quite silly. The problem, however, is not that the proof is incorrect, but that the starting hypothesis "let N be the largest natural number" is false, since there is no largest natural number. ■

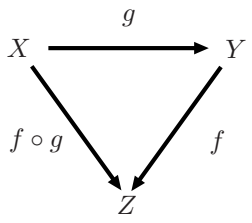

Figure 1.2: Commutative diagram of function composition.

1.3 Functions and maps

A function

$$f : X \to Y$$

maps the set X (known as the domain of f) into the set Y (known as the co-domain of f). This means that for all $x \in X$ there exists a unique $y \in Y$ such that $f(x) = y$. The element $f(x) \in Y$ is known as the value of f at x. The set

$$\{y \in Y \mid \exists x \in X \ : \ f(x) = y\} \subseteq Y$$

is called the range of f (sometimes denoted by $f(X)$) and the set

$$\{(x, y) \in X \times Y \mid y = f(x)\} \subseteq X \times Y$$

is called the graph of f.

Definition 1.4 *A function $f : X \to Y$ is called:*

1. *Injective (or one-to-one) if and only if $f(x_1) = f(x_2)$ implies that $x_1 = x_2$.*
2. *Surjective (or onto) if and only if for all $y \in Y$ there exists $x \in X$ such that $y = f(x)$.*
3. *Bijective if and only if it is both injective and surjective, i.e. for all $y \in Y$ there exists a unique $x \in X$ such that $y = f(x)$.*

Given two functions $g : X \to Y$ and $f : Y \to Z$ their composition is the function $(f \circ g) : X \to Z$ defined by

$$(f \circ g)(x) = f(g(x)).$$

Commutative diagrams help visualize function compositions (Figure 1.2).

Exercise 1.4 Show that composition is associative. In other words, for any three functions $g : X \to Y$, $f : Y \to Z$ and $h : W \to X$ and for all $w \in W$, $f \circ (g \circ h)(w) = (f \circ g) \circ h(w)$.

By virtue of this associativity property, we will simply use $f \circ g \circ h : W \to Y$ to denote the composition of three (or more) functions.

A special function that can always be defined on any set is the identity function, also called the identity map, or simply the identity.

Definition 1.5 *The identity map on X is the function $1_X : X \to X$ defined by $1_X(x) = x$ for all $x \in X$.*

Exercise 1.5 Show that the identity map is bijective.

Using the identity map one can also define various inverses of functions.

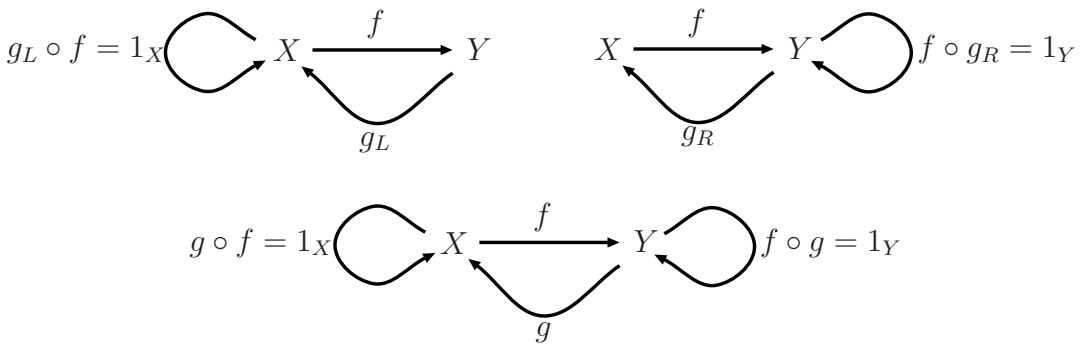

Figure 1.3: Commutative diagram of function inverses.

Definition 1.6 *Consider a function $f : X \to Y$.*

1. *The function $g_L : Y \to X$ is called a left inverse of f if and only if $g_L \circ f = 1_X$.*
2. *The function $g_R : Y \to X$ is called a right inverse of f if and only if $f \circ g_R = 1_Y$.*
3. *The function $g : Y \to X$ is called an inverse of f if and only if it is both a left inverse and a right inverse of f, i.e. $(g \circ f = 1_X) \wedge (f \circ g = 1_Y)$.*

f is called invertible if an inverse of f exists.

The commutative diagrams for the different types of inverses are shown in Figure 1.3. It turns out that these different notions of inverse are intimately related to the infectivity and subjectivity of the function f.

Theorem 1.8 *Consider two sets X and Y and a function $f : X \to Y$.*

1. *f has a left inverse if and only if it is injective.*
2. *f has a right inverse if and only if it is surjective.*
3. *f is invertible if and only if it is bijective.*
4. *If f is invertible then any two inverses (left-, right- or both) coincide.*

Proof: Parts 1-3 are left as an exercise (Problem 1.2). For part 4, assume, for the sake of contradiction, that f is invertible but there exist two different inverses, $g_1 : Y \to X$ and $g_2 : Y \to X$ (a similar argument applies to left- and right- inverses). Since the inverses are different, there must exist $y \in Y$ such that $g_1(y) \neq g_2(y)$. Let $x_1 = g_1(y)$ and $x_2 = g_2(y)$ and note that $x_1 \neq x_2$. Then

$$\begin{aligned} x_1 = g_1(y) = 1_X \circ g_1(y) &= (g_2 \circ f) \circ g_1(y) \quad (g_2 \text{ inverse of } f) \\ &= g_2 \circ (f \circ g_1)(y) \quad (\text{composition is associative}) \\ &= g_2 \circ 1_Y(y) \quad (g_1 \text{ inverse of } f) \\ &= g_2(y) = x_2. \end{aligned}$$

This contradicts the assumption that $x_1 \neq x_2$. ∎

Problems for chapter 1

Problem 1.1 (Even and odd numbers) Show that every $n \in \mathbb{N}$ is either even or odd, but not both.

Problem 1.2 (Inverses of functions) Consider two sets X and Y and a function $f : X \to Y$. Show that:

1. f has a left inverse if and only if it is injective.
2. f has a right inverse if and only if it is surjective.
3. f is invertible if and only if it is bijective.

Chapter 2

Introduction to Algebra

2.1 Groups

Definition 2.1 *A group $(G, *)$ is a set G equipped with a binary operation $* : G \times G \to G$ such that:*

1. *The operation $*$ is associative: $\forall a, b, c \in G$, $a * (b * c) = (a * b) * c$.*
2. *There exists an identity element: $\exists e \in G$, $\forall a \in G$, $a * e = e * a = a$.*
3. *Every element has an inverse element: $\forall a \in G$, $\exists a^{-1} \in G$, $a * a^{-1} = a^{-1} * a = e$.*

$(G, *)$ is called commutative (or Abelian) if and only if in addition to 1-3 *above*

4. *$*$ is commutative: $\forall a, b \in G$, $a * b = b * a$.*

Example (Common groups)

$(\mathbb{R}, +)$ is a commutative group. What is the identity element? What is the inverse?

$(\mathbb{R}, \cdot)$ is not a group, since 0 has no inverse.

The set $(\{0, 1, 2\}, +\text{mod}3)$ is a group. What is the identity element? What is the inverse? Is it commutative? Recall that $(0 + 1)\text{mod}3 = 1$, $(1 + 2)\text{mod}3 = 0$, $(2 + 2)\text{mod}3 = 1$, etc.

The set of rotations of $\mathbb{R}^2$ (usually denoted by SO(2), or U(1) or S(1)) given by

$$\left(\left\{ \left[\begin{array}{cc} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{array} \right] \middle| \, \theta \in (-\pi, \pi] \right\}, \cdot \right)$$

with the usual operation of matrix multiplication is a group[1]. What is the identity? What is the inverse? ■

Fact 2.1 *For a group $(G, *)$ the identity element, e, is unique. Moreover, for all $a \in G$ the inverse element, a^{-1}, is unique.*

[1]For the time being, the reader is asked to excuse the use of matrices in the examples. Matrices will be formally defined in the next section, but will be used in the meantime for informal illustrations.

Proof: To show the first statement, assume, for the sake of contradiction, that there exist two identity elements $e, e' \in G$ with $e \neq e'$. Then for all $a \in G$, $e * a = a * e = a$ and $e' * a = a * e' = a$. Then:

$$e = e * e' = e'$$

which contradicts the assumption that $e \neq e'$.

To show the second statement, assume, for the sake of contradiction, that there exists $a \in G$ with two inverse elements, say a_1 and a_2 with $a_1 \neq a_2$. Then

$$a_1 = a_1 * e = a_1 * (a * a_2) = (a_1 * a) * a_2 = e * a_2 = a_2,$$

which contradicts the assumption that $a_1 \neq a_2$. ■

2.2 Rings and fields

Definition 2.2 *A ring $(R, +, \cdot)$ is a set R equipped with two binary operations, $+ : R \times R \to R$ (called addition) and $\cdot : R \times R \to R$ (called multiplication) such that:*

1. *Addition satisfies the following properties:*
   - *It is associative: $\forall a, b, c \in R$, $a + (b + c) = (a + b) + c$.*
   - *It is commutative: $\forall a, b \in R$, $a + b = b + a$.*
   - *There exists an identity element: $\exists 0 \in R$, $\forall a \in R$, $a + 0 = a$.*
   - *Every element has an inverse element: $\forall a \in R$, $\exists (-a) \in R$, $a + (-a) = 0$.*
2. *Multiplication satisfies the following properties:*
   - *It is associative: $\forall a, b, c \in R$, $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.*
   - *There exists an identity element: $\exists 1 \in R$, $\forall a \in R$, $1 \cdot a = a \cdot 1 = a$.*
3. *Multiplication is distributive with respect to addition: $\forall a, b, c \in R$, $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(b + c) \cdot a = b \cdot a + c \cdot a$.*

$(R, +, \cdot)$ is called commutative if in addition $\forall a, b \in R$, $a \cdot b = b \cdot a$.

Example (Common rings)

$(\mathbb{R}, +, \cdot)$ is a commutative ring.

$(\mathbb{R}^{n \times n}, +, \cdot)$ with the usual operations of matrix addition and multiplication is a non-commutative ring.

The set of rotations

$$\left(\left\{ \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \middle| \theta \in (-\pi, \pi] \right\}, +, \cdot \right)$$

with the same operations is not a ring, since it is not closed under addition.

$(\mathbb{R}[s], +, \cdot)$, the set of polynomials of s with real coefficients, i.e. $a_n s^n + a_{n-1} s^{n-1} + \ldots + a_0$ for some $n \in \mathbb{N}$ and $a_0, \ldots, a_n \in \mathbb{R}$ is a commutative ring.

$(\mathbb{R}(s), +, \cdot)$, the set of rational functions of s with real coefficients, i.e.

$$\frac{a_m s^m + a_{m-1} s^{m-1} + \ldots + a_0}{b_n s^n + b_{n-1} s^{n-1} + \ldots + b_0}$$

for some $n, m \in \mathbb{N}$ and $a_0, \ldots, a_m, b_0, \ldots, b_n \in \mathbb{R}$ with $b_n \neq 0$ is a commutative ring. We implicitly assume here that the numerator and denominator polynomials are co-prime, that is they do not have any common factors; if they do one can simply cancel these factors until the two polynomials are co-prime. For example, it is easy to see that with such cancellations any rational function of the form

$$\frac{0}{b_n s^n + b_{n-1} s^{n-1} + \ldots + b_0}$$

can be identified with the rational function $0/1$, which is the identity element of addition for this ring.

$(\mathbb{R}_p(s), +, \cdot)$, the set of proper rational functions of s with real coefficients, i.e.

$$\frac{a_n s^n + a_{n-1} s^{n-1} + \ldots + a_0}{b_n s^n + b_{n-1} s^{n-1} + \ldots + b_0}$$

for some $n \in \mathbb{N}$ with $a_0, \ldots, a_n, b_0, \ldots, b_n \in \mathbb{R}$ with $b_n \neq 0$ is a commutative ring. Note that $a_n = 0$ is allowed, i.e. it is possible for the degree of the numerator polynomial to be less than or equal to that of the denominator polynomial. ■

Exercise 2.1 Show that for every ring $(R, +, \cdot)$ the identity elements 0 and 1 are unique. Moreover, for all $a \in R$ the inverse element $(-a)$ is unique.

Fact 2.2 *If $(R, +, \cdot)$ is a ring then:*

1. *For all $a \in R$, $a \cdot 0 = 0 \cdot a = 0$.*
2. *For all $a, b \in R$, $(-a) \cdot b = -(a \cdot b) = a \cdot (-b)$.*

Proof: To show the first statement note that

$$\begin{aligned} a + 0 = a \Rightarrow a \cdot (a + 0) = a \cdot a \Rightarrow a \cdot a + a \cdot 0 = a \cdot a \\ \Rightarrow -(a \cdot a) + a \cdot a + a \cdot 0 = -(a \cdot a) + a \cdot a \\ \Rightarrow 0 + a \cdot 0 = 0 \Rightarrow a \cdot 0 = 0. \end{aligned}$$

The second equation is similar. For the second statement note that

$$0 = 0 \cdot b = (a + (-a)) \cdot b = a \cdot b + (-a) \cdot b \Rightarrow -(a \cdot b) = (-a) \cdot b.$$

The second equation is again similar. ■

Definition 2.3 *A <u>field</u> $(F, +, \cdot)$ is a commutative ring that in addition satisfies*

- *Multiplication inverse: $\forall a \in F$ with $a \neq 0$, $\exists a^{-1} \in F$, $a \cdot a^{-1} = 1$.*

Example (Common fields)

$(\mathbb{R}, +, \cdot)$ is a field.

$(\mathbb{R}^{n \times n}, +, \cdot)$ is not a field, since singular matrices have no inverse.

$(\{A \in \mathbb{R}^{n \times n} \mid \textsc{Det}(A) \neq 0\}, +, \cdot)$ is not a field, since it is not closed under addition.

The set of rotations

$$\left(\left\{ \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \middle| \, \theta \in (-\pi, \pi] \right\}, +\cdot \right)$$

is not a field, it is not even a ring.

$(\mathbb{R}[s], +, \cdot)$ is not a field, since the multiplicative inverse of a polynomial is not a polynomial but a rational function.

$(\mathbb{R}(s), +, \cdot)$ is a field.

$(\mathbb{R}_p(s), +, \cdot)$ is not a field, since the multiplicative inverse of a proper rational function is not necessarily proper. ■

Exercise 2.2 If $(F, +, \cdot)$ is a field then for all $a \in R$ the multiplicative inverse element a^{-1} is unique.

Exercise 2.3 Show that if $(F, +, \cdot)$ is a field and $a \neq 0$ then $a \cdot b = a \cdot c \Leftrightarrow b = c$. Is the same true for a ring? Illustrate using an example from $\mathbb{R}^{2\times 2}$.

Given a field $(F, +, \cdot)$ and integers $n, m \in \mathbb{N}$ one can define matrices

$$A = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} \in F^{n\times m}$$

with $a_{11}, \dots, a_{nm} \in F$. One can then define the usual matrix operations in the usual way: matrix multiplication, matrix addition, determinant, identity matrix, inverse matrix, adjoint matrix, ... We assume that the reader is familiar with these operations from basic linear algebra, a brief summary is provided in Appendix B.

Matrices can also be formed from the elements of a commutative ring $(R, +, \cdot)$.

Example (Transfer function matrices) As we will see in Chapter 5, the set of "transfer functions" of time invariant linear systems with m inputs and p outputs are $p \times m$ matrices whose elements are proper rational functions,

$$\left.\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned}\right\} \Rightarrow G(s) = C(sI - A)^{-1}B + D \in \mathbb{R}_p(s)^{p\times m}.$$

■

There are, however, subtle differences.

Fact 2.3 *Assume that $(R, +, \cdot)$ is a commutative ring, $A \in R^{n\times n}$, and $\text{DET}(A) \neq 0$. It is not always the case that A^{-1} exists.*

Roughly speaking, what goes wrong is that even though $\text{DET}A \neq 0$, the inverse $(\text{DET}A)^{-1}$ may not exist in R. Then the inverse of the matrix defined as $A^{-1} = (\text{DET}(A))^{-1}\text{ADJ}(A)$ is not defined.

The example of the transfer function matrices given above illustrates the point. Consider, for example, the case $m = p = 2$ and the matrix

$$G(s) = \begin{bmatrix} \frac{1}{s+1} & \frac{1}{s+1} \\ -\frac{1}{s+1} & \frac{1}{s+1} \end{bmatrix} \in \mathbb{R}_p(s)^{2\times 2}.$$

Note that $\text{DET}[G(s)] = \frac{2}{(s+1)^2} \neq 0$ but

$$G(s)^{-1} = \frac{\text{ADJ}[G(s)]}{\text{DET}[G(s)]} = \begin{bmatrix} \frac{s+1}{2} & -\frac{s+1}{2} \\ \frac{s+1}{2} & \frac{s+1}{2} \end{bmatrix} \notin \mathbb{R}_p(s)^{2\times 2}.$$

In other words, the elements of the inverse of a transfer function matrix are not necessarily proper rational functions, and hence the inverse matrix cannot be the transfer function of a linear time invariant system.

2.3 Linear spaces

We now come to the algebraic object of greatest interest for linear systems, namely linear (or vector) spaces.

Definition 2.4 *A linear space $(V, F, \oplus, \odot)$ is a set V (of vectors) and a field $(F, +, \cdot)$ (of scalars) equipped with two binary operations, $\oplus : V \times V \to V$ (called vector addition) and $\odot : F \times V \to V$ (called scalar multiplication) such that:*

1. *Vector addition satisfies the following properties:*
   - *It is associative: $\forall x, y, z \in V$, $x \oplus (y \oplus z) = (x \oplus y) \oplus z$.*
   - *It is commutative: $\forall x, y \in V$, $x \oplus y = y \oplus x$.*
   - *There exists an identity element: $\exists \theta \in V$, $\forall x \in V$, $x \oplus \theta = x$.*
   - *For every element there exists an inverse element: $\forall x \in V$, $\exists (\ominus x) \in V$, $x \oplus (\ominus x) = \theta$.*
2. *Scalar multiplication satisfies the following properties:*
   - *It is associative: $\forall a, b \in F$, $x \in V$, $a \odot (b \odot x) = (a \cdot b) \odot x$.*
   - *Multiplication by the multiplication identity of F leaves elements unchanged: $\forall x \in V$, $1 \odot x = x$.*
3. *Scalar multiplication is distributive with respect to vector addition: $\forall a, b \in F$, $\forall x, y \in V$, $(a + b) \odot x = (a \odot x) \oplus (b \odot x)$ and $a \odot (x \oplus y) = (a \odot x) \oplus (a \odot y)$.*

Exercise 2.4 Let $(F, +, \cdot)$ be a field. Show that $(F, F, +, \cdot)$ is a linear space.

As for groups, rings and fields the following fact is immediate.

Exercise 2.5 For every linear space $(V, F, \oplus, \odot)$ the identity element θ is unique. Moreover, for all $x \in V$ there exists a unique inverse element $\ominus x$.

The following relations can also be established.

Fact 2.4 *If $(V, F, \oplus, \odot)$ is a linear space and 0 is the addition identity element of F then for all $x \in V$, $0 \odot x = \theta$. Moreover, for all $a \in F$, $x \in V$, $(-a) \odot x = \ominus(a \odot x) = a \odot (\ominus x)$.*

The proof is left as an exercise (Problem 2.3).

Once we have found a few linear spaces we can always generate more by forming the so called product spaces.

Definition 2.5 *If $(V, F, \oplus_V, \odot_V)$ and $(W, F, \oplus_W, \odot_W)$ are linear spaces over the same field, the product space $(V \times W, F, \oplus, \odot)$ is the linear space comprising all pairs $(v, w) \in V \times W$ with $\oplus$ defined by $(v_1, w_1) \oplus (v_2, w_2) = (v_1 \oplus_V v_2, w_1 \oplus_W w_2)$, and $\odot$ defined by $a \odot (v, w) = (a \odot_V v, a \odot_W w)$.*

Exercise 2.6 Show that $(V \times W, F, \oplus, \odot)$ is a linear space. What is the identity element for addition? What is the inverse element?

Two types of linear spaces will play a central role in these notes. The first is constructed by taking repeatedly the product of a field with itself.

Example (Finite product of a field) For any field, $(F, +, \cdot)$, consider the product space F^n. Let $x = (x_1, \ldots, x_n) \in F^n$, $y = (y_1, \ldots, y_n) \in F^n$ and $a \in F$ and define $\oplus : F^n \times F^n \to F^n$ by

$$x \oplus y = (x_1 + y_1, \ldots, x_n + y_n)$$

and $\odot : F \times F^n \to F^n$ by

$$a \odot x = (a \cdot x_1, \ldots, a \cdot x_n).$$

Note that both operations are well defined since a, $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ all take values in the same field, F.

Exercise 2.7 Show that $(F^n, F, \oplus, \odot)$ is a linear space. What is the identity element θ? What is the inverse element $\ominus x$ of $x \in F^n$?

The most important instance of this type of linear space in these notes will be $(\mathbb{R}^n, \mathbb{R}, +, \cdot)$ with the usual addition and scalar multiplication for vectors. The state, input, and output spaces of linear systems will be linear spaces of this type. ■

The second class of linear spaces that will play a key role in linear system theory are function spaces.

Example (Function spaces) Let $(V, F, \oplus_V, \odot_V)$ be a linear space and D be any set. Let $\mathcal{F}(D, V)$ denote the set of functions of the form $f : D \to V$. Consider $f, g \in \mathcal{F}(D, V)$ and $a \in F$ and define $\oplus : \mathcal{F}(D, V) \times \mathcal{F}(D, V) \to \mathcal{F}(D, V)$ by

$$(f \oplus g) : D \to V \text{ such that } (f \oplus g)(d) = f(d) \oplus_V g(d) \ \ \forall d \in D$$

and $\odot : F \times \mathcal{F}(D, V) \to \mathcal{F}(D, V)$ by

$$(a \odot f) : D \to V \text{ such that } (a \odot f)(d) = a \odot_V f(d) \ \ \forall d \in D$$

Note that both operations are well defined since $a \in F$, $f(d), g(d) \in V$ and $(V, F, \oplus_V, \odot_V)$ is a linear space.

Exercise 2.8 Show that $(\mathcal{F}(D, V), F, \oplus, \odot)$ is a linear space. What is the identity element? What is the inverse element?

The most important instance of this type of linear space in these notes will be $(\mathcal{F}([t_0, t_1], \mathbb{R}^n), \mathbb{R}, +, \cdot)$ for real numbers $t_0 < t_1$. The trajectories of the state, input, and output of the dynamical systems we consider will take values in linear spaces of this type. The state, input and output trajectories will differ in terms of their "smoothness" as functions of time. We will use the following notation to distinguish the level of smoothness of the function in question:

- $C([t_0, t_1], \mathbb{R}^n)$ will be the linear space of continuous functions $f : [t_0, t_1] \to \mathbb{R}^n$.
- $C^1([t_0, t_1], \mathbb{R}^n)$ will be the linear space of differentiable functions $f : [t_0, t_1] \to \mathbb{R}^n$.
- $C^k([t_0, t_1], \mathbb{R}^n)$ will be the linear space of k-times differentiable functions $f : [t_0, t_1] \to \mathbb{R}^n$.
- $C^\infty([t_0, t_1], \mathbb{R}^n)$ will be the linear space of infinitely differentiable functions $f : [t_0, t_1] \to \mathbb{R}^n$.
- $C^\omega([t_0, t_1], \mathbb{R}^n)$ will be the linear space of analytic functions $f : [t_0, t_1] \to \mathbb{R}^n$, i.e. functions which are infinitely differentiable and whose Taylor series expansion converges for all $t \in [t_0, t_1]$.

Exercise 2.9 Show that all of these sets are linear spaces. You only need to check that they are closed under addition and scalar multiplication. E.g. if f and g are differentiable, then so is $f \oplus g$.

Exercise 2.10 Show that for all $k = 2, 3, \ldots$

$$\begin{aligned} C^\omega([t_0, t_1], \mathbb{R}^n) \subset C^\infty([t_0, t_1], \mathbb{R}^n) \subset C^k([t_0, t_1], \mathbb{R}^n) \subset C^{k-1}([t_0, t_1], \mathbb{R}^n) \\ \subset C([t_0, t_1], \mathbb{R}^n) \subset (\mathcal{F}([t_0, t_1], \mathbb{R}^n), \mathbb{R}, +, \cdot). \end{aligned}$$

Note that all subsets are strict, so there must be functions that belong to one set but not the previous one. Try to think of examples. ■

To simplify the notation, unless there is special reason to distinguish the operations and identity element of a linear space from those of the field, from now on we will use the regular symbols $+$ and $\cdot$ instead of $\oplus$ and $\odot$ for the linear space operations of vector addition and scalar multiplication respectively; in fact as for real numbers we will mostly ommit $\cdot$ and simply write av instead of $a \odot v$ for $a \in F$, $v \in V$. Likewise, unless explicitly needed we will also use 0 instead of θ to denote the identity element of addition. Finally we will stop writing the operations explicitly when we define the vector space and write (V, F) or simply V whenever the field is clear from the context.

2.4 Subspaces and bases

Definition 2.6 *Let (V, F) be a linear space and $W \subseteq V$. (W, F) is a linear subspace of V if and only if it is itself a linear space, i.e. for all $w_1, w_2 \in W$ and all $a_1, a_2 \in F$, we have that $a_1 w_1 + a_2 w_2 \in W$.*

Note that by definition a linear space and all its subspaces are linear spaces over the same field. The equation provides a way of testing whether a given set W is a subspace or not: One needs to check whether *linear combinations* of elements of W with coefficients in F are also elements of W.

Exercise 2.11 Show that if W is a subspace then for all $n \in \mathbb{N}$, and $a_i \in F$, $w_i \in W$ for $i = 1, \dots, n$

$$\sum_{i=1}^{n} a_i w_i \in W.$$

Show further that $\theta_V \in W$. Hence show that $\theta_W = \theta_V$.

Example (Linear subspaces) In $\mathbb{R}^2$, the set $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 = 0\}$ is subspace. So is the set $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 = x_2\}$. But the set $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 = x_1 + 1\}$ is not a subspace and neither is the set $\{(x_1, x_2) \in \mathbb{R}^2 \mid (x_1 = 0) \vee (x_2 = 0)\}$.

In $\mathbb{R}^3$, all subspaces are:

1. $\mathbb{R}^3$
2. 2D planes through the origin.
3. 1D lines through the origin.
4. $\{0\}$.

For examples of subspace of function spaces consider $(\mathbb{R}[t], \mathbb{R})$ (polynomials of $t \in \mathbb{R}$ with real coefficients). This is a linear subspace of $C^\infty(\mathbb{R}, \mathbb{R})$, which in turn is a linear subspace of $C(\mathbb{R}, \mathbb{R})$. The set

$$\{f : \mathbb{R} \to \mathbb{R} \mid \forall t \in \mathbb{R},\ |f(t)| \leq 1\}$$

on the other hand is not a subspace of $\mathcal{F}(\mathbb{R}, \mathbb{R})$. ■

Exercise 2.12 Show that $\{f : \mathbb{R} \to \mathbb{R} \mid \forall t \in \mathbb{R},\ |f(t)| \leq 1\}$ is not a subspace. How about $\{f : \mathbb{R} \to \mathbb{R} \mid \exists M > 0,\ \forall t \in \mathbb{R},\ |f(t)| \leq M\}$?

It is easy to see that the family of subspaces of a given a linear space is closed under finite addition and intersection.

Exercise 2.13 Let $\{(W_i, F)\}_{i=1}^n$ be a finite family of subspaces of a linear space (V, F). Show that $(\cap_{i=1}^n W_i, F)$ is also a subspace. Is $(\cup_{i=1}^n W_i, F)$ a subspace?

Exercise 2.14 Let (W_1, F) and (W_2, F) be subspaces of (V, F) and define

$$W_1 + W_2 = \{w_1 + w_2 \mid w_1 \in W_1, w_2 \in W_2\}.$$

Show that $(W_1 + W_2, F)$ is a subspace of (V, F).

A subset of a linear space will of course not be a subspace in general. Each subset of a linear space does, however, generate a subspace in a natural way.

Definition 2.7 *Let (V, F) be a linear space and $S \subseteq V$. The linear subspace of (V, F) generated by S is the smallest subspace of (V, F) containing S.*

Here, "smallest" is to be understood in the sense of set inclusion.

Exercise 2.15 What is the subspace generated by $\{(x_1, x_2) \in \mathbb{R}^2 \mid (x_1 = 0) \vee (x_2 = 0)\}$? What is the subspace generated by $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 = x_1 + 1\}$? What is the subspace of $\mathbb{R}^2$ generated by $\{(1, 2)\}$?

Definition 2.8 *Let (V, F) be a linear space and $S \subseteq V$. The span of S is the set defined by*

$$\textsc{Span}(S) = \left\{ \sum_{i=1}^n a_i v_i \;\middle|\; n \in \mathbb{N}, a_i \in F, v_i \in S, i = 1, \ldots, n \right\}.$$

Fact 2.5 *Let (V, F) be a linear space and $S \subseteq V$. The subspace generated by S coincides with* $\textsc{Span}(S)$.

Proof: The fact that $\textsc{Span}(S)$ is a subspace and contains S is easy to check. To show that is is the smallest subspace containing S, consider another subspace, W, that contains S and an arbitrary $v \in \textsc{Span}(S)$; we will show that $v \in W$ and hence $\textsc{Span}(S) \subseteq W$. Since $v \in \textsc{Span}(S)$ it can be written as

$$v = \sum_{i=1}^n a_i v_i$$

for some $n \in \mathbb{N}$ and $a_i \in F$, $v_i \in S$, $i = 1, \ldots, n$. Since $S \subseteq W$ we must also have $v_i \in W$, $i = 1, \ldots, n$ and hence $v \in W$ (since W is a subspace). ∎

The elements of $\textsc{Span}(S)$ are known as *linear combinations* of elements of S. Notice that in general the set S may contain an infinite number of elements; this was for example the case for the set $\{(x_1, x_2) \in \mathbb{R}^2 \mid (x_1 = 0) \vee (x_2 = 0)\}$ in Exercise 2.15. The span of S, however, is defined as the set of all finite linear combinations of elements of S.

Definition 2.9 *Let (V, F) be a linear space. A set $S \subseteq V$ is called linearly independent if and only if for all $n \in \mathbb{N}$, $v_i \in S$ for $i = 1, \ldots, n$ with $v_i \neq v_j$ if $i \neq j$,*

$$\sum_{i=1}^n a_i v_i = 0 \Leftrightarrow a_i = 0, \ \forall i = 1, \ldots, n.$$

A set which is not linearly independent is called linearly dependent.

Note again that the set S may be infinite, but we only consider finite linear combinations to define linear independence.

Exercise 2.16 Show that a set $S \subseteq V$ is linearly independent if and only if none of its elements can be written as a finite linear combination of other elements in S. Show that every set S that contains the identity element of vector addition is linearly dependent.

Example (Linearly independent set in $\mathbb{R}^n$) The following finite family of vectors are linearly independent in $(\mathbb{R}^3, \mathbb{R})$:

$$\{(1,0,0),(0,1,0),(0,0,1)\}.$$

It is easy to show that any linear combination of these vectors is equal to $(0,0,0)$ if and only if all three coefficients are equal to 0. ■

Example (Linearly independent functions) Consider $C([-1,1],\mathbb{R})$ and let $f_k(t) = t^k$ for all $t \in [-1,1]$ and $k \in \mathbb{N}$. Clearly $f_k(t) \in C([-1,1],\mathbb{R})$.

Fact 2.6 *The vectors of the collection $\{f_k(t)\}_{k=0}^{\infty}$ are linearly independent.*

Proof: We need to show that for all $n \in \mathbb{N}$

$$\sum_{i=0}^{n} a_i f_i = \theta \Leftrightarrow a_i = 0, \forall i = 0, \ldots, N.$$

Here $\theta \in C([-1,1],\mathbb{R})$ denotes the addition identity element for $C([-1,1],\mathbb{R})$, i.e. the function $\theta : [-1,1] \to \mathbb{R}$ such that $\theta(t) = 0$ for all $t \in [-1,1]$.

One direction ($\Leftarrow$) is obvious. For the other ($\Rightarrow$) note that

$$\begin{aligned} f(t) = \sum_{i=0}^{n} a_n f_n(t) = a_0 + a_1 t + \ldots + a_n t^n = \theta \Rightarrow & f(t) = 0 \ \forall t \in [-1,1] \\ \Rightarrow & f(t) = \frac{d}{dt} f(t) = \ldots = \frac{d^n}{dt^n} f(t) = 0 \ \forall t \in [-1,1] \\ \Rightarrow & f(0) = \frac{d}{dt} f(0) = \ldots = \frac{d^n}{dt^n} f(0) = 0 \\ \Rightarrow & a_0 = a_1 = \ldots = n! a_n = 0. \end{aligned}$$

Hence $a_0 = a_1 = \ldots = a_n = 0$ and $\{f_k(t)\}_{k=0}^{\infty}$ are linearly independent. ■

Notice that in this case the collection $\{f_k(t)\}_{k=0}^{\infty}$ contains an infinite number of linearly independent elements. ■

At this stage one may be tempted to ask how many linearly independent vectors can be found in a given vector space. This number is a fundamental property of the vector space, and is closely related to the notion of a basis.

Definition 2.10 *Let (V,F) be a linear space. A set of vectors $S \subseteq V$ is a <u>basis of (V,F)</u> if and only if they are linearly independent and* $\textsc{Span}(S) = V$.

One can then show the following.

Fact 2.7 *Let (V,F) be a linear space. If a basis of (V,F) with a finite number of elements exists, then every other basis of (V,F) has the same number of elements.*

Proof: Consider two bases of (V,F), S and S' and assume that S has a finite number of elements $\{v_i\}_{i=1}^{n}$. Assume, for the sake of contradiction that S' does not have the same number of elements.

Without loss of generality assume further that S' has more elements than S; if not, interchange S and S' since in this case S' must also have a finite number of elements. Take $n+1$ elements from S', $\{v_i'\}_{i=1}^{n+1}$ and recall that since S' is a basis they must be linearly independent. Since $\{v_i\}_{i=1}^{n}$ generate V we can write v_1' as a linear combination

$$v_1' = a_1 v_1 + \ldots + a_n v_n.$$

Note that at least one of the a_i must be non-zero; otherwise $v_1' = 0$ and the set $\{v_i'\}_{i=1}^{n+1}$ cannot be linearly independent (Exercise 2.16). Assume, without loss of generality that $a_1 \neq 0$ and write

$$v_1 = \frac{1}{a_1} v_1' - \frac{a_2}{a_1} v_2 - \ldots - \frac{a_n}{a_1} v_n$$

(where we make use of the identities in Problem 2.3). Since $\{v_i\}_{i=1}^{n}$ is a basis any element $v \in V$ can be expressed as a linear combination of v_1, v_2, …, v_n and hence, by the above equation, as a linear combination of v_1', v_2, …, v_n. In particular, we can write

$$v_2' = b_1 v_1' + b_2 v_2 + \ldots + b_n v_n.$$

Note again that the b_2, …, b_n cannot all be zero; otherwise $v_2' = b_1 v_1'$ and $\{v_i'\}_{i=1}^{n+1}$ cannot be linearly independent. Assume, without loss of generality, that $b_2 \neq 0$ and write

$$v_2 = \frac{1}{b_2} v_2' - \frac{b_1}{b_2} v_1' - \ldots - \frac{b_n}{b_2} v_n.$$

This shows that every vector is V can be written as a linear combination of v_1', v_2', v_3, …, v_n. Repeat for v_3', etc. until finally

$$v_{n+1}' = c_1 v_1' + c_2 v_2' + \ldots + c_n v_n'.$$

This, however, contradicts the assumption that the set $\{v_i\}_{i=1}^{n}$ is linearly independent. ∎

The fact provides conditions under which the number of elements of a basis of a linear space in well defined, and independent of the choice of the basis. A similar statement can be made for linear spaces which do not have a basis with a finite number of elements. In this case there are again families of vectors whose span covers "almost all" elements of the vector space. The proof, however, is considerably more involved. It also relies on concepts covered in Chapters 3 and 7 that go beyond the purely algebraic structure considered in this chapter.

Definition 2.11 *Let (V, F) be a linear space. If a basis of (V, F) with a finite number of elements exists, the number of elements of this basis is called the dimension of (V, F) and (V, F) is called finite dimensional. If not, (V, F) is called infinite dimensional.*

Exercise 2.17 If (V, F) has dimension n then any set of $n+1$ or more vectors is linearly dependent.

Example (Bases) The vectors $\{(1,0,0),(0,1,0),(0,0,1)\}$ form a basis for the linear space $\mathbb{R}^3$ over the field $\mathbb{R}$. This is called the "canonical basis of $\mathbb{R}^3$" and is usually denoted by $\{e_1, e_2, e_3\}$. This shows that $\mathbb{R}^3$ is finite dimensional and of dimension 3. Other choices of basis are of course possible, for example $\{(1,1,0),(0,1,0),(0,1,1)\}$. In fact any three linearly independent vectors will form a basis for $\mathbb{R}^3$.

In the same way, the vectors $\{(1,0,0),(0,1,0),(0,0,1)\}$ form the canonical basis for the linear space $\mathbb{C}^3$ over the field $\mathbb{C}$; therefore $\mathbb{C}^3$ over $\mathbb{C}$ has dimension 3. On the other hand, $\mathbb{C}^3$ is a linear space also over the field $\mathbb{R}$; in this case it has dimension 6, the following being a basis: $\{(1,0,0),(i,0,0),(0,1,0),(0,i,0),(0,0,1),(0,0,i)\}$.

Exercise 2.18 Find a basis for $\mathbb{R}^n$ over the field $\mathbb{R}$. Hence, show that $\mathbb{R}^n$ has dimension n over $\mathbb{R}$. Show that $\mathbb{C}^n$ has dimension n over $\mathbb{C}$ and $2n$ over $\mathbb{R}$.

The linear space $\mathcal{F}([-1,1],\mathbb{R})$ is infinite dimensional. We have already shown that the collection $\{t^k \mid k \in \mathbb{N}\} \subseteq \mathcal{F}([-1,1],\mathbb{R})$ is linearly independent. The collection contains an infinite number of elements and may or may not span $\mathcal{F}([-1,1],\mathbb{R})$. Therefore any basis of $\mathcal{F}([-1,1],\mathbb{R})$ (which must by definition span the set) must contain at least as many elements. ■

Let $\{b_1, b_2, \ldots, b_n\}$ be a basis of a finite dimensional linear space (V, F). By definition,

$$\textsc{Span}(\{b_1,\ldots,b_n\}) = V \text{ therefore } \forall x \in V\ \exists \xi_1,\ldots,\xi_n \in F\ :\ x = \sum_{i=1}^{n} \xi_i b_i.$$

The vector $\xi = (\xi_1, \ldots, \xi_n) \in F^n$ is called the representation of $x \in V$ with respect to the basis $\{b_1, b_2, \ldots, b_n\}$.

Fact 2.8 *The representation of a given $x \in V$ with respect to a basis $\{b_1, \ldots, b_n\}$ is unique.*

The proof is left as an exercise (Problem 2.6).

Representations of the same vector with respect to different bases can of course be different.

Example (Representations) Let $x = (x_1, x_2, x_3) \in (\mathbb{R}^3, \mathbb{R})$. The representation of x with respect to the canonical basis is simply $\xi = (x_1, x_2, x_3)$. The representation with respect to the basis $\{(1,1,0),(0,1,0),(0,1,1)\}$, however, is $\xi' = (x_1, x_2 - x_1 - x_3, x_3)$ since

$$x = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (x_2 - x_1 - x_3) \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

■

Representations can also be defined for infinite dimensional spaces, but we will not get into the details here. As an example, consider $f(t) \in C^\omega([-1,1],\mathbb{R})$. One can consider a "representation" of $f(t)$ with respect to the basis $\{t^k \mid k \in \mathbb{N}\}$ defined through the Taylor series expansion. For example, expansion about $t = 0$ gives

$$f(t) = f(0) + \frac{df}{dt}(0)t + \frac{1}{2}\frac{d^2f}{dt^2}(0)t^2 + \ldots.$$

which suggests that the representation of $f(t)$ is $\xi = (f(0), \frac{df}{dt}(0), \frac{1}{2}\frac{d^2f}{dt^2}(0), \ldots)$. Making this statement formal, however, is beyond the scope of these notes.

It turns out that all representations of an element of a linear space are related to one another: Knowing one we can compute all others. To do this we need the concept of linear maps.

2.5 Linear maps

Definition 2.12 *Let (U, F) and (V, F) be two linear spaces. The function $\mathcal{A} : U \to V$ is called linear if and only if $\forall u_1, u_2 \in U$, $a_1, a_2 \in F$*

$$\mathcal{A}(a_1u_1 + a_2u_2) = a_1\mathcal{A}(u_1) + a_2\mathcal{A}(u_2).$$

Note that both linear spaces have to be defined over the same field. For clarity we will sometimes write $\mathcal{A} : (U, F) \to (V, F)$ if we need to specify the field over which the linear spaces are defined.

Example (Linear maps) Let $(U, F) = (\mathbb{R}^n, \mathbb{R})$, $(V, F) = (\mathbb{R}^m, \mathbb{R})$ and consider a matrix $A \in \mathbb{R}^{m\times n}$. Define

$$\begin{array}{rccc} \mathcal{A}: & \mathbb{R}^n & \to & \mathbb{R}^m \\ & u & \mapsto & A \cdot u. \end{array}$$

It is easy to show that $\mathcal{A}$ is a linear map. Indeed:

$$\mathcal{A}(a_1u_1 + a_2u_2) = A \cdot (a_1u_1 + a_2u_2) = a_1A \cdot u_1 + a_2A \cdot u_2 = a_1\mathcal{A}(u_1) + a_2\mathcal{A}(u_2).$$

Consider now $f \in \mathcal{C}([0,1], \mathbb{R})$ and define the functions

$$\begin{aligned} \mathcal{A} : (\mathcal{C}([0,1],\mathbb{R}),\mathbb{R}) &\to (\mathbb{R},\mathbb{R}) \\ f &\mapsto \int_0^1 f(t)dt \quad \text{(integration)} \\ \mathcal{A}' : (\mathcal{C}([0,1],\mathbb{R}),\mathbb{R}) &\to (\mathcal{C}([0,1],\mathbb{R}),\mathbb{R}) \\ f &\mapsto g(t) = \int_0^t e^{-a(t-\tau)} f(\tau)d\tau \quad \text{(convolution with } e^{-at}\text{)}. \end{aligned}$$

Exercise 2.19 Show that the functions $\mathcal{A}$ and $\mathcal{A}'$ are both linear.

■

It is easy to see that linear maps map the zero element of their domain to the zero element of their co-domain.

Exercise 2.20 Show that if $\mathcal{A} : U \to V$ is linear then $\mathcal{A}(\theta_U) = \theta_V$.

Other elements of the domain may also be mapped to the zero element of the co-domain, however.

Definition 2.13 *Let $\mathcal{A} : U \to V$ linear. The null space of $\mathcal{A}$ is the set*

$$\text{Null}(\mathcal{A}) = \{u \in U \mid \mathcal{A}(u) = \theta_V\} \subseteq U$$

and the range space of $\mathcal{A}$ is the set

$$\text{Range}(\mathcal{A}) = \{v \in V \mid \exists u \in U \ : \ v = \mathcal{A}(u)\} \subseteq V.$$

The word "space" in "null space" and "range space" is of course not accidental.

Fact 2.9 *Show that* $\text{Null}(\mathcal{A})$ *is a linear subspace of* (U, F) *and* $\text{Range}(\mathcal{A})$ *is a linear subspace of* (V, F).

The proof is left as an exercise (Problem 2.5). It is easy to see that the properties of the null and range spaces are closely related to the injectivity and surjectivity of the corresponding linear map, and hence its invertibility (Problem 1.2).

Theorem 2.1 *Let $\mathcal{A} : U \to V$ be a linear map and let $b \in V$.*

1. *A vector $u \in U$ such that $\mathcal{A}(u) = b$ exists if and only if $b \in \text{Range}(\mathcal{A})$. In particular, $\mathcal{A}$ is surjective if and only if $\text{Range}(\mathcal{A}) = V$.*
2. *If $b \in \text{Range}(\mathcal{A})$ and for some $u_0 \in U$ we have that $\mathcal{A}(u_0) = b$ then for all $u \in U$:*

$$\mathcal{A}(u) = b \Leftrightarrow u = u_0 + z \text{ with } z \in \text{Null}(\mathcal{A}).$$

3. *If $b \in \text{Range}(\mathcal{A})$ there exists a unique $u \in U$ such that $\mathcal{A}(u) = b$ if and only if $\text{Null}(\mathcal{A}) = \{\theta_U\}$. In other words, $\mathcal{A}$ is injective if and only if $\text{Null}(\mathcal{A}) = \{\theta_U\}$.*

Proof: Part 1 follows by definition.

Part 2, ($\Leftarrow$):

$$u = u_0 + z \Rightarrow \mathcal{A}(u) = \mathcal{A}(u_0 + z) = \mathcal{A}(u_0) + \mathcal{A}(z) = \mathcal{A}(u_0) = b.$$

Part 2, ($\Rightarrow$):

$$\mathcal{A}(u) = \mathcal{A}(u_0) = b \Rightarrow \mathcal{A}(u - u_0) = \theta_V \Rightarrow z = (u - u_0) \in \text{NULL}(\mathcal{A}).$$

Part 3 follows from part 2. ∎

Finally, we generalise the concept of an eigenvalue to more general linear maps of a (potentially infinite dimensional) linear space.

Definition 2.14 *Let (V, F) be a linear space and consider a linear map $\mathcal{A} : V \to V$. An element $\lambda \in F$ is called an eigenvalue of $\mathcal{A}$ if and only if there exists $v \in V$ such that $v \neq \theta_V$ and $\mathcal{A}(v) = \lambda \cdot v$. In this case, v is called an eigenvector of $\mathcal{A}$ for the eigenvalue λ.*

Example (Eigenvalues) For maps between finite dimensional spaces defined by matrices, the interpretation of eigenvalues is the familiar one from linear algebra. Since eigenvalues and eigenvectors are in general complex numbers/vectors we consider matrices as linear maps between complex finite dimensional spaces (even if the entries of the matrix itself are real). For example, consider the linear space $(\mathbb{C}^2, \mathbb{C})$ and the linear map $\mathcal{A} : \mathbb{C}^2 \to \mathbb{C}^2$ defined by the matrix

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

through matrix multiplication; in other words, for all $x \in \mathbb{C}^2$, $\mathcal{A}(x) = Ax$.

It is easy to see that the eigenvalues of $\mathcal{A}$ are $\lambda_1 = j$ and $\lambda_2 = -j$. Moreover, any vector of the form $c \begin{bmatrix} j \\ -1 \end{bmatrix}$ for any $c \in \mathbb{C}$ with $c \neq 0$ is an eigenvector of λ_1 and any vector of the form $c \begin{bmatrix} j \\ 1 \end{bmatrix}$ is an eigenvector of λ_2.

Definition 2.14 also applies to infinite dimensional spaces, however. Consider, for example, the linear space $(C^\infty([t_0, t_1], \mathbb{R}), \mathbb{R})$ of infinitely differentiable real valued functions of the interval $[t_0, t_1]$. Consider also the function $\mathcal{A} : C^\infty([t_0, t_1], \mathbb{R}) \to C^\infty([t_0, t_1], \mathbb{R})$ defined by differentiation, i.e. for all $f : [t_0, t_1] \to \mathbb{R}$ infinitely differentiable define

$$(\mathcal{A}(f))(t) = \frac{df}{dt}(t), \quad \forall t \in [t_0, t_1].$$

Exercise 2.21 Show that $\mathcal{A}$ is well defined, i.e. if $f \in C^\infty([t_0, t_1], \mathbb{R})$ then $\mathcal{A}(f) \in C^\infty([t_0, t_1], \mathbb{R})$. Show further that $\mathcal{A}$ is linear.

One can see that in this case the linear map $\mathcal{A}$ has infinitely many eigenvalues. Indeed, any function $f : [t_0, t_1] \to \mathbb{R}$ of the form $f(t) = e^{\lambda t}$ for $\lambda \in \mathbb{R}$ is an eigenvector with eigenvalue λ, since

$$(\mathcal{A}(f))(t) = \frac{d}{dt} e^{\lambda t} = \lambda e^{\lambda t} = \lambda f(t), \quad \forall t \in [t_0, t_1]$$

which is equivalent to $\mathcal{A}(f) = \lambda \cdot f$. ∎

Exercise 2.22 Let $\mathcal{A} : (V, F) \to (V, F)$ be a linear map and consider any $\lambda \in F$. Show that the set $\{v \in V \mid \mathcal{A}(v) = \lambda v\}$ is a subspace of V.

2.6 Linear maps generated by matrices

Two of the examples in the previous section made use of matrices to define linear maps between finite dimensional linear spaces. It is easy to see that this construction is general: Given a field F, every matrix $A \in F^{m\times n}$ defines a linear map $\mathcal{A} : (F^n, F) \to (F^m, F)$ by matrix multiplication, i.e. for all $x = (x_1, \dots, x_n) \in F^n$,

$$\mathcal{A}(x) = Ax = \begin{bmatrix} \sum_{j=1}^n a_{1j}x_j \\ \vdots \\ \sum_{j=1}^n a_{mj}x_j \end{bmatrix}.$$

Naturally, for linear maps between finite dimensional spaces that are generated by matrices a close relation between the properties of the linear map and those of the matrix exist.

Definition 2.15 *Let F be a field, $A \in F^{n\times m}$ be a matrix and $\mathcal{A} : (F^m, F) \to (F^n, F)$ the linear map defined by $\mathcal{A}(x) = Ax$ for all $x \in F^m$. The rank of the matrix A (denoted by $\textsc{Rank}(A)$) is the dimension of the range space, $\textsc{Range}(\mathcal{A})$, of $\mathcal{A}$. The nullity of A (denoted by $\textsc{Nullity}(A)$) is the dimension of the null space, $\textsc{Null}(\mathcal{A})$, of $\mathcal{A}$.*

The following facts can be used to link the nullity and rank of matrices. They will prove useful when manipulating matrices later on.

Theorem 2.2 *Let $A \in F^{n\times m}$ and $B \in F^{m\times p}$.*

1. $\textsc{Rank}(A) + \textsc{Nullity}(A) = m$.
2. $0 \le \textsc{Rank}(A) \le \min\{m, n\}$.
3. $\textsc{Rank}(A) + \textsc{Rank}(B) - m \le \textsc{Rank}(AB) \le \min\{\textsc{Rank}(A), \textsc{Rank}(B)\}$.
4. *If $P \in F^{m\times m}$ and $Q \in F^{n\times n}$ are invertible then*

$$\begin{aligned} &\textsc{Rank}(A) = \textsc{Rank}(AP) = \textsc{Rank}(QA) = \textsc{Rank}(QAP) \\ &\textsc{Nullity}(A) = \textsc{Nullity}(AP) = \textsc{Nullity}(QA) = \textsc{Nullity}(QAP). \end{aligned}$$

The proof is left as an exercise (Problem 2.7).

For square matrices there exists a close connection between the properties of $\mathcal{A}$ and the invertibility of the matrix A.

Theorem 2.3 *Let F be a field, $A \in F^{n\times n}$ be a matrix, and $\mathcal{A} : F^n \to F^n$ the linear map defined by $\mathcal{A}(x) = Ax$ for all $x \in F^n$. The following statements are equivalent:*

1. *A is invertible.*
2. *$\mathcal{A}$ is bijective.*
3. *$\mathcal{A}$ is injective.*
4. *$\mathcal{A}$ is surjective.*
5. $\textsc{Rank}(A) = n$.
6. $\textsc{Nullity}(A) = 0$.
7. *The columns $a_{\bullet j} = (a_{1j}, \dots, a_{nj}) \in F^n$ form a linearly independent set $\{a_{\bullet j}\}_{j=1}^n$.*

8. *The rows* $a_{i\bullet} = (a_{i1}, \ldots, a_{in}) \in F^n$ *form a linearly independent set* $\{a_{i\bullet}\}_{i=1}^n$.

The proof is left as an exercise (Problem 2.8).

Finally, for linear maps defined through matrices, one can give the usual interpretation to eigenvalues and eigenvectors.

Theorem 2.4 *Let* F *be a field,* $A \in F^{n\times n}$ *be a matrix, and* $\mathcal{A} : F^n \to F^n$ *be the linear map defined by* $\mathcal{A}(x) = Ax$ *for all* $x \in F^n$. *The following statements are equivalent:*

1. $\lambda \in \mathbb{C}$ *is an eigenvalue of* $\mathcal{A}$.
2. $\textsc{Det}(\lambda I - A) = 0$.
3. *There exists* $v \in \mathbb{C}^n$ *such that* $v \neq 0$ *and* $Av = \lambda v$. *Such a* v *is called a* right eigenvector *of* A *for the eigenvalue* λ.
4. *There exists* $\eta \in \mathbb{C}^n$ *such that* $\eta \neq 0$ *and* $\eta^T A = \lambda \eta^T$. *Such an* η *is called a* left eigenvector *of* A *for the eigenvalue* λ.

The proof is left as an exercise (Problem 2.11). The theorem shows (among other things) that the eigenvalues of a linear map defined by a square matrix $A \in F^{n\times n}$ can be computed by finding the roots of the characteristic polynomial of the matrix, defined by

$$\chi_A(\lambda) = \textsc{Det}[\lambda I - A] = s^n + \chi_1 s^{n-1} + \ldots + \chi_{n-1} s + \chi_n.$$

Notice that by definition the characteristic polynomial is a monic polynomial (the leading coefficient is equal to 1) of degree n.

Definition 2.16 *The* spectrum *of a matrix*[2] $A \in F^{n\times n}$ *is the list of eigen values of* A, *denoted by* $\textsc{Spec}[A] = \{\lambda_1, \ldots, \lambda_n\}$.

Finally, combining the notion of eigenvalue with Theorem 2.3 leads to the following condition for the invertibility of a matrix.

Theorem 2.5 *A matrix* $A \in F^{n\times n}$ *is invertible if and only if none of its eigenvalues are equal to the zero element* $0 \in F$.

The proof is left as an exercise (Problem 2.8).

2.7 Matrix representation of linear maps

In the previous section we saw that every matrix $A \in F^{m\times n}$ defines a linear map $\mathcal{A} : F^n \to F^m$ by simple matrix multiplication. In fact it can be shown that the opposite is also true: Any linear map between two finite dimensional linear spaces can be represented as a matrix by fixing bases for the two spaces.

Consider a linear map

$$\mathcal{A} : U \longrightarrow V$$

between two finite dimensional linear spaces (U, F) and (V, F). Assume that the dimension of (U, F) is n and the dimension of (V, F) is m. Fix bases $\{u_j\}_{j=1}^n$ for (U, F) and $\{v_i\}_{i=1}^m$ for (V, F). Let

$$y_j = \mathcal{A}(u_j) \in V \text{ for } j = 1, \ldots, n.$$

[2]Sometimes a distinction is made between the *spectrum* of the matrix (list of eigenvalues with repeated eigenvalues included) and the *spectra list* of a matrix (list of eigenvalues without repetitions). We will however not make this distinction here.

The vectors $y_j \in V$ can all be represented with respect to the basis $\{v_i\}_{i=1}^m$ of (V, F). In other words for all $j = 1, \dots, n$ there exist unique $a_{ij} \in F$ such that

$$y_j = \mathcal{A}(u_j) = \sum_{i=1}^{m} a_{ij} v_i.$$

The $a_{ij} \in F$ can then be used to form a matrix

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \in F^{m \times n}.$$

Since representations are unique (Fact 2.8), the linear map $\mathcal{A}$ and the bases $\{u_j\}_{j=1}^n$ and $\{v_i\}_{i=1}^m$ uniquely define the matrix $A \in F^{m \times n}$.

Consider now an arbitrary $x \in U$. Again there exists unique representation $\xi \in F^n$ of x with respect to the basis $\{u_j\}_{j=1}^n$,

$$x = \sum_{j=1}^{n} \xi_j u_j$$

Let us now see what happens to this representation if we apply the linear map $\mathcal{A}$ to x. Clearly $\mathcal{A}(x) \in V$, therefore there exists a unique representation $\eta \in F^m$ of $\mathcal{A}(x)$ with respect to the basis $\{v_i\}_{i=1}^m$. It turns out that the two representations of x and $\mathcal{A}(x)$ are related by matrix multiplication.

Fact 2.10 *$\xi \in F^n$ of a vector $x \in U$ with respect to the basis $\{u_j\}_{j=1}^n$ and $\eta \in F^m$ of $\mathcal{A}(x) \in V$ with respect to the basis $\{v_i\}_{i=1}^m$. Then $\eta = A \cdot \xi$, where $\cdot$ denotes standard matrix multiplication.*

Proof: By definition

$$\mathcal{A}(x) = \sum_{i=1}^{m} \eta_i v_i.$$

Recall that

$$\mathcal{A}(x) = \mathcal{A}\left(\sum_{j=1}^{n} \xi_j u_j\right) = \sum_{j=1}^{n} \xi_j \mathcal{A}(u_j) = \sum_{j=1}^{n} \xi_j \sum_{i=1}^{m} a_{ij} v_i = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} \xi_j\right) v_i.$$

By uniqueness of representation

$$\eta_i = \sum_{j=1}^{n} a_{ij} \xi_j \Rightarrow \eta = A \cdot \xi,$$

where $\cdot$ denotes the standard matrix multiplication. ∎

Therefore, when one looks at the representations of vectors with respect to given bases, application of the linear map $\mathcal{A}$ to $x \in U$ is equivalent to multiplication of its representation (an element of F^n) with the matrix $A \in F^{m \times n}$. To illustrate this fact we will write things like

$$\begin{array}{rcl} (U, F) & \xrightarrow{\mathcal{A}} & (V, F) \\ x & \longmapsto & \mathcal{A}(x) \\ \{u_j\}_{j=1}^n & \xrightarrow{A \in F^{m \times n}} & \{v_i\}_{i=1}^m \\ \xi \in F^n & \longmapsto & A\xi \in F^m \end{array}$$

Theorem 2.6 *The following relations between linear map operations and the corresponding matrix representations hold:*

1. *Consider linear maps $\mathcal{B} : (U, F) \to (V, F)$ and $\mathcal{A} : (V, F) \to (W, F)$ where U, V and W are finite dimensional linear spaces of dimensions n, m and p respectively. Then composition $\mathcal{C} = \mathcal{A} \circ \mathcal{B} : (U, F) \to (W, F)$ is also a linear map. Moreover, if we fix bases $\{u_k\}_{k=1}^n$, $\{v_i\}_{i=1}^m$ and $\{w_j\}_{j=1}^p$ for the three spaces and*

$$\begin{array}{ccc} (U,F) & \xrightarrow{\mathcal{B}} & (V,F) \\ \{u_k\}_{k=1}^n & \xrightarrow[B\in F^{m\times n}]{} & \{v_i\}_{i=1}^m \end{array} \quad \text{and} \quad \begin{array}{ccc} (V,F) & \xrightarrow{\mathcal{A}} & (W,F) \\ \{v_i\}_{i=1}^m & \xrightarrow[A\in F^{p\times m}]{} & \{w_j\}_{j=1}^p \end{array}$$

   then

$$\begin{array}{ccc} (U,F) & \xrightarrow{\mathcal{C}=\mathcal{A}\circ\mathcal{B}} & (W,F) \\ \{u_k\}_{k=1}^n & \xrightarrow{C=A\cdot B\in F^{p\times n}} & \{w_j\}_{j=1}^p \end{array}$$

   where $\cdot$ denotes the standard matrix multiplication.

2. *Consider an invertible linear map $\mathcal{A} : (V, F) \to (V, F)$ on an n-dimensional linear space V and let $\mathcal{A}^{-1} : (V, F) \to (V, F)$ denote its inverse. If A is the representation of $\mathcal{A}$ with respect to a given basis of V, then A^{-1} is the representation of $\mathcal{A}^{-1}$ with respect to the same basis.*

The proof is left as an exercise (Problem 2.9). Analogous statements can of course be made about the representations of linear maps obtained by adding, or scaling other linear maps.

2.8 Change of basis

Given a linear map, $\mathcal{A} : (U, F) \to (V, F)$, selecting different bases for the linear spaces (U, F) and (V, F) leads to different representations.

$$\begin{array}{ccc} (U,F) & \xrightarrow{\mathcal{A}} & (V,F) \\ \{u_j\}_{j=1}^n & \xrightarrow{A\in F^{m\times n}} & \{v_i\}_{i=1}^m \\ \{\tilde{u}_j\}_{j=1}^n & \xrightarrow{\tilde{A}\in F^{m\times n}} & \{\tilde{v}_i\}_{i=1}^m \end{array}$$

In this section we investigate the relation between the two representations A and $\tilde{A}$.

Recall first that changing basis changes the representations of all vectors in the linear spaces. It is therefore expected that the representation of a linear map will also change.

Example (Change of basis) Consider $x = (x_1, x_2) \in \mathbb{R}^2$. The representation of x with respect to the canonical basis $\{u_1, u_2\} = \{(1, 0), (0, 1)\}$ is simply $\xi = (x_1, x_2)$. The representation with respect to the basis $\{\tilde{u}_1, \tilde{u}_2\} = \{(1, 0), (1, 1)\}$ is $\tilde{\xi} = (x_1 - x_2, x_2)$ since

$$x = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = (x_1 - x_2) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

■

To derive the relation between A and $\tilde{A}$, consider first the identity map

$$\begin{array}{ccc} (U,F) & \xrightarrow{1_U} & (U,F) \\ x & \longmapsto & 1_U(x) = x \\ \{u_j\}_{j=1}^n & \xrightarrow{I\in F^{n\times n}} & \{u_j\}_{j=1}^n \\ \{\tilde{u}_j\}_{j=1}^n & \xrightarrow{Q\in F^{n\times n}} & \{u_j\}_{j=1}^n \end{array}$$

I denotes the usual identity matrix in $F^{n\times n}$

$$I = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \in F^{n\times n}$$

where 0 and 1 are the addition and multiplication identity of F. The argument used to derive the representation of a linear map as a matrix in Section 2.7 suggests that the elements of $Q \in F^{n\times n}$ are simply the representations of $1_U(\tilde{u}_j) = \tilde{u}_j$ (i.e. the elements of the basis $\{\tilde{u}_j\}_{j=1}^n$) with respect to the basis $\{u_j\}_{j=1}^n$. Likewise

$$\begin{array}{rcl} (V,F) & \xrightarrow{1_V} & (V,F) \\ x & \longmapsto & 1_V(x) = x \\ \{v_i\}_{i=1}^m & \xrightarrow{I\in F^{m\times m}} & \{v_i\}_{i=1}^m \\ \{v_i\}_{i=1}^m & \xrightarrow{P\in F^{m\times m}} & \{\tilde{v}_i\}_{i=1}^m \end{array}$$

Exercise 2.23 Show that the matrices $Q \in F^{n\times n}$ and $P \in F^{m\times m}$ are invertible. (Recall that 1_U and 1_V are bijective functions.)

Example (Change of basis (cont.)) Consider the identity map

$$\begin{array}{rcl} \mathbb{R}^2 & \xrightarrow{1_{\mathbb{R}^2}} & \mathbb{R}^2 \\ x & \longmapsto & x \\ & I=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\in\mathbb{R}^{2\times 2} & \\ \{(1,0),(0,1)\} & \longrightarrow & \{(1,0),(0,1)\} \\ (x_1,x_2) & \longmapsto & (x_1,x_2) \end{array}$$

On the other hand,

$$\begin{array}{rcl} \mathbb{R}^2 & \xrightarrow{1_{\mathbb{R}^2}} & \mathbb{R}^2 \\ x & \longmapsto & x \\ & Q=\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\in\mathbb{R}^{2\times 2} & \\ \{(1,0),(1,1)\} & \longrightarrow & \{(1,0),(0,1)\} \\ (\tilde{x}_1,\tilde{x}_2) & \longmapsto & (x_1,x_2) = (\tilde{x}_1+\tilde{x}_2,\tilde{x}_2) \end{array}$$

and

$$\begin{array}{rcl} \mathbb{R}^2 & \xrightarrow{1_{\mathbb{R}^2}} & \mathbb{R}^2 \\ x & \longmapsto & x \\ & Q^{-1}=\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}\in\mathbb{R}^{2\times 2} & \\ \{(1,0),(0,1)\} & \longrightarrow & \{(1,0),(1,1)\} \\ (x_1,x_2) & \longmapsto & (\tilde{x}_1,\tilde{x}_2) = (x_1-x_2,x_2) \end{array}$$

■

Now notice that $\mathcal{A} = 1_V \circ \mathcal{A} \circ 1_U$, therefore

$$\begin{array}{ccccccc} (U,F) & \xrightarrow[Q\in F^{n\times n}]{1_U} & (U,F) & \xrightarrow[A\in F^{m\times n}]{\mathcal{A}} & (V,F) & \xrightarrow[P\in F^{m\times m}]{1_V} & (V,F) \\ \{\tilde{u}_j\}_{j=1}^n & \longrightarrow & \{u_j\}_{j=1}^n & \longrightarrow & \{v_i\}_{i=1}^m & \longrightarrow & \{\tilde{v}_i\}_{i=1}^m. \end{array}$$

By Theorem 2.6 the representation of the linear map

$$\mathcal{A} = 1_V \circ \mathcal{A} \circ 1_U : (U,F) \to (V,F)$$

with respect to the bases $\{\tilde{u}_j\}_{j=1}^n$ and $\{\tilde{v}_i\}_{i=1}^m$ is

$$\tilde{A} = P \cdot A \cdot Q \in F^{m\times n}$$

where A is the representation of $\mathcal{A}$ with respect to the bases $\{u_j\}_{j=1}^n$ and $\{v_i\}_{i=1}^m$ and $\cdot$ denotes ordinary matrix multiplication. Since $P \in F^{m\times m}$ and $Q \in F^{n\times n}$ are invertible it is also true that

$$A = P^{-1} \cdot \tilde{A} \cdot Q^{-1}.$$

Matrices that are related to each other in this way will play a central role in subsequent calculations. We therefore give them a special name.

