Linear Control Theory and Structured Markov Chains

Yoni Nazarathy

Lecture Notes for a Course in the 2016 AMSI Summer School
Based on a book draft co-authored with
Sophie Hautphenne, Erjen Lefeber and Peter Taylor.

Last Updated: February 2, 2016.

Whole Book Draft

Preface

This booklet contains lecture notes and exercises for a 2016 AMSI Summer School Course: "Linear Control Theory and Structured Markov Chains" taught at RMIT in Melbourne by Yoni Nazarathy. The notes are based on a subset of a draft book about a similar subject by Sophie Hautphenne, Erjen Lefeber, Yoni Nazarathy and Peter Taylor. The course includes 28 lecture hours spread over 3.5 weeks. The course includes assignments, short in-class quizzes and a take-home exam. These assement items are to appear in the notes as well.

The associated book is designed to teach readers, elements of linear control theory and structured Markov chains. These two fields rarely receive a unified treatment as is given here. It is assumed that the readers have a minimal knowledge of calculus, linear algebra and probability, yet most of the needed facts are summarized in the appendix, with the exception of basic calculus. Nevertheless, the level of mathematical maturity assumed is that of a person who has covered 2-4 years of applied mathematics, computer science and/or analytic engineering courses.

Linear control theory is all about mathematical models of systems that abstract dynamic behavior governed by actuatotors and sensed by sensors. By designing state feedback controllers, one is often able to modify the behavior of a system which otherwise would operate in an undesirable manner. The underlying mathematical models are inherently deterministic, as is suited for many real life systems governed by elementary physical laws. The general constructs are system models, feedback control, observers and optimal control under quadratic costs. The basic theory covered in this book has reached relative maturity nearly half a century ago: the 1960's, following some of the contributions by Kalman and others. The working mathematics needed to master basic linear control theory is centered around linear algebra and basic integral transforms. The theory relies heavily on eigenvalues, eigenvectors and others aspects related to the spectral decomposition of matrices.

Markov chains are naturally related to linear dynamical systems and hence linear control theory, since the state transition probabilities of Markov chains evolve as a linear dynamical system. In addition the use of spectral decompositions of matrices, the matrix exponential and other related features also resembles linear dynamical systems. The field of structured Markov chains, also referred to as Matrix Analytic Methods, goes back to the mid 1970's, yet has gained popularity in the teletraffic, operations research

and applied probability community only in the past two decades. It is unarguably a more esoteric branch of applied mathematics in comparison to linear control theory and it is currently not applied as abundantly as the former field.

A few books at a similar level to this one focus on dynamical systems and show that the probabilistic evolution of Markov chains over finite state spaces behaves as linear dynamical systems. This appears most notably in [**?**]. Yet, structured Markov chains are more specialized and posses more miracles. In certain cases, one is able to analyze the behavior of Markov chains on infinite state spaces, by using their structure. E.g. underlying matrices may be of block diagonal form. This field of research often focuses on finding effective algorithms for solutions of the underlying performance analysis problems. In this book we simply illustrate the basic ideas and methods of the field. It should be noted that structured Markov chains (as Markov chains in general) often make heavy use of non-negative matrix theory (e.g. the celebrated Perron-Frobenius Theorem). This aspect of linear algebra does not play a role in the classic linear control theory that we present here, yet appears in the more specialized study of control of non-negative systems.

Besides the mathematical relation between linear control theory and structured Markov chains, there is also a much more practical relation which we stress in this book. Both fields, together with their underlying methods, are geared for improving the way we understand and operate dynamical systems. Such systems may be physical, chemical, biological, electronic or human. With its styled models, the field of linear control theory allows us to find good ways to actually control such systems, on-line. With its ability to capture truly random behavior, the field of structured Markov chains allows us to both describe some significant behaviors governed by randomness, as well as to efficiently quantify (solve) their behaviors. But control does not really play a role.

With the exception of a few places around the world (e.g. the Mechanical Engineering Department at Eindhoven University of Technology), these two fields are rarely taught simultaneously. Our goal is to facilitate such action through this book. Such a unified treatment will allow applied mathematicians and systems engineers to understand the underlying concepts of both fields in parallel, building on the connections between the two.

Below is a detailed outline of the structure of the book. Our choice of material to cover was such as to demonstrate most of the basic features of both linear control theory and structured Markov chains, in a treatment that is as unified as possible.

Outline of the contents:

The notes contains a few chapters and some appendices. The chapters are best read sequentially. Notation is introduced sequentially. The chapters contain embedded short exercises. These are meant to help the reader as she progresses through the book, yet at the same time may serve as mini-theorems. That is, these exercises are both deductive

and informative. They often contain statements that are useful in their own right. The end of each chapter contains a few additional exercises. Some of these exercises are often more demanding, either requiring computer computation or deeper thought. We do not refer to computer commands related to the methods and algorithms in he book explicitly. Nevertheless, in several selected places, we have illustrated example MATLAB code that can be used.

For the 2016 AMSI summer school, we have indicated besides each chapter the in-class duration that this chapter will receive in hours.

Chapter 1 (2h) is an elementary introduction to systems modeling and processes. In this chapter we introduce the types of mathematical objects that are analyzed, give a feel for some applications, and describe the various use-cases in which such an analysis can be carried out. By a use-case we mean an activity carried out by a person analyzing such processes. Such use cases include "performance evaluation", "controller design", "optimization" as well as more refined tasks such as stability analysis, pole placement or evaluation of hitting time distributions.

Chapter 2 (7h) deals with two elementary concepts: Linear Time Invariant (LTI) Systems and Probability Distributions. LTI systems are presented from the viewpoint of an engineering-based "signals and systems" course. A signal is essentially a time function and system is an operator on functional space. Operators that have the linearity and time-invariance property are LTI and are described neatly by either their impulse response, step response, or integral transforms of one of these (the transfer function). It is here that the convolution of two signals plays a key role. Signals can also be used to describe probability distributions. A probability distribution is essentially an integrable non-negative signal. Basic relations between signals, systems and probability distributions are introduced. In passing we also describe an input–output form of stability: BIBO stability, standing for "bounded input results in bounded output". We also present feedback configurations of LTI systems, showing the usefulness of the frequency domain (s-plane) representation of such systems.

Chapter 3 (13h) moves onto dynamical models. It is here that the notion of state is introduced. The chapter begins by introducing linear (deterministic) dynamical systems. These are basically solutions to systems of linear differential equations where the free variable represents time. Solutions are characterized by matrix powers in discrete time and matrix exponentials in continuous time. Evaluation of matrix powers and matrix exponentials is a subject of its right as it has to do with the spectral properties of matrices, this is surveyed as well. The chapter then moves onto systems with discrete countable (finite or infinite) state spaces evolving stochastically: Markov chains. The basics of discrete time and continuous time Markov chains are surveyed. In doing this a

few example systems are presented. We then move onto presenting input–state–output systems, which we refer to as (A, B, C, D) systems. These again are deterministic objects. This notation is often used in control theory and we adopt it throughout the book. The matrices A and B describe the effect on input on state. The matrices C and D are used to describe the effect on state and input on the output. After describing (A, B, C, D) systems we move onto distributions that are commonly called Matrix Exponential distributions. These can be shown to be directly related to (A, B, C, D) systems. We then move onto the special case of phase type (PH) distributions that are matrix exponential distributions that have a probabilistic interpretation related to absorbing Markov chains. In presenting PH distributions we also show parameterized special cases.

Chapter 4 (0h) is not taught as part of the course. This chapter dives into the heart of Matrix Analytic Modeling and analysis, describing quasi birth and deaths processes, Markovian arrival processes and Markovian Binary trees, together with the algorithms for such models. The chapter begins by describing QBDs both in discrete and continuous time. Then moves onto Matrix Geometric Solutions for the stationary distribution showing the importance of the matrices G and R. The chapter then shows elementary algorithms to solve for G and R focusing on the probabilistic interpretation of iterations of the algorithms. State of the art methods are summarized but are not described in detail. Markovian Arrival Point Processes and their various sub-classes are also survyed. As examples, the chapter considers the M/PH/1 queue, PH/M/1 queue as well as the PH/PH/1 generalization. The idea is to illustrate the power of algorithmic analysis of stochastic systems.

Chapter 5 (4h) focuses on (A, B, C, D) systems as used in control theory. Two main concepts are introduced and analyzed: state feedback control and observers. These are cast in the theoretical framework of basic linear control theory, showing the notions of controllability and observabillity. The chapter begins by introducing two physical examples of (A, B, C, D) systems. The chapter also introduces canonical forms of (A, B, C, D) systems.

Chapter 6 (2h) deals with stability of both deterministic and stochastic systems. Notions and conditions for stability were alluded to in previous chapters, yet this chapter gives a comprehensive treatment. At first stability conditions for general deterministic dynamical systems are presented. The concept of a Lyapounov function is introduced. This is the applied to linear systems and after that stability of arbitrary systems by means of linearization is introduced. Following this, examples of setting stabilizing feedback control rules are given. We then move onto stability of stochastic systems (essentially positive recurrence). The concept of a Foster-Lyapounov function is given for showing positive recurrence of Markov chains. We then apply it to quasi-birth-death processes

proving some of the stability conditions given in Chapter 4 hold. Further stability conditions of QBD's are also given. The chapter also contains the Routh-Hourwitz and Jury criterions.

Chapter 7 (0h) is not taught as part of the course. is about optimal linear quadratic control. At first Bellman's dynamic programming principle is introduced in generality, and then it is formulated for systems with linear dynamics and quadratic costs of state and control efforts. The linear quadratic regulator (LQR) is introduced together with its state feedback control mechanism, obtained by solving Ricaati equations. Relations to stability are overviewed. The chapter then moves onto Model-predictive control and constrained LQR.

Chapter 8 (0h) is not taught as part of the course. This chapter deals with fluid buffers. The chapter involves both results from applied probability (and MAM), as well as a few optimal control examples for deterministic fluid systems controlled by a switching server. The chapter begins with an account of the classic fluid model of Anick, Mitra and Sondhi. It then moves onto additional models including deterministic switching models.

Chapter 9 (0h) is not taught as part of the course. This chapter introduces methods for dealing with deterministic models with additive noise. As opposed to Markov chain models, such models behave according to deterministic laws, e.g. (A, B, C, D) systems, but are subject to (relatively small) stochastic disturbances as well as to measurement errors that are stochastic. After introducing basic concepts of estimation, the chapter introduces the celebrated Kalman filter. There is also brief mention of linear quadratic Gaussian control (LQG).

The notes also contains an extensive appendix **which the students are required to cover by themselves as demand arises**. The appendix contains proofs of results in cases where we believe that understanding the proof is instructive to understanding the general development in the text. In other cases, proofs are omitted.

Appendix A touches on a variety of basics: Sets, Counting, Number Systems (including complex numbers), Polynomials and basic operations on vectors and matrices.

Appendix B covers the basic results of linear algebra, dealing with vector spaces, linear transformations and their associated spaces, linear independence, bases, determinants and basics of characteristic polynomials, eigenvalues and eigenvectors including the Jordan Canonical Form.

Appendix C covers additional needed results of linear algebra.

Appendix D contains probabilistic background.

Appendix E contains further Markov chain results, complementing the results presented in the book.

Appendix F deals with integral transforms, convolutions and generalized functions. At first convolutions are presented, motivated by the need to know the distribution of the sum of two independent random variables. Then generalized functions (e.g. the delta function) are introduced in an informal manner, related to convolutions. We then present the Laplace transform (one sided) and the Laplace-Stiltijes Transform. Also dealing with the region of convergence (ROC). In here we also present an elementary treatment of partial fraction expansions, a method often used for inverting rational Laplace transforms. The special case of the Fourier transform is briefly surveyed, together with a discussion of the characteristic function of a probability distribution and the moment generating function. We then briefly outline results of the z-transform and of probability generating functions.

Besides thanking Sophie, Erjen and Peter, my co-authors for the book on which these notes are based, I would also like to thank (on their behalf) to several colleagues and students for valuable input that helped improve the book. Mark Fackrell and Nigel Bean's analysis of Matrix Exponential Distributions has motivated us to treat the subjects of this book in a unified treatment. Guy Latouche was helpful with comments dealing with MAM. Giang Nugyen taught jointly with Sophie Hautphenene a course in Vietnam covering some of the subjects. A Master's student from Eindhoven, Kay Peeters, visiting Brisbane and Melbourne for 3 months and prepared a variety of numerical examples and illustrations, on which some of the current illustrations are based. Also thanks to Azam Asanjarani and to Darcy Bermingham. The backbone of the book originated while the authors were teaching an AMSI summer school course, in Melbourne during January 2013. Comments from a few students such as Jessica Yue Ze Chan were helpful.

I hope you find these notes useful,
Yoni.

Contents

Chapter 1

Introduction (2h)

A *process* is a function of time describing the behavior of some system. In this book we deal with several types of processes. Our aim is to essentially cover processes coming from two fields of research:

1. Deterministic linear systems and control.

2. Markovian stochastic systems with a structured state-space.

The first field is sometimes termed *systems and control theory*. Today it lies on the intersection of engineering and applied mathematics. The second field is called *Matrix Analytic Methods* (MAM), it is a sub-field of *Applied Probability* (which is sometimes viewed as a branch of *Operations Research*). MAM mostly deals with the analysis of specific types of *structured Markov models*.

Control and systems theory advanced greatly in the 1960's due to the American and Soviet space programs. Matrix Analytic Methods is a newer area of research. It became a "recognized" subfield of applied probability sometime in the past 25 years. Thousands of researchers (and many more practitioners including control engineers) are aware and knowledgeable of systems and control theory. As opposed to that, MAM still remains a rather specialized area. At the basis of systems and control theory, lies the study of *linear control theory* (LCT). In this book we teach MAM and LCT together, presenting a unified exposition of the two fields where possible.

Our motivation for this unification is that both LCT and MAM use similar mathematical structures, patterns and results from *linear algebra* to describe models, methods and their properties. Further, both fields can sometimes be used to approach the same type of application, yet from different viewpoints. LCT yields efficient methods for designing automatic feedback controllers to systems. MAM yields efficient computational methods for performance analysis of a rich class of stochastic models.

In this introductory chapter informally introduce a variety of basic terms. In doing so, we do not describe LCT nor MAM further. We also motivate the study of dynamical

models, namely models that describe the evolution of processes over time. Further, we survey the remainder of the book as well as the *mathematical background* appendix.

1.1 Types of Processes

The dynamical processes arising in LCT and MAM can essentially be classified into four types. These types differ based on the time-index (continuous or discrete) and their values (uncountable or countable). We generally use the following notation:

- $\{\boldsymbol{x}(t)\}$ with $t \in \mathbb{R}$ and $\boldsymbol{x}(t) \in \mathbb{R}^n$.
- $\{X(t)\}$ with $t \in \mathbb{R}$ and $X(t) \in \mathcal{S}$, where $\mathcal{S}$ is some countable (finite or infinite set).
- $\{\boldsymbol{x}(\ell)\}$ with $\ell \in \mathbb{Z}$ and $\boldsymbol{x}(\ell) \in \mathbb{R}^n$.
- $\{X(\ell)\}$ with $\ell \in \mathbb{Z}$ and $X(\ell) \in \mathcal{S}$, where $\mathcal{S}$ is some countable (finite or infinite set).

The processes $\{\boldsymbol{x}(t)\}$ and $\{X(t)\}$ are *continuous time* while the processes $\{\boldsymbol{x}(\ell)\}$ and $\{X(\ell)\}$ are *discrete time*. Considering the values that the processes take, $\{\boldsymbol{x}(t)\}$ and $\{\boldsymbol{x}(\ell)\}$ take on values in some Euclidean vector space (uncountable), as opposed to that, $\{X(t)\}$ and $\{X(\ell)\}$ take on values in some countable set.

In some instances the processes are viewed as *deterministic*. By this we mean their *trajectory* is fixed and does not involve randomness. Alternatively they are modelled as *stochastic*. This implies that their evolution involves some chance behaviour that can be formally specified through a *probability space*. This means that there is not one unique possible trajectory (also known as *sample path* in the stochastic case) of the process but rather a collection (typically infinite collection) of possible realizations:

$$\{X_\omega(\cdot), \quad \omega \in \Omega\}.$$

It is then a matter of the *probability law* of the process to indicate which specific realization is taking place in practice.

Most of the LCT models that we cover in this book are of a deterministic nature. As opposed to that, all of the MAM models that we cover are stochastic. The basic MAM models that we introduce are based on *Markov chains* on countable state space (with the exception of Chapter 8 on fluid queues). Hence we consider the processes $X(\cdot)$ as stochastic. Similarly the processes $\boldsymbol{x}(\cdot)$ are considered deterministic.

1.1.1 Representations of Countable State Spaces

Since the *state space*, $\mathcal{S}$ of the discrete-state stochastic processes, $X(\cdot)$, is countable, we can often treat it as $\{1, \ldots, N\}$ for some finite N or $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ depending on if

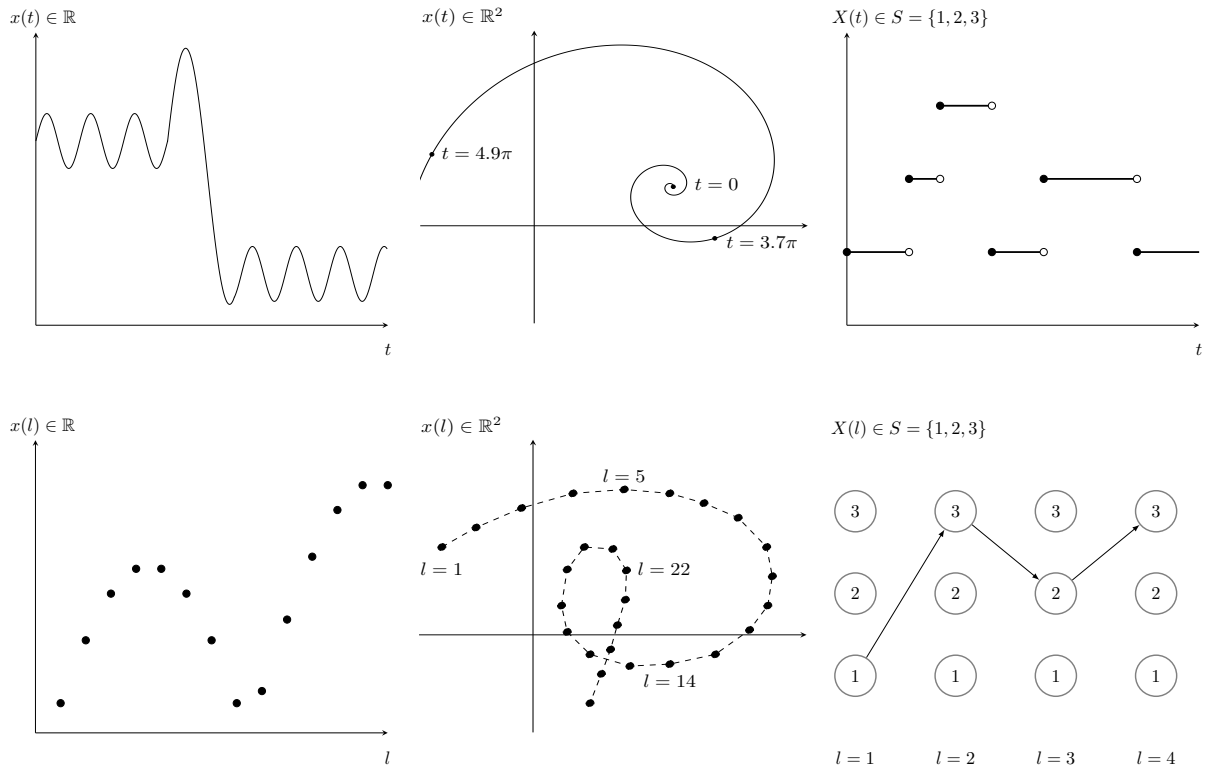


Figure 1.1: Illustration of realizations of different types of processes

it is finite or infinite. Nevertheless, for many of the stochastic processes that we shall consider it will be useful to represent $\mathcal{S}$ as $\mathbb{Z}_+^2$ or some subset of it. In that case we shall call one coordinate of $s \in \mathcal{S}$ as the *level* and the other coordinate as the *phase*. Further, since the process is now vector valued we will denote it by $\{\boldsymbol{X}(t)\}$ in the continuous time case and $\{\boldsymbol{X}(\ell)\}$ in the discrete time case.

1.1.2 Other Variations of Processes (omitted from course)

We shall also touch variations of the types of process, 1–4, detailed above. Which we informally discuss now. One such variation is taking a process with inherently deterministic dynamics, $\boldsymbol{x}(\cdot)$, and adding stochastic "perturbations" to it. In discrete time this is typically done by adding "noise terms" at each of the steps of the process. In continuous time it is typically done by means of a *stochastic differential equation*. Both of these cases are important, yet they are out of the scope of this book.

Another variation is a continuous time, uncountable state (referred to as *continuous*

state) stochastic process that has *piece-wise linear* trajectories taking values in $\mathbb{R}$. In that case, one way to describe a trajectory of the process is based on a sequence of time points,

$$T_0 < T_1 < T_2, \ldots,$$

where the values of $X(t)$ for $t = T_\ell$, $\ell = 0, 1, 2, \ldots$ is given. Then for time points,

$$t \notin \{T_0, T_1, \ldots\},$$

we have,

$$X(t) = X(T_\ell) + (t - T_\ell)\frac{X(T_{\ell+1}) - X(T_\ell)}{T_{\ell+1} - T_\ell} \quad \text{if} \quad t \in (T_\ell, T_{\ell+1}).$$

1.1.3 Behaviours

We shall informally refer to the *behavior* of $\boldsymbol{x}(\cdot)$ or $X(\cdot)$ as a description of the possible trajectories that these processes take. Some researchers have tried to formalize this in what is called the *behavioral approach* to systems. We do not discuss this further. The next section describes what we aim to do with respect to the *behaviors* of processes.

1.2 Use-cases: Modeling, Simulation, Computation, Analysis, Optimization and Control

What do we do with these processes, $\boldsymbol{x}(\cdot)$ or $X(\cdot)$ in their various forms? Well, they typically arise as *models* of true physical situations. Concrete non-trivial examples are in the section below.

We now describe *use-cases* of models. I.e. the actions that we (as applied mathematicians) do with respect to models of processes. Each of these use-cases has an ultimate purpose of helping reach some goal (typically in applications).

1.2.1 Modelling

We shall refer to the action of *modeling* as taking a true physical situation and setting up a deterministic process $\boldsymbol{x}(\cdot)$ or a stochastic process $X(\cdot)$ to describe it. Note that "physical" should be interpreted in the general sense, i.e. it can be monetary, social or related to bits on digital computers. The result of the modeling process is a *model* which is essentially $\boldsymbol{x}(\cdot)$ or $X(\cdot)$ or a family of such processes parameterized in some manner.

Example 1.2.1. *Assume a population of individuals where it is observed (or believed):*

Every year the population doubles.

Assume that at onset there are 10 *individuals.*
Here are some suggested models:

1. $x(0) = 10$ *and*
$$x(\ell + 1) = 2x(\ell).$$

2. $x(0) = 10$ *and*
$$\dot{x}(t) = (\log 2)x(t),$$
where we use the notation $\dot{x}(t) :=: \frac{d}{dt}x(t)$ *and* log *is with the natural base.*

3. $\mathbb{P}(X(0) = 10) = 1$ *and*
$$X(\ell + 1) = \sum_{k=1}^{X(\ell)} \xi_{\ell,k},$$
with $\xi_{\ell,k}$ *i.i.d. non-negative random variables with a specified distribution satisfying* $\mathbb{E}[\xi_{1,1}] = 2$.

4. *A continuous time* branching process *model with a behavior similar to 3 in the same way that the behavior of 2 is similar to 1. We do not specify this model further now.*

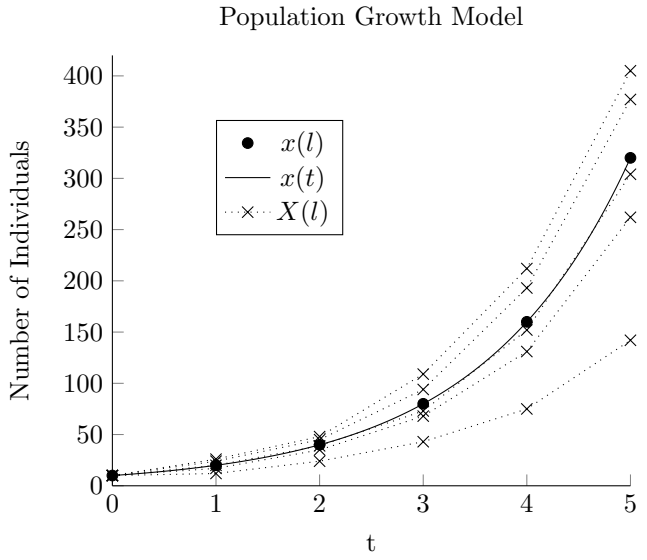


Figure 1.2: Different types of processes that can describe population growth.

As can be seen from the example above we have 4 different models that can be used to describe the same physical situation. The logical reasoning of which *model is best* is part of the action of modeling.

Exercise 1.2.2. *Suggest another model that can describe the same situation. There is obviously not one correct answer.*

1.2.2 Simulation

The action of *simulation* is the action of generating numeric realizations of a given model. For deterministic models it implies plotting $\boldsymbol{x}(\cdot)$ in some manner or generating an array that represents a sample of its values. For stochastic models there is not one single realization, so it implies *generating* one or more realizations of $X(\cdot)$ by means of *Monte-Carlo*. That is, by using pseudo-random number generation and methods of *stochastic simulation*.

Simulation is useful for *visualization* but also for *computation and analysis* as we describe below.

Exercise 1.2.3. *Simulate the trajectories of models (1) and (2) from Example 1.2.1. For model (3), simulate 4 sample trajectories. Plot all 6 realizations on one graph.*

1.2.3 Computation and Analysis

The action of *computation* is all about finding descriptors related to the underlying models (or the underlying processes). Computation may be done by generating closed formulas for descriptors, by running algorithms, or by conducting deterministic or stochastic simulations of $\boldsymbol{x}(\cdot)$ or $X(\cdot)$ respectively.

For example. A computation associated with model (1) of Example 1.2.1 is solving the difference equation to get,

$$x(\ell) = 10 \cdot 2^{\ell}. \tag{1.1}$$

In this case, the computation results in an *analytical* solution.

Exercise 1.2.4. *What is the solution of model (2) of Example 1.2.1? How does it compare to* (1.1)*?*

Getting explicit analytical solutions to differential equations is not always possible. Hence the difference between analysis and computation.

The action of *analyzing* is all about understanding the behaviors of the processes resulting from the model. In a concrete numerical setting it may mean comparing values for different parameters. For example, assume the parameter "twice" in Example 1.2.1 was replaced by α. Alternatively it may mean proving theorems about the behaviors. This is perhaps the difference between practice and research, although the distinction is vague.

A synonymous term that encompasses both computation and analysis is *performance analysis*. Associated with the behaviors of $x(\cdot)$ or $X(\cdot)$ we often have *performance measures*. Here are some typical performance measures that may be of interest. Some of these are qualitative and some are qunatiativie:

1. **Stability**
2. **Fixed point**
3. **Mean**
4. **Variance**
5. **Distribution**
6. **Hitting times**

Computation and analysis is typically is done with respect to performance measures such as the ones above or others.

1.2.4 Optimization

Making models is often so that we can optimize the underlying physical process. The idea is that trying the underlying process for all possible combinations is typically not possible, so optimizing the model is may be preferred. In a sense optimization may be viewed as a decoupled step from the above, since one can often formulate some optimization problem in terms of objects that come out of performance measures of the process.

1.2.5 Control

Optimization is typically considered to be something that we do over a slow time scale, while control implies intervening with the physical process continuously with a hope of making the behavior more suitable to requirements. The modeling type of action done here is the *design of the control law*. This in fact, yields a modified model, with modified behaviors.

Example 1.2.5. *We continue with the simple population growth example. Assume that culling is applied when ever the population reaches a certain level, d. In that case, individuals are removed bringing the population down to level c where $c < d$.*

This is a control policy. Here the aim of the control is obviously to keep the "population at bay". The values c and d are parameters of the control policy (also called the "control" or the "controller").

Exercise 1.2.6. *Repeat Exercise 1.2.3 with this policy where $c = 10$ and $d = 300$.*

Exercise 1.2.7. *Formulate some non-trivial optimization problem on the parameters of the control policy. For this you need to "make up some story" of costs etc...*

1.2.6 Our Scope

In this book we focus on quite specific processes. For the stochastic ones we carry out analysis (and show methods to do so) - but do not deal with control. For the deterministic ones we do both analysis and control. The reason for "getting more" out of the deterministic models is that they are in fact simpler. So why use stochastic models if we do not talk about control? Using them for performance measures can be quite fruitful and can perhaps give better models of the physical reality than the deterministic models (in some situations).

1.3 Application Examples

Moving away from the population growth example of the previous section, we now introduce four general examples that we will vaguely follow throughout the book. We discuss the underlying "physics" of these examples and will continue to refer to them in the chapters that follow.

1.3.1 An Inverted Pendulum on a Cart

Consider a cart fixed on train tracks on which there is a tall vertical rod above the cart, connected to the cart on a joint. The cart can move forward and backwards on the train tracks. The rod tends to fall to one of the sides – it has 180 degrees of movement.

For simplicity we assume that there is no friction for the cart on the train tracks and that there is no friction for the rod. That is there is no friction on the joint between the rod and the cart and there is no air friction when the rod falls down.

We assume there are two controlled motors in the system. The first can be used to apply force on the cart pushing it forward or backwards on the train tracks. The second can be used to apply a torque on the rod at the joint.

This idealized physical description is already a *physical model*. It is a matter of physical modeling to associate this model (perhaps after mild modifications or generalizations) to certain applications. Such applications may be a "Segway Machine" or the firing of a missile vertically up to the sky.

This physical model can be described by differential equations based on Newton's laws (we will do so later on). Such a mathematical model describes the physical system well and can then be used for simulation, computation, analysis, optimization and control.

It is with respect to this last use-case (control) that the inverted pendulum on a cart is so interesting. Indeed if forces are not applied through the motor and if the rod is not at rest in an angle of either 0, 90 or 180 degrees, then it will tend to fall down to the angles of 0 or 180. That is, it is unstable. Yet with proper "balancing" through the motors, the rod may be stabilized at 90 degrees. As we discuss control theory, we will

see how to do this and analyze this system further.

1.3.2 A Chemical Engineering Processes

Consider a cylindrical fluid tank containing water and a dissolved chemical in the water. Assume that there is a stirring propeller inside the tank that is stirring it well. The tank is fed by two input flows. One of pure water and one of water with the chemical dissolved in it. There is output flow from the tank at the bottom. It is known that the output flow rate is proportional to the square root of the height of the water level in the tank.

The system operator may control the incoming flow of pure water, the incoming flow of water with dissolved chemical, and the concentration of dissolved chemical coming in.

Two goals that the operator wants to achieve are:

1. Keep the fluid level in tank within bounds. I.e. not to let it underflow and not to let it overflow.

2. Maintain a constant (or almost constant) concentration of the chemical in the outgoing flow.

Here also we will see how such a model can be described and controlled well by means of linear control theory. Further, this model has some flavor of a *queueing model*. Queueing models play a central role in MAM.

1.3.3 A Manufacturing Line

Consider a manufacturing process in which items move from one operating station to the next until completion. Think of the items as cars in a car manufacturing plant. Frames arrive to the line from outside and then cars pass through stations one by one until they pass the last station and are fully assembled and ready. At each station assume there is one operator which serves the items that have arrived to it sequentially - one after the other. Thus, each station in isolation is in fact a queue of items waiting to be served. In practice there are often room limitations: most stations may only accommodate a finite number of items. If a station is full, the station "upstream to it" can not pass completed items down, etc.

Industrial engineers managing, optimizing and controlling such processes often try to minimize randomness and uncertainty in such processes, yet this is not always possible:

- Service stations break down occasionally, often at random durations.

- The arrivals of raw materials is not always controlled.

- There is variability in the service times of items at individual stations. Thus the output from one station to the next is a variable process also.

Besides the fact that variability plays a key role, this application example is further different from the previous two in that items are discrete. Compare this to the previous two applications where momentum, speed, concentration, fluid flows and volume are all purely continuous quantities.

A mathematical model based on MAM can be applied to this application example. Especially to each of the individual stations in isolation (aggregating the whole model using an approximation). Yet, if item processing durations are short enough and there is generally a non-negligible amount of items, then the process may also be amenable to control design based on LCT.

1.3.4 A Communication Router

A Communication router receives packets from n incoming sources and passes each to m outgoing destinations. Upon arrival of a packet it is known to which output port (destination) it should go, yet if that port is busy (because another packet is being transmitted on it) then the incoming packets needs to be queued in memory. In practice such systems sometimes work in discrete time enforced by the design of the router.

Here packet arrivals are random and bursty and it is often important to make models that capture the essential statistics of such arrival processes. This is handled well by MAM. Further, the queueing phenomena that occur are often different than those of the manufacturing line due to the high level of variability in packet arrivals.

Bibliographic Remarks

.

There are a few books focusing primarily on MAM. The first of these was [**?**] which was followed by [**?**]. A newer manuscript which gives a comprehensive treatment of methods and algorithms is [**?**]. Certain chapters of [**?**] also deal with MAM. Other MAM books are [**?**].

Exercises

1. Choose one of the four application examples appearing in Section 1.3 (Inverted Pendulum, Chemical Plant, Manufacturing Line, Communication Router). For this example do the following:

(a) Describe the application in your own words while stating the importance of having a mathematical model for this application. Use a figure if necessary. Your description should be half a page to two pages long.

(b) Suggest the flavor of the type of mathematical model (or models) that you would use to analyze, optimize and control this example. Justify your choice.

(c) Refer to the uses cases appearing in Section 1.2. Suggest how each of these applies to the application example and to the model.

(d) Consider the performance analysis measures described under the use case "computation and Analysis" in Section 1.2. How does each of these use cases apply to the application example and model that you selected?

Chapter 2

LTI Systems and Probability Distributions (7h)

Throughout this book we refer to the time functions that we analyze as *processes*, yet in this chapter it is better to use the term *signals* as to agree with classic systems theory (systems theory based on input–output relations of systems).

A *linear time invariant system* (LTI system) is an operator acting on signals (time functions) in some function class, where the operator adheres to both the *linearity property* and the *time invariance* property. An LTI system can be characterized by its *impulse response* or *step response*. These are the outputs of the system resulting from a delta function or a step function respectively. Instead of looking at the impulse response or step response an integral transform of one of these functions may be used.

A *probability distribution* is simply the *probabilistic law* of a *random variable*. It can be represented in terms of the *cumulative distribution function*,

$$F(t) := \mathbb{P}(X \leq t),$$

where X denotes a random variable. We concentrate on *non-negative random variables*.

Just like the impulse or step response of a system, a *probability distribution* may be represented by an integral transform. For example, the Laplace-Stieltjes Transform (LST) of a probability distribution $F(\cdot)$ is

$$\hat{F}(s) = \int_0^\infty e^{-st} dF(t).$$

In this chapter we describe both LTI systems and probability distributions and discuss some straight forward relationships between the two.

2.1 Signals

The term *signal* is essentially synonymous with a function, yet a possible difference is that a signal can be described by various different representations, each of which is a different function.

Signals may be of a *discrete time* type or a *continuous time* type. Although in practical applications these days, signals are often "digitized", for mathematical purposes we consider signals to be real (or complex). Signals may be either scalar or vector.

It is typical and often convenient to consider a signal through an integral transform (e.g. the Laplace transform) when the transform exists.

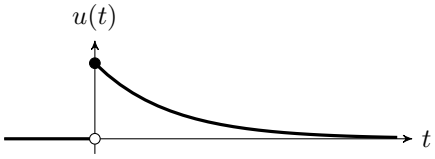
Example 2.1.1. *Consider the signal,*

$$u(t) = \begin{cases} 0, & t < 0, \\ e^{-t}, & 0 \le t. \end{cases}$$

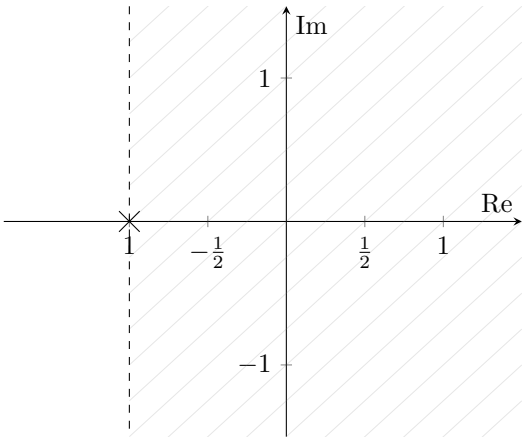
The Laplace transform of the signal is,

$$\hat{u}(s) = \int_0^\infty e^{-st} e^{-t} dt = \frac{1}{s+1}, \quad \textit{for } Re(s) > -1.$$

In this case, both $u(t)$ and $\hat{u}(s)$ represent the same signal. We often say that $u(t)$ is the time-domain *representation of the signal where as $\hat{u}(s)$ is the* frequency-domain *representation.*



The signal $u(t)$ of Example 2.1.1



A Pole Zero Plot, and Region of Convergence for the Laplace Transform

2.1.1 Operations on Signals

It is common to do operations on signals. Here are a few very common examples:

- $\tilde{u}(t) = \alpha_1 u_1(t) + \alpha_2 u_2(t)$: Add, subtract, scale or more generally take linear combinations.
- $\tilde{u}(t) = u(t - \tau)$: Translation. Shift forward in case $\tau > 0$ (delay) by τ.
- $\tilde{u}(t) = u(-t)$: Reverse time.
- $\tilde{u}(t) = u(\alpha t)$: Time scaling. Stretch when $0 < \alpha < 1$. Compress when $1 < \alpha$.
- $\tilde{u}(\ell) = u(\ell T)$: Sample to create a discrete time signal from a continuous time signal (sampling period is T).
- $\tilde{u}(t) = \sum_\ell u(\ell) K\left(\frac{t-\ell T}{T}\right)$, where $K(\cdot)$ is an *interpolation function*. I.e. it has the properties $K(0) = 1, K(\ell) = 0$ for other integers $\ell \neq 0$. This creates a continuous time signal, $\tilde{u}(\cdot)$ from a discrete time signal, $u(\cdot)$.

Exercise 2.1.2. *Find the $K(\cdot)$ that will do* linear interpolation*, i.e. connect the dots. Illustrate how this works on a small example.*

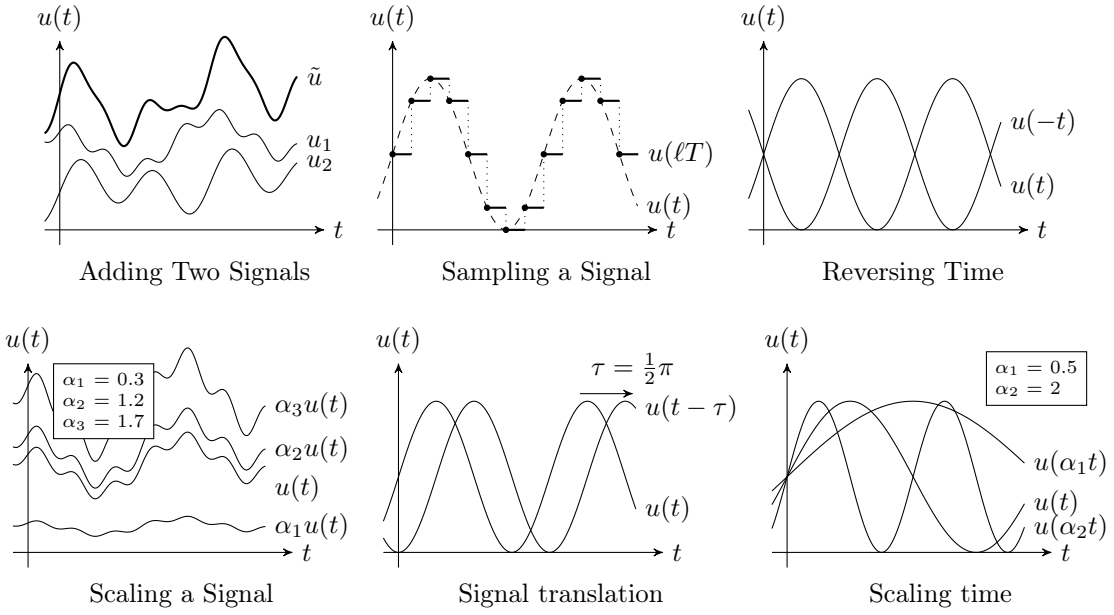


Figure 2.1: Operations on Signals.

2.1.2 Signal Spaces

It is common to consider *signal spaces* (function spaces). For example $\mathcal{L}^2$ is the space of all continuous time signals $\{u(t)\}$, such that $||u||_2 < \infty$. Here $||\cdot||_2$ is the usual $\mathcal{L}^2$ norm:

$$||u||_2 := \sqrt{\int_{-\infty}^{\infty} u(t)^2 dt}.$$

Other useful norms that we consider are the $\mathcal{L}^1$ norm, $||\cdot||_1$:

$$||u||_1 := \int_{-\infty}^{\infty} \big|u(t)\big| dt,$$

which induces the $\mathcal{L}^1$ space and the $\mathcal{L}^\infty$ space, induced by the $\mathcal{L}^\infty$ norm, $||\cdot||_\infty$ norm:

$$||u||_\infty = \sup_{t\in\mathbb{R}} |u(t)|.$$

Signals in $\mathcal{L}^\infty$ are bounded from above and below.

For discrete time signals, the space ℓ^2 is the space of all discrete time signals $\{u(\ell)\}$ (do not confuse our typical time index ℓ with the ℓ denoting the space) such that $||u||_2 < \infty$. In this case $||\cdot||_2$ is the usual ℓ^2 norm:

$$||u||_2 := \sqrt{\sum_{\ell=-\infty}^{\infty} u(\ell)^2}.$$

Similarly, the ℓ^1 and ℓ^∞ norms can be defined. Note that the above definitions of the norms are for real valued signals. In the complex valued cases replace $u(t)^2$ by $u(t)\overline{u(t)}$ (and similarly for discrete time). We don't deal much with complex valued signals.

Many other types of signals spaces can be considered. Other than talking about *bounded signals*, that is signals from $\mathcal{L}^\infty$ or ℓ^∞, we will not be too concerned with signal spaces in this book.

2.1.3 Generalized Signals

Besides the signals discussed above, we shall also be concerned with *generalized signals* (generalized functions). The archetypal such signal is the *delta function* (also called the *impulse*). In discrete time there is no need to consider it as a generalized function since this object is denoted by $\delta[\ell]$ (observe the square brackets) and is defined as:

$$\delta[\ell] := \begin{cases} 1 & \ell = 0, \\ 0 & \ell \neq 0. \end{cases}$$

In continuous time we are interested in an analog: a signal, $\{\delta(t)\}$ that is 0 everywhere except for at the time point 0 and satisfies,

$$\int_{-\infty}^{\infty} \delta(t)dt = 1.$$

Such a function does not exist in the normal sense, yet the mathematical object of a generalized function may be defined for this purpose (this is part of Schwartz's theory of distributions). More details and properties of the Delta function (and related generalized functions) are in the appendix. To understand the basics of LTI systems only a few basic properties need to be considered.

The main property of $\delta(t)$ that we need is that for any *test function*, $\phi(\cdot)$:

$$\int_{-\infty}^{\infty} \delta(t)\phi(t)dt = \phi(0).$$

2.2 Input Output LTI Systems - Definitions and Categorization

A *system* is a mapping of an input signal to an output signal. When the signals are scalars the system is called SISO (Single Input Single Output). When inputs are vectors and outputs are vectors the system is called MIMO (Multi Input Multi Output). Other combinations are MISO (not the soup) and SIMO. We concentrate on SISO systems in this chapter.

We denote input-output systems by $\mathcal{O}(\cdot)$. Formally these objects are operators on signal spaces. For example we may denote $\mathcal{O} : \mathcal{L}^2 \to \mathcal{L}^2$. Yet for our purposes this type of formalism is not necessary. As in the figure below, we typically denote the output of the system by $\{y(t)\}$. I.e.,

$$y(\cdot) = \mathcal{O}\big(u(\cdot)\big).$$

In most of the subsequent chapters, we will associate a *state* with the system (denoted by $\boldsymbol{x}(\cdot)$ or $X(\cdot)$) and sometimes ignore the input and the output. As described in the introductory chapter it is the state processes that are the main focus of this book. Yet in this chapter when we consider *input-output systems*, the notion of state still does not play a role.

In general it is **not true** that $y(t)$ is determined solely by $u(t)$, it can depend on $u(\cdot)$ at other time points. In the special case where the output at time t depends only on the input at time t we say the system is *memoryless*. I.e. for memoryless systems, the output at time t depends only on the input at time t. This means that there exists some function $g : \mathbb{R} \to \mathbb{R}$ such that $y(t) = g\big(u(t)\big)$. These systems are typically quite boring.

A system is *non-anticipating* (or *causal*) if the output at time t depends only on the inputs during times up to time t. This is defined formally by requiring that for all t_0,

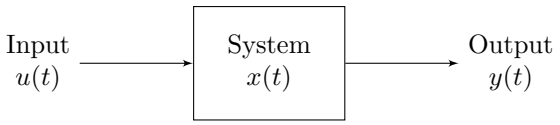


Figure 2.2: A system operates on an input signal $u(\cdot)$ to generate an output signal $y(\cdot) = \mathcal{O}\big(u(\cdot)\big)$. The system may have a *state*, $\{\boldsymbol{x}(t)\}$. Looking at state is not our focus now. The notation in the figure is for continuous time. Discrete time analogs ($u(\ell)$, $\boldsymbol{x}(\ell)$, $y(\ell)$) hold.

whenever the inputs u_1 and u_2 obey $u_1(t) = u_2(t)$ for all $t \leq t_0$, the corresponding outputs y_1 and $y_2(t)$ obey $y_1(t) = y_2(t)$ for all $t \leq t_0$.

A system is *time invariant* if its behaviour does not depend on the actual current time. To formally define this, let $y(t)$ be the output corresponding to $u(t)$. The system is *time invariant* if the output corresponding to $u(t-\tau)$ is $y(t-\tau)$, for any time shift τ.

A system is *linear* if the output corresponding to the input $\alpha_1 u_1(t) + \alpha_2 u_2(t)$ is $\alpha_1 y_1(t) + \alpha_2 y_2(t)$, where y_i is the corresponding input to u_i and α_i are arbitrary constants.

Exercise 2.2.1. *Prove that the linearity property generalises to inputs of the form* $\sum_{i=1}^{N} \alpha_i u_i(t)$.

Systems that are both *linear and time invariant* posses a variety of important properties. We abbreviate such systems with the acronym LTI. Such systems are extremely useful in both control and signal processing. The LTI systems appearing in control theory are typically casual while those of signal processing are sometimes not.

Exercise 2.2.2. *For discrete time input $u(\ell)$ define,*

$$y(\ell) = \frac{1}{N+M+1} \sum_{m=-M}^{N} \big(u(\ell+m)\big)^{\alpha+\beta\cos(\ell)}.$$

When $\alpha = 1$ and $\beta = 0$ this system is called a sliding window averager. *It is very useful and abundant in time-series analysis and related fields. Otherwise, there is not much practical meaning for the system other than the current exercise.*

Determine when the system is memoryless, casual, linear, time invariant based on the parameters N, M, α, β.

A final general notion of systems that we shall consider is *BIBO stability*. BIBO stands for bounded-input-bounded-output. A system is defined to be BIBO stable if whenever the input u satisfies $||u||_\infty < \infty$ then the output satisfies $||y||_\infty < \infty$. We will see in the section below that this property is well characterised for LTI systems.

2.3 LTI Systems - Relations to Convolutions

We shall now show how the operation of convolution naturally appears in LTI systems. It is recommended that the reader briefly reviews the appendix section on convolutions.

