Mathematics of Engineering Systems

Serdar Yüksel
Queen's University, Mathematics and Statistics

Lecture Notes

April 3, 2024

This document contains supplemental lecture notes which has been used primarily at Queen's University for the course *MTHE 335/MATH 335 Mathematics of Engineering Systems* offered in the Department of Mathematics and Statistics for the Mathematics and Engineering Program.

Contents

1

Introduction

1.1 Introduction

In a *differential equations* course, one studies quantitative and qualitative behaviours of solutions to differential equations. For such equations, under mild regularity conditions, a given initial condition (in the absence of disturbances) leads to a unique solution/output. One could generalize this to difference equations for discrete-time equations.

For such setups, we can view the solution as a map from a set of initial conditions to an appropriate set of solutions/outcomes such as a set of paths; e.g. continuous functions for the former (continuous-time setup), and discrete-time signals for the latter (discrete-time setup).

This solution map may be regarded as a *system* mapping the inputs to the outputs.

Such an approach has remarkable implications in engineering and applied mathematics. In many engineering or applied mathematics areas, one may also need to consider the presence of noise/disturbance terms (which are typically external/exogenous inputs) or one may also have the liberty to affect the solutions through introducing an external *control* term. Accordingly, one should view the aforementioned map to be from some set of initial states, some set of disturbances and some set of external inputs, to some output set.

Systems theory is concerned with rigorously studying, defining and analyzing, as well as shaping the input-output behaviour of such maps (which we will call *system*s). See Figure 1.1 for a generic depiction.

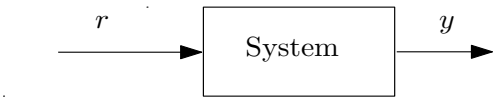

Fig. 1.1: A (control-free) system

In this course, we will study systems theory and through our development, we will also present a detailed analysis on signals spaces, representation of signals using signal bases, and their optimal approximations, and we will introduce some aspects of optimization. There are many applications that we will study in our course, which will primarily concern signal processing, communications, and control; but we will also find occasions to touch on many related applications involving signal spaces and systems design.

In the context of systems which are linear (a rigorous definition is to be given later; these systems are essentially linear functions from a linear space of inputs to a linear space of outputs), causal (where the output at any given time cannot depend on inputs occurring at later time stages) and time-invariant (where a time-delay in the input leads to an equivalent time delay in the output), we will see that the input-output relation admits efficient representational properties when the signals are expressed in terms of complex harmonics. This will motivate the Fourier Transform, and its generalizations (the Laplace Transform and the Z-transform), which will be studied in detail in our course.

To rigorously study the Dirac Delta function, which will let us define the impulse response and the frequency response of such systems, we will introduce distribution theory and the Schwartz space of signals.

1.2 Applications

1.2.1 Applications in Control Theory

In control systems, the goal is to shape the input-output behaviour by possibly utilizing feedback from system outputs (Figure 1.3 and Figure 1.2) under various design criteria and constraints. Commonly considered criteria are system stability (e.g. convergence to a point or a set with respect to initial state conditions, or boundedness of the output corresponding to any bounded input), reference tracking, robustness to incorrect models (unspecified system dynamics) and presence of disturbance (which may appear in the system itself or in the measurements available at the controller: that is, either a system noise or a measurement noise), and optimal control. Figure 1.5 depicts some of these considerations. A common, and one of the earliest modern examples, of control systems is the thermostat-based temperature control system depicted in Figure 1.4.

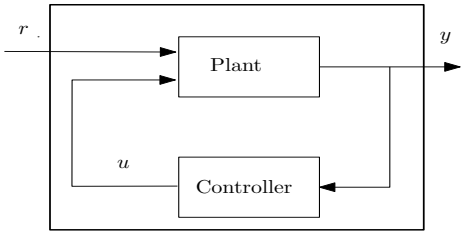

Fig. 1.2: A feedback control diagram; here r is the input and the controller depends on the output y directly (and not just through an error term as in Figure 1.3). The controller unit is designed to have the system output respond to r in a desirable fashion under various criteria that will be studied further.

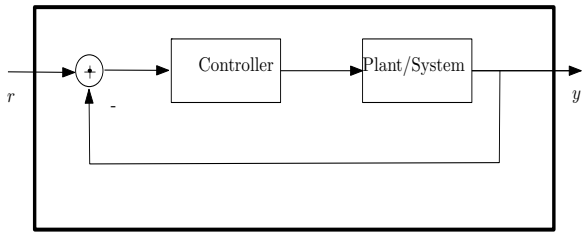

Fig. 1.3: A system shaped by error feedback control. Here, the control uses the feedback only through the error between an input signal, r, and the system output, y. This setup is a very common configuration.

Let us demonstrate some of the common design criteria, starting with a conceptually simple example. Consider

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

where $a \in \mathbb{R}_+$ is a scalar, $u(t)$ is the control input, that can be selected given the information $\{x(s), s < t\}$ and $n(t)$ is some disturbance/noise acting on the system. The disturbance is external, that is, the controller has nothing to do with it.

If $a > 0$, then in the absence of control, the solution to the system is given with

$$x(t) = e^{at}x(0) + \int_0^t e^{a(t-s)}n(s)ds$$

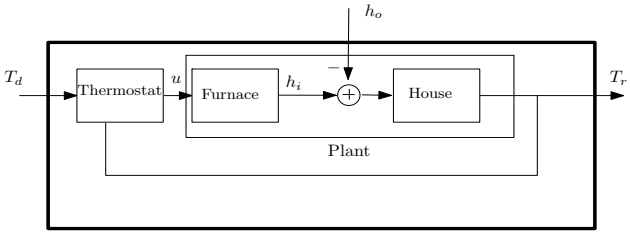

Fig. 1.4: A thermostat heating system is a popular example of the configuration given in Figure 1.3. Here T_d is a desired temperature, T_r is the (actual) room temperature, h_i is the heat from the furnace and h_o represents the heat leaked outside (or the cold air entering the house). Here, the thermostat is the controller which decides on whether the gas valve should be turned on or off based on the desired temperature and the sensed actual room temperature; u is the control input representing these turn-on or turn-off signals. The furnace is the *actuator* which maps the control input to the heat input, h_i, entering the system. House is the process or the system, whose output (the room temperature) is to be controlled. Often one lumps the actuator and the system (house in this case) as a single unit and calls it a *plant*. Thus, one typically considers a controller and a plant (to be controlled), in a control system together with the external inputs to the system (in this case: the desired temperature and the heat leakage). One could also consider a setup where the room temperature is recorded (by the sensor in the thermostat unit) with measurement error; in this case the measurement error/noise in the sensor should also be considered as an external input.

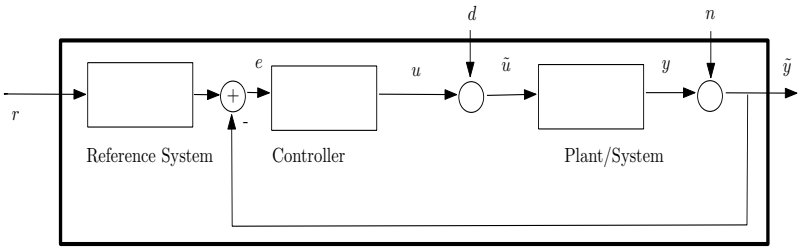

Fig. 1.5: A further control system flow diagram; here r is a reference and d and n are system/load and measurement disturbances, and are the inputs to the system. The controller unit is designed to have the system output respond to these inputs in a desirable fashion under various criteria that will be studied further in these notes.

In particular, if $n(s) = K \neq 0$ for some constant K, then even when $x(0) = 0$, we have that $\lim_{t\to\infty} |x(t)| = \infty$.

On the other hand, if we use the control input (using the feedback from the *state* of the system) $u(t) = -(a+1)x(t)$, we obtain the equation

$$\frac{dx}{dt} = -x(t) + n(t)$$

with the solution

$$x(t) = e^{-t}x(0) + \int_0^t e^{-(t-s)}n(s)ds,$$

which remains bounded if $\sup_s |n(s)| < \infty$. Thus, with control utilizing *feedback* we have achieved some notion of stability which will be termed as *bounded-input-bounded-output* stability. Such a setup can be represented by Figure 1.2, where the input is n, the output is $y = x$ and $u(t) = -(a+1)x(t)$.

The setup depicted in Figure 1.4 can be considered as a reference-tracking example, where the desired temperature process is the reference signal that the system output is designed to be tracking.

Control can also be used to steer the *state* of the system from some initial condition to some final condition. If the final condition is an equilibrium point, often this task is called *stabilization*. If the goal is to steer the state to some arbitrary point in the state space, the task is called *reachability* (from an initial state) or *controllability* (with respect to a final state). A related concept is *observability*, which is a crucial concept in particular when the information available at the controller with regard to the state is perturbed by some measurement noise.

1.2.2 Applications in Signal Processing Theory

Applications in signal processing has enabled much of modern technology. Two primary applications are in filter design, which allows for estimation and denoising, and sampling theories which allow for discrete-time processing of continuous-time signals.

Consider a signal x which is perturbed by noise w. A filter is a system which takes the noisy signal $y = x + w$ as input and provides a cleaner signal $\hat{x}$ as its output.

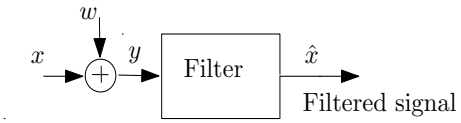

Fig. 1.6: Filtering of noisy signals. Here w is a noise, x is a signal to be reconstructed and $\hat{x}$ is the output of the reconstruction given the noisy input y.

Many systems in practice are continuous, but they need to be processed by computers, which inevitably have to work with discrete-time/discrete-space signals (as the ultimate language of transistors/chips are binary in terms of 0s and 1s). Therefore, one needs to first sample a continuous-time signal and work with such signals in discrete-time, before processing them, and interpolating them back to continuous-time.

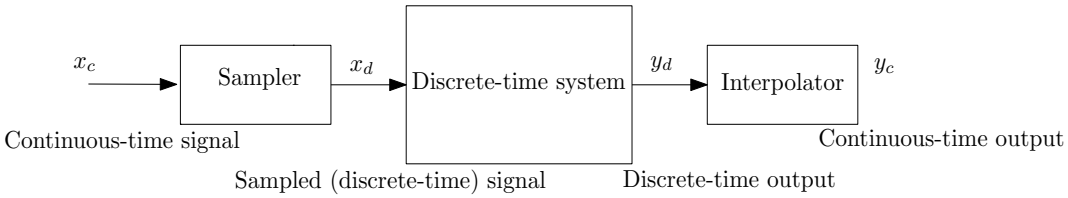

Fig. 1.7: Discrete-time processing of continous-time systems.

1.2.3 Applications in Communications and Information Theory

Modern engineering systems are typically highly interconnected with their environment which necessitates the presence of data-links between various components of a system. Typically such systems require finite representations of uncountable or large state space valued signals (in addition to discrete-time representation/approximation of continuous-time signals). These include, quantization, coding and decoding of signals over communication channels (with or without feedback). Each of the individual components, such as encoders, channels and decoders, may separately be viewed as systems, though typically by the term *communication system*, we will refer to the entire ensemble mapping the source symbol and stochastic noise (in the channel), to the decoder output.

Many other systems, however, operate in continuous-time. An example is classical analog radio communications, where signals are modulated to carrier signals with targeted frequency waves, transmitted over wireless or wired media, and demodulated upon reception by a decoder.

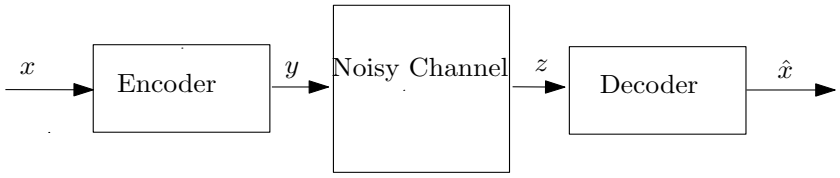

Fig. 1.8: Communication of a signal across a noisy channel.

1.3 Linearization

Even though the primary focus of these notes is on linear models, we will see that linearization of non-linear models, around a point of interest leads to a design method where a design based on the linear model achieves satisfactory performance for the non-linear system in a local sense to be studied later in the notes.

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable function. Let $g(x) = \begin{bmatrix} g^1(x) & \dots & g^m(x) \end{bmatrix}$ and $x \in \mathbb{R}^n$ be written as $x = \begin{bmatrix} x^1 & \dots & x^n \end{bmatrix}$. The Jacobian matrix of f at x, $J^x(g)$, is an $m \times n$ matrix function consisting of partial derivatives of g such that

$$J^x(g)(i,j) = \frac{\partial g^i}{\partial x^j}(x), \qquad i = 1, 2, \cdots, m; j = 1, 2, \cdots, n$$

Now, let $x \in \mathbb{R}^m$ and $u \in \mathbb{R}^p$ and

$$\frac{dx}{dt} = f(x, u)$$

be such that $f(\bar{x}, \bar{u}) = 0$. In this case, we say that x is at equilibrium at $\bar{x}$ with constant control input $\bar{u}$. Suppose that we slightly perturb x and u around the equilibrium $(\bar{x}, \bar{u})$. Let us write $x(t) = \bar{x} + \tilde{x}(t)$ and $u = \bar{u} + \tilde{u}(t)$, where $\tilde{x}$ and $\tilde{u}$ are small. Then,

$$\frac{d(\bar{x} + \tilde{x})}{dt} = f(\bar{x} + \tilde{x}, \bar{u} + \tilde{u})$$

Notice that $\frac{d(\bar{x}+\tilde{x})}{dt} = \frac{d\tilde{x}}{dt}$. If f is continuously differentiable, it follows that

$$f(\bar{x} + \tilde{x}, \bar{u} + \tilde{u}) \approx f(\bar{x}, \bar{u}) + J^f_x(\bar{x}, \bar{u})\tilde{x} + J^f_u(\bar{x}, \bar{u})\tilde{u},$$

where $J^f_x(\bar{x}, \bar{u})$ is the Jacobian of $f(\cdot, \bar{u}) : \mathbb{R}^m \to \mathbb{R}^m$ at fixed $\bar{u}$ and $J^f_u(\bar{x}, \bar{u})$ is the Jacobian of $f(\bar{x}, \cdot) : \mathbb{R}^p \to \mathbb{R}^m$ at fixed $\bar{x}$. Let

$$J^f_x(\bar{x}, \bar{u}) =: A, \qquad J^f_u(\bar{x}, \bar{u}) =: B,$$

we obtain

$$\frac{d\tilde{x}}{dt} = A\tilde{x} + B\tilde{u},$$

as an approximate linear description of the system at around the equilibrium point $(\bar{x}, \bar{u})$. We will observe that such a linearization is very useful in systems design.

Consider the following example involving an inverted (non-linear) pendulum over a cart system (see Figure 12.1), with masses of the pendulum and cart given with m and M, respectively. The goal is to keep the inverted pendulum (locally) stable around $\theta = 0$ by the control acting horizontally on the cart with mass M.

The non-linear mechanical/rotational dynamics equations can be derived as:

$$\begin{aligned} u &= M\frac{d^2y}{dt^2} + m\frac{d^2}{dt^2}(y + l\sin(\theta)) = M\frac{d^2y}{dt^2} + m\frac{d^2y}{dt^2} + ml\cos(\theta)\frac{d^2\theta}{dt^2} - ml(\frac{d\theta}{dt})^2\sin(\theta) \\ ml^2\frac{d^2\theta}{dt^2} &= mg\sin(\theta)l - m\frac{d^2y}{dt^2}\cos(\theta)l \end{aligned} \tag{1.1}$$

Around $\theta = 0$, $\frac{d\theta}{dt} = 0$, we apply the linear approximations $\sin(\theta) \approx \theta$ and $\cos(\theta) \approx 1$, and $(\frac{d\theta}{dt})^2 \approx 0$ to arrive at

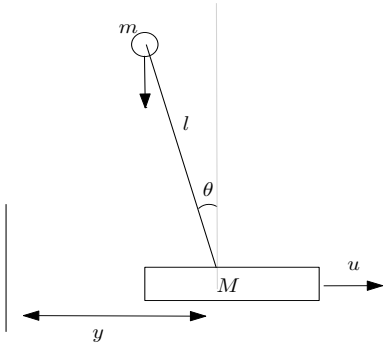

Fig. 1.9

$$\begin{aligned} M\frac{d^2y}{dt^2} &= u - (m\frac{d^2y}{dt^2} + ml\frac{d^2\theta}{dt^2}) \\ l\frac{d^2\theta}{dt^2} &= g\theta - \frac{d^2y}{dt^2} \end{aligned} \tag{1.2}$$

Finally, writing $x_1 = y, x_2 = \frac{dy}{dt}, x_3 = \theta, x_4 = \frac{d\theta}{dt}$, we arrive at the linear model in state space form

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{(M+m)g}{Ml} & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ \frac{-1}{Ml} \end{bmatrix} u,$$

where $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$.

A remarkable implication is the following: Using some systems theoretic analysis, it can be shown that linearization will let us construct a control function/policy/law that makes the idealized linear system stable, which in turn makes the original non-linear system locally stable around the (open-loop unstable) equilibrium point.

1.4 Mathematics of Systems

Given the introductory discussion presented, in the following we will first develop a rigorous study of a class of signal spaces which arise in systems theory and applications. We will then investigate signal expansions and approximations. This will also serve as an introduction to Fourier theory.

We will then study systems and their various regularity, structural and stability properties. We will, in this course, particularly focus on linear systems. A primary motivation, as we saw earlier in the previous section, is that many physical systems are either linear or locally almost linear (in the sense that a design based on a linear approximation leads to satisfactory performance).

Fourier theory occupies a dominant domain in linear systems theory: In the historical theory of systems (and control), one often reads about classical design and modern design: classical design is with regard to methods based on frequency-domain analysis of systems, and modern design (or state-space design) refers to methods based on time-domain analysis. In our course, we will discuss both approaches extensively. Fourier theory (and its generalizations via Laplace and Z-transforms) facilitate the frequency-domain analysis.

We will then study several applications in further detail. The course will lay the foundations for further study on the applications considered here, as well as those not studied here, in addition to many related areas in both engineering and applied mathematics.

2

Signal Spaces: Linear, Banach and Hilbert Spaces, and Basis Expansions

In this chapter, we present a general review of signal spaces.

2.1 Normed Linear (Vector) Spaces and Metric Spaces

Definition 2.1.1 *A linear (vector) space $\mathbb{X}$ is a space which is closed under addition and scalar multiplication: In particular, we define an addition operation $+$, and a scalar multiplication operation $\cdot$ such that*

$$+ : \mathbb{X} \times \mathbb{X} \to \mathbb{X}$$

$$\cdot : \mathbb{C} \times \mathbb{X} \to \mathbb{X}$$

with the following properties (we note that we may take the scalars to be either real or complex numbers). The following are satisfied for $x, y \in \mathbb{X}$ and α, β scalars:

(i) $x + y = y + x$

(ii) $(x + y) + z = x + (y + z)$.

(iii) $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$.

(iv) $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$.

(v) There is a null vector $\underline{0}$ such that $x + \underline{0} = x$.

(vi) $\alpha \cdot (\beta \cdot x) = (\alpha\beta) \cdot x$

(vii) $1 \cdot x = x$

(viii) For every $x \in \mathbb{X}$, there exists an element, called the (additive) inverse of x and denoted with $-x$ with the property $x + (-x) = \underline{0}$.

Example 2.1. (i) The space $\mathbb{R}^n$ with pointwise additional and scalar multiplication is a linear space. The null vector is $\underline{0} = (0, 0, \cdots, 0) \in \mathbb{R}^n$.
(ii) Consider the interval $[a, b]$. The collection of real-valued continuous functions on $[a, b]$, with pointwise addition and scalar multiplication is a linear space. The null element $\underline{0}$ is the function which is identically 0. This space is called the space of real-valued continuous functions on $[a, b]$
(iii) The set of all infinite sequences of real numbers having only a finite number of terms not equal to zero is a vector space. If one adds two such sequences, the sum also belongs to this space. This space is called the space of finitely many non-zero sequences.
(iv) The collection of all polynomial functions defined on an interval $[a, b]$ with complex coefficients forms a complex linear space. Note that the sum of polynomials is another polynomial.

Definition 2.1.2 *A non-empty subset M of a (real) linear vector space $\mathbb{X}$ is called a subspace of $\mathbb{X}$ if*

$$\alpha x + \beta y \in M, \quad \forall x, y \in M \quad \textit{and} \quad \alpha, \beta \in \mathbb{R}.$$

In particular, the null element $\underline{0}$ is an element of every subspace. For M, N two subspaces of a vector space $\mathbb{X}$, $M \cap N$ is also a subspace of $\mathbb{X}$.

Definition 2.1.3 *A normed linear space X is a linear vector space on which a map from X to $\mathbb{R}_+$, called its norm, is defined such that:*

- $||x|| \geq 0 \quad \forall x \in X, \quad ||x|| = 0$ *if and only if x is the null element (under addition and multiplication) of X.*
- $||x + y|| \leq ||x|| + ||y||$
- $||\alpha x|| = |\alpha| ||x||, \quad \forall \alpha \in \mathbb{R}, \quad \forall x \in X$

Definition 2.1.4 *In a normed linear space X, an infinite sequence of elements $\{x_n\}$ converges to an element x if the sequence $\{||x_n - x||\}$ converges to zero.*

Example 2.2. a) The space $C([a, b]; \mathbb{R})$ of continuous functions from $[a, b]$ to R with the norm $||x|| = \max_{\{a \leq t \leq b\}} |x(t)|$ is a normed linear space.

b) $l_p(\mathbb{Z}_+; \mathbb{R}) := \{x \in \Gamma(\mathbb{Z}_+; \mathbb{R}) : ||x||_p = \left(\sum_{i \in \mathbb{Z}_+} |x(i)|^p \right)^{\frac{1}{p}} < \infty\}$ is a normed linear space for all $1 \leq p < \infty$. c) Recall that if S is a set of real numbers bounded above, then there is a smallest real number y such that $x \leq y$ for all $x \in S$. The number y is called the *least upper bound* or *supremum* of S. If S is not bounded from above, then the supremum is ∞. In view of this, for $p = \infty$, we define

$$l_\infty(\mathbb{Z}_+; \mathbb{R}) := \{x \in \Gamma(\mathbb{Z}_+; \mathbb{R}) : ||x||_\infty = \sup_{i \in \mathbb{Z}_+} |x(i)| < \infty\}$$

d) $L_p([a, b]; \mathbb{R}) = \{\{x \in \Gamma([a, b]; \mathbb{R}) : ||x||_p = \left(\int_a^b |x(i)|^p \right)^{\frac{1}{p}} < \infty\}$ is a normed linear space. For $p = \infty$, we typically write: $L_\infty([a, b]; \mathbb{R}) := \{x \in \Gamma([a, b]; \mathbb{R}) : ||x||_\infty = \sup_{t \in [a,b]} |x(t)| < \infty\}$. However, for $1 \leq p < \infty$, to satisfy the condition that $||x||_p = 0$ implies that $x(t) = 0$, we need to assume that functions which are equal to zero almost everywhere are equivalent; for $p = \infty$ the definition is often revised with essential supremum instead of supremum so that

$$||x||_\infty = \inf_{y : y(t) = x(t) \text{a.e.}} \sup_{t \in [a,b]} |y(t)|$$

This subtle difference needs to be made explicit in some applications.

To show that l_p defined above is a normed linear space, we need to show that $||x + y||_p \leq ||x||_p + ||y||_p$.

Theorem 2.1.1 (Minkowski's Inequality) *For $1 \leq p \leq \infty$*

$$||x + y||_p \leq ||x||_p + ||y||_p$$

See Exercise 2.4.2, which also studies a proof of Hölder's Inequality, that is critically used in the proof of Minkowski's inequality:

Theorem 2.1.2 (Hölder's Inequality)

$$\sum x(k) y(k) \leq ||x||_p ||y||_q,$$

with $1/p + 1/q = 1$ and $1 \leq p, q \leq \infty$.

Definition 2.1.5 *A metric defined on a set X, is a function $d : X \times X \to \mathbb{R}$ such that:*

- $d(x, y) \geq 0, \quad \forall x, y \in X$ *and* $d(x, y) = 0$ *if and only if* $x = y$.
- $d(x, y) = d(y, x), \quad \forall x, y \in X$.
- $d(x, y) \leq d(x, z) + d(z, y), \quad \forall x, y, z \in X$.

Definition 2.1.6 *A metric space (X, d) is a set equipped with a metric d.*

A normed linear space is also a metric space, with metric

$$d(x, y) = ||x - y||.$$

Definition 2.1.7 *Let X and Y be two normed linear spaces, and let $B \subset X$ be a subset of X. A law (rule, relation) T which relates with every element of B an element in Y, is called a transformation from X to Y with domain B. The relation is often expressed as $x \mapsto y = T(x)$.*

If for every $y \in Y$ there is an x such that $y = T(x)$, the transformation is said to be *onto* (or *surjective*). If for every element of Y, there is at most one x such that $y = T(x)$, the transformation is said to be *one-to-one* (or *injective*). If these two properties hold simultaneously, the transformation is said to be *bijective*.

Definition 2.1.8 *A transformation $T : X \to Y$ (or $T \in \Gamma(X; Y)$) is linear if for every $x_1, x_2 \in X$ and $\alpha_1, \alpha_2 \in \mathbb{R}$, we have $T(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 T(x_1) + \alpha_2 T(x_2)$.*

Definition 2.1.9 *A transformation $T : X \to Y$ for normed linear spaces X, Y is continuous at $x_0 \in X$, if for every $\epsilon > 0, \exists \delta > 0$ such that $||x - x_0|| \leq \delta$ implies that $||T(x) - T(x_0)|| \leq \epsilon$ (Here the norms depend on the vector spaces X and Y). T is said to be continuous, if it is continuous at every $x_0 \in X$.*

Definition 2.1.10 *A transformation $T : X \to Y$ is sequentially continuous at $x_0 \in X$, if $x_n \to x$ implies that $T(x_n) \to T(y)$.*

Theorem 2.1.3 *Sequential continuity and continuity are equivalent for normed linear spaces.*

Theorem 2.1.4 *If the transformation T is a linear one, then continuity is equivalent to being continuous at the null element.*

For some applications, sequential continuity may be more convenient to work with as one may not need to quantify ϵ, δ pairs to verify continuity.

An important class of normed spaces that is widely used in optimization and engineering problems are Banach spaces:

Definition 2.1.11 *A sequence $\{x_n\}$ in a normed space X is Cauchy if for every $\epsilon > 0$, there exists an N such that $||x_n - x_m|| \leq \epsilon$, for all $n, m \geq N$.*

An important observation on Cauchy sequences is that every converging sequence is Cauchy, however, not all Cauchy sequences are convergent: This is because the limit might not live in the original space where the sequence elements take values in. This motivates the property of completeness:

Definition 2.1.12 *A normed linear space X is complete, if every Cauchy sequence in X has a limit in X. A complete normed linear space is called Banach.*

Banach spaces are important for many reasons including the following one: In many mathematical applications (such as existence of and numerical methods for solutions to differential equations), machine learning problems (such as iterative updates of data driven dynamics), stochastic analysis, or optimization problems (for which a sequence of approximating solutions may be obtained), sometimes we would like to see if a given sequence converges, without knowing what the limit of the sequence may be. Banach spaces allow us to use Cauchy sequence arguments to claim the existence of limits and some of their properties.

An example is the following: consider the solutions to the equation $Ax = b$ for A a square matrix and b a vector. One can identify conditions on an iteration of the form $x_{k+1} = (I - A)x_k + b$ to form a Cauchy sequence and converge to a solution through the *contraction principle*. As noted above, existence of solutions to ordinary differential equations also follow from Cauchy sequence arguments.

In applications, we will also discuss completeness of a subset. A subset of a Banach space X is complete if and only if it is closed. If it is not closed, one can provide a counterexample sequence which does not converge. If the set is closed, every Cauchy sequence in this set has a limit in X and this limit should be a member of this set, hence the set is complete.

The real space $\mathbb{R}$ is a complete space.

Theorem 2.1.5 $l_p(\mathbb{Z}_+;\mathbb{R}) := \{x \in \Gamma(\mathbb{Z}_+;\mathbb{R}) : \|x\|_p = \left(\sum_{i\in\mathbb{N}_+} |x(i)|^p\right)^{\frac{1}{p}} < \infty\}$ *is a Banach space for all* $1 \leq p \leq \infty$.

Proof. (i) Let $\{x_n\}$ be Cauchy. This implies that for every $\epsilon > 0$, $\exists N$ such that for all $n, m \geq N$ $\|x_n - x_m\|_p \leq \epsilon$. This also implies that for all $n > N$, $\|x_n\|_p \leq \|x_N\|_p + \epsilon$. Now let us denote x_n by the vector $\{x_1^n, x_2^n, x_3^n \dots, \}$. It follows that for every k the sequence $\{x_k^n\}$ is also Cauchy. Since $x_k^n \in \mathbb{R}$, and $\mathbb{R}$ is complete, $x_k^n \to x_k$ for some x_k. Thus, the sequence x_n *pointwise* converges to some vector x_*.

(ii) Is $x_* \in l_p(\mathbb{Z}_+;\mathbb{R})$? Define

$$x_{n,K} = \{x_1^n, x_2^n, \dots, x_{K-1}^n, x_K^n, 0, 0, \dots\},$$

that is, the vector which truncates after the Kth coordinate. Now, it follows that

$$\|x_{n,K}\|_p \leq \|x_N\|_p + \epsilon,$$

for every $n \geq N$ and K and

$$\lim_{n\to\infty} \|x_{n,K}\|_p^p = \lim_{n\to\infty} \sum_{i=1}^{K} |x_i^n|^p = \sum_{i=1}^{K} |x_i|^p,$$

since there are only finitely many elements in the summation. We have

$$\|x_{n,K}\|_p \leq \|x_N\|_p + \epsilon,$$

and thus

$$\lim_{n\to\infty} \|x_{n,K}\|_p = \|x_{*,K}\|_p \leq \|x_N\|_p + \epsilon,$$

Let us take another limit, by the monotone convergence theorem (recall that this theorem states that a monotonically increasing sequence which is bounded has a limit).

$$\lim_{K\to\infty} \|x_{*,K}\|_p^p = \lim_{K\to\infty} \sum_{i=1}^{K} |x_i|^p = \|x_*\|_p^p \leq (\|x_N\|_p + \epsilon)^p.$$

(iii) The final question is: Does $\|x_n - x_*\|_p \to 0$? Since the sequence is Cauchy, it follows that for $n, m \geq N$

$$\|x_n - x_m\|_p \leq \epsilon$$

Thus, for every $K \in \mathbb{N}$, $\|x_{n,K} - x_{m,K}\|_p \leq \epsilon$, and since K is finite

$$\lim_{m\to\infty} \|x_{n,K} - x_{m,K}\|_p = \|x_{n,K} - x_{*,K}\|_p \leq \epsilon$$

Now, we take another limit

$$\lim_{K\to\infty} \|x_{n,K} - x_{*,K}\|_p \leq \epsilon$$

By the monotone convergence theorem again,

$$\lim_{K\to\infty} \|x_{n,K} - x_{*,K}\|_p = \|x_n - x\|_p \leq \epsilon$$

Hence, $\|x_n - x\|_p \to 0$. □

The above spaces are also denoted $l_p(\mathbb{Z}_+)$, when the range space is clear from context.

Theorem 2.1.6 *The space of bounded functions* $\{x : [0,1] \to \mathbb{R}, \sup_{t\in[0,1]} |x(t)| < \infty\}$ *is a Banach space.*

The above space is often denoted by $L_\infty([0,1];\mathbb{R})$ or $L_\infty([0,1])$.

Remark 2.3. A brief remark on notations: When the range space is $\mathbb{R}$, the notation $l_p(\Omega)$ denotes $l_p(\Omega;\mathbb{R})$ for a discrete-time index set Ω and likewise for a continuous-time index set Ω, $L_p(\Omega)$ denotes $L_p(\Omega;\mathbb{R})$.

2.2 Hilbert Spaces

We first define pre-Hilbert spaces.

Definition 2.2.1 *A pre-Hilbert space X is a linear vector space where an inner product is defined on* $X \times X$. *Corresponding to each pair* $x, y \in X$ *the inner product* $\langle x, y\rangle$ *is a scalar (that is real-valued or complex-valued). The inner product satisfies the following axioms:*

1. $\langle x, y\rangle = \langle y, x\rangle^*$ *(the superscript denotes the complex conjugate) (we will also use* $\overline{\langle y, x\rangle}$ *to denote the complex conjugate)*
2. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
3. $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$
4. $\langle x, x\rangle \geq 0$*, equals* 0 *iff* x *is the null element.*

The following is a crucial result in such a space, known as the Cauchy-Schwarz inequality.

Theorem 2.2.1 *For* $x, y \in X$,

$$\langle x, y\rangle \leq \sqrt{\langle x, x\rangle}\sqrt{\langle y, y\rangle},$$

where equality occurs if and only if $x = \alpha y$ *for some scalar* α.

Exercise 2.2.1 *In a pre-Hilbert space* $\langle x, x\rangle$ *defines a norm:* $||x|| = \sqrt{\langle x, x\rangle}$

The proof for the result requires one to show that $\sqrt{\langle x, x\rangle}$ satisfies the triangle inequality, that is

$$||x + y|| \leq ||x|| + ||y||,$$

which can be proven by an application of the Cauchy-Schwarz inequality.

Not all spaces admit an inner product. In particular, however, $l_2(\mathbb{N}_+;\mathbb{R})$ admits an inner product with $\langle x,y\rangle = \sum_{t\in\mathbb{N}_+} x(n)y(n)$ for $x,y \in l_2(\mathbb{N}_+;\mathbb{R})$. Furthermore, $||x|| = \sqrt{\langle x,x\rangle}$ defines a norm in $l_2(\mathbb{N}_+;\mathbb{R})$.

The inner product, in the special case of $\mathbb{R}^N$, is the usual inner vector product; hence $\mathbb{R}^N$ is a pre-Hilbert space with the usual inner-product.

Definition 2.2.2 *A complete pre-Hilbert space, is called a Hilbert space.*

Hence, a Hilbert space is a Banach space, endowed with an inner product, which induces its norm.

Proposition 2.2.1 *The inner product is continuous: if $x_n \to x$, and $y_n \to y$, then $\langle x_n, y_n\rangle \to \langle x,y\rangle$ for x_n, y_n in a Hilbert space.*

2.2.1 Why are we interested in Hilbert Spaces?

Hilbert spaces allow us to do the following:

1. We can state a Projection Theorem, and hence a notion of optimality by providing a definition for orthogonality. The geometric insights presented carry over to more general optimization problems.
2. If a Hilbert space is separable (to be defined shortly), there exists a countably (or sometimes only finitely) many sequence of orthonormal vectors which can be used as basis to represent all the members in this space.
3. A Hilbert space formulation allows us to develop approximations of signals using a finite number of basis signals. This is used in many practical expansions, such as the Fourier expansion among others.

Proposition 2.2.2 *In a Hilbert space X, two vectors $x,y \in X$ are orthogonal if $\langle x,y\rangle = 0$. A vector x is orthogonal to a set $S \subset X$ if $\langle x,y\rangle = 0 \quad \forall y \in S$.*

A set X is closed if it contains every limit of any converging sequence taking values in X.

Theorem 2.2.2 (Projection Theorem) *Let H be a Hilbert space and B a subspace of H. Consider the problem:*

$$\inf_{m\in B} \|x-m\| \tag{2.1}$$

(i) A necessary and sufficient condition for $m^* \in B$ to be the minimizing element in B so that

$$\inf_{m\in B} \|x-m\| = \|x-m^*\| \tag{2.2}$$

is that, $x-m^*$ be orthogonal B; that is

$$\|x-m^*\| \le \|x-y\|, \qquad \forall y \in B.$$

If exists, such an m^* is unique.

(ii) Let H be a Hilbert space and B a closed subspace of H. For any vector $x \in H$, there is a unique vector $m^* \in B$ satisfying (2.2).

Proof. For (i), suppose that $m_0 \in B$ is such that $\exists m \in B$ with $\langle x-m_0, m\rangle > 0$. Without any loss, take $\|m\| = 1$ and $\langle x-m_0, m\rangle = \delta$. We can show that with $m_1 = m_0 + \delta m$, we will have $\|x-m_1\|^2 = \|x-m_0\|^2 - \delta^2 < \|x-m_0\|^2$, and thus m_0 cannot be a minimizer of (2.1).

On the other hand, if $\langle x - m^*, M\rangle = 0$, we have that for any $m \in B$ $\|x - m\|^2 = \|x - m^* + (m^* - m)\|^2 = \|x - m^*\|^2 + \|m - m^*\|^2 \geq \|x - m^*\|^2$, and thus m^* is indeed the minimizer over all $m \in B$. Uniqueness of such a minimizer also follows from this argument since if there were another minimizer $\bar{m} \neq m^*$, we would have

$$\|x - \bar{m}\|^2 = \|x - m^*\|^2 + \|\bar{m} - m^*\|^2 > \|x - m^*\|^2,$$

a contradiction.

For (ii), let

$$\delta = \inf_{m \in M} \|x - m\| \tag{2.3}$$

Let $\{m_k, k \in \mathbb{N}\}$ be so that $\{\|x - m_k\|\} \to \delta$. Observe that

$$\begin{aligned}&\langle x - m_k + x - m_n, x - m_k + x - m_n\rangle + \langle x - m_k - (x - m_n), x - m_k - (x - m_n)\rangle \\ &= 2\langle x - m_k, x - m_k\rangle + 2\langle x - m_n, x - m_n\rangle\end{aligned} \tag{2.4}$$

Write

$$\begin{aligned}&\langle x - m_k + x - m_n, x - m_k + x - m_n\rangle = \langle 2(x - \frac{m_k + m_n}{2}, 2(x - \frac{m_k + m_n}{2})\rangle \\ &= 4\langle x - \frac{m_k + m_n}{2}, x - \frac{m_k + m_n}{2}\rangle\end{aligned} \tag{2.5}$$

Since $\frac{m_k + m_n}{2} \in B$, by (2.3),

$$\langle x - m_k + x - m_n, x - m_k + x - m_n\rangle \geq \delta,$$

we have that

$$\|m_k - m_n\|^2 \leq 2\|x - m_n\|^2 + 2\|x - m_k\|^2 - 4\delta^2$$

As a result, as $|x - m_n\|^2 \to \delta^2$, we have that m_k is Cauchy. Since M is closed, it has a limit; call the limit $\tilde{m}$. We claim that the limit is optimal and hence is m^*: Consider the difference:

$$\langle x - m_n, x - m_n\rangle - \langle x - \tilde{m}, x - \tilde{m}\rangle$$

We claim that the difference goes to zero. Indeed,

$$\begin{aligned}&|\langle x - m_n, x - m_n\rangle - \langle x - \tilde{m}, x - \tilde{m}\rangle| \\ &= |\langle x - m_n, x - m_n\rangle - \langle x - m_n, x - \tilde{m}\rangle \\ &\qquad + \langle x - m_n, x - \tilde{m}\rangle - \langle x - \tilde{m}, x - \tilde{m}|\rangle \\ &\leq |\langle x - m_n, x - m_n\rangle - \langle x - m_n, x - \tilde{m}\rangle| \\ &\qquad + |\langle x - m_n, x - \tilde{m}\rangle - \langle x - \tilde{m}, x - \tilde{m}|\rangle \\ &= |\langle x - m_n, m_n - \tilde{m}\rangle| \\ &\qquad + |\langle m_n - \tilde{m}, x - \tilde{m}|\rangle \\ &\leq \|x - m_n\|\|m_n - \tilde{m}\| + \|m_n - \tilde{m}\|\|x - \tilde{m}\|\end{aligned} \tag{2.6}$$

where the final inequality is due to Cauchy-Schwarz. Now, since $\|x - m_n\| \to \delta$, we have that $\|x - m_n\|$ is bounded. Finally, as $\|m_n - \tilde{m}\| \to 0$, both terms in the final line (2.6) go to zero and we conclude that

$$\delta = \lim_{n \to \infty} \|x - m_n\| = \|x - \tilde{m}\|$$

And therefore $\tilde{m}$ is a minimizing vector. By part (i), this has to be the only minimizing vector in B.

□

We will see applications of the projection theorem during the semester while studying optimal filter design and signal representation.

2.3 Approximations and Signal Expansions

2.3.1 Orthogonality

Definition 2.3.1 *A set of vectors in a Hilbert space S is orthogonal if all elements of this set are orthogonal to each other. The set is orthonormal if each vector in this set has norm equal to one.*

The Gram-Schmidt orthogonalization procedure can be invoked to generate a set of orthonormal sequences. This procedure states that given a sequence $\{x_i\}$ is linearly independent vectors, there exists an orthonormal sequence of vectors $\{e_i\}$ such that for every $x_k, \alpha_k, 1 \leq k \leq n$, there exists $\beta_k, 1 \leq k \leq n$ with

$$\sum_{k=1}^{n} \alpha_k x_k = \sum_{k=1}^{n} \beta_k e_k,$$

that is the linear span of $\{x_k, 1 \leq k \leq n\}$ is equal to the linear span of $\{e_k, 1 \leq k \leq n\}$ for every $n \in \mathbb{N}$.

Recall that a set of vectors $\{e_i\}$ is linearly dependent if there exists a complex-valued vector $\mathbf{c} = \{c_1, c_2, \ldots, c_N\}$ such that $\sum_i^N c_i e_i = 0$ with at least one coefficient $c_i \neq 0$. The following is a simple exercise.

Proposition 2.3.1 *A sequence of orthonormal vectors is a linearly independent collection.*

We say that a sequence $\sum_{i=1}^{n} \epsilon_i e_i$ converges to x, if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $||x - \sum_{i=1}^{n} \epsilon_i e_i|| < \epsilon$, for all $n \geq N$.

One of the important results while studying Hilbert spaces is the following:

Theorem 2.3.1 *Let $\{e_i\}$ be a sequence of orthonormal vectors in a Hilbert space H. Let $\{x_n = \sum_{i=1}^{n} \epsilon_i e_i\}$ be a sequence of vectors in H. The sequence converges to a vector x if and only if*

$$\sum_{i=1}^{\infty} |\epsilon_i|^2 < \infty.$$

In this case $\langle x, e_i \rangle = \epsilon_i$.

2.3.2 Separable Hilbert Spaces and Countable Expansions

Definition 2.3.2 *Given a normed linear space X, a subset $D \subset X$ is dense in X, if for every $x \in X$, and each $\epsilon > 0$, there exists a member $d \in D$ such that $||x - d|| \leq \epsilon$.*

Definition 2.3.3 *A set is countable if every element of the set can be associated with an integer via an ordered mapping.*

Examples of countable spaces are finite sets and the set $\mathbb{Q}$ of rational numbers. An example of uncountable sets is the set $\mathbb{R}$ of real numbers.

The following was proven in class:

Theorem 2.3.2 *(a) A countable union of countable sets is countable.*

(b) A finite Cartesian product of countable sets is countable.

(c) Infinite Cartesian products of countable sets may not be countable. The same holds if each of the sets is even finite.

(d) $[0,1]$ *is not countable.*

Cantor's diagonal argument and the triangular enumeration are important steps in proving the theorem above.

Definition 2.3.4 *A space* X *is separable, if it contains a countable dense set.*

Separability states that it suffices to work with a countable set, when a set is uncountable, for computational purposes. Examples of separable sets are $\mathbb{R}$, and the set of continuous and bounded functions on a compact set metrized with the maximum distance between the functions.

Theorem 2.3.3 *Let* H *be a separable Hilbert space. Then, every orthonormal system of vectors in* H *has a finite or countably infinite number of elements.*

Proof. Let $D = \{x_1, x_2, \cdots, \}$ be a countable set dense in H. Let $\{e_\alpha\}$ be a set of orthonormal vectors. Observe that, for $\alpha \neq \beta$, $\|e_\alpha - e_\beta\|^2 = \|e_\alpha\|^2 + \|e_\beta\|^2 = 2$ and hence $\|e_\alpha - e_\beta\| = \sqrt{2}$.

By denseness of D, for every e_α, there exists an element x_{k_α} such that $\|e_\alpha - x_{k_\alpha}\| < \frac{1}{\sqrt{2}}$. This implies that, for any other e_β, we have, by the relation, $\|x\| - \|y\| \leq \|x - y\|$,

$$\|e_\beta - e_\alpha\| - \|(e_\alpha - x_{k_\alpha}\| \leq \|e_\beta - e_\alpha - \bigg((e_\alpha - x_{k_\alpha}\bigg)\| = \|e_\beta - x_{k_\alpha}\|,$$

and thus

$$\sqrt{2} - \frac{1}{\sqrt{2}} \leq \|e_\beta - x_{k_\alpha}\|$$

Therefore, for every e_α there is one and only one x_{α_k} which is strictly inside a distance of $\frac{1}{\sqrt{2}}$. Thus, we can associate with every element in $\{e_\alpha\}$ and unique element in the countable set D. Thus, $\{e_\alpha\}$ is countable. □

Definition 2.3.5 *An orthonormal sequence in a Hilbert space* H *is complete if the only vector in* H *which is orthogonal to each of the vectors in the sequence is the null vector.*

Theorem 2.3.4 *A Hilbert space* H *contains a complete orthonormal sequence (that is, a countable collection of such vectors) if and only if it is separable.*

Proof. (i) Let H be separable. Then, there exists a countable dense subset $D = \{x_1, x_2, \cdots, \}$. Apply the Gram-Schmidt procedure to obtain $\{e_1, e_2, \cdots, \}$, an orthonormal collection. We claim that this set is a complete orthonormal sequence. Suppose not; that is, let $h \in H$ be so that $\|h\| \neq 0$ and yet $\langle h, e_k \rangle = 0$ for every $k \in \mathbb{N}$. Now, for every $\epsilon > 0$, there exists $x_m \in D$ with $\|h - x_m\| \leq \epsilon$ and observe that $x_m = \sum_{k=1}^{m} \alpha_i e_i$ since the span of the vectors $\{e_1, e_2, \cdots, e_m\}$ contains x_m. Then,

$$\|h\|^2 = \langle h, h \rangle = \langle h - \sum_{k=1}^{m} \alpha_i e_i, h \rangle = \langle h - x_m, h \rangle \leq \|h - x_m\| \|h\| \leq \epsilon \|h\|,$$

which implies that $\|h\| \leq \epsilon$. Since $\epsilon > 0$ is arbitrary; this completes the proof that $\|h\| = 0$ and h is the null element.