Definition 2.17 *Two matrices $A \in F^{m\times n}$ and $\tilde{A} \in F^{m\times n}$ are equivalent if and only if there exist $Q \in F^{n\times n}$, $P \in F^{m\times m}$ both invertible such that $\tilde{A} = P \cdot A \cdot Q$.*

The discussion above leads immediately to the following conclusion.

Theorem 2.7 *Two matrices are equivalent if and only if they are representations of the same linear map.*

Proof: The "if" part follows from the discussion above. For the "only if" part use the matrices to define linear maps from F^n to F^m. ∎

As a special case, consider the situation where $(U,F) = (V,F)$

$$\begin{array}{ccc} (U,F) & \xrightarrow{\mathcal{A}} & (U,F) \\ \{u_j\}_{j=1}^n & \xrightarrow[A\in F^{n\times n}]{} & \{u_j\}_{j=1}^n \\ \{\tilde{u}_j\}_{j=1}^n & \xrightarrow[\tilde{A}\in F^{n\times n}]{} & \{\tilde{u}_j\}_{j=1}^n \end{array}$$

As before

$$\begin{array}{ccccccc} (U,F) & \xrightarrow[Q\in F^{n\times n}]{1_U} & (U,F) & \xrightarrow[A\in F^{n\times n}]{\mathcal{A}} & (U,F) & \xrightarrow[Q^{-1}\in F^{n\times n}]{1_U} & (U,F) \\ \{\tilde{u}_j\}_{j=1}^n & \longrightarrow & \{u_j\}_{j=1}^n & \longrightarrow & \{u_j\}_{j=1}^n & \longrightarrow & \{\tilde{u}_j\}_{j=1}^n \end{array}$$

therefore

$$\tilde{A} = Q^{-1} \cdot A \cdot Q.$$

The last equation is usually referred to as the *change of basis* formula; the matrix Q is variously referred to as a *similarity transformation*, a *change of basis*, a *coordinate transformation*, or a *change of coordinates*. It can be seen that many of the matrix operations from linear algebra involve such changes of basis: Elementary row/column operations, echelon forms, diagonalization, rotations, and many others.

We shall see that many of the operations/properties of interest in linear system theory are unaffected by change of basis in the state input or output space: Controllability, observability, stability, transfer function, etc. This is theoretically expected, since these are properties of the system in question and not the way we choose to represent it through a choice of basis. It is also very useful in practice: Since we are effectively free when choosing the basis in which to represent our system, we will often use specific bases that make the calculations easier.

Problems for chapter 2

Problem 2.1 (Groups)

1. Consider an arbitrary set S. Show that the set of bijective functions $f : S \to S$ forms a group under the operation of function composition. What does this group correspond to in the case where S is finite?

2. Draw an equilateral triangle in $\mathbb{R}^2$, say with vertices at $(1, 0)$, $(-\frac{1}{2}, \frac{\sqrt{3}}{2})$ and $(-\frac{1}{2}, \frac{-\sqrt{3}}{2})$. Consider the set of symmetries of this triangle, i.e. the set of all functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ that map the triangle onto itself. Show that this set forms a group under the operation of function composition. How many elements does this group contain? What is the identity element? Repeat for the case of a square and a circle.

Problem 2.2 (Rings and Fields)

1. Let $(R, +, \cdot)$ be a ring and consider elements $\alpha, \beta \in R$. Show that $\alpha \cdot 0 = 0 \cdot \alpha = 0$ and $(-\alpha) \cdot \beta = \alpha \cdot (-\beta) = -(\alpha \cdot \beta)$ where as usual $0 \in R$ denotes the addition identity and $-*$ the addition inverse of element $* \in R$.

2. Let $(R, +, \cdot)$ be a ring. Show that there exists a unique element $0 \in R$ such that for all $\alpha \in R$, $\alpha + 0 = 0 + \alpha = \alpha$. Moreover, show that for all $\alpha \in R$ there exists a unique $(-\alpha) \in R$ such that $\alpha + (-\alpha) = (-\alpha) + \alpha = 0$.

3. Let $(F, +, \cdot)$ be a field and consider $\alpha, \beta, \gamma \in F$. Show that

$$\forall \alpha \neq 0, \quad \alpha \cdot \beta = \alpha \cdot \gamma \Leftrightarrow \beta = \gamma.$$

   Is the same true for a ring? Justify your answer.

Problem 2.3 (Identity and inverse properties) Let $(V, F, \oplus, \odot)$ be a linear space and 0 be the addition identity element of F. Show that for all $x \in V$, $0 \odot x = \theta$. Moreover, show that for all $a \in F$, $x \in V$, $(-a) \odot x = \ominus(a \odot x) = a \odot (\ominus x)$.

Problem 2.4 (Examples of linear spaces) Let $(F, +, \cdot)$ be a field.

1. Consider $x = (\xi_1, \ldots, \xi_n) \in F^n$ and $y = (\eta_1, \ldots, \eta_n) \in F^n$ and $\alpha \in F$. Define $\oplus : F^n \times F^n \to F^n$ and $\odot : F \times F^n \to F^n$ by

$$\begin{aligned} x \oplus y &= (\xi_1 + \eta_1, \ldots, \xi_n + \eta_n) \\ \alpha \odot x &= (\alpha \cdot \xi_1, \ldots, \alpha \cdot \xi_n). \end{aligned}$$

   Show that $(F^n, F, \oplus, \odot)$ is a linear space. What is the addition identity? What is the addition inverse?

2. Let now $(V, F, +, \cdot)$ be a linear space and D an arbitrary set. Let $\mathcal{F}(D, V)$ denote the set of all functions $f : D \to V$. For $f, g \in \mathcal{F}(D, V)$ and $\alpha \in F$ define $f \oplus g : D \to V$ and $\alpha \odot f : D \to V$ by

$$\begin{aligned} (f \oplus g)(d) &= f(d) + g(d) \\ (\alpha \odot f)(d) &= \alpha \cdot f(d) \end{aligned}$$

   for all $d \in D$. Show that $(\mathcal{F}(D, V), F, \oplus, \odot)$ is a linear space. What is the addition identity? What is the addition inverse.

Problem 2.5 (Subspaces) 1. Let U and V be linear spaces and let $\mathcal{L}(U,V)$ denote the set of linear functions $\mathcal{A}: U \to V$. Show that $\mathcal{L}(U,V)$ is a linear subspace of $\mathcal{F}(U,V)$, the space of all functions mapping U into V with the usual operations of function addition and scalar multiplication.

2. Let U and V be linear spaces and $\mathcal{A}: U \to V$ be a linear function. Show that $\textsc{Range}(\mathcal{A})$ is a subspace of V and $\textsc{Null}(\mathcal{A})$ is a subspace of U.

3. Let $\{W_i\}_{i=1}^n$ be a finite family of subspaces of V. Show that the the intersection and the direct sum of these subspaces

$$\bigcap_{i=1}^{n} W_i = \{v \in V \mid \forall i = 1,\dots,n,\ v \in W_i\}$$
$$\bigoplus_{i=1}^{n} W_i = \{v \in V \mid \exists w_i \in W_i, i = 1,\dots n,\ v = w_1 + \dots w_n\}$$

are themselves subspaces of V.

Problem 2.6 (Basis and vector representation) Let V be a finite dimensional linear space.

1. Let W be a subspace of V. Show that W is also finite dimensional and its dimension can be no greater than that of V.

2. Show that the representation of a given $x \in V$ with respect to a basis $\{b_1,\dots,b_n\}$ is unique.

Problem 2.7 (Rank and nullity) Let $(F,+,\cdot)$ be a field and consider the linear maps $\mathcal{A}: (F^n,F) \to (F^m,F)$ and $\mathcal{B}: (F^p,F) \to (F^n,F)$ represented by matrices $A \in F^{m\times n}$ and $B \in F^{p\times n}$ respectively. Show that:

1. $0 \le \textsc{Rank}(A) \le \min\{n,m\}$ and $\textsc{Rank}(A) + \textsc{Nullity}(A) = n$.

2. $\textsc{Rank}(A) + \textsc{Rank}(B) - n \le \textsc{Rank}(BA) \le \min\{\textsc{Rank}(A), \textsc{Rank}(B)\}$.

   (Hint: Let $\mathcal{A}': \textsc{Range}(\mathcal{B}) \to F^m$ be the restriction of $\mathcal{A}$ to $\textsc{Range}(\mathcal{B})$. Then:

   (a) $\textsc{Range}(\mathcal{A}\circ\mathcal{B}) = \textsc{Range}(\mathcal{A}') \subseteq \textsc{Range}(\mathcal{A})$,

   (b) $\textsc{Null}(\mathcal{A}') \subseteq \textsc{Null}(\mathcal{A})$.

   To show part 2 apply the result from part 1 to $\mathcal{A}'$.)

Problem 2.8 (Invertible matrices) Let F be a field, $A \in F^{n\times n}$ be a matrix, and $\mathcal{A}: F^n \to F^n$ the linear map defined by $\mathcal{A}(x) = Ax$ for all $x \in F^n$. Show that the following statements are equivalent:

1. A is invertible.

2. None of the eigenvalues of A is equal to zero.

3. $\mathcal{A}$ is bijective.

4. $\mathcal{A}$ in injective.

5. $\mathcal{A}$ is surjective.

6. $\textsc{Rank}(A) = n$.

7. $\textsc{Nullity}(A) = 0$.

8. The columns $a_j = (a_{1j}, \ldots, a_{nj}) \in F^n$ form a linearly independent set $\{a_j\}_{j=1}^n$.
9. The rows $a'_i = (a_{i1}, \ldots, a_{in}) \in F^n$ form a linearly independent set $\{a'_i\}_{i=1}^n$.

Problem 2.9 (Matrix representation properties)

1. Consider linear maps $\mathcal{A} : (V, F) \to (W, F)$ and $\mathcal{B} : (U, F) \to (V, F)$. Assume that U, V, W have finite dimensions m, n, p respectively, and that $\mathcal{A}$ and $\mathcal{B}$ have representations $A \in F^{p\times n}$ and $B \in F^{n\times m}$ with respect to given bases for the three spaces. Show that the composition $\mathcal{C} = \mathcal{A} \circ \mathcal{B} : (U, F) \to (W, F)$ has representation $C = AB$ with respect to the same bases.
2. Consider a linear map $\mathcal{A} : (U, F) \to (U, F)$ where U has finite dimension n. Assume that $\mathcal{A}$ has representation $A \in F^{n\times n}$ with respect to a given basis for U. Show that if $\mathcal{A}$ is invertible, then $\mathcal{A}^{-1}$ has representation A^{-1} with respect to the same basis.

Problem 2.10 (Matrix representation examples)

1. Consider a linear map $\mathcal{A} : (U, F) \to (U, F)$ where U has finite dimension n. Assume there exists a vector $b \in U$ such that the collection $\{b, \mathcal{A}(b), \mathcal{A} \circ \mathcal{A}(b), \ldots, \mathcal{A}^{n-1}(b)\}$ forms a basis for U. Derive the representation of $\mathcal{A}$ and b with respect to this basis.
2. Consider a linear map $\mathcal{A} : (U, F) \to (U, F)$ where U has finite dimension n. Assume there exists a basis b_i, $i = 1, \ldots, n$ for U such that $\mathcal{A}(b_n) = \lambda b_n$ and $\mathcal{A}(b_i) = \lambda b_i + b_{i+1}$, $i = 1, \ldots, n-1$. Derive the representation of $\mathcal{A}$ with respect to this basis.
3. Consider two matrices $A, \tilde{A} \in \mathbb{R}^{n\times n}$ related by a similarity transformation; i.e. there exists $Q \in \mathbb{R}^{n\times n}$ invertible such that $\tilde{A} = Q^{-1}AQ$. Show that $\textsc{Spec}[A] = \textsc{Spec}[\tilde{A}]$.

Problem 2.11 (Matrix eigenvalues) Let F be a field, $A \in F^{n\times n}$ be a matrix, and $\mathcal{A} : F^n \to F^n$ be the linear map defined by $\mathcal{A}(x) = Ax$ for all $x \in F^n$. The following statements are equivalent:

1. $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathcal{A}$.
2. $\textsc{Det}(\lambda I - A) = 0$.
3. There exists $v \in \mathbb{C}^n$ such that $v \neq 0$ and $Av = \lambda v$.
4. There exists $\eta \in \mathbb{C}^n$ such that $\eta \neq 0$ and $\eta^T A = \lambda \eta^T$.

Problem 2.12 (Linear function spaces) Show that linear functions $\mathcal{A} : U \to V$ between two linear spaces (U, F) and (V, F) form a linear space over the field F under the usual operations of function addition and scalar multiplication.

Chapter 3

Introduction to Analysis

Consider a linear space (V, F) and assume the $F = \mathbb{R}$ or $F = \mathbb{C}$ so that for $a \in F$ the absolute value (or modulus) $|a|$ is well defined.

3.1 Norms and continuity

Definition 3.1 *A norm on a linear space (V, F) is a function $\|\cdot\| : V \to \mathbb{R}_+$ such that:*

1. $\forall v_1, v_2 \in V$, $\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$ *(triangle inequality).*
2. $\forall v \in V$, $\forall a \in F$, $\|av\| = |a| \cdot \|v\|$.
3. $\|v\| = 0 \Leftrightarrow v = 0$.

A linear space equipped with such a norm is called a normed linear space and is denoted by $(V, F, \|\cdot\|)$.

$v = 0$ in the last line refers of course to the zero vector in V. A norm provides a notion of "length" for an element of a linear space. The norm can also be used to define a notion of "distance" between elements of a linear space; one can think of $\|v_1 - v_2\|$ as the distance between two vectors $v_1, v_2 \in V$.

Example (Normed spaces) In $(\mathbb{R}^n, \mathbb{R})$, the following are examples of norms:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \text{(1-norm)}$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \quad \text{(Euclidean or 2-norm)}$$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \quad \text{for } p \geq 1, \quad \text{(p-norm)}$$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \quad \text{(infinity norm)}$$

Exercise 3.1 Show that $\|x\|_1$, $\|x\|_2$, and $\|x\|_\infty$ satisfy the axioms of a norm.

■

In fact the same definitions would also hold for the linear space $(\mathbb{C}^n, \mathbb{C})$. Different norms of course lead to different notions of distance for the same linear space. One way to visualise these different distance notions is through the so-called open balls.

Definition 3.2 *If $(V, F, \|\cdot\|)$ is a normed linear space, the (open) ball of radius $r \in \mathbb{R}_+$ centered at $v \in V$ is the set*

$$B(v, r) = \{v' \in V \mid \|v - v'\| < r\}.$$

The ball $B(0, 1)$ is called the unit ball of $(V, F, \|\cdot\|)$.

Useful properties of the ball in a given norm are provided in Problem 3.2.

Exercise 3.2 Draw the unit balls of the normed linear spaces $(\mathbb{R}^2, \mathbb{R}, \|\cdot\|_1)$, $(\mathbb{R}^2, \mathbb{R}, \|\cdot\|_2)$, and $(\mathbb{R}^2, \mathbb{R}, \|\cdot\|_\infty)$.

Definition 3.3 *A set $S \subseteq V$ is called bounded if $S \subseteq B(0, r)$ for some r.*

For example, any ball $B(v, r)$ is bounded, because $B(v, r) \subseteq B(0, \|v\| + r + 1)$.

Definition 3.4 *Let $(V, F, \|\cdot\|)$ be a normed space. A function $v : \mathbb{N} \to V$ is called a sequence in V. The sequence converges to a point $\overline{v} \in V$ if and only if*

$$\forall \epsilon > 0 \; \exists N \in \mathbb{N} \; m \geq N \Rightarrow \|v(m) - \overline{v}\| < \epsilon.$$

In this case $\overline{v}$ is called the limit of the sequence.

To denote sequences, we will mostly use the notation $\{v_i\}_{i=0}^{\infty}$, where $i \in \mathbb{N}$ plays the role of the index variable. To indicate that a sequence converges to some vector $v \in V$ we write $v_i \overset{i \to \infty}{\longrightarrow} \overline{v}$ or $\lim_{i \to \infty} v_i = \overline{v}$. Note that by the definition of the norm, the statement "the sequence $\{v_i\}_{i=0}^{\infty}$ converges to $\overline{v} \in V$" is equivalent to $\lim_{i \to \infty} \|v_i - \overline{v}\| = 0$ in the usual sense in $\mathbb{R}$.

Using the definition one can also define open and closed sets.

Definition 3.5 *Let $(V, F, \|\cdot\|)$ be a normed space. A set $K \subseteq V$ is called closed if and only if it contains all its limit points; i.e. if and only if for all sequences $\{v_i\}_{i=1}^{\infty} \subseteq K$ if $v_i \to v \in V$ then $v \in K$. K is called open if and only if its complement $V \setminus K$ is closed. A set that is both closed and bounded is called compact.*

Some useful properties of open and closed sets are stated in Problem 3.3.

Consider now functions between two normed spaces

$$f : (U, F, \|\cdot\|_U) \to (V, F, \|\cdot\|_V)$$

Definition 3.6 *f is called continuous at $u \in U$ if and only if*

$$\forall \epsilon > 0 \; \exists \delta > 0 \text{ such that } \|u - u'\|_U < \delta \Rightarrow \|f(u) - f(u')\|_V < \epsilon.$$

The function is called continuous on U if and only if it is continuous at every $u \in U$.

In other words, f is continuous at u if and only if

$$\forall \epsilon > 0 \; \exists \delta > 0 \text{ such that } u' \in B(u, \delta) \Rightarrow f(u') \in B(f(u), \epsilon).$$

or, with the usual notation, $\lim_{u' \to u} f(u') = f(u)$. Continuity expresses the requirement that small changes in $u \in U$ should lead to small changes in $f(u) \in V$. Useful properties of continuous functions are stated in Problem 3.4.

Exercise 3.3 Show that the set of continuous functions $f : (U, F, \|\cdot\|_U) \to (V, F, \|\cdot\|_V)$ is a linear subspace of $\mathcal{F}(U, V)$.

Fact 3.1 *The norm $\|\cdot\|$, as a function between the normed linear spaces $(V, F, \|\cdot\|)$ and $(\mathbb{R}, \mathbb{R}, |\cdot|)$, is continuous on V.*

Proof: The triangular inequality implies $\|x\| = \|x - x_0 + x_0\| \leq \|x - x_0\| + \|x_0\|$, and therefore $\|x\| - \|x_0\| \leq \|x - x_0\|$. Interchanging the role of x and x_0, we also get $|\|x\| - \|x_0\|| \leq \|x - x_0\|$. Now in the definition of continuity of $\|\cdot\|$ at x_0:

$$\forall \epsilon > 0 \exists \delta > 0 \; \|x - x_0\| < \delta \Rightarrow |\|x\| - \|x_0\|| < \epsilon$$

it sufficies to choose $\delta = \epsilon$ to obtain $|\|x\| - \|x_0\|| \leq \|x - x_0\| < \delta = \epsilon$. ∎

Continuity of functions has an intimate relationship with the convergence of sequences and hence with open and closed sets. This relationship is highlighted in Problem 3.4.

3.2 Equivalent norms

Note that in the definitions of convergence and continuity we have left open the choice of the norm. One may wonder if, by choosing another norm, a a convergent sequence may cease to converge, or a continuous function can cease to be continuous. To investigate this question we need the notion of equivalent norms.

Definition 3.7 *Consider a linear space (V, F) with two norms, $\|\cdot\|_a$ and $\|\cdot\|_b$. The two norms are equivalent if and only if*

$$\exists m_u \geq m_l > 0, \forall v \in V \;\; m_l \|v\|_a \leq \|v\|_b \leq m_u \|v\|_a.$$

Exercise 3.4 Show that in this case it is also true that there exist $m'_l, m'_u \in \mathbb{R}_+$ such that $\forall v \in V$, $m'_l \|v\|_b \leq \|v\|_a \leq m'_u \|v\|_b$. Show further that if $\|\cdot\|_a$ is equivalent to $\|\cdot\|_b$ and $\|\cdot\|_b$ is equivalent to $\|\cdot\|_c$, then $\|\cdot\|_a$ is equivalent to $\|\cdot\|_c$

Example (Equivalent norms) Consider $x \in \mathbb{R}^n$ and the $\|x\|_1$ and $\|x\|_\infty$ norms defined above. Then

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \leq \sum_{i=1}^n |x_i| = \|x\|_1$$
$$\|x\|_1 = \sum_{i=1}^n |x_i| \leq \sum_{i=1}^n \left(\max_{j=1,\ldots,n} |x_j| \right) = n \max_{j=1,\ldots,n} |x_i| = n \|x\|_\infty.$$

Therefore $\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty$ and the two norms are equivalent (take $m_l = 1$ and $m_u = n$). ∎

Fact 3.2 *Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two equivalent norms on a linear space (V, F) with $F = \mathbb{R}$ or $F = \mathbb{C}$. A sequence $\{v_i\}_{i=0}^{\infty} \subseteq V$ converges to some $v \in V$ in $(V, F, \|\cdot\|_a)$ if and only if it converges to v in $(V, F, \|\cdot\|_b)$.*

Proof: Suppose that the sequence $\{v_i\}_{i=0}^{\infty}$ converges to $v \in V$ with respect to the norm $\|\cdot\|_a$, that is for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|v_m - v\|_a < \epsilon$ for all $m \geq N$. Due to equivalence of the norms, there exists $m_u > 0$ such that $m_l\|v\|_a \leq \|v\|_b \leq m_u\|v\|_a$ for all v. Fix an arbitrary $\epsilon > 0$. Then there exists $N \in \mathbb{N}$ such that, for all $m \geq N$, $\|v_m - v\|_a < \frac{\epsilon}{m_u}$; But then, with the same N, for all $m \geq N$

$$\|v_m - v\|_b \leq m_u\|v_m - v\|_a < m_u\frac{\epsilon}{m_u} = \epsilon.$$

■

Since function continuity, open and closed sets were all defined in terms of convergence of sequences, this fact further implies that open/closed sets defined using one norm remain open/closed for any other equivalent norm. Likewise, continuous functions defined with respect to a pair of norms remain continuous with respect to any other pair of respectively equivalent norms.

The fact that $\|x\|_1$ and $\|x\|_\infty$ are equivalent norms on $\mathbb{R}^n$ is not a coincidence. A remarkable result states, indeed, that any two norms on a finite-dimensional space are equivalent. To show this, we will use the following fact, which is indeed a corollary of two fundamental theorems in real analysis.

Fact 3.3 *(A corollary to the Weierstrass Theorem). A continuous function $f : S \to \mathbb{R}$ defined on a subset $S \subseteq \mathbb{R}^n$ that is compact in $(\mathbb{R}^n, \|\cdot\|_2)$ attains a minimum on S.*

In other words, if the function $f : S \to \mathbb{R}$ is continuous and the set S is compact there exist $x_m \in S$ and a real number m such that $f(x_m) = m \leq f(x)$ for all $x \in S$, or

$$m = \inf_{x\in S} f(x) = \min_{x\in S} f(x) = f(x_m) > -\infty.$$

The proof of this fact can be found in [16]. Recall that the infimum $\inf_{x\in S} f(x)$ is the greatest lower bound of f on S, i.e. the largest number $m \in \mathbb{R}$ such that $f(x) \geq m$ for all $x \in S$; likewise, the supremum $\sup_{x\in S} f(x)$ is the least upper bound of f on S, i.e. the smallest number $M \in \mathbb{R}$ such that $f(x) \leq M$ for all $x \in S$. Fact 3.3 states that if the function f is continuous and the set S is compact the infimum (and by adding a minus sign also the supremum) is finite and attained for some $x_m \in S$; in this case the infimum coincides with the minimum $\min_{x\in S} f(x)$ of the function (and the supremum with the maximum $\max_{x\in S} f(x)$). This is not the case in general of course. For example the function $f : \mathbb{R}_+ \to \mathbb{R}$ defined by $f(x) = e^{-x}$ is continuous and defined over the closed but unbounded (hence not compact) set $[0, \infty)$. The maximum and supremum of the function coincide $1 = \sup_{x\in\mathbb{R}_+} f(x) = \max_{x\in\mathbb{R}_+} f(x) = f(0)$. The infimum of the function, on the other hand, is $0 = \inf_{x\in\mathbb{R}_+} f(x)$, but is not attained for any $x \in [0, \infty)$; hence the minimum is not defined. Likewise, the function

$$f(x) = \begin{cases} -1 - x & \text{if } x \in [-1, 0) \\ 1 - x & \text{if } x \in [0, 1] \end{cases}$$

is defined over the compact set $S = [-1, 1]$ but us discontinuous at 0. Again $\sup_{x\in[-1,1]} f(x) = \max_{x\in[-1,1]} f(x) = f(0) = 1$ but $\inf_{x\in[-1,1]} f(x) = -1$ is not attained for any $x \in [-1, 1]$ and $\min_{x\in[-1,1]} f(x)$ is undefined. Finally, for the function $f(x) = -1/x$ on $(0, 1]$ the infimum is not a finite number, since the function tends to $-\infty$ as x tends to 0; this is precisely the situation we need to exclude in the proof of Proposition 3.1 below.

For completeness we also recall the following fact.

Fact 3.4 *(Cauchy Inequality). For $a_i \in \mathbb{R}$ $b_i \in \mathbb{R}$, $i = 1, \ldots, n$,*

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

The proof is left as an exercise.

Theorem 3.1 *Any two norms on a finite-dimensional space V are equivalent.*

Proof: For simplicity we assume that $F = \mathbb{R}$; the proof for $F = \mathbb{C}$ is similar, e.g. by identifying $\mathbb{C}$ with $\mathbb{R}^2$. Assume V is finite dimensional of dimension n and let $\{v_i\}_{i=1}^n$ be a basis. For an arbitrary element $x \in V$ consider the representation $\xi \in \mathbb{R}^n$, i.e. $x = \sum_{i=1}^n \xi_i v_i$ and define

$$\|x\|_a = \sqrt{\sum_{i=1}^n |\xi_i|^2}$$

One can show that $\|\cdot\|_a : V \to \mathbb{R}_+$ is indeed a norm on V (along the lines of Exercise 3.1). By Exercise 3.4 it suffices to show that an arbitrary norm $\|\cdot\|_b$ is equivalent to $\|\cdot\|_a$, i.e. there exist $m_u > m_l > 0$ such that for all $x \in V$, $m_l\|x\|_a \le \|x\|_b \le m_u\|x\|_a$.

By the norm axioms for $\|\cdot\|_b$ and Fact 3.4

$$\|x\|_b = \left\|\sum_{i=1}^n \xi_i v_i\right\|_b \le \sum_{i=1}^n |\xi_i| \cdot \|v_i\|_b \le \sqrt{\sum_{i=1}^n |\xi_i|^2}\sqrt{\sum_{i=1}^n \|v_i\|_b^2} \le m_u\|x\|_a.$$

where we have set $m_u = \sqrt{\sum_{i=1}^n \|v_i\|_b^2}$.

Consider now the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(\alpha) = \left\|\sum_{i=1}^n \alpha_i v_i\right\|_b$$

for $\alpha \in \mathbb{R}^n$. We show that f is continuous as a function from $(\mathbb{R}^n, \mathbb{R}, \|\cdot\|_2)$ to $(\mathbb{R}, \mathbb{R}, |\cdot|)$. Indeed, given two elements $\alpha, \alpha' \in \mathbb{R}^n$

$$\begin{aligned}
|f(\alpha) - f(\alpha')| &= \left| \left\|\sum_{i=1}^n \alpha_i v_i\right\|_b - \left\|\sum_{i=1}^n \alpha_i' v_i\right\|_b \right| \le \left\|\sum_{i=1}^n (\alpha_i - \alpha_i') v_i\right\|_b \quad \text{see proof of Fact 3.1} \\
&\le \sum_{i=1}^n |\alpha_i - \alpha_i'| \, \|v_i\|_b \text{ by the properties of the norm } \|\cdot\|_b \\
&\le \sqrt{\sum_{i=1}^n |\alpha_i - \alpha_i'|^2}\sqrt{\sum_{i=1}^n \|v_i\|_b^2} \text{ by Fact 3.4} \\
&= m_u\|\alpha - \alpha'\|.
\end{aligned}$$

Therefore, for any $\alpha \in \mathbb{R}^n$ and any $\epsilon > 0$ if we select $\delta = \epsilon/m_u$ then for all $\alpha' \in \mathbb{R}^n$ such that $\|\alpha - \alpha'\|_2 < \delta$ it is true that $|f(\alpha) - f(\alpha')| < \epsilon$; hence f is continuous.

Finally, consider the set $S = \{\alpha \in \mathbb{R}^n \mid \sum_{i=1}^n \alpha_i^2 = 1\} \subseteq \mathbb{R}^n$. Clearly $\|\alpha\|_2 \le 1$ for all $\alpha \in S$, hence S is bounded in $(\mathbb{R}^n, \mathbb{R}, \|\cdot\|_2)$. Moreover, S is also closed in $(\mathbb{R}^n, \mathbb{R}, \|\cdot\|_2)$ as it is the inverse image of the closed set $\{1\}$ under the continuous (by Fact 3.1) function $\alpha \mapsto \sqrt{\sum_{i=1}^n \alpha^2}$ mapping $(\mathbb{R}^n, \mathbb{R}, \|\cdot\|_2)$ into $(\mathbb{R}, \mathbb{R}, |\cdot|)$ (see Problem 3.4). Hence, S is compact in $(\mathbb{R}^n, \mathbb{R}, \|\cdot\|_2)$ and the continuous function $f : S \to \mathbb{R}$ attains a minimum m for some $\alpha^* \in S$. Then for any $x \in V$ with a

representation $\xi \in \mathbb{R}^n$ with respect to the basis v_i,

$$\begin{aligned}\|x\|_b &= \left\|\sum_{i=1}^n \xi_i v_i\right\|_b = \sqrt{\sum_{i=1}^n |\xi_i|^2}\ \left\|\sum_{i=1}^n \frac{\xi_i}{\sqrt{\sum_{i=1}^n |\xi_i|^2}} v_i\right\| \\ &= \sqrt{\sum_{i=1}^n |\xi_i|^2}\ \left\|\sum_{i=1}^n \alpha_i v_i\right\| \text{ where we set } \alpha_i = \frac{\xi_i}{\sqrt{\sum_{i=1}^n |\xi_i|^2}} \\ &= \|x\|_a f(\alpha) \geq m\|x\|_a\end{aligned}$$

The last inequality follows since by construction $\sum_{i=1}^n \alpha_i^2 = 1$, hence $\alpha \in S$ and f is lower bounded by m. Setting $m_l = m$ completes the proof. ∎

3.3 Infinite-dimensional normed spaces

Consider now real numbers $t_0 \leq t_1$ and the linear space $C([t_0, t_1], \mathbb{R}^n)$ of continuous functions $f : [t_0, t_1] \to \mathbb{R}^n$. For each $t \in [t_0, t_1]$, $f(t) \in \mathbb{R}^n$; consider its standard 2-norm $\|f(t)\|_2$ as a vector in $\mathbb{R}^n$ and define the function $\|\cdot\|_\infty : C([t_0, t_1], \mathbb{R}^n) \to \mathbb{R}_+$ by

$$\|f\|_\infty = \max_{t \in [t_0, t_1]} \|f(t)\|_2.$$

Note that, by Fact 3.3, the maximum is indeed attained, that is there exists some $t^* \in [t_0, t_1]$ such that $\|f\|_\infty = \|f(t^*)\|_2$. More generally, e.g. if the functions can be discontinuous or the domain is not compact, one can use the supremum instead of the maximum in the definition of $\|f\|_\infty$.

One can show that the function $\|f\|_\infty$ defined in this way is a norm on $(C([t_0, t_1], \mathbb{R}^n), \mathbb{R})$. Indeed, for all continuous functions f, $\|f\|_\infty$ is greater than or equal to zero and it is trivially equal to zero if and only if f is the zero function (i.e. $f(t) = 0 \in' Re^n$ for all $t \in [t_0, t_1]$). Moreover,

$$\begin{aligned}\|\alpha f\|_\infty &= \max_{t \in [t_0, t_1]} \|\alpha f(t)\|_2 = |\alpha| \max_{t \in [t_0, t_1]} \|f(t)\|_2 = |\alpha|\ \|f\|_\infty, \\ \|f + g\|_\infty &= \max_{t \in [t_0, t_1]} \|f(t) + g(t)\|_2 \leq \max_{t \in [t_0, t_1]} \|f(t)\|_2 + \max_{t \in [t_0, t_1]} \|g(t)\|_2 = \|f\|_\infty + \|g\|_\infty.\end{aligned}$$

Norms on function spaces can also be defined using integration. For example, for $f \in C([t_0, t_1], \mathbb{R}^n)$ we can define the following:

$$\begin{aligned}\|f\|_1 &= \int_{t_0}^{t_1} \|f(t)\|_2\ dt \\ \|f\|_2 &= \sqrt{\int_{t_0}^{t_1} \|f(t)\|_2^2\ dt} \\ \|f\|_p &= \left(\int_{t_0}^{t_1} \|f(t)\|_2^p\ dt\right)^{\frac{1}{p}}\end{aligned}$$

Here, the integral takes the role of the finite summation in the definition of the corresponding norms on finite-dimensional spaces. Since the 2-norm that appears in the integrands is a continuous function (Fact3.1), the integrands are also continuous functions of the variable t, and all these quantities are well-defined. Moreover, all of them are norms. Take for example the first one. Of course we have $\|f\|_1 \geq 0$ for all $f \in C([t_0, t_1], \mathbb{R}^n)$; the norm of the zero-function is of course equal to zero; on the other hand, given a continuous function f, if f is nonzero at a point it must be nonzero in a whole

subinterval of $[t_0, t_1]$, therefore $\|f\|_1$ is also nonzero. (Summing up, $\|f\|_1 = 0$ if and only if $f = 0$.) Moreover,

$$\|\alpha f\|_1 = \int_{t_0}^{t_1} \|\alpha f(t)\|_2 \, dt = |\alpha| \int_{t_0}^{t_1} \|f(t)\|_2 \, dt = |\alpha| \, \|f\|_1$$
$$\|f + g\|_1 = \int_{t_0}^{t_1} \|f(t) + g(t)\|_2 \, dt \leq \int_{t_0}^{t_1} (\|f(t)\|_2 + \|g(t)\|_2) \, dt = \|f\|_1 + \|g\|_1$$

It is also easy to see that the use of the 2-norm in the integrands (in the definitions of $\|f\|_1$, $\|f\|_2$, $\|f\|_p$) and in the maximum (in the definition of $\|f\|_\infty$) is arbitrary: Since norms in $\mathbb{R}^n$ are all equivalent one could have used any other norm (say the 1- or the ∞-norm), leading to other norms on $C([t_0, t_1], \mathbb{R}^n)$.

Exercise 3.5 Show that the function norms obtained by replacing $\|f(t)\|_2$ by another norm on $\mathbb{R}^n$ in the integrands and the maximum are equivalent to the ones defined above.

Motivated by this fact and in analogy to finite-dimensional spaces, the reader may be tempted to think that the all the norms introduced above are equivalent. Unfortunately, this is not the case.

Example (Non-equivalent norms) Consider the functions $f \in C([0, 1], \mathbb{R})$ and the two norms

$$\|f\|_\infty = \max_{t \in [0,1]} |f(t)| \text{ and } \|f\|_1 = \int_0^1 |f(t)| dt$$

We will show that the two norms are not equivalent. Assume, for the sake of contradiction, that there exist $m_l, m_u \in \mathbb{R}_+$ such that for all $f \in C([0, 1], \mathbb{R})$ we have $m_l \|f\|_1 \leq \|f\|_\infty \leq m_u \|f\|_1$. Since

$$\|f\|_1 = \int_0^1 |f(t)| \, dt \leq \int_0^1 \|f\|_\infty \, dt = \|f\|_\infty,$$

one can indeed choose $m_l = 1$. On the other hand, it is impossible to find m_u that satisfies the other inequality for all f. Assume, for the sake of contradiction, that such an m_u exists and consider the family of functions $f_n(t) = t^n \in C([0, 1], \mathbb{R})$ for $n \in \mathbb{N}$. Clearly $\|f_n\|_\infty = 1$ and $\|f_n\|_1 = \frac{1}{n+1}$. Therefore $\|f_a\|_\infty = (n + 1)\|f_a\|_1$. Selecting $n + 1 > m_u$ leads to a contradiction, showing that the two norms are not equivalent. ■

3.4 Completeness

In our treatment of the solution of differential equations below we will be faced with the task of establishing the convergence of a series in a certain linear space. In principle, one should be able to do so using Definition 3.4. In practice, however, this is not always possible as the limit may not be known. From the definition it is, however, apparent that if the sequence $\{v_i\}_{i=0}^\infty$ converges to some $v \in V$ the v_i need to be getting closer and closer to v, and therefore closer and closer to each other. Can we decide whether the sequence converges just by looking at $\|v_i - v_j\|$?

Definition 3.8 *A sequence $\{v_i\}_{i=0}^\infty$ is called a Cauchy sequence if and only if*

$$\forall \epsilon > 0 \; \exists N \in \mathbb{N} \; \forall m \geq N, \;\; \|v_m - v_N\| < \epsilon.$$

Exercise 3.6 Show that a sequence is Cauchy if and only if for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|v_m - v_n\| < \epsilon$ for all $m, n \geq N$. Show further that every convergent sequence is Cauchy.

The converse however is not always true: there may exist Cauchy sequences that do not converge to a point in the linear space of interest. Linear spaces for which all Cauchy sequences converge to a point in the space are known as Banach (or complete) spaces.

Definition 3.9 *The normed space $(V, F, \|\cdot\|)$ is called complete (or Banach) if and only if every Cauchy sequence converges.*

Fortunately many of the spaces we are interested in are complete.

Fact 3.5 *The linear space $(\mathbb{R}, \mathbb{R}, |\cdot|)$ is a Banach space.*

This observation is in fact the reason why the real number system was introduced in the first place, and is so fundamental that in many expositions of analysis it is taken as an axiom. It is also the main tool used to establish completeness of many other spaces.

Theorem 3.2 *Let $F = \mathbb{R}$ or $F = \mathbb{C}$ and assume that (V, F) is finite-dimensional. Then $(V, F, \|\cdot\|)$ is a Banach space for any norm $\|\cdot\|$.*

Proof: Suppose that $F = \mathbb{R}$. Let $\{b_1, \cdots, b_d\}$ be a basis of V, so that every vector $v \in V$ can be represented in an unique way as $v = \sum_{k=1}^{d} x_k b_k$ for some $x_k \in \mathbb{R}$. On $\mathbb{R}^d$ let the function $\|\cdot\|_*$ be defined by

$$\|x\|_* = \|(x_1, \cdots, x_d)\|_* = \left\| \sum_{k=1}^{d} x_k b_k \right\|$$

$\|\cdot\|_*$ is a norm on $\mathbb{R}^d$; moreover, $\{v_n\}_{n=0}^{\infty}$ is Cauchy or convergent in $(V, F, \|\cdot\|)$ if and only if the corresponding sequence of d-tuples $\{x^n\}_{n=0}^{\infty}$ is Cauchy or convergent in $(\mathbb{R}^d, \mathbb{R}, \|\cdot\|_*)$ (justify these assertions). Thus, it suffices to show that $(\mathbb{R}^d, \mathbb{R}, \|\cdot\|_*)$ is complete.

Indeed, since $\mathbb{R}^d$ is finite-dimensional, $\|\cdot\|_*$ is equivalent to $\|\cdot\|_\infty$, and Cauchy or convergent sequences with respect to one norm are Cauchy or convergent also with respect to the other. Suppose therefore that $\{x^n\}_{n=0}^{\infty}$ is a Cauchy sequence. This implies that

$$\forall \epsilon > 0 \ \exists N \in \mathbb{N} \ \forall m, n \geq N, \ \|x^n - x^m\|_\infty < \epsilon. \tag{3.1}$$

But then, we have the same property for each of the k-th components:

$$\forall \epsilon > 0 \ \exists N \in \mathbb{N} \ \forall m, n \geq N, \ |x_k^n - x_k^m| < \epsilon.$$

In other words, the k-th components $\{x_k^n\}_{n=0}^{\infty}$ form a Cauchy sequence in $\mathbb{R}$. Since $\mathbb{R}$ is complete, this sequence converges. Define $\bar{x} = (\bar{x}_1, \cdots, \bar{x}_k, \cdots, \bar{x}_d)$, where

$$\bar{x}_k = \lim_{n \to \infty} x_k^n$$

Finally, in 3.1, fix N and n, and let $m \to \infty$:

$$\forall \epsilon > 0 \ \exists N \in \mathbb{N} \ \forall n \geq N, \ \|x^n - \bar{x}\|_\infty < \epsilon.$$

(we can take the limit within $\|\cdot\|_\infty$ because the norm is continuous). Therefore, $\{x^n\}_{n=0}^{\infty}$ is convergent, and the same happens for $\{v_n\}_{n=0}^{\infty}$; since the latter is arbitrary, (V, F) is Banach.

To establish the same result for complex spaces ($F = \mathbb{C}$) one can repeat the same proof after showing that $\mathbb{C}$ itself is complete. This should now be easy, and is left as an exercise. ∎

Corollary 3.1 *Any finite-dimensional subspace W of a linear space $(V, F, \|\cdot\|)$ is a closed subset of V.*

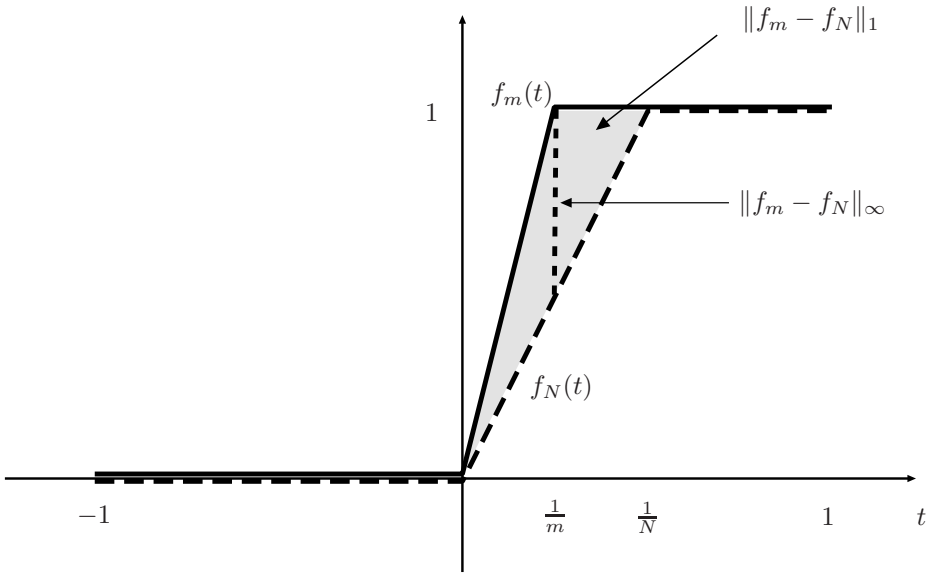

Figure 3.1: Examples of functions in non-Banach space.

Proof: According to Definition 3.5, W is closed if for all sequences $\{w_i\}_{i=1}^{\infty} \subseteq W$, if $w_i \to w \in V$, then $w \in W$. Now if an arbitrary sequence $\{w_i\}_{i=1}^{\infty} \subseteq W$ is convergent in V, then it is Cauchy both in V and in W. Since W is Banach, its limit point w belongs to W. ∎

This fact will come in handy in Chapter 7 in the context of inner product spaces. The state, input and output spaces of our linear systems will be of the form $\mathbb{R}^n$, and will all be Banach spaces. One may be tempted to think that this is more generally true. Unfortunately infinite-dimensional spaces are less well-behaved.

Example (Non-Banach space) Consider the normed space $(C([-1,1],\mathbb{R}),\mathbb{R},\|\cdot\|_1)$ of continuous functions $f:[-1,1]\to\mathbb{R}$ with the $1-$norm

$$\|f\|_1 = \int_{-1}^{1} |f(t)|dt.$$

For $i = 1, 2, \ldots$ consider the sequence of functions (Figure 3.1)

$$f_i(t) = \begin{cases} 0 & \text{if } t < 0 \\ it & \text{if } 0 \le t < 1/i \\ 1 & \text{if } t \ge 1/i. \end{cases} \tag{3.2}$$

It is easy to see that the sequence is Cauchy. Indeed, if we take $N, m \in \{1, 2, \ldots\}$ with $m \ge N$,

$$\begin{aligned} \|f_m - f_N\|_1 &= \int_{-1}^{1} |f_m(t) - f_N(t)|dt = \int_0^{1/m} (mt - Nt)dt + \int_{1/m}^{1/N} (1 - Nt)dt \\ &= (m-N)\left[\frac{t^2}{2}\right]_0^{1/m} + \left[1 - N\frac{t^2}{2}\right]_{1/m}^{1/N} = \frac{m-N}{2mN} \le \frac{1}{2N}. \end{aligned}$$

Therefore, given any $\epsilon > 0$ by selecting $N > 1/(2\epsilon)$ we can ensure that for all $m \ge N$, $\|f_m - f_N\|_1 < \epsilon$. One can guess that the sequence $\{f_i\}_{i=1}^{\infty}$ converges to the function

$$f(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \ge 0. \end{cases}$$

Indeed

$$\|f_i - f\|_1 = \int_0^{1/i} (1 - it)dt = \frac{1}{2i} \to 0.$$

The problem is that the limit function $f \in \mathcal{F}([-1,1],\mathbb{R})$ is discontinuous and therefore $f \notin C([-1,1],\mathbb{R})$. Hence $C([-1,1],\mathbb{R})$ is not a Banach space since the Cauchy sequence $\{f_i\}_{i=1}^n$ does not converge to an element of $C([-1,1],\mathbb{R})$. ∎

Fortunately there are several infinite dimensional linear spaces that are Banach. The important one for this chapter is the space of continuous functions under the infinity norm.

Theorem 3.3 *$(C([t_0,t_1],\mathbb{R}^n),\mathbb{R},\|\cdot\|_\infty)$ is a Banach space.*

Proof: We articulate the proof in three steps: First, given a sequence of continuous functions which is Cauchy in the norm $\|\cdot\|_\infty$, we define a "pointwise limit" function; second, we prove that the sequence converges to this function; third, we prove that this function is continuous.

Let $\{f_n\}_{n=1}^\infty$ be a Cauchy sequence of continuous functions $f : [t_0,t_1] \to \mathbb{R}^n$. For each $t \in [t_0,t_1]$, $\{f_n(t)\}_{n=1}^\infty$ is a Cauchy sequence of vectors in $\mathbb{R}^n$ (why?). Since $\mathbb{R}^n$ is a Banach space (Theorem 3.2), every such sequence has a limit. We define the function f as follows:

$$f(t) = \lim_{n\to\infty} f_n(t)$$

Next, we show that $\{f_n\}_{n=1}^\infty$ converges to f also with respect to the norm $\|\cdot\|_\infty$. Indeed, the fact that $\{f_n\}_{n=1}^\infty$ is Cauchy means that, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$\forall n, m \geq N, \forall t \in [t_0,t_1],\ \|f_n(t) - f_m(t)\| \leq \|f_n - f_m\|_\infty < \epsilon$$

In this equation, fix t and n, and let $m \to \infty$:

$$\forall n \geq N, \forall t \in [t_0,t_1],\ \|f_n(t) - f(t)\| \leq \epsilon$$

Taking the supremum over t,

$$\forall n \geq N, \|f_n - f\|_\infty \leq \epsilon$$

hence $\|f_n - f\|_\infty \to 0$.

It remains to show that f is continuous. Let $\bar{t} \in [t_0,t_1]$. It holds:

$$\begin{aligned}\|f(t) - f(\bar{t})\| &= \|f(t) - f_n(t) + f_n(t) - f_n(\bar{t}) + f_n(\bar{t}) - f(\bar{t})\| \\ &\leq \|f(t) - f_n(t)\| + \|f_n(t) - f_n(\bar{t})\| + \|f_n(\bar{t}) - f(\bar{t})\|\end{aligned}$$

Now fix an $\epsilon > 0$. Since $\{f_n\}_{n=1}^\infty$ converges to f with respect to $\|\cdot\|_\infty$, there exists n such that $\|f - f_n\|_\infty < \epsilon/3$, and therefore $\|f(t) - f_n(t)\| < \epsilon/3$ for all t and in particular $\|f_n(\bar{t}) - f(\bar{t})\| < \epsilon/3$. On the other hand, since each function of the sequence is continuous, and in particular so is f_n, there exists $\delta > 0$ such that $\|f_n(t) - f_n(\bar{t})\| < \epsilon/3$ whenever $|t - \bar{t}| < \delta$. Thus, for all $\epsilon > 0$ there exists (n and) $\delta > 0$ such that, if $|t - \bar{t}| < \delta$, then

$$\|f(t) - f(\bar{t})\| \leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon,$$

and f is continuous.

Summarizing, given an arbitrary Cauchy sequence $\{f_n\}_{n=1}^\infty$ in $C([t_0,t_1],\mathbb{R},\mathbb{R}^n)$, we can construct a function $f \in C([t_0,t_1],\mathbb{R},\mathbb{R}^n)$ such that $\|f_n - f\|_\infty \to 0$. ∎

Exercise 3.7 Show that the sequence $\{f_i\}_{i=1}^\infty$ defined in (3.2) is not Cauchy in $(C([-1,1],\mathbb{R}),\mathbb{R},\|\cdot\|_\infty)$.

The fact that $(C([t_0,t_1],\mathbb{R}^n),\mathbb{R},\|\cdot\|_\infty)$ is a Banach space will be exploited below for the proof of existence of solutions for ordinary differential equations. In Chapter 7 we will encounter another infinite dimensional linear space, the so-called space of square integrable functions, which will play a central role in the discussion of controllability and observability.

3.5 Induced norms and matrix norms

Consider now the space $(\mathbb{R}^{m\times n}, \mathbb{R})$.

Exercise 3.8 Show that $(\mathbb{R}^{m\times n}, \mathbb{R})$ is a linear space with the usual operations of matrix addition and scalar multiplication. (Hint: One way to do this is to identify $\mathbb{R}^{m\times n}$ with $\mathbb{R}^{nm}$.)

The following are examples of norms on $(\mathbb{R}^{m\times n}, \mathbb{R})$:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}| \quad \text{(cf. 1 norm in } \mathbb{R}^{nm})$$

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2\right)^{\frac{1}{2}} \quad \text{(Frobenius norm, cf. 2 norm in } \mathbb{R}^{nm})$$

$$\max_{i=1,\dots,m}\max_{j=1,\dots,n}|a_{ij}| \quad \text{(cf. } \infty \text{ norm in } \mathbb{R}^{nm})$$

More commonly used are the norms derived when one considers matrices as linear maps between linear spaces. We start with a more general definition.

Definition 3.10 *Consider the linear space of continuous functions $f : (U, F, \|\cdot\|_U) \to (V, F, \|\cdot\|_V)$ between two normed spaces. The induced norm of f is defined as*

$$\|f\| = \sup_{u\neq 0}\frac{\|f(u)\|_V}{\|u\|_U}.$$

One can check that, whenever the supremum is finite, it indeed defines a norm on the space of continuous functions (Problem 3.6). Notice that the induced norm depends not only on the function, but also on the norms imposed on the two spaces.

For continuous linear functions between normed space, the definition of the induced norm simplifies somewhat. It turns out that it is not necessary to take the supremum over all non-zero vectors in U; it suffices to consider vectors with norm equal to 1.

Fact 3.6 *Consider two normed spaces $(U, F, \|\cdot\|_U)$ and $(V, F, \|\cdot\|_V)$ and a continuous linear function $\mathcal{A} : U \to V$. Then*

$$\|\mathcal{A}\| = \sup_{\|u\|_U=1}\|\mathcal{A}(u)\|_V.$$

The proof is left as an exercise (Problem 3.6).

Example (Induced matrix norms) Consider $A \in F^{m\times n}$ and consider the linear map $\mathcal{A} : F^n \to F^m$ defined by $\mathcal{A}(x) = A \cdot x$ for all $x \in F^n$. By considering different norms on F^m and F^n different induced norms for the linear map (and hence the matrix) can be defined:

$$\|A\|_p = \sup_{x\in F^n}\frac{\|Ax\|_p}{\|x\|_p}$$

$$\|A\|_1 = \max_{j=1,\dots,n}\sum_{i=1}^{m}|a_{ij}| \quad \text{(maximum column sum)}$$

$$\|A\|_2 = \max_{\lambda\in \text{SPEC}[A^TA]}\sqrt{\lambda} \quad \text{(maximum singular value)}$$

$$\|A\|_\infty = \max_{i=1,\dots,m}\sum_{j=1}^{n}|a_{ij}| \quad \text{(maximum row sum)}$$

■

It turns out that the induced norm of linear maps is intimately related to their continuity.

Theorem 3.4 *Consider two normed spaces $(U, F, \|\cdot\|_U)$ and $(V, F, \|\cdot\|_V)$ and a linear function $\mathcal{A} : U \to V$. The following statements are equivalent*

1. *$\mathcal{A}$ is continuous.*
2. *$\mathcal{A}$ is continuous at 0.*
3. *$\sup_{\|u\|_U=1} \|\mathcal{A}(u)\|_V < \infty$ and the induced norm $\|\mathcal{A}\|$ is well-defined.*

Proof: $1 \Rightarrow 2$: By definition a function that is continuous everywhere is also continuous at 0.

$2 \Rightarrow 3$: By contraposition. Assume that $\sup_{\|u\|_U=1} \|\mathcal{A}(u)\|_V$ is infinite. Then for all $M > 0$ there exists $u_M \in U$ with $\|u_M\|_U = 1$ such that $\|\mathcal{A}(u_M)\|_V \geq M$. In particular, for all $n \in \mathbb{N}$ there exists $u_n \in U$ with $\|u_n\|_U = 1$ such that $\|\mathcal{A}(u_n)\|_V \geq n$. Consider $\hat{u}_n = u_n/n \in U$. Clearly, $\|\hat{u}_n\|_U = \|u_n\|_U/n = 1/n \to 0$, hence, by the properties and continuity of the norm $\lim_{n\to\infty} \hat{u}_n = 0$. On the other hand,

$$\|\mathcal{A}(\hat{u}_n)\|_V = \|\mathcal{A}(u_n/n)\|_V = \frac{1}{n}\|\mathcal{A}(u_n)\|_V \geq \frac{n}{n} = 1$$

where the second equality follows by linearity of $\mathcal{A}$ and the norm axioms. Since $\mathcal{A}(0) = 0$ by linearity of $\mathcal{A}$,

$$\lim_{n\to\infty} \|\mathcal{A}(\hat{u}_n)\|_V \neq 0 = \mathcal{A}(0)$$

and the function cannot be continuous at 0.

$3 \Rightarrow 1$: For simplicity we revert to the original (equivalent by Fact 3.6) definition of the induced norm and assume that

$$\sup_{u\neq 0} \frac{\|\mathcal{A}(u)\|_V}{\|u\|_U} = M < \infty.$$

This implies that $\|\mathcal{A}(u)\|_V \leq M\|u\|_U$ for all $u \neq 0$; moreover, since $\mathcal{A}(0) = 0$ by linearity of $\mathcal{A}$, it the inequality also holds for $u = 0$. Therefore for all $u, u' \in U$

$$\|\mathcal{A}(u) - \mathcal{A}(u')\|_V = \|\mathcal{A}(u - u')\|_V \leq M\|u - u'\|_U.$$

Therefore, for all $u \in U$ and for all $\epsilon > 0$ there exists $\delta > 0$ such that $\|\mathcal{A}(u) - \mathcal{A}(u')\|_V < \epsilon$ for all $u' \in U$ with $\|u - u'\|_U < \delta$ (select any $0 < \delta < \epsilon/M$). Hence $\mathcal{A}$ is continuous at all $u \in U$.

In summary, $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 1$ and the three statements are equivalent. ■

An easy corollary is that linear functions between finite dimensional spaces are always continuous.

Corollary 3.2 *All linear functions $\mathcal{A} : U \to V$ between two finite dimensional normed spaces $(U, F, \|\cdot\|_U)$ and $(V, F, \|\cdot\|_V)$ are continuous.*

Proof: Fix bases for the two spaces so that the linear function $\mathcal{A}$ is represented by multiplication by a matrix A. Since both spaces are finite dimensional all norms are equivalent; for simplicity let $\|\cdot\|_U$ and $\|\cdot\|_V$ both be the corresponding infinity norms.

Note that if $A = 0$ the function $\mathcal{A}$ is constant and hence continuous. Otherwise, consider arbitrary an $u \in U$ and note that $\|\mathcal{A}(u) - \mathcal{A}(u')\|_\infty \leq \|A\|_\infty \|u - u'\|_\infty$. Clearly $\|A\|_\infty$ (the maximum row sum) is a finite number. Therefore, for all $\epsilon > 0$ there exists $\delta = \epsilon/\|A\|_\infty$ such that $\|u - u'\|_\infty < \delta$ implies that $\|\mathcal{A}(u) - \mathcal{A}(u')\|_\infty < \epsilon$. Hence the function is continuous. ■

This is of course not the case for general linear functions. The following example provides a preview of our discussion on stability of linear systems (Chapter 6).

Example (Discontinuous linear functions) Consider the function $\mathcal{A}$ defined by

$$\begin{array}{rcl} \mathcal{A} : (\mathbb{R}^n, \mathbb{R}, \|\cdot\|_2) & \longrightarrow & (C([0,\infty), \mathbb{R}^n), \mathbb{R}, \|\cdot\|_\infty) \\ x_0 & \longmapsto & \mathcal{A}(x_0) = e^t x_0 \end{array}$$

Note that $\mathcal{A}$ a finite dimensional space ($\mathbb{R}^n$) into an infinite dimensional space $C([0,\infty), \mathbb{R}^n)$ and the the 2-norm is used in the finite dimensional space whereas the infinity norm

$$\sup_{t\in[0,\infty)} \|f(t)\|_2 \text{ for } f(\cdot) \in C([0,\infty), \mathbb{R}^n)$$

is used on the infinite dimensional one. It is easy to check that the function $\mathcal{A}$ is indeed linear. Moreover, $\mathcal{A}(0)$ is the zero function hence

$$\|\mathcal{A}(0)\|_\infty = 0.$$

For any $x_0 \neq 0$, however, $\mathcal{A}(x_0) = e^t x_0$ is unbounded and $\|\mathcal{A}(x_0)\|_\infty$ is infinite. Hence the function is discontinuous and its induced norm is indeed not defined. This will be the situation we encounter for unstable linear systems in Chapter 6.

By contrast, the function

$$\begin{array}{rcl} \mathcal{A}' : (\mathbb{R}^n, \mathbb{R}, \|\cdot\|_2) & \longrightarrow & (C([0,\infty), \mathbb{R}^n), \mathbb{R}, \|\cdot\|_\infty) \\ x_0 & \longmapsto & \mathcal{A}(x_0) = e^{-t} x_0 \end{array}$$

is also linear. Moreover, for any x_0,

$$\|\mathcal{A}(x_0)\|_\infty = \sup_{t\in[0,\infty)} \|e^{-t}x_0\|_2 = \|x_0\|_2$$

hence the function is continuous at 0. Its induced norm is well defined

$$\|\mathcal{A}\| = \sup_{\|x_0\|_2=1} \|\mathcal{A}(x_0)\|_\infty = \sup_{\|x_0\|_2=1} \|x_0\|_2 = 1.$$

This will be the situation we encounter for stable linear systems in Chapter 6. ■

For future reference we summarise some of the basic properties of the induced norm in a theorem. These properties are the main reason why the induced norm is used extensively for linear functions between normed spaces (instead of the more obvious matrix norms outlined in the beginning of the section).

Theorem 3.5 *Consider continuous linear functions $\mathcal{A}, \tilde{\mathcal{A}} : (V, F, \|\cdot\|_V) \to (W, F, \|\cdot\|_W)$ and $\mathcal{B} : (U, F, \|\cdot\|_U) \to (V, F, \|\cdot\|_V)$ between normed linear spaces and let $\|\cdot\|$ denote the corresponding induced norms.*

1. *For all $v \in V$, $\|\mathcal{A}(v)\|_W \leq \|\mathcal{A}\| \cdot \|v\|_V$.*
2. *For all $a \in F$, $\|a\mathcal{A}\| = |a| \cdot \|\mathcal{A}\|$.*
3. *$\|\mathcal{A} + \tilde{\mathcal{A}}\| \leq \|\mathcal{A}\| + \|\tilde{\mathcal{A}}\|$.*
4. *$\|\mathcal{A}\| = 0 \Leftrightarrow \mathcal{A}(v) = 0$ for all $v \in V$ (zero map).*
5. *$\|\mathcal{A} \circ \mathcal{B}\| \leq \|\mathcal{A}\| \cdot \|\mathcal{B}\|$.*

Proof: With the exception of the last statement, all others are immediate consequences of the definition of the induced norm and the fact that it is indeed a norm (Problem 3.6). To show the last statement, consider

$$\begin{aligned}\|\mathcal{A}\circ\mathcal{B}\| &= \sup_{\|u\|_U=1}\|(\mathcal{A}\circ\mathcal{B})(u)\|_W = \sup_{\|u\|_U=1}\|\mathcal{A}(\mathcal{B}(u))\|_W\\ &\leq \sup_{\|u\|_U=1}\|\mathcal{A}\|\cdot\|\mathcal{B}(u)\|_V \text{ (by the first statement)}\\ &=\|\mathcal{A}\|\cdot\sup_{\|u\|_U=1}\|\mathcal{B}(u)\|_V = \|\mathcal{A}\|\cdot\|\mathcal{B}\|.\end{aligned}$$

■

3.6 Ordinary differential equations

The main topic of these notes are dynamical systems of the form

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{3.3}$$
$$y(t) = C(t)x(t) + D(t)u(t) \tag{3.4}$$

where

$$t\in\mathbb{R},\ x(t)\in\mathbb{R}^n,\ u(t)\in\mathbb{R}^m,\ y(t)\in\mathbb{R}^p$$
$$A(t)\in\mathbb{R}^{n\times n},\ B(t)\in\mathbb{R}^{n\times m},\ C(t)\in\mathbb{R}^{p\times n},\ D(t)\in\mathbb{R}^{p\times m}$$

The difficult part conceptually is equation (3.3), a linear ordinary differential equation (ODE) with time varying coefficients $A(t)$ and $B(t)$. Equation (3.3) is a special case of the (generally non-linear) ODE

$$\dot{x}(t) = f(x(t),u(t),t) \tag{3.5}$$

with

$$t\in\mathbb{R},\ x(t)\in\mathbb{R}^n,\ u(t)\in\mathbb{R}^m$$
$$f:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\longrightarrow\mathbb{R}^n.$$

The only difference is that for linear systems the function $f(x,u,t)$ is a linear function of x and u for all $t\in\mathbb{R}$.

In this section we are interested in finding "solutions" (also known as "trajectories", "flows", ...) of the ODE (3.5). In other words:

- Given:

$$f:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\to\mathbb{R}^n \quad \text{dynamics}$$
$$(t_0,x_0)\in\mathbb{R}\times\mathbb{R}^n \quad \text{"initial" condition}$$
$$u(\cdot):\mathbb{R}\to\mathbb{R}^m \quad \text{input trajectory}$$

- Find:

$$x(\cdot):\mathbb{R}\to\mathbb{R}^n \quad \text{state trajectory}$$

- Such that:

$$x(t_0) = x_0$$
$$\frac{d}{dt}x(t) = f(x(t),u(t),t)\ \ \forall t\in\mathbb{R}$$

While this definition is acceptable mathematically it tends to be too restrictive in practice. The problem is that according to this definition $x(t)$ should not only be continuous as a function of time, but also differentiable; since the definition makes use of the derivative $dx(t)/dt$, it implicitly assumes that the derivative is well defined. This will in general disallow input trajectories $u(t)$ which are discontinuous as a function of time. We could in principle only allow continuous inputs (in which case the above definition of a solution should be sufficient) but unfortunately many interesting input functions turn out to be discontinuous.

Example (Hasty driver) Consider a car moving on a road. Let $y \in \mathbb{R}$ denote the position of the car with respect to a fixed point on the road, $v \in \mathbb{R}$ denote the velocity of the car and $a \in \mathbb{R}$ denote its acceleration. We can then write a "state space" model for the car by defining

$$x = \begin{bmatrix} y \\ v \end{bmatrix} \in \mathbb{R}^2, \; u = a \in \mathbb{R}$$

and observing that

$$\dot{x}(t) = \begin{bmatrix} \dot{y}(t) \\ \dot{v}(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ u(t) \end{bmatrix}.$$

Defining

$$f(x,u) = \begin{bmatrix} x_2 \\ u \end{bmatrix}$$

the dynamics of our car can then be described by the (linear, time invariant) ODE

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

Assume now that we would like to select the input $u(t)$ to take the car from the initial state

$$x(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

to the terminal state

$$x(T) = \begin{bmatrix} y_F \\ 0 \end{bmatrix}$$

in the shortest time possible (T) while respecting constraints on the speed and acceleration

$$v(t) \in [0, V_{max}] \text{ and } a(t) \in [a_{min}, a_{max}] \;\; \forall t \in [0,T]$$

for some $V_{max} > 0$, $a_{min} < 0 < a_{max}$. It turns out that the optimal solution for $u(t)$ is discontinuous. Assuming that y_F is large enough it involves three phases:

1. Accelerate as much as possible ($u(t) = a_{max}$) until the maximum speed is reached ($x_2(t) = V_{max}$).
2. "Coast" ($u(t) = 0$) until just before reaching y_F.
3. Decelerate as much as possible ($u(t) = a_{min}$) and stop exactly at y_F.

Unfortunately this optimal solution is not allowed by our definition of solution given above. ■

To make the definition of solution more relevant in practice we would like to allow discontinuous $u(t)$, albeit ones that are not too "wild". Measurable functions provide the best notion of how "wild" input functions are allowed to be and still give rise to reasonable solutions for differential equations. Unfortunately the proper definition of measurable functions requires some exposure to measure theory and is beyond the scope of these notes; the interested reader is referred to [18] for a treatment of dynamical systems from this perspective. For our purposes the following, somewhat simpler definition will suffice.