2.3.1 Discrete Time Systems

Consider the discrete time setting: $y(\cdot) = \mathcal{O}\big(u(\cdot)\big)$. Observe that we may represent the input signal, $\{u(\ell)\}$ as follows:

$$u(\ell) = \sum_{k=-\infty}^{\infty} \delta[\ell - k]u(k).$$

This is merely a representation of a discrete time signal $u(\ell)$ using the shifted (by ℓ) discrete delta function,

$$\delta[\ell - k] = \begin{cases} 1 & \ell = k, \\ 0 & \ell \neq k. \end{cases}$$

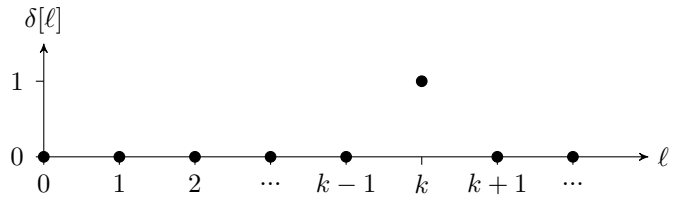


Figure 2.3: Discrete Delta Function $\delta[\ell - k]$

We now have,

$$y(\ell) = \mathcal{O}\Big(u(\ell)\Big) = \mathcal{O}\Big(\sum_{k=-\infty}^{\infty} \delta[\ell - k]u(k)\Big) = \sum_{k=-\infty}^{\infty} u(k)\mathcal{O}\Big(\delta[\ell - k]\Big).$$

Now denote,

$$h(\ell) := \mathcal{O}\big(\delta[\ell]\big).$$

Since the system is time invariant we have that $\mathcal{O}\big(\delta[\ell - k]\big) = h(\ell - k)$. So we have arrived at:

$$y(\ell) = \sum_{k=-\infty}^{\infty} u(k)h(\ell - k) = \big(u * h\big)(\ell).$$

This very nice fact shows that the output of LTI systems can in fact be described by the convolution of the input with the function $h(\cdot)$. This function deserves a special name: *impulse response*. We summarize the above in a theorem:

Theorem 2.3.1. *The output of a discrete time LTI-SISO system, $\mathcal{O}(\cdot)$, resulting from an input $u(\cdot)$ is the convolution of $u(\cdot)$ with the system's impulse response, defined as: $h(\cdot) := \mathcal{O}\big(u(\cdot)\big)$.*

2.3.2 Continuous Time Systems

For continuous time systems the same argument essentially follows. Here the *impulse response* is defined as,

$$h(\cdot) := \mathcal{O}\big(\delta(\cdot)\big).$$

Theorem 2.3.2. *The output of a continuous time LTI-SISO system, $\mathcal{O}(\cdot)$, resulting from an input $u(\cdot)$ is the convolution of $u(\cdot)$ with the system's impulse response.*

Proof. By the defining property of the Dirac delta function (see the appendix on generalized functions),

$$\int_{-\infty}^{\infty} \delta(\tau)u(t-\tau)d\tau = u(t-0) = u(t). \tag{2.1}$$

Using (F.2), we have,

$$\begin{aligned} y(t) &= \mathcal{O}\Big(u(t)\Big) = \mathcal{O}\Big(\int_{-\infty}^{\infty} \delta(t)u(t-\tau)d\tau\Big) = \mathcal{O}\Big(\int_{-\infty}^{\infty} u(\tau)\delta(t-\tau)d\tau\Big) \\ &= \int_{-\infty}^{\infty} u(\tau)\mathcal{O}\big(\delta(t-\tau)\big)d\tau = \int_{-\infty}^{\infty} u(\tau)h(t-\tau)d\tau = \big(u*h\big)(t). \end{aligned}$$

Observe that in the above we assume that the system is linear in the sense that,

$$\mathcal{O}\Big(\int_{-\infty}^{\infty} \alpha_s u_s ds\Big) = \int_{-\infty}^{\infty} \alpha_s \mathcal{O}\big(u_s\big)ds.$$

□

2.3.3 Characterisations based on the Impulse Response

The implications of Theorems 2.3.1 and 2.3.2 are that LTI SISO systems are fully characterized by their impulse response: Knowing the impulse response of $\mathcal{O}(\cdot)$, $h(\cdot)$, uniquely identifies $\mathcal{O}(\cdot)$. This is usefull since the operation of the whole system is summarized by one signal! This also means that to every signal there corresponds a system. So systems and signals are essentially the same thing.

Now based on the impulse response we may determine if an LTI system is memoryless, causal and BIBO-stable:

Exercise 2.3.3. *Show that an LTI system is memory less if and only if the impulse response has the form $h(t) = K\delta(t)$ for some constant scalar K.*

Exercise 2.3.4. *Show that an LTI system is causal if and only if $h(t) = 0$ for all $t < 0$.*

Exercise 2.3.5. *Consider the sliding window averager of exercise 2.2.2 with $\alpha = 1$ and $\beta = 0$. Find it's impulse response and find the parameters for which it is casual.*

Theorem 2.3.6. *An LTI system with impulse response $h(\cdot)$ is BIBO stable if and only if,*

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

Further if this holds then,

$$||y||_\infty \leq ||h||_1 \, ||u||_\infty, \tag{2.2}$$

for every bounded input.

Proof. The proof is for discrete-time (the continuous time case is analogous). Assume first that $||h||_1 < \infty$. To show the system is BIBO stable we need to show that if $||u||_\infty < \infty$ then $||y||_\infty < \infty$:

$$|y(\ell)| = \Big| \sum_{k=-\infty}^{\infty} h(\ell - k)u(k) \Big| \leq \sum_{k=-\infty}^{\infty} |h(\ell - k)| \, |u(k)| \leq \Big(\sum_{k=-\infty}^{\infty} |h(\ell - k)| \Big) ||u||_\infty$$

So,

$$||y||_\infty \leq ||h||_1 \, ||u||_\infty < \infty.$$

Now to prove that $||h||_1 < \infty$ is also a necessary condition. We choose the input,

$$u(\ell) = \operatorname{sign}\big(h(-\ell)\big).$$

So,

$$y(0) = \sum_{k=-\infty}^{\infty} h(0 - k)u(k) = \sum_{k=-\infty}^{\infty} |h(-k)| = ||h||_1.$$

Thus if $||h||_1 = \infty$ the output for this (bounded) input, $u(\cdot)$, is unbounded. Hence if $||h||_1 = \infty$ the system is not BIBO stable. Hence $||h||_1 < \infty$ is a necessary condition for BIBO stability. □

Exercise 2.3.7. *What input signal achieves equality in (2.2)?*

Exercise 2.3.8. *Prove the continuous time version of the above.*

Exercise 2.3.9. *Prove the above for signals that are in general complex valued.*

2.3.4 The Step Response

It is sometimes useful to represent systems by means of their *step response* instead of their impulse response. The step response is defined as follows:

$$H(t) := \int_{-\infty}^{t} h(\tau)d\tau, \qquad H(\ell) := \sum_{k=-\infty}^{\ell} h(k).$$

Knowing the impulse response we can get the step response by integration or summation (depending if the context is discrete or continuous time) and we can get the impulse response by,

$$h(t) = \frac{d}{dt}H(t), \qquad h(\ell) = H(\ell) - H(\ell - 1).$$

It should be noted that in many systems theory texts, $H(\cdot)$ is reserved for the transfer function (to be defined in the sequel). Yet in our context we choose to use the h, H notation so as to illustrate similarities with the f, F notation apparent in probability distributions.

Where does the name *step response* come from? Consider the input to the system: $u(t) = \mathbf{1}(t)$, *the unit-step*. Then by Theorem 2.3.2 the output is,

$$y(t) = \int_{-\infty}^{\infty} \mathbf{1}(t-\tau)h(\tau)d\tau = \int_{-\infty}^{t} h(\tau)d\tau = H(t).$$

2.4 Probability Distributions Generated by LTI Hitting Times

A *probability distribution function* of a non-negative random variable is a function,

$$F(\cdot) : \mathbb{R} \to [0, 1],$$

satisfying:

1. $F(t) = 0, \forall t \in (-\infty, 0)$.
2. $F(\cdot)$ is monotonic non-decreasing
3. $\lim_{t\to\infty} F(t) = 1$.

Denoting the random variable by X, the probabilistic meaning of $F(\cdot)$ is

$$F(t) = \mathbb{P}(X \leq t).$$

For example if X is a *uniform random variable* with support $[0, 1]$ we have,

$$F(t) = \begin{cases} 0, & t < 0, \\ t, & 0 \leq t \leq 1, \\ 1, & 1 < t. \end{cases} \tag{2.3}$$

2.4.1 The Inverse Probability Transform

If $F(\cdot)$ is both continuous and strictly increasing on $(0,\infty)$ then it corresponds to a random variable with support $[0,\infty)$ (it can get any value in this range) that is continuous on $(0,\infty)$. In this case, if $F(0) > 0$ we say that the random variable has an *atom at 0*. If $F(\cdot)$ is strictly increasing on $(0,\infty)$ the inverse function,

$$F^{-1} : [0,1] \to [0,\infty),$$

exists. We call this function the *inverse probability transform*. In this case we have the following:

Theorem 2.4.1. *Let $F(\cdot)$ be a probability distribution function of a nonnegative random variable that is strictly increasing on $(0,\infty)$ with inverse probability transform $F^{-1}(\cdot)$. Let U denote a uniform random variable with support $[0,1]$. Then the random variable,*

$$X = F^{-1}(U),$$

has distribution function $F(\cdot)$.

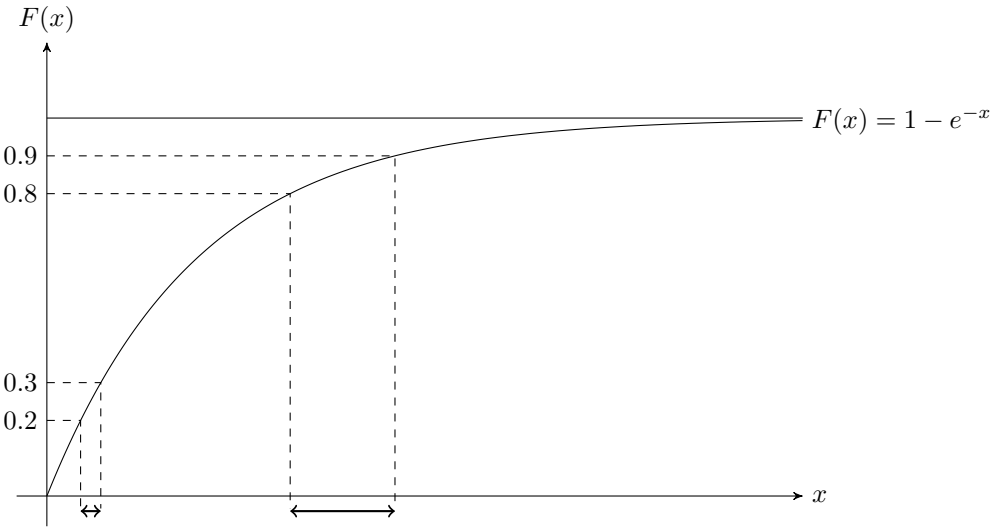


Figure 2.4: The CDF of an exponential distribution with unit mean. The inverse probability transform operates by generating uniform variables on the the $[0,1]$ subset of the y-axis.

Proof. Denote,

$$\tilde{F}(t) := \mathbb{P}(X \le t).$$

We wish to show that $\tilde{F}(\cdot) = F(\cdot)$:

$$\tilde{F}(t) = \mathbb{P}\big(F^{-1}(U) \le t\big) = \mathbb{P}\big(U \le F(t)\big) = F(t).$$

The second equality follows from the fact that $F^{-1}(\cdot)$ is monotonic. The third equality follows from the distribution function of uniform $[0,1]$ random variables, (2.3). $\square$

Note: The requirement that $F(\cdot)$ be that of a non-negative random variable with support $[0,\infty)$ can be easily relaxed, yet for the purpose of our presentation the statement above is preferred.

Exercise 2.4.2. *Let $F(\cdot)$ be an arbitrary distribution function. Formulate and prove and adapted version of this theorem.*

The inverse probability transform yields a recipe for generating random variables of arbitrary distribution using the *Monte Carlo Method*. All that is needed is a method to generate uniform random variables on $[0,1]$. The common method is the use of digital computers together with *pseudo-random number generators*.

Exercise 2.4.3. *Say you want to generate exponentially distributed random variables (see Section 2.4.5) with parameter (inverse of the mean), $\lambda > 0$. How would you do that given uniform random variables on [0,1]?*

Implement your method in computer software for $\lambda = 2$ and verity the mean and variance of your Monte-Carlo generated random variables. You can do this by generating 10^5 instances and taking sample mean and sample variance and then comparing to the theoretical desired values.

2.4.2 Hitting Times of LTI Step Responses

Consider now causal BIBO stable LTI systems in continuous time. Since the unit-step is a bounded signal it implies that the step response of such systems is bounded. Further since the system is causal we have that $H(0) = 0$. In such cases define the *step response support* to be the interval $[\underline{\mathrm{H}}, \overline{H}]$ where,

$$\underline{\mathrm{H}} := \inf\{H(t) : t \in [0,\infty)\}, \qquad \overline{H} := \sup\{H(t) : t \in [0,\infty)\}.$$

Consider now $x \in (\underline{\mathrm{H}}, \overline{H})$ and define,

$$\tau(x) := \inf\{t \geq 0 : H(t) = x\}.$$

We refer to this function as the *step response hitting time* of value x. We can now define a class of probability distributions associated with continuous time casual BIBO stable LTI systems:

Definition 2.4.4. *Consider a continuous time casual BIBO stable LTI system with step response support $[\underline{H}, \overline{H}]$. Let U be a uniformly distributed random variable $[\underline{H}, \overline{H}]$ and define,*

$$F(t) = \mathbb{P}\big(\tau(U) \leq t\big).$$

Then $F(\cdot)$ is called an LTI step response hitting time distribution.

2.4.3 Step Responses That are Distribution Functions

Consider now causal LTI systems whose step response, $H(\cdot)$ satisfies the properties of a distribution function. In this case the step response support is $[0, 1]$ and the *LTI step response hitting time distribution*, $F(\cdot)$, defined in 2.4.4 in fact equals the step response. That is $F(\cdot) = H(\cdot)$. We summarize this idea in the theorem below:

Theorem 2.4.5. *Consider an LTI system whose step response $H(\cdot)$ satisfies the properties of a distribution function. Assume U is a uniform $[0, 1]$ random variable. Assume this system is subject to input $u(t) = \mathbf{1}(t)$ and X denotes the time at which the output $y(t)$ hits U. Then X is distributed as $H(\cdot)$.*

Proof. The result follows from the definitions above and Theorem 2.4.1. □

Further note that since the impulse response is the derivative of the step response, in the case of the theorem above it plays the role of the density of the random variable X.

2.4.4 The Transform of the Probability Distribution

Given a probability distribution function of a non-negative random variable, $F(x)$, the Laplace-Stieltjes transform (LST) associated with the distribution function is:

$$\hat{f}(s) = \int_0^\infty e^{-st} dF(t). \tag{2.4}$$

If $F(\cdot)$ is absolutely continuous, with density $f(\cdot)$, i.e.,

$$F(t) = \int_0^t f(s)ds,$$

then the LST is simply the Laplace transform of the density:

$$\hat{f}(s) = \int_0^\infty e^{-st} f(t)dt.$$

If $F(\cdot)$ is continuous on $(0, \infty)$ yet has an atom at zero ($\alpha_0 := F(0) > 0$), then,

$$\hat{f}(s) = \alpha_0 + \int_0^\infty e^{-st} f(t)dt. \tag{2.5}$$

The use of the LST in (2.4) generalizes this case yet for our purposes in this book (2.5) suffices. Further details about integral transforms are in the appendix.

The use of Laplace transforms in applied probability is abundant primarily due to the fact that the Laplace transform of a sum of independent random variables is the product of their Laplace transforms (see the appendix). Other nice features include the fact that

moments can be easily obtained from the Laplace transform as well asymptotic properties of distributions.

Related to the LST $\hat{f}(s)$ of a random variable X we also have the Moment Generating Function (MGF): $\phi_1(s) = \mathbb{E}[e^{sX}]$, the characteristic function $\phi_2(\omega) = \mathbb{E}[e^{i\,\omega\,X}]$ and the probability generating function (PGF) $\phi_3(z) = \mathbb{E}[z^X]$. For a given distribution, $F(\cdot)$, the function $\phi_1(\cdot)$, $\phi_2(\cdot)$ and $\phi_3(\cdot)$ are intimately related to each other and to the LST, $\hat{f}(\cdot)$. Each is common and useful in a slightly different context. In this text, we mostly focus on the LST.

Exercise 2.4.6. *Show that for a random variable X, with LST, $\hat{f}(\cdot)$, the k'th moment satisfies:*

$$\mathbb{E}[X^k] = (-1)^k \frac{d^k}{ds^k} \hat{f}(s)\big|_{s=0}.$$

2.4.5 The Exponential Distribution (and System)

A random variable T has an exponential distribution with a parameter $\lambda > 0$, denoted by $T \sim \exp(\lambda)$, if its distribution function is

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t}, & \text{for } t \geq 0, \\ 0, & \text{for } t < 0. \end{cases}$$

It follows that the probability density function of T is

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, & \text{for } t \geq 0, \\ 0, & \text{for } t < 0, \end{cases}$$

and the LST is,

$$\hat{f}_T(s) = \frac{1}{\lambda + s}, \qquad Re(s) > -\lambda.$$

Exercise 2.4.7. *Use the LST to show that the mean and variance of T are $1/\lambda$ and $1/\lambda^2$ respectively.*

We digress momentarily to discuss hazard rates. For an absolutely continuous random variable $X \geq 0$ with distribution function F and probability density function f, the *hazard* (or *failure*) *rate function* is given by

$$r(t) = \frac{f(t)}{1 - F(t)}.$$

We can think of $r(t)\delta t$ as the probability that $X \in (t, t + \delta t]$ conditional on $X > t$.

The value $r(t)$ of the hazard function at the point t is thus the rate that the lifetime expires in the next short time after t given that it has survived that long. It is a common description of probability distributions in the field of reliability analysis.

Exercise 2.4.8. *Show that give a hazard rate function, $r(t)$, the CDF can be reconstructed by:*

$$F(t) = 1 - e^{-\int_0^t r(u)\,du}.$$

Exercise 2.4.9. *Show that for an exponential random variable $T \sim \exp(\lambda)$, the hazard rate is constant: $r_T(t) = \lambda$.*

Intimately related to the constant hazard rate property is the *Lack of memory* (or *memoryless* – yet do not confuse with the same term for LTI systems) property which characterises exponential random variables:

$$\mathbb{P}(T > t + s | T > t) = \mathbb{P}(T > s).$$

Exercise 2.4.10. *Describe in word the meaning of the memoryless property of exponential random variables, treating T as the lifetime of a component in a working device.*

Exercise 2.4.11. *Prove that exponential random variables are memoryless. Further sketch a proof showing that any continuous random variable with support $[0, \infty)$ which is memoryless must be exponential. In doing so, assume that the only (function) solution to $g(s + t) = g(s)g(t)$ is $g(u) = e^{a\,u}$ for some a.*

We are often required to consider a "race" between several exponential random variables. For example, consider the case in reliability analysis where a working device is composed of several components whose lifetimes are the random variables, $T_1, \ldots, T_k$ and the device requires all components to be operating (not failing) for it to be operating. In this case lifetime of such a device has value $M := \min(T_1, \ldots, T_k)$. Further, the index of the first component that fails is,

$$I := \big\{ i \in \{1, \ldots, k\} \ : \ T_i = M \big\}.$$

Note that the set I is a contains a single element w.p. 1; it is the element $\in\in \{1, \ldots, k\}$ that "won" the race.

Often such lifetime random variables are taken to be of constant hazard rate (exponential) and assumed independent. In this case, the following is very userful:

Theorem 2.4.12. *In the exponential race denote $\Lambda = \lambda_1 + \ldots + \lambda_k$ we have:*

1. *$M \sim \exp(\Lambda)$.*
2. *I is a discrete random variables on $\{1, \ldots, k\}$ with $\mathbb{P}(I = i) = \lambda_i / \Lambda$.*
3. *M and I are independent.*

We will find this theorem extremely useful for Continuous Time Markov Chains (CTMC) as well (presented in the next Chapter).

Exercise 2.4.13. *Prove the above for $k = 2$.*

Exercise 2.4.14. *Use induction to carry the above proof for arbitrary k.*

2.5 LTI Systems - Transfer Functions

Having seen Laplace transforms in probability distributions, let us now look at their role in LTI systems.

2.5.1 Response to Sinusoidal Inputs

It is now useful to consider our LTI SISO systems as operating on complex valued signals. Consider now an input of the form $u(t) = e^{-st}$ where $s \in \mathbb{C}$. We shall denote $s = \sigma + i\omega$, i.e. $\sigma = Re(s)$ and $\omega = Im(s)$. We now have,

$$y(t) = \int_{-\infty}^{\infty} h(\tau)u(t-\tau)d\tau = \int_{-\infty}^{\infty} h(\tau)e^{s(t-\tau)}d\tau = \Big(\int_{-\infty}^{\infty} h(\tau)e^{-s\tau}d\tau\Big)e^{st}.$$

Denoting $\hat{h}(s) = \int_{-\infty}^{\infty} h(\tau)e^{-s\tau}d\tau$ we found that for exponential input, e^{st}, the output is simply a multiplication by the complex constant (with respect to t), $\hat{h}(s)$:

$$y(t) = \hat{h}(s)e^{st}.$$

Observe that $\hat{h}(s)$ is exactly the Laplace transform of the impulse response. It is central to control and system theory and deserves a name: the *transfer function*. Thus he transfer function tells us by which scalar (complex scalar) we multiply inputs of the form $u(t) = e^{st}$.

When the input signal under consideration has real part $\sigma = 0$, i.e. $u(t) = e^{i\omega t}$ then the output can still be represented in terms of the transfer function:

$$y(t) = \hat{h}(i\omega)e^{iwt}$$

In this case $y(t)$ is referred to as the *frequency response* of the *harmonic* input $e^{i\omega t}$ at frequency ω. And further, $\hat{\hat{h}}(\omega) := \hat{h}(i\omega)$ is called the *Fourier transform* of the impulse response at frequency ω. Note that both the Fourier and Laplace transform are referred to in practice as the transfer function.

For discrete time systems an analog of the Laplace transform is the Z-transform:

$$\hat{f}(z) = \sum_{\ell=-\infty}^{\infty} f(z)z^{-\ell}.$$

2.5.2 The Action of the Transfer Function

Since we have seen that $y(t) = \big(u * h\big)(t)$, we can use the convolution property of transforms to obtain,

$$\hat{y}(s) = \hat{u}(s)\hat{h}(s),$$

where in continuous time, $\hat{u}(\cdot)$ and $\hat{y}(\cdot)$ are the Laplace transforms of the input and output respectively. In discrete time they are the z-transforms.

Note that the Laplace transform of $\delta(t)$ is a constant 1 and this agrees with the above equation.

Hence LTI systems have the attractive property that the action of the system on an input signal $u(\cdot)$ may be easily viewed (in the frequency domain) by multiplication of the transfer function.

Consider now the integrator LTI system,

$$y(t) = \int_0^t u(s)\, ds, \qquad \text{or} \qquad y(\ell) = \sum_{k=0}^{\ell} u(k).$$

The impulse response of these systems is $h(t) = \mathbf{1}(t)$, where if we consider discrete time and replace t by ℓ the meaning of $\mathbf{1}(\ell)$ is that it is defined only on integers.

The transfer function of these systems is,

$$\hat{h}(s) = \int_0^\infty e^{-st} dt = \frac{1}{s}, \quad 0 < Re(s),$$

for continuous time and

$$\hat{h}(z) = \sum_{k=0}^{\infty} \big(\frac{1}{z}\big)^k = \frac{1}{1 - z^{-1}} = \frac{z}{1 - z}, \quad 1 < |z|,$$

for discrete time.

Exercise 2.5.1. *Verify the above calculations.*

Consider now an arbitrary causal LTI system, $\mathcal{O}_a$ and denote the integrator system by $\mathcal{O}_I$. Then the system,

$$y(t) = \mathcal{O}_I\Big(\mathcal{O}_a\big(u(t)\big)\Big),$$

is the system that first applies $\mathcal{O}_a$ and the integrates the output. If the impulse response of $\mathcal{O}_a$ is $h_a(\cdot)$ then the step response is $\mathcal{O}_I\Big(\mathcal{O}_a\big(h_a(\cdot)\big)\Big)$. Now since we know the transfer function of the integrators we have:

Theorem 2.5.2. *For a system with transfer function $\hat{h}(s)$ (continuous time system) or $\hat{h}(z)$ (discrete time system), the respective functions of the step response are:*

$$\frac{1}{s}\hat{h}(s), \qquad \textit{or} \qquad \frac{1}{1 - z^{-1}}\hat{h}(z).$$

It is also sometimes useful to represent the transfer function at a (complex) frequency s, by the ratio:

$$\hat{h}(s) = \frac{\hat{y}(s)}{\hat{u}(s)}. \tag{2.6}$$

This representation is often useful when modelling a system as a differential equation. For example, consider a physical system where the output $y(\cdot)$ is related to the input $u(\cdot)$ by,

$$\dot{y}(t) + ay(t) = Cu(t),$$

for some constant a and initial conditions (say at time $t = 0$) $y(0) = 0$. Applying the Laplace transform and noting that for a function $f(t)$ the Laplace transform of the derivative $\dot{f}(t)$ is $s\hat{f}(s) - f(0)$ we get,

$$s\hat{y}(s) + a\hat{y}(s) = \frac{1}{C}\hat{u}(s),$$

or,

$$\frac{\hat{y}(s)}{\hat{u}(s)} = \frac{C}{s+a},$$

hence the transfer function for this system, is

$$\frac{\hat{y}(s)}{\hat{u}(s)} = \frac{C}{s+a},$$

which upon inversion (e.g. from a Laplace transform table) yields impulse response,

$$h(t) = Ce^{-at}.$$

Hence for $a > 0$ and $C = a$, this system is equivalent to an exponential distribution.

Similar transforms also hold for higher order (linear) differential equations. In basic control theory, this allows to model physical input output systems by differential equations and then almost "read out" the transfer function.

2.5.3 Joint Configurations of LTI SISO Systems

See Figures 2.5 and **??** for basic combinations that may be performed with systems.

One of the virtues of using transfer functions instead of convolutions with impulse responses, is that such a representation allows us to look at the LTI system resulting from a control feedback loop:

Much of classic control theory deals with the design and calibration of LTI systems, $\hat{g}_1$ and $\hat{g}_2$ placed in a configuration as in Figure **??**, supporting feedback to the *plant* $\hat{p}(s)$. The whole system relating output y to input *reference* r is then also LTI and may be analyzed in the frequency (or 's') domain easily.

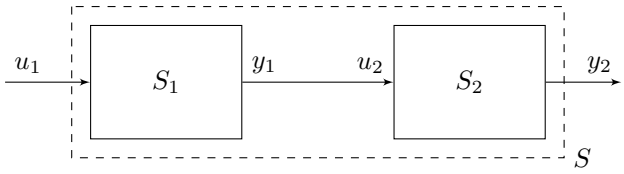


Figure 2.5: Two systems in series

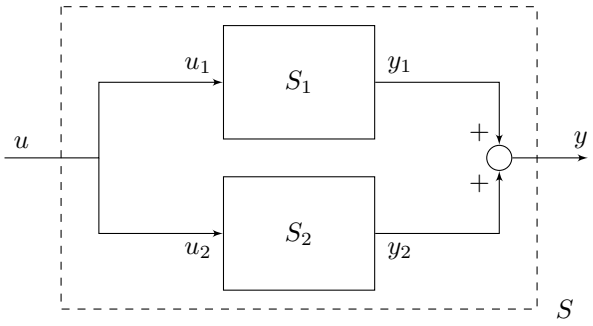


Figure 2.6: Two systems in parallel

The idea is to find the $\hat{h}$ that satisfies,

$$\hat{y}(s) = \hat{r}(s)\hat{h}(s).$$

This can be done easily:

$$\hat{y}(s) = \hat{u}(s)\hat{p}(s) = \hat{e}(s)\hat{g}_1(s)\hat{p}(s) = \big(\hat{r}(s) - \hat{y}_m(s)\big)\hat{g}_1(s)\hat{p}(s) = \big(\hat{r}(s) - \hat{g}_2(s)\hat{y}(s)\big)\hat{g}_1(s)\hat{p}(s).$$

Solving for $\hat{y}(s)$ we have,

$$\hat{y}(s) = \hat{r}(s)\frac{\hat{g}_1(s)\hat{p}(s)}{1 + \hat{g}_2(s)\hat{g}_1(s)\hat{p}(s)}.$$

Hence the feedback system is:

$$\tilde{h}(s) = \frac{\hat{g}_1(s)\hat{p}(s)}{1 + \hat{g}_2(s)\hat{g}_1(s)\hat{p}(s)}.$$

Exercise 2.5.3. *What would be the feedback system if there was* positive feedback *instead of negative. I.e. if the circle in the figure would have a '+' instead of '-'?*

Studies of feedback loop of this type constitute classic engineering control and are not the focus of our book. Yet for illustration we show the action of a PID (proportional – integral – derivative) controller on the inverted pendulum.

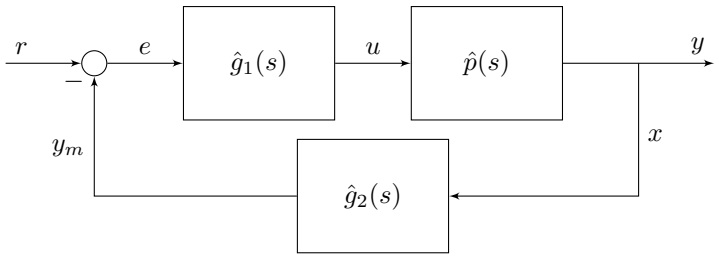


Figure 2.7: A *plant*, $\hat{p}(s)$ is controlled by the blocks $\hat{g}_1(s)$ and $\hat{g}_2(s)$ they are both optional (i.e. may be set to be some constant K or even 1).

2.6 Probability Distributions with Rational Laplace-Stieltjes Transforms

In the next chapter we will encounter probability distributions whose LST is a rational function. It will also be apparent that many LTI systems (those having a finite state space representation) have a rational $\hat{h}(\cdot)$. Having made a connection between probability distributions and LTI systems in Theorem 2.4.5, we will want to view such probability distributions as step response outputs of corresponding LTI systems.

For now, it is good to get acquainted with a few examples of such probability distributions and their LSTs:

Exercise 2.6.1. *What is the LST of an exponential distribution with parameter λ?*

Exercise 2.6.2. *Calculate the LST of a sum of n independent random exponential random variables, each with parameter λ. This is a special case of the Gamma distribution and is sometimes called an Erlang distribution.*

Exercise 2.6.3. *Consider n exponential random variables, $X_1, \ldots, X_n$ where the parameter of the i'th variable is λ_i. Let $p_1, \ldots, p_n$ be a sequence of positive values such that, $\sum_{i=1}^{n} p_i = 1$. Let Z be a random variable that equals X_i with probability p_i. Calculate the LST of Z. Such a "mixture" of exponential random variables is sometimes called an* hyper-exponential random variable.

Bibliographic Remarks

Exercises

1. Consider the function $f(t) = e^{at} + e^{bt}$ with $a, b, t \in \mathbb{R}$.
   (a) Find the Laplace transform of $f(\cdot)$.
   (b) Find the Laplace transform of $g_1(t) := \frac{d}{dt} f(t)$

(c) Find the Laplace transform of $g_2(t) := \int_0^t f(\tau)d\tau$

2. Prove that the Laplace transform of the convolution of two functions is the product of the Laplace transforms of the individual functions.

3. Consider Theorem 2.10 about BIBO stability. Prove this theorem for discrete time complex valued signals.

4. Carry out exercise 2.16 from Section 2.4.

5. Consider the differential equation:

$$\dot{y}(t) + ay(t) = u(t), \qquad y(0) = 0.$$

Treat the differential equation as a system, $y(\cdot) = \mathcal{O}\big(u(\cdot)\big)$.

(a) Is it an LTI system?

(b) If so, find the system's transfer function.

(c) Assume the system is a *plant* controlled in feedback as described in Section 2.5, with $g_1(s) = 1$ and $g_2(s) = K$ for some constant K. Plot (using software) the step response of the resulting closed loop system for $a = 1$ and for various values of K (you choose the values).

6. Consider a sequence of n systems in tandem where the output of one system is input to the next. Assume each of the systems has the impulse response $h(t) = e^{-t}\mathbf{1}(t)$. As input to the first system take $u(t) = h(t)$.

(a) What is the output from this sequence of systems? I.e. find $y(t)$, such that,

$$y(\cdot) = \mathcal{O}\big(\mathcal{O}\big(\mathcal{O}\big(\ldots\ldots\ldots\mathcal{O}\big(u(t)\big)\ldots\big)\big)\big),$$

such that the composition is repeated n times.

(b) Relate your result to Exercise 2.21 in Section 2.6.

(c) Assume n grows large, what can you say about the output of the sequence of systems? (Hint: Consider the Central Limit Theorem).

Chapter 3

Linear Dynamical Systems and Markov Chains (13h)

In Chapter 1 we introduced four basic types of processes. These included the continuous time processes:

1. $\{\boldsymbol{x}(t)\}$ with $t \in \mathbb{R}$ and $\boldsymbol{x}(t) \in \mathbb{R}^n$.
2. $\{X(t)\}$ with $t \in \mathbb{R}$ and $X(t) \in \mathcal{S}$, where $\mathcal{S}$ is some countable (finite or infinite set).

As well as their discrete time counter-parts:

3. $\{\boldsymbol{x}(\ell)\}$ with $\ell \in \mathbb{Z}$ and $\boldsymbol{x}(\ell) \in \mathbb{R}^n$.
4. $\{X(\ell)\}$ with $\ell \in \mathbb{Z}$ and $X(\ell) \in \mathcal{S}$, where $\mathcal{S}$ is some countable (finite or infinite set).

We typically take the continuous valued processes (1) and (3) to be deterministic and the discrete valued processes (2) and (4) to be stochastic.

In this chapter we introduce the dynamics of the most fundamental classes of such processes. In the deterministic case we introduce the behaviors associated with linear dynamics. These types of processes are defined by means of linear differential or difference equations. In the stochastic case we introduce the behaviors associated with the Markovian property. In such cases we introduce the processes as Markov chains in continuous or discrete time.

In the previous chapter we treated systems as input-output relationships, generally ignoring the notion of their *state*. This chapter differs in the sense that it is almost all about state. We now treat the values of the processes as a *state* of a system.

After introducing the behaviors associated with linear dynamical systems and Markov chains, we move on to introduce some of the basic objects that will appear in the

continuation of the book. These include (A, B, C, D) linear input-output systems. Such objects combine the notion of state with input and output. Further we show matrix exponential probability distributions which are closely related to (A, B, C, D) systems. We close with phase type distributions which are a special case of matrix exponential probability distributions that is defined by means of a Markov chain.

As a note, the reader should observe that most of the processes introduced in this chapter (and in the remainder of the book) are *time invariant*. This concept was defined in terms of SISO LTI systems in Chapter 2. In the more general setting it implies that the behavior of the process is not influenced by the current time. Such processes are some times called *time homogenous* or *stationary*. Yet we caution the reader about the use of the term "stationary" since it has a different meaning in different contexts.

3.1 Linear Dynamical Systems

We now consider deterministic processes of the form:

1. $\{\boldsymbol{x}(t)\}$ with $t \in \mathbb{R}_+$ and $\boldsymbol{x}(t) \in \mathbb{R}^n$.

3. $\{\boldsymbol{x}(\ell)\}$ with $\ell \in \mathbb{Z}_+$ and $\boldsymbol{x}(\ell) \in \mathbb{R}^n$.

Observe the slight difference from our previous definition: We consider the time index as starting at 0, i.e. $\mathbb{R}_+$ is the set of nonnegative reals and similarly for $\mathbb{Z}_+$. This is useful since we will describe some initial value.

The standard way to describe the behavior of such processes is to suggest some Lipschitz continuous $f : \mathbb{R}^n \to \mathbb{R}^n$ and set:

$$\dot{\boldsymbol{x}}(t) = f\big(\boldsymbol{x}(t)\big) \quad \text{or} \quad \boldsymbol{x}(\ell+1) = f\big(\boldsymbol{x}(\ell)\big), \tag{3.1}$$

together with a specified *initial value* $\boldsymbol{x}(0) = \boldsymbol{x}_0$. The continuous time or discrete time equation (3.1) together with an initial value is sometimes referred to as an *initial value problem*.

Such processes are generally referred to as *autonomous dynamical systems*. In the dynamical system context, the use of the phrase "autonomous" is due to the fact that the evolution does not depend on time (as opposed to $\dot{\boldsymbol{x}}(t) = f\big(t, \boldsymbol{x}(t)\big)$ for example). Observe also that the use of the phrase "system" here is not in the input-output context used in Chapter 2. Rather the system is essentially the process $\boldsymbol{x}(\cdot)$ and its behaviors.

An alternative to looking at the differential/difference equation occurring in (3.1) is to look at the integral/summation version:

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 + \int_0^t f\big(\boldsymbol{x}(s)\big)ds \quad \text{or} \quad \boldsymbol{x}(\ell) = \boldsymbol{x}_0 + \sum_{k=0}^{\ell-1} \Big(f\big(\boldsymbol{x}(k)\big) - \boldsymbol{x}(k)\Big).$$

Some of the theory of dynamical systems (and differential equations) deals with the existence and uniqueness of the continuous time system appearing in (3.1) (in the discrete time setting there are no such issues). To illustrate possible uniqueness problems, consider the following:

Example 3.1.1. *Take* $\dot{\boldsymbol{x}}(t) = \boldsymbol{x}(t)^{1/3}$, $\boldsymbol{x}(0) = 0$, *then there are at least two solutions:*

$$\boldsymbol{x}(t) = 0 \quad \textit{and} \quad \boldsymbol{x}(t) = \Big(\frac{4}{9}\, t\Big)^{3/2}.$$

We do not consider uniqueness and existence issues any further since our interest is in the special case:

$$f(\boldsymbol{x}) = A\boldsymbol{x},$$

for $A \in \mathbb{R}^{n\times n}$. That is, we consider *linear dynamical systems* of the form:

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) \quad \text{or} \quad \boldsymbol{x}(\ell+1) = A\boldsymbol{x}(\ell), \tag{3.2}$$

together with,

$$\boldsymbol{x}(0) = \boldsymbol{x}_0. \tag{3.3}$$

For these systems, uniqueness and existence is not an issue:

Theorem 3.1.2. *For any* $\boldsymbol{x}_0 \in \mathbb{R}^n$ *and* $A \in \mathbb{R}^{n\times n}$ *there exists a unique* $\boldsymbol{x}(\cdot)$ *satisfying* (3.2) *and* (3.3).

The proof for the discrete time case is immediate. We do not prove this result from first principles for the continuous time case, yet rather construct the unique solution in the sequel.

We now show two generic examples that bear significant importance in their own right. Further application examples appear in the exercises of this chapter.

Example 3.1.3. Linearization around an equilibrium point: *Consider a general (non-linear) dynamical system,*

$$\dot{\boldsymbol{x}}(t) = f\big(\boldsymbol{x}(t)\big),$$

where $f(\cdot)$ *is Lipschitz continuous. An* equilibrium point *of the system is a point* $\overline{\boldsymbol{x}} \in \mathbb{R}^n$ *such that,*

$$f\big(\overline{\boldsymbol{x}}\big) = \mathbf{0}.$$

Taking the Taylor series expansion of $f(\cdot)$ *at the point* $\overline{\boldsymbol{x}}$, *we have,*

$$f(\boldsymbol{x}) = f(\overline{\boldsymbol{x}}) + J(\overline{\boldsymbol{x}})(\boldsymbol{x} - \overline{\boldsymbol{x}}) + o\big(||\boldsymbol{x} - \overline{\boldsymbol{x}}||\big).$$

where $J(\cdot)$ *is the Jacobian matrix of* $f(\cdot)$. *We can then analyze the linear dynamical system,*

$$\dot{\boldsymbol{y}}(t) = A\,\boldsymbol{y}(t), \quad \textit{with} \quad A = J(\overline{\boldsymbol{x}}),$$

with initial value, $\boldsymbol{y}(0) = \boldsymbol{x}_0 - \overline{\boldsymbol{x}}$. *Then* $\{\boldsymbol{y}(t)\}$ *approximates* $\{\boldsymbol{x}(t) - \overline{\boldsymbol{x}}\}$ *at the vicinity of* $\mathbf{0}$.

Example 3.1.4. Higher order derivatives: *Consider the linear, autonomous, homogeneous ordinary scalar differential equation of order n:*

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \ldots + a_1 y^{(1)}(t) + a_0 y(t) = 0. \tag{3.4}$$

Denote now,

$$x_1(\cdot) := y(\cdot),\ \ x_2(\cdot) := y^{(1)}(\cdot),\ \ \ldots\ \ x_{n-1}(\cdot) := y^{(n-2)}(\cdot),\ \ x_n(\cdot) := y^{(n-1)}(\cdot),$$

and consider the autonomous system,

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \vdots \\ \dot{x}_{n-1}(t) \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & \cdots & -a_{n-1} & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_{n-1}(t) \\ x_n(t) \end{bmatrix}.$$

Then it is clear that solutions of the n'th dimensional system also satisfy (3.4)*. Note that the above matrix is called a* companion matrix *associated with $(a_0, \ldots, a_{n-1})$.*
The idea of maintaining higher order derivatives as part of the state also comes up naturally in the modeling use case. For example, it may be very natural in certain cases to have the state record the location, speed and acceleration of a physical object.

3.1.1 Example Models

Example 3.1.5. *Consider the following publication scenario for an academic researcher: Each year, each published paper of the researcher yields one new research direction which results in a new submission. Further each submitted paper becomes published.*
Let $x_1(\ell)$ and $x_2(\ell)$ denote the number of submitted and published papers of the academic in her ℓ'th year of research respectively. Then,

$$x_1(\ell) = x_2(\ell - 1), \qquad \textit{and} \qquad x_2(\ell) = x_1(\ell - 1) + x_2(\ell - 1).$$

or,

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

3.1.2 Finding the trajectory

Given a system (3.2)-(3.3), one way of finding the behavior of $\boldsymbol{x}(\cdot)$ is by successive iterations in the discrete time case or by approximation methods in the continuous time

case (we do not cover such methods here, see the bibliographic remarks at the end of the chapter). Alternatively (and often preferably) we would like to find a more insightful mathematical description of the solution. In the discrete time case this is elementary:

$$\boldsymbol{x}(\ell) = A\boldsymbol{x}(\ell-1) = AA\boldsymbol{x}(\ell-2) = A^3\boldsymbol{x}(\ell-3) = \ldots = A^\ell\boldsymbol{x}(\ell-\ell) = A^\ell\boldsymbol{x}_0.$$

Hence,

$$\boldsymbol{x}(\ell) = A^\ell\boldsymbol{x}_0 \tag{3.5}$$

In the continuous time case we need to introduce the continuous analog of the *matrix power* A^ℓ. We call this object the *matrix exponential* and denote it by e^{At}. It is formally constructed below.

Picard Iterations and Matrix Exponential

Given a continuous initial value problem (generally time varying):

$$\dot{\boldsymbol{x}}(t) = f\big(t, \boldsymbol{x}(t)\big), \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0,$$

a *Picard iteration sequence* is a sequence of functions constructed as follows:

$$\begin{aligned} \boldsymbol{\phi}_0(t) &= \boldsymbol{x}_0 \\ \boldsymbol{\phi}_{m+1}(t) &= \boldsymbol{x}_0 + \int_0^t f\big(s, \boldsymbol{\phi}_m(s)\big)ds, \quad m = 0, 1, 2, 3, \ldots \end{aligned}$$

It can be shown that if $f(\cdot)$ satisfies some Lipschitz conditions then the successive approximations $\boldsymbol{\phi}_m(\cdot)$, $m = 0, 1, 2, \ldots$ exist, are continuous and converge uniformly as $m \to \infty$ to the unique solution which we denote $\boldsymbol{\phi}(\cdot)$. That is, for every $\epsilon > 0$ there exists an N such that for all t in the specified domain,

$$||\boldsymbol{\phi}(t) - \boldsymbol{\phi}_m(t)|| < \epsilon,$$

whenever $m > N$.

To illustrate this, It is useful to briefly consider the time-dependent (non-autonomous) system (specified at initial time t_0, not necessarily 0):

$$\dot{\boldsymbol{x}}(t) = A(t)\boldsymbol{x}(t), \qquad \boldsymbol{x}(t_0) = \boldsymbol{x}_0. \tag{3.6}$$

By taking successive Picard iterations, it is observed that the solution of the above system is:

$$\boldsymbol{\phi}(t) = \Phi(t, t_0)\boldsymbol{x}_0, \tag{3.7}$$

where the *state transition* matrix $\Phi(t, t_0)$ is defined as follows:

$$\begin{aligned}\Phi(t,t_0) \\ := \; & I + \int_{t_0}^{t} A(s_1)ds_1 + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)ds_2\, ds_1 + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)\int_{t_0}^{s_2} A(s_3)ds_3\, ds_2\, ds_1 + \dots \\ & \dots + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)\dots\int_{t_0}^{s_{m-1}} A(s_m)ds_m\, ds_{m-1}\dots ds_1 + \dots\end{aligned}$$

The above expression is sometimes called the *Peano-Baker series*. Note that, $\Phi(t,t) = I$. Note that we can differentiate the Peano-Baker series with respect to t to get:

$$\dot{\Phi}(t, t_0) = A(t)\Phi(t, t_0).$$

In the time-independent case of $A(t) = A$, the m'th term in the Peano-Baker series reduces to:

$$A^m \int_{t_0}^{t}\int_{t_0}^{s_1}\int_{t_0}^{s_2}\dots\int_{t_0}^{s_{m-1}} 1 ds_m \dots ds_1 = \frac{(t-t_0)^m}{m!}A^m.$$

Hence in this case, the state transition matrix reduces to the form,

$$\Phi(t, t_0) = \sum_{k=0}^{\infty} \frac{(t-t_0)^k}{k!}A^k \tag{3.8}$$

From the theory of differential equations and the result about the convergence of the Picard iteration sequence, the following can be deduced:

Theorem 3.1.6. *Let $A \in \mathbb{R}^{n\times n}$. Denote,*

$$S_m(t) = \sum_{k=0}^{m} \frac{t^k}{k!}A^k.$$

Then each element of the matrix $S_m(t)$ converges absolutely and uniformly on finite interval containing 0, as $m \to \infty$.

We can thus define the *matrix exponential* for any $t \in \mathbb{R}$ as,

$$e^{At} := \sum_{k=0}^{\infty} \frac{t^k}{k!}A^k.$$

Thus for the linear autonomous system, from (3.8) we have,

$$\Phi(t, t_0) = e^{A(t-t_0)}. \tag{3.9}$$

Hence the behavior of $\boldsymbol{x}(t) = A\boldsymbol{x}$, with initial value $\boldsymbol{x}(0) = \boldsymbol{x}_0$ is,

$$\boldsymbol{x}(t) = e^{At}\boldsymbol{x}_0. \tag{3.10}$$

Exercise 3.1.7. *Show the following elementary properties of the matrix exponential:*

1. $e^0 = I$.
2. *For scalar* t_1, t_2, $e^{At_1}e^{At_2} = e^{A(t_1+t_2)}$.
3. $e^{A'} = \big(e^A\big)'$.

Here are some further properties of the matrix-exponential

Theorem 3.1.8. *The following holds:*

1. *For* λ *and eignevalue of* A *it holds that* e^λ *is an eigenvalue of* e^A.
2. $\det\big(e^A\big) = e^{tr(A)}$
3. $A^p e^{At} = e^{At}A^p$ *for integer* p.
4. *If* $AB = BA$ *then,*
$$e^{A+B} = e^A e^B.$$
5. $\frac{d}{dt}e^{At} = Ae^{At}$.
6. *For non-singular* A,
$$\int_0^t e^{A\tau}\, d\,\tau = (e^{At} - I)A^{-1}.$$

Exercise 3.1.9. *Show by example that* $e^{A+B} = e^A e^B$ *does not necessarily imply that* $AB = BA$.

3.2 Evaluation of the Matrix Exponential

The straight forward method to compute A^ℓ is by carrying out $\ell - 1$ successive multiplications of the matrix A. This can be reduced to $O\big(\log(\ell)\big)$ matrix multiplications by carrying out successive evaluations of,

$$A_2 := AA, \quad A_4 := A_2A_2, \quad A_8 := A_4A_4, \quad \ldots \quad A_{2^k} := A_{2^{k-1}}A_{2^{k-1}},$$

up to $k = \lfloor \log_2 \ell \rfloor$, and then multiplying A_{2^k} by A (or other A_{2^i}) a few more times to "complete" the product from A^{2^k} to A^ℓ.

While this sort of algorithmic "divide and concur" approach yields a significant computation improvement, it still does not yield any insight about the structure of the sequence of matrices $\{A^\ell,\ \ell = 0, 1, 2, \ldots\}$.

The straight forward method to approximately compute e^{At} is to choose a large K and evaluate the finite sum,

$$\sum_{k=0}^{K} \frac{(At)^k}{k!}.$$

Since e^{At} always exists, the finite sum converges to the correct value as $K \to \infty$. Here one can look for basic computational improvements. For example by using the relation,

$$\frac{(At)^{k+1}}{(k+1)!} = \Big(\frac{At}{k+1}\Big)\frac{(At)^k}{k!},$$

to compute the $k+1$'st term in the summation based on the k'th term. But here again, such computational improvements do not yield insight on the structure of $\{e^{At},\ t \geq 0\}$.

We now consider more powerful and insightful linear-algebraic methods for effectively evaluating A^ℓ and e^{At} as well as for gaining insight on the behavior of these matrix sequences.

The following exercise shows that in some cases, evaluation is straightforward and explicit.

Exercise 3.2.1. *Take,*

$$A = \begin{bmatrix} 0 & 0 \\ \gamma & 0 \end{bmatrix},$$

and find $\{A^\ell,\ \ell = 0, 1, 2, \ldots\}$ *and* $\{e^{At},\ t \geq 0\}$ *explicitly.*

3.2.1 The Similarity Transformation

Given $P \in \mathbb{R}^{n\times n}$, with $det(P) \neq 0$, the matrices A and $\tilde{A}$ are said to be *similar* if,

$$\tilde{A} = P^{-1}AP \qquad \text{or alternatively} \qquad A = P\tilde{A}P^{-1}. \tag{3.11}$$

Here the action of replacing A by $P\tilde{A}P^{-1}$ is called the *similarity transformation.*

Assume now that we can find P such that evaluation of $\tilde{A}^\ell$ is in some way easier than A^ℓ. In this case, carrying out a similarity transformation is beneficial since,

$$\begin{aligned} A^\ell &= \big(P\tilde{A}P^{-1}\big)^\ell \\ &= \big(P\tilde{A}P^{-1}\big)\cdot\big(P\tilde{A}P^{-1}\big)\cdot\ldots\ldots\cdot\big(P\tilde{A}P^{-1}\big) \\ &= P\tilde{A}\big(P^{-1}P\big)\tilde{A}\big(P^{-1}P\big)\cdot\ldots\ldots\cdot\big(P^{-1}P\big)\tilde{A}P^{-1} \\ &= P\tilde{A}^\ell P^{-1}. \end{aligned}$$

Similarly, as a direct consequence, for the matrix exponential we have,

$$e^{At} = Pe^{\tilde{A}t}P^{-1}.$$

We now arrive at the question of what is a simple $\tilde{A}$? Well the simplest is a *diagonal matrix*.