(ii) Now, let H have a complete orthonormal sequence $\{e_1, e_2, \cdots\}$. We will show that

$$D = \bigcup_{n \in \mathbb{N}} \{x \in H : x = \sum_{k=1}^{n} \alpha_i e_i, \alpha_i \in \mathbb{Q}\},$$

is a countable dense subset in H, and this H is separable. That D is separable follows from the fact that for every n, the set $\{x \in H : x = \sum_{k=1}^{n} \alpha_i e_i, \alpha_i \in \mathbb{Q}\}$ is countable as a Cartesian product of finitely many countable sets, and thus the countable union over $n \in \mathbb{N}$ leads to a countable set. We now show that this set is dense in H.

Consider the vector $\sum_{i=1}^{\infty}\langle x, e_i\rangle e_i$. Let $h = x - \sum_{i=1}^{\infty}\langle x, e_i\rangle e_i$. We claim that $\|h\| = 0$. For any $m \in \mathbb{N}$,

$$\begin{aligned}
\langle h, e_m\rangle &= \langle x - \sum_{i=1}^{\infty}\langle x, e_i\rangle e_i, e_m\rangle \\
&= \langle x, e_m\rangle - \langle \lim_{n\to\infty}\sum_{i=1}^{n}\langle x, e_i\rangle e_i, e_m\rangle \\
&= \langle x, e_m\rangle - \lim_{n\to\infty}\langle \sum_{i=1}^{n}\langle x, e_i\rangle e_i, e_m\rangle \\
&= \langle x, e_m\rangle - \langle x, e_m\rangle = 0. \qquad (2.7)
\end{aligned}$$

But since $\{e_1, e_2, \cdots\}$ is a complete orthonormal sequence, it must be that $\|h\| = 0$. Hence, $x = \sum_{i=1}^{\infty}\langle x, e_i\rangle e_i$. Now approximate this vector, by first truncating the sum so that for any $\epsilon > 0$,

$$\|x - \sum_{i=1}^{N}\langle x, e_i\rangle e_i\| \leq \epsilon/2,$$

and then approximating the coefficients by rational numbers so that $\|\sum_{i=1}^{N}\langle x, e_i\rangle e_i - \sum_{i=1}^{N}\alpha_i e_i\| \leq \epsilon/2$ with $\alpha_i \in \mathbb{Q}$. This implies that

$$\begin{aligned}
&\|x - \sum_{i=1}^{N}\alpha_i e_i\| \\
&\leq \|x - \sum_{i=1}^{N}\langle x, e_i\rangle e_i\| + \|\sum_{i=1}^{N}\langle x, e_i\rangle e_i - \sum_{i=1}^{N}\alpha_i e_i\| \\
&\leq \epsilon \qquad (2.8)
\end{aligned}$$

Since for every ϵ such an approximation can be made by some element in D, this completes the proof. □

The proof above also showed that in a Hilbert space H, a complete orthonormal sequence e_n defines a *basis* so that for any $x \in H$, we have

$$x = \lim_{N\to\infty}\sum_{i=1}^{N}\langle x, e_i\rangle e_i$$

2.3.3 Separability of l_2 and L_2 spaces

In view of Theorem 2.3.4, the following result builds on the fact that the sequence of orthonormal vectors

$$\left\{e_n, n \in \mathbb{N} : \quad e_n : \mathbb{Z}_+ \to \mathbb{R}, \quad e_n(m) = 1_{\{m=n\}}, \quad m \in \mathbb{Z}_+\right\}$$

is a countable complete orthonormal set in $l_2(\mathbb{Z}_+; \mathbb{R})$: Note that for any $h = \{h(1), h(2), \cdots, \} \in l_2(\mathbb{Z}_+; \mathbb{R})$, $\langle h, e_n\rangle = h(n)$ and hence for any vector $v \in l_2(\mathbb{Z}_+; \mathbb{R})$

$$\langle v, e_n\rangle = 0 \quad \forall n \in \mathbb{Z}_+ \implies \|v\| = 0.$$

Theorem 2.3.5 *The Hilbert space $l_2(\mathbb{Z}_+;\mathbb{R})$ with inner product*

$$\langle h_1, h_2\rangle = \sum_{n\in\mathbb{Z}_+} h_1(n)h_2(n),$$

is separable.

Next, we will show that $L_2([a,b];\mathbb{R})$ is separable for $a,b\in\mathbb{R}$. To establish this result, we will review some useful facts. In the following, we will need to use some basic properties of Lebesgue integration; the reader may find Appendix A useful. The analysis regarding the proof of this result is optional for our course.

Theorem 2.3.6 (Bernstein-Weierstrass' Theorem) *Any function in $C([0,1];\mathbb{R})$ can be approximated arbitrarily well by a polynomial under the supremum norm.*

Proof. This can be proven by construction. Define for some $f\in C([0,1])$ and $n\in\mathbb{N}$ the Bernstein polynomial of order n as:

$$B_{n,f}(t) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k}t^k(1-t)^{n-k}.$$

It can be proven that

$$\lim_{n\to\infty}\sup_{t\in[0,1]}|f(t)-B_{n,f}(t)| = 0.$$

See Exercise 2.4.14 for a probability theoretic proof, by noting that if X_k is a collection of $\{0,1\}$ valued independent and identically distributed random variables with $P(X_k=1)=t$, we have that

$$E[f(\frac{1}{N}\sum_{k=1}^{N}X_k)]\to f(t),$$

where this holds uniformly over $t\in[0,1]$. This expectation operation defines a polynomial on $[0,1]$. See Exercise 2.4.14 for further details. ◇

Theorem 2.3.7 *The set $C([0,1];\mathbb{R})$, of continuous functions, is dense in $L_2([0,1];\mathbb{R})$.*

Proof Sketch. First let us consider $f\in L_2([0,1];\mathbb{R})$ that is bounded. Such a function can be approximated by a simple function with an arbitrarily small error under the supremum norm, by the construction of the Lebesgue integration (see Appendix A). By a result known as Urysohn's Lemma, for any Borel E, we can approximate any indicator function $1_{\{x\in E\}}$ with a continuous function g so that $\int_{-K}^{K}|1_{\{x\in E\}}-g(x)|dx\le\epsilon$ for any $\epsilon>0$: This follows first by noting that for any measurable E, for any $\epsilon>0$ there exist F closed and G open with $F\subset E\subset G$ and $\lambda(G\setminus F)<\epsilon$ where λ is the Lebesgue measure, and then noting that one can construct a continuous function which takes the value 1 on F and the value 0 outside G (e.g., $g(x)=1-\frac{d(x,F)}{\max(d(x,F),d(x,[0,1]\setminus G))}$). Thus, given a finite number of indicator functions, their sum can be approximated by so many continuous functions with an arbitrarily small error. If the function f is not bounded, it can be first represented as a sum of positive and negative parts, and each (say the positive part) can be approximated with a bounded function, since for a non-negative valued function f:

$$\lim_{N\to\infty}\int_{t\in[0,1]}|f(t)-\min(N,f(t))|^2dt = 0.$$

This argument follows from the fact that $|f(t)-\min(N,f(t))|^2\le f^2(t)$, and the dominated convergence theorem due to the fact that $f^2(t)$ is integrable. Thus, for any ϵ, a bounded function can be used to approximate f. For such a function, the method presented above can be used to establish the denseness of continuous functions. ◇

Theorem 2.3.8 *The space $L_2([0,1];\mathbb{R})$ is separable.*

Proof. Given Theorem 2.3.7, we can show that polynomials $\{t^k, k \in \mathbb{Z}\}$ defined on $[0,1]$ can be used to construct a complete orthogonal sequence. In view of Theorem 2.3.4, this will imply separability.

To see this, let $\langle h, t^k \rangle = 0$ for all $k \in \mathbb{Z}_+$ but $\|h\| \neq 0$. Then for any finite n and $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{R}$:

$$\langle h, h \rangle = \langle h - \sum_{i=1}^{n} \alpha_i t^i, h \rangle \leq \|h - \sum_{i=1}^{n} \alpha_i t^i\| \|h\| \tag{2.9}$$

The expression $\|h - \sum_{i=1}^{n} \alpha_i t^i\|$ can be made arbitrarily small by first approximating h with a continuous function with a given approximation error (by Theorem 2.3.7) and then approximating that fixed continuous function with a polynomial of some finite degree with another arbitrarily small approximation error (by Theorem 2.3.6). Thus $\|h\|$ must be less than any positive real number. Therefore, there exists a complete sequence. This sequence can be made orthonormal by the Gram-Schmidt procedure. ◇

Thus, Bernstein polynomials are dense in $C([0,1])$, but these are not orthogonal in $L_2([0,1];\mathbb{R})$; and if we apply the Gram-Schmidt procedure to $\{1, t, t^2, \cdots\}$, we can obtain an orthonormal collection of polynomials that is a complete sequence in $L_2([0,1];\mathbb{R})$ as we saw in the proof of Theorem 2.3.8. These orthonormal polynomials are called *Legendre polynomials*. For $L_2([0,1];\mathbb{R})$, there exist further complete orthonormal sequences, to be discussed shortly.

We now discuss the separability of $L_2(\mathbb{R};\mathbb{R})$.

Theorem 2.3.9 *$L_2(\mathbb{R}_+;\mathbb{R})$ is separable.*

Proof Sketch. Consider the set of functions: $F_K = \{g : g(x) = f(x)1_{\{|x| \leq K\}}, f \in L_2(\mathbb{R}_+;\mathbb{R})\}$. Each F_K is separable under the L_2 norm and there exists a countable dense subset in F_K, call them N_K. Furthermore, for every $f \in L_2(\mathbb{R}_+;\mathbb{R})$ and every $\epsilon > 0$, there exists some K and some $g \in F_K$ such that $\|f - g\| \leq \epsilon$ (this follows from the dominated convergence theorem, or the monotone convergence theorem depending on how one may use either). Now, let K range over positive integers, and observe that $\cup_{K \in \mathbb{N}} N_K$ is countable, as a countable union of countable sets. Hence, this set is a countable dense subset in $L_2(\mathbb{R}_+;\mathbb{R})$.

We will study the applications of these results in Fourier transformations, filter design and estimation.

Two further results are presented in the following.

Theorem 2.3.10 *The set $L_2([1,\infty);\mathbb{R})$ is dense in $L_1([1,\infty);\mathbb{R})$.*

Proof: Let $g \in L_1([1,\infty);\mathbb{R})$. Let $g_K(t) = g(t)1_{\{|t| \leq K\}}, t \in \mathbb{R}$. It follows that $g_K \in L_2([1,\infty);\mathbb{R})$ and $||g - g_K||_1 \leq \epsilon$. ◇

Theorem 2.3.11 *Let C_c denote the space of continuous functions with compact support. C_c is dense in $L_1(\mathbb{R};\mathbb{R})$.*

Proof Sketch. Recall first that the support or a function f is defined as the closure of the collection of points on which f is non-zero. First let us assume that the bounded domain $[-K, K]$, for some finite K, contains the support of f. In this domain, we can approximate any $f \in L_1(\mathbb{R};\mathbb{R})$ with a simple function with an arbitrarily small L_1-error, by the construction of the Lebesgue integration (see Appendix A). Again by Urysohn's Lemma, we can approximate any indicator function $1_{\{x \in E\}}$ with a continuous function g with compact support so that $\int_{-K}^{K} |1_{\{x \in E\}} - g(x)| \leq \epsilon$. Thus, given a finite number of indicator functions, their sum can be approximated by so many continuous functions with an arbitrarily small error. If the function f has unbounded support, then we can truncate it first to obtain a function with a finite support, with some (arbitrarily small) approximation error. ◇

The above is also important in that, it shows that in $L_p(\mathbb{R}_+), 1 \leq p < \infty$, the mass of a function cannot escape to infinity. We will revisit this important characteristic occasionally in particular while discussing Fourier transforms.

2.3.4 Signal expansions in $L_2([a,b];\mathbb{R})$ or $L_2([a,b];\mathbb{C})$: Fourier, Haar and Polynomial Bases

Fourier Signals as Basis Vectors and Fourier Series

Fourier series is a very important class of orthonormal sequences which are used to represent both discrete-time and continuous-time signals. These will be studied later on in much detail. In particular, we will soon see that in $L_2([0,P];\mathbb{C})$ the family of complex exponentials

$$\{e_k : e_k(t) = \frac{1}{\sqrt{P}} e^{i2\pi \frac{k}{P} t}, k \in \mathbb{Z}\},$$

provides a complete orthonormal sequence.

Accordingly, for any $x \in L_2([0,P];\mathbb{C})$, we can write

$$x = \sum_{k\in\mathbb{Z}} \langle x, e_k\rangle e_k$$

or by expanding the inner-product, we have

$$x(t) = \sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} x(s) \frac{1}{\sqrt{P}} e^{-i2\pi \frac{k}{P} s} ds \right) \frac{1}{\sqrt{P}} e^{i2\pi \frac{k}{P} t}$$

where the convergence of the infinite sum is in the L_2 sense.

This expansion is precisely the Fourier series expansion of the function x in $L_2([0,P];\mathbb{C})$. The inner-product $\langle x, e_k\rangle$ defines the Fourier transform.

Legendre Polynomials as Basis Vectors

We have seen, in the context of Theorems 2.3.6 and 2.3.7 (see the proof of 2.3.8), the functions $\{t^k, k \in \mathbb{Z}_+\}$ can be used to constructor an orthonormal collection of signals which is complete in $L_2([a,b];\mathbb{R})$. These complete orthonormal polynomials are called Legendre polynomials.

Haar Functions as Basis Vectors

One further practically very important basis is the class of Haar functions (known as wavelets). Define

$$\Psi_{0,0}(x) = \begin{cases} 1, & \text{if } 0 \le x \le 1 \\ 0 & \text{else} \end{cases} \tag{2.10}$$

and for $n \in \mathbb{Z}_+, k \in \{0,1,2,\ldots,2^n-1\}$,

$$\Phi_{n,k}(x) = \begin{cases} 2^{n/2}, & \text{if } k2^{-n} \le x < (k+1/2)2^{-n} \\ -2^{n/2}, & \text{if } (k+1/2)2^{-n} \le x \le (k+1)2^{-n} \\ 0 & \text{else} \end{cases} \tag{2.11}$$

Theorem 2.3.12 *The (Haar) set of vectors*

$$\{\Psi_{0,0}, \Phi_{n,k}, n \in \mathbb{Z}_+, k \in \{0,1,2,\ldots,2^n-1\}\}$$

is a complete orthonormal sequence in $L_2([0,1];\mathbb{R})$.

The important observation to note here is that, different expansions might be suited for different engineering applications: for instance Haar series are occasionally used in image processing with certain *edge* behaviours, whereas Fourier expansion is extensively used in speech processing and communications theoretic applications.

2.3.5 Approximations

Approximations allow us to represent data using finitely many vectors. The basis expansions studied above can be used to obtain the best approximation of a signal up to finitely many terms to be used in an approximation: This can be posed as a projection problem, and we have seen that the best approximation is one in which the approximation error is orthogonal to all the vectors used in the approximation (defining an approximation subspace).

2.4 Exercises

Exercise 2.4.1 *a) The set $C^\infty(\mathbb{R})$, which is the set of all functions from $\mathbb{R}$ to $\mathbb{R}$ that are infinitely differentiable, together with the operations of addition and scalar multiplication defined as follows, is a vector space: For any $f_1, f_2 \in C^\infty(\mathbb{R})$*

$$(f_1 + f_2)(x) = f_1(x) + f_2(x), \qquad x \in \mathbb{R}$$

and for any $\alpha \in \mathbb{R}$ and $f \in C^\infty(\mathbb{R})$

$$(\alpha \cdot f)(x) = \alpha f(x), \qquad x \in \mathbb{R}$$

i) Now, consider $\mathcal{P}(\mathbb{R})$ to be the set of all (polynomial) functions that maps $\mathbb{R}$ to $\mathbb{R}$ such that any $f \in \mathcal{P}(\mathbb{R})$ can be written as $f(x) = \sum_{i=0}^{n} a_i x^i$ for some $n \in \mathbb{N}$ with $a_0, a_1, \cdots, a_n \in \mathbb{R}$. Suppose that we define the same addition and scalar multiplication operations as defined above. Is $\mathcal{P}(\mathbb{R})$ a subspace in $C^\infty(\mathbb{R})$?

ii) Show that the space of all functions in $C^\infty(\mathbb{R})$ which map $\mathbb{R}$ to $\mathbb{R}$ which satisfy $f(10) = 0$ is a vector space with addition and multiplication defined as above.

b) Consider the set $\mathbb{R}^n$. On $\mathbb{R}^n$, define an addition operation and a scalar multiplication operation as follows:

$$(x_1, x_2, \cdots, x_n) + (y_1, y_2, \cdots, y_n) = (x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n)$$

$$\alpha \cdot (x_1, x_2, \cdots, x_n) = (\alpha x_1, \alpha x_2, \cdots, \alpha x_n)$$

Show that, with these operations, $\mathbb{R}^n$ is a vector space.

c) Consider the set

$$\mathbb{W} = \{(x, y) : x \in \mathbb{R}, x \in \mathbb{R}, x > 0, y > 0\}$$

On this set, define an addition operation and a scalar multiplication operation as follows:

$$(x_1, y_1) + (x_2, y_2) = (x_1 y_1, x_2 y_2)$$

$$\alpha \cdot (x, y) = (x^\alpha, y^\alpha)$$

Show that, with these operations, $\mathbb{W}$ is a vector space. Hint: Consider a bijection between $\mathbb{W}$ and the space $\mathbb{R}^2$ with $\mathbb{W} \ni (x, y) \mapsto (\log(x), \log(y)) \in \mathbb{R}^2$.

Exercise 2.4.2 (Hölder's inequality) *Let $1 \leq p, q \leq \infty$ with $1/p + 1/q = 1$. Let $x \in l_p(\mathbb{Z}_+)$ and $y \in l_q(\mathbb{Z}_+)$. Then,*

$$\sum_{i=0}^{\infty} |x_i y_i| \leq ||x||_p ||y||_q$$

This is known as Hölder's inequality. Equality holds if and only if

$$(\frac{x_i}{||x||_p})^{(1/q)} = (\frac{y_i}{||y||_q})^{(1/p)},$$

for each $i \in \mathbb{Z}_+$.

To prove this, perform the following: a) Show that for $a \geq 0, b \geq 0$, $c \in (0,1)$*:* $a^c b^{1-c} \leq ca + (1-c)b$ *with equality if and only if* $a = b$*. To show this, you may consider the function* $f(t) = t^c - ct + c - 1$ *and see how it behaves for* $t \geq 0$ *and let* $t = a/b$*.*

b) Apply the inequality $a^c b^{1-c} \leq ca + (1-c)b$ *to the numbers:*

$$a = (\frac{|x_i|}{||x||_p})^p, \quad b = (\frac{|y_i|}{||y||_q})^q, \quad c = 1/p$$

Hölder's inequality is useful to prove Minkowski's inequality which states that for $1 < p < \infty$*,*

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p$$

This proceeds as follows:

$$\begin{aligned}
&\sum_{i=1}^n |x_i + y_i|^p \leq \sum_{i=1}^n |x_i + y_i|^{p-1}|x_i + y_i| \leq \sum_{i=1}^n |x_i + y_i|^{p-1}|x_i| + \sum_{i=1}^n |x_i + y_i|^{p-1}|y_i| \\
&= \left(\sum_{i=1}^n |x_i + y_i|^{(p-1)q}\right)^{1/q} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |x_i + y_i|^{(p-1)q}\right)^{1/q} \left(\sum_{i=1}^n |y_i|^p\right)^{1/p} \\
&= \left(\sum_{i=1}^n |x_i + y_i|^p\right)^{1/q} \left(\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p\right)^{1/p}\right)
\end{aligned} \tag{2.12}$$

Thus, using that $1 - 1/q = 1/p$*,*

$$\left(\sum_{i=1}^n |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p\right)^{1/p},$$

Now, the above holds for every n*. Taking the limit* $n \to \infty$ *(first on the right and then on the left), it follows that*

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p$$

which is the desired inequality.

Exercise 2.4.3 *a) Let* $C([0,1];\mathbb{R})$ *be the space of continuous functions in* $\Gamma([0,1];\mathbb{R})$ *with the norm*

$$||f|| = \sup_{t \in [0,1]} |f(t)|.$$

Is this space a complete normed linear space?

b) In class we will show that under the norm $||f|| = \int_0^1 |f(t)|dt$*, the space of continuous functions* $C([0,1];\mathbb{R})$ *is not complete. Let us revisit this property.*

Consider the sequence

$$x_n(t) = \begin{cases} 1, & \text{if } 0 \le t \le 1/2 \\ -2^n(t-1/2)+1 & \text{if } 1/2 < t < (1/2)+(1/2)^n \\ 0, & \text{if } (1/2)+(1/2)^n \le t \le 1 \end{cases} \tag{2.13}$$

Is this sequence Cauchy under the described norm? Does the sequence have a limit which is continuous?

Exercise 2.4.4 *Given a normed linear space $(X, ||.||)$, introduce a map $n : X \times X \to \mathbb{R}$:*

$$n(x,y) = \frac{||x-y||}{1+||x-y||}$$

Show that $n(x,y)$ is a metric: That is, it satisfies the triangle inequality:

$$n(x,y) \le n(x,z) + n(z,y), \quad \forall x,y,z \in X,$$

and that $n(x,y) = 0$ iff $x = y$, and finally $n(x,y) = n(y,x)$.

Exercise 2.4.5 *Let $\{e_n, n \in \mathbb{N}\}$ be a complete orthonormal sequence in a real Hilbert space H. Let $\mathcal{M}$ be a subspace of H, spanned by $\{e_k, k \in S\}$, for some finite set $S \subset \mathbb{N}$. That is,*

$$\mathcal{M} = \{v \in H : \exists \alpha_k \in \mathbb{R}, k \in S, v = \sum_{k \in S} \alpha_k e_k\}$$

Let $x \in H$ be given. Find $x^* \in \mathcal{M}$ which is the solution to the following:

$$\min_{x_0 \in \mathcal{M}} ||x - x_0||,$$

in terms of x, and $\{e_n, n \in \mathbb{N}\}$.

Hint: Any vector in H can be written as $x = \sum_{n \in \mathbb{N}} \langle x, e_n \rangle e_n$.

Exercise 2.4.6 *Let $T : L_2(\mathbb{R}_+; \mathbb{R}) \to \mathbb{R}$ be a mapping given by:*

$$T(f) = \int_1^\infty f(t) \frac{1+t^2}{t^4} dt$$

Is T continuous at any given $f_0 \in L_2(\mathbb{R}_+; \mathbb{R})$?

Provide precise arguments. What does it mean to be continuous at f_0?

Exercise 2.4.7 *Consider an inner-product defined by:*

$$\langle x, y \rangle = \lim_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} x(t) y(t)$$

Is the resulting inner-product (pre-Hilbert) space separable*?*

Exercise 2.4.8 *Let $x, y \in f(\mathbb{Z}; \mathbb{R})$; that is, x, y map $\mathbb{Z}$ to $\mathbb{R}$, such that $x = \{\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots\}$ and $y = \{\dots, y_{-2}, y_{-1}, y_0, y_1, y_2, \dots\}$ and $x_k, y_k \in \mathbb{R}$, for all $k \in \mathbb{Z}$.*

For a)-c) below, state if the following are true or false with justifications in a few sentences:

a) $\langle x, y \rangle = \sum_{i \in \mathbb{Z}} i^2 x_i y_i$ is an inner-product.

b) $\langle x, y\rangle = \sum_{i\in\mathbb{Z}} x_i y_i$ *is an inner-product.*
c) $\{x : ||x||_2^2 < \infty\}$ *contains a complete orthonormal sequence, where* $||x||_2 = \sqrt{\sum_{i\in\mathbb{Z}} |x(i)|^2}$.

Exercise 2.4.9 *Let* $\mathbb{X}$ *be a Hilbert space and* $x, y \in \mathbb{X}$. *Prove the following:*

a)

$$||x+y||^2 + ||x-y||^2 = 2||x||^2 + 2||y||^2.$$

b)

$$|\langle x, y\rangle| \leq (||x||)(||y||).$$

c) [10 Points]

$$2|\langle x, y\rangle| \leq ||x||^2 + ||y||^2.$$

Exercise 2.4.10 *Let* H *be a finite dimensional Hilbert space and* $\{v_1, v_2\}$ *be two linearly independent vectors in* H.

Let $b_1, b_2 \in \mathbb{R}$. *Show that, among all vectors* $x \in H$, *which satisfies*

$$\langle x, v_1\rangle = b_1,$$

$$\langle x, v_2\rangle = b_2,$$

the vector $x^* \in H$ *has the minimum norm if* x^* *satisfies:*

$$x^* = \alpha_1 v_1 + \alpha_2 v_2,$$

with

$$\langle v_1, v_1\rangle\alpha_1 + \langle v_2, v_1\rangle\alpha_2 = b_1,$$
$$\langle v_1, v_2\rangle\alpha_1 + \langle v_2, v_2\rangle\alpha_2 = b_2.$$

Exercise 2.4.11 *Let* H *be a Hilbert space and* $C \subset H$ *be a dense subset of* H. *Suppose that any element* h_C *in* C *is such that for every* $\epsilon > 0$, *there exist* $n \in \mathbb{N}$ *and* $\beta_i \in \mathbb{R}, i \in \mathbb{N}$ *so that*

$$\|\sum_{i=0}^{n} \beta_i e_i - h_C\| \leq \epsilon$$

where $\{e_\alpha, \alpha \in \mathbb{N}\}$ *is a countable sequence of orthonormal vectors in* H.

Is it the case that H *is* separable*?*

Exercise 2.4.12 *Let* x *be in the real Hilbert space* $L_2([0,1];\mathbb{R})$ *with the inner product*

$$\langle x, y\rangle = \int_0^1 x(t)y(t)dt.$$

We would like to express x *in terms of the following two signals (which belong to the Haar signal space)*

$$u_1(t) = 1_{\{t\in[0,1/2)\}} - 1_{\{t\in[1/2,1]\}}, \quad t \in [0,1]$$

$$u_2(t) = 1_{\{t\in[0,1]\}}, \quad t \in [0,1]$$

such that

$$\int_0^1 |x(t) - \sum_{i=1}^{2} \alpha_i u_i(t)|^2 dt$$

is minimized, for $\{\alpha_1, \alpha_2 \ \in \mathbb{R}\}$.

a) Using the Gram-Schmidt procedure, obtain two orthonormal vectors $\{e_1(t), e_2(t)\}$ *such that these vectors linearly span the same space spanned by* $\{u_1(t), u_2(t)\}$.

b) State the problem as a projection theorem problem by clearly identifying the Hilbert space and the projected subspace.

c) Let $x(t) = 1_{\{t\in[1/2,1]\}}$. *Find the minimizing* α_1, α_2 *values.*

Exercise 2.4.13 *Alice and Bob are approached by a generous company and asked to solve the following problem: The company wishes to store any signal* f *in* $L_2(\mathbb{R}_+;\mathbb{R})$ *in a computer with a given error of* $\epsilon > 0$*, that is for every* $f \in L_2(\mathbb{R}_+)$*, there exists some signal* $h \in H$ *such that* $||f - h||_2 \leq \epsilon$ *(thus the error is uniform over all possible signals), where* H *is the stored family of signals (in the computer's memory).*

To achieve this, they encourage Alice or Bob to use a finite or a countable expansion to represent the signal and later store this signal in an arbitrarily large memory. Hence, they allow Alice or Bob to purchase as much memory as they would like for a given error value of ϵ.

Alice turns down the offer and says it is impossible to do that for any ϵ *with a finite memory and argues then she needs infinite memory, which is impossible.*

Bob accepts the offer and says he may need a very large, but finite, memory for any given $\epsilon > 0$*; thus, the task is possible.*

Which one is the accurate assessment?

a) If you think Alice is right, which further conditions can she impose to make this possible? Why is she right?

b) If you think Bob is right, can you suggest a method? Why is he right?

Exercise 2.4.14 *Prove Theorem 2.3.6 using a probability theoretic method. Proceed as follows: The number*

$$B_{n,f}(t) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k}t^k(1-t)^{n-k}.$$

can be expressed as the expectation $E[f_t(\frac{S_n}{n})]$*, where* $S_n = X_1 + X_2 + \cdots + X_n$*, where* X_i *is an i.i.d. collection of Bernoulli random variables where* $X_i = 1$ *with probability* t *and* $X_i = 0$ *with probability* $1 - t$*. Here, observe that the sum* S_n *has a binomial distribution. Thus,*

$$\sup_{t\in[0,1]} |f(t) - B_{n,f}(t)| = \sup_{t\in[0,1]} |E_t[f(\frac{S_n}{n})] - f(t)|,$$

where E_t*, for each* t*, denotes the expectation with respect to the i.i.d. Bernoulli random variables* X_i *each with* $P(X_i = 1) = t$*. Let* P_t *denote the probability measure induced by these* t*-parametrized i.i.d. sequence of Bernoulli random variables.*

Since f *is continuous and* $[0, 1]$ *is compact,* f *is uniformly continuous. Thus, for every* $\epsilon > 0$*, there exists* $\delta > 0$ *such that* $|x - y| < \delta$ *implies that* $|f(x) - f(y)| \leq \epsilon$*. Thus,*

$$\begin{aligned}
&|E_t[f(\frac{S_n}{n})] - f(t)| \\
&= \int_{\omega:|\frac{S_n}{n}-t|\leq\delta} |f(\frac{S_n}{n})] - f(t)|P(d\omega) + \int_{\omega:|\frac{S_n}{n}-t|>\delta} |f(\frac{S_n}{n})] - f(t)|P(d\omega) \\
&\leq \epsilon + 2 \sup_{y\in[0,1]} |f(y)|P_t(|\frac{S_n}{n} - t| > \delta) \qquad (2.14)
\end{aligned}$$

The last term converges to ϵ *as* $n \to \infty$ *by the law of large numbers. The above holds for every* $\epsilon > 0$.

Now, one needs to show that this convergence is uniform in t: For this show that for all $t \in [0,1]$, via Markov's inequality and the independence of X_i

$$P_t(|\frac{S_n}{n} - t| > \delta) = P_t(|\frac{S_n}{n} - t|^2 > \delta^2) \leq \frac{1}{4n\delta^2},$$

establishing uniform convergence (over $t \in [0,1]$), and thus complete the proof.

Exercise 2.4.15 (A useful result on countability properties) *Let $F : \mathbb{R} \to \mathbb{R}$ be a monotonically increasing function (that is, $x_1 \leq x_2$ implies that $F(x_1) \leq F(x_2)$). Show that F can have at most countably many points of discontinuity.*

Hint. If a point is such that F is discontinuous, then there exists $n \in \mathbb{N}$ with $F(x^+) := \liminf_{x_n \downarrow x} F(x)$, $F(x^-) := \limsup_{x_n \uparrow x} F(x)$, $F(x^+) - F(x^-) > \frac{1}{n}$. Express $\mathbb{R} = \cup_{m \in \mathbb{Z}}(m, m+1]$. Let $B_n^m := \{x \in (m, m+1] : F(x^+) - F(x^-) > \frac{1}{n}\}$. It must be that B_n^m is finite for otherwise the jump would be unbounded in the interval $(m, m+1]$. Then, the countable union $\cup_n B_n^m$ will be countable. Finally $\cup_m B_n^m$ is also countable.

Exercise 2.4.16 *Prove Theorem 2.3.12.*

3

Dual Spaces, the Schwartz Space and Distribution Theory, and the Dirac Delta Function

A complete understanding of Fourier transforms is possible through an investigation building on distributions: A distribution is a continuous, linear function on a space of test functions. The space of test functions we will consider will prove to be very useful.

To gain some intuition, consider the function $\sin(nt)$. This function does not have a pointwise limit as $n \to \infty$. However, with f an arbitrary continuous function, the integral $\int \sin(nt) f(t) dt$ has a well-defined limit, which is zero. In this sense, $\sin(nt)$ admits a limit which is equivalent to the constant function with value 0. This will motivate us to introduce distribution theory.

There will be some additional useful properties: Every distribution is differentiable, and the differentiation is continuous. Most importantly, a function whose Fourier transform is not defined as a function might have a transform in a distributional sense.

Perhaps, it will not be immediately evident that the study of such a theory is needed in engineering practice. However, the patient student will realize the importance of this topic, and versatility in introduces, both this semester, in the context of Fourier transformations, as well as next year, or afterwards, while studying topics in optimization, control, and probability.

In our course, one important application which arises while studying linear systems as well as Laplace and Fourier transforms is with regard to the use of the impulse (or Dirac delta) function. Such functions do not live in the set of $\mathbb{R}$-valued functions, and hence many operations such as integration, become ill-stated. However, the Dirac delta function is such an important and crucial object that one has to know how to work with it even in the most elementary applications in signal processing, circuit analysis, control, and communications, in addition to many other areas of engineering and applied mathematics. We will see that the appropriate way to study the impulse function is to always work *under an integral*. We will make this discussion more precise in the following.

3.1 Dual Space of a Normed Linear Space

Let f be a linear functional on a normed linear space X. We say f is bounded (in the operator norm) if there is a constant M such that $|f(x)| \leq M\|x\|$ for all $x \in X$. The smallest such M is called the norm of f and is denoted by $\|f\|$, also given by:

$$\|f\| := \sup_{x:\|x\|\neq 0} \frac{|f(x)|}{\|x\|}. \tag{3.1}$$

The space of all bounded, linear functionals on X is called the (topological) dual space of X. This is equivalent to the space of all continuous and linear functions, as continuity and boundedness imply each other:

Exercise 3.1.1 *Show that a linear functional on a normed linear space is bounded if and only if it is continuous.*

Let X be a normed space of signals and let us define the dual space of X as the set of linear and bounded functions on X to $\mathbb{R}$ or $\mathbb{C}$, and let us denote this space by X^*. This space is a linear space, under pointwise addition and scalar multiplication of functions in it.

Furthermore, the space of continuous and linear functions is itself a normed space with the norm given above in (3.1).

Exercise 3.1.2 *Show that $(X^*, \|\cdot\|)$ is a Banach space.*

Remark 3.1. We note that the above holds even if X itself is not Banach.

A key result for identifying the dual spaces for $l_p(\mathbb{Z}_+;\mathbb{R})$ or $L_p(\mathbb{R}_+;\mathbb{R})$ spaces is Hölder's inequality (see Theorem 2.1.2): Let $1 \leq p, q < \infty$ or possibly ∞. Then,

$$\sum_{i\in\mathbb{Z}_+} x_i y_i \leq \|x\|_p \|y\|_q,$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Theorem 3.1.1 (Riesz Representation Theorem for $l_p(\mathbb{Z}_+;\mathbb{R})$ and $L_p(\mathbb{R}_+;\mathbb{R})$ spaces) *Every linear bounded function on $l_p(\mathbb{Z}_+;\mathbb{R})$, $1 \leq p < \infty$, is representable uniquely in the form*

$$f(x) = \sum_{i=0}^{\infty} \eta_i x_i,$$

where $\eta = \{\eta_i\}$ is in l_q. Furthermore, every element in l_q defines a member of $l_p(\mathbb{Z}_+;\mathbb{R})^*$, and

$$\|f\| = \|\eta\|_q, \tag{3.2}$$

This also applies to $L_p(\mathbb{R}_+;\mathbb{R})$ spaces.

Riesz Representation Theorem tells us that while studying spaces such as $L_p(\mathbb{R}_+;\mathbb{R})$ or $l_p(\mathbb{N};\mathbb{R})$, we can use an inner-product like (but not really an inner-product in the way we defined Hilbert spaces) expression to represent the set of all linear functions on X by:

$$\langle \cdot, y\rangle : x \mapsto \langle x, y\rangle = \int_{\mathbb{R}} x(t)y(t)dt$$

where $\langle x, y\rangle$ is a continuous linear function on X, but this is equivalent to a function $y \in X^*$ having an inner-product like function with $x \in X$. Likewise, for a discrete-time signal:

$$\langle \cdot, y\rangle : x \mapsto \langle x, y\rangle = \sum_{i=1}^{\infty} x(i)y(i)dt,$$

is a linear function on X.

For example if $X = L_p(\mathbb{R}_+;\mathbb{R})$ for $1 \leq p < \infty$, we can show that the dual space of X is representable by elements in $L_q(\mathbb{R}_+;\mathbb{R})$ where $\frac{1}{p} + \frac{1}{q} = 1$.

In the special case of $p = 2$ we have the space $L_2(\mathbb{R}_+;\mathbb{R})$, which has its dual space as itself.

The following is a general result for Hilbert spaces.

Theorem 3.1.2 (Riesz Representation Theorem for Hilbert Spaces) *Every linear bounded function on a Hilbert space H admits a representation of the form:*

$$f(x) = \langle x, y\rangle$$

for some $y \in H$.

We say that $x \in X$ and $x^* \in X^*$ are aligned if

$$\langle x, x^* \rangle = \|x\| \|x^*\|.$$

Remark 3.2. Some observations beyond the scope of our course follow.

(i) The dual space of $l_\infty(\mathbb{Z}_+; \mathbb{R})$ or $L_\infty(\mathbb{R}_+; \mathbb{R})$ is more complicated (due to the fact that such functions do not converge to zero as the index goes unbounded), and will not be considered in this course. On the other hand, let $c_0 \in \Gamma(\mathbb{Z}_+; \mathbb{R})$ be the set of signals which decay to zero. The dual of this space is (associated with, in the sense of the representation result presented earlier) $l_1(\mathbb{Z}; \mathbb{R})$.

(ii) The dual of $C([a, b]; \mathbb{R})$ can be associated with the space of signed measures with bounded *total variation*. Likewise, let $C_0(\mathbb{R}; \mathbb{R})$ denote the space of continuous functions f which satisfy $\lim_{|x| \to \infty} f(x) = 0$. The dual of this space is (associated with) the space of finite signed measures with bounded *total variation*.

(iii)Those of you who will take further courses on probability will study the concept of weak convergence of probability measures. A sequence of probability measures μ_n converges to some probability measure μ weakly if for every f in $C_b(\mathbb{R}; \mathbb{R})$ (that is the set of continuous and bounded functions on $\mathbb{R}$)

$$\int \mu_n(dx) f(x) \to \int \mu(dx) f(x).$$

If we had replaced $C_b(\mathbb{R}; \mathbb{R})$ with $C_0(\mathbb{R}; \mathbb{R})$ here, note that this would coincide with the weak*-convergence of $\mu_n \to \mu$ (to be studied in the following). Nonetheless, in probability theory the convergence stated above is so important that this is simply called *weak convergence*.

3.1.1 Weak and Weak* Convergence

Earlier, we discussed that in a normed space X, a sequence of vectors $\{x_n\}$ converges to a vector x if

$$\|x_n - x\| \to 0.$$

The above is also called strong convergence.

Definition 3.1.1 *A sequence $\{x_n\}$ in X is said to converge weakly to x if*

$$f(x_n) \to f(x)$$

for all $f \in X^*$.

Exercise 3.1.3 *Let $x \in l_2(\mathbb{N}; \mathbb{R})$. Show that if*

$$x \to x^*,$$

then

$$\langle x, f \rangle \to \langle x^*, f \rangle \qquad \forall f \in l_2(\mathbb{N}; \mathbb{R})$$

We note however that, weak convergence does not imply strong convergence.

A related convergence notion, one that we will adopt while studying distributions, is that of weak* convergence, defined next.

Definition 3.1.2 *A sequence $\{f_n\}$ in X^* is said to converge in the weak* sense to f if*

$$f_n(x) \to f(x)$$

for all $x \in X$.

We note that such a convergence notion is useful in the study of solutions to differential equations (ordinary and partial), optimal control theory, and probability theory as well, even though we will not be able to discuss these in our course.

3.2 Distribution Theory

A distribution is a *linear* and *continuous* $\mathbb{R}-$valued function (that is, a functional) on a space of test functions. Thus, a distribution can be viewed to be an element of the dual space of a linear space test functions (even though we will see that the linear space of test functions does not need to form a normed linear space).

Studying distributions and sets of test functions present many benefits for our course. For example, the delta function has a natural representation as a distribution. Furthermore, Fourier analysis will be observed to be a bijective mapping from a space of test functions to another one, and this space of test functions is rich enough to approximate many functions that we encounter in applications sufficiently well. Furthermore, we will define the Fourier transform first on a space of test functions and extend it from this space to larger spaces, such as $L_2(\mathbb{R};\mathbb{C})$.

3.2.1 Space $\mathcal{D}$ and $\mathcal{S}$ of Test Functions

Let $\mathcal{D}$ denote a set of test functions from $\mathbb{R}$ to $\mathbb{R}$, which are smooth (infinitely differentiable) and which have bounded support sets. Such functions exist, for example

$$f(t) = 1_{\{|t|\leq 1\}} e^{\frac{1}{t^2-1}},$$

is one such function.

We say a sequence of signals $\{x_i\}$ in $\mathcal{D}$ converges to the null element $\underline{0}$ if a) For every $i \in \mathbb{N}$, there exists a compact, continuous-time domain $T \subset \mathbb{R}$ such that the support set of x_i is contained T (we define the support for a function f to be the closure of the set of points $\{t : f(t) > 0\}$).

b) For every $\epsilon > 0$, and k there exists an $N_k \in \mathbb{Z}_+$ such that for all $n \geq N_k$, $p_k(x) \leq \epsilon$, where $p_k = \sup_{t\in\mathbb{R}} |\frac{d^k}{dt^k} x(t)|$ (that is, all the derivatives of x converge to zero uniformly on $\mathbb{R}$).

In applications we usually encounter signals with unbounded support. Hence, a theory based on the above test functions might not be sufficient. Furthermore, the Fourier transform of a function in $\mathcal{D}$ is not in the same space (a topic to be discussed further). As such, we will find it convenient to slightly extend the space of test signals.

Definition 3.2.1 (Schwartz Signal Space $\mathcal{S}$) *An infinitely differentiable signal $\phi : \mathbb{R} \to \mathbb{R}$ is in the Schwartz Signal space, denoted with $\mathcal{S}$, if for each $k \in \mathbb{Z}_+$ and for each $l \in \mathbb{Z}_+$*

$$\sup_{t\in\mathbb{R}} |t^l \phi^{(k)}(t)| < \infty,$$

where $\phi^{(k)}(t) = \frac{d^k}{dt^k}\phi(t)$.

For example the function $\phi(t) = e^{-t^2}$ is a Schwartz signal.

One can equip $\mathcal{S}$ with the topology generated by a countable number of *semi-norms*:

$$p_{\alpha,\beta}(\phi) := \sup_t |t^\alpha \frac{d^\beta}{d^\beta}\phi(t)|,$$

for $\alpha, \beta \in \mathbb{N}$. That is, we say, a sequence of functions ϕ_n in $\mathcal{S}$ converges to another one ϕ if

$$\lim_{n\to\infty} p_{\alpha,\beta}(\phi_n - \phi) = 0, \quad (\alpha,\beta) \in \mathbb{Z}_+ \times \mathbb{Z}_+.$$

With the above, we could define a metric by working with the above seminorms: for $x, y \in \mathcal{S}$, let us define a metric between the two vectors as:

$$d(x,y) = \sum_n \frac{1}{2^n} \frac{p_n(x)}{1 + p_n(x)}, \tag{3.3}$$

where n is a countable enumeration of the pairs $(\alpha,\beta) \in \mathbb{Z}_+ \times \mathbb{Z}_+$.

The Schwartz space of signals equipped with such a metric will be a complete space. Furthermore, differentiation operator becomes a continuous operation in $\mathcal{S}$, under this metric; a topic which we will discuss further.

As had been discussed before (slightly generalizing Theorem 2.1.3), a functional T from $\mathcal{S} \to C$ is continuous if and only if for every convergent sequence in $\mathcal{S}$, $\phi_n \to \phi$, we have $T(\phi_n) \to T(\phi)$. We note that checking sequential continuity is typically easier than continuity, since in the space $\mathcal{S}$, it is not convenient to compute the distance between two vectors given the quite involved construction of the metric in (3.3).

Definition 3.2.2 *A distribution is a linear, continuous functional on the space of test functions $\mathcal{S}$.*

Thus, a distribution is an element of the **dual space** of $\mathcal{S}$ (that is, $\mathcal{S}^*$), even though $\mathcal{S}$ is not defined as a normed space, but as a metric space which is nonetheless a linear space.

General Distributions and Singular Distributions

Distributions can be regular and singular. Regular distributions can be expressed as an integral of a test function and a locally integrable function (that is a function which has a finite absolute integral on any compact domain on which it is defined). For example if $\gamma(t)$ is a real-valued integrable function on $\mathbb{R}$, and $\phi \in \mathcal{S}$ the distribution given by

$$\bar{\gamma}(\phi) := \int_{\mathbb{R}} \gamma(t)\phi(t)dt \tag{3.4}$$

is a regular distribution on $\mathcal{S}$, **represented** by a regular, integrable, function $\gamma(t)$.

Definition 3.2.3 *A tempered signal, $x(t)$ is one which satisfies small growth, that is, for some $\beta, \gamma \in \mathbb{R}$, $N \in \mathbb{Z}_+$:*

$$|x(t)| \leq \beta|t|^N + \gamma, \quad \forall t \in \mathbb{R}$$

Any tempered signal can represent a regular distribution.

Singular distributions do not admit such a representation. For example the Dirac delta distribution $\bar{\delta}$, defined for all $\phi \in \mathcal{S}$:

$$\bar{\delta}(\phi) = \phi(0),$$

does not admit a representation in the form $\int g(t)\phi(t) = \phi(0)$. Even when there is no function which can be used to represent a singular distribution, one occasionally represents a singular distribution as if such a signal exists and call the representing function a singular or a generalized function. The informal expression $\int \delta(t)\phi(t) = \phi(0)$ is a common example for this, where δ is the generalized impulse function which takes the value ∞ at 0, and zero elsewhere.

Proposition 3.2.1 *The map $\bar{\delta}(\phi) = \phi(0)$ defines a distribution on $\mathcal{S}$. This distribution is called the Dirac delta distribution.*

Proof: We need to show that the $\bar{\delta}(\phi)$ is linear and continuous. Linearity is immediate. To show continuity, let $\phi_n \to \phi$. This implies that $\sup_t |\phi_n(t) - \phi(t)| = 0$, by the definition of convergence. Hence

$$\bar{\delta}(\phi_n) = \phi_n(0) \to \phi(0) = \bar{\delta}(\phi)$$

$\diamond$

Exercise 3.2.1 *Show that the relation*

$$\bar{f}(\phi) = \int_0^\infty t^2\phi(t)dt, \qquad \phi \in \mathcal{S}$$

defines a distribution $\bar{f} \in \mathcal{S}^*$.

While performing operations on distributions, we first study regular distributions and try to check if the operation is consistent with such distributions.

Equivalence and Convergence of Distributions

Two distributions $\bar{\gamma}$ and $\bar{\zeta}$ are equal if

$$\bar{\gamma}(f) = \bar{\zeta}(f), \quad \forall f \in \mathcal{S}$$

Definition 3.2.4 *A sequence of distributions* $\{\bar{\gamma}_n\}$ *converges to a distribution* $\bar{\gamma}$ *if*

$$\bar{\gamma}_n(f) \to \bar{\gamma}(f), \quad \forall f \in \mathcal{S}$$

Observe that the above notion is identical to the weak* convergence notion discussed earlier.