Definition 3.11 *A function $u : \mathbb{R} \to \mathbb{R}^m$ is piecewise continuous if and only if it is continuous at all $t \in \mathbb{R}$ except those in a set of discontinuity points $D \subseteq \mathbb{R}$ that satisfy:*

1. *$\forall \tau \in D$ left and right limits of u exist, i.e. $\lim_{t\to\tau^+} u(t)$ and $\lim_{t\to\tau^-} u(t)$ exist and are finite. Moreover, $u(\tau) = \lim_{t\to\tau^+} u(t)$.*
2. *$\forall t_0, t_1 \in \mathbb{R}$ with $t_0 < t_1$, $D \cap [t_0, t_1]$ contains a finite number of points.*

The symbols $\lim_{t\to\tau^+} u(t)$ and $\lim_{t\to\tau^-} u(t)$ indicate the limit of $u(t)$ as t approaches τ from above ($t \geq \tau$) and from below ($t \leq \tau$). We will use the symbol $PC([t_0, t_1], \mathbb{R}^m)$ to denote the linear space of piecewise continuous functions $f : [t_0, t_1] \to \mathbb{R}^m$ (and similarly for $PC(\mathbb{R}, \mathbb{R}^m)$).

Exercise 3.9 Show that $(PC([t_0, t_1], \mathbb{R}^m), \mathbb{R})$ is a linear space under the usual operations of function addition and scalar multiplication.

Definition 3.11 includes all continuous functions, the solution to the hasty driver example, square waves, etc. Functions that are not included are things like $1/t$ or $\tan(t)$ (that go to infinity for some $t \in \mathbb{R}$), and obscure constructs like

$$u(t) = \begin{cases} 0 & (t \geq 1/2) \vee (t \leq 0) \\ -1 & t \in \left[\frac{1}{2k+1}, \frac{1}{2k}\right) \\ 1 & t \in \left[\frac{1}{2(k+1)}, \frac{1}{2k+1}\right) \end{cases} \quad k = 1, 2, \ldots$$

that have an infinite number of discontinuity points in the finite interval $[0, 1/2]$.

Let us now return to our differential equation (3.5). Let us generalize the picture somewhat by considering

$$\dot{x}(t) = p(x(t), t) \tag{3.6}$$

where we have obscured the presence of $u(t)$ by defining $p(x(t), t) = f(x(t), u(t), t)$. We will impose the following assumption of p.

Assumption 3.1 *The function $p : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is piecewise continuous in its second argument, i.e. there exists a set of discontinuity points $D \subseteq \mathbb{R}$ such that for all $x \in \mathbb{R}^n$*

1. *$p(x, \cdot) : \mathbb{R} \to \mathbb{R}^n$ is continuous for all $t \in \mathbb{R} \setminus D$.*
2. *For all $\tau \in D$, $\lim_{t\to\tau^+} p(x, t)$ and $\lim_{t\to\tau^-} p(x, t)$ exist and are finite and $p(x, \tau) = \lim_{t\to\tau^+} p(x, t)$.*
3. *$\forall t_0, t_1 \in \mathbb{R}$ with $t_0 < t_1$, $D \cap [t_0, t_1]$ contains a finite number of points.*

Exercise 3.10 Show that if $u : \mathbb{R} \to \mathbb{R}^m$ is piecewise continuous (according to Definition 3.11) and $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^n$ is continuous then $p(x, t) = f(x, u(t), t)$ satisfies the conditions of Assumption 3.1.

We are now in a position to provide a formal definition of the solution of ODE (3.6).

Definition 3.12 *Consider a function $p : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ satisfying the conditions of Assumption 3.1. A continuous function $\phi : \mathbb{R} \to \mathbb{R}^n$ is called a solution of (3.6) passing through $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$ if and only if*

1. *$\phi(t_0) = x_0$, and*
2. *$\forall t \in \mathbb{R} \setminus D$, ϕ is differentiable at t, and $\frac{d}{dt}\phi(t) = p(\phi(t), t)$.*

The solution only needs to be continuous and differentiable "almost everywhere", i.e. everywhere except at the set of discontinuity points, D, of p. This definition allows us to include input trajectories that are discontinuous as functions of time, but still reasonably well behaved.

Is this a good definition? For example, is it certain that there exists a function satisfying the conditions of the definition? And if such a function does exist, is it unique, or can there be other functions that also satisfy the same conditions? These are clearly important questions in control theory where the differential equations usually represent models of physical systems (airplanes, cars, circuits etc.). In this case the physical system being modeled is bound to do something as time progresses. If the differential equation used to model the system does not have solutions or has many, this is an indication that the model we developed is inadequate, since it cannot tell us what this "something" is going to be.

The answer to these questions very much depends on the function p. Notice that in all subsequent examples the function $p(x,t)$ is independent of t; therefore all potential problems have noting to do with the fact that p is piecewise continuous as a function of time.

Example (No solutions) Consider the one dimensional system

$$\dot{x}(t) = -\operatorname{sgn}(x(t)) = \begin{cases} -1 & \text{if } x(t) \geq 0 \\ 1 & \text{if } x(t) < 0, \end{cases}$$

For the initial condition $(t_0, x_0) = (0, c)$ with $c > 0$, Definition 3.12 allows us to define the solution $\phi(t) = c - t$ for $t \leq c$. There does not exist a function, however, that satisfies Definition 3.12 for $t > c$.

Exercise 3.11 Verify this.

In particular, for the initial condition $(t_0, x_0) = (0, 0)$ the solution is undefined for all $t > 0$. ■

The problem here is that the function $p(x,t)$ is discontinuous in x. The example suggests that to be able to use Definition 3.12 effectively we need to exclude $p(x,t)$ that are discontinuous in x. Another alternative would be to relax the Definition 3.12 further to allow the so-called Filippov solutions [8, 20], but this is beyond the scope of these notes.

Is the existence of a solution guaranteed if we restrict our attention to $p(x,t)$ that are continuous in x? Unfortunately not!

Example (Finite Escape Time) Consider the one dimensional system

$$\dot{x}(t) = x(t)^2$$

with $(t_0, x_0) = (0, c)$ for some $c > 0$. The function

$$x(t) = \frac{c}{1 - tc}$$

is a solution of this differential equation.

Exercise 3.12 Verify this.

Notice that as $t \to 1/c$ the solution escapes to infinity; no solution is defined for $t \geq 1/c$. As in the previous example, Definition 3.12 allows us to define a solution until a certain point in time, but no further. There is a subtle difference between the two examples, however. In the latter case the solution can always be extended for a little extra time; it is defined over the right-open interval $(-\infty, 1/c)$. In the former case, on the other hand, a "dead end" is reached in finite time; for initial conditions $(0, c)$ the solution is only defined over the closed interval $(-\infty, c]$. ■

The problem here is that $p(x,t)$ grows too fast as a function of x. To use Definition 3.12 we will therefore need to exclude functions that grow too fast.

Even if these "exclusions" work and we manage to ensure that a solution according to Definition 3.12 exists, is this solution guaranteed to be unique, or can there be many functions $\phi(t)$ satisfying the conditions of the definition for the same (t_0, x_0)? Unfortunately the answer is "yes".

Example (Multiple Solutions) Consider the one dimensional system

$$\dot{x}(t) = 2|x(t)|^{\frac{1}{2}} \operatorname{sgn}(x(t))$$

with $(t_0, x_0) = (0,0)$. For all $a \geq 0$ the function

$$x(t) = \begin{cases} \pm(t-a)^2 & t \geq a \\ 0 & t < a \end{cases}$$

is a solution of the differential equation.

Exercise 3.13 Verify this.

Notice that in this case the solution is not unique. In fact there are infinitely many solutions, one for each $a \geq 0$. ■

The problem here is that $p(x,t)$ is "too steep", since its slope goes to infinity as x tends to 0. To use Definition 3.12 we will therefore need to also exclude functions that are too steep.

Functions that are discontinuous, "grow too fast" or are "too steep" can all be excluded at once by the following definition.

Definition 3.13 *The function $p : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is globally Lipschitz in x if and only if there exists a piecewise continuous function $k : \mathbb{R} \to \mathbb{R}_+$ such that*

$$\forall x, x' \in \mathbb{R}^n, \forall t \in \mathbb{R} \ \ \|p(x,t) - p(x',t)\| \leq k(t)\|x - x'\|.$$

$k(t)$ is called the Lipschitz constant of p at $t \in \mathbb{R}$.

Example (Lipschitz functions) One can easily verify that linear functions are Lipschitz; we will do so when we introduce linear systems. All differentiable functions with bounded derivatives are also Lipschitz. However, not all Lipschitz functions are differentiable. For example, the absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}$ is Lipschitz (with Lipschitz constant 1) but not differentiable at $x = 0$. All Lipschitz functions are continuous, but not all continuous functions are Lipschitz. For example, the functions x^2 and $\sqrt{|x|}$ from $\mathbb{R}$ to $\mathbb{R}$ are both continuous, but neither is Lipschitz. x^2 is locally Lipschitz, i.e. for all $x_0 \in \mathbb{R}$, there exist $\epsilon > 0$ and $k_{x_0}(\cdot) : \mathbb{R} \to \mathbb{R}_+$ piecewise continuous such that for all $x \in \mathbb{R}$:

$$\text{if } \|x - x_0\| < \epsilon \text{ then } \|p(x,t) - p(x_0,t)\| \leq k_{x_0}(t)\|x - x_0\|$$

But there is no $k(\cdot) : \mathbb{R} \to \mathbb{R}_+$ that will work for all x_0. $\sqrt{x}$ is not even locally Lipschitz, a finite $k_{x_0}(\cdot) : \mathbb{R} \to \mathbb{R}_+$ does not exist for $x_0 = 0$. ■

In the following section we will show that global Lipschitz continuity is sufficient to ensure the existence and uniqueness of solutions for Global Lipshitz continuity is indeed a tight sufficient condition for existence and uniqueness of solutions, albeit not a necessary one. The examples outlined earlier in this section demonstrate that there exist differential equation defined by non-Lipschitz functions that do not posess unique solutions. On the other hand, one can also find differential equation

defined by non-Lipschitz functions that do posess unique solutions for all intial conditions; to establish this fact, however, more work is needed on a case by case basis. The existence and uniqueness results discussed here can be further fine tuned (assuming for example local Lipschitz continuity to derive local existence of solutions). In the Chapter 4, however, we will show that linear differential equations, the main topic of these notes, always satisfy the global Lipschitz assumption. We will therefore not pursue such refinements here, instead we refer the interested reader to [12, 17].

3.7 Existence and uniqueness of solutions

We are now in a position to state and prove a fundamental fact about the solutions of ordinary differential equations.

Theorem 3.6 (Existence and uniqueness) *Assume $p : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is piecewise continuous with respect to its second argument (with discontinuity set $D \subseteq \mathbb{R}$) and globally Lipschitz with respect to its first argument. Then for all $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$ there exists a unique continuous function $\phi : \mathbb{R} \to \mathbb{R}^n$ such that:*

1. $\phi(t_0) = x_0$.
2. $\forall t \in \mathbb{R} \setminus D$, $\frac{d}{dt}\phi(t) = p(\phi(t), t)$.

The proof of this theorem is rather sophisticated. We will build it up in three steps:

1. Background lemmas.
2. Proof of existence (construction of a solution).
3. Proof of uniqueness.

3.7.1 Background lemmas

For a function $f : [t_0, t_1] \to \mathbb{R}^n$ with $f(t) = (f_1(t), \dots, f_n(t))$ define

$$\int_{t_0}^{t_1} f(t)dt = \begin{bmatrix} \int_{t_0}^{t_1} f_1(t)dt \\ \vdots \\ \int_{t_0}^{t_1} f_n(t)dt \end{bmatrix}.$$

Fact 3.7 *Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$. Then for all $t_0, t_1 \in \mathbb{R}$,*

$$\left\| \int_{t_0}^{t_1} f(t)dt \right\| \leq \left| \int_{t_0}^{t_1} \|f(t)\| \, dt \right|.$$

Proof:(Sketch) Roughly speaking one can approximate the integral by a sum, use triangle inequality on the sum and take a limit. Note that the absolute value on the right had side is necessary, since the integral there will be negative if $t_1 < t_0$. ∎

Recall that $m! = 1 \cdot 2 \cdot \ldots \cdot m$ denotes the factorial of a natural number $m \in \mathbb{N}$.

Fact 3.8 *The following hold:*

1. $\forall m, k \in \mathbb{N}$, $(m + k)! \geq m! \cdot k!$.

2. $\forall c \in \mathbb{R}$, $\lim_{m\to\infty} \frac{c^m}{m!} = 0$.

Proof: Part 1 is easy to prove by induction (Theorem 1.2). For part 2, the case $|c| < 1$ is trivial, since $c^m \to 0$ while $m! \to \infty$. For $c > 1$, take $M > 2c$ integer. Then for $m > M$

$$\frac{c^m}{m!} = \frac{c^M c^{m-M}}{M!(M+1)(M+2)\dots m} \leq \frac{c^M}{M!} \cdot \frac{c^{m-M}}{(2c)^{m-M}} = \frac{c^M}{M!} \cdot \frac{1}{2^{m-M}} \to 0.$$

The proof for $c < -1$ is similar. ∎

Theorem 3.7 (Fundamental theorem of calculus) *Let $g : \mathbb{R} \to \mathbb{R}$ piecewise continuous with discontinuity set $D \subseteq \mathbb{R}$. Then for all $t_0 \in \mathbb{R}$ the function $f(t) = \int_{t_0}^t g(\tau)d\tau$ is continuous and for all $t \in \mathbb{R} \setminus D$,*

$$\frac{d}{dt} f(t) = g(t).$$

Theorem 3.8 (Gronwall Lemma) *Consider $u(\cdot), k(\cdot) : \mathbb{R} \to \mathbb{R}_+$ piecewise continuous, $c_1 \geq 0$, and $t_0 \in \mathbb{R}$. If for all $t \in \mathbb{R}$*

$$u(t) \leq c_1 + \left| \int_{t_0}^t k(\tau)u(\tau)d\tau \right|$$

then for all $t \in \mathbb{R}$

$$u(t) \leq c_1 \exp \left| \int_{t_0}^t k(\tau)d\tau \right|.$$

Proof: Consider $t > t_0$ (the proof for $t < t_0$ is symmetric and gives rise to the absolute values in the theorem statement). Let

$$U(t) = c_1 + \int_{t_0}^t k(\tau)u(\tau)d\tau.$$

Notice that $u(t) \leq U(t)$, since for $t \geq t_0$ the absolute value is redundant as $u(\cdot)$ and $k(\cdot)$ are non-negative. By the fundamental theorem of calculus U is continuous and wherever k and u are continuous

$$\frac{d}{dt} U(t) = k(t)u(t).$$

Then

$$\begin{aligned}
u(t) \leq U(t) &\Rightarrow u(t)k(t)e^{-\int_{t_0}^t k(\tau)d\tau} \leq U(t)k(t)e^{-\int_{t_0}^t k(\tau)d\tau} \\
&\Rightarrow \left(\frac{d}{dt}U(t)\right) e^{-\int_{t_0}^t k(\tau)d\tau} - U(t)k(t)e^{-\int_{t_0}^t k(\tau)d\tau} \leq 0 \\
&\Rightarrow \left(\frac{d}{dt}U(t)\right) e^{-\int_{t_0}^t k(\tau)d\tau} + U(t)\frac{d}{dt}\left(e^{-\int_{t_0}^t k(\tau)d\tau}\right) \leq 0 \\
&\Rightarrow \frac{d}{dt}\left(U(t)e^{-\int_{t_0}^t k(\tau)d\tau}\right) \leq 0 \\
&\Rightarrow U(t)e^{-\int_{t_0}^t k(\tau)d\tau} \text{ decreases as } t \text{ increases} \\
&\Rightarrow U(t)e^{-\int_{t_0}^t k(\tau)d\tau} \leq U(t_0)e^{-\int_{t_0}^{t_0} k(\tau)d\tau} \ \ \forall t \geq t_0 \\
&\Rightarrow U(t)e^{-\int_{t_0}^t k(\tau)d\tau} \leq U(t_0) = c_1 \ \ \forall t \geq t_0 \\
&\Rightarrow u(t) \leq U(t) \leq c_1 e^{\int_{t_0}^t k(\tau)d\tau}
\end{aligned}$$

which concludes the proof. ∎

3.7.2 Proof of existence

We will now construct a solution for the differential equation

$$\dot{x}(t) = p(x(t), t)$$

passing through $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$ using an iterative procedure. This second step of our existence-uniqueness proof will itself involve three steps:

1. Construct a sequence of functions $x_m(\cdot) : \mathbb{R} \to \mathbb{R}^n$ for $m = 1, 2, \ldots$.
2. Show that for all $t_1 \leq t_0 \leq t_2$ this sequence is a Cauchy sequence in the Banach space $C([t_1, t_2], \mathbb{R}^n), \|\cdot\|_\infty)$. Therefore the sequence converges to a limit $\phi(\cdot) \in C([t_1, t_2], \mathbb{R}^n)$.
3. Show that the limit $\phi(\cdot)$ is a solution to the differential equation.

Step 2.1: We construct a sequence of functions $x_m(\cdot) : \mathbb{R} \to \mathbb{R}^n$ for $m = 1, 2, \ldots$ by the so called Picard iteration:

1. $x_0(t) = x_0 \ \forall t \in \mathbb{R}$
2. $x_{m+1}(t) = x_0 + \int_{t_0}^{t} p(x_m(\tau), \tau) d\tau, \quad \forall t \in \mathbb{R}.$

The generated sequence of functions is known at the Picard Iteration. Notice that all the functions $x_m(\cdot)$ generated in this way are continuous by construction.

Consider any $t_1, t_2 \in \mathbb{R}$ such that $t_1 \leq t_0 \leq t_2$. Let

$$\overline{k} = \sup_{t \in [t_1, t_2]} k(t) \text{ and } T = t_2 - t_1.$$

Notice that under the conditions of the theorem both $\overline{k}$ and T are non-negative and finite. Let $\|\cdot\|$ be the infinity norm on $\mathbb{R}^n$. Then for all $t \in [t_1, t_2]$

$$\begin{aligned}
\|x_{m+1}(t) - x_m(t)\| &= \left\| x_0 + \int_{t_0}^{t} p(x_m(\tau), \tau) d\tau - x_0 - \int_{t_0}^{t} p(x_{m-1}(\tau), \tau) d\tau \right\| \\
&= \left\| \int_{t_0}^{t} [p(x_m(\tau), \tau) - p(x_{m-1}(\tau), \tau)] d\tau \right\| \\
&\leq \left| \int_{t_0}^{t} \|p(x_m(\tau), \tau) - p(x_{m-1}(\tau), \tau)\| \, d\tau \right| \quad \text{(Fact 3.7)} \\
&\leq \left| \int_{t_0}^{t} k(\tau) \|x_m(\tau) - x_{m-1}(\tau)\| \, d\tau \right| \quad (p \text{ is Lipschitz in } x) \\
&\leq \overline{k} \left| \int_{t_0}^{t} \|x_m(\tau) - x_{m-1}(\tau)\| \, d\tau \right|
\end{aligned}$$

For $m = 0$

$$\|x_1(t) - x_0(t)\| \leq \left| \int_{t_0}^{t} \|p(x_0, \tau)\| \, d\tau \right| \leq \left| \int_{t_1}^{t_2} \|p(x_0, \tau)\| \, d\tau \right| = M,$$

for some non-negative, finite number M (which of course depends on t_1 and t_2). For $m = 1$

$$\begin{aligned}
\|x_2(t) - x_1(t)\| &\leq \overline{k} \left| \int_{t_0}^{t} \|x_1(\tau) - x_0(\tau)\| \, d\tau \right| \\
&\leq \overline{k} \left| \int_{t_0}^{t} M d\tau \right| = \overline{k} M |t - t_0|.
\end{aligned}$$

For $m = 2$,

$$\begin{aligned}\|x_3(t) - x_2(t)\| &\leq \overline{k} \left| \int_{t_0}^{t} \|x_2(\tau) - x_1(\tau)\| \, d\tau \right| \\ &\leq \overline{k} \left| \int_{t_0}^{t} \overline{k} M |\tau - t_0| d\tau \right| = M \frac{\overline{k}^2 (t - t_0)^2}{2}.\end{aligned}$$

In general, for all $t \in [t_1, t_2]$,

$$\|x_{m+1}(t) - x_m(t)\| \leq M \frac{[\overline{k}|t - t_0|]^m}{m!} \leq M \frac{[\overline{k}T]^m}{m!}.$$

Step 2.2: We show that the sequence $x_m(\cdot)$ is a Cauchy sequence in the Banach space $(C([t_1, t_2], \mathbb{R}^n), \|\cdot\|_\infty)$. Note that $x_m(\cdot)$ is continuous by construction by the fundamental theorem of calculus. We therefore need to show that $x_m(\cdot)$ has bounded infinity norm for all m and that

$$\forall \epsilon > 0 \; \exists N \in \mathbb{N} \; \forall m \geq N, \|x_m(\cdot) - x_N(\cdot)\|_\infty \leq \epsilon.$$

We start with the second statement. Take $m \geq N \geq 0$ integers and consider

$$\begin{aligned}\|x_m(\cdot) - x_N(\cdot)\|_\infty &= \|x_m(\cdot) - x_{m-1}(\cdot) + x_{m-1}(\cdot) - \ldots + x_{N+1}(\cdot) - x_N(\cdot)\|_\infty \\ &= \left\| \sum_{i=0}^{m-N-1} [x_{i+N+1}(\cdot) - x_{i+N}(\cdot)] \right\|_\infty \\ &\leq \sum_{i=0}^{m-N-1} \|x_{i+N+1}(\cdot) - x_{i+N}(\cdot)\|_\infty \\ &\leq M \sum_{i=0}^{m-N-1} \frac{(\overline{k}T)^{i+N}}{(i+N)!} \quad \text{(Step 2.1)} \\ &\leq M \sum_{i=0}^{m-N-1} \frac{(\overline{k}T)^N}{N!} \frac{(\overline{k}T)^i}{i!} \quad \text{(Fact 3.8)} \\ &= M \frac{(\overline{k}T)^N}{N!} \sum_{i=0}^{m-N-1} \frac{(\overline{k}T)^i}{i!} \\ &\leq M \frac{(\overline{k}T)^N}{N!} \sum_{i=0}^{\infty} \frac{(\overline{k}T)^i}{i!} \\ &= M \frac{(\overline{k}T)^N}{N!} e^{\overline{k}T}.\end{aligned}$$

Note that, in particular, if we take $N = 0$ this implies that

$$\|x_m(\cdot) - x_0(\cdot)\|_\infty \leq M e^{\overline{k}T}.$$

This, in turn, by the triangle inequality and the fact that $x_0(t) = x_0$ for all $t \in [t_1, t_2]$ implies that $\|x_m(\cdot)\|_\infty \leq \|x_0\| + M e^{\overline{k}T} < \infty$, which establishes that $x_m(\cdot)$ has bounded infinity norm hence $x_m(\cdot) \in (C([t_1, t_2], \mathbb{R}^n), \|\cdot\|_\infty)$ as required. Moreover, by Fact 3.8

$$\lim_{N \to \infty} \frac{(\overline{k}T)^N}{N!} = 0,$$

therefore,

$$\forall \epsilon > 0 \; \exists N \in \mathbb{N} \; \forall m \geq N, \|x_m(\cdot) - x_N(\cdot)\|_\infty \leq \epsilon$$

and the sequence $\{x_m(\cdot)\}_{m=0}^{\infty}$ is Cauchy in the Banach space $(C([t_1,t_2],\mathbb{R}^n),\|\cdot\|_\infty)$. Hence it converges in the infinity norm to some continuous function $\phi(\cdot):[t_1,t_2]\to\mathbb{R}^n$ with $\|\phi(\cdot)\|_\infty<\infty$. Notice that convergence in the infinity norm implies that

$$\begin{aligned}\|x_m(\cdot)-\phi(\cdot)\|_\infty \overset{m\to\infty}{\longrightarrow} 0 &\Rightarrow \sup_{t\in[t_1,t_2]}\|x_m(t)-\phi(t)\| \overset{m\to\infty}{\longrightarrow} 0\\ &\Rightarrow \|x_m(t)-\phi(t)\| \overset{m\to\infty}{\longrightarrow} 0 \ \ \forall t\in[t_1,t_2]\\ &\Rightarrow x_m(t) \overset{m\to\infty}{\longrightarrow} \phi(t) \ \ \forall t\in[t_1,t_2]\end{aligned}$$

Step 2.3: Finally, we show that the limit function $\phi(\cdot):[t_1,t_2]\to\mathbb{R}^n$ solves the differential equation. We need to verify that:

1. $\phi(t_0)=x_0$; and
2. $\forall t\in[t_1,t_2]\setminus D$, $\frac{d}{dt}\phi(t)=p(\phi(t),t)$.

For the first part, notice that by construction $x_0(t_0)=x_0$ and for all $m>0$ since

$$x_{m+1}(t_0)=x_0+\int_{t_0}^{t_0}p(x_m(\tau),\tau)d\tau=x_0.$$

Therefore $x_m(t_0)=x_0$ for all $m\in\mathbb{N}$ and $\lim_{m\to\infty}x_m(t_0)=\phi(t_0)=x_0$.

For the second part, for $t\in[t_1,t_2]$ consider

$$\begin{aligned}\left\|\int_{t_0}^{t}[p(x_m(\tau),\tau)-p(\phi(\tau),\tau)]d\tau\right\| &\le \left|\int_{t_0}^{t}\|p(x_m(\tau),\tau)-p(\phi(\tau),\tau)\|\,d\tau\right| \quad \text{(Fact 3.7)}\\ &\le \left|\int_{t_0}^{t}k(\tau)\|x_m(\tau)-\phi(\tau)\|d\tau\right| \quad (p \text{ is Lipschitz})\\ &\le \overline{k}\left|\int_{t_0}^{t}\sup_{t\in[t_1,t_2]}\|x_m(t)-\phi(t)\|d\tau\right|\\ &\le \overline{k}\|x_m(\cdot)-\phi(\cdot)\|_\infty\left|\int_{t_0}^{t}d\tau\right|\\ &= \overline{k}\|x_m(\cdot)-\phi(\cdot)\|_\infty|t-t_0|\\ &\le \overline{k}\|x_m(\cdot)-\phi(\cdot)\|_\infty T.\end{aligned}$$

Since, by Step 2.2, $\|x_m(\cdot)-\phi(\cdot)\|_\infty \overset{m\to\infty}{\longrightarrow} 0$,

$$\int_{t_0}^{t}p(x_m(\tau),\tau)d(\tau)\overset{m\to\infty}{\longrightarrow}\int_{t_0}^{t}p(\phi(\tau),\tau)d\tau.$$

Therefore

$$\begin{array}{ccccc} x_{m+1}(t) & = & x_0 & + & \int_{t_0}^{t}p(x_m(\tau),\tau)d\tau\\ \downarrow & & \downarrow & & \downarrow\\ \phi(t) & = & x_0 & + & \int_{t_0}^{t}p(\phi(\tau),\tau)d\tau\end{array}$$

By the fundamental theorem of calculus ϕ is continuous and

$$\frac{d}{dt}\phi(t)=p(\phi(t),t) \ \ \forall t\in[t_1,t_2]\setminus D.$$

Therefore, our iteration converges to a solution of the differential equation for all $t\in[t_1,t_2]$. Since $[t_1,t_2]$ are arbitrary, we can see that our iteration converges to a solution of the differential equation for all $t\in\mathbb{R}$, by selecting t_1 small enough and t_2 large enough to ensure that $t\in[t_1,t_2]$.

3.7.3 Proof of uniqueness

Can there be more solutions besides the ϕ that we constructed in Section 3.7.2? It turns out that this is not the case. Assume, for the sake of contradiction, that there are two different solutions $\phi(\cdot), \psi(\cdot) : \mathbb{R} \to \mathbb{R}^n$. In other words

1. $\phi(t_0) = \psi(t_0) = x_0$;
2. $\forall \tau \in \mathbb{R} \setminus D$, $\frac{d}{d\tau}\phi(\tau) = p(\phi(\tau), \tau)$ and $\frac{d}{d\tau}\psi(\tau) = p(\psi(\tau), \tau)$; and
3. there exits $\hat{t} \in \mathbb{R}$ such that $\psi(\hat{t}) \neq \phi(\hat{t})$.

By construction $\hat{t} \neq t_0$ and $\phi(t_0) - \psi(t_0) = x_0 - x_0 = 0$. Moreover,

$$\begin{aligned}
&\frac{d}{d\tau}(\phi(\tau) - \psi(\tau)) = p(\phi(\tau), \tau) - p(\psi(\tau), \tau), \quad \forall \tau \in \mathbb{R} \setminus D \\
&\qquad \Rightarrow \phi(t) - \psi(t) = \int_{t_0}^{t} [p(\phi(\tau), \tau) - p(\psi(\tau), \tau)] d\tau, \quad \forall t \in \mathbb{R} \\
&\qquad \Rightarrow \|\phi(t) - \psi(t)\| \leq \left| \int_{t_0}^{t} \|p(\phi(\tau), \tau) - p(\psi(\tau), \tau)\| d\tau \right| \leq \left| \int_{t_0}^{t} k(\tau) \|\phi(\tau) - \psi(\tau)\| d\tau \right| \\
&\qquad \Rightarrow \|\phi(t) - \psi(t)\| \leq c_1 + \left| \int_{t_0}^{t} k(\tau) \|\phi(\tau) - \psi(\tau)\| d\tau \right| \quad \forall c_1 \geq 0.
\end{aligned}$$

Letting $u(t) = \|\phi(t) - \psi(t)\|$ and applying the Gronwall lemma leads to

$$0 \leq \|\phi(t) - \psi(t)\| \leq c_1 e^{\int_{t_0}^{t} k(\tau) d\tau}, \quad \forall c_1 \geq 0.$$

Letting $c_1 \to 0$ leads to

$$\|\phi(t) - \psi(t)\| = 0 \Rightarrow \phi(t) = \psi(t) \quad \forall t \in \mathbb{R}$$

which contradicts the assumption that $\psi(\hat{t}) \neq \phi(\hat{t})$.

This concludes the proof of existence and uniqueness of solutions of ordinary differential equations. From now on we can talk about "the solution" of the differential equation, as long as the conditions of the theorem are satisfied. It turns out that the solution has several other nice properties, for example it varies continuously as a function of the initial condition and parameters of the function p.

Unfortunately, the solution cannot usually be computed explicitly as a function of time. In this case we have to rely on simulation algorithms to approximate the solution using a computer. The nice properties of the solution (its guarantees of existence, uniqueness and continuity) come very handy in this case, since they allow one to design algorithms to numerically approximate the solutions and rigorously evaluate their convergence properties. Unfortunately for more general classes of systems, such as hybrid systems, one cannot rely on such properties and the task of simulation becomes much more challenging.

One exception to the general rule that solutions cannot be explicitly computed is the case of linear systems, where the extra structure afforded by the linearity of the function p allows us to study the solution in greater detail. We pursue this direction in Chapter 4.

Problems for chapter 3

Problem 3.1 (Norms) Let F be either $\mathbb{R}$ or $\mathcal{C}$. Show that the following are well-defined norms for the linear spaces (F^n, F), $(F^{m \times n}, F)$ and $\big(\mathcal{C}([t_0, t_1], F^n), F\big)$, respectively:

1. $||x||_\infty = \max_i |x_i|$, where $x = (x_1, \ldots, x_n) \in F^n$;
2. $||A||_\infty = \max_i \sum_j |a_{i,j}|$, where $A = [a_{i,j}] \in F^{m \times n}$;
3. $||f||_\infty = \max_{t \in [t_0, t_1]} ||f(t)||_\infty$, where $f \in \mathcal{C}([t_0, t_1], F^n)$.

For $x \in F^n$, show in addition that the norms $||x||_\infty$, $||x||_1$ and $||x||_2$ are equivalent (Hint: you may assume Schwarz's inequality: $\sum_{i=1}^n (|x_i| \cdot |y_i|) \leq ||x||_2 \cdot ||y||_2$, $\forall x, y \in F^n$).

Problem 3.2 (Ball of a given norm) Consider a normed vector space $(V, F, ||\cdot||)$ and $v \in V$ and $r \in \mathbb{R}_+$ define the open and closed balls centered at v with radius r as the sets

$$B(v, r) = \{v' \in V \mid \|v - v'\| < r\} \text{ and } \overline{B(v, r)} = \{v' \in V \mid \|v - v'\| < r\} \text{ respectively.}$$

Show that:

1. $B(v, r)$ is open and $\overline{B(a, r)}$ is closed.
2. $v_1, v_2 \in B(v, r) \Rightarrow \lambda v_1 + (1 - \lambda) v_2 \in B(v, r), \forall \lambda \in [0, 1]$ ($B(v, r)$ is convex);
3. $v \in B(0, r) \Rightarrow -v \in B(0, r)$ ($B(0, r)$ is balanced);
4. $\forall v' \in V \exists r \in (0, +\infty)$ such that $v' \in B(0, r)$.

Problem 3.3 Let $(V, F, \|\cdot\|)$ be a normed space. Show that:

1. The sets V and $\emptyset$ are both open and closed.
2. If $K_1, K_2 \subseteq V$ are open sets then $K_1 \cap K_2$ is an open set.
3. If $K_1, K_2 \subseteq V$ are closed sets then $K_1 \cap K_2$ is a closed set.
4. Let $\{K_i \subseteq V \mid i \in I\}$ be a collection of open sets, where I is an arbitrary (finite, or infinite) index set. Then $\cup_{i \in I} K_i$ is an open set.
5. Let $\{K_i \subseteq V \mid i \in I\}$ be a collection of closed sets, where I is an arbitrary (finite, or infinite) index set. Then $\cap_{i \in I} K_i$ is a closed set.

(Hint: Show 1, 3 and 5, then show that 3 implies 2 and 5 implies 4.)

Problem 3.4 (Continuity) Let $f : (U, F, \|\cdot\|_U) \to (V, F, \|\cdot\|_V)$ be a function between two normed spaces. Show that the following statements are equivalent:

1. f is continuous.
2. For all sequences $\{u_i\}_{i=1}^\infty \subseteq U$, $\lim_{i \to \infty} u_i = u \Rightarrow \lim_{i \to \infty} f(u_i) = f(u)$.
3. For all $K \subseteq V$ open, the set $f^{-1}(K) = \{u \in U \mid f(u) \in K\}$ is open.
4. For all $K \subseteq V$ closed, the set $f^{-1}(K)$ is closed.

Problem 3.5 (Equivalent Norms) Let (V, F) be a linear space. Let $||\cdot||_a$ and $||\cdot||_b$ be equivalent norms on (V, F). Let $v \in V$, $X \subseteq V$ and let $\{v_i\}_{i \in \mathbb{N}}$ be a sequence of elements of V. Show that:

1. $\{v_i\}_{i \in \mathbb{N}}$ is Cauchy w.r.t. $||\cdot||_a \Leftrightarrow \{v_i\}_{i \in \mathbb{N}}$ is Cauchy w.r.t. $||\cdot||_b$;
2. $v_i \overset{i \to \infty}{\longrightarrow} v$ w.r.t. $||\cdot||_a \Leftrightarrow v_i \overset{i \to \infty}{\longrightarrow} v$ w.r.t. $||\cdot||_b$;
3. X is dense in V w.r.t. $||\cdot||_a \Leftrightarrow X$ is dense in V w.r.t. $||\cdot||_b$.

Problem 3.6 (Induced norms) Let $(U, F, ||\cdot||_U)$ and $(V, F, ||\cdot||_V)$ be normed spaces, and let θ_U be the zero vector of U.

1. Show that the induced norm

$$||\mathcal{F}|| = \sup_{u \in U:\ u \neq \theta_U} \frac{||\mathcal{F}(u)||_V}{||u||_U}$$

   is a well-defined norm for the space of continuous operators $\mathcal{F} : U \to V$ (you may assume that the space of operators is a linear space over F).

2. Let $\mathcal{A} : U \to V$ be a linear operator. For the induced norm $||\mathcal{A}||$, show that

$$||\mathcal{A}|| = \sup_{u:\ ||u||_U = 1} ||\mathcal{A}(u)||_V .$$

Problem 3.7 (ODE solution properties) Consider $p : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ Lipschitz continuous in its first argument and piecewise continuous in its second. For t, $t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$, let $s(t, t_0, x_0)$ denote the unique solution of the differential equation

$$\frac{\partial}{\partial t} s(t, t_o, x_0) = p(s(t, t_0, x_0), t), \text{ with } s(t_0, t_0, x_0) = x_0.$$

Consider arbitrary $t, t_1, t_0 \in \mathbb{R}$. Show that:

1. For all $x_0 \in \mathbb{R}^n$, $s(t, t_1, s(t_1, t_0, x_0)) = s(t, t_o, x_0)$.
2. For all $x_0 \in \mathbb{R}^n$, $s(t_0, t, s(t, t_0, x_0)) = x_0$.
3. The function $s(t, t_0, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is continuous.

Problem 3.8 (Population dynamics)

Problem 3.9 (Discontinuous dynamics)

Chapter 4

Time varying linear systems: Solutions

We now return to the system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{4.1}$$
$$y(t) = C(t)x(t) + D(t)u(t) \tag{4.2}$$

where

$$t \in \mathbb{R},\ x(t) \in \mathbb{R}^n,\ u(t) \in \mathbb{R}^m,\ y(t) \in \mathbb{R}^p$$

and

$$A(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times n},\ \ B(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times m},$$
$$C(\cdot) : \mathbb{R} \to \mathbb{R}^{p\times n},\ \ D(\cdot) : \mathbb{R} \to \mathbb{R}^{p\times m}$$

which will concern us from now on.

4.1 Motivation: Linearization about a trajectory

Why should one worry about time varying linear systems? In addition to the fact that many physical systems are naturally modeled of the form (4.1)–(4.2), time varying linear systems naturally arise when one linearizes non-linear systems about a trajectory.

Consider a non-linear system modeled by an ODE

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)) \\ y(t) &= h(x(t), u(t)) \end{aligned} \tag{4.3}$$

with $t \in \mathbb{R}$, $x(t) \in \mathbb{R}^n$, $u(t) \in U \subseteq \mathbb{R}^m$, $f : \mathbb{R}^n \times U \to \mathbb{R}^n$. To ensure that the solution to (4.3) is well defined assume that f is globally Lipschitz in its first argument and continuous in its second argument and restrict attention to input trajectories $u(\cdot) : \mathbb{R} \to U$ which are piecewise continuous.

Assume that the system starts at a given initial state $x_0 \in \mathbb{R}^n$ at time $t = 0$ and we would like to drive it to a given terminal state $x_F \in \mathbb{R}^n$ at some given time $T > 0$. Moreover, we would like to accomplish this with the minimum amount of "effort". More formally, given a function

$$l : \mathbb{R}^n \times U \to \mathbb{R} \text{ (known as the running cost or Lagrangian)}$$

we would like to solve the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \int_0^T l(x(t),u(t))dt \\ \text{over} & u(\cdot):[0,T]\to U \text{ piecewise continuous} \\ \text{subject to} & x(0)=x_0 \\ & x(T)=x_F \\ & x(\cdot) \text{ is the solution of } \dot{x}(t)=f(x(t),u(t)). \end{array}$$

Essentially $l(x,u)$ encodes the "effort" of applying input u at state x.

The most immediate way to solve this optimization problem is to apply a theorem known as the maximum principle [15, 13], which leads to an optimal input trajectory $u(\cdot)^*:[0,T]\to U$ and the corresponding optimal state trajectory $x^*(\cdot):[0,T]\to\mathbb{R}^n$ such that $x^*(0)=x_0$, $x^*(T)=x_F$ and $\dot{x}^*(t)=f(x^*(t),u^*(t))$ for almost all $t\in[0,T]$.

Notice that the optimal input trajectory is "open loop". If we were to apply it to the real system the resulting state trajectory, $x(t)$, will most likely be far from the expected optimal state trajectory $x^*(t)$. The reason is that the differential equation model (4.3) of the system is bound to include modeling approximation, ignore minor dynamics, disturbance inputs, etc. In fact one can show that the resulting trajectory $x(t)$ may even diverge exponentially from the optimal trajectory $x^*(t)$ as a function of t, even for very small errors in the dynamics $f(x,u)$.

To solve this problem and force the real system to track the optimal trajectory we need to introduce feedback. Several ways of doing this exist. The most direct is to solve the optimization problem explicitly over feedback functions, i.e. determine a function $g:\mathbb{R}^n\times\mathbb{R}\to U$ such that $u(t)=g(x(t),t)$ is the optimal input if the system finds itself at state $x(t)$ at time t. This can be done using the theory of dynamic programming [3, 4]. The main advantage is that the resulting input will explicitly be in feedback form. The disadvantage is that the computation needed to do this is rather wasteful; in particular it requires one to solve the problem for all initial conditions rather than just the single (known) initial condition x_0. In most cases the computational complexity associated with doing this is prohibitive.

A sub-optimal but much more tractable approach is to linearize the non-linear system about the optimal trajectory $(x^*(\cdot),u^*(\cdot))$. We consider perturbations about the optimal trajectory

$$x(t)=x^*(t)+\delta x(t) \text{ and } u(t)=u^*(t)+\delta u(t).$$

We assume that $\delta x(t)\in\mathbb{R}^n$ and $\delta u(t)\in\mathbb{R}^m$ are small (i.e. we start close to the optimal trajectory). The plan is to use $\delta u(t)$ to ensure that $\delta x(t)$ remains small (i.e. we stay close to the optimal trajectory). Note that

$$x(t)=x^*(t)+\delta x(t)\Rightarrow\dot{x}(t)=\dot{x}^*(t)+\frac{d}{dt}(\delta x(t))=f(x^*(t),u^*(t))+\frac{d}{dt}(\delta x(t)).$$

Assume that

$$f(x,u)=\begin{bmatrix} f_1(x,u) \\ \vdots \\ f_n(x,u) \end{bmatrix}$$

where the functions $f_i:\mathbb{R}^n\times U\to\mathbb{R}$ for $i=1,\dots,n$ are differentiable. For $t\in[0,T]$ define

$$\frac{\partial f}{\partial x}(x^*(t),u^*(t))=\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x^*(t),u^*(t)) & \dots & \frac{\partial f_1}{\partial x_n}(x^*(t),u^*(t)) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(x^*(t),u^*(t)) & \dots & \frac{\partial f_n}{\partial x_n}(x^*(t),u^*(t)) \end{bmatrix}=A(t)\in\mathbb{R}^{n\times n}$$

$$\frac{\partial f}{\partial u}(x^*(t),u^*(t))=\begin{bmatrix} \frac{\partial f_1}{\partial u_1}(x^*(t),u^*(t)) & \dots & \frac{\partial f_1}{\partial u_m}(x^*(t),u^*(t)) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial u_1}(x^*(t),u^*(t)) & \dots & \frac{\partial f_n}{\partial u_m}(x^*(t),u^*(t)) \end{bmatrix}=B(t)\in\mathbb{R}^{n\times m}.$$

Then

$$\begin{aligned}\dot{x}(t) &= f(x(t),u(t)) = f(x^*(t)+\delta x(t), u^*(t)+\delta u(t)) \\ &= f(x^*(t),u^*(t)) + \frac{\partial f}{\partial x}(x^*(t),u^*(t))\delta x(t) + \frac{\partial f}{\partial u}(x^*(t),u^*(t))\delta u(t) + \text{higher order terms}\end{aligned}$$

If $\delta x(t)$ and $\delta u(t)$ are small then the higher order terms are small compared to the terms linear in $\delta x(t)$ and $\delta u(t)$ and the evolution of $\delta x(t)$ is approximately described by the linear time varying system

$$\frac{d}{dt}(\delta x(t)) = A(t)\delta x(t) + B(t)\delta u(t).$$

We can now use the theory developed in subsequent chapters to ensure that $\delta x(t)$ remains small and hence the nonlinear system tracks the optimal trajectory $x^*(t)$ closely.

4.2 Existence and structure of solutions

More formally, let $(X,\mathbb{R})$, $(U,\mathbb{R})$, and $(Y,\mathbb{R})$ be finite dimensional linear spaces, of dimensions n, m, and p respectively[1]. Consider families of linear functions,

$$\begin{aligned}\mathcal{A}(t) &: X \to X & \mathcal{B}(t) &: U \to X \\ \mathcal{C}(t) &: X \to Y & \mathcal{D}(t) &: U \to Y\end{aligned}$$

parametrized by a real number $t \in \mathbb{R}$. Fix bases, $\{e_i\}_{i=1}^n$ for $(X,\mathbb{R})$, $\{f_i\}_{i=1}^m$ for $(U,\mathbb{R})$, and $\{g_i\}_{i=1}^p$ for $(Y,\mathbb{R})$. Let $A(t)$, $B(t)$, $C(t)$, and $D(t)$ denote respectively the representation of the linear maps $\mathcal{A}(t)$, $\mathcal{B}(t)$, $\mathcal{C}(t)$, and $\mathcal{D}(t)$ with respect to those bases,

$$\begin{array}{ccccc} (X,\mathbb{R}) & \xrightarrow{\mathcal{A}(t)} & (X,\mathbb{R}) & \qquad (U,\mathbb{R}) & \xrightarrow{\mathcal{B}(t)} & (X,\mathbb{R}) \\ \{e_i\}_{i=1}^n & \xrightarrow[A(t)\in\mathbb{R}^{n\times n}]{} & \{e_i\}_{i=1}^n & \qquad \{f_i\}_{i=1}^m & \xrightarrow[B(t)\in\mathbb{R}^{n\times m}]{} & \{e_i\}_{i=1}^n \end{array}$$

$$\begin{array}{ccccc} (X,\mathbb{R}) & \xrightarrow{\mathcal{C}(t)} & (Y,\mathbb{R}) & \qquad (U,\mathbb{R}) & \xrightarrow{\mathcal{D}(t)} & (Y,\mathbb{R}) \\ \{e_i\}_{i=1}^n & \xrightarrow[C(t)\in\mathbb{R}^{p\times n}]{} & \{g_i\}_{i=1}^p & \qquad \{f_i\}_{i=1}^m & \xrightarrow[D(t)\in\mathbb{R}^{p\times m}]{} & \{g_i\}_{i=1}^p. \end{array}$$

Here we will be interested in dynamical systems of the form

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{4.4}$$
$$y(t) = C(t)x(t) + D(t)u(t) \tag{4.5}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$ denote the representations of elements of $(X,\mathbb{R})$, $(U,\mathbb{R})$, and $(Y,\mathbb{R})$ with respect to the bases $\{e_i\}_{i=1}^n$, $\{f_i\}_{i=1}^m$, and $\{g_i\}_{i=1}^p$ respectively.

To ensure that the system (4.4)–(4.5) is well-posed we will from now on impose following assumption.

Assumption 4.1 *$A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $D(\cdot)$ and $u(\cdot)$ are piecewise continuous.*

It is easy to see that this assumption ensures that the solution of (4.4)–(4.5) is well defined.

Fact 4.1 *For all $u(\cdot) : \mathbb{R} \to \mathbb{R}^m$ piecewise continuous and all $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$ there exists a unique solution $x(\cdot) : \mathbb{R} \to \mathbb{R}^n$ and $y(\cdot) : \mathbb{R} \to \mathbb{R}^p$ for the system (4.4)–(4.5).*

[1]The complex numbers, $\mathbb{C}$, can also be used as the field, at the expense of some additional complication in dimension counting. For simplicity we will think of linear spaces as defined over the field of real numbers, unless otherwise specified (e.g. for eigenvalue calculations).

Proof:(Sketch) Define $p : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ by

$$p(x,t) = A(t)x + B(t)u(t).$$

Exercise 4.1 Show that under Assumption 4.1 p satisfies the conditions of Theorem 3.6. (Take D to be the union of the discontinuity sets of $A(\cdot)$, $B(\cdot)$, and $u(\cdot)$).

The existence and uniqueness of $x(\cdot) : \mathbb{R} \to \mathbb{R}^n$ then follows by Theorem 3.6. Defining $y(\cdot) : \mathbb{R} \to \mathbb{R}^p$ by

$$y(t) = C(t)x(t) + D(t)u(t) \quad \forall t \in \mathbb{R}$$

completes the proof. ∎

The unique solution of (4.4)–(4.5) defines two functions

$$\begin{aligned} x(t) &= s(t, t_0, x_0, u) \quad \underline{\text{state transition map}} \\ y(t) &= \rho(t, t_0, x_0, u) \quad \underline{\text{output response map}} \end{aligned}$$

mapping the input trajectory $u(\cdot) : \mathbb{R} \to \mathbb{R}^m$ and initial condition $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$ to the state and output at time $t \in \mathbb{R}$ respectively. It is easy to see that the function s and the input u also implicitly define the function ρ through

$$\rho(t, t_0, x_0, u) = C(t)s(t, t_0, x_0, u) + D(t)u(t).$$

Therefore the main properties of the solution functions can be understood by analysing the properties of the state solution function s.

Theorem 4.1 *Let D_x be the union of the discontinuity sets of $A(\cdot)$, $B(\cdot)$ and $u(\cdot)$ and D_y the union of the discontinuity sets of $C(\cdot)$, $D(\cdot)$ and $u(\cdot)$.*

1. *For all $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$, $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$*
   - *$x(\cdot) = s(\cdot, t_0, x_0, u) : \mathbb{R} \to \mathbb{R}^n$ is continuous and differentiable for all $t \in \mathbb{R} \setminus D_x$.*
   - *$y(\cdot) = \rho(\cdot, t_0, x_0, u) : \mathbb{R} \to \mathbb{R}^p$ is piecewise continuous with discontinuity set D_y.*
2. *For all $t, t_0 \in \mathbb{R}$, $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$, $x(\cdot) = s(t, t_0, \cdot, u) : \mathbb{R}^n \to \mathbb{R}^n$ and $\rho(t, t_0, \cdot, u) : \mathbb{R}^n \to \mathbb{R}^p$ are continuous.*
3. *For all $t, t_0 \in \mathbb{R}$, $x_{01}, x_{02} \in \mathbb{R}^n$, $u_1(\cdot), u_2(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$, $a_1, a_2 \in \mathbb{R}$*

$$\begin{aligned} s(t, t_0, a_1x_{01} + a_2x_{02}, a_1u_1 + a_2u_2) &= a_1s(t, t_0, x_{01}, u_1) + a_2s(t, t_0, x_{02}, u_2) \\ \rho(t, t_0, a_1x_{01} + a_2x_{02}, a_1u_1 + a_2u_2) &= a_1\rho(t, t_0, x_{01}, u_1) + a_2\rho(t, t_0, x_{02}, u_2). \end{aligned}$$

4. *For all $t, t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$, $u \in PC(\mathbb{R}, \mathbb{R}^m)$,*

$$\begin{aligned} s(t, t_0, x_0, u) &= s(t, t_0, x_0, 0) + s(t, t_0, 0, u) \\ \rho(t, t_0, x_0, u) &= \rho(t, t_0, x_0, 0) + \rho(t, t_0, 0, u) \end{aligned}$$

The last statement requires some care, as 0 is used in two different ways: As the zero element in $\mathbb{R}^n$ ($\theta_{\mathbb{R}^n} = (0, \ldots, 0)$ in the notation of Chapter 3) and as the zero element in the space of piecewise continuous function ($\theta_{\mathrm{PC}}(t) = (0, \ldots, 0)$ for all $t \in \mathbb{R}$ in the notation of Chapter 3). The interpretation should hopefully be clear from the location of 0 in the list of arguments of s and ρ.

Proof: Part 1 follows from the definition of the solution. Part 4 follows from Part 3, by setting $u_1 = 0$, $u_2 = u$, $x_{01} = x_0$, $x_{02} = 0$, and $a_1 = a_2 = 1$. Part 2 follows from Part 4, by noting that $s(t, t_0, \cdot, u) = s(t, t_0, \cdot, 0) + s(t, t_0, 0, u)$ and $s(t, t_0, \cdot, 0) : \mathbb{R}^n \to \mathbb{R}^n$ is a linear function between finite dimensional linear spaces (and hence continuous by Corollary 3.2); the argument for ρ is similar. So we only need to establish Part 3.

Let $x_1(t) = s(t, t_0, x_{01}, u_1)$, $x_2(t) = s(t, t_0, x_{02}, u_2)$, $x(t) = s(t, t_0, a_1 x_{01} + a_2 x_{02}, a_1 u_1 + a_2 u_2)$, and $\phi(t) = a_1 x_1(t) + a_2 x_2(t)$. We would like to show that $x(t) = \phi(t)$ for all $t \in \mathbb{R}$. By definition

$$x(t_0) = a_1 x_{01} + a_2 x_{02} = a_1 x_1(t_0) + a_2 x_2(t_0) = \phi(t_0).$$

Moreover, if we let $u(t) = a_1 u_1(t) + a_2 u(t)$ then for all $t \in \mathbb{R} \setminus D$

$$\begin{aligned}
\dot{x}(t) &= A(t)x(t) + B(t)u(t) \\
\dot{\phi}(t) &= a_1 \dot{x}_1(t) + a_2 \dot{x}_2(t) \\
&= a_1(A(t)x_1(t) + B(t)u_1(t)) + a_2(A(t)x_2(t) + B(t)u_2(t)) \\
&= A(t)(a_1 x_1(t) + a_2 x_2(t)) + B(t)(a_1 u_1(t) + a_2 u_2(t)) \\
&= A(t)\phi(t) + B(t)u(t).
\end{aligned}$$

Therefore $x(t) = \phi(t)$ since the solution to the linear ODE is unique. Linearity of ρ follows from the fact that $C(t)x + D(t)u$ is linear in x and u. ∎

4.3 State transition matrix

By Part 4 of Theorem 4.1 the solution of the system can be partitioned into two distinct components:

$$\begin{array}{ccccc}
s(t, t_0, x_0, u) & = & s(t, t_0, x_0, 0) & + & s(t, t_0, 0, u) \\
\text{state transition} & = & \text{zero input transition} & + & \text{zero state transition}
\end{array}$$

$$\begin{array}{ccccc}
\rho(t, t_0, x_0, u) & = & \rho(t, t_0, x_0, 0) & + & \rho(t, t_0, 0, u) \\
\text{output response} & = & \text{zero input response} & + & \text{zero state response.}
\end{array}$$

Moreover, by Part 3 of Theorem 4.1, the zero input components $s(t, t_0, x_0, 0)$ and $\rho(t, t_0, x_0, 0)$ are linear in $x_0 \in \mathbb{R}^n$. Therefore, in the basis $\{e_i\}_{i=1}^n$ used for the representation of $A(\cdot)$, the linear map $s(t, t_0, \cdot, 0) : \mathbb{R}^n \to \mathbb{R}^n$ has a matrix representation. This representation, that will of course depend on t and t_0 in general, is called the state transition matrix and is denoted by $\Phi(t, t_0)$. Therefore, assuming $s(t, t_0, x_0, 0)$ refers to the representation of the solution with respect to the basis $\{e_i\}_{i=1}^n$,

$$s(t, t_0, x_0, 0) = \Phi(t, t_0)x_0. \tag{4.6}$$

Exercise 4.2 Show that the representation of $\rho(t, t_0, \cdot, 0) : \mathbb{R}^n \to \mathbb{R}^p$ with respect to the bases $\{e_i\}_{i=1}^n$ and $\{g_i\}_{i=1}^p$ is given by $C(t)\Phi(t, t_0)$; in other words $\rho(t, t_0, x_0, 0) = C(t)\Phi(t, t_0)x_0$.

Therefore the state transition matrix $\Phi(t, t_0)$ completely characterizes the zero input state transition and output response. We will soon see that, together with the input trajectory $u(\cdot)$, it also characterizes the complete state transition and output response.

Theorem 4.2 *$\Phi(t, t_0)$ has the following properties:*

1. *$\Phi(\cdot, t_0) : \mathbb{R} \to \mathbb{R}^{n \times n}$ is the unique solution of the linear matrix ordinary differential equation*

$$\frac{\partial}{\partial t}\Phi(t, t_0) = A(t)\Phi(t, t_0) \text{ with } \Phi(t_0, t_0) = I.$$

Hence it is continuous for all $t \in \mathbb{R}$ and differentiable everywhere except at the discontinuity points of $A(t)$.

2. *For all $t, t_0, t_1 \in \mathbb{R}$, $\Phi(t, t_0) = \Phi(t, t_1)\Phi(t_1, t_0)$.*
3. *For all $t_1, t_0 \in \mathbb{R}$, $\Phi(t_1, t_0)$ is invertible and its inverse is $[\Phi(t_1, t_0)]^{-1} = \Phi(t_0, t_1)$.*

Proof: Part 1. Recall that $s(t, t_0, x_0, 0) = \Phi(t, t_0)x_0$ is the solution to the linear differential equation $\dot{x}(t) = A(t)x(t)$ with $x(t_0) = x_0$. For $i = 1, \dots, n$ consider the solution $x^i(\cdot) = s(\cdot, t_0, x^i(t_0), 0)$ to the linear differential equation starting at the representation $x^i(t_0) = (0, \dots, 1, \dots, 0) \in \mathbb{R}^n$ of the basis vector e_i; in other words,

$$\dot{x}^i(t) = A(t)x^i(t) \text{ with } x^i(t_0) = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \text{ hence, by (4.6), } x^i(t) = \Phi(t, t_0) \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}.$$

Note that $x^i(t)$ is equal to the i^{th} column of $\Phi(t, t_0)$. Putting the columns next to each other

$$\Phi(t, t_0) = \begin{bmatrix} x^1(t) & x^2(t) & \dots & x^n(t) \end{bmatrix}$$

shows that

$$\begin{aligned} \frac{\partial}{\partial t}\Phi(t, t_0) &= \begin{bmatrix} \frac{d}{dt}x^1(t) & \frac{d}{dt}x^2(t) & \dots & \frac{d}{dt}x^n(t) \end{bmatrix} \\ &= \begin{bmatrix} A(t)x^1(t) & A(t)x^2(t) & \dots & A(t)x^n(t) \end{bmatrix} \\ &= A(t) \begin{bmatrix} x^1(t) & x^2(t) & \dots & x^n(t) \end{bmatrix} = A(t)\Phi(t, t_0). \end{aligned}$$

Moreover,

$$\Phi(t_0, t_0) = \begin{bmatrix} x^1(t_0) & x^2(t_0) & \dots & x^n(t_0) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = I \in \mathbb{R}^{n \times n}.$$

Part 1 follows.

Part 2. Consider arbitrary $t_0, t_1 \in \mathbb{R}$ and let $L(t) = \Phi(t, t_0)$ and $R(t) = \Phi(t, t_1)\Phi(t_1, t_0)$. We would like to show that $L(t) = R(t)$ for all $t \in \mathbb{R}$. Note that (by Part 1)

$$\begin{aligned} L(t_1) &= \Phi(t_1, t_0) \\ R(t_1) &= \Phi(t_1, t_1)\Phi(t_1, t_0) = I \cdot \Phi(t_1, t_0) = \Phi(t_1, t_0). \end{aligned}$$

Therefore $L(t_1) = R(t_1)$. Moreover, (also by Part 1)

$$\begin{aligned} \frac{d}{dt}L(t) &= \frac{d}{dt}\Phi(t, t_0) = A(t)\Phi(t, t_0) = A(t)L(t) \\ \frac{d}{dt}R(t) &= \frac{d}{dt}[\Phi(t, t_1)\Phi(t_1, t_0)] = \frac{d}{dt}[\Phi(t, t_1)]\Phi(t_1, t_0) = A(t)\Phi(t, t_1)\Phi(t_1, t_0) = A(t)R(t) \end{aligned}$$

Therefore $L(t) = R(t)$ by existence and uniqueness of solutions of linear differential equations.

Part 3. First we show that $\Phi(t, t_0)$ is nonsingular for all $t, t_0 \in \mathbb{R}$. Assume, for the sake of contradiction, that it is not, i.e. there exists $t, t_0 \in \mathbb{R}$ such that $\Phi(t, t_0)$ is singular. Then the columns of $\Phi(t, t_0)$ are linearly dependent (Theorem 2.3) and there exists $x_0 \in \mathbb{R}^n$ with $x_0 \neq 0$ such that $\Phi(t, t_0)x_0 = 0$.

Let $x(\tau) = \Phi(\tau, t_0)x_0$. Notice that $x(t) = \Phi(t, t_0)x_0 = 0$ and

$$\frac{d}{d\tau}x(\tau) = \frac{d}{d\tau}\Phi(\tau, t_0)x_0 = A(\tau)\Phi(\tau, t_0)x_0 = A(\tau)x(\tau).$$

Therefore $x(\tau)$ is the unique solution to the differential equation:

$$\frac{d}{d\tau}x(\tau) = A(\tau)x(\tau) \text{ with } x(t) = 0. \tag{4.7}$$

The function $x(\tau) = 0$ for all $\tau \in \mathbb{R}$ clearly satisfies (4.7); therefore it is the unique solution to (4.7).

Let now $\tau = t_0$. Then

$$0 = x(t_0) = \Phi(t_0, t_0)x_0 = I \cdot x_0 = x_0$$

which contradicts the fact that $x_0 \neq 0$. Therefore $\Phi(t, t_0)$ cannot be singular.

To determine its inverse, recall that for all $t, t_0, t_1 \in \mathbb{R}$,

$$\Phi(t, t_1)\Phi(t_1, t_0) = \Phi(t, t_0)$$

and let $t = t_0$. Then

$$\Phi(t_0, t_1)\Phi(t_1, t_0) = \Phi(t_0, t_0) = I \Rightarrow [\Phi(t_1, t_0)]^{-1} = \Phi(t_0, t_1).$$

■

In addition to these, the state transition matrix also has several other interesting properties, some of which can be found in the exercises. We can now show that the state transition matrix $\Phi(t, t_0)$ completely characterizes the solution of linear time varying differential equations.

Theorem 4.3 *For all $t, t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$, $u(\cdot)) \in PC(\mathbb{R}, \mathbb{R}^m)$,*

$$\begin{array}{ccccc} s(t, t_0, x_0, u) & = & \Phi(t, t_0)x_0 & + & \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau \\ \textit{state transition} & = & \textit{zero input transition} & + & \textit{zero state transition} \end{array}$$

$$\begin{array}{ccccc} \rho(t, t_0, x_0, u) & = & C(t)\Phi(t, t_0)x_0 & + & C(t)\int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau + D(t)u(t) \\ \textit{output response} & = & \textit{zero input response} & + & \textit{zero state response.} \end{array}$$

Proof: Several methods exist for proving this fact. The simplest is to invoke the rule of Leibniz for differentiating integrals.

$$\frac{d}{dt}\left[\int_{a(t)}^{b(t)} f(t, \tau)d\tau\right] = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t}f(t, \tau)d\tau + f(t, b(t))\frac{d}{dt}b(t) - f(t, a(t))\frac{d}{dt}a(t).$$

(for the sake of comparison, notice that the fundamental theorem of calculus is the special case $a(t) = t_0$, $b(t) = t$ and $f(t, \tau) = f(\tau)$ independent of t.)

We start by showing that

$$s(t, t_0, x_0, u) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau$$

If we let $L(t) = s(t, t_0, x_0, u)$ and $R(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau$ we would like to show that for all $t \in \mathbb{R}$, $L(t) = R(t)$. Notice that by definition $L(t_0) = s(t_0, t_0, x_0, u) = x_0$ and

$$\frac{d}{dt}L(t) = \frac{d}{dt}s(t, t_0, x_0, u) = A(t)s(t, t_0, x_0, u) + B(t)u(t) = A(t)L(t) + B(t)u(t).$$

We will show that $R(t)$ satisfies the same differential equation with the same initial condition; the claim then follows by the existence and uniqueness theorem.

Note first that

$$R(t_0) = \Phi(t_0, t_0)x_0 + \int_{t_0}^{t_0} \Phi(t, \tau)B(\tau)u(\tau)d\tau = I \cdot x_0 + 0 = x_0 = L(t_0).$$

Moreover, by the Leibniz rule

$$\begin{aligned}
\frac{d}{dt}R(t) &= \frac{d}{dt}\left[\Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau\right] \\
&= \left[\frac{d}{dt}\Phi(t, t_0)\right]x_0 + \frac{d}{dt}\left[\int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau\right] \\
&= A(t)\Phi(t, t_0)x_0 + \int_{t_0}^{t} \frac{\partial}{\partial t}\Phi(t, \tau)B(\tau)u(\tau)d\tau + \Phi(t, t)B(t)u(t)\frac{d}{dt}t - \Phi(t_0, t_0)B(t_0)u(t_0)\frac{d}{dt}t_0 \\
&= A(t)\Phi(t, t_0)x_0 + \int_{t_0}^{t} A(t)\Phi(t, \tau)B(\tau)u(\tau)d\tau + I \cdot B(t)u(t) \\
&= A(t)\left[\Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau\right] + B(t)u(t) \\
&= A(t)R(t) + B(t)u(t).
\end{aligned}$$

Therefore $R(t)$ and $L(t)$ satisfy the same linear differential equation for the same initial condition, hence they are equal for all t by uniqueness of solutions.

To obtain the formula for $\rho(t, t_0, x_0, u)$ simply substitute the formula for $s(t, t_0, x_0, u)$ into $y(t) = C(t)x(t) + D(t)u(t)$. ∎

Let us now analyze the zero state transition and response in greater detail. By Theorem 4.1, the zero state transition and the zero state response are both linear functions $s(t, t_0, 0, \cdot) : PC(\mathbb{R}, \mathbb{R}^m) \to \mathbb{R}^n$ and $\rho(t, t_0, 0, \cdot) : PC(\mathbb{R}, \mathbb{R}^m) \to \mathbb{R}^p$ respectively.

$$\begin{array}{ccc}
(PC(\mathbb{R}, \mathbb{R}^m), \mathbb{R}) & \overset{s(t,t_0,0,\cdot)}{\longrightarrow} & (\mathbb{R}^n, \mathbb{R}) \\
u(\cdot) & \longmapsto & \displaystyle\int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau
\end{array}$$

and

$$\begin{array}{ccc}
(PC(\mathbb{R}, \mathbb{R}^m), \mathbb{R}) & \overset{\rho(t,t_0,0,\cdot)}{\longrightarrow} & (\mathbb{R}^p, \mathbb{R}) \\
u(\cdot) & \longmapsto & \displaystyle C(t)\int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau + D(t)u(t).
\end{array}$$

Fix the basis $\{f_j\}_{j=1}^m$ for $(\mathbb{R}^m, \mathbb{R})$ used in the representation of $B(t) \in \mathbb{R}^{n\times m}$. Fix $\sigma \geq t_0$ and consider a family of functions $\delta_{(\sigma,\epsilon)}(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$ parametrized by $\epsilon > 0$ and defined by

$$\delta_{(\sigma,\epsilon)}(t) = \begin{cases} 0 & \text{if } t < \sigma \\ \frac{1}{\epsilon} & \text{if } \sigma \leq t < \sigma + \epsilon \\ 0 & \text{if } t \geq \sigma + \epsilon. \end{cases}$$

For $j = 1, \ldots, m$, consider the zero state transition[2] $s(t, t_0, 0, \delta_{(\sigma,\epsilon)}(t)f_j)$ under input $\delta_{(\sigma,\epsilon)}(t)f_j$. Since $s(t_0, t_0, 0, \delta_{(\sigma,\epsilon)}(t)f_j) = 0$ and the input is zero until $t = \sigma$,

$$s(t, t_0, 0, \delta_{(\sigma,\epsilon)}(t)f_j) = 0 \quad \forall t < \sigma.$$

[2]Strictly speaking, we should use the $\mathbb{R}^m$ representation of the basis vector $f_i \in U$ instead of f_i itself. The reader is asked to excuse this slight abuse of the notation.