Observe that the eigenvalues of $\tilde{A}$ and A are the same because the characteristic polynomial is the same:

$$\det\big(P^{-1}AP-\lambda I\big) = \det\big(P^{-1}(A-\lambda I)P\big) = \det(P^{-1})\det\big(A-\lambda I\big)\det(P) = \det\big(A-\lambda I\big).$$

3.2.2 Diagonalizable Matrices

If $\tilde{A}$ is a *diagonal matrix*,

$$\tilde{A} = \begin{bmatrix} \tilde{a}_1 & 0 & \dots & 0 \\ 0 & \tilde{a}_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \tilde{a}_n \end{bmatrix} := \mathrm{diag}(\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_n),$$

then it is easy to see that, $\tilde{A}^\ell = \mathrm{diag}(\tilde{a}_1^\ell, \tilde{a}_2^\ell, \dots, \tilde{a}_n^\ell)$. Further finding the matrix exponential of a diagonal matrix is also simple:

Exercise 3.2.2. *Show that if as above,* $\tilde{A} = diag\big(\tilde{a}_1, \dots, \tilde{a}_n\big)$ *, then,*

$$e^{\tilde{A}t} = diag\Big(e^{\tilde{a}_1 t}, \dots\dots, e^{\tilde{a}_n t}\Big).$$

We are thus motivated to find a similarity transformation matrix P that will yield a diagonal $\tilde{A}$. When this can be done, the matrix P that *diagonalizes* A, can be constructed by taking columns to be *eigenvectors* of A, each corresponding to a different eigenvalue. We illustrate the basic idea now.

The similarity transformation (3.11) may be read as,

$$AP = P\tilde{A}. \tag{3.12}$$

Now if the columns of P are eigenvectors of A and we impose on $\tilde{A}$ to be diagonal, then (3.12) is read as,

$$A\boldsymbol{p}_{\cdot,i} = \tilde{a}_i \boldsymbol{p}_{\cdot,i}, \quad i = 1, \dots, n,$$

where $\boldsymbol{p}_{\cdot,i}$ denotes the i'th column of P. In this case the diagonal elements of $\tilde{A}$ are nothing but the *eigenvalues* of A.

When is this possible? For start we have the following:

Theorem 3.2.3. *If for* $A \in \mathbb{R}^{n\times n}$ *there are* n *distinct eigenvalues then* A *is diagonalizable.*

Obviously having distinct eigenvalues is not a necessary condition for A to be diagonalizable (consider for example certain diagonal matrices):

Exercise 3.2.4. *Give an example of a a matrix A with non-distinct eigenvalues that is still diagonalizable.*

The *algebraic multiplicity* of an eigenvalue, λ_0, denoted $m_a(\lambda)$ is the multiplicity of the root λ_0 in in the characteristic equation $p_A(\cdot) = 0$. Namely, the polynomial $p_A(\lambda)$ is divisible by exactly $m_a(\lambda_0)$ powers of $(\lambda - \lambda_0)$.

Exercise 3.2.5. *Argue why $n = \sum_i m_a(\lambda_i)$.*

The *geometric multiplicity*, denoted $m_g(\lambda_0)$ is the dimension of the subspace E_{λ_0} (spanning eigenvectors associated with λ_0.

Theorem 3.2.6. *For a matrix A with eigenvalue λ:*

$$1 \leq m_g(\lambda) \leq m_a(\lambda).$$

Theorem 3.2.7. *A matrix is diagonalizable if and only if for all its eigenvalues $m_a = m_g$.*

3.2.3 Jordan's Canonical Form

If A is not diagonalizable, what can be done? We first define the *power vector*, $\boldsymbol{w}$, (sometimes called *generalized eignenvector* of a matrix A, if for some scalar λ and positive integer p:

$$(A - \lambda I)^p \boldsymbol{w} = 0.$$

The subspace spanned by power vectors corresponding to λ is called the *power space* of λ. The *order* of a power vector is said to be p if,

$$(A - \lambda I)^p \boldsymbol{w} = 0, \qquad \text{but} \qquad (A - \lambda I)^{p-1} \boldsymbol{w} \neq 0.$$

Exercise 3.2.8. *Show that if $\boldsymbol{w}$ is a power vector of order p with λ, then $(L - \lambda I)\boldsymbol{w}$ is a power vector of order $p - 1$, $(L - \lambda I)^2 \boldsymbol{w}$ is a power vector of order $p - 2$ and so on through to $(L - \lambda I)^{p-1} \boldsymbol{w}$ which is an eigenvector.*

For eigenvalues λ with eigenvector $\boldsymbol{v}$ we have that $e^{At} v = e^{\lambda t} v$. For in the case of a powervector $\boldsymbol{w}$ we have:

$$e^{At} w = e^{(A-\lambda I)t + \lambda t I} w = e^{\lambda t} e^{(A-\lambda I)t} w = e^{\lambda t} \sum_{k=0}^{\infty} \frac{1}{k!} t^k (A - \lambda I)^k w = e^{\lambda t} \sum_{k=0}^{p-1} \frac{1}{k!} t^k (A - \lambda I)^k w.$$

Example 3.2.9. *Consider,*

$$B = \begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix}.$$

It holds that $p_A(\lambda) = (\lambda - 1)^2$. *We have that* $(1, -1)'$ *is an eigenvector with eigenvalue* 1 *and* $(1, 1)'$ *is a power vector of order* 2 *corresponding to the eigenvalue* 1. *In this case,*

$$e^{Bt}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = e^t \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$
$$e^{Bt}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = e^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 2te^t \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Theorem 3.2.10. *Let* A *be a square matrix:*

1. *A collection of power vectors, each corresponding to distinct eigenvalues is linearly independent.*
2. *Every power vector has order less than or equal to* $m_a(\lambda)$.
3. *The power space corresponding to* λ *is the same as the kernel of* $(A - \lambda I)^{m_a(\lambda)}$.
4. *The dimension of the power space corresponding to* λ *is exactly* $m_a(\lambda)$.
5. *There exists a basis consisting of power vectors of* A.

Now using power vectors, we can construct the *Jordan Canonical Form* of an arbitrary matrix A (diagonalizable or not). Given an eignevalue λ, a *Jordan Block* is a block of size m is the $m \times m$ matrix.

$$I_m(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda & 1 & & & \\ \vdots & & \lambda & 1 & & \\ \vdots & & & \ddots & \ddots & \\ \vdots & & & & \lambda & 1 \\ 0 & 0 & \cdots & \cdots & \cdots & \lambda \end{bmatrix}.$$

Note that a Jordan block of size 1 is the scalar λ.

Now the *Jordan canonical form* of a matrix A with r distinict eigenvalues, $\lambda_1, \ldots, \lambda_r$ each with geometric multplicity denoted by m_i is the matrix,

$$\tilde{A} = \begin{bmatrix} I_{m_1}(\lambda_1) & & & & \\ & I_{m_2}(\lambda_1) & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I_{m_r}(\lambda_r) \end{bmatrix}.$$

Exercise 3.2.11. *Explain why the Jordan canonical form of a matrix reduces to a diagonal matrix if a matrix is diagonalizable.*

Theorem 3.2.12. *There always exists a non-singular P such that,*

$$\tilde{A} = P^{-1}AP,$$

where $\tilde{A}$ is the Jordan canonical form of A.

Now in general we have $e^A = P^{-1}e^{\tilde{A}}P$ and since $\tilde{A}$ is either diagonal or block-diagonal (in the non-diagonalizable) case, it is "easier" to evaluate its matrix exponential (similarly for matrix powers).

3.2.4 The Resolvent

We now wish to get the Laplace transform of the impulse response matrix. As a first step, let us consider the autonomous system $\dot{\boldsymbol{x}} = A\boldsymbol{x}$ with $\boldsymbol{x}(0) = \boldsymbol{x}_0$ (studied in Section 3.1). Using the derivative property of Laplace transforms we have:

$$s\hat{\boldsymbol{x}}(s) - \boldsymbol{x}_0 = A\hat{\boldsymbol{x}}(s), \tag{3.13}$$

and thus for s that are not eigenvalues of A,

$$\hat{\boldsymbol{x}}(s) = (sI - A)^{-1}\boldsymbol{x}_0.$$

Hence the Laplace transform matrix of e^{At} is $(sI - A)^{-1}$. This is called the *resolvent* of the system.

Note that the resolvent yields an additional method for computing e^{At}. Here is an example:

Example 3.2.13. *Consider,*

$$A = \begin{bmatrix} -1 & 3 \\ 0 & 1 \end{bmatrix}.$$

Then,

$$(sI-A)^{-1} = \begin{bmatrix} s+1 & -3 \\ 0 & s-1 \end{bmatrix}^{-1} = \frac{1}{(s+1)(s-1)} \begin{bmatrix} s-1 & 3 \\ 0 & s+1 \end{bmatrix} = \begin{bmatrix} \frac{1}{s+1} & \frac{3/2}{s-1} - \frac{3/2}{s+1} \\ 0 & \frac{1}{s-1} \end{bmatrix}$$

So,

$$e^{At} = \begin{bmatrix} e^{-t} & \frac{3}{2}(e^t - e^{-t}) \\ 0 & e^t \end{bmatrix}.$$

3.2.5 More on Matrix Exponential Computation

3.3 Markov Chains in Discrete Time

Moving from the deterministic objects $\boldsymbol{x}(\cdot)$ to the stochastic ones $X(\cdot)$, we now consider Markov chains. Specifically, *discrete time Markov chains* (DTMC) and *continuous time Markov chains* (CTMC).

3.3.1 Markov Chain Basics

A *stochastic process* is a random function $X(\ell, \omega)$ where say $t \in \mathbb{R}$ (or $\ell \in \mathbb{Z}$) represents time and $\omega \in \Omega$ is a point in the probability sample space. An alternative view, is to think of a stochastic process as a family (sequence) of random variables $\{X(t, \omega),\, t \in \mathbb{R}\}$ (or $\ell \in \mathbb{Z}$). Stochastic processes get interesting when the random variables are not independent. I.e. when there is some dependence structure between them. In the sequel we omit the fact that $X(\cdot, \omega)$ depends on ω from the notation, but keep in mind it is always there.

When analysing a stochastic process, we sometimes use the term *sample path* or alternatively *realisation* to refer to one instance of the time function $X(\cdot, \omega)$ associated with a single ω.

An elementary, but highly useful stochastic process is the *time homogenous finite state space discrete time Markov chain* (*finite DTMC* for short). This is a sequence of random variables indexed by $\ell \in \mathbb{Z}_+$ with the following three properties:

1. Lack of memory (Markovian property):

$$\mathbb{P}(X(\ell+1) = j \mid X(\ell) = i_t, \ldots, X(0) = i_0) = \mathbb{P}(X(\ell+1) = j \mid X(\ell) = i_t).$$

2. Time Homogeneity (this makes the probability law of the the process time-homogenous):

$$\mathbb{P}(X(\ell+1) = j \mid X(\ell) = i) = \mathbb{P}(X(1) = j \mid X(0) = i) := p_{i,j}.$$

3. Finite state space: There is some finite set (state space), $\mathcal{S}$, such that,

$$\mathbb{P}(X(\ell) \notin \mathcal{S}) = 0, \quad \forall \ell.$$

Since we are considering finite state space Markov chains, we may think of $\mathcal{S} = \{1, \ldots, N\}$ for some fixed integer $N \geq 2$. At the end of section we briefly also discuss infinite (but still countable) state-spaces. As you read this, it may be a good idea that you occasionally ask yourself, where (and how) the finite state space assumption is used.

Based on properties (1) and (2) above, it can be seen that in order to specify the probability law of the evolution of $\{X(\ell)\}$ we need to specify, $p_{i,j}$ for $i, j \in \mathcal{S}$ as well as the distribution of $X(0)$ (the *initial distribution*). The convenient way to specify the *transition probabilities* is by an $N \times N$ matrix $P = [p_{i,j}]$ with non-negative elements and with row sums $= 1$. I.e. each row i can be treated as a PMF indicating the distribution of $X(\ell+1)$ given that $X(\ell) = i$. A convenient way to is specify the initial distribution is by a row vector, $\mathbf{r}(0)$ of length N having non-negative elements and summing to 1 with i'th entry, $\mathbf{r}_i(0)$ meaning: $\mathbb{P}(X(0) = i) = \mathbf{r}_i(0)$. This is can again be viewed as a PMF.

Note that a non-negative matrix with row sums equal to 1 is called a *stochastic matrix*. Don't let the name confuse you; it isn't a random variable or a random matrix, it is a deterministic object.

Exercise 3.3.1. *Check that* $\mathbb{P}(C|AB) = \mathbb{P}(C|A) \Leftrightarrow \mathbb{P}(CB|A) = \mathbb{P}(C|A)\mathbb{P}(B|A)$.

Now using basic conditional probability and the law of total probability we can get some very nice properties. First for $\ell = 0, 1, 2, \ldots$, denote,

$$p_{i,j}^{(\ell)} = \mathbb{P}(X(\ell) = j \mid X(0) = i),$$

and the matrix of these probabilities by $P^{(\ell)} = [p_{i,j}^{(\ell)}]$. Also denote,

$$r_i(\ell) = \mathbb{P}(X(\ell) = i),$$

with $\mathbf{r}(\ell)$ being the row vector of these probabilities.

Exercise 3.3.2. *The basic dynamics of DTMCs is given by the following:*

1. *Show that* $P^{(0)}$ *is the identity matrix.*
2. *Show (arguing probabilistically) that* $P^{(\ell)}$ *is a stochastic matrix for any* $\ell \in \mathbb{Z}_+$.
3. *Show the* Chapman-Kolmogorov equations *hold:*

$$p_{i,j}^{(m+n)} = \sum_{k=1}^{N} p_{i,k}^{(m)}\, p_{k,j}^{(n)}.$$

4. *Show that* $P^{(\ell)} = P^{\ell}$. *I.e.* $P^{\ell} = P \cdot P \cdot \ldots \cdot P$, *where the product is of* ℓ *matrices.*
5. *Show that* $\mathbf{r}(\ell) = \mathbf{r}(0)P^{\ell}$ *(the right hand side here is a row vector multiplied by a matrix).*

The next exercise, will ensure you got the point. I hope you are in the mood.

Exercise 3.3.3. *Make a model of your feelings. Say* $1 \equiv$ *"happy",* $2 \equiv$ *"indifferent",* $3 \equiv$ *"sad". Assume that you are Markovian (i.e. the way you feel at day* $\ell + 1$ *is not affected by days prior to day* ℓ*, if the feelings at day* ℓ *are known)*

1. *Specify the transition probabilities matrix* P *which you think matches you best.*
2. *Assume that at day* 0 *you are sad with probability* 1. *What is the probability of being happy in day* 3.
3. *Assume that at day* 0 *you have a (discrete) uniform distribution of feelings, what is the probability of being happy in day* 3.
4. *Assuming again, that the initial distribution is uniform, what is the probability of "happy, happy, sad, sad, happy" (a sequence of 5 values on times* $\ell = 0, 1, \ldots, 4$*).*

Markov chains generalised i.i.d. sequences:

Exercise 3.3.4. *Assume you are given a PMF $p_X(\cdot)$ with support $\{1, \ldots, N\}$. How can you make a Markov chain such that $\{X(\ell)\}$ is an i.i.d. sequence of that PMF? I.e. what matrix P will you use? Explain.*

The fact that $\mathbf{r}(\ell) = \mathbf{r}(0)P^\ell$ is remarkable and beautiful. But in general it is quite hard to have an explicit analytic expression for P^ℓ. With some effort, you can do this for a two-state Markov chain:

Exercise 3.3.5. *Consider the Markov chain over $\mathcal{S} = \{1, 2\}$.*

1. *How many free parameters are in this model (i.e. how many numbers specify $\boldsymbol{r}(0)$ and P)?*
2. *Write an expression for P^ℓ in terms of the parameters (e.g. do this by diagonalising the matrix P so that you can evaluate matrix powers easily).*
3. *Write an expression for $\boldsymbol{r}(\ell)$.*
4. *What happens to $\boldsymbol{r}(\ell)$ as $\ell \to \infty$?*
5. *Do you have any intuition on the previous result?*

3.3.2 First-Step Analysis

Consider a gambler; one of those hard-core TAB types. She has $X(\ell)$ dollars at day ℓ. Her goal is to reach L dollars, since this is the amount needed for the new tattoo she wants[1]. She attends the bookies daily and is determined to gamble her one dollar a day, until she reaches either L or goes broke, reaching 0. On each gamble (in each day) she has a chance of p of earning a dollar and a chance of $1 - p$ of loosing a dollar.

This problem is sometimes called the *gambler's ruin* problem. We can view her fortune as the state of a Markov chain on state space, $\mathcal{S} = \{0, 1, 2 \ldots, L-1, L\}$.

Exercise 3.3.6. *Specify the transition probabilities $p_{i,j}$ associated with this model.*

At day $\ell = 0$, our brave gambler begins with $X(0) = x_0$ dollars. As she drives to the bookies, Jimmy texts her: "Hey babe, I was wondering what is the the chance you will eventually reach the desired L dollars?". She thinks while driving, but can't concentrate, so she stops the car by the side of the road and sketches out the following in writing: Define,

$$\tau_0 := \inf\{\ell \geq 0 \ : \ X(\ell) = 0\}, \qquad \tau_L := \inf\{\ell \geq 0 \ : \ X(\ell) = L\}.$$

[1]The tattoo will feature the name of her boyfriend, "Jimmy" together with a picture of a Holden.

These two objects are random variables which are called *hitting times* (the time it takes till hitting a state for the first time). They are random because different realisations of $X(\cdot, \omega)$ imply different values for τ_0 or τ_L. Note that the infimum of the empty set is defined to be ∞. So if our gambler, for example reaches L, then $\tau_0 = \infty$ and similarly if the other case occurs.

In terms of hitting times, Jimmy's question to our gambler, was to evaluate:

$$q_i := \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i\big), \qquad \text{with} \qquad i = x_0.$$

We define q_i for all states i, because to evaluate q_{x_0} we will need the other q_i also. It is obvious that $q_0 = 0$ and $q_L = 1$ but what if $i \in \{1, \dots, L-1\}$? Well here we can partition the event $\{\tau_L > \tau_0\}$ based on the *first step*:

$$\begin{aligned}
q_i &= \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i\big) \\
&= \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i,\, X(1) = i+1\big)\, p_{i,i+1} + \mathbb{P}\big(\tau_L < \tau_0,\, \mid X(0) = i,\, X(1) = i-1\big)\, p_{i,i-1} \\
&= \mathbb{P}\big(\tau_L < \tau_0 \mid X(1) = i+1\big)\, p_{i,i+1} + \mathbb{P}\big(\tau_L < \tau_0,\, \mid X(1) = i-1\big)\, p_{i,i-1} \\
&= \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i+1\big)\, p_{i,i+1} + \mathbb{P}\big(\tau_L < \tau_0,\, \mid X(0) = i-1\big)\, p_{i,i-1} \\
&= q_{i+1}\, p + q_{i-1}(1-p).
\end{aligned}$$

So using this *first step analysis* we end up with $L+1$ equations for the $L+1$ unknowns $q_0, q_1, \dots, q_L$:

$$\begin{bmatrix}
1 & 0 & & & & & 0 \\
(1-p) & -1 & p & & & & \\
0 & (1-p) & -1 & p & & & \\
& \ddots & \ddots & \ddots & \ddots & & \\
& & & (1-p) & -1 & p & 0 \\
& & & & (1-p) & -1 & p \\
0 & & & & & 0 & 1
\end{bmatrix}
\begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ \vdots \\ \vdots \\ q_{L-2} \\ q_{L-1} \\ q_L \end{bmatrix}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

The unique solution to these equations happens to be,

$$q_i = \begin{cases} i/L & \text{if } p = 1/2, \\ \frac{1 - \left(\frac{1-p}{p}\right)^i}{1 - \left(\frac{1-p}{p}\right)^L} & \text{if } p \neq 1/2. \end{cases} \tag{3.14}$$

Exercise 3.3.7. *Verify that the solution given above is correct.*

1. *(Analytically) – Plug it in the equations and see it satisfies them.*
2. *(Numerically) – Make a* 10×10 *matrix in matlab (or anything else) and see that the vector* q_i *solves the equations above.*

3. *(Simulation) – Simulate this gamblers ruin problem for some given parameters (say with $L = 9$) to verify that q_i is indeed correct. Basically do this by generating sample paths $X(\cdot, \omega)$ for all times ℓ, till* $\min\{\tau_0, \tau_L\}$.

Exercise 3.3.8. *Assume you didn't know the formula in* (3.14). *Think of methods in which you can obtain it. Outline your methods. Try to start with $p = 1/2$ and then move onto $p \neq 1/2$.*

The concept of *first step analysis* goes hand in hand with Markov chains and is useful for a variety of settings. When our gambler finished the calculations above, she texted Jimmy the result (q_{x_0}) and drove off. But then she got another text: "Honey love, for how many more days will you do this? Can't wait babe!". She thinks, and then figures out that Jimmy wants to know,

$$m_i := \mathbb{E}[\min\{\tau_0, \tau_L\} \mid X(0) = i] \qquad \text{with} \qquad i = x_0.$$

By now our gambler knows how to do first step analysis, even while driving. She formulates the following: First,

$$m_0 = 0 \qquad \text{and} \qquad m_L = 0.$$

Even Jimmy can do this part. But further for $i \in \{1, 2, \ldots, L-1\}$:

$$\begin{aligned} m_i &= p_{i,i+1}(1 + m_{i+1}) + p_{i,i-1}(1 + m_{i-1}) \\ &= 1 + p_{i,i+1}m_{i+1} + p_{i,i-1}m_{i-1} \\ &= 1 + p\, m_{i+1} + (1-p)\, m_{i-1} \end{aligned}$$

So again we have $L + 1$ equations with $L + 1$ unknowns.

Exercise 3.3.9. *Find the solution when $p = 1/2$.*

Exercise 3.3.10. *Find the solution when $p \neq 1/2$.*

3.3.3 Class Structure, Periodicity, Transience and Recurrence

Note: Some of the derivations in this section are informal. Nevertheless, the reader should know that without much extra effort, all of the results can be proved in a precise manner.

One way to visualise the transition matrix of a finite DTMC is by drawing the weighted graph associated with P. Edges associated with (i, j) such that $p_{i,j} = 0$ are omitted . If you ignore the weights you simply get a directed graph. What does this graph tell you? Well, by studying it, you can see which paths the process may possibly take, and which paths are never possible. Of course, if $p_{i,j} > 0$ for all state pairs, then there is nothing to do because you have a complete graph. But in applications and theory, we often have

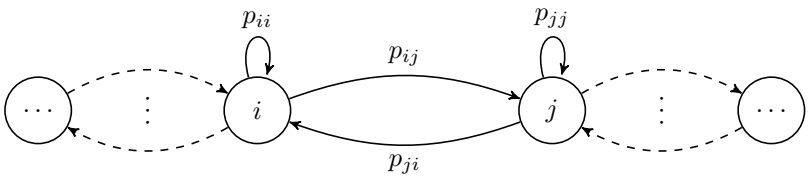


Figure 3.1: DTMC Transition Diagram

$p_{i,j} = 0$ for a significant portion of the tuples (i, j). This allows us to study the *directed graph* that has edge (i, j) only when $p_{i,j} > 0$. This graph obviously doesn't specify all of the information about the DTMC, but it does tell us the *class structure*. We describe this now.

We say that two states, i and j *communicate* if there are two non-negative integers t_1 and t_2 such that $p_{i,j}^{(t_1)} > 0$ and $p_{j,i}^{(t_2)} > 0$. This implies there is a path (in the directed graph) from i to j and a path from j to i. We denote communication of i and j by $i \leftrightarrow j$. The relation of communication is an equivalence relation[2] over the set of states. Namely: $i \leftrightarrow i$ (reflexivity); if $i \leftrightarrow j$ then $j \leftrightarrow i$ (symmetry); and finally if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$ (transitivity).

Exercise 3.3.11. *Use the Chapman-Kolmogorov equations to prove transitivity.*

The implication of the fact that $\leftrightarrow$ is an equivalence relation is that it induces equivalence classes, $\mathcal{C}_1, \mathcal{C}_2, \ldots$ that are a partition of $\mathcal{S}$. That is, $\mathcal{C}_i$ and $\mathcal{C}_j$ are disjoint for $i \neq j$ and $\cup_i \mathcal{C}_i = \mathcal{S}$. All states within class $\mathcal{C}_i$ communicate with each other, but do not communicate with states that are not in $\mathcal{C}_i$. Obviously for finite state spaces of size N, there can be at most N classes and this upper bound is achieved only when $P = I$, the identity matrix. At the other extreme, we are often interested in Markov chains with only one class. Such Markov chains are said to be *irreducible*.

A state i is said to have a period of d if $p_{i,i}^{(\ell)} = 0$ for all integers ℓ that are not divisible by d, and further d is the greatest integer with this property. E.g, assume, that $p_{i,i}^{(3)} > 0$, $p_{i,i}^{(6)} > 0$, $p_{i,i}^{(9)} > 0$ etc... and further $p_{i,i}^{(\ell)} = 0$ for $\ell \notin \{3, 6, 9, \ldots\}$. So if we start at time 0 in state i we can only expect to be in state i at the times $3, 6, 9, \ldots$. It isn't guaranteed that at those times we visit state i, but we know that if we do visit state i, it is only at those times. It can be shown that all states in the same class have the same period. But we won't ponder on that. In general, we aren't so interested in periodic behaviour, but we need to be aware of it. In particular note that if $p_{i,i} > 0$ for all states i, then the Markov chain is guaranteed to be non-periodic.

[2]If for some reason you don't know what an *equivalence relation* is, don't stress. You'll understand from the text.

Define now, the hitting time[3] (starting at 1): $\tau_i = \inf\{\ell \geq 1 \mid X(\ell) = i\}$ and define,

$$f_{i,j}^{(\ell)} = \begin{cases} \mathbb{P}(\tau_j = \ell \mid X(0) = i) & \text{if } \ell \geq 1, \\ 0 & \text{if } \ell = 0. \end{cases}$$

Further define $f_{i,j} = \sum_{\ell=1}^{\infty} f_{i,j}^{(\ell)}$. This is the probability of ever making a transition into state j, when starting at state i:

$$f_{i,j} = \mathbb{P}\Big(\sum_{\ell=1}^{\infty} \mathbf{1}\{X(\ell) = i\} \geq 1 \mid X(0) = i \Big).$$

A state i is said to be *recurrent* if $f_{i,i} = 1$. This means that if $X(0) = i$ we will continue visiting the state again and again. A state that is not recurrent is *transient*; i.e. i.e., $f_{i,i} < 1$ then there is a non-zero chance $(1 - f_{i,i})$ that we never return to the state.

Exercise 3.3.12. *Assume that $X(0) = i$ and state i is transient. Explain why the distribution of the number of visits to state i after time 0, is geometric with success probability $1 - f_{i,i}$ and mean $1/(1 - f_{i,i})$. I.e.,*

$$\mathbb{P}\big(\sum_{\ell=1}^{\infty} 1\{X(\ell) = i\} \;=\; n \mid X(0) = i \big) = (1 - f_{i,i})(f_{i,i})^n, \qquad n = 0, 1, 2, \ldots$$

Further, write an expression (in terms of $f_{i,j}$ values) for,

$$\mathbb{P}\big(\sum_{\ell=1}^{\infty} 1\{X(\ell) = j\} \;=\; n \mid X(0) = i \big).$$

In certain cases, it is obvious to see the values of $f_{i,j}$:

Exercise 3.3.13. *Consider the Markov chain with transition matrix,*

$$P = \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}.$$

1. *What are the classes of the Markov chain.*
2. *Which states are transient, and which are recurrent.*
3. *What are $f_{i,j}$ for all i, j?*

[3]Some authors refer to the case starting at time 1 as as a first passage time and to the case starting at time 0 as a *hitting time*. This distinction only matters if the initial state is i itself.

Consider now the following example,

$$P = \begin{bmatrix} 0.1 & 0.7 & 0.2 & 0 \\ 0.4 & 0.3 & 0 & 0.3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.15}$$

The classes of this example are $\mathcal{C}_1 = \{1, 2\}$, $\mathcal{C}_2 = \{3\}$ and $\mathcal{C}_3 = \{4\}$. Here without doing any calculations it is already obvious that $f_{3,3} = 1$ and $f_{4,4} = 1$, since states 3 and 4 are recurrent. They are even called *absorbing*, because once you get to state 3 or state 4, you never leave. So $f_{3,i} = 0$ for $i \neq 3$ and further $f_{4,i} = 0$ for $i \neq 4$. But the values $f_{i,j}$ with $i \in \{1, 2\}$ are not as clear. Starting in state 1, for example, there is a 0.2 chance of absorbing in 3 and and with the complement there is a chance of staying within the class $\mathcal{C}_1$. So how does this affect $f_{1,i}$?
The general mechanism we can use is first step analysis. This is the basic equation:

$$\begin{aligned} f_{i,j} &= \mathbb{P}\big(\sum_{\ell=1}^{\infty} 1\{X(\ell) = j\} \geq 1 \mid X(0) = i\big) \\ &= \sum_{k \neq j} \mathbb{P}\big(\sum_{\ell=1}^{\infty} 1\{X(\ell) = j\} \geq 1 \mid X(0) = i,\ X(1) = k\big)\, p_{i,k} \\ &+ \mathbb{P}\big(\sum_{\ell=1}^{\infty} 1\{X(\ell) = j\} \geq 1 \mid X(0) = i,\ X(1) = j\big)\, p_{i,j} \\ &= \sum_{k \neq j} f_{k,j}\, p_{i,k} + p_{i,j} \\ &= \sum_{k \neq j} p_{i,k} f_{k,j} + p_{i,j}. \end{aligned}$$

Exercise 3.3.14. *This exercise relates to the matrix P in* (3.15).

1. *Find $f_{1,3}$ and $f_{1,4}$ (you'll need to find out other $f_{i,j}$ values for this).*
2. *Explain why $f_{1,3} + f_{1,4} = 1$.*
3. *Run a simulation to verify your calculated value of $f_{1,3}$.*

There are many characterisations of recurrent and transient states. One neat characterisation is the following:

$$\text{State } i \text{ is recurrent if and only if } \sum_{\ell=0}^{\infty} p_{i,i}^{(\ell)} = \infty. \tag{3.16}$$

The idea of the derivation looks at the expected number of visits to the state:

$$\mathbb{E}\big[\sum_{\ell=0}^{\infty} 1\{X(\ell)=i\} \mid X(0)=i\big] = \sum_{\ell=0}^{\infty} \mathbb{E}\big[1\{X(\ell)=i\} \mid X(0)=i\big] = \sum_{\ell=0}^{\infty} p_{i,i}^{(\ell)}$$

Now for a recurrent state, we know that $\sum_{\ell=0}^{\infty} \mathbf{1}\{X(\ell)=i\} = \infty$ and thus the expectation of this random variable should also be ∞. So this shows the direction $\Longleftarrow$. For the other direction assume that state i is transient (the contrapositive). In this case we saw that $\sum_{\ell=0}^{\infty} \mathbf{1}\{X(\ell)=i\}$ is a geometric random variable with finite expectation, so $\sum_{\ell=0}^{\infty} p_{i,i}^{(\ell)} < \infty$.

In many cases, we can't explicitly compute $p_{i,i}^{(\ell)}$ so there isn't much computational use for (3.16). But one classic fascinating example is the *simple random walk*. For this we assume now a state is $\mathcal{S} = \mathbb{Z}$ (not finite any more!). Take $p \in [0,1]$ and set,

$$p_{i,j} = \begin{cases} p & \text{if } j = i+1, \\ (1-p) & \text{if } j = i-1, \\ 0 & \text{otherwise.} \end{cases}$$

The example is called a random walk because at every time step the walker takes either a step up with probability p or a step down with probability $1-p$. It is called simple, because the change at each time point is a random variable with support $\{-1,1\}$. In the general random walk, steps would be of arbitrary magnitude.

A nice feature of this model is that we can actually calculate $p_{i,i}^{(\ell)}$.

Exercise 3.3.15. *Verify the following:*

1. *If $p=0$ or $p=1$ there is an infinite number of classes, but if $p \in (0,1)$ the model is irreducible.*
   For the rest of the items below, assume $p \in (0,1)$.

2. *The model is periodic with period 2.*
   So now we will consider $p_{i,i}^{(2\ell)}$, since for $\ell \in \{1,3,5,7,\ldots\}$, $p_{i,i}^{(\ell)} = 0$.

3. *Explain why:*
   $$p_{i,i}^{(2\ell)} = \binom{2\ell}{\ell} p^{\ell}(1-p)^{\ell}.$$

4. *Now use the Stirling approximation for $\ell!$ (see Appendix) to show,*
   $$p_{i,i}^{(2\ell)} \sim \frac{\big(4p(1-p)\big)^{\ell}}{\sqrt{\pi\ell}},$$
   where the symbol $\sim$ implies that as $\ell \to \infty$ the ratio of the left hand side and the right hand side goes to 1.

5. *Verify (using the definition of convergence of a series), that if $a_\ell \sim b_\ell$ then $\sum_\ell a_\ell < \infty$ if and only if $\sum_\ell b_\ell < \infty$.*

6. *Verify that*
$$\sum_{\ell=0}^{\infty} \frac{\big(4p(1-p)\big)^\ell}{\sqrt{\pi\ell}} = \infty,$$
if and only if $p = 1/2$ (otherwise the series converges).

With the results of the above exercise we know that state i (for any i) is recurrent if and only if $p = 1/2$. That is if $p \neq 1/2$ then all states are transient. Loosely speaking, the chain will "drift off" towards $+\infty$ if $p > 1/2$ and towards $-\infty$ if $p < 1/2$. States may be revisited, but ultimately, each state i will be revisited only a finite number of times.

In finite Markov chains, we can't have all states transient:

Exercise 3.3.16. *Argue why a finite DTMC, must have at least one recurrent state.*

In the infinite state space case, we can sometimes have that,
$$\mathbb{E}[\tau_i \mid X(0) = i] = \infty,$$
even when state i is recurrent. Such is actually the case for the simple random walk in the symmetric case ($p = 1/2$). This cannot happen when the state space is finite. This phenomenon is called *null-recurrence*. The other case,
$$\mathbb{E}[\tau_i \mid X(0) = i] < \infty,$$
is referred to as *positive-recurrence*. In finite state space DTMC all recurrent states are positive-recurrent. Further, in the finite state space case, if the DTMC is irreducible then all states are recurrent and thus all states are positive-recurrent. Further on this is in Chapter 6 dealing with stability.

3.3.4 Limiting Probabilities

We are often interested in the behaviour of $\{X(\ell)\}$ over long time periods. In applied mathematics, infinity, is a good approximation for "long". There is much to say here and we will only cover a small portion of the results and cases. Specifically, let us now assume that our DTMC has finite state-space, that it is irreducible, and that it is aperiodic (all states have a period of 1). Limiting probability results often hold when these assumptions are partially relaxed, but one needs to take more care in specifying the results.

To illustrate the main idea we return to exercise (3.3.3). If your example chain for that exercise had $p_{i,i} \in (0, 1)$ then the above conditions are satisfied. Let us assume that this is the case. Now ask[4],

[4]Beware of such questions if your current age is $10 * n \pm \epsilon$ where ϵ is small. Such thoughts can throw you on soul adventures that you may end up regretting – or maybe not.

"Over the long range, in what proportion of my days am I happy?"

Remembering that our code for "happy" was 1, the question can be posed as finding

$$\pi_1 := \lim_{\ell\to\infty} \mathbb{E}\Big[\frac{\sum_{\ell=0}^{\ell} \mathbf{1}\{X(\ell)=1\}}{\ell}\Big].$$

The value π_1 is then referred to as the limiting probability of being in state 1. I should hope that for your Markov chain of exercise (3.3.3), π_1 is high (close to 1). How can we evaluate it? The key result is that we can solve the system of equations:

$$\begin{aligned}
\pi_1 &= \pi_1\, p_{1,1} + \pi_2\, p_{2,1} + \pi_3\, p_{3,1},\\
\pi_2 &= \pi_1\, p_{1,2} + \pi_2\, p_{2,2} + \pi_3\, p_{3,2},\\
\pi_3 &= \pi_1\, p_{1,3} + \pi_2\, p_{2,3} + \pi_3\, p_{3,3},\\
1 &= \pi_1 + \pi_2 + \pi_3.
\end{aligned}$$

Now the unique solution, $[\pi_1,\, \pi_2,\, \pi_3]$ gives the long range proportion during which state i is occupied. Note that we have 4 equations with only 3 unknowns, but we should in fact omit one (any one) of the first 3 equations (this is a consequence of the fact P is a singular matrix). These equations are called the *balance equations*. In matrix form they are compactly written with $\boldsymbol{\pi}$ taken as a row vector and $\mathbf{1}$ a column vector of 1's.

$$\begin{aligned}
\boldsymbol{\pi} &= \boldsymbol{\pi} P,\\
1 &= \boldsymbol{\pi}\mathbf{1}.
\end{aligned} \tag{3.17}$$

Remember that, $\boldsymbol{r}(\ell+1) = \boldsymbol{r}(\ell)\, P$. What is a fixed point of this linear dynamical system? Fixed points $\overline{\boldsymbol{r}}$ need to satisfy: $\overline{\boldsymbol{r}} = \overline{\boldsymbol{r}} P$. One obvious such fixed point is $\mathbf{0}'$. But this fixed point is not a probability distribution. Is it the only fixed point? What if P has an eigenvalue equal to 1? In this case any (left) eigenvector corresponding to the eigenvalue 1 is a fixed point. One such (left) eigenvector is $\boldsymbol{\pi}$. Indeed the Perron-Frobenious theorem implies that P has an eigenvalue of 1.

Exercise 3.3.17. *Consider your matrix P of exercise* (3.3.3)*. Use a computer for the following:*

1. *Solve the balance equations for $\boldsymbol{\pi}$.*
2. *Run a single simulation of the DTMC for $T = 10,000$ time points. Choose any initial distribution for $X(0)$. Evaluate for $i \in \{1, 2, 3\}$,*

$$\hat{\pi}_i := \frac{\sum_{\ell=0}^{T} \mathbf{1}\{X(\ell)=i\}}{T},$$

compare these values to the answer of item 1.

3. *Compute P^5, P^{10},P^{20} and P^{100}. Compare the rows of these matrices with the answer of item 1.*

4. *The numerical illustration of the previous item, indicates that the rows all converge to $\boldsymbol{\pi}$. If this is indeed true (which it is), argue that for any initial distribution, $\boldsymbol{r}(0)$,*
$$\lim_{\ell\to\infty} \boldsymbol{r}(\ell) = \boldsymbol{\pi}.$$

The numerics of the above exercise, indicate the validity of the following (we omit the proof – note also that there are much more general formulations):

Theorem 3.3.18. *Consider a finite DTMC that is irreducible and non-periodic. Then,*

1. *The balance equations* (3.17) *have a unique solution with $\pi_i \in (0,1)$.*

2. *It holds that for any $i \in \mathcal{S}$,*
$$\lim_{\ell\to\infty} p_{i,j}^{(\ell)} = \pi_j.$$

3. *It holds that,*
$$\pi_i = \frac{1}{\mathbb{E}[\tau_i \mid X(0) = i]}.$$

4. *For any function, $f : \mathcal{S} \to \mathbb{R}$, we have with probability one,*
$$\lim_{\ell\to\infty} \frac{\sum_{k=0}^{\ell} f\big(X(k)\big)}{\ell} = \sum_{i\in\mathcal{S}} \pi_i\, f(i).$$

So basically, knowing $\boldsymbol{\pi}$ gives us much information about the *long run* or *steady state* behaviour of the system. When talking about long range behaviour it is $\boldsymbol{\pi}$ that matters; the initial distribution, $\boldsymbol{r}(0)$ becomes insignificant. Item 4 (also called the *ergodic property*) shows that long range behaviour can be summarised in terms of $\boldsymbol{\pi}$.

One of the names of the distribution $\boldsymbol{\pi}$ is the *stationary distribution* also known as the *invariant distribution*. A process $\{X(\cdot)\}$ (in discrete or continuous time) is stationary if for any integer $k \geq 0$ and any time values, $t_1, \ldots, t_k$, and any integer τ,
$$\mathbb{P}(X(t_1) = i_1,\, \ldots,\, X(t_k) = i_k) = \mathbb{P}(X(t_1+\tau) = i_1,\, \ldots,\, X(t_k+\tau) = i_k).$$

Exercise 3.3.19. *Use the equations describing $\boldsymbol{\pi}$ to show:*

1. *If we start at time 0 with $\boldsymbol{r}(0) = \boldsymbol{\pi}$, then $\boldsymbol{r}(1) = \boldsymbol{\pi}$ and this holds for all $\boldsymbol{r}(\ell)$.*

2. *More generally, show that if we start at time 0 with $\boldsymbol{r}(0) = \boldsymbol{\pi}$ then the process is stationary.*

So when we look at a DTMC, we can consider the *stationary version* where we choose $\boldsymbol{r}(0) = \boldsymbol{\pi}$. This means we are looking at the system which is already in "statistical equilibrium". Such systems may not exactly occur in practice, but it is often a very sensible approximation for systems that have been running for some time.

If on the other hand $\boldsymbol{r}(0) \neq \boldsymbol{\pi}$, then the DTMC is not stationary. But still, if we let it run for some time, it can be approximately considered to be stationary. This is due to item 2 of the theorem above.

3.4 Markov Chains in Continuous Time

Note that we use the phrase *Markov chain* for the case when the state space is countable, reserving the phrase *Markov process* for the case when the state space is continuous (this usage is not universal). We now discuss *Continuous Time Markov Chains* (CTMC).

3.4.1 Continuous Time Basics

Informally a finite state CTMC, is a process $\{X(t)\}$ in continuous time that satisfies:

1. Lack of memory (Markovian property):

$$\begin{aligned}&\mathbb{P}\big(X(t+s) = j \mid X(t) = i \text{ and further info about } X(u) \text{ for } u \in [0,t)\big)\\ &= \mathbb{P}\big(X(t+s) = j \mid X(t) = i\big).\end{aligned}$$

2. Time Homogeneity:

$$\mathbb{P}(X(t+s) = j \mid X(t) = i) = \mathbb{P}(X(s) = j \mid X(0) = i).$$

3. Finite state space: There is some finite set (state space), $\mathcal{S}$, such that,

$$\mathbb{P}(X(t) \notin \mathcal{S}) = 0, \quad \forall t.$$

In case $\mathcal{S}$ is not finite, but rather countably infinite, the process is still a CTMC and many of the results hold.

Suppose $X(0) = j$ and $T(j)$ is the first time the CTMC leaves j. Then

$$\begin{aligned}&\mathbb{P}(T(j) > t+s \mid T(j) > s)\\ &= \mathbb{P}(X(v) = j, 0 \le v \le t+s \mid X(u) = j, 0 \le u \le s)\\ &= \mathbb{P}(X(v) = j, s < v \le t+s \mid X(s) = j) \text{ (Markov)}\\ &= \mathbb{P}(X(v) = j, 0 < v \le t \mid X(0) = j) \text{ (homogeneous)}\\ &= \mathbb{P}(T(j) > t)\end{aligned}$$

So $T(j)$ has the memoryless property, and is thus exponentially distributed (see the extensive discussion about the exponential distribution in Chapter 2).

As in discrete time, we can specify an initial probability (row) vector, $\boldsymbol{r}$ with

$$r_i = \mathbb{P}(X(0) = i).$$

But how do we specify the transition rule? Instead of a probability transition matrix P, what governs the evolution of a continuous-time Markov chain is an *infinitestimal generator* $Q = [q_{ij}]_{i,j\in\mathcal{S}}$, with components q_{ij}, for $i \neq j$, being the instantaneous transition rate of going from i to j, for $i, j \in \mathcal{S}$. In other words, for a sufficiently small interval of time $h > 0$

$$\mathbb{P}\big(X(t+h) = j \mid X(t) = i\big) = q_{ij}h + o(h),$$

where $o(h)$ goes to zero faster than h does. The matrix Q has non-positive elements on the diagonal ($q_{ii} \leq 0$ for $i \in \mathcal{S}$), and nonnegative elements off-diagonal ($q_{ij} \geq 0$ for $i \neq j$). Further, $Q\mathbf{1} = \mathbf{0}$. Since each row sums to 0 it implies that,

$$q_{j,j} = -\sum_{k\neq j} q_{j,k}.$$

A consequence is that starting at $X(0) = i$, the Markov chain stays in this state for an exponentially-distributed amount of time, with rate $-q_{ii}$, then moves to k with probability $q_{ik}/(-q_{ii})$ (see the discussion of a race between independent exponential random variables in Chapter 2). Then, it stays in state k for an exponentially-distributed amount of time, with rate $-q_{kk}$, so on. Given that the chain is in state i, the exponential duration of time the chain will state in a state and the next state to be jumped to are independent. Note that CTMCs are sometimes called Markov Jump Processes (MJP).

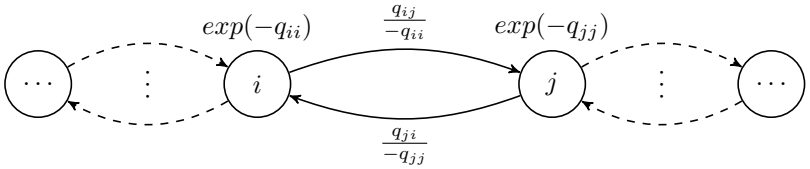


Figure 3.2: A CTMC Transition Diagram with probabilities of transitions on the arcs. The alternative is to mark transition rates on the arcs.

For small h,

$$\begin{aligned}
\mathbb{P}\big(X(t+h) = k \mid X(t) = j\big) &= p_{j,k}^{(h)} \\
&\approx (I + h\,Q)_{jk} \\
&= \begin{cases} h\,q_{jk}, & \text{if } j \neq k, \\ 1 + hq_{jj}, & \text{if } j = k. \end{cases}
\end{aligned}$$

So we can think of $q_{j,k}$ as the rate of transition from j to k, with

$$q_{j,k} \quad \text{being} \quad \begin{cases} \geq 0 \text{ if } k \neq j, \\ \leq 0 \text{ if } k = j. \end{cases}$$

The total rate of leaving state j is $\sum_{k \neq j} q_{j,k} = -q_{j,j}$, so the exponential duration of time spent in state j has parameter $\lambda_j = -q_{jj}$.

To see where the CTMC moves upon leaving state j, observe that, for $k \neq j$,

$$\mathbb{P}(X(h) = k \mid X(h) \neq j, X(0) = j) = \frac{p_{j,k}^{(h)}}{\sum_{l \neq j} p_{j,l}^{(h)}} \to \frac{q_{j,k}}{-q_{j,j}} \quad \text{as} \quad h \to 0.$$

That is, when the CTMC leaves state j, it has probability $-q_{j,k}/q_{j,j}$ of moving to state k.

An alternative way of thinking about a CTMC involves *competing alarm clocks* that ring at exponentially distributed times with different rates: for each state i, we set up a clock C_{ij} for every state $j \neq i$ to which the system can move from i in one transition. Each clock C_{ij} rings after an interval of time exponentially distributed with rate q_{ij}, and we assume that these random intervals are mutually independent. Then, from i the chain moves to whichever j whose clock C_{ij} rings first. As the minimum of independent exponential random variables is an exponential random variable, the time that system remains in i is also exponentially distributed, with rate $-q_{ii} = \sum_{i \neq j} q_{ij}$. Thanks to the memoryless property of exponential distributions, we do not need to reset the clocks after the system moves to some state $k \neq i$—we just need to consider another set of clocks C_{kj}.

Note that the index of the winner of the clocks is independent of the value it got. This allows a basic mechanism to for simulating a CTMC.

Exercise 3.4.1. *Describe how to simulate a CTMC based on generation of exponential random variables and generation of random variables I_i, $i \in \mathcal{S}$ each with support $j \in \mathcal{S} \setminus \{i\}$, distributed according to the probabilities $\{q_{i,j}/ - q_{i,i}\}$.*

Observe that

$$\begin{aligned} p_{i,j}^{(s+t)} &= \sum_{k \in \mathcal{S}} \mathbb{P}(X(s+t) = j | X(s) = k, X(0) = i) \mathbb{P}(X(s) = k | X(0) = i) \\ &= \sum_{k \in \mathcal{S}} p_{i,k}^{(s)} p_{k,j}^{(t)}. \end{aligned}$$

These are the *Chapman-Kolmogorov equations* for a CTMC. In matrix form, we write $P^{(t)} = [p_{j,k}^{(t)}]$. Then, for $s, t \geq 0$, the Chapman-Kolmogorov equations can be expressed in the form,

$$P^{(t+s)} = P^{(t)} P^{(s)}.$$

For non-explosive CTMCs, the matrix Q determines the transition probability completely by solving the backward or forward equations to get

$$P^{(t)} \;=\; e^{Qt}.$$

subject to $P^{(0)} = I$.
As a consequence,

$$\mathbb{P}\big(X(t) = j \mid X(0) = i\big) = [e^{Qt}]_{ij} \quad \text{for } i, j \in \mathcal{S}, \tag{3.18}$$

and the distribution vector $\boldsymbol{r}(t)$ with components $r_i^{(t)} = \mathbb{P}\big(X(t) = i\big)$ is given by

$$\boldsymbol{r}(t) = \boldsymbol{r}(0)e^{Qt}. \tag{3.19}$$

Notice that the matrix $e^{Q\,t}$ is stochastic, and plays a similar role as the probability transition matrix P in a discrete-time Markov chain.