Exercise 3.2.2 *Let for* $j \in \mathbb{Z}_+, j > 0$

$$f_j(t) = \begin{cases} j, & \text{if } 0 \le t \le \frac{1}{j} \\ 0 & \text{else} \end{cases} \tag{3.5}$$

a) For any real-valued function $g \in \mathcal{S}$*, define*

$$\bar{f}_j(g) := \int_0^\infty f_j(t)g(t)dt.$$

Show that $\bar{f}_j(.)$ *is a distribution on* $\mathcal{S}$ *for every* $j \in \mathbb{N}$.

b) Show that

$$\lim_{j\to\infty} \int_0^\infty f_j(t)g(t)dt = \bar{\delta}(g) = g(0).$$

Conclude that, the sequence of regular *distributions* $\bar{f}_j(.)$*, represented by a real-valued, integrable function* $f_j(t)$*, converges to the Dirac delta distribution* $\bar{\delta}(.)$ *on the space of test signals in* $\mathcal{S}$.

In fact, we can find many other signals which can define regular distributions whose limit is the delta distribution. This motivates the following section.

3.3 Approximate Identity Sequences

Definition 3.3.1 *Let* $\psi_n : \mathbb{R} \to \mathbb{R}$ *be a sequence such that*

- $\psi_n(t) \geq 0, \qquad t \in \mathbb{R}, n \in \mathbb{N}.$
- $\int \psi_n(t)dt = 1, \qquad n \in \mathbb{N}.$
- $\lim_{n\to\infty} \int_{\delta \leq |t|} \psi_n(t)dt = 0, \qquad \forall \delta > 0.$

Such ψ_n sequences are called approximate identity sequences.

We have seen one example in (3.5). The result discussed generalizes to any approximate identity sequence:

Theorem 3.3.1 *Distributions represented by approximate identity sequences converge to the Dirac delta distribution as $n \to \infty$.*

Proof. Let $\phi \in \mathcal{S}$. Then,

$$\begin{aligned} &= \int \psi_n(t)\phi(t)dt - \phi(0) \\ &= \int \psi_n(t)\phi(t)dt - \int \psi_n(t)\phi(0)dt \\ &= \int \psi_n(t)(\phi(t) - \phi(0))dt \end{aligned}$$

Since ϕ is continuous, for every $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that $|\phi(t) - \phi(0)| \leq \epsilon$. Accordingly,

$$\begin{aligned} &\limsup_{n\to\infty} |\int \psi_n(t)\phi(t)dt - \phi(0)| \\ &= \limsup_{n\to\infty} |\int \psi_n(t)(\phi(t) - \phi(0))dt| \\ &\leq \limsup_{n\to\infty} |\int_{t:|t|>\delta_\epsilon} \psi_n(t)(\phi(t) - \phi(0))dt| \\ &+ \qquad \limsup_{n\to\infty} |\int_{t:|t|\leq\delta_\epsilon} \psi_n(t)(\phi(t) - \phi(0))dt| \\ &\leq \limsup_{n\to\infty} \int_{t:|t|>\delta_\epsilon} \psi_n(t)|(\phi(t) - \phi(0))|dt \\ &+ \qquad \limsup_{n\to\infty} \int_{t:|t|\leq\delta_\epsilon} \psi_n(t)|\phi(t) - \phi(0)|dt \\ &\leq (2\sup_{t\in\mathbb{R}} |\phi(t)|) \limsup_{n\to\infty} \int_{t:|t|>\delta_\epsilon} \psi_n(t)dt \\ &+ \qquad \limsup_{n\to\infty} \int_{t:|t|\leq\delta_\epsilon} \psi_n(t)\epsilon dt \\ &\leq \epsilon, \end{aligned} \tag{3.6}$$

where we use the properties of the approximate identity sequences above. Since $\epsilon > 0$ is arbitrary; the result follows so that the limit above exists and is zero. Thus, we conclude that

$$\int \psi_n(t)\phi(t)dt \to \bar{\delta}(\phi) = \phi(0).$$

□

One example for such f_n functions is the following *Gaussian* sequence given by.

$$f_n(t) = \frac{1}{\sqrt{2\pi\frac{1}{n}}} e^{\frac{-1}{2}\frac{-t^2}{\frac{1}{n}}} \tag{3.7}$$

Observe that such an f_n sequence lives in $\mathcal{S}$.

We state a further example below.

Exercise 3.3.1 *Consider the sequence*

$$\psi_n(x) = c_n(1+\cos(x))^n 1_{\{|x|\leq\pi\}}$$

where c_n is so that make $\int \psi_n(x)dx = 1$. Show that

$$\lim_{n\to\infty} \int_{|x|\geq\delta} \psi_n(x)dx = 0 \qquad \forall \delta > 0.$$

One useful sequence, which does not satisfy the non-negativity property above, but that satisfies the convergence property (to $\bar{\delta}$) is the following sequence:

$$\psi_n(x) = \frac{\sin(nx)}{\pi x}$$

Theorem 3.3.2 *For any $\phi \in \mathcal{S}$*

$$\lim_{n\to\infty} \int \psi_n(dx)\phi(x) = \phi(0) \tag{3.8}$$

Proof. We use the following supporting results:

(i)

$$\lim_{R\to\infty} \int_{-R}^{R} \frac{\sin(x)}{x} dx = \pi$$

(ii) Riemann-Lebesgue Lemma (see Theorem 5.3.3): For any integrable function g, $\lim_{|f|\to\infty} \int g(x)e^{ifx}dx = 0$.

(iii)We have $\lim_{|x|\to 0} \frac{\phi(x)-\phi(0)}{x} = h(x)$ for some smooth h.

Let us express the integration as

$$\int_{|x|\geq 1} \psi_n(x)\phi(x) + \int_{|x|\leq 1} \psi_n(dx)\phi(x)$$

The first expression goes to zero, by the Riemann-Lebesgue Lemma. For the second term, we write

$$\int_{|x|\leq 1} \psi_n(x)\phi(x) = \int_{|x|\leq 1} \phi(0)\psi_n(x)dx + \int_{|x|\leq 1} \frac{\phi(x)-\phi(0)}{\pi x} \sin(nx)dx.$$

The second term in this expression goes to zero, by the Riemann-Lebesgue Lemma through the relation $\lim_{|x|\to 0} \frac{\phi(x)-\phi(0)}{x} = h(x)$ for some smooth h and splitting the integral to an arbitrarily small interval $(-\delta, \delta)$ and its complement $[-1,1]\backslash(-\delta,\delta)$ and studying these separately, noting the convergence of the smaller interval integration to that involving the smooth function h. The first term $\int_{|x|\geq 1} \phi(0)\psi_n(x)dx$ converges to $\phi(0)$: By writing $u = nx$, we obtain

$$\int_{|x|\leq 1} \frac{\sin(nx)}{\pi x} dx = \int_{|u|\leq n} \frac{\sin(u)}{\pi u} du$$

and using the fact that $\lim_{n\to\infty} \int_{-n}^{n} \frac{\sin(x)}{\pi x} dx = 1$, the result follows. □

3.3.1 Convolution and its use in approximations

The convolution of two functions (whenever this integration is well-defined) is defined as:

$$(\psi * \phi)(t) = \int \psi(\tau)\phi(t-\tau)d\tau = \int \phi(\tau)\psi(t-\tau)d\tau$$

The convolution can be defined for any pair of functions which are in $L_2(\mathbb{R};\mathbb{R})$. The convolution of two functions in $\mathcal{S}$ is also in $\mathcal{S}$.

A very useful result is the following.

Theorem 3.3.3 *If ψ_n is an approximate identity sequence, Then,*

$$(\psi_n * f)(t) \to f(t),$$

for every continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$, uniformly on compact sets $[a, b] \subset \mathbb{R}$.

Proof. Write

$$\begin{aligned}
&(\psi_n * f)(t) - f(t) \\
&= \int \psi_n(\tau) f(t-\tau)d\tau - f(t) \\
&= \int \psi_n(\tau) f(t-\tau)d\tau - \int \psi_n(\tau) f(t)d\tau \\
&= \int \psi_n(\tau)\Big(f(t-\tau) - f(t)\Big)d\tau
\end{aligned}$$

Since f is continuous, for every $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that $|f(t) - f(t-\delta_\epsilon)| \le \epsilon$. Accordingly,

$$\begin{aligned}
&\limsup_{n\to\infty} |(\psi_n * f)(t) - f(t)| \\
&= \limsup_{n\to\infty} \Big| \int \psi_n(\tau)\Big(f(t-\tau) - f(t)\Big)d\tau \Big| \\
&\le \limsup_{n\to\infty} \Big| \int_{\tau:|\tau|>\delta_\epsilon} \psi_n(\tau)\Big(f(t-\tau) - f(t)\Big)d\tau \Big| \\
&+ \quad \limsup_{n\to\infty} \Big| \int_{\tau:|\tau|\le\delta_\epsilon} \psi_n(\tau)\Big(f(t-\tau) - f(t)\Big)d\tau \Big| \\
&\le \limsup_{n\to\infty} \int_{\tau:|\tau|>\delta_\epsilon} \psi_n(\tau)|f(t-\tau) - f(t)|d\tau \\
&+ \quad \limsup_{n\to\infty} \int_{\tau:|\tau|\le\delta_\epsilon} \psi_n(\tau)|f(t-\tau) - f(t)|d\tau \\
&\le (2\sup_{t\in\mathbb{R}} |f(t)|) \limsup_{n\to\infty} \int_{\tau:|\tau|>\delta_\epsilon} \psi_n(\tau)d\tau \\
&+ \quad \limsup_{n\to\infty} \int_{\tau:|\tau|\le\delta_\epsilon} \psi_n(\tau)\epsilon d\tau \\
&\le \epsilon
\end{aligned} \tag{3.9}$$

Since $\epsilon > 0$ is arbitrary; the result follows so that the limit above exists and is zero. Note that the for compact K, the δ, ϵ pair can be taken to be uniform for all $t \in K$, that is, for every $\epsilon > 0$, there exists a $\delta_{\epsilon,K} > 0$ such that $|f(t) - f(t-\delta_\epsilon)| \le \epsilon$ for all $t \in K$. Therefore, the convergence is uniform over compact sets, in the sense that

$$\lim_{n\to\infty}\sup_{t\in K}|(\psi_n * f)(t) - f(t)| = 0.$$

□

Note that with ψ_n defined as in (3.7), $(\psi_n \star \phi)$ is always infinitely differentiable, and one may conclude the following:

Corollary 3.3.1 *The space of smooth signals are dense in the space of continuous functions with a compact support under the supremum norm.*

3.3.2 Complenetess of complex exponentials in $L_2([-\pi,\pi];\mathbb{C})$

Using Theorem 3.3.3, with

$$\psi_n(t) = c_n(1+\cos(t))^n,$$

(which is an approximate identity sequence as shown in Exercise 3.3.1, when c_n is picked so that $\int \psi_n(t)dt = 1$), we can prove the following:

Theorem 3.3.4 *The family of complex exponentials in $L_2([-\pi,\pi];\mathbb{C})$:*

$$\{e_n(t)\} = \{\frac{1}{\sqrt{2\pi}}e^{int}, \quad n \in Z\}$$

forms an orthonormal sequence which is complete.

Proof. The proof follows from Theorem 3.3.3, by writing

$$(\psi_n * f)(t) = \int \psi_n(t-\tau)f(\tau) = \int c_n(1+\cos(t-\tau))^n f(\tau)d\tau$$

Now,

$$\begin{aligned}(1+\cos(t-\tau))^n &= \sum_{k=0}^{n}\binom{n}{k}(\cos(t-\tau))^k \\ &= \sum_{k=0}^{n}\binom{n}{k}(\frac{e^{i(t-\tau)}+e^{-i(t-\tau)}}{2})^k \\ &= \sum_{k=-n}^{n}\binom{n}{k}b_{n,k}e^{ik(t-\tau)} \end{aligned} \tag{3.10}$$

for some collection of coefficients $b_{n,k}$, $k = -n, -n+1, \cdots, n$. Thus,

$$(\psi_n * f)(t) = \int \sum_{k=-n}^{n}\binom{n}{k}b_{n,k}e^{ik(t-\tau)}f(\tau)d\tau = \sum_{k=-n}^{n}e^{ikt}a_{n,k}$$

where

$$a_{n,k} := b_{n,k}\int e^{-ik\tau}f(\tau)d\tau, \qquad k = -n, -n+1, \cdots, n$$

That is, we have that, as $n \to \infty$,

$$\sum_{k=-n}^{n}e^{ikt}a_{n,k} \to f(t)$$

uniformly over $t \in [-\pi,\pi]$. Accordingly, via the arguments as in (2.9), it follows that the only vector in $L_2([-\pi,\pi];\mathbb{C})$ which is orthogonal to $\{e^{ikt}, t \in [-\pi,\pi]\}$ for $k \in \mathbb{Z}$ is the null vector and thus, the collection

$$\{e_n(t)\} = \{\frac{1}{\sqrt{2\pi}} e^{int}, \quad n \in Z\},$$

forms a complete orthonormal sequence in $L_2([-\pi, \pi]; \mathbb{C})$. □

This sequence is used for the Fourier expansion of functions in $L_2([0, 2\pi]; \mathbb{C})$; see Section 2.3.4.

3.4 Some Operations on Distributions [Optional]

While studying several properties of distributions, one typically first starts with a generalized distribution and tries to extend the properties to singular distributions.

One important property of distributions is that, every distribution has a derivative. Furthermore, we will also be taking the Fourier transform of distributions, but the derivative, once again, will have a meaning as a distribution; that is it will only have a meaning when it is applied to a class of test functions.

Definition 3.4.1 *The derivative of a distribution $\bar{\gamma} \in \mathcal{S}^*$ is defined as:*

$$(D\bar{\gamma})(\phi) = -\bar{\gamma}(\frac{d\phi}{dt}), \qquad \phi \in \mathcal{S}.$$

We can check if this definition is consistent with a distribution represented by a regular function. Consider (3.4) and note that through, integration by parts

$$\int \frac{d}{dt}\gamma(t)\phi(t)dt = -\int \gamma(t)\frac{d}{dt}\phi(t) = -\bar{\gamma}(\frac{d\phi}{dt}).$$

We can now verify that the Dirac delta distribution is the derivative of the step distribution, $\bar{u} : \mathcal{S} \to \mathbb{R}$, defined as

$$\bar{u}(f) = \int_0^\infty f(t)dt, \quad \forall f \in \mathcal{S},$$

which is an important relationship in engineering applications: e.g., the step function often models a turn-on event for a switch in circuit theory.

Let $\bar{T}$ be a distribution given by $\bar{F}(\phi) = \int F(t)\phi(t)dt$.

The convolution of F with $\phi \in \mathcal{S}$ would be:

$$\int F(\tau)\phi(t-\tau)d\tau$$

We can interpret this as a distribution in the following sense. Let $T_t(\phi)(\tau) = \phi(\tau - t)$ be the shifting operator and $Rg(x) = g(-x)$ be the inversing operator: Then,

$$\int F(\tau)\phi(t-\tau)d\tau = \int F(\tau)(RT_t\phi)(\tau) = \bar{F}((RT_t\phi))$$

This then motivates the following: The convolution of a function ϕ in $\mathcal{S}$ and a distribution $\bar{f}$ is defined by:

$$(\phi * \bar{f})(t) = \int f(\tau)\phi(t-\tau)d\tau = \bar{f}(RT_t\phi),$$

where, as before, $Rg(x) = g(-x)$ is the inversing operator and $T_t(\phi)(\tau) = \phi(\tau - t)$ is the shifting operator.

Theorem 3.4.1 *For any distribution $\bar{f}$ and ϕ in $\mathcal{S}$, $\phi * \bar{f}$ is an infinitely differentiable function and can be used to represent a regular distribution.*

Let $\bar{f}, \bar{g}$ be two regular distributions represented by f, g, respectively. The convolution of $\bar{f} \star \bar{g}$ is given by the relation, whenever this is well-defined:

$$\left(\bar{f} \star \bar{g}\right)(\phi) = \bar{f}(h_g(\phi)), \quad \forall \phi \in \mathcal{D},$$

with

$$h_g(\phi) = \int_{-\infty}^{\infty} g(\tau - t)\phi(\tau)d\tau$$

It should be observed that, with the above definition:

$$\left(\bar{f} \star \bar{\delta}\right)(\phi) = \bar{f}(\phi), \quad \forall \phi \in \mathcal{D},$$

that is the delta distribution is the identity element in distributions under the operation of convolution.

We state the following very useful result, without a formal proof.

Theorem 3.4.2 *For any singular distribution, there exists a sequence of regular distributions represented by a signal in $\mathcal{S}$ which converges to the singular distribution.*

Sketch of Proof. Let f_n be the approximate identity sequence given by (3.7), so that $f_n \in \mathcal{S}$. Then, it can be shown that for any singular $\bar{g}$, $f_n * \bar{g}$ is a smooth function, and can be used to represent a regular distribution such that $(f_n * \bar{g})(\phi) = \bar{g}(\int f_n(t - \tau)\phi(t)dt) \to \bar{g}(\phi)$ for any $\phi \in \mathcal{S}$. Furthermore, for any $\epsilon > 0$, $(f_n * \bar{g})(t)e^{-\epsilon t^2}$ is in $\mathcal{S}$ and as $n \to \infty$ and $\epsilon \to 0$,

$$(f_n * \bar{g}e^{-\epsilon t^2})(\phi) \to \bar{g}(\phi).$$

$\diamond$

3.5 Fourier Transform of Schwartz signals

We will continue the discussion of Schwartz signals in the context of Fourier transforms. One appealing aspect of Schwartz signals is that, the Fourier transform of a Schwartz signal lives in the space of Schwartz signals. In fact, the Fourier transform on the space of Schwartz signals is both onto and one to one (hence a bijection). This will be proven later. Since the space of continuous functions is dense in the space of square integrable functions, and $\mathcal{S}$ is dense in the space of continuous functions under the supremum norm by Theorem 3.3.3, we will use the bijection property of the Fourier transform on $\mathcal{S}$ to define the Fourier transform of square integrable functions.

3.6 Appendix

3.6.1 Optional: Application to Optimization Problems and the Generalization of the Projection Theorem [11]

The duality results above and Hölder's inequality are important in applications to optimization problems. The geometric ideas we reviewed in the context of the projection theorem apply very similarly to such spaces, where the inner-product is replaced by the duality pairings. Let us make this more explicit: Let for a subspace M,

$$M^{\perp} := \{x^* : \langle m, x^* \rangle = 0, \forall m \in M\}.$$

Theorem 3.6.1

(i) Let x be an element in a real normed space X and let d denote its distance from a subspace M. Then,

$$d = \inf_{m\in M} \|x - m\| = \max_{\{\|x^*\|\leq 1, x^*\in M^\perp\}} \langle x, x^*\rangle$$

If the infimum is achieved, then the maximum on the right is achieved for some x_0^* such that $x - m_0$ is aligned with x_0^*.

(ii) In particular, if m_0 satisfies

$$\|x - m_0\| \leq \|x - m\|, \forall m \in M,$$

there must be a non-zero vector $x^* \in X^*$ such that $\langle m, x_0^*\rangle = 0$ for all m and x^* is aligned with $x - m_0$.

We note that by alignment of x^* and x, it is meant that $x^*(x) = \|x^*\|\|x\|$.

Proof.

(i) For any $\epsilon > 0$, let $m_\epsilon \in M$ be such that $\|x - m_\epsilon\| \leq d + \epsilon$. For any $x^* \in M^\perp$ with $\|x^*\| \leq 1$, we have

$$\langle x, x^*\rangle = \langle x - m_\epsilon, x^*\rangle \leq \|x^*\|\|x - m_\epsilon\| \leq d + \epsilon.$$

Since $\epsilon > 0$ is arbitrary, it follows that $\langle x, x^*\rangle \leq d$.

It remains to be shown if we could find an x_0^* for which $\langle x, x_0^*\rangle = d$.

Let N be the subspace spanned by the vectors in the collection $x + M$; every vector in N can be written as $n = \alpha x + m$ for some $m \in M$ and $\alpha \in \mathbb{R}$. One can define a linear function on N by the equation

$$f(n) = \alpha d,$$

and note that this basically assigns the value zero for those vectors in N that are strictly in M.

We have:

$$\|f\| = \sup_{n\in N} \frac{|f(n)|}{\|n\|} = \sup \frac{|\alpha| d}{\|\alpha x + m\|} = \sup \frac{|\alpha| d}{|\alpha|\|x + \frac{m}{\alpha}\|} = 1,$$

since $\inf \|x + \frac{m}{\alpha}\| = d$. Now, define an extension of f from N to all of X (this is possible due to an extension theorem known as Hahn-Banach Extension theorem). Call this extension x_0^* and such an extension can be made so that $\|x_0^*\| = \|f\| = 1$ and $x_0^*(n) = f(n)$ for $n \in N$. Since $f(m) = 0$ for $m \in M$, we have that $x^*(m) = 0$ as well for $m \in M$ (so $x_0^* \in M^\perp$). Thus, $\langle x, x_0^*\rangle = d$ (since $x = x + \underline{0}$ with the view that $\underline{0} \in M$).

(ii) Now let m_0 exist so that $\|x - m_0\| = d$ and let x_0^* be any element so that $x_0^* \in M^\perp$, $\|x_0^*\| = 1$ and $\langle x, x_0^*\rangle = d$ (the construction above is an example). Then,

$$\langle x - m_0, x_0^*\rangle = \langle x, x_0^*\rangle = d = \|x_0^*\|\|x - m_0\|,$$

where the last equality holds since $\|x_0^*\| = 1$ and $\|x - m_0\| = d$. Thus, x_0^* is aligned with $x - m_0$.

$\diamond$

Theorem 3.6.2 *Let M be a subspace in a real normed space X. Let $x^* \in X^*$ be at a distance d from $M^\perp$. (i)*

$$d = \min_{m^*\in M^\perp} \|x^* - m^*\| = \sup_{x\in M, \|x\|\leq 1} \langle x, x^*\rangle,$$

where the minimum on the left is achieved for $m_0^* \in M^\perp$. (ii) If the supremum on the right is achieved for some $x_0 \in M$, then $x^* - m_0^*$ is aligned with x_0.

Proof. (i) For any $m^* \in M^\perp$, we have

$$\|x^* - m^*\| = \sup_{\|x\|\leq 1} \langle x, x^* - m^*\rangle \geq \sup_{\|x\|\leq 1, x\in M} \langle x, x^* - m^*\rangle = \sup_{\|x\|\leq 1, x\in M} \langle x, x^*\rangle =: \|x^*\|_M.$$

Thus, $\|x^* - m^*\| \geq \|x^*\|_M$. We now seek for $m_0^* \in M^\perp$ giving an equality. Consider x^* restricted to M. Let y^* be the (Hahn-Banach) extension of the restriction of x^* to M, to the whole X. Thus, $\|y^*\|_M = \|x^*\|_M$ and $(x^* - y^*)(m) = 0$ for $m \in M$. Define $m_0^* = x^* - y^*$. Then, $m_0^* \in M^\perp$ and $\|x^* - m_0^*\| = \|y^*\| = \|x^*\|_M$.

(ii) If the supremum on the right-hand side is achieved for some $x_0 \in M$, then with $\|x_0\| = 1$, $\|x^* - m_0^*\| = \langle x_0, x^* \rangle = \langle x_0, x^* - m_0^* \rangle$. Thus, $x^* - m_0^*$ is aligned with x_0. $\diamond$

An Application: Constrained Dual Optimization Problems.

Consider the following constrained optimization problem:

$$d = \min_{x^* : \langle y_i, x^* \rangle = c_i, 1 \leq i \leq n} \|x^*\|$$

Observe that if $\bar{x}^*$ is any vector satisfying the constraints, then

$$d = \min_{x^* : \langle y_i, x^* \rangle = c_i, 1 \leq i \leq n} \|x^*\| = \min_{m^* \in M^\perp} \|\bar{x}^* - m^*\|,$$

where M denotes the space spanned by $\{y_1, y_2, \cdots, y_n\}$ and $\bar{x}^*$ is some vector satisfying the constraints.

From Theorem 3.6.2, we have that

$$d = \min_{m^* \in M^\perp} \|\bar{x}^* - m^*\| = \sup_{x \in M, \|x\| \leq 1} \langle x, \bar{x}^* \rangle,$$

Now, any vector in M is of the form $m = Ya$ where $Y = [y_1 \quad y_2 \quad \cdots \quad y_n]$ is a matrix and a is a column vector. Thus,

$$d = \min_{x^* : \langle y_i, x^* \rangle = c_i, 1 \leq i \leq n} \|x^*| = \sup_{\|Ya\| \leq 1} \langle Ya, \bar{x}^* \rangle = \sup_{\|Ya\| \leq 1} c^T a,$$

where the last equality follows because $\bar{x}^*$ satisfies the constraints and that

$$\langle Ya, \bar{x}^* \rangle = \langle a, Y^T \bar{x}^* \rangle = c^T a$$

Thus, the optimal solution to the constrained problem can be written as

$$\sup_{\|Ya\| \leq 1} c^T a,$$

where the optimal x^* is aligned with the optimal Ya.

3.7 Exercises

Exercise 3.7.1 *Does there exist a sequence of functions $\{f_j\}$ in $L_2(\mathbb{R}_+; \mathbb{R})$ such that a sequence of distributions $\bar{f}_j$ represented by f_j on the set of Schwartz signals $\mathcal{S}$ converges to zero in a distributional sense, but f_j does not converge to zero (that is, in the L_2 norm). That is, does there exist a sequence of functions $\{f_j\}$ in $L_2(\mathbb{R}_+; \mathbb{R})$ such that*

$$\lim_{j \to \infty} \left(\int_0^\infty |f_j(t)|^2 dt \right)$$

is not zero, but

$$\lim_{j \to \infty} \left(\int_0^\infty f_j(t) \phi(t) dt \right) = 0, \quad \forall \phi \in \mathcal{S}.$$

If there exists one, give an example. If there does not exist one, explain why.

Exercise 3.7.2 *a) Let T be a mapping from $L_2(\mathbb{R}_+;\mathbb{R})$ to $\mathbb{R}$ (extended to possibly include $-\infty,\infty$) given by:*

$$T(f)=\int_{\mathbb{R}_+} f(t)\frac{t}{1+t^2}dt$$

Let $f_0\in L_2(\mathbb{R}_+;\mathbb{R})$ be given by:

$$f_0(t)=\frac{1}{t^2+1},\quad \forall t\in\mathbb{R}_+.$$

Is T continuous on $L_2(\mathbb{R}_+;\mathbb{R})$ at f_0?

b) Let $\mathcal{S}$ be the space of Schwartz signals. Let $T:\mathcal{S}\to\mathbb{R}$ be a mapping given by:

$$T(\phi)=\phi'(0),\quad \phi\in\mathcal{S},$$

where

$$\phi'(t)=\frac{d}{dt}\phi(t)\quad \forall t.$$

Is T a distribution on $\mathcal{S}$? That is, is T continuous and linear on $\mathcal{S}$?

Exercise 3.7.3 *Let $T:\mathcal{S}\to[-\infty,\infty]$ be a mapping defined by:*

$$T(\phi)=\limsup_{A\to\infty}\int_{-A}^{A}\phi(t)e^{t^2}dt$$

Is T continuous on $\mathcal{S}$? Prove your argument.

Hint: The function $g(t)=e^{-at^2}$ is in $\mathcal{S}$, for any $a>0$.

Exercise 3.7.4 *Let for $j\in\mathbb{N}$,*

$$f_j(t)=\begin{cases} j, & \text{if } 0\le t\le \frac{1}{j}\\ 0 & \text{else}\end{cases}$$

For $g\in\mathcal{S}$, define

$$\bar{f}_j(g):=\int_0^\infty f_j(t)g(t)dt.$$

Show that $\bar{f}_j(.)$ is a distribution on $\mathcal{S}$. Show that

$$\lim_{j\to\infty}\int_0^\infty f_j(t)g(t)dt=\bar{\delta}(g)=g(0).$$

Conclude that, the sequence of regular *distributions $\bar{f}_j(.)$, represented by a real-valued, integrable function $f_j(t)$, converges to the delta distribution $\bar{\delta}(.)$ on the space of test signals $\mathcal{S}$.*

Exercise 3.7.5 *Let $\mathcal{S}$ be the space of Schwartz signals. Let $T:\mathcal{S}\to\mathbb{R}$ be a mapping given by:*

$$T(\phi)=\phi'(0),\quad \phi\in\mathcal{S},$$

where

$$\phi'(t) = \frac{d}{dt}\phi(t) \quad \forall t.$$

Is T a distribution on $\mathcal{S}$? That is, is T continuous and linear on $\mathcal{S}$?

4

Systems

An input-output system is defined by an input signal set $\mathcal{U}$, an output set $\mathcal{Y}$ and a subset $\mathcal{R} \subset \mathcal{U} \times \mathcal{Y}$, called the rule (relation) of the system. Hence, $\mathcal{R}$ consists of the input-output pairs in the system. We often find it convenient to associate with $\mathcal{R}$ a transformation or map $\mathcal{T}$ so that $y = \mathcal{T}(u)$ and thus $\mathcal{R} = \{(u, \mathcal{T}(u)), \quad u \in \mathcal{U}\}$.

Accordingly, we have that if $(u, y^1) \in \mathcal{R}$ and $(u, y^2) \in \mathcal{R}$ then $y^1 = y^2$.

Let T_1, T_2 be time-index sets; and U, Y be signal range spaces such that $\mathcal{U} = U^{T_1}, \mathcal{Y} = Y^{T_2}$, that is:

$$\mathcal{U} = \{f : T_1 \to U\},$$

$$\mathcal{Y} = \{f : T_2 \to Y\}.$$

If $\mathcal{U}$ and $\mathcal{Y}$ consist of signals with discrete-time indices, then the system is said to be a discrete-time (DT) system. Of the indices are both continuous, then the system is a continuous-time (CT) system. If one of them is discrete and the other continuous, the system is said to be hybrid. Often, we have $T = T_1 = T_2$, which will be assumed in the following.

4.1 System Properties

Memorylessness. Let U be a input signal range, Y an output signal range, and a time index T. If any input output pair $(u, \mathcal{T}(u))$ can be written component-wise as

$$y_t = \Psi(t, u_t), \quad t \in T$$

for some fixed function $\Psi : T \times U \to Y$, then the system is memoryless.

Causality/Non-anticipativeness. If the output at any time t is not dependent on the input signal values at time $s > t$, then the system is non-anticipative (causal). That is, let $u^1 = \{u_t^1, t \in T\}$ and $u^2 = \{u_t^2, t \in T\}$. Let $(u^1, y^1) \in \mathcal{R}$ and $(u^2, y^2) \in \mathcal{R}$. If it is that $u_s^1 = u_s^2$ for $s \leq t$, then, for a causal system, it must be that $y_t^1 = y_t^2$.

Example 4.1. Let a relation be given by $y_t = ax_{t+1} + x_t + bx_{t-1}$. Such a system is causal if $a = 0$; it is memoryless if $a = 0, b = 0$.

Time-Invariance. A system is time-invariant if for every input-output pair ($(u, y) \in \mathcal{R}$): and time-shift in the input leads to the same time-shift in the output:

$$(\sigma^\theta u, \sigma^\theta y) \in \mathcal{R},$$

where we define a time-shift as follows: With $T = \mathbb{Z}$ or $\mathbb{R}$, let $\theta \in T$. We define $\sigma^\theta : \mathcal{U} \to \mathcal{U}$ with

$$\Big(\sigma^\theta(u)\Big)_t = u_{t+\theta}, \quad \forall t \in T$$

You are encouraged to visualize the time shift above: σ^θ pushes a signal to the left by θ.

4.2 Linear Systems

4.2.1 Representation of Discrete-Time Signals in terms of Unit Pulses

Let $x \in \Gamma(\{0, 1, \ldots, N-1\}; \mathbb{R})$. Then, one can represent x pointwise as

$$x(n) = \sum_{i=0}^{N-1} x(i)\delta(n-i)$$

or

$$x(n) = \sum_{i=0}^{N-1} x(i)\delta_i(n)$$

where δ_i is the shifted unit pulse given by:

$$\delta_i(n) = 1_{\{n=i\}}, \quad n \in \mathbb{Z},$$

with $1_{\{.\}}$ denoting the indicator function.

4.2.2 Linear Systems

Linear systems have important engineering practice. Many physical systems are locally linear, as we have seen earlier. An input-output system is linear if $\mathcal{U}, \mathcal{Y}, \mathcal{R}$ are all linear vector spaces.

In this course, we will say that a discrete-time (DT) system is linear if the input output relation can be written as:

$$y(n) = \sum_{m=-\infty}^{\infty} k(n,m)u(m). \tag{4.1}$$

In the above, $k(n,m)$ is called the kernel of the system. The value $k(n,m)$ reveals to effect of an input at time m to the output at time n. If in addition to linearity, one imposes time-invariance, the resulting system becomes a convolution system, as we will shortly observe.

We say a continuous-time (CT) system is linear if the input output relation can be expressed as

$$y(t) = \int_{\tau=-\infty}^{\infty} h(t,\tau)u(\tau)d\tau$$

We note here that a precise characterization for linearity (for a system as in (4.1)) would require the interpretation of a system as a (bounded) linear operator from one space to another space. One can obtain a Riesz representation theorem type characterization leading to (4.1), provided that $\mathcal{U}$ and $\mathcal{Y}$ satisfy certain properties, and the system is continuous and linear. The following exercise is an example.

Exercise 4.2. Let $\mathcal{T}$ be a linear system mapping $l_1(\mathbb{Z}; \mathbb{R})$ to $l_1(\mathbb{Z}; \mathbb{R})$. Show that if this system is linear and continuous, it can be written so that $y = \mathcal{T}(u)$;

$$y(n) = \sum_{m\in\mathbb{Z}} h(n,m)u(m), \quad n \in \mathbb{Z}$$

for some $h : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$.

Since $u \in l_1(\mathbb{Z}; \mathbb{R})$, we can write $u = \lim_{N\to\infty} u_N$ (that is, $\|u - u_N\|_1 \to 0$ as $n \to \infty$), where $u_N = \sum_{i=-N}^{N} u_i\delta_i$ with

$$\delta_i(n) = 1_{\{n=i\}}$$

Since T is linear,

$$T(u) = \mathcal{T}(\lim_{N\to\infty} u_N) = \lim_{N\to\infty} \mathcal{T}(u_N) = \lim_{N\to\infty} \sum_{i=-N}^{N} u_i T(\delta_i).$$

By continuity, we have that the limit exists and is

$$y(n) = \lim_{N\to\infty} \sum_{i=-N}^{N} u_i T(\delta_i)(n).$$

Writing $h(n,m) = T(\delta_m)(n)$, the result follows. Observe that $h(n,m) = T(\delta_m)(n)$ is the output of the system at time n when the input to the system is a unit pulse applied at time m.

4.3 Linear and Time-Invariant (Convolution) Systems

If, in addition to linearity, we wish to have time-invariance, then one can show that

$$y(n) = \sum_{k=-\infty}^{\infty} k(n,m)u(m),$$

will have to be such that $k(n,m)$ should be dependent only on $n-m$. This follows from the fact that a shift in the input would have to lead to the same shift in the output, implying that $k(n,m) = k(n+\theta, m+\theta)$ for any $\theta \in \mathbb{Z}$.

Let us discuss this further. Suppose a linear system described by

$$y(n) = \sum_{m} k(n,m)u(m), \tag{4.2}$$

is time-invariant. Let, for some $\theta \in \mathbb{Z}$,

$$v = \sigma^{-\theta}(u)$$

$v(m) = u(m-\theta)$. Let the signal g be the output of the system when the input is the discrete-time signal v. It follows that

$$\begin{aligned} g(n) &= \sum_{m\in\mathbb{Z}} k(n,m)v(m) \\ &= \sum_{m\in\mathbb{Z}} k(n,m)u(m-\theta) \\ &= \sum_{m'\in\mathbb{Z}} k(n,m'+\theta)u(m') \end{aligned}$$

By time-invariance, it must be that $g = \sigma^{-\theta}(y)$. That is, $g(n) = y(n-\theta)$ or $g(n+\theta) = y(n)$. Thus,

$$g(n+\theta) = \sum_{m'\in\mathbb{Z}} k(n+\theta, m'+\theta)u(m') = y(n) \tag{4.3}$$

Since the equivalence in (4.2)-(4.3) above has to hold for every input signal, it must be that $k(n+\theta, m+\theta) = k(n,m)$ for all n, m values, and for all θ values. Therefore $k(n,m)$ should only be a function of the difference $n-m$. Hence, a linear system is time-invariant if and only if the input-output relation can be written as:

$$y(n) = \sum_{m=-\infty}^{\infty} h(n-m)u(m)$$

for some function $h : \mathbb{Z} \to \mathbb{R}$. This function is called the impulse-response of the system since, if $u = \delta_0$, then

$$y(n) = \sum_{m=-\infty}^{\infty} h(n-m)\delta_0(m) = h(n)$$

Due to this representation, linear time-invariant systems are also called convolution systems.

One can show that a convolution system is non-anticipative if $h(n) = 0$ for $n < 0$.

Similar discussions apply to continuous-time systems by replacing the summation with integrals:

$$y(t) = \int_{\tau=-\infty}^{\infty} h(t-\tau)u(\tau)d\tau$$

The function h is the output of the system when the input is an impulse function. This is why h is called the *impulse response* of a convolution system.

Exercise 4.3.1 *Let $x(t) \in \mathbb{R}^N$ and $t \geq 0$ and real-valued. Recall that the solution to the following differential equation:*

$$x'(t) = Ax(t) + Bu(t),$$

$$y(t) = Cx(t),$$

with the initial condition $x(t_0) = x_0$ is given by

$$x(t) = e^{A(t-t_0)}x_{t_0} + \int_{\tau=t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau, \quad t \geq 0$$

(a) Suppose that $x(t_0) = 0$ and all eigenvalues of A have their real parts as negative and $\|u\|_\infty < \infty$. Let $t_0 \to -\infty$. Show that if one is to represent $x(t) = (h * u)(t)$, we have

$$h(t) = Ce^{At}B1_{\{t\geq 0\}}.$$

(b) Alternatively, we could skip the condition that the eigenvalues of A have their real parts as negative, but require that $x(0) = 0$ and $u(t) = 0$ for $t < 0$. Express the solution as a convolution

$$x(t) = (h * u)(t),$$

and find $h(t)$.

(c) Let $y(t) = Cx(t) + Du(t)$. Repeat the above.

Exercise 4.3.2 *Let $x(n) \in \mathbb{R}^N$ and $n \in \mathbb{Z}$. Consider a linear system given by*

$$x(n+1) = Ax(n) + Bu(n)$$

$$y(n) = Cx(n), \quad n \geq 0$$

with the initial condition $x(n_0) = 0$ for some n_0. a) Suppose all the eigenvalues of A are strictly inside the unit disk in the complex plane and $\|u\|_\infty < \infty$. Let $n_0 \to -\infty$. Express the solution $y(n)$ as a convolution

$$y(n) = (h * u)(n),$$

and find that

$$h(n) = CA^{n-1}B1_{\{n\geq 1\}}.$$

b) Alternatively, we could skip the condition that the eigenvalues of A are strictly inside the unit disk in the complex plane, but require that $n_0 = 0$ so that $x(0) = 0$ and also $u(n) = 0$ for $n < 0$. Express the solution as a convolution

$$x(n) = (h * u)(n),$$

and find $h(n)$.

(c) Let $y(n) = Cx(n) + Du(n)$. Repeat the above.

4.4 Bounded-Input-Bounded-Output (BIBO) Stability of Convolution Systems

A DT system is BIBO stable if $||u||_\infty := \sup_{m\in Z} |u(m)| < \infty$ implies that $||y||_\infty := \sup_{m\in Z} |y(m)| < \infty$.

A CT system is BIBO stable if $||u||_\infty := \sup_{t\in R} |u(m)| < \infty$ implies that $||y||_\infty := \sup_{t\in R} |y(t)| < \infty$.

Theorem 4.4.1 *A convolution system is BIBO stable if and only if*

$$||h||_1 < \infty$$

4.5 The Frequency Response (or Transfer) Function of Linear Time-Invariant Systems

A very important property of convolution systems is that, if the input is a harmonic function, so is the output:

Let $u \in L_\infty(\mathbb{R};\mathbb{R})$. If

$$u(t) = e^{i2\pi ft},$$

then

$$y(t) = \left(\int_{-\infty}^{\infty} h(s)e^{-i2\pi fs}ds\right)e^{i2\pi ft}$$

We define:

$$\hat{h}(f) := \left(\int_{-\infty}^{\infty} h(s)e^{-i2\pi fs}ds\right),$$

and call this value the *frequency response* of the system for frequency f, whenever it exists. We often call this function the *transfer function* of the system.

A similar discussion applies for a discrete-time system.

Let $h \in l_1(\mathbb{Z};\mathbb{R})$. If $u(n) = e^{i2\pi fn}$, then

$$y(n) = \left(\sum_{m=-\infty}^{\infty} h(m)e^{-i2\pi fm}\right)e^{i2\pi fm}$$

with the frequency response function

$$\hat{h}(f) := \left(\sum_{m=-\infty}^{\infty} h(m)e^{-i2\pi fm}\right)$$

Convolution systems are used as filters through the characteristics of the frequency response. In class, examples will be presented involving the resistor-capacitor circuits and resistor-capacitor-inductor circuits.

4.6 Steady-State vs. Transient Solutions

Let $x(t) \in \mathbb{R}^N$. Consider a system defined with the relation:

$$x'(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + Du(t), \qquad t \geq t_0,$$

for some fixed t_0. Consider an input $u(t) = e^{st}, t \geq t_0$ for some $s \in \mathbb{C}$ (for the time being, assume that $s = i2\pi f$ for some $f \in \mathbb{R}$). Suppose that s is not an eigenvalue of A. Using the relation

$$x(t) = e^{A(t-t_0)}x(t_0) + e^{At}\int_{t_0}^{t} e^{-A\tau}Be^{s\tau}d\tau = e^{A(t-t_0)}x(t_0) + e^{At}\int_{t_0}^{t} e^{s\tau}e^{-A\tau}Bd\tau$$

we obtain

$$x(t) = \left(e^{A(t-t_0)}x(t_0) - e^{A(t-t_0)}e^{At_0}(sI-A)^{-1}e^{-At_0}e^{st_0}B\right) + \left(e^{At}(sI-A)^{-1}e^{-At}e^{st}B\right)$$

Using the property that for any t

$$e^{At}(sI-A)^{-1}e^{-At} = (sI-A)^{-1},$$

we obtain

$$y(t) = Ce^{A(t-t_0)}\left(x(t_0) - (sI-A)^{-1}e^{st_0}B\right) + \left(C(sI-A)^{-1}B + D\right)e^{st}, \qquad t \in \mathbb{R}_+$$

The first term is called the transient response of the system and the second term is called the steady-state response.

If A is a stable matrix, with all its eigenvalues inside the unit circle, the first term decays to zero as t increases (or with fixed t, as $t_0 \to -\infty$). Alternatively, if we set $t_0 = 0$ and write

$$x(0) = (sI-A)^{-1}B,$$

the output becomes

$$y(t) = \left(C(sI-A)^{-1}B + D\right)e^{st}, \qquad t \geq t_0$$

The map $C(sI-A)^{-1}B + D$ is called the *transfer function* of the system. When $s = i2\pi f$, this is the frequency response.

Bu direct computation in the integration formula, we can show that if s were an eigenvalue of A, then the steady-state output would be $p_s(t)e^{At}$ for some polynomial p_s.

The case with $s = i2\pi ft$ is crucial for stable systems. Later on we will investigate the more general case with $s \in \mathbb{C}$.

4.7 Bode Plots for Studying System Response to Harmonic Inputs

If we apply $u(t) = e^{i2\pi ft}$ or $e^{i\omega t}$, we observed in the above that the output would be $\hat{h}(f)e^{i2\pi ft}$.

Bode plots allow us to efficiently visualize $\hat{h}(f)$ by depicting the magnitude and phase, in a logarithmic scale; in the pre-digital era under the absence of computers such plots were effective means to represent transfer functions with the logarithmic scale in mid-20th century.

Observe that since $\hat{h}(f) = \overline{\hat{h}(-f)}$, it suffices to consider only $f \geq 0$. Let $\omega = 2\pi f$. Let $i = 1, 2, \cdots, 5$ and $s_i = r_i e^{i\theta_i}$ where $r_i = |s_i|$ and θ_i is the phase of s_i.

$$h(i\omega) = \frac{s_1 s_2}{s_3 s_4 s_5} = \left(\frac{r_1 r_2}{r_3 r_4 r_5}\right)e^{i(\theta_1+\theta_2-\theta_3-\theta_4-\theta_5)}$$

Thus,

$$|h(i\omega)| = \frac{r_1 r_2}{r_3 r_4 r_5}$$

and

$$\log(|h(i\omega)|) = \log(r_1) + \log(r_2) - \log(r_3) - \log(r_4) - \log(r_5)$$

Note also that

$$\log(e^{i(\theta_1+\theta_2-\theta_3-\theta_4-\theta_5)}) = i(\theta_1 + \theta_2 - \theta_3 - \theta_4 - \theta_5)$$

so that

$$\angle h(i\omega) = \theta_1 + \theta_2 - \theta_3 - \theta_4 - \theta_5$$

Thus, the logarithms allow us to consider the contributions of each complex number in an additive fashion both for the magnitude and the phase.

Now, consider

$$h(i\omega) = K(i\omega)^n \frac{1 + i\frac{\omega}{\omega_0}}{(i\frac{\omega}{\omega_n})^2 + 2\zeta i\frac{i\omega}{\omega_n} + 1}$$

We can thus consider the contributions of $K(i\omega)^n$, $1 + i\frac{\omega}{\omega_0}$, and $(i\frac{\omega}{\omega_n})^2 + 2\zeta i\frac{i\omega}{\omega_n} + 1$ separately.

For $K(i\omega)^n$, we note that

$$\log|K(i\omega)^n| = \log(|K|) + n\log(\omega)$$

and

$$\angle K(i\omega)^n = \angle K + n\frac{\pi}{2}$$

For $1 + i\frac{\omega}{\omega_0}$, we use the following approximations: for $\omega \approx 0$, $1 + i\frac{\omega}{\omega_0} \approx 1$. For $\omega = \omega_0$, $1 + i\frac{\omega}{\omega_0} = 1 + i$. For $\omega \gg \omega_0$, $|1 + i\frac{\omega}{\omega_0}| \approx \frac{|\omega|}{\omega_0}$

Likewise, for the angle: for $\omega \approx 0$:

$$\angle 1 + i\frac{\omega}{\omega_0} \approx 0$$

For $\omega \gg \omega_0$,

$$\angle 1 + i\frac{\omega}{\omega_0} \approx \frac{\pi}{2}$$

At $\omega = \omega_0$, $\angle 1 + i\frac{\omega}{\omega_0} = \frac{\pi}{4}$.