Exercise 4.3 Show this by invoking the existence-uniqueness theorem.

For $t \geq \sigma + \epsilon$ and assuming ϵ is small

$$\begin{aligned} s(t, t_0, 0, \delta_{(\sigma,\epsilon)}(t) f_j) &= \int_{t_0}^{t} \Phi(t,\tau) B(\tau) \delta_{(\sigma,\epsilon)}(\tau) f_j d\tau \\ &= \int_{\sigma}^{\sigma+\epsilon} \Phi(t,\tau) B(\tau) \frac{1}{\epsilon} f_j d\tau \\ &= \frac{1}{\epsilon} \int_{\sigma}^{\sigma+\epsilon} \Phi(t,\sigma+\epsilon)\Phi(\sigma+\epsilon,\tau) B(\tau) f_j d\tau \\ &= \frac{\Phi(t,\sigma+\epsilon)}{\epsilon} \int_{\sigma}^{\sigma+\epsilon} \Phi(\sigma+\epsilon,\tau) B(\tau) f_j d\tau \\ &\approx \frac{\Phi(t,\sigma+\epsilon)}{\epsilon} \left[\Phi(\sigma+\epsilon,\sigma) B(\sigma) f_j\right] \epsilon \\ &\xrightarrow{\epsilon \to 0} \Phi(t,\sigma)\Phi(\sigma,\sigma) B(\sigma) f_j = \Phi(t,\sigma) B(\sigma) f_j \end{aligned}$$

Therefore

$$\lim_{\epsilon \to 0} s(t, t_0, 0, \delta_{(\sigma,\epsilon)}(t) f_j) = \Phi(t,\sigma) B(\sigma) f_j.$$

Formally, if we pass the limit inside the function s and define

$$\delta_\sigma(t) = \lim_{\epsilon \to 0} \delta_{(\sigma,\epsilon)}(t)$$

we obtain

$$s(t, t_0, 0, \delta_\sigma(t) f_j) = \Phi(t,\sigma) B(\sigma) f_j \in \mathbb{R}^m.$$

The statement is "formal" since to pass the limit inside the function we first need to ensure that the function is continuous.

Exercise 4.4 We already know that the function $s(t, t_0, 0, \cdot) : PC(\mathbb{R}, \mathbb{R}^m) \to \mathbb{R}^n$ is linear. What more do we need to check to make sure that it is continuous?

Moreover, strictly speaking $\delta_\sigma(t)$ is not an acceptable input function, since it is equal to infinity at $t = \sigma$ and hence not piecewise continuous. Indeed, $\delta_\sigma(t)$ is not a real valued function at all, it just serves as a mathematical abstraction for an input pulse of arbitrarily small length. This mathematical abstraction is known as the impulse function or the Dirac pulse. Even though in practice the response of a real system to such an impulsive input cannot be observed, by applying as input piecewise continuous functions $\delta_{(\sigma,\epsilon)}(t)$ for ϵ small enough one can approximate the state transition $s(t, t_0, 0, \delta_\sigma(t) f_j)$ arbitrarily closely. Notice also that for $t \geq \sigma$

$$s(t, t_0, 0, \delta_\sigma(t) f_j) = s(t, \sigma, B(\sigma) f_j, 0)$$

i.e. the zero state transition due to the impulse $\delta_\sigma(t) f_j$ is also a zero input transition starting with state $B(\sigma) f_j$ at time σ.

Repeating the process for all $\{f_j\}_{j=1}^m$ leads to m vectors $\Phi(t,\sigma) B(\sigma) f_j$ for $j = 1, \ldots, m$. Ordering these vectors according to their index j and putting them one next to the other leads to the impulse transition matrix, $K(t,\sigma) \in \mathbb{R}^{n \times m}$, defined by

$$K(t,\sigma) = \begin{cases} \Phi(t,\sigma) B(\sigma) & \text{if } t \geq \sigma \\ 0 & \text{if } t < \sigma. \end{cases}$$

The (i,j) element of $K(t,\sigma)$ contains the trajectory of state x_i when the impulse function $\delta_\sigma(t)$ is applied to input u_j. Note that, even though these elements cannot be measured in practice, the impulse transition matrix is still a well defined, matrix valued function for all $t, \sigma \in \mathbb{R}$.

Substituting $s(t,0,0,\delta_\sigma(t)f_j)$ into the equation $y(t)=C(t)x(t)+D(t)u(t)$ leads to

$$\rho(t,t_0,0,\delta_\sigma(t)f_j)=C(t)\Phi(t,\sigma)B(\sigma)f_j+D(t)f_j\delta_\sigma(t)\in\mathbb{R}^m.$$

and the impulse response matrix, $H(t,\sigma)\in\mathbb{R}^{p\times m}$, defined by

$$H(t,\sigma)=\begin{cases} C(t)\Phi(t,\sigma)B(\sigma)+D(t)\delta_\sigma(t) & \text{if } t\geq\sigma \\ 0 & \text{if } t<\sigma. \end{cases}$$

Note that, unlike $K(t,\sigma)$, $H(t,\sigma)$ is in general not a well defined matrix valued function since it in general contains an impulse in its definition (unless of course $D(t)=0$ for all $t\in\mathbb{R}$).

Problems for chapter 4

Problem 4.1 (Invariant Subspaces) Let $L:V\to V$ be a linear map on an n-dimensional vector space V over the field F. Recall that a subspace $M\subset V$ is called L-invariant if $Lx\in M$ for every $x\in M$. Suppose that V is a direct sum of two subspaces M_1 and M_2, i.e., $M_1\cap M_2=\{0\}$, and $M_1+M_2=V$. If both M_1 and M_2 are L-invariant, show that there exists a matrix representation $A\in F^{n\times n}$ of the form

$$A=\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22}\end{bmatrix}$$

with $\text{DIM}(A_{11})=\text{DIM}(M_1)$ and $\text{DIM}(A_{22})=\text{DIM}(M_2)$. (Recall that the sum of subspaces M and N of a vector space X is the set of all vectors of the form $m+n$ where $m\in M$ and $n\in N$. A vector space X is the direct sum of two subspaces M and N if every vector $x\in X$ has a unique prepresentation of the form $x=m+n$ where $m\in M$ and $n\in N$; we write $X=M\oplus N$.)

Problem 4.2 (Eigenvalues and Invariant Subspaces) Let A be a real-valued $n\times n$ matrix. Suppose that $\lambda+i\mu$ is a complex eigenvalue of A and $x+iy$ is the corresponding complex eigenvector, where $\lambda,\mu\in\mathbb{R}$ and $x,y\in\mathbb{R}^n$. Show that $x-iy$ is also an eigenvector with eigenvalue $\lambda-i\mu$. Let V be the 2-dimensional subspace spanned by x and y, i.e., V is the set of linear combinations with real-valued coefficients of the real-valued vectors x and y. Show that V is an invariant subspace of A, namely, if $z\in V$ then we have $Az\in V$.

Problem 4.3 (ODEs)

1. Let $f:\mathbb{R}^n\to\mathbb{R}^n$ be Lipschitz with Lipschitz constant $K\in[0,+\infty)$. For $t\in\mathbb{R}$, let $x(t)$ be the solution of $\dot{x}(t)=f(x(t))$, with $x(0)=x_0$. Let $\bar{x}\in\mathbb{R}^n$ be such that $f(\bar{x})=0$. Show that $||x(t)-\bar{x}||\leq e^{Kt}||x_0-\bar{x}||$, $\forall t\in\mathbb{R}_+$. (Here $||\cdot||$ is the Euclidean norm on $(\mathbb{R}^n,\mathbb{R})$ for which K is defined. Hint: use the Gronwall Lemma).

2. Let $A:\mathbb{R}_+\to\mathbb{R}^{n\times n}$, $B:\mathbb{R}_+\to\mathbb{R}^{n\times m}$ and $u:\mathbb{R}_+\to\mathbb{R}^m$ be piecewise continuous functions. Show that, for any $x_0\in\mathbb{R}^n$, the linear ODE

$$\begin{cases}\dot{x}(t)=A(t)x(t)+B(t)u(t), & t\in\mathbb{R}_+, \\ x(0)=x_0\end{cases}$$

has a unique solution $x(\cdot):\mathbb{R}_+\to\mathbb{R}^n$. (Hint: you may assume that if $A(t)$ is piecewise continuous then so is its induced norm).

Problem 4.4 (Linear ODEs) Let $A(\cdot):\mathbb{R}\to\mathbb{R}^{n\times n}$ be piecewise continuous. Consider the following linear ODE:

$$\dot{x}(t)=A(t)x(t), \tag{4.8}$$

and let $\Phi(t,t_0)$ be the state transition matrix.

1. Show that $\frac{\partial}{\partial t}\Phi(t_0, t) = -\Phi(t_0, t)A(t)$.
   (Hint: differentiate the identity $\Phi(t_0, t)\Phi(t, t_0) = I$.)

2. Let $\mathbf{X}_0$ be a convex set (i.e. $x, x' \in \mathbf{X}_0 \Rightarrow \lambda x + (1-\lambda)x' \in \mathbf{X}_0,\ \forall \lambda \in [0, 1]$). Let $s(t, t_0, x_0)$ be the solution of (1) associated with the initial condition $x(t_0) = x_0$. Show that the set

   $$\mathbf{X}(t) = \{s(t, t_0, x_0) :\ x_0 \in \mathbf{X}_0\}$$

   is convex for all $t \in \mathbb{R}$.

3. A function $X(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times n}$ is said to be a fundamental matrix for the matrix differential equation $\dot{X}(t) = A(t)X(t)$ if it is a solution and it is nonsingular for all $t \in \mathbb{R}$. Show that, given any fundamental matrix $X(\cdot)$, it holds that $\Phi(t, t_0) = X(t) \cdot X(t_0)^{-1}$.

Problem 4.5 (Change of basis)

Chapter 5

Time invariant linear systems: Solutions and transfer functions

Let us now turn to the special case where all the matrices involved in the system dynamics are constant, in other words

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{5.1}$$
$$y(t) = Cx(t) + Du(t) \tag{5.2}$$

with $t \in \mathbb{R}_+$, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$ and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$ are constant matrices.

5.1 Time domain solution

Define the exponential of the matrix $A \in \mathbb{R}^{n\times n}$ by

$$e^{At} = I + At + \frac{A^2t^2}{2!} + \ldots + \frac{A^kt^k}{k!} + \ldots \in \mathbb{R}^{n\times n}. \tag{5.3}$$

Theorem 5.1 *For all $t, t_0 \in \mathbb{R}_+$, $\Phi(t, t_0) = e^{A(t-t_0)}$.*

Proof: Exercise. Show that $e^{A(t-t_0)}$ satisfies the conditions of Part 1 of Theorem 4.2 by taking the derivative of the expansion in (5.3) with respect to t. The result follows by uniqueness of solutions of ordinary differential equations. ∎

The above theorem together with the different properties of $\Phi(t, t_0)$ established in Chapter 4 immediately lead to the following corollary.

Corollary 5.1 *The state transition matrix, solution, impulse transition, and impulse response of a time invariant linear system satisfy the following properties:*

1. *For all $t, t_1, t_0 \in \mathbb{R}$, $e^{At_1}e^{At_2} = e^{A(t_1+t_2)}$ and $\left[e^{At}\right]^{-1} = e^{-At}$.*
2. *For all $t, t_0 \in \mathbb{R}$, $\Phi(t, t_0) = \Phi(t - t_0, 0)$.*

3. For all $t, t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$, $u(\cdot) \in \mathrm{PC}(\mathbb{R}, \mathbb{R}^m)$,

$$s(t, t_0, x_0, u) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau$$
$$\rho(t, t_0, x_0, u) = Ce^{A(t-t_0)}x_0 + C\int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau + Du(t).$$

4. For all $t, \sigma \in \mathbb{R}$ the

$$K(t,\sigma) = K(t-\sigma, 0) = \begin{cases} e^{A(t-\sigma)}B & \text{if } t \geq \sigma \\ 0 & \text{if } t < \sigma. \end{cases}$$
$$H(t,\sigma) = H(t-\sigma, 0) = \begin{cases} Ce^{A(t-\sigma)}B + D\delta_0(t-\sigma) & \text{if } t \geq \sigma \\ 0 & \text{if } t < \sigma. \end{cases}$$

From the above it becomes clear that for linear time invariant systems the solution is independent of the initial time t_0; all that matters is how much time has elapsed since then, i.e. $t - t_0$. Without loss of generality we will therefore take $t_0 = 0$ and write

$$x(t) = s(t, 0, x_0, u) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}B(\tau)u(\tau)d\tau$$
$$y(t) = \rho(t, 0, x_0, u) = Ce^{At}x_0 + C\int_0^t e^{A(t-\tau)}B(\tau)u(\tau)d\tau + Du(t)$$
$$K(t) = K(t, 0) = \begin{cases} e^{At}B & \text{if } t \geq 0 \\ 0 & \text{if } t < 0. \end{cases}$$
$$H(t) = H(t, 0) = \begin{cases} Ce^{At}B + D\delta_0(t) & \text{if } t \geq 0 \\ 0 & \text{if } t < 0. \end{cases}$$

Notice that in this case the integral that appears in the state transition and output response is simply the convolution of the input $u(\cdot)$ with the impulse transition and impulse response matrices respectively,

$$x(t) = e^{At}x_0 + (K * u)(t)$$
$$y(t) = Ce^{At}x_0 + (H * u)(t).$$

Exercise 5.1 Verify this.

5.2 Semi-simple matrices

For the time invariant case, the state transition matrix e^{At} can be computed explicitly. In some cases this can be done directly from the infinite series; this is the case, for example, for nilpotent matrices, i.e. matrices for which there exists $N \in \mathbb{N}$ such that $A^N = 0$; conditions to determine when this is the case will be given in Section 5.4. More generally, one can use Laplace transforms or eigenvectors to do this. We concentrate primarily on the latter method; Laplace transforms will be briefly discussed in Section 5.4. We start with the simpler case of the so-called semi-simple matrices.

Definition 5.1 *A matrix $A \in \mathbb{R}^{n\times n}$ is semi-simple if and only if its right eigenvectors $\{v_i\}_{i=1}^n \subseteq \mathbb{C}^n$ are linearly independent in the linear space $(\mathbb{C}^n, \mathbb{C})$.*

Theorem 5.2 *Matrix $A \in \mathbb{R}^{n\times n}$ is semi-simple if and only if there exists a nonsingular matrix $T \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{C}^{n\times n}$ such that $A = T^{-1}\Lambda T$.*

Proof: ($\Rightarrow$) Recall that if $\{v_i\}_{i=1}^n$ are linearly independent then the matrix $[v_1 \dots v_n] \in \mathbb{C}^{n\times n}$ is invertible. Let

$$T = [v_1\ v_2\ \dots\ v_n]^{-1} \in \mathbb{C}^{n\times n}$$

Then

$$AT^{-1} = [Av_1\ Av_2\ \dots\ Av_n] = [\lambda_1 v_1\ \lambda_2 v_2\ \dots\ \lambda_n v_n] = T^{-1}\Lambda$$

where $\lambda_i \in \mathbb{C}$ are the corresponding eigenvalues. Multiplying on the right by T leads to

$$A = T^{-1}\Lambda T.$$

($\Leftarrow$) Assume that there exists matrices $T \in \mathbb{C}^{n\times n}$ nonsingular and $\Lambda \in \mathbb{C}^{n\times n}$ diagonal such that

$$A = T^{-1}\Lambda T \Rightarrow AT^{-1} = T^{-1}\Lambda.$$

Let $T^{-1} = [w_1\ \dots w_n]$ where $w_i \in \mathbb{C}^n$ denoted the i^{th} column of T^{-1} and

$$\Lambda = \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n \end{bmatrix}$$

with $\sigma_i \in \mathbb{C}$. Then

$$Aw_i = \sigma_i w_i$$

and therefore w_i is a right eigenvector of A with eigenvalue σ_i. Since T^{-1} is invertible its columns (and eigenvectors of A) $\{w_i\}_{i=1}^n$ are linearly independent. ■

Now let us see how this fact helps is compute the state transition matrix e^{At}. Recall that e^{At} is related to the solution of the differential equation $\dot{x}(t) = Ax(t)$; in particular the solution of the differential equation starting at $x(0) = x_0$ can be written as $x(t) = e^{At}x_0$. Recall also that $A \in \mathbb{R}^{n\times n}$ can be thought of as the representation of some linear operator $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ with respect to some (usually the canonical) basis $\{e_i\}_{i=1}^n$. If A is semi-simple, then its eigenvectors are linearly independent and can also be used as a basis. Let us see what the representation of the linear operator $\mathcal{A}$ with respect to this basis is.

$$\begin{array}{lcl} (\mathbb{C}^n, \mathbb{C}) & \xrightarrow{\mathcal{A}} & (\mathbb{C}^n, \mathbb{C}) \\ \{e_i\}_{i=1}^n & \xrightarrow{A\in\mathbb{R}^{n\times n}} & \{e_i\}_{i=1}^n \text{ (basis leading to representation of } \mathcal{A} \text{ by } A) \\ \{v_i\}_{i=1}^n & \xrightarrow{\tilde{A}=TAT^{-1}=\Lambda} & \{v_i\}_{i=1}^n \text{ (eigenvector basis)} \end{array}$$

Recall that if $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is the representation of x with respect to the basis $\{e_i\}_{i=1}^n$ its representation with respect to the complex basis $\{v_i\}_{i=1}^n$ will be the complex vector $\tilde{x} = Tx \in \mathbb{C}^n$. The above formula simply states that

$$\dot{\tilde{x}} = T\dot{x} = TAx = TAT^{-1}\tilde{x} = \Lambda\tilde{x}.$$

Therefore, if A is semi-simple, its representation with respect to the basis of its eigenvectors is the diagonal matrix Λ of its eigenvalues. Notice that even though A is a real matrix, its representation Λ is in general complex, since the basis $\{v_i\}_{i=1}^n$ is also complex.

What about the state transition matrix?

Fact 5.1 *If A is semi-simple*

$$e^{At} = T^{-1}e^{\Lambda t}T = T^{-1}\begin{bmatrix} e^{\lambda_1 t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e^{\lambda_n t} \end{bmatrix} T \tag{5.4}$$

Proof: Exercise. Show that $A^k = T^{-1}\Lambda^k T$ and substitute into the expansion (5.3). ∎

In other words:

$$\begin{array}{rcl} (\mathbb{R}^n, \mathbb{R}) & \overset{s(t,0,\cdot,\theta_U)}{\longrightarrow} & (\mathbb{R}^n, \mathbb{R}) \\ x_0 & \longmapsto & x(t) \\ \{e_i\}_{i=1}^n & \overset{e^{At}\in\mathbb{R}^{n\times n}}{\longrightarrow} & \{e_i\}_{i=1}^n \text{ (basis leading to representation of } \mathcal{A} \text{ by } A) \\ \{v_i\}_{i=1}^n & \overset{e^{\Lambda t}=Te^{At}T^{-1}}{\longrightarrow} & \{v_i\}_{i=1}^n \text{ (eigenvector basis).} \end{array}$$

Once again, note that the matrices T, T^{-1} and $e^{\Lambda t}$ will general be complex. Fact 5.1, however, shows that when taking their product the imaginary parts will all cancel and we will be left with a real matrix.

Fact 5.1 shows that if a matrix is semi-simple the calculation of the matrix exponential is rather straightforward. It would therefore be desirable to establish conditions under which a matrix is semisimple.

Definition 5.2 *A matrix $A \in \mathbb{R}^{n\times n}$ is simple if and only if its eigenvalues are distinct, i.e. $\lambda_i \neq \lambda_j$ for all $i \neq j$.*

Theorem 5.3 *All simple matrices are semi-simple.*

Proof: Assume, for the sake of contradiction, that $A \in \mathbb{R}^{n\times n}$ is simple, but not semi-simple. Then $\lambda_i \neq \lambda_j$ for all $i \neq j$ but $\{v_i\}_{i=1}^n$ are linearly dependent in $(\mathbb{C}^n, \mathbb{C})$. Hence, there exist $a_1, \ldots, a_n \in \mathbb{C}$ not all zero, such that

$$\sum_{i=1}^n a_i v_i = 0.$$

Without loss of generality, assume that $a_1 \neq 0$ and multiply the above identity by $(A-\lambda_2 I)(A-\lambda_3 I)\ldots(A-\lambda_n I)$ on the left. Then

$$a_1(A-\lambda_2 I)(A-\lambda_3 I)\ldots(A-\lambda_n I)v_1 + \sum_{i=2}^n a_i(A-\lambda_2 I)(A-\lambda_3 I)\ldots(A-\lambda_n I)v_i = 0.$$

Concentrating on the first product and unraveling it from the right leads to

$$\begin{aligned} a_1(A-\lambda_2 I)(A-\lambda_3 I)\ldots(A-\lambda_n I)v_1 =& a_1(A-\lambda_2 I)(A-\lambda_3 I)\ldots(Av_1-\lambda_n v_1) \\ =& a_1(A-\lambda_2 I)(A-\lambda_3 I)\ldots(\lambda_1 v_1-\lambda_n v_1) \\ =& a_1(A-\lambda_2 I)(A-\lambda_3 I)\ldots v_1(\lambda_1-\lambda_n) \\ =& a_1(A-\lambda_2 I)(A-\lambda_3 I)\ldots(Av_1-\lambda_{n-1}v_1)(\lambda_1-\lambda_n) \\ & \text{etc.} \end{aligned}$$

which leads to

$$a_1(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)\ldots(\lambda_1-\lambda_n)v_1 + \sum_{i=2}^n a_i(\lambda_i-\lambda_2)(\lambda_i-\lambda_3)\ldots(\lambda_i-\lambda_n)v_i = 0.$$

Each term of the sum on the right will contain a term of the form $(\lambda_i - \lambda_i) = 0$. Hence the sum on the right is zero, leading to

$$a_1(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)\ldots(\lambda_1-\lambda_n)v_1 = 0.$$

But, $v_1 \neq 0$ (since it is an eigenvector) and $a_1 \neq 0$ (by the assumption that $\{v_i\}_{i=1}^n$ are linearly dependent) and $(\lambda_1 - \lambda_i) \neq 0$ for $i = 2, \ldots, n$ (since the eigenvalues are distinct). This leads to a contradiction. ∎

In addition to simple matrices, other matrices are semi-simple, for example diagonal matrices and orthogonal matrices. In fact one can show that semi-simple matrices are "dense" in $\mathbb{R}^{n\times n}$, in the sense that

$$\forall A \in \mathbb{R}^{n\times n},\ \forall \epsilon > 0,\ \exists A' \in \mathbb{R}^{n\times n} \text{ semi-simple } : \|A - A'\| < \epsilon.$$

(Here $\|\cdot\|$ denotes any matrix norm, they are all equivalent). The reason is that for a matrix to be non semi-simple the matrix of its eigenvectors $T \in \mathbb{C}^{n\times n}$ must be singular, i.e. we must have $\textsc{Det}[T] = 0$. But this is a fragile condition, as arbitrary small perturbations of the matrix A will in general lead to $\textsc{Det}[T] \neq 0$. In fact it is not hard to convince ourselves that "almost all" matrices in $\mathbb{R}^{n\times n}$ are semi-simple, since the condition $\textsc{Det}[T] = 0$ imposes a single constraint on the n^2 dimensional space of complex matrices[1].

In summary, almost all matrices are semi-simple, though not all.

Example (Non semi-simple matrices) The matrix

$$A_1 = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$$

is not semi-simple. Its eigenvalues are $\lambda_1 = \lambda_2 = \lambda_3 = \lambda$ (hence it is not simple) but there is only one eigenvector

$$\begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \Rightarrow \left\{ \begin{array}{c} \lambda x_1 + x_2 = \lambda x_1 \\ \lambda x_2 + x_3 = \lambda x_2 \\ \lambda x_3 = \lambda x_3 \end{array} \right\} \Rightarrow v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix};$$

all other eigenvectors will be multiples of v_1. Note that A_1 has the same eigenvalues as the matrices

$$A_2 = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \text{ and } A_3 = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}.$$

A_2 is also not semi-simple, it has only two linearly independent eigenvectors $v_1 = (1, 0, 0)$ and $v_3 = (0, 0, 1)$. A_3 is semi-simple, with the canonical basis as eigenvectors. Notice that none of the three matrices is simple. ■

5.3 Jordan form

Can non semi-simple matrices be written in a form that simplifies the computation of their state transition matrix e^{At}? This is possible through a change of basis that involves their eigenvectors. The problem is that since in this case there are not enough linearly independent eigenvectors, the basis needs to be completed with additional vectors. It turns out that there is a special way of doing this using the so called generalized eigenvectors, so that the representation of the matrix A in the resulting basis has a particularly simple form, the so-called Jordan canonical form.

To start with, notice that for the eigenvector $v_i \in \mathbb{C}^n$ corresponding to eigenvalue $\lambda_i \in \mathbb{C}$ of $A \in \mathbb{R}^{n\times n}$

$$Av_i = \lambda_i v_i \Leftrightarrow (A - \lambda_i I)v_i = 0 \Leftrightarrow v_i \in \textsc{Null}[A - \lambda_i I].$$

Recall that $\textsc{Null}[A - \lambda_i I]$ is a subspace of $\mathbb{C}^n$.

Definition 5.3 *The algebraic mutiplicity of an eigenvalue $\lambda \in \mathbb{C}$ of a matrix $A \in \mathbb{R}^{n\times n}$ is the number of times λ appears in the spectrum* $\textsc{Spec}[A]$. *The geometric multiplicity of λ is the dimension of* $\textsc{Null}[A - \lambda I]$.

[1] As an aside, these statements can also be made for simple and invertible matrices.

Example (Non semi-simple matrices (cont.)) For the matrices considered above

$$\begin{aligned}
&\textsc{Null}[A_1 - \lambda I] = \textsc{Span}\{(1,0,0)\} \Rightarrow \textsc{Dim}[\textsc{Null}[A_1 - \lambda I]] = 1\\
&\textsc{Null}[A_2 - \lambda I] = \textsc{Span}\{(1,0,0),(0,0,1)\} \Rightarrow \textsc{Dim}[\textsc{Null}[A_2 - \lambda I]] = 2\\
&\textsc{Null}[A_3 - \lambda I] = \textsc{Span}\{(1,0,0),(0,1,0),(0,0,1)\} \Rightarrow \textsc{Dim}[\textsc{Null}[A_3 - \lambda I]] = 3.
\end{aligned}$$

Notice that in all three cases the algebraic multiplicity of the eigenvalue λ is 3. The problem with A_1 and A_2 is that the geometric multiplicity of λ is smaller than its algebraic multiplicity. This implies that there are not enough eigenvectors associated with eigenvalue λ (in other words linearly independent vectors in the null space) to form a basis. ■

The above example suggests that the reason matrices fail to be semi-simple is a discrepancy between their algebraic and geometric multiplicities. To alleviate this problem we need to complete the basis with additional vectors. For this purpose we consider the so called generalized eigenvectors.

Definition 5.4 *A Jordan chain of length $\mu \in \mathbb{N}$ at eigenvalue $\lambda \in \mathbb{C}$ is a family of vectors $\{v^j\}_{j=1}^{\mu} \subseteq \mathbb{C}^n$ such that*

1. *$\{v^j\}_{j=1}^{\mu}$ are linearly independent, and*
2. *$[A - \lambda I]v^1 = 0$ and $[A - \lambda I]v^j = v^{j-1}$ for $j = 2, \ldots, \mu$.*

A Jordan chain $\{v^j\}_{j=0}^{\mu}$ is called maximal if it cannot be extended, i.e. there does not exist $v \in \mathbb{C}^n$ linearly independent from $\{v^j\}_{j=1}^{\mu}$ such that $[A-\lambda I]v = v^{\mu}$. The elements of all the maximal Jordan chains at λ are the generalized eigenvectors of λ.

Fact 5.2 *Let $\{v^j\}_{j=0}^{\mu} \subseteq \mathbb{C}^n$ be a Jordan chain of length μ at eigenvalue $\lambda \in \mathbb{C}$ of the matrix $A \in \mathbb{R}^{n\times n}$:*

1. *v^1 is an eigenvector of A with eigenvalue λ.*
2. *$v^j \in \textsc{Null}[(A - \lambda I)^j]$ for $j = 1, \ldots, \mu$.*
3. *$\textsc{Null}[(A - \lambda I)^j] \subseteq \textsc{Null}[(A - \lambda I)^{j+1}]$ for any $j = 1, 2, \ldots$.*

Proof: Part 1: By definition $[A - \lambda I]v^1 = v^0 = 0$, hence $Av^1 = \lambda v^1$.

Part 2: $[A - \lambda I]^j v^j = [A - \lambda I]^{j-1} v^{j-1} = \ldots = [A - \lambda I]^1 v^1 = v^0 = 0$.

Part 3: Let $v \in \textsc{Null}[(A - \lambda I)^j]$, i.e. $[A - \lambda I]^j v = 0$. Then $[A - \lambda I]^{j+1} v = 0$ and hence $v \in \textsc{Null}[(A - \lambda I)^{j+1}]$. ■

Example (Non semi-simple matrices (cont.)) For the matrices considered above, A_1 has one maximal Jordan chain of length $\mu = 3$ at λ, with

$$v^1 = (1,0,0), \;\; v^2 = (0,1,0), \;\; v^3 = (0,0,1).$$

A_2 has two maximal Jordan chains at λ, one of length $\mu_1 = 2$ and the other of length $\mu_2 = 1$, with

$$v_1^1 = (1,0,0), \;\; v_1^2 = (0,1,0), \;\; v_2^1 = (0,0,1).$$

Finally, A_3 has three maximal Jordan chains at λ each of length $\mu_1 = \mu_2 = \mu_3 = 1$, with

$$v_1^1 = (1,0,0), \;\; v_2^1 = (0,1,0), \;\; v_3^1 = (0,0,1).$$

Notice that in all three cases the generalized eigenvectors are the same, but they are partitioned differently into chains. Note also that in all three cases the generalized eigenvectors taken all together form a linearly independent family of $n = 3$ vectors. ■

It can be shown that the last observation is not a coincidence: The collection of all generalised eigenvectors is always linearly independent.

Lemma 5.1 *Assume that the matrix $A \in \mathbb{R}^{n\times n}$ has k linearly independent eigenvectors $v_1, \dots, v_k \in \mathbb{C}^n$ with corresponding maximal Jordan chains $\{v_i^j\}_{j=0}^{\mu_i} \subseteq \mathbb{C}^n$, $i = 1, \dots, k$. Then the matrix $\begin{bmatrix} v_1^1 & \dots & v_1^{\mu_1} & \dots & v_k^1 & \dots & v_k^{\mu_k} & \dots \end{bmatrix} \in \mathbb{C}^{n\times n}$ is invertible. In particular, $\sum_{i=1}^k \mu_i = n$.*

The proof of this fact is rather tedious and will be omitted, see [19]. Consider now a change of basis

$$T = \begin{bmatrix} v_1^1 & \dots & v_1^{\mu_1} & v_2^1 & \dots & v_2^{\mu_2} & \dots \end{bmatrix}^{-1} \in \mathbb{C}^{n\times n} \tag{5.5}$$

comprising the generalised eigenvectors as the columns of the matrix T^{-1}.

Theorem 5.4 *With the definition of T in equation (5.5), the matrix $A \in \mathbb{R}^{n\times n}$ can be written as $A = T^{-1}JT$ where $J \in \mathbb{C}^{n\times n}$ is block-diagonal*

$$J = \begin{bmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & J_k \end{bmatrix} \in \mathbb{C}^{n\times n}, \quad J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_i & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_i & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda_i \end{bmatrix} \in \mathbb{C}^{\mu_i\times\mu_i}, \quad i = 1, \dots, k$$

and $\lambda_i \in \mathbb{C}$ is the eigenvalue corresponding to the Jordan chain $\{v_i^j\}_{j=0}^{\mu_i}$.

Notice that there may be multiple Jordan chains for the same eigenvalue λ, in fact their number will be the same as the number of linearly independent eigenvectors associated with λ. If $k = n$ (equivalently, all Jordan chains have length 1) then the matrix is semi-simple, T^{-1} is the matrix of eigenvectors of A, and $J = \Lambda$.

The theorem demonstrates that any matrix can be brought into a special, block diagonal form using its generalised eigenvectors as a change of basis. This special block diagonal form is known as the Jordan canonical form.

Definition 5.5 *The block diagonal matrix J in Theorem 5.4 is the called the Jordan canonical form of the matrix A. The matrices J_i are known as the Jordan blocks of the matrix A.*

Example (Non semi-simple matrices (cont.)) In the above example, the three matrices A_1, A_2 and A_3 are already in Jordan canonical form. A_1 comprises one Jordan block of size 3, A_2 two Jordan blocks of sizes 2 and 1 and A_3 three Jordan blocks, each of size 1. ■

How does this help with the computation of e^{At}?

Theorem 5.5 $e^{At} = T^{-1}e^{Jt}T$ *where*

$$e^{Jt} = \begin{bmatrix} e^{J_1 t} & 0 & \dots & 0 \\ 0 & e^{J_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{J_k t} \end{bmatrix} \text{ and } e^{J_i t} = \begin{bmatrix} e^{\lambda_i t} & te^{\lambda_i t} & \frac{t^2 e^{\lambda_i t}}{2!} & \dots & \frac{t^{\mu_i-1} e^{\lambda_i t}}{(\mu_i-1)!} \\ 0 & e^{\lambda_i t} & te^{\lambda_i t} & \dots & \frac{t^{\mu_i-2} e^{\lambda_i t}}{(\mu_i-2)!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda_i t} \end{bmatrix}, \quad i = 1, \dots, k.$$

Proof: Exercise. Show that $A^j = T^{-1}J^jT$, then that

$$J^j = \begin{bmatrix} J_1^j & 0 & \dots & 0 \\ 0 & J_2^j & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & J_k^j \end{bmatrix}, \text{ and hence that } e^{Jt} = \begin{bmatrix} e^{J_1 t} & 0 & \dots & 0 \\ 0 & e^{J_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{J_k t} \end{bmatrix}.$$

Finally show that

$$e^{J_i t} = \begin{bmatrix} e^{\lambda_i t} & te^{\lambda_i t} & \frac{t^2 e^{\lambda_i t}}{2!} & \dots & \frac{t^{\mu_i - 1} e^{\lambda_i t}}{(\mu_i - 1)!} \\ 0 & e^{\lambda_i t} & te^{\lambda_i t} & \dots & \frac{t^{\mu_i - 2} e^{\lambda_i t}}{(\mu_i - 2)!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda_i t} \end{bmatrix}$$

by differentiating with respect to t, showing that the result is equal to $J_i e^{J_i t}$ and invoking uniqueness (or in a number of other ways). ∎

So the computation of the matrix exponential becomes easy once again. Notice that if $k = n$ then we are back to the semi-simple case. As in the semi-simple case, all matrices involved in the product will in general be complex. However, the theorem ensures that when taking the product the imaginary parts will cancel and the result will be a real matrix.

Example (Non semi-simple matrices (cont.)) In the above example,

$$e^{A_1 t} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2} e^{\lambda t} \\ 0 & e^{\lambda t} & te^{\lambda t} \\ 0 & 0 & e^{\lambda t} \end{bmatrix}, \quad e^{A_2 t} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} & 0 \\ 0 & e^{\lambda t} & 0 \\ 0 & 0 & e^{\lambda t} \end{bmatrix}, \quad e^{A_3 t} = \begin{bmatrix} e^{\lambda t} & 0 & 0 \\ 0 & e^{\lambda t} & 0 \\ 0 & 0 & e^{\lambda t} \end{bmatrix},$$

∎

Notice that in all cases e^{At} consists of linear combinations of elements of the form

$$e^{\lambda_i t}, te^{\lambda_i t}, \dots, t^{\mu_i - 1} e^{\lambda_i t}$$

for $\lambda_i \in \text{SPEC}[A]$ and μ_i the length of the longest Jordan chain at λ_i. In other words

$$e^{At} = \sum_{\lambda \in \text{SPEC}[A]} \Pi_\lambda(t) e^{\lambda t} \tag{5.6}$$

where for each $\lambda \in \text{SPEC}[A]$, $\Pi_\lambda(t) \in \mathbb{C}[t]^{n \times n}$ is a matrix of polynomials of t with complex coefficients and degree at most equal to the length of the longest Jordan chain at λ. In particular, if A is semi-simple all Jordan chains have length equal to 1 and the matrix exponential reduces to

$$e^{At} = \sum_{\lambda \in \text{SPEC}[A]} \Pi_\lambda e^{\lambda t}$$

where $\Pi_\lambda \in \mathbb{C}^{n \times n}$ are constant complex matrices. Notice again that even though in general both the eigenvalues λ and the coefficients of the corresponding polynomials $\Pi_\lambda(t)$ are complex numbers, because the eigenvalues appear in complex conjugate pairs the imaginary parts for the sum cancel out and the result is the real matrix $e^{At} \in \mathbb{R}^{n \times n}$.

5.4 Laplace transforms

To establish a connection to more conventional control notation, we recall the definition of the Laplace transform of a signal $f(\cdot) : \mathbb{R}_+ \to \mathbb{R}^{n \times m}$ mapping the non-negative real numbers to the

linear space of $n \times m$ real matrices:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty f(t)e^{-st}dt \in \mathbb{C}^{n\times m}$$

where the integral is interpreted element by element and we assume that it is well defined (for a careful discussion of this point see, for example, [7]). The Laplace transform $\mathcal{L}\{f(t)\}$ transforms the real matrix valued function $f(t) \in \mathbb{R}^{n\times m}$ of the real number $t \in \mathbb{R}_+$ to the complex matrix valued function $F(s) \in \mathbb{C}^{n\times m}$ of the complex number $s \in \mathbb{C}$. The inverse Laplace transform $\mathcal{L}^{-1}\{F(s)\}$ performs the inverse operation; it can also be expressed as an integral, even though in the calculations considered here one mostly encounters functions $F(s)$ that are recognisable Laplace transforms of known functions $f(t)$; in particular the functions $F(s)$ will typically be proper rational functions of s whose inverse Laplace transform can be computed by partial fraction expansion.

Fact 5.3 *The Laplace transform (assuming that it is defined for all functions concerned) has the following properties:*

1. *It is linear, i.e. for all $A_1, A_2 \in \mathbb{R}^{p\times n}$ and all $f_1(\cdot) : \mathbb{R}_+ \to \mathbb{R}^{n\times m}$, $f_2(\cdot) : \mathbb{R}_+ \to \mathbb{R}^{n\times m}$*

$$\mathcal{L}\{A_1f_1(t) + A_2f_2(t)\} = A_1\mathcal{L}\{f_1(t)\} + A_2\mathcal{L}\{f_2(t)\} = A_1F_1(s) + A_2F_2(s)$$

2. *$\mathcal{L}\left\{\frac{d}{dt}f(t)\right\} = sF(s) - f(0)$.*
3. *$\mathcal{L}\{(f * g)(t)\} = F(s)G(s)$ where $(f * g)(\cdot) : \mathbb{R}_+ \to \mathbb{R}^{p\times m}$ denotes the convolution of $f(\cdot) : \mathbb{R}_+ \to \mathbb{R}^{p\times n}$ and $g(\cdot) : \mathbb{R}_+ \to \mathbb{R}^{n\times m}$ defined by*

$$(f * g)(t) = \int_0^t f(t-\tau)g(\tau)d\tau.$$

Proof: Exercise, just use the definition and elementary calculus. ∎

Fact 5.4 *For all $A \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{R}_+$, $\mathcal{L}\left\{e^{At}\right\} = (sI - A)^{-1}$.*

Proof: Recall that

$$\begin{aligned}
\frac{d}{dt}e^{At} = Ae^{At} &\Rightarrow \mathcal{L}\left\{\frac{d}{dt}e^{At}\right\} = \mathcal{L}\left\{Ae^{At}\right\} \\
&\Rightarrow s\mathcal{L}\left\{e^{At}\right\} - e^{A0} = A\mathcal{L}\left\{e^{At}\right\} \\
&\Rightarrow s\mathcal{L}\left\{e^{At}\right\} - I = A\mathcal{L}\left\{e^{At}\right\} \\
&\Rightarrow (sI - A)\mathcal{L}\left\{e^{At}\right\} = I.
\end{aligned}$$

The claim follows by multiplying on the left by $(sI - A)^{-1}$; notice that the matrix is invertible for all $s \in \mathbb{C}$, except the eigenvalues of A. ∎

Let us look somewhat more closely to the structure of the Laplace transform, $(sI - A)^{-1}$, of the state transition matrix. This is an $n \times n$ matrix of strictly proper rational functions of s, which as we saw in Chapter 2 form a sub-ring of $(\mathbb{R}_p(s), +, \cdot)$. To see this recall that by definition

$$(sI - A)^{-1} = \frac{\text{Adj}[sI - A]}{\text{Det}[sI - A]}$$

The denominator is simply the characteristic polynomial,

$$\chi_A(s) = s^n + \chi_1 s^{n-1} + \ldots + \chi_n \in \mathbb{R}[s]$$

of the matrix A, where $\chi_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

In the numerator is the "adjoint" of the matrix $(sI - A)$. Recall that the (i, j) element of this $n \times n$ matrix is equal to the determinant (together with the corresponding sign $(-1)^i(-1)^j$) of the sub-matrix of $(sI - A)$ formed by removing its j^{th} row and the i^{th} column. Since $sI - A$ has all terms in s on the diagonal when we eliminate one row and one column of $sI - A$ we eliminate at least one term containing s. Therefore the resulting sub-determinants will be polynomials of order at most $n - 1$ in s, in other words

$$\textsc{Adj}[sI - A] \in \mathbb{R}[s]^{n\times n} \text{ and } (sI - A)^{-1} \in \mathbb{R}_p(s)^{n\times n}.$$

Given this structure, let us write $(sI - A)^{-1}$ more explicitly as

$$(sI - A)^{-1} = \frac{M(s)}{\chi_A(s)} \tag{5.7}$$

where $M(s) = M_0 s^{n-1} + \ldots + M_{n-2}s + M_{n-1}$ with $M_i \in \mathbb{R}^{n\times n}$ for $i = 0, \ldots, n - 1$.

Theorem 5.6 *The matrices M_i satisfy*

$$\begin{aligned} M_0 &= I \\ M_i &= M_{i-1}A + \chi_i I \quad \textit{for } i = 1, \ldots, n-1 \\ M_{n-1}A + \chi_n I &= 0. \end{aligned}$$

Proof: Post multiplying equation (5.7) by $(sI - A)\chi_A(s)$ leads to

$$\chi_A(s)I = M(s)(sI - A) \Rightarrow (s^n + \chi_1 s^{n-1} + \ldots + \chi_n)I = \left(M_0 s^{n-1} + \ldots + M_{n-2}s + M_{n-1}\right)(sI - A).$$

Since the last identity must hold for all $s \in \mathbb{C}$ the coefficients of the two polynomials on the left and on the right must be equal. Equating the coefficients for s^n leads to the formula for M_0, for s^{n-1} leads to the formula for M_1, etc. ∎

The theorem provides an easily implementable algorithm for computing the Laplace transform of the state transition matrix without having to invert any matrices. The only thing that is needed is the computation of the characteristic polynomial of A and some matrix multiplications. In addition, the following useful fact about square matrices can be deduced as a corollary.

Theorem 5.7 *(Cayley-Hamilton) Every square matrix $A \in \mathbb{R}^{n\times n}$ satisfies its characteristic polynomial, i.e.*

$$\chi_A(A) = A^n + \chi_1 A^{n-1} + \ldots + \chi_n I = 0 \in \mathbb{R}^{n\times n}.$$

Proof: By the last equation of Theorem 5.6, $M_{n-1}A + \chi_n I = 0$. From the next to last equation, $M_{n-1} = M_{n-2}A + \chi_{n-1}I$; substituting this into the last equation leads to

$$M_{n-2}A^2 + \chi_{n-1}A + \chi_n I = 0.$$

Substituting M_{n-2} from the third to last equation, etc. leads to the claim. ∎

The Cayley-Hamilton Theorem has a number of interesting and useful consequences. We state two of these here and will return to them in Chapter 8.

Corollary 5.2 *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. For any $k \in \mathbb{N}$, A^k can be written as a linear combination of $\{I, A, A^2, \ldots, A^{n-1}\}$.*

The proof is left as an exercise. Stated another way, the corollary shows that all powers of an n-dimensional square matrix live in a low-dimansional subspace of the n^2 dimensional linear space of square matrices; this subspace has dimension at most n and is spanned by the matrices $\{I, A, A^2, \ldots, A^{n-1}\}$. To state the second corollary we first recall the definition of a nilpotent matrix.

Definition 5.6 *A matrix $A \in \mathbb{R}^{n\times n}$ is nilpotent if and only if $A^N = 0$ for some $N \in \mathbb{N}$.*

Corollary 5.3 *The following statements are equivalent:*

1. *$A \in \mathbb{R}^{n\times n}$ is nilpotent.*
2. *$A^n = 0$.*
3. *$\text{SPEC}[A] = \{0, \ldots, 0\}$.*

Proof: $2 \Rightarrow 1$: Obvious, simply take $N = n$ in the definition of nilpotent matrix.

$3 \Rightarrow 2$: Note that if $\text{SPEC}[A] = \{0, \ldots, 0\}$ the characteristic polynomial of A is $\chi_A(\lambda) = \lambda^n$. By the Cayley-Hamilton Theorem, $\chi_A(A) = A^n = 0$.

$1 \Rightarrow 3$: By contraposition. By Theorem 5.4 A can be written in Jordan canonical form $A = T^{-1}JT$. It is easy to see that $A^N = T^{-1}J^NT$, therefore, since T is invertible, $A^N = 0$ if and only if $J^N = 0$. Moreover,

$$J^N = \begin{bmatrix} J_1^N & 0 & \ldots & 0 \\ 0 & J_2^j & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & J_k^N \end{bmatrix}$$

hence $J^N = 0$ if and only if $J_i^N = 0$ for each $i = 1, \ldots, k$. Finally, since each particular Jordan block J_i, $i = 1, \ldots, k$ is upper triangular it is easy to see that

$$J_i^N = \begin{bmatrix} \lambda_i^N & * & * & \ldots & * & * \\ 0 & \lambda_i^N & * & \ldots & * & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_i^N & * \\ 0 & 0 & 0 & \ldots & 0 & \lambda_i^N \end{bmatrix},$$

where $*$ stands for some complex number (possibly equal to zero). This implies that A^N cannot be zero unless each $\lambda_i = 0$. ∎

The Laplace transform can also be used to compute the response of the system. Notice that taking Laplace transforms of both sides of the differential equation governing the evolution of $x(t)$ leads to

$$\dot{x}(t) = Ax(t) + Bu(t) \overset{\mathcal{L}}{\Longrightarrow} sX(s) - x_0 = AX(s) + BU(s).$$

Hence

$$X(s) = (sI - A)^{-1}x_0 + (sI - A)^{-1}BU(s). \tag{5.8}$$

How does this relate to the solution of the differential equation that we have already derived? We have shown that

$$x(t) = e^{At}x_0 + \int_0^t e^{At-\tau}Bu(\tau)d\tau = e^{At}x_0 + (K * u)(t)$$

where $(K * u)(t)$ denotes the convolution of the impulse state transition

$$K(t) = e^{At}B$$

with the input $u(t)$. Taking Laplace transforms we obtain

$$\begin{aligned} X(s) = \mathcal{L}\{x(t)\} =& \mathcal{L}\{e^{At}x_0 + (K * u)(t)\} \\ =& \mathcal{L}\{e^{At}\}x_0 + \mathcal{L}\{(K * u)(t)\} \\ =& (sI - A)^{-1}x_0 + \mathcal{L}\{e^{At}B\}\mathcal{L}\{u(t)\} \\ =& (sI - A)^{-1}x_0 + (sI - A)^{-1}BU(s) \end{aligned}$$

which coincides with equation (5.8), as expected.

Equation (5.8) provides an indirect, purely algebraic way of computing the solution of the differential equation, without having to compute a matrix exponential or a convolution integral. One can form the Laplace transform of the solution, $X(s)$, by inverting $(sI - A)$ (using, for example, the matrix multiplication of Theorem 5.6) and substituting into equation (5.8). From there the solution $x(t)$ can be computed by taking an inverse Laplace transform. Since $(sI-A)^{-1} \in \mathbb{R}_p(s)^{n\times n}$ is a matrix of strictly proper rational functions, $X(s) \in \mathbb{R}_p(s)^n$ will be a vector of strictly proper rational functions, with the characteristic polynomial in the denominator. The inverse Laplace transform can therefore be computed by partial fraction expansions, at least for many reasonable input functions (constants, sines and cosines, exponentials, ramps, polynomials, and combinations thereof).

Taking the Laplace transform of the output equation leads to

$$y(t) = Cx(t) + Du(t) \overset{\mathcal{L}}{\Longrightarrow} Y(s) = CX(s) + DU(s).$$

By (5.8)

$$Y(s) = C(sI - A)^{-1}x_0 + (sI - A)^{-1}BU(s) + DU(s)$$

which for $x_0 = 0$ (zero state response) reduces to

$$Y(s) = C(sI - A)^{-1}BU(s) + DU(s) = G(s)U(s). \tag{5.9}$$

Definition 5.7 *The function $G(\cdot) : \mathbb{C} \to \mathbb{C}^{p\times m}$ defined by*

$$G(s) = C(sI - A)^{-1}B + D \tag{5.10}$$

is called the transfer function of the system.

Comparing equation (5.9) with the zero state response that we computed earlier

$$y(t) = C\int_0^t e^{A(t-\tau)}Bu(\tau)d\tau + Du(t) = (H * u)(t)$$

it is clear that the transfer function is the Laplace transform of the impulse response $H(t)$ of the system

$$G(s) = \mathcal{L}\{H(t)\} = \mathcal{L}\{Ce^{At}B + D\delta_0(t)\} = C(sI - A)^{-1}B + D.$$

Substituting equation (5.7) into (5.10) we obtain

$$G(s) = C\frac{M(s)}{\chi_A(s)}B + D = \frac{CM(s)B + D\chi_A(s)}{\chi_A(s)}. \tag{5.11}$$

Since $K(s)$ is a matrix of polynomials of degree at most $n - 1$ and $\chi_A(s)$ is a polynomial of degree n we see that $G(s) \in \mathbb{R}_p(s)^{p\times m}$ is a matrix of proper rational functions. If moreover $D = 0$ then the rational functions are strictly proper.

Definition 5.8 *The poles of the system are the values of $s \in \mathbb{C}$ are the roots of the denominator polynomial of $G(s)$.*

From equation (5.11) it becomes apparent that all poles of the system are eigenvalues (i.e. are contained in the spectrum) of the matrix A. Note, however, that not all eigenvalues of A are necessarily poles of the system, since there may be cancellations of common factors in the numerator and denominator when forming the fraction (5.11). It turns out that such cancellations are related to the controllability and observability properties of the system. We will return to this point in Chapter 8, after introducing these notions.

Problems for chapter 5

Problem 5.1 (Change of basis) Let $\{u_i\}_{i=1}^m, \{x_i\}_{i=1}^n, \{y_i\}_{i=1}^p$ be bases of the linear spaces $(\mathbb{R}^m, \mathbb{R})$, $(\mathbb{R}^n, \mathbb{R})$ and $(\mathbb{R}^p, \mathbb{R})$, respectively. Let $u(\cdot) : \mathbb{R}_+ \to \mathbb{R}^m$ be piecewise continuous. For $t \in \mathbb{R}_+$, consider the linear time-invariant system:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases} \tag{5.12}$$

with $x(t_0) = x_0$, where all matrix representations are given w.r.t. $\{u_i\}_{i=1}^m, \{x_i\}_{i=1}^n, \{y_i\}_{i=1}^p$. Now let $\{\widetilde{x}_i\}_{i=1}^n$ be another basis of $(\mathbb{R}^n, \mathbb{R})$ and let $T \in \mathbb{R}^{n\times n}$ represent the change of basis from $\{x_i\}_{i=1}^n$ to $\{\widetilde{x}_i\}_{i=1}^n$.

1. Derive the representation of the system w.r.t. bases $\{u_i\}_{i=1}^m, \{\widetilde{x}_i\}_{i=1}^n, \{y_i\}_{i=1}^p$.
2. Compute the transition map $\widetilde{\Phi}(t, t_0)$ and the impulse response matrix $\widetilde{H}(t, \tau)$ w.r.t. the new representation. How do they compare with the corresponding quantities $\Phi(t, t_0)$ and $H(t, \tau)$ in the original representation?

Problem 5.2 (Time-invariant Systems) Consider the linear time-invariant system of Problem 5.1.

1. Show that
$$\Phi(t, t_0) = \exp\big(A(t - t_0)\big) = \sum_{k=0}^{\infty} \frac{\big(A(t - t_0)\big)^k}{k!}.$$
2. Given two matrices $A_1, A_2 \in \mathbb{R}^{n\times n}$ show that, if $A_1A_2 = A_2A_1$, then $A_2 \exp(A_1t) = \exp(A_1t)A_2$ and $\exp\big((A_1 + A_2)t\big) = \exp(A_1t)\exp(A_2t)$. Also show that these properties may not hold if $A_1A_2 \neq A_2A_1$.
3. Show that the impulse response matrix satisfies $H(t, \tau) = H(t - \tau, 0)$ (i.e. it depends only on the difference $t - \tau$).

Problem 5.3 (Discretization of Continuous-time Systems)

1. Consider the *time-varying* linear system
$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t) + D(t)u(t), \end{aligned} \tag{$\triangleleft$}$$
with initial condition $x(t_0) = x_0 \in \mathbb{R}^n$, $t \geq t_0$. Consider a set of time instants t_k, with $k = 0, 1, 2, \dots$, such that $t_k < t_{k+1}$ for all k. Let $u(t)$ be constant between subsequent time instants: $u(t) = u_k$ $\forall k \in \mathbb{N}$, $\forall t \in [t_k, t_{k+1})$. Let $\bar{x}_{k+1} = x(t_{k+1})$ and $\bar{y}_{k+1} = y(t_{k+1})$ be the state and the output of system ($\triangleleft$) *sampled* at times t_k. Show that there exist matrices $\bar{A}_k$, $\bar{B}_k$, $\bar{C}_k$ and $\bar{D}_k$ such that
$$\begin{aligned} \bar{x}_{k+1} &= \bar{A}_k\bar{x}_k + \bar{B}_ku_k, \\ \bar{y}_k &= \bar{C}_k\bar{x}_k + \bar{D}_ku_k. \end{aligned} \tag{$\square$}$$

2. Now assume that ($\triangleleft$) is *time-invariant*, i.e. $\big(A(t), B(t), C(t), D(t)\big) = (A, B, C, D)$, $\forall t \geq t_0$, and that there exists a fixed $T > 0$, $t_{k+1} - t_k = T$, $\forall k$. Provide simplified expressions for $\bar{A}_k$, $\bar{B}_k$, $\bar{C}_k$ and $\bar{D}_k$ and show that they are independent of k.

Problem 5.4 (Realization) Consider the following n-th order scalar differential equation with constant coefficients:

$$y^{(n)}(t) + a_1 y^{(n-1)}(t) + \ldots + a_{n-1} y^{(1)}(t) + a_n y(t) = u(t), \quad t \in \mathbb{R}_+, \tag{5.13}$$

where $y^{(i)}(t)$ denotes the i-th derivative of y at t, $\{a_i\} \subset \mathbb{R}$ and $u(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is a piecewise continuous input. Show that (5.13) can be put in the form (5.12) for an appropriate definition of the state $x(t) \in \mathbb{R}^n$ and of matrices A, B, C, D.

Problem 5.5 (Jordan Blocks) Let $\lambda, \lambda_1, \lambda_2 \in F$, with $F = \mathbb{R}$ or $F = \mathcal{C}$. Compute $\exp(A \cdot t)$ for the following definitions of matrix A:

$$1.\quad A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}; \qquad 2.\quad A = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}; \qquad 3.\quad A = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ & & & \lambda \end{bmatrix} \in F^{n \times n},$$

where the elements not shown are zeroes. (Hint: in 3, consider the decomposition $A = \lambda I + N$, and make use of $(\lambda I + N)^k = \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} (\lambda I)^i \cdot N^{k-i}$.)

Problem 5.6 (Jordan blocks and matrix exponential) For $i = 1, \ldots, m$, let $\Lambda_i \in \mathcal{C}^{n_i \times n_i}$. Define $n = n_1 + \ldots + n_m$ and the block diagonal matrix

$$\Lambda = \operatorname{diag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_m) = \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_m \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{$\circ$}$$

1. Show that $\exp(\Lambda) = \operatorname{diag}\big(\exp(\Lambda_1), \exp(\Lambda_2), \ldots, \exp(\Lambda_m)\big)$.
2. Compute $\exp(\Lambda_i)$ for the following definitions of Λ_i: (assume each entry is real)

$$\text{(a)}\quad \Lambda_i = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{n_i} \end{bmatrix}; \qquad \text{(b)}\quad \Lambda_i = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ & & & \lambda \end{bmatrix};$$

$$\text{(c)}\quad \Lambda_i = \begin{bmatrix} \omega & \sigma \\ -\sigma & \omega \end{bmatrix} \in \mathbb{R}^{2 \times 2},$$

   where the elements not shown are zeroes. [Hint: in (b), consider the decomposition $\Lambda = \lambda I + N$, and make use of $(\lambda I + N)^k = \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} (\lambda I)^i \cdot N^{k-i}$.]

3. Assume that $v = x + iy$ and $v^* = x - iy$, with $x, y \in \mathbb{R}^n$, are complex eigenvectors of a matrix $A \in \mathbb{R}^{n \times n}$, i.e. $Av = \lambda v$ and $Av^* = \lambda^* v^*$ for some $\lambda = \sigma + i\omega$ with $\sigma, \omega \in \mathbb{R}$. Let ($\circ$), with $\Lambda_1 \in \mathcal{C}^{2 \times 2}$, be the Jordan decomposition of $A \in \mathbb{R}^{n \times n}$ corresponding to a basis of the form $\{v, v^*, v_3, v_4, \ldots v_n\}$.
   (a) Write the expression of Λ_1.
   (b) Find a new basis and the corresponding change of basis T such that $T\Lambda T^{-1} = \operatorname{diag}(\widetilde{\Lambda}_1, \Lambda_2, \ldots, \Lambda_m)$ with $\widetilde{\Lambda}_1 \in \mathbb{R}^{2 \times 2}$ (real). What is the expression of $\widetilde{\Lambda}_1$?

Problem 5.7 (Modal Analysis) For $t \in \mathbb{R}_+$, consider the ODE $\dot{x} = Ax$, $x(0) = x_0$. Compute $\exp(At)$ for the cases listed below. In each case provide a rough plot of the parametric curves $x_2(t)$ vs. $x_1(t)$ for some initial conditions x_0.

1. $A = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix}$, for $\omega < 0$ and $\omega > 0$.
2. $A = \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix}$, for $\sigma < 0$ and $\sigma > 0$.
3. $A = \begin{bmatrix} \lambda_1 & \lambda_2 - \lambda_1 \\ 0 & \lambda_2 \end{bmatrix}$, for $\lambda_1 < 0 < \lambda_2$, $\lambda_1, \lambda_2 < 0$ and $\lambda_1, \lambda_2 > 0$.
4. $A = \begin{bmatrix} \lambda - 1 & 1 \\ -1 & \lambda + 1 \end{bmatrix}$, for $\lambda < 0$, $\lambda > 0$ and $\lambda = 0$.

Problem 5.8 (Matrix powers) Consider $A \in \mathbb{R}^{n \times n}$, $C \in \mathbb{R}^{p \times n}$ and $B \in \mathbb{R}^{n \times m}$. Show that

1. For any $k \in \mathbb{N}$, A^k can be written as a linear combination of $\{I, A, A^2, \ldots, A^{n-1}\}$.
2. $CA^kB = 0$ for all $k \in \mathbb{N}$ if and only if $CA^kB = 0$ for $k = 0, 1, \ldots, n-1$.

Problem 5.9 (Nilpotent matrices) Show that a matrix $A \in \mathbb{R}^{n \times n}$ is nilpotent (i.e. there exists $k \in \mathbb{N}$ such that $A^k = 0$) if and only if all of its eigenvalues are equal to zero. Show further that in this case $k \leq n$.

Problem 5.10 (Transfer function) Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$ and $D \in \mathbb{R}^{p \times m}$. Let $u(\cdot) : \mathbb{R}_+ \to U$ be piecewise continuous. For $t \in \mathbb{R}_+$, consider the linear time-invariant system:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t). \end{cases}$$

1. Consider a change of basis $\tilde{x} = Tx$. Compute the transfer function with respect to the new basis and compare it to the transfer function in the original basis.
2. Assume now that A is semisimple. Use your answer in part 1 to provide a simple formula for $G(s)$ in terms of the eigenvalues of A.

Chapter 6

Stability

Consider again the time varying linear system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$
$$y(t) = C(t)x(t) + D(t)u(t).$$

Stability addresses the question of what happens to the solutions of this system as time, t, increases. Do they remain bounded, will they get progressively smaller, or will they diverge to infinity. Stability deals first and foremost with the properties of the differential equation. We will therefore ignore the output equation to start with. We will also start with the zero input case ($u = \theta_U$, or $u(t) = 0$ for all $t \in \mathbb{R}$), i.e. by considering the solutions of

$$\dot{x}(t) = A(t)x(t). \tag{6.1}$$

We will then return to inputs and outputs in Section 6.4. As we did for the definition of the solutions of the differential equation, we will start by considering general nonlinear systems (Section 6.1), then specialize to linear time varying systems (Section 6.2), then specialize further to linear time invariant systems (Section 6.3).

6.1 Nonlinear systems: Basic definitions

Consider again a general nonlinear, time varying system defined by a differential equation

$$\dot{x}(t) = p(x(t), t) \tag{6.2}$$

for $t \in \mathbb{R}$, $x(t) \in \mathbb{R}^n$ and $p : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$. To ensure existence and uniqueness of solutions assume that p is Lipschitz continuous in its first argument and piecewise continuous in its second. Let $s(t, t_0, x_0)$ denote the unique solution of (6.2) at time $t \in \mathbb{R}$ passing through $x_0 \in \mathbb{R}^n$ at time $t_0 \in \mathbb{R}$.

Though the computation of the solution function $s : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is impossible in general, for some x_0 the solution function may become particularly simple.

Definition 6.1 *A state $\hat{x} \in \mathbb{R}^n$ is called an equilibrium of system (6.2) if and only if $p(\hat{x}, t) = 0$ for all $t \in \mathbb{R}$.*

The following fact is an immediate consequence of this observation.

Fact 6.1 *If $\hat{x} \in \mathbb{R}^n$ is an equilibrium of system (6.2) then $s(t, t_0, \hat{x}) = \hat{x}$ for all $t, t_0 \in \mathbb{R}$.*

Proof: Note that for the proposed solution $s(t_0, t_0, \hat{x}) = \hat{x}$ and

$$\frac{d}{dt}s(t, t_0, \hat{x}) = \frac{d}{dt}\hat{x} = 0 = p(\hat{x}, t) = p(s(t, t_0, \hat{x}), t).$$

The conclusion follows by existence and uniqueness of solutions. ∎

The fact shows that a solution which passes through an equilibrium, $\hat{x}$, at some point in time is forced to stay on the equilibrium for all times. We call this constant solution the *equilibrium solution* defined by the equilibrium $\hat{x}$.

What if a solution passes close to the equilibrium, but not exactly through it? Clearly such a solution will no longer be identically equal to the equilibrium, but will it move away from the equilibrium, or will it remain close to it? Will it converge to the equilibrium and if so at what rate?

To answer these questions we first need to fix a norm on $\mathbb{R}^n$ to be able to measure distances. Any norm will do since they are all equivalent, for simplicity we will use the Euclidean norm throughout. Equiped with this norm, we can now formalize the above questions in the following definition.