There are, obviously, differences and similarities between discrete-time Markov chains (DTMC) and continuous-time Markov chains (CTMC). In both cases, given its current state, where the system will jump to next does not depend on its past trajectory; however, the time between two successive transitions is one unit of time for a DTMC, and is exponentially distributed with a state-dependent rate for a CTMC. Furthermore, the general concepts of limiting behaviours and state properties carry from discrete-time to continuous-time context, but the associated mathematical expressions differ.
The *forward equations* are:

$$\frac{d}{dt}P^{(t)} = Q\,P^{(t)},$$

and the *backward equations* are:

$$\frac{d}{dt}P^{(t)} = P^{(t)}\,Q.$$

This summary table is for discrete and continuos time:

	DTMC	**CTMC**
Unit of time	One step	"dt"
Basic info	P	Q
Distribution propagation	$P^{(\ell)} = P^{\ell}$	$P^{(t)} = e^{Q\,t}$
Evolution	geometric times+jumps	exponential times+jumps
Stationarity	$\boldsymbol{\pi}P = \boldsymbol{\pi}$	$\boldsymbol{\pi}Q = 0$

3.4.2 Further Continuous Time Properties

The concepts in discrete-time Markov chains of one state being accessible from another, of one state communicating with another, and of the chains being irreducible or reducible, are still applicable in continuous time.

A state i is said to be *absorbing* if, once entering this state, the system remains in this state forever. That is, $q_{ij} = 0$ for all $j \in \mathcal{S}$.

Definition 3.4.2. *We say that a vector $\boldsymbol{\pi} = (\pi_i)_{i\in\mathcal{S}}$ is a stationary distribution of a continuous-time Markov chain $\{X(t)\}$ if $\boldsymbol{\pi}$ satisfies the conditions*

$$\boldsymbol{\pi}Q = \mathbf{0}', \tag{3.20}$$

$$\boldsymbol{\pi}\mathbf{1} = 1. \tag{3.21}$$

One cannot define periodicity for continuous-time chains in a similar fashion to discrete-time chains; somewhat unexpectedly, this means we have a stronger version of convergence to stationary for irreducible CTMCs.

Theorem 3.4.3. *Every irreducible finite-state continuous-time Markov chain has a unique stationary distribution vector $\boldsymbol{\pi}$, which is also the limiting distribution of the chain:*

$$\lim_{t\to\infty} \boldsymbol{r}(t) = \boldsymbol{\pi}, \tag{3.22}$$

for every initial distribution $\boldsymbol{r}(0)$, where $\boldsymbol{r}(t)$ denotes the probability distribution vector at time t.

Other results regarding the stationary distribution also hold. Namely time averages and the relation between mean return time to a state and the stationary distribution.

A phenomena that may occur in CTMCs with infinite (countable) state spaces is explosion. This means that the chain makes an infinite number of transitions in finite time.

3.5 Elementary Structured Markov Models

We now consider the most basic types of structured Markov Models.

3.5.1 Birth-and-Death Processes

A continuous-time *Birth-and-Death process* is a Markov chain $\{X(t) : t \geq 0\}$ on the countably infinite state space $\mathcal{S} = \{0, 1, 2, 3, \ldots\}$ (i.e. $\mathcal{S} = \mathbb{Z}_+$) of which the generator

has the following *tridiagonal* structure

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -\mu_1 - \lambda_1 & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -\mu_2 - \lambda_2 & \lambda_2 & \dots \\ 0 & 0 & \mu_3 & -\mu_3 - \lambda_3 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.23}$$

where λ_n and μ_n are real nonnegative numbers for all n.

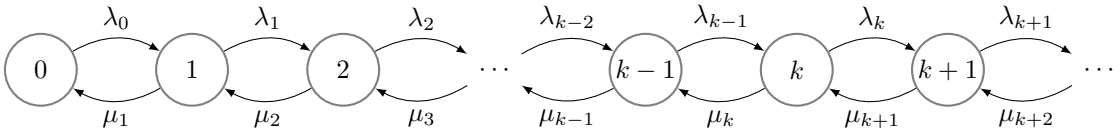


Figure 3.3: Birth Death Process on $\mathcal{S} = \{0, 1, \dots\}$

In this process, the only possible transitions from a given state n are to the state $n-1$ with rate μ_n, or to the state $n+1$ with rate λ_n. So, the process stays in state $n \geq 1$ for an exponentially distributed period of time with parameter $(\mu_n + \lambda_n)$, at the end of which it moves to state $n-1$ with probability $\mu_n/(\mu_n + \lambda_n)$ (this corresponds to a "death" event), or to state $n+1$ with probability $\lambda_n/(\mu_n + \lambda_n)$ (this corresponds to a "birth" event). When in state 0, the process moves to state 1 with probability one after an exponentially distributed period of time with parameter λ_0.

3.5.2 The Poisson Process

Before discussing a few examples of Birth-and-Death processes, we start with an important example of a *pure birth* process called the *Poisson process*.

Consider a system with no death events and where the births correspond to the arrival of events occurring *independently of each other* and such that the interarrival times are identically distributed according to an exponential distribution with parameter λ. Let $X(t)$ represent the number of arrivals in the system in the time interval $[0, t]$. The process $\{X(t) : t \geq 0\}$ is a *counting* process called a *Poisson* process with parameter (or rate) λ. It is a Markov chain with generator Q as in (4.6.1) where $\mu_n = 0$ and $\lambda_n = \lambda > 0$ for all n. Indeed, recall from Section 3.3 that for any $n \geq 0$, the entry $q_{n,n+1}$ of the generator Q is defined as

$$\mathbb{P}\big(X(t+h) = n+1 \,|\, X(t) = n\big) = q_{n,n+1}\, h + o(h), \quad \text{for small } h.$$

Then

$$\begin{aligned}\mathbb{P}[X(t+h) = n+1 \,|\, X(t) = n] &= \int_0^h (e^{-\lambda(h-u)})\,\lambda\, e^{-\lambda u} du \\ &= \lambda h\, e^{-\lambda h} \\ &= \lambda h\,(1 - \lambda\, h + o(h)) \\ &= \lambda h + o(h),\end{aligned}$$

so that $q_{n,n+1} = \lambda$ for all n. The Markov chain $\{X(t) : t \geq 0\}$ is transient, there is therefore no stationary distribution.

Theorem 3.5.1. *The number of arrivals in the system in the time interval* $[0, t]$*,* $X(t)$*, is Poisson distributed with parameter* λt*.*

Exercise 3.5.2. *Show that the above holds using the forward equations.*

The Poisson process is a natural modelling tool for a variety of phenomena such as the arrival of phone calls at a switchboard, the particles emission by a radioactive substance, the arrival of cars at a roundabout, the arrival of items at a station of a manufacturing process (see the third application example we discussed in Section 1.3), or the arrival of customers at a counter.

There are many other basic properties of a Poisson process that one often studies in an introductory stochastic processes course. We do not go into further details here, but rather list these points (further details are to appear in the Markov Chains Appendix):

- A special (central) case within the class of Renewal-Processes where the time-stationary and event-stationary versions agree.
- Can be defined as the only simple counting process that is Levy.
- The uniform (order statistic) property.
- Poisson superposition.
- Poisson splitting.

Some generalizations of the Poisson Poisson are the *compound Poisson process*, *time-varying Poisson process*, *doubly stochastic Poisson process* (Cox Process) and general Poisson processes on Metric spaces. We do not discuss these further here.

The Markovian branching process. Consider a population model where all individuals behave independently of each other and according to the same rules: each individual lives for an exponentially distributed time with parameter μ and generates new individuals during its lifetime according to a Poisson process with rate λ (that is,

there is one Poisson process per living individual). Let $X(t)$ represent the population size at time t. Then $\{X(t), t \geq 0\}$ is a Birth-and-Death process with $\mu_n = n\mu$ and $\lambda_n = n\lambda$ for $n \geq 0$.

We can show that the mean population size at time t, $m(t) = \mathbb{E}[X(t)]$, satisfies the ordinary differential equation

$$\dot{m}(t) = (\lambda - \mu)\, m(t),$$

so that $m(t) = e^{(\lambda-\mu)t}\, m(0)$. We thus see that the mean population size explodes if and only if $\lambda > \mu$.

Exercise 3.5.3. *Draw a parallel between the previous result and Example 1.1.*

Note that in a branching process in absence of immigration, the state 0 is absorbing and we can show that all other states $n \geq 1$ are transient. There is therefore no stationary distribution.

One quantity of interest in branching processes is the probability that, starting from a given state (initial population size) $n \geq 1$, the process eventually reaches the absorbing state 0 (that is, the population eventually becomes extinct). This probability is referred to the *extinction probability* of the branching process.

Branching processes form a fascinating branch of applied probability, and it is out of the scope of the present book to study them in more details here.

3.5.3 The Birth-Death Stationary Distribution.

Assume $\lambda_n, \mu_n > 0$ for all n. The infinite stationary distribution vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$ satisfies the (infinite) system of equations

$$\begin{aligned} \boldsymbol{\pi}\, Q &= \mathbf{0}', \\ \boldsymbol{\pi}\, \mathbf{1} &= 1, \end{aligned}$$

of which the solution exists if and only if the process is positive recurrent.

Theorem 3.5.4. *The stationary distribution vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$ of the Birth-and-Death process with generator (4.6.1) satisfies the recurrence*

$$\pi_n = \pi_{n-1}\, \frac{\lambda_{n-1}}{\mu_n}, \quad \textit{for } n \geq 1.$$

As a consequence, π_n can be expressed in terms of π_0 for any $n \geq 1$:

$$\pi_n = \pi_0\, \frac{\lambda_0\, \lambda_1\, \ldots\, \lambda_{n-1}}{\mu_1\, \mu_2\, \ldots\, \mu_n}. \tag{3.24}$$

The Birth-and-Death process is positive recurrent if and only if

$$\sum_{n \geq 1} \frac{\lambda_0\, \lambda_1\, \ldots\, \lambda_{n-1}}{\mu_1\, \mu_2\, \ldots\, \mu_n} < \infty. \tag{3.25}$$

Proof: The system of equations

$$\begin{aligned} \boldsymbol{\pi}\, Q &= \mathbf{0}', \\ \boldsymbol{\pi}\, \mathbf{1} &= 1, \end{aligned}$$

becomes here

$$\begin{aligned} \mu_1 \pi_1 &= \lambda_0 \pi_0, && (3.26) \\ (\mu_n + \lambda_n)\pi_n &= \lambda_{n-1}\pi_{n-1} + \mu_{n+1}\pi_{n+1}, \quad n \geq 1, && (3.27) \\ \pi_0 + \pi_1 + \pi_2 + \ldots &= 1. && (3.28) \end{aligned}$$

From (3.26) and (3.27) we obtain that the stationary probabilities satisfy the recurrence

$$\pi_n = \pi_{n-1}\, \frac{\lambda_{n-1}}{\mu_n}, \quad \text{for } n \geq 1,$$

from which the expression (3.24) of π_n in terms of π_0 is straightforward. As a consequence, (3.28) can be rewritten as

$$\pi_0 \left[1 + \sum_{n \geq 1} \frac{\lambda_0\, \lambda_1\, \ldots\, \lambda_{n-1}}{\mu_1\, \mu_2\, \ldots\, \mu_n} \right] = 1,$$

or equivalently

$$\pi_0 = \left[1 + \sum_{n \geq 1} \frac{\lambda_0\, \lambda_1\, \ldots\, \lambda_{n-1}}{\mu_1\, \mu_2\, \ldots\, \mu_n} \right]^{-1},$$

and the process is positive recurrent if and only if

$$\sum_{n \geq 1} \frac{\lambda_0\, \lambda_1\, \ldots\, \lambda_{n-1}}{\mu_1\, \mu_2\, \ldots\, \mu_n} < \infty.$$

□

3.5.4 Simple Queueing Models

Queueing models constitute some of the most basic (and usefull) examples of birth-death processes. In general, a queueing system is composed of an *arrival process*, *a service mechanism* and of other rules that describe the operation of the system.

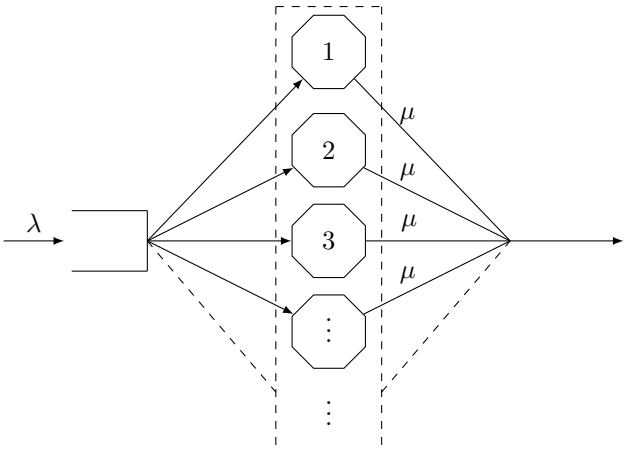


Figure 3.4: An infinite server system.

The M/M/1 queue. Consider a queueing system with a single server, in which customers arrive according to a Poisson process with rate λ and service times have an exponential distribution with parameter μ. Let $X(t)$ denote the number of customers present in the system at time t, including the one in service (if there is one). The process $\{X(t) : t \geq 0\}$ is a Birth-and-Death process with $\mu_n = \mu$ and $\lambda_n = \lambda$ for all n.
Let $\rho = \lambda/\mu$ be the *traffic intensity*; this ratio represents the average number of new arrivals in the system during the service time of one customer. From (3.24), the stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$ of the queue length satisfies

$$\pi_n = \pi_0 \, \rho^n. \tag{3.29}$$

The process is positive recurrent (or *stable*) if and only if

$$\sum_{n \geq 1} \rho^n < \infty \quad \Leftrightarrow \quad \lambda < \mu,$$

that is, if on average arrivals happen slower than service completions. In that case, from (3.28) and (3.29), π_0 satisfies

$$\pi_0 \left[1 + \sum_{n \geq 1} \rho^n \right] = 1 \quad \Rightarrow \quad \pi_0 = 1 - \rho,$$

and we finally obtain

$$\pi_n = (1 - \rho) \, \rho^n, \quad \text{for } n \geq 0. \tag{3.30}$$

We thus see that the stationary queue length has a *geometric* distribution with parameter $1 - \rho$. The steady-state mean queue length is then given by

$$\sum_{n=0}^{\infty} n \, \pi_n = \frac{\rho}{1 - \rho},$$

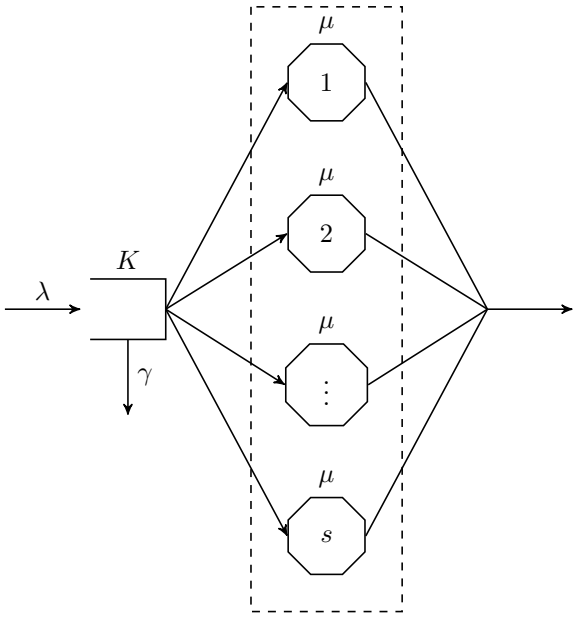


Figure 3.5: A finite service system with reneggering.

which has a vertical asymptote at $\rho = 1$.

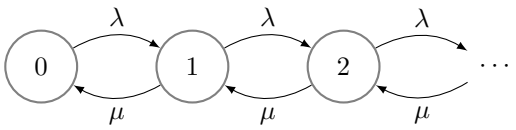


Figure 3.6: The M/M/1 Transition Diagram.

The M/M/∞ queue. Consider a queueing system with infinitely many servers operating in parallel and independently of each other, so that every arriving customer is served immediately (there is no waiting time).

This model corresponds to a Birth-and-Death process with $\lambda_n = \lambda$ and $\mu_n = n\mu$ for all n.

The M/M/c queue. Consider a multi-server queueing system with $c \geq 1$ servers operating in parallel and independently of each other, in which arrivals and service times follow the same rules as in the M/M/1 queue.

This model corresponds to a Birth-and-Death process with $\lambda_n = \lambda$ and $\mu_n = \min(n, c)\mu$ for all n.

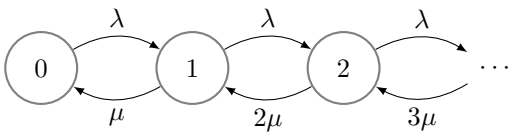


Figure 3.7: The M/M/∞ BD Chain

Exercise 3.5.5. *Determine the stationary distribution and the stability condition for the M/M/c queue.*

The M/M/c/K queue. We can assume that a queueing system with $c \geq 1$ servers has a finite capacity K (with $K \geq c$). If the state of the system is K, every new arrival is considered as lost.

Exercise 3.5.6. *Show that this queueing model corresponds to a Birth-and-Death process (by specifying λ_n and μ_n for all n), and determine its stationary distribution and the stability condition.*

Other birth-death Queueing Systems

The M/M/s/K+M feature customer abandonments at rate γ. The generator matrix is of this form:

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ s \\ (s+1) \\ (s+2) \\ \vdots \\ (s+K-1) \\ (s+K) \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & \cdots & s & (s+1) & (s+2) & \cdots & (s+K-1) & (s+K) \end{array} \\ \left[\begin{array}{cccccccccccc} -\lambda & \lambda & . & . & \cdots & . & . & . & \cdots & . & . \\ \mu & -(\lambda+\mu) & \lambda & . & \cdots & . & . & . & \cdots & . & . \\ . & 2\mu & -(\lambda+2\mu) & \lambda & \cdots & . & . & . & \cdots & . & . \\ . & . & 3\mu & -(\lambda+3\mu) & \ddots & . & . & . & \ddots & . & . \\ \hline \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \hline . & . & . & . & \ddots & -(\lambda+s\mu) & \lambda & . & \ddots & \vdots & \vdots \\ . & . & . & . & \cdots & (s\mu+\gamma) & -(\lambda+s\mu+\gamma) & \lambda & & . & . \\ . & . & . & . & \ddots & . & (s\mu+2\gamma) & -(\lambda+s\mu+2\gamma) & \ddots & . & . \\ \hline \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \hline . & . & . & . & \ddots & . & . & . & \ddots & -(\lambda+s\mu+(K-1)\gamma) & \lambda \\ . & . & . & . & \cdots & . & . & . & \cdots & (s\mu+K\gamma) & -(s\mu+K\gamma) \end{array}\right] \end{array}$$

Assume now that the system starts empty. Then the transient distributions follow:

Exercise 3.5.7. *Consider a finite population queue: Inhabitants of an Island with a population of* 15 *occasionally go to use an internet stand on the Island, queueing when the stand is occupied. The rate of desire to use the stand is λ. The service rate is μ. Assume exponential service times (and further typical assumptions). What is the stationary distribution?*

Exercise 3.5.8. *Compare the above to the Erlang Loss System.*

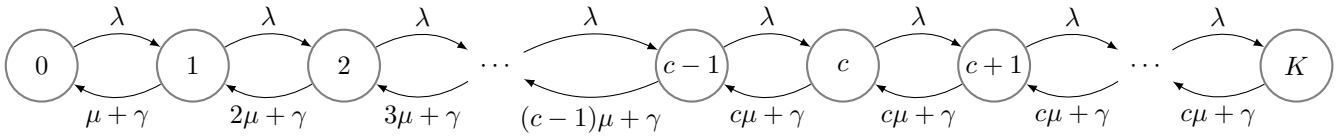


Figure 3.8: The M/M/c/K+M BD Chain

3.6 (A, B, C, D) Linear Input-Output Systems

In Chapter 2 we looked at input-output LTI systems (SISO versions). Then earlier in the current chapter (starting in Section 3.1) we looked at linear dynamical systems. In that case we did not consider input and output, instead we considered the *state* of the process. We now combine the two types of objects to get systems that we call (A, B, C, D) *Linear Input-Output Systems*. Among other things, these systems will serve as the basis for linear control theory to be studied in Chapter 5

These systems relate 3 processes: *input* $\boldsymbol{u}(\cdot)$, *state* $\boldsymbol{x}(\cdot)$ and *output* $\boldsymbol{y}(\cdot)$. As their name[5] suggests, (A, B, C, D) systems are parameterized by 4 matrices: $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. So the dimension of the state is n, the dimension of the input is m and the dimension of the output is p. The SISO case is when $m = 1$ and $p = 1$ (yet does not require that $n = 1$).

The *input – state – output* evolution of such systems is defined as follows:

$$\begin{array}{rcl} \dot{\boldsymbol{x}}(t) &=& A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &=& C\boldsymbol{x}(t) + D\boldsymbol{u}(t) \end{array} \quad \text{or} \quad \begin{array}{rcl} \boldsymbol{x}(\ell+1) &=& A\boldsymbol{x}(\ell) + B\boldsymbol{u}(\ell) \\ \boldsymbol{y}(\ell) &=& C\boldsymbol{x}(\ell) + D\boldsymbol{u}(\ell) \end{array}, \tag{3.31}$$

for the continuous time and discrete time cases respectively. Our presentation here focuses primarily on the continuous time case. Observe that if $B = 0$ and/or $\boldsymbol{u}(\cdot) \equiv 0$, then the *state evolution* is as that of an autonomous linear dynamical system (as the systems presented in Section 3.1). Yet if $B \neq 0$ and $\boldsymbol{u}(\cdot)$ takes non-zero values then the state-evolution is modified/altered by the input.

In control theory applications (handled in Chapter 2), the matrix A indicates the "untouched behavior of the *plant*", the matrix B indicates the "effect of *actuators* on the plant", the matrix C indicates the "*sensors* in the system" and the matrix D indicates the "effect that the input has directly on the output".

Observe that if $C = I$ and $D = 0$ then $\boldsymbol{y}(\cdot) = \boldsymbol{x}(\cdot)$. I.e. the output of the system is exactly the state. Yet in applications, C is typically a "flat long" matrix ($p < n$) while B is a "tall thin" matrix ($n > m$). Such dimensions represent the fact that "there are not many sensors" and "there are not many actuators" respectively ("not many" is compared to the number of state variables). In the extreme SISO case, C is in fact a row vector (which we shall denote by $\boldsymbol{c}'$) and B is a column vector (which we shall denote by $\boldsymbol{b}$).

[5]This is a name we have given, it is not a standard name.

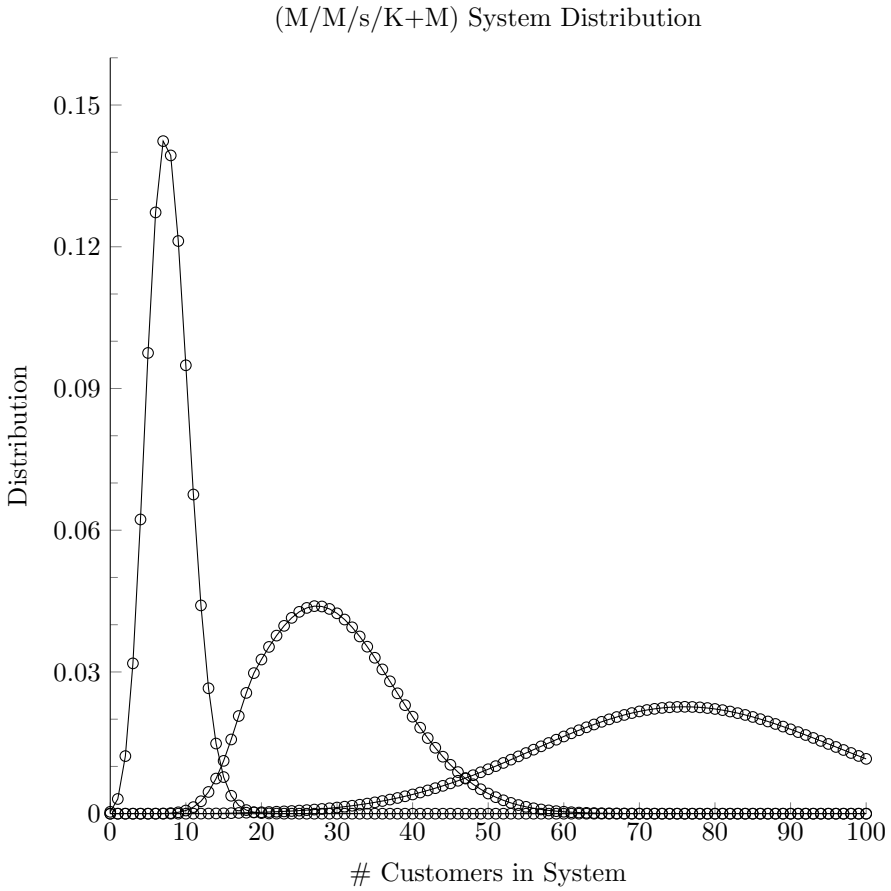


Figure 3.9: Evolution of $\boldsymbol{r}(t)$, for $t = [10, 100, 100000]$.

Further, d is a scalar. In that case the $(A, \boldsymbol{b}, \boldsymbol{c}', d)$ system dynamics are,

$$\begin{array}{rcl} \dot{\boldsymbol{x}}(t) & = & A\boldsymbol{x}(t) + \boldsymbol{b}u(t), \\ y(t) & = & \boldsymbol{c}'\boldsymbol{x}(t) + du(t), \end{array} \tag{3.32}$$

or similarly for the discrete time case.

3.6.1 Time-Domain Representation of the Output

Specify now initial conditions,

$$\boldsymbol{x}(0) = \boldsymbol{x}_0.$$

If $\boldsymbol{x}_0 = \mathbf{0}$ we say the system is *starting at rest*. We now have,

Theorem 3.6.1.

$$\boldsymbol{y}(t) = Ce^{At}\boldsymbol{x}_0 + C\int_0^t e^{A(t-s)}B\boldsymbol{u}(s)ds + D\boldsymbol{u}(t), \tag{3.33}$$

$$\boldsymbol{y}(\ell) = CA^{\ell}\boldsymbol{x}_0 + C\sum_{k=0}^{\ell-1} A^{\ell-(k+1)}B\boldsymbol{u}(k) + D\boldsymbol{u}(\ell). \tag{3.34}$$

The continuous time equation (3.33) can be obtained from Picard iterations. The discrete time equation is easily obtained by recursing the stem equations:

Exercise 3.6.2. *Prove* (3.34).

We can now verify using Theorem 3.6.1 that the mapping,

$$\boldsymbol{y}(\cdot) = \mathcal{O}\big(\boldsymbol{u}(\cdot)\big)$$

is LTI if the system is starting at rest. That is, (A, B, C, D) systems yield linear time invariant input output systems (in the sense of Chapter 2).

In this MIMO-LTI setting the impulse response generalizes to the *matrix impulse response*. Focusing on the continuous time version, we assume it admits an *integral representation*,

$$\boldsymbol{y}(t) = \mathcal{O}\big(\boldsymbol{u}(\cdot)\big)(t) = \int_{-\infty}^{\infty} h(t-\tau)\boldsymbol{u}(\tau)d\tau = (h * \boldsymbol{u})(t),$$

with $h(t) \in \mathbb{R}^{p\times m}$ being the *impulse response matrix*. Note that for inputs $\{\boldsymbol{u}(t)\}$ that have coordinates 0 except for the j'th coordinate, $u_j(t)$, the i'th component of the output has the form,

$$y_i(t) = \int_{-\infty}^{\infty} h_{ij}(t-\tau)u_j(\tau)d\tau,$$

as a SISO system with impulse response $h_{ij}(t)$.

Any MIMO LTI system is said to be *causal* if and only if $h(t) = 0_{p\times n}$ for $t < 0$ and thus for inputs with positive support,

$$\boldsymbol{y}(t) = \int_0^t h(t-\tau)\boldsymbol{u}(\tau)d\tau.$$

We can further get the following useful representations:

$$h(t) = \mathbf{1}_{p\times p}(t)\Big(Ce^{At}B + D\delta_{m\times m}(t)\Big), \tag{3.35}$$

where we use a diagonal matrix of m delta-functions, $\delta_{m\times m}(t)$.

Exercise 3.6.3. *Argue the validity of* (3.35) *based on Theorem 3.6.1.*

In the SISO case, (3.35) reads,

$$h(t) = \boldsymbol{c}' e^{At} \boldsymbol{b} + d\delta(t). \tag{3.36}$$

Further in this case, the step response is:

$$H(t) = \int_0^t h(s)ds = d + \boldsymbol{c}' \Big(\int_0^t e^{As} ds \Big) \boldsymbol{b}.$$

Hence when A is non-singular, the step response in the SISO case reads:

$$H(t) = d - \boldsymbol{c}' A^{-1} \boldsymbol{b} + \boldsymbol{c}' e^{At} A^{-1} \boldsymbol{b}. \tag{3.37}$$

3.6.2 The Transfer Function Matrix

The relation of convolutions and Laplace transforms carries over easily to the non-scalar version here. If the matrix Laplace transform, $\hat{h}(s)$ of $h(\cdot)$ exists then,

$$\hat{\boldsymbol{y}}(s) = \hat{h}(s)\hat{\boldsymbol{u}}(s).$$

A matrix Laplace transform such as this is simply a Laplace transform of each of the elements. In this case, $\hat{h}(s)$ is the *transfer function matrix.*

Building on the idea of the resolvent, the transfer function takes on a very specific form for (A, B, C, D) systems. We can extend the resolvent to (A, B, C, D) systems by mimicking (3.13), this time for $\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t)$ (starting at rest):

$$s\hat{\boldsymbol{x}}(s) - \boldsymbol{0} = A\hat{\boldsymbol{x}} + B\hat{\boldsymbol{u}}.$$

This yields (for s values that are not eigenvalues of A):

$$\hat{\boldsymbol{x}}(s) = (sI - A)^{-1} B\hat{\boldsymbol{u}}(s).$$

Substitution in $\boldsymbol{y}(t) = C\boldsymbol{x}(t) + D\boldsymbol{u}(t)$ we get,

$$\hat{\boldsymbol{y}}(s) = \big(C(sI - A)^{-1} B + D\big)\hat{\boldsymbol{u}}(s).$$

We have thus derived the following representation of the transfer function matrix for (A, B, C, D) systems:

Theorem 3.6.4.

$$\hat{h}(s) = C(sI - A)^{-1} B + D. \tag{3.38}$$

In the SISO case, (3.38) *is a scalar function and reads.*

$$\hat{h}(s) = \boldsymbol{c}'(sI - A)^{-1} \boldsymbol{b} + d. \tag{3.39}$$

Exercise 3.6.5. *Explain why the elements of* (3.38) *as well as* (3.39) *are rational functions.*

We have just shown that all (A, B, C, D) systems have rational Laplace transforms. We further have the following (without proof):

Theorem 3.6.6. *Any matrix of rational functions (or a single scalar rational function) treated as a transfer function, is the transfer function of an* (A, B, C, D) *system.*

The action of finding an (A, B, C, D) system that has a given rational transfer function is called *realization* (do not confuse this with the other meaning of the word "realization" that is synonymous with a *sample path* of a random process). In practice one often tries to *realize* a system by choosing "physical components" that have a given behavior (specified by the transfer function). A bit more on this is in Chapter 5.

Note that there is not a unique (A, B, C, D) system corresponding to a transfer function. This is illustrated now through equivalent representations.

3.6.3 Equivalent Representations of Systems

Given $P \in \mathbb{R}^{n\times n}$, with $det(P) \neq 0$, we can *change the coordinates* of the state-space as follows:

$$P\tilde{x} = x.$$

By substitution in the system equations, we see that resulting system is,

$$\left(\tilde{A},\ \tilde{B},\ \tilde{C},\ \tilde{D}\right) = \left(P^{-1}AP,\ P^{-1}B,\ CP, D\right). \tag{3.40}$$

Both systems have the same external representations (i.e. same impulse response/transfer function) and are thus called *equivalent systems.*

Note that the matrices A and $\tilde{A}$ are similar, hence making such a change of coordinates is sometimes referred to as performing a *similarity transform* on the system.

Exercise 3.6.7. *Prove* (3.40) *using either Theorem 3.6.1 or Theorem 3.6.4.*

3.6.4 Rational Laplace-Stieltjes Transforms Revisited

In Chapter 2 we saw the straight forward connection between LTI input output systems and probability distributions. That chapter ended with a few example probability distributions whose Laplace transform is rational. In the previous section we have seen that all rational transfer functions may be *realized* as (A, B, C, D) systems. The class of such systems whose step response, $H(t)$ is a probability distribution is called a *matrix exponential distribution* (ME). We define this in detail now.

Given $A \in \mathbb{R}^{n\times n}$ with $det(A) \neq 0$, vectors $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$ and a scalar $c_0 \geq 0$, consider the following function:

$$F(t) = \begin{cases} 0, & t < 0, \\ d, & t = 0, \\ d - \boldsymbol{c}' A^{-1}\boldsymbol{b} + \boldsymbol{c}' e^{At} A^{-1}\boldsymbol{b}, & t > 0. \end{cases}$$

If $F(t)$ satisfies the properties of a distribution then $F(\cdot)$ is said to be a *matrix exponential distribution* of order n.

The Laplace-Stieltjes transform (LST) is given by

$$\hat{F}(s) \;=\; \boldsymbol{c}'(sI - A)^{-1}\,\boldsymbol{b} + d. \tag{3.41}$$

We have the following:

Theorem 3.6.8. *If a distribution has a rational LST then it can be represented as a ME distribution.*

In the next section we look at a special sub-set of ME distributions whose parameters bear probabilistic meaning.

Note: Given $\hat{F}(\cdot)$ that is rational. It is not a trivial matter to check if it corresponds to a distribution (i.e. it is not easy to verify if the corresponding step-response is monotonic). If we know it corresponds to a distribution, then that distribution is ME.

3.7 Phase–Type (PH) Distributions

Having seen the family of matrix exponential distributions we now define a sub-class of these distributions whose construction is based on hitting times of Markov chains. We call such distributions phase–type distributions.

Phase–type distributions have had a profound effect on applied probability and stochastic modeling in the past few decades. As will be demonstrated in the next chapter, they essentially allow to incorporate behaviors of arbitrary distributions in stochastic models that are governed by CTMCs (remember that the "basic distribution" in CTMCs is the exponential distribution – this is quite restrictive from a modeling point of view).

3.7.1 The Absorption Time in an Absorbing CTMC

Consider a CTMC $\{X(t) : t \geq 0\}$ on the finite state space $\mathcal{S} = \{0, 1, 2, \ldots, m\}$, where state 0 is absorbing and states $\{1, 2, \ldots, m\}$ are transient. We denote the generator of the CTMC as the $(m+1) \times (m+1)$ matrix Q and define it shortly.

Let τ denote the hitting time of state 0:

$$\tau := \inf\{t : X(t) = 0\}.$$

Regardless the initial (transient) state, the Markov chain will eventually hit the absorbing state in a finite time with probability one, therefore $\tau < \infty$ almost surely (that is, the distribution of τ is *nondefective* or *proper*).

To construct, Q, take the vector $\boldsymbol{c} \in \mathbb{R}_+^m$ with $\boldsymbol{c}'\mathbf{1} \leq 1$ and let $(c_0, \boldsymbol{c}')'$ denote the probability distribution (row vector) of $X(0)$. That is,

$$c_0 = 1 - \boldsymbol{c}'\mathbf{1},$$

and,

$$\mathbb{P}\big(X(0) = i\big) = c_i.$$

Now take $A \in R^{m\times m}$ with $det(A) \neq 0$, negative entries on the diagonal positions, nonnegative entries on the off-diagonal positions and $A\mathbf{1} \leq 0$. Such a matrix is called a *sub-generator*. Construct now Q as follows:

$$Q = \left[\begin{array}{c|c} 0 & \mathbf{0}' \\ \hline \boldsymbol{b} & A \end{array}\right],$$

where the column vector $\boldsymbol{b} := -A\mathbf{1}$ can be interpreted as the *absorption rate vector* (in state 0).

Exercise 3.7.1. *Argue why Q is a generator matrix where state 0 is absorbing (and thus recurrent) and states $\{1, \ldots, m\}$ are transient.*

We call the distribution of the random variable τ, a *phase type* distribution. It is parameterized by $\boldsymbol{c}$ and A. This is because for every $\boldsymbol{c}$ and A we have a CTMC and every CTMC implies the behavior of the hitting time random variable. We thus use the notation $PH(\boldsymbol{c}', A)$.

Let $F(t) := \mathbb{P}(\tau \leq t)$ denote the distribution function of $PH(\boldsymbol{c}', A)$. We have the following:

Theorem 3.7.2.

$$F(t) = \begin{cases} 0, & t < 0, \\ c_0, & t = 0, \\ 1 - \boldsymbol{c}'\, e^{At}\, \mathbf{1}, & t > 0. \end{cases}$$

Exercise 3.7.3. *Show that the density (excluding the possible atom c_0 at 0) is*

$$f(t) = \boldsymbol{c}'\ e^{At}\, \boldsymbol{b}.$$

Exercise 3.7.4. *Show that the LST is*

$$\hat{f}(s) = c_0 + \boldsymbol{c}'(sI - A)^{-1}\, \boldsymbol{b}.$$

Using the LST, it is a standard matter to obtain the moments:

Theorem 3.7.5. *Let $\tau \sim PH(\boldsymbol{c}', A)$, then*

$$\mathbb{E}[\tau^k] = (-1)^k\ k!\ \boldsymbol{c}'\ A^{-k}\ \mathbf{1}.$$

3.7.2 Examples

- An exponential distribution with parameter λ is a very special case of a PH distribution with $n = 1$, $c = 1$, $A = -\lambda$, and $b = \lambda$.

- Let Z be distributed according to an Erlang $E(n, \lambda)$; in other words, Z represents the sum of n independent exponential random variables with parameter λ. Then the distribution of Z can be seen as particular PH distribution of order n, that is, $Z \sim PH(\boldsymbol{c}, A)$ with

$$\boldsymbol{c}' = [1, 0, \ldots, 0], \quad A = \begin{bmatrix} -\lambda & \lambda & & \\ & -\lambda & \lambda & \\ & & \ddots & \\ & & & -\lambda \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \lambda \end{bmatrix}.$$

- Let Z be distributed according to an hyperexponential distribution with density

$$f_Z(z) = \sum_{1 \le k \le n} c_k \, \lambda_k e^{-\lambda_k z}, \qquad \text{where } c_k > 0 \text{ for all } k, \text{ and } \sum_{1 \le k \le n} c_k = 1.$$

  Then Z is the convex mixture of n exponential random variables, and $Z \sim PH(\boldsymbol{c}', A)$ with

$$\boldsymbol{c}' = [c_1, c_2, \ldots, c_n], \quad A = \begin{bmatrix} -\lambda_1 & & & \\ & -\lambda_2 & & \\ & & \ddots & \\ & & & -\lambda_n \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix}.$$

3.7.3 A Dense Family of Distributions

Any non-negative distribution may be approximated by a PH distribution. The approximation becomes better as the number of phases grows. In fact, if the space of all non-negative distributions is taken as a metric space (one needs to define a metric for this – we omit the details), it can be shown that the set of PH distributions is *dense* in that space (analogy: the rational numbers are dense in the reals).

Exercise 3.7.6. *Think how to approximate an arbitrary non-negative distribution by using mixtures of Erlang distributions where the number of phases in the Erlang distributions is large.*

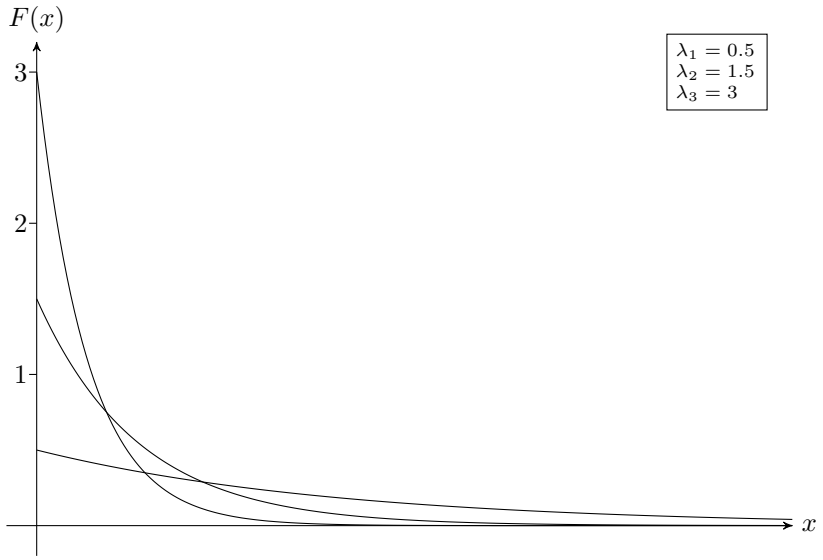


Figure 3.10: Exponential Distribution PDF

3.7.4 Relationship to ME Distributions

Note that PH distributions form a special class of ME distributions since both distributions share the same structural properties. Yet for PH distributions, the vectors $\boldsymbol{c}$ and $\boldsymbol{b}$, and the matrix A characterizing the PH distribution are respectively probability mass vectors and a sub-generator matrix (i.e. they have probabilistic meaning). As opposed to that, the parameters of ME distributions bear no probabilistic meaning.

Exercise 3.7.7. *In what way is a PH distribution a special case of an ME distribution? That is, given a $PH(\boldsymbol{c}', A)$ distribution function, represent it as an $ME(\cdot)$ distribution function.*

Theorem 3.7.8. *There exists distributions that are ME, yet are not PH.*

3.7.5 Operations on PH Random Variables

The class of PH distributions is closed with respect to:

1. Multiplication by a constant
2. Addition (of independent random variables)
3. Mixtures
4. Minimum

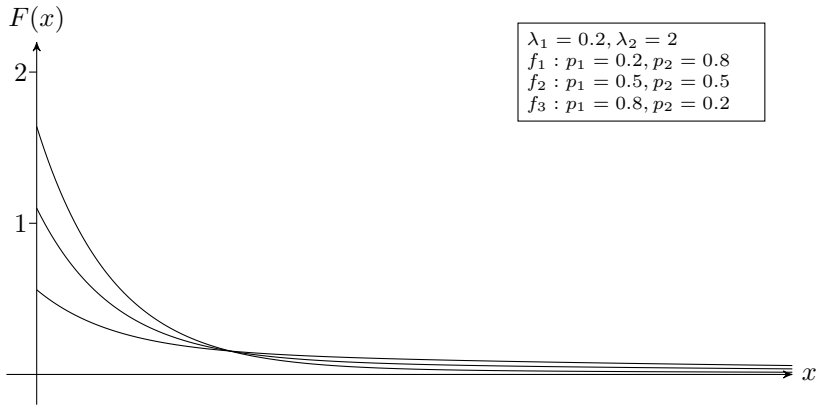


Figure 3.11: Hyperexponential Distribution PDF

3.7.6 Moment Matching

When modelling, we are often faced with the task of choosing a distribution based on some observed moments. In the simplest case we are given the mean and perhaps the variance. An alternative view is to use the squared coefficient of variation c^2. How can we match PH distributions for this?

$c_v^2 < 1$ Generalized Erlang Disitribution Paramatrization

When processes with coefficients under 1 are considered, they can be modelled with the help of Generalized Erlang Distributions. This type of distribution is different from regular Erlang Distributions in the sense that it does not use n exponentially distributed steps with parameter μ, but one step with parameter μ_1 and (n-1) steps with parameter μ_2. These means are as follows:

$$\mu_1 = \frac{n}{1 + \sqrt{(n-1)(nc^2 - 1)}}$$
$$\mu_2 = \mu_1 \frac{n-1}{\mu_1 - 1}$$

In which the number of steps is based on c_v^2 as follows: $n = \frac{1}{c_v^2}$.

$c_v^2 > 1$ Hyper Exponential Paramatrization

In order to generate distributions with a coefficient of variation $c_v^2 > 1$, a Hyper Exponential distribution can be used. In order to visualise the effects of an increasing c_v^2 on

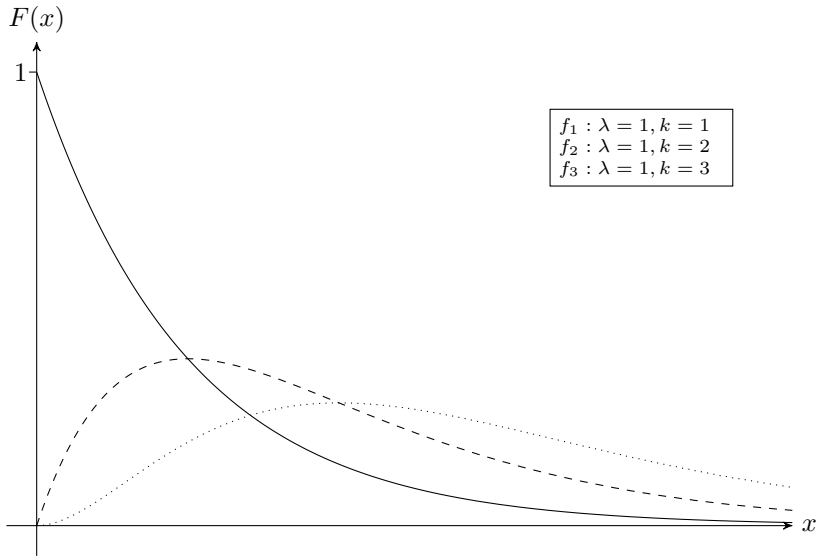


Figure 3.12: Erlang Distribution PDF

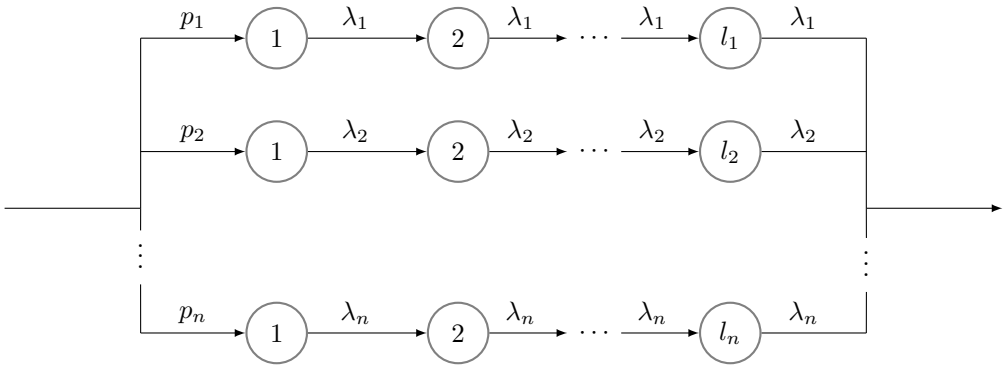


Figure 3.13: Mixture of Erlang Random Variables

the queue length, the mean of the service process is fixed to be 1. By then varying the mean of the arrival process up to 1, it is possible to generate values for the queue length for various utilizations of the system.