Finally, for

$$((i\frac{\omega}{\omega_n})^2 + 2i\zeta\frac{\omega}{\omega_n} + 1)^{-1}$$

For $\omega \approx 0$, the magnitude is approximately 1, with its logarithm approximately 0. For $\omega = \omega_n$, the magnitude is $\frac{1}{2\zeta}$. For $\omega \gg \omega_n$, the magnitude decays as $-2\log(|\omega|)$.

For the phase: for $\omega \approx 0$, the phase is approximately 0. For $\omega = \omega_n$, the phase is $-\frac{\pi}{2}$. For $\omega \gg \omega_n$, the phase is close to π.

Bode plots approximate these expressions in a log-log plot (for the magnitude).

4.8 Interconnections of Systems and Feedback Control Systems

We will discuss serial connections, parallel connections, output and error feedback connections.

Control systems are those who input-output behaviour is shaped by control laws (typically through using system outputs to generate the control inputs –termed, output feedback–) so that desired system properties such as stability, robustness to incorrect models, robustness (to system or measurement noise) -also called, disturbance rejection-, tracking a given reference signal, and ultimately, optimal performance are attained. These will be made precise as control theoretic applications are investigated further.

4.9 State-Space Description of Linear Systems

We will study state-space realizations of linear time-invariant systems in further detail in *Chapter* 9. We provide a brief discussion in the following.

4.9.1 Principle of superposition

For a linear time invariant system, if (u, y) is an input-output pair, then $\sigma_\theta u, \sigma_\theta y$ is also an input-output pair and thus, $a_1 u + b_1 \sigma_\theta u, a_1 y + b_1 \sigma_\theta y$ is also such a pair.

4.9.2 State-space description of input-output systems

The notion of a state. Suppose that we wish to compute the output of a system at $t \geq t_0$ for some t_0. In a general (causal) system, we need to use all the past applied input terms $u(s), s \leq t_0$ and all the past output values $y(s), s < t_0$ to compute the output at t_0. The *state* of a system summarizes all the past relevant data that is sufficient to compute the future paths. Some systems admit a finite-dimensional state representation, some do not. In the following, we consider continuous-time and discrete-time systems where a finite-dimensional state representation can be made.

Consider a continuous-time system given by:

$$\sum_{k=0}^{N} a_k \frac{d^k}{dt^k} y(t) = \sum_{m=0}^{N-1} b_m \frac{d^m}{dt^m} u(t),$$

with $a_N = 1$. Such a system can be written in the form:

$$\frac{d}{dt} x(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & 0 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{N-1} \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

$$C = \begin{bmatrix} b_N & b_{N-1} & \cdots & b_1 \end{bmatrix}$$

Likewise, consider a discrete-time system of the form:

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{m=1}^{N} b_m u(n-m)$$

with $a_0 = 1$, can be written in the form

$$x(n+1) = Ax(n) + Bu(n), \qquad y(n) = Cx(n)$$

where

$$x_N(n) = y(n), x_{N-1}(n) = y(n-1), \quad \cdots, x_1(n) = y(n-(N-1))$$

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & 1 \\ -a_N & -a_{N-1} & -a_{N-2} & \cdots & -a_1 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

$$C = \begin{bmatrix} b_N & b_{N-1} & \cdots & b_1 \end{bmatrix}$$

4.9.3 Stability of linear systems described by state equations

Theorem 4.9.1 *For a linear differential equation*

$$x' = Ax + u,$$

the system is BIBO stable if and only if

$$\max_{\lambda_i} \{\mathrm{Re}\{\lambda_i\}\} < 0,$$

where $\mathrm{Re}\{.\}$ *denotes the real part of a complex number, and* λ_i *denotes the eigenvalues of A.*

Theorem 4.9.2 *For a linear differential equation*

$$x(n+1) = Ax(n) + u(n),$$

the system is BIBO stable if and only if

$$\max_{\lambda_i} \{|\lambda_i|\} < 1,$$

where λ_i *denotes the eigenvalues of A.*

4.10 Exercises

Exercise 4.10.1 *Consider a linear system described by the relation:*

$$y(n) = \sum_{m \in \mathbb{Z}} h(n,m)u(m), \quad n \in \mathbb{Z}$$

for some $h : \mathbb{Z} \times \mathbb{Z} \to \mathbb{C}$.

a) When is such a system causal?

b) Show that such a system is time-invariant if and only if it is a convolution system.

Exercise 4.10.2 *Let $x(t) \in \mathbb{R}^N$ and $t \geq 0$ and real-valued. Recall that the solution to the following differential equation:*

$$x'(t) = Ax(t) + Bu(t)$$

with the initial condition $x(0) = x_0$ is given by

$$x(t) = e^{A(t)}x_0 + \int_{\tau=0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau, \quad t \geq t_0$$

Suppose $x(0) = 0$ and $u(t) = 0$ for $t < 0$. Express the solution as a convolution

$$x(t) = (h * u)(t),$$

and find $h(t)$.

Note: With the assumption that the system is stable, we can avoid the condition that $u(t) = 0$ for $t < 0$. In this case, we are able to write

$$x(t) = e^{At-t_0}x(t_0) + \int_{\tau=t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau,$$

and take the limit as $t_0 \to -\infty$, leading to $h(t) = e^{At}B1_{\{t\geq 0\}}$.

Exercise 4.10.3 *Let $x(n) \in \mathbb{R}^N$ and $n \in \mathbb{Z}$. Consider a linear system given by*

$$x(n+1) = Ax(n) + Bu(n), \quad n \geq 0$$

with the initial condition $x(0) = 0$. Suppose $x(0) = 0$ and $u(n) = 0$ for $n < 0$. Express the solution $x(n)$ as a convolution

$$x(n) = (h * u)(n),$$

and find $h(n)$.

Note: With the assumption that the system is stable, we can avoid the condition that $u(n) = 0$ for $n < 0$.In this case, we can write

$$x(n) = A^{n-n_0}x(n_0) + \sum_{m=n_0}^{n-1} A^{n-m-1}Bu(m),$$

and take the limit as $n_0 \to -\infty$ leading to $h(n) = A^{n-1}B1_{\{n\geq 1\}}$.

Exercise 4.10.4 *Consider a continuous-time system described by the equation:*

$$\frac{dy(t)}{dt} = ay(t) + u(t), \quad t \in \mathbb{R},$$

where $a < 0$.

a) Find the impulse response of the system. Is the system bounded-input-bounded-output (BIBO) stable?

b) Suppose that the input to this system is given by $\cos(2\pi f_0 t)$. Let y_{f_0} be the output of the system. Find $y_{f_0}(t)$.

c) If exists, find

$$\lim_{f_0 \to \infty} y_{f_0}(t),$$

for all $t \in \mathbb{R}_+$.

Exercise 4.10.5 *Consider a discrete-time system described by the equation:*

$$y(n+1) = a_1 y(n) + a_2 y(n-1) + u(n), \quad n \in \mathbb{Z},$$

a) Is this system linear? Time-invariant?

b) For what values of a_1, a_2 is the system BIBO (bounded-input-bounded-output) stable?

Exercise 4.10.6 (Stability of Linear Time-Varying Systems) *Let T be a linear system mapping with the representation;*

$$y(n) = \sum_{m \in \mathbb{Z}} h(n,m)u(m), \quad n \in \mathbb{Z}$$

for some $h : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$. Show that this system is BIBO stable if

$$\sup_n \sum_m |h(n,m)| < \infty.$$

Let us define a system to be regularly BIBO stable if for $\epsilon > 0$, $\exists \delta > 0$ such that $\|u\|_\infty \leq \delta$ implies that $\|y\|_{infty} \leq \epsilon$. Show that the system above is regularly BIBO stable if and only if

$$\sup_n \sum_m |h(n,m)| < \infty.$$

Exercise 4.10.7 *Let T be a linear system mapping $l_1(\mathbb{Z};\mathbb{R})$ to $l_1(\mathbb{Z};\mathbb{R})$. Show that this system is linear and continuous only if the system can be written so that with $y = T(u)$;*

$$y(n) = \sum_{m \in \mathbb{Z}} h(n,m)u(m), \quad n \in \mathbb{Z}$$

for some $h : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$.

5

The Fourier Transformation

We have seen in *Chapter*s 2 and 3 that complex exponentials can be used to approximate any square integrable signal arbitrarily well. We saw in *Chapter* 4 that complex exponentials possess the eigenfunction property for linear time-invariant systems.

The above motivate the use of the representation of signals in terms of complex exponentials and the spectral properties of transfer functions. This study is achieved by Fourier transforms.

Accordingly, Fourier transforms play a significant role in systems theory and applied mathematics at large. There are four types of Fourier transformations. Discrete-to-Discrete (DDFT), Continuous-to-Discrete (CDFT), Discrete-to-Continuous (DCFT), Continuous-to-Continuous (CCFT).

We will start with the first two. Before we proceed, recall that a bijective transformation $\mathcal{T}$ is a map from a signal set X to another one $\hat{X}$, such that $\mathcal{T}$ is onto and one-to-one; the Fourier transform will constitute examples of such transformations with further very useful structural and regularity properties to be studied in this chapter and beyond.

5.1 Discrete-to-Discrete (DDFT) and Continuous-to-Discrete (CDFT) Fourier transforms

5.1.1 Fourier Series Expansions

Discrete Time

The N-dimensional complex vector space $l_2(\{0, 1, 2 \ldots, N-1\}; \mathbb{C})$ is a Hilbert space with the inner product:

$$\langle h_1, h_2 \rangle = \sum_{n=0}^{N-1} h_1(n)\overline{h_2(n)},$$

where the bar notation $\overline{(.)}$ denotes the complex conjugate of its argument.

The set of complex harmonic signals:

$$\frac{1}{\sqrt{N}} e^{i2\pi kn}, \quad k \in \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\},$$

provides a complete orthonormal sequence, hence, provides a basis for $l_2(\{0, 1, 2 \ldots, N-1\}; \mathbb{C})$. The Fourier series expansion is given by:

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \hat{x}(\frac{k}{N}) e^{i2\pi \frac{k}{N} n}, \quad n \in \{0, 1, \ldots, N-1\}$$

where

$$\hat{x}(\frac{k}{N}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) e^{-i2\pi \frac{k}{N} n}, \quad k \in \{0, 1, \ldots, N-1\} \tag{5.1}$$

Continuous Time

The complex vector space $L_2([0, P]; \mathbb{C})$ is a Hilbert space with the inner product:

$$\langle h_1, h_2 \rangle = \int_0^P h_1(t) \overline{h_2(t)} dt,$$

where the bar denotes the complex conjugate.

The countably infinite sequence of complex harmonic signals:

$$\frac{1}{\sqrt{P}} e^{i2\pi f n}, \quad f \in \{\frac{k}{P}, k \in Z\},$$

provides a complete orthogonal sequence, hence, provides a basis for $L_2([0, P]; \mathbb{C})$.

The completeness of Fourier series in $L_2[0, P]$ is based on the argument that trigonometric polynomials are dense in $L_2([0, P]; \mathbb{C})$ (see Theorem 3.3.4) and the discussion in Section 3.3.2.

The Fourier series expansion is given by:

$$x(t) = \frac{1}{\sqrt{P}} \sum_k \hat{x}(\frac{k}{P}) e^{i2\pi \frac{k}{P} t}, \quad t \in [0, P]$$

where

$$\hat{x}(f) = \frac{1}{\sqrt{P}} \int_0^P x(t) e^{-i2\pi f t} dt, \quad f \in \{\frac{k}{P}, k \in \mathbb{Z}\} \tag{5.2}$$

Thus, in the context of Section 2.3, a Fourier series expansion is the representation of a signal in terms of the collection of a complete orthonormal harmonic signal sequence.

5.1.2 Discrete-to-Discrete (DDFT) and Continuous-to-Discrete (CDFT) Fourier transforms

In view of the above, we define Discrete-to-Discrete (DDFT), Continuous-to-Discrete (CDFT) as follows:

DDFT:

$$\mathcal{F}_{DD} : l_2(\{0, 1, \ldots, N-1\}; \mathbb{C}) \to l_2(\{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N-1}{N}\}; \mathbb{C})$$
$$\hat{x} = \mathcal{F}_{DD}(x)$$

and

$$\hat{x}(f) = \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} x(n) e^{-i2\pi f n}, \quad f \in \{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$$

CDFT:

$$\mathcal{F}_{CD} : L_2([0,P];\mathbb{C}) \to l_2(\mathbb{Z}\frac{1}{P};\mathbb{C})$$
$$\hat{x} = \mathcal{F}_{CD}(x)$$

and

$$\hat{x}(f) = \int_{\tau=0}^{P} \frac{1}{\sqrt{P}} x(\tau) e^{-i2\pi f\tau} d\tau, \quad f \in \{\frac{k}{P}, k \in \mathbb{Z}\}$$

The inverses of these can also be defined:

Inverse DDFT:

$$\mathcal{F}_{DD}^{-1} : l_2(\{0/N, 1/N, \ldots, (N-1)/N\};\mathbb{C}) \to l_2(\{0,1,\ldots,N-1\};\mathbb{C})$$
$$x = \mathcal{F}_{DD}^{-1}(\hat{x})$$

and

$$x(n) = \sum_{k=0}^{N-1} \frac{1}{\sqrt{N}} \hat{x}(\frac{k}{N}) e^{i2\pi \frac{k}{N} n}, \quad n \in \{0,1,\ldots,N-1\}\}$$

Inverse CDFT:

$$\mathcal{F}_{DD}^{-1} : l_2(\mathbb{Z}\frac{1}{P};\mathbb{C}) \to L_2([0,P];\mathbb{C})$$
$$x = \mathcal{F}_{CD}^{-1}(\hat{x})$$

and

$$x(t) = \sum_{k\in Z} \frac{1}{\sqrt{P}} \hat{x}(\frac{k}{P}) e^{i2\pi \frac{k}{P} t} dt, t \in [0,P]$$

Exercise 5.1.1 *a) For some $N \in \mathbb{N}$, let $x \in l_2\Big((-N,-(N-1),\cdots,N-1,N);\mathbb{C}\Big)$ with*

$$x(n) = 1_{\{|n|\le N_1\}}$$

Find the Fourier series expansion of x. Sudy the case with $N_1 = N$, and the case with $N_1 = 0$.

b) For some $T \in \mathbb{R}_+$, let $x \in L_2([-\frac{T}{2},\frac{T}{2}];\mathbb{C})$ with

$$x(t) = 1_{\{|t|\le T_1\}}$$

Find the Fourier series expansion of x. Study the case with $T_1 = \frac{T}{2}$, and the case with $T_1 = \frac{1}{n}$,

$$x_n(t) = n1_{\{|t|\le T_1\}},$$

as $n \to \infty$.

Solution. *a) We have for $k = 0,1,\cdots,2N+1$,*

$$\hat{x}(\frac{k}{2N+1}) = \sum_{n=-N_1}^{N_1} \frac{1}{2N+1} e^{-i\frac{2\pi k}{2N+1}n} = \frac{1}{2N+1} e^{-i\frac{2\pi k}{2N+1}N_1} \sum_{n=-0}^{2N_1} e^{-i\frac{2\pi k}{2N+1}n}$$

For $k \neq 0$, we have

$$\hat{x}(\frac{k}{2N+1}) = \frac{1}{2N+1} e^{-i\frac{2\pi k N_1}{2N+1}} \frac{(1 - e^{-i\frac{2\pi(2N_1+1)}{2N+1}})}{1 - e^{-i\frac{2\pi k}{2N+1}}}$$

For $k = 0$, we have

$$\hat{x}(\frac{0}{2N+1}) = \frac{2N_1+1}{\sqrt{2N+1}}$$

If $N_1 = N$, we have: For $k \neq 0$

$$\hat{x}(\frac{k}{2N+1}) = 0$$

and for $k = 0$, we have

$$\hat{x}(\frac{0}{2N+1}) = \sqrt{2N+1}$$

For the other extreme, if $N_1 = 0$, we have for all $k \in \{0, \cdots, 2N+1\}$

$$\hat{x}(\frac{k}{2N+1}) = \frac{1}{\sqrt{2N+1}}$$

b) We have the expansion (in the L_2-sense):

$$x(t) = \sum_{k\in\mathbb{Z}} \hat{x}(\frac{k}{T}) \frac{1}{\sqrt{T}} e^{i\frac{2\pi k}{T}t}$$

with

$$\hat{x}(\frac{k}{T}) = \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \frac{1}{\sqrt{T}} e^{-i\frac{2\pi k}{T}t} dt$$

Thus, with x as given, we have

$$\hat{x}(\frac{k}{T}) = \int_{-T_1}^{T_1} \frac{1}{\sqrt{T}} e^{i\frac{2\pi k}{T}t} = \frac{e^{i\frac{2\pi k T_1}{T}} - e^{-i\frac{2\pi k T_1}{T}}}{i2\pi/\sqrt{T}} = \frac{\sin(2\pi k\frac{T_1}{T})}{2\frac{T_1}{T}} \frac{2T_1}{\sqrt{T}}$$

If $T_1 = \frac{T}{2}$, we have that $\hat{x}(\frac{k}{T}) = 0$ for all $k \neq 0$ and $\hat{x}(\frac{0}{T}) = \sqrt{T}$.

If $x_n(t) = n1_{\{|t|\leq T_1\}}$, then we observe that $\hat{x}_n(\frac{k}{T}) \to \frac{2}{\sqrt{T}}$ for all $k \in \mathbb{Z}$.

We observe from these examples that, as a general insight (which can be made more rigorous under additional conditions), if we expand the signal in time domain, we shrink it in the frequency domain; and if we shrink it in time domain, we expand it in the frequency domain.

5.1.3 Properties of the Discrete Fourier Transforms

Theorem 5.1.1 (Parseval's Equality) *The transformations $\mathcal{F}_{DD}$ and $\mathcal{F}_{CD}$ are unitary, that is:*

$$\langle x, x\rangle = \langle \hat{x}, \hat{x}\rangle$$

Proof. We prove this for $\mathcal{F}_{CD}$ ($\mathcal{F}_{DD}$ follows more directly using the following arguments). Let $x \in L_2([0,P];\mathbb{C})$. Define $x_K := \sum_{-K}^{K} \hat{x}(\frac{k}{P}) \frac{1}{\sqrt{P}} e^{i2\pi\frac{k}{P}t}$. We know that $\lim_{K\to\infty} \|x_K - x\|_2 = 0$. Consider

$$\begin{aligned}
\langle x, x\rangle &= \lim_{K\to\infty} \langle x, x_K\rangle \\
&= \lim_{K\to\infty} \int_0^P x(t) \overline{\sum_{k=-K}^{K} \hat{x}(\frac{k}{P})\frac{1}{\sqrt{P}} e^{i2\pi\frac{k}{P}t}} \\
&= \lim_{K\to\infty} \sum_{k=-K}^{K} \overline{\hat{x}(\frac{k}{P})} \int_0^P x(t)\frac{1}{\sqrt{P}} e^{-i2\pi\frac{k}{P}t} \\
&= \lim_{K\to\infty} \sum_{k=-K}^{K} \overline{\hat{x}(\frac{k}{P})}\hat{x}(\frac{k}{P}) \\
&= \langle \hat{x}, \hat{x}\rangle
\end{aligned} \tag{5.3}$$

□

Time-shift, periodicity and differentiation will also be discussed.

An important property is with regard to convolution. The transform of a convolution is equal to the point-wise product of two signals. The following will be discussed in class:

Theorem 5.1.2 *Let $f, g \in L_2([0, P])$. Then,*

$$\mathcal{F}_{CD}(f * g)(k) = \sqrt{P}\mathcal{F}_{CD}(f)(k)\mathcal{F}_{CD}(g)(k), \quad k \in \mathbb{Z}(\frac{1}{P}).$$

That is, convolution is equaivalent to point-wise multiplication in the frequency domain.

5.1.4 Computational Aspects: The FFT Algorithnm

The Fast Fourier Transform (FFT) is a very important algorithm to implement DTFT ($\mathcal{F}_{DD}$) in a computationally efficient fashion in practice. The *fft* command in Matlab generates the transform.

Observe that for the operations described in (5.1)

$$\mathcal{F}_{DD} : l_2(\{0, 1, \ldots, N-1\}; \mathbb{C}) \to l_2(\{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N-1}{N}\}; \mathbb{C})$$

so that

$$\hat{x} = \mathcal{F}_{DD}(x),$$

with

$$\hat{x}(f) = \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} x(n) e^{-i2\pi fn}, \quad f \in \{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N-1}{N}\},$$

there are N complex multiplications and N complex additions for each f (and thus there will be N such computations). Thus, the FFT algorithm in the form above has the computational complexity of N^2 complex operations (with one complex operation being equivalent to one complex addition and one complex multiplication).

If N is even, we can write

$$\begin{aligned}
\hat{x}(\frac{k}{N}) &= \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} x(n) e^{-i2\pi\frac{k}{N}n} \\
&= \sum_{m=0}^{\frac{N}{2}-1} \frac{1}{\sqrt{N}} x(2m) e^{-i2\pi\frac{k}{N}2m} + \sum_{m=0}^{\frac{N}{2}-1} \frac{1}{\sqrt{N}} x(2m+1) e^{-i2\pi\frac{k}{N}(2m+1)}
\end{aligned}$$

$$= \sum_{m=0}^{\frac{N}{2}-1} \frac{1}{\sqrt{N}} x(2m) e^{-i2\pi \frac{k}{N/2} m} + \Big(\sum_{m=0}^{\frac{N}{2}-1} \frac{1}{\sqrt{N}} x(2m+1) e^{-i2\pi \frac{k}{N/2} m} \Big) e^{-i2\pi \frac{k}{N}} \tag{5.4}$$

Thus, we if define $x_0(m) = x(2m)$ and $x_1(m) = x(2m+1)$, for $m = 0, 1, \cdots, \frac{N}{2} - 1$. Thus, the above leads to

$$\hat{x}(\frac{k}{N}) = \hat{x}_0(\frac{k}{\frac{N}{2}}) + e^{-i2\pi \frac{k}{N}} \hat{x}_1(\frac{k}{\frac{N}{2}})$$

Note that the Fourier transform Thus, as we see, the above leads to a parallel processing of two smaller length transforms. If N is a power of 2, we can continue with this approach to a building block of length $N = 2$. By inductively splitting the summations in the expansions as above, the FFT algorithm then reduces the computational complexity for the $\mathcal{F}_{DD}$ from N^2 complex operations to $N \log_2(N)$ (with N being a power of 2) such operations.

5.2 The Discrete-to-Continuous Fourier Transform (DCFT): $\mathcal{F}_{DC}$

The Discrete-to-Continuous Fourier Transform can be viewed as the inverse of $\mathcal{F}_{CD}$ (with P taken to be 1): A signal $x \in l_2(\mathbb{Z}; \mathbb{R})$ may be expanded as:

$$x(n) = \int_0^1 \hat{x}(f) e^{i2\pi f n}, \quad n \in Z$$

with

$$\hat{x}(f) = \sum_{n \in Z} x(n) e^{-i2\pi f n}, \quad f \in [0, 1)$$

5.3 The CCFT: $\mathcal{F}_{CC}$ on $\mathcal{S}$ and its extension to $L_2(\mathbb{R})$

We will define the CCFT by the relation

$$x(t) = \int_{f=-\infty}^{\infty} \hat{x}(f) e^{i2\pi f t} df, f \in R$$

with

$$\hat{x}(f) = \int_{\tau=-\infty}^{\infty} x(t) e^{-i2\pi f t} dt, \quad f \in R$$

Two important properties are given in the following.

Theorem 5.3.1 *For $\phi \in \mathcal{S}$ and $m \in \mathbb{Z}_+$:*

a) $\mathcal{F}_{CC}(\frac{d^m}{dt^m}\phi)(f) = (i2\pi f)^m \mathcal{F}_{CC}(\phi)(f)$.

b) $\frac{d^m}{df^m}\mathcal{F}_{CC}(\phi)(f) = \mathcal{F}_{CC}((-i2\pi t)^m \phi)(f)$

Proof. a) Take $m = 1$, applying integration by parts,

$$\int \frac{d}{dt}\phi(t) e^{-i2\pi f t} dt = \phi(t) e^{-i2\pi f t} \,|_{-\infty}^{\infty} - \int \phi(t)(-i2\pi f)\phi(t) e^{-i2\pi f t} dt$$

Since $\phi \in \mathcal{S}$, the first term on the right is zero and the result follows. For $m > 1$, the results follows by repeating the above. b)

$$\begin{aligned}
&\frac{d}{df}\int \phi(t)e^{-i2\pi ft}dt = \lim_{h\to 0}\frac{\int \phi(t)(e^{-i2\pi(f+h)t} - e^{-i2\pi ft})dt}{h}\\
&= \lim_{h\to 0}\frac{\int \phi(t)(e^{-i2\pi ht} - 1)e^{-i2\pi ft}dt}{h}\\
&= \lim_{h\to 0}\frac{\int \phi(t)\Big((\cos(2\pi ht) - 1) + -i(\sin(2\pi ht) - 0)\Big)e^{-i2\pi ft}dt}{h}\\
&= \int \phi(t)(-i2\pi t)e^{-i2\pi ft}dt, \qquad (5.5)
\end{aligned}$$

where the last line follows from the dominated convergence theorem as in the proof of Theorem A.2.1 in the Appendix. For $m > 1$, the result follows by repeating the above. □

Theorem 5.3.2 *$\mathcal{F}_{CC}$ is a continuous linear map on $\mathcal{S}$ to $\mathcal{S}$.*

Proof. Let $\hat{\phi} = \mathcal{F}_{CC}(\phi)$. We first show that $\hat{\phi} \in \mathcal{S}$. We use Theorem 5.3.1 in the following. We have

$$\begin{aligned}
&\sup_{f\in\mathbb{R}} |f^m \frac{d^k}{df^k}\hat{\phi}(f)| = \sup_{f\in\mathbb{R}} |f^m \mathcal{F}_{CC}((-i2\pi t)^k\phi(t))(f)|\\
&= \sup_{f\in\mathbb{R}} |\mathcal{F}_{CC}\left(\frac{d^m}{dt^m}\left((\frac{1}{i2\pi})^m(-i2\pi t)^k\phi(t)\right)\right)(f)| < \infty \qquad (5.6)
\end{aligned}$$

Note that $\left(\frac{d^m}{dt^m}\left((2\pi t)^k\phi(t)\right)\right. \in \mathcal{S}$. It follows that the term above is bounded for every $m, k \in \mathbb{Z}_+$ and we have that $\hat{\phi} \in \mathcal{S}$.

We now discuss continuity. With $\hat{\phi}_n = \mathcal{F}_{CC}(\phi_n)$, let $\phi_n \to \phi \equiv 0$ (that is $\phi(t) = 0$ for all t) in the Schwartz space $\mathcal{S}$. Then, building on the analysis above for every $m, k \in \mathbb{Z}_+$,

$$\sup_{f\in\mathbb{R}} |f^m \frac{d^k}{df^k}\hat{\phi}_n(f)| \to 0.$$

It follows that $\mathcal{F}_{CC}(\phi_n) \to \hat{\phi}$ (in the Schwartz space) where $\hat{\phi}(f) = 0$ for all $f \in \mathbb{R}$. □

We will see in the following that the CCFT is also a transformation from $L_2(\mathbb{R};\mathbb{C}) \to L_2(\mathbb{R};\mathbb{C})$.

5.3.1 The Inverse Transform

The following is known as the Riemann-Lebesgue lemma.

Theorem 5.3.3 *Let $g \in L_1(\mathbb{R};\mathbb{R})$. Then,*

$$\lim_{f\to\infty}\int g(t)e^{-i2\pi ft}dt = 0$$

Proof. By Theorem 2.3.11, for every $\epsilon > 0$ there exists $f_c \in C_c$ (that is f_c is continuous with compact support) such that $||f - f_c|| \le \epsilon/2$. Furthermore, we can approximate a continuous function f_c with a function g_c which is differentiable (even with continuous derivatives, a function in C_c^1 -say by polynomials-) so that $||f - g_c|| \le \epsilon$. Now, by integration by parts

$$|\int_a^b g_c(t)e^{-i2\pi ft}dt| \le |\frac{i}{2\pi f}g_c(t)e^{-2i\pi ft}\Big|_a^b| + |\frac{1}{2\pi f}\int_a^b |g_c'(t)dt|$$

The right hand side converges to 0 as $|f| \to \infty$. This holds for every $\epsilon > 0$. □

Recall Exercise 3.3.1, where Theorem 5.3.3 was utilized to arrive at the following:

Lemma 5.3.1 *The following holds for all $\phi \in \mathcal{S}$:*

$$\lim_{n\to\infty} \int \frac{\sin(nx)}{\pi x}\phi(x)dx \to \bar{\delta}(\phi) = \phi(0)$$

With this discussion, we can show that the inverse $\mathcal{F}_{CC}^{-1}$ is defined on $\mathcal{S}$ as:

$$\mathcal{F}_{CC}^{-1}(\hat{\phi}) = \int \hat{\phi}(f)e^{i2\pi ft}df.$$

Observe that:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \hat{\phi}(f)e^{i2\pi ft}df \\
&\lim_{A\to\infty} \int_{f=-A}^{A} \hat{\phi}(f)e^{i2\pi ft}df \\
&= \lim_{A\to\infty} \int_{f=-A}^{A} \left(\int \phi(\tau)e^{-i2\pi f\tau}d\tau \right) e^{i2\pi ft}df \\
&= \lim_{A\to\infty} \int \phi(\tau) \int_{f=-A}^{A} e^{-i2\pi f\tau}e^{i2\pi ft}d\tau df \qquad (5.7) \\
&= \lim_{A\to\infty} \int \phi(\tau) \left(\int_{f=-A}^{A} e^{-i2\pi f(\tau-t)}df \right) d\tau
\end{aligned}$$

For every fixed A, we can justify (5.7) by (Fubini's) Theorem A.3.1. By Lemma 5.3.1, the fact that $\int_{f=-A}^{A} e^{-i2\pi f(\tau-t)}df = \frac{1}{\pi(t-\tau)}\sin(2A\pi(t-\tau))$ represents a distribution which converges to the $\bar{\delta}_t$ distribution (which then satisfies $\int \phi(\tau)\bar{\delta}_t(\tau) \equiv \int \phi(\tau)\delta(t-\tau) = \int \phi(t-u)\delta(u) = \phi(t)$) leads to the desired result.

5.3.2 Plancherel's Identity / Parseval's Theorem

Theorem 5.3.4 *For every $h, g \in \mathcal{S}$ (as well as $L_2(\mathbb{R};\mathbb{C})$, the Fourier transforms $\hat{f}, \hat{g}$ satisfy:*

$$\langle \hat{h}, \hat{g} \rangle = \langle h, g \rangle$$

The proof of this follows from the following: First suppose $h, g \in \mathcal{S}$ are Schwartz signals (where we allow such signals to take complex values). Then, observe that

$$\begin{aligned}
\int \hat{h}(f)\overline{\hat{g}(f)}df &= \int_f \left(\int_t h(t)e^{-i2\pi ft}dt \right) \overline{\hat{g}(f)}df \\
&= \int_t h(t)\left(\int_f \overline{\hat{g}(f)}e^{-i2\pi ft}df \right)dt \qquad (5.8) \\
&= \int_t h(t)\overline{\int_f \hat{g}(f)e^{i2\pi ft}df}\,dt \\
&= \int_t h(t)\overline{g(t)}dt \qquad (5.9)
\end{aligned}$$

Here, (5.8) follows from Fubini's theorem. Hence, it follows that

$$\langle \hat{h}, \hat{g} \rangle = \langle h, g \rangle,$$

5.3.3 Extension of $\mathcal{F}_{CC}$ on $L_2(\mathbb{R};\mathbb{C})$ and Plancherel's theorem

The above discussion applies for for $h, g \in \mathcal{S}$ (where we also allow for $\mathbb{C}$-valued functions in S). We will show that one can extend the Fourier transform for signals in $L_2(\mathbb{R};\mathbb{C})$. Furthermore, as not every function in $L_2(\mathbb{R})$ is also in $L_1(\mathbb{R})$ (for example: $f(t) = \frac{1}{\sqrt{1+t^2}}$ is in L_2 but not in L_1), one cannot define a Fourier transform of a general function in $L_2(\mathbb{R})$ pointwise directly by an integral.

Recall from Theorem 2.3.7, and the discussions following it, that continuous functions are dense in integrable functions, and from Corollary 3.3.1 one can approximate a continuous function by its convolution with a smooth approximate identity sequence leading to a smooth approximation. These lead to the following.

Theorem 5.3.5 *$\mathcal{S}$ is dense in $L_2(\mathbb{R})$.*

With this theorem, our goal will be to define the CCFT of a signal f in $L_2(\mathbb{R};\mathbb{C})$ is as follows. Let $\{\phi_n\}$ be a sequence of functions in $\mathcal{S}$ converging to f (This is possible by the denseness of $\mathcal{S}$ in L_2). We *define* the CCFT of f as the L_2 limit of a sequence of CCFTs of $\{\phi_n\}$. Such an extension defines the *unique extension* of $\mathcal{F}_{CC}$ from $\mathcal{S}$ to $\mathcal{S}$ which is continuous on L_2. This result builds on the following theorem.

Theorem 5.3.6 *Let $T : M \to Y$ be a linear mapping, Y a Banach space, M a dense linear subspace of a normed linear space X. Furthermore, let T be bounded in the sense that*

$$\|T\| = \sup_{x\in M, x\neq 0} \frac{||Tx||}{||x||} < \infty.$$

Then, there exists a unique extension of T on X to Y, denoted by $\bar{T} : X \to Y$, such that $\bar{T}(x) = T(x)$ for $x \in M$ (that is, T and $\bar{T}$ are in agreement on M). Furthermore, $||T|| = ||\bar{T}||$.

Proof. For every x there exists a sequence $x_n \in M$ so that $x_n \to x$ in X. We want to define for every $x \in X$

$$\bar{T}(x) = \lim_{x_n \to x} T(x_n) = \lim_{n\to\infty} T(x_n)$$

First note that x_n is Cauchy, as it is converging to x, and therefore since for every $\epsilon > 0$ there exists N so that for $n, m \geq N$, $\|x_n - x_m\| \leq \epsilon$. It follows that sequence $T(x_n)$ is also Cauchy since for every $\epsilon' = \|T\|\epsilon > 0$ there exists N so that for $n, m \geq N$, $|T(x_n) - T(x_m)| = |T(x_n - x_m)| \leq \|T\|\|x_n - x_m\| \leq \|T\|\epsilon = \epsilon'$. Note that $\bar{T}(x)$ is well-defined, in the sense that for any other sequence $y_n \to x$, we will have that

$$|\lim_{n\to\infty} T(x_n) - \lim_{n\to\infty} T(y_n)| = |\lim_{n\to\infty} T(x_n - y_n)| \leq \lim_{n\to\infty} \|T\||x_n - x| \leq \lim_{n\to\infty} \|T\|(|x_n - x| + |y_n - x|) = 0,$$

since both x_n and y_n converge to x. Finally, we show that $\|\bar{T}\| = \|T\|$. First, for any $x \in X$, with $M \ni x_n \to x$,

$$|\bar{T}(x)| = |\lim_{n\to\infty} T(x_n)| \leq \|T\| \lim_{n\to\infty} \|x_n\| = \|T\|\|x\|,$$

and thus $\|\bar{T}\| \leq \|T\|$. On the other hand, $\bar{T}(x) = T(x)$ for $x \in M$ and thus we must have $\|T\| \leq \|\bar{T}\|$. Thus, the norms must be equal. □

Now, let $M = \mathcal{S}$, $X = Y = L_2(\mathbb{R};\mathbb{C})$, $T = \mathcal{F}_{CC}$ and note by Theorem 5.3.4 that $||\mathcal{F}_{CC}|| = 1$ (**when viewed as a mapping from a subset of** $L_2(\mathbb{R};\mathbb{C})$). In view of this, **we define** F_{CC} **on** $L_2(\mathbb{R};\mathbb{C})$ **to be the unique extension of** $\mathcal{F}_{CC}$ **from** $\mathcal{S} \to L_2(\mathbb{R};\mathbb{C})$. Thus, for $h \in L_2(\mathbb{R};\mathbb{C})$,

$$\langle h, h\rangle = \langle \hat{h}, \hat{h}\rangle, \quad \hat{h} = \mathcal{F}_{CC}(h).$$

In view of the above, we have that Plancherel's Theorem also applies to signals in $L_2(\mathbb{R};\mathbb{C})$.

$$\langle g, h\rangle = \langle \hat{g}, \hat{h}\rangle, \qquad h, g \in \mathcal{S}$$

that

$$||\mathcal{F}_{CC}|| = 1.$$

By the stated (extension) theorem, it follows that $\mathcal{F}_{CC}$ is also unitary on $L_2(\mathbb{R};\mathbb{C})$ with the same operator norm, that is:

$$\langle g, h\rangle = \langle \hat{g}, \hat{h}\rangle, \qquad h, g \in L_2(\mathbb{R};\mathbb{C}).$$

To verify this, let g and h be in $L_2(\mathbb{R};\mathbb{C})$ and $g_n \to g$, and let $h_n \to h$ with $g_n, h_n \in \mathcal{S}$ being Schwartz signals. By Plancherel's identity:

$$\langle g_n, h_n\rangle = \langle \hat{g}_n, \hat{h}_n\rangle$$

Let us take the limit as $n \to \infty$ on both sides. For the left hand side the Cauchy-Schwarz inequality implies that

$$\lim_{n\to\infty} \langle g_n, h_n\rangle = \langle g, h\rangle,$$

because

$$|\langle g_n, h_n\rangle - \langle g, h\rangle| = |\langle g_n - g, h_n\rangle + \langle g, h_n - h\rangle| \le ||g_n - g||||h_n|| + ||g||h_n - h|| \to 0$$

Likewise

$$|\langle \hat{g}_n, \hat{h}_n\rangle - \langle \hat{g}, \hat{h}\rangle| = |\langle \hat{g}_n - \hat{g}, \hat{h}_n\rangle + \langle \hat{g}, \hat{h}_n - \hat{h}\rangle| \le ||\hat{g}_n - \hat{g}||||\hat{h}_n|| + ||\hat{g}||||\hat{h}_n - \hat{h}|| \to 0$$

Here, the convergence to zero follows from the fact that $||\hat{g}_n - \hat{g}|| \to 0$ (since $||g_n - g|| = ||\hat{g}_n - \hat{g}|| \to 0$ by the discussion above) and that $||\hat{h}_n||$ is bounded. This generalizes Plancherel's identity for signals in $L_2(\mathbb{R};\mathbb{C})$.

5.4 Fourier Transform of Distributions ($\mathcal{F}_{CC}$ on $\mathcal{S}^*$)

Recall that $\mathcal{S}^*$ is the dual space on $\mathcal{S}$, that is the space of distributions (linear and continuous functions) on $\mathcal{S}$.

The Fourier transform of a distribution is defined by the following relation: Let $T \in \mathcal{S}^*$. Then, with $\hat{T} = \mathcal{F}_{CC}(T)$, we have

$$\langle \hat{T}(\phi) = T(\hat{\phi}) \quad \phi \in \mathcal{S} \tag{5.10}$$

The inverse $\mathcal{F}_{CC}^{-1}$ of a distribution is defined with the relation

$$\mathcal{F}_{CC}^{-1}(T)(\phi) = T(\mathcal{F}_{CC}^{-1}(\phi)) \quad \phi \in \mathcal{S}$$

With the above, we can conclude that $\hat{T}$ itself is a distribution. Just as the CCFT is a map from $\mathcal{S}$ to itself, the CCFT is also a mapping from $\mathcal{S}^*$ to itself. Thus, every distribution has a Fourier Transform. Furthermore, the map $\mathcal{F}_{CC} : \mathcal{S}^* \to \mathcal{S}^*$ is continuous, linear, and one-to-one. The continuity follows from the definition of the Fourier transform and continuity in $\mathcal{S}^*$.

This definition is consistent with the Fourier transform of a regular distribution (represented by some function $\psi \in \mathcal{S}$) being a distribution which is represented by the Fourier transform of ψ. That is,

$$\int \hat{\psi}(f)\phi(f) = \int \psi(t)\hat{\phi}(t).$$

This equality follows from Fubini's theorem by expressing $\hat{\psi}(f) = \int \psi(t)e^{-i2\pi ft}dt$.

Since any singular distribution can be expressed as a weak* limit of such regular distributions (represented by signals in $\mathcal{S}$), the definition above in (5.10) is consistent with the $\mathcal{F}_{CC}$ of regular distributions.

Exercise 5.4.1 *Show that $\bar{\delta}$ has its CCFT as a distribution represented by the function $h(f) = 1$ for all $f \in \mathbb{R}$.*

Example 5.1. We can compute the Fourier transform of $\cos(2\pi f_0 t)$ by viewing it as a distribution, in the sense that it represents a distribution. Observing that $\cos(2\pi f_0 f) = \frac{1}{2}e^{i2\pi f_0 f} + \frac{1}{2}e^{-i2\pi f_0 f}$, we consider:

$$\begin{aligned}
&\lim_{T\to\infty}\int_{-T}^{T} e^{i2\pi f_0 f}\hat{\phi}(f)df \\
&= \lim_{T\to\infty}\int_{-T}^{T} e^{i2\pi f_0 f}\left(\int \phi(t)e^{-i2\pi ft}dt\right)df \\
&= \lim_{T\to\infty}\int \phi(t)\left(\int_{-T}^{T} e^{i2\pi f_0 f}e^{-i2\pi ft}df\right)dt \\
&= \lim_{T\to\infty}\int \phi(t)\left(\int_{-T}^{T} e^{i2\pi (f_0-t) f}df\right)dt \\
&= \lim_{T\to\infty}\int \phi(t)\frac{\sin(2\pi(f_0-t)T)}{\pi(f_0-t)}dt \\
&= \bar{\delta}_{f_0}(\phi)\left(\equiv \int \phi(t)\delta_{f_0}(t)dt\right)
\end{aligned} \tag{5.11}$$

In the analysis above, we use the fact that $\hat{\phi} \in \mathcal{S}$, use Fubini's Theorem to change the order of the integrations and finally invoke (3.8). Thus, the CCFT of a cosine will be $1/2\delta_{f_0}(t) + 1/2\delta_{-f_0}(t)$. Here, the last equation in brackets is meant to be in an intuitive sense.

Exercise 5.4.2 *Compute the CCFT of the unit step function, viewed as a distribution.*

5.5 $\mathcal{F}_{CC}$ of periodic signals

The CCFT of a periodic signal can also be viewed as a distribution. Let $x(t)$ be continuous and periodic with period P. Then,

$$\begin{aligned}
&\langle \hat{x}, \phi\rangle = \langle x, \hat{\phi}\rangle \\
&= \lim_{T\to\infty}\int_{-T}^{T} x(f)\hat{\phi}(f)df \\
&= \lim_{T\to\infty}\lim_{N\to\infty}\int_{-T}^{T}\sum_{k=-N}^{N}\hat{x}(\frac{k}{P})\frac{1}{\sqrt{P}}e^{i2\pi\frac{k}{P}f}\hat{\phi}(f)df \\
&= \lim_{N\to\infty}\lim_{T\to\infty}\int_{-T}^{T}\sum_{k=-N}^{N}\hat{x}(\frac{k}{P})\frac{1}{\sqrt{P}}e^{i2\pi\frac{k}{P}f}\hat{\phi}(f)df \\
&= \lim_{N\to\infty}\lim_{T\to\infty}\int_{-T}^{T}\sum_{k=-N}^{N}\hat{x}(\frac{k}{P})\frac{1}{\sqrt{P}}e^{i2\pi\frac{k}{P}f}\left(\int \phi(t)e^{-i2\pi ft}dt\right)df \\
&= \lim_{N\to\infty}\lim_{T\to\infty}\int \phi(t)\left(\sum_{k=-N}^{N}\hat{x}(\frac{k}{P})\frac{1}{\sqrt{P}}\int_{-T}^{T} e^{i2\pi\frac{k}{P}f}e^{-i2\pi ft}df\right)dt \\
&= \lim_{N\to\infty}\lim_{T\to\infty}\int \phi(t)\left(\sum_{k=-N}^{N}\hat{x}(\frac{k}{P})\frac{1}{\sqrt{P}}\frac{\sin(2\pi(\frac{k}{P}-t)T)}{\pi(\frac{k}{P}-t)}\right)dt \\
&= \sum_{k=-\infty}^{\infty}\hat{x}(\frac{k}{P})\frac{1}{\sqrt{P}}\bar{\delta}_{\frac{k}{P}}(\phi)
\end{aligned} \tag{5.12}$$

$$\left(\equiv \int \phi(t) \sum_{k=-\infty}^{\infty} \hat{x}(\frac{k}{P}) \frac{1}{\sqrt{P}} \delta_{\frac{k}{P}}(t) dt \right) \tag{5.13}$$

In the above, in (5.12) we can justify the change in the orders since the integration can be thought to be essentially over a compact domain (with the contribution of the integral from outside a compact domain to be made arbitrarily small with a sufficiently large compact set, uniformly over N) as $\hat{\phi} \in \mathcal{S}$.

Thus, we can essentially first view a periodic signal with its CDFT and then replace the values at $\frac{k}{P}$ with $\delta_{\frac{k}{P}}$. This is in agreement with an engineering insight: If one is to express a periodic signal with its $\mathcal{F}_{CD}$ expression:

$$x(t) = \sum_{k \in \mathbb{Z}} \hat{x}(\frac{k}{P}) \frac{1}{\sqrt{P}} e^{i2\pi \frac{k}{P} t}$$

one would expect that this would be equivalent to the integral form:

$$x(t) = \int \sum_{k \in \mathbb{Z}} \hat{x}(\frac{k}{P}) \frac{1}{\sqrt{P}} \delta_{\frac{k}{P}}(t) e^{i2\pi f t} df,$$

whose equivalence is made precise through a distributional approach presented above.

5.6 Band-limited vs Time-limited Functions

Let $\{x(t), \quad t \in \mathbb{R}\}$ be a CT signal. If this signal has a finite bandwidth B, that is if

$$\hat{x}(f) = 0, \quad |f| > B,$$

then it is not possible for the signal to have a bounded support.