Definition 6.2 *Let $\hat{x} \in \mathbb{R}^n$ be an equilibrium of system (6.2). This equilibrium is called:*

1. *Stable if and only if for all $t_0 \in \mathbb{R}$, and all $\epsilon > 0$, there exists $\delta > 0$ such that*
$$\|x_0 - \hat{x}\| < \delta \Rightarrow \|s(t, t_0, x_0) - \hat{x}\| < \epsilon, \ \ \forall t \geq t_0.$$
2. *Unstable if and only if it is not stable.*
3. *Uniformly stable if and only if for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $t_0 \in \mathbb{R}$*
$$\|x_0 - \hat{x}\| < \delta \Rightarrow \|s(t, t_0, x_0) - \hat{x}\| < \epsilon, \ \ \forall t \geq t_0.$$
4. *Locally asymptotically stable if and only if it is stable and for all $t_0 \in \mathbb{R}$ there exists $M > 0$ such that*
$$\|x_0 - \hat{x}\| \leq M \Rightarrow \lim_{t\to\infty} \|s(t, t_0, x_0) - \hat{x}\| = 0.$$
5. *Globally asymptotically stable if and only if it is stable and for all $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$*
$$\lim_{t\to\infty} \|s(t, t_0, x_0) - \hat{x}\| = 0.$$
6. *Locally exponentially stable if and only if for all $t_0 \in \mathbb{R}$ there exist $\alpha, m, M > 0$ such that for all all $t \geq t_0$*
$$\|x_0 - \hat{x}\| \leq M \Rightarrow \|s(t, t_0, x_0) - \hat{x}\| \leq m\|x_0 - \hat{x}\|e^{-\alpha(t-t_0)}.$$
7. *Globally exponentially stable if and only if for all $t_0 \in \mathbb{R}$ there exist $\alpha, m > 0$ such that for all $x_0 \in \mathbb{R}^n$ and all $t \geq t_0$*
$$\|s(t, t_0, x_0) - \hat{x}\| \leq m\|x_0 - \hat{x}\|e^{-\alpha(t-t_0)}.$$

Special care is needed in the above definition: The order of the quantifiers is very important. Note for example that the definition of stability implicitly allows δ to depend on t_0 and ϵ; one sometimes writes $\delta(t_0, \epsilon)$ to highlight this dependence. On the other hand, in the definition of uniform stability δ can depend on ϵ but not on t_0, i.e. the same δ must work for all t_0; one sometimes uses the notation $\delta(\epsilon)$ to highlight this fact. Likewise, the definition of global exponential stability requires α and m to be independent of x_0, i.e. the same α and m must work for all $x_0 \in \mathbb{R}^n$; a variant of this definition where m and α are allowed to depend on x_0 is sometimes referred to as *semi-global exponential stability*.

The definition distinguishes stability concepts along three axes. The most fundamental distinction deals with the convergence of nearby solutions to the equilibrium. The equilibrium is called unstable if we cannot keep solutions close to it by starting sufficiently close, stable if we can keep solutions as close as we want by starting them sufficiently close, asymptotically stable if in addition nearby solutions converge to the equilibrium, and exponentially stable if they do so at an exponential rate. The second distinction deals with how these properties depend on the starting time, t_0: For uniform stability the starting time is irrelevant, the property holds the same way irrespective of when we look at the system. The third distinction deals with how these properties depend on the starting state, x_0: "Local" implies that the property holds provided we start close enough to the equilibrium, whereas global requires that the property holds irrespective of where we start. Note that this distinction is irrelevant for stability and uniform stability, since the conditions listed in the definition are required to hold provided we start close enough. One can also combinatorially mix these qualities to define other variants of stability notions: Uniform local asymptotic stability (where the equilibrium is uniformly stable and the convergence rate is independent of the starting time), uniform global exponential stability, etc. We will not pursue these variants of the definitions here, since most of them turn out to be irrelevant when dealing with linear systems.

It is easy to see that the notions of stability introduced in Definition 6.2 are progressively stronger.

Fact 6.2 *Consider an equilibrium of system (6.2). Then the following statements are true:*

If the equilibrium is	*then it is also*
uniformly stable	*stable*
locally asymptotically stable	*stable*
globally asymptotically stable	*locally asymptotically stable*
locally exponentially stable	*locally asymptotically stable*
globally exponentally stable	*globally asymptotically stable*
globally exponentally stable	*locally exponentially stable*

Proof: Most of the statements are obvious from the definition. Asymptotic stability requires stability, global asymptotic stability implies that the conditions of local asymptotic stability hold for any $M > 0$, etc. The only part that requires any work is showing that local/global exponential stability implies local/global asymptotic stability.

Consider a globally[1] exponentially stable equilibrium $\hat{x}$, i.e. assume that for all t_0 there exist $\alpha, m > 0$ such that for all x_0, $\|s(t, t_0, x_0) - \hat{x}\| \leq m\|x_0 - \hat{x}\|e^{-\alpha(t-t_0)}$ for all $t \geq t_0$. For $t_0 \in \mathbb{R}$ and $\epsilon > 0$ take $\delta = \epsilon/m$. Then for all $x_0 \in \mathbb{R}^n$ such that $\|x_0 - \hat{x}\| < \delta$ and all $t \geq t_0$

$$\|s(t, t_0, x_0) - \hat{x}\| \leq m\|x_0 - \hat{x}\|e^{-\alpha(t-t_0)} < m\delta e^{-\alpha(t-t_0)} = \epsilon e^{-\alpha(t-t_0)} \leq \epsilon.$$

Hence the equilibrium is stable. Moreover, since by the properties of the norm $\|s(t, t_0, x_0) - \hat{x}\| \geq 0$,

$$0 \leq \lim_{t\to\infty} \|s(t, t_0, x_0) - \hat{x}_0\| \leq \lim_{t\to\infty} m\|x_0 - \hat{x}\|e^{-\alpha(t-t_0)} = 0.$$

Hence $\lim_{t\to\infty} \|s(t, t_0, x_0) - \hat{x}\| = 0$ and the equilibrium is asymptotically stable. ∎

It is also easy to see that the stability notions of Definition 6.2 are strictly stronger one from the other; in other words the converse implications in the table of Fact 6.2 are in general not true. We show this through a series of counter-examples.

Example (Stable, non-uniformly stable equilibrium) For $x(t) \in \mathbb{R}$ consider the linear, time varying system

$$\dot{x}(t) = -\frac{2t}{1+t^2}x(t) \tag{6.3}$$

[1]The argument for local exponential stability is effectively the same.

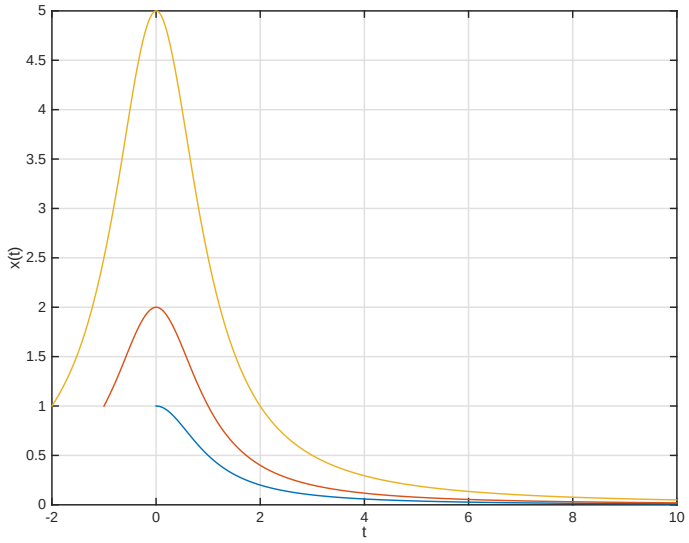

Figure 6.1: Three trajectories of the linear time varying system of equation (6.3) with initial condition $x(t_0) = 1$ and $t_0 = 0$, -1 and -2 respectively.

Exercise 6.1 Show that the system has a unique equilibrium at $\hat{x} = 0$. Show further that

$$s(t, t_0, x_0) = \frac{1 + t_0^2}{1 + t^2} x_0$$

by differentiating and invoking existence and uniqueness of solutions.

Typical trajectories of the system for $x_0 = 1$ and different values of t_0 are shown in Figure 6.1.

It is easy to see that $\hat{x} = 0$ is a stable equilibrium. Indeed, given $t_0 \in \mathbb{R}$ and $\epsilon > 0$ let $\delta = \epsilon/(1 + t_0^2)$. Then for all $x_0 \in \mathbb{R}^n$ such that $\|x_0\| < \delta$,

$$\|s(t, t_0, x_0)\| = \left\| \frac{1 + t_0^2}{1 + t^2} x_0 \right\| < \frac{\epsilon}{1 + t^2} \leq \epsilon.$$

However, the equilibrium is not uniformly stable: For a given ϵ we cannot find a δ that works for all $t_0 \in \mathbb{R}$. To see this, notice that for $t_0 \leq 0$, $\|s(t, t_0, x_0)\|$ reaches a maximum of $(1 + t_0^2)\|x_0\|$ at $t = 0$. Hence to ensure that $\|s(t, t_0, x_0)\| < \epsilon$ we need to ensure that $(1 + t_0^2)\|x_0\| < \epsilon$ which is impossible to do by restricting x_0 alone; for any $0 < \delta < \epsilon$ and $\|x_0\| < \delta$ we can make $\|s(0, t_0, x_0)\| > \epsilon$ by taking $t_0 < -\sqrt{\epsilon/\delta - 1}$. ■

Example (Stable, non asymptotically stable equilibrium) For $x(t) \in \mathbb{R}^2$ consider the linear, time invariant system

$$\dot{x}(t) = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix} x(t) \text{ with } x(0) = x_0 = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}. \tag{6.4}$$

Since the system is linear time invariant we can take $t_0 = 0$ without loss of generality.

Exercise 6.2 Show that the system has a unique equilibrium $\hat{x} = 0$. Show further that

$$\Phi(t, 0) = \begin{bmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{bmatrix}$$

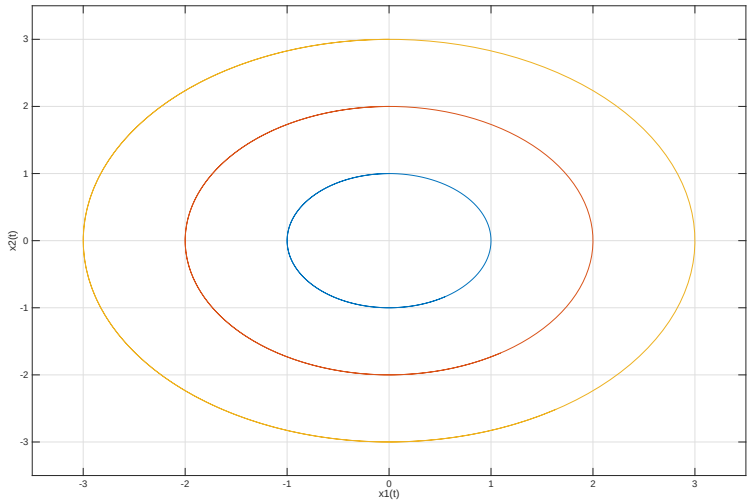

Figure 6.2: Three trajectories of the linear time invariant system of equation (6.4) with initial condition $x(0) = (0, 1)$, $(0, 2)$ and $(0, 3)$ respectively.

by differentiating and invoking Theorem 4.2.

Typical trajectories of the system for different values of x_0 are shown in Figure 6.2. Clearly

$$x(t) = \Phi(t, 0)x_0 = \begin{bmatrix} x_{01}\cos(\omega t) - x_{02}\sin(\omega t) \\ x_{01}\sin(\omega t) + x_{02}\cos(\omega t) \end{bmatrix}.$$

Using the 2-norm leads to

$$\|x(t)\|^2 = \|x_0\|^2.$$

Therefore the system is uniformly stable (take $\delta = \epsilon$) but not asymptotically stable (since in general $\lim_{t\to\infty} \|x(t)\| = \|x_0\| \neq 0$). ■

Exercise 6.3 Show that linear time invariant system $\dot{x}(t) = 0$ with $x(t) \in \mathbb{R}$ also has a stable, but not asymptotically stable equilibrium $\hat{x} = 0$. Does this system have any other equilibria? Are they stable? Asymptotically stable?

Example (Asymptotically stable, non exponentially stable equilibrium) Let us return to the system of equation (6.3). Recall that $\hat{x} = 0$ is a stable equilibrium. Moreover,

$$\lim_{t\to\infty} \|s(t, t_0, x_0)\| = \|x_0\|(1 + t_0^2) \lim_{t\to\infty} \frac{1}{1 + t^2} = 0$$

for all t_0, x_0. Hence the equilibrium is globally asymptotically stable. It is, however not exponentially stable (not even locally). Assume, for the sake of contradiction, that for all $t_0 \in \mathbb{R}$ there exist $\alpha, m, M > 0$ such that for all $x_0 \in \mathbb{R}$ with $\|x_0\| \leq M$ and all $t \geq t_0$,

$$\|s(t, t_0, x_0)\| \leq m\|x_0\|e^{-\alpha(t-t_0)}$$

In particular, for $t_0 = 0$ this would imply that for all $t \geq 0$

$$\|s(t, 0, 1)\| \leq m\|x_0\|e^{-\alpha t} \Rightarrow \frac{e^{\alpha t}}{1 + t^2} < m.$$

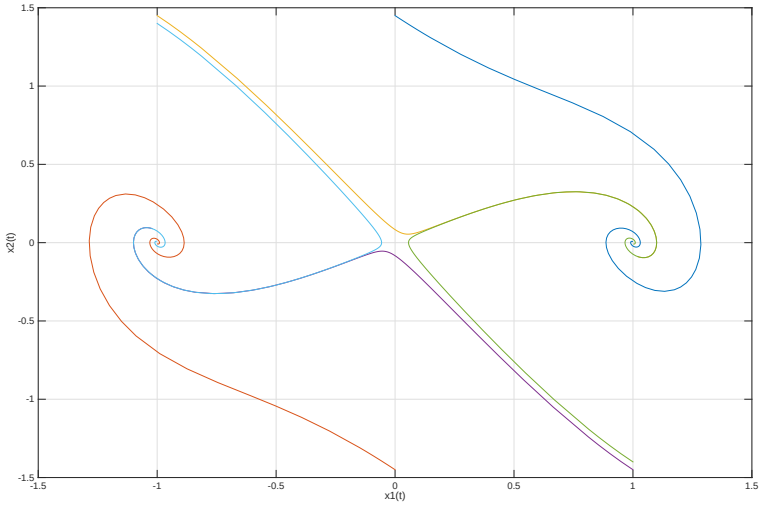

Figure 6.3: Trajectories of the nonlinear system of equation (6.5).

Since $\frac{e^{\alpha t}}{1+t^2} \to \infty$ as $t \to \infty$ this is a contradiction. ■

Notice that so far all the counter-examples have involved linear systems (sometimes time varying). As we will see in the next section, for linear systems local and global asymptotic stability are equivalent concepts. To distinguish between them we therefore need to resort to non-linear systems.

Example (Locally, non-globally asymptotically stable equilibrium) Consider a two dimensional state vector $x = (x_1, x_2) \in \mathbb{R}^2$ whose evolution is governed by the following differential equations

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ x_1(t) - x_1(t)^3 - x_2(t) \end{bmatrix}. \tag{6.5}$$

It is easy to see that this system has three equilibria, at $(0,0)$, $(1,0)$ and $(-1,0)$ respectively.

Exercise 6.4 Verify that these are indeed the only equilibria of this system. Is the system linear or nonlinear? What is the function p in $\dot{x}(t) = p(x(t), t)$? Is it globally Lipshitz in its first argument?

Clearly none of the three equilibria can be globally asymptotically stable: It is impossible for all trajectories to converge to a particular equilibrium, since those starting at another equilibrium will stay put. To study whether some of these equilibria are stable or locally asymptotically stable one can compute the linearisation of the system about each of the equilibria and study the stability of the resulting linear system using the methods presented later in this chapter. It is reasonable to assume that nearby the equilibrium, where the terms neglected in the linearisation are small, the behaviour of the nonlinear system will be similar to that of its linearisation. Hence if the linearisation is asymptotically stable one would expect the equilibrium to be locally asymptotically stable for the nonlinear system. This argument can in fact be formalized, leading to the so-called Lyapunov Indirect Method for checking local asymptotic stability, or instability of nonlinear systems. The interested reader is referred to [17] or [12] for more information on this topic.

For this system, linearisation suggests that one would expect the equilibria $(1,0)$ and $(-1,0)$ to be locally asymptotically stable and the equilibrium $(0,0)$ to be unstable (Problem 6.5). This can be visually confirmed by simulating the system for various initial conditions and plotting the trajectories

$x(t)$. The most informative way of doing this is to generate a parametric plot of $x_1(t)$ against $x_2(t)$ parametrized by t. This so-called *phase plane* plot for this system is shown in Figure 6.3. ■

Before restricting our attention to linear system we point out two more general facts about the stability concepts introduced in Definition 6.2. The first is an intimate relation between stability and continuity. To expose this link we need to think of the function mapping initial conditions to state trajectories from the initial time t_0 onwards. Since state trajectories are continuous functions of time, for each $t_0 \in \mathbb{R}$ one can think of this function as a map between the state space $\mathbb{R}^n$ and the space of continuous functions $C([t_0, \infty), \mathbb{R}^n)$

$$\begin{aligned} s(\cdot, t_0, \odot) : \mathbb{R}^n &\longrightarrow C([t_0, \infty), \mathbb{R}^n) \\ x_0 &\longmapsto \{s(\cdot, t_0, x_0) : [t_0, \infty) \to \mathbb{R}^n\}. \end{aligned}$$

The strange notation is meant to alert the reader to the fact that we consider $s(\cdot, t_0, \odot)$ for fixed t_0 as a function mapping a vector (denoted by the placeholder $\odot$) to a function of time (denoted by the placeholder $\cdot$, left over after x_0 is substituted for $\odot$).

Recall that for the stability definitions we have equipped $\mathbb{R}^n$ with a norm $\|\cdot\|$. We now equip $C([t_0, \infty), \mathbb{R}^n)$ with the corresponding infinity norm

$$\|s(\cdot, t_0, x_0)\|_{t_0,\infty} = \sup_{t \geq t_0} \|s(t, t_0, x_0)\|, \tag{6.6}$$

where we include t_0 in the notation to make the dependence on initial time explicit. Notice that the first norm in Equation (6.6) is a norm on the infinite dimensional function space (i.e., $s(\cdot, t_0, x_0) \in C([t_0, \infty), \mathbb{R}^n)$ is though of as a function of time), whereas the second norm is a norm on the finite dimensional state space (i.e., $s(t, t_0, x_0) \in \mathbb{R}^n$ is the value of this function for the specific time $t \in [t_0, \infty)$).

Fact 6.3 *An equilibrium, $\hat{x}$, of system (6.2) is stable if and only if for all $t_0 \in \mathbb{R}$ the function $s(\cdot, t_0, \odot)$ mapping the normed space $(\mathbb{R}^n, \|\cdot\|)$ into the normed space $(C([t_0, \infty), \mathbb{R}^n), \|\cdot\|_{t_0,\infty})$ is continuous at $\hat{x}$.*

Proof: The statement is effectively a tautology. Fix $t_0 \in \mathbb{R}$ and recall that, according to Definition 3.6, $s(\cdot, t_0, \odot)$ is continuous at $\hat{x}$ if and only for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|x_0 - \hat{x}\| < \delta \Rightarrow \|s(\cdot, t_0, x_0) - s(\cdot, t_0, \hat{x})\|_{t_0,\infty} < \epsilon.$$

By Equation (6.6) this is equivalent to

$$\|x_0 - \hat{x}\| < \delta \Rightarrow \|s(t, t_0, x_0) - s(t, t_0, \hat{x})\| < \epsilon \ \ \forall t \geq t_0,$$

which, recalling that $s(t, t_0, \hat{x}) = \hat{x}$ for all $t \geq t_0$ is in turn equivallent to

$$\|x_0 - \hat{x}\| < \delta \Rightarrow \|s(t, t_0, x_0) - \hat{x}\| < \epsilon \ \ \forall t \geq t_0,$$

which is precisely the definition of stability. ■

A similar relation between uniform stability and uniform continuity (where the δ above is independent of t_0) can also be derived in the same way.

The second general fact relates to the possible rate of convergence. The strongest notion of stability in Definition 6.2, namely exponential stability, requires that solutions converge to the equilibrium exponentially (i.e. rather quickly) in time. Could they converge even faster? Could we, for example, introduce another meaningful stability definition that requires solutions to coverge with a rate of $e^{-\alpha t^2}$ for some $\alpha > 0$? And if not, can we at least increase α in the exponential convergence? The following fact reveals that Lipschitz continuity imposes a fundamental limit on how fast convergence can be.

Fact 6.4 *Let $\hat{x}$ be an equilibrium of system (6.2) and assume that there exists $k > 0$ such that for all $x, x' \in \mathbb{R}^n$, $\|p(x,t) - p(x',t)\| \leq k\|x - x'\|$. Then for all $t_0 \in \mathbb{R}$ and all $t \geq t_0$*

$$\|x_0 - \hat{x}\|e^{-k(t-t_0)} \leq \|s(t,t_0,x_0) - \hat{x}\| \leq \|x_0 - \hat{x}\|e^{k(t-t_0)}$$

Proof: If $x_0 = \hat{x}$ the claim is trivially true, we therefore restrict attention to the case $x_0 \neq \hat{x}$. Note that in this case we must have $s(t,t_0,x_0) \neq \hat{x}$ for all t; if $s(t,t_0,x_0) = \hat{x}$ for some t then $s(t,t_0,x_0)$ must be the equilibrium solution and $s(t,t_0,x_0) = \hat{x}$ for all t which, setting $t = t_0$ contradicts the fact that $x_0 \neq \hat{x}$.

Recall that for simplicity we are using the Euclidean norm. Hence $\|s(t,t_0,x_0) - \hat{x}\|^2 = (s(t,t_0,x_0) - \hat{x})^T(s(t,t_0,x_0) - \hat{x})$ and

$$\begin{aligned}
\left|\frac{d}{dt}\|s(t,t_0,x_0) - \hat{x}\|^2\right| &= \left|\frac{d}{dt}s(t,t_0,x_0)^T(s(t,t_0,x_0) - \hat{x}) + (s(t,t_0,x_0) - \hat{x})^T\frac{d}{dt}s(t,t_0,x_0)\right| \\
&= \left|p(s(t,t_0,x_0),t)^T(s(t,t_0,x_0) - \hat{x}) + (s(t,t_0,x_0) - \hat{x})^T p(s(t,t_0,x_0),t)\right| \\
&\leq \left|p(s(t,t_0,x_0),t)^T(s(t,t_0,x_0) - \hat{x})\right| + \left|(s(t,t_0,x_0) - \hat{x})^T p(s(t,t_0,x_0),t)\right| \\
&\leq \|p(s(t,t_0,x_0),t)^T\| \cdot \|s(t,t_0,x_0) - \hat{x}\| + \|(s(t,t_0,x_0) - \hat{x})^T\| \cdot \|p(s(t,t_0,x_0),t)\| \\
&= 2\|s(t,t_0,x_0) - \hat{x}\| \cdot \|p(s(t,t_0,x_0),t)\| \\
&= 2\|s(t,t_0,x_0) - \hat{x}\| \cdot \|p(s(t,t_0,x_0),t) - p(\hat{x},t)\| \\
&\leq 2k\|s(t,t_0,x_0) - \hat{x}\| \cdot \|s(t,t_0,x_0) - \hat{x}\|.
\end{aligned}$$

On the other hand,

$$\left|\frac{d}{dt}\|s(t,t_0,x_0) - \hat{x}\|^2\right| = \left|2\|s(t,t_0,x_0) - \hat{x}\|\frac{d}{dt}\|s(t,t_0,x_0) - \hat{x}\|\right|.$$

Since $s(t,t_0,x_0) \neq \hat{x}$, combining the two equations we must have

$$\left|\frac{d}{dt}\|s(t,t_0,x_0) - \hat{x}\|\right| \leq k\|s(t,t_0,x_0) - \hat{x}\|$$

or in other words

$$-k\|s(t,t_0,x_0) - \hat{x}\| \leq \frac{d}{dt}\|s(t,t_0,x_0) - \hat{x}\| \leq k\|s(t,t_0,x_0) - \hat{x}\|.$$

Applying the Gronwall Lemma (Theorem 3.8) to the right inequality leads to

$$\|s(t,t_0,x_0) - \hat{x}\| \leq \|x_0 - \hat{x}\|e^{k(t-t_0)}.$$

From the right inequality (adapting the steps of the proof of the Gronwall Lemma) we have

$$\begin{aligned}
\frac{d}{dt}\left(\|s(t,t_0,x_0) - \hat{x}\|e^{k(t-t_0)}\right) &= \frac{d}{dt}\left(\|s(t,t_0,x_0) - \hat{x}\|\right)e^{k(t-t_0)} + \|s(t,t_0,x_0) - \hat{x}\|\frac{d}{dt}e^{k(t-t_0)} \\
&\geq -k\|s(t,t_0,x_0) - \hat{x}\|e^{k(t-t_0)} + \|s(t,t_0,x_0) - \hat{x}\|ke^{k(t-t_0)} = 0.
\end{aligned}$$

Hence for all $t \geq t_0$,

$$\|s(t,t_0,x_0) - \hat{x}\|e^{k(t-t_0)} \geq \|s(t_0,t_0,x_0) - \hat{x}\|e^{k(t_0-t_0)} = \|x_0 - \hat{x}\|,$$

which leads to

$$\|s(t,t_0,x_0) - \hat{x}\| \geq \|x_0 - \hat{x}\|e^{-k(t-t_0)}.$$

■

In summary, convergence to an equilibrium can be at most exponential. The fact also shows that if an equilibrium is unstable divergence cannot be any faster than exponential. Even though a fixed Lipshitz constant is assumed to simplify the proof it is easy to see that the claim still holds if the Lipschitz constant is time varying but bounded from above and below; one simply needs to replace k by its lower bound in the left inequality and its upper bound in the right inequality. The lower bound on the Lipschitz constant also provides a bound for the rate of exponential convergence.

6.2 Linear time varying systems

We note that Definition 6.2 is very general and works also for nonlinear systems. Since for linear systems we know something more about the solution of the system it turns out that the conditions of the definition are somewhat redundant in this case. Consider now the linear time varying system

$$\dot{x}(t) = A(t)x(t) \tag{6.7}$$

and let $s(t, t_0, x_0)$ denote the solution at time t starting at x_0 at time t_0. Since $s(t, t_0, x_0) = \Phi(t, t_0)x_0$, the solution is linear with respect to the initial state and all the stability definitions reduce to checking properties of the state transition matrix $\Phi(t, t_0)$.

First note that for all $t_0 \in \mathbb{R}_+$ if $x_0 = 0$ then

$$s(t, t_0, 0) = \Phi(t, t_0)x_0 = 0 \in \mathbb{R}^n \quad \forall t \in \mathbb{R}_+$$

is the solution of (6.7). Another way to think of this observation is that if $x(t_0) = 0$ for some $t_0 \in \mathbb{R}_+$, then

$$\dot{x}(t_0) = A(t)x(t_0) = 0$$

therefore the solution of the differential equation does not move from 0. Either way, the solution of (6.7) that passes through the state $x(t_0) = 0$ at some time $t_0 \in \mathbb{R}_+$ will be identically equal to zero for all times, and $\hat{x} = 0$ is an equilibrium of (6.7).

Exercise 6.5 Can there be other $x_0 \neq 0$ such that $s(t, t_0, x_0) = x_0$ for all $t \in \mathbb{R}_+$?

Theorem 6.1 *Let $\|\Phi(t, 0)\|$ denote the norm of the matrix $\Phi(t, 0) \in \mathbb{R}^{n\times n}$ induced by the Euclidean norm in $\mathbb{R}^n$. The equilibrium $\hat{x} = 0$ of (6.7) is:*

1. *Stable if and only if for all $t_0 \in \mathbb{R}$, there exists $K > 0$ such that*

$$\|\Phi(t, 0)\| \leq K \text{ for all } t \geq 0.$$

2. *Locally asymptotically stable if and only if $\lim_{t\to\infty} \|\Phi(t, 0)\| = 0$.*

Proof: Part 1: We first show that if there exists $K > 0$ such that $\|\Phi(t, 0)\| \leq K$ for all $t \geq 0$ then the equilibrium $\hat{x} = 0$ is stable; without loss of generality, we can take $K > 1$. Fix $t_0 \in \mathbb{R}$ and, for simplicity, distinguish two cases:

1. $t_0 < 0$. In this case let $M(t_0) = \sup_{\tau\in[t_0,0]} \|\Phi(\tau, t_0)\|$.
2. $t_0 \geq 0$. In this case let $M(t_0) = \sup_{\tau\in[0,t_0]} \|\Phi(\tau, t_0)\|$.

In the first case, if $t \in [t_0, 0]$ note that

$$\|s(t, t_0, x_0)\| \leq \|\Phi(t, t_0)\| \cdot \|x_0\| \leq \sup_{\tau\in[t_0,0]} \|\Phi(\tau, t_0)\| \cdot \|x_0\| = M(t_0)\|x_0\|.$$

If $t > 0$,

$$\begin{aligned}\|s(t, t_0, x_0)\| &= \|\Phi(t, t_0)x_0\| = \|\Phi(t, 0)\Phi(0, t_0)x_0\| \\ &\leq \|\Phi(t, 0)\| \cdot \|\Phi(0, t_0)\| \cdot \|x_0\| \leq \|\Phi(t, 0)\| \sup_{\tau\in[t_0,0]} \|\Phi(\tau, t_0)\| \cdot \|x_0\| \\ &= KM(t_0)\|x_0\|.\end{aligned}$$

Similarly, in the second case for all $t \geq t_0 \geq 0$,

$$\begin{aligned}\|s(t,t_0,x_0)\| &= \|\Phi(t,0)\Phi(0,t_0)x_0\| \leq \|\Phi(t,0)\| \cdot \|\Phi(0,t_0)\| \cdot \|x_0\| \\ &\leq \|\Phi(t,0)\| \sup_{\tau\in[0,t_0]} \|\Phi(\tau,t_0)\| \cdot \|x_0\| \\ &= KM(t_0)\|x_0\|.\end{aligned}$$

In all cases,

$$\begin{aligned}\|s(t,t_0,x_0)\| \leq KM(t_0)\|x_0\| &\text{ for all } t \geq t_0 \\ \Rightarrow \sup_{t\geq t_0} \|s(t,t_0,x_0)\| &\leq KM(t_0)\|x_0\| \\ \Rightarrow \|s(t,t_0,x_0)\|_{t_0,\infty} &\leq KM(t_0)\|x_0\| \\ \Rightarrow \sup_{\|x_0\|=1} \frac{\|s(t,t_0,x_0)\|_{t_0,\infty}}{\|x_0\|} &\leq KM(t_0)\end{aligned}$$

Hence the induced norm of the function

$$\begin{array}{rccc} s(\cdot,t_0,\odot,0): & (\mathbb{R}^n,\|\cdot\|) & \longrightarrow & (C([t_0,\infty),\mathbb{R}^n),\|\cdot\|_{t_0,\infty}) \\ & x_0 & \longmapsto & s(\cdot,t_0,x_0,0)=\Phi(\cdot,t_0)x_0 \end{array}$$

is finite, the function is continuous and the equilibrium $\hat{x} = 0$ is stable by Fact 6.3.

Conversely, we show by contraposition that if $\|\Phi(t,0)\|$ is unbounded (i.e., for all $K > 0$ there exists $t \geq 0$ such that $\|\Phi(t,0)\| > K$) then the equilibrium $\hat{x} = 0$ cannot be stable (i.e., there exist t_0 and $\epsilon > 0$ such that for all $\delta > 0$ there exists $x_0 \in \mathbb{R}^n$ with $\|x_0\| < \delta$ and $t \geq t_0$ such that $\|s(t,t_0,x_0)\| \geq \epsilon$). For simplicity, take $t_0 = 0$ and $\epsilon = 1$. For any $\delta > 0$ pick $K = 2/\delta$ and find the $t \geq 0$ such that $\|\Phi(t,0)\| > K$. Recall that

$$\|\Phi(t,0)\| = \sup_{\|x\|=1} \|\Phi(t,0)x\| > K.$$

Therefore, there exists $x \in \mathbb{R}$ such that $\|x\| = 1$ and $\|\Phi(t,0)x\| > K$. Let $x_0 = x\delta/2$. Then $\|x_0\| = \|x\|\delta/2 = \delta/2 < \delta$ and

$$\|s(t,0,x_0)\| = \|\Phi(t,0)x\delta/2\| = \|\Phi(t,0)x\|\delta/2 > K\delta/2 = 1.$$

Part 2: Assume first that $\lim_{t\to\infty} \|\Phi(t,0)\| = 0$ and show that the equilibrium solution is asymptotically stable. For all $(x_0,t_0) \in \mathbb{R}^n \times \mathbb{R}_+$

$$\|s(t,t_0,x_0)\| = \|\Phi(t,t_0)x_0\| = \|\Phi(t,0)\Phi(0,t_0)x_0\| \leq \|\Phi(t,0)\| \cdot \|\Phi(0,t_0)x_0\|.$$

Therefore, for all $x_0 \in \mathbb{R}^n$, $t_0 \in \mathbb{R}$ the second term is constant and

$$\lim_{t\to\infty} \|s(t,t_0,x_0)\| = 0.$$

To establish that the equilibrium is asymptotically stable it therefore suffices to show that it is stable. Note that the function $\|\Phi(\cdot,0)\| : \mathbb{R}_+ \to \mathbb{R}$ is continuous (by the properties of Φ and the continuity of the norm). Therefore, since $\lim_{t\to\infty} \|\Phi(t,0)\| = 0$, we must have that $\|\Phi(\cdot,0)\|$ is bounded and the equilibrium is stable by Part 1.

Conversely, assume that the equilibrium is locally asymptotically stable. Then there exists $M > 0$ such that for all $t \in \mathbb{R}$ and all $x_0 \in \mathbb{R}^n$ with $\|x_0\| \leq M$,

$$\lim_{t\to\infty} s(t,t_0,x_0) = \lim_{t\to\infty} \Phi(t,t_0)x_0 = \lim_{t\to\infty} (\Phi(t,0)\Phi(0,t_0)x_0) = \lim_{t\to\infty} (\Phi(t,0))\,\Phi(0,t_0)x_0 = 0 \qquad (6.8)$$

Consider the canonical basis $\{e_i\}_{i=1}^n$ for $\mathbb{R}^n$ and take $t_0 = 0$. Letting $x_0 = e_i$ in (6.8) shows that the i^{th} column of $\Phi(t,0)$ tends to zero. Repeating for $i = 1,\ldots,n$ establishes the claim. ∎

The use of the matrix norm induced by the Euclidean norm on $\mathbb{R}^n$ is again optional, any other induced norm would also work. From this theorem one can further conclude that when it comes to linear systems (even linear time varying systems) local stability notions are equivalent to global ones.

Fact 6.5 *The equilibrium solution $\hat{x} = 0$ of (6.1) is*

1. *Globally asymptotically stable if and only if it is locally asymptotically stable.*
2. *Globally exponentially stable if and only if it is locally exponentially stable.*

The proof is left as an exercise (Problem 6.6).

6.3 Linear time invariant systems

For linear time invariant systems the stability discussion simplifies even further. Consider the linear time invariant system

$$\dot{x}(t) = Ax(t) \tag{6.9}$$

and recall that in this case the solution at time t starting at x_0 at time t_0 is given by

$$s(t, t_0, x_0) = e^{A(t-t_0)}x_0.$$

It is easy to see that in this case there is no difference between uniform and non-uniform stability notions.

Fact 6.6 *The equilibrium $\hat{x} = 0$ of the linear time invariant system $\dot{x}(t) = Ax(t)$ is uniformly stable if and only if it is stable.*

The proof is left as an exercise (Problem 6.7).

We have seen that for linear systems there is no distinction between local and global asymptotic/exponential stability (Fact 6.5). Asymptotic stability, however, is still a weaker property than exponential stability, even for time varying linear systems (see the example of equation (6.3). For time invariant linear systems it turns out that asymptotic stability and exponential stability are also equivalent to each other. Moreover, one can determine whether a time invariant linear system is exponentially stable through a simple algebraic calculation.

Theorem 6.2 *For linear time invariant systems the following statements are equivalent:*

1. *The equilibrium $\hat{x} = 0$ is asymptotically stable.*
2. *The equilibrium $\hat{x} = 0$ is exponentially stable.*
3. *For all $\lambda \in \text{Spec}[A]$, $\text{Re}[\lambda] < 0$.*

We start the proof by establishing the following lemma.

Lemma 6.1 *For all $\epsilon > 0$ there exists $m > 0$ such that for all $t \in \mathbb{R}_+$*

$$\|e^{At}\| \leq me^{(\mu+\epsilon)t}$$

where $\|\cdot\|$ denotes an induced norm on $\mathbb{R}^{n\times n}$ and $\mu = \max\{\text{Re}[\lambda] \mid \lambda \in \text{Spec}[A]\}$.

Proof: Recall that the existence of the Jordan canonical form implies that

$$e^{At} = \sum_{\lambda \in \text{SPEC}[A]} \Pi_\lambda(t) e^{\lambda t}$$

where $\Pi_\lambda(t) \in \mathbb{C}[t]^{n \times n}$ are $n \times n$ matrices of polynomials in t with complex coefficients. Consider (for simplicity) the infinity induced norm $\|\cdot\|_\infty$ for $\mathbb{C}^{n \times n}$. Then

$$\|e^{At}\|_\infty = \left\| \sum_{\lambda \in \text{SPEC}[A]} \Pi_\lambda(t) e^{\lambda t} \right\|_\infty \leq \sum_{\lambda \in \text{SPEC}[A]} \|\Pi_\lambda(t)\|_\infty \cdot \left| e^{\lambda t} \right|.$$

For $\lambda \in \text{SPEC}[A]$ let $\lambda = \sigma + j\omega$ with $\sigma, \omega \in \mathbb{R}$; if $\lambda \in \mathbb{R}$ simply set $\omega = 0$. Note that

$$\left| e^{\lambda t} \right| = \left| e^{(\sigma + j\omega)t} \right| = \left| e^{\sigma t} \right| \cdot \left| e^{j\omega t} \right| = e^{\sigma t} \cdot |\cos(\omega t) + j \sin(\omega t)| = e^{\sigma t}.$$

Therefore, since $\sigma = \text{RE}[\lambda] \leq \mu$,

$$\|e^{At}\|_\infty \leq \left(\sum_{\lambda \in \text{SPEC}[A]} \|\Pi_\lambda(t)\|_\infty \right) e^{\mu t}.$$

Recall that $\|\Pi_\lambda(t)\|_\infty$ is the maximum among the rows of $\Pi_\lambda(t)$ of the sum of the magnitudes of the elements in the row. Since all entries are polynomial, then there exists a polynomial $p_\lambda(t) \in \mathbb{R}[t]$ such that $p_\lambda(t) \geq \|\Pi_\lambda(t)\|_\infty$ for all $t \in \mathbb{R}_+$. If we define the polynomial $p(t) \in \mathbb{R}[t]$ by

$$p(t) = \sum_{\lambda \in \text{SPEC}[A]} p_\lambda(t),$$

then

$$\|e^{At}\|_\infty \leq p(t) e^{\mu t}. \tag{6.10}$$

Since $p(t)$ is a polynomial in $t \in \mathbb{R}_+$, for any $\epsilon > 0$ the function $p(t)e^{-\epsilon t}$ is continuous and $\lim_{t \to \infty} p(t)e^{-\epsilon t} = 0$. Therefore $p(t)e^{-\epsilon t}$ is bounded for $t \in \mathbb{R}_+$, i.e. there exists $m > 0$ such that $p(t)e^{-\epsilon t} \leq m$ for all $t \in \mathbb{R}_+$. Substituting this into equation (6.10) leads to

$$\forall \epsilon > 0 \;\; \exists m > 0, \;\; \|e^{At}\|_\infty \leq m e^{(\mu + \epsilon)t}.$$

■

Proof: (Of Theorem 6.2) We have already seen that $2 \Rightarrow 1$ (Fact 6.2).

$3 \Rightarrow 2$: If all eigenvalues have negative real part then

$$\mu = \max\{\text{RE}[\lambda] \mid \lambda \in \text{SPEC}[A]\} < 0.$$

Consider $\epsilon \in (0, -\mu)$ and set $\alpha = -(\mu + \epsilon) > 0$. By Lemma 6.1 there exists $m > 0$ such that

$$\|e^{At}\| \leq m e^{-\alpha t}.$$

Therefore for all $(x_0, t_0) \in \mathbb{R}^m \times \mathbb{R}_+$ and all $t \geq t_0$

$$\|s(t, t_0, x_0)\| = \left\| e^{A(t - t_0)} x_0 \right\| \leq \left\| e^{A(t - t_0)} \right\| \cdot \|x_0\| \leq m \|x_0\| e^{-\alpha(t - t_0)}.$$

Hence the equilibrium solution is exponentially stable.

$1 \Rightarrow 3$: By contraposition. Assume there exists $\lambda \in \text{SPEC}[A]$ such that $\text{RE}[\lambda] \geq 0$ and let $v \in \mathbb{C}^n$ denote the corresponding eigenvector. Without loss of generality take $t_0 = 0$ and note that

$$s(t, 0, v) = e^{\lambda t} v \in \mathbb{C}^n.$$

This follows by existence and uniqueness of solutions, since $e^{\lambda 0}v = v$ and

$$\frac{d}{dt}e^{\lambda t}v = e^{\lambda t}\lambda v = e^{\lambda t}Av = Ae^{\lambda t}v.$$

For clarity of exposition we distinguish two cases.

First case: λ is real. Then

$$\|s(t,0,v)\| = \left|e^{\lambda t}\right| \cdot \|v\|.$$

Since $\lambda \geq 0$, $\|s(t,0,v)\|$ is either constant (if $\lambda = 0$), or tends to infinity as t tends to infinity (if $\lambda > 0$). In either case the equilibrium solution cannot be asymptotically stable.

Second case: λ is complex. Let $\lambda = \sigma + j\omega \in \mathbb{C}$ with $\sigma, \omega \in \mathbb{R}$, $\sigma \geq 0$ and $\omega \neq 0$. Let also $v = v_1 + jv_2 \in \mathbb{C}^n$ for $v_1, v_2 \in \mathbb{R}^n$. Recall that since $A \in \mathbb{R}^{n\times n}$ eigenvalues come in complex conjugate pairs, so $\lambda' = \sigma - j\omega$ is also an eigenvalue of A with eigenvector $v' = v_1 - jv_2$. Since v is an eigenvector by definition $v \neq 0$ and either $v_1 \neq 0$ or $v_2 \neq 0$. Together with the fact that $\omega \neq 0$ this implies that $v_2 \neq 0$; otherwise $A(v_1 + jv_2) = (\sigma + j\omega)(v_1 + jv_2)$ implies that $Av_1 = (\sigma + j\omega)v_1$ which cannot be satisfied by any non-zero real vector v_1. Therefore, without loss of generality we can assume that $\|v_2\| = 1$. Since

$$v_2 = \frac{-j(v_1 + jv_2) + j(v_1 - jv_2)}{2} = \frac{-jv + jv'}{2}$$

by the linearity of the solution

$$\begin{aligned}
s(t,0,2v_2) &= -js(t,0,v_1 + jv_2) + js(t,0,v_1 - jv_2) \\
&= -je^{(\sigma+j\omega)t}(v_1 + jv_2) + je^{(\sigma-j\omega)t}(v_1 - jv_2) \\
&= e^{\sigma t}\left[-je^{j\omega t}(v_1 + jv_2) + je^{-j\omega t}(v_1 - jv_2)\right] \\
&= e^{\sigma t}\left[-j(\cos(\omega t) + j\sin(\omega t))(v_1 + jv_2) + j(\cos(\omega t) - j\sin(\omega t))(v_1 - jv_2)\right] \\
&= 2e^{\sigma t}\left[v_2\cos(\omega t) + v_1\sin(\omega t)\right].
\end{aligned}$$

Hence

$$\|s(t,0,2v_2)\| = 2e^{\sigma t}\left\|v_2\cos(\omega t) + v_1\sin(\omega t)\right\|.$$

Note that

$$\|v_2\cos(\omega t) + v_1\sin(\omega t)\| = \|v_2\| = 1$$

whenever $t = \pi k/\omega$ for $k \in \mathbb{N}$. Since $\sigma \geq 0$ the sequence

$$\{\|s(\pi k/\omega, 0, 2v_2)\|\}_{k=0}^{\infty} = \left\{2e^{\sigma k\pi/\omega}\right\}_{k=0}^{\infty}$$

is bounded away from zero; it is either constant (if $\sigma = 0$) or diverges to infinity (if $\sigma > 0$). Hence $\|s(t,0,2v_1)\|$ cannot converge to 0 as t tends to infinity and the equilibrium solution cannot be asymptotically stable.

We have shown that $1 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1$. Hence the three statements are equivalent. ∎

Theorem 6.2 does not hold is the system is time varying: We have already seen an example of a system that is asymptotically but not exponentially stable, so the three statements cannot be equivalent. How about $3 \Rightarrow 1$? In other words if the eigenvalues of the matrix $A(t) \in \mathbb{R}^{n\times n}$ have negative real parts for all $t \in \mathbb{R}$ is the equilibrium solution of $\dot{x}(t) = A(t)x(t)$ asymptotically stable? This would provide an easy test of stability for time varying systems. Unfortunately, however, this statement is not true.

Example (Unstable time varying system) For $a \in (1,2)$ consider the matrix

$$A(t) = \begin{bmatrix} -1 + a\cos^2(t) & 1 - a\cos(t)\sin(t) \\ -1 + a\cos(t)\sin(t) & -1 + a\sin^2(t) \end{bmatrix}$$

Exercise 6.6 Show that the eigenvalues λ_1, λ_2 of $A(t)$ have $\text{RE}[\lambda_i] = -\frac{2-\alpha}{2}$ for $i = 1, 2$. Moreover,

$$\Phi(t,0) = \begin{bmatrix} e^{(a-1)t}\cos(t) & e^{-t}\sin(t) \\ -e^{(a-1)t}\sin(t) & e^{-t}\cos(t) \end{bmatrix}.$$

Therefore, if we take $x_0 = (1, 0)$ then

$$s(t,0,x_0) = \begin{bmatrix} e^{(a-1)t}\cos(t) \\ -e^{(a-1)t}\sin(t) \end{bmatrix} \Rightarrow \|s(t,0,x_0)\|_2 = e^{(a-1)t}.$$

Since $a \in (1, 2)$, $\text{RE}[\lambda_i] < 0$ for all $t \in \mathbb{R}_+$ but nonetheless $\|s(t,0,x_0)\|_2 \to \infty$ and the equilibrium solution is not stable. ■

As the example suggests, determining the stability of time varying linear systems is in general difficult. Some results exist (for example, if $A(t)$ has negative eigenvalues, is bounded as a function of time and varies "slowly enough") but unfortunately, unlike time invariant linear systems, there are no simple, general purpose methods that can be used to investigate the stability of time varying linear systems.

An argument similar to that of Theorem 6.2 can be used to derive stability conditions for time invariant systems in terms of the eigenvalues of the matrix A.

Theorem 6.3 *The equilibrium $\hat{x} = 0$ of a linear time invariant system is stable if and only if the following two conditions are met:*

1. *For all $\lambda \in \text{SPEC}[A]$, $\text{RE}[\lambda] \leq 0$.*
2. *The algebraic and geometric multiplicity of all $\lambda \in \text{SPEC}[A]$ such that $\text{RE}[\lambda] = 0$ are equal.*

The proof is left as an exercise (Problem 6.7).

6.4 Systems with inputs and outputs

Consider now the full linear time varying system

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned}$$

with piecewise continuous input function $u(\cdot) : \mathbb{R} \to \mathbb{R}^m$. Based on the structure of the solutions $s(t, t_0, x_0, u)$ and $\rho(t, t_0, x_0, u)$ it should come as no surprise that the properties of the zero input solution $s(t, t_0, x_0, 0)$ (i.e. the solutions of (6.7)) to a large extend also determine the properties of the solution under non-zero inputs. Recall that for simplicity we consider the Euclidean norm for $\mathbb{R}^n$ and for $f(\cdot) : \mathbb{R} \to \mathbb{R}^n$ we defined

$$\|f(\cdot)\|_{t_0,\infty} = \sup_{t \geq t_0} \|f(t)\|.$$

Similar definitions can be given for the matrix valued functions $A(t)$, $B(t)$, etc. using the corresponding induced norms. Recall also that for existence and uniqueness of solutions we assume that $A(t)$, $B(t)$, etc. are piecewise continuous functions of time.

Theorem 6.4 *Assume that*

1. *The equilibrium solution is exponentially stable, i.e. there exists $m, \alpha > 0$ such that for all $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}_+$ and all $t \in \mathbb{R}_+$, $\|s(t, t_0, x_0, 0)\| \leq m\|x_0\|e^{-\alpha(t-t_0)}$.*

2. For all $t_0 \in \mathbb{R}$, $\|A(\cdot)\|_{t_0,\infty}$, $\|B(\cdot)\|_{t_0,\infty}$, $\|C(\cdot)\|_{t_0,\infty}$, $\|D(\cdot)\|_{t_0,\infty}$ *are bounded.*

Then for all $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}$ *and all* $u(\cdot) : [t_0, \infty) \to \mathbb{R}^m$ *with* $\|u(\cdot)\|_{t_0,\infty}$ *bounded,*

$$\|s(\cdot, t_0, x_0, u)\|_{t_0,\infty} \leq m\|x_0\|e^{\alpha t_0} + \frac{m}{\alpha}\|B(\cdot)\|_{t_0,\infty}\|u(\cdot)\|_{t_0,\infty}$$
$$\|\rho(\cdot, t_0, x_0, u)\|_{t_0,\infty} \leq m\|C(\cdot)\|_{t_0,\infty}\|x_0\|e^{\alpha t_0} + \left[\frac{m}{\alpha}\|C(\cdot)\|_{t_0,\infty}\|B(\cdot)\|_{t_0,\infty} + \|D(\cdot)\|_{t_0,\infty}\right]\|u(\cdot)\|_{t_0,\infty}.$$

If in addition $\lim_{t\to\infty} u(t) = 0$ *then* $\lim_{t\to\infty} s(t, t_0, x_0, u) = 0$ *and* $\lim_{t\to\infty} \rho(t, t_0, x_0, u) = 0$.

The proof requires only some manipulation of norm inequalities and is left as an exercise (Problem 6.8).

One can see that also in the case of systems with inputs there is an intimate relation between stability and continuity of the solution functions in an appropriate function space. To see this consider the functions

$$\begin{aligned} s(\cdot, t_0, 0, \odot) : \mathrm{PC}([t_0, \infty), \mathbb{R}^m) &\longrightarrow C([t_0, \infty), \mathbb{R}^n) \\ \{u(\cdot) : [t_0, \infty) \to \mathbb{R}^m\} &\longmapsto \{s(\cdot, t_0, 0, u) : [t_0, \infty) \to \mathbb{R}^n\} \\ \rho(\cdot, t_0, 0, \odot) : \mathrm{PC}([t_0, \infty), \mathbb{R}^m) &\longrightarrow \mathrm{PC}([t_0, \infty), \mathbb{R}^p) \\ \{u(\cdot) : [t_0, \infty) \to \mathbb{R}^m\} &\longmapsto \{\rho(\cdot, t_0, 0, u) : [t_0, \infty) \to \mathbb{R}^p\} \end{aligned}$$

The strange notation is again meant to alert the reader to the fact that we consider $s(\cdot, t_0, 0, \odot)$ for fixed t_0 as a function mapping a piecewise continuous function of time (denoted by the placeholder $\odot$) to a continuous function of time (denoted by the placeholder $\cdot$, left over after u is substituted for $\odot$). Then Theorem 6.4 directly implies the following.

Corollary 6.1 *If the linear time varying system is exponentially stable then the functions* $s(\cdot, t_0, 0, \odot)$ *and* $\rho(\cdot, t_0, 0, \odot)$ *are continuous.*

The proof is left as an exercise (Problem 6.8).

The properties guaranteed by Theorem 6.4 are known as the bounded input bounded state (BIBS) and the bounded input bounded output (BIBO) properties. We just note that exponential stability is in general necessary, since asymptotic stability is not enough.

Example (Unstable with input, asymptotically stable without) For $x(t), u(t) \in \mathbb{R}$ and $t \in \mathbb{R}_+$ consider the system

$$\dot{x}(t) = -\frac{1}{1+t}x + u.$$

Exercise 6.7 Show that for $t_0 \geq 0$, $\Phi(t, t_0) = (1 + t_0)/(1 + t)$.

Let $x(0) = 0$ and apply the constant input $u(t) = 1$ for all $t \in \mathbb{R}_+$. Then

$$s(t, 0, 0, 1) = \int_0^t \Phi(t, \tau)u(\tau)d\tau = \int_0^t \frac{1+\tau}{1+t}d\tau = \frac{t + t^2/2}{1+t} \to \infty$$

even though the equilibrium solution is asymptotically stable and the input is bounded since $\|u(\cdot)\|_\infty = 1 < \infty$. ∎

6.5 Lyapunov equation

In addition to computing eigenvalues, there is a second algebraic test that allows us to determine the asymptotic stability of linear time invariant systems, by solving the linear equation

$$A^T P + PA = -Q$$

(known as the Lyapunov equation). To formally state this fact recall that a matrix $P \in \mathbb{R}^{n\times n}$ is called symmetric if and only if $P^T = P$. A symmetric matrix, $P = P^T \in \mathbb{R}^{n\times n}$ is called positive definite if and only if for all $x \in \mathbb{R}^n$ with $x \neq 0$, $x^T P x > 0$; we then write $P = P^T > 0$.

Theorem 6.5 *The following statements are equivalent:*

1. *The equilibrium solution of $\dot{x}(t) = Ax(t)$ is asymptotically stable.*
2. *For all $Q = Q^T > 0$ the equation $A^T P + PA = -Q$ admits a unique solution $P = P^T > 0$.*

The proof of this fact is deferred to the next section. For the time being we simply show what can go wrong with the Lyapunov equation if the system is not asymptotically stable.

Example (Lyapunov equation for two dimensional systems) Let $x(t) \in \mathbb{R}^2$ and consider the linear system $\dot{x}(t) = Ax(t)$ for some generic

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

We fix a matrix $Q = 2I$ and since P is symmetric we have

$$P = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} \in \mathbb{R}^{2\times 2} \text{ and } Q = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

The Lyapunov equation

$$\begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} + \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

reduces to a set of three linear equations

$$\begin{aligned} a_{11}p_1 + a_{21}p_2 &= -1 \\ a_{12}p_1 + (a_{11} + a_{22})p_2 + a_{21}p_3 &= 0 \\ a_{12}p_2 + a_{22}p_3 &= -1 \end{aligned}$$

in the three unknowns p_1, p_2, p_3 (the top-right and bottom-left equations are identical).

To gain some insight into the conditions of Theorem 6.5 let us see what happens if we replace some trivial stable, asymptotically stable and unstable matrices A into these equations. For example, if we take

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$$

(asymptotically stable with eigenvalues at -1 and -2) the system has a unique solution $p_1 = 1$, $p_2 = 0$, $p_3 = 1/2$ leading to a positive definite

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$$

as expected. If we take

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$$

(unstable with eigenvalues at 1 and -2) the system has a unique solution $p_1 = -1$, $p_2 = 0$, $p_3 = 1/2$ leading to

$$P = \begin{bmatrix} -1 & 0 \\ 0 & 1/2 \end{bmatrix}$$

which is not positive definite. If we take

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(unstable with two eigenvalues at 1) the system has infinitely many solutions ($p_1 = -1 = p_3$ but p_2 is arbitrary). Finally, if we take

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

(stable but not asymptotically stable with eigenvalues at $\pm j$) the system has no solutions (we must have $p_2 = 1$ and $p_2 = -1$ at the same time). ■

Problems for chapter 6

Problem 6.1 (Stability) For $t \in \mathbb{R}_+$, consider the ODE $\dot{x}(t) = A(t)x(t)$, where, for $a \in \mathbb{R}$,

$$A(t) = \begin{bmatrix} -1 + a\cos^2(t) & 1 - a\cos(t)\sin(t) \\ -1 - a\cos(t)\sin(t) & -1 + a\sin^2(t), \end{bmatrix}$$

1. Show that
$$\Phi(t,0) = \begin{bmatrix} e^{(a-1)t}\cos(t) & e^{-t}\sin(t) \\ -e^{(a-1)t}\sin(t) & e^{-t}\cos(t). \end{bmatrix}$$
2. Deduce how the stability properties of the system change with the value of a.
3. Find the values of a for which the eigenvalues of $A(t)$ have negative real part for all $t \geq 0$. Compare the result with that of 2.

Problem 6.2 (Laplace Transform) Consider the time-invariant system:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t). \end{cases}$$

1. Let $x(0) = 0$. Make use of the Laplace transform to derive an input-output (ARMA) model of the system of the following form: for $a_i \in \mathbb{R}$ and $B_i \in \mathbb{R}^{p\times m}$, $i = 0, 1, \ldots, n$,
$$a_0 y^{(n)}(t) + a_1 y^{(n-1)}(t) + \ldots + a_n y(t) = B_0 u^{(n)}(t) + B_1 u^{(n-1)}(t) + \ldots + B_n u(t),$$
where $y^{(i)}(t)$ (resp. $u^{(i)}(t)$) denotes the i-th derivative of y (resp. of u) at t. You may assume that $y^{(i)}(0) = 0$ and $u^{(i)}(0) = 0$ for $i = 0, \ldots n$.
2. Let $A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ B_2 \end{bmatrix}$ and $C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$, where, for $1 \leq n_2 \leq n$, $A_{22} \in \mathbb{R}^{n_2\times n_2}$, $B_2 \in \mathbb{R}^{n_2\times m}$, $C_2 \in \mathbb{R}^{p\times n_2}$ and all other matrix blocks have consistent dimensions. Compute the transfer function of the system and use the result to build a state-space representation of order (i.e. size of the state evolution matrix) n_2 having the same transfer function.
(Hint: verify that the inverse of an invertible block triangular matrix $M = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix}$, with M_{11} and M_{22} square, is also block triangular, i.e. $M^{-1} = \begin{bmatrix} M^*_{11} & 0 \\ M^*_{21} & M^*_{22} \end{bmatrix}$; in particular, M^*_{22} must be equal to... .)

Problem 6.3 (Stability)

1. Consider a system $\dot{x}(t) = f(x(t))$. Suppose that we have $\frac{d}{dt}\big(x(t)^T P x(t)\big) \leq -x(t)^T Q x(t)$, where P and Q are symmetric positive definite matrices (here the time-derivative is taken along solutions of the system). Prove that under this condition the system is exponentially stable, in the sense that its solutions satisfy $\|x(t)\| \leq c\mathrm{e}^{-\mu t}\|x(0)\|$ for some $c, \mu > 0$. Note that this statement is true whether the system is linear or not. Further, if $f(x) = Ax$, then show that the above condition is equivalent to $A^T P + PA \leq -Q$.

2. Consider the system $\dot{x}(t) = Ax(t) + Bu(t)$ such that $\left\|\mathrm{e}^{At}\right\| \leq c\mathrm{e}^{-\mu t}$ for some $c, \mu > 0$.

   (a) Prove that if u is bounded over all time (in the sense that $\sup_{0\leq t<\infty}\|u(t)\| \leq M$ for some M), then x is also bounded, for any initial condition.

   (b) Now restrict attention to the zero initial condition ($x_0 = 0$). We can view the system above as a linear operator from the normed linear space of bounded functions $u : [0, \infty[\to \mathbb{R}^m$ with norm $\|u\| := \sup_{0\leq t<\infty}\|u(t)\|$, to the normed linear space of functions $x : [0, \infty[\to \mathbb{R}^n$ with the norm $\|x\| := \sup_{0\leq t<\infty}\|x(t)\|$. What can you say about the induced norm of this operator, using the calculations you made in (a)? What can you say about its continuity?

Problem 6.4 (Discrete-time Systems) Consider a discrete-time linear system $x(k+1) = Ax(k) + Bu(k)$, $k = 0, 1, \ldots$.

1. Write the formula for the solution $x(k)$ at time k starting from some initial state $x(0)$ at time 0.

2. Assume A is semi-simple. Under what conditions on the eigenvalues of A is the discrete-time system $x(k+1) = Ax(k)$ (no controls) asymptotically stable? Stable? Justify your answers. (Stability definitions are the same as for continuous-time systems, just replace t by k.)

3. Lyapunov's second method for discrete-time system $x(k+1) = f(x(k))$ involves the difference $\Delta V(x) := V(f(x)) - V(x)$ instead of the derivative $\dot{V}(x)$; with this substitution, the statement is the same as in the continuous-time case. Derive the counterpart of the Lyapunov equation for the LTI discrete-time system $x(k+1) = Ax(k)$.

Problem 6.5 (Linearisation example) Consider a two dimensional state vector $x = (x_1, x_2) \in \mathbb{R}^2$ whose evolution is governed by the following differential equations

$$\begin{aligned}\dot{x}_1(t) &= x_2(t)\\ \dot{x}_2(t) &= x_1(t) - x_1(t)^3 - x_2(t).\end{aligned}$$

1. Compute all equilibria of the system.

2. Compute the linearisation of the system about its equilibria. The calculation is the same as the one in Section 4.1, replacing the "optimal trajectory", $x^*(t)$, with the equilibrium solution $x^*(t) = \hat{x}$ for all t, where $\hat{x}$ is each of the equilibria of the system.

3. Compute the eigenvalues of the matrices of the resulting linearisation and speculate about the stability of the equilibria of the nonlinear system by invoking Theorem 6.2. Simulate the system to confirm your intuition.

Problem 6.6 (When local is global) The equilibrium solution $\hat{x} = 0$ of the linear time varying system $\dot{x}(t) = A(t)x(t)$ is

1. Globally asymptotically stable if and only if it is locally asymptotically stable.

2. Globally exponentially stable if and only if it is locally exponentially stable.

Is it possible for the linear time varying system to have more than one equilibria? If yes, is it possible for any of them to be locally asymptotically stable in such a case?

Problem 6.7 (When all is uniform) Show that the equilibrium $\hat{x} = 0$ of the linear time invariant system $\dot{x}(t) = Ax(t)$ is uniformly stable if and only if it is stable. Show further that this is the case if and only if the following two conditions are met:

1. For all $\lambda \in \text{SPEC}[A]$, $\text{RE}[\lambda] \leq 0$.
2. For all $\lambda \in \text{SPEC}[A]$ such that $\text{RE}[\lambda] = 0$, all Jordan blocks of λ have dimension 1.

Problem 6.8 (BIBO and BIBS stability) Consider the time varying linear system

$$\begin{aligned}\dot{x}(t) &= A(t)x(t) + B(t)u(t)\\ y(t) &= C(t)x(t) + D(t)u(t).\end{aligned}$$

with $A(t)$, $B(t)$, $C(t)$, $D(t)$ piecewise continuous. Assume that

1. The equilibrium solution is exponentially stable, i.e. there exists $m, \alpha > 0$ such that for all $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}_+$ and all $t \in \mathbb{R}_+$, $\|s(t, t_0, x_0, \theta_U)\| \leq m\|x_0\|e^{-\alpha(t-t_0)}$.
2. For all $t_0 \in \mathbb{R}$, $\|A(\cdot)\|_{t_0,\infty}$, $\|B(\cdot)\|_{t_0,\infty}$, $\|C(\cdot)\|_{t_0,\infty}$, $\|D(\cdot)\|_{t_0,\infty}$ are bounded.

Show that:

1. For all $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}$ and all $u(\cdot) : [t_0, \infty) \to \mathbb{R}^m$ with $\|u(\cdot)\|_{t_0,\infty}$ bounded,

$$\begin{aligned}\|s(\cdot, t_0, x_0, u)\|_{t_0,\infty} &\leq m\|x_0\|e^{\alpha t_0} + \frac{m}{\alpha}\|B(\cdot)\|_{t_0,\infty}\|u(\cdot)\|_{t_0,\infty}\\ \|\rho(\cdot, t_0, x_0, u)\|_{t_0,\infty} &\leq m\|C(\cdot)\|_{t_0,\infty}\|x_0\|e^{\alpha t_0} + \left[\frac{m}{\alpha}\|C(\cdot)\|_{t_0,\infty}\|B(\cdot)\|_{t_0,\infty} + \|D(\cdot)\|_{t_0,\infty}\right]\|u(\cdot)\|_{t_0,\infty}.\end{aligned}$$

2. For all $t_0 \in \mathbb{R}$ the functions

$$\begin{aligned}s(\cdot, t_0, 0, \odot) : \text{PC}([t_0, \infty), \mathbb{R}^m) &\longrightarrow C([t_0, \infty), \mathbb{R}^n)\\ \{u(\cdot) : [t_0, \infty) \to \mathbb{R}^m\} &\longmapsto \{s(\cdot, t_0, 0, u) : [t_0, \infty) \to \mathbb{R}^n\}\\ \rho(\cdot, t_0, 0, \odot) : \text{PC}([t_0, \infty), \mathbb{R}^m) &\longrightarrow \text{PC}([t_0, \infty), \mathbb{R}^p)\\ \{u(\cdot) : [t_0, \infty) \to \mathbb{R}^m\} &\longmapsto \{\rho(\cdot, t_0, 0, u) : [t_0, \infty) \to \mathbb{R}^p\}\end{aligned}$$

   are continuous.

Chapter 7

Inner product spaces

We return briefly to abstract vector spaces to introduce the notion of inner products, that will form the basis of our discussion on controllability and observability.

7.1 Inner product

Consider a field F, either $\mathbb{R}$ or $\mathbb{C}$. If $F = \mathbb{C}$ let $|a| = \sqrt{a_1^2 + a_2^2}$ denote the magnitude and $\overline{a} = a_1 - ja_2$ denote the complex conjugate of $a = a_1 + ja_2 \in F$. If $F = \mathbb{R}$ let $|a|$ denote the absolute value of $a \in F$; for simplicity define $\overline{a} = a$ in this case.

Definition 7.1 *Let (H, F) be a linear space. A function $\langle \cdot, \cdot \rangle : H \times H \to F$ is called an inner product if and only if for all $x, y, z \in H$, $\alpha \in F$,*

1. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$.
2. $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$.
3. $\langle x, x \rangle$ *is real and positive for all* $x \neq 0$.
4. $\langle x, y \rangle = \overline{\langle y, x \rangle}$ *(complex conjugate).*

$(H, F, \langle \cdot, \cdot \rangle)$ is then called an inner product space.

Exercise 7.1 For all $x, y, z \in H$ and all $a \in F$, $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ and $\langle ax, y \rangle = \overline{a} \langle x, y \rangle$. Moreover, $\langle x, 0 \rangle = \langle 0, x \rangle = 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$.

Fact 7.1 *If $(H, F, \langle \cdot, \cdot \rangle)$ is an inner product space then the function $\| \cdot \| : H \to F$ defined by $\|x\| = \sqrt{\langle x, x \rangle}$ is a norm on (H, F).*

Note that the function is well defined by property 3 of the inner product definition. The proof is based on the following fact.