In the Hyper Exponential case, two nodes are considered with rate parameters μ_1 and μ_2. The first node is entered with probability p, while the second is entered with probability $1-p$. As mentioned before, the mean of this process is fixed to be 1. This mean can be expressed as a function of the means of the nodes as follows:

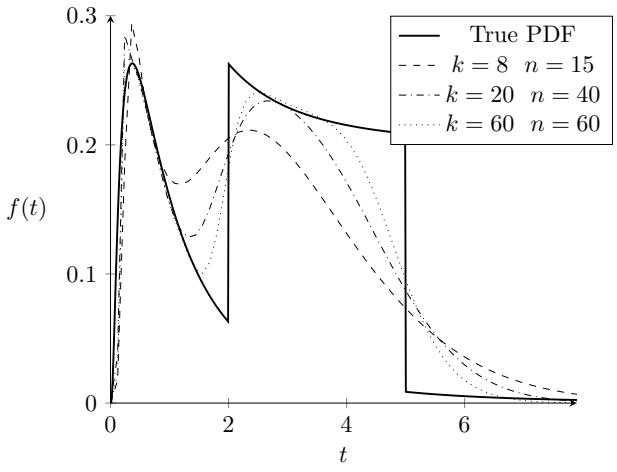


Figure 3.14: A mixture of a Uniform random variable on (2,5), with p=0.7, and a Log Normal with $\mu = 0$ and $\sigma = 1$, with $1 - p = 0.3$ is approximated by k Erlangs, each the sum of n exponentials.

$$\frac{1}{\mu} = 1 = \frac{p}{\mu_1} + \frac{(1-p)}{\mu_2} \tag{3.42}$$

Additionally, the coefficient of variation is a function of the variance and the mean: $c_v^2 = \frac{\sigma^2}{\mu^2}$. However, with the mean set to be one, this equation simplifies to the variance alone. The formula can then be expressed as follows:

$$c^2 = 2\left(\frac{p}{\mu_1^2} + \frac{(1-p)}{\mu_2^2}\right) - 1 \tag{3.43}$$

$$\frac{c^2+1}{2} = \frac{p}{\mu_1^2} + \frac{(1-p)}{\mu_2^2} \tag{3.44}$$

From 3.42 the following expression for μ_2 is derived:

$$\mu_2 = \frac{(1-p)}{(1-\frac{p}{\mu_1})} \tag{3.45}$$

Next, from 3.43 an expression for p can be derived, by filling in the previous equation:

$$p = \frac{(c_v^2 - 1)}{(c_v^2 + 1 + \frac{2}{\mu_1^2} - \frac{4}{\mu_1})} \tag{3.46}$$

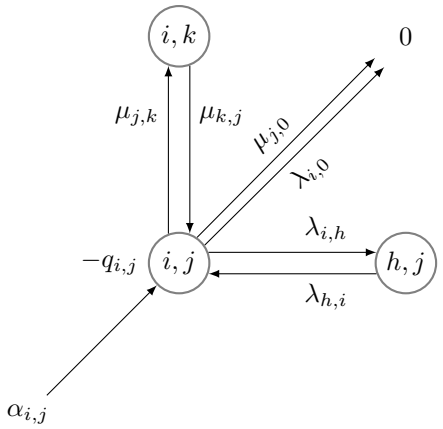


Figure 3.15: Transitions of a State $(i, j) \in S_1 \times S_2$

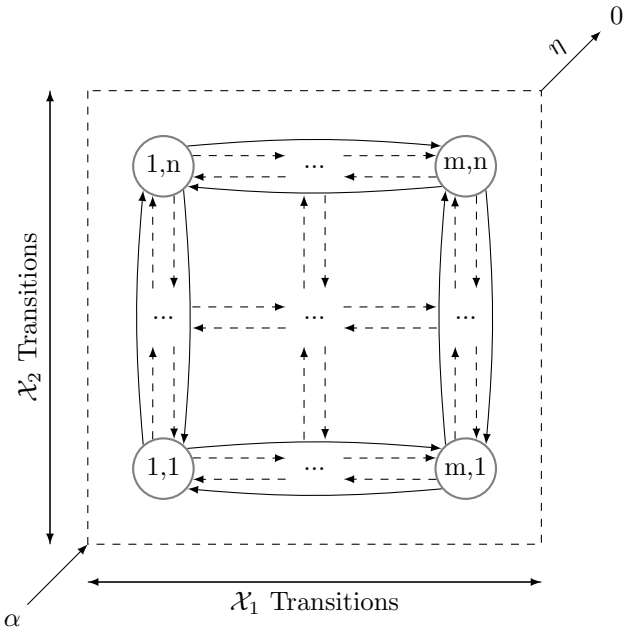


Figure 3.16: Two Dimensional Representation of Minimum of Two Phasetypes

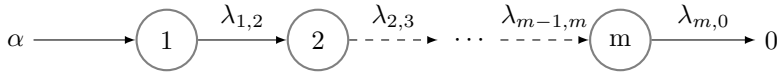


Figure 3.17: PH Representation of Generalised Erlang Random Variable

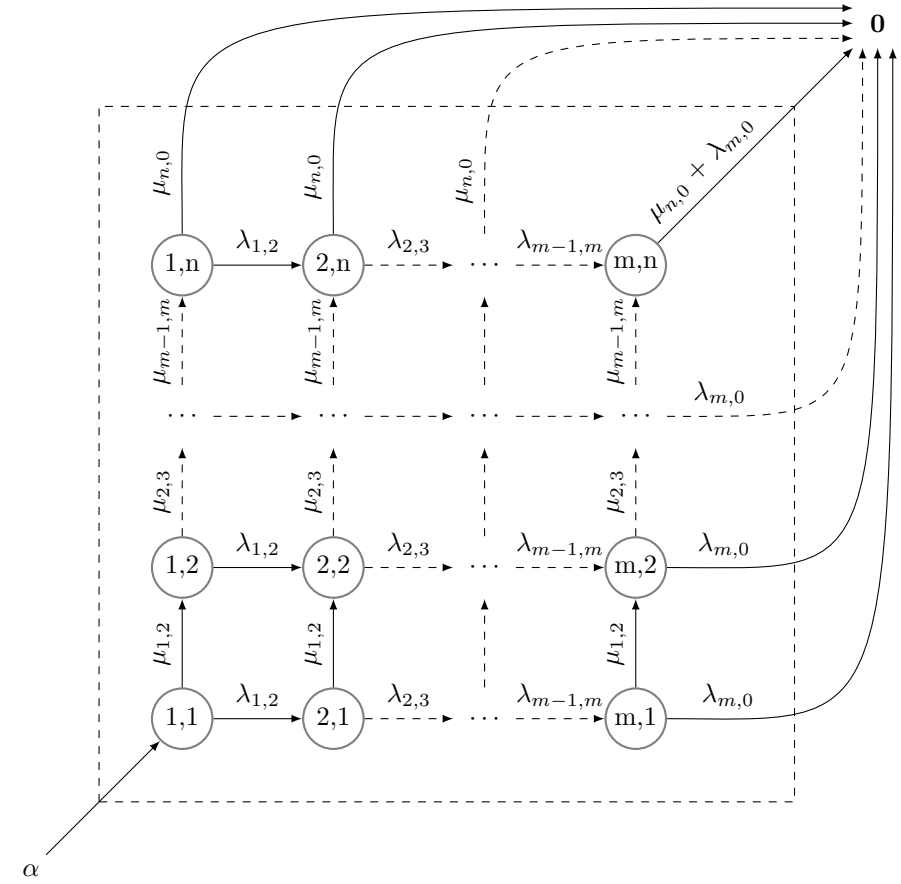


Figure 3.18: Two Dimensional PH Representation of Minimum of Two Generalised Erlang Random Variables

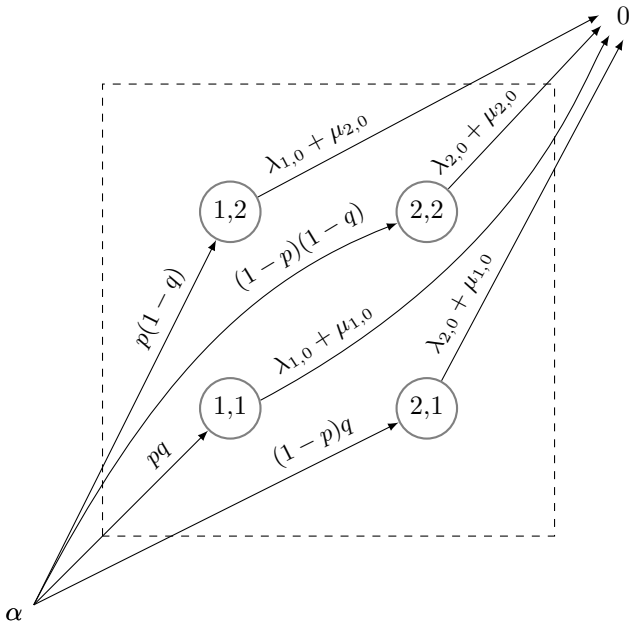


Figure 3.19: Two Dimensional PH Representation of Minimum of Two Hyperexponential Random Variables

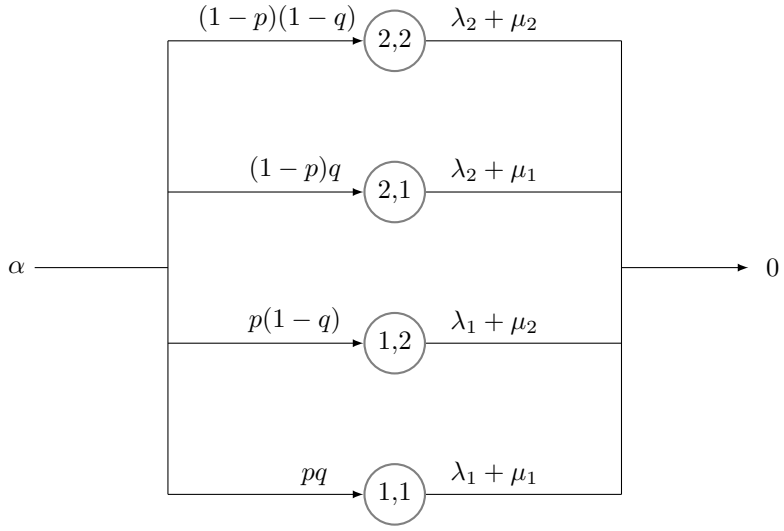


Figure 3.20: The Minimum of Two Hyperexponential Random Variables is also Hyperexponential

This expression can then be used on 3.45 to derive an expression that solely depends on μ_1 and c^2. This equation, together with 3.46, yields the following set of expressions:

$$p = \frac{(c_v^2 - 1)}{(c_v^2 + 1 + \frac{2}{\mu_1^2} - \frac{4}{\mu_1})} \tag{3.47}$$

$$\mu_2 = \frac{2(1 - \mu_1)}{2 - \mu_1(c_v^2 + 1)} \tag{3.48}$$

Here it should be noted that an extra restriction on μ_1 should be imposed as $2 - \mu_1(c_v^2 + 1) > 0$. This restriction results in the following inequality:

$$\mu_1 < \frac{2}{c_v^2 + 1} \tag{3.49}$$

This set of equations then allows us to choose values for any given c_v^2. As an example one can consider a process with mean $\mu = 1$ and $c_v^2 = 7$. The inequality then shows that $\mu_1 < \frac{1}{4}$, so $\mu_1 = \frac{1}{5}$ for example. Next $p = \frac{3}{19}$ and $\mu_2 = 4$ are obtained. These values can be verified by filling them in for the definitions of the mean and coefficient of variation:

$$\mu = \frac{p}{\mu_1} + \frac{(1-p)}{\mu_2} = \frac{15}{19} + \frac{4}{19} = 1$$

$$c_v^2 = 2\begin{bmatrix} p & (1-p) \end{bmatrix} \begin{bmatrix} \frac{1}{\mu_1^2} & 0 \\ 0 & \frac{1}{\mu_2^2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 1$$

$$c_v^2 = 2\left(\frac{\frac{3}{19}}{\left(\frac{1}{5}\right)^2} + \frac{\frac{16}{19}}{16} \right) - 1 = 2\left(\frac{76}{19}\right) - 1 = 7$$

It should be noted that any value for c_v^2 can be approximated with the aforementioned μ_1, as long as $c_v^2 < \left(\frac{2-\mu_1}{\mu_1}\right)$ holds. Alternatively, a very small μ_1 allows for a large range of c_v^2 to be approximated.

3.8 Relations Between Discrete and Continuous Time

We now explore relationships between CTMCs and related DTMCs as well as between (A, B, C, D) systems and PH/ME distributions of both discrete and continuous time.

3.8.1 Different Discretizations of a CTMC

When considering a CTMC, $\{X(t)\}$ with generator matrix Q there are several ways in which we can associate a DTMC, $\{\tilde{X}(\ell)\}$ with this DTMC. The first way is *discrete*

time sampling. Here the DTMC is the CTMC sampled every T time units, that is,

$$\tilde{X}(\ell) = X(\ell\, T).$$

In this case, we have that the transition probability kernel of the DTMC $\tilde{X}(\cdot)$ is,

$$P_1 = e^{Q\,T}.$$

A second way is to consider the so called *embedded Markov chain* (also known as the *jump chain*). If the CTMC is irreducible we have that $q_{i,i} < 0$ (strictly) and thus we can define,

$$P_2 = I - \mathrm{diag}(Q)^{-1} Q.$$

This is basically a stochastic matrix where for $i \neq j$, $P_{i,j} = q_{i,j}/ - q_{i,i}$ and for $i = j$ we have $P_{i,j} = 0$. More generally if we have that $q_{i,i} = 0$ for some i (this corresponds to an absorbing state i in the CTMC), we should set in the embedded Markov chain, $P_{i,i} = 1$. In summary, the embedded Markov chain represents the CTMC sampled at jump points.

A third way corresponds to CTMCs where there is an upper bound to $\{-q_{i,i}\}$:

$$\max_{i \in \mathcal{S}} -q_{i,i} \leq \gamma.$$

This always holds when $\mathcal{S}$ is finite. In this case, a *uniformized chain* is a DTMC with,

$$P_3 = I + \frac{1}{\gamma} Q.$$

Exercise 3.8.1. *Show that P_1, P_2 and P_3 are stochastic matrices.*

Note that in the uniformized chain, transitions from a state to itself are possible (in a single discrete time step). This is not the case for the embedded Markov chain. The idea of uniformization is to have a single Poisson process at rate γ that marks the transitions of all types. Whenever the "clock" of this Poisson process "rings" a transition is made according to P_3 (sometimes allowing transitions from a state to itself).

Let $\boldsymbol{\pi}$ be the stationary distribution of P_3. Then,

$$\boldsymbol{\pi}(I + \frac{1}{\gamma} Q) = \boldsymbol{\pi},$$

and hence,

$$\boldsymbol{\pi} Q = \mathbf{0}.$$

So the stationary distribution of P_3 and Q is the same. This in general does not hold for P_2 nor for P_1:

Exercise 3.8.2. *Show by example that P_2 and P_1 in general have a different stationary distribution than Q.*

3.8.2 Sampling a Continuous Time (A, B, C, D) System

Consider a continuous time (A, B, C, D) system that is *sampled* at time intervals of T. Assume that a piecewise constant input is applied: $u(t) = u(\ell)$ for $t \in [\ell T, (\ell + 1)T)$. In this case, the discrete time system,

$$\left(\tilde{A},\ \tilde{B},\ \tilde{C},\ \tilde{D}\right) := \left(e^{AT},\ \int_0^T e^{A\tau} d\tau B,\ C,\ D\right),$$

agrees with the continuous time system (A, B, C, D) at the sampling points.

Exercise 3.8.3. *Prove this.*

3.8.3 Discrete/Continuous, PH/ME Distributions Relationships

In similar vein to Markov chains and (A, B, C, D) systems, there are obvious relationships between discrete and continous PH/ME distributions. At this point we only focus on the exponential and geometric distribution:

Exercise 3.8.4. *Let $X \sim \exp(\lambda)$. Denote $N = \lfloor X \rfloor$. Show that N is geometrically distributed and find its parameter.*

Bibliographic Remarks

Exercises

Suppose $\mathbb{P}(X(0) = 1) = 1/3$, $\mathbb{P}(X(0) = 2) = 0$, $\mathbb{P}(X(0) = 3) = 1/2$, $\mathbb{P}(X(0) = 4) = 1/6$ and

$$P = \begin{pmatrix} 1/4 & 0 & 1/4 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 2/3 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix}.$$

- Find the distribution of $X(1)$,
- Calculate $\mathbb{P}(X(\ell + 2) = 2 | X(\ell) = 4)$, and
- Calculate $\mathbb{P}(X(3) = 2, X(2) = 3, X(1) = 1)$.

Sometimes we want to model a physical system where the future does depend on part of the past. Consider following example. A sequence of random variables $\{X_n\}$ describes the weather at a particular location, with $X_n = 1$ if it is sunny and $X_n = 2$ if it is rainy on day n.

Suppose that the weather on day $n+1$ depends on the weather conditions on days $n-1$ and n as is shown below:

$n-1$	n	$n+1$	prob
rain	rain	rain	0.6
sunny	sunny	sunny	0.8
sunny	rain	rain	0.5
rain	sunny	sunny	0.75

If we put $Y(\ell) = (X_{\ell-1}, X_\ell)$, then $Y(\cdot)$ is a DTMC. The possible states are $1' = (1,1)$, $2' = (1,2)$, $3' = (2,1)$ and $4' = (2,2)$.
We see that $\{Y(\ell) : \ \ell \geq 1\}$ is a DTMC with transition matrix

$$P = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.75 & 0.25 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \end{pmatrix}.$$

Example 3.8.5. *Give a criterion for ergodicity of the DTMC with state space* $\{0, 1, 2, \cdots\}$ *and transition matrix*

$$P = \begin{pmatrix} q & p & 0 & 0 & 0 & \ddots \\ q & 0 & p & 0 & 0 & \ddots \\ 0 & q & 0 & p & 0 & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

When the DTMC is ergodic, derive its stationary distribution.
We saw that this DTMC is irreducible, aperiodic and recurrent when $p \leq q$. *Solve the linear equations*

$$(\pi_0, \pi_1, \cdots) = (\pi_0, \pi_1, \cdots)P$$

to get $\pi_k = (p/q)^k \pi_0$.
We also need $\sum_{k\geq 0} \pi_k = 1$. *The sum on the left hand side is finite only if* $p < q$, *in which case* $\pi_0 = 1 - (p/q)$ *and so* $\pi_k = [1 - (p/q)]\,(p/q)^k$. *So there is a solution to* $\boldsymbol{\pi} = \boldsymbol{\pi}P$ *with* $\sum_{k\geq 0} \pi_k = 1$, *and hence the DTMC is ergodic, only if* $p < q$, *in which case*

$$\mu_k = \frac{1}{(p/q)^k(1 - (p/q))}.$$

A manufacturing machine at a factory is required in the production process non-stop (24 hours a day and 7 days a week). Nevertheless, the machine experiences both "off periods" and "on periods", where in the former it is not operating due to maintenance or malfunction and in the later it is operating as needed.

In analyzing the performance of the factory, an elementary model for the machine is that of an alternating sequence of independent random variables,

$$X_1, Y_1, X_2, Y_2, X_3, Y_3, \ldots,$$

where $X_i \sim F_X(\cdot)$ represents an "on period" and $Y_i \sim F_Y(0)$ represents "off periods". It is known that at time $t = 0$ the machine has just changed from "off" to "on". In such a case, the state of the machine at time t is represented by $X(t)$ (where say 0 implies "off" and 1 implies "on").

As a first step it is assumed that, $F_X(t) = 1 - e^{-\mu t}$ and $F_Y(t) = 1 - e^{-\lambda t}$. In this case:

2. Argue why $X(t)$ is a CTMC. What is the generator matrix?

3. Simulate a random sample path of $\{X(t), t \in [0, 20]\}$ with $\mu = 2$ and $\lambda = 1$. Plot the trajectory that you have simulated.

4. Calculate the long term proportion of time (i.e. over $t \in [0, \infty)$) during which the machine is "on" (respectively "off"). State your result in terms of the symbols μ and λ.

5. Simulate a long trajectory and use your simulation result to verify your answer to the question above.

6. Let $q(t) = \mathbb{P}\big(\text{"on" at time } t\big)$. Estimate $\{q(t), t \in [0, 10]\}$ by means of simulation. Plot your result.

7. Now calculate $\{q(t), t \in [0, 10]\}$ numerically (without simulation). You may compare to the result above.

8. Now try to find a precise analytic expression for $q(t)$ (in terms of λ and μ), compare your expression to the result above.

9. Is the information that a change occurred "exactly at time 0" important or are the results the same if it were simply stated that the machine is "on" at time 0? Explain your result.

Exercise 3.8.6. *Determine whether each of the following matrices is the generator of a Markov chain and, if yes, describe how the CTMC evolves (λ and μ are nonnegative):*

$$Q_1 = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix};\ Q_2 = \begin{pmatrix} -2 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & 2 & -3 \end{pmatrix};\ Q_3 = \begin{pmatrix} -2 & 1 & 1 \\ 0 & -1 & 1 \\ 1 & 2 & -3 \end{pmatrix}.$$

Assume further that after the machine is in "off" state it needs to be in "warmup" state before moving to "on". Thus the operation of the machine is determined by the sequence,

$$X_1, Y_1, Z_1, X_2, Y_2, Z_2, X_3, Y_3, Z_3 \ldots,$$

where X_i and Y_i are distributed as before and the "warmup" periods Z_i are as follows: $Z_i \sim F_Z(\cdot)$ with $F_Z(t) = 1 - e^{-\gamma t}$.

13. Repeat now questions 5, 7, 10, 11 (assuming $\gamma = 3$ for questions 6 and 7).

It was found now that there is a chance of p that at the end of the warmup period the machine will enter "off" instead of "on".

14. Repeat now questions 5, 7. Leave your answer symbolic in terms of p.

The above CTMC model appears restrictive as it assumes that the distribution of "on", "off" and "warmup" durations is exponential. Comparison to data indicates that it is plausible to assume "on" and "off" durations are exponentially distributed, yet this is not the case for "warmup". In that case, it is suggested to use a PH distribution with m phases, $PH(\boldsymbol{c}', A)$.

15. Incorporate the assumption about the PH distribution of "warmup" in the CTMC model. You should now have a Markov chain where $|\mathcal{S}| = m + 2$. Write down the generator matrix of this CTMC.

The last exercise illustrated one of the strengths of PH distributions: They allow to incorporate the distribution of arbitrary behavior in a CTMC. Often the construction of the PH distribution constitutes modeling in its own right:

16. Assume that the "warmup duration" is either "long" or "short". In the "long" case it is exponentially distributed with γ_1. In the "short" case it is exponentially distributed with γ_2. We have $\gamma_1 < \gamma_2$. There is a chance of $r \in (0, 1)$ that it is long, and a chance of $1 - r$ that it is short. This is an hyper-exponential distribution. Show how it is a special case of the PH distribution and incorporate it in the CTMC model.

17. Assume now that it is measured that "warm up periods" have a mean of m and a squared coefficient of variation of $c^2 > 1$ (the squared coefficient of variation of a random variable is the variance divided by the mean squared). Show how to incorporate this in the CTMC by means of a PH distribution of order 2 yielding arbitrary mean and arbitrary squared coefficient of variation > 1.

18. Why is the restriction of $c^2 > 1$ important? Can you answer the case of $c^2 \in (0, 1)$ with only 2 phases? If not argue why not. As a bonus you may try to find a PH distribution of higher order for this.

	(1,1)	(1,2)	···	(1,n)	(2,1)	(2,2)	···	(2,n)	···	(m,1)	(m,2)	···	(m,n)	(0)
(1,1)	$-q_{1,1}$	$\mu_{1,2}$	$\cdots$	$\mu_{1,n}$	$\lambda_{1,2}$					$\lambda_{1,m}$				$\eta_{1,1}$
(1,2)	$\mu_{2,1}$	$-q_{1,2}$	$\cdots$	$\mu_{2,n}$		$\lambda_{1,2}$					$\lambda_{1,m}$			$\eta_{1,2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$			$\ddots$		$\cdots$			$\ddots$		$\vdots$
(1,n)	$\mu_{n,1}$	$\mu_{n,2}$	$\cdots$	$-q_{1,n}$				$\lambda_{1,2}$					$\lambda_{1,m}$	$\eta_{1,n}$
(2,1)	$\lambda_{2,1}$				$-q_{2,1}$	$\mu_{1,2}$	$\cdots$	$\mu_{1,n}$		$\lambda_{2,m}$				$\eta_{2,1}$
(2,2)		$\lambda_{2,1}$			$\mu_{2,1}$	$-q_{2,2}$	$\cdots$	$\mu_{2,n}$			$\lambda_{2,m}$			$\eta_{2,2}$
$\vdots$			$\ddots$		$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\cdots$			$\ddots$		$\vdots$
(2,n)				$\lambda_{2,1}$	$\mu_{n,1}$	$\mu_{n,2}$	$\cdots$	$-q_{2,n}$					$\lambda_{2,m}$	$\eta_{2,n}$
$\vdots$			$\vdots$				$\vdots$		$\ddots$			$\vdots$		$\vdots$
(m,1)	$\lambda_{m,1}$				$\lambda_{m,2}$					$-q_{m,1}$	$\mu_{1,2}$	$\cdots$	$\mu_{1,n}$	$\eta_{m,1}$
(m,2)		$\lambda_{m,1}$				$\lambda_{m,2}$				$\mu_{2,1}$	$-q_{m,2}$	$\cdots$	$\mu_{2,n}$	$\eta_{m,2}$
$\vdots$			$\ddots$				$\ddots$		$\cdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
(m,n)				$\lambda_{m,1}$				$\lambda_{m,2}$		$\mu_{n,1}$	$\mu_{n,2}$	$\cdots$	$-q_{m,n}$	$\eta_{m,n}$
(0)	0	0	0	0	0	0	0	0	$\cdots$	0	0	0	0	0

Figure 3.21: Generator of Minimum of Two PH Random Variables

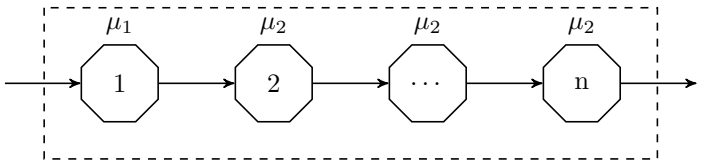


Figure 3.22: Generalized Erlang System

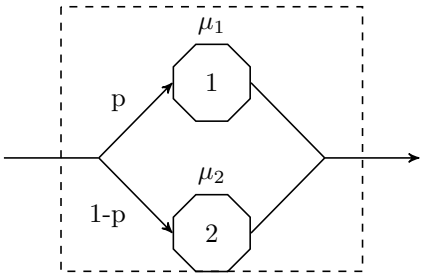


Figure 3.23: 2-Hyperexponential Distribution

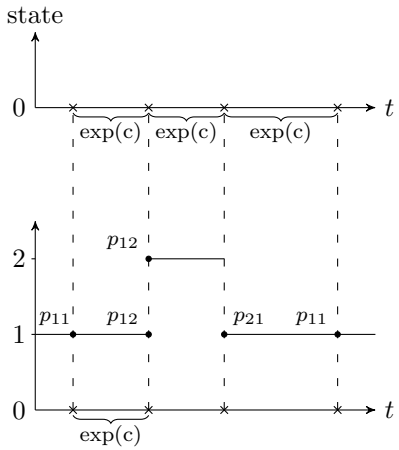


Figure 3.24: CTMC Uniformization

Chapter 4

Structured Markov Chains

In this chapter, we study special types of structured Markov chains with infinitely many states called *Quasi-Birth-and-Death processes* (QBDs). QBDs are the matrix generalisations of simple Birth-and-Death processes, in a similar way as PH distributions are the matrix generalisations of exponential distributions. We thus start the chapter with a section on Birth-and-Death processes, after which we define QBDs and we discuss the computation of their stationary distribution.

We assume that the reader is familiar with both discrete-time and continuous-time Markov chains (see Section 3.3). We first introduce Birth-and-Death processes and QBDs in continuous-time, which are more natural from a modelling point of view. Then we switch to QBDs in discrete-time, which are more suitable for a probabilistic interpretation of the results.

We end the chapter with a section on (continuous-time) Markovian arrival processes, which are the matrix generalisations of Poisson processes, and an illustrative example in queueing theory.

4.1 Quasi-Birth-and-Death Processes

4.1.1 Motivation

Suppose that we want to model a single server queue where the arrivals follow a Poisson process with rate λ and the service times follow a $PH(\boldsymbol{c}', A)$ distribution, generalizing the exponential distribution; this type of queueing system is denoted as the $M/PH/1$ queue.

Recall from Section 3.6 that the service time therefore corresponds to the time until absorption of a Markov chain $\{\varphi(t) : t \geq 0\}$ with one absorbing state 0 and m transient states $\{1, 2, \ldots, m\}$, an initial probability vector $(c_0, \boldsymbol{c}')'$, and a generator with the

following structure

$$\left[\begin{array}{c|c} 0 & \mathbf{0}' \\ \hline \boldsymbol{b} & A \end{array}\right],$$

where $\boldsymbol{b} = -A\mathbf{1}$.

Let $N(t)$ denote the number of customers in the queueing system at time t (including the customer being served, if there is one). The transition from n customers to $n+1$ corresponds to the arrival of one customer to the system, which happens at rate λ. The transition from n customers to $n-1$ corresponds to the service completion of one customer, with rate depending on the current phase of the underlying Markov chain $\{\varphi(t) : t \geq 0\}$ defining the PH distribution. To be able to fully characterise the state transitions of the $M/PH/1$ queue, we thus need to keep track of the phases of the underlying process $\{\varphi(t) : t \geq 0\}$. Therefore, the evolution of the $M/PH/1$ queue can be modelled by the *two-dimensional* Markov chain $\{\boldsymbol{X}(t) = (N(t), \varphi(t)) : t \geq 0\}$, with state space

$$\mathcal{S} = (0, \cdot) \cup \{(n, i) : n \geq 1, 1 \leq i \leq m\},$$

in which the states are ordered using the lexicographic order.

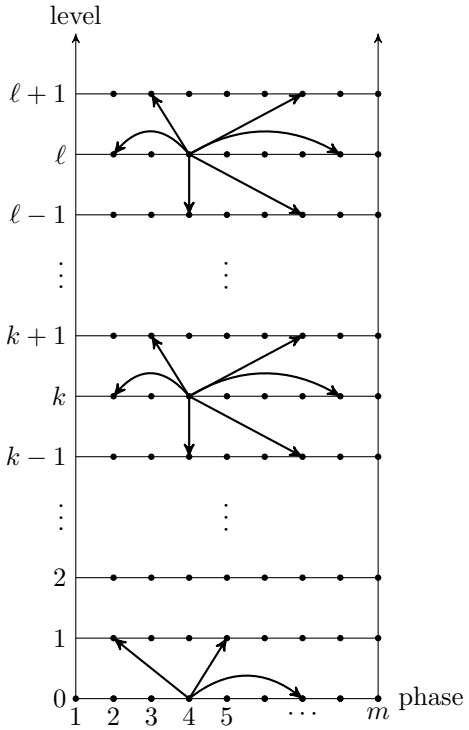


Figure 4.1: Generalized QBD Transition Diagram

As for Birth-and-Death processes, only a few transitions are allowed from each state of the Markov chain $\{\boldsymbol{X}(t) : t \geq 0\}$. For $1 \leq i, j \leq m$, the only possible transitions

from state $(0,\cdot)$ are to the states $(1,i)$, with rate λc_i, and for $n \geq 1$, the only possible transitions from state (n,i) are to the states

- $(n+1,i)$, with rate λ,
- (n,j), $j \neq i$, with rate $a_{i,j}$,
- $(n-1,j)$, with rate $b_i\, c_j$ if $n > 1$, or to the state $(0,\cdot)$ with rate b_i if $n = 1$.

The generator of $\{\boldsymbol{X}(t) : t \geq 0\}$ has the following tridiagonal block structure

$$Q = \begin{bmatrix} -\lambda & \lambda\,\boldsymbol{c}' & \boldsymbol{0}' & \boldsymbol{0}' & \dots \\ \boldsymbol{b} & A - \lambda I & \lambda I & 0 & \dots \\ \boldsymbol{0} & \boldsymbol{b}\cdot\boldsymbol{c}' & A - \lambda I & \lambda I & \dots \\ \boldsymbol{0} & 0 & \boldsymbol{b}\cdot\boldsymbol{c}' & A - \lambda I & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where $\boldsymbol{b}\cdot\boldsymbol{c}'$ is the *outer product* of the column vector $\boldsymbol{b}$ with the row vector $\boldsymbol{c}'$, and I is the identity matrix of required order. Note that the three blocks $-\lambda$, $\lambda\,\boldsymbol{c}'$ and $\boldsymbol{b}$ have a dimension different from the other blocks because there is only one state associated with an empty queue (the state $(0,\cdot)$), while there are m states associated with a queue of size $n > 0$ (the states $\{(n,i) : 1 \leq i \leq m\}$).

The Markov chain $\{\boldsymbol{X}(t) : t \geq 0\}$ is an example of a continuous-time *Quasi-Birth-and-Death Process* (QBD), that is, a two-dimensional Markov chain of which the generator has a tridiagonal block structure.

4.1.2 Discrete-time QBDs

In the rest of the chapter, we consider the discrete-time case, which will be more appropriate for probabilistic interpretation. Generally speaking, a discrete-time QDB process is a two-dimensional Markov chain $\{\boldsymbol{X}(\ell) = (N(\ell), \varphi(\ell)) : \ell \in \mathbb{Z}_+\}$ on the state space $\mathcal{S} = \mathbb{Z}_+ \times \{1, 2, \dots, m\}$ where m is finite, and

- $N(\ell) \in \mathbb{Z}_+$ is called the *level*,
- $\varphi(\ell) \in \{1, 2, \dots, m\}$ is called the *phase*,
- the only possible transitions from the state (n,i) are to the states
  - $(n+1,j)$ (one level up),
  - (n,j) (the same level),
  - $(n-1,j)$ (one level down),

  for $1 \leq i, j \leq m$.

The corresponding transition probabilities are given by

- $(A_1)_{i,j}$, to go up, from (n, i) to $(n{+}1, j)$,
- $(A_0)_{i,j}$ if $n > 0$, or $b_{i,j}$ if $n = 0$, to remain in the same level, from (n, i) to (n, j),
- $(A_{-1})_{i,j}$ to go down, from (n, i) to $(n{-}1, j)$.

The transition probability matrix can be written in the form

$$P = \begin{bmatrix} B & A_1 & 0 & 0 & \dots \\ A_{-1} & A_0 & A_1 & 0 & \dots \\ 0 & A_{-1} & A_0 & A_1 & \dots \\ 0 & 0 & A_{-1} & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{4.1}$$

As the transitions are homogeneous with respect to the levels, we say that the QBD process is *level-independent*. Level-dependent QBDs are more difficult to analyse and will not be considered here.

Let $A = A_{-1} + A_0 + A_1$ be the (stochastic) probability transition matrix of the phase process. The following result is a recipe to check whether a QBD is positive recurrent. The proof of this result is out of the scope of the present course.

Theorem 4.1.1. *If the matrix A is irreducible, then the QBD is positive recurrent if and only if $\zeta = \boldsymbol{\eta}\, A_1\, \mathbf{1} - \boldsymbol{\eta}\, A_{-1}\, \mathbf{1} < 0$, where $\boldsymbol{\eta}$ is the stationary probability vector of A. The QBD is null recurrent if $\zeta = 0$, and it is transient if $\zeta > 0$.*

We define the stationary probability vector of the QBD as $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots]$, where

$$(\boldsymbol{\pi}_n)_i = \lim_{k\to\infty} \mathbb{P}[(X_k, \varphi_k) = (n, i)].$$

The stationary probability vector $\boldsymbol{\pi}$ satisfies the infinite system of equations

$$\boldsymbol{\pi}\, P = \boldsymbol{\pi}, \tag{4.2}$$

$$\boldsymbol{\pi}\, \mathbf{1} = 1, \tag{4.3}$$

where P is given in (4.1). The solution of the system exists if and only if the QBD is positive recurrent.

4.2 Matrix Geometric Solutions

4.2.1 Matrix-geometric property of the stationary distribution

In this section, we continue with discrete-time QBDs. We shall show that the stationary probability vector $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots]$ of a QBD satisfies a *matrix geometric* property,

similar to the geometric property satisfied by the stationary distribution of M/M/1 queues.

Let $\mathcal{L}(n)$ denote the level n of the QBD, that is,

$$\mathcal{L}(n) = \{(n,i) : 1 \leq i \leq m\}.$$

Theorem 4.2.1. *If the QBD is positive recurrent, then there exists a nonnegative matrix N of size $m \times m$ such that*

$$\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n \, A_1 \, N \quad \textit{for } n \geq 0.$$

The matrix N is such that $n_{i,j}$ $(1 \leq i,j \leq m)$ is equal to the expected number of visits to the state (n,j), starting from the state (n,i), before the first visit to any of the states in $\mathcal{L}(n-1)$, and is independent of $n \geq 1$.

We may also write

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n \quad \textit{for } n \geq 0,$$

where $R = A_1 N$ is such that, for any $n \geq 0$, $r_{i,j}$ $(1 \leq i,j \leq m)$ is equal to the expected number of visits to $(n+1,j)$, before a return to $\mathcal{L}(n)$, given that the process starts in (n,i).

Proof: Recall that the stationary probability vector $\boldsymbol{\pi}$ satisfies (4.2)–(4.3). Let us partition the state space $\mathcal{S}$ of the QBD into E and $E^c = \mathcal{S} \setminus E$. The first equation (4.2) then becomes

$$\left[\begin{array}{cc} \boldsymbol{\pi}_E & \boldsymbol{\pi}_{E^c} \end{array}\right] \left[\begin{array}{c|c} P_E & P_{EE^c} \\ \hline P_{E^cE} & P_{E^c} \end{array}\right] = \left[\begin{array}{cc} \boldsymbol{\pi}_E & \boldsymbol{\pi}_{E^c} \end{array}\right],$$

that is,

$$\begin{aligned} \boldsymbol{\pi}_E P_E + \boldsymbol{\pi}_{E^c} P_{E^cE} &= \boldsymbol{\pi}_E \\ \boldsymbol{\pi}_E P_{EE^c} + \boldsymbol{\pi}_{E^c} P_{E^c} &= \boldsymbol{\pi}_{E^c}, \end{aligned}$$

where P_E is the submatrix of transition probabilities between states of E, P_{EE^c} is the submatrix of transition probabilities from states of E to states of E^c, etc. For consistent notation, we shall adopt the following convention and write $\sum_{n \geq 0} M^n = (I - M)^{-1}$ for any substochastic matrix M such that the series converges, even when M is infinite dimensional. With this, the above system is equivalent to

$$\boldsymbol{\pi}_E \left(P_E + P_{EE^c}(I - P_{E^c})^{-1} P_{E^cE}\right) = \boldsymbol{\pi}_E \tag{4.4}$$

$$\boldsymbol{\pi}_{E^c} = \boldsymbol{\pi}_E \, P_{EE^c} \, (I - P_{E^c})^{-1}. \tag{4.5}$$

Exercise 4.2.2. *Show that the matrix N_{E^c} of expected sojourn time in the subset E^c, before the first passage to the complementary subset E, is given by*

$$N_{E^c} = (I - P_{E^c})^{-1}.$$

Next, we choose $E = \mathcal{L}(0)$, $E^c = \{\mathcal{L}(1), \mathcal{L}(2), \mathcal{L}(3), \ldots, \}$, so that

$$P = \left[\begin{array}{c|c} P_E & P_{EE^c} \\ \hline P_{E^cE} & P_{E^c} \end{array}\right] = \left[\begin{array}{c|cccc} B & A_1 & 0 & 0 & \cdots \\ \hline A_{-1} & A_0 & A_1 & & \\ 0 & A_{-1} & A_0 & A_1 & \\ 0 & & A_{-1} & A_0 & \ddots \\ \vdots & & & \ddots & \ddots \end{array}\right].$$

Equation (4.5) then becomes

$$\left[\begin{array}{ccc} \boldsymbol{\pi}_1 & \boldsymbol{\pi}_2 & \ldots \end{array}\right] = \boldsymbol{\pi}_0 \left[\begin{array}{cccc} A_1 & 0 & 0 & \cdots \end{array}\right] \underbrace{\left[\begin{array}{cccc} I - A_0 & -A_1 & & \\ -A_{-1} & I - A_0 & -A_1 & \\ & -A_{-1} & I - A_0 & \ddots \\ & & \ddots & \ddots \end{array}\right]^{-1}}_{N}.$$

We obtain for $\boldsymbol{\pi}_1$

$$\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0 A_1 N_{11},$$

where N_{11} is the upper left block of the matrix $N_{E^c} = (I - P_{E^c})^{-1}$.

Exercise 4.2.3. *Justify the fact that the (i,j)th entry of the matrix N_{11} can be interpreted as the expected number of visits of state $(1,j)$, starting from $(1,i)$, before returning to $\mathcal{L}(0)$.*

Similarly, for any $n \geq 1$, choose

$$\begin{aligned} E &= \{\mathcal{L}(0), \mathcal{L}(1), \ldots, \mathcal{L}(n)\}, \\ E^c &= \{\mathcal{L}(n+1), \mathcal{L}(n+2), \ldots\}. \end{aligned}$$

Since the QBD is level-independent, the matrices P_{E^c} are the same irrespective of the value of n chosen in the partition. Therefore, the matrix N_{11} (still defined as the upper left block of $(I - P_{E^c})^{-1}$) is independent of n, and its (i,j)th entry can be generally interpreted as the expected number of visits of state $(n+1,j)$, starting from $(n+1,i)$, before returning to $\{\mathcal{L}(0), \mathcal{L}(1), \ldots, \mathcal{L}(n)\}$, for any $n \geq 0$.

By the same argument as above, we get for all $n \geq 0$

$$\begin{aligned} \boldsymbol{\pi}_{n+1} &= \boldsymbol{\pi}_n A_1 N_{11} \\ &= \boldsymbol{\pi}_0 (A_1 N_{11})^{n+1}. \end{aligned}$$

In summary, the stationary probability vector $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots]$ satisfies

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n \quad \text{for all } n \geq 0, \tag{4.6}$$

where $R = A_1 N_{11}$ and $r_{i,j}$ can be interpreted as the expected number of visits to $(n+1,j)$ starting from (n,i) before the first return to $\mathcal{L}(n)$, for any $n \geq 0$.

□

4.2.2 Characterisation of $\boldsymbol{\pi}_0$ and R

To complete the characterisation of the stationary distribution, it remains to characterise $\boldsymbol{\pi}_0$ and R. Take again $E = \mathcal{L}(0)$. From Equation (4.4) we get

$$\boldsymbol{\pi}_0(B + A_1\, N_{11}\, A_{-1}) = \boldsymbol{\pi}_0.$$

From the normalization constraint $\boldsymbol{\pi}\,\mathbf{1} = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i\,\mathbf{1} = 1$ and (4.6), we have

$$\sum_{i=0}^{\infty} \boldsymbol{\pi}_i\,\mathbf{1} = \boldsymbol{\pi}_0 \sum_{i=0}^{\infty} R^i\,\mathbf{1} = 1.$$

We thus see that if the QBD is positive recurrent, then the spectral radius $\mathrm{sp}(R)$ of the matrix R is strictly less than 1 and $\sum_{i=0}^{\infty} R^i = (I - R)^{-1}$ — actually, we can show that the positive recurrence condition is not only sufficient but also necessary, but this requires more time. Therefore, in the positive recurrent case, $\boldsymbol{\pi}_0$ is the unique solution of the system

$$\left\{ \begin{array}{rcl} \boldsymbol{\pi}_0(B + R\,A_{-1}) & = & \boldsymbol{\pi}_0 \\ \boldsymbol{\pi}_0\,(I - R)^{-1}\,\mathbf{1} & = & 1. \end{array} \right.$$

Next, from $\boldsymbol{\pi}\,P = \boldsymbol{\pi}$ with the matrix P given by (4.1), we obtain

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}A_1 + \boldsymbol{\pi}_n A_0 + \boldsymbol{\pi}_{n+1}A_{-1}, \qquad \text{for all } n \geq 1,$$

which, with $\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}\,R$, leads to the following quadratic fixed-point matrix equation for R

$$R = A_1 + RA_0 + R^2 A_{-1}, \tag{4.7}$$

for which there is generally no explicit solution.

4.2.3 The probability matrix G

There are several other equivalent expressions for R. One practical way to compute the matrix R is to express it with the help of a probability matrix G defined as follows: $g_{i,j}$ is the probability of eventually moving to the state $(0, j)$, starting from the state $(1, i)$, that is, G records the probability to reach $\mathcal{L}(0)$ in a finite time, starting from $\mathcal{L}(1)$. Similarly, from the level-independence assumption, G records the probability to reach $\mathcal{L}(n-1)$ in a finite time, starting from $\mathcal{L}(n)$, for any $n \geq 1$.

The matrix G satisfies the (matrix) quadratic equation

$$G = A_{-1} + A_0 G + A_1 G^2, \tag{4.8}$$

which is easier to interpret than (4.7). Indeed, G records the probability to reach $\mathcal{L}(0)$ in a finite time, starting from $\mathcal{L}(1)$, and the right-hand side of (4.8) decomposes this probability according to the first transition from $\mathcal{L}(1)$. The first term corresponds to the

QQQQ Missing PDF.

Figure 4.2: Visualization of G_{ij}

case where the QBD directly moves from $\mathcal{L}(1)$ to $\mathcal{L}(0)$ in one transition, with probabilities recorded in A_{-1}. The second term corresponds to the case where the first transition from $\mathcal{L}(1)$ is within the same level, with probabilities recorded in A_0, from which the QBD still has to move to $\mathcal{L}(0)$, with probabilities recorded in G. Finally, the third term corresponds to the case where the first transition from $\mathcal{L}(1)$ is to $\mathcal{L}(2)$, with probabilities recorded in A_1, from which the QBD has first to return to $\mathcal{L}(1)$, with probabilities recorded in G, and then to $\mathcal{L}(0)$, with probabilities recorded in G, whence the factor G^2.

We generally have to resort to numerical techniques to solve Equation (4.8) as it has an explicit solution only in a few special cases.

In order to describe the relationship between the matrices R and G, let us now consider the probabilistic interpretation of the matrix A_0+A_1G: $(A_0+A_1G)_{i,j}$ is the probability of visiting $(1, j)$ from $(1, i)$ avoiding $\mathcal{L}(0)$. Therefore, $[I-(A_0+A_1G)]^{-1}$ is the mean number of visits of $(1, j)$ starting from $(1, i)$ and avoiding $\mathcal{L}(0)$. The probabilistic interpretations of $[I - (A_0 + A_1G)]^{-1}$ and R lead to the following link between the matrices R and G:

$$R = A_1\big[I - (A_0 + A_1G)\big]^{-1}. \tag{4.9}$$

Therefore, if the matrix G is known, then we can determine R by using the above relationship.

It can be shown that the QBD is positive recurrent if and only if G is stochastic and $\mathrm{sp}(A_0 + 2A_1G) < 1$. In the next section, we shall focus on the numerical solution of Equation (4.8) for the matrix G of positive recurrent QBDs, instead of Equation (4.7) for the matrix R, since it is known a priori that the solution must satisfy $G\mathbf{1} = \mathbf{1}$.

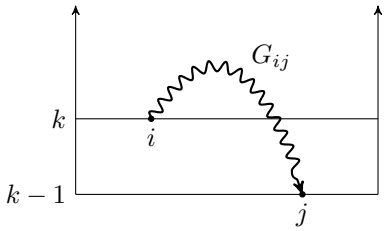


Figure 4.3: Visualization of G_{ij}

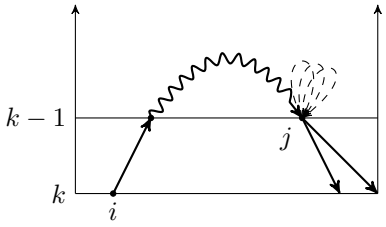


Figure 4.4: Visualization of R_{ij}

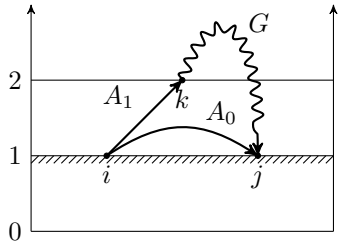


Figure 4.5: Visualization of $(A_0 + A_1 G)_{ij}$

4.2.4 Remark on continuous-time QBDs

The results mentioned previously still hold for continuous-time QBDs. The only differences are

- the interpretation of the matrix R which is a bit more involved;
- the equations for R and G, which become respectively

$$\begin{aligned} 0 &= A_1 + RA_0 + R^2 A_{-1} \\ 0 &= A_{-1} + A_0 G + A_1 G^2, \end{aligned}$$

  the last equation being equivalent to

$$G = (-A_0)^{-1} A_{-1} + (-A_0)^{-1} A_1 G^2 : \tag{4.10}$$

- the link between R and G, which becomes

$$R = A_1 \big[-(A_0 + A_1 G) \big]^{-1}, \tag{4.11}$$

  and
- the system of equations satisfied by $\boldsymbol{\pi}_0$, which becomes

$$\begin{aligned} \boldsymbol{\pi}_0 (B + R\, A_{-1}) &= \mathbf{0}' & (4.12) \\ \boldsymbol{\pi}_0 \, (I - R)^{-1} \, \mathbf{1} &= 1. & (4.13) \end{aligned}$$

4.3 Algorithmic Solutions

In this section, we discuss a basic algorithm to solve for the matrix G in the discrete-time setting.

The fixed point equation for G,

$$X = A_{-1} + A_0 X + A_1 X^2,$$

can be equivalently rewritten as

$$X = \left[I - (A_0 + A_1 X)\right]^{-1} A_{-1}. \tag{4.14}$$

A first approach is to solve (4.14) using functional iteration, which leads to the following linear algorithm. We start with $G_0 = 0$ and iterate

$$G(k) = \left[I - (A_0 + A_1 G(k-1))\right]^{-1} A_{-1}, \tag{4.15}$$

or equivalently

$$G(k) = A_{-1} + A_0 G(k) + A_1 G(k-1) G(k). \tag{4.16}$$

The matrix $G(k)$ computed at the kth iteration has the following probabilistic interpretation: $g_{i,j}(k)$ is the probability that the QBD moves from the state $(1, i)$ in $\mathcal{L}(1)$ to $\mathcal{L}(0)$ in a finite amount of time by visiting the specific state $(0, j)$, and does so *without going to* $\mathcal{L}(k+1)$ *and higher levels*. With this interpretation, we can show that $G(k)$ indeed satisfies (4.16) using the same arguments as those used in the previous section to show that G satisfies (4.8).

In other words, at the kth iteration, the QBD is allowed to move freely among k levels only. As k increases, the restriction on the permitted levels disappears. The sequence $\{G(k)\}_{k\geq 0}$ is monotonically increasing and converges to the matrix G.

If the QBD is recurrent, then G is stochastic and we can stop the iteration when $||\mathbf{1} - G(k)\mathbf{1}||_\infty < \epsilon$ for a predefined tolerance ϵ.