Theorem 5.6.1 *Let g, with $\|g\|_2 \neq 0$ have a finite bandwidth. Then, g cannot have a finite support.*

Sketch of Proof. The proof uses ideas from complex analysis, and the Paley-Wiener Theorem. Let g have finite support. Consider the integral

$$\int g(t) e^{zt} dt : \mathbb{C} \to \mathbb{C},$$

which is an extension from the real-line to $\mathbb{C}$ of the Fourier transform $\int g(t) e^{-i2\pi f t} dt : \mathbb{R} \to \mathbb{C}$. Then,

$$\frac{d^k}{dz^k} \int g(t) e^{-zt} dt,$$

would be finite for every $k \in \mathbb{N}$ due to the finite support condition on g. Since the integral $\int g(t) e^{-zt} dt$ is a complex number, having finite derivatives for all k implies that the function is analytic and the Taylor series can be used to express the signal in some neighborhood of any given point. Since this argument holds for every $z \in \mathbb{C}$, the integral is in fact an *entire function* and thus the Taylor series expansion must converge everywhere: Thus, if $\int g(t) e^{-zt} dt$ is zero in a continuum of points, then the integral has to be identically zero for all z, leading to a contradiction which would contradict the condition that $\|g\|_2 = \|\hat{g}\|_2 \neq 0$. $\square$

5.7 Exercises

Exercise 5.7.1 *a) Let $x \in L_2([0,2]; \mathbb{C})$ be given by:*

$$x(t) = 1_{\{t\leq 1\}}t + 1_{\{1\leq t\leq 2\}}(1-t)$$

Let $\hat{x}(\frac{k}{2})$ *denote the Fourier series coefficient corresponding to* $\{\frac{1}{\sqrt{2}}e^{i2\pi\frac{k}{2}t}, t\in[0,2]\}$.

With Matlab, generate the plot of the signal

$$x_N(t) = \sum_{k=-N}^{N}\hat{x}(\frac{k}{2})\frac{1}{\sqrt{2}}e^{i2\pi\frac{k}{2}t}, \quad t\in[0,2]$$

for $N = 5, 10$ *and* 15. *Here* $\hat{x}(\frac{k}{P})$ *are the Fourier Series expansion coefficients.*

Observe that, the signal looks more and more like the original signal as N *gets larger.*

b) Prove that $\lim_{N\to\infty}\int(x(t)-x_N(t))^2dt = 0$.

Hint: Use the properties of Hilbert spaces and the fact that $\{\frac{1}{\sqrt{P}}e^{i2\pi\frac{k}{P}t}\}$ *forms a complete orthonormal sequence. You could invoke this result directly in your argument.*

c) Does for a general $x\in L_2([0,2];\mathbb{C})$,

$$\sup_{t\in[0,2]}|x(t)-x_N(t)|\to 0,$$

as $N\to\infty$*? Explain your argument.*

Exercise 5.7.2 *CCFT is a map from from* $\mathcal{S}$ *to* $\mathcal{S}$. *One typical example is the Gaussian signal* $e^{-at^2/2}$ *for some* $a>0$*:*

Show that for $a>0$, *the CCFT of*

$$\phi(t) = e^{-at^2/2}$$

is equal to

$$\hat{\phi}(f) = Ke^{-2\pi^2f^2/a},$$

for some K *and conclude that* $\hat{\phi}$ *is also a Schwartz signal.*

Show that K *is independent of* f. *Find* K.

Exercise 5.7.3 *Show that CCFT is a unitary transformation from* $L_2(\mathbb{R};\mathbb{C})$ *to itself. That is, show that Placherel's Identity holds for functions in* $L_2(\mathbb{R};\mathbb{C})$.

Exercise 5.7.4 *a) Show that CCFT is a continuous map from* $\mathcal{S}^*$ *to* $\mathcal{S}^*$. *That is, CCFT maps distributions to distributions and this is continuous map on* $\mathcal{S}^*$.

b) Show that the CCFT of the δ*-distribution is another distribution represented by a function which is equal to 1 for all* f*: That is,*

$$\mathcal{F}_{CC}(\bar{\delta}) = H,$$

with

$$H(\phi) = \int\phi(t)dt$$

Observe that this is in agreement with the general understanding that $\hat{\delta}(f) = 1$ *for all* f.

Exercise 5.7.5 *Consider an impulse train defined by:*

$$w_P(t) = \sum_{n\in\mathbb{Z}}\delta(t+nP)$$

so that the distribution that we can associate with this impulse train would be defined by:

$$\overline{w_P}(\phi) = \sum_{n\in\mathbb{Z}} \phi(nP),$$

for $\phi \in \mathcal{S}$.

a) Show that $\overline{w_P}$ is a distribution.

b) Show that

$$\hat{\overline{w_P}}(\phi) = \int \frac{1}{P} w_{\frac{1}{P}}(t)\phi(t)dt,$$

that is, the $\mathcal{F}_{CC}$ of this train is another impulse train.

Exercise 5.7.6 *Consider a square-integrable signal with non-zero L_2 norm, with bounded support. That is, there exists a compact set, outside of which this signal is identically zero. Can the CCFT of such a signal, with a bounded support in time-domain, also have bounded support in frequency domain?*

6

Frequency Domain Analysis of Linear Time-Invariant (LTI) Systems

6.1 Input-Output Relations for Linear Time-Invariant Systems via Fourier Analysis

As we discussed in Section 4.5, a very important property of convolution systems is that if the input is a harmonic function, so is the output: Let $u \in L_\infty(\mathbb{R};\mathbb{C})$ given with

$$u(t) = e^{i2\pi ft},$$

be the input to a linear time-invariant system

$$y(t) = \int_{\tau=-\infty}^{\infty} h(t-\tau)u(\tau)d\tau = \int_{\tau=-\infty}^{\infty} h(\tau)u(t-\tau)d\tau$$

Then, the output satisfies

$$y(t) = \left(\int_{-\infty}^{\infty} h(s)e^{-i2\pi fs}ds\right)e^{i2\pi ft}$$

The integral

$$\hat{h}(f) := \left(\int h(t)e^{-i2\pi ft}dt\right),$$

is $\mathcal{F}_{CC}(h)$ evaluated at f. We call this value, the frequency response of the system at frequency f, whenever it exists.

A similar discussion applies for a discrete-time system: Let $h \in l_1(\mathbb{Z};\mathbb{C})$. If $u(n) = e^{i2\pi fn}$ is the input to a linear time-invariant system given with

$$y(n) = \sum_{m=-\infty}^{\infty} h(n-m)u(m) = \sum_{m=-\infty}^{\infty} h(m)u(n-m).$$

then

$$y(n) = \left(\sum_{m=-\infty}^{\infty} h(m)e^{-i2\pi fm}\right)e^{i2\pi fm}.$$

We recognize that

$$\hat{h}(f) := \sum_{m=-\infty}^{\infty} h(m)e^{-i2\pi fm} = \Big(\mathcal{F}_{DC}(h)\Big)(f),$$

and call $\hat{h}$ the frequency response function.

Convolution systems are used as filters through the characteristics of the frequency response.

Some Properties.

Recall the following properties of $\mathcal{F}_{CC}$.

(i) Let $u, v \in \mathcal{S}$. We have that

$$\Big(\mathcal{F}_{CC}(u * v)\Big)(f) = \hat{u}(f)\hat{v}(f)$$

(ii) If $v = \frac{d}{dt}u$, then $\hat{v}(f) = i2\pi f\hat{u}(f)$.

(iii)Let $v = \sigma^{\theta}(u)$ for some $\theta \in \mathbb{Z}$, that is $v(n) = u(n+\theta)$. Then,

$$\hat{v}(f) = \sum_{n\in\mathbb{Z}} u(n+\theta)e^{-i2\pi fn} = e^{i2\pi\theta f}\hat{u}(f)$$

The above will be very useful properties for studying LTI systems. We can also obtain converse differentiation properties, which will be considered in further detail in Section 7.2 while studying the Z and the Laplace transformations. Nonetheless, we will present two such properties in the following (see Section A.2 for a justification on changing the order of differentiations and summations/integrations):

Let

$$\mathcal{F}_{DC}(x)(f) = \hat{x}(f) = \sum_{n} x(n)e^{-i2\pi fn}$$

Then, through changing the order of limit and summation:

$$\frac{d\hat{x}(f)}{df} = \sum_{n} \frac{dx(n)e^{-i2\pi fn}}{df} = \sum_{n}(-2i\pi nx(n))e^{-i2\pi n}$$

This leads to the conclusion that with $v(n) = -nx(n)$, with v absolutely summable,

$$\mathcal{F}_{DC}(v)(f) = \frac{1}{i2\pi}\frac{d\hat{x}(f)}{df}$$

Likewise, for the continuous-time case with $x \in \mathcal{S}$

$$\mathcal{F}_{CC}(x)(f) = \hat{x}(f) = \int_{t} x(t)e^{-i2\pi ft}dt$$

Via the analysis in Section A.2, through changing the order of limit and integration,

$$\frac{d\hat{x}(f)}{df} = \int \frac{dx(t)e^{-i2\pi ft}}{df}dt = \int(-2i\pi tx(t))e^{-i2\pi t}dt \tag{6.1}$$

This leads to the conclusion that with $v(t) = -tx(t)$, with $v(t)$ (absolutely) integrable,

$$\mathcal{F}_{CC}(v)(f) = \frac{1}{i2\pi}\frac{d\hat{x}(f)}{df}$$

In the context of LTI systems, we will occasionally build on the following properties: If $u(t) = 1_{\{t\geq 0\}}e^{at}$, with $a < 0$, then then $\hat{u}(f) = \frac{1}{-a+i2\pi f}$.

Likewise for $\mathcal{F}_{DC}$, for $|a| < 1$, if $u(n) = a^{n-1}1_{\{n\geq 1\}}$, then $\hat{u}(f) = e^{-i2\pi f}\frac{1}{1-ae^{-i2\pi f}}$.

The properties above are crucial, and typically sufficient, for studying a large class of linear time invariant systems described by differential and difference equations (convolution systems).

6.2 Transfer Functions and their Computation for Convolution Systems via Fourier Transforms

In applications for control, communications, and signal processing, one may design systems or filters using the properties of the frequency response functions.

Consider the following continuous-time system with input u and output y:

$$\sum_{k=0}^{N} a_k \frac{d^k}{dt^k} y(t) = \sum_{m=0}^{M} b_m \frac{d^m}{dt^m} u(t)$$

Taking the $\mathcal{F}_{CC}$ of both sides, we obtain

$$\left(\sum_{k=0}^{N} a_k (i2\pi f)^k \right) \hat{y}(f) = \left(\sum_{m=0}^{M} b_m (i2\pi f)^m \right) \hat{u}(f)$$

This leads to:

$$\hat{h}(f) = \frac{\sum_{m=0}^{M} b_m (i2\pi f)^m}{\sum_{k=0}^{N} a_k (i2\pi f)^k}$$

As an example, let us consider

$$\frac{dy}{dt} = -ay(t) + u(t), \quad a > 0$$

For this system, we obtain by taking the $\mathcal{F}_{CC}$ of both sides (assuming this exists), we have

$$\hat{h}(f) = \frac{1}{a + i2\pi f}$$

and by the discussion earlier,

$$h(t) = e^{-at} 1_{\{t \geq 0\}}.$$

Likewise, for discrete-time systems:

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{m=0}^{M} b_m u(n-m)$$

Taking the $\mathcal{F}_{DC}$ of both sides (assuming the $\mathcal{F}_{DC}$ exist), we obtain

$$\hat{h}(f) = \frac{\hat{y}(f)}{\hat{u}(f)} = \frac{\sum_{m=0}^{M} b_m e^{-i2\pi mf}}{\sum_{k=0}^{N} a_k e^{-i2\pi kf}}$$

As an example, consider

$$y(n+1) = ay(n) + u(n), \quad |a| < 1$$

For this system, we obtain by taking the $\mathcal{F}_{DC}$ of both sides (assuming this exists), we arrive at $(e^{i2\pi f} - a)\hat{y}(f) = \hat{u}(f)$ and thus

$$\hat{h}(f) = \frac{1}{e^{i2\pi f} - a} = \frac{e^{-i2\pi f}}{1 - ae^{-i2\pi f}}$$

As discussed earlier, this is the $\mathcal{F}_{DC}$ of

$$h(n) = a^{n-1} 1_{\{n-1 \geq 0\}}$$

Exercise 6.2.1 *Consider the R-C circuit considered in class with the equations:*

$$\frac{dV_C(t)}{dt} = -\frac{1}{RC} V_C(t) + \frac{1}{RC} u(t)$$

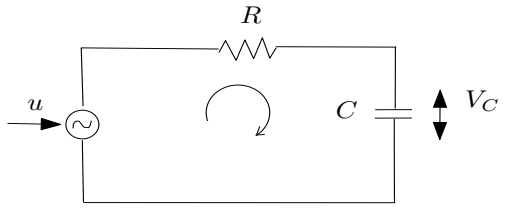

Fig. 6.1: An RC-circuit as an input-output system

a) Viewed as a linear time-invariant system, where u is the input and V_C is the output, find the impulse response and the frequency response.

b) Qualitatively, plot the Bode diagram.

Exercise 6.2.2 *Consider the R-L-C circuit considered in class, with the dynamics*

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = u(t)$$

Note $V_C = Q/C$.

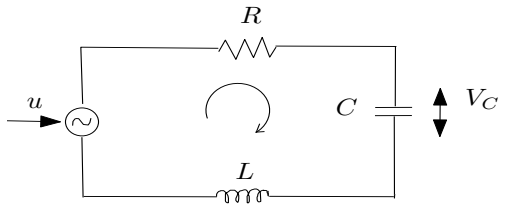

Fig. 6.2: An RLC-circuit as an input-output system

a) Viewed as a linear time-invariant system, where u is the input and V_C is the output, find the impulse response and the frequency response.

b) Qualitatively, plot the Bode diagram in the setup when R is very small.

c) Show that when R is very small, the f value which maximizes the amplitude of the frequency response is around $\frac{1}{2\pi\sqrt{LC}}$. Such a model is often used as an antenna of a radio receiver with the value of the capacitance denoting a tuning parameter.

How to compute the inverse transform? One can, using $\hat{h}$, compute $h(t)$ or $h(n)$, if one is able to compute the inverse transform. A useful method is the partial fraction expansion method. More general techniques will be discussed in the following chapter. Let

$$R(\lambda) = \frac{P(\lambda)}{Q(\lambda)} = \frac{p_0 + p_1\lambda + \cdots + p_M\lambda^M}{q_0 + q_1\lambda + \cdots + q_N\lambda^N}, \lambda \in \mathbb{C}$$

If $M < N$, we call this fraction strictly proper. If $M \leq N$, the fraction is called proper and if $M > N$, it is called improper.

If $M > N$, we can write

$$R(\lambda) = T(\lambda) + \tilde{R}(\lambda),$$

where T has degree $M - N$ and $\tilde{R}(\lambda)$ is strictly proper. We can in particular write:

$$\tilde{R}(\lambda) = \sum_{i=1}^{K}\left(\sum_{k=1}^{m_i}\frac{A_{ik}}{(\lambda - \lambda_i)^k}\right)$$

where λ_i are the roots of Q and m_i is the multiplicity of λ_i.

This is important because we can use the expansion and the aforementioned properties of $\mathcal{F}_{CC}$ and $\mathcal{F}_{DC}$ to compute the inverse transforms. Such approaches will be studied in further detail in the following chapter.

6.3 Exercises

Exercise 6.3.1 *Consider a continuous-time system described by the equation:*

$$\frac{dy(t)}{dt} = ay(t) + u(t), \quad t \in \mathbb{R},$$

where $a < 0$.

a) Find the impulse response of this system.

b) Suppose that the input to this system is given by $\cos(2\pi f_0 t)$. *Let* y_{f_0} *be the output of the system. Find* $y_{f_0}(t)$.

c) If exists, find

$$\lim_{f_0 \to \infty} y_{f_0}(t),$$

for all $t \in \mathbb{R}_+$.

Exercise 6.3.2 *Let a system be described by:*

$$y(n+1) = ay(n) + bu(n) + cu(n-1), \quad n \in \mathbb{Z}.$$

a) For what values of a, b, c *is this system bounded-input-bounded-output (BIBO) stable?*

b) Let $a = 2, b = 1, c = 1$. *Compute the impulse response of the system.*

c) With $a = 2, b = 1, c = 1$*; find the output as a function of* n*, when the input is*

$$u(n) = 1_{\{n \geq 0\}}$$

Exercise 6.3.3 *Consider a Linear Time Invariant (LTI) system characterized by:*

$$y^{(1)}(t) = -ay(t) + u(t), \qquad t \in \mathbb{R}$$

with $a > 0$.

a) Find the impulse response of this system.

b) Find the frequency response of the system.

c) Let $u(t) = e^{-t} 1_{(t \geq 0)}$. *Find* $y(t)$.

Exercise 6.3.4 *Consider a continuous time LTI system with a frequency response*

$$\hat{h}(f) = 1_{(|f| < f_0)} \quad f \in \mathbb{R}$$

a) Find the impulse response of the system; that is the output of the system when the input is the signal representing the $\bar{\delta}$ *distribution.*

b) Find the CCFT of the output, when the input is given by

$$u(t) = e^{-t} \cos(f_1 t) 1_{(t \geq 0)}$$

Exercise 6.3.5 *a) Let $x \in l_2(\mathbb{Z};\mathbb{C})$. Compute the DCFT of*

$$x(n) = a^n 1_{(n\geq 0)},$$

with $|a| < 1$.

b) Compute the DCFT of x:

$$x(n) = \cos(3\pi f_0 n)$$

Exercise 6.3.6 *Many signals take values in multi-dimensional spaces. If you were to define a CDFT for signals in $L_2([0, P_1] \times [0, P_2];\mathbb{C})$, for given $P_1, P_2 \in \mathbb{R}_+$, how would you define it?*

Exercise 6.3.7 *Let a non-anticipative LTI system be given by:*

$$y(n) = \frac{3}{4}y(n-1) - \frac{1}{8}y(n-2) + u(n)$$

a) Compute the frequency response of this system.

b) Compute the impulse response of the system.

c) Find the output when the input is

$$u(n) = (\frac{1}{2})^n 1_{(n\geq 0)}$$

7

The Laplace and Z-Transformations

7.1 Introduction

The powerful tools of Fourier transforms do not directly generalize to signals which are not square integrable. Laplace and Z-transforms allow for the generalization of the machinery developed for *Fourier transformable* signals to a more general class of signals. For example, applications in control systems often require the design of control policies/laws which may turn an open-loop unstable system into a stable system; for studying such unstable signals it is essential to expand the class of signals which can be studied using frequency domain methods. Furthermore, one-sided Laplace and Z-transforms will be seen to be useful in studying systems which have non-zero initial conditions.

The Laplace transform generalizes the CCFT and the Z-transform generalizes the DCFT: If a signal is x is not in $L_1(\mathbb{R};\mathbb{R})$, $y(t) = x(t)e^{-rt}$ may be in $L_1(\mathbb{R};\mathbb{R})$ for some $r > 0$. Likewise, if a signal is x is not in $l_1(\mathbb{Z};\mathbb{R})$, $y(n) = x(n)r^{-n}$ may be in $L_1(\mathbb{R};\mathbb{R})$ for some $r > 1$. The Fourier transforms of these scaled signals correspond to the Laplace and the Z-transforms of x.

A signal is said to be of at-most-exponential growth, if there exist real numbers M, α with $|x(t)| \leq M1_{\{t\geq 0\}}e^{\alpha t}$ for some $M \in \mathbb{R}, \alpha \in \mathbb{R}$. For such signals, the Laplace transform will be defined for certain parameter values. A similar discussion applies for the Z-transform such that if $|x(n)| \leq M1_{\{n\geq 0\}}r^n$ for some $M \in \mathbb{R}, r \in \mathbb{R}$, the Z-transform will be defined for a range of values. These will be detailed further in the following.

7.1.1 The Two-sided Laplace Transform

The two-sided Laplace transform of a continuous-time signal is defined through the pointwise relation:

$$X = \mathcal{L}(x)$$

with

$$X(s) = \int_{t\in\mathbb{R}} x(t)e^{-st}, \quad s \in \mathbb{C}$$

The set $\left\{ s \in \mathbb{C} : \int_{t\in\mathbb{R}} |x(t)e^{-st}| < \infty \right\}$ is called the region of convergence (ROC).

7.1.2 The Two-sided Z-Transform

The two-sided Z-transform of a discrete-time signal is defined through the pointwise relation:

$$X = \mathcal{Z}(x)$$

with

$$X(z) = \sum_{n\in\mathbb{Z}} x(n)z^{-n}, \quad z \in \mathbb{C}$$

The set $\{z \in \mathbb{C} : \sum_{n\in\mathbb{Z}} |x(n)z^{-n}| < \infty\}$ is called the region of convergence (ROC).

7.1.3 The One-sided Laplace Transform

The one-sided Laplace transform of a continuous-time signal is defined through the pointwise relation:

$$X_+ = \mathcal{L}_+(x)$$

with

$$X_+(s) = \int_{t\in\mathbb{R}_+} x(t)e^{-st}$$

The set $\{s \in \mathbb{C} : \int_{t\in\mathbb{R}_+} |x(t)e^{-st}| < \infty\}$ is called the region of convergence (ROC).

7.1.4 The One-sided Z-Transform

The one-sided Z-transform of a discrete-time signal is defined through the pointwise relation:

$$X = \mathcal{Z}_+(x)$$

with

$$X_+(z) = \sum_{n\in\mathbb{Z}_+} x(n)z^{-n}$$

The set $\{z \in \mathbb{C} : \sum_{n\in\mathbb{Z}_+} |x(n)z^{-n}| < \infty\}$ is called the region of convergence (ROC).

7.2 Properties

7.2.1 Linearity

Provided that z is in the ROC for both of the signals x and y

$$(\mathcal{Z}(x + y))(z) = (\mathcal{Z}(x))(z) + (\mathcal{Z}(y))(z)$$

This property applies for the other transforms as well.

7.2.2 Convolution

Provided that z is in the ROC for both of the signals x and y

$$(\mathcal{Z}(x * y))(z) = (\mathcal{Z}(x))(z)(\mathcal{Z}(y))(z)$$

This property applies for the other transforms as well.

7.2.3 Shift Property

Let $y(n) = x(n+m)$, then

$$(\mathcal{Z}(y))(z) = (\mathcal{Z}(x))(z)z^m$$

For the one-sided transform, however, the following holds. Let $m = 1$. Then,

$$(\mathcal{Z}_+(y))(z) = \Big((\mathcal{Z}_+(x))(z) - x(0)\Big)z$$

The general case for $m \in \mathbb{Z}_+$ can be computed accordingly. For example, let $y(n) = x(n-1)$, then

$$(\mathcal{Z}_+(y))(z) = (\mathcal{Z}_+(x))(z)z^{-1} - x(-1).$$

Likewise, let $y(t) = x(t-\theta)$. Then,

$$(\mathcal{L}(y))(s) = (\mathcal{L}(x))(s)e^{-s\theta} \tag{7.1}$$

For the one-sided transform, let $y(t) = x(t-\theta)1_{\{t\geq\theta\}}$. Then,

$$(\mathcal{L}_+(y))(s) = (\mathcal{L}_+(x))(s)e^{-s\theta}$$

7.2.4 Converse Shift Property

Let $y(n) = x(n)a^n 1_{\{n\geq 0\}}$, then

$$(\mathcal{Z}(y))(z) = (\mathcal{Z}(x))(\frac{z}{a}),$$

provided that $\frac{z}{a} \in ROC$ for x.

Let $y(t) = x(t)e^{at}1_{\{t\geq 0\}}$, then

$$(\mathcal{L}(y))(s) = (\mathcal{L}(x))(s-a),$$

provided that $s - a \in ROC$ for x.

7.2.5 Differentiation Property (in time domain)

Let $D(x)$ denote $\frac{dx}{dt}$ (assumed to exist), and let $x(t) = 0$ for $t \leq b$ some $b \in \mathbb{R}$ and that $|x(t)| \leq Me^{at}$. Then,

$$\mathcal{L}(Dx)(s) = s\mathcal{L}(x)(s)$$

$$\mathcal{L}_+(Dx)(s) = s\mathcal{L}_+(x)(s) - x(0),$$

for $\mathrm{Re}\{s\} > a$.

7.2.6 Converse Differentiation

Suppose that $\limsup_{n\to\infty} |x(n)|^{(\frac{1}{n})} \leq R$ for some $R \in \mathbb{R}$. This implies that $\{z : |z| > R\}$ is in the ROC. To see this one should note that for every $\delta > 0$, there exists N_δ such that for $n > N_\delta$ we have $|x(n)|^{(\frac{1}{n})} < R + \delta$. Then, take δ to be less that $|z| - R$ for any given $|z| > R$.

Now, let $y(n) = -nx(n)$. Then,

$$\mathcal{Z}_+(y)(z) = z\frac{d}{dz}(\mathcal{Z}_+(x))(z),$$

for $|z| > R$.

The proof of this result uses the fact that with

$$X(z) = \sum_{n\geq 0} x(n) z^{-n},$$

for $|z| > R$,

$$X'(z) = \sum_{n\geq 0} -x(n) n z^{-n-1}.$$

We now verify this result. First observe that

$$K(z) = \sum_{n\geq 0} -x(n) n z^{-n-1},$$

is also absolutely convergent for $|z| > R$ (by the δ, N_δ argument noted above). We now show that $K(z)$ is indeed the derivative of $X(z)$. Consider

$$\begin{aligned}
&\frac{d}{dz}X(z) - K(z) = \lim_{h\to 0} \frac{X(z+h) - X(z)}{h} - K(z) \\
&= \lim_{h\to 0} \sum_{n\geq 0} x(n) \frac{(z+h)^{-n} - z^{-n}}{h} - K(z) \\
&= \lim_{h\to 0} \sum_{n=0}^{N} x(n) \frac{(z+h)^{-n} - z^{-n}}{h} + \sum_{n=N+1}^{\infty} x(n) \frac{(z+h)^{-n} - z^{-n}}{h} - K(z)
\end{aligned}$$

With $|z| - |h| > R$, observe that

$$(\frac{1}{z+h})^n - (\frac{1}{z})^n = (\frac{1}{z+h} - \frac{1}{z}) \sum_{k=0}^{n-1} (\frac{1}{z+h})^k (\frac{1}{z})^{n-1-k}$$

Since $|z+h| \geq |z| - |h| > R$, we have that

$$|\frac{\frac{1}{z+h} - \frac{1}{z}}{h}| \leq |\frac{1}{z+h}\frac{1}{z}| \leq (R+\delta)^{-2}$$

for some $\delta > 0$. and

$$\frac{1}{h}\left|(\frac{1}{z+h})^n - (\frac{1}{z})^n\right| \leq n(R+\delta)^{-2}(R+\delta)^{-(n-1)} = n(R+\delta)^{-(n-1)}$$

Thus, for all h sufficiently small, we have that

$$\sum_{n=N+1}^{\infty} x(n) \frac{(z+h)^{-n} - z^{-n}}{h} \leq \sum_{n\geq N+1} x(n) n (R+\delta)^{-(n-1)},$$

and this term can be made arbitrarily small (uniformly over sufficiently small h) by picking a large enough N since $\limsup_{n\to\infty} (|x(n)|)^{(\frac{1}{n})} \leq R$. For the first term, we write that

$$\lim_{h\to 0} \sum_{n=0}^{N} x(n) \frac{(z+h)^{-n} - z^{-n}}{h} = \sum_{n=0}^{N} -x(n) n z^{-n-1},$$

since there are only finitely many terms. Since $\sum_{n=0}^{N} -x(n) n z^{-n-1} - K(z) \to 0$ as N goes to ∞, the result follows.

The derivative rule for the Laplace transforms follows from a similar reasoning. Let $|x(t)| \leq Me^{at}$ for some $M, a \in \mathbb{R}$. Let $y(t) = -tx(t)$. Then, for s with $Re\{s\} > a$, we have

$$\mathcal{L}_+(y)(s) = \frac{d}{ds}(\mathcal{L}_+(x))(s)$$

To see this, consider with $\mathcal{L}_+(y)(s) = X(s)$

$$\begin{aligned}
\frac{d}{ds}X(s) &= \lim_{h\to 0}\frac{X(s+h)-X(s)}{h} \\
&= \lim_{h\to 0}\frac{\int_{t\geq 0} x(t)e^{-(s+h)t} - e^{-st}}{h}dt \\
&= \lim_{h\to 0}\int_{t\geq 0} x(t)e^{-st}\frac{e^{-ht}-1}{h}dt \\
&= \lim_{h\to 0}\int_{t\geq 0} x(t)e^{-st}\frac{\sum_{k=0}^{\infty}\frac{(-ht)^k}{k!}-1}{h}dt \\
&= \lim_{h\to 0}\int_{t\geq 0} -x(t)e^{-st}t\sum_{k=0}^{\infty}\frac{(-ht)^k}{(k+1)!}dt
\end{aligned} \tag{7.2}$$

Since $\sum_{k=0}^{\infty}\frac{(-ht)^k}{(k+1)!} \leq \sum_{k=0}^{\infty}\frac{|-ht|^k}{k!} \leq e^{|ht|}$, and $\lim_{h\to 0} e^{|ht|} = 1$ for all t, and that $\int x(t)te^{-st}e^{|h|t}dt$ is integrable when s is in the ROC (by taking h sufficiently small), the dominated convergence theorem implies that

$$\frac{d}{ds}X(s) = \int_{t\geq 0} -tx(t)e^{-st}dt = \mathcal{L}_+\{(-tx(t))\}(s).$$

Remark 7.1. Note that for $\mathcal{F}_{CC}$, a more subtle argument is needed for pushing the derivative inside the integration in (6.1) via applying Theorem A.2.1 to imaginary and real parts of an exponential separately. For the Z and Laplace transforms above, we are using the liberty of the region of convergence being outside the critical curves/lines (as in $|z| > R$).

7.2.7 Scaling

If $y(t) = x(\alpha t)$, then

$$\mathcal{L}(y)(s) = \frac{1}{|\alpha|}\mathcal{L}(x)(\frac{s}{\alpha}),$$

provided that $\frac{s}{\alpha} \in ROC$ for x.

7.2.8 Initial Value Theorem

Let $x(n) \leq Ma^n$ for all $n \in \mathbb{Z}$ and for some $M, a \in \mathbb{R}$. Then,

$$\lim_{z\to\infty} X_+(z) = x(0),$$

for $|z| > a$.

Let $x(t) \leq Me^{at}$ and $\frac{d}{dt}x(t) \leq Me^{at}$ for all $t \in \mathbb{R}$ and for some $M, a \in \mathbb{R}$. Then,

$$\lim_{s\to\infty, \mathrm{Re}\{s\}>a} sX_+(s) = x(0),$$

for $\mathrm{Re}\{s\} > a$. The proof of this result follows from the differentiation property (in time domain) (and an application of the dominated convergence theorem (see Theorem A.1.5) if $Res \to \infty$ or the Riemann-Lebesgue Lemma (see Theorem 5.3.3) if $\mathrm{Im}(s) \to \infty$ but the real part does not converge to infinity).

7.2.9 Final Value Theorem

If $\lim_{t\to\infty} x(t) =: M < \infty$, then

$$\lim_{t\to\infty} x(t) = \lim_{s\to 0, Re\{s\}>0} sX_+(s),$$

Proof. Let $M = \lim_{t\to\infty} x(t)$. Then with $M = s\int Me^{-st}dt$, it suffices to show that

$$\lim_{s\to 0, Re\{s\}>0} \int (x(t) - M)e^{-st}dt = 0. \tag{7.3}$$

For K with for all $t > K$: $|x(t) - M| \le \epsilon$, it follows that

$$\lim_{s\to 0, Re\{s\}>0} |s\int_K^\infty (x(t) - M)e^{-st}dt| \le \lim_{s\to 0, Re\{s\}>0} |s\epsilon\frac{\epsilon e^{-Re\{s\}K}}{s}| \le \epsilon.$$

On the other hand with this K fixed, the remainder $\lim_{s\to 0, Re\{s\}>0} s\int_0^K (x(t) - M)e^{-st} = 0$. Thus, (7.3) holds.

To be able to apply the Final Value Theorem, it is important to ensure that the finiteness condition, $\lim_{t\to\infty} x(t) =: M < \infty$, holds. **Note that if we have that all poles of $sX_+(s)$ are in the left half plane, this ensures that $\lim_{t\to\infty} x(t)$ exists and is finite.**

For a discrete-time signal, if $\lim_{n\to\infty} x(n) < \infty$, then

$$\lim_{n\to\infty} x(n) = \lim_{z\to 1, |z|>1} (1 - z^{-1})X_+(z)$$

The proof follows from the same arguments used in the proof above for the Laplace setup. Once again, note that if we have that all poles of $(1 - z^{-1})X_+(z)$ are strictly inside the unit circle, then $\lim_{n\to\infty} x(n)$ exists and is finite.

7.3 Computing the Inverse Transforms

There are usually three methods that can be applied depending on a particular problem. One is through the partial fraction expansion and using the properties of the transforms. Typically, for linear systems, this is the most direct approach. All is required is to know that

$$\mathcal{Z}(a^n 1_{n\ge 0})(z) = \frac{1}{1 - az^{-1}}$$

with $z \in \{z : |z| > a\}$ and

$$\mathcal{L}(e^{at}1_{t\ge 0})(s) = \frac{1}{s - a}$$

with $s \in \{s : Re\{s\} > a\}$, together with the properties we discussed above. One needs to pay particular attention to the regions of convergence: for examples both of the signals $x_1(t) = e^{at}1_{\{t\ge 0\}}$ and $x_2(t) = -e^{at}1_{\{t<0\}}$ have their Laplace transforms as $\frac{1}{s-a}$, but the first one is defined for $Re\{s\} > a$ and the second one for $Re\{s\} < a$.

A second method is to try to expand the transforms using power series (Laurent series) and match the components in the series with the signal itself.

Example 7.2. Compute the inverse transform of $X(z) = \frac{1}{z^2-1}$ where the region of convergence is defined to be $\{z : |z| > 1\}$. You may want to first write $X(z) = z^{-2}\frac{1}{1-z^{-2}}$.

The most general method is to compute a contour integration along the unit circle or the imaginary line of a scaled signal. In this case, for Z-transforms:

$$x(n) = \frac{1}{i2\pi} \int_c X(z) z^{n-1} dz$$

where the contour integral is taken along a circle in the region of convergence in a counter-clockwise fashion. For the Laplace transform:

$$x(t) = \frac{1}{i2\pi} \int_c X(s) e^{st} ds,$$

where the integral is taken along the line $Re\{s\} = R$ which is in the region of convergence. Cauchy's Integral Formula (see Theorem B.0.2) may be employed to obtain solutions.

However, for applications considered in this course, the partial fraction expansion is the most direct approach.

7.4 Systems Analysis using the Laplace and the Z Transforms

For a given convolution system, the property that if u is an input, h is the impulse response and y the output

$$(\mathcal{L}(y))(s) = (\mathcal{L}(u))(s)(\mathcal{L}(h))(s)$$

leads to the fact that

$$H(s) = \frac{Y(s)}{U(s)}$$

Likewise,

$$(\mathcal{Z}(y))(z) = (\mathcal{Z}(u))(z)(\mathcal{Z}(h))(z)$$

leads to the fact that

$$H(z) = \frac{Y(z)}{U(z)}$$

Besides being able to compute the impulse response and transfer functions for such systems, we can obtain useful properties of a convolution system through the use of Laplace and Z transforms.

7.5 Causality (Realizability), Stability and Minimum-Phase Systems

A convolution system is causal (realizable) if $h(n) = 0$ for $n < 0$. This implies that if $r \in \mathbb{R}_+$ is in the region of convergence, so is R for any $R > r$. Therefore, the region of convergence must contain the entire area outside some circle if it is non-empty.

A convolution system is BIBO stable if and only if $\sum_n |h(n)| < \infty$, which implies that $|z| = 1$ must be in the region of convergence.

Therefore, a causal convolution system is BIBO stable if and only if the region of convergence is of the form $\{z : |z| > R\}$ for some $R < 1$.

In particular, let $P(z)$ and $Q(z)$ be polynomials in z. Let the transfer function of a discrete-time LTI system be given by

$$H(z) = \frac{P(z)}{Q(z)}$$

This system is stable and causal if and only if: the degree of P is less than or equal to the degree of Q and all poles of H (that is, the zeros of Q which do not cancel with the zeros of P) are inside the unit circle.

A similar discussion applies for continuous-time systems: Such a system is BIBO stable if the imaginary axis is in the region of convergence. Such a system is causal if the region of convergence includes $\{s : Re\{s\} > R\}$ for some R,

provided that the region of convergence is non-empty. Therefore, a continuous-time system is BIBO stable and causal if the region of convergence is of the form $\{s : Re\{s\} > R\}$ for some $R < 0$. Thus, if $P(s)$ and $Q(s)$ be polynomials in z and the transfer function of a continuous-time LTI system be given by

$$H(s) = \frac{P(s)}{Q(s)},$$

this system is stable if poles of H are in the left-half plane.

A practically important property of stable and causal convolution systems is whether the inverse of the transfer function is realizable through also a causal and stable system: Such systems are called *minimum-phase* systems. Thus, a system is minimum-phase if all of its zeros and poles are inside the unit circle. A similar discussion applies for continuous-time systems, all the poles and zeros in this case belong to the left-half plane. Such systems are called minimum phase since every system transfer function can be written as a product of a minimum phase system transfer function and a transfer function which has unit magnitude for a purely harmonic input but which has a larger (positive) phase change as the frequency is varied. To make this more concrete, consider

$$G_1(s) = \frac{s-1}{s+5}$$

This system is not minimum-phase. Now, write this system as:

$$G_1(s) = \left(\frac{s+1}{s+5}\right)\left(\frac{s-1}{s+1}\right)$$

Here, $G_2(s) := \frac{s+1}{s+5}$ is minimum-phase. $G_3(s) := \frac{s-1}{s+1}$ is so that is magnitude on the imaginary axis is always 1. However, this term contributes to a positive phase. This can be observed by plotting the Bode diagram, as for small ω values, the signal has a phase close to π which gradually decays to zero.

Another way to observe this added delay is the following: Write $\frac{s-1}{s+1} = 1 - \frac{2}{s+1}$, and observe that the inverse Laplace of this term is the Dirac delta impulse minus the effect of the inverse Laplace of $-\frac{2}{s+1}$; this latter term is $b(t) := -2e^{-t}1_{\{t\geq 0\}}$ or in terms of a linear system it is the solution of a causal system with input u whose output is $\int b(t-\tau)u(\tau)d\tau$ (or generally $\int^t Ce^{A(t-s)}Bu(s)$ for appropriate matrices A, B, C): This term adds a delayed response compared with the effect of the Dirac delta impulse.

Yet, another interpretation is the following: Let $\theta > 0$. One has the approximation

$$\frac{1 - s\theta/2}{1 + s\theta/2} \approx e^{-s\theta}$$

for small $s = i\omega$ values through expanding the exponential term. By our analysis earlier in (7.1), a negative complex exponential in the Laplace transform contributes to a positive time delay. The approximation above is known as a first-order Padé approximation.

Thus, non-minimum-phase systems have higher delay properties in their impulse responses compared to minimum-phase systems.

7.6 Initial Value Problems using the Laplace and Z Transforms

The one-sided transforms are very useful in obtaining solutions to differential equations with initial conditions, as well as difference equations with initial conditions.

7.7 Exercises

Exercise 7.7.1 *a) Compute the (two-sided) Z-transform of*

$$x(n) = 2^n 1_{\{n \geq 0\}}$$

Note that you should find the Region of Convergence as well.

b) Compute the (two-sided) Laplace-transform of

$$x(t) = e^{2t} 1_{\{t \geq 0\}}$$

Find the regions in the complex plane, where the transforms are finite valued.

c) Show that the one-sided Laplace transform of $\cos(\alpha t)$ *satisfies*

$$\mathcal{L}_+\{\cos \alpha t\} = \frac{s}{s^2 + \alpha^2}, \quad \mathrm{Re(s)} > 0$$

d) Compute the inverse Laplace transform of

$$\frac{s^2 + 9s + 2}{(s-1)^2(s+3)}, \quad \mathrm{Re(s)} > 1$$

Hint: Use partial fraction expansion and the properties of the derivative of a Laplace transform.

Exercise 7.7.2 *Find the inverse Z-transform of:*

$$X(z) = \frac{3 - \frac{5}{6}z^{-1}}{(1 - \frac{1}{4}z^{-1})(1 - \frac{1}{3}z^{-1}}, \quad |z| > 1$$

Exercise 7.7.3 *Let* $P(z)$ *and* $Q(z)$ *be polynomials in* z*. Let the transfer function of a discrete-time LTI system be given by*

$$H(z) = \frac{P(z)}{Q(z)}$$

a) Suppose the system is BIBO stable. Show that the system is causal (non-anticipative) if and only if $\frac{P(z)}{Q(z)}$ *is a proper fraction (that is the degree of the polynomial in the numerator cannot be greater than the one of the denominator).*

b) Show that the system is BIBO stable if and only if the Region of Convergence of the transfer function contains the unit circle. Thus, for a system to be both causal and stable, what are the conditions on the roots of $Q(z)$*?*

Exercise 7.7.4 *Let a system be described by:*

$$y(n+2) = 3y(n+1) - 2y(n) + u(n), \quad n \in \mathbb{Z}.$$

a) Is this system non-anticipative? Bounded-input-bounded-output (BIBO) stable?

b) Compute the transfer function of this system.

c) Compute the impulse response of the system.

d) Compute the output when the input is

$$u(n) = (-1)^n 1_{\{n \geq 0\}}$$

Exercise 7.7.5 *Let a system be described by:*

$$y(n+2) = 3y(n+1) - 2y(n) + u(n), \quad n \in \mathbb{Z}.$$

a) Is this system non-anticipative? Bounded-input-bounded-output (BIBO) stable?

b) Compute the transfer function of this system.

c) Compute the impulse response of the system.

d) Compute the output when the input is

$$u(n) = (-1)^n 1_{\{n \geq 0\}}$$

Exercise 7.7.6 *a) Let $H(z) = \frac{1}{1-z^2}$. Given that H represents the transfer function (Z-transform of the impulse response) of a causal filter, find $h(n)$.*

b) Find the solution to the following sequence of equations:

$$y(n+2) = 2y(n+1) + y(n), n \geq 0$$

with initial conditions:

$$y(0) = 0, \qquad y(1) = 1.$$

That is, find $y(n), n \geq 0$.

8

Control Analysis and Design through Frequency Domain Methods

8.1 Transfer Function Shaping through Control: Closed-Loop vs. Open-Loop

In Section 1.2.1, we discussed some control theoretic configurations mapping an input to an output and how the map can be shaped by control design. Two common architectures are depicted in Figure 1.2 (as a general output feedback; here the control depends on the output of the system) and in Figure 1.3 (as an error output feedback control system; here the control depends on the error between the external input and the system output). More general configurations are also possible, as discussed in Section 1.2.1. For consistency, we will focus on a particular architecture noting that the analysis to follow can be generalized to any of these models.

Consider the (error) feedback loop given in Figure 8.4. Here $P(s)$ denotes the transfer function of the system to be controlled. By writing $Y(s) = P(s)C(s)(R(s) - Y(s))$, it follows that

$$\frac{Y(s)}{R(s)} = \frac{P(s)C(s)}{1 + P(s)C(s)} \tag{8.1}$$

is the closed-loop transfer function (under negative unity feedback). Compare this with the setup where there is no feedback: in this case, the transfer function would have been $P(s)C(s)$. This latter expression is often called the (open) loop-transfer function. The goal is to shape (8.1) via the control characterized with $C(s)$ in the frequency domain.

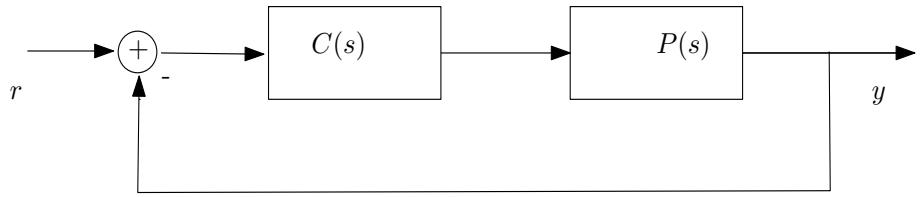

Fig. 8.1

8.1.1 Some motivation via a common class of controllers: PID controllers

By Laplace transforms, we know that differentiation in time domain entails a multiplication with s, and integration involves multiplication with $\frac{1}{s}$. In the context of the setup of Figure 8.4, let $e(t) = y(t) - r(t)$. A popular and practical type of control structure involves:

$$u(t) = k_i \int_0^t e(t)dt + k_d \frac{de}{dt} + k_p e(t)$$

which writes as, in Laplace domain,

$$U(s) = \left(\frac{k_i}{s} + k_d s + k_p\right) E(s).$$

Thus, the control uses both the error itself (proportional), its integration, and its derivative; leading to the term **PID** control.

8.2 Bode-Plot Analysis

Bode plots were studied earlier in class. With Bode plots, we observed that we can identify the transfer function of a system, when a system is already stable. However, the Bode plot does not provide insights on how to design a control system or how to adjust a controller so that stability is attained.

8.3 The Root Locus Method

The set $\{s : 1 + C(s)P(s) = 0\}$ consists of the poles of the transfer function. If one associates with the controller a gain parameter K, the root locus method traces the set of all poles as K ranges from 0 to ∞ (and often from 0 to $-\infty$ as well), so that $\{s : 1 + KC(s)P(s) = 0\}$ is traced. Thus, this method provides a design technique in identifying desirable values for the parameter K.

The root locus method allows one to identify the poles. The pole information clearly identifies BIBO stability properties. Additionally, it lets one select desirable pole values: for example, poles with imaginary components lead to significant transient fluctuations and poles with real parts closer to the origin (on the left half plane) dominate the behaviour involving the response characteristics. A control engineer/designer may reasons to choose certain poles over others.

For the approach to be practical, the following key mathematical result is to be noted.

Theorem 8.3.1 *Consider the polynomial $a(s) + Kb(s) = 0$ where a and b are polynomials. The roots of this polynomial vary continuously as $K \in \mathbb{R}$ changes.*

Proof Sketch. We will follow a contrapositive argument showing that if the roots do not converge to the roots of the polynomial, the polynomial cannot converge either, as K approaches a fixed number (say 1, without any loss). Consider the following three steps:
a) First show that for any polynomial $p(s) = a_0 + a_1 s + \cdots + a_{n-1}s^{n-1} + s^n$, the matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}$$

has $p(s)$ as its characteristic polynomial.