Theorem 7.1 (Schwarz inequality) *With $\| \cdot \|$ defined as in Fact 7.1, $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$ for all $x, y \in H$.*

Proof: If $x = 0$ or $y = 0$ the claim is obvious by Exercise 7.1. Otherwise, select $\alpha \in F$ such that $|\alpha| = 1$ and $\alpha \langle x, y \rangle = |\langle x, y \rangle|$; i.e. if $F = \mathbb{R}$ take α to be the sign of $\langle x, y \rangle$ and if $F = \mathbb{C}$ take

$\alpha = \frac{\langle y,x\rangle}{|\langle x,y\rangle|}$. Then for all $\lambda \in \mathbb{R}$

$$\begin{aligned}
\|\lambda x + \alpha y\|^2 &= \langle \lambda x + \alpha y, \lambda x + \alpha y\rangle \\
&= \langle \lambda x + \alpha y, \lambda x\rangle + \langle \lambda x + \alpha y, \alpha y\rangle \\
&= \lambda\langle \lambda x + \alpha y, x\rangle + \alpha\langle \lambda x + \alpha y, y\rangle \\
&= \lambda\overline{\langle x, \lambda x + \alpha y\rangle} + \alpha\overline{\langle y, \lambda x + \alpha y\rangle} \\
&= \lambda\overline{(\lambda\langle x, x\rangle + \alpha\langle x, y\rangle)} + \alpha\overline{(\lambda\langle y, x\rangle + \alpha\langle y, y\rangle)} \\
&= \lambda^2\|x\|^2 + \lambda|\langle x, y\rangle| + \lambda|\langle x, y\rangle| + |\alpha|^2\|y\|^2 \\
&= \lambda^2\|x\|^2 + 2\lambda|\langle x, y\rangle| + \|y\|^2.
\end{aligned}$$

Since by definition $\langle \lambda x + \alpha y, \lambda x + \alpha y\rangle \geq 0$ we must have

$$\lambda^2\|x\|^2 + 2\lambda|\langle x, y\rangle| + \|y\|^2 \geq 0.$$

This is a quadratic in λ that must be non-negative for all $\lambda \in \mathbb{R}$. This will be the case if and only if it is non-negative at its minimum point. Differentiating with respect to λ shows that the minimum occurs when

$$\lambda = -\frac{|\langle x, y\rangle|}{\|x\|^2}$$

Substituting this back into the quadratic we see that

$$\frac{|\langle x, y\rangle|^2}{\|x\|^4}\|x\|^2 - 2\frac{|\langle x, y\rangle|^2}{\|x\|^2} + \|y\|^2 \geq 0 \Rightarrow |\langle x, y\rangle|^2 \leq \|x\|^2 \cdot \|y\|^2$$

■

Exercise 7.2 Prove Fact 7.1 using the Schwarz inequality.

Definition 7.2 *Let $(H, F, \langle\cdot,\cdot\rangle)$ be an inner product space. The norm defined by $\|x\|^2 = \langle x, x\rangle$ for all $x \in H$ is called the norm defined by the inner product. If the normed space $(H, F, \|\cdot\|)$ is complete (a Banach space) then $(H, F, \langle\cdot,\cdot\rangle)$ is called a Hilbert space.*

To demonstrate the above definitions we consider two examples of inner products, one for finite dimensional and one for infinite dimensional spaces; the latter brings us to the discussion of the Hilbert space of square integrable functions, that will play a central role in the discussion of controllability and observability in Chapter 8.

Example (Finite dimensional inner product space) For $F = \mathbb{R}$ or $F = \mathbb{C}$, consider the linear space (F^n, F). Define the inner product $\langle\cdot,\cdot\rangle : F^n \times F^n \to F$ by

$$\langle x, y\rangle = \sum_{i=1}^{n} \overline{x_i} y_i = \overline{x}^T \cdot y$$

for all $x, y \in F^n$, where $\overline{x}^T$ denotes complex conjugate (element-wise) transpose.

Exercise 7.3 Show that this satisfies the axioms of the inner product.

It is easy to see that the norm defined by this inner product

$$\|x\|^2 = \langle x, x\rangle = \sum_{i=1}^{n} |x_i|^2 = \|x\|_2^2$$

is the Euclidean norm. We have already seen that $(F^n, F, \|\cdot\|_2)$ is a complete normed space (Theorem 3.2), hence $(F^n, F, \langle\cdot,\cdot\rangle)$ is a Hilbert space. ■

7.2 The space of square-integrable functions

For $F = \mathbb{R}$ or $F = \mathbb{C}$, $t_0, t_1 \in \mathbb{R}$ with $t_0 \leq t_1$ consider the family of square integrable functions $f(\cdot) : [t_0, t_1] \to F^n$, i.e. all functions such that

$$\int_{t_0}^{t_1} \|f(t)\|_2^2 dt < \infty.$$

Note that the discussion can be extended to the case where $t_0 = -\infty$, or $t_1 = +\infty$, or both simply by considering the domain of the function to be respectively $(t_0, t_1]$, $[t_0, t_1)$ or (t_0, t_1) and requiring that the integral be finite.

Exercise 7.4 Show that the sum of square-integrable functions is also square-integrable and hence square integrable functions form a subspace.

On this space we can define the L^2 inner product

$$\langle f, g \rangle = \int_{t_0}^{t_1} \overline{f(t)}^T g(t) dt$$

where as before $\overline{f(t)}^T$ denotes complex conjugate transpose of $f(t) \in F^n$. It is easy to see that the norm defined by this inner product is the 2−norm, already introduced for continuous functions in Chapter 3

$$\|f(\cdot)\|_2 = \left(\int_{t_0}^{t_1} \|f(t)\|_2^2 dt \right)^{\frac{1}{2}}.$$

Since square integrable functions also contain discontinuous functions, however, we now run into difficulties with our norm definition. Consider for example the function f that takes the value 1 at $t = 0$ and 0 everywhere else. Then $\int_0^{+\infty} |f(t)|^2 \, dt = 0$, but f is not the zero function, which violates the axioms of the norm (Definition 3.1). To resolve this issue we can identify functions that are equal except for "a few" points with each other and consider them as the same function. More formally, we can define an equivalence relation between functions: Functions $f_1(\cdot) : [t_0, t_1] \to F^n$ and $f_2(\cdot) : [t_0, t_1] \to F^n$ will be called equivalent if and only if

$$\int_{t_0}^{t_1} \|f_1(t) - f_2(t)\|_2^2 dt = 0.$$

Notice that two functions are equivalent if they are equal "almost everywhere" (except for finitely many points, countably many points, etc.) We can then identify equivalent functions with each other (formally, identify all functions in the same equivalence class with the equivalence class itself) and consider as our space of interest the set of non-equivalent square integrable functions (formally, the quotient space of the equivalence relation). One can check that this space is indeed a linear space over the field F under the usual operations of function addition and scalar multiplication. Moreover, $\|f(\cdot)\|_2$ is now a well defined norm on this space, since all functions such that $\int_{t_0}^{t_1} \|f(t)\|_2^2 dt = 0$ are now identified with the zero function.

Unfortunately, one can see that the resulting space is not complete. Loosely speaking, the limit of a Cauchy sequence of square integrable functions in the 2-norm may be a function whose integral is not defined (a non-integrable function in the sense of Riemann). The canonical way in which this problem is solved is by extending the definition of integral. The *Lebesgue integral* of a function is a technical construction that generalizes in many ways the notion of integral that the reader remembers from calculus courses (the *Riemann integral*). When a function f is integrable in the sense of Riemann (either over $[t_0, t_1]$ or over $\mathbb{R}_+$), the Lebesgue integral of f always exists and coincides with its Riemann integral. But Lebesgue integration allows for a much larger class of functions to be integrable, comprising functions way, way "wilder" than just discontinuous at finitely many

points. For example, the function $d : [0,1] \to \mathbb{R}$ such that $d(t) = 1$ if $t \in \mathbb{Q}$, and $d(t) = 0$ elsewhere is discontinuous at all points and is *not* integrable according to Riemann; but we mention, albeit without any justification, that the Lebesgue integral $\int_0^1 d(t)dt$ exists and is equal to 0. Lebesgue integration also provides more general convergence theorems, and an easier way to reduce multiple integrals to iterated integrals of one variable (Fubini's theorem).

If the reader proceeds in the field of System Theory, he or she will have to seriously study Lebesgue integration at some point. But since the long excursus into measure theory required by this study leads far beyond the scope of this course, we will avoid this discussion altogether, leave to the reader the bare statement that the Lebesgue integral is "a more general notion of integral", and instead rely, without providing a proof, on the following fundamental result of mathematical analysis:

Fact 7.2 *Given a normed linear space $(V, F, \|\cdot\|)$, there exist a* complete *normed linear space $(\bar{V}, F, \|\cdot\|)$ and a norm-preserving linear function $\varphi : V \to \bar{V}$ (that is, $\|\varphi(x)\| = \|x\|$ for all $x \in V$) such that every point in $\bar{V}$ is a limit point of some sequence of points in $\varphi(V)$. $\bar{V}$ is called the* completion *of V.*

In other words, every normed space V can be mapped to a "dense" subspace of a complete space $\bar{V}$; the map preserves the distances between points (it is an "isometry"), so that V and the "dense" subspace can be identified to all practical purposes. To gain intuition, the reader can think at the completion of V as the union of V and the set of all the "missing limit points" of the non-convergent Cauchy sequences of V. For example, $\mathbb{R}$ is the completion of $\mathbb{Q}$. On the other hand, if V is already complete, then the completion of V is V itself (hence, the completion $\mathbb{R}$ is $\mathbb{R}$). What is now of greatest interest for us is the following

Fact 7.3 *The completion of $(C([t_0, t_1], F^n), \|\cdot\|_2)$ is the space of Lebesgue square-integrable functions over $[t_0, t_1]$, that is the set of all those functions $f : [t_0, t_1] \to F^n$ for which the integral $\int_{t_0}^{t_1} |f(t)|^2\, dt$ exists (in the sense of Lebesgue) and is finite; this space is a complete normed space with respect to the norm $\|f\|_2 := \sqrt{\int_{t_0}^{t_1} |f(t)|^2\, dt}$, provided that functions which are equal "almost everywhere" are identified. All the functions f whose square is integrable in the sense of Riemann belong to this space, and their norm according to both the definitions of integral coincide. This completion is denoted $L^2([t_0, t_1], F^n)$.*

$L^2([t_0, t_1], F^n)$ is now a true Banach spaces; together with the L^2 inner product it is therefore a Hilbert space.

7.3 Orthogonal complement

Definition 7.3 *Let $(H, F, \langle\cdot,\cdot\rangle)$ be an inner product space. $x, y \in H$ are called orthogonal if and only if $\langle x, y\rangle = 0$.*

Example (Orthogonal vectors) Consider the inner product space $(\mathbb{R}^n, \mathbb{R}, \langle\cdot,\cdot\rangle)$ with the inner product $\langle x, y\rangle = \overline{x}^T y$ defined above. Given two non-zero vectors $x, y \in \mathbb{R}^n$ one can define the angle, θ, between them by

$$\theta = \cos^{-1}\left(\frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}\right)$$

x, y are orthogonal if and only if $\theta = (2k+1)\pi/2$ for $k \in \mathbb{Z}$.

Exercise 7.5 Show that θ is well defined by the Schwarz inequality. Show further that $\theta = (2k+1)\pi/2$ if and only if x and y are orthogonal.

■

This interpretation of the "angle between two vectors" motivates the following generalisation of a well known theorem in geometry.

Theorem 7.2 (Pythagoras theorem) *Let $(H, F, \langle\cdot,\cdot\rangle)$ be an inner product space. If $x, y \in H$ are orthogonal then $\|x+y\|^2 = \|x\|^2 + \|y\|^2$, where $\|\cdot\|$ is the norm defined by the inner product.*

Proof: Exercise. ■

Note that H does not need to be $\mathbb{R}^3$ (or even finite dimensional) for the Pythagoras theorem to hold.

Given a subspace $M \subseteq H$ of a linear space, consider now the set of all vectors $y \in H$ which are orthogonal to all vectors $x \in M$.

Definition 7.4 *The orthogonal complement of a subspace, M, of an inner product space $(H, F, \langle\cdot,\cdot\rangle)$ is the set*

$$M^\perp = \{y \in H \mid \langle x, y\rangle = 0 \ \ \forall x \in M\}.$$

Consider now an inner product space $(H, F, \langle\cdot,\cdot\rangle)$ with the norm $\|y\|^2 = \langle y, y\rangle$ defined by the inner product. Recall that a sequence $\{y_i\}_{i=0}^\infty$ is said to converge to a point $y \in H$ if an only if $\lim_{i\to\infty} \|y - y_i\| = 0$ (Definition 3.4). In this case y is called the limit point of the sequence $\{y_i\}_{i=0}^\infty$. Recall also that a subset $K \subseteq H$ is called closed if and only if it contains the limit points of all the sequences $\{y_i\}_{i=0}^\infty \subseteq K$ (Definition 3.5).

Fact 7.4 *Let M be a subspace of the inner product space $(H, F, \langle\cdot,\cdot\rangle)$. $M^\perp$ is a closed subspace of H and $M \cap M^\perp = \{0\}$.*

Proof: To show that $M^\perp$ is a subspace consider $y_1, y_2 \in M^\perp$, $a_1, a_2 \in F$ and show that $a_1y_1 + a_2y_2 \in M^\perp$. Indeed, for all $x \in M$, $\langle x, a_1y_1 + a_2y_2\rangle = a_1\langle x, y_1\rangle + a_2\langle x, y_2\rangle = 0$.

To show that $M^\perp$ is closed, consider a sequence $\{y_i\}_{i=0}^\infty \subseteq M^\perp$ and assume that it converges to some $y \in H$. We need to show that $y \in M^\perp$. Note that, since $y_i \in M^\perp$, by definition $\langle x, y_i\rangle = 0$ for all $x \in M$. Consider an arbitrary $x \in M$ and note that

$$\langle x, y\rangle = \langle x, y_i + y - y_i\rangle = \langle x, y_i\rangle + \langle x, y - y_i\rangle = \langle x, y - y_i\rangle.$$

By the Schwarz inequality,

$$0 \le |\langle x, y\rangle| = |\langle x, y - y_i\rangle| \le \|x\| \cdot \|y - y_i\|.$$

Since $\lim_{i\to\infty} \|y - y_i\| = 0$ we must have that $|\langle x, y\rangle| = 0$. Hence $y \in M^\perp$.

Finally, consider $y \in M \cap M^\perp$. Since $y \in M^\perp$, $\langle x, y\rangle = 0$ for all $x \in M$. Since for y itself we have that $y \in M$, $\langle y, y\rangle = \|y\|^2 = 0$. By the axioms of the norm (Definition 3.1) this is equivalent to $y = 0$. ■

Definition 7.5 *Let M, N be subspaces of a linear space (H, F). The sum of M and N is the set*

$$M + N = \{w \mid \exists u \in M, v \in N \text{ such that } w = u + v\}.$$

If in addition $M \cap N = \{0\}$ then $M + N$ is called the direct sum of M and N and is denoted by $M \oplus N$.

Exercise 7.6 Show that $M + N$ is a subspace of H.

Fact 7.5 *$V = M \oplus N$ if and only if for all $x \in V$ there exists unique $u \in M$ and $v \in N$ such that $x = u + v$.*

Proof: (exercise) ($\Rightarrow$). Let $V = M \oplus N$. Then for all $x \in V$ there exist $u \in M$ and $v \in N$ such that $x = u + v$. It remains to show that the u and v are unique. Assume, for the sake of contradiction, that they are not. Then there exist $u' \in M$, $v' \in N$ with $(u', v') \neq (u, v)$ such that $x = u' + v'$. Then

$$u' + v' = u + v \Rightarrow u - u' = v - v'.$$

Moreover, $u - u' \in M$ and $v - v' \in N$ (since M, N are subspaces). Hence $M \cap N \supseteq \{u - u', 0\} \neq \{0\}$ which contradicts the fact that $M \cap N = \{0\}$.

($\Leftarrow$). Assume that for all $x \in V$ there exist unique $u \in M$ and $v \in N$ such that $x = u + v$. Then $V = M + N$ by definition. We need to show that $M \cap N = \{0\}$. Assume, for the sake of contradiction that this is not the case, i.e. there exists $y \neq 0$ such that $y \in M \cap N$. Consider and arbitrary $x \in V$ and the unique $u \in M$ and $v \in N$ such that $x = u + v$. Define $u' = u + y$ and $v' = v - y$. Note that $u' \in M$ and $v' \in N$ since M and N are subspaces and $y \in M \cap N$. Moreover, $u' + v' = u + y + v - y = u + v = x$ but $u \neq u'$ and $v \neq v'$ since $y \neq 0$. This contradicts the uniqueness of u and v. ∎

Theorem 7.3 *Let M be a closed subspace of a Hilbert space $(H, F, \langle \cdot, \cdot \rangle)$. Then:*

1. *$H = M \oplus M^\perp$ (denoted by $M \overset{\perp}{\oplus} M^\perp$).*
2. *For all $x \in H$ there exists a unique $y \in M$ such that $x - y \in M^\perp$. This y is called the orthogonal projection of x onto M.*
3. *For all $x \in H$ the orthogonal projection $y \in M$ is the unique element of M that achieves the minimum*
$$\|x - y\| = \inf\{\|x - u\| \mid u \in M\}.$$

7.4 Adjoint of a linear map

Definition 7.6 *Let $(U, F, \langle \cdot, \cdot \rangle_U)$ and $(V, F, \langle \cdot, \cdot \rangle_V)$ be Hilbert spaces and $\mathcal{A} : U \to V$ a continuous linear map. The adjoint of $\mathcal{A}$ is the map $\mathcal{A}^* : V \to U$ defined by*

$$\langle v, \mathcal{A}u \rangle_V = \langle \mathcal{A}^* v, u \rangle_U$$

for all $u \in U$, $v \in V$.

Theorem 7.4 *Let $(U, F, \langle \cdot, \cdot \rangle_U)$, $(V, F, \langle \cdot, \cdot \rangle_V)$ and be $(W, F, \langle \cdot, \cdot \rangle_W)$ Hilbert spaces, $\mathcal{A} : U \to V$, $\mathcal{B} : U \to V$ and $\mathcal{C} : W \to U$ be continuous linear maps and $a \in F$. The following hold:*

1. *$\mathcal{A}^*$ is well defined, linear and continuous.*
2. *$(\mathcal{A} + \mathcal{B})^* = \mathcal{A}^* + \mathcal{B}^*$.*
3. *$(a\mathcal{A})^* = \overline{a}\mathcal{A}^*$.*
4. *$(\mathcal{A} \circ \mathcal{C})^* = \mathcal{C}^* \circ \mathcal{A}^*$*
5. *If $\mathcal{A}$ is invertible then $(\mathcal{A}^{-1})^* = (\mathcal{A}^*)^{-1}$.*
6. *$\|\mathcal{A}^*\| = \|\mathcal{A}\|$ where $\|\cdot\|$ denotes the induced norms defined by the inner products.*
7. *$(\mathcal{A}^*)^* = \mathcal{A}$.*

Example (Finite dimensional adjoint) Consider $U = F^n$, $V = F^m$, and $A = [a_{ij}] \in F^{m\times n}$ both with the inner product giving rise to the Euclidean norm and define a linear map $\mathcal{A} : U \to V$ by $\mathcal{A}(x) = Ax$ for all $x \in U$. Then for all $x \in U$, $y \in V$

$$\begin{aligned}\langle y, \mathcal{A}(x)\rangle_{F^m} &= \overline{y}^T Ax = \sum_{i=1}^{m} \overline{y_i}(Ax)_i = \sum_{i=1}^{m} \overline{y_i} \sum_{j=1}^{n} a_{ij}x_j = \sum_{j=1}^{n} \overline{\sum_{i=1}^{m} \overline{a_{ij}} y_i} x_j \\ &= \sum_{j=1}^{n} \overline{(\overline{A}^T y)_j} x_j = \overline{\overline{A}y}^T x = \langle \overline{A}^T y, x\rangle_{F^n} = \langle \mathcal{A}^*(y), x\rangle_{F^n}.\end{aligned}$$

Therefore the adjoint of the linear map defined by the matrix $A \in F^{m\times n}$ is the linear map defined by the matrix $\overline{A}^T = [\overline{a_{ji}}] \in F^{n\times m}$, the complex conjugate transpose (known as the Hermitian transpose) of A. If in addition $F = \mathbb{R}$, then the adjoint of the linear map defined by the matrix A is the linear map defined by the matrix $A^T = [a_{ji}] \in F^{n\times m}$, i.e. simply the transpose of A. ∎

As we have seen in Chapter 2, any linear map between finite dimensional vector spaces can be represented by a matrix, by fixing bases for the domain and co-domain spaces. The above example therefore demonstrates that when it comes to linear maps between finite dimensional spaces the adjoint operation always involves taking the complex conjugate transpose of a matrix. Recall also that linear maps between finite dimensional spaces are always continuous (Theorem 3.4). For infinite dimensional spaces the situation is in general more complicated.

Example (Infinite dimensional adjoint) Let $U = (L^2([t_0, t_1], F^m)$ be the space of square integrable functions $u : [t_0, t_1] \to F^m$ and let $V = (F^n, F, \langle\cdot,\cdot\rangle_{F^n})$ be a finite dimensional space with the inner product giving rise to the Euclidean norm. Consider $G(\cdot) \in L^2([t_0, t_1], F^{n\times m})$ (think for example of $G(t) = \Phi(t_1, t)B(t)$ for $t \in [t_0, t_1]$ for a linear system). Define a function $\mathcal{A} : U \to V$ by

$$\mathcal{A}(u(\cdot)) = \int_{t_0}^{t_1} G(\tau)u(\tau)d\tau.$$

It is easy to see that $\mathcal{A}$ is well defined and linear. Assume also that $G(t)$ is well behaved enough (e.g. continuous and bounded) so that that $\mathcal{A}$ is also continuous. For arbitrary $x \in V$, $u(\cdot) \in U$

$$\begin{aligned}\langle x, \mathcal{A}(u(\cdot))\rangle_{F^m} &= \overline{x}^T \left(\int_{t_0}^{t_1} G(\tau)u(\tau)d\tau\right) \\ &= \int_{t_0}^{t_1} \overline{x}^T G(\tau)u(\tau)d\tau \\ &= \int_{t_0}^{t_1} \overline{\overline{G(\tau)}^T x}^T u(\tau)d\tau \\ &= \langle \overline{G}^T x, u\rangle_2.\end{aligned}$$

Therefore the adjoint $\mathcal{A}^* : F^n \to L^2([t_0, t_1], F^m)$ of the linear map $\mathcal{A}$ is given by

$$(\mathcal{A}^*(x))(\cdot) = \overline{G(\cdot)}^T x : [t_0, t_1] \to F^m$$

where as before $\overline{G(\tau)}^T \in F^{m\times n}$ denotes the Hermitian transpose of $G(\tau) \in F^{n\times m}$. ∎

Definition 7.7 *Let $(H, F, \langle\cdot,\cdot\rangle)$ be a Hilbert space and $\mathcal{A} : H \to H$ be linear and continuous. $\mathcal{A}$ is called self-adjoint if and only if $\mathcal{A}^* = \mathcal{A}$, in other words for all $x, y \in H$*

$$\langle x, \mathcal{A}(y)\rangle = \langle \mathcal{A}(x), y\rangle.$$

Example (Finite dimensional self-adjoint map) Let $H = F^n$ with the standard inner product and $A = [a_{ij}] \in F^{n\times n}$. The linear map $\mathcal{A} : F^n \to F^n$ defined by $\mathcal{A}(x) = Ax$ is self adjoint if and only if $\overline{A}^T = A$, or in other words $a_{ij} = \overline{a_{ji}}$ for all $i, j = 1, \dots n$. Such matrices are called Hermitian. If in addition $F = \mathbb{R}$, $\mathcal{A}$ is self-adjoint if and only if $a_{ij} = a_{ji}$, i.e. $A^T = A$. Such matrices are called symmetric. ■

Example (Infinite dimensional self-adjoint map) Let $H = L^2([t_0, t_1], \mathbb{R})$ and $K(\cdot, \cdot) : [t_0, t_1] \times [t_0, t_1] \to \mathbb{R}$ such that

$$\int_{t_0}^{t_1} \int_{t_0}^{t_1} |K(t, \tau)|^2 dt d\tau < \infty$$

(think for example of $\Phi(t, \tau)B(\tau)$.) Define $\mathcal{A} : L^2([t_0, t_1], \mathbb{R}) \to L^2([t_0, t_1], \mathbb{R})$ by

$$(\mathcal{A}(u(\cdot)))(t) = \int_{t_0}^{t_1} K(t, \tau)u(\tau)d\tau \text{ for all } t \in [t_0, t_1].$$

Exercise 7.7 Show that $\mathcal{A}$ is linear, continuous and self-adjoint.

■

Let (H, F) be a linear space and $\mathcal{A} : H \to H$ a linear map. Recall that (Definition 2.14) an element $\lambda \in F$ is called an eigenvalue of $\mathcal{A}$ if and only if there exists $v \in H$ with $v \neq 0$ such that $\mathcal{A}(v) = \lambda v$; in this case v is called an eigenvector corresponding to λ.

Fact 7.6 *Let $(H, F, \langle \cdot, \cdot \rangle)$ be a Hilbert space and $\mathcal{A} : H \to H$ be linear, continuous and self-adjoint. Then*

1. *All the eigenvalues of $\mathcal{A}$ are real.*
2. *If λ_i and λ_j are eigenvalues with corresponding eigenvectors $v_i, v_j \in H$ and $\lambda_i \neq \lambda_j$ then v_i is orthogonal to v_j.*

Proof: Part 1: Let λ be an eigenvalue and $v \in H$ the corresponding eigenvector. Then

$$\langle v, \mathcal{A}(v) \rangle = \langle v, \lambda v \rangle = \lambda \langle v, v \rangle = \lambda \|v\|^2$$

Since $\mathcal{A} = \mathcal{A}^*$, however, we also have that

$$\langle v, \mathcal{A}(v) \rangle = \langle \mathcal{A}(v), v \rangle = \overline{\langle v, \mathcal{A}(v) \rangle} = \overline{\langle v, \lambda v \rangle} = \overline{\lambda} \|v\|^2.$$

Since v is an eigenvector we must have $v \neq 0$, hence $\lambda = \overline{\lambda}$ and λ is real.

Part 2. By definition

$$\left.\begin{array}{rcl} \mathcal{A}(v_i) = \lambda_i v_i & \Rightarrow & \langle v_j, \mathcal{A}(v_i) \rangle = \lambda_i \langle v_j, v_i \rangle \\ \mathcal{A}(v_j) = \lambda_j v_j & \Rightarrow & \langle \mathcal{A}(v_j), v_i \rangle = \lambda_j \langle v_j, v_i \rangle \\ \mathcal{A}^* = \mathcal{A} & \Rightarrow & \langle v_j, \mathcal{A}(v_i) \rangle = \langle \mathcal{A}(v_j), v_i \rangle \end{array}\right\} \Rightarrow (\lambda_i - \lambda_j)\langle v_j, v_i \rangle = 0.$$

Therefore if $\lambda_i \neq \lambda_j$ we must have $\langle v_j, v_i \rangle = 0$ and v_i and v_j are orthogonal. ■

7.5 Finite rank lemma

Let U, V be linear spaces and consider a linear map $\mathcal{A} : U \to V$. Recall that the range space, $\textsc{Range}(\mathcal{A})$, and null space, $\textsc{Null}(\mathcal{A})$, of $\mathcal{A}$ defined by

$$\begin{aligned}\textsc{Range}(\mathcal{A}) &= \{v \in V \mid \exists u \in U, v = \mathcal{A}(u)\}\\ \textsc{Null}(\mathcal{A}) &= \{u \in U \mid \mathcal{A}(u) = 0\}\end{aligned}$$

are linear subspaces of V and U respectively.

Theorem 7.5 (Finite Rank Lemma) *Let $F = \mathbb{R}$ or $F = \mathbb{C}$, let $(H, F, \langle\cdot,\cdot\rangle)$ be a Hilbert space and recall that $(F^m, F, \langle\cdot,\cdot\rangle_{F^m})$ is a finite dimensional Hilbert space. Let $\mathcal{A} : H \to F^m$ be a continuous linear map and $\mathcal{A}^* : F^m \to H$ be its adjoint. Then:*

1. *$\mathcal{A} \circ \mathcal{A}^* : F^m \to F^m$ and $\mathcal{A}^* \circ \mathcal{A} : H \to H$ are linear, continuous and self adjoint.*
2. *$H = \textsc{Range}(\mathcal{A}^*) \overset{\perp}{\oplus} \textsc{Null}(\mathcal{A})$, i.e. $\textsc{Range}(\mathcal{A}^*) \cap \textsc{Null}(\mathcal{A}) = \{0\}$, $\textsc{Range}(\mathcal{A}^*) = (\textsc{Null}(\mathcal{A}))^\perp$ and $H = \textsc{Range}(\mathcal{A}) \oplus \textsc{Null}(\mathcal{A})$. Likewise, $F^m = \textsc{Range}(\mathcal{A}) \overset{\perp}{\oplus} \textsc{Null}(\mathcal{A}^*)$.*
3. *The restriction of the linear map $\mathcal{A}$ to the range space of $\mathcal{A}^*$,*

   $$\mathcal{A}|_{\textsc{Range}(\mathcal{A}^*)} : \textsc{Range}(\mathcal{A}^*) \to F^m$$

   is a bijection between $\textsc{Range}(\mathcal{A}^*)$ and $\textsc{Range}(\mathcal{A})$.
4. *$\textsc{Null}(\mathcal{A} \circ \mathcal{A}^*) = \textsc{Null}(\mathcal{A}^*)$ and $\textsc{Range}(\mathcal{A} \circ \mathcal{A}^*) = \textsc{Range}(\mathcal{A})$.*
5. *The restriction of the linear map $\mathcal{A}^*$ to the range space of $\mathcal{A}$,*

   $$\mathcal{A}^*|_{\textsc{Range}(\mathcal{A})} : \textsc{Range}(\mathcal{A}) \to H$$

   is a bijection between $\textsc{Range}(\mathcal{A})$ and $\textsc{Range}(\mathcal{A}^*)$.
6. *$\textsc{Null}(\mathcal{A}^* \circ \mathcal{A}) = \textsc{Null}(\mathcal{A})$ and $\textsc{Range}(\mathcal{A}^* \circ \mathcal{A}) = \textsc{Range}(\mathcal{A}^*)$.*

Proof: Part 1: $\mathcal{A}^*$ is linear and continuous and hence $\mathcal{A} \circ \mathcal{A}^*$ and $\mathcal{A}^* \circ \mathcal{A}$ are linear, continuous and self-adjoint by Theorem 7.4.

Part 2: Recall that $\textsc{Range}(\mathcal{A}) \subseteq F^m$ and therefore is finite dimensional. Moreover, $\textsc{Dim}[\textsc{Range}(\mathcal{A}^*)] \leq \textsc{Dim}[F^m] = m$, therefore $\textsc{Range}(\mathcal{A}^*)$ is also finite dimensional. Therefore both $\textsc{Range}(\mathcal{A})$ and $\textsc{Range}(\mathcal{A}^*)$ are closed (see Problem 3.3), hence by Theorem 7.3

$$H = \textsc{Range}(\mathcal{A}^*) \oplus \textsc{Range}(\mathcal{A}^*)^\perp \text{ and } F^m = \textsc{Range}(\mathcal{A}) \oplus \textsc{Range}(\mathcal{A})^\perp.$$

But

$$\begin{aligned}x \in \textsc{Range}(\mathcal{A})^\perp &\Leftrightarrow \langle x, v\rangle_{F^m} = 0 \ \ \forall v \in \textsc{Range}(\mathcal{A})\\ &\Leftrightarrow \langle x, \mathcal{A}(y)\rangle_{F^m} = 0 \ \ \forall y \in H\\ &\Leftrightarrow \langle \mathcal{A}^*(x), y\rangle_H = 0 \ \ \forall y \in H\\ &\Leftrightarrow \mathcal{A}^*(x) = 0 \text{ (e.g. take } \{y_i\} \text{ a basis for } H)\\ &\Leftrightarrow x \in \textsc{Null}(\mathcal{A}^*).\end{aligned}$$

Therefore, $\textsc{Null}(\mathcal{A}^*) = \textsc{Range}(\mathcal{A})^\perp$ therefore $F^m = \textsc{Range}(\mathcal{A}) \overset{\perp}{\oplus} \textsc{Null}(\mathcal{A}^*)$. The proof for H is similar.

Part 3: Consider the restriction of the linear map $\mathcal{A}$ to the range space of $\mathcal{A}^*$

$$\mathcal{A}|_{\text{RANGE}(\mathcal{A}^*)} : \text{RANGE}(\mathcal{A}^*) \to F^m.$$

Clearly for all $x \in \text{RANGE}(\mathcal{A}^*)$, $\mathcal{A}|_{\text{RANGE}(\mathcal{A}^*)}(x) = \mathcal{A}(x) \in \text{RANGE}(\mathcal{A})$ therefore

$$\mathcal{A}|_{\text{RANGE}(\mathcal{A}^*)} : \text{RANGE}(\mathcal{A}^*) \to \text{RANGE}(\mathcal{A}).$$

We need to show that this map is injective and surjective.

To show that it is surjective, note that for all $y \in \text{RANGE}(\mathcal{A})$ there exists $x \in H$ such that $y = \mathcal{A}(x)$. But since $H = \text{RANGE}(\mathcal{A}^*) \overset{\perp}{\oplus} \text{NULL}(\mathcal{A})$, $x = x_1 + x_2$ for some $x_1 \in \text{RANGE}(\mathcal{A}^*)$, $x_2 \in \text{NULL}(\mathcal{A})$. Then

$$y = \mathcal{A}(x_1 + x_2) = \mathcal{A}(x_1) + \mathcal{A}(x_2) = \mathcal{A}(x_1) + 0 = \mathcal{A}(x_1).$$

Hence, for all $y \in \text{RANGE}(\mathcal{A})$ there exists $x_1 \in \text{RANGE}(\mathcal{A}^*)$ such that $y = \mathcal{A}(x_1)$ and the map is surjective.

To show that the map is injective, recall that this is the case if and only if $\text{NULL}(\mathcal{A}|_{\text{RANGE}(\mathcal{A}^*)}) = \{0\}$. Consider an arbitrary $y \in \text{NULL}(\mathcal{A}|_{\text{RANGE}(\mathcal{A}^*)})$. Then $y \in \text{RANGE}(\mathcal{A}^*)$ and there exists $x \in F^m$ such that $y = \mathcal{A}^*(x)$ and moreover $\mathcal{A}(y) = 0$. Therefore $y \in \text{NULL}(\mathcal{A}) \cap \text{RANGE}(\mathcal{A}^*) = \{0\}$ since $H = \text{RANGE}(\mathcal{A}^*) \overset{\perp}{\oplus} \text{NULL}(\mathcal{A})$ and the map is injective.

Part 4: To show that $\text{NULL}(\mathcal{A} \circ \mathcal{A}^*) = \text{NULL}(\mathcal{A}^*)$ consider first an arbitrary $x \in \text{NULL}(\mathcal{A} \circ \mathcal{A}^*)$. Then

$$\begin{aligned}
\mathcal{A} \circ \mathcal{A}^*(x) = 0 &\Rightarrow \langle x, \mathcal{A} \circ \mathcal{A}^*(x) \rangle_{F^m} = 0 \\
&\Rightarrow \langle \mathcal{A}^*(x), \mathcal{A}^*(x) \rangle_H = 0 \\
&\Rightarrow \|\mathcal{A}^*(x)\|_H^2 = 0 \\
&\Rightarrow \mathcal{A}^*(x) = 0 \\
&\Rightarrow x \in \text{NULL}(\mathcal{A}^*).
\end{aligned}$$

Hence $\text{NULL}(\mathcal{A} \circ \mathcal{A}^*) \subseteq \text{NULL}(\mathcal{A}^*)$. Conversely, consider an arbitrary $x \in \text{NULL}(\mathcal{A}^*)$. Then

$$\mathcal{A}^* x = 0 \Rightarrow \mathcal{A} \circ \mathcal{A}^*(x) = 0 \Rightarrow x \in \text{NULL}(\mathcal{A} \circ \mathcal{A}^*)$$

and hence $\text{NULL}(\mathcal{A}^*) \subseteq \text{NULL}(\mathcal{A} \circ \mathcal{A}^*)$. Overall, $\text{NULL}(\mathcal{A} \circ \mathcal{A}^*) = \text{NULL}(\mathcal{A}^*)$.

Finally, to show that $\text{RANGE}(\mathcal{A}) = \text{RANGE}(\mathcal{A} \circ \mathcal{A}^*)$ note that

$$\begin{aligned}
\text{RANGE}(\mathcal{A}) &= \{y \in F^m \mid \exists x \in H, y = \mathcal{A}(x)\} \\
&= \{y \in F^m \mid \exists x \in \text{RANGE}(\mathcal{A}^*), y = \mathcal{A}(x)\} \text{ (by Part 3)} \\
&= \{y \in F^m \mid \exists u \in F^m, y = \mathcal{A}(\mathcal{A}^*(u))\} \\
&= \{y \in F^m \mid \exists u \in F^m, y = (\mathcal{A} \circ \mathcal{A}^*)(u)\} \\
&= \text{RANGE}(\mathcal{A} \circ \mathcal{A}^*).
\end{aligned}$$

The proofs of Part 5 and Part 6 are analogous to those of Part 3 and Part 4 respectively. ∎

Notice that the assumption that the co-domain of $\mathcal{A}$ is finite dimensional is only used to establish that the ranges of $\mathcal{A}$ and $\mathcal{A}^*$ are closed sets, to be able to invoke Theorem 7.3. In fact, Theorem 7.5 holds more generally if we replace the spaces by their closures.

7.6 Application: Matrix pseudo-inverse

Consider the finite dimensional Hilbert spaces $(\mathbb{R}^n, \mathbb{R}, \langle\cdot,\cdot\rangle_{\mathbb{R}^n})$ and $(\mathbb{R}^m, \mathbb{R}, \langle\cdot,\cdot\rangle_{\mathbb{R}^m})$ with the inner product giving rise to the corresponding Euclidean norms. Consider a matrix $A \in \mathbb{R}^{m\times n}$ and the linear map

$$\begin{array}{rccc} \mathcal{A}: & \mathbb{R}^n & \to & \mathbb{R}^m \\ & x & \mapsto & A\cdot x. \end{array}$$

Consider also the adjoint map and recall that it is represented by the transpose of A,

$$\begin{array}{rccc} \mathcal{A}^*: & \mathbb{R}^n & \to & \mathbb{R}^m \\ & y & \mapsto & A^T\cdot y. \end{array}$$

Let us also introduce some notation for the rows and columns of the matrix A

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad a_{\bullet i} = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{mi} \end{bmatrix}, \quad a_{j\bullet} = \begin{bmatrix} a_{j1} & \cdots & a_{jn} \end{bmatrix}$$

for $i = 1,\dots,n$ and $j = 1,\dots,m$. Note that by definition

$$\textsc{Range}(\mathcal{A}) = \textsc{Span}\{a_{\bullet 1},\dots,a_{\bullet n}\} \text{ and } \textsc{Range}(\mathcal{A}^*) = \textsc{Span}\{a_{1\bullet}^T,\dots,a_{m\bullet}^T\}.$$

Given $b \in \mathbb{R}^m$ we seek solutions $x \in \mathbb{R}^n$ to the set of linear equations $Ax = b$.

Exercise 7.8 Show that an $x \in \mathbb{R}^n$ such that $Ax = b$ exists if and only if $b \in \textsc{Range}(\mathcal{A})$. If such an x exists, then it is unique if and only if $\textsc{Null}(\mathcal{A}) = \{0\}$.

If $m = n$ and the matrix A is invertible, then $\mathcal{A}$ is bijective (Theorem 2.3) and a unique solution exists for all $b \in \mathbb{R}^n$. This unique solution is given by $x = A^{-1}b$.

What if $m \neq n$ or A is not invertible? In this case one could have multiple solutions (if $\textsc{Null}(\mathcal{A}) \neq \{0\}$), or no solutions (if $b \notin \textsc{Range}(\mathcal{A}) \neq \mathbb{R}^m$). Still, the Finite Rank Lemma allows us to say something about the linear equation $Ax = b$ even in this case.

Consider first the case where $\textsc{Range}(\mathcal{A}) = \mathbb{R}^m$.

Exercise 7.9 Show that if $\textsc{Range}(\mathcal{A}) = \mathbb{R}^m$ then $m \leq n$, $\textsc{Null}(\mathcal{A}^*) = \{0\}$, and $\{a_{j\bullet}^T\}_{j=1}^m$ are linearly independent.

By the Finite Rank Lemma the map

$$\mathcal{A}|_{\textsc{Range}(\mathcal{A}^*)} : \textsc{Range}(\mathcal{A}^*) \to \textsc{Range}(\mathcal{A}) = \mathbb{R}^m$$

is bijective. Therefore for a given $b \in \mathbb{R}^m$ there exists a unique $\tilde{x} \in \textsc{Range}(\mathcal{A}^*)$ such that $A\tilde{x} = b$. Note that, even though $\tilde{x}$ is the only element of $\textsc{Range}(\mathcal{A}^*)$ with this property, it is not necessarily the only element in $\mathbb{R}^n$. If $\textsc{Null}(\mathcal{A}) \neq \{0\}$, one can pick any $\hat{x} \in \textsc{Null}(\mathcal{A})$ and add it to $\tilde{x}$. The resulting $x = \tilde{x} + \hat{x}$ will satisfy

$$\mathcal{A}(\tilde{x} + \hat{x}) = \mathcal{A}(\tilde{x}) + \mathcal{A}(\hat{x}) = A\tilde{x} + 0 = b.$$

The resulting x, however, will have norm at least as big as $\tilde{x}$. Since by construction $\tilde{x} \in \textsc{Range}(\mathcal{A}^*)$ and $\hat{x} \in \textsc{Null}(\mathcal{A})$ and by the Finite Rank Lemma $\textsc{Range}(\mathcal{A}^*)$ and $\textsc{Null}(\mathcal{A})$ are orthogonal to each other,

$$\|x\|^2 = \langle x, x\rangle = \langle \tilde{x} + \hat{x}, \tilde{x} + \hat{x}\rangle = \langle \tilde{x}, \tilde{x}\rangle + \langle \tilde{x}, \hat{x}\rangle + \langle \hat{x}, \tilde{x}\rangle + \langle \hat{x}, \hat{x}\rangle = \|\tilde{x}\|^2 + \|\hat{x}\|^2 \geq \|\tilde{x}\|^2.$$

In other words, among the (possibly infinitely many) $x \in \mathbb{R}^n$ for which $Ax = b$ the unique $\tilde{x} \in \text{RANGE}(\mathcal{A}^*)$ is the one with the smallest magnitude.

Can we find this unique $\tilde{x}$ corresponding to a given $b \in \mathbb{R}^m$? By the Finite Rank Lemma $\text{RANGE}(\mathcal{A} \circ \mathcal{A}^*) = \text{RANGE}(\mathcal{A}) = \mathbb{R}^m$ and $\text{NULL}(\mathcal{A} \circ \mathcal{A}^*) = \text{NULL}(\mathcal{A}^*) = \{0\}$. Hence, by Theorem 2.3, $\mathcal{A} \circ \mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^m$ is bijective. Since by Theorem 2.6 composition of linear maps corresponds to multiplication of the corresponding matrices

$$\begin{array}{rccc} \mathcal{A} \circ \mathcal{A}^* : & \mathbb{R}^m & \to & \mathbb{R}^m \\ & y & \mapsto & AA^T y. \end{array}$$

and the matrix $AA^T \in \mathbb{R}^{m \times m}$ is invertible. Consider then

$$\tilde{x} = A^T (AA^T)^{-1} b.$$

It is easy to see that this is the unique $\tilde{x} \in \text{RANGE}(\mathcal{A})$ that satisfies $A\tilde{x} = b$. We summarise the discussion in the following theorem.

Theorem 7.6 *Let $A \in \mathbb{R}^{m \times n}$ and consider the linear map $\mathcal{A}(x) = Ax$ and some $b \in \mathbb{R}^m$. Assume that $\text{RANGE}(\mathcal{A}) = \mathbb{R}^m$ and define $\tilde{x} = A^T (AA^T)^{-1} b$. Then*

1. *$\tilde{x}$ is the unique element of $\text{RANGE}(\mathcal{A}^*)$ that satisfies $A\tilde{x} = b$.*
2. *$\tilde{x}$ is the orthogonal projection onto $\text{RANGE}(\mathcal{A}^*)$ of any $x \in \mathbb{R}^n$ such that $Ax = b$.*
3. *$\tilde{x}$ is unique minimiser of $\|x\|$ subject to $Ax = b$.*

Proof: For Part 1, let $y = (AA^T)^{-1} b \in \mathbb{R}^m$ and note that $\tilde{x} = A^T y$; hence $\tilde{x} \in \text{RANGE}(\mathcal{A})$. Moreover,

$$\mathcal{A}(\tilde{x}) = A \left(A^T (AA^T)^{-1} b \right) = (AA^T)(AA^T)^{-1} b = b.$$

Hence $A\tilde{x} = b$. Uniqueness follows since, by the Finite Rank Lemma, $\mathcal{A}|_{\text{RANGE}(\mathcal{A}^*)} : \text{RANGE}(\mathcal{A}^*) \to \text{RANGE}(\mathcal{A})$ is bijective.

For Part 2, consider any $x \in \mathbb{R}^n$ such that $Ax = b$. Then

$$\mathcal{A}(x - \tilde{x}) = Ax - A\tilde{x} = b - b = 0$$

and $x - \tilde{x} \in \text{NULL}(\mathcal{A})$. By the Finite Rank Lemma, $\mathbb{R}^n = \text{RANGE}(\mathcal{A}^*) \overset{\perp}{\oplus} \text{NULL}(\mathcal{A})$ hence by Theorem 7.3, $\tilde{x}$ is the orthogonal projection of x onto $\text{RANGE}(\mathcal{A}^*)$ (take $H = \mathbb{R}^n$, $M = \text{RANGE}(\mathcal{A}^*)$ and $M^\perp = \text{NULL}(\mathcal{A})$ in Part 2 of Theorem 7.3).

Part 3 is an immediate consequence of Part 2 and of Part 3 of Theorem 7.3. ∎

The matrix $A^\dagger = A^T (AA^T)^{-1}$ is called the right pseudo-inverse of the matrix A, since $AA^\dagger = 1$. This calculation is indeed very similar to the computation of minimum energy controls that we will encounter in Chapter 8. The only difference is that the finite dimensional Hilbert space $(\mathbb{R}^n, \mathbb{R}, \langle \cdot, \cdot \rangle_{\mathbb{R}^n})$ will be replaced in Chapter 8 by the infinite dimensional Hilbert space $L^2([t_0, t_1], \mathbb{R}^m)$.

Let us now consider the case where $\text{NULL}(\mathcal{A}) = \{0\}$.

Exercise 7.10 Show that if $\text{NULL}(\mathcal{A}^*) = \{0\}$ then $\text{RANGE}(\mathcal{A}^*) = \mathbb{R}^n$, $m \geq n$, and $\{a_{\bullet i}\}_{i=1}^n$ are linearly independent.

Since in this case it is possible that $b \notin \text{RANGE}(\mathcal{A})$ there may not exist an $x \in \mathbb{R}^n$ such that $Ax = b$. A weaker requirement is to find an $x \in \mathbb{R}^n$ that minimises the difference $\|Ax - b\|$; in other words, find an $\tilde{x} \in \mathbb{R}^n$ such that $\|A\tilde{x} - b\| \leq \|Ax - b\|$ for all $x \in \mathbb{R}^n$. Natural questions are of course

whether such an $\tilde{x}$ exists (i.e. whether the minimum is attained) and whether it is unique. We show here that the answer to both is "yes" and provide a way of computing $\tilde{x}$ from A and b.

By the Finite Rank Lemma, $\mathbb{R}^m = \text{RANGE}(\mathcal{A}) \overset{\perp}{\oplus} \text{NULL}(\mathcal{A}^*)$. Therefore, by Theorem 7.3, for the given $b \in \mathbb{R}^m$ there exist unique $\tilde{y} \in \text{RANGE}(\mathcal{A})$ and $\hat{y} \in \text{NULL}(\mathcal{A}^*)$ such that $b = \tilde{y} + \hat{y}$; moreover, $\tilde{y}$ is the unique element of $\text{RANGE}(\mathcal{A})$ that achieves the minimum $\inf\{\|b - y\| \mid y \in \text{RANGE}(\mathcal{A})\}$. By the Finite Rank Lemma, since $\text{RANGE}(\mathcal{A}^*) = \mathbb{R}^n$, the linear map

$$\mathcal{A} : \mathbb{R}^n \to \text{RANGE}(\mathcal{A})$$

is bijective. Therefore, there exists a unique $\tilde{x} \in \mathbb{R}^n$ such that $A\tilde{x} = \tilde{y}$. By the Finite Rank Lemma $\text{RANGE}(\mathcal{A}^* \circ \mathcal{A}) = \text{RANGE}(\mathcal{A}^*) = \mathbb{R}^n$ and $\text{NULL}(\mathcal{A}^* \circ \mathcal{A}) = \text{NULL}(\mathcal{A}) = \{0\}$. Hence, by Theorem 2.3, $\mathcal{A}^* \circ \mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ is bijective. Since by Theorem 2.6 composition of linear maps corresponds to multiplication of the corresponding matrices

$$\begin{array}{rccc} \mathcal{A}^* \circ \mathcal{A} : & \mathbb{R}^n & \to & \mathbb{R}^n \\ & x & \mapsto & A^T A x \end{array}$$

and the matrix $A^T A \in \mathbb{R}^{n \times n}$ is invertible. We claim that

$$\tilde{x} = (A^T A)^{-1} A^T b$$

is the solution to our problem.

Theorem 7.7 *Let $A \in \mathbb{R}^{m \times n}$ and consider the linear map $\mathcal{A}(x) = Ax$ and some $b \in \mathbb{R}^m$. Assume that $\text{NULL}(\mathcal{A}) = \{0\}$ and define $\tilde{x} = (A^T A)^{-1} A^T b$. Then*

1. *$\tilde{x}$ is the unique element of $\mathbb{R}^n$ such that $A\tilde{x}$ is the orthogonal projection of b onto $\text{RANGE}(\mathcal{A})$.*
2. *$\tilde{x}$ is unique minimiser of $\{\|Ax - b\| \mid x \in \mathbb{R}^n\}$.*

Proof: For Part 1, note that

$$A^T(b - A\tilde{x}) = A^T b - A^T A(A^T A)^{-1} A^T b = A^T b - A^T b = 0.$$

Hence $b - A\tilde{x} \in \text{NULL}(\mathcal{A}^*)$ and, since by the Finite Rank Lemma $\mathbb{R}^m = \text{RANGE}(\mathcal{A}) \overset{\perp}{\oplus} \text{NULL}(\mathcal{A}^*)$, $A\tilde{x}$ is the orthogonal projection of b onto $\text{RANGE}(\mathcal{A})$ by Theorem 2.6. The fact that $\tilde{x}$ is the unique element of $\mathbb{R}^n$ with this property follows by the Finite Rank Lemma (bijectivity of $\mathcal{A} : \mathbb{R}^n \to \text{RANGE}(\mathcal{A})$).

Part 2 is an immediate consequence of Part 1 and of Part 3 of Theorem 7.3. ∎

The matrix $A^\ddagger = (A^T A)^{-1} A^T$ is called the left pseudo-inverse of the matrix A, since $A^\ddagger A = 1$. This calculation is indeed very similar to the one we will use in Chapter 8 to determine the initial state that best matches a given zero input response, in the context of our discussion on observability. The only difference is that the finite dimensional Hilbert space $(\mathbb{R}^m, \mathbb{R}, \langle \cdot, \cdot \rangle_{\mathbb{R}^m})$ will be replaced in Chapter 8 by the infinite dimensional Hilbert space $L^2([t_0, t_1], \mathbb{R}^p)$.

Finally, we note that the two calculations, for the case when $\text{RANGE}(\mathcal{A}) = \mathbb{R}^m$ and the case when $\text{NULL}(\mathcal{A}) = \{0\}$, are duals of each other, in the sense that they simply involve invoking the symmetric statements of the Finite Rank Lemma. This duality will also be observed in the case of controllability and observability in Chapter 8 and will allow us to easily extend statements proved for one property to corresponding statements for the other.

Problems for chapter 7

Problem 7.1 (An Example of a Hilbert Space) Let $E = \mathrm{L}_2([t_0, t_1], \mathbb{R})$ be the Hilbert space of square integrable real-valued functions on $[t_0, t_1]$. Let $K : [t_0, t_1] \times [t_0, t_1] \to \mathbb{R}$ be such that

$$\int_{t_0}^{t_1} \int_{t_0}^{t_1} |K(t, \tau)|^2 \, \mathrm{d}t \, \mathrm{d}\tau < \infty. \tag{$\dagger$}$$

Define $A : \mathrm{L}_2([t_0, t_1], \mathbb{R}) \to \mathrm{L}_2([t_0, t_1], \mathbb{R})$ by

$$\big(A(u)\big)(t) = \int_{t_0}^{t_1} K(t, \tau) u(\tau) \mathrm{d}\tau \quad \forall\, t \in [t_0, t_1].$$

Prove that A is linear and continuous. Also prove that A is self-adjoint if K is symmetric.

Chapter 8

Controllability and observability

Let us return now to time varying linear systems

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$
$$y(t) = C(t)x(t) + D(t)u(t).$$

We will investigate the following questions

- Can the input u be used to steer the state of the system to an arbitrary value?
- Can we infer the value of the state by observing the input and output?

We will again start by a general discussion of these questions for non-linear systems. We will then specialize to the case of time varying linear systems and then further to the case of time invariant linear systems.

8.1 Nonlinear systems

Consider the nonlinear system

$$\dot{x}(t) = f(x(t), u(t), t) \tag{8.1}$$
$$y(t) = r(x(t), u(t), t) \tag{8.2}$$

where $t \in \mathbb{R}_+$, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}^n$ and $r : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}^p$ and assume that r is continuous, that f is Lipschitz in x and continuous in u and t, and that $u : \mathbb{R}_+ \to \mathbb{R}^m$ is piecewise continuous. Under these conditions there exist continuous solution maps such that for $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}_+$, $t_1 \geq t_0$, $u(\cdot) : [t_0, t_1] \to \mathbb{R}^m$ and $t \in [t_0, t_1]$ return the solution

$$x(t) = s(t, t_0, x_0, u), \quad y(t) = \rho(t, t_0, x_0, u)$$

of system (8.1)–(8.2). For simplicity, assume further that r is continuous in x and piecewise continuous in t.

Definition 8.1 *Consider $t_0, t_1 \in \mathbb{R}$ with $t_0 \leq t_1$. The input trajectory $u(\cdot) \in PC([t_0, t_1], \mathbb{R}^m)$ steers $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}$ to $(x_1, t_1) \in \mathbb{R}^n \times \mathbb{R}$ if and only if $s(t_1, t_0, x_0, u) = x_1$. The system (8.1)–(8.2) is controllable on $[t_0, t_1]$ if and only if for all $x_0, x_1 \in \mathbb{R}^n$ there exists $u(\cdot) \in PC([t_0, t_1], \mathbb{R}^m)$ that steers (x_0, t_0) to (x_1, t_1).*

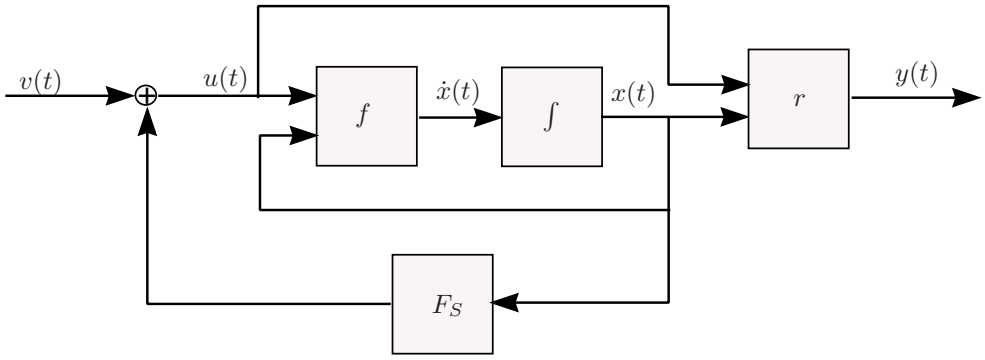

Figure 8.1: State feedback.

Notice that controllability has nothing to do with the output of the system, it is purely an input to state relationship; we will therefore talk about "controllability of the system (8.1)" ignoring Equation (8.2) for the time being. The following fact is an immediate consequence of the definition.

Fact 8.1 *The system (8.1)–(8.2) is controllable on $[t_0, t_1]$ if and only if for all $x_0 \in \mathbb{R}^n$ the function $s(t_1, t_0, x_0, \cdot) : PC([t_0, t_1], \mathbb{R}^m) \to \mathbb{R}^n$ is surjective.*

It is easy to see that controllability is preserved under state and output feedback. Consider a state feedback map $F_S : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$ and a new input variable $v(t) \in \mathbb{R}^m$. Let $u(t) = v(t) + F_S(x(t), t)$ and define the state feedback system

$$\dot{x}(t) = f(x(t), v(t) + F_S(x(t), t), t) = f_S(x(t), v(t), t) \tag{8.3}$$

To prevent technical difficulties assume that F_S is continuous in x and piecewise continuous in t. One can picture the action of the feedback map F_S as shown in Figure 8.1.

Likewise, consider a output feedback map $F_O : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^m$ and assume that r does not explicitly depend on u, i.e. $y(t) = r(x(t), t)$. Consider a new input variable $w(t) \in \mathbb{R}^m$, let $u(t) = w(t) + F_O(y(t), t)$ and define the output feedback system

$$\dot{x}(t) = f(x(t), w(t) + F_O(y(t), t), t) = f(x(t), w(t) + F_O(r(x(t), t), t), t) = f_O(x(t), w(t), t) \tag{8.4}$$

To prevent technical difficulties assume that F_O is continuous in y and piecewise continuous in t. One can picture the action of the feedback map F_O as shown in Figure 8.2.

Theorem 8.1 *Consider $t_o, t_1 \in \mathbb{R}$ with $t_0 \leq t_0$. The following statements are equivalent:*

1. *The system of Equation (8.1) is controllable on $[t_0, t_1]$.*
2. *The system of Equation (8.3) is controllable on $[t_0, t_1]$.*
3. *The system of Equation (8.4) is controllable on $[t_0, t_1]$.*

Proof: The proof requires only careful application of the definitions. For example, to show that if the system of Equation (8.1) is controllable on $[t_0, t_1]$ then so is the system of Equation (8.3), consider arbitrary $x_0, x_1 \in \mathbb{R}^n$ and look for $v(\cdot) \in PC([t_0, t_1], \mathbb{R}^m)$ that steers (x_0, t_0) to (x_1, t_1). Since (8.1) is controllable on $[t_0, t_1]$ there exists $u(\cdot) \in PC([t_0, t_1], \mathbb{R}^m)$ such that the unique solution, $s(t, t_0, x_0, u)$, of (8.1) satisfies $s(t_1, t_0, x_0, u) = x_1$. Define $v(t) = u(t) - F_S(s(t, t_0, x_0, u), t)$; clearly $v(\cdot) : [t_0, t_1] \to \mathbb{R}^m$ is piecewise continuous since u is piecewise continuous, F_S is assumed to be continuous in its

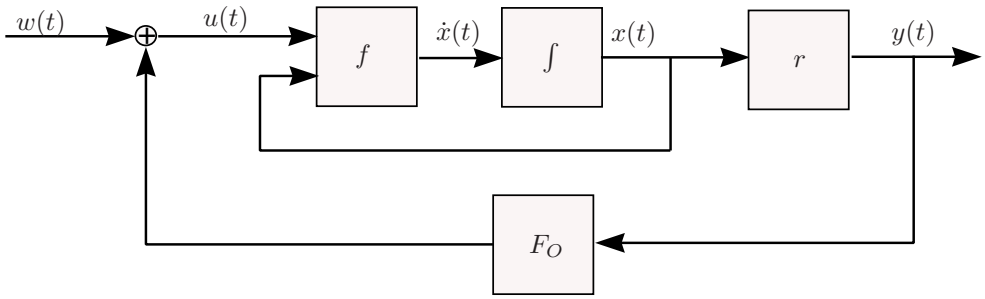

Figure 8.2: Output feedback.

first argument and piecewise continuous in the second, and, by definition, $s(\cdot, t_0, x_0, u)$ is continuous. Moreover, under $v(\cdot)$ to system (8.3) evolves according to

$$\dot{x}(t) = f(x(t), v(t) + F_S(x(t), t), t) = f(x(t), u(t) - F_S(s(t, t_0, x_0, u), t) + F_S(x(t), t), t)$$

starting at $x(t_0) = x_0$. Clearly $x(t) = s(t, t_0, x_0, u)$ satisfies both the initial condition and the differential equation and therefore is also the unique solution for system (8.3) under the input $v(\cdot)$. Hence, the proposed input $v(\cdot)$ steers (x_0, t_0) to (x_1, t_1).

The remaining parts are similar and are left as an exercise. ∎

Dual to the definition of controllability is that of observability.

Definition 8.2 *Consider $t_0, t_1 \in \mathbb{R}$ with $t_0 \leq t_1$. The system (8.1)–(8.2) is observable on $[t_0, t_1]$ if and only if for all $x_0 \in \mathbb{R}^n$ and all $u(\cdot) \in PC([t_0, t_1], \mathbb{R}^m)$, given $u(\cdot) : [t_0, t_1] \to \mathbb{R}^m$ and the corresponding output $y(\cdot) = \rho(\cdot, t_0, x_0, u) : [t_0, t_1] \to \mathbb{R}^p$ the value of x_0 can be uniquely determined.*

A slight rephrasing of the definition leads to the following fact.

Fact 8.2 *The system (8.1)–(8.2) is observable on $[t_0, t_1]$ if and only if for all $u(\cdot) \in PC([t_0, t_1], \mathbb{R}^m)$ the function*

$$\begin{aligned} \rho(\cdot, t_0, \odot, u) : \mathbb{R}^n &\longrightarrow PC([t_0, t_1], \mathbb{R}^p) \\ x_0 &\longmapsto \rho(\cdot, t_0, x_0, u) : [t_0, t_1] \to \mathbb{R}^p \end{aligned}$$

is injective.

It is easy to see that if we establish the value of $x_0 \in \mathbb{R}^n$ we can in fact reconstruct the value of $x(t)$ for all $t \in [t_0, t_1]$; this is because, by uniqueness, $x(t) = s(t, t_0, x_0, u)$ is uniquely determined once we know t, t_0, x_0 and $u : [t_0, t] \to \mathbb{R}^m$.

It is easy to see that observability is preserved under output feedback and input feed-forward. Consider a input feed-forward map $F_F : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^p$ and a new output variable $z(t) \in \mathbb{R}^p$. Let $z(t) = y(t) + F_F(u(t), t)$ and define the input feed-forward system with

$$z(t) = r(x(t), u(t), t) + F_F(u(t), t) = r_F(x(t), u(t), t). \tag{8.5}$$

To prevent technical difficulties assume that F_F is continuous in u and piecewise continuous in t. One can picture the action of the feed-forward map F_F as shown in Figure 8.3.

Theorem 8.2 *The following statements are equivalent:*

1. *The system (8.1)–(8.2) is observable.*

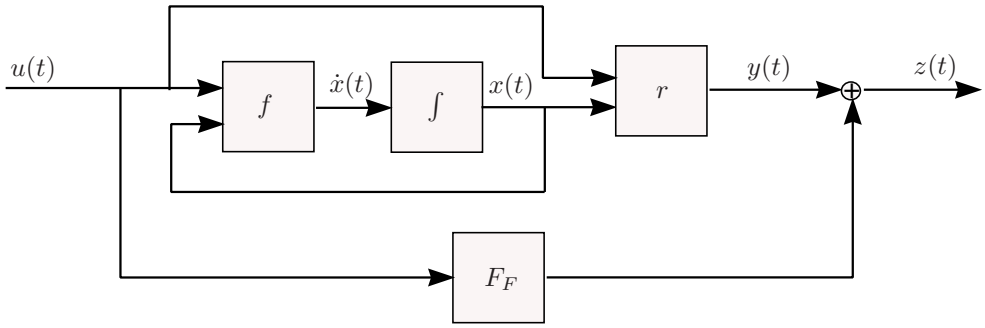

Figure 8.3: Input feed-forward.

2. *The system* (8.4)–(8.2) *is observable.*
3. *The system* (8.1)–(8.5) *is observable.*

Proof: Exercise. ∎

Notice that observability is not preserved by state feedback. If system (8.1)–(8.2) is observable, system (8.3)–(8.2) may or may not be observable and vice versa.

Exercise 8.1 Provide a simple example of a linear system whose observability properties are altered by state feedback.

In nonlinear control, feedback linearization provides a dramatic demonstration of the fact that state feedback can alter observability. In feedback linearization state feedback and a coordinate transformation is used to alter the dynamics of some of the states, so that system behaves like a linear system. The price to pay for this is that the remaining states (the so-called zero dynamics) are disconnected from the output, i.e. are rendered unobservable; the reader is referred to [17, 10, 14] for more information on this topic.

8.2 Linear time varying systems: Controllability

Let us now see how the additional structure afforded by linearity can be used to derive precise conditions for controllability and observability. For most of the discussion we will need to assume that the input trajectories of the system take values in an appropriate Hilbert space, to be able to apply the Finite Rank Lemma and perform projections. Since the space of piecewise continuous input trajectories used up to now (e.g. for the existence and uniqueness arguments) is not complete we will assume that input trajectories take values in the somewhat larger Hilbert space of square integrable functions. More specifically, given $t_0, t_1 \in \mathbb{R}$ with $t_0 \leq t_1$ we will consider input trajectories $u(\cdot) : [t_0, t_1] \to \mathbb{R}^m$ and assume that $u \in (L^2([t_0, t_1], \mathbb{R}^m), \langle \cdot, \cdot \rangle_2)$ with the inner product defined by

$$\langle u, \hat{u} \rangle_2 = \int_{t_0}^{t_1} u(t)^T \hat{u}(t) dt.$$

for $u, \hat{u} \in L^2([t_0, t_1], \mathbb{R}^m)$. Note that by considering this extended space we can no longer assume that the input trajectories will be piecewise continuous, and hence are no longer covered by the existence uniqueness argument of Theorem 3.6. We will alleviate this difficulty, however, by showing that the properties of interest here (e.g. steering the system from a given initial state to a desired final state) can indeed be achieved by piecewise continuous trajectories.