The following theorem is a consequence of the above iteration.

Theorem 4.3.1. *The matrix G is the minimal nonnegative solution of the matrix equation*

$$X = A_{-1} + A_0 X + A_1 X^2.$$

Proof. We rewrite Equation (4.14) as $X = F(X)$ where $F(X) = \left[I - (A_0 + A_1 X)\right]^{-1} A_{-1} = \sum_{n\geq 0} (A_0 + A_1 X)^n A_{-1}$. Since A_{-1}, A_0 and A_1 are nonnegative matrices, $F(X)$ is monotonically increasing in X, that is, if $X \leq Y$, then $F(X) \leq F(Y)$.

Assume that G^* is another nonnegative solution of $X = F(X)$. Then, since $G^* \geq 0$, we have $G^* = F(G^*) \geq F(0) = G(1)$, and by induction, we find $G^* \geq G(k)$ for all $k \geq 1$. By letting $k \to \infty$, we obtain $G^* \geq G$, so G is the minimal nonnegative solution of Equation (4.14). □

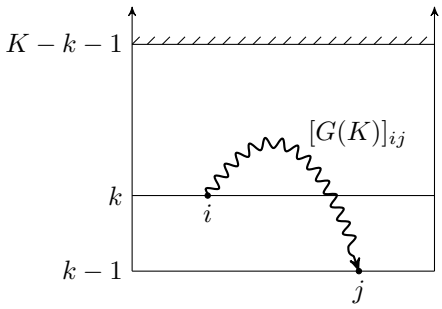


Figure 4.6: Visualization of $[G(K)]_{ij}$

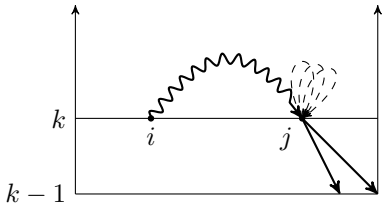


Figure 4.7: Visualization of $[I-(A_0+A_1G)]^{-1}_{ij}$

4.3.1 Remark on continuous-time QBDs

The algorithm can be easily modified in order to become applicable to continuous-time QBDs. It is left as an exercise to show that Equation (4.16) becomes

$$G(k) = (-A_0)^{-1}A_{-1} + (-A_0)^{-1}A_1G(k-1)G(k),$$

which can be explicitly rewritten as

$$G(k) = -[A_0 + A_1G(k-1)]^{-1}A_{-1},$$

where $G(k)$ has the same probabilistic interpretation as in the discrete-time setting.

4.4 Markovian arrival processes

In this section, we come back to the continuous-time setting.

4.4.1 Markovian arrival processes

In the same way as a Poisson process is a particular Birth-and-Death process with no death event (it is actually called a *pure birth process*, see Section 4.1), a *Markovian*

arrival process (MAP) is a particular level-independent QBD with no transitions to any lower level, that is, such that $A_{-1} = 0$. The generator of the MAP has thus the following block structure

$$Q = \begin{bmatrix} D_0 & D_1 & 0 & 0 & \dots \\ 0 & D_0 & D_1 & 0 & \dots \\ 0 & 0 & D_0 & D_1 & \dots \\ 0 & 0 & 0 & D_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Like a QBD, a MAP is therefore a two-dimensional Markov chain $\boldsymbol{X}(t) = \{(N(t), \varphi(t)) : t \geq 0\}$ on the state space $\mathbb{Z}_+ \times \{0, 1, \dots, m\}$, where m is finite. The process $N(t)$ counts the number of arrivals in $[0, t]$ and is called the *level* process. The process $\varphi(t)$ is a continuous-time Markov process, called the *phase* process. The matrix D_0 records the phase transition rates which are not associated to an arrival, and the matrix D_1 records the phase transition rates associated to an arrival; more precisely, $(D_1)_{i,j}$ is the rate at which there is an arrival *and* the phase moves from i to j.

A MAP is a more general counting process than a Poisson process: the interarrival times are not necessarily independent of each other, nor exponentially distributed. MAPs are dense in the set of point processes on the real line, which makes them a powerful modelling tool. In queueing theory, MAPs are often used to model the arrival of customers to a queue.

Exercise 4.4.1. *Describe the MAP/M/1 queue and give its generator.*

Exercise 4.4.2. *Describe the MAP/PH/1 queue and give its generator.*

4.4.2 PH renewal processes

An important special case of a MAP is the *PH renewal process*. It is essentially the same as a MAP except that in a PH renewal process the interarrival times are *independent* of each other and $PH(\boldsymbol{c}', A)$ distributed. In this case, $D_0 = A$, and D_1 can be decomposed as $D_1 = \boldsymbol{b} \cdot \boldsymbol{c}'$ where $\boldsymbol{b} = -A\mathbf{1}$ and $\boldsymbol{c}'$ is the initial probability (row) vector of the PH distribution.

4.5 Illustrative examples: The PH/M/1, M/PH/1 and PH/PH/1 Queues

In this section, we illustrate the previous results with an example of a single server queue in which arrivals occur according to a PH renewal process and the service time is exponentially distributed. Such a queueing system is called a PH/M/1 queue.

More precisely, we assume that the service time is exponentially distributed with parameter μ, and that the interarrival times have the Erlang distribution $E(3, \lambda)$, that

is, they are the sum of three independent exponential random variables with parameter λ. Recall from Section 3.6 that the distribution $E(3,\lambda)$ corresponds to a $PH(\boldsymbol{c}', A)$ distribution of order 3 such that

$$\boldsymbol{c}' = [1,0,0], \quad A = \begin{bmatrix} -\lambda & \lambda & 0 \\ 0 & -\lambda & \lambda \\ 0 & 0 & -\lambda \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ \lambda \end{bmatrix}.$$

In this particular case, we talk about the Erlang/M/1 queue.

The Erlang/M/1 queue can be modelled as a continuous-time QBD $\{\boldsymbol{X}(t) = (N(t), \varphi(t)) : t \geq 0\}$, where $N(t)$ represents the number of customers in the system at time t, and $\varphi(t)$ corresponds to the phase of the underlying Markov chain describing the Erlang distribution. Its generator is given by

$$Q = \begin{bmatrix} A & \boldsymbol{b}\cdot\boldsymbol{c}' & 0 & 0 & \dots \\ \mu I & A-\mu I & \boldsymbol{b}\cdot\boldsymbol{c}' & 0 & \dots \\ \mathbf{0} & \mu I & A-\mu I & \boldsymbol{b}\cdot\boldsymbol{c}' & \dots \\ \mathbf{0} & 0 & \mu I & A-\mu I & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The matrices B, A_{-1}, A_0 and A_1 defined in Section 4.2 are given by $B = A$, $A_{-1} = \mu I$, $A_0 = A - \mu I$, and $A_1 = \boldsymbol{b}\cdot\boldsymbol{c}'$.

In order to determine for which values of λ and μ the process is positive recurrent, we compute the stationary probability vector $\boldsymbol{\eta}$ of the irreducible generator $\tilde{A} = A_{-1} + A_0 + A_1$ given by

$$\tilde{A} = \begin{bmatrix} -\lambda & \lambda & 0 \\ 0 & -\lambda & \lambda \\ \lambda & 0 & -\lambda \end{bmatrix}.$$

Exercise 4.5.1. *Show that the stationary probability vector of $\tilde{A}$ is given by $\boldsymbol{\eta} = [1/3, 1/3, 1/3]$ irrespective of the value of λ, and that consequently the process is positive recurrent if and only if $\mu > \lambda/3$. Give an interpretation of this result.*

In what follows, we choose $\lambda = 2$ and $\mu = 1$. Using the algorithm described in Section 4.4 implemented in Matlab, we obtain

$$G = \begin{bmatrix} 0.40 & 0.33 & 0.27 \\ 0.11 & 0.49 & 0.40 \\ 0.16 & 0.24 & 0.60 \end{bmatrix}.$$

Using the relation (4.11) between R and G, we obtain

$$R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.81 & 0.66 & 0.53 \end{bmatrix}.$$

Finally, solving (4.12)–(4.13), we obtain

$$\boldsymbol{\pi_0} = [0.06,\ 0.11,\ 0.16],$$

from which we can compute $\boldsymbol{\pi_n}$ for any n using (4.6).

Exercise 4.5.2. *Implement the algorithm described in Section 4.4 to compute the matrices G and R, and the stationary distribution $\boldsymbol{\pi}$ for other values of λ and μ.*

MERGE IN (TEXT BELOW IS FROM KAY) – PH/M/1:
When looking at (PH/M/1), the arrival process is of a PH type distribution. In this example we consider an arrival process consisting of an arrival phase 2 with distribution parameter λ, and a second phase 1 with distribution parameter λ' that is entered with probability q. This generator is visualized in the figure below:

This is from the perspective of an item moving through the arrival process. This can be represented as follows on a system level, with a single machine with distribution parameter μ:

From this system it is possible to create a transition diagram, according to the following possible transitions:

From	*To*	*Rate*	*Conditions*
(k,1)	(k,2)	$q\lambda$	for $k \geq 0$
(k,1)	(k,2)	$p\lambda$	for $k \geq 0$
(k,1)	(k-1,1)	μ	for $k \geq 1$
(k,2)	(k+1,1)	λ'	for $k \geq 0$
(k,2)	(k-1,2)	μ	for $k \geq 1$

The transition diagram for this system is then as follows:

The corresponding generator Q can then be derived by taking the transition rates. The diagonal contains the distribution parameter for the corresponding state, so that each row sums up to 0.
With Q known, B, A_{-1}, A_0 and A_1 are as follows:

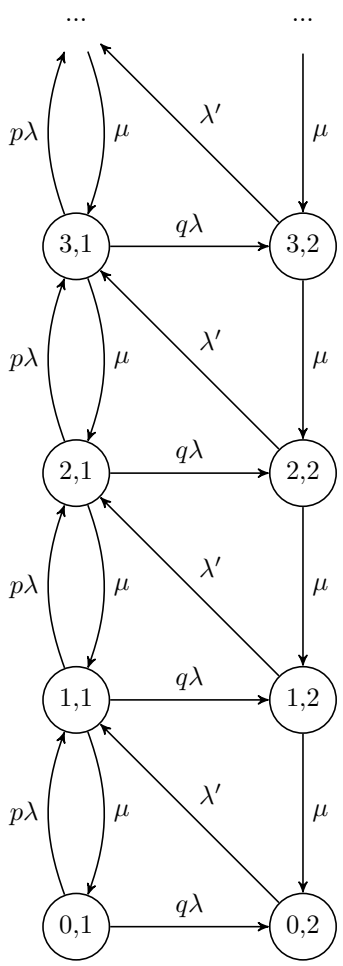


Figure 4.8: $PH_2/M/1$ Transition Diagram

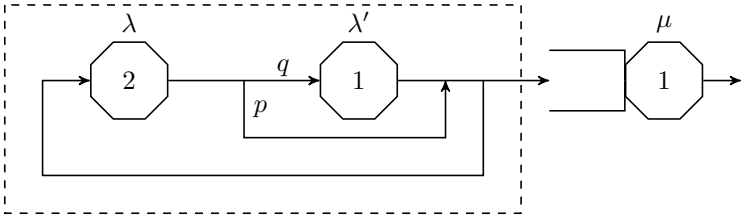


Figure 4.9: $PH_2/M/1$ System

$$\mathbf{c}' = \begin{bmatrix} 1 & 0 \end{bmatrix}$$
$$A = \begin{bmatrix} -\lambda & q\lambda \\ 0 & \lambda' \end{bmatrix}$$
$$\mathbf{b} = A\mathbf{1} = \begin{bmatrix} p\lambda \\ \lambda' \end{bmatrix}$$
$$B = A$$
$$A_{-1} = \mu I$$
$$A_0 = A - \mu I$$
$$A_1 = \mathbf{b}\,\boldsymbol{\alpha}' = \begin{bmatrix} p\lambda & 0 \\ \end{bmatrix}$$

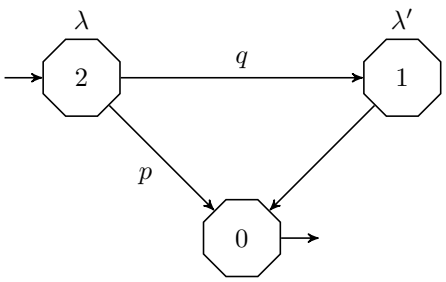


Figure 4.10: PH_2 Diagram

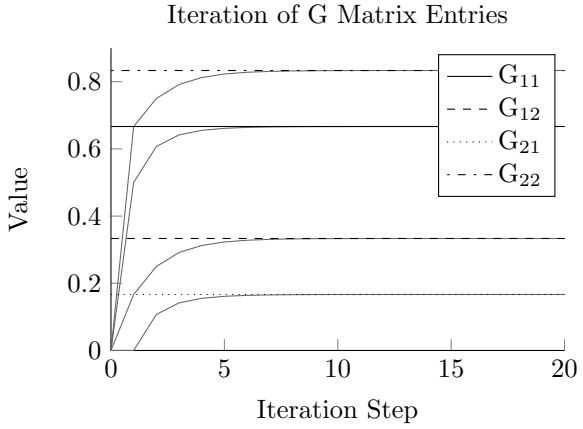


Figure 4.11: Iteration progression of the probability matrix G

In turn, this yields the following matrix $\tilde{A}$:

$$\tilde{A} = A_{-1} + A_0 + A_1 = A + \mathbf{b} \cdot \mathbf{c}' = \begin{bmatrix} -q\lambda & q\lambda \\ \lambda' & -\lambda' \end{bmatrix}$$

Filling this matrix in $\eta A = 0$, together with $\eta \mathbf{1} = 1$, yields the following set of equations:

$$\begin{aligned} (-q\lambda)\pi_1 + \lambda'\pi_2 &= 0 \\ (q\lambda)\pi_1 - \lambda'\pi_2 &= 0 \\ \pi_1 + \pi_2 &= 1 \end{aligned}$$

Which can be rewritten to the following:

$$
Q = \begin{array}{c} \\ (0,1) \\ (0,2) \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \\ \vdots \end{array}
\begin{array}{c}
\begin{array}{cccccc} (0,1) & (0,2) & (1,1) & (1,2) & (2,1) & (2,2) \quad \ldots \end{array} \\
\left[\begin{array}{cc|cc|cc|c}
q_1^o & q\lambda & p\lambda & . & . & . & \ldots \\
. & q_2^o & \lambda' & . & . & . & \ldots \\
\hline
\mu & . & q_1^* & q\lambda & p\lambda & . & \ldots \\
. & \mu & . & q_2^* & \lambda' & . & \ldots \\
\hline
. & . & \mu & . & q_1^* & q\lambda & \ldots \\
. & . & . & \mu & . & q_2^* & \ldots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}\right]
\end{array}
\tag{4.17}
$$

Figure 4.12: $PH_2/M/1$ Generator

$$
\begin{aligned}
\pi_1 &= \frac{\lambda'}{(q\lambda + \lambda')} \\
\pi_2 &= \frac{q\lambda}{(q\lambda + \lambda')} \\
\frac{\pi_1}{\pi_2} &= \frac{\lambda'}{q\lambda}
\end{aligned}
$$

The QBD is positive recurrent when:

$$\zeta = \eta A_1 \mathbf{1} - \eta A_{-1} \mathbf{1} < 0$$

Or after rewriting:

$$\frac{\lambda\lambda'}{q\lambda + \lambda'} < \mu$$

If we then assume the average time spent in the two phases is the same ($\pi_1 = \pi_2$), with $\lambda' = 2$ and $\lambda = 1$, it follows that $q = p = 0.5$. Filling this in yields $1 < \mu$, which we can choose freely as long as the aforementioned equation holds. Here $\mu = 2$ is taken. The iteration for G(k) can then begin, where:

Figure 4.13: Iteration

$$G(k) = -[A_0 + A_1 G(k-1)]^{-1} A_{-1}$$

The implementation in Matlab then yields (for $G(0) = 0$) the following graph:

Where it is clear that $G(k) \to G$ in less than 10 steps. With
$R = A_1[-(A_0 + A_1 G)]^{-1}$, R becomes:

$$R = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{6} \end{bmatrix}$$

The following set of equations can now be solved.

$$\begin{aligned} \boldsymbol{\pi}_0(B + RA_{-1}) &= \mathbf{0}' \\ \boldsymbol{\pi}_0(I - R)^{-1}\mathbf{1} &= 1 \end{aligned}$$

$$\begin{aligned} (B + RA_{-1})'\boldsymbol{\pi}_0' &= \mathbf{0} \\ [(I - R)^{-1}\mathbf{1}]'\boldsymbol{\pi}_0' &= 1 \end{aligned}$$

Where the second set of equations has been rewritten for Matlab implementation. The resulting output is the following stationary probability distribution:

$$\boldsymbol{\pi}_0 = \begin{bmatrix} \frac{1}{12} & \frac{1}{6} \end{bmatrix}$$

MERGE IN (TEXT BELOW IS FROM KAY) – M/PH/1:

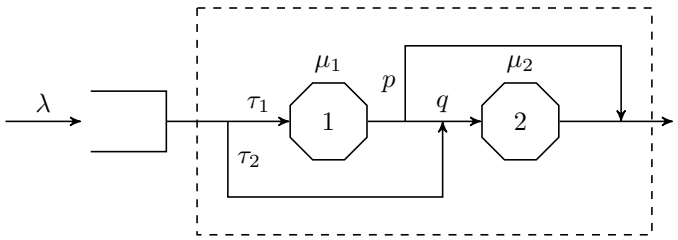


Figure 4.14: (M/PH$_2$/1) System

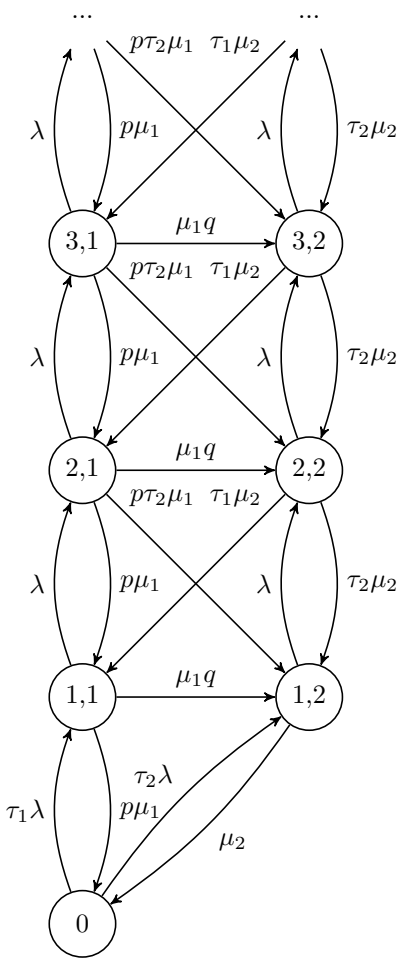


Figure 4.15: Transition Diagram M/PH$_2$/1 Example

Figure 4.16: $ER_2/ER_2/1$ System

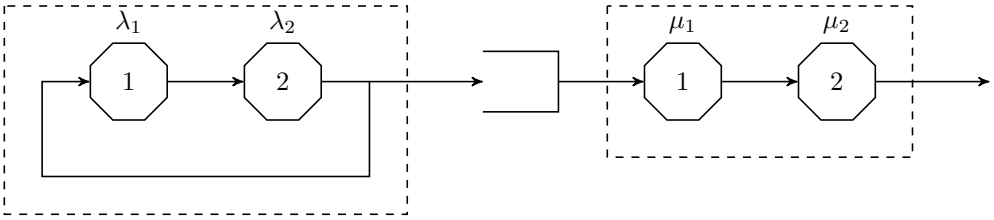


Figure 4.17: $PH_a/PH_s/1$

$$Q = \begin{array}{c} \\ (0) \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} (0) & (1,1) & (1,2) & (2,1) & (2,2) & \ldots \end{array} \\ \left[\begin{array}{c|cc|cc|c} q_0 & \tau_1\lambda & \tau_2\lambda & . & . & \ldots \\ \hline p\mu_1 & q_1 & q\mu_1 & \lambda & . & \ldots \\ p\mu_2 & . & q_2 & . & \lambda & \ldots \\ \hline . & \tau_1 p\mu_1 & \tau_2 p\mu_1 & q1 & q\mu_1 & \ldots \\ . & \tau_1\mu_2 & \tau_2\mu_2 & . & q_2 & \ldots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right] \end{array} \tag{4.18}$$

MERGE IN (TEXT BELOW IS FROM KAY) – PH/PH/1:

From	*To*	*Rate*	*Conditions*
(0,1)	(0,2)	λ_1	
(0,2)	(1,1,1)	λ_2	
(1,2,1)	(0,1)	μ_2	
(1,2,2)	(0,2)	μ_2	
(k,1,1)	(k,1,2)	λ_1	for $k \geq 1$
(k,2,1)	(k,2,2)	λ_1	for $k \geq 1$
(k,1,1)	(k,2,1)	μ_1	for $k \geq 1$
(k,1,2)	(k,2,2)	μ_1	for $k \geq 1$
(k,1,2)	($k+1$,1,1)	λ_2	for $k \geq 1$
(k,2,2)	($k+1$,2,1)	λ_2	for $k \geq 1$
(k,2,1)	($k-1$,1,1)	μ_2	for $k \geq 2$
(k,2,2)	($k-1$,1,2)	μ_2	for $k \geq 2$

MERGED FROM ANOTHER FILE OF KAY...

The goal of the algorithm is to obtain the number of customers in the system as a function of c_a^2, c_s^2 and the utilization ρ. The number of customers in the queue can be

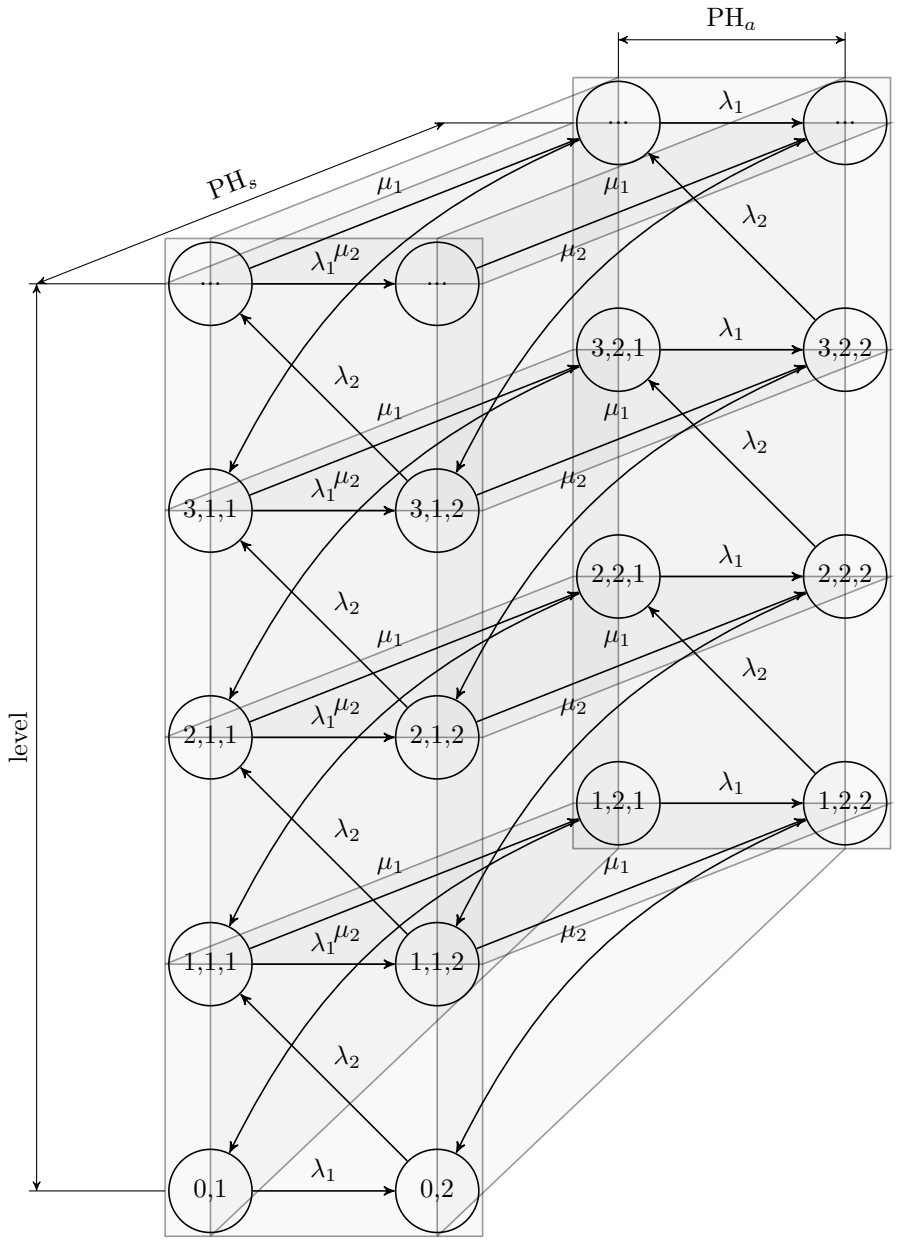


Figure 4.18: PH$_a$/PH$_s$/1 Transition Diagram

$$
Q = \begin{array}{c} \\ (0,1) \\ (0,2) \\ (1,1,1) \\ (1,1,2) \\ (1,2,1) \\ (1,2,2) \\ (2,1,1) \\ (2,1,2) \\ (2,2,1) \\ (2,2,2) \\ \vdots \end{array}
\begin{array}{c}
\begin{array}{cccccccccc} (0,1) & (0,2) & (1,1,1) & (1,1,2) & (1,2,1) & (1,2,2) & (2,1,1) & (2,1,2) & (2,2,1) & (2,2,2) & \ldots \end{array} \\
\left[\begin{array}{cc|cccc|cccc|c}
-\lambda & \lambda & . & . & . & . & . & . & . & . & \ldots \\
. & -\lambda & \lambda & . & . & . & . & . & . & . & \ldots \\
\hline
. & . & -(\mu+\lambda) & \mu & \lambda & . & . & . & . & . & \ldots \\
\mu & . & . & -(\mu+\lambda) & . & \lambda & . & . & . & . & \ldots \\
. & . & . & . & -(\mu+\lambda) & \mu & \lambda & . & . & . & \ldots \\
. & \mu & . & . & . & -(\mu+\lambda) & . & \lambda & . & . & \ldots \\
\hline
. & . & . & . & . & . & -(\mu+\lambda) & \mu & \lambda & . & \ldots \\
. & . & \mu & . & . & . & . & -(\mu+\lambda) & . & \lambda & \ldots \\
. & . & . & . & . & . & . & . & -(\mu+\lambda) & \mu & \ldots \\
. & . & . & . & \mu & . & . & . & . & -(\mu+\lambda) & \ldots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}\right]
\end{array}
$$

Figure 4.19: $\mathrm{PH}_a/\mathrm{PH}_s/1$ Generator

determined by summing up the probability the system is at a level, multiplied by that level:

$$EQ = \sum_{i=1}^{m} \pi_i R^i I$$

In this summation the stopping point is determined by the following criterion rather than a fixed number m: continue while $\pi_i R^i I > M$. Here M determines the accuracy and length of the iteration process and can be chosen by the user. However, this summation requires the stationary probability distribution π and matrix R. To this end a distinction between various c_v^2 is made:

$$c_v^2 = \begin{cases} ER_{\frac{1}{c_v^2}}, & c_v^2 < 1 \\ M, & c_v^2 = 1 \\ HE_2, & c_v^2 > 1 \end{cases}$$

Based on the input values, the type of distribution for the arrival and departure processes can then be determined. After the corresponding initial distribution vector, subgenerator and exit-vector have been determined, the various sub-matrices of the generator Q can be derived. Q and its submatrices are of the following form:

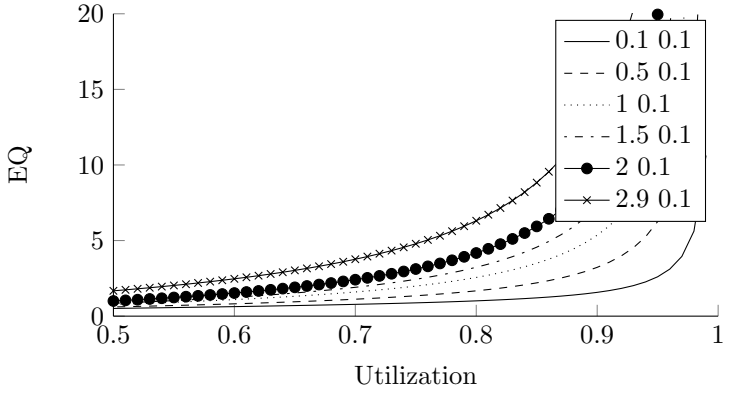


Figure 4.20: $EQ(\rho)$ for $c_s^2 = 0.1$ and various c_a^2

$$Q = \begin{bmatrix} B_0 & B_1 & . & . & \dots \\ B_{-1} & A_0 & A_1 & . & \dots \\ . & A_{-1} & A_0 & A_1 & \dots \\ . & . & A_{-1} & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\begin{aligned} A_1 &= b \cdot c \otimes I \\ A_{-1} &= I \otimes t \cdot u \\ A_0 &= A \otimes I + I \otimes S \\ B_1 &= b \cdot c \otimes u \\ B_{-1} &= I \otimes t \\ B_0 &= A \end{aligned}$$

At this point something interesting occurs, as the dimension of the matrices B can be of a different dimension than the matrices A. This means the matrix R is still based on A, but the equation $\pi_n = pi_0 R^n$ is no longer guaranteed to hold. Therefore we explicitly write the balance equation $\pi Q = 0$, with the aforementioned guess for R. These equations can be rewritten to the following form:

$$\pi_0(B_0 + RA_{-1}) = 0 \pi_i(A1 + RA_0 + R^2 A_{-1}) \qquad = 0$$

For $i >= 0$. Where this common matrix can be rewritten to the iteration scheme $R_i = -(A_1 + R_{i-1}^2 A_{-1})A_0^{-1}$. However, in order to derive the stationary probability vectors the first two equations of this set have to be used. This results in the following set of equations:

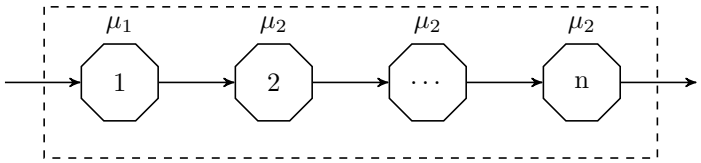


Figure 4.21: Generalized Erlang distribution ER_n

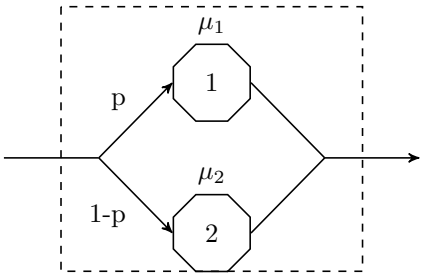


Figure 4.22: Hyper-exponential distribution HE_2

$$\begin{bmatrix}\pi_0 & \pi_1\end{bmatrix}\begin{bmatrix}B_0 & B_1\\ B_{-1} & (A_0+RA_{-1})\end{bmatrix} = \begin{bmatrix}0 & \dots & 0\end{bmatrix}$$

Additionally, the normalising equation $\begin{bmatrix}\pi_0 & \pi_1\end{bmatrix}\begin{bmatrix}\mathbf{1}\\ (I-R)^{-1}\cdot\mathbf{1}\end{bmatrix} = 1$ yields another equation. This is the last equation needed to obtain the stationary distribution vectors, which are obtained by filling in this equation for the first column of the previous set of equations and solving them. The resulting Matlab implementation is shown below, where utilization values of up to 0.99 are considered for the (c_a^2, c_v^2) combinations listed in C and the resulting queue lengths saved in the matrix plotdata.

4.6 Branching Processes and the Markovian Binary Tree

Around 1870, when studying the problem of family names extinction in British peerages, Galton and Watson showed for the first time how the computation of probabilities could explain the effects of randomness in the development of families or populations. They proposed a mathematical model that went unnoticed for many years, and that reappeared in isolated papers in the 1920's and 1930's.

Galton and Watson's model, and its many extensions, became widely studied in the

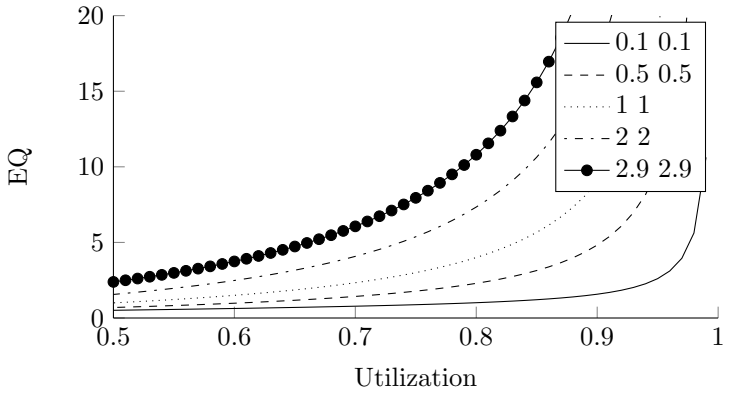


Figure 4.23: $EQ(\rho)$ for various $c_a^2 = c_s^2$

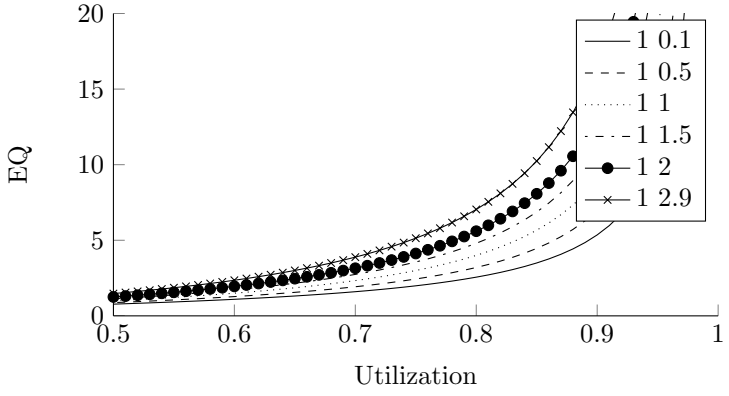


Figure 4.24: $EQ(\rho)$ for $c_a^2 = 1$ and various c_s^2

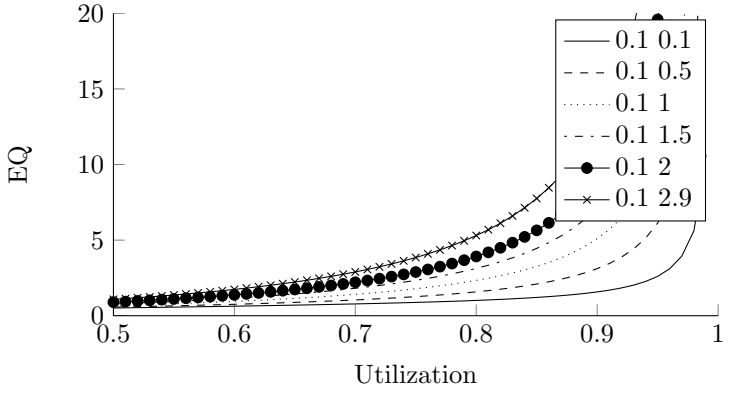


Figure 4.25: $EQ(\rho)$ for $c_a^2 = 0.1$ and various c_s^2

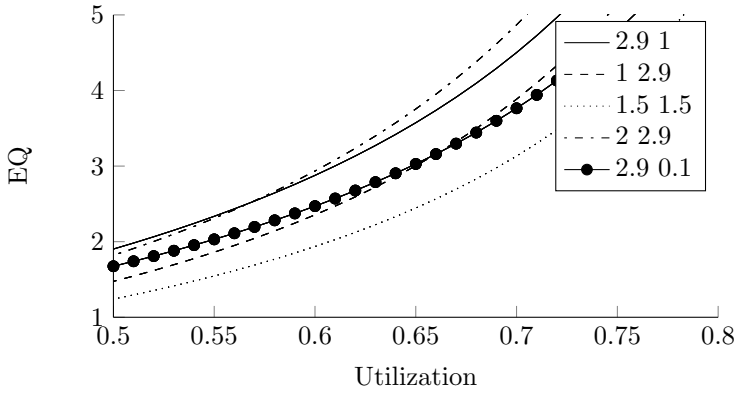


Figure 4.26: $EQ(\rho)$ showing crossing behaviour

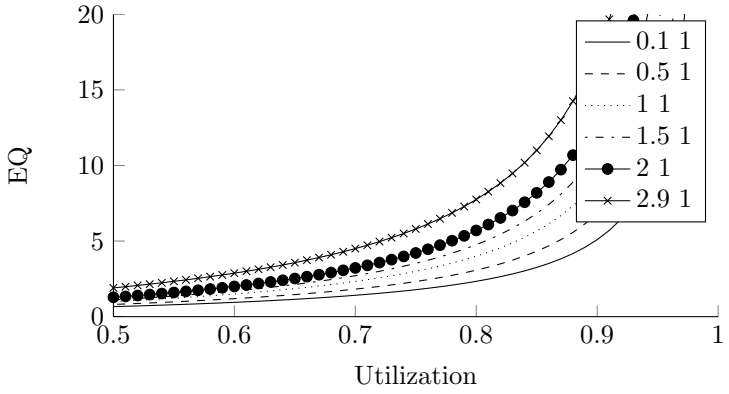


Figure 4.27: $EQ(\rho)$ for $c_s^2 = 1$ and various c_a^2

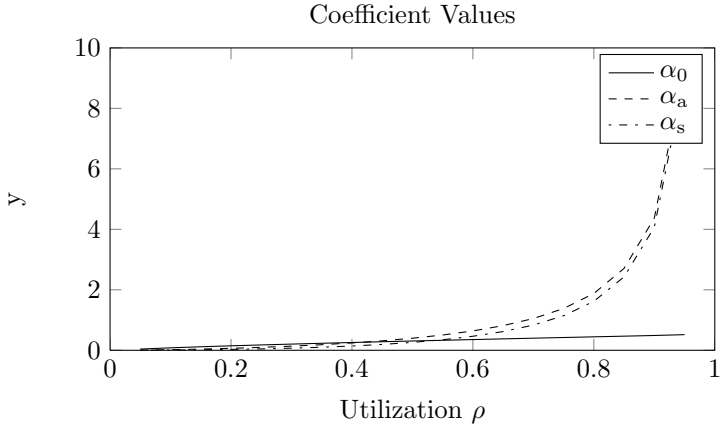


Figure 4.28: $\alpha_0(\rho)$, $\alpha_a(\rho)$ and $\alpha_s(\rho)$, from $EQ = \alpha_0 + \alpha_a c_a^2 + \alpha_s c_s^2$

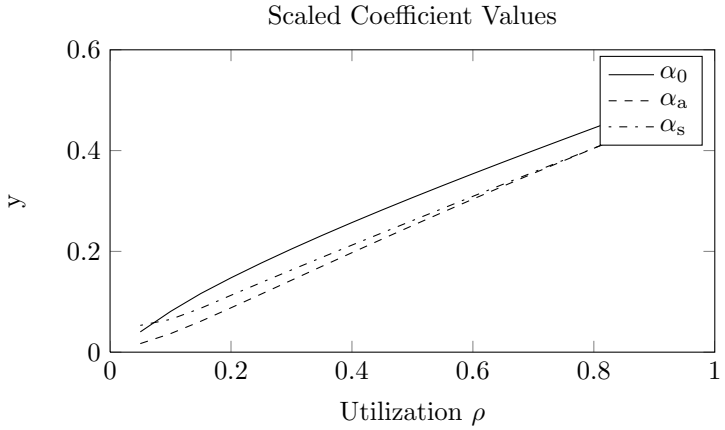


Figure 4.29: $\alpha_0(\rho)$, $\frac{\alpha_a(\rho)}{1-\rho}$ and $\frac{\alpha_s(\rho)}{1-\rho}$

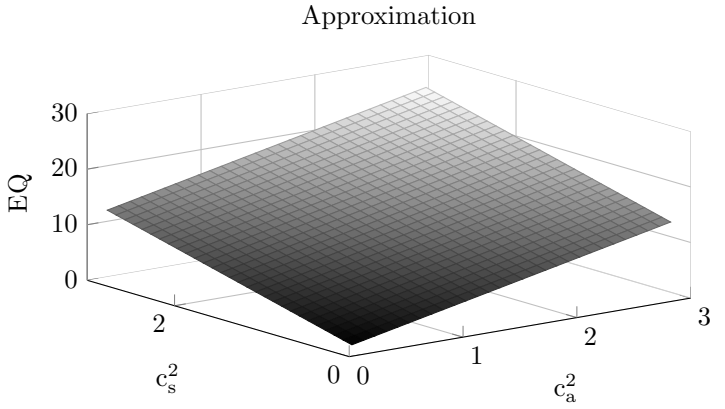


Figure 4.30: $EQ(\rho, c_a^2, c_s^2$ for $EQ = \alpha_0 + \alpha_a c_a^2 + \alpha_s c_s^2$

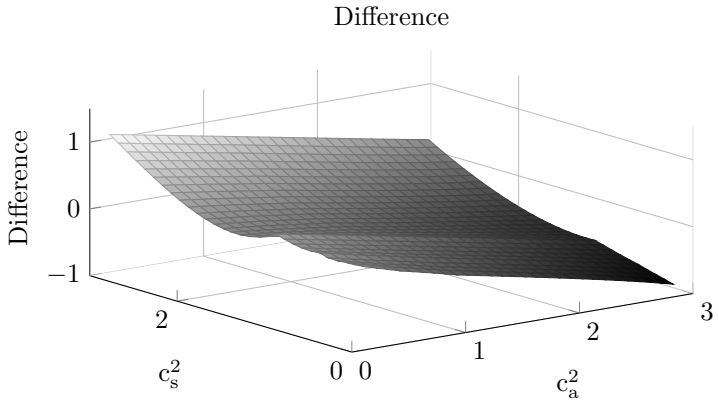


Figure 4.31: Difference between simulation and approximation of EQ

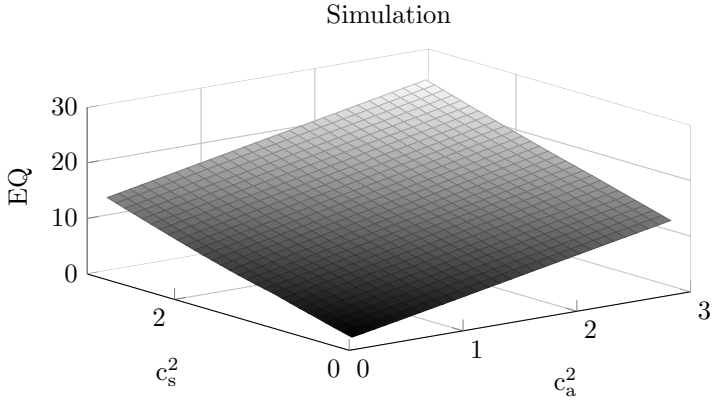


Figure 4.32: $EQ(\rho, c_a^2, c_s^2$ by simulation

1940's, both from a strictly theoretical and from a more practical point of view. Applications ranged from the evolution of genes populations to chain reaction of neutrons, and cosmic rays. This body of work was brought under the name of *branching processes,* which nowadays still form a lively field of research.

Branching processes can be seen as stochastic processes describing the dynamics of a population of individuals which reproduce and die independently, according to some specific probability distributions.

- The individuals may all be identical, or they may belong to one of several types differing by their reproduction and mortality rates – called a *multitype branching process.*

- The individuals evolve in either discrete or continuous time, with exponential or general lifetime distributions, respectively called the *Markovian branching process* and the *age-dependent branching process.*

- The reproduction rules of an individual may depend on the actual size of the population, in a so-called *population size dependent branching process.*

In this lecture, we will first focus on Markovian branching processes, and then we will consider a matrix generalisation of these processes, called *Markovian binary trees.*

4.6.1 Markovian branching processes

Assume that the lifetime of an individual is exponentially distributed with parameter μ and that, during its lifetime, the individual reproduces according to a Poisson process with rate λ, giving birth to one child at a time. All new individuals behave independently of each other, following the same rules as their parent.

Population size at time t

Let Z_t denote the population size in the branching process at time t. The process $\{Z_t\}$, $t \in [0, \infty)$ is called a Markovian branching process; it is a continuous-time Markov chain with state space $E = \{0, 1, 2, 3, \ldots\}$, where state 0 is absorbing, and with associated transition rates (for $i \geq 1$)

$$\begin{array}{rcl} q_{i,i-1} & = & i\,\mu \\ q_{i,i+k} & = & i\,\lambda \\ q_{ii} & = & -i\,(\mu+\lambda) \end{array} \quad .$$

The generator has the following tridiagonal structure

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \dots \\ 0 & 2\mu & -2(\mu+\lambda) & 2\lambda & \dots \\ 0 & 0 & 3\mu & -3(\mu+\lambda) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Such a Markov chain is also called a *linear birth and death process.*
Let $F(s,t)$ be the probability generating function (p.g.f.) of Z_t:

$$F(s,t) = \sum_{k\geq 0} \mathrm{P}[Z_t = k]\, s^k, \quad s \in [0,1].$$

Assume that there is only one individual in the population at time $t = 0$. In order to obtain an equation for $F(s,t)$, we condition on the first event happening to the initial individual before time t:

- either the initial individual is still living at time t and has not reproduced yet, which occurs with probability $e^{-(\mu+\lambda)t}$. In that case, the population at time t is made up of the initial individual only, therefore the p.g.f. of the population size at time t is s,
- or the initial individual dies before reproducing in the time interval $(u, u+du)$ where $u \leq t$, which occurs with probability $\mu\, e^{-(\mu+\lambda)u}\, du$. In that case, the population is empty at time t, therefore the p.g.f. of the population size at time t is 1.
- or, finally, the initial individual reproduces in the time interval $(u, u+du)$ where $u \leq t$, which occurs with probability $\lambda\, e^{-(\mu+\lambda)u}\, du$. In that case, the population at time t is made up of the descendants of both the initial individual and its child after $t-u$ time units, both evolving independently of each other; the p.g.f. of the population size at time t is therefore $F^2(s, t-u)$.

We thus have

$$\begin{aligned} F(s,t) &= s\, e^{-(\mu+\lambda)t} + \int_0^t 1\, \mu\, e^{-(\mu+\lambda)u}\, du + \int_0^t F^2(s,t-u)\, \lambda\, e^{-(\mu+\lambda)u}\, du \\ &= e^{-(\mu+\lambda)t} \left[s + \int_0^t \mu\, e^{(\mu+\lambda)v}\, dv + \int_0^t F^2(s,v)\, \lambda\, e^{(\mu+\lambda)v}\, dv \right]. \end{aligned}$$

By differentiating the last equation with respect to t, we finally obtain the following differential equation for $F(s,t)$

$$\frac{\partial F(s,t)}{\partial t} = \mu - (\mu+\lambda)\, F(s,t) + \lambda F^2(s,t) \tag{4.19}$$

with initial condition $F(s,0) = s$ (since we assume that the branching process starts at time 0 with one individual). This Riccati differential equation corresponds to the backward Kolmogorov equation for the Markov chain Z_t and can be solved explicitly:

$$F(s,t) = \begin{cases} 1 + \dfrac{(\lambda-\mu)(s-1)}{(\lambda-\mu)\,e^{(\mu-\lambda)t} - \lambda(s-1)}, & \text{if } \lambda \neq \mu \\ 1 + \dfrac{(s-1)}{1-\lambda t(s-1)}, & \text{if } \lambda = \mu. \end{cases}$$

The mean population size at time t, denoted as $M(t)$, is obtained by differentiating (4.19) with respect to s at $s = 1$, which gives

$$\frac{dM(t)}{dt} = (\lambda - \mu)\, M(t)$$

with the initial condition $M(0) = 1$. Therefore

$$M(t) = e^{(\lambda-\mu)\,t}. \tag{4.20}$$

We see here that the difference $(\lambda - \mu)$ between the reproduction and death rates plays an important role in the evolution of the branching process. Indeed,

- if $\lambda > \mu$, then $\lim_{t\to\infty} M(t) = \infty$, the population explodes on the average (this case is called the *supercritical* case),
- if $\lambda = \mu$, then $M(t) = 1$ for all t, the mean size of the population stays constant equal to 1, but as we shall see later, the process almost surely becomes extinct (*critical* case),
- if $\lambda < \mu$, then $\lim_{t\to\infty} M(t) = 0$, the population almost surely becomes extinct (*subcritical* case).