We can show this by directly by computing the characteristic polynomial (and by an inductive argument): The result holds for $n = 2$. Now, let this hold for $n - 1 \geq 2$. We show that it also holds for n. Observe that the determinant of the matrix

$$\lambda I - A = \begin{bmatrix} \lambda & -1 & 0 & \cdots & 0 \\ 0 & \lambda & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \\ a_0 & a_1 & a_2 & \cdots & \lambda + a_{n-1} \end{bmatrix}$$

is

$$\lambda(a_1 + \cdots + a_{n-1}\lambda^{n-2}) + (-1)^{n-1}a_0(-1)^{n-1} = a_0 + a_1\lambda + \cdots + a_{n-1}\lambda^{n-1} + \lambda^n$$

Alternatively, (given earlier studies involving differential equations) you can show this by realizing that the matrix above appears in the first-order ODE reduction of the nth order constant coefficient differential equation:

$$\frac{d^n y(t)}{dt^n} + a_{n-1}\frac{d^{n-1}y(t)}{dt^{n-1}} + a_{n-2}\frac{d^{n-2}y(t)}{dt^{n-2}} + \cdots a_1\frac{dy(t)}{dt} + a_0 y(t) = 0, \tag{8.2}$$

by writing

$$x(t) = \begin{bmatrix} y(t) & \frac{dy(t)}{dt} & \cdots & \frac{d^{n-1}y(t)}{dt^{n-1}} \end{bmatrix}^T \tag{8.3}$$

Then, realize that $e^{\lambda_i t}c$ (for some constant c) solves the differential equation (8.2) and thus obtain a solution for (8.3), leading to an eigenvalue equation: $A\begin{bmatrix} e^{\lambda_i t} \\ \lambda_i e^{\lambda_i t} \\ \vdots \\ \lambda_i^{n-1}e^{\lambda_i t} \end{bmatrix} = \lambda_i \begin{bmatrix} e^{\lambda_i t} \\ \lambda_i e^{\lambda_i t} \\ \vdots \\ \lambda_i^{n-1}e^{\lambda_i t} \end{bmatrix}$. For a repeated eigenvalue we need to consider $e^{\lambda_i t}t$ or further functions, and a slightly more involved analysis, to arrive at generalized eigenvectors.

b) From this, one can show through some algebraic analysis that eigenvalues are uniformly bounded where the bound continuously changes with the polynomial coefficients: Consider an eigenvalue λ with eigenvector v with $Av = \lambda v$ and $v = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^T$. Then, for every i, we have that

$$|\lambda||v_i| = |\sum_j A(i,j)v(j)| \leq \left(\max_i \sum_j |A(i,j)|\right) \max_j |v_j|$$

and hence

$$|\lambda| \max_i |v_i| \leq \left(\max_i \sum_j |A(i,j)|\right) \max_j |v_j|,$$

and thus $|\lambda| \leq \max_i \sum_j |A(i,j)|$ for all eigenvalues. As a result $|\lambda_j| \leq \max(1, \sum_i |a_i|)$ for each $1 \leq j \leq n$. Alternatively, you can use a very useful result known as Gershgorin circle theorem. What matters is that the bound (on the eigenvalues) is uniformly bounded (as K changes), since the obtained bound above changes continuously with K.

c) Now, as K changes, the coefficients of the polynomial continuously change. Therefore, the roots of the polynomial are uniformly bounded for K sufficiently close to any fixed K^*. This implies that for every sequence $K_m \to K^*$, the corresponding sequence of roots must contain a converging subsequence (since the ordering of the roots may be arbitrary one can consider the metric between two vectors defined as the smallest l_2 distance among all possible permutations[1]).Then, we can arrive at a contradiction for the following contrapositive argument: suppose that K approaches K^* along some sequence but the roots do not converge to the roots of the polynomial with $K = K^*$.

Let $\{s_1^K, \cdots, s_n^K\}$ be the roots of the polynomial $p_K(s)$ for a given K. For any sequence $K_m \to K^*$, by part a), the family of roots will take values from a compact set. Thus, there must exist a converging subsequence, call such a subsequential limit $\{\bar{s}_1, \cdots, \bar{s}_n\}$. Now, if this limit is not the same as the roots of the polynomial with $K = 1$, then, the value of the polynomial at $K = K^*$, $p_{K^*}(s)$ must be different then the value $\prod_{m=1}^n (s - \bar{s}_m)$ for some $s \in \mathbb{C}$. But, by continuity of the polynomial itself in K, $p_K(s) \to p_1(s)$ for all $s \in \mathbb{C}$, a contradiction. Thus, every converging subsequence must converge to the roots of the polynomial $p_{K^*}(s)$. For the metric for convergence, as noted, we consider the smallest possible l_2 metric on $\mathbb{C}^n$ among all permutations.

Finally, we now show that in fact every sequence itself must be converging: suppose not, then there exists ϵ and a subsequence $\{\bar{s}_1, \cdots, \bar{s}_n\}$ such that $\{\bar{s}_1, \cdots, \bar{s}_n\}$ is ϵ away from the roots of the $p_{K^*}(s)$ under the metric considered above. But this subsequence itself must contain a converging subsequence, by the arguments above, and the limit has to be the roots of $p_{K^*}(s)$. Hence, a contradiction. □

[1] As an exercise, show that this permutation does not violate the conditions of being a metric: $\bar{d}(x,y) := \min_\sigma(d(x,\sigma(y))$, where σ permutes the order of the components of the vector y.

We note that an elementary proof along this direction is given in [13].

Remark 8.1. The above also directly establishes the very useful result that when one is given a matrix, the eigenvalues of the matrix are continuous in (pointwise perturbations of) its entries.

Exercise 8.3.1 *Consider Figure 8.2.*

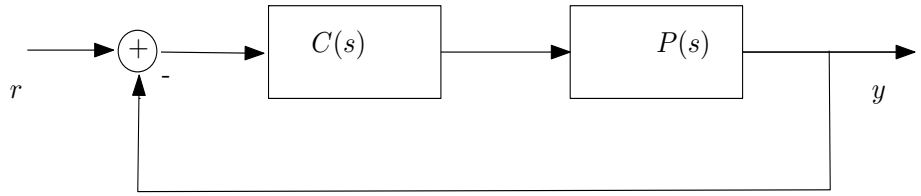

Fig. 8.2

a) Let the plant and controller be given with $P(s) = \frac{1}{s^2}$ (double integrator dynamics) and $C(s) = k_p$ (such a controller is known as a proportional controller). Find the root locus as k_p changes from 0 to ∞.

b) [PD Control] Consider $P(s) = \frac{1}{s^2}$ (double integrator dynamics), and $C(s) = k_p + k_d s$ (the term PD-controller means proportional plus derivative control). Let $k_p = k_d = K$. Find the root locus as K changes from 0 to ∞. Conclude, while comparing with part a above, that the addition of the derivative controller has pushed the poles to the left-half plane (thus, leading to stability!)

c) [Reference Tracking] For the system with the controller in part b), let $K > 0$: Let $r(t) = A1_{t\geq 0}$ for some $A \in \mathbb{R}$. Find $\lim_{t\to\infty} y(t)$. Hint: Apply the Final Value Theorem. We have that $R(s) = A\frac{1}{s}$ and with $Y(s) = A\frac{Ks+K}{s(s^2+Ks+K)}$ we have that $sY(s)$ has all poles on the left-half plane. By the final value theorem, the limit is A. Thus, the output asymptotically tracks the input signal.

Some engineering interpretation. *$\frac{1}{s^2}$ can be viewed as a map from acceleration to position: $\frac{d^2y}{dt^2} = u$; part a) in the above suggests that if we only use position error we cannot have a stable tracking system; but if we use position and derivative (that is, velocity) information, then we can make the system stable. Furthermore, if we have a reference tracking problem, the output will indeed track the reference path.*

8.4 Nyquist Stability Criterion

Consider a feedback control system with negative unity feedback as in Figure 8.4, with the loop-transfer function $P(s)C(s)$ and the closed-loop transfer function $\frac{P(s)C(s)}{1+P(s)C(s)}$.

With the Bode plot, we observed that we can identify the transfer function when a system is already stable. However, the Bode plot does not provide insights on how to adjust the controller so that stability is attained. The Root Locus method allows for parametrically adjusting the instability region. Complementing the Root Locus method, the Nyquist plot provides further insight on controller design and its robustness properties to parameter variations, to be discussed further below.

Recall first that a right-half plane pole of $1 + P(s)C(s)$ implies instability. In general, it is not difficult to identify the poles of the (open-loop) transfer function $P(s)C(s)$. Therefore, in the following we will assume that we know the number of right-half plane poles $P(s)C(s)$. Note also that the poles of $P(s)C(s)$ are the same as the poles of $1 + P(s)C(s)$; thus, we will assume that we know the number of right-half plane poles of $1 + P(s)C(s)$.

For the Nyquist plot, we will construct a clockwise contour starting from $-iR$ to the origin and then to $+iR$ and then along a semi-circle of radius R to close the curve. Later on we will take $R \to \infty$.

We will refer to this contour as a Nyquist contour.

If there is a pole of $L(s)$ on the imaginary axis, we should carefully exclude this from our contour: to exclude these, we divert the path along a semi-circle of a very small radius r around the pole in a counter clock-wise fashion (which will later be taken to be arbitrarily close to 0). The exclusion of such a pole will not make a difference in the stability analysis: we focus on the zeroes of $1 + P(s)C(s)$ in the right-half plane (as such a pole cannot make $1 + P(s)C(s) = 0$).

Theorem 8.4.1 Nyquist Criterion. *Consider a closed loop system with the loop transfer function $L(s) = C(s)P(s)$. Suppose that $L(s)$ has P poles in the region encircled by the Nyquist contour. Let N be the number of clockwise encirclements of -1 by L(s) when s encircles the Nyquist contour Γ clock-wise. Then, the closed loop has $N + P$ poles in the right-half plane.*

Note: One can alternatively trace the contours counterclockwise and then count N as the number of counterclockwise encirclements. The result will be the same.

The proof builds on what is known as the *principle of variation of the argument* theorem, which we state and prove next.

Theorem 8.4.2 *Let D be a closed region in the complex plane with Γ its boundary. Let $f : \mathbb{C} \to \mathbb{C}$ be complex differentiable (and hence analytic) on D (and on Γ) except at a finite number of poles and with a finite number of zeroes, all in the interior of D. Then, the change in the argument of f (normalized by 2π) over Γ (known as the winding number w_n) is given by:*

$$w_n = \frac{1}{2\pi}\Delta_\Gamma arg f = \frac{1}{i2\pi}\int_\Gamma \frac{f'(z)}{f(z)}dz = Z - P,$$

where Δ_Γ is the net variation in the angle (or argument) of f when z traces the contour Γ in the counter-clock-wise direction; Z is the number of zeroes, and P is the number of poles (with multiplicities counted).

Proof.
a) Let $z = p$ be a zero of multiplicity m. Then, in small neighbourhoods of p, we have

$$f(z) = (z-p)^m g(z),$$

where g is analytic and non-zero. Now,

$$\frac{f'(z)}{f(z)} = \frac{m(z-p)^{m-1}g(z) + (z-p)^m g'(z)}{(z-p)^m g(z)} = \frac{m}{z-p} + \frac{g'(z)}{g(z)}$$

Thus, $\frac{f'(z)}{f(z)}$ has a single pole at p and the integration (normalized with $\frac{1}{i2\pi}$) will be m (by Cauchy's integral formula Theorem B.0.2). Thus, the sum of all residues at the zeros of f will be Z.

Now, make the same reasoning for the poles, say q, with $f(z) = (z-q)^{-m}g(z)$ and observe that that the sum of the residues at the poles is $-P$.

Hence,

$$Z - P = \frac{1}{i2\pi}\int \frac{f'(z)}{f(z)}dz \tag{8.4}$$

b) Let Γ be parametrized as $\gamma(t), a \le t \le b$, with $\gamma(a) = \gamma(b)$. Recall that f is complex differentiable on all of Γ. Now, write

$$\begin{aligned}\int_\Gamma \frac{f'(z)}{f(z)}dz &= \int_a^b \frac{f'(\gamma(t))}{f(\gamma(t))}\gamma'(t)dt \\ &= \log(f(\gamma(t)))\Big|_a^b = \log(|f(\gamma(t))|)\Big|_a^b + \Big(i2\pi arg(f(\gamma(t)))\Big)\Big|_a^b,\end{aligned}$$

since $\log(f(z)) = \log(|f(z)|) + iarg(f(z))$. As $|f(\gamma(b))| = |f(\gamma(a))|$, we have that

$$\int_{\Gamma} \frac{f'(z)}{f(z)} dz = i2\pi arg f(\gamma(t))\Big|_a^b = i2\pi \Delta_{\Gamma} arg$$

The proof then follows from (8.4). □

Now, to apply this theorem, consider $f(s) = 1 + P(s)C(s)$, as noted earlier observe that the poles of $1 + P(s)C(s)$ are the same as the poles of $P(s)C(s)$. We are interested in the zeroes of $1 + P(s)C(s)$, and whether they are in the right-half plane. So, all we need to compute is the number of zeroes of $1 + P(s)C(s)$ through the difference $N = Z - P$ given by the number of encirclements: Note now that the number encirclements of $1 + P(s)C(s)$ around 0 is the same as the number of encirclements of $P(s)C(s)$ around -1. So, the number of zeroes in the right-half plane of $1 + P(s)C(s)$ will be P plus the winding number. As a final note, in the Nyquist analysis, we construct the contour clockwise and apply the argument principle accordingly (so, the number of encirclements around -1 should be counted *clock-wise*). So, the number of interest is $N + P$, as claimed.

To compute the number of clock-wise encirclements, compute $L(i\omega)$ starting from $\omega = 0$ and increase ω to ∞. Observe that $L(i\omega) = \overline{L(-i\omega)}$, computing $L(i\omega)$ for $\omega > 0$ is sufficient to compute the values for $\omega < 0$. Finally, to compute $L(s)$ as $|s| \to \infty$, we note that often $|L(s)|$ converges to a constant as $|s| \to \infty$, and the essence of the encirclements is given by the changes occurred as s traces the imaginary axis.

Exercise 8.4.1 *a) Consider $C(s) = K$, $P(s) = \frac{1}{(s+1)^2}$. Is this system stable for a given $K > 0$. Explain through the Nyquist stability criterion.*

b) Consider $P(s)C(s) = \frac{1}{(s+a)^3}$ with the controller in an error feedback form so that the closed loop transfer function is given by $\frac{P(s)C(s)}{1+P(s)C(s)}$. Is this system stable? Explain through the Nyquist stability criterion.

c) Let $P(s)C(s) = \frac{3}{(s+1)^3}$. Compute the gain stability margin. Draw the phase stability margin on the Nyquist curve.

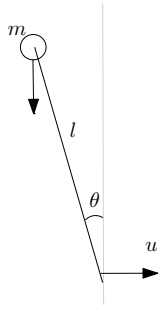

Fig. 8.3

Exercise 8.4.2 *Consider the inverted pendulum displayed in Figure 8.3, where a torque of u is applied to maintain the pendulum around $\theta = 0$. The dynamics can be derived as*

$$\frac{d^2\theta(t)}{dt^2} = \frac{g}{l}\sin(\theta(t)) + \frac{u(t)\cos(\theta(t))}{ml^2}$$

For simplicity, let us assume the coefficients (m, l) are selected so that the above simplifies to:

$$\frac{d^2\theta(t)}{dt^2} = \sin(\theta(t)) + u(t)\cos(\theta(t))$$

Consider the linearization around $\theta = 0, \frac{d\theta}{dt} = 0$ (with the approximation $\sin(\theta) \approx \theta, \cos(\theta) \approx 1$).

Then, it follows that the Plant, modeling the linearized inverted pendulum, would have its transfer function as

$$P(s) = \frac{1}{s^2 - 1}$$

Now, suppose that we apply the control

$$C(s) = k(s + 1),$$

with an error feedback control configuration as in Figure 8.2.

Via the Nyquist crierion, find conditions on k so that the closed-loop linearized system is BIBO stable.

Hint. Note that $P(s)$ has a right-half plane pole, so the Nyquist criterion has to encircle -1 clock-wise.

Robustness

Nyquist's criterion also suggests a robustness analysis: Gain and phase margins, mainly as a way to measure how far $P(s)L(s)$ is from -1 in both the magnitude (1) and phase ((π)) terms.

Roughly speaking, for systems which hit the real-line only once, in the complex plane the angle between -1 and the location where the Nyquist plot hits the unit circle (magnitude equaling 1) is called the phase stability margin. The ratio between -1 and the point where the Nyquist plot hits the negative x-axis is called the the gain stability margin.

8.4.1 System gain, passivity and the small gain theorem

Consider a linear system with feedback, which we assume to be stable: We generalize the observation above by viewing the input as one in $L_2(\mathbb{R};\mathbb{C})$. Consider then the gain of a linear system with:

$$\gamma := \sup_{u \in L_2(\mathbb{R};\mathbb{C}):\|u\|_2 \neq 0} \frac{\|y\|_2}{\|u\|_2}$$

We know, by Parseval's theorem (Theorem 5.3.4), that

$$\gamma := \sup_{u \in L_2(\mathbb{R};\mathbb{C}):\|u\|_2 \neq 0} \frac{\|\mathcal{F}_{CC}(y)\|_2}{\|\mathcal{F}_{CC}(u)\|_2}$$

which is equal to, by writing $\mathcal{F}_{CC}(y)(i\omega) = \mathcal{F}_{CC}(u)(i\omega)G(i\omega)$, where G is the closed-loop transfer function. It can then be shown by noting that

$$\int_\omega |\mathcal{F}_{CC}(u)(i\omega)|^2 G(i\omega)^2 | d\omega \leq \left(\sup_\omega |G(i\omega)|^2 \right) \int |\mathcal{F}_{CC}(u)(i\omega)|^2,$$

the following holds:

$$\gamma := \sup_{u \in L_2(\mathbb{R};\mathbb{C}):\|u\|_2 \neq 0} \sqrt{\frac{\int_\omega |\mathcal{F}_{CC}(u)(i\omega)G(i\omega)|^2 d\omega}{\|\mathcal{F}_{CC}(u)\|^2}} = \sup_{\omega \in \mathbb{R}} |G(i\omega)| =: \|G\|_\infty$$

We write $\sup_{u \in L_2(\mathbb{R};\mathbb{C})}$, since a maximizing frequency may not exist or a maximizing input, even if a maximizing frequency ω^* exists, $u(t) = e^{i\omega^* t}$, will not be square integrable. However, this can be approximated arbitrarily well by truncation of the input and the output: Let $\omega^* = 2\pi f^*$ and $u_K(t) = 1_{\{-\frac{K}{2} \leq t \leq \frac{K}{2}\}} e^{i2\pi f^* t}$. As $K \to \infty$, the gain of the system will approximate γ arbitrarily well.

Then, we define the gain of a linear system as:

$$\gamma := ||G(i\omega)||_\infty = \sup_{\omega \in \mathbb{R}} |G(i\omega)|$$

We note that the same definition can also be applied to multi-input multi-output systems, in which case, it follows that

$$\gamma := ||G(i\omega)||_\infty = \sup_{\omega\in\mathbb{R}} \lambda_{\max}(G(i\omega)^H G(i\omega)).$$

This formula combines the insights presented in the above two formulations.

Define a system to be L_2-stable if a bounded input, in the L_2-sense, leads to a bounded output in the L_2-sense. BIBO stability of a linear system implies L_2-stability: BIBO stability implies $\|h\|_1 < \infty$, which implies that $H(i\omega)$ is uniformly bounded for $\omega \in \mathbb{R}$ so that $|H(i\omega)| \leq \|h\|_1$: Let $u^m \in \mathcal{S}$. Then, observe that the output $\|y^m\|_2^2 = \int |Y^m(i\omega)^2| \frac{1}{2\pi} d\omega = \int |H^2(i\omega)(U^m)^2(i\omega)| \frac{1}{2\pi} d\omega \leq \|h\|_1^2 \int |(U^m)^2(i\omega)| \frac{1}{2\pi} d\omega$. Thus, following Theorem 5.3.6, we can extend the domain (input function space) to be the entire L_2 space: Now, approximate any u with $\|u\|_2 < \infty$ with the sequence $\{u^m \in \mathcal{S}, m \in \mathbb{N}\}$, taking the limit of the output sequence of which lets us conclude that the output $\|y\|_2$ is also bounded. Thus $\|y\|_2^2 = \int |Y(i\omega)^2| \frac{1}{2\pi} d\omega = \int |H^2(i\omega)U^2(i\omega)| \frac{1}{2\pi} d\omega < \infty$.

We next state a useful result on the verification of stability.

Theorem 8.4.3 (Small Gain Theorem) *Consider a feedback control system with closed-loop transfer function* $G(s) = \frac{H_1(s)}{1+H_1(s)H_2(s)}$*, where* H_1 *and* H_2 *are stable. Suppose further that the gains of* H_1 *and* H_2 *are* γ_1 *and* γ_2*, respectively. Then, if* $\gamma_1\gamma_2 < 1$*, the closed-loop system is stable.*

Note that $\sup_\omega \frac{H_1(i\omega)}{1+H_1(i\omega)H_2(i\omega)}$ is uniformly bounded as $1 + H_1(i\omega)H_2(i\omega)$ is uniformly bounded away from 0. The proof follows from Nyquist's criterion: Since H_1, H_2 are stable, they have no poles in the right half-plane. Furthermore, since $\gamma_1\gamma_2 < 1$, $|C(s)P(s)|$ (here: $H_1(i\omega)H_2(i\omega)$) will be uniformly away from the point -1, thus a positive gain margin will be maintained and -1 will not be encircled. The system is then stable.

The above concept does not involve phase properties, and it may be conservative for certain applications. On phase properties, there is a further commonly used concept of *passivity*: By Nyquist's criterion, for stable P and C, if $P(s)C(s)$ is so that the phase is in $(-\pi, \pi)$ for all $s = i\omega, \omega \in \mathbb{R}$, then the closed loop system will be stable.

8.5 Exercises

Exercise 8.5.1 *Let* $C(s) = K$, $P(s) = \frac{s+1}{s(\frac{s}{10}-1)}$.

Study stability properties using the root locus method as K is varied from 0 *to* ∞.

Exercise 8.5.2 *Consider Figure 8.4.*

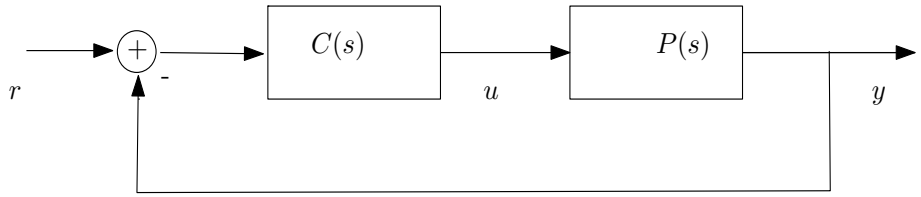

Fig. 8.4

The plant $P(s)$ *is a linearized inverted pendulum and suppose for simplicity that its transfer function is* $P(s) = \frac{1}{s^2-1}$*. Suppose that the controller applied is given with* $C(s) = k(s + 2)$ *(and thus, it is a P-D controller) for some parameter* $k \in \mathbb{R}_+$.

a) Write the plant P *in state space form, where the input is* u *and the output is* y.

b) By writing u as a function of $r(t) - y(t)$, express the overall (closed-loop) system as a linear map from r to y.

c) Find conditions on k for the system to be BIBO stable.

d) Verify that this result is consistent with a Nyquist or root-locus method analysis.

9

Realizability and State Space Representation

So far in the course, we considered the input-output approach to study control systems, which in the convolution (linear time-invariant) system setup, resulted in frequency domain methods through a transfer function analysis (such as in arriving at the root locus / Nyquist stability / Bode plot methods). These methods are often referred to as *classical control design* tools.

In this chapter, we will introduce state-space based methods.

The notion of a state. Suppose that, given $t \in \mathbb{R}$ (or $\mathbb{Z}$), we wish to compute the output of a system at $t \geq t_0$. In a general causal system, we may need to use all the past applied input terms $u(s); s \leq t$ and/or all the past output values $y(s); s < t$ to compute the output at t. The *state* of a system summarizes all the past relevant data that is sufficient to compute the future paths in the sense that if the state at t_0, $x(t_0)$ is given, then to compute $y(t)$, one would only need to use $\{u(s) \quad s \in [t_0, t]\}, \{y(s) \quad s \in [t_0, t)\}$ and $x(t_0)$. In particular, the past $\{y(s), u_s; \quad s < t_0\}$ would not be needed.

Some systems admit a finite-dimensional state representation, some do not.

Control design based on state-space methods is called *modern control design*.

Consider a linear system in state-space form:

$$\frac{dx}{dt} = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + Du(t) \tag{9.1}$$

We will say that such a system is given by the 4-tuple: (A, B, C, D).

We know that the solution to this system is given with

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-s)}Bu(s)ds$$

and

$$y(t) = Ce^{At}x(0) + C\int_0^t e^{A(t-s)}Bu(s)ds + Du(t)$$

Taking the (one-sided) Laplace transform of both sides in (9.1), we obtain for s in the ROC,

$$Y_+(s) = C(sI - A)^{-1}(x(0) + BU_+(s)) + DU_+(s)$$

Assuming $x(0) = 0$, we have the following as the transfer function

$$Y_+(s) = (C(sI - A)^{-1}B + D)U(s)$$

Note that we could also have taken the two-sided Laplace transform (see the analysis in Section 4.6).

9.1 Realizations: Controllable, Observable and Modal Forms

A transfer function $H(s)$ is state-space realizable if there exists *finite dimensional* (A, B, C, D) so that we can write for $s \in \mathbb{C}$, whenever well-defined,

$$H(s) = C(sI - A)^{-1}B + D$$

Theorem 9.1.1 *A transfer function of a linear time-invariant system $H(s)$ is realizable if and only if it is a proper rational fraction (that is, $H(s) = \frac{P(s)}{Q(s)}$ where both the numerator P and the denominator Q are polynomials, and with degree of P less than or equal to the degree of Q).*

Proof. (i) If realizable, then L is proper and rational: Let L be realizable, that is

$$L(s) = C(sI - A)^{-1}B + D$$

for finite-dimensional matrices (A, B, C, D). We will show that this implies that L is proper and that L is a rational fraction.

Now, for a square matrix E which is invertible, we know that $\det(E) = \sum_j E(i,j)(-1)^{i+j}M(i,j)$, where $M(i,j)$ is the determinant of the matrix obtained by deleting the ith row and jth column from E. Let F be the co-factor matrix given by $F(i,j) = (-1)^{i+j}M(i,j)$. It then follows that

$$E^{-1} = \frac{1}{\det(A)}F^T$$

You can verify this by noting that the diagonals of EE^{-1} would be 1 (as the determinant would appear both in the numerator and denominator), and the off-diagonals would be so that the product of each row-column pair would be equivalent to the determinant of a square matrix whose rows are repeated, leading to a zero. Let us call F the adjoint of A. Then,

$$C(sI - A)^{-1}B = \frac{1}{\det(sI - A)}C[Adj(sI - A)]B$$

Since the denominator is a polynomial of order n, and the adjoint matrix consists of polynomials of order at most $n - 1$, it follows that the expression is a fraction, which is in fact also proper.

To verify properness, we can also have the following reasoning: for $|s| > \max_i\{|\lambda_i|\}$, with λ_i the eigenvalues of A, we have that

$$(sI - A)^{-1} = (s(I - s^{-1}A))^{-1} = s^{-1}(I - s^{-1}A)^{-1} = s^{-1}\sum_{k=0}^{\infty}(s^{-1}A)^k$$

Since $\lim_{|s|\to\infty} |s^{-1}\sum_{k=0}^{\infty}(s^{-1}A)^k| = 0$, it must be that $C(sI - A)^{-1}B$ is strictly proper.

The presence of a non-zero D is what may lead the transfer function to be proper, and not strictly proper.

(ii) If L is proper and rational, then it is realizable: The realizations will be constructed explicitly below under various setups; through various canonical realizable forms. □

Exercise 9.1.1 *Can you construct a causal system which does not admit a rational transfer function? Hint: Consider $y(t) = au(t-1)$, where $t \in \mathbb{R}$ and a is a scalar. Or in a feedback loop: $y(t) = a(r(t-1) + y(t-1))$, $t \in \mathbb{R}$. Such systems with no rational transfer function are sometimes called distributed parameter systems.*

9.1.1 Controllable canonical realization

Consider a continuous-time system given by:

$$\sum_{k=0}^{N} a_k \frac{d^k}{dt^k}y(t) = \sum_{m=0}^{N} b_m \frac{d^m}{dt^m}u(t), \tag{9.2}$$

with $a_N = 1$. Taking the Laplace transform, we know that the transfer function writes as

$$H(s) = \frac{\sum_{m=0}^{N} b_m s^m}{\sum_{k=0}^{N} a_k s^s}$$

Suppose that the system is strictly proper. Such a system can be realized with the form:

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

$$A_c = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{N-1} \end{bmatrix}$$

$$B_c = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

$$C_c = \begin{bmatrix} b_0 & b_1 & \cdots & b_{N-1} \end{bmatrix}$$

If the system is proper, but not strictly proper, then, we will also have $D_c = d$, where d is the remainder term in the partial fraction expansion,

$$H(s) = d + \sum_{i=1}^{K} \left(\sum_{k=1}^{m_i} \frac{A_{ik}}{(s - p_i)^k} \right) \tag{9.3}$$

where p_i are the roots of $s^n + \sum_{k=0}^{N-1} a_k s^k$ and m_i is the multiplicity of p_i.

9.1.2 Observable canonical realization

Consider (9.2). This system can also be realized as

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

with

$$A = \begin{bmatrix} -a_{N-1} & 1 & 0 & \cdots & 0 \\ -a_{N-2} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ -a_1 & 0 & 0 & \cdots & 1 \\ -a_0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} b_{N-1} \\ b_{N-2} \\ \vdots \\ b_0 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$$

If we reverse the order of the coordinates of x, we arrive at;

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

with

$$A_o = \begin{bmatrix} 0 & 0 & 0 & \cdots & -a_0 \\ 1 & 0 & 0 & \cdots & -a_1 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & -a_{N-2} \\ 0 & 0 & 0 & \cdots & -a_{N-1} \end{bmatrix}$$

$$B_o = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{N-1} \end{bmatrix}$$

$$C_o = \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}$$

Observe that $A_o = A_c^T, B_o = C_c^T, C_o = B_c^T$. This is the *standard observable canonical form*.

Exercise 9.1.2 *Show that the transfer functions under the controllable and observable realization canonical forms are equivalent directly by comparing $C_c(sI - A_c)^{-1}B_c$ and $C_o(sI - A_o)^{-1}B_o$.*

9.1.3 Modal realization

Consider a partial fraction expansion (9.3) with only simple poles:

$$H(s) = \frac{\sum_{m=0}^{N} b_m s^m}{\sum_{k=0}^{N} a_k s^s} = d + \sum_{i}^{N} \frac{k_i}{s - p_i}$$

In this case, we can realize the system as the sum of decoupled modes:

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

with

$$A = \begin{bmatrix} p_1 & 0 & 0 & \cdots & 0 \\ 0 & p_2 & 0 & \cdots & \cdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & p_N \end{bmatrix}$$

$$B = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_N \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$$

If in the partial fraction expansion is more general as in (9.3), then the corresponding structure can be realized also: this will lead to a Jordan form for the matrix A, since, e.g.,

$$\frac{1}{(s - p_i)^2} = \frac{1}{s - p_i}\frac{1}{s - p_i}$$

will define a serial connection of two modal blocks; the first one with $x_2' = p_i x_2 + u$ and the second one $x_1' = p_i x_1 + x_2$.

Please see our class notes where we presented diagrams depicting each of the forms above.

Discrete-time setup. The above also apply to the discrete-time setup. For example, a discrete-time system of the form

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{m=1}^{N} b_m u(n-m)$$

with $a_0 = 1$, can be written in the controllable canonical form

$$x(n+1) = Ax(n) + Bu(n), \qquad y_t = Cx_t$$

where

$$x_N(n) = y(n), x_{N-1}(n) = y(n-1), \quad \cdots, x_1(n) = y(n-(N-1))$$

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \cdots & 1 \\ -a_N & -a_{N-1} & -a_{N-2} & \cdots & -a_1 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

$$C = \begin{bmatrix} b_N & b_{N-1} & \cdots & b_1 \end{bmatrix}$$

Observable and modal canonical forms follow similarly.

9.2 Zero-State Equivalence and Algebraic Equivalence

We say that two systems (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are **zero-state equivalent** if the induced transfer functions are equal, that is

$$C(sI - A)^{-1}B + D = \tilde{C}(sI - \tilde{A})^{-1}\tilde{B} + \tilde{D}$$

Theorem 9.2.1 *Two linear time-invariant state-space models (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are zero-state equivalent if and only if $D = \tilde{D}$ and $CA^mB = \tilde{C}\tilde{A}^m\tilde{B}$ for all $m \in \mathbb{Z}_+$.*

Proof. For $|s| > \max_i\{|\lambda_i|\}$, with λ_i the eigenvalues of A, we have that $s^{-1}A$ will have all of its eigenvalues less than 1.

$$(sI - A)^{-1} = (s(I - s^{-1}A))^{-1} = s^{-1}(I - s^{-1}A)^{-1} = s^{-1}\sum_{k=0}^{\infty}(s^{-1}A)^k$$

Then,

$$C(sI - A)^{-1}B + D = \left(\sum_{k\in\mathbb{Z}_+} Cs^{-1}(s^{-1}A)^kB\right) + D = s^{-1}\left(\sum_{k\in\mathbb{Z}_+} C(s^{-1}A)^kB\right) + D$$

and

$$\tilde{C}(sI - \tilde{A})^{-1}\tilde{B} + \tilde{D} = \left(\sum_{k\in\mathbb{Z}_+} \tilde{C}s^{-1}(s^{-1}\tilde{A})^k\tilde{B}\right) + \tilde{D} = s^{-1}\left(\sum_{k\in\mathbb{Z}_+} \tilde{C}(s^{-1}\tilde{A})^k\tilde{B}\right) + \tilde{D}$$

Since for all sufficiently large $|s|$ values, s is in the region of convergence and the above are equal functions of s, meaning their difference is the zero function identically; this implies that all the coefficients in the expansions must be equal. Prove

this last statement as an exercise (e.g., by writing $\bar{s} = s^{-1}$ and study the behaviour of the difference between two absolutely summable polynomials $\sum_{k=0}^{\infty} \alpha_k \bar{s}^k - \sum_{k=0}^{\infty} \beta_k \bar{s}^k = 0$ for all $\bar{s}$ in a neighborhood around the origin, and conclude that $\alpha_k = \beta_k$ must be equal). Note that $\tilde{D} = D$ also follows as a result.

□

Note that by the Cayley-Hamilton theorem, it suffices to test the relation $CA^m B = \tilde{C}\tilde{A}^m\tilde{B}$ for $m = 0, \cdots, n-1$ where n is the larger of the dimensions of A or $\tilde{A}$.

There is an alternative notion, called **algebraic equivalence**: Let P be invertible and let us define a transformation through $\tilde{x} = Px$. Then, we can write the model (9.1) as

$$\frac{d\tilde{x}}{dt} = \tilde{A}\tilde{x}(t) + \tilde{B}u(t), \qquad \tilde{y}(t) = \tilde{C}\tilde{x}(t) + \tilde{D}u(t),$$

with $\tilde{A} = PAP^{-1}, \tilde{B} = PB, \tilde{C} = CP^{-1}, \tilde{D} = D$. In this case, we say that (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are algebraically equivalent.

Theorem 9.2.2 *Show that algebraic equivalence implies zero-state equivalence but not vice versa.*

Proof. Observe that for every $k \in \mathbb{Z}_+$ $\tilde{C}\tilde{A}^k\tilde{B} = CP^{-1}\Big(PA^kP^{-1}\Big)PB = CA^m = kB$. The reverse implication is not correct. In particular, one can always artificially add further state variables to arrive at a larger matrix A which, through zeroes in A and C so that the new component has no impact on the output variable y. □

9.3 Discretization

Consider

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

Suppose that we apply piece-wise constant control actions u which are varied only at the discrete time instances given with $\{t : t = kT, k \in \mathbb{Z}_+\}$ so that $u(t) = u(kT)$ for $t \in [kT, (k+1)T)$.

We write

$$x((k+1)T) = e^{AT}x(kT) + \left(\int_{kT}^{(k+1)T} e^{A((k+1)T-s)}Bu(s)ds\right)$$

Writing $\tau = (k+1)T - s$ with $d\tau = -ds$, we arrive at

$$x((k+1)T) = e^{AT}x(kT) + \left(\int_0^T e^{A\tau}Bd\tau\right)u(kT)$$

With $x_k := x(kT)$ and $u_k := u(kT)$, we arrive at

$$x_{k+1} = A_d x_k + B_d u_k$$

where

$$A_d = e^{AT}$$

$$B_d = \int_0^T e^{A\tau}Bd\tau$$

If A is invertible, the integration of $\int e^{A\tau}d\tau$ leads to $A^{-1}(e^{AT} - I)B$.

10

The Sampling Theorem

One important requirement for signal transmission and storage is the need for discretization of the signal, both in time-index sets and in signal-range sets. The former is called sampling; the latter is called quantization. In the following, we discuss sampling.

10.1 The Sampling Theorem

Samplers perform the discretization in the time-index. We discuss a very important theorem in the following.

10.1.1 Sampling of a Continuous-Time (CT) Signal

Theorem 10.1.1 (Shannon-Nyquist Sampling Theorem) *Let* $\{x(t), \quad t \in \mathbb{R}\}$ *be a CT signal in* $L_2(\mathbb{R};\mathbb{R})$ *and let* $\hat{x} = \mathcal{F}_{CC}(x)$. *If this signal has a finite bandwidth* B, *that is if*

$$\hat{x}(f) = 0, \quad |f| > B,$$

(that is, the support of $\hat{x}$ *is contained in* $[-B, B]$*) then it is possible to reconstruct this signal by samples of this signal (with arbitrarily small error in the* $L_2(\mathbb{R};\mathbb{R})$ *sense, i.e., in the* $\|\cdot\|_2$ *norm), where the sampling period* T *that allows this satisfies*

$$\frac{1}{2T} > B.$$

If this condition holds, the recovery is satisfied by the relation

$$\tilde{x}(t) = \sum_{n\in\mathbb{Z}} x(nT) \frac{\sin(\pi \frac{t-nT}{T})}{\pi \frac{t-nT}{T}},$$

where $\{x(nT), n \in \mathbb{Z}\}$ *denotes the samples of the signal* x.

This is a remarkable result which paves the way to communications, signal processing, and digital control technologies, which are ubiquitous in our daily lives. For example, human voice range is typically about 20 Hz to 20 kHz (female voice typically has a higher frequency band than male voice), but primarily most of the content is between 2 to 4 KHz. The bandwidth allocated for a telephone transmission channel is about 4 kHz. These suggest that one can recover human voice quite well with large enough samples taken: $T = \frac{1}{8000}$ seconds.

We now present a sketch of the proof of the result above. Let $x \in \mathcal{S}$: We know that, by Theorem 5.3.5, any $x \in L_2(\mathbb{R};\mathbb{R})$ can be approximated arbitrarily well by a signal in $\mathcal{S}$ and therefore, in the following, we can assume that $x \in \mathcal{S}$ (with an arbitrarily small approximation error in the $\|\cdot\|_2$ norm).

Consider the sampled signal

$$x_p(t) = \int x(t) \sum_{k\in\mathbb{Z}} \delta(t-kT), t \in \mathbb{R},$$

where $\sum_{k\in\mathbb{Z}} \delta(t-kT)$ is called an *impulse train*. Observe that, the impulse train is a distribution, and recall from Section 5.4 that the CCFT of a distribution is itself a distribution. We will consider $\hat{x}_p(f)$, which denotes the CCFT of the sampled signal, which now needs to be viewed as a distribution.

Now let us discuss what the CCFT for the impulse train $\sum_{k\in\mathbb{Z}} \delta(t-kT)$ should be:

Exercise 10.1.1 *Consider an impulse train defined by:*

$$w_P(t) = \sum_{n\in\mathbb{Z}} \delta(t+nP)$$

so that the distribution that we can associate with this impulse train would be defined by:

$$\overline{w_P}(\phi) = \sum_{n\in\mathbb{Z}} \phi(nP) = \lim_{N\to\infty} \sum_{n=-N}^{N} \phi(nP),$$

for $\phi \in \mathcal{S}$.

a) Show that $\overline{w_P}$ is a distribution.

b) Show that

$$\hat{\overline{w_P}}(\phi) = \int \frac{1}{P} w_{\frac{1}{P}}(t)\phi(t)dt,$$

that is, the $\mathcal{F}_{CC}$ of this train is another impulse train.

Solution. *a) Let $\phi_n \to 0$ in $\mathcal{S}$. Then, since $\sup_t |\phi_n(t)(1+t^2)| \to 0$,*

$$\begin{aligned} |\overline{w_P}(\phi)| &= |\sum_{k\in\mathbb{Z}} \phi_n(kP)| \\ &= |\sum_{k\in\mathbb{Z}} \frac{1}{1+k^2P^2}\Big(\phi_n(kP)(1+k^2P^2)\Big)| \le \sum_{k\in\mathbb{Z}} \frac{1}{1+k^2P^2} \sup_{k\in\mathbb{Z}}\Big(\phi_n(kP)(1+k^2P^2)\Big) \to 0 \end{aligned} \tag{10.1}$$

Thus, $\overline{w_P}$ is continuous on $\mathcal{S}$. Furthermore, $\overline{w_P}$ is linear on on $\mathcal{S}$ since for any real α, β and $\phi_1, \phi_2 \in \mathcal{S}$,

$$\overline{w_P}(\alpha\phi_1 + \beta\phi_2) = \alpha\overline{w_P}(\phi_1) + \beta\overline{w_P}(\phi_2).$$

Thus, $\overline{w_P}$ is a distribution.

b) By Section 5.4, we have that $\hat{\overline{w_P}}(\phi) = \overline{w_P}(\hat{\phi})$. Then,

$$\begin{aligned} \hat{\overline{w_P}}(\phi) = \overline{w_P}(\hat{\phi}) &= \lim_{N\to\infty} \sum_{k=-N}^{N} \hat{\phi}(kP) \\ &= \lim_{N\to\infty} \sum_{k=-N}^{N} (\int \phi(t) e^{-i2\pi kPt} dt) \\ &= \lim_{N\to\infty} \int \phi(t) \Big(\sum_{k=-N}^{N} e^{-i2\pi kPt} \Big) dt \\ &= \lim_{N\to\infty} \int \phi(t) \Big(e^{i2\pi NPt} \Big(\sum_{k=0}^{2N} e^{-i2\pi kPt} \Big) dt \end{aligned}$$

$$
\begin{aligned}
&= \lim_{N\to\infty} \int \phi(t) \left(e^{i2\pi NPt} \frac{1 - e^{-i2\pi(2N+1)Pt}}{1 - e^{-i2\pi Pt}} \right) dt \\
&= \lim_{N\to\infty} \int \phi(t) \left(e^{i2\pi NPt} \frac{e^{i2\pi(1/2)Pt} - e^{-i2\pi(2N+1/2)Pt}}{e^{i\pi Pt} - e^{-i\pi Pt}} \right) dt \\
&= \lim_{N\to\infty} \int \phi(t) \left(\frac{e^{i2\pi(N+1/2)Pt} - e^{-i2\pi(N+1/2)Pt}}{e^{i\pi Pt} - e^{-i\pi Pt}} \right) dt \\
&= \lim_{N\to\infty} \int \phi(t) \left(\frac{\sin(2\pi(N+1/2)Pt)}{\sin(\pi Pt)} \right) dt \\
&= \lim_{N\to\infty} \int \phi(t) \frac{1}{P} \left(\frac{\pi Pt}{\sin(\pi Pt)} \frac{\sin(2\pi(N+1/2)Pt)}{\pi t} \right) dt \\
&= \lim_{N\to\infty} \int \phi(t) \frac{1}{P} \left(\frac{\pi Pt}{\sin(\pi Pt)} \frac{\sin(2\pi(N+1/2)Pt)}{\pi t} \right) dt \\
&= \lim_{N\to\infty} \sum_{k=-N}^{N} \frac{1}{P} \phi(\frac{k}{P}) \\
&= \int \frac{1}{P} w_{\frac{1}{P}}(t) \phi(t) dt
\end{aligned} \tag{10.2}
$$

where we use Fubini's theorem in the second equality. Observe that $\frac{\sin(2\pi(N+1/2)Pt)}{\sin(\pi Pt)}$ *is periodic with period* $1/P$*. For* $t \in [-\frac{1}{2P}, \frac{1}{2P}]$*, this function can be shown, as we studied earlier (see Theorem 3.3.2) to (as a distribution) converge to a delta function. Due to the periodicity, the result follows.* □

Remark 10.1. As we observed earlier and above, the sequence of partial sums $\left(\sum_{k=-N}^{N} e^{-i2\pi kPt} \right)$ of partial sums converges to a periodic Dirac delta function as a distribution (see also Theorem 3.3.2). This sequence is known as the Dirichlet kernel sequence.

It then follows that, the $\mathcal{F}_{CC}$ of $\sum_{k\in\mathbb{Z}} \delta(t - kT)$ is another impulse train given as:

$$\hat{I}(f) = \frac{1}{T} \sum_k \delta(f - \frac{k}{T}).$$

This result is in agreement with an engineering intuition that if one considers I as a periodic signal, and computes its $\mathcal{F}_{CD}$ as $\hat{I}(\frac{k}{T}) = \frac{1}{\sqrt{T}} 1$ for all $k \in \mathbb{Z}$, this leads to

$$I(t) = \sum_k \frac{1}{\sqrt{T}} (\frac{1}{\sqrt{T}}) e^{i2\pi \frac{k}{T} t}$$

Expressing this in an integral form, we obtain that

$$\hat{I}(f) = \frac{1}{T} \sum_k \delta(f - \frac{k}{T}).$$

The $\mathcal{F}_{CC}$ of $x_p(t)$ then satisfies, by the properties of convolution in frequency domain corresponding to a time-wise product in time domain, the following:

$$\hat{x}_p(f) = (\hat{I} * \hat{x})(f).$$

It follows after some analysis the following relation holds

$$\hat{x}_p(f) = \frac{1}{T} \sum_{m\in\mathbb{Z}} \hat{x}(f - m\frac{1}{T}),$$

which is an addition of shifted versions of $\hat{x}$.

If T is small enough, then the signal

$$\hat{\hat{x}}(f) = T1_{(|f|<\frac{1}{2T})}\hat{x}_p(f),$$

is exactly equal to the $\mathcal{F}_{CC}$ of the original signal. Hence, a low-pass filter which takes out the repetitions of the frequency shifted due to the convolution with the impulse sequence, except for the primary component, leads to the reconstruction of the original signal.

Since the rectangular low-pass filter $T1_{(|f|<\frac{1}{2T})}$ corresponds to

$$h(t) = \frac{\sin(\pi\frac{t}{T})}{\pi\frac{t}{T}},$$

the reconstruction can then be written as:

$$\tilde{x}(t) = \sum_{n\in\mathbb{Z}} \hat{x}_p(nT)\frac{\sin(\pi\frac{t-nT}{T})}{\pi\frac{t-nT}{T}}.$$

It also follows that if $T > \frac{1}{2B}$, the exact reconstruction will not be possible since the shifted frequency components will overlap with each other. This is known as *aliasing*.