Since controllability is an input to state property only the matrices $A(\cdot)$ and $B(\cdot)$ come into play. To simplify the notation we give the following definition.

Definition 8.3 *The pair $(A(\cdot), B(\cdot))$ is controllable on $[t_0, t_1]$ if and only if for all $x_0, x_1 \in \mathbb{R}^n$ there exists $u : [t_0, t_1] \to \mathbb{R}^m$ that steers (x_0, t_0) to (x_1, t_1), i.e.*

$$x_1 = \Phi(t_1, t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau.$$

Lemma 8.1 *The following statements are equivalent:*

1. *$(A(\cdot), B(\cdot))$ is controllable on $[t_0, t_1]$.*
2. *For all $x_0 \in \mathbb{R}^n$ there exists $u : [t_0, t_1] \to \mathbb{R}^m$ that steers (x_0, t_0) to $(0, t_1)$.*
3. *For all $x_1 \in \mathbb{R}^n$ there exists $u : [t_0, t_1] \to \mathbb{R}^m$ that steers $(0, t_0)$ to (x_1, t_1).*

Proof: (Sketch). $1 \Rightarrow 2$ and $1 \Rightarrow 3$: Obvious.

$2 \Rightarrow 1$: For arbitrary $x_0, x_1 \in \mathbb{R}^n$ let $\hat{x}_0 = x_0 - \Phi(t_0, t_1)x_1$ and consider the input $\hat{u} : [t_0, t_1] \to \mathbb{R}^m$ that steers $(\hat{x}_0, t_0)$ to $(0, t_1)$. It is easy to see that the same input steers (x_0, t_0) to (x_1, t_1).

$3 \Rightarrow 1$: Exercise. ∎

Exercise 8.2 The three statements of Lemma 8.1 are not equivalent for nonlinear systems. Where does linearity come into play in the proof?

Lemma 8.1 states that for linear systems controllability is equivalent to controllability to zero and to reachability from zero. In the subsequent discussion will use the last of the three equivalent statements to simplify the analysis; Lemma 8.1 implies that this can be done without loss of generality. In line with this, we introduce the following definition.

Definition 8.4 *A state $x_1 \in \mathbb{R}^n$ is reachable on $[t_0, t_1]$ by the pair $(A(\cdot), B(\cdot))$ if and only if there exists $u(\cdot) \in L^2([t_0, t_1], \mathbb{R}^m)$ that steers $(0, t_0)$ to (x_1, t_1). The reachability map on $[t_0, t_1]$ of the pair $(A(\cdot), B(\cdot))$ is the function*

$$\begin{array}{rccc} \mathcal{L}_r : & L^2([t_0, t_1], \mathbb{R}^m) & \longrightarrow & \mathbb{R}^n \\ & u & \longmapsto & \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau. \end{array}$$

Lemma 8.2 *$\mathcal{L}_r$ is linear and continuous. The set of reachable states is the linear subspace* $\textsc{Range}(\mathcal{L}_r)$.

Proof: The fact that $\mathcal{L}_r$ is linear is obvious from the properties of the integral. The fact that reachable states are equal to $\textsc{Range}(\mathcal{L}_r)$ is immediate from the definition of reachable states. To show that $\mathcal{L}_r$ is continuous let $\|\cdot\|$ denote the 2-norm in $\mathbb{R}^n$ and note that for all $u(\cdot) \in L^2([t_0, t_1], \mathbb{R}^m)$

$$\begin{aligned} \|\mathcal{L}_r(u)\|^2 &= \left\| \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau \right\|^2 \\ &\leq \int_{t_0}^{t_1} \|\Phi(t_1, \tau)B(\tau)\|^2 \cdot \|u(\tau)\|^2 d\tau. \end{aligned}$$

Since $\Phi(t_1, \cdot)B(\cdot)$ is a piecewise continuous function on $[t_0, t_1]$, there exists some $M > 0$ such that $\|\Phi(t_1, \tau)B(\tau)\| \leq M$ for all $\tau \in [t_0, t_1]$. Hence

$$\|\mathcal{L}_r(u)\|^2 \leq M^2 \int_{t_0}^{t_1} \|u(\tau)\|^2 d\tau = M^2 \|u(\cdot)\|_2^2$$

i.e. the induced norm of $\mathcal{L}_r$ is finite and the function is continuous. ∎

Definition 8.5 *The controllability gramian of the pair $(A(\cdot), B(\cdot))$ on $[t_0, t_1]$ is the matrix*

$$W_r(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_1, \tau) B(\tau) B(\tau)^T \Phi(t_1, \tau)^T d\tau \in \mathbb{R}^{n \times n}.$$

The controllability gramian will be our primary vehicle for studying the controllability of time varying linear systems. We start by establishing some of its basic properties.

Lemma 8.3 *The controllability gramian, $W_r(t_0, t_1)$, has the following properties:*

1. *It is symmetric.*
2. *It is positive semi-definite.*
3. *For all $t_0' \leq t_0$, $W_r(t_0', t_1) \geq W_r(t_0, t_1)$ in the sense that $x^T[W_r(t_0', t_1) - W_r(t_0, t_1)]x \geq 0$ for all $x \in \mathbb{R}^n$.*

Proof: For Part 1, note that

$$\begin{aligned} W_r(t_0, t_1)^T &= \left(\int_{t_0}^{t_1} \Phi(t_1, \tau) B(\tau) B(\tau)^T \Phi(t_1, \tau)^T d\tau \right)^T \\ &= \int_{t_0}^{t_1} \left(\Phi(t_1, \tau) B(\tau) B(\tau)^T \Phi(t_1, \tau)^T \right)^T d\tau \\ &= \int_{t_0}^{t_1} \left(\Phi(t_1, \tau)^T \right)^T \left(B(\tau)^T \right)^T B(\tau)^T \Phi(t_1, \tau)^T d\tau \\ &= W_r(t_0, t_1). \end{aligned}$$

For Part 2 note that for all $x \in \mathbb{R}^n$

$$x^T W_r(t_0, t_1) x = \int_{t_0}^{t_1} x^T \Phi(t_1, \tau) B(\tau) B(\tau)^T \Phi(t_1, \tau)^T x d\tau = \int_{t_0}^{t_1} \left\| B(\tau)^T \Phi(t_1, \tau)^T x \right\|^2 d\tau \geq 0.$$

Finally, for Part 3, note that

$$x^T [W_r(t_0', t_1) - W_r(t_0, t_1)] x = \int_{t_0'}^{t_0} x^T \Phi(t_1, \tau) B(\tau) B(\tau)^T \Phi(t_1, \tau)^T x d\tau \geq 0$$

(as shown in Part 2). ∎

The following theorem allows us to answer controllability questions by simply checking the rank of the controllability gramian.

Theorem 8.3 *The following statements are equivalent:*

1. *$(A(\cdot), B(\cdot))$ is controllable on $[t_0, t_1]$.*
2. *$\textsc{Range}(\mathcal{L}_r) = \mathbb{R}^n$.*
3. *$\textsc{Range}(\mathcal{L}_r \circ \mathcal{L}_r^*) = \mathbb{R}^n$.*
4. *$\textsc{Det}[W_r(t_0, t_1)] \neq 0$.*

Proof: 1 $\Leftrightarrow$ 2: By Lemma 8.1, $(A(\cdot), B(\cdot))$ is controllable on $[t_0, t_1]$ if and only if all states are reachable from 0 on $[t_0, t_1]$, i.e. if and only if $\mathcal{L}_r$ is surjective, i.e. if and only if $\textsc{Range}(\mathcal{L}_r) = \mathbb{R}^n$.

2 $\Leftrightarrow$ 3: By the Finite Rank Lemma.

3 $\Leftrightarrow$ 4: Notice that $\mathcal{L}_r \circ \mathcal{L}_r^* : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map between two finite dimensional spaces. Therefore it admits a matrix representation. We will show that $W_r(t_0, t_1) \in \mathbb{R}^{n\times n}$ is the matrix representation of this linear map. Then $\mathcal{L}_r \circ \mathcal{L}_r^*$ is surjective if and only if $W_r(t_0, t_1)$ is invertible i.e. if and only if $\textsc{Det}[W_r(t_0, t_1)] \neq 0$.

Recall that

$$\mathcal{L}_r : u \mapsto \int_{t_0}^{t_1} \Phi(t_1, \tau) B(\tau) u(\tau) d\tau$$

and $\langle \mathcal{L}_r^* x, u\rangle = \langle x, \mathcal{L}_r u\rangle$. We have already seen that (Section 7.4)

$$[\mathcal{L}_r^*(x)](\tau) = B(\tau)^T \Phi(t_1, \tau)^T x, \text{ for all } \tau \in [t_0, t_1], x \in \mathbb{R}^n$$

Therefore

$$[\mathcal{L}_r \circ \mathcal{L}_r^*](x) = \left[\int_{t_0}^{t_1} \Phi(t_1, \tau) B(\tau) B(\tau)^T \Phi(t_1, \tau)^T d\tau\right] x = W_r(t_0, t_1) x.$$

Therefore the matrix representation of $\mathcal{L}_r \circ \mathcal{L}_r^* : \mathbb{R}^n \to \mathbb{R}^n$ is the matrix $W_r(t_0, t_1) \in \mathbb{R}^{n\times n}$. ∎

Exercise 8.3 Show that $(A(\cdot), B(\cdot))$ is controllable on $[t_0, t_1]$ if an only if $W_r(t_0, t_1)$ is positive definite (i.e. $x^T W_r(t_0, t_1) x > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$).

8.3 Linear time varying systems: Minimum energy control

We have just seen that for all $x_0, x_1 \in \mathbb{R}^n$ there exists $u : [t_0, t_1] \to \mathbb{R}^m$ that steers (x_0, t_0) to (x_1, t_1) if and only if the matrix $W_r(t_0, t_1) \in \mathbb{R}^{n\times n}$ is invertible. The question now becomes can we compute such a u? And if so, can we find a piecewise continuous one, as required for existence and uniqueness of solutions? Or even a continuous one?

First of all we note that if such an input exists it will not be unique. To start with let us restrict our attention to the case where $x_0 = 0$ (reachability); we have already seen this is equivalent to the general controllability case ($x_0 \neq 0$) to which we will return shortly. Recall that by the Finite Rank Lemma

$$L^2([t_0, t_1], \mathbb{R}^m) = \textsc{Range}(\mathcal{L}_r^*) \overset{\perp}{\oplus} \textsc{Null}(\mathcal{L}_r).$$

Recall also that $L^2([t_0, t_1], \mathbb{R}^m)$ is infinite dimensional while $\textsc{Range}(\mathcal{L}_r^*)$ is finite dimensional (of dimension at most n), since it is a subspace of $\mathbb{R}^n$ which is finite dimensional. Therefore $\textsc{Null}(\mathcal{L}_r) \neq \{0\}$, in fact it will be an infinite dimensional subspace of $L^2([t_0, t_1], \mathbb{R}^m)$.

For a given $x_1 \in \mathbb{R}^n$ consider $u \in L^2([t_0, t_1], \mathbb{R}^m)$ that steers $(0, t_0)$ to (x_1, t_1), i.e.

$$\mathcal{L}_r(u) = x_1.$$

Then for every $\hat{u} \in L^2([t_0, t_1], \mathbb{R}^m)$ with $\hat{u} \in \textsc{Null}(\mathcal{L}_r)$ we have

$$\mathcal{L}_r(u + \hat{u}) = \mathcal{L}_r(u) + \mathcal{L}_r(\hat{u}) = x_1 + 0 = x_1.$$

Therefore for any input u that steers $(0, t_0)$ to (x_1, t_1) any other input of the form $u + \hat{u}$ with $\hat{u} \in \textsc{Null}(\mathcal{L}_r)$ will do the same. So in general there will be an infinite number of inputs that get the job done. The question now becomes can we somehow find a "canonical" one? It turns out that this can be done by a projection onto $\textsc{Range}(\mathcal{L}_r^*)$.

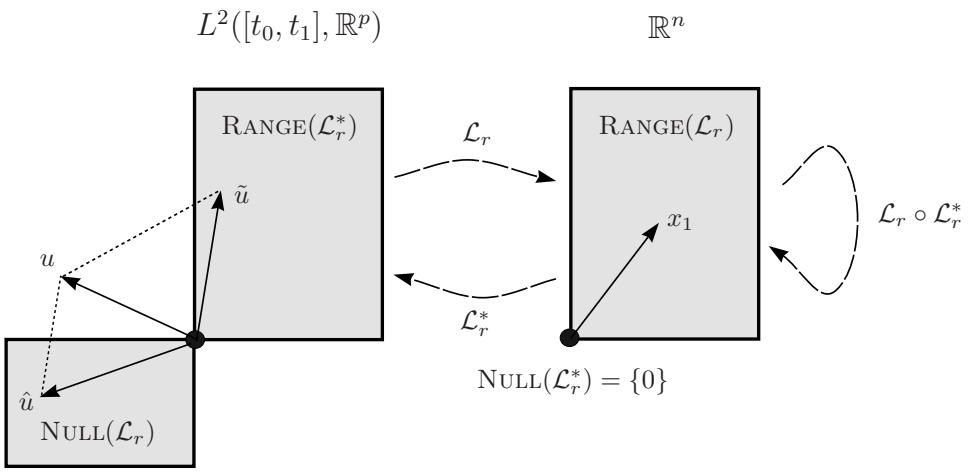

Figure 8.4: Linear space decomposition induced by reachability map.

Recall that by the Finite Rank Lemma

$$\mathcal{L}_r|_{\text{RANGE}(\mathcal{L}_r^*)} : \text{RANGE}(\mathcal{L}_r^*) \to \text{RANGE}(\mathcal{L}_r)$$

is a bijection. Therefore, for all $x_1 \in \text{RANGE}(\mathcal{L}_r)$ there exists a unique $\tilde{u} \in \text{RANGE}(\mathcal{L}_r^*)$ such that $\mathcal{L}_r(\tilde{u}) = x_1$. We have seen that for any $\hat{u} \in \text{NULL}(\mathcal{L}_r)$ if we let $u = \tilde{u} + \hat{u}$ then $\mathcal{L}_r(u) = x_1$. However,

$$\begin{aligned}\|u\|_2^2 &= \langle u, u\rangle_2 = \langle \tilde{u} + \hat{u}, \tilde{u} + \hat{u}\rangle_2 \\ &= \langle \tilde{u}, \tilde{u}\rangle_2 + \langle \hat{u}, \hat{u}\rangle_2 + \langle \tilde{u}, \hat{u}\rangle_2 + \langle \hat{u}, \tilde{u}\rangle_2.\end{aligned}$$

But $\tilde{u} \in \text{RANGE}(\mathcal{L}_r^*)$, $\hat{u} \in \text{NULL}(\mathcal{L}_r^*)$ and by the finite rank lemma $\text{RANGE}(\mathcal{L}_r^*)$ is orthogonal to $\text{NULL}(\mathcal{L}_r)$. Therefore $\langle \tilde{u}, \hat{u}\rangle_2 = \langle \hat{u}, \tilde{u}\rangle_2 = 0$ and

$$\|u\|_2^2 = \|\tilde{u}\|_2^2 + \|\hat{u}\|_2^2 \geq \|\tilde{u}\|_2^2$$

since $\|\hat{u}\|_2^2 \geq 0$. Therefore, among all the $u \in L^2([t_0,t_1],\mathbb{R}^m)$ that steer $(0,t_0)$ to (x_1,t_1), the unique $\tilde{u} \in \text{RANGE}(\mathcal{L}_r^*)$ is the one with the minimum 2−norm. The relation of the different spaces involved is depicted in Figure 8.4.

We now return to the general controllability case and formally state our findings.

Theorem 8.4 *Assume $(A(\cdot), B(\cdot))$ are controllable in $[t_0,t_1]$. Given $x_0, x_1 \in \mathbb{R}^n$ define $\tilde{u} : [t_0,t_1] \to \mathbb{R}^m$ by $\tilde{u} = \mathcal{L}_r^* \circ (\mathcal{L}_r \circ \mathcal{L}_r^*)^{-1}[x_1 - \Phi(t_1,t_0)x_0]$, i.e.*

$$\tilde{u}(t) = B(t)^T \Phi(t_1,t)^T W_r(t_0,t_1)^{-1}[x_1 - \Phi(t_1,t_0)x_0] \quad \forall t \in [t_0,t_1].$$

Then

1. *$\tilde{u}$ steers (x_0,t_0) to (x_1,t_1).*
2. *$\tilde{u}(\cdot) : [t_0,t_1] \to \mathbb{R}^m$ is piecewise-continuous, with discontinuity points the same as those of $B(\cdot) : [t_0,t_1] \to \mathbb{R}^{n\times m}$, In particular, $\tilde{u}(\cdot)$ is continuous if and only if $B(\cdot)$ is continuous.*
3. *$\|\tilde{u}\|_2^2 = [x_1 - \Phi(t_1,t_0)x_0]^T W_r(t_0,t_1)^{-1}[x_1 - \Phi(t_1,t_0)x_0]$.*
4. *If $u : [t_0,t_1] \to \mathbb{R}^m$ steers (x_0,t_0) to (x_1,t_1) then $\|u\|_2 \geq \|\tilde{u}\|_2$.*

Proof:

Part 1:

$$\begin{aligned}
x(t_1) &= \Phi(t_1,t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_1,t)B(t)u(t)dt \\
&= \Phi(t_1,t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_1,t)B(t)B(t)^T\Phi(t_1,t)^T W_r(t_0,t_1)^{-1}\left[x_1 - \Phi(t_1,t_0)x_0\right] dt \\
&= \Phi(t_1,t_0)x_0 + \left[\int_{t_0}^{t_1} \Phi(t_1,t)B(t)B(t)^T\Phi(t_1,t)^T dt\right] W_r(t_0,t_1)^{-1}\left[x_1 - \Phi(t_1,t_0)x_0\right] \\
&= \Phi(t_1,t_0)x_0 + W_r(t_0,t_1)W_r(t_0,t_1)^{-1}\left[x_1 - \Phi(t_1,t_0)x_0\right] \\
&= x_1.
\end{aligned}$$

Part 2 is obvious from the formula, since $B(t)$ is piecewise continuous, $\Phi(t_1,t)$ is continuous and all other quantities are constant. In fact the only discontinuity points are those of $B(t)$. Hence if $B(t)$ is continuous (e.g. for time invariant linear systems) then the minimum energy controls are continuous.

Part 3:

$$\begin{aligned}
\|\tilde{u}\|_2^2 &= \int_{t_0}^{t_1} u(t)^T u(t)dt \\
&= \left[x_1 - \Phi(t_1,t_0)x_0\right]^T \left(W_r(t_0,t_1)^{-1}\right)^T \left[\int_{t_0}^{t_1} \Phi(t_1,t)B(t)B(t)^T\Phi(t_1,t)^T dt\right] \\
&\quad W_r(t_0,t_1)^{-1}\left[x_1 - \Phi(t_1,t_0)x_0\right] \\
&= \left[x_1 - \Phi(t_1,t_0)x_0\right]^T \left(W_r(t_0,t_1)^T\right)^{-1} W_r(t_0,t_1)W_r(t_0,t_1)^{-1}\left[x_1 - \Phi(t_1,t_0)x_0\right] \\
&= \left[x_1 - \Phi(t_1,t_0)x_0\right]^T W_r(t_0,t_1)^{-1}\left[x_1 - \Phi(t_1,t_0)x_0\right]
\end{aligned}$$

since $W_r(t_0,t_1)$ is self-adjoint by the finite rank lemma.

Part 4: Notice that by its definition $\tilde{u} \in \textsc{Range}(\mathcal{L}_r^*)$. The claim follows by the discussion leading up to the statement of the theorem. ∎

The perceptive reader will have noticed that this calculation is effectively the same as the one carried out in Section 7.6 to compute the pseudo-inverse of a matrix. The only difference is that here one of the two spaces is infinite dimensional. Compare also Figure 8.4 with Figure **??**.

The theorem provides several interesting insights into the structure of these so-called *minimum energy controls*. For one, note that even though for mathematical reasons we were forced to allow more general control inputs $u \in L^2([t_0,t_1],\mathbb{R}^m)$ to apply the Finite Rank Lemma, the minimum energy controls (the unique $\tilde{u} \in \textsc{Range}(\mathcal{L}_r^*)$ that steers (x_0,t_0) to (x_1,t_1)) happens to be piecewise continuous. In other words, if there is a control in $u \in L^2([t_0,t_1],\mathbb{R}^m)$ that will "get the job done" there is also one in $\mathrm{PC}([t_0,t_1],\mathbb{R}^m)$ (in fact, at least as good a one) that will do the same. This is good to know, since we do not have to worry about existence of solutions.

The structure of the minimum energy itself, $\|\tilde{u}\|_2^2$, is also revealing. $\|\tilde{u}\|_2^2$ is a quadratic function of $x_1 - \Phi(t_1,t_0)x_0$, a term which reflects the distance between where we want the system to be at time t_1 (namely x_1) and where it would end up if we left it alone (namely $\Phi(t_1,t_0)x_0$). In other words, the further we want to push the system from its natural course the more energy we need to expend. The term in the middle, $W_r(t_0,t_1)^{-1}$, in a sense reflects how controllable the system is. By Theorem 8.3, if the system is uncontrollable $W_r(t_0,t_1)$ will be singular and, loosely speaking, $W_r(t_0,t_1)^{-1}$ would be infinite, suggesting that we may not be able to push the system where we want however hard we try. If the system is weakly controllable $W_r(t_0,t_1)$ will be positive definite but "small", hence $W_r(t_0,t_1)^{-1}$ will be "large". In other words, the less controllable a system is the more energy we need to expend to get it where we want. The observation that $W_r(t_0,t_1) \leq W_r(t_0',t_1)$ for $t_0' \leq t_0$ and

$t_1' \geq t_1$ (Fact 8.3) also admits a natural interpretation: The more time we have to take the system where we want the less energy we need to do so.

The proof of the theorem reveals that an even more general statement is possible. Even if the system is not controllable, if it happens that $x_1 - \Phi(t_1, t_0)x_0 \in \text{RANGE}(\mathcal{L}_r)$ then we can still find a minimum energy input that steers (x_0, t_0) to (x_1, t_1). The construction will be somewhat different (in this case $W_r(t_0, t_1)$ will not be invertible) but the general idea is the same: Consider the restricted map

$$\mathcal{L}_r|_{\text{RANGE}(\mathcal{L}_r^*)} : \text{RANGE}(\mathcal{L}_r^*) \to \text{RANGE}(\mathcal{L}_r)$$

which is bijective even if $\text{RANGE}(\mathcal{L}_r) \neq \mathbb{R}^n$ and select the unique $\tilde{u} \in \text{RANGE}(\mathcal{L}_r^*)$ for which $\mathcal{L}_r(\tilde{u}) = x_1 - \Phi(t_1, t_0)x_0$. Indeed, even if $x_1 - \Phi(t_1, t_0)x_0 \notin \text{RANGE}(\mathcal{L}_r)$ we can still formulate a more general problem where the objective is to minimize $\|x_1 - x(t_1)\|$ using a minimum energy input. The construction will be even more complicated in this case, but the idea is again similar to the pseudo-inverse calculations: Project $x_1 - \Phi(t_1, t_0)x_0$ onto $\text{RANGE}(\mathcal{L}_r)$ and among the u that map to the projection select the unique one that belongs to $\text{RANGE}(\mathcal{L}_r^*)$.

Finally, we note that Theorem 8.4 provides a simple example of an optimal control problem. The $\tilde{u}$ of the theorem is the input that drives the system from (x_0, t_0) to (x_1, t_1) and minimizes

$$\|u\|_2^2 = \int_{t_0}^{t_1} u(t)^T u(t) dt.$$

This is the starting point for more general optimal control problems where costs of the form

$$\int_{t_0}^{t_1} \left(x^T(t)Q(t)x(t) + u(t)^T R(t)u(t) \right) dt$$

for some $Q(t) \in \mathbb{R}^{n \times n}$ and $R(t) \in \mathbb{R}^{m \times m}$ symmetric and positive definite are minimized, subject to the constraints imposed by the dynamics of the system and the desired initial and final states. An example are the so-called Linear Quadratic Regulator (LQR) problems. The interested reader is referred to [4, 13] for more information on this topic.

8.4 Linear time varying systems: Observability and duality

We now turn to observability of time varying linear systems. As will soon become apparent, observability depends only on the properties of the matrices $A(\cdot)$ and $C(\cdot)$; $B(\cdot)$ and $D(\cdot)$ play no role. We therefore start by specializing Definition 8.2 to the linear time varying context.

Definition 8.6 *The pair of matrices $(C(\cdot), A(\cdot))$ is called observable on $[t_0, t_1]$ if and only if for all $x_0 \in \mathbb{R}^m$ and all $u : [t_0, t_1] \to \mathbb{R}^m$, one can uniquely determine x_0 from the information $\{(u(t), y(t)) \mid t \in [t_0, t_1]\}$.*

Note that once we know x_0 and $u : [t_0, t] \to \mathbb{R}^m$ we can reconstruct $x(t)$ for all $t \in [t_0, t_1]$ by

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau.$$

Moreover, since for all $t \in [t_0, t_1]$

$$y(t) = C(t)\Phi(t, t_0)x_0 + C(t) \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau + D(t)u(t),$$

the last two terms in the summation can be reconstructed if we know $u : [t_0, t_1] \to \mathbb{R}^m$. Therefore the difficult part in establishing observability is to determine x_0 from the zero input response $C(t)\Phi(t, t_0)x_0$ with $t \in [t_0, t_1]$. Without loss of generality we will therefore restrict our attention to the case $u(t) = 0$ for all $t \in [t_0, t_1]$.

Definition 8.7 *A state $x_0 \in \mathbb{R}^n$ is called unobservable on $[t_0, t_1]$ if and only if $C(t)\Phi(t, t_0)x_0 = 0$ for all $t \in [t_0, t_1]$.*

Clearly the state $x_0 = 0$ is unobservable. The question then becomes whether there are any $x_0 \neq 0$ that are also unobservable. This would be bad, since the zero input response from these states would be indistinguishable from that of the 0 state.

Definition 8.8 *The observability map of the pair $(C(\cdot), A(\cdot))$ is the function*

$$\begin{aligned} \mathcal{L}_o : \mathbb{R}^n &\longrightarrow L_2([t_0, t_1], \mathbb{R}^p) \\ x_0 &\longmapsto C(t)\Phi(t, t_0)x_0 \quad \forall t \in [t_0, t_1]. \end{aligned}$$

The following fact is a direct consequence of the definition.

Lemma 8.4 *$\mathcal{L}_o$ is a linear, continuous function of x_0. Moreover, for all $x_0 \in \mathbb{R}^n$, $\mathcal{L}_o(x_0) \in PC([t_0, t_1], \mathbb{R}^p)$. The state $x_0 \in \mathbb{R}^n$ is unobservable if and only if $x_0 \in \text{NULL}(\mathcal{L}_o)$. The pair of matrices $(C(\cdot), A(\cdot))$ is observable if and only if $\text{NULL}(\mathcal{L}_o) = \{0\}$.*

The proof is left as an exercise. Notice that to be able to apply the Finite Rank Lemma we are forced to consider functions $y(\cdot) : [t_0, t_1] \to \mathbb{R}^p$ that belong to the space $L_2([t_0, t_1], \mathbb{R}^p)$ in Definition 8.8. Lemma 8.4, however, shows that the only functions we will encounter will live in $\text{RANGE}(\mathcal{L}_o)$ and will be piecewise continuous with discontinuity points the same as those of $C(\cdot) : [t_0, t_1] \to \mathbb{R}^{p \times n}$.

Definition 8.9 *The observability gramian of the pair $(C(\cdot), A(\cdot))$ is the matrix*

$$W_o(t_0, t_1) = \int_{t_0}^{t_1} \Phi(\tau, t_0)^T C(\tau)^T C(\tau)\Phi(\tau, t_0) d\tau \in \mathbb{R}^{n \times n}.$$

The following theorem provides a complete characterization of observability of linear time varying systems in terms of the properties of the matrix $W_o(t_0, t_1)$.

Theorem 8.5 *The following statements are equivalent:*

1. *The pair of matrices $(C(\cdot), A(\cdot))$ is observable on $[t_0, t_1]$.*
2. $\text{NULL}(\mathcal{L}_o) = \{0\}$.
3. $\text{NULL}(\mathcal{L}_o^* \circ \mathcal{L}_o) = \{0\}$.
4. $\text{DET}[W_o(t_0, t_1)] \neq 0$.

Proof: The proof is effectively the same as that of Theorem 8.3 with $\mathcal{L}_o$ in place of $\mathcal{L}_r^*$. Indeed, $W_o(t_0, t_1)$ turns out to be the representation of the linear map $\mathcal{L}_o^* \circ \mathcal{L}_o : \mathbb{R}^n \to \mathbb{R}^n$ in the basis used for the representation of the matrix $A(\cdot)$. ∎

The similarity between the controllability and observability theorems is further highlighted by the decomposition into subspaces induced by $\mathcal{L}_o$, as shown in Figure 8.5. Comparing Figure 8.5 to Figure 8.4, it becomes apparent that controllability and observability are in fact dual concepts, they are just the two faces of the Finite Rank Lemma. To formalize this statement, note that given the system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{8.6}$$
$$y(t) = C(t)x(t) + D(t)u(t) \tag{8.7}$$

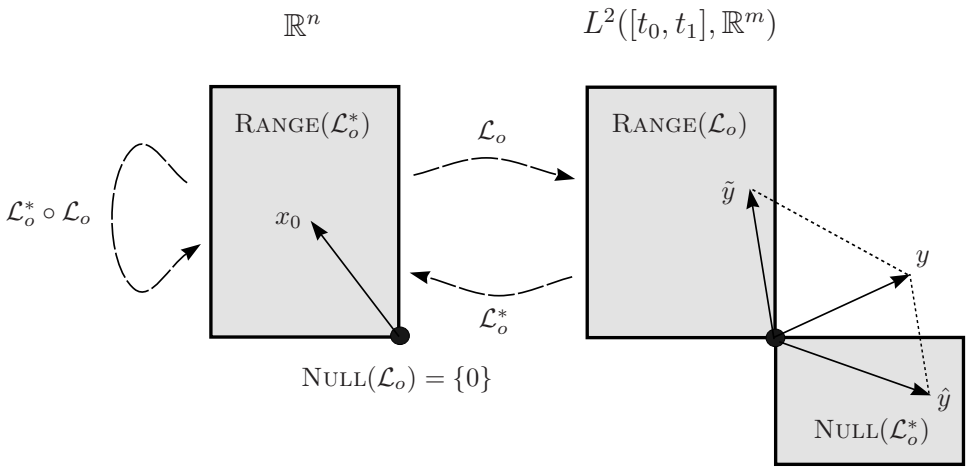

Figure 8.5: Linear space decomposition induced by observability map.

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, one can define the *dual system*

$$\dot{\bar{x}}(t) = -A(t)^T \bar{x}(t) - C(t)^T \bar{u}(t) \tag{8.8}$$
$$\bar{y}(t) = B(t)^T \bar{x}(t) + D(t)^T \bar{u}(t) \tag{8.9}$$

with $\bar{x}(t) \in \mathbb{R}^n$, $\bar{u}(t) \in \mathbb{R}^p$ and $\bar{y}(t) \in \mathbb{R}^m$. The two systems are closely related in terms of their controllability and observability properties.

Theorem 8.6 (Duality theorem) *Let $\Phi(t,t_0), \Psi(t,t_0) \in \mathbb{R}^{n\times n}$ denote the state transition matrices of system (8.6)–(8.7) and system (8.8)–(8.9) respectively.*

1. $\Psi(t,t_0) = \Phi(t_0,t)^T$
2. *System system (8.6)–(8.7) is controllable on $[t_0,t_1]$ if and only if system (8.8)–(8.9) is observable on $[t_0,t_1]$.*
3. *System system (8.6)–(8.7) is observable on $[t_0,t_1]$ if and only if system (8.8)–(8.9) is controllable on $[t_0,t_1]$.*

Proof: For Part 1, recall that by Theorem 4.2, $\Psi(t_0,t_0) = \Phi(t_0,t_0)^T = I$ and $\frac{\partial}{\partial t}\Psi(t,t_0) = -A(t)^T\Psi(t,t_0)$. Moreover

$$\begin{aligned}\Phi(t_0,t)\Phi(t,t_0) = \Phi(t_0,t_0) &\Rightarrow \frac{\partial}{\partial t}\left[\Phi(t_0,t)\Phi(t,t_0)\right] = \frac{\partial}{\partial t}\left[\Phi(t_0,t)\right]\Phi(t,t_0) + \Phi(t_0,t)\frac{\partial}{\partial t}\Phi(t,t_0) = 0\\ &\Rightarrow \frac{\partial}{\partial t}\left[\Phi(t_0,t)\right]\Phi(t,t_0) = -\Phi(t_0,t)A(t)\Phi(t,t_0)\\ &\Rightarrow \frac{\partial}{\partial t}\Phi(t_0,t)^T = -A(t)^T\Phi(t_0,t)^T.\end{aligned}$$

The conclusion follows by existence and uniqueness.

For Part 2, by Theorem 8.5 system (8.8)–(8.9) is observable on $[t_0,t_1]$ if and only if the matrix $\int_{t_0}^{t_1}\Psi(\tau,t_0)^T\left(B(\tau)^T\right)^T B(\tau)^T\Psi(\tau,t_0)d\tau$ is invertible. Substituting $\Psi(t,t_0)$ from Part 1 this matrix is the same as

$$\int_{t_0}^{t_1}\Phi(t_0,\tau)B(\tau)B(\tau)^T\Phi(t_0,\tau)^T d\tau = \Phi(t_0,t_1))\int_{t_0}^{t_1}\Phi(t_1,\tau)B(\tau)B(\tau)^T\Phi(t_1,\tau)^T d\tau\Phi(t_0,t_1)^T.$$

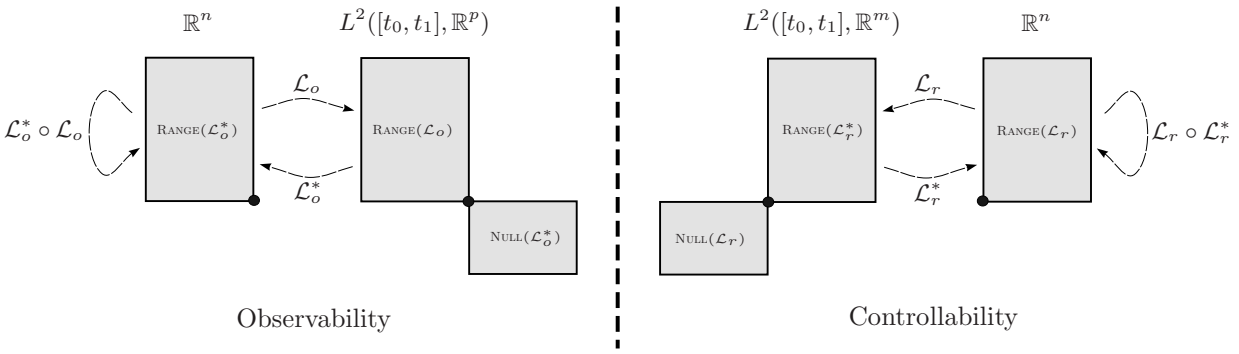

Figure 8.6: duality.

Since $\Phi(t_0,t_1)$ is invertible by Theorem 4.2, the above is invertible if and only if

$$\int_{t_0}^{t_1} \Phi(t_1,\tau)B(\tau)B(\tau)^T\Phi(t_1,\tau)^T d\tau$$

is invertible, i.e. if and only if system (8.6)–(8.7) is controllable (by Theorem 8.3).

The proof of Part 3 is similar to that of Part 2 and is left as an exercise. ■

Duality is illustrated by the reflection symmetry of Figure 8.6. It allows us to immediately extend results proved for controllability to related observability results and vice versa. For example, the following dual statement of the minimum energy control theorem allows us to estimate the value of the initial condition x_0 that gave rise to a particular zero input response trajectory $y(\cdot) : [t_0,t_1] \to \mathbb{R}^p$.

Theorem 8.7 *Assume that the pair of matrices $(C(\cdot),A(\cdot))$ is observable and consider an arbitrary $y \in L^2([t_0,t_1],\mathbb{R}^p)$. Then*

$$x_0 = (\mathcal{L}_o^* \circ \mathcal{L}_o)^{-1}\mathcal{L}_o^*(y) = [W_o(t_0,t_1)]^{-1}\int_{t_0}^{t_1} \Phi(\tau,t_0)^T C(\tau)^T y(\tau)d\tau$$

is the unique minimizer of $\|y - \mathcal{L}_o(x)\|_2$ over $x \in \mathbb{R}^n$. Moreover,

$$\min_{x\in\mathbb{R}^n} \|y - \mathcal{L}_o(x)\|_2^2 = \|y\|_2^2 - x_0^T W_o(t_0,t_1)x_0$$

Proof: By the Finite Rank Lemma, y uniquely decomposes into a sum $y = \tilde{y}+\hat{y}$ with $\tilde{y} \in$ RANGE$(\mathcal{L}_o)$ and $\hat{y} \in$ NULL$(\mathcal{L}_o^*)$ (see Figure 8.5). For all $x \in \mathbb{R}^n$, we clearly have $\mathcal{L}_o(x) \in$ RANGE$(\mathcal{L}_o)$ and hence $\tilde{y} - \mathcal{L}_o(x) \in$ RANGE$(\mathcal{L}_o)$ (recall that RANGE$(\mathcal{L}_o)$ is a subspace of $L^2([t_0,t_1],\mathbb{R}^p)$). Since RANGE$(\mathcal{L}_o)$ and NULL$(\mathcal{L}_o^*)$ are orthogonal

$$\|y - \mathcal{L}_o(x)\|_2^2 = \|\tilde{y} - \mathcal{L}_o(x)\|_2^2 + \|\hat{y}\|_2^2 \geq \|\hat{y}\|_2^2. \tag{8.10}$$

Therefore $\min_{x\in\mathbb{R}^n} \|y - \mathcal{L}_o(x)\|_2$ is achieved for all $x \in \mathbb{R}^n$ such that $\mathcal{L}_o(x) = \tilde{y}$.

Because the system is observable NULL$(\mathcal{L}_o) = \{0\}$ and the function $\mathcal{L}_o^* \circ \mathcal{L}_o : \mathbb{R}^n \to \mathbb{R}^n$ is invertible (after all $W_o(t_0,t_1)$ is the representation of this map and DET$[W_o(t_0,t_1)] \neq 0$). Moreover, RANGE$(\mathcal{L}_o^*) = \mathbb{R}^n$) and by the Finite Rank Lemma, the function

$$\mathcal{L}_o(x) : \mathbb{R}^n \to \text{RANGE}(\mathcal{L}_o)$$

is bijective . Therefore, there exists a unique $x_0 \in \mathbb{R}^n$ such that $\mathcal{L}_o(x_0) = \tilde{y}$. Note that

$$\begin{aligned}(\mathcal{L}_o^* \circ \mathcal{L}_o)^{-1}\mathcal{L}_o^*(y) &= (\mathcal{L}_o^* \circ \mathcal{L}_o)^{-1}\mathcal{L}_o^*(\tilde{y} + \hat{y}) \\ &= (\mathcal{L}_o^* \circ \mathcal{L}_o)^{-1}\mathcal{L}_o^*(\tilde{y}) \quad (\text{since } \hat{y} \in \textsc{Null}(\mathcal{L}_o^*)) \\ &= (\mathcal{L}_o^* \circ \mathcal{L}_o)^{-1}\mathcal{L}_o^* \circ \mathcal{L}_o(x_0) \\ &= x_0.\end{aligned}$$

Writing the above in the coordinates used in the representation of $A(\cdot)$ and recalling the infinite dimensional adjoint calculation in Section 7.4 shows that

$$x_0 = [W_o(t_0, t_1)]^{-1} \int_{t_0}^{t_1} \Phi(\tau, t_0)^T C(\tau)^T y(\tau) d\tau$$

is the unique minimizer of $\|y - \mathcal{L}_o(x)\|_2$

To compute the minimum, note that $\tilde{y}(t) = [\mathcal{L}_o(x_0)](t) = C(t)\Phi(t, t_0)x_0$. Hence,

$$\|\tilde{y}\|_2^2 = \int_{t_0}^{t_1} \tilde{y}(t)^T \tilde{y}(t) dt = x_0^T \int_{t_0}^{t_1} \Phi(t, t_0)^T C(t)^T C(t)\Phi(t, t_0) dt x_0 = x_0^T W_o(t_0, t_1) x_0.$$

The value of the minimum follows by substituting into (8.10). ∎

Notice that the theorem does not require the observed output, $y(\cdot)$, to lie in $\textsc{Range}(\mathcal{L}_o)$, i.e. to correspond exactly to the zero input response of some initial state $x_0 \in \mathbb{R}^n$. In fact, the observed $y(\cdot)$ does even have to be piecewise continuous, as one would expect from a zero input response. The formula provided in the theorem generates the x_0 whose zero input responce most closely matches the observed $y(\cdot)$ in the 2-norm sense. This is very useful, since in practice the measured system response will be corrupted by noise and other disturbances, hence is unlikely to correspond exactly to some initial state.

As in the case of controllability, the above calculation can also be extended to the case where the system is not observable. The minimizer will not be unique in this case and the computation will be more involved since $W_0(t_0, t_1)$ will not be invertible, but the idea is similar: Find the projection of the measured output y onto $\textsc{Range}(\mathcal{L}_o)$ and the minimum norm x_0 that corresponds to this projection.

8.5 Linear time invariant systems: Observability

Next, we restrict our attention further to the special case

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t).\end{aligned}$$

Define the *observability matrix* by

$$O = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{np \times n}.$$

The following fact summaries some basic properties of the observability matrix.

Lemma 8.5 *The observability matrix O has the following properties:*

1. $\textsc{Null}(O) = \textsc{Null}(\mathcal{L}_o) = \{x_0 \in \mathbb{R}^n \mid \textit{unobservable}\}$.

2. $\textsc{Null}(O)$ *is an* A *invariant subspace, i.e. if* $x \in \textsc{Null}(O)$ *then* $Ax \in \textsc{Null}(O)$.

Proof: Part 1: x_0 is unobservable if and only if $x_0 \in \textsc{Null}(\mathcal{L}_o)$, i.e. if and only if

$$\begin{aligned} C\Phi(t,0)x_0 = 0 \ \ \forall t \in [0,t_1] &\Leftrightarrow Ce^{At}x_0 = 0 \ \ \forall t \in [0,t_1] \\ &\Leftrightarrow C\left(I + At + \frac{A^2t^2}{2} + \dots\right)x_0 = 0 \ \ \forall t \in [0,t_1] \\ &\Leftrightarrow CA^k x_0 = 0 \ \forall k \in \mathbb{N} \ \ (t^k \text{ linearly independent by Fact 2.6}) \\ &\Leftrightarrow CA^k x_0 = 0 \ \forall k = 0,\dots,n-1 \ \ \text{(by Cayley-Hamilton theorem)} \\ &\Leftrightarrow \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x_0 = 0 \\ &\Leftrightarrow x_0 \in \textsc{Null}(O). \end{aligned}$$

Part 2: Consider $x_0 \in \textsc{Null}(O)$, i.e. $CA^k x_0 = 0$ for all $k = 0,\dots,n-1$. We would like to show that $x = Ax_0 \in \textsc{Null}(O)$. Indeed $Cx = CAx_0 = 0$, $CAx = CA^2x_0 = 0$, …, until

$$CA^{n-1}x = CA^n x_0 = -C(\chi_n I + \chi_{n-1}A + \dots + \chi_1 A^{n-1})x_0 = 0$$

(where we make use of the Cayley Hamilton theorem). ∎

We are now in a position to establish conditions that allow us to easily determine whether a linear time invariant system is observable or not.

Theorem 8.8 *For any* $[t_0,t_1]$ *the following statements are equivalent:*

1. *The pair of matrices* (C,A) *is observable on* $[t_0,t_1]$.
2. $\textsc{Rank}(O) = n$.
3. *For all* $\lambda \in Co$
$$\textsc{Rank}\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n.$$

Proof: Consider any $[t_0,t_1]$ and without loss of generality assume that $t_0 = 0$.

To show 1 is equivalent to 2, note that

$$\begin{aligned} (C,A) \text{ observable on } [t_0,t_1] &\Leftrightarrow \textsc{Null}(\mathcal{L}_o) = \{0\} \\ &\Leftrightarrow \textsc{Null}(O) = \{0\} \\ &\Leftrightarrow \textsc{Dim}(\textsc{Null}(O)) = \textsc{Nullity}(O) = 0 \\ &\Leftrightarrow \textsc{Rank}(O) = n - \textsc{Nullity}(O) = n \ \ (\text{ by Theorem 2.2}). \end{aligned}$$

We show that 1 implies 3 by contraposition. Note that since $A \in \mathbb{R}^{n\times n}$

$$\textsc{Rank}\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} \leq n.$$

Assume that there exists $\lambda \in \mathbb{C}$ such that

$$\textsc{Rank}\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} < n.$$

Then the columns of the matrix are linearly dependent and there exists $v \in \mathbb{C}^n$ with $v \neq 0$ such that

$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} v = \begin{bmatrix} (\lambda I - A)v \\ Cv \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Therefore, λ must be an eigenvalue of A with v the corresponding eigenvector and $Cv = 0$. Recall that if $v \in \mathbb{C}^n$ is an eigenvector of A then

$$s(t, 0, v, \theta_U) = e^{At}\hat{e} = e^{\lambda t}v$$

(see, for example, the proof of Theorem 6.2). Therefore

$$Ce^{At}\hat{e} = e^{\lambda t}Cv = 0, \ \ \forall t \geq 0$$

and $v \neq 0$ is unobservable. Hence the pair of matrices (C, A) is unobservable[1]

Finally, we show that 3 implies 1 also by contraposition. Assume that C, A is unobservable. Then $\textsc{Null}(O) \neq \{0\}$ and $\textsc{Dim}(\textsc{Null}(O)) = r > 0$ for some $0 < r \leq n$. Choose a basis for $\textsc{Null}(O)$, say $\{v_1, \ldots, v_r\}$. Choose also a basis for $\textsc{Null}(O)^\perp$, say $\{w_1, \ldots, w_{n-r}\}$. Since $\textsc{Null}(O)$ is finite dimensional (hence closed) by Theorem 7.3,

$$\mathbb{R}^n = \textsc{Null}(O) \overset{\perp}{\oplus} \textsc{Null}(O)^\perp$$

and $\{w_1, \ldots, w_{n-r}, v_1, \ldots, v_r\}$ form a basis for $\mathbb{R}^n$. The representation of $x \in \mathbb{R}^n$ with respect to this basis decomposes into two parts, $x = (x_1, x_2)$, with $x_1 \in \mathbb{R}^{n-r}$ ("observable") and $x_2 \in \mathbb{R}^r$ ("unobservable"). For the rest of this proof we assume that all vectors are represented with respect to this basis.

Recall that, from Part 2, $\textsc{Null}(O)$ is an $A-$invariant subspace, i.e. for all $x \in \textsc{Null}(O)$, $Ax \in \textsc{Null}(O)$. Note that

$$x \in \textsc{Null}(O) \Rightarrow x = \begin{bmatrix} 0 \\ x_2 \end{bmatrix}.$$

Moreover,

$$Ax = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} 0 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_{12}x_2 \\ A_{22}x_2 \end{bmatrix}$$

Therefore, since $Ax \in \textsc{Null}(O)$ we must have that $A_{12}x_2 = 0$ for all $x_2 \in \mathbb{R}^r$, which implies that $A_{12} = 0 \in \mathbb{R}^{(n-r)\times r}$ (think of using the elements of the canonical basis of $\mathbb{R}^r$ as x_2 one after the other). Notice further that if $x \in \textsc{Null}(O)$ then in particular $x \in \textsc{Null}(C)$ and

$$Cx = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} 0 \\ x_2 \end{bmatrix} = C_2x_2 = 0 \ \ \forall x_2 \in \mathbb{R}^r.$$

Hence $C_2 = 0 \in \mathbb{R}^{p\times r}$.

In summary, in the new coordinates the system representation is

$$\begin{array}{rcl} \dot{x}_1(t) &=& A_{11}x_1(t) \qquad\qquad +B_1u(t) \\ \dot{x}_2(t) &=& A_{21}x_1(t) \ +A_{22}x_2(t) \ +B_2u(t) \\ y(t) &=& C_1x_1(t) \qquad\qquad +Du(t). \end{array}$$

Take $\lambda \in \mathbb{C}$ an eigenvalue of $A_{22} \in \mathbb{R}^{r\times r}$ and let $v \in \mathbb{C}^r$ with $v \neq 0$ be the corresponding eigenvector. Then

$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} \begin{bmatrix} 0 \\ v \end{bmatrix} = \begin{bmatrix} \lambda I - A_{11} & 0 \\ -A_{21} & \lambda I - A_{22} \\ C_1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda(I - A_{22})v \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

[1] This argument is fine as long as λ is real. If λ is complex, strictly speaking, we have not constructed a real vector $x_0 \in \mathbb{R}^n$ which is unobservable. This can be done, however, by taking linear combinations of the complex eigenvectors, as in the proof of Theorem 6.2. The construction is left as an exercise.

(We note in passing that the above shows that λ will also be an eigenvalue of $A \in \mathbb{R}^{n\times n}$ with eigenvector $(0, v)$). Hence the columns of the matrix are linearly dependent and

$$\textsc{Rank} \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} < n.$$

■

The theorem provides two easy ways of checking the observability of a time invariant linear system, by checking the rank of matrices. Notice that the conditions of the theorem are independent of $[t_0, t_1]$. Therefore, (C, A) is observable on some $[t_0, t_1]$ with $t_0 < t_1$ if and only if it is observable on all $[t_0, t_1]$. To put it another way, observations $\{(u(t), y(t) \mid t \in [0, \epsilon]\}$ over an arbitrary small interval are sufficient to reconstruct the initial condition x_0 and hence all future states of the system (at least in theory).

The most commonly used test for observability involves testing the rank of the matrix O. Though easy to implement for small systems, this test may lead to numerical conditioning problems for larger systems, since it requires the computation of powers of A up to A^{n-1}. Testing the rank of

$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$$

tends to be more robust numerically on the other hand. Notice that even though the theorem suggests that this condition needs to be tested for all $\lambda \in \mathbb{C}$ (clearly an impossible task) the proof indicates that the condition need only be checked for $\lambda \in \textsc{Spec}[A]$.

The decomposition of the state developed in the proof of Part 4 also shows that for any (C, A), the state of the system decomposes into an observable and and unobservable part. An immediate corollary of the proof of the theorem is that the observable part is indeed observable.

Corollary 8.1 *For any matrices $C \in \mathbb{R}^{p\times n}$ and $A \in \mathbb{R}^{n\times n}$ there exists a change of basis $T \in \mathbb{R}^{n\times n}$ with $\textsc{Det}[T] \neq 0$, such that in the new coordinates the representation of the matrices decomposes into*

$$\hat{A} = TAT^{-1} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \quad \text{and} \quad \hat{C} = CT^{-1} = \begin{bmatrix} C_1 & 0 \end{bmatrix}$$

and the pair of matrices (C_1, A_{11}) is observable.

Proof: The decomposition was already established in the proof of Theorem 8.8. To show that (C_1, A_{11}) are observable assume, for the sake of contradiction, that they are not. Then there exists $x_1 \in \mathbb{R}^{(n-r)\times r}$ with $x_1 \neq 0$ such that

$$\begin{bmatrix} C_1 \\ C_1A_{11} \\ \vdots \\ C_1A_{11}^{n-r-1} \end{bmatrix} x_1 = 0.$$

The structure of the matrices implies that

$$O \begin{bmatrix} x_1 \\ 0 \end{bmatrix} = \begin{bmatrix} C_1 & 0 \\ C_1A_{11} & 0 \\ \vdots & \vdots \\ C_1A_{11}^{n-r-1} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ 0 \end{bmatrix} = 0.$$

Hence $(x_1, 0) \in \textsc{Null}(O)$. By the choice of basis, however, $(x_1, 0) \in \textsc{Span}\{w_1, \ldots, w_{n-r}\} = \textsc{Null}(O)^{\perp}$. Since $(x_1, 0) \neq 0$ this is a contradiction (by Fact 7.4). ■

Note also that because

$$\textsc{Det}\begin{bmatrix} \lambda I_1 - A_{11} & 0 \\ -A_{21} & \lambda I_2 - A_{22} \end{bmatrix} = \textsc{Det}\left[\lambda I_1 - A_{11}\right]\textsc{Det}\left[\lambda I_2 - A_{22}\right]$$

(where I_1 and I_2 are identity matrices of appropriate dimensions) the spectrum of A decomposes into

$$\textsc{Spec}[A] = \textsc{Spec}[A_{11}] \cup \textsc{Spec}[A_{22}]$$

where the former set contains eigenvalues whose eigenvectors are observable (known as *observable modes*) while the latter contains eigenvalues whose eigenvectors are unobservable (known as *unobservable modes*). An immediate danger with unobservable systems is that the state may be diverging to infinity without any indication of this at the output. Indeed, if the system is unobservable and one of the eigenvalues of the matrix A_{22} above has positive real part the state $x_2(t)$ may go to infinity as t increases. Since, however, $(0, x_2)$ is in the nullspace of O, the output $y(t)$ will be unaffected by this (either directly or through the state x_1) and may remain bounded. For unobservable systems one would at least hope that the unobservable modes (the eigenvalues of the matrix A_{22}) are stable. In this case, if the output of the system is bounded one can be sure that the state is also bounded. This requirement, which is weaker than observability, is known as *detectability*.

8.6 Linear time invariant systems: Controllability

As expected, the picture for controllability of time invariant linear systems is dual to that of observability. One can define the *controllability matrix* by

$$P = \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n \times nm}$$

and use it to test whether the systems is controllable. The main facts in this direction are summarized in the following theorem.

Theorem 8.9 *For any $[t_0, t_1]$ the following hold:*

1. $\textsc{Range}(P) = \textsc{Range}(\mathcal{L}_r) = \{x \in \mathbb{R}^n \mid \textit{reachable}\}$.
2. *The pair of matrices (A, B) is controllable on $[t_0, t_1]$ if and only if* $\textsc{Rank}(P) = n$.
3. $\textsc{Range}(P)$ *is an A invariant subspace.*
4. *The pair of matrices (A, B) is controllable on $[t_0, t_1]$ if and only if for all $\lambda \in \mathbb{C}$*

$$\textsc{Rank}\begin{bmatrix} \lambda I - A & B \end{bmatrix} = n.$$

The proof is effectively the same as that of Theorem 8.8 by duality. As before, the condition of Part 4 need only be tested for $\lambda \in \textsc{Spec}[A]$. Moreover, for all (A, B) the state decomposes into a controllable and an uncontrollable part. For an appropriate choice of basis the system representation becomes:

$$\begin{array}{rcccc} \dot{x}_1(t) & = & A_{11}x_1(t) & +A_{12}x_2(t) & +B_1u(t) \\ \dot{x}_2(t) & = & & +A_{22}x_2(t) & \\ y(t) & = & C_1x_1(t) & C_2x_2(t) & +Du(t) \end{array}$$

where the pair of matrices (A_{11}, B_1) is controllable. Likewise, the spectrum of A decomposes into

$$\textsc{Spec}[A] = \textsc{Spec}[A_{11}] \cup \textsc{Spec}[A_{22}]$$

where the former set contains eigenvalues whose eigenvectors are reachable (*controllable modes*) while the latter contains eigenvalues whose eigenvectors are unreachable (*uncontrollable modes*).

Dually to observability, an immediate danger with uncontrollable systems is that the state may be diverging to infinity without the input being able to prevent it. If the system is uncontrollable and one of the eigenvalues of the matrix A_{22} above has positive real part the state $x_2(t)$ may go to infinity as t increases. Since the evolution of this state is not affected by the input (neither directly nor through the state x_1) there is nothing the input can do to prevent this. For uncontrollable systems one would at least hope that the uncontrollable modes (the eigenvalues of the matrix A_{22}) are stable. This requirement, which is weaker than controllability, is known as *stabilizability*.

8.7 Kalman decomposition

Applying the two theorems one after the other leads to the Kalman decomposition theorem.

Theorem 8.10 (Kalman decomposition) *For an appropriate change of basis for $\mathbb{R}^n$ the state vector is partitioned into*

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \begin{matrix} \textit{controllable, observable} \\ \textit{controllable, unobservable} \\ \textit{uncontrollable, observable} \\ \textit{uncontrollable, unobservable} \end{matrix}.$$

In these coordinates the system matrices get partitioned into

$$A = \begin{bmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & A_{43} & A_{44} \end{bmatrix} \quad B = \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix}$$
$$C = \begin{bmatrix} C_1 & 0 & C_3 & 0 \end{bmatrix} \quad D$$

with

$$\left(\begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \right) \textit{ controllable and } \left(\begin{bmatrix} C_1 & C_3 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{bmatrix} \right) \textit{ observable.}$$

The theorem shows that the spectrum of A gets partitioned into

$$\textsc{Spec}[A] = \textsc{Spec}[A_{11}] \cup \textsc{Spec}[A_{22}] \cup \textsc{Spec}[A_{33}] \cup \textsc{Spec}[A_{44}]$$

where the first set contains controllable and observable modes, the second set controllable and unobservable modes, the third set uncontrollable and observable modes and the fourth set uncontrollable and unobservable modes.

Pictotially the interdependencies of the sub-systems revealed by the Kalman decomposition are shown in Figure 8.7. Note that this figure is not a block diagram, the arrows represent dependencies between subsystems rather than signals. The labels next to the arrows indicate which of the matrices in Theorem 8.10 is responsible for the corresponding dependence.

Problems for chapter 8

Problem 8.1 (Controllability Gramian) Let

$$W_r(t_1, t_0) = \int_{t_0}^{t_1} \phi(t_1, \tau) B(\tau) B(\tau)^T \phi(t_1, \tau)^T \mathrm{d}\tau$$

denote the controllability Gramian of the linear time varying system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$

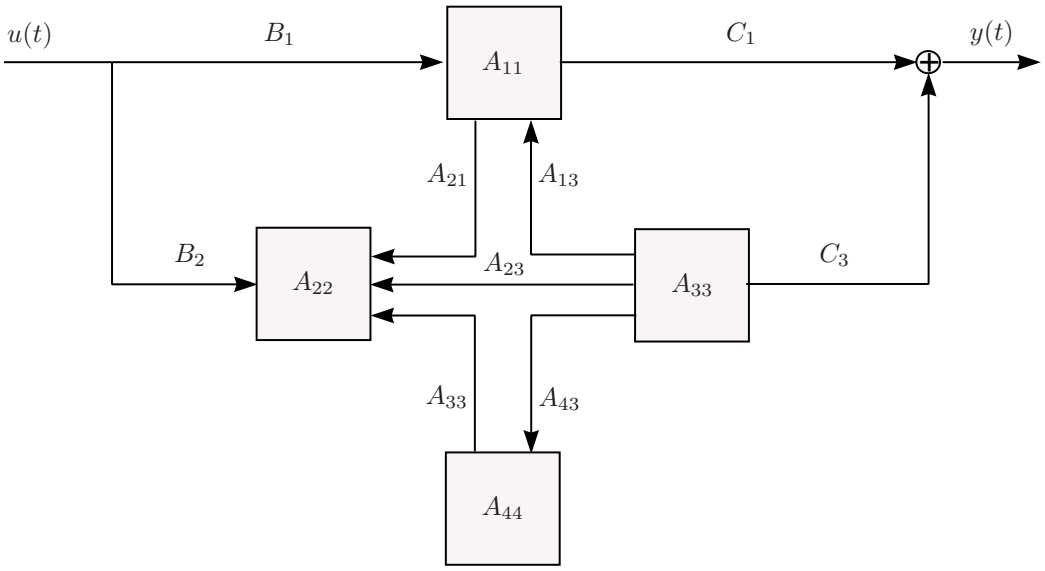

Figure 8.7: Kalman decomposition.

1. Show that $W_r(\cdot, t_0)$ is the unique solution of the matrix differential equation

$$\dot{X}(t) = A(t)X(t) + X(t)A^T(t) + B(t)B^T(t) \qquad (\ddagger)$$

with $X(t_0) = 0 \in \mathbb{R}^{n\times n}$.

2. Show that $W_r(t_1, t_0) \geq 0$ and that

$$W_r(t_1, t_0) - W_r(t_1, t_0') \geq 0 \qquad \forall\, t_0' \geq t_0. \qquad (*)$$

[Here $P \geq 0$ means that $P \in \mathbb{R}^{n\times n}$ is positive semi-definite.]

Problem 8.2 (Controllability) Consider a harmonic oscillator with control input u, satisfying the equation

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}u(t)$$

Suppose we want to drive the system from the state $[1,0]^\mathsf{T}$ to the state $[0,0]^\mathsf{T}$ in 2π units of time.

1. Does there exist a control u which makes this transfer?
2. Now suppose u is to be a piecewise constant function of time of the form

$$u(t) = \begin{cases} u_1 & 0 \leq t < 2\pi/3,\\ u_2 & 2\pi/3 \leq t < 4\pi/3,\\ u_3 & 4\pi/3 \leq t < 2\pi \end{cases}$$

Do there exist constants u_1, u_2, u_3 such that we can make the transfer from $[1,0]^\mathsf{T}$ at $t = 0$ to $[0,0]^\mathsf{T}$ at time $t = 2\pi$?

Problem 8.3 (Controllability) Let A be an $n \times n$ matrix and B be an $n \times m$ matrix, both with real entries. Assume that the pair (A, B) is controllable. Prove or disprove the following statements (a counterexample suffices to disprove a statement):

1. The pair (A^2, B) is controllable.

2. Given that the system
$$\dot{x}(t) = Ax(t) + Bu(t)$$
has the initial condition $x(0) = x_0 \neq 0$, it is possible to find a piecewise continuous control $u : [0, +\infty) \to \mathbb{R}^m$ such that the system is brought to *rest* at $t = 1$ (i.e., $x(t) = 0$ for all $t \geq 1$).

3. Suppose that the system is initially at *rest*, i.e. x(0)=0, and for $\bar{x} \in \mathbb{R}^n$ we wish to find a piecewise continuous control $u : [0, +\infty) \to \mathbb{R}^m$ such that $x(t) = \bar{x}$ for all $t \geq 1$. Such a control can be found for all $\bar{x} \in \mathbb{R}^n$.

Problem 8.4 (Observability) For some matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, $C \in \mathbb{R}^{1\times n}$ and $D \in \mathbb{R}$, consider a single-input, single-output time-invariant system

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t),\\ y(t) &= Cx(t) + Du(t).\end{aligned}$$

Assume that (A, C) is observable.

1. For a generic $t \geq 0$, provide an expression for $x(t)$ in terms of the derivatives $\left\{y^{(i)}(t); i = 0, 1, \ldots, n-1\right\}$ and $\left\{u^{(i)}(t); i = 0, 1, \ldots, n-1\right\}$. [Hint: Consider the output equation $y(t) = Cx(t) + Du(t)$ and differentiate repeatedly.]

2. Discuss why for time-invariant systems the observation of y and u over an arbitrarily small interval $t \in [0, \epsilon]$, with $\epsilon > 0$, suffices to reconstruct $x(0)$.

Problem 8.5 (Duality) Let $\phi(t, \tau)$ denote the state-transition matrix of

$$\begin{aligned}\dot{x}(t) &= A(t)x(t) + B(t)u(t)\\ y(t) &= C(t)x(t) + D(t)u(t)\end{aligned}$$

and $\psi(t, \tau)$ that of the dual system

$$\begin{aligned}\dot{\widetilde{x}}(t) &= -A(t)^*\widetilde{x}(t) - C(t)^*\widetilde{u}(t)\\ \widetilde{y}(t) &= B(t)^*\widetilde{x}(t) + D(t)^*\widetilde{u}(t)\end{aligned}$$

Show that $\psi(t, \tau) = \phi(\tau, t)^*$.

Chapter 9

State Feedback and Observer Design

Knowing whether a system is controllable or observable is usually not an end in itself. One wants to subsequently exploit this knowledge to design controllers that will make the system behave in some desired way: Ensure stability, force the output to track some desired trajectory, force the state to converge to an equilibrium at some desired rate, etc. We have already seen how for controllable systems one can generate input trajectories to steer the system from its initial state to some desired final state (Theorem 8.4). In practice, however, this may not be sufficient. Due to small deviations in the initial state or in the matrices involved in the dynamics, these so-called *open loop* input trajectories will invariably fail to steer the system as expected. Indeed, if the system is unstable the deviation may get arbitrarily large as time goes to infinity. A much more robust way of steering the system is to measure its state as we go along and if we find that it deviates from the desired trajectory adapt the open loop inputs to correct the deviation. The simplest case of such a *feedback control* involves stabilization, where the desired trajectory is simply an equilibrium solution and the objective is to design a controller to make the system asymptotically stable. Similar (dual) comments can also be made for the merits of "closed loop" state reconstruction by incorporating output measurements on-line, as opposed to the "batch processing" approach of Theorem 8.7.

In this chapter we investigate how such controllers and state estimators can be designed. We first consider the so-called state feedback controllers, where one assumes that the entire state of the system can be measured and used when making decisions. We then extend the approach to output feedback, where one assumes that only the outputs of the system are available for measurement. In particular, we consider the case of observer based output feedback, where an observer is used to reconstruct the value of the state from the output measurements; a state feedback controller then uses the reconstructed state to steer the system.

The discussion is restricted to linear time invariant systems. In this case, one makes use of the controllability and observability matrices derived in Chapter 8 to transform the system into special forms for which controllers and observers can easily be designed. The transformations involve changes of basis using invertible matrices derived from the controllability and observability matrices. For clarity of exposition we first recall some facts about basis changes.