Time until extinction

Let T_e denote the time until extinction of the branching process and $F(t)$ its distribution, that is, $F(t) = \mathrm{P}[T_e < t] = \mathrm{P}[Z_t = 0] = F(0,t)$. Thus by taking $s = 0$ in (4.19), we obtain a similar Riccati differential equation for $F(t)$:

$$\frac{\partial F(t)}{\partial t} = \mu - (\mu + \lambda)\, F(t) + \lambda\, F^2(t) \tag{4.21}$$

with initial condition $F(0) = 0$. Its solution is given by

$$F(t) = \begin{cases} 1 + \dfrac{(\mu-\lambda)}{(\lambda-\mu)\,e^{(\mu-\lambda)t} + \lambda}, & \text{if } \lambda \neq \mu \\ 1 - \dfrac{1}{1+\lambda t}, & \text{if } \lambda = \mu. \end{cases}$$

Extinction probability

Let $q = \mathrm{P}[T_e < \infty] = \lim_{t\to\infty} F(t)$ be the probability that the branching process eventually becomes extinct. By taking $t \to \infty$ in (4.21), we obtain that q satisfies the following quadratic equation:

$$0 = \mu - (\mu + \lambda)\, s + \lambda\, s^2.$$

We can rewrite this equation in order to give it a probabilistic interpretation in terms of the branching process. Observe that the total number of children generated by an individual during its lifetime is geometrically distributed with parameter $\mu/(\lambda+\mu)$. Let

$$P(s) = \sum_{n\geq 0} \left(\frac{\lambda}{\mu+\lambda}\right)^n \left(\frac{\mu}{\mu+\lambda}\right) s^n, \quad s \in [0,1]$$

be the p.g.f. of the total number of children generated by an individual during its lifetime. It is also called the *progeny generating function* of an individual. Then

$$\begin{aligned} 0 = \mu - (\mu+\lambda)\, s + \lambda\, s^2 &\Leftrightarrow s = \frac{\mu}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda} s^2 \\ &\Leftrightarrow s = \left(1 - \frac{\lambda}{\mu+\lambda} s\right)^{-1} \frac{\mu}{\mu+\lambda} \\ &\Leftrightarrow s = \sum_{n\geq 0} \left(\frac{\lambda}{\mu+\lambda}\right)^n \left(\frac{\mu}{\mu+\lambda}\right) s^n \\ &\Leftrightarrow s = P(s). \end{aligned}$$

So $q = P(q)$ means that the population generated by the first individual eventually becomes extinct if and only if all the children of this individual generate populations which eventually become extinct independently of each other. Note that the fixed-point equation $s = P(s)$ has always the trivial solution $s = 1$. We arrive to a fundamental result in the theory of branching processes (see for instance [**?**]).

Theorem 4.6.1. *The extinction probability q is the minimal nonnegative solution of the fixed-point equation*

$$s = P(s)$$

where $P(s)$ is the progeny generating function of the individuals.

In our case, we can solve the fixed point equation explicitly, and we find

$$q = \min(1, \mu/\lambda).$$

Note that this solution can also be obtained by taking $t \to \infty$ in the explicit expressions of $F(t)$. We see that

$$q < 1 \Leftrightarrow \lambda > \mu,$$

which again shows that the process has a chance to explode only when $\lambda > \mu$.

In the next section we investigate the matrix generalisation of the Markovian branching process, called the *Markovian binary tree*.

4.6.2 Markovian binary tree

Recall that the matrix generalisation of the exponential distribution is the phase-type distribution. In a Markovian binary tree (MBT), the lifetime of the individuals is distributed according to a PH (α, D) distribution with n transient phases and one absorbing phase. Let $\boldsymbol{d} = -D\mathbf{1}$ be the absorption (or exit) rate vector. In our context, $\boldsymbol{d}$ can thus be interpreted as the death rate vector. It thus means that, during its lifetime, an individual makes hidden transitions among the transient states of an underlying Markov chain and dies once the Markov chain enters its absorbing phase.

By decomposing the transition matrix D of the PH distribution into two parts, the underlying Makov chain can also be used to describe the reproduction process of an individual. Define the $n \times n$ transition rate matrix D_0 and the $n \times n^2$ birth rate matrix B as follows:

- $(D_0)_{ij}$ $(i \neq j)$ is the rate at which an individual moves from phase i to phase j *without* giving birth;
- $B_{i,kj}$ is the rate at which an individual moves from phase i to phase j *and gives birth* to a child which starts its own lifetime chain in phase k.

With this, $D_{ij} = (D_0)_{ij} + \sum_k B_{i,kj}$, that is, in matrix form, $D = D_0 + B(\mathbf{1} \otimes I)$ where $\otimes$ denotes the Kronecker product and I the identity matrix (of size n when not otherwise specified). The diagonal elements of D_0 are strictly negative and $|(D_0)_{ii}|$ is the parameter of the exponential distribution of the sojourn time of an individual in phase i before one of the following events occurs: a phase transition, the birth of a child, or the death of the individual. The matrices and vector satisfy $D_0\mathbf{1} + B\mathbf{1} + \boldsymbol{d} = \mathbf{0}$.

Recall that PH distributions are dense in the set of distributions with nonnegative support. The reproduction process of the individuals is a generalisation of the Poisson process called a (transient) Markovian arrival process (see Latouche *et al.*) which is also dense in the set of point processes [**?**]. These properties make the MBTs a fairly general class of branching processes, with the nice property that they are still Markovian processes.

In the next three subsections, we derive the matrix analogue of the performance measures described in Section 2. We refer to [**?**, **?**, **?**, **?**, **?**] for more details.

4.6.3 Population size at time t

Let $\boldsymbol{Z}(t) = [Z_1(t), \ldots, Z_n(t)]$ be the population size vector at time t in each of the n transient phases, that is, $Z_i(t)$ counts the number of individuals in phase i at time t. We define the conditional p.g.f. of the population size at time t, given that the MBT starts with a first individual in phase i:

$$F_i(\boldsymbol{s}, t) = \sum_{\boldsymbol{k} \geq \mathbf{0}} P[\boldsymbol{Z}(t) = \boldsymbol{k} | \boldsymbol{Z}(0) = \boldsymbol{e}_i] \boldsymbol{s}^{\boldsymbol{k}},$$

where $\boldsymbol{k} = (k_1, \ldots, k_n)$, $k_i \in \mathbb{N}$, $\boldsymbol{e}_i$ is the vector with one 1 at the ith entry and zeros elsewhere, $\boldsymbol{s} = (s_1, \ldots, s_n)^T$, $|s_i| \leq 1$ and $\boldsymbol{s}^{\boldsymbol{k}} = s_1^{k_1} \cdots s_n^{k_n}$. Let $\boldsymbol{F}(\boldsymbol{s}, t) = (F_1(\boldsymbol{s}, t), \ldots, F_n(\boldsymbol{s}, t))^T$ denote the conditional population size generating vector.

To obtain an equation for the vector function $\boldsymbol{F}(\boldsymbol{s}, t)$, we proceed like in the scalar case and condition on the first *observable* event that happens to the initial individual before time t:

- either the initial individual is still living at time t and has not reproduced yet, which occurs with probability $e^{D_0 t}$. In that case, the population at time t is made up of the initial individual only, therefore the p.g.f. of the population size at time t is $\boldsymbol{s}$,
- or the initial individual dies before reproducing in the time interval $(u, u + du)$ where $u \leq t$, which occurs with probability $e^{D_0 u}\, \boldsymbol{d}\, du$. In that case, the population is empty at time t, therefore the p.g.f. of the population size at time t is 1.
- or, finally, the initial individual reproduces in the time interval $(u, u + du)$ where $u \leq t$, which occurs with probability $e^{D_0 u}\, B\, du$. In that case, the population at time t is made up of the descendants of both the initial individual and its child after $t - u$ time units, both evolving independently of each other; the p.g.f. of the population size at time t is therefore $\boldsymbol{F}(s, t-u) \otimes \boldsymbol{F}(s, t-u)$.

We thus have

$$\begin{aligned} F(s,t) &= e^{D_0 t}\, \boldsymbol{s} + \int_0^t e^{D_0 u}\, \boldsymbol{d}\, du + \int_0^t e^{D_0 u}\, B\, (\boldsymbol{F}(s, t-u) \otimes \boldsymbol{F}(s, t-u))\, du \\ &= e^{D_0 t} \left[\boldsymbol{s} + \int_0^t e^{-D_0 v}\, \boldsymbol{d}\, dv + \int_0^t e^{-D_0 v}\, B\, (\boldsymbol{F}(s, v) \otimes \boldsymbol{F}(s, v))\, dv \right]. \end{aligned}$$

By differentiating the last equation with respect to t, we obtain the (backward Kolmogorov) differential equation for the vector function $\boldsymbol{F}(\boldsymbol{s}, t)$

$$\frac{\partial \boldsymbol{F}(\boldsymbol{s}, t)}{\partial t} = \boldsymbol{d} + D_0\, \boldsymbol{F}(\boldsymbol{s}, t) + B\, (\boldsymbol{F}(\boldsymbol{s}, t) \otimes \boldsymbol{F}(\boldsymbol{s}, t)), \tag{4.22}$$

with $\boldsymbol{F}(\boldsymbol{s}, 0) = \boldsymbol{s}$. In this case there is no explicit solution to this matrix quadratic differential equation, but it can be solved numerically using for instance the function `ode45` in Matlab.

Let $M_{ij}(t) = E[Z_j(t) | \boldsymbol{Z}(0) = \boldsymbol{e}_i]$ be the conditional mean number of individuals in phase j in the population at time t given that the process started at time $t = 0$ with one individual in phase i, and $M(t) = (M_{ij}(t))$. We have

$$M_{ij}(t) = \left(\frac{\partial F_i(\boldsymbol{s}, t)}{\partial s_j} \right) \Bigg|_{\boldsymbol{s} = \boldsymbol{1}}. \tag{4.23}$$

Using (4.22) and (4.23), we obtain the matrix differential equation for $M(t)$:

$$\frac{\partial M(t)}{\partial t} = \Omega\, M(t), \tag{4.24}$$

with $M(0) = I$, where

$$\Omega = D_0 + B\,(\mathbf{1} \otimes I + I \otimes \mathbf{1}). \tag{4.25}$$

By solving (4.24) we obtain that the matrix of the mean population size at time t is given by

$$M(t) = e^{\Omega t}. \tag{4.26}$$

In this case, the dynamics of the branching process depends on the dominant eigenvalue ω of the matrix Ω, which plays a role analogue to $\lambda - \mu$ in the Markovian branching process: if $\omega > 0$, then the mean population size explodes, while if $\omega \leq 0$, the mean population size stays bounded.

4.6.4 Time until extinction

Let $F_i(t)$ denote the conditional probability that the population becomes extinct before time t, given that it started at time $t = 0$ with one individual in phase i, and $\boldsymbol{F}(t) = (F_1(t), \ldots, F_n(t))^T$. Similar to Section 2.2, we have $\boldsymbol{F}(t) = \boldsymbol{F}(\mathbf{0}, t)$, so by taking $\boldsymbol{s} = \mathbf{0}$ in (4.22), we obtain

$$\frac{\partial \boldsymbol{F}(t)}{\partial t} \;=\; \boldsymbol{d} + D_0\, \boldsymbol{F}(t) + B\,(\boldsymbol{F}(t) \otimes \boldsymbol{F}(t)), \tag{4.27}$$

with $\boldsymbol{F}(0) = \mathbf{0}$. As for $\boldsymbol{F}(\boldsymbol{s}, t)$, numerical tools are necessary for solving this matrix quadratic differential equation.

4.6.5 Extinction probability

Let q_i denote the conditional probability that the population eventually becomes extinct, given that it started at time $t = 0$ with one individual in phase i, and let $\boldsymbol{q} = (q_1, \ldots, q_n)^T$ be the extinction probability vector. We have $\boldsymbol{q} = \lim_{t\to\infty} \boldsymbol{F}(t)$, so by taking $t \to \infty$ in (4.27), we obtain an equation for $\boldsymbol{q}$:

$$\mathbf{0} = \boldsymbol{d} + D_0\, \boldsymbol{q} + B\,(\boldsymbol{q} \otimes \boldsymbol{q}). \tag{4.28}$$

Define $\boldsymbol{\theta} = (-D_0)^{-1}\, \boldsymbol{d}$ and $\Psi = (-D_0)^{-1}\, B$. The entry θ_i is the pobability that an individual in phase i eventually dies before giving birth, and $\Psi_{i,kj}$ is the probability that an individual in phase i reproduces before dying by giving birth to a child starting in phase k while the parent is in phase j after the birth. Equation (4.28) can be rewritten as

$$\boldsymbol{q} = \boldsymbol{\theta} + \Psi(\boldsymbol{q} \otimes \boldsymbol{q}) \tag{4.29}$$

which has the following interpretation: the population eventually becomes extinct if and only if the initial individual dies before reproducing, or if it reproduces and both the child and the parent generate populations which eventually become extinct (independently of each other).

Using the following property of Kronecker products

$$\Psi(\boldsymbol{q} \otimes \boldsymbol{q}) = \Psi(\boldsymbol{q} \otimes I)\, \boldsymbol{q} = \Psi(I \otimes \boldsymbol{q})\, \boldsymbol{q},$$

we can rewrite (4.29) equivalently as

$$\boldsymbol{q} = [I - \Psi(\boldsymbol{q} \otimes I)]^{-1}\, \boldsymbol{\theta} \tag{4.30}$$

or

$$\boldsymbol{q} = [I - \Psi(I \otimes \boldsymbol{q})]^{-1}\, \boldsymbol{\theta}. \tag{4.31}$$

Note that $\boldsymbol{P}(\boldsymbol{s}) = [I - \Psi(\boldsymbol{s} \otimes I)]^{-1}\, \boldsymbol{\theta}$ is the conditional progeny generating function of an individual, given its initial phase. Therefore Equation (4.30) can be rewritten as $\boldsymbol{q} = \boldsymbol{P}(\boldsymbol{q})$. None of the equivalent matrix equations (4.29), (4.30), and (4.31) can be solved analytically, but they can be solved numerically using for instance

- the *Depth algorithm*, which is a linear algorithm based on the functional iteration method applied to (4.29) (see [**?**]),
- the *Order algorithm*, which is a linear algorithm based on the functional iteration method applied to (4.30) or to (4.31), and works faster than the Depth algorithm (see [**?**]),
- the *Thicknessess algorithm*, which is a linear algorithm based on the functional iteration method applied *alternately* to (4.30) and (4.31), and works faster than the Order algorithm in some circumstances (see [**?**]),
- the *Newton algorithms*, which are quadratic algorithms based on the Newton method applied to (4.29) (see [**?**]), and to (4.30) or (4.31) (see [**?**]).

Note that all of these algorithms have a probabilistic interpretation in terms of the branching process. For instance, the kth iteration of the order algoritm computes the probability that the MBT becomes extinct before the kth generation.

Recall that ω is the dominant eigenvalue of the matrix Ω defined in (4.25). The following theorem provides an extinction criterion for the MBT.

Theorem 4.6.2. $\boldsymbol{q} = \mathbf{1} \Leftrightarrow \omega \leq 0$.

Therefore, the algorithms listed above are useful only in the case where $\omega > 0$ (super-critical case).

4.6.6 Sensitivity analysis

In this section, we perform some sensitivity analysis of some performance measures of an MBT (such as the mean population size or the extinction probability) with respect to perturbations or errors on the parameters of the model.

Let p be a parameter of the model (for instance $p = d_i$ or $p = B_{i,kj}$), and let X be a performance measure obtained from the model (for instance $X = M_{ij}(t)$, or $X = q_i$). We define the *sensitivity* of X with respect to p as the local slope of X, considered as a function of p:

$$\partial_p X = \frac{\partial X}{\partial p}. \tag{4.32}$$

The scale of X and p may be different; it is therefore convenient to consider proportional perturbations instead of absolute ones. The proportional response to a proportional perturbation is the *elasticity*. The elasticity of X with respect to p is defined by the ratio of the relative increase of X to the relative increase of p:

$$\frac{\partial \log X}{\partial \log p} = \partial_p X \, \frac{p}{X}. \tag{4.33}$$

The interpretation of the elasticity is as follows: if $\partial \log X / \partial \log p = a$, it follows that if p increases by 1 percent, then X increases by approximatively $100[\exp(a/100) - 1]$ percent. Note that $100[\exp(a/100) - 1)] \approx a$ when a is small.

Sensitivity of the mean population size at time t

We derive an explicit formula for $\partial_p M(t)$, where $M(t) = \exp(\Omega t)$.

Let us consider the following system of differential equations

$$\left\{ \begin{array}{rcl} \partial_t M(t) & = & \Omega \, M(t) \\ \partial_t \partial_p M(t) & = & \Omega \, \partial_p M(t) + \partial_p \Omega \, M(t), \end{array} \right.$$

where the first equation is the differential equation (4.24) satisfied by $M(t)$, and the second differential equation is obtained by differentiating the first one with respect to p. This system may be equivalently rewritten as

$$\partial_t \left[\begin{array}{c} \partial_p M(t) \\ M(t) \end{array} \right] = \left[\begin{array}{cc} \Omega & \partial_p \Omega \\ 0 & \Omega \end{array} \right] \cdot \left[\begin{array}{c} \partial_p M(t) \\ M(t) \end{array} \right],$$

with initial condition $[\partial_p M(0), M(0)]^T = [0, I]^T$. This is a new differential equation of the form $\partial_t Y(t) = A\, Y(t)$, of which the solution is given by $Y(t) = \exp(A\, t)\, Y(0)$. Therefore,

$$\partial_p M(t) = [I, 0] \cdot \exp\left(\left[\begin{array}{cc} \Omega & \partial_p \Omega \\ 0 & \Omega \end{array} \right] t \right) \cdot \left[\begin{array}{c} 0 \\ I \end{array} \right]. \tag{4.34}$$

Sensitivity of the extinction probability

We derive an explicit formula for $\partial_p \boldsymbol{q}$ when $\boldsymbol{q} \neq \mathbf{1}$. By differentiating (4.28) with respect to p, we obtain

$$\begin{aligned} \mathbf{0} &= \partial_p \boldsymbol{d} + \partial_p D_0\, \boldsymbol{q} + D_0\, \partial_p \boldsymbol{q} + \partial_p B\, (\boldsymbol{q} \otimes \boldsymbol{q}) + B\, (\partial_p \boldsymbol{q} \otimes \boldsymbol{q}) + B\, (\boldsymbol{q} \otimes \partial_p \boldsymbol{q}) \\ &= [D_0 + B(\boldsymbol{q} \otimes I + I \otimes \boldsymbol{q}]\, \partial_p \boldsymbol{q} + \partial_p \boldsymbol{d} + \partial_p D_0\, \boldsymbol{q} + \partial_p B\, (\boldsymbol{q} \otimes \boldsymbol{q}). \end{aligned}$$

If $[D_0 + B(\boldsymbol{q} \otimes I + I \otimes \boldsymbol{q}]$ is invertible, then we obtain

$$\partial_p \boldsymbol{q} = -[D_0 + B(\boldsymbol{q} \otimes I + I \otimes \boldsymbol{q}]^{-1} [\partial_p \boldsymbol{d} + \partial_p D_0\, \boldsymbol{q} + \partial_p B\, (\boldsymbol{q} \otimes \boldsymbol{q})].$$

4.6.7 Application in demography

The transient phases of the Markov chain controlling the lifetime of the individuals in the MBT may be purely fictitious, or they may have some physical interpretation, such as the age or the health condition of the individuals.

We model female families in several countries with MBTs, in which the transient phases correspond to successive age classes, and reproduction events correspond to the birth of daughters only. We adapt the data on age-specific fertility and mortality rates to the parameters of the MBT as explained below.

The United Nations web sites [**?**, **?**] give fertility and mortality rates for five-yearly age classes, in addition to infant mortality rates. This had led us to model the lifetime of a woman in such a way that one transient phase corresponds to one age interval. Therefore, we have $n = 22$ phases in all, which correspond to the age classes $5-9$, $10-14$, $\ldots$, $95-99$, there is one class for the newborn (age 0), one for the class $1-4$, and finally one for women aged 100 and above. The interval $0-4$ is split in two in order to make use of the available infant mortality rates.

The time unit is the year, and the matrix D_0 of hidden transition rates is given by

$$D_0 = \begin{bmatrix} * & 1 & & & & \\ & * & 1/4 & & & \\ & & * & 1/5 & & \\ & & & \ddots & & \\ & & & & * & 1/5 \\ & & & & & * \end{bmatrix}$$

where a $*$ on the diagonal indicates a number such that $D_0\mathbf{1} + B\mathbf{1} + \boldsymbol{d} = \mathbf{0}$. It means that, in the absence of death, a woman spends an expected amount of time of one year at age 0, four years in the interval $1-4$, and 5 years thereafter, until being over 100.

The age-specific fertility rate in age class i is defined as the number living births during the year, according to the age class i of the mother, for each 1000 women of the same

age class i. Since the fertility data available to us does not distinguish between the birth of girls and the birth of boys, we use the sex ratio (defined as the ratio between the number of births of boys and the number of births of girls) to adapt the global fertility rates from [**?**]. Therefore the female birth rate β_i per individual in phase i is

$$\beta_i = \frac{\text{age-specific fertility rate in } i}{1000 \cdot (\text{sex ratio} + 1)},$$

and the birth rate matrix $B = (\boldsymbol{e}_1 \otimes \text{diag}(\boldsymbol{\beta}))$.

Finally, the age-specific mortality rate d_i in age class i is defined as the number of deaths during the year of women in age class i divided by the population in the age class i, and $\boldsymbol{d} = (d_1, \ldots, d_{22})^T$.

Mean family size

The mean total size at time t of the family generated by a woman born at time 0 is thus given by $m(t) = [e^{\Omega t}\,\mathbf{1}]_1$. We plot $m(t)$ as a function of t in Figure 4.33 for five countries. We see how fast the Congolese family grows compared to the other countries. Also, we see that during the first 70 years, the mean family sizes in Turkey and in Morocco are the same, but that they follow diverging paths after that, the Turkish family size, which is nearly critical, growing very slowly.

Figure 4.33: Mean family size generated by a new-born woman as a function of time.

Time until extinction

We plot in Figure 4.34 the distribution of the time until extinction of the family generated by a new-born woman, that is $F_1(t)$, for t between 0 and 180.

We note a big difference in the shape of the curves for Congo and for South Africa, compared to the other countries. We interpret this as a youth mortality effect; it reflects also the fact that if a Congolese family eventually becomes extinct, it happens quite quickly, before it has had time to grow. So it seems that if the first mother and her young daughters survive, then the family has a high probability to be maintained, which explains why the curve of Congo is already almost constant, and significantly below 1, after 100 years only.

Extinction probability

Here, $\boldsymbol{q}$ represents the conditional probability that the female family generated by a single woman eventually becomes extinct, given the age class of this woman at time 0.

Figure 4.35 shows each entry of $\boldsymbol{q}$. We only plot the results for supercritical countries, as the extinction probability is $\boldsymbol{q} = \mathbf{1}$ for subcritical and critical countries, regardless of

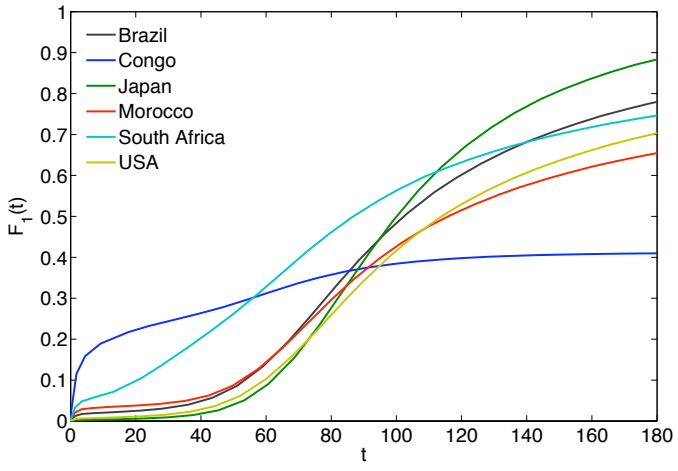


Figure 4.34: Distribution function of the time until extinction of the family generated by one new-born woman.

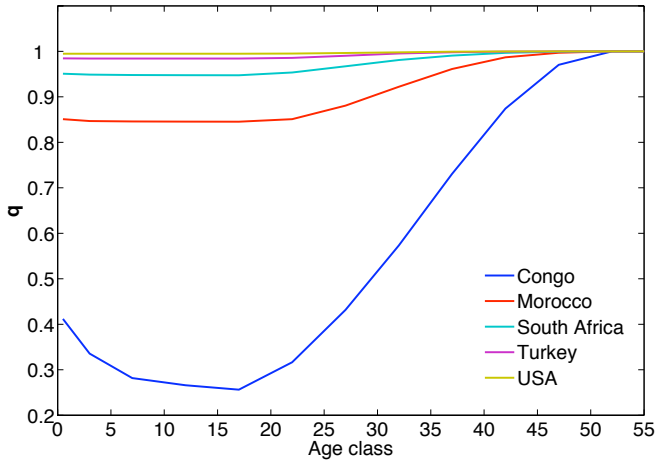


Figure 4.35: Extinction probability of the family generated by first woman as a function of her age class.

the age class of the initial woman. It is especially useful to compare countries and to see the effect of the age class of the initial woman on the potential survival of her female family.

We notice the infant and the juvenile mortality effects: for instance, the family generated by a Congolese girl aged between 1 and 4 years has a higher probability of eventually becoming extinct than for a girl aged between 15 and 19 years, because the first girl has a significantly lower probability than the second one of reaching adulthood. Most of the other countries also have a detectable dip, which is to be expected since in every country

there is some rate of mortality for girls before they reach reproductive maturity.

Finally, observe that as time increases, the curves in Figure 4.34 do tend to the extinction probabilities q_1.

More details on the MBT model in demography can be found in [**?**].

Sensitivity analysis

We show in Figures 4.36 and 4.37 the elasticity of the mean family size and of the extinction probability vector, respectively, with respect to the infant mortality rate d_1. We see that the absolute value of the elasticities are higher for Congo than for other countries, which shows that a small perturbation of the infant mortality rate in Congo would have a higher impact on the dynamics of the population than in another country. On the contrary, a small perturbation in the infant mortality rate would have almost no impact on the family evolution in Turkey and in the US.

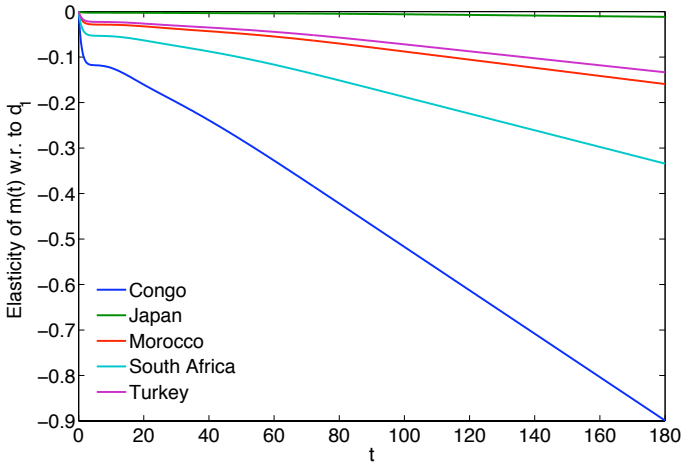


Figure 4.36: Elasticity of the mean total population size at time t with respect to the infant mortality rate d_1.

Bibliographic Remarks

Exercises

Tandem queueing system. We consider a system of two single-server queues in tandem, as described in Figure QQQQ..

The assumptions are as follows. New customers arrive according to a Poisson process with rate λ and join a first waiting room in front of sever 1. They receive a first service,

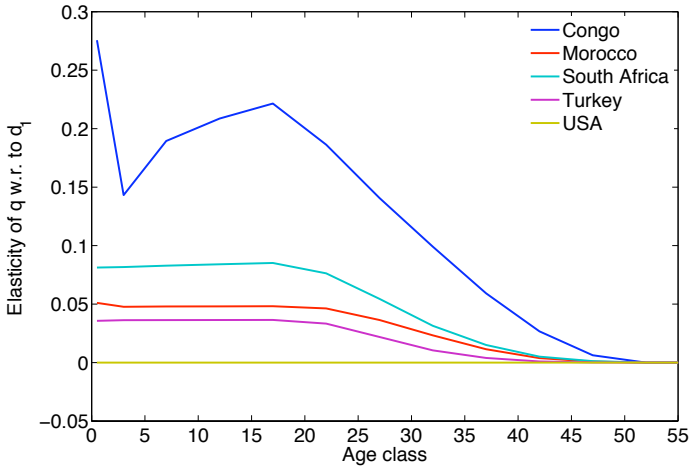


Figure 4.37: Elasticity of the extinction probability vector with respect to the infant mortality rate d_1.

exponentially distributed with rate μ_1, then they move to a second waiting room and eventually receive a second service, exponentially distributed with rate μ_2. The first waiting room has finite capacity $K-1$, so that there can be at most K customers in the first half of the system (which corresponds to an M/M/1/K queue). The second waiting room does not have any capacity restriction.

1. Model this tandem queueing system as a level-independent QBD by appropriately defining the level and the phase of the process at time t.
   (Hint: the number of customers in each sub-system may correspond to one dimension of the QBD).

2. Describe the block matrices B, A_{-1}, A_0, and A_1 involved in the generator of the QBD.

3. Determine a necessary and sufficient condition for positive recurrence of the QBD in terms of K, λ, μ_1, and μ_2.

4. Take $K = 3$, $\lambda = 1$, $\mu_1 = 1/2$, and $\mu_2 = 1$, and compute the stationary distribution vector $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots]$ of the QBD (provide $\boldsymbol{\pi}_n$ for a few values of n).

   For that purpose, you first need to compute the matrix G using the algorithm presented in Section 4.4 of the lecture notes, then the matrix R, and finally, the stationary probability vector at level 0, $\boldsymbol{\pi}_0$ (note that we are in the continuous-time setting).

5. Using your results from (d), compute (numerically) the asymptotic (steady state) mean number of customers who have received the first but not the second service (this sub-question should give you a second hint for sub-question (a)).

6. Same questions as in (d) and (e) with $\mu_2 = 1/2$. Comment on your results.

Chapter 5

State Feedback, Observers and Separation in Their Design (4h)

We now move away from Structured Markov Chains (QBDs) and return to the (A, B, C, D) systems introduced in Section 3.4:

$$\begin{array}{rcl} \dot{\boldsymbol{x}}(t) &=& A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &=& C\boldsymbol{x}(t) + D\boldsymbol{u}(t) \end{array} \quad \text{or} \quad \begin{array}{rcl} \boldsymbol{x}(\ell+1) &=& A\boldsymbol{x}(\ell) + B\boldsymbol{u}(\ell) \\ \boldsymbol{y}(\ell) &=& C\boldsymbol{x}(\ell) + D\boldsymbol{u}(\ell) \end{array} . \tag{5.1}$$

As described in that section, these systems relate 3 processes: *input* $\boldsymbol{u}(t) \in \mathbb{R}^m$, *state* $\boldsymbol{x}(t) \in \mathbb{R}^n$ and *output* $\boldsymbol{y}(t) \in \mathbb{R}^p$, or their discrete time versions (with ℓ instead of t). The SISO case ($m = 1$ and $p = 1$) remains of special interest.

In Chapter 3 such systems were introduced in order to build up the description of PH distributions: Their step response has a matrix exponential component, identical to that of a *matrix exponential* distribution which generalizes the PH distributions. Now, in the current chapter, we consider these systems in their own right. They are useful for modeling a variety of situations, often physical.

After illustrating how the pendulum and water tank examples can be modeled as (A, B, C, D) systems, we move on to discuss the *control* and *state measurement* properties of (A, B, C, D) systems. In this respect, the general idea is to design a mechanism for updating the input of the system, $\boldsymbol{u}(\cdot)$ based on either the current state, or alternatively based on indirect state measurements through the output, $\boldsymbol{y}(\cdot)$. This discussion starts out with regularity conditions (*controllability* and *observability*), and after discussing basic transformations of (A, B, C, D) systems (canonical forms), we move onto demonstrate the two basic operational constructs of Linear Control Theory (LCT): *state feedback* and *state observers*.

As stated at the onset of this course, LCT is a well established field that has had profound impact on many engineering applications in the past 50 years. This chapter only aims to give a brief mathematical glimpse by illustrating the basic concepts.

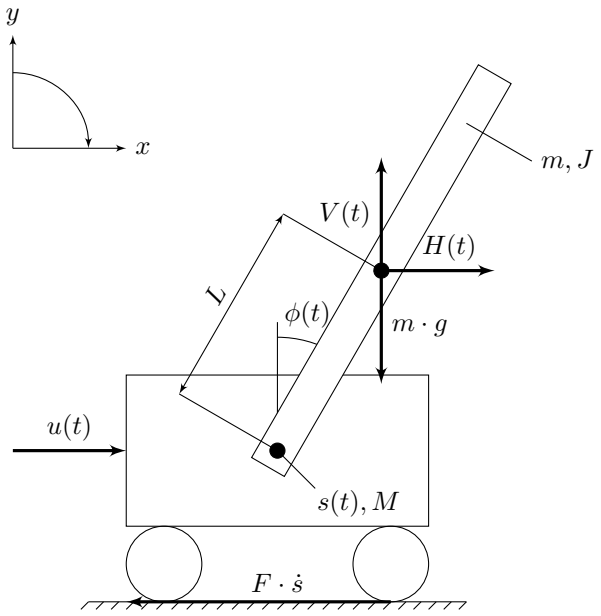


Figure 5.1: Visualization of the dynamics of the inverted pendulum on a cart

5.1 Examples of (A, B, C, D) Systems Needing Control

We now present models of the pendulum and fluid tank applications, first introduced in Section 1.3. The purpose is to hint (or illustrate) at the *modeling* steps that one takes in order to produce (A, B, C, D) models. We note that carrying out such modeling requires plenty of care and thought. This is not the essence of our current course and may require one or more specialized engineering, operations research or science courses.

In both examples we begin with basic physical principles, apply linearization and other transformations and finally reach an (A, B, C, D) model.

The "Inverted Pendulum on a Cart" Example

The pivot of the pendulum is mounted on a carriage that can move horizontally. The carriage is driven by a motor that exerts a force $u(t)$ on the carriage.

The displacement of the pivot (center of the carriage) at time t is $s(t)$ and the angular rotation at time t is $\phi(t)$. The carriage has mass M. The mass of the pendulum is m and the distance from its pivot to its center of gravity is L. The moment of inertia with respect to the center of gravity is J. Friction is accounted for only in the motion of the carriage and is assumed proportional to speed with coefficient F. It is not accounted for at the pivot. We further assume that m is small with respect to M and neglect the horizontal reaction force on the motion of the carriage.

The forces exerted on the pendulum are:

1. The force mg at the center of gravity (g is the gravitational acceleration).
2. A horizontal reaction force, $H(t)$
3. A vertical reaction force, $V(t)$

Based on Newton's second law for motion on a line (force = mass × acceleration) and Newton's second law for angular motion (torque = moment of inertia × angular acceleration), together with basic trigonometry and "physical thinking" we get the following equations:

$$\begin{aligned} m\frac{d^2}{dt^2}\big(s(t) + L\sin\phi(t)\big) &= H(t), && (5.2)\\ m\frac{d^2}{dt^2}L\cos\phi(t) &= V(t) - mg, && (5.3)\\ J\frac{d^2\phi(t)}{dt^2} &= LV(t)\sin\phi(t) - LH(t)\cos\phi(t), && (5.4)\\ M\frac{d^2s(t)}{dt^2} &= u(t) - F\frac{ds(t)}{dt}. && (5.5) \end{aligned}$$

Carrying out the derivatives above and rearranging, we get the two equations,

$$\ddot{\phi}(t) - \frac{g}{L'}\sin\phi(t) + \frac{1}{L'}\ddot{s}(t)\cos\phi(t) = 0, \tag{5.6}$$

and,

$$M\ddot{s}(t) = u(t) - F\dot{s}(t), \tag{5.7}$$

where L' is the "effective pendulum length":

$$L' := \frac{J + mL^3}{mL}.$$

A solution to this system is $u(t), s(t), \phi(t) \equiv 0$. We can now get rid of the non-linear terms by linearization (Taylor series expansion of the sin and cos in the first equation). This makes the first equation:

$$\ddot{\phi}(t) - \frac{g}{L'}\phi(t) + \frac{1}{L'}\ddot{s}(t) = 0. \tag{5.8}$$

Now having the linear ODEs with constant coefficients, (5.7) and (5.8), at hand, we are essentially ready to define *state*, *input* and *output*:

- For the **state** set:

  $$x_1(t) := s(t), \quad x_2(t) := \dot{s}(t), \quad x_3(t) := s(t) + L'\phi(t), \quad x_4(t) := \dot{s}(t) + L'\dot{\phi}(t).$$

  Note that the third component is the a linearized approximation to the displacement of a point of the pendulum at distance L'.

- As **input** to the system, simply take $u(t)$ (the force exerted on the carriage).
- As **output** to the system take,

$$y_1(t) := \phi(t), \quad y_2(t) := s(t).$$

We may now represent the system as an (A, B, C, D) system:

$$\dot{\boldsymbol{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{F}{M} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{g}{L'} & 0 & \frac{g}{L'} & 0 \end{bmatrix} \boldsymbol{x}(t) + \begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ 0 \end{bmatrix} u(t), \tag{5.9a}$$

$$\boldsymbol{y}(t) = \begin{bmatrix} -\frac{1}{L'} & 0 & \frac{1}{L'} & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \boldsymbol{x}(t). \tag{5.9b}$$

Observe that (as is often the case) $D = 0$. Further the matrices, A, B and C are generally quite sparse.

Exercise 5.1.1. *Complete the missing steps in the example above, to get the* (A, B, C, D) *system specified above.*

The "Fluid Tank" in a Chemical Engineering Process Example

Consider a fluid tank with two incoming feeds with time-varying flow rates, $F_1(t)$ and $F_2(t)$. Both feeds contain dissolved material with constant concentrations c_1 and c_2 respectively. The outgoing flow (from a drain at the bottom of the tank) has flow rate $F(t)$. The tank is assumed to be "well-stirred" so that the concentration of the outgoing flow equals the concentration $c(t)$ in the tank. Let $V(t)$ be the volume of the fluid in the tank. Assume that the tank has constant cross-sectional area, S, so that the height of the fluid level, $h(t)$, follows,

$$h(t) = \frac{V(t)}{S}.$$

We can now postulate (model) that,

$$F(t) = k\sqrt{\frac{V(t)}{S}},$$

where k is an experimental constant. We thus have the "mass-balance" equations:

$$\frac{dV(t)}{dt} = F_1(t) + F_2(t) - k\sqrt{\frac{V(t)}{S}}, \tag{5.10}$$

$$\frac{d}{dt}\big(c(t)V(t)\big) = c_1F_1(t) + c_2F_2(t) - c(t)k\sqrt{\frac{V(t)}{S}}. \tag{5.11}$$

The second equation is for the quantity of the dissolved material, $c(t)V(t)$.

One often takes equations of the above form, and looks for an *operating point* (or a fixed point, or equilibrium point). The idea is to assume all quantities are constant and get equations of the form,

$$\begin{aligned} 0 &= F_{10} + F_{20} - F_0, \\ 0 &= c_1 F_{10} + c_2 F_{20} - c_0 F_0, \\ F_0 &= k\sqrt{\frac{V_0}{S}}. \end{aligned}$$

Here F_{10} and F_{20} are nominal flow rates of the two feeds. Similarly F_0 is the nominal outgoing rate and c_0 is the nominal outgoing concentration. Under these nominal values, the volume of the tank is V_0. For a desired c_0 and F_0, it is then a *static problem* to find $V_0 > 0$ and $F_{10}, F_{20} \geq 0$ that solve the equations. We do not concern our self here with methods of finding such an operating point.

Once an operating point is found, it is convenient to define,

$$\theta := \frac{V_0}{F_0},$$

this is the "holdup time" of the fluid tank. We now define,

$$\begin{aligned} F_1(t) &= F_{10} + u_1(t), \\ F_2(t) &= F_{20} + u_2(t), \\ V(t) &= V_0 + x_1(t), \\ c(t) &= C_0 + x_2(t). \end{aligned}$$

We then take $\boldsymbol{u}(t) = [u_1(t),\ u_2(t)]'$ as **input** to the system and $\boldsymbol{x}(t) = [x_1(t),\ x_2(t)]'$ as the **state**. As output we take $\boldsymbol{y}(t) = [y_1(t),\ y_2(t)]'$ with,

$$y_1(t) = F(t) - F_0, \qquad y_2(t) = c(t) - c_0 = x_2(t).$$

Equations (5.10) and (5.11) are non-linear. Hoping that we shall operate our system near the operating point, we linearize the equations. Combining with the definitions of state, input and output we get the following (A, B, C, D) system:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \begin{bmatrix} -\frac{1}{2\theta} & 0 \\ 0 & -\frac{1}{\theta} \end{bmatrix} \boldsymbol{x}(t) + \begin{bmatrix} 1 & 1 \\ \frac{c_1 - c_0}{V_0} & \frac{c_2 - c_0}{V_0} \end{bmatrix} \boldsymbol{u}(t), \\ \boldsymbol{y}(t) &= \begin{bmatrix} \frac{1}{2\theta} & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{x}(t). \end{aligned}$$

5.2 Controllability and Observability Conditions

We now introduce two *regularity conditions*: *controllability* and *observability*. The first deals with the ability to "drive the state" of the system to any given point. The second deals with the ability to "reconstruct the state" based on input and output observations (without seeing the state). For (A, B, C, D) systems, these conditions have very clear algebraic characterizations.

5.2.1 Controllability

A state $\boldsymbol{x}_d \in \mathbb{R}^n$ is said to be *reachable* (synonymous with *controllable-from-the-origin*) if there exists an input $\boldsymbol{u}(\cdot)$ that transfers $\boldsymbol{x}(\cdot)$ from the zero state ($\boldsymbol{x}(0) = \boldsymbol{0}$) to $\boldsymbol{x}_d$ in some finite time. A state $\boldsymbol{x}_s \in \mathbb{R}^n$ is said to be *controllable* if there exists an input that transfers the state from $\boldsymbol{x}_s$ to the zero state in some finite time. I.e. if $\exists \boldsymbol{u}(\cdot)$, such that when $\boldsymbol{x}(0) = \boldsymbol{x}_s$ we have, $\boldsymbol{x}(\tau) = \boldsymbol{0}$ for some $\tau < \infty$. These definitions are applicable to both discrete and continuous time systems. If a system is not controllable we say it is *uncontrollable*.

We note (without proof here) that while reachability always implies controllability, controllability implies reachability only when the state transition matrix (A^ℓ or e^{At}) is non-singular. This is always true for continuous time systems but for discrete time systems it is required that A is non-singular. We will mostly ignore discrete time systems with singular A and thus treat reachability of a state and controllability of a state as essentially synonymous terms.

The set of all reachable/controllable states is called the *reachable / controllable subspace* of the system. It indeed holds that this set is a linear sub-space of $\mathbb{R}^n$ (again without proof here).

We say the whole system is *reachable / controllable* if any state is *reachable /controllable*, i.e. if the reachable / controllable subspace is $\mathbb{R}^n$. In this case we may also say that the pair (A, B) is *reachable / controllable*. Observe that the notions of reachabillity/controllability do not involve (C, D).

A key structure in the development is the matrix $\text{con}_k(A, B)$, defined for positive integer k as follows:

$$\text{con}_k(A, B) := \left[B, AB, A^2B, \ldots, A^{k-1}B\right] \in \mathbb{R}^{n \times mk}.$$

To see the source of the $\text{con}_k(A, B)$ matrix, consider the discrete time system with k-step input sequence reversed in time:

$$\overline{\boldsymbol{u}}_k = \left[\boldsymbol{u}(k-1)', \boldsymbol{u}(k-2)', \ldots, \boldsymbol{u}(0)'\right]' \in \mathbb{R}^{km}.$$

Since the evolution of state is,

$$\boldsymbol{x}(\ell) = A^\ell \boldsymbol{x}(0) + \sum_{i=0}^{\ell-1} A^{\ell-(i+1)} B \boldsymbol{u}(i),$$

we have that with input $\overline{\boldsymbol{u}}$ over time steps, $0, 1, \ldots, k-1$, the state at time k can be represented by:

$$\boldsymbol{x}(k) = A^k \boldsymbol{x}(0) + \mathrm{con}_k(A, B)\overline{\boldsymbol{u}}_k. \tag{5.12}$$

Exercise 5.2.1. *Carry out the (block)-matrix operations to obtain* (5.12).

Hence the $\mathrm{con}_k(A, B)$ matrix captures the propagation of state in discrete time systems. As we shall see, it is also used in continuous time systems.
The following lemma summarizes important properties of $\mathrm{con}_k(A, B)$:

Lemma 5.2.2. *If* $k < n$,

$$range\Big(con_k(A, B)\Big) \subset range\Big(con_n(A, B)\Big).$$

If $k \geq n$,

$$range\Big(con_k(A, B)\Big) = range\Big(con_n(A, B)\Big).$$

Proof. The statement for $k < n$ is obvious as adding columns to a matrix can only increase the dimension of its range.
Now (as a reminder) the Cayley-Hamilton theorem states that for arbitrary $A \in \mathbb{R}^{n\times n}$, with characteristic polynomial,

$$p(s) := \det(sI - A) = s^n + p_{n-1}s^{n-1} + \ldots + p_1 s + p_0,$$

we have the following matrix identity,

$$A^n + p_{n-1}A^{n-1} + \ldots + p_1 A + p_0 I = 0_{n\times n}.$$

Hence,

$$A^n = -p_{n-1}A^{n-1} - \ldots - p_1 A - p_0 I.$$

Alternatively,

$$A^n B = -p_0 B - p_1 AB - \ldots - p_{n-1}A^{n-1}B.$$

So the additional m columns in $\mathrm{con}_{n+1}(A, B)$ compared to $\mathrm{con}_n(A, B)$ (these are the columns of $A^n B$) are linear combinations of the columns of $\mathrm{con}_n(A, B)$. Further the additional m columns in $\mathrm{con}_{n+2}(A, B)$ (these are $A^{n+1}B$) are,

$$AA^n B = -p_0 AB - p_1 A^2 B + \ldots + -p_{n-2}A^{n-1}B. - p_{n-1}A^n B.$$

and these are linear combinations of columns of $\mathrm{con}_{n+1}(A, B)$. Continuing by induction the result is proved. □

We are now ready to specify a necessary and sufficient condition for reachability/controllability based on the so-called *controllability matrix* which we define as follows:

$$\mathrm{con}(A,B) := \mathrm{con}_n(A,B).$$

I.e. it is the matrix that can be used to examine the state propagation over inputs for a number of time steps equal to the dimension of the state of the system.

Theorem 5.2.3. *A discrete time (A,B,C,D) system is reachable if and only if*

$$\mathit{rank}\Big(\mathit{con}(A,B)\Big) = n \qquad \mathit{or\ alternatively} \qquad \mathit{range}\Big(\mathit{con}(A,B)\Big) = \mathbb{R}^n.$$

Proof. It is possible to transfer from state $\boldsymbol{x}_s$ to state $\boldsymbol{x}_d$ in k steps if and only if there exists an input sequence, $\overline{\boldsymbol{u}}$ such that

$$\mathrm{con}_k(A,B)\,\overline{\boldsymbol{u}} = \boldsymbol{x}_d - A^k\boldsymbol{x}_s.$$

That is for reachability, set $\boldsymbol{x}_s = 0$ and the system is reachable if and only if there is an integer k, such that,

$$\boldsymbol{x}_d \in \mathrm{range}\Big(\mathrm{con}_k(A,B)\Big).$$

Now if $\mathrm{rank}\Big(\mathrm{con}(A,B)\Big) = n$ then any $\boldsymbol{x}_d$ can be reached in n steps and thus it is reachable.
Conversely if it is reachable, since $\boldsymbol{x}_d$ is arbitrary, there is a k for which,

$$\mathrm{range}\Big(\mathrm{con}_k(A,B)\Big) = \mathbb{R}^n \qquad \text{or alternatively} \qquad \mathrm{rank}\Big(\mathrm{con}_k(A,B)\Big) = n.$$

If then $k \leq n$ we must have by the first part of Lemma 5.2.2 that,

$$n \geq \mathrm{rank}\Big(\mathrm{con}(A,B)\Big),$$

and hence $\mathrm{rank}\Big(\mathrm{con}(A,B)\Big) = n$. Alternatively, if $k \geq n$ then by the second part of Lemma 5.2.2 we have,

$$n = \mathrm{rank}\Big(\mathrm{con}(A,B)\Big).$$

Hence having the controllability matrix be full-rank is a necessary condition for controllability. □

Exercise 5.2.4. *Explain why every state in a controllable discrete time system may be reached in n steps or less. Give an example of a system and a state that can not be reached faster than n steps (hint: Think of the manufacturing line example and about "filling up work in the buffers").*

5.2.2 Continuous Time

For continuous time systems, $\text{con}_k(A, B)$ does not have the same direct meaning as in (5.12) yet plays a central role. Assume $\boldsymbol{x}(0) = \boldsymbol{x}_s$ and an input $\{\boldsymbol{u}(t),\, t \in [0, T]\}$ is applied such that $\boldsymbol{x}(T) = \boldsymbol{x}_d$, then,

$$\boldsymbol{x}_d = e^{AT}\boldsymbol{x}_s + \int_0^T e^{A(T-\tau)}B\boldsymbol{u}(\tau)d\tau.$$

The *reachability sub-space during time* $[0, T]$ is then:

$$\mathcal{R}_T := \big\{\boldsymbol{x} \in \mathbb{R}^n : \exists\ \{\boldsymbol{u}(t),\, t \in [0, T]\},\ \text{such that},\ \boldsymbol{x} = \int_0^T e^{A(T-\tau)}B\boldsymbol{u}(\tau)d\tau\big\}.$$

Lemma 5.2.5. *For any $T > 0$,*

$$\mathcal{R}_T \subset \textit{range}\Big(\textit{con}(A, B)\Big).$$

Theorem 5.2.6. *A continuous time (A, B, C, D) system is reachable/controllable if and only if*

$$\textit{rank}\Big(\textit{con}(A, B)\Big) = n.$$

Exercise 5.2.7. *Show that the pendulum example from Section 5.1 is controllable.*

Exercise 5.2.8. *Show that the fluid tank example from Section 5.1 is controllable if and only if $c_1 \neq c_2$. Explain why this makes sense.*

5.2.3 Observability

A system is said to be *observable* if knowledge of the outputs and the inputs over some finite time interval is enough to determine the initial state $\boldsymbol{x}(0)$. For a discrete time system this means that $\boldsymbol{x}(0)$ can be uniquely identified based on $\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(\ell_f - 1)$ and $\boldsymbol{u}(0), \ldots, \boldsymbol{u}(\ell_f - 1)$ for some finite ℓ_f. For continuous time systems it means that $\boldsymbol{x}(0)$ can be uniquely identified by $\{\boldsymbol{y}(t),\, t \in [0, t_f]\}$ and $\{\boldsymbol{u}(t),\, t \in [0, t_f]\}$ for some finite t_f.