Remark 10.2. To see that the informational content of a finite bandwidth signal is captured by sampling, consider the following (Shannon's argument): Since $\hat{x}$ has bounded support $[-B, B]$, $x(\frac{k}{2B}) = \int \hat{x}(f)\frac{1}{2B}e^{-i\frac{2\pi}{2B}fk}df$ can be viewed to be the $\mathcal{F}_{CD}$ of $\hat{f}$, but now in time domain. Having access to $x(\frac{k}{2B}), k \in \mathbb{Z}$ then is sufficient to recover $\hat{x}$ which can then be used to recover the original signal. The more detailed analysis above provides an explicit construction, and also further insight on the approximation error when there is aliasing. The same analysis applies for the discrete-time setting to be discussed next.

10.1.2 Sampling of a Discrete-Time (DT) Signal

A similar discussion applies to DT signals. We can view a discrete signal as a sampled continuous-time signal and hence as a distribution; and consider the sampling of a discrete-time signal as the sampling of a distribution. The discussion, leaving the details involving the convolution operators of distributions, leads to the following: Let $\{x(n)\}$ be a DT signal. If this signal has a finite bandwidth as B, that is if

$$\hat{x}(f) = 0, \quad |f| > B, \quad f \in [0, 1),$$

then it is possible to reconstruct exactly this signal by samples of this signal, where the sampling period N that allows this satisfies:

$$\frac{1}{2N} > B$$

As in the CT-case, the function that we use for sampling is the discrete-time impulse train given by.

$$I(n) = \sum_{k\in\mathbb{Z}} \delta(n - kN)$$

We can take the truncated sum

$$I^{(M)}(n) = \sum_{k\in -M}^{M} \delta(n - kN)$$

and work with this signal, which is in $l_2(\mathbb{Z}; \mathbb{R})$ and construct an approximate signal (via Parseval's theorem after reconstruction of the signal via the low-pass filter) through the idealized analysis noted below.

As earlier, using the fact that $\sum_{-M}^{M} e^{-i2\pi fkN}$, converges as a distribution, as $M \to \infty$, to an impulse train; the DCFT of I sequence is equal to the distribution represented by:

$$\hat{I}(f) = \frac{1}{N} \sum_{k \in \mathbb{Z}} \delta(f - \frac{k}{N})$$

Therefore, let $x_N(n) = x(n) \times I(n)$. It follows that

$$\hat{x}_N(f) = \frac{1}{N} \sum_{k \in \mathbb{Z}} \hat{x}(f - \frac{k}{N})$$

Thus, the same discussion we had above for the CT case applies here also. Note that if we had truncated the sum as in $I^{(M)}$ above, the reconstruction would be nearly identical to the term above with the error bounded via a Parseval analysis.

Applying a low-pass filter with Fourier transform given by $N1_{(|f|<\frac{1}{2N})}$, whose inverse Fourier is the kernel is

$$h(n) = \frac{\sin(\frac{\pi n}{N})}{\frac{\pi n}{N}}$$

leads to the reconstruction to be written as:

$$\tilde{x}(n) = \sum_{k \in \mathbb{Z}} x(kN) \frac{\sin(\pi \frac{n-kN}{N})}{\pi \frac{n-kN}{N}}.$$

One further remark follows: When a signal is sampled, the unsampled components become zero. Since it is already known that these values are zero, they can be taken out. This is known as decimation. Decimation has the effect of enlarging the bandwidth of the sampled signal, but it does not lead to an information loss. That is, let $x_d(n) = x_N(nN) = x(nN)$ for all N where x satisfies the conditions noted above to allow for recovery from its samples. Then, we have

$$\tilde{x}(n) = \sum_{k \in \mathbb{Z}} x_d(k) \frac{\sin(\pi \frac{n-kN}{N})}{\pi \frac{n-kN}{N}}.$$

Furthermore, we have that

$$\begin{aligned} \hat{x}_d(f) &= \sum_{n \in \mathbb{Z}} x_d(n) e^{-i2\pi fn} = \sum_{n \in \mathbb{Z}} x_N(nN) e^{-i2\pi fn} \\ &= \sum_{n \in \mathbb{Z}} x_N(nN) e^{-i2\pi \frac{f}{N} nN} = \sum_{k \in \mathbb{Z}} x_N(k) e^{-i2\pi \frac{f}{N} k} = \hat{x}_N(\frac{f}{N}) \end{aligned} \tag{10.3}$$

10.2 Exercises

Exercise 10.2.1 *Suppose that we have a continuous time signal $x \in L_2(\mathbb{R}; \mathbb{R})$ given by:*

$$x(t) = m1_{\{|t| \leq \frac{1}{m}\}},$$

where $m \in \mathbb{R}_+$ is a constant. Note that this signal has a bounded support.

Suppose that we sample this signal with a period T. Our goal is to try to reconstruct this signal from its samples after passing the sampled signal through a low-pass filter with a frequency response

$$\hat{h}(f) = T1_{\{|f| \leq \frac{1}{2T}\}}$$

Let $\bar{x}$ denote the reconstructed signal.

a) Given m and T, is perfect reconstruction possible (in the $L_2(\mathbb{R};\mathbb{R})$ sense)? For what values of m, T?

b) Given an arbitrary m and T, find an upper bound on

$$\int_{\mathbb{R}} |x(t) - \bar{x}(t)|^2 dt$$

as a function of the $\mathcal{F}_{CC}$ of x.

Hint: For part b, you may use Parseval's Theorem. One may expect that if the energy of the signal frequencies outside $[-1/2T, 1/2T]$ band is small, the reconstruction error may also be small.

Exercise 10.2.2 *a) Consider a discrete-time signal $\{x(n)\}$ with a bandwidth B. A discrete-time sampler samples this signal with a period N such that the sampled signal satisfies*

$$x_p(n) = \begin{cases} x(n) & \text{if } 0 \equiv n \mod N, \\ 0 & \text{else.} \end{cases}$$

Following this, a decimator is applied to the system to obtain the signal:

$$x_d(n) = x_p(nN).$$

This new signal is stored in a storage device such as a recorder. Later, the original signal is attempted to be recovered from the storage device. What should the relation between B and N be such that, such a recovery is perfect, that is

$$\sup_n |x(n) - \tilde{x}(n)| = 0,$$

where $\{\tilde{x}(n)\}$ denotes the reconstructed signal.

Identify the steps such that $\{\tilde{x}(n)\}$ is recovered from $\{x_d(n)\}$.

b) Typically human voice has a bandwidth of 4kHz. Suppose we wish to store a speech signal with bandwidth equal to 4kHz with a recorder. Since the recorder has finite memory, one needs to sample the signal. What is the maximum sampling period (in seconds) to be able to reconstruct this signal with no error.

Exercise 10.2.3 *Consider an impulse train defined by:*

$$w_P(t) = \sum_{n\in\mathbb{Z}} \delta(t + nP)$$

so that the distribution that we can associate with this impulse train would be defined by:

$$\overline{w_P}(\phi) = \sum_{n\in\mathbb{Z}} \phi(nP),$$

for $\phi \in \mathcal{S}$.

a) Show that $\overline{w_P}$ is a distribution.

b) Show that

$$\hat{w_P}(\phi) = \int \frac{1}{P} w_{\frac{1}{P}}(t)\phi(t)dt,$$

that is, the $\mathcal{F}_{CC}$ of this train is another impulse train.

Exercise 10.2.4 *Consider a discrete-time signal $\{h(n)\}$ with DCFT as:*

$$\hat{h}(f) = 1_{(f \in ([0,\frac{1}{8}] \cup (\frac{7}{8},1)))} \quad f \in [0,1).$$

Determine the DCFT of the signal $g(n) = h(2n), n \in \mathbb{Z}$.

Exercise 10.2.5 *a) Let $m(t)$ be a real-valued signal with a bandwidth B. One can use a transformation, known as the Hilbert transform, to further compress the signal. This transform makes use of the fact that, the Fourier transform of a real signal is conjugate symmetric.*

Let $\bar{x}$ denote the Hilbert transform of a signal x in L_2, the space of square integrable functions with the usual inner product. The CCFT of the Hilbert transform of a signal is given by:

$$\hat{\bar{x}}(f) = -i\mathrm{sign(f)\hat{x}(f)}.$$

Using this relation, prove that the Hilbert transform of a signal is orthogonal to the signal itself.

b) Briefly describe the (double-sideband) Amplitude Modulation (AM) and the Frequency Modulation (FM) techniques for radio communications.

Using the result in part a), one can further suppress the bandwidth requirement for the double-sideband Amplitude Modulation (AM) technique, leading to a single-sideband AM signal.

Exercise 10.2.6 *Consider a square-integrable signal with non-zero L_2 norm, with bounded support. That is, there exists a compact set, outside of which this signal is identically zero. Can the CCFT of such a signal, with a bounded support in time-domain, also have bounded support in frequency domain?*

You may use any method you wish to use to answer this question.

11

Stability and Lyapunov's Method

11.1 Introduction

In many engineering applications, one wants to make sure the system behaves well in the long-run; without worrying too much about the particular path the system takes (so long as these paths are acceptable). Stability is the characterization ensuring that a system behaves well in the long-run. We will make this statement precise in the following.

11.2 Stability Criteria

Consider $\frac{dx}{dt} = f(x)$, with initial condition $x(0)$. Let $f(0) = 0$. We will consider three definitions of stability:

(i) Local Stability (around an equilibrium point): For every $\epsilon > 0$, there exists $\delta > 0$ such that $\|x(0)\| < \delta$ implies that $\|x(t)\| < \epsilon$, $\forall t \geq 0$ (This is also known as *stability in the sense of Lyapunov*).

(ii) Local Asymptotic Stability (around an equilibrium point): $\exists \delta > 0$ such that $\|x(0)\| < \delta$ implies that $\lim_{t\to\infty} \|x(t)\| = 0$.

(iii) Global Asymptotic Stability: For every $x(0) \in \mathbb{R}^n$, $\lim_{t\to\infty} \|x(t)\| = 0$. Hence, here, for any initial condition, the system converges to 0.

It should be added that, stability does not necessarily need to be defined with regard to 0; that is stability can be defined around any $z \in \mathbb{R}^n$ with $f(z) = 0$. In this case, the above norms above should be replaced with $\|x(t) - z\|$ (such that, for example for the asymptotic stability case, $x(t)$ will converge to z).

One could consider an inverted pendulum as an example of a system which is not locally stable.

11.2.1 Linear Systems

Consider an initial value problem $\frac{dx}{dt} = Ax$, with initial conditions $x(0) = x_0$. As studied earlier, the solution is

$$x(t) = e^{At}x_0$$

where

$$e^{At} = I + At + A^2\frac{t^2}{2} + \dots A^n\frac{t^n}{n!} + \dots$$

is the matrix exponential (see Exercise 11.6.1). We now briefly review how to compute matrix exponentials with a focus on stability properties. Consider first a 2×2 matrix

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

In this case, $e^{At} = I + It + I^2\frac{t^2}{2!} + \cdots + I^n\frac{t^n}{n!} + \ldots$, and

$$e^{At} = \begin{bmatrix} e^t & 0 \\ 0 & e^t \end{bmatrix}$$

With similar arguments, if A is diagonal with

$$A = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix},$$

we obtain

$$e^{At} = \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 \\ 0 & e^{\lambda_2 t} & 0 \\ 0 & 0 & e^{\lambda_3 t} \end{bmatrix},$$

Hence, it is very easy to compute the exponential when the matrix is diagonal.

Note now that if $AB = BA$, that is, if A and B commute, then (see Exercise 11.6.2)

$$e^{(A+B)} = e^A e^B$$

We will use this to compute the matrix exponential in the case where the matrix is in a Jordan form. Let us write

$$A = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix}$$

as $B + C$, where

$$B = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_1 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

We note that $BC = CB$, for B is the identity matrix multiplied by a scalar number. Hence, $e^{At} = e^{Bt}e^{Ct}$. All we need to compute is e^{Ct}, as we have already discussed how to compute e^{Bt}. Here, one should note that $C^3 = 0$.

More generally, for a Jordan matrix where the number of 1s off the diagonal of a Jordan block is $k - 1$, the kth power is equal to 0.

Therefore,

$$e^{Ct} = I + Ct + C^2\frac{t^2}{2!} + C^3\frac{t^3}{3!} + \ldots,$$

becomes

$$e^{Ct} = I + Ct + C^2\frac{t^2}{2!} = \begin{bmatrix} 1 & t & t^2/2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}$$

Hence,

$$e^{At} = \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 \\ 0 & e^{\lambda_1 t} & 0 \\ 0 & 0 & e^{\lambda_1 t} \end{bmatrix} \begin{bmatrix} 1 & t & t^2/2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} & te^{\lambda_1 t} & \frac{t^2}{2}e^{\lambda_1 t} \\ 0 & e^{\lambda_1 t} & te^{\lambda_1 t} \\ 0 & 0 & e^{\lambda_1 t} \end{bmatrix}$$

Now that we know how to compute the exponential of a Jordan form, we can proceed to study a general matrix.

Let $A = PBP^{-1}$, where B is in a Jordan form. Then,

$$A^k = (PBP^{-1})^k = PB^kP$$

Finally,

$$e^A = Pe^B P^{-1}$$

and

$$e^{At} = Pe^{Bt} P^{-1}$$

Hence, once we obtain a diagonal matrix or a Jordan form matrix B, we can compute the exponential e^{At} very efficiently.

The main insight here is that the eigenvalues determine whether the system remains bounded or not. In case, we have a repeated eigenvalue of 0, then the Jordan form determines whether the solution remains bounded or not. We state the following theorem.

Theorem 11.2.1 *For a linear differential equation*

$$x' = Ax,$$

the solution is locally and globally asymptotically stable if and only if

$$\max_{\lambda_i}\{\mathrm{Re}\{\lambda_\mathrm{i}\}\} < 0,$$

where $\mathrm{Re}\{.\}$ *denotes the real part of a complex number, and* λ_i *denotes the* i*th eigenvalue of A.*

Theorem 11.2.2 *For a linear differential equation*

$$x' = Ax,$$

the system is locally stable if and only if the following two conditions hold

(i) $\max_{\lambda_i}\{\mathrm{Re}\{\lambda_\mathrm{i}\}\} \leq 0,$

(ii) if $\mathrm{Re}\{\lambda_\mathrm{i}\} = 0$*, for some* λ_i*, the algebraic multiplicity of this eigenvalue should be the same as the geometric multiplicity.*

In practice, many systems are not linear, and hence the above theorems are not applicable.

11.3 A General Approach: Lyapunov's Method

A very versatile and effective approach to stabilization is via *Lyapunov functions* (this approach is often called Lyapunov's second method, the first one being an analysis based on linearization to be considered after this section). Let $\Omega \subset \mathbb{R}^n$ be an open set containing the equilibrium point, taken to be 0 here without any loss. A function : $\mathbb{R}^n \to \mathbb{R}$ is called a Lyapunov function if

1. $V(x) > 0, \forall x \neq 0, x \in \Omega$,
2. $V(x) = 0$, if $x = 0$,
3. V is continuous, and has continuous partial derivatives.

First we present results on local asymptotic stability. As above, let $\Omega \in \mathbb{R}^n$ be an open set containing 0.

Theorem 11.3.1 *a) For a given differential equation* $x'(t) = f(x(t))$ *with* $f(0) = 0$*, and continuous* f*, if we can find a Lyapunov function* V *such that*

$$\frac{d}{dt}V(x(t)) \leq 0,$$

for all $x(t) = x \in \Omega$*, then, the system is locally stable (stable in the sense of Lyapunov).*
b) For a given differential equation $x'(t) = f(x(t))$ *with* $f(0) = 0$*, and continuous* f*, if we can find a Lyapunov function* $V(x)$ *such that*

$$\frac{d}{dt}V(x(t)) < 0,$$

for $x(t) = x \in \Omega \setminus \{0\}$, the system is locally asymptotically stable.
c) If b) holds for V so that $\lim_{\|x\|\to\infty} V(x) = \infty$, and

$$\frac{d}{dt}V(x(t)) < 0,$$

for $x(t) = z \in \mathbb{R}^n \setminus \{0\}$, then the system is globally asymptotically stable.

Proof. a) Let $\epsilon > 0$ be given, which we may assume to be so that $\{x : \|x\| = \epsilon\} \subset \Omega$ (this is where we use the condition that Ω is an open set containing the equilibrium point). We will show the existence of $\delta > 0$ such that for all $\|x(0)\| \leq \delta$, $\|x(t)\| < \epsilon$ for all $t \geq 0$.

Define $m = \min_{\{x:\|x\|=\epsilon\}} V(x)$ (such an m exists by the continuity of V). Let $0 < \delta \leq \epsilon$ so that for all $\|x\| \leq \delta$ it follows that $V(x) < m$. Such a δ exists by continuity of V. Now, for any $\|x(0)\| \leq \delta$, $V(x(0)) < m$ and furthermore by the condition $\frac{d}{dt}V(x(t)) \leq 0$, $V(x(t)) < m$ as long as $x(t) \in \Omega$ (and thus also as long as $\|x(t)\| \leq \epsilon$).

This then implies that $\|x(t)\| < \epsilon$ since otherwise, there would have to be some time t_1 so that $\|x(t_1)\| = \epsilon$ and $\|x(t)\| < \epsilon$ for $t \leq t_1$, which would result in $V(x(t_1)) \geq m$ while $x(t) \in \Omega$ until t_1 so that

$$V(x(t_1)) = V(x(0)) + \int_{s=0}^{t_1} \frac{d}{ds}V(x(s))ds \leq V(x(0)) < m$$

leading to a contradiction.

b) $\epsilon = r > 0$, $\delta > 0$ satisfy the condition in part a so that with $\|x(0)\| \leq \delta$, $\|x(t)\| \leq r$ for all t so that $x(t) \in \Omega$. Since we have that $\frac{d}{dt}V(x(t)) < 0$, $V(x(t))$ is a monotonically decreasing family of non-negative numbers and it therefore has a limit. Call this limit c. We will show that $c = 0$. Suppose $c \neq 0$. Let $\alpha > 0$ be such that for all $\|x\| < \alpha$, $V(x) < c$. It then follows that for all t, $x(t) \in \{x : \alpha \leq \|x\| \leq r\}$. Define $a = \max_{\{x:\alpha\leq\|x\|\leq r\}} \left(\nabla_x V(x)\right) f(x)$. The number a exists since the functions considered are continuous, maximized over a compact set. This a is attained, it must be that $a < 0$. This implies then that $V(x(t)) = V(x(0)) + \int_{s=0}^{t} \frac{d}{ds}V(x(s))ds \leq V(x(0)) + at$, which implies that $V(x(t))$ will be negative after a finite t. This can't be true, therefore c cannot be non-zero.

c) The proof in b) applies with Ω being the entire state space: for any given $x(0) = z$, the set $\{x : V(x) \leq z\}$ will be compact and the proof will follow identically. □

Theorem 11.3.2 *[Region of asymptotic stability] Let us, in addition to the conditions noted in Theorem 11.3.1(b), further impose that for some l, $\Omega_l := \{x : V(x) \leq l\}$ is a bounded set and $\Omega_l \subset \Omega$ where Ω satisfies Theorem 11.3.1(b). Then, every solution of the system with initial state $x(0) \in \Omega_l$ converges to equilibrium.*

By following (and slightly modifying) the proof of Theorem 11.3.1(b), we can conclude that Ω_l is a region of attraction for the equilibrium point, which is defined as a set of initial states whose corresponding solutions converge to the equilibrium point: $\{x(0) : \lim_{t\to\infty} x(t) = 0\}$.

Remark 11.1. For local stability, by restricting the analysis to Ω, we can allow the Lyapunov function V to take even negative values outside Ω or not necessarily be continuous outside Ω. In Theorem 11.3.1, we used such properties of V only on Ω.

Example 11.2. Show that $x' = -x^3$ is locally asymptotically stable, by picking $V(x) = x^2$ as a Lyapunov function. Is this solution globally asymptotically stable as well?

Example 11.3. Show that $x' = -2x + x^3$ is locally asymptotically stable, by picking $V(x) = x^2$ as a Lyapunov function. Is this solution globally asymptotically stable? Find a region of attraction for local stability.

Remark 11.4. One should note that BIBO stability and the stability notions considered in this chapter have very different contexts; BIBO stability is concerned with the input-output behaviour of systems and the criteria considered in this chapter are with regard to the effects of initial conditions (also called *internal stability*). The conditions are also slightly different for the linear setup: for continuous-time linear systems, BIBO stability requires all the eigenvalues to have strictly negative real parts. Stability itself, however, may require more relaxed conditions.

11.3.1 Revisiting the linear case

Recall that an $n \times n$ real matrix P is positive definite if P is symmetric and $x^T P x > 0$ for all $x \neq 0$. It is positive semi-definite of $x^T P x \geq 0$ for all $x \in \mathbb{R}^n$. Note that being symmetric is part of the definition.

We state the following important theorem.

Theorem 11.3.3 *All eigenvalues of a square matrix A have negative real parts if and only if for any given positive definite N, the (Lyapunov) equation*

$$A^T M + MA = -N$$

has a unique solution M, where the solution is positive definite.

Proof. a) Let A have eigenvalues with negative real parts: we will show that $M = \int_0^\infty e^{A^T t} N e^{At} dt$ is a solution (whose uniqueness will also be established shortly). Since A is stable, we have

$$A^T M + MA = \int_0^\infty A^T e^{A^T t} N e^{At} dt + \int_0^\infty e^{A^T t} N e^{At} A dt = \int_0^\infty \frac{d}{dt}(e^{A^T t} N e^{At}) dt = -N,$$

where the last line follows from the fact that $\lim_{t\to\infty} e^{At} = 0$.

M is symmetric, which can be shown by verifying that $M^T = M$. Furthermore, M is positive definite: for any $x \neq 0$, we have that $x^T M x = \int_0^\infty x^T e^{A^T t} N e^{At} x dt = \int_0^\infty \|\sqrt{N} e^{At} x\|^2 dt > 0$, where $\sqrt{N}$ is the invertible matrix that solves the equation $\sqrt{N}\sqrt{N} = N$: with $N = U \Lambda U^T$ where Λ is the diagonal matrix containing the (all positive real) eigenvalues of N on the diagonal, we have that $\sqrt{N} = U\sqrt{\Lambda}U^T$, where $\sqrt{\Lambda}$ is a diagonal matrix consisting of square roots of the eigenvalues of N on the diagonal.

For uniqueness, consider the following: let $\bar{M}$ also solve $A^T \bar{M} + \bar{M} A = -N$. Then, since A has all eigenvalues with negative real parts, it follows from the fundamental theorem of calculus that

$$\bar{M} = -\int_0^\infty \frac{d}{dt}(e^{A^T t} \bar{M} e^{At}) dt$$

but

$$\begin{aligned}
\bar{M} &= -\int_0^\infty \frac{d}{dt}(e^{A^T t} \bar{M} e^{At}) dt = -\int_0^\infty (e^{A^T t} A^T \bar{M} e^{At} + e^{A^T t} \bar{M} A e^{At}) dt \\
&= -\int_0^\infty e^{A^T t} \Big(A^T \bar{M} + \bar{M} A \Big) e^{At} dt = -\int_0^\infty e^{A^T t} (-N) e^{At} dt = \int_0^\infty e^{A^T t} N e^{At} dt \\
&= M
\end{aligned}$$

b) For the reverse direction, consider the Lyapunov function $V(x) = x^T M x$. Then,

$$\frac{d}{dt} V(x(t)) = \frac{d}{dt}(x^T(t) M x(t)) = x^T M A x + x^T A^T M x = x^T (MA + A^T M) x = -x^T N x$$

which is strictly negative for all $x \neq 0$. Then, Theorem 11.3.1(b) would imply that the system $x' = Ax$ is globally asymptotically stable and hence, by Theorem 11.2.1, the eigenvalues of A have to have negative real parts. □

11.4 Non-Linear Systems and Linearization

Theorem 11.4.1 *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable. Consider $x' = f(x)$ and let $f(x^*) = 0$ for some $x^* \in \mathbb{R}^n$. Let $A := Df(x^*)$ be the Jacobian of f at x^* (that is, with $f(x) = \begin{bmatrix} f^1(x) \cdots f^n(x) \end{bmatrix}^T$ the linearization with $A(i,j) = \frac{\partial f^i}{\partial x^j}(x^*)$). Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. If $Re\{\lambda_i\} < 0$ for $i = 1, \cdots, n$, then x^* is locally asymptotically stable.*

Direct Proof. Without any loss, we can assume that $x^* = 0$ (since the linearization is with respect to the perturbation regardless of the fixed (equilibrium) point). Since A has all its eigenvalues strictly in the left half plane, there exists $\epsilon > 0$ so that $A + \epsilon I$ also has all the eigenvalues in the left half plane. You can show this via the continuity result in Theorem 8.3.1 (see, in particular, Remark 8.1).

Therefore, $e^{(A+\epsilon I)t}$ will remain bounded: $\|e^{tA}e^{\epsilon t}\| \leq K$ for some constant K for all $t \geq 0$. Here, for a matrix, we use the norm $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|}$. In the following, we drop the subscript $_2$.

Now write

$$x'(t) = Ax(t) + E(x(t)), \tag{11.1}$$

with $E(x(t)) := f(x(t)) - A(x(t))$. Let $r > 0$ be such that $\|z\| \leq r$ implies that $\|E(z)\| \leq \beta\|z\|$, where we will take β sufficiently small (to be determined at the end of the proof). Let $B := \{z : \|z\| \leq \frac{r}{2}\}$. The idea is to show that if $x(0) \in B$, then $x(t) \in B$ for all $t \geq 0$ and $\lim_{t\to\infty} x(t) = 0$. Recall that the solution to (11.1) is given with

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-s)}E(x(s))ds$$

and for $x(0) \in B$, we have, as long as $x(\tau) \in B$ for all $\tau \leq t$, that

$$\|x(t)\| \leq K\left(e^{-t\epsilon}\|x(0)\| + \beta\int_0^t e^{-\epsilon(t-s)}\|x(s)\|ds\right)$$

Write $\alpha = K\beta$ and write the above as

$$\|x(t)\| \leq e^{-t\epsilon}\|x(0)\| + \alpha'\int_0^t e^{-\epsilon(t-s)}\|x(s)\|ds$$

Define $g(t) = e^{t\epsilon}\|x(t)\|$. Then, the above writes as

$$g(t) \leq K\|x(0)\| + \alpha\int_0^t g(s)ds \tag{11.2}$$

This then implies the following

$$g(t) \leq e^{\alpha t}K\|x(0)\|,$$

(by an argument known as Grönwall's inequality). *Let us now prove this intermediate step: Write:*

$$v(s) = e^{-\alpha s}\int_0^s \alpha g(r)dr$$

and

$$v'(s) = \left(-\alpha e^{-\alpha s}\int_0^s \alpha g(r)dr + e^{-\alpha s}\alpha g(s)\right) = e^{-\alpha s}\alpha\left(g(s) - \alpha\int_0^s g(r)dr\right) \leq K\|x(0)\|\alpha e^{-\alpha s}$$

This implies that, since $v(0) = 0$

$$v(t) \leq \int_0^t K\|x(0)\|\alpha e^{-\alpha s} ds = K\|x(0)\|(1 - e^{-\alpha t})$$

and thus $e^{-\alpha t} \int_0^t \alpha g(r) dr \leq K\|x(0)\|(1 - e^{-\alpha t})$ *leading to*

$$\int_0^t \alpha g(r) dr \leq K\|x(0)\|e^{\alpha t} - K\|x(0)\|$$

and using (11.2), we arrive at

$$g(t) \leq K\|x(0)\|e^{\alpha t}$$

Then,

$$\|x(t)\| \leq e^{\alpha t - \epsilon t} K\|x(0)\|$$

If we now take $K\beta = \alpha < \epsilon$, then $x(t) \to 0$ and the proof is complete. □

An alternative, and a more direct, argument can be made through a Lyapunov analysis. As noted earlier, this method is sometimes called Lyapunov's first method.

Proof of Theorem 11.4.1 via Lyapunov's Method. Take $x^* = 0$, we have that $A := Df(0)$ is stable, where $Df(x)$ is the linearization of f at x. Then, by part b) of the proof of Theorem 11.3.3, there exists $M > 0$ so that the Lyapunov function $V(x) := x^T M x$ satisfies $\frac{dV(x(t))}{dt} = -x^T(t) N x(t)$, for the linearized model $x' = Ax$, where N is positive-definite.

By continuous-differentiability, we have that the non-linear system satisfies $Ax + E(x) = Ax + g(x)x$ with $||g(x)|| \to 0$ as $x \to 0$. To see this, apply the definition of the derivative and use continuity of this derivative.

We can then show that the Lyapunov function is also decreasing for the non-linear model as long as x is close enough to the equilibrium. That is,

$$\begin{aligned} &\frac{dV(x(t))}{dt} = \frac{d(x^T(t) M x(t))}{dt} \\ &= x^T(t)(MA + A^T M)x(t) + x^T(t) M g(x(t)) x(t) + x^T(t)(g(x(t)))^T M x(t) \\ &= -x^T(t) N x(t) + x^T(t)(M g(x(t)) + g(x(t))^T M) x(t) \end{aligned} \tag{11.3}$$

It follows that $-x^T(t) N x(t) \leq -\lambda_{\min}\|x\|^2$, where $\lambda_{\min}$ is the minimum eigenvalue of N. To show this, observe that $N - \lambda_{\min} I$ is positive semi-definite. Now, since $Mg(x(t)) + (g(x(t)))^T M \to 0$ as $x(t) \to 0$, we have that

$$\frac{dV(x(t))}{dt} \leq -\lambda_{\min}\|x(t)\|^2 + x^T(t)(M g(x(t)) + (g(x(t)))^T M) x(t) < 0,$$

for $x(0)$ small enough (note that all eigenvalues of the symmetric matrix $(Mg(x(t)) + (g(x(t)))^T M)$ will converge to zero, by the continuity of the eigenvalues in the matrix entries). This establishes local asymptotic stability. □

The above shows that linearization can be a very effective method. However, when the linearization leads to a matrix with an eigenvalue having a zero real part, the analysis (above based on linearization) is inconclusive and further analysis would be required. To make this observation explicit, consider two systems

$$x' = -x^5$$

$$x' = x^5$$

which have the same linearization around 0. By a Lyapunov stability argument with taking $V(x) = x^2$, the first system can be shown to be locally and globally stable, whereas the second one is not (which can be verified by solving the equation directly: show that $x(t)$ blows up in finite time!).

11.5 Discrete-time Setup

The stability results presented for continuous-time linear systems have essentially identical generalizations for the discrete-time setup. In this case, we require the eigenvalues to be strictly inside the unit disk for asymptotic stability (local and global); and for local stability we additionally have the relaxation that the Jordan form corresponding to an eigenvalue on the unit circle is to be strictly diagonal: Any Jordan form block J of size $N \times N$, with eigenvalue λ_i, can be written as

$$\lambda_i I + E$$

where E is a matrix which has all terms zero, except the super-diagonal (the points right above the diagonal), at which points the value is 1. The second term E is such that $E^N = 0$. Finally, we use the power expansion and using the fact that any matrix commutes with the identity matrix:

$$(\lambda_i I + E)^n = \sum_{k=0}^{n} \binom{n}{k} \lambda_i^n I E^{n-k}.$$

Since $E^N = 0$, we have

$$(\lambda_i I + E)^n = \sum_{k=0}^{N-1} \binom{n}{k} \lambda_i^{n-k} I E^k$$

One can have discrete-time generalizations of Lyapunov functions.

Theorem 11.5.1 *Consider*

$$x_{k+1} = Ax_k.$$

All eigenvalues of A have magnitudes strictly less than 1 if and only if for any given positive definite matrix N or for $N = P^T P$ where P is any given $m \times n$ matrix with $m < n$, the discrete Lyapunov equation

$$M - A^T M A = N$$

has a unique solution which is also positive definite.

The solution in the theorem statement is $M = \sum_{k \in \mathbb{Z}_+} (A^T)^k N A^k$.

11.6 Exercises

Exercise 11.6.1 *Let A be a square matrix. Show that $\frac{d}{dt} e^{At} = Ae^{At}$.*

Solution. *Let t, h be scalars. Since At and Ah commute so that $(At)(Ah) = (Ah)(At)$ we have that $e^{A(t+h)} = e^{At} e^{Ah}$. Then,*

$$\begin{aligned}
\frac{d}{dt} e^{At} &= \lim_{h \to 0} \frac{e^{A(t+h)} - e^{At}}{h} \\
&= e^{At} \lim_{h \to 0} \frac{e^{Ah} - I}{h} \\
&= e^{At} \lim_{h \to 0} \left(\sum_{k=1}^{\infty} \frac{A^k h^k}{h(k!)} \right) \\
&= e^{At} \lim_{h \to 0} \left(A + h \sum_{k=2}^{\infty} \frac{A^k h^{k-2}}{k!} \right) \\
&= Ae^{At}
\end{aligned} \tag{11.4}$$

In the analysis above, the final line follows from the fact that the sum converges to zero as $h \to 0$: let $\tilde{A}(i,j) = |A(i,j)|$ be the matrix consisting of the absolute value of the entries of A, then for each entry (i,j) of the matrix

$$\left|\left(h\sum_{k=2}^{\infty}\frac{A^k h^{k-2}}{k!}\right)(i,j)\right| \leq |h|\left(\sum_{k=2}^{\infty}\frac{\tilde{A}^k|h|^{k-2}}{k!}\right)(i,j) \leq |h|\left(\tilde{A}^2\sum_{k=2}^{\infty}\frac{\tilde{A}^{k-2}|h|^{k-2}}{(k-2)!}\right)(i,j)$$

We have that for an arbitrary $\epsilon > 0$, the following analysis applies:

$$\begin{aligned}
&\lim_{h\to 0}|h|\left(\tilde{A}^2\sum_{k=2}^{\infty}\frac{\tilde{A}^{k-2}|h|^{k-2}}{(k-2)!}\right)(i,j)\\
&\leq \lim_{h\to 0}|h|\left(\tilde{A}^2\sum_{k=2}^{\infty}\frac{\tilde{A}^{k-2}|\epsilon|^{k-2}}{(k-2)!}\right)(i,j)\\
&= \lim_{h\to 0}|h|\left(\tilde{A}^2\sum_{k=0}^{\infty}\frac{\tilde{A}^{k}|\epsilon|^{k}}{k!}\right)(i,j)\\
&= \lim_{h\to 0}|h|\tilde{A}^2 e^{\tilde{A}\epsilon}(i,j)\\
&= 0
\end{aligned} \tag{11.5}$$

The result follows.

Exercise 11.6.2 *Show that for square matrices A and B, which commute, that is*

$$AB = BA,$$

it follows that

$$e^{(A+B)} = e^A e^B.$$

Solution. *Recall that*

$$e^A = \lim_{T\to\infty}\sum_{k=0}^{T}\frac{A^k}{k!} = \sum_{k=0}^{\infty}\frac{A^k}{k!},$$

with the definition that $A^0 = I$. It follows that

$$\begin{aligned}
e^A e^B &= (\lim_{T\to\infty}\sum_{k=0}^{T}\frac{A^k}{k!})(\lim_{T\to\infty}\sum_{l=0}^{T}\frac{B^l}{l!})\\
&= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\frac{A^k}{k!}\frac{B^l}{l!} = \sum_{k=0}^{\infty}\sum_{u=k}^{\infty}\frac{1}{k!(u-k)!}A^k B^{u-k} &(11.6)\\
&= \sum_{u=0}^{\infty}\sum_{k=0}^{u}\frac{1}{k!(u-k)!}A^k B^{u-k} &(11.7)\\
&= \sum_{u=0}^{\infty}\sum_{k=0}^{u}\frac{1}{u!}\frac{u!}{k!(u-k)!}A^k B^{u-k} = \sum_{u=0}^{\infty}\frac{1}{u!}\sum_{k=0}^{u}\frac{u!}{k!(u-k)!}A^k B^{u-k}\\
&= \sum_{u=0}^{\infty}\frac{1}{u!}\sum_{k=0}^{u}\binom{u}{k}A^k B^{u-k} = \sum_{u=0}^{\infty}\frac{1}{u!}(A+B)^u &(11.8)\\
&= e^{(A+B)} &(11.9)
\end{aligned}$$

In the above, (11.6) follows by defining $u = k + l$, and (11.7) follows from re-expressing the summation. Finally, (11.8) follows from $AB = BA$ and the following:

$$(A+B)^k = \sum_{m=0}^{k} \binom{k}{m} A^m B^{k-m}.$$

We now prove this last statement. Clearly for $k = 1$,

$$(A+B)^1 = \binom{1}{0} B + \binom{1}{1} A$$

Let us proceed by induction. Suppose this is true for k. Let us show that it is true for $k+1$.

$$(A+B)^k (A+B) = \sum_{m=0}^{k} \binom{k}{m} A^{m+1} B^{k-m} + \sum_{m=0}^{k} \binom{k}{m} A^m B^{k+1-m}.$$

Let us separate out the terms involving $A^p B^{k+1-p}$ for $0 \leq p \leq k+1$. It turns out that we obtain:

$$\sum_{p=0}^{k+1} A^p B^{k+1-p} \left(\binom{k}{p} + \binom{k}{p-1} \right)$$

Now,

$$\begin{aligned} &\binom{k}{p} + \binom{k}{p-1} = \frac{k!}{p!(k-p)!} + \frac{k!}{(p-1)!(k+1-p)!} \\ &= \frac{k!}{(p-1)!(k-p)!} \left(\frac{1}{p} + \frac{1}{(k+1-p)} \right) = \frac{k!}{(p-1)!(k-p)!} \left(\frac{k+1}{p(k+1-p)} \right) \\ &= \frac{(k+1)!}{(p)!(k+1-p)!} = \binom{k+1}{p} \end{aligned} \tag{11.10}$$

This completes the proof.

Exercise 11.6.3 *Let $x(t)$ satisfy*

$$\frac{dx}{dt} = -x^7$$

Is $x(t)$ (locally) asymptotically stable?

Exercise 11.6.4 *Consider $x'' + x' + x = 0$. Is this system asymptotically stable?*

Hint: Convert this equation into a system of first-order differential equations, via $x_1 = x$ and $x_2 = x_1'$, $x_2' = -x_1 - x_2$. Then apply $V(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^2$ as a candidate Lyapunov function.

Exercise 11.6.5 (A Further Lyapunov Stability Theorem and Barbalat's Lemma) *Prove the following theorem:*

Theorem 11.6.1 *Consider*

$$x' = f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and $f(0) = 0$. Let $V : \mathbb{R}^n \to \mathbb{R}_+$ be continuously differentiable. Prove the following: Suppose that there exists a continuous function $W : \mathbb{R}^n \to \mathbb{R}_+$ such that

$$\frac{d}{dt} V(x(t)) \leq -W(x(t)) \leq 0$$

Then, provided $x(t)$ remains bounded, $W(x(t)) \to 0$.

Hint: Write

$$V(x(t)) = V(x_0) + \int_0^t \frac{dV(x(s))}{ds} ds \leq V(x_0) - \int_0^t W(x(s)) ds$$

and conclude that $\int_0^t W(x(s))ds \leq V(x_0)$ for all $t \geq 0$ and by the non-negativity of W, we have that $\int_0^\infty W(x(s))ds \leq V(x_0)$. From here, we want to establish that $W(x(t)) \to 0$, provided that (by hypothesis) $x(t)$ remains bounded. Complete the proof.

Prove and use Barbalat's lemma: Let $f : K \to \mathbb{R}_+$ be uniformly continuous over K. Then, if $x(t)$ remains in K and if $\int_0^\infty f(x(s))ds$ is finite, then $f(x(t)) \to 0$.

In the following two exercises, we will see two applications.

Note: The above result also implies an important stability theorem known as Lasalle's invariance principle.

Exercise 11.6.6 (Application to formation control, consensus algorithms or opinion dynamics) *Consider a network of N agents which are connected over a graph. We say that A is an adjacency graph if $A(i,j) = 1$ if Agents i and Agent j are connected and $A(i,j) = 0$ otherwise. For each agent $i \in \{1, \cdots, N\}$ define $d_i = \sum_{j=1}^N A(i,j)$ to be the degree of the agent. Now define $L = A - D$ where D is a diagonal matrix with $D(i,i) = d_i$. Such a matrix L is called a Laplacian.*

Now, suppose that the agents update their states by the following equation:

$$\frac{dx}{dt} = -Lx$$

Observe that L is a positive semi-definite matrix and if the graph is connected the only eigenvector corresponding to the zero eigenvalue is $\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^T$. This you can see by noting that $x^T Lx = \sum A(i,j)(x_i - x_j)^2$.

In this case, define the following Lyapunov function:

$$V(x) = x^T x$$

Then,

$$\frac{d}{dt}(V(x(t))) = (\frac{dx}{dt})^T x(t) + x^T(t)\frac{dx}{dt} = -x^T(t)Lx(t) \leq 0$$

The above ensures that $x(t)$ remains bounded. Now, invoke Barbalat's Lemma to conclude that $x^T(t)Lx(t) \to 0$. Since the only eigenvalue corresponding to $Lx(t) = 0$ is $x(t) = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^T$ and throughout the updates the sum $\frac{1}{N}(x_1(t) + x_2(t) + \cdots x_N(t))$ is a constant (as the sum does not change), we have that $x(t) \to \frac{1}{N}(x_1(0) + x_2(0) + \cdots x_N(0))$

Exercise 11.6.7 (Application to adaptive control) *Consider*

$$\frac{dx}{dt} = ax + u$$

Suppose that our goal is to have $\lim_{t\to\infty} x(t) = 0$. We know that if we select $u(t) = -(a + \kappa)x$ for any $\kappa > 0$, the system is stable. In particular, let $\kappa = 1$.

In many engineering applications, the value of a is unknown.

Adaptive control theory is the sub-field of control theory studying such problems. The goal is to allow the controller to learn the system to be able to achieve the desired goal.

Suppose that the controller runs the following policy:

$$u(t) = -(\hat{a}(t) + 1)x(t),$$

which leads to $x' = (a - \hat{a}(t) - 1)x(t)$, where $\hat{a}(t)$ is an estimate of a. Suppose that we take

$$\hat{a}'(t) = x^2(t)$$

In this case, consider the Lyapunov function:

$$V(x,\hat{a}) = \frac{x^2}{2} + \frac{(\hat{a}-a)^2}{2}$$

Then,

$$\frac{d}{dt}V(x(t),\hat{a}(t)) = x(t)x'(t) + (\hat{a}(t)-a)\hat{a}'(t) = x^2(t)(a-\hat{a}(t)-1) + (\hat{a}(t)-a)x^2(t) = -x^2(t).$$

Now, strictly speaking this derivative analysis does not satisfy the conditions given in Theorem 11.3.1. However, it does satisfy Theorem 11.6.1, with $W(x) = x^2$ *and* $V(x,\hat{a})$ *as given. For this, we can conclude that* $x(t) \to 0$ *(note that we have that* $x(t)$ *is bounded by the condition that* $V(x,\hat{a})$ *is positive and its derivative is non-increasing).*

Note that there is no claim that $\hat{a}(t) \to a$*. This was not part of the design goal, only that* $x(t) \to 0$*. However, this is achieved by adaptive control.*

12

Controllability and Observability

12.1 Controllability

Consider

$$\frac{dx}{dt} = Ax(t) + Bu(t),$$

where A is $n \times n$ and B is $n \times p$. However, for simplicity, in the derivations below throughout the chapter, we will assume that $p = 1$.

Definition 12.1.1 *The pair (A,B) is said to be controllable if for any $x(0) = x_0 \in \mathbb{R}^n$ and $x_f \in \mathbb{R}^n$, there exists $T < \infty$ and a control input $\{u_s, 0 \leq s \leq T\}$ so that $x_T = x_f$.*

Consider the following:

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 1 \\ -4 & -5 \end{bmatrix} x + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} u$$

In this case, if b_1, b_2 are both 0, it is evident that the system is not controllable: for every given $x(0)$, the future paths are uniquely determined.

Consider, now the more interesting case with $b_1 = 1 = -b_2$. In this case, if the initial condition $x(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$ takes values from the subspace determined by the line $x_1(0) + x_2(0) = 0$, then, for all $t > 0$, the state remains in this subspace. To see this, note that $\frac{d(x_1(t)+x_2(t))}{dt} = \frac{dx_1(t)+dx_2(t)}{dt} = 0$ so that the sum of the state components does not change and thus $x_1(t) + x_2(t)$ remains 0. Thus, this subspace, which is a strict subset of $\mathbb{R}^2$, is invariant no matter what control is applied: this system is not controllable.

Theorem 12.1.1 *Conditions (i), (ii), (iii), and (iv) below are equivalent:*

(i) (A, B) *is controllable.*

(ii) The $n \times n$ matrix

$$W_c(t) = \int_0^t e^{As} BB^T e^{A^T s} ds = \int_0^t e^{A(t-s)} BB^T e^{A^T (t-s)} ds$$

is full-rank for every $t > 0$.

(iii)The controllability matrix

$$\mathcal{C} := \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$$

is full-rank.

(iv) The matrix

$$\begin{bmatrix} A - \lambda I & B \end{bmatrix}$$

has full rank (i.e., rank n) at every eigenvalue λ of A.

The matrix W_c above is called the controllability Grammian of (A, B).

Proof. (i) $\leftrightarrow$ (ii). Let $W_c(t)$ be invertible for $t > 0$ and let x_1 be the target state to be arrived at in some finite time. In fact, we will show that we can get to x_1 at any $t_1 > 0$: Write

$$x(t_1) = e^{At_1}x(0) + \int_0^{t_1} e^{A(t_1-s)}Bu(s)ds$$

Let

$$u(s) = -B^T e^{A^T(t_1-s)}W_c^{-1}(t_1)\left(e^{At_1}x(0) - x_1\right)$$

Then,

$$\begin{aligned} x(t_1) &= e^{At_1}x(0) - \int_0^{t_1} e^{A(t_1-s)}BB^T e^{A^T(t_1-s)}W_c^{-1}(t_1)\left(e^{At_1}x(0) - x_1\right)ds \\ &= e^{At_1}x(0) - \int_0^{t_1} e^{A(t_1-s)}BB^T e^{A^T(t_1-s)}W_c^{-1}(t_1)ds\left(e^{At_1}x(0) - x_1\right) \\ &= e^{At_1}x(0) - W_c(t_1)W_c^{-1}(t_1)\left(e^{At_1}x(0) - x_1\right) = x_1 \end{aligned} \tag{12.1}$$

Thus, invertibility of $W_c(t)$ implies controllability.