9.1 Revision: Change of basis

Consider a linear, time invariant system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{9.1}$$
$$y(t) = Cx(t) + Du(t), \tag{9.2}$$

where as usual, $t \in \mathbb{R}_+$, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. Consider also a change of basis $\tilde{x}(t) = Tx(t)$ where $\tilde{x}(t) \in \mathbb{R}^n$ and $T \in \mathbb{R}^{n\times n}$ is an invertible matrix. Recall that in the new coordinates the evolution of system is also described by linear equations

$$\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}u(t) \tag{9.3}$$
$$y(t) = \tilde{C}\tilde{x}(t) + \tilde{D}u(t), \tag{9.4}$$

with $\tilde{A} \in \mathbb{R}^{n\times n}$, $\tilde{B} \in \mathbb{R}^{n\times m}$, $\tilde{C} \in \mathbb{R}^{p\times n}$ and $\tilde{D} \in \mathbb{R}^{p\times m}$. Throughout these notes several facts regarding the relation between these two descriptions of the system have been established. We summarize the most important ones in the following theorem.

Theorem 9.1 *Consider the linear time invariant systems (9.1)–(9.2) and (9.3)–(9.4) related through the change of coordinates $\tilde{x}(t) = Tx(t)$ for all $t \in \mathbb{R}_+$, where $T \in \mathbb{R}^{n\times n}$ is an invertible matrix. The following hold:*

1. *The matrices in (9.3)–(9.4) are given by $\tilde{A} = TAT^{-1}$, $\tilde{B} = TB$, $\tilde{C} = CT^{-1}$, $\tilde{D} = D$.*
2. *The matrices A and $\tilde{A}$ have the same eigenvalues, i.e.* $\textsc{Spec}[\tilde{A}] = \textsc{Spec}[A]$.
3. *The two systems have the same transfer function, i.e.*
$$\tilde{G}(s) = \tilde{C}(sI - \tilde{A})^{-1}\tilde{B} + \tilde{D} = C(sI - A)^{-1}B + D = G(s).$$
4. *The two systems have the same impulse response matrix, i.e.*
$$H(t) = Ce^{At}B + D\delta_0(t) = \tilde{C}e^{\tilde{A}t}\tilde{B} + \tilde{D}\delta_0(t).$$
5. *$(\tilde{A}, \tilde{B})$ is controllable if and only if (A, B) is controllable.*
6. *$(\tilde{C}, \tilde{A})$ is observable if and only if (C, A) is observable.*

The proof is left as an exercise.

The theorem shows that changes of coordinates do not affect the fundamental properties of the system. This is to be expected, since a change of coordinates does not change the system itself, just the representation of the system. So the only things affected by the change of coordinates are those having to do with the representation of the matrices and not fundamental properties such as stability, controllability, or observability.

Exercise 9.1 Are the state transition and impulse state transition matrices of the two systems the same?

In this chapter we will repeatedly use the properties of coordinate transformations to bring the system equations into special forms that make it easier to design controllers and observers. We will then exploit the properties listed in Theorem 9.1 to transfer the design to the original coordinates. The two forms we will especially be interested in are the so-called "controllable canonical form" and "observable canonical form". Both of these will first be presented for the easier case of single input, single output systems and then generalized.

9.2 Linear state feedback for single input systems

Consider first a linear time invariant system with only one input

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{9.5}$$

where $u(t) \in \mathbb{R}$ and $B \in \mathbb{R}^n$ is a vector. The output equations will play no role in this section and will be omitted for simplicity, we will return to them in Section 9.3.

Recall that the system (9.5) is controllable if and only if the matrix

$$P = \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n \times n} \tag{9.6}$$

has full rank, or, equivalently (since the matrix is square), if and only if it is invertible, or, equivalently, if and only if its columns are linearly independent. We will use the last fact to construct an invertible matrix $T \in \mathbb{R}^{n \times n}$ to use as a change of coordinates. We will start by defining a matrix $S \in \mathbb{R}^{n \times n}$ that will be invertible whenever the system is controllable and will eventually become T^{-1}.

Let the characteristic polynomial of the matrix $A \in \mathbb{R}^{n \times n}$ be

$$\textsc{Det}[\lambda I - A] = \lambda^n + \chi_1 \lambda^{n-1} + \dots + \chi_{n-1}\lambda + \chi_n.$$

Consider a family of n vectors, $s_i \in \mathbb{R}^n$ for $i = 1, \dots, n$, defined last-to-first as follows:

$$\begin{aligned} s_n &= B \\ s_{n-1} &= As_n + \chi_1 B = AB + \chi_1 B \\ s_{n-2} &= As_{n-1} + \chi_2 B = A^2 B + \chi_1 AB + \chi_2 B \\ &\vdots \\ s_1 &= As_2 + \chi_{n-1} B = A^{n-1}B + \chi_1 A^{n-2} B + \dots + \chi_{n-1} B. \end{aligned}$$

Lemma 9.1 *$As_1 + \chi_n B = 0$. The matrix $S = [s_1 \ \dots \ s_n] \in \mathbb{R}^{n \times n}$ is invertible if and only if the system (9.5) is controllable.*

Proof: By definition

$$\begin{aligned} As_1 + \chi_n B &= A^n B + \chi_1 A^{n-1} B + \dots + \chi_{n-1} AB + \chi_n B \\ &= \left(A^n + \chi_1 A^{n-1} + \dots + \chi_{n-1} A + \chi_n I \right) B = 0 \end{aligned}$$

by the Cayley-Hamilton theorem (Theorem 5.7).

Recall that S is invertible if and only if its columns are linearly independent. Notice that s_i is the sum of $A^{n-i}B$ plus a linear combination of $A^{n-i-1}B$, …, B. The $A^{n-i}B$ terms are linearly independent of each other if and only if the system is controllable, hence the claim follows. Or, more formally, note that

$$S = \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \cdot \begin{bmatrix} \chi_{n-1} & \chi_{n-2} & \cdots & \chi_1 & 1 \\ \chi_{n-2} & \chi_{n-3} & \dots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \chi_1 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

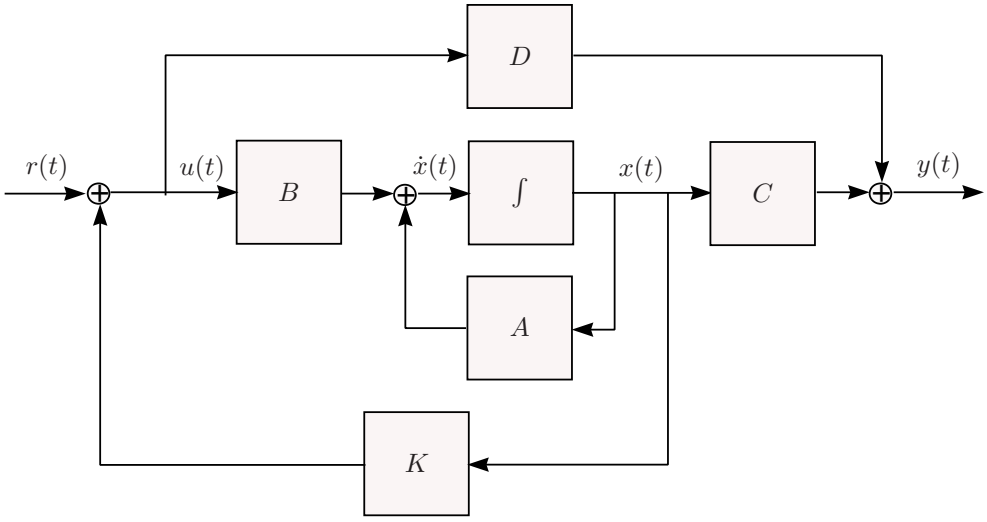

Figure 9.1: Linear state feedback.

Therefore, recalling the definition of the matrix P in (9.6),

$$\textsc{Det}[S] = \textsc{Det}[P] \cdot \textsc{Det} \begin{bmatrix} \chi_{n-1} & \chi_{n-2} & \cdots & \chi_1 & 1 \\ \chi_{n-2} & \chi_{n-3} & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \chi_1 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{bmatrix},$$

which is not equal to zero if and only if the system is controllable; notice that the determinant of the matrix on the right is equal to ± 1. ∎

The lemma shows that whenever system (9.5) is controllable we can use S to define a change of coordinates. It turns out that in the new coordinates the system matrices take a particularly useful form.

Theorem 9.2 *The system (9.5) is controllable if and only if there exists a change of coordinates $\tilde{x}(t) = Tx(t)$ with $T \in \mathbb{R}^{n\times n}$ invertible, such that the matrices $\tilde{A} = TAT^{-1}$ and $\tilde{B} = TB$ satisfy:*

$$\tilde{A} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\chi_n & -\chi_{n-1} & -\chi_{n-2} & \cdots & -\chi_1 \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{9.7}$$

Proof: ($\Rightarrow$): Assume that the system is controllable. Define $S \in \mathbb{R}^{n\times n}$ as above and let $T = S^{-1}$. Note that

$$B = S \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \Rightarrow \tilde{B} = TB = S^{-1}B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Consider next the matrix $AS = A\,[s_1 \ \dots \ s_n]$. Reading the columns from the right we have

$$As_n = s_{n-1} - \chi_1 B = S \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -\chi_1 \end{bmatrix}, \quad As_{n-1} = s_{n-2} - \chi_2 B = S \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ -\chi_2 \end{bmatrix}, \text{ etc. until}$$

$$As_1 = -\chi_n B = S \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ -\chi_n \end{bmatrix} \text{ (by Lemma 9.1).}$$

Hence,

$$AS = S \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\chi_n & -\chi_{n-1} & -\chi_{n-2} & \cdots & -\chi_1 \end{bmatrix}$$

which implies that

$$\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\chi_n & -\chi_{n-1} & -\chi_{n-2} & \cdots & -\chi_1 \end{bmatrix} = S^{-1}AS = TAT^{-1} = \tilde{A}.$$

($\Leftarrow$): Assume that there exists a change of coordinates bringing the systems in the form (9.7). Note that in the new coordinates

$$\tilde{P} = \left[\tilde{B} \ \dots \ \tilde{A}^{n-1}\tilde{B}\right] = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & * \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 1 & \dots & * & * \\ 1 & -\chi_1 & \dots & * & * \end{bmatrix}$$

where $*$ stands for some number which depends on $\chi_1, \dots, \chi_n$. Hence $\textsc{Det}\left[\tilde{P}\right] = 1$ and the system in the new coordinates is controllable. Theorem 9.1 then implies that the system in the original coordinates is also controllable. ∎

The system representation of equation (9.7) is known as the *controllable canonical form*. The term "canonical" refers to the fact that any controllable system can be brought to this form through a change of coordinates, as Theorem 9.2 shows. The controllable canonical form is useful for designing state feedback controllers to modify the behavior of system (9.5). A state feedback controller is a method for selecting the inputs of the system as a function of the state to force the system to exhibit some desired behavior. Here we will restrict our attention to linear time invariant state feedback controllers of the form

$$u(t) = Kx(t) + r(t), \tag{9.8}$$

where $K \in \mathbb{R}^{m\times n}$ is known as the *gain matrix*, and $r(t) \in \mathbb{R}^m$ is an external input vector; in the single input case considered in this section $K = [k_1 \ \dots \ k_n] \in \mathbb{R}^{1\times n}$ is just a row vector.

The system obtained by connecting a state feedback controller of the form (9.8) with a system of the form (9.5) is known as the *closed loop system* (Figure 9.1). Its dynamics are described by the linear state equations

$$\dot{x}(t) = (A + BK)x(t) + Br(t). \tag{9.9}$$

We would like to select the gain matrix K such that the closed loop systems exhibits some desired behavior, for example is asymptotically stable and its state tends to zero sufficiently fast. To ensure properties like these one typically needs to select K so that the eigenvalues of the closed loop system matrix $A + BK$ coincide with some desired set $\{\lambda_1, \dots, \lambda_n\} \subseteq \mathbb{C}$. Clearly for this to be possible the set must contain the complex conjugates of all its members, i.e. $\{\overline{\lambda_1}, \dots, \overline{\lambda_n}\} = \{\lambda_1, \dots, \lambda_n\}$; we call such a set of complex numbers a *complex conjugate set*.

Theorem 9.3 *System (9.5) is controllable if and only if for all complex conjugate sets $\{\lambda_1, \dots, \lambda_n\} \subseteq \mathbb{C}$ there exists $K \in \mathbb{R}^{m \times n}$ such that* $\textsc{Spec}[A + BK] = \{\lambda_1, \dots, \lambda_n\}$.

Proof: ($\Rightarrow$). Assume that (9.5) is controllable. Given a desired complex conjugate set $\{\lambda_1, \dots, \lambda_n\} \subseteq \mathbb{C}$ we would like to select feedback gains K such that the closed loop matrix $A + BK$ has this set as eigenvalues. In other words, we would like

$$\begin{aligned} \textsc{Det}[\lambda I - (A + BK)] &= (\lambda - \lambda_1)(\lambda - \lambda_2) \dots (\lambda - \lambda_n) \\ &= \lambda^n + d_1\lambda^{n-1} + \dots + d_{n-1}\lambda + d_n \end{aligned} \tag{9.10}$$

where $d_1, \dots, d_n$ are real numbers uniquely determined by $\lambda_1, \dots, \lambda_n$.

By Theorem 9.2 there exists a change of coordinates $\tilde{x}(t) = Tx(t)$ to bring the system into controllable canonical form

$$\dot{\tilde{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\chi_n & -\chi_{n-1} & -\chi_{n-2} & \dots & -\chi_1 \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u(t). \tag{9.11}$$

In the new coordinates, the feedback function can be written as

$$u(t) = Kx(t) + r(t) = KT^{-1}Tx(t) + r(t) = \tilde{K}\tilde{x}(t) + r(t)$$

where $\tilde{K} = \begin{bmatrix} \tilde{k}_n & \dots & \tilde{k}_1 \end{bmatrix} \in \mathbb{R}^{1 \times n}$ is the representation of the gain matrix in the new coordinates. Substituting into (9.11) we obtain

$$\begin{aligned} \dot{\tilde{x}}(t) &= (\tilde{A} + \tilde{B}\tilde{K})\tilde{x}(t) + \tilde{B}r(t) \\ &= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \tilde{k}_n - \chi_n & \tilde{k}_{n-1} - \chi_{n-1} & \tilde{k}_{n-2} - \chi_{n-2} & \dots & \tilde{k}_1 - \chi_1 \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} r(t). \end{aligned}$$

Notice that the system is still in controllable canonical form. Therefore the terms in the last row are the coefficients of the characteristic polynomial

$$\textsc{Det}[\lambda I - (\tilde{A} + \tilde{B}\tilde{K})] = \lambda^n - (\tilde{k}_1 - \chi_1)\lambda^{n-1} - \dots - (\tilde{k}_{n-1} - \chi_{n-1})\lambda - (\tilde{k}_n - \chi_n).$$

If we set

$$\tilde{k}_1 = \chi_1 - d_1, \ \dots, \ \tilde{k}_n = \chi_n - d_n$$

this becomes identical to the desired characteristic polynomial (9.10). Hence with this selection for the matrix $\tilde{K}$ we have made the eigenvalues of $\tilde{A} + \tilde{B}\tilde{K}$ equal to the desired set $\{\lambda_1, \dots, \lambda_n\}$. Finally,

since eigenvalues are not affected by coordinate transformations (Theorem 9.1) setting $K = \tilde{K}T$ ensures that with the state feedback $u(t) = \tilde{K}Tx(t) + r(t)$ the eigenvalues of the matrix $A + BK$ of the closed loop system are also equal to the desired list.

($\Leftarrow$). By contraposition. Assume that the system is not controllable. Then, by Theorem 8.9, there exists a change of coordinates $\tilde{x}(t) = Tx(t)$ such that

$$\dot{\tilde{x}}(t) = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} \tilde{B}_1 \\ 0 \end{bmatrix} u(t)$$

and the pair of matrices $(\tilde{A}_{11}, \tilde{B}_1)$ is controllable. In these coordinates a feedback function $u(t) = Kx(t) + r(t)$ can be written as

$$u(t) = KT^{-1}Tx(t) + r(t) = \begin{bmatrix} \tilde{K}_1 & \tilde{K}_2 \end{bmatrix} \tilde{x}(t) + r(t).$$

The closed loop system then becomes

$$\dot{\tilde{x}}(t) = \begin{bmatrix} \tilde{A}_{11} + \tilde{B}_1\tilde{K}_1 & \tilde{A}_{12} + \tilde{B}_1\tilde{K}_2 \\ 0 & \tilde{A}_{22} \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} \tilde{B}_1 \\ 0 \end{bmatrix} r(t)$$

and its characteristic polynomial

$$\textsc{Det} \begin{bmatrix} \lambda I_1 - (\tilde{A}_{11} + \tilde{B}_1\tilde{K}_1) & -(\tilde{A}_{12} + \tilde{B}_1\tilde{K}_2) \\ 0 & \lambda I_2 - \tilde{A}_{22} \end{bmatrix} = \textsc{Det}[\lambda I_1 - (\tilde{A}_{11} + \tilde{B}_1\tilde{K}_1)]\textsc{Det}[\lambda I_2 - \tilde{A}_{22}],$$

where I_1 and I_2 denote identity matrices of appropriate dimensions. Hence

$$\textsc{Spec}[\tilde{A} + \tilde{B}\tilde{K}] = \textsc{Spec}[\tilde{A}_{11} + \tilde{B}_1\tilde{K}_1] \cup \textsc{Spec}(\tilde{A}_{22}).$$

Notice that part of the spectrum (the eigenvalues of $\tilde{A}_{22}$) is not affected by the feedback gains $\tilde{K}$. Therefore given a list of desired eigenvalues $\{\lambda_1, \ldots, \lambda_n\} \subseteq \mathbb{C}$ it is impossible, in general, to select $\tilde{K}$ such that $\textsc{Spec}[\tilde{A} + \tilde{B}\tilde{K}] = \{\lambda_1, \ldots, \lambda_n\}$; since the pair of matrices $(\tilde{A}_{11}, \tilde{B}_1)$ is controllable this is possible if and only if $\textsc{Spec}[\tilde{A}_{22}] \subseteq \{\lambda_1, \ldots, \lambda_n\}$. Setting $K = \tilde{K}T$ and noting that $\textsc{Spec}[A+BK] = \textsc{Spec}[\tilde{A} + \tilde{B}\tilde{K}]$ (Theorem 9.1) completes the proof. ∎

Theorem 9.3 provides a method for moving the eigenvalues of a controllable linear system to arbitrary, complex conjugate values by selecting a feedback gain matrix K and applying linear state feedback $u(t) = Kx(t) + r(t)$. This method is known as *pole placement*, *eigenvalue placement*, or *eigenvalue assignment*. Even though for the proof of the theorem we first have to bring the system into controllable canonical form, in practice this is not necessary. Given matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ we can directly compute the characteristic polynomial $\textsc{Det}[\lambda I - (A + BK)]$, treating the entries in the matrix K as variables. The coefficients of the characteristic polynomial will turn out to be functions of the entries of the gain matrix $K \in \mathbb{R}^{m\times n}$. Equating these coefficients with those of the desired characteristic polynomial (9.10) leads to a system of equations where the entries of K play the role of the unknowns. Theorem 9.3 then ensures that this system will have a solution which, if used in the feedback function $u(t) = Kx(t) + r(t)$, will lead to a closed loop system with the desired eigenvalues. As we will see below, this procedure works more generally, even if $m > 1$. The difference is that for single input systems ($m = 1$) the system of equations has the same number of equations as unknowns (both n). Thus (assuming controllability) there is a unique solution to this system of equations and hence a unique feedback gain matrix $K \in \mathbb{R}^{1\times n}$ that will do the job. For multi-input systems ($m > 1$), on the other hand, there are more unknowns (nm) than equations (n). Therefore there will be multiple choices of K that will lead to the same eigenvalues for the close loop system. Unless other considerations are present (minimizing some cost criterion, etc.) one can eliminate the redundant degrees of freedom by setting some of the elements of K equal to zero. This leads to a sparser feedback gain matrix, which both simplifies the calculations and makes the feedback function easier to implement in practice.

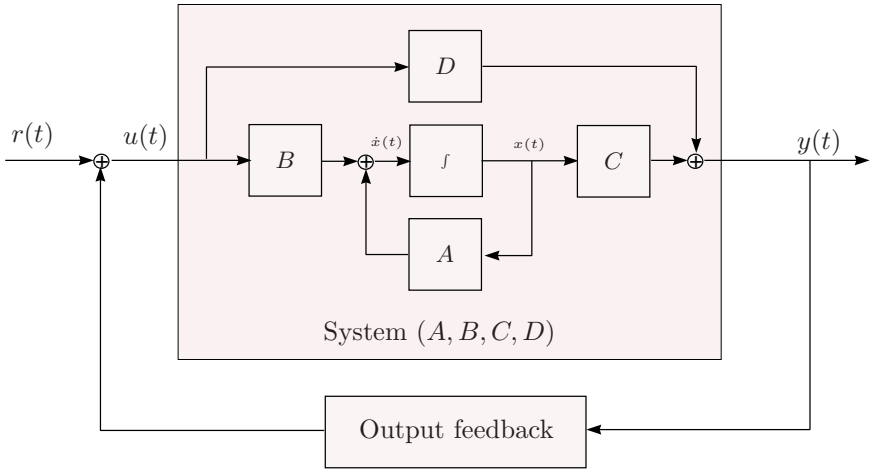

Figure 9.2: Output feedback.

Theorem 9.3 also shows that the procedure outlined above for designing feedback gain matrices will in general fail if the system is not controllable. The same steps can again be executed, but in this case the system of equations in the entries of K will not have any solution in general. The reason is the presence of the uncontrollable modes (the elements of $\text{SPEC}[\tilde{A}_{22}]$ in the proof of Theorem 9.3) which are not affected by our choice of K. If $\text{SPEC}[\tilde{A}_{22}]$ contains some eigenvalues with positive real part, whatever gain matrix K we select, it will not be possible to obtain a stable closed loop system. If stability is what we are after, then drastic system re-design is needed. We may, for example, try to add, or modify one of the system actuators to augment the matrix B and make the offending modes controllable. If, on the other hand, $\text{SPEC}[\tilde{A}_{22}]$ lies entirely on the left half of the complex plane (i.e. the system is stabilizable) it is still possible to stabilize the system. We just have to resign ourselves to the fact that the eigenvalues in $\text{SPEC}[\tilde{A}_{22}]$ will stay the same whatever we do and select K to move the remaining eigenvalues (the controllable modes in $\text{SPEC}[\tilde{A}_{11}]$) to desired locations. To do this, we first we have to determine the uncontrollable modes (e.g. using the rank test in Theorem 8.9, condition 4), then construct a list of desired eigenvalues containing the uncontrollable modes, compute the characteristic polynomial $\text{DET}[\lambda I - (A + BK)]$, and equate its coefficients with those of the desired characteristic polynomial. The proof of Theorem 9.3 together with the fact that the uncontrollable modes are included in the list of desired eigenvalues will ensure that the resulting system of equations will have a solution for the entries of K. The solution will not be unique, however, even for single input systems: Some of the degrees of freedom (those corresponding to $\tilde{K}_2$ in the proof of Theorem 9.3) will be redundant. Unless other considerations are present, one can again set these redundant degrees of freedom to zero to make the matrix K as sparse as possible.

9.3 Linear state observers for single output systems

The method of pole placement described above is powerful and theoretically elegant. One can argue, however, it is not entirely practical. In real systems it is often the case that not all states are available for measurement. The set up of Figure 9.1 is therefore optimistic, because in reality one only has access to information about the output of the system (measured through appropriate sensors) when trying to make decisions about what inputs to apply. A more realistic feedback arrangement should therefore look more like Figure 9.2. The problem now is how to design a feedback function that uses the measurements of the outputs (past and present) to decide what input to apply to the system.

Since we already have a powerful method for designing state feedback controllers, one idea that presents itself is to augment this by an algorithm that generates an estimate of the state using past and present inputs and outputs. Then, instead of the true value of the state (which is not

available) we can use the estimate of the state and our state feedback controller to compute a value for the input to apply. An algorithm that generates an estimate of the state using the past and present inputs and outputs is known as a *state estimator* or a *state observer*. The resulting feedback arrangement will look like Figure 9.3.

It turns out that for linear systems it suffices to consider linear state observers. Analogously to the case of state feedback design we will start by describing the design of such observers for single output systems ($p = 1$) and return to the general case of multi-output systems ($p > 1$) toward the end of this chapter. We will consider the usual linear system equations

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{9.12}$$
$$y(t) = Cx(t) + Du(t) \tag{9.13}$$

where $C \in \mathbb{R}^{1\times n}$ and $D \in \mathbb{R}^{1\times m}$ are just row vectors.

A linear state observer generates an estimate of the state, $\hat{x}(t)$, at time $t \in \mathbb{R}_+$ using the values of the input and output for all times up to t, $\{(u(\tau), y(\tau)) \mid \tau \in [0, t]\}$. The observer can itself be thought of as a linear system, with n states ($\hat{x}(t) \in \mathbb{R}^n$), $m + p$ inputs ($(u(t), y(t)) \in \mathbb{R}^{m+p}$), and n outputs (the state estimate itself $\hat{x}(t) \in \mathbb{R}^n$). The equations of the observer try to mimic those of the linear system (9.12)–(9.13), with a correction term to account for any mismatch between the measured output, $y(t)$, at time t and the output predicted by the observer (denoted by $\hat{y}(t)$) based on its own state estimate, $\hat{x}(t)$. More precisely,

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t)) \tag{9.14}$$
$$\hat{y}(t) = C\hat{x}(t) + Du(t). \tag{9.15}$$

Notice that at time t the observer only makes use of its own estimate of the state, $\hat{x}(t)$, the input $u(t)$ and the output $y(t)$ to update its estimate of the state; in particular the unknown value of the state, $x(t)$, is not used anywhere. The linear correction term, $L(y(t) - \hat{y}(t))$, introduces an indirect estimate of the difference between the estimated and the true value of the state which, as we will see below, can be used to make the two converge to each other. The matrix $L \in \mathbb{R}^{n\times p}$ is known as the *observer gain matrix*. Note also that only the present values of $y(t)$ and $u(t)$ are used in Equations (9.14)–(9.15). The state estimate $\hat{x}(t)$ at time t, however, will depend on all of the past values of the inputs and outputs, $\{(u(\tau), y(\tau)) \mid \tau \in [0, t]\}$, as well as the initial condition $\hat{x}_0$ that we select. In a sense, $\hat{x}(t)$ encapsulates all the relevant state information contained in the input and output trajectories up to time t.

Under what conditions will $\hat{x}(t)$ be a good estimate of $x(t)$? To study this question we can look at the estimation error, $e(t) = x(t) - \hat{x}(t) \in \mathbb{R}^n$. If this error (or more precisely its norm) is small then the state estimate generated by the observer is accurate. Clearly, since initially we know nothing about the true value of $x(0)$ our initial estimate, $\hat{x}(0)$, and hence the initial error, $e(0)$, can be arbitrarily bad. The question is whether the estimate will get better as time goes on and more information is collected through $u(t)$ and $y(t)$. To answer this question we can look at how the error, $e(t)$, evolves over time:

$$\begin{aligned}
\dot{e}(t) &= \dot{x}(t) - \dot{\hat{x}}(t) \\
&= Ax(t) + Bu(t) - A\hat{x}(t) - Bu(t) - L(y(t) - \hat{y}(t)) \\
&= A(x(t) - \hat{x}(t)) - L(Cx(t) + Du(t) - Cx(t) - Du(t)) \\
&= (A - LC)(x(t) - \hat{x}(t)) \\
&= (A - LC)e(t).
\end{aligned}$$

Notice that the evolution of the estimation error is also governed by a linear system without any inputs. The estimation error will therefore converge to zero if and only if the eigenvalues of the matrix $A - LC$ have negative real part. Otherwise, as time goes on our estimate of the state will remain as bad as our initial guess, or get even worse, despite the fact that we have collected more information through the output of the system.

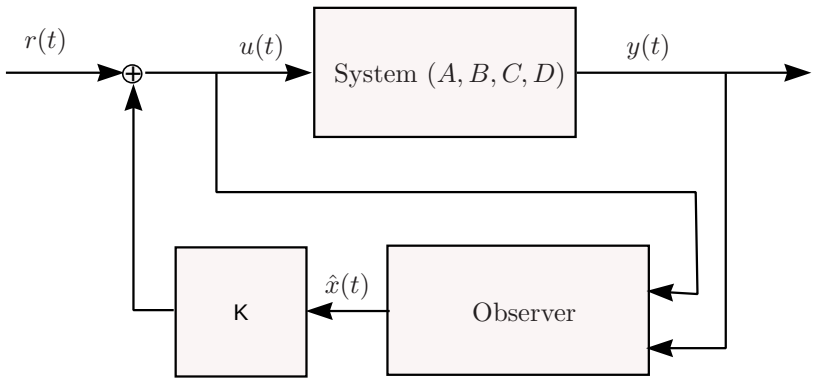

Figure 9.3: Linear feedback with state estimation. $\hat{x}(t)$ denotes the estimate of the state

The question now becomes can we select the observer gain matrix L so that the eigenvalues of $A-LC$ have negative real parts? It is easy to see that this question is the dual of the question addressed in Section 9.2. If the system is observable, one can in fact mimic the pole placement method step by step to develop a method for placing the eigenvalues of the estimation error dynamics in some desired, complex conjugate values. First, one can write the system in *observable canonical form* using a change of coordinates.

Theorem 9.4 *The system (9.12)–(9.13) is observable if and only if there exists a change of coordinates $\tilde{x}(t) = Tx(t)$ with $T \in \mathbb{R}^{n\times n}$ invertible, such that the matrices $\tilde{A} = TAT^{-1}$ and $\tilde{C} = BT^{-1}$ satisfy:*

$$\tilde{A} = \begin{bmatrix} 0 & 0 & \dots & 0 & -\chi_n \\ 1 & 0 & \dots & 0 & -\chi_{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & -\chi_1 \end{bmatrix},$$
$$\tilde{C} = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

The proof is the dual to that of Theorem 9.2: One uses the rows of the observability matrix

$$O = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{n\times n}$$

which under the observability assumption are linearly independent to generate a change of coordinates and bring the system to the desired form. One can then use the observable canonical form to show the following.

Theorem 9.5 *System (9.12)–(9.13) is observable if and only if for all complex conjugate sets $\{\lambda_1, \dots, \lambda_n\} \subseteq \mathbb{C}$ there exists $L \in \mathbb{R}^{m\times n}$ such that* $\textsc{Spec}[A - LC] = \{\lambda_1, \dots, \lambda_n\}$.

The procedure for selecting the observer gain matrix is also similar to the one for selecting the controller gain matrix. One uses the desired eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ to generate a desired characteristic polynomial. Then one computes the polynomial $\textsc{Det}[\lambda I - (A - LC)]$ and equates its coefficients to those of the desired polynomial, giving rise to a system of n equations in np unknowns (the entries of the observer gain matrix $L \in \mathbb{R}^{n\times p}$). If the system (9.12)–(9.13) is observable, then the system of equations has a unique solution (if $p = 1$) or multiple solutions (if $p > 1$), for which the eigenvalues of $A-LC$ are equal to the desired ones. If the system (9.12)–(9.13) is not observable, on

the other hand, the system of equations will in general have no solution. If the system is not even detectable (i.e. some of the unobservable modes have non-negative real parts) then there is no hope of designing a stable observer that will generate a correct estimate of the state, even asymptotically. If the system is detectable one can still design a stable observer by computing the unobservable modes, including them in the list of desired eigenvalues, and placing the remaining eigenvalues in some desired locations as above.

9.4 Output feedback and the separation principle

We have seen how, assuming that the value of the state is known, one can design a linear feedback controller so that the eigenvalues of the closed loop system will be at some desired position. We have also seen how, in cases where the value of the state is not fully measured, one can design an observer that will asymptotically reconstruct the state. Presumably when one puts the two designs (controller and observer) together as in Figure 9.3 the closed loop system will work as planned: In the long run the state estimation error will converge to zero, the observer will provide the feedback controller with the right value of the state, the controller will apply an appropriate input for this value of the state, and everything will work. Or will it? Could something go catastrophically wrong in the transients? Could the initial state estimation error cause the controller to make the wrong decisions and destabilize the system, leading to greater estimation error, greater mistakes in the control inputs, etc. so that eventually the whole thing breaks down?

To answer this question we now turn our attention to the dynamics of the closed loop system of Figure 9.3. Collecting all the relations from Sections 9.2 and 9.3 we have:

$$\begin{aligned}
\dot{x}(t) &= Ax(t) + Bu(t)\\
y(t) &= Cx(t) + Du(t)\\
u(t) &= K\hat{x}(t) + r(t)\\
\dot{\hat{x}}(t) &= A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t))\\
\hat{y}(t) &= C\hat{x}(t) + Du(t).
\end{aligned}$$

Substituting one into the other leads to

$$\begin{aligned}
\dot{x}(t) &= Ax(t) + BK\hat{x}(t) + Br(t)\\
\dot{\hat{x}}(t) &= LCx(t) + (A + BK - LC)\hat{x}(t) + Br(t)\\
y(t) &= Cx(t) + DK\hat{x}(t) + Dr(t).
\end{aligned}$$

The closed loop system is therefore itself a linear time invariant system, with $2n$ states $(x(t), \hat{x}(t)) \in \mathbb{R}^{2n}$), m inputs ($r(t) \in \mathbb{R}^m$) and p outputs ($y(t) \in \mathbb{R}^p$) and state equations

$$\begin{aligned}
\begin{bmatrix} \dot{x}(t) \\ \dot{\hat{x}}(t) \end{bmatrix} &= \begin{bmatrix} A & BK \\ LC & A + BK - LC \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} + \begin{bmatrix} B \\ B \end{bmatrix} r(t)\\
y(t) &= \begin{bmatrix} C & DK \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} + Dr(t).
\end{aligned}$$

To determine whether this system is stable it is easier to start with a coordinate transformation, making the state estimation error one of the states. Note that

$$\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = \begin{bmatrix} x(t) \\ x(t) - \hat{x}(t) \end{bmatrix} = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} = T \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix}$$

where I stands for the $n \times n$ identity matrix.

Exercise 9.2 Show that $T = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix}$ is invertible and that $T^{-1} = T$.

In the new coordinates

$$\begin{bmatrix} \dot{x}(t) \\ \dot{e}(t) \end{bmatrix} = \begin{bmatrix} A+BK & -BK \\ 0 & A-LC \end{bmatrix} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} r(t)$$
$$y(t) = \begin{bmatrix} C+DK & -DK \end{bmatrix} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} + Dr(t).$$

Note that

$$\textsc{Det} \begin{bmatrix} \lambda I - (A+BK) & BK \\ 0 & \lambda I - (A-LC) \end{bmatrix} = \textsc{Det}[\lambda I - (A+BK)]\textsc{Det}[\lambda I - (A-LC)].$$

hence

$$\textsc{Spec} \begin{bmatrix} A+BK & -BK \\ 0 & A-LC \end{bmatrix} = \textsc{Spec}[A+BK] \cup \textsc{Spec}[A-LC].$$

In summary, the $2n$ eigenvalues of the closed loop system coincide with the n eigenvalues of the system with perfect state feedback and the n eigenvalues of the state estimation error dynamics. This fact greatly simplifies the design of output feedback controllers for linear systems. One can design the observer and the state feedback gain matrix separately, put them together, and the resulting closed loop system will work as planned. This separation of the controller from the observer design, known as the *principle of separation*, is a fundamental property of linear systems. Unfortunately, even though related principles have been established for other classes of systems, this nice property does not always hold.

From the above discussion it appears that, assuming that the system is controllable and observable, the eigenvalues of the closed loop system can be arbitrarily placed. Therefore in principle one can make the closed loop system respond arbitrarily fast, by making the real part of all its eigenvalues sufficiently negative. There are several good practical reasons, however, why this temptation should be resisted. The most important is modelling inaccuracies. Any model of a system used for design or analysis is an approximate mathematical abstraction of a physical process; in fact a linear model is usually a rather crude approximation of reality, valid only for certain values of the state and input. The real physical process is bound to exhibit nonlinearities, additional dynamics, noise in the sensors and actuators, delays, and other phenomena not adequately captured by the model. Faster response, small eigenvalues and high gains tend to make the closed loop system more sensitive to all these unmodeled factors.

In the end, the choice of the eigenvalues of the closed loop system (and hence the gains of the controller and observer) usually comes down to a trade-off between several such considerations. For linear systems, methods that in some cases allow one to establish optimal trade-offs have been developed, in areas such as optimal filtering, linear quadratic Gaussian control, robust control, etc. Very often, however, insight, intuition, and trial-and-error play a central role.

9.5 The multi-input, multi-output case

Finally, we turn our attention to systems with multiple inputs ($m > 1$) and/or multiple outputs ($p > 1$). The situation is very similar in this case. Careful consideration of the development in the previous sections indeed reveals that the only place where the single input (respectively single output) assumption is used is in the development of the controllable canonical form in Theorem 9.2 (respectively observable canonical form in Theorem 9.4). This in turn is used in the "only if" part of Theorem 9.3, to establish that if the system is controllable then its eigenvalues can be moved to arbitrary locations by state feedback) (respectively Theorem 9.5). The rest of the argument (including the entire discussion in Section 9.4) remains unaffected, even if $m > 1$, or $p > 1$. In this section we will discuss how the missing parts of Theorem 9.2 and Theorem 9.3 can be filled in for multi-input systems; the discussion of observer design for multi-output systems is dual and will be omitted.

Consider again system (9.5) with $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, possibly with $m > 1$. Consider the columns of $B = \begin{bmatrix} b_1 & \dots & b_m \end{bmatrix}$ and assume that they are linearly independent. This assumption can be made without loss of generality. Indeed, if the columns of B are linearly dependent there exists a change of coordinates in $\mathbb{R}^m$ $\hat{u}(t) = \hat{T}u(t)$ with $\hat{T} \in \mathbb{R}^{m\times n}$ invertible such that for $\hat{B} = B\hat{T}^{-1}$ and an appropriate partition of $\hat{u}(t)$:

$$Bu(t) = \hat{B}\hat{u} = \begin{bmatrix} \hat{B}_1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \hat{u}_1(t) \\ \hat{u}_2(t) \end{bmatrix}$$

and the columns of $\hat{B}_1$ are linearly independent; 0 denotes a matrix of appropriate dimensions all of whose elements are equal to zero. Hence in the new coordinates the inputs $\hat{u}_2(t)$ are redundant and we can work with just the reduced system

$$\dot{x}(t) = Ax(t) + \hat{B}_1\hat{u}_1(t)$$

where the columns of $\hat{B}_1$ are linearly independent.

Assuming that the system is controllable, the matrix

$$P = \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n\times nm}$$

has rank equal to n. As for the single input case, the idea is to use this matrix to generate a change of coordinates, $\tilde{x}(t) = Tx(t)$ for some $T \in \mathbb{R}^{n\times n}$ invertible, so that the resulting matrices $\tilde{A} = TAT^{-1}$ and $\tilde{B} = TB$ are such that one can easily design feedback controllers for them. Since the matrix has rank n, out of its nm columns,

$$\{b_1, \dots, b_m, Ab_1, \dots, Ab_m, \dots, A^{n-1}b_1, \dots, A^{n-1}b_m\}$$

we can select n linearly independent ones. The way we do the selection is not particularly important. Different choices will lead to somewhat different structures in the matrices $\tilde{A}$ and $\tilde{B}$, hence the resulting form we obtain in the end will be somewhat less "canonical" than the controllable canonical form is for single input systems. The most common choice is to start selecting the columns from the left, skipping any columns that are linearly dependent on the previously selected ones, until n linearly independent columns have been selected. Since we have assumed that the columns of B are linearly independent, b_1, …, b_m will be among the selected columns. In addition, for each $j = 1, \dots, m$ there is a maximal integer $k_j = 1, \dots, n$ such that $A^{k_j-1}b_j$ will be among the selected columns. Or, equivalently, a minimal integer such that $A^{k_j}b_j$ can be written as a linear combination of the previously selected columns.

Exercise 9.3 Show that if $A^{k_j-1}b_j$ is among the selected columns for some k_j then so is $A^{k_j-i}b_j$ for all $i = 1, \dots, k_j$. Or, equivalently, that if $A^{k_j}b_j$ can be written as a linear combination of the previously selected columns then so can A^ib_j for all $i \geq k_j$.

In summary, the selected columns will come in chains of the form

$$\begin{aligned} &b_1, \dots, A^{k_1-1}b_1 \\ &b_2, \dots, A^{k_2-1}b_2 \\ &\quad\vdots \\ &b_m, \dots, A^{k_m-1}b_m. \end{aligned}$$

The integers $k_1, \dots, k_m$ are known as the *controllability indices* of the system.

Exercise 9.4 Show that the sum of the controllability indices is $k_1 + \dots + k_m = n$.

We will assume that the controllability indices come in increasing order $k_1 \leq k_2 \leq \ldots \leq k_m$; if not, reorder the columns of B and $\hat{P}$ below so that this is the case.

Arrange the selected columns in a new matrix

$$\hat{P} = \begin{bmatrix} b_1 & \ldots & A^{k_1-1}b_1 & b_2 & \ldots & A^{k_2-1}b_2 & \ldots & b_m & \ldots & A^{k_m-1}b_m \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Note that $\hat{P}$ is invertible since, by construction, its columns are linearly independent. Compute the inverse of $\hat{P}$ and consider its rows

$$\hat{P}^{-1} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} \in \mathbb{R}^{n\times n}, \;\; p_1, \ldots, p_n \in \mathbb{R}^{1\times n}.$$

Among these rows select the ones corresponding to the controllability indices p_{k_1}, $p_{k_1+k_2}$, $\ldots$, $p_{k_1+\ldots+k_m} = p_n$ and form the matrix

$$T = \begin{bmatrix} p_{k_1} \\ p_{k_1}A \\ \vdots \\ p_{k_1}A^{k_1-1} \\ \vdots \\ p_n \\ p_nA \\ \vdots \\ p_nA^{k_m-1} \end{bmatrix} \in \mathbb{R}^{(k_1+\ldots+k_m)\times n} = \mathbb{R}^{n\times n}.$$

Lemma 9.2 *$p_{k_1}A^{k_1-1}b_1 = p_{k_1+k_2}A^{k_2-1}b_2 = \ldots = p_nA^{k_m-1}b_m = 1$. Moreover, $p_{k_1}A^ib_1 = 0$ for $i = 0, \ldots, k_1 - 2$ and $p_{k_1}A^ib_j = 0$ for $j = 2, \ldots, m$ and $i = 0, \ldots, k_j - 1$. Likewise, $p_{k_1+k_2}A^ib_2 = 0$ for $i = 0, \ldots, k_2 - 2$ and $p_{k_1+k_2}A^ib_j = 0$ for $j = 1, 3, \ldots, m$ and $i = 0, \ldots, k_j$, etc. The matrix T is invertible.*

Proof: By definition

$$\hat{P}^{-1}\hat{P} = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \cdot \begin{bmatrix} b_1 & \ldots & A^{k_1-1}b_1 & \ldots & b_m & \ldots & A^{k_m-1}b_m \end{bmatrix}$$

$$= \begin{bmatrix} p_1b_1 & \ldots & p_1A^{k_1-1}b_1 & \ldots & p_1A^{k_m-1}b_m \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{k_1}b_1 & \ldots & p_{k_1}A^{k_1-1}b_1 & \ldots & p_{k_1}A^{k_m-1}b_m \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_nb_1 & \ldots & p_nA^{k_1-1}b_1 & \ldots & p_nA^{k_m-1}b_m \end{bmatrix} = I.$$

Therefore, the elements on the diagonal are all equal to 1; in particular,

$$p_{k_1}A^{k_1-1}b_1 = p_{k_1+k_2}A^{k_2-1}b_2 = \ldots = p_nA^{k_m-1}b_m = 1.$$

Likewise, the off-diagonal elements are all equal to 0; in particular

$$p_{k_1}b_1 = \ldots = p_{k_1}A^{k_1-2}b_1 = p_{k_1}b_2 = \ldots = p_{k_1}A^{k_2-1}b_2 = \ldots = p_{k_1}b_m = \ldots = p_{k_1}A^{k_m-1}b_m = 0$$

$$\vdots$$

$$p_nb_1 = \ldots = p_nA^{k_1-1}b_1 = p_nb_2 = \ldots = p_nA^{k_2-1}b_2 = \ldots = p_nb_m = \ldots = p_nA^{k_m-2}b_m = 0.$$

To see that the matrix T is invertible, consider

$$T\hat{P} = \begin{bmatrix} p_{k_1} \\ \vdots \\ p_{k_1}A^{k_1-1} \\ \vdots \\ p_n \\ \vdots \\ p_nA^{k_m-1} \end{bmatrix} \cdot \begin{bmatrix} b_1 & \dots & A^{k_1-1}b_1 & \dots & b_m & \dots & A^{k_m-1}b_m \end{bmatrix}$$

$$= \begin{bmatrix} p_{k_1}b_1 & \dots & p_{k_1}A^{k_1-1}b_1 & \dots & p_{k_1}A^{k_m-1}b_m \\ p_{k_1}Ab_1 & \dots & p_{k_1}A^{k_1}b_1 & \dots & p_{k_1}A^{k_m}b_m \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{k_1}A^{k_1-1}b_1 & \dots & p_{k_1}A^{2(k_1-2)}b_1 & \dots & p_{k_1}A^{k_1+k_m-2}b_m \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_nb_1 & \dots & p_nA^{k_1-1}b_1 & \dots & p_nA^{k_m-1}b_m \\ p_nAb_1 & \dots & p_nA^{k_1}b_1 & \dots & p_nA^{k_m}b_m \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_nA^{k_m-1}b_1 & \dots & p_nA^{k_1+k_m-2)}b_1 & \dots & p_nA^{2(k_m-1)}b_m \end{bmatrix} = \begin{bmatrix} P_{11} & \dots & P_{1m} \\ P_{21} & \dots & P_{2m} \\ \vdots & \ddots & \vdots \\ P_{m1} & \dots & P_{mm} \end{bmatrix}.$$

In the last row we have decomposed the $n \times n$ matrix into blocks $P_{ij} \in \mathbb{R}^{k_i \times k_j}$. By the first part of the lemma, every block of the form P_{ii}, for $i = 1, \dots, m$ has 1 on its anti-diagonal (elements $(1, k_i)$, $(2, k_i-1)$, …, $(k_i, 1)$) and 0 everywhere else. Every block of the form P_{ij} for $i \neq j$, on the other hand has all its elements equal to zero. Therefore the last matrix is invertible (indeed, its determinant is equal to either $+1$ or to -1). Since $\textsc{Det}[T\hat{P}] = \textsc{Det}[T]\textsc{Det}[\hat{P}]$ and $\hat{P}$ is invertible by construction (recall that its columns were selected to be linearly independent), we must have that $\textsc{Det}[T] \neq 0$. Hence the matrix T is invertible. ∎

Consider now the coordinate transformation $\hat{x}(t) = Tx(t)$. It is easier to think of the matrices $\tilde{A} = TAT^{-1}$ and $\tilde{B} = TB$ as decomposed into blocks

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \dots & \tilde{A}_{1m} \\ \vdots & \ddots & \vdots \\ \tilde{A}_{m1} & \dots & \tilde{A}_{mm} \end{bmatrix} \text{ and } \tilde{B} = \begin{bmatrix} \tilde{B}_1 \\ \vdots \\ \tilde{B}_m \end{bmatrix},$$

with $\tilde{A}_{ij} \in \mathbb{R}^{k_i \times k_j}$ and $\tilde{B}_i \in \mathbb{R}^{k_i \times m}$, $i, j = 1, \dots, m$.

Let us first turn our attention to the matrix $\tilde{B} = TB$:

$$\tilde{B} = \begin{bmatrix} p_{k_1} \\ \vdots \\ p_{k_1}A^{k_1-1} \\ \vdots \\ p_n \\ \vdots \\ p_nA^{k_m-1} \end{bmatrix} \cdot \begin{bmatrix} b_1 & \dots & b_m \end{bmatrix} = \begin{bmatrix} p_{k_1}b_1 & \dots & p_{k_1}b_m \\ \vdots & \ddots & \vdots \\ p_{k_1}A^{k_1-1}b_1 & \dots & p_{k_1}A^{k_1-1}b_m \\ \vdots & \ddots & \vdots \\ p_nb_1 & \dots & p_nb_m \\ \vdots & \ddots & \vdots \\ p_nA^{k_m-1}b_1 & \dots & p_{k_1}A^{k_1-1}b_m \end{bmatrix}.$$

Scanning the elements of $\tilde{B}$ and comparing to Lemma 9.2 it becomes apparent that all of them are equal to 0, except one element in each column, which is equal to 1, namely

$$\tilde{b}_{k_1 1} = \tilde{b}_{(k_1+k_2)2} = \dots = \tilde{b}_{nm} = 1.$$

For the matrix $\tilde{A}$, the fact that $\tilde{A}T = TA$ implies that

$$\begin{bmatrix} \tilde{A}_{11} & \dots & \tilde{A}_{1m} \\ \vdots & \ddots & \vdots \\ \tilde{A}_{m1} & \dots & \tilde{A}_{mm} \end{bmatrix} \cdot \begin{bmatrix} p_{k_1} \\ \vdots \\ p_{k_1}A^{k_1-1} \\ \vdots \\ p_n \\ \vdots \\ p_nA^{k_m-1} \end{bmatrix} = \begin{bmatrix} p_{k_1} \\ \vdots \\ p_{k_1}A^{k_1-1} \\ \vdots \\ p_n \\ \vdots \\ p_nA^{k_m-1} \end{bmatrix} A$$

$$\Rightarrow \begin{bmatrix} \tilde{A}_{11} \begin{bmatrix} p_{k_1} \\ \vdots \\ p_{k_1}A^{k_1-1} \end{bmatrix} + \ldots + \tilde{A}_{1m} \begin{bmatrix} p_n \\ \vdots \\ p_nA^{k_m-1} \end{bmatrix} \\ \vdots \\ \tilde{A}_{m1} \begin{bmatrix} p_{k_1} \\ \vdots \\ p_{k_1}A^{k_1-1} \end{bmatrix} + \ldots + \tilde{A}_{mm} \begin{bmatrix} p_n \\ \vdots \\ p_nA^{k_m-1} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} p_{k_1}A \\ \vdots \\ p_{k_1}A^{k_1} \\ \vdots \\ p_nA \\ \vdots \\ p_nA^{k_m} \end{bmatrix}.$$

Since the matrix T is invertible its rows (which can be thought of as the vectors $p_{k_1}^T$, …, $(p_{k_1}A^{k_1-1})^T$, …, $(p_nA^{k_m-1})^T$) are linearly independent. Reading the rows of the matrix one by one and equating their coefficients on the left and on the right shows that:

- The first row of $\tilde{A}_{11}$ will have its second entry equal to 1 (the coefficient of $p_{k_1}A$), and all other entries equal to 0 (the coefficients of p_{k_1}, $p_{k_1}A^2$ etc.).
- The second row of $\tilde{A}_{11}$ will have its third entry equal to 1 (the coefficient of $p_{k_1}A^2$) and all other entries equal to 0, etc. until,
- row $k_1 - 1$ of $\tilde{A}_{11}$, which will have its last entry equal to 1 and all other entries equal to 0. Likewise,
- the first row of $\tilde{A}_{12}$ will have all its entries equal to 0 (coefficients of $p_{k_1+k_2}$, …, $p_{k_1+k_2}A^{k_2-1}$), etc. until,
- Row $k_1 - 1$ of $\tilde{A}_{1m}$, which will have all its entries equal to 0.

Row k_1 is equal to $p_{k_1}A^{k_1}$ and (since the rows of T are linearly independent and span $\mathbb{R}^{1\times n}$) can be written as a linear combination

$$p_{k_1}A^{k_1} = \tilde{a}_{k_1 1}p_{k_1} + \tilde{a}_{k_1 2}p_{k_1}A + \ldots \tilde{a}_{k_1 n}p_nA^{k_m-1}$$

for some $\tilde{a}_{k_1 1}$, …, $\tilde{a}_{k_1 n} \in \mathbb{R}$; we will not bother too much with the exact form of these coefficients.

Repeating the process for the remaining blocks shows that in the new coordinates the system matrices $\tilde{A}$ and $\tilde{B}$ become respectively

$$\begin{bmatrix} 0 & 1 & \dots & 0 & & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & & 0 & 0 & \dots & 0 \\ \tilde{a}_{k_1 1} & \tilde{a}_{k_1 2} & \dots & \tilde{a}_{k_1 k_1} & & \tilde{a}_{k_1(n-k_m+1)} & \tilde{a}_{k_1(n-k_m+2)} & \dots & \tilde{a}_{k_1 n} \\ & \vdots & & & \ddots & & \vdots & & \\ 0 & 0 & \dots & 0 & & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & & 0 & 0 & \dots & 1 \\ \tilde{a}_{n1} & \tilde{a}_{n2} & \dots & \tilde{a}_{nk_1} & & \tilde{a}_{n(n-k_m+1)} & \tilde{a}_{n(n-k_m+2)} & \dots & \tilde{a}_{nn} \end{bmatrix}, \quad \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \\ 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \\ 0 & \dots & 1 \end{bmatrix}.$$

This special form is known as the Brunowski normal form and it the analogue of the controllable canonical form for multi-input systems. We note, however, that, unlike the controllable canonical form, the Brunowski normal form is not unique, since it depends on the choice of columns of the matrix $\hat{P}$, the way these columns are rearranged, etc.

How does the Brunowski normal form help us to design controllers for the system? Assume that we are given a complex conjugate set of desired eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. As in the single input case we would like to select a linear state feedback gain matrix $K \in \mathbb{R}^{m \times n}$ so that under the feedback

$$u(t) = Kx(t) + r(t)$$

the closed loop matrix $A+BK$ has the desired eigenvalues. To see that this is possible first bring the system in Brunowski normal form through a coordinate transformation $\tilde{x}(t) = Tx(t)$ and determine the controllability indices k_1, ..., k_m. Write also the state feedback in the new coordinates

$$u(t) = KT^{-1}Tx(t) + r(t) = \tilde{K}\tilde{x}(t) + r(t)$$

and consider the rows of $\tilde{K}$

$$\tilde{K} = \begin{bmatrix} \tilde{K}_1 \\ \vdots \\ \tilde{K}_m \end{bmatrix} = \begin{bmatrix} \tilde{k}_{11} & \ldots & \tilde{k}_{1n} \\ \vdots & \ddots & \vdots \\ \tilde{k}_{m1} & \ldots & \tilde{k}_{mn} \end{bmatrix}.$$

Split the set of desired eigenvalues into m subsets $\{\lambda_1, \ldots, \lambda_{k_1}\}$, ..., $\{\lambda_{n-k_m+1}, \ldots, \lambda_n\}$. For simplicity assume that this can be done so that each of the subsets is itself complex conjugate. Form the m characteristic polynomials

$$\begin{aligned} (\lambda - \lambda_1)\ldots(\lambda - \lambda_{k_1}) &= \lambda^{k_1} + d_{11}\lambda^{k_1-1} + \ldots + d_{1k_1} \\ &\ldots \\ (\lambda - \lambda_{n-k_m+1})\ldots(\lambda - \lambda_n) &= \lambda^{k_m} + d_{m1}\lambda^{k_m-1} + \ldots + d_{mk_m}. \end{aligned}$$

The idea is to use each row $\tilde{K}_i$ of the feedback matrix to ensure that the subsystem $\tilde{A}_{ii}$ has the corresponding characteristic polynomial. To do this select

$$\begin{aligned} &\tilde{k}_{11} = -(\tilde{a}_{k_1 1} + d_{1k_1}), \ldots, \tilde{k}_{1k_1} = -(\tilde{a}_{k_1 k_1} + d_{11}), \ldots, \tilde{k}_{1(k_1+1)} = -\tilde{a}_{k_1(k_1+1)}, \ldots, \tilde{k}_{1n} = -\tilde{a}_{k_1 n} \\ &\quad \ldots \\ &\tilde{k}_{m1} = -\tilde{a}_{n1}, \ldots, \tilde{k}_{m(n-k_m)} = -\tilde{a}_{n(n-k_m)}, \ldots, \tilde{k}_{mn} = -(\tilde{a}_{k_1 n} + d_{m1}). \end{aligned}$$

The elements of the off-diagonal blocks $\tilde{A}_{ij}$ for $i \neq j$ are then eliminated and the resulting closed loop system matrix becomes

$$\tilde{A} + \tilde{B}\tilde{K} = \begin{bmatrix} 0 & 1 & \ldots & 0 & & & 0 & 0 & \ldots & 0 & & \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & \ldots & 1 & \ldots & & 0 & 0 & \ldots & 0 & & \\ -d_{1k_1} & -d_{1(k_1-1)} & \ldots & -d_{11} & & & 0 & 0 & \ldots & 0 & & \\ & \vdots & & & \ddots & & & \vdots & & & & \\ 0 & 0 & \ldots & 0 & & 0 & 1 & & \ldots & 0 & & \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & & \ddots & \vdots & & \\ 0 & 0 & \ldots & 0 & \ldots & 0 & 0 & & \ldots & 1 & & \\ 0 & 0 & \ldots & 0 & & -d_{mk_m} & d_{m(k_m-1)} & & \ldots & -d_{m1} & & \end{bmatrix}.$$

Notice that the matrix is block-diagonal and that each of the diagonal blocks is in controllable

canonical form. Therefore $\textsc{Det}[\lambda I - (\tilde{A} + \tilde{B}\tilde{K})]$ decomposes into

$$\begin{aligned}
&\textsc{Det}\begin{bmatrix} 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \\ d_{1k_1} & d_{1(k_1-1)} & \dots & d_{11} \end{bmatrix} \cdot \dots \cdot \textsc{Det}\begin{bmatrix} 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \\ d_{mk_m} & d_{m(k_m-1)} & \dots & d_{m1} \end{bmatrix} \\
&= (\lambda^{k_1} + d_{11}\lambda^{k_1-1} + \dots + d_{1k_1}) \dots (\lambda^{k_1} + d_{m1}\lambda^{k_1-1} + \dots + d_{mk_m}) \\
&= (\lambda - \lambda_1) \dots (\lambda - \lambda_{k_1}) \dots (\lambda - \lambda_{n-k_m+1}) \dots (\lambda - \lambda_n)
\end{aligned}$$

which is the desired characteristic polynomial. Setting $K = \tilde{K}T$ completes the design (recall that coordinate transformations do not affect the eigenvalues by Theorem 9.1).

As one can imagine from the above discussion, bringing a system into Brunowski normal form can be rather tedious. Fortunately this is not necessary if the end goal is to design feedback controllers for the system. As for single input systems, to do this it suffices to form the characteristic polynomial of the matrix $(A+BK)$, treating the elements of the feedback gain matrix $K \in \mathbb{R}^{m\times n}$ as unknowns. Equating the coefficients of this polynomial to those of the desired characteristic polynomial $(\lambda - \lambda_1) \dots (\lambda - \lambda_n)$ leads to a system of n equations with the nm elements of K as unknowns. The Brunowski normal form construction guarantees that if the system is controllable this system of equations will have a solution; for multi-input systems it will in fact have an infinite number of solutions, since the number of equations is smaller than the number of unknowns. If the system is not controllable but merely stabilizable then one has to make sure that the uncontrollable (but stable) modes are included in the set of desired eigenvalues and repeat the process.

In a similar way one can also construct observers and implement output feedback controllers for the system. The construction is just dual to that of the Brunowski normal form and will not be given in detail.

Problems for chapter 9

Problem 9.1 (Pole placement) Consider the system

$$\begin{aligned}
\dot{x}(t) &= \begin{bmatrix} 0 & 0 & 2 \\ 1 & 0 & 0 \\ 0 & 2 & 1 \end{bmatrix} x(t) + \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} u(t) = Ax(t) + Bu(t), \\
y(t) &= \begin{bmatrix} 0 & 0 & 2 \end{bmatrix} x(t) = Cx(t).
\end{aligned} \tag{$\square$}$$

1. Is the system observable? Is it controllable? Justify your answer in each case.
2. Design a state feedback $u = Kx$ such that the closed loop system has three poles at $s = -2$.
3. Recall that a state-observer provides an estimate $\hat{x}(t)$ of the state $x(t)$ of ($\square$) by means of the differential equation

$$\begin{aligned}
\dot{\hat{x}}(t) &= A\hat{x}(t) + Bu(t) + L\big(\hat{y}(t) - y(t)\big), \\
\hat{y}(t) &= C\hat{x}(t).
\end{aligned}$$

   Compute the observer gain L such that the dynamics of the error $e := x - \hat{x}$ have three poles at $s = -3$.

Problem 9.2 (Controllable canonical form) Use the procedure of Section 9.5 to re-derive the controllable canonical form for single input systems. Explain how the coefficients in the last row of matrix $\tilde{A}$ arise.

Problem 9.3 (Controllability, observability and state feedback) Consider the linear state feedback arrangement of Figure 9.1.

1. Show that the closed loop system is controllable if and only if the open loop system is controllable.
2. Assume now that the open loop system is observable. Is it true that the closed loop system will always be observable? Provide a proof, or a counter-example.

Problem 9.4 (Controllability, observability and observer feedback) Consider the observer feedback arrangement of Figure 9.3. Assume that the open loop system is observable and controllable. Will the closed loop system always be observable? Will it be controllable? In both cases provide either a proof, or a counter-example.

Appendix A

Notation

A.1 Shorthands

- Def. = Definition
- Thm. = Theorem
- Ex./ = Example or exercise
- iff = if and only if
- wrt = with respect to
- wlog = without loss of generality
- ftsoc = for the sake of contradiction
- $\therefore$ = therefore
- $\rightarrow\leftarrow$ = contradiction
- : = such that
- $\square$ = Q.E.D. (quod erat demonstrandum)

A.2 Sets

- $\in$, $\notin$, $\subseteq$, $\not\subseteq$, $\cap$, $\cup$, $\emptyset$
- For $A, B \subseteq X$, A^c stands for the complement of a set and $\setminus$ for set difference, i.e. $A^c = X \setminus A$ and $A \setminus B = A \cap B^c$.
- $\mathbb{R}$ real numbers, $\mathbb{R}_+$ non-negative real numbers, $[a, b]$, $[a, b)$, $(a, b]$, (a, b) intervals
- $\mathbb{Q}$ rational numbers
- $\mathbb{Z}$ integers, $\mathbb{N}$ non negative integers (natural numbers)
- $\mathbb{C}$ complex numbers, $\mathbb{C}_+$ complex numbers with non-negative real part
- Sets usually defined through their properties as in: For $a < b$ real

$$[a, b] = \{x \in \mathbb{R}^n \mid a \leq x \leq b\}, \text{ or } [a, b) = \{x \in \mathbb{R}^n \mid a \leq x < b\}, \text{ etc.}$$

- Cartesian products of sets: X, Y two set, $X \times Y$ is the set of ordered pairs (x, y) such that $x \in X$ and $y \in Y$.

Example ($\mathbb{R}^n$) $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$ (n times).

$$x \in \mathbb{R}^n \Rightarrow x = (x_1, x_2, \ldots, x_n) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \text{ with } x_1, \ldots, x_n \in \mathbb{R}$$

■

A.3 Logic

- $\Rightarrow$, $\Leftarrow$, $\Leftrightarrow$, $\exists$, $\forall$.
- $\exists!$ = exists unique.
- $\wedge$ = and
- $\vee$ = or
- $\neg$ = not

Exercise A.1 Is the statement $\neg(\exists!\ x \in \mathbb{R}\ :\ x^2 = 1)$ true or false? What about the statement $(\exists\, x \in \mathbb{R}\ :\ x^2 = -1)$?

Appendix B

Basic linear algebra

Appendix C

Basic calculus

Bibliography

[1] Panos J. Antsaklis and Anthony N. Michel. *Linear Systems*. Birkhäuser, Boston, MA, U.S.A., 2006.

[2] Panos J. Antsaklis and Anthony N. Michel. *Linear Systems Primer*. Birkhäuser, Boston, MA, U.S.A., 2007.

[3] Richard E. Bellman. *Dynamic programming*. Princeton University Press, Princeton, NJ, U.S.A., 1957.

[4] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Vol. 1, 3^{rd} ed., 2005, Vol. 2, 4^{th} ed., 2012.

[5] Roger W. Brockett. *Finite-Dimensional Linear Systems*. Academic Press, 1979.

[6] Frank M. Callier and Charles A. Desoer. *Linear System Theory, 2nd. ed.* Springer Science and Business Media, New York, NY, U.S.A., 1991.

[7] Phil Dyke. *An Introduction to Laplace Transforms and Fourier Series*. Springer-Verlag, London, U.K., 2001.

[8] A.F. Filippov. *Sliding Modes in Control and Optimization*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2010.

[9] Joao P. Hespanha. *Linear Systems Theory*. Princeton University Press, Princeton, NJ, U.S.A., 2009.

[10] Alberto Isidori. *Nonlinear Control Systems*. Springer-Verlag, London, U.K., 1995.

[11] Thomas Kailath. *Linear Systems*. Prentice Hall, Upper Saddle River, NJ, U.S.A., 1980.

[12] Hassan K. Khalil. *Nonlinear Control*. Prentice Hall, 2015.

[13] Daniel Liberzon. *Calculus of Variations and Optimal Control Theory*. Princeton University Press, Princeton, NJ, U.S.A., 2012.

[14] Henk Nijmeijer and Arjan van der Schaft. *Nonlinear Dynamical Control Systems*. Springer-Verlag, New York, NY, U.S.A., 1990.

[15] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidge, and E.F. Mishchenko. *The Mathematical Theory of Optimal Processes*. Wiley, 1962.

[16] Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Companies, Inc., 1976.

[17] S. Shankar Sastry. *Nonlinear Systems: Analysis Stability and Control*. Springer-Verlag, New York, NY, U.S.A., 1999.

[18] Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer-Verlag, New York, NY, U.S.A., 1998.

[19] Gilbert Strang. *Linear Algebra and Its Applications*. Brooks/Cole/Cengage, 4 edition, 2006.

[20] Vadim I. Utkin. *Sliding Modes in Control and Optimization*. Springer-Verlag, New York, NY, U.S.A., 1992.