We will now show that controllability implies that $W_c(t)$ is invertible for every $t > 0$. Now, suppose that (A, B) is controllable but $W_c(t)$ is not invertible. Then, there exists a vector $v \neq 0$ so that

$$v^T W_c(t)v = 0$$

or

$$\int_0^t v^T e^{A(t-s)}BB^T e^{A^T(t-s)}vds = 0$$

or

$$\int_0^t \|B^T e^{A^T(t-s)}v\|_2^2 ds = 0$$

implying that for $s \in [0, t]$

$$B^T e^{A^T(t-s)}v = 0, \qquad s \in [0, t] \tag{12.2}$$

or $B^T e^{A^T s}v = 0, \qquad s \in [0, t]$. Now, if $W_c(t)$ is non-singular for some $t > 0$, then, since $e^{At}B$ is an analytic function, by (12.2), the expression $B^T e^{A^T t}v = 0$ for all $t \in \mathbb{R}$.

But if (A, B) is controllable, there exists a control input u that steers the system from $x(0) = 0$ to $x(t_1) = v$ for some t_1, and as a result

$$v = e^{At_1}0 + \int_0^{t_1} e^{A(t_1-s)}Bu(s)ds = 0$$

and multiplying by v^T,

$$v^T v = \int_0^{t_1} v^T e^{A(t_1-s)}Bu(s)ds = 0$$

But, by (12.2) above

$$\int_0^{t_1} v^T e^{As} B u(s) ds = 0$$

Thus,

$$v^T v = 0$$

which is a contradiction since $v \neq 0$.

(ii) $\leftrightarrow$ (iii). Suppose that $\mathcal{C}$ does not have full-rank. Then, there exists v such that

$$v^T A^k B = 0, \qquad 0 \leq k \leq n-1$$

Since $e^{At}B$ can be written as a linear combination of $\{A^k B, 0 \leq k \leq n-1\}$ (by the Cayley-Hamilton theorem), it follows that $v^T e^{At} B = 0$ for all $t > 0$. This implies then that $W_c(t)$ is not invertible.

Now, suppose that $\mathcal{C}$ is full-rank but $W_c(t)$ is not invertible. Then, there exists a non-zero v so that $v^T e^{At} B = 0$ for all $t \geq 0$ (by the argument earlier leading to (12.2). With $t = 0$, we have $v^T B = 0$, taking the derivative at $t = 0$ leads to $v^T AB = 0$, and taking up to $n-1$st derivative at $t = 0$, we have $v^T A^k B = 0, \quad 0 \leq k \leq n-1$. Thus, $\mathcal{C}$ cannot be full-rank.

An implication of this relationship will be presented further below.

(iii) $\leftrightarrow$ (iv). If $\mathcal{C}$ has full-rank, then the matrix

$$\begin{bmatrix} A - \lambda I & B \end{bmatrix}$$

has full rank. Suppose not: then there exists $v \neq 0$ so that

$$v^T \begin{bmatrix} A - \lambda I & B \end{bmatrix} = 0$$

and that $v^T A = \lambda v^T$ and $v^T B = 0$. Thus, v is a left eigenvector and λ is an eigenvalue of A. Then, $v^T A^2 = \lambda^2 v^T$, and for all $1 \leq k \leq n-1$: $v^T A^k = \lambda^k v^T$. As a result,

$$v^T \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = \begin{bmatrix} v^T B & \lambda v^T B & \cdots & \lambda^{n-1} v^T B \end{bmatrix} = 0$$

and $\mathcal{C}$ does not have full-rank.

Conversely, suppose that $\mathcal{C}$ does not have full-rank. For this statement, we present a direct argument, but borrowing a result to be presented later in the chapter. We will see in Section 12.5 that if the system is not controllable, then there exists a transformation so that

$$A_c = PAP^{-1}, \qquad B_c = PB,$$

is such that for some vector $x = \begin{bmatrix} 0^T & w^T \end{bmatrix}$

$$xA_c = 0, \quad xB_c = 0$$

or

$$x[PAP^{-1}, PB] = 0$$

meaning that

$$xP[AP^{-1}, B] = 0$$

so that xP is orthogonal to the range of both A and B, which leads to the result. $\square$

Reachable Set (from origin) and the Controllable Subspace An implication of the proof of the equivalence between (ii) and (iii) is that the range space of $\mathcal{C}$ and the range space of the linear operator from the space of integrable control inputs $\mathcal{U} = \{u : \mathbb{R}_+ \to \mathbb{R}, \|u\|_1 < \infty\}$ to $\mathbb{R}^n$ defined with

$$\bigcup_{t \geq 0} \{ \int_0^t e^{A(t-s)} B u(s) ds, \quad u \in \mathcal{U} \}$$

are equal. This set is called the *reachable* (from the origin) set. This set is also called the controllable subspace.

Exercise 12.1.1 *The controllability property is invariant under an algebraically equivalent transformation of the coordinates: $\tilde{x} = Px$ for some invertible P.*

Hint: Use the rank condition and show that with $\frac{d\tilde{x}}{dt} = \tilde{A}\tilde{x}(t) + \tilde{B}u(t)$ and $\tilde{A} = PAP^{-1}, \tilde{B} = PB$, and with $\mathcal{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$, we have that the transformed controllability matrix writes as $\tilde{\mathcal{C}} = \begin{bmatrix} \tilde{B} & \tilde{A}\tilde{B} & \cdots & \tilde{A}^{n-1}\tilde{B} \end{bmatrix} = P\mathcal{C}$.

12.2 Observability

In many problems a controller has access to only the inputs applied and outputs measured. A very important question is whether the controller can recover the state of the system through this information.

Consider

$$\frac{dx}{dt} = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + Du(t)$$

Definition 12.2.1 *The pair (A,C) is said to be observable if for any $x(0) = x_0 \in \mathbb{R}^n$, there exists $T < \infty$ such that the knowledge of $\{(y_s, u_s), \quad 0 \leq s \leq T\}$ is sufficient to uniquely determine $x(0)$.*

In the above, we could consider without any loss that $u(t) = 0$ for all t, since the control terms appear in additive forms whose effects can be cancelled from the measurements.

Consider then

$$\frac{dx}{dt} = Ax(t), \qquad y(t) = Cx(t)$$

The measurement at time t writes as:

$$y(t) = Ce^{At}x(0)$$

taking the derivative

$$\frac{dy(t)}{dt} = CAe^{At}x(0)$$

and taking the derivatives up to order $n-1$, we obtain for $1 \leq k \leq n-1$

$$\frac{d^{k-1}y(t)}{dt^{k-1}} = CA^{k-1}e^{At}x(0)$$

In matrix form, we can write the above as

$$\begin{bmatrix} y(t) \\ \frac{dy(t)}{dt} \\ \vdots \\ \frac{d^{n-1}y(t)}{dt^{n-1}} \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} e^{At}x(0)$$

Thus, the question of being able to recover $x(0)$ from the measurements becomes that of whether the observability matrix $\mathcal{O} := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$ is full-rank or not. Note that adding further rows to this matrix does not increase the rank by the Cayley-Hamilton theorem. Thus, we can recover the initial state if the observability matrix is full-rank.

Furthermore, we have that Ce^{At} is a linear combination of $\{CA^k, k = 0, 1, \cdots, n-1\}$. Therefore, if x_0 is orthogonal to $\{CA^k, k = 0, 1, \cdots, n-1\}$, then it is also orthogonal to Ce^{At}. In particular, if the observability matrix is not full-rank, then there exists a non-zero x_0 so that $Ce^{At}x_0 = 0$. Thus, we cannot distinguish between x_0 and the 0 vector in $\mathbb{R}^n$ and thus the system is not observable.

Then, we have the following theorem:

Theorem 12.2.1 *The system*

$$\frac{dx}{dt} = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + Du(t)$$

is observable if and only if

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

is full-rank.

The null-space of $\mathcal{O}$, that is, $\{v \in \mathbb{R}^n : \mathcal{O}v = 0\}$ is called the unobservable subspace.

The structure of the observability matrix $\mathcal{O}$ and the controllability matrix $\mathcal{C}$ leads to the following very important and useful duality result.

Theorem 12.2.2 (A, C) *is observable if and only if* (A^T, C^T) *is controllable.*

In view of Theorem 12.1.1 (and in particular, now that we have related observability to a condition of the form given in Theorem 12.1.1(iii)), we have the following immediate result:

Theorem 12.2.3 (A, C) *is observable if and only if*

$$W_o(t) = \int_0^t e^{A^T s} C^T C e^{As} ds$$

is invertible for all $t > 0$.

12.3 Feedback and Pole Placement

Consider $u = -Kx$. Then,

$$\frac{dx}{dt} = Ax(t) + Bu(t) = (A - BK)x(t)$$

Theorem 12.3.1 *The eigenvalues of* $A - BK$ *can be placed arbitrarily if and only if* (A, B) *is controllable.*

To see this result, first consider a system in the controllable canonical realization form (see Section 9.1.1) with

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

$$A_c = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{N-1} \end{bmatrix}$$

$$B_c = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

Note that, the eigenvalues of A solve the characteristic polynomial whose coefficients are located in the bottom row of A_c (see the proof of Theorem 8.3.1).

Now, apply $u = -Kx$ so that $u = \sum_{i=1}^{N} -k_i x_i$, leading to

$$A - BK = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & 0 \\ -(a_0 + k_1) & -(a_1 + k_2) & -(a_2 + k_3) & \cdots & -(a_{N-1} + k_N) \end{bmatrix}$$

Once again, since the eigenvalues of this matrix is solves the characteristic polynomial whose coefficients are located in the bottom row (by the proof of Theorem 8.3.1), and these coefficients can be placed by selecting the scalars k_i, we can arbitrarily place the eigenvalues of the closed-loop matrix by feedback.

Through a coordinate transformation $\tilde{x} = Px$, every controllable system $x' = Ax + Bu$ can be transformed to an algebraically equivalent linear system in the controllable canonical realization form (A_c, B_c) above. As we saw, for a system in this form, a control can be found so that all the eigenvalues of the closed loop system are on the left-half plane. Finally, the system can be moved back to the original coordinates.

We now see how this (transformation into a controllable canonical realization form) is possible. With $\tilde{x} = Px$, we have that

$$\frac{d\tilde{x}}{dt} = \tilde{A}\tilde{x}(t) + \tilde{B}u(t)$$

with $\tilde{A} = PAP^{-1}, \tilde{B} = PB$. Now, if (A, B) is controllable, we know that $\mathcal{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$ is full-rank. The transformed controllability matrix writes as: $\tilde{\mathcal{C}} = \begin{bmatrix} \tilde{B} & \tilde{A}\tilde{B} & \cdots & \tilde{A}^{n-1}\tilde{B} \end{bmatrix} = P\mathcal{C}$. As a result,

$$P = \tilde{\mathcal{C}}\mathcal{C}^{-1}$$

whose validity follows from the fact that $\mathcal{C}$ is invertible. This leads us to the following conclusion.

Theorem 12.3.2 *Consider $x' = Ax + Bu$ where $u \in \mathbb{R}$. Every such system, provided that (A, B) is controllable, can be transformed into a system $z' = \tilde{A}z + \tilde{B}u$ with the transformation $z = Px$ so that $(\tilde{A}, \tilde{B})$ is in the controllable canonical realization form.*

The above then suggests a method to achieve stabilization through feedback: First transform into a controllable canonical realization form, place the eigenvalues through feedback, and transform the system back to the original coordinate.

12.4 Observers and Observer Feedback

Consider

$$\frac{dx}{dt} = Ax + Bu, \qquad y = Cx$$

Suppose that the controller intends to track the state. A candidate for such a purpose is to write an *observer* system of the form

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + L(y - C\hat{x})$$

We then obtain with $e = x - \hat{x}$, and subtracting the above two equations from one another

$$\frac{de}{dt} = Ae - LCe = (A - LC)e$$

Then, the question whether $e(t) \to 0$ is determined whether the eigenvalues of $A - LC$ can be pushed to the left-half plane with some appropriate L. If the system is observable, then this is possible, with the same arguments applicable to the pole

placement analysis presented in the previous section (note that controllability and observability are related to each other with a simple duality property that was presented in Theorem 12.2.2: that is L^T can be selected so that $A^T - C^T L^T$ has all eigenvalue in the left-half plane, which will also imply that $A - LC$ will have the same property).

Now that under observability we have that the controller can track the state with asymptotically vanishing error, suppose that we consider

$$\frac{dx}{dt} = Ax + Bu, \qquad y = Cx$$

with the goal of stabilizing the actual system state $x(t)$.

Suppose that we run an observer, and that we consider the following feedback control policy

$$u(t) = -K\hat{x}(t)$$

where K is what we used for pole placement, and $\hat{x}$ is what we used in our observer. In this case, we obtain the following relation:

$$\begin{bmatrix} \frac{dx}{dt} \\ \frac{de}{dt} \end{bmatrix} = \begin{bmatrix} A - BK & BK \\ 0 & A - LC \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix}$$

Due to the upper triangular form, we conclude that $\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} \to 0$ if both $A - BK$ and $A - LC$ are stable matrices; two conditions that we have already established under controllability and observability properties. Such a design leads to the *separation principle* for linear control systems: run an observer and apply the control as if the observer state is the actual state. This design is stabilizing.

12.5 Canonical Forms

Theorem 12.1. *(i) If $v \in \mathbb{R}^n$ is in the controllable subspace, then so is Av.*

(ii) If $v \in \mathbb{R}^n$ is in the unobservable subspace, then so is Av.

That is, the controllable and unobservable subspaces are A-invariant.

If a model is not controllable, then we can construct a state transformation $\tilde{x} = Px$ with the form $\tilde{x} = \begin{bmatrix} \tilde{x}_c \\ \tilde{x}_{\bar{c}} \end{bmatrix}$ with

$$\frac{d\tilde{x}}{dt} = \begin{bmatrix} A_c & A_{12} \\ 0 & A_{uc} \end{bmatrix} \begin{bmatrix} \tilde{x}_c \\ \tilde{x}_{uc} \end{bmatrix} + \begin{bmatrix} B_c \\ 0 \end{bmatrix} u$$

$$y = \begin{bmatrix} C_c & C_{uc} \end{bmatrix} \begin{bmatrix} \tilde{x}_c \\ \tilde{x}_{uc} \end{bmatrix}$$

In the above (A_c, B_c) is controllable. In the above, A_{12} is some submatrix. The form above is called a *controllable canonical form.*

The matrix P can be obtained with constructing P^{-1} to consist of the following: Let n_1 be the rank of the controllability matrix $\mathcal{C}$. Then take the first n_1 columns of P^{-1} to be n_1 linearly independent columns of $\mathcal{C}$, and the remaining $n - n_1$ columns are arbitrary vectors which make P^{-1} invertible. If we write

$$A_c = PAP^{-1}, \qquad B_c = PB,$$

we have that

$$P^{-1}A_c = AP^{-1}, \qquad P^{-1}B_c = B$$

Using the fact that the controllable subspace if A invariant, it follows that the structure of A_c has to have the given structure.

An implication of the above analysis is that

$$C(sI - A)^1 B + D = C_c(sI - A_c)^{-1} B_c + D$$

A similar construction applies for *observable canonical forms*.

$$\frac{d\tilde{x}}{dt} = \begin{bmatrix} A_o & 0 \\ A_{21} & A_{uo} \end{bmatrix} \begin{bmatrix} \tilde{x}_o \\ \tilde{x}_{uo} \end{bmatrix} + \begin{bmatrix} B_o \\ B_{uo} \end{bmatrix} u$$

$$y = \begin{bmatrix} C_o & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_o \\ \tilde{x}_{uo} \end{bmatrix}$$

with the property that (A_o, C_o) is observable.

An implication of the above analysis is that

$$C(sI - A)^1 B + D = C_o(sI - A_o)^{-1} B_o + D$$

One can apply a joint construction, known as Kalman's decomposition. There exists a coordinate transformation so that

$$z = Px$$

with

$$z = \begin{bmatrix} x_{co} \\ x_{c/uo} \\ x_{uc/o} \\ x_{uc/uo} \end{bmatrix}$$

$$\bar{A} = PAP^{-1}$$

leads to

$$\frac{dz}{dt} = \begin{bmatrix} A_{c/o} & 0 & A_{x/o}0 & \\ A_{c/x} & A_{x/uo} & A_{x/x} & A_{x/uo} \\ 0 & 0 & A_{uc/o} & 0 \\ 0 & 0 & A_{uc/x} & A_{uc/uo} \end{bmatrix} z + \begin{bmatrix} B_{c/o} \\ B_{c/uo} \\ 0 \\ 0 \end{bmatrix} u$$

$$y = \begin{bmatrix} C_{c/o} & 0 & C_{uc/o} & 0 \end{bmatrix} z + Du,$$

where $(A_{c/o}, B_{c/o}, C_{c/o})$ is both controllable and observable. Furthermore,

$$C(sI - A)^{-1} B + D = C_{c/o}(sI - A_{c/o})^{-1} B_{c/o} + D$$

A corollary of the above discussion is that the *minimal realization*; that is, the state-space realization with the smallest dimensions involving matrices, is attained when the system is both controllable and observable, as there are no redundant state variables.

From the controllable canonical form, we can also establish the following result.

A linear system is stabilizable (in the sense of local or global asymptotic stability) by control if and only if A_{uc}, whenever exists, is a stable matrix (i.e., with eigenvalues strictly in the left half plane).

Define a control-free system to be detectable if whenever $y(t) \to 0$ then $x(t) \to 0$. A consequence of the observable canonical form is that a system is detectable if and only if A_{uo}, whenever exists, is a stable matrix.

12.6 Using Riccati Equations to Find Stabilizing Linear Controllers [Optional]

While controllability and observability properties reveal what is possible or impossible with regard to stabilization, they don't directly present an easy-to-compute or constructive method for arriving at design.

One effective method is through Riccati equations. We will present the discussion for discrete-time, but the approach is essentially identical for continuous-time (with the stability conditions of linear systems, as noted earlier, being different).

12.6.1 Controller design via Riccati equations

Consider the following linear system

$$x_{t+1} = Ax_t + Bu_t, \tag{12.3}$$

where $x \in \mathbb{R}^n, u \in \mathbb{R}^m$.

Suppose that we would like to minimize the expression over all control laws:

$$\sum_{t=0}^{\infty} x_t^T Q x_t + u_t^T R u_t \tag{12.4}$$

with $R > 0, Q \geq 0$.

Theorem 12.6.1 *Consider (12.3).*

(i) If (A, B) is controllable there exists a solution to the Riccati equation

$$P = Q + A^T PA - A^T PB(B^T PB + R)^{-1} B^T PA. \tag{12.5}$$

(ii) if (A, B) is controllable and, with $Q = C^T C$, (A, C) is observable; as $t \to -\infty$, the sequence of Riccati recursions, for $P_0 = \bar{P}$ with $\bar{P}$ arbitrary,

$$P_t = Q + A^T P_{t+1} A - A^T P_{t+1} B(B^T P_{t+1} B + R)^{-1} B^T P_{t+1} A, \tag{12.6}$$

converges to some limit P that satisfies (12.5). That is, convergence takes place for any initial condition $\bar{P}$. Furthermore, such a P is unique, and is positive definite. Finally, under the control policy

$$u_t = -(B^T PB + R)^{-1} B^T PAx_t,$$

$\{x_t\}$ is stable.

(iii)Under the conditions of part (ii), the control minimizes (12.4),

In the above, we established a method to find K so that $A - BK$ is stable: Run the recursions, for any arbitrary initial condition, (12.6), find the limit P and select $u_t = -Kx_t$ with

$$K = (B^T PB + R)^{-1} B^T PA \tag{12.7}$$

This controller will be stabilizing.

12.6.2 Observer design via Riccati equations

A similar phenomenon as applies for observer design. In fact, with the above discussion, using the duality analysis presented earlier, we can directly design an observer so that the matrix $A - LC$ is stable. By writing the condition as the stability of $A^T - C^T L^T$, the question becomes that of finding L^T for which $A^T - C^T L^T$ is a stable matrix.

Let (A, C) be observable. In Theorem 12.6.1, if we replace A with A^T, B with C^T, and defining $W = BB^T$ for any B with (A, B) controllable, we obtain:

$$S = W + ASA^T - ASC^T(CSC^T + R)^{-1}CSA^T.$$

or the Riccati equations

$$S_{t+1} = W + AS_tA^T - AS_tC^T(CS_tC^T + R)^{-1}CS_tA^T.$$

whose limit as $t \to \infty$ for any initial S_0 will converge to a unique limit. Finally, taking

$$L^T = (CSC^T + R)^{-1}CSA^T \tag{12.8}$$

will lead to the conclusion that $A - LC$ is stable.

12.6.3 Putting controller and observer design together

Accordingly, all we need for the system:

$$x_{t+1} = Ax_t + Bu_t, \qquad y_t = Cx_t \tag{12.9}$$

is that (A, B) be controllable and (A, C) be observable. With this, via (12.7)-(12.8) we can find K and L so that the system

$$x_{k+1} = Ax_k - BK\hat{x}_k$$

$$\hat{x}_{k+1} = Ax_k + L(Cx_k - C\hat{x}_k)$$

or, equivalently, with $e_k = x_k - \hat{x}_k$, the system defined with

$$\begin{bmatrix} x_{k+1} \\ e_{k+1} \end{bmatrix} = \begin{bmatrix} A - BK & BK \\ 0 & A - LC \end{bmatrix} \begin{bmatrix} x_k \\ e_k \end{bmatrix}$$

is stable.

In the above, the conditions on (A, B) being controllable and (A, C) being observable can be relaxed: controllability can be replaced with stabilizability and observability can be relaxed to detectability. While stability will be maintained, the only difference would be that P or S would not be guaranteed to be positive-definite.

12.6.4 Continuous-time case

A similar discussion as above applies for the continuous-time setup. We only discuss the control design, as the observer design follows from duality, as shown above.

Consider

$$\frac{dx}{dt} = Ax + Bu$$

Let $Q \geq 0, R > 0$. The only difference with the continuous-time is that the discrete-time Riccati equations above are replaced by a corresponding Riccati differential equation:

$$-\frac{dP}{dt} = Q + A^T P + PA - PBR^{-1}B^T P.$$

If (A, B) is controllable and, with $Q = C^T C$, (A, C) is observable, then there exists a unique positive-definite matrix P such that the following algebraic Riccati equation is satisfied:

$$Q + A^T P + PA - PBR^{-1}B^T P = 0$$

With this P, the control given by

$$u = -Kx = -R^{-1}B^T Px$$

is so that $A - BK$ is stable.

12.7 Applications and Exercises

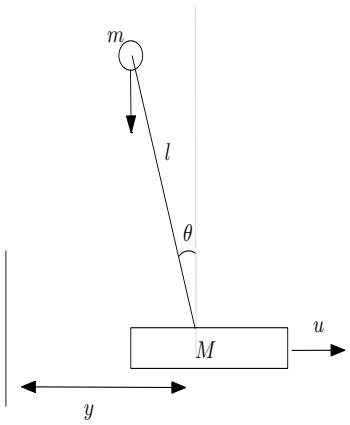

Fig. 12.1

Exercise 12.7.1 *Recall that we had studied the controlled pendulum on a cart (see Figure 12.1). The non-linear mechanical/rotational dynamics equations were found to be*

$$\begin{aligned} M\frac{d^2y}{dt^2} &= u - m\frac{d^2}{dt^2}(y + l\sin(\theta)) = u - m\frac{d^2y}{dt^2} + ml\frac{d^2\theta}{dt^2} - ml(\frac{d\theta}{dt})^2\sin(\theta) \\ m\frac{d^2\theta}{dt^2} &= \frac{mg}{l}\sin(\theta) - \frac{m}{l}\frac{d^2y}{dt^2}\cos(\theta) \end{aligned} \tag{12.10}$$

Around $\theta = 0, \frac{d\theta}{dt} = 0$*, we apply the linear approximations* $\sin(\theta) \approx \theta$ *and* $\cos(\theta) \approx 1$*, and* $(\frac{d\theta}{dt})^2 \approx 0$ *to arrive at*

$$\begin{aligned} M\frac{d^2y}{dt^2} &= u - (m\frac{d^2y}{dt^2} + ml\frac{d^2\theta}{dt^2}) \\ l\frac{d^2\theta}{dt^2} &= g\theta - \frac{d^2y}{dt^2} \end{aligned} \tag{12.11}$$

Finally, writing $x_1 = y, x_2 = \frac{dy}{dt}, x_3 = \theta, x_4 = \frac{d\theta}{dt}$*, we arrive at the linear model in state space form*

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{(M+m)g}{Ml} & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ \frac{-1}{Ml} \end{bmatrix} u,$$

where $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$.

a) When is the linearized model controllable?

b) Does there exist a control policy with $u = -Kx$ *that makes the closed loop linearized system stable? Select specific values for* M, m, l *so that controllability holds, and accordingly find an explicit* K.

c) With the controller in part b), can you conclude that through the arguments presented in the previous chapter (e.g. Theorem 11.4.1), that your (original non-linear) system is locally asymptotically stable?

Hint: a) With

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{(M+m)g}{Ml} & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ \frac{-1}{Ml} \end{bmatrix}$$

we have that

$$\begin{bmatrix} B & AB & A^2B & A^3B \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{M} & 0 & \frac{mg}{M^2l} \\ \frac{1}{M} & 0 & \frac{mg}{M^2l} & 0 \\ 0 & \frac{-1}{Ml} & 0 & \frac{-(M+m)g}{M^2l^2} \\ \frac{-1}{Ml} & 0 & \frac{-(M+m)g}{M^2l^2} & 0 \end{bmatrix}$$

You will be asked to find the condition for this system to be invertible in your homework assignment.

b) By controllability, we can place the eigenvalues of the matrix arbitrarily. Find an explicit K. *You can use the method presented earlier in the chapter, or try to explicitly arrive at a stabilizing control matrix.*

c) Then, by Theorem 11.4.1, the system is locally stable around the equilibrium point. Precisely explain why this is the case.

Exercise 12.7.2 *Consider the linear system*

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u$$

Is this system controllable? Does there exist a matrix K *so that with* $u = Kx$, *the eigenvalues of the closed-loop matrix:* $\left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} K \right)$ *can be arbitrarily assigned?*

Exercise 12.7.3 *Consider*

$$\frac{dx}{dt} = Ax + Bu, \qquad y = Cx$$

with

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} x$$

a) Is this system observable? b) Is this system controllable? c) Provide a stabilizing feedback control policy by running an observer.

Hint: a) and b) Yes. c) The system is both controllable and observable. If the system state were available, we could have $u = -Kx$ *and select* K *so that* $A - BK$ *is stable. Find such a* K. *Now, we can run an observer as explained in Section*

12.4:

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + L(y - C\hat{x})$$

with the property that $A - LC$ is stable. Find such an L. Then, the control to be applied would be: $u_t = -K\hat{x}_t$. Find explicit values.

Exercise 12.7.4 *a) Show that controllability is invariant under an algebraically equivalent transformation of the coordinates: $\tilde{x} = Px$ for some invertible P.*

b) Consider

$$\frac{dx}{dt} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u$$

Express, through a transformation, this system in a controllable canonical realization form.

Exercise 12.7.5 *Consider*

$$\frac{dx}{dt} = Ax + Bu, \qquad y = Cx$$

with

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$
$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} x$$

a) Is this system observable? b) Is this system controllable? c) Provide a stabilizing feedback control policy by running an observer.

Note. The model here and the model for P given in Exercise 8.5.2 are related.

Solution. *The system is both controllable and observable. If the system state were available, we could have $u = -Kx$ and select K so that $A - BK$ is stable. Find such a K.*

Now, we can run an observer as explain in the lecture notes:

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + L(y - C\hat{x})$$

with the property that $A - LC$ is stable. Find such an L.

Then, the control to be applied would be: $u_t = -K\hat{x}_t$.

A

Integration and Some Useful Properties

A.1 Measurable Space

Let $\mathbb{X}$ be a collection of points. Let $\mathcal{F}$ be a collection of subsets of $\mathbb{X}$ with the following properties such that $\mathcal{F}$ is a σ-field (also called a σ-algebra), that is:

- $\mathbb{X} \in \mathcal{F}$
- If $A \in \mathcal{F}$, then $\mathbb{X} \setminus A \in \mathcal{F}$
- If $A_k \in \mathcal{F}, k = 1, 2, 3, \ldots$, then $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$ (that is, the collection is closed under countably many unions).

By De Morgan's laws, and set properties, it can be shown that the collection has to be closed under countable intersections as well.

If the third item above holds for only finitely many unions or intersections, then, the collection of subsets is said to be a *field* or $algebra$.

With the above, $(\mathbb{X}, \mathcal{F})$ is termed a measurable space (that is we can associate a measure to this space; which we will discuss shortly). For example the full power-set of any set is a σ-field.

A $\sigma-$field $\mathcal{J}$ is generated by a collection of sets $\mathcal{A}$, if $\mathcal{J}$ is the smallest $\sigma-$field containing the sets in $\mathcal{A}$, and in this case, we write $\mathcal{J} = \sigma(\mathcal{A})$.

A.1.1 Borel $\sigma-$field

An important class of σ-fields is the Borel $\sigma-$field on a metric (or more generally, topological) space. Such a $\sigma-$field is the one which is generated by open sets. The term *open* naturally depends on the space being considered. For this course, we will mainly consider spaces which are complete, separable and metric spaces (such as the space of real numbers $\mathbb{R}$, or countable sets). Recall that in a metric space with metric d, a set U is open if for every $x \in U$, there exists some $\epsilon > 0$ such that $\{y : d(x, y) < \epsilon\} \subset U$. We note also that the empty set is a special open set.

The Borel $\sigma-$field on $\mathbb{R}$ is then the one generated by sets of the form $(a, b) \subset \mathbb{R}$, that is, open intervals We will denote the Borel $\sigma-$field on a space $\mathbb{X}$ as $\mathcal{B}(\mathbb{X})$.

A.1.2 Measurable Function

If $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ are measurable spaces; we say a mapping from $h : \mathbb{X} \to \mathbb{Y}$ is measurable if

$$h^{-1}(B) = \{x : h(x) \in B\} \in \mathcal{B}(\mathbb{X}), \quad \forall B \in \mathcal{B}(\mathbb{Y})$$

Theorem A.1.1 *To show that a function is measurable, it is sufficient to check the measurability of the inverses of sets that generate the σ-algebra on the image space.*

Proof. Observe that set operations satisfy that for any $B \in \mathcal{B}(\mathbb{Y})$: $h^{-1}(\mathbb{Y} \setminus B) = \mathbb{X} \setminus h^{-1}(B)$ and

$$h^{-1}(\cup_{i=1}^{\infty} B_i) = \cup_{i=1}^{\infty} h^{-1}(B_i), \quad h^{-1}(\cap_{i=1}^{\infty} B_i) = \cap_{i=1}^{\infty} h^{-1}(B_i).$$

Thus, the set of all subsets whose inverses are Borel:

$$\mathcal{M} = \{B \subset \mathbb{Y} : h^{-1}(B) \in \mathcal{B}(\mathbb{X})\}$$

is a σ-algebra over $\mathbb{Y}$ and this set contains the open sets. Thus, $\mathcal{B}(\mathbb{X}) \subset \mathcal{M}$. □

Therefore, for Borel measurability, it suffices to check the measurability of the inverse images of open sets. Furthermore, for real valued functions, to check the measurability of the inverse images of open sets, it suffices to check the measurability of the inverse images sets of the form $\{(-\infty, a], a \in \mathbb{R}\}$, $\{(-\infty, a), a \in \mathbb{R}\}$, $\{(a, \infty), a \in \mathbb{R}\}$ or $\{[a, -\infty), a \in \mathbb{R}\}$, since each of these generate the Borel σ-field on $\mathbb{R}$. In fact, here we can restrict a to be $\mathbb{Q}$-valued, where $\mathbb{Q}$ is the set of rational numbers.

A.1.3 Measure

A positive measure μ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ is a map from $\mathcal{B}(\mathbb{X})$ to $[0, \infty]$ which is *countably additive* such that for $A_k \in \mathcal{B}(\mathbb{X})$ and $A_k \cap A_j = \emptyset$:

$$\mu\Big(\cup_{k=1}^{\infty} A_k \Big) = \sum_{k=1}^{\infty} \mu(A_k).$$

Definition A.1.1 *μ is a probability measure if it is positive and $\mu(\mathbb{X}) = 1$.*

Definition A.1.2 *A measure μ is finite if $\mu(\mathbb{X}) < \infty$, and $\sigma-$finite if there exist a collection of subsets such that $X = \cup_{k=1}^{\infty} A_k$ with $\mu(A_k) < \infty$ for all k.*

On the real line $\mathbb{R}$, the Lebesgue measure is defined on the Borel $\sigma-$field (in fact on a somewhat larger field obtained through adding all subsets of Borel sets of measure zero: this is known as *completion* of a σ-field) such that for $A = (a, b)$, $\mu(A) = b - a$. Borel field of subsets is a subset of *Lebesgue measurable* sets, that is there exist Lebesgue measurable sets which are not Borel sets. There exist Lebesgue measurable sets of measure zero which contain uncountably many elements; an example is the Cantor set.

A.1.4 The Extension Theorem

Theorem A.1.2 *[The Extension Theorem (Carathéodory)] Let $\mathcal{M}$ be an algebra over $\mathbb{X}$, and suppose that there exists a map (called a pre-measure) $P : \mathcal{M} \to \mathbb{R}_+$ so that for any (possibly countably infinitely many) pairwise disjoint sets $A_n \in \mathcal{M}$, if the countable union $\cup_n A_n \in \mathcal{M}$, then $P(\cup_n A_n) = \sum_n P(A_n)$. Suppose also that there exists a countable collection of sets B_n with $\mathbb{X} = \cup_n B_n$, each with $P(B_n) < \infty$ (that is P is σ-finite). Then, there exists a unique measure P' on the $\sigma-$field generated by $\mathcal{M}$, $\sigma(\mathcal{M})$, which is consistent with P on $\mathcal{M}$.*

The above is useful since, when one states that two measures are equal it suffices to check if they are equal on the set of sets which generate the $\sigma-$algebra, and not necessarily on the entire $\sigma-$field.

We can construct the Lebesgue measure on $\mathcal{B}(\mathbb{R})$ by defining it on finitely many unions and intersections of intervals of the form (a, b), $[a, b)$, $(a, b]$ and $[a, b]$, and the empty set, thus forming a field, and extending this to the Borel $\sigma-$field. Thus, the relation $\mu(a, b) = b - a$ for $b > a$ is sufficient to define the Lebesgue measure.

A.1.5 Integration

Let h be a non-negative measurable function from $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The Lebesgue integral of h with respect to a measure μ can be defined in three steps:

First, for $A \in \mathcal{B}(\mathbb{X})$, define $1_{\{x \in A\}}$ (or $1_{(x \in A)}$, or $1_A(x)$) as an indicator function for event $x \in A$, that is the value that the function takes is 1 if $x \in A$, and 0 otherwise. In this case, define

$$\int_{\mathbb{X}} 1_{\{x \in A\}} \mu(dx) := \mu(A).$$

Now, let us define simple functions h such that, there exist $A_1, A_2, \ldots, A_n$ all in $\mathcal{B}(\mathbb{X})$ and positive numbers $b_1, b_2, \ldots, b_n$ such that $h_n(x) = \sum_{k=1}^{n} b_k 1_{\{x \in A_k\}}$. For such functions, define

$$\int_{\mathbb{X}} h_n(x) \mu(dx) := \sum_{k=1}^{n} b_k \mu(A_k).$$

Now, for any given measurable h, there exists a sequence of simple functions h_n such that $h_n(x) \uparrow h(x)$ monotonically, that is $h_{n+1}(x) \geq h_n(x)$ (for a construction, if h only takes non-negative values, consider partitioning the positive real line to two intervals $[0, n)$ and $[n, \infty)$, and partition $[0, n)$ to $n2^n$ uniform intervals, define $h_n(x)$ to be the lower floor of the interval that contains $h(x)$: thus

$$h_n(x) = k2^{-n}, \quad \text{if} \quad k2^{-n} \leq h(x) < (k+1)2^{-n}, \quad k = 0, 1, \cdots, n2^n - 1,$$

and $h_n(x) = n$ for $h(x) \geq n$. By definition, and since $h^{-1}([k2^{-n}, (k+1)2^{-n}))$ is Borel, h_n is a simple function. If the function takes also negative values, write $h(x) = h_+(x) - h_-(x)$, where h_+ is the non-negative part and $-h_-$ is the negative part, and construct the same for $h_-(x)$. We *define* the limit (which exists as a real valued monotonically increasing sequence) as the Lebesgue integral:

$$\lim_{n \to \infty} \int h_n(x) \mu(dx) =: \int h(x) \mu(dx)$$

We note that the notation $\int h d\mu$ or $\int h(x) d\mu(x)$ can also be used in place of $\int h(x) \mu(dx)$.

There are three important convergence theorems.

A.1.6 Fatou's Lemma, the Monotone Convergence Theorem and the Dominated Convergence Theorem

Theorem A.1.3 (Monotone Convergence Theorem) *If μ is a $\sigma-$finite positive measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $\{f_n, n \in \mathbb{Z}_+\}$ is a sequence of measurable functions from $\mathbb{X}$ to $\mathbb{R}$ which pointwise, monotonically, converges to f so that $0 \leq f_n(x) \leq f_{n+1}(x)$ for all n, and*

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

for $\mu-$almost every x, then

$$\int_{\mathbb{X}} f(x) \mu(dx) = \lim_{n \to \infty} \int_{\mathbb{X}} f_n(x) \mu(dx)$$

The following is a consequence of the monotone convergence theorem, but is a critical result which will be utilized in the notes.

Theorem A.1.4 (Fatou's Lemma) *If μ is a $\sigma-$finite positive measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $\{f_n, n \in \mathbb{Z}_+\}$ is a sequence of measurable functions, bounded from below, from $\mathbb{X}$ to $\mathbb{R}$, then*

$$\int_{\mathbb{X}} \liminf_{n \to \infty} f_n(x) \mu(dx) \leq \liminf_{n \to \infty} \int_{\mathbb{X}} f_n(x) \mu(dx)$$

Theorem A.1.5 (Dominated Convergence Theorem) *If (i) μ is a $\sigma-$finite positive measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, (ii) g is a Borel measurable function with*

$$\int_{\mathbb{X}} g(x)\mu(dx) < \infty,$$

and (iii) $\{f_n, n \in \mathbb{Z}_+\}$ is a sequence of measurable functions from $\mathbb{X}$ to $\mathbb{R}$ which satisfy $|f_n(x)| \leq g(x)$ for $\mu-$almost every x, and $\lim_{n\to\infty} f_n(x) = f(x)$, then:

$$\int_{\mathbb{X}} f(x)\mu(dx) = \lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(dx)$$

Note that for the monotone convergence theorem, there is no restriction on boundedness; whereas for the dominated convergence theorem, there is a boundedness condition. On the other hand, for the dominated convergence theorem, the pointwise convergence does not have to be monotone.

A.2 Differentiation under an Integral

Consider an integral that depends on two parameters of the form:

$$J(r) = \int_{\mathbb{R}} g(r,t)dt$$

for some integrable $g(r,\cdot)$ for every $r \in [a,b]$.

Theorem A.2.1 *Suppose that $r \in (a,b)$ for some $a, b \in \mathbb{R}$ so that (i) for all t, $g(r,t)$ is continuously differentiable and (ii) there exists an integrable function h so that for all $r \in (a,b)$,*

$$|\frac{\partial}{\partial r} g(r,t)| \leq h(t)$$

almost everywhere. Then,

$$\frac{d}{dr}\int_{\mathbb{R}} g(r,t)dt = \int_{\mathbb{R}} \frac{\partial}{\partial r} g(r,t)dt$$

Proof.

$$\begin{aligned}
&\frac{d}{dr}\int_{\mathbb{R}} g(r,t)dt \\
&= \lim_{s\to 0}\int_{\mathbb{R}} \frac{g(r+s,t)-g(r,t)}{s}dt \\
&= \lim_{s\to 0}\int_{\mathbb{R}} \frac{\partial}{\partial r} g(r+\tau(r,s),t)dt \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial r} g(r,t)dt
\end{aligned} \tag{A.1}$$

Here, $r \leq \tau(r,s) \leq r+s$ by the mean value theorem; and observe that $\lim_{s\to 0}\tau(r,s) = r$. The last equality follows from the dominated convergence theorem since $g(r+\tau(r,s),t)$ is dominated by h (for sufficiently small s values) and converges pointwise. □

The above also applies for complex-valued functions, by considering the real and the imaginary parts separately.

As an example, consider the CCFT of a function g:

$$\hat{g}(f) = \int g(t)e^{-i2\pi ft}dt$$

We observe that provided $\int |g(t)|t < \infty$,

$$\frac{d}{df}\hat{g}(f) = -i2\pi \int g(t)te^{-i2\pi ft}dt$$

To see this, express the derivative $\frac{d}{df}\hat{g}(f)$ as the limit $\lim_{h\to 0}\int g(t)e^{-i2\pi ft}\frac{e^{-i2\pi ht}-1}{h}dt$ and write

$$\frac{(e^{-i2\pi ht}-1)}{h} = \frac{\cos(2\pi ht)-1)}{h} - \frac{i\sin(2\pi ht)-0}{h} = 2\pi t(\sin(2\pi \bar{h}t) + i\cos(2\pi \tilde{h}t))$$

for some $\bar{h} \in [0,h], \tilde{h} \in [0,h]$, via the mean-value theorem for both expressions. Now, apply Theorem A.2.1.

A.3 Fubini's Theorem (also Fubini-Tonelli's Theorem)

Theorem A.3.1 *Let $f : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ and E, F be Borel sets. Then,*
(i) If f is non-negative on E, F,

$$\int_{E\times F} f(x,y)dxdy = \int_E \left(\int_F f(x,y)dy\right)dx = \int_F \left(\int_E f(x,y)dx\right)dy$$

(ii) If f is integrable on E, F,

$$\int_{E\times F} f(x,y)dxdy = \int_E \left(\int_F f(x,y)dy\right)dx = \int_F \left(\int_E f(x,y)dx\right)dy$$

We note that f is integrable if and only if $\int_{E\times F}|f(x,y)|dxdy$ is finite, which implies that to check for integrability it suffices to show that one of the integrals $\int_E \left(\int_F |f(x,y)|dy\right)dx$ or $\int_F \left(\int_E |f(x,y)|dx\right)dy$ is finite. Note that if the function considered is non-negative, we don't seek integrability. This theorem also applies to summations. This theorem is a very useful result while working with transformations, as we do extensively in this course.

B

Cauchy's Integral Formula

Let $\Omega \subset \mathbb{C}$ be an open, connected set. Let $f : \Omega \to \mathbb{C}$ be a holomorphic function; that is, the limit

$$f'(p) = \lim_{s \to p} \frac{f(s) - f(p)}{s - p}$$

exists (which is the derivative of f at p) at every $p \in \Omega$. We note that a complex function is holomorphic if and only if it is analytic (unlike the real function setup).

Theorem B.0.1 (Cauchy's Integral Theorem) *Let $\gamma : [0, 1] \to \Omega$ be differentiable with $\gamma(0) = \gamma(1)$ and let Γ be the closed contour traced by γ. Then,*

$$\int_0^1 f(\gamma(t))\gamma'(t)dt = \int_\Gamma f(z)dz = 0$$

With the above, we will now derive a very important result in complex analysis as is relevant in our course.

Theorem B.0.2 (Cauchy's Integral Formula) *Let Γ be a contour that encircles the point $p \in \mathbb{C}$ (counterclockwise) only once and f be as above. Then,*

$$\int_\Gamma \frac{f(s)}{s - p} ds = 2\pi i f(p)$$

Note that this formula implies that if we know the values of a function along the boundaries of a contour, we can uniquely recover the values of the function inside the contour, by selecting any p inside Γ and apply the result above.

Proof. a) First, observe that the value of the integral does not depend on the path as long as the path encircles p, by the Cauchy integral theorem.

In particular, we can take $C_\epsilon = p + \epsilon e^{i2\pi t}$ as t ranges from 0 to 1.

Then,

$$\int_\Gamma \frac{f(s)}{s - p} ds = \int_{C_\epsilon} \frac{f(s)}{s - p} ds$$

b) Write $s = p + \epsilon e^{i2\pi t}$ and $ds = \epsilon i 2\pi e^{i2\pi t}$. Now, first take $f(s) \equiv 1$, and observe that

$$\int_{C_\epsilon} \frac{1}{s - p} ds = \epsilon i 2\pi \int_0^1 \frac{1}{\epsilon e^{i2\pi t}} \epsilon e^{i2\pi t} dt = 2\pi i$$

c) We have then that

$$\int_{C_\epsilon} \frac{f(s)}{s - p} ds - 2\pi i f(p) = \int_{C_\epsilon} \frac{f(s) - f(p)}{s - p} ds.$$

We will show that this difference is zero. Write

$$|\int_{C_\epsilon} \frac{f(s)-f(p)}{s-p} ds| \leq \int_0^1 |\frac{f(p+\epsilon e^{i2\pi t})-f(p)}{\epsilon} \epsilon e^{i2\pi t}| dt \leq \max_{t\in[0,1]} |f(p+\epsilon e^{i2\pi t})-f(p)|$$

The above holds for every $\epsilon > 0$, and as can be taken arbitrarily close to zero, the result follows. □

References

1. K. J. Astrom and R. M. Murray, *Feedback Systems*, Princeton University Press, 2008.
2. T. Başar, S. P. Meyn, W. R. Perkins, *Control System Theory and Design*, University of Illinois at Urbana-Champaign, Lecture Notes, arXiv: 2007.01367)
3. C.-T. Chen, *Linear System Theory and Design*, Oxford University Press, Inc., 1998.
4. G. F. Franklin, J. D. Powell and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, Prentice Hall Press, 2014.
5. Z. Gajic, Zoran, *Linear Dynamic Systems and Signals*, Prentice Hall/Pearson Education, 2003.
6. C. Gasquet and P. Witomski, *Fourier Analysis and Applications: Filtering, Numerical Computation, Wavelets*, Springer (translated by R. Ryan).
7. J. K. Hunter and B. Nachtergaele, *Applied Analysis*, World Scientific Publishing Co. Inc., River Edge, NJ, 2001.
8. D. Liberzon, *ECE 517: Nonlinear and Adaptive Control*, University of Illinois at Urbana-Champaign, Lecture Notes.
9. H. Kwakernaak and R. Sivan, *Modern Signals and Systems*, Prentice Hall, 1991.
10. A. D. Lewis, *A Mathematical Approach to Classical Control*, MTHE 332 Lecture Notes, Queen's University.
11. D. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.
12. A. V. Oppenheim, A. S. Willsky and S. Nawab, *Signals and Systems*, Prentice-Hall, Inc., Upper Saddle River, NJ; 2nd Edition, 1996.
13. D. J. Uherka and A. M. Sergott. On the continuous dependence of the roots of a polynomial on its coefficients. The American mathematical monthly, 84(5), 368-370, 1977.