The development of observability criteria, generally parallels that of controllability. For discrete time systems,

$$\boldsymbol{y}(\ell) = CA^k\boldsymbol{x}(0) + \sum_{i=0}^{\ell-1} CA^{\ell-(i+1)}B\boldsymbol{u}(i) + D\boldsymbol{u}(\ell).$$

or alternatively define,

$$\tilde{\boldsymbol{y}}(k) = \boldsymbol{y}(k) - \big(\sum_{i=0}^{k-1} CA^{k-(i+1)}B\boldsymbol{u}(i) + D\boldsymbol{u}(k)\big),$$

and,

$$\mathrm{obs}_k(A,C) = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{k-1} \end{bmatrix} \in \mathbb{R}^{pk\times n}.$$

Then,

$$\mathrm{obs}_k(A,C)\,\boldsymbol{x}(0) = \begin{bmatrix} \tilde{\boldsymbol{y}}(0) \\ \tilde{\boldsymbol{y}}(1) \\ \vdots \\ \tilde{\boldsymbol{y}}(k-1) \end{bmatrix}. \tag{5.13}$$

The system is thus observable in k time units if (5.13) has the same unique solution, $\boldsymbol{x}(0)$ for any k.
We define the *observability matrix* as :

$$\mathrm{obs}(A,C) := \mathrm{obs}_n(A,C).$$

Theorem 5.2.9. *A discrete or continuous (A,B,C,D) system is observable if and only if,*

$$rank\Big(obs(A,C)\Big) = n.$$

We omit the proof.

Exercise 5.2.10. *Try to complete the proof above, following similar lines to the proof of the controllability condition.*

5.2.4 Duality between Controllability and Observability

Consider the (A,B,C,D) system,

$$\begin{array}{rcl} \dot{\boldsymbol{x}}(t) &=& A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &=& C\boldsymbol{x}(t) + D\boldsymbol{u}(t) \end{array}. \tag{5.14}$$

The *dual system* is defined as,

$$\begin{array}{rcl} \dot{\boldsymbol{x}}(t) &=& A'\boldsymbol{x}(t) + C'\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &=& B'\boldsymbol{x}(t) + D'\boldsymbol{u}(t) \end{array}. \tag{5.15}$$

Notice that in the dual system, the state dimension is still n, but the dimensions of the input and the output were switched: The new input dimension is p and the new output dimension is m. The same definition holds for discrete time systems.

Theorem 5.2.11. *The system* (5.14) *is controllable if and only if the dual system* (5.15) *is observable. Similarly the system* (5.14) *is observable if and only if the dual system* (5.15) *is controllable.*

Proof. We have that,

$$\mathrm{con}(A, B) = \mathrm{obs}(A', B')', \quad \mathrm{obs}(A, C)' = \mathrm{con}(A', C').$$

□

The same definitions and result hold for discrete time systems.

5.2.5 Uncontrollable and Unobservable Systems

It sometimes occurs in practice and or theory that systems are not controllable. In practice one typically tries to design the actuators of the system as to make it controllable, yet sometimes this is either not possible or not needed. In theory, we shall need results of such systems in the sequel (e.g. to show the importance of controllability for state-feedback).
Similarly, as in the exercise below, it sometimes occurs that systems are not observable.

Exercise 5.2.12. *Consider the pendulum* (A, B, C, D) *system from Section 5.1. For that system,*

$$C = \begin{bmatrix} -\frac{1}{L'} & 0 & \frac{1}{L'} & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Assume that the second row of C *is removed. That is the sensor for reading the displacement of the system,* $s(t)$ *is removed. Show that the system is not observable.*

In such cases of uncontrollability and unobservability, there are a variety of results dealing with partitioning of the state space $\mathbb{R}^n$ into linear sub-spaces, that are controllable/uncontrollable, observable/unobservable. The most general result of this flavor is referred to as *The Kalman Decomposition*, we shall not cover it in generality here. One of the sub-cases of the Kalman Decomposition is the theorem below (it is used in the sequel).

Theorem 5.2.13. *If the pair* (A, B) *is uncontrollable there exists* $P \in \mathbb{R}^{n\times n}$ *with* $\det(P) \neq 0$ *such that,*

$$P^{-1}AP = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix}, \qquad P^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

with $A_1 \in \mathbb{R}^{n_r\times n_r}$, $B_1 \in \mathbb{R}^{n_r\times m}$, *where* $n_r < n$ *and the pair* (A_1, B_1) *is controllable.*

Exercise 5.2.14. *Explain the dynamics of the* $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ *system with* $(\tilde{A}, \tilde{B})$ *as above. Why is it obvious that the coordinates* $x_{n_r+1}, \ldots, x_n$ *are not controllable?*

5.3 Canonical Forms

In Section 3.4, we saw that by setting $\tilde{\boldsymbol{x}} = P^{-1}\boldsymbol{x}$ for some $P \in \mathbb{R}^{n\times n}$ with $\det(P) \neq 0$ we get a system, equivalent to (A, B, C, D),

$$\left(\tilde{A},\ \tilde{B},\ \tilde{C},\ \tilde{D}\right) = \left(P^{-1}AP,\ P^{-1}B,\ CP, D\right).$$

The equivalence is in the sense that the input-output mapping that the system induces (e.g. the step response) is not altered by the change of coordinates of the state space. See Exercise 3.6.7.

In this section we see how to find useful P matrices so as to put the equivalent $(\tilde{A},\ \tilde{B},\ \tilde{C},\ \tilde{D})$ system in a useful form. We mostly focus on SISO systems and search for equivalent representations of systems following very special forms (canonical forms). These forms will useful in designing feedback controllers and observers in the sections below.

Since the matrices A and $\tilde{A}$ are similar, an important property of the similarity transform is that the similar matrices, A and $\tilde{A} = P^{-1}AP$ share the same eigenvalues or alternatively the same characteristic polynomial.

Exercise 5.3.1. *Show that indeed,*

$$\det\left(sI - A\right) = \det\left(sI - P^1AP\right).$$

Controller Canonical Form (The SISO case)

A SISO system, $(\tilde{A}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{c}}', \tilde{d})$ is said to be in *controller form* if,

$$\tilde{A} = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & 1 \\ -p_0 & -p_1 & \dots & \dots & \dots & -p_{n-1} \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{b}} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

There is no restriction on the structure of the $\tilde{\boldsymbol{c}}'$ and $\tilde{d}$ components.

Exercise 5.3.2. *Show that the coefficients of the characteristic polynomial of $\tilde{A}$ are, $p_0, p_1, \dots, p_{n-1}, 1$, i.e.:*

$$\det\left(sI - \tilde{A}\right) = s^n + p_{n-1}s^{n-1} + \dots + p_1 s + p_0.$$

Theorem 5.3.3. *If $(A, \boldsymbol{b})$ is controllable then there exists, P such that $(A, \boldsymbol{b}, \boldsymbol{c}', d)$ has an equivalent representation in controller form.*

Proof. Since $(A, \boldsymbol{b})$ is controllable, the $n \times n$ (SISO case) controllability matrix, $\text{con}(A, \boldsymbol{b})$ is non-singular. Denote now the last row of the inverse, $\text{con}(A, \boldsymbol{b})^{-1}$ by $\boldsymbol{q}'$, i.e,

$$[B \; BA \; BA^2 \; \dots \; BA^{n-1}]^{-1} = \begin{bmatrix} * \\ \boldsymbol{q}' \end{bmatrix}.$$

We set now,

$$P = \begin{bmatrix} \boldsymbol{q}' \\ \boldsymbol{q}'A \\ \boldsymbol{q}'A^2 \\ \vdots \\ \boldsymbol{q}'A^{n-1} \end{bmatrix}^{-1},$$

and using the equivalent system representation (3.40), we can show that $\tilde{A}$ and $\tilde{\boldsymbol{b}}$ are in the desired controller form. □

Exercise 5.3.4. *Complete the details of the above proof. In the process, show that P is non-singular.*

Observer Canonical Form (The SISO case)

In a similar manner to the controller canonical form, a SISO system, $(\tilde{A}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{c}}', \tilde{d})$ is said to be in *observer form* if,

$$\tilde{A} = \begin{bmatrix} 0 & \dots & \dots & \dots & 0 & -p_0 \\ 1 & 0 & \dots & \dots & \dots & -p_1 \\ 0 & 1 & 0 & \dots & \dots & -p_2 \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \dots & 0 & 1 & 0 & -p_{n-2} \\ 0 & \dots & \dots & 0 & 1 & -p_{n-1} \end{bmatrix} \quad \text{and}$$

$$\tilde{\boldsymbol{c}}' = \begin{bmatrix} 0 & \dots & \dots & \dots & 0 & 1 \end{bmatrix}.$$

Theorem 5.3.5. *If (A, C) is controllable then there exists, P such that (A, B, C, D) has an equivalent representation in observer form.*

Exercise 5.3.6. *Try to complete the proof of the above theorem.*

Extensions to MIMO Systems

There are extensions of the observer canonical form and controller canonical form to MIMO systems. We do not cover these here.

5.4 State Feedback Control

Having an (A, B, C, D) system at hand, we may wish to modify its behavior. This is after all the main use case of linear control theory (see Section 1.3). One goal may be to stabilize it (more on that in Chapter 6). A mathematically general way to do this is called *state feedback control.* (or linear state feedback control).

The idea is to select a matrix $K_f \in \mathbb{R}^{m\times n}$ and set the input to the system as follows:

$$\boldsymbol{u}(t) = -K_f\boldsymbol{x}(t) + \boldsymbol{r}(t).$$

Here $\boldsymbol{r}(\cdot)$, stands for a *reference input.*

Now the dynamics of the system under state feedback control become:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= (A - BK_f)\boldsymbol{x}(t) + B\boldsymbol{r}(t), && (5.16)\\ \boldsymbol{y}(t) &= (C - DK_f)\boldsymbol{x}(t) + D\boldsymbol{r}(t), && (5.17) \end{aligned}$$

and similarly for the discrete time case. Note that in the SISO case, F is a row vector and we denote it by $\boldsymbol{k}'$. In this case, the dynamics of the state feedback control are denoted:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= (A - \boldsymbol{b}\boldsymbol{k}_f')\boldsymbol{x}(t) + \boldsymbol{b}r(t), && (5.18)\\ y(t) &= (\boldsymbol{c}' - d\boldsymbol{k}_f')\boldsymbol{x}(t) + dr(t), && (5.19) \end{aligned}$$

Exercise 5.4.1. *Show how* (5.19) *and* (5.19) *is obtained.*

We thus see that under state feedback control we get a new system that responds to the reference input, $\boldsymbol{r}(\cdot)$ like the (A, B, C, D) system,

$$\big(A - BK_f,\ B,\ C - DK_f,\ D\big),$$

and similarly if using the SISO notation. One typical control application of this is to set $\boldsymbol{r}(t) \equiv 0$. Thus the state evolution becomes,

$$\dot{\boldsymbol{x}}(t) = (A - BK_f)\boldsymbol{x}(t).$$

Now a goal of a *controller design* is to find the state feedback matrix K such that the solution,

$$\boldsymbol{x}(t) = e^{(A-BK_f)t}\boldsymbol{x}(0),$$

behaves as desired. In practice, the first "desirable behavior" is that of stability as will be discussed in Chapter 6, yet other criteria may include responsiveness, robustness to errors in parameters, and reduction of oscillatory behavior. We do not touch these issues further here, yet note that these constitute a good part of an engineering oriented control course.

The theorem below shows that for controllable systems, we may find state-feedback laws so as to achieve arbitrary desired behavior.

Theorem 5.4.2. *Given $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ there exists $K_f \in \mathbb{R}^{m\times n}$ such that the n eigenvalues of $A - BK_f$ are assigned to arbitrary, real or complex conjugate locations if and only if (A, B) is a controllable pair.*

5.5 Observers

We now show how to design a system based on the original system whose state is denoted by $\hat{\boldsymbol{x}}(\cdot)$ and is designed so that $\hat{\boldsymbol{x}}(t)$ is an estimate of $\boldsymbol{x}(t)$. This simple (yet very powerful idea) is called the *Luenberger observer*. The basic equation in the design of the "observer system" is this:

$$\dot{\hat{\boldsymbol{x}}}(t) = A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) - K_o\big(\hat{\boldsymbol{y}}(t) - \boldsymbol{y}(t)\big),$$

where $K_o \in \mathbb{R}^{n\times p}$ and

$$\hat{\boldsymbol{y}}(t) = C\hat{\boldsymbol{x}}(t) + D\boldsymbol{u}(t).$$

Combining we have,

$$\dot{\hat{\boldsymbol{x}}}(t) = (A - K_oC)\hat{\boldsymbol{x}}(t) + \big[B - K_oD, K_o\big] \begin{bmatrix} \boldsymbol{u}(t) \\ \boldsymbol{y}(t) \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = C\hat{\boldsymbol{x}}(t) + \big[D, 0\big] \begin{bmatrix} \boldsymbol{u}(t) \\ \boldsymbol{y}(t) \end{bmatrix}$$

Thus the Luenberger observer system, associated with the system (A, B, C, D) is the system,

$$\big(A - K_oC,\ \big[B - K_oD, K_o\big],\ C,\ \big[D, 0\big]\big),$$

whose input is $[\boldsymbol{u}'(\cdot), \boldsymbol{y}'(\cdot)]'$, i.e. the input of the original system together with the output of the original system. See Figure 5.2.

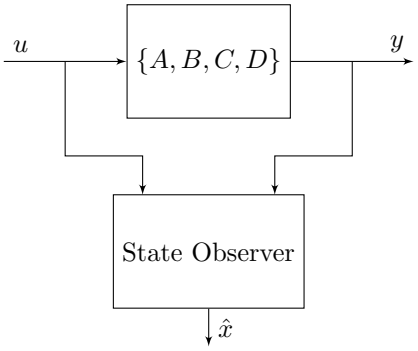


Figure 5.2: A system with an observer

As opposed to the original system which in typical applications has some physical manifestation, the observer is typically implemented in one way or another (often using digital

computers). The *state* of the observer, $\hat{\boldsymbol{x}}(\cdot)$ is thus **accessible by design** and as we show now can yield a very good estimate of the actual (non-fully accessible) state, $\boldsymbol{x}(\cdot)$.

Exercise 5.5.1. *Show that if $\hat{\boldsymbol{x}}(0) = \boldsymbol{x}(0)$ then $\hat{\boldsymbol{x}}(t) = \boldsymbol{x}(t)$ for all $t \geq 0$.*

The *estimation error* between the state and the estimate is

$$\boldsymbol{e}(t) := \boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t),$$

. Thus,

$$\begin{aligned}
\dot{\boldsymbol{e}}(t) &= \dot{\boldsymbol{x}}(t) - \dot{\hat{\boldsymbol{x}}}(t) \\
&= (A\boldsymbol{x}(t) + B\boldsymbol{u}(t)) - \big(A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) - K_o(\hat{\boldsymbol{y}}(t) - \boldsymbol{y}(t)\big) \\
&= (A\boldsymbol{x}(t) + B\boldsymbol{u}(t)) - \big(A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) - K_o((C\hat{\boldsymbol{x}}(t) + D\boldsymbol{u}(t)) - (C\boldsymbol{x}(t) + D\boldsymbol{u}(t))\big) \\
&= (A - K_oC)(\boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)) \\
&= (A - K_oC)\dot{\boldsymbol{e}}(t).
\end{aligned}$$

Hence the estimation error associated with the Luenberger observer behaves like the linear dynamical (autonomous) system,

$$\dot{\boldsymbol{e}}(t) = (A - K_oC)\boldsymbol{e}(t).$$

Hence we can design the behavior of the estimation error of the error term based on K_o. More on this on stability, yet at this point note that if K_o is designed so that $(A - K_oC)$ has eigenvalues strictly in the LHP then $\boldsymbol{e}(t) \to 0$ as $t \to \infty$ yielding an *asymptotic state estimator*. I.e. the estimation error would vanish as time progresses!!!! This is for any initial condition of both the system, $\boldsymbol{x}(0)$ and the observer $\hat{\boldsymbol{x}}(t)$.

It turns out that the *observability* condition is exactly the condition that specifies when the autonomous system $(A - K_oC)$ can be shaped arbitrarily. The following theorem is a parallel of Theorem 5.4.2

Theorem 5.5.2. *There is a $K_o \in \mathbb{R}^{n\times p}$ so that eigenvalues of $A - K_oC$ are assigned to arbitrary locations if and only if the pair (A, C) is observable.*

Proof. The eigenvalues of $(A - K_oC)' = A' - C'K_o'$ are arbitrarily assigned via K_o' if and only if the pair (A', C') is controllable (Theorem 5.4.2). This by duality (Theorem 5.2.11) occurs if and only if (A, C) is observable. □

A related concept is the Kalman filter.

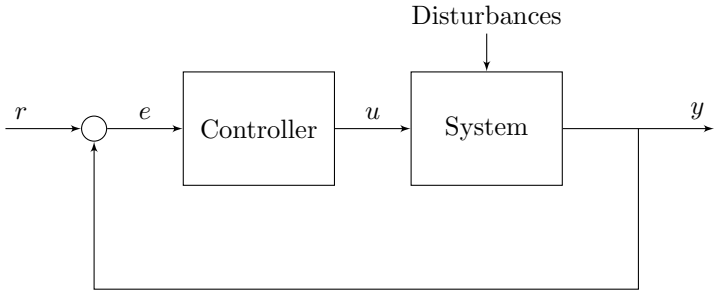


Figure 5.3: A controlled system

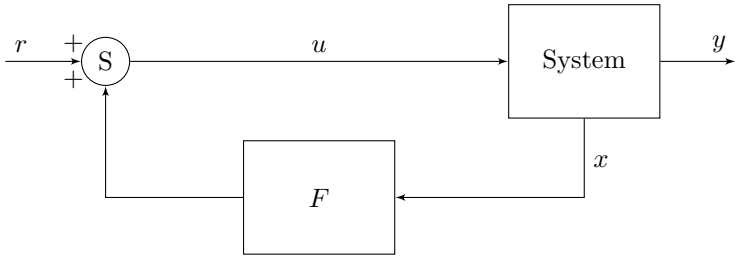


Figure 5.4: A system with feedback

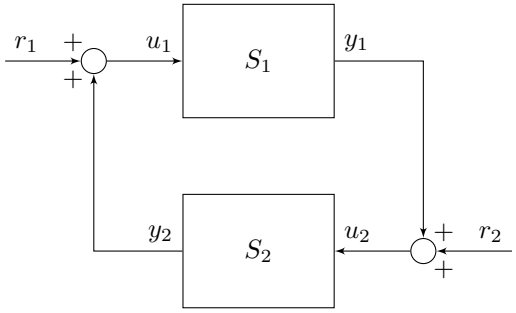


Figure 5.5: feedbacka

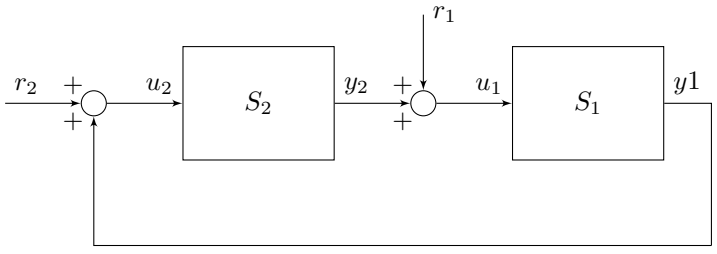


Figure 5.6: feedbackb

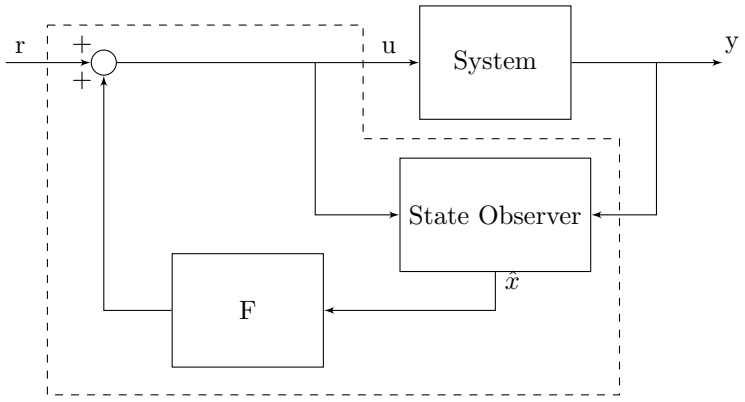


Figure 5.7: A system with a controller and observer

5.6 The Separation Principle

Now that we know about state feedback and observers, we can combine them practically into a controlled system that has an observer for generating $\hat{\boldsymbol{x}}(\cdot)$ and then uses $\hat{\boldsymbol{x}}(\cdot)$ as input to a "state feedback" controller. This means that the input is,

$$\boldsymbol{u}(t) = -K_f\hat{\boldsymbol{x}}(t) + \boldsymbol{r}(t). \tag{5.20}$$

where as before, the observer follows,

$$\dot{\hat{\boldsymbol{x}}}(t) = A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) - K_o\big(\hat{\boldsymbol{y}}(t) - \boldsymbol{y}(t)\big).$$

with,

$$\hat{\boldsymbol{y}}(t) = C\hat{\boldsymbol{x}}(t) + D\boldsymbol{u}(t).$$

Combining the above with $\boldsymbol{y}(t) = C\boldsymbol{x}(t) + D\boldsymbol{u}(t)$ we get,

$$\dot{\hat{\boldsymbol{x}}}(t) = (A - K_oC)\hat{\boldsymbol{x}}(t) + K_oC\boldsymbol{x}(t) + B\boldsymbol{u}(t).$$

Hence if we now combine (5.20) and look at the *compensated system* (original plant together with a state feedback law operating on an observer estimate), we get:

$$\begin{aligned}
\begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ \dot{\hat{\boldsymbol{x}}}(t) \end{bmatrix} &= \begin{bmatrix} A & -BK_f \\ K_oC & A - K_oC - BK_f \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \hat{\boldsymbol{x}}(t) \end{bmatrix} + \begin{bmatrix} B \\ B \end{bmatrix} \boldsymbol{r}(t), \\
\boldsymbol{y}(t) &= \begin{bmatrix} C & -DK_f \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \hat{\boldsymbol{x}}(t) \end{bmatrix} + D\boldsymbol{r}(t).
\end{aligned}$$

Thus the compensated system is of state dimension $2n$ and has as state variables both the state variables of the system $\boldsymbol{x}(\cdot)$ and the observer "virtual"-state variables $\hat{\boldsymbol{x}}(\cdot)$.

It is useful to apply the following similarity transform to the system:

$$P \begin{bmatrix} \boldsymbol{x} \\ \hat{\boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \hat{\boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{e} \end{bmatrix}$$

Hence as in (3.40), the resulting system is:

$$\begin{aligned} \begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ \dot{\boldsymbol{e}}(t) \end{bmatrix} &= \begin{bmatrix} A - BK_f & -BK_f \\ 0 & A - K_oC \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{e}(t) \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} \boldsymbol{r}(t) \\ \boldsymbol{y}(t) &= \begin{bmatrix} C - DK_f & -DK_f \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{e}(t) \end{bmatrix} + D\boldsymbol{r}(t) \end{aligned}$$

Exercise 5.6.1. *Show that the system above (with state ($\boldsymbol{x}$,$\boldsymbol{e}$)) is not controllable.*

The reason for not being fully controllable is that the state at the coordinates of corresponding to $\boldsymbol{e}(\cdot)$ should converge to $\mathbf{0}$, independently of $\boldsymbol{r}(\cdot)$.
Notice now that,

$$\det\left(sI_{2n} - \begin{bmatrix} A - BK_f & -BK_f \\ 0 & A - K_oC \end{bmatrix}\right) = \det\big(sI_n - (A - BK_f)\big)\ \det\big(sI_n - (A - K_oC)\big).$$

This implies that the behavior (determined by the eigenvalues) of the compensated system can be fully determined by selecting K_f and K_o separately! I.e,

char' poly' of $\boldsymbol{x}$ and $\boldsymbol{e}$ dynamics =
char' poly' resulting from choice of K_f $\times$ char' poly' resulting from choice of K_o.

This is called the *separation principle* and it has far reaching implications: **One may design the controller and the state estimator in separation and then combine. The dynamics of one will not affect the dynamics of the other.**

5.7 Examples of Control

Bibliographic Remarks

Exercises

This exercise is concerned with robustness. Loosely speaking, we call a controlled system robust if small errors in the model or in the controller have small effects on the controlled behavior. In this exercise, we consider robustness both with respect to measurement errors and with respect to parameter uncertainty. Consider the input-output system

$$6y(t) - 5\dot{y}(t) + \ddot{y}(t) = u(t). \tag{5.21}$$

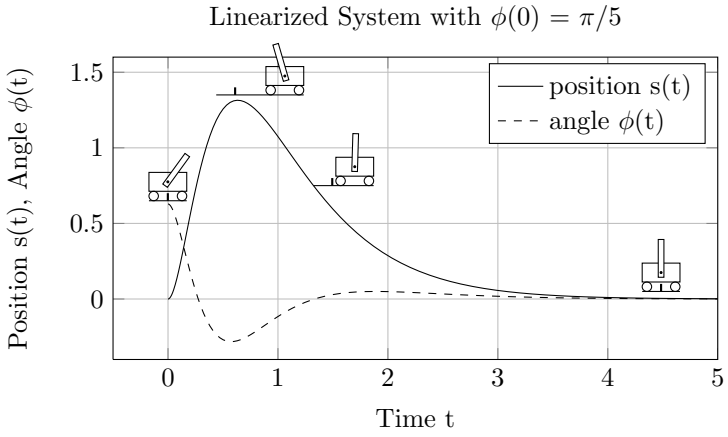


Figure 5.8: Illustrated simulation results, for $\phi(0) = \frac{\pi}{5}$

1. Show that this system is open loop ($u(t) = 0$) unstable. I.e. show that even for the input $u(t) = 0$, the output diverges.

Assume that we want to stabilize the system using feedback control (as is in Section 2.5 in the course reader). Our first attempt is

$$u(t) = -5\dot{y}(t) + \ddot{y}(t). \tag{5.22}$$

It appears that this yields an extremely fast and accurate controller, since the system output is

$$y(t) = 0.$$

We now investigate whether the proposed controller is indeed such a superior controller. If we were able to implement the controller with infinite precision, then, there seems to be no problem. Suppose, however, that this controller is implemented by means of a sensor that does not measure $y(t)$ exactly. Assume that the sensor output is $y(t) + v(t)$, where $v(t)$ is a (deterministic) noise term (also known as a disturbance). The controller is then given by

$$u(t) = -5\big(\dot{y}(t) + \dot{v}(t)\big) + \ddot{y}(t) + \ddot{v}(t).$$

2. Determine the output $y(t)$ for the case that $v(t) = \epsilon \sin(2\pi f t)$, $\epsilon > 0$, $f \in \mathbb{R}$. Conclude that an arbitrarily small disturbance can have a significant impact if f is sufficiently large. Thus, the controller (5.22) is not robust with respect to measurement noise.

3. Determine the controller canonical form for the system (5.21). I.e. propose a state representation and describe the system as an (A, B, C, D) system in controller canonical form.

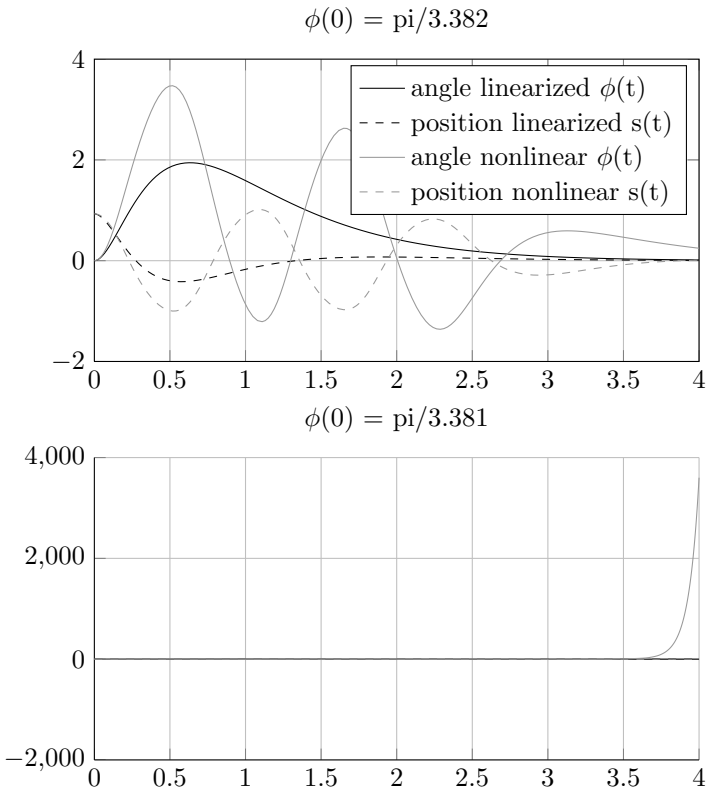


Figure 5.9: Illustration of the tipping point at which K_f is no longer a stabilizing controller for the non-linear system

4. Prove that the system (5.21) can not be stabilized by static output feedback, that is by a feedback of the form $u(t) = -ky(t)$.

5. Determine now a state feedback that assigns the closed-loop poles to -1; -2.

6. Design an observer with observer poles equal to -3; -4.

7. Combine the controller and the observer to obtain a feedback compensator with poles at -1; -2; -3; -4.

8. Suppose that this observer has the noisy sensor output as input. The observer equation then becomes

$$\dot{\hat{\boldsymbol{x}}}(t) = A\hat{\boldsymbol{x}} + \boldsymbol{b}u(t) - K_o\boldsymbol{c}'\hat{\boldsymbol{x}}(t) + K_o\big((y(t) + v(t)\big)$$

Does this observer lead to an acceptable controlled system? Compare your conclusion with the one obtained in part 2.

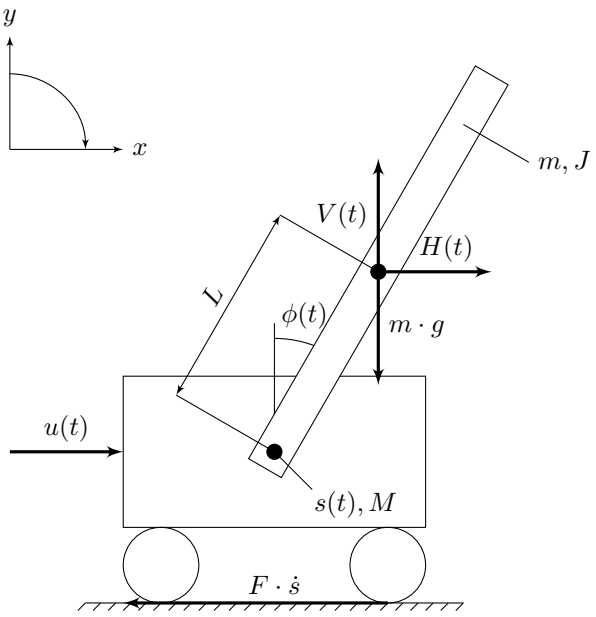


Figure 5.10: Visualization of the dynamics of the inverted pendulum on a cart

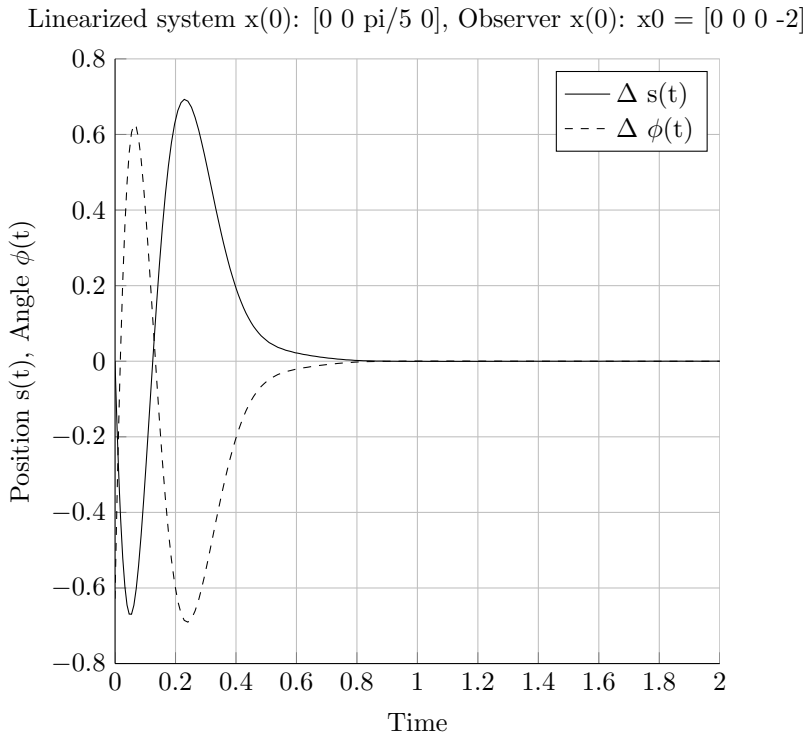


Figure 5.11: Difference between the observer and true dynamics for $x(0) = [0, 0, \frac{\pi}{5}, 0]$ and $\hat{x}(0) = [0, 0, 0, -2]$

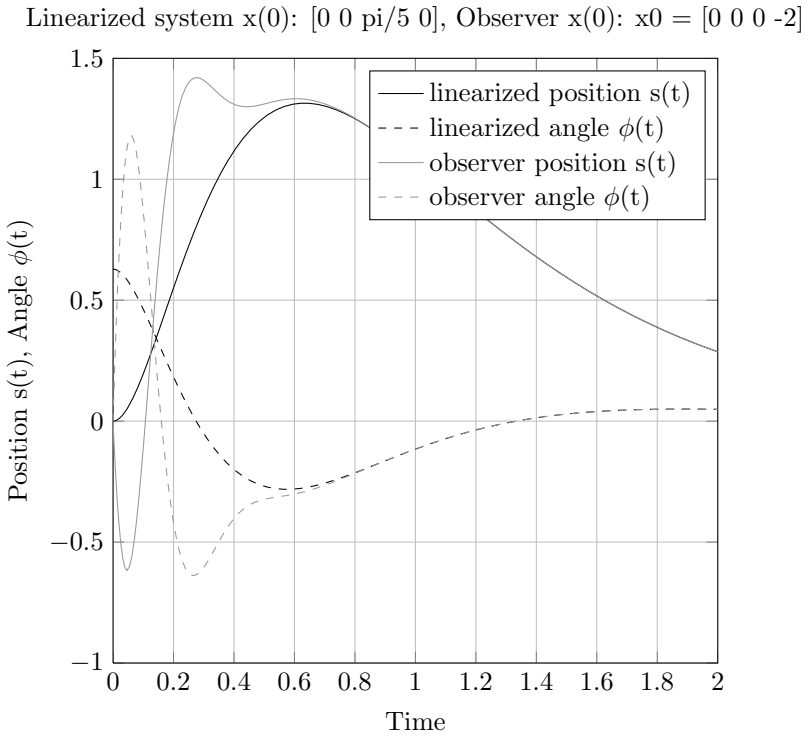


Figure 5.12: Illustration of the controlled linearized system and the observer for $x(0) = [0, 0, \frac{\pi}{5}, 0]$ and $\hat{x}(0) = [0, 0, 0, -2]$

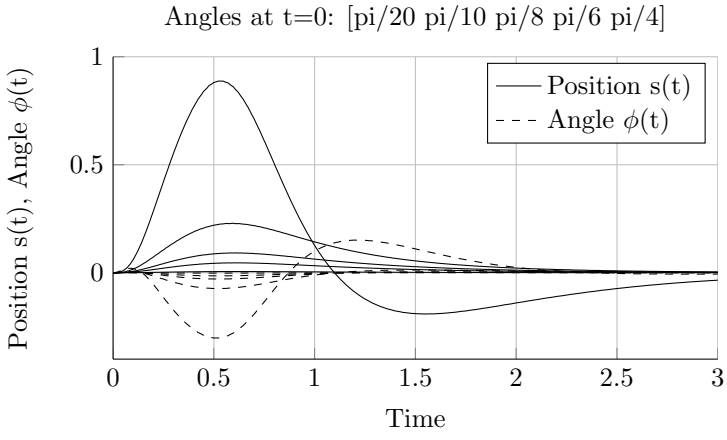


Figure 5.13: Difference between $s(t)$ and $\phi(t)$ between the controlled linearized and non-linear systems for $\phi(0) = [\frac{\pi}{20}, \frac{\pi}{10}, \frac{\pi}{8}, \frac{\pi}{6}, \frac{\pi}{4}]$

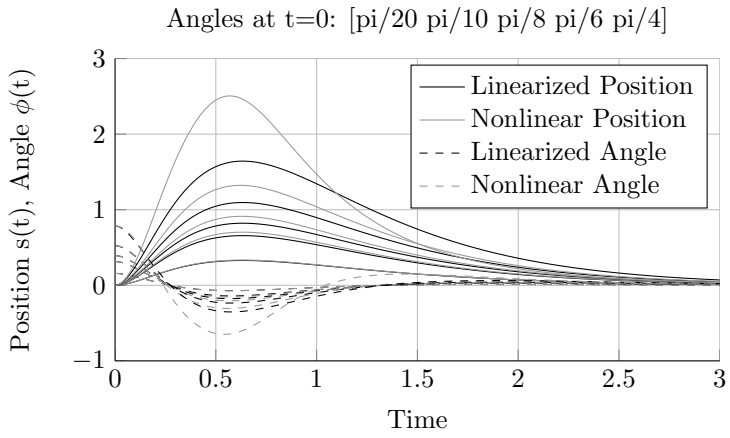


Figure 5.14: $s(t)$ and $\phi(t)$ of the controlled linearized and non-linear system, for $\phi(0) = [\frac{\pi}{20}, \frac{\pi}{10}, \frac{\pi}{8}, \frac{\pi}{6}, \frac{\pi}{4}]$

Chapter 6

Stability (2h)

We have loosely touched the concept of *stability* several times in this course. In this chapter we fill in some of the missing details. We first deal with general deterministic dynamical systems (not necessarily linear) and then specialize to the linear cases. Our discussion is mostly for the continuous time case (discrete time analogs exist yet are skipped here). We then briefly touch on stability of Markov chains by means of a Foster-Lyapunov function.

6.1 Equilibrium Points and Stability of Linear Dynamical Systems

6.2 Stability of General Deterministic Systems

Although we mainly considered linear dynamical systems, in this section we start from more general *nonlinear* time invariant dynamical systems, as introduced in Example 3.1.3:

$$\dot{\boldsymbol{x}}(t) = f\big(\boldsymbol{x}(t)\big), \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{6.1}$$

where $f(\cdot)$ is a Lipschitz continuous function, and $f(\mathbf{0}) = \mathbf{0}$.

The assumption that $f(\cdot)$ is a Lipschitz continuous function guarantees that the system (6.1) has a unique solution for any given initial condition. Furthermore, since we are interested in stability, we should have at least one $\bar{\boldsymbol{x}}$ such that $f\big(\bar{\boldsymbol{x}}\big) = \mathbf{0}$, and since we can always apply the change of coordinates $\tilde{\boldsymbol{x}} := \boldsymbol{x} - \bar{\boldsymbol{x}}$, we assume without loss of generality that $f(\mathbf{0}) = \mathbf{0}$.

We can now consider stability of the process $\{\boldsymbol{x}(t)\}$, or the system (6.1).

Definition 6.2.1. *We call the equilibrium point* $\boldsymbol{x} = \mathbf{0}$ *of* (6.1) locally (marginally)

stable *(in the sense of Lyapunov) if there exists a ball with radius* $r > 0$,

$$B_r := \{\boldsymbol{x} \mid \|\boldsymbol{x}\| < r\},$$

and a bound M *such that all solutions* $\boldsymbol{x}(t)$ *starting in that ball remain bounded, i.e,* $\|\boldsymbol{x}(t)\| \leq M$ *for all* $t \geq 0$.
If the bound M *holds for all* $\boldsymbol{x}_0 \in \mathbb{R}$, *then the origin is* globally (marginally) stable.

Definition 6.2.2. *The equilibrium point* $\boldsymbol{x} = \mathbf{0}$ *of* (6.1) *is said to be* locally asymptotically stable (LAS) *if it is locally stable and for all solutions starting in the ball* B_r*:* $\lim_{t\to\infty} \|\boldsymbol{x}(t)\| = 0$.
Similarly, the equilibrium point $\boldsymbol{x} = \mathbf{0}$ *of* (6.1) *is said to be* globally asymptotically stable (GAS) *if it is globally stable and for all solutions:* $\lim_{t\to\infty} \|\boldsymbol{x}(t)\| = \mathbf{0}$.

Definition 6.2.3. *We call the equilibrium point* $\boldsymbol{x} = \mathbf{0}$ *of* (6.1) unstable *if it is not locally stable.*

Named after the Russian mathematician Aleksandr Lyapunov (1857-1918), Lyapunov functions are important to stability theory and control theory. A similar concept appears in the theory of general state space Markov Chains, usually under the name Foster-Lyapunov functions. In this section we restrict ourselves to Lyapunov functions for analyzing stability of the nonlinear time invariant system (6.1). But before we do so, we first introduce the following terminology:

Definition 6.2.4. *A continuous function* $V : \mathbb{R}^n \to \mathbb{R}_+$ *is a* locally positive definite function *if for some* $\epsilon > 0$ *and some continuous, strictly increasing function* $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$*:*

$$V(\mathbf{0}) = 0 \qquad \textit{and} \qquad V(\boldsymbol{x}) \geq \alpha(\|\boldsymbol{x}\|), \qquad \forall \boldsymbol{x} \in B_\epsilon.$$

A continuous function $V : \mathbb{R}^n \to \mathbb{R}_+$ *is a* positive definite function *if it is a locally positive definite function and in addition* $\lim_{x\to\infty} \alpha(x) = \infty$.

Theorem 6.2.5. *For the system* (6.1), *let* $V : \mathbb{R}^n \to \mathbb{R}_+$ *be a function with derivative* $\dot{V}(\boldsymbol{x}) = \frac{d}{dt}V(\boldsymbol{x}) = \frac{dV}{dx}(\boldsymbol{x}) \cdot f(\boldsymbol{x})$ *along trajectories of* (6.1).

- *If* $V(\boldsymbol{x})$ *is locally positive definite and* $\dot{V}(x) \leq 0$ *locally in* $\boldsymbol{x}$, *then the origin of the system* (6.1) *is locally stable (in the sense of Lyapunov).*
- *If* $V(\boldsymbol{x})$ *is locally positive definite and* $-\dot{V}(\boldsymbol{x})$ *is locally positive definite, then the origin of the system* (6.1) *is locally asymptotically stable (in the sense of Lyapunov).*
- *If* $V(\boldsymbol{x})$ *is positive definite and* $-\dot{V}(\boldsymbol{x})$ *is positive definite, then the origin of the system* (6.1) *is globally asymptotically stable (in the sense of Lyapunov).*

Once for a given system a Lyapunov function $V(\cdot)$ has been found, one can conclude asymptotic stability. However, coming up with a Lyapunov function is difficult in general, though typically the energy of the system would be a good starting point. Fortunately, for an asymptotically stable system the search for a Lyapunov function is not futile, since the converse of Theorem 6.2.5 also exists: if an equilibrium point is (locally/globally/asymptotically) stable, then there exists a function $V(\cdot)$ satisfying the conditions of Theorem 6.2.5.

We now apply these results to study stability of autonomous linear dynamical systems, i.e., stability of the system

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t), \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0. \tag{6.2}$$

As a Lyapunov function candidate we take the quadratic form $V(\boldsymbol{x}) = \boldsymbol{x}'P\boldsymbol{x}$ where P is a positive definite matrix. Then we get

$$\dot{V}(\boldsymbol{x}) = \dot{\boldsymbol{x}}'P\boldsymbol{x} + \boldsymbol{x}'P\dot{\boldsymbol{x}} = \big(A\boldsymbol{x}\big)'P\boldsymbol{x} + \boldsymbol{x}'PA\boldsymbol{x} = \boldsymbol{x}'\big(A'P + PA\big)\boldsymbol{x} = -\boldsymbol{x}Q\boldsymbol{x}$$

So stability of linear time invariant systems can be studied by considering the Lyapunov equation

$$A'P + PA = -Q \tag{6.3}$$

For $Q \geq 0$ we can conclude that the origin of (6.2) is (globally) stable, and for $Q > 0$ we can conclude that the origin of (6.2) is (globally) asymptotically stable. Since the origin must be the only equilibrium point of the system, we typically say that the *system* (rather than just the equilibrium point) is asymptotically stable.

To find a quadratic Lyapunov function for the system (6.2), we might pick a $Q > 0$ and then try to solve (6.3) for a positive definite P. Therefore, (6.3) is also called the *Lyapunov equation*. If the system (6.2) is (globally) asymptotically stable, this approach will always work, as stated in the following:

Theorem 6.2.6. *Given the linear dynamical system* (6.2) *and a positive definite matrix Q, then there exists a (unique) positive definite matrix P satisfying* (6.3) *if and only if the system* (6.2) *is (globally) asymptotically stable.*

Proof. If P is a positive definite solution of (6.3), then $V(\boldsymbol{x}) = \boldsymbol{x}'P\boldsymbol{x}$ is a Lyapunov function for the system (6.2). From Theorem 6.2.5, global asymptotic stability follows.

For the "only if" part, let $Q > 0$ be given for (6.2) which is GAS. Take $P = \int_0^\infty e^{A't}Qe^{At}dt$, which is positive definite since $e^{A't}Qe^{At}$ for all t (Q is positive definite). Furthermore P is well defined, since the integral converges due to the fact that A is asymptotically stable.

Now we show that P satisfies (6.3):

$$\begin{aligned} A'P + PA &= \int_0^\infty \left(A'e^{A't}Qe^{At} + e^{A't}Qe^{At}A \right) dt \\ &= \int_0^\infty \frac{d}{dt}\left(e^{A't}Qe^{At} \right) dt \\ &= e^{A't}Qe^{At}\Big|_0^\infty = -Q. \end{aligned}$$

To prove uniqueness, let $\tilde{P}$ be an other solution to (6.3). Then

$$\begin{aligned} \tilde{P} &= -\int_0^\infty \frac{d}{dt}\left(e^{A't}\tilde{P}e^{At} \right) dt \\ &= -\int_0^\infty e^{A't}\left(A'\tilde{P} + \tilde{P}A \right) e^{At} dt \\ &= \int_0^\infty e^{A't}Qe^{At} dt = P \end{aligned}$$

□

Exercise 6.2.7. *Verify the identity* $\tilde{P} = -\int_0^\infty \frac{d}{dt}\left(e^{A't}\tilde{P}e^{At} \right) dt$.

Theorem 6.2.6 provides us with a way for checking asymptotic stability of a linear time invariant dynamical system. Efficient algorithms ($O(n^3)$) for determining P for given A and Q have been implemented in the Matlab function `lyap`. However, one can also give conditions for the stability based on the eigenvalues of A.

In order to state stability conditions in terms of the eigenvalues of the matrix A, we need to introduce the notion of semisimple eigenvalues. Those are eigenvalues for which the algebraic multiplicity (the multiplicity of the corresponding root of the characteristic polynomial) equals the geometric multiplicity (dimension of the associated eigenspace, i.e., the number of independent eigenvectors with that eigenvalue). Eigenvalues with multiplicity 1 are always semisimple. Consider the matrices

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Both matrices have 0 as a double eigenvalue. Nevertheless, this eigenvalue is semisimple in the first case and not in the second case. Note that the matrix A is diagonalizable if and only if all its eigenvalues are semisimple.

Now we can characterize stability of the system (6.2) as follows:

Theorem 6.2.8. *The system* (6.2) *is*

asymptotically stable *if and only if the eigenvalues of A have negative real part,*

stable *if and only if A has no eigenvalues with positive real part and all eigenvalues with zero real part are semisimple,*

unstable *if and only if A has either an eigenvalue with positive real part or a non-semisimple one with zero real part.*

Note that only the eigenvalues with zero real part need to be semisimple for stability. Eigenvalues with negative real part are allowed to be nonsemisimple. Therefore, for neither stability nor asymptotic stability diagonalizability of the matrix A is required.

6.3 Stability by Linearization

Consider the nonlinear time invariant system (6.1). Let $A = \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})\big|_{\boldsymbol{x}=\mathbf{0}}$ be the Jacobian matrix of f evaluated in $\mathbf{0}$, so that $\dot{\tilde{\boldsymbol{x}}} = A\tilde{\boldsymbol{x}}$ is the linearization of (6.1) around $\boldsymbol{x} = \mathbf{0}$.

Theorem 6.3.1. • *If the linearization is asymptotically stable, then the nonlinear system is locally asymptotically stable.*

• *If the linearization is unstable, then the nonlinear system is unstable.*

Proof. • Define $g(\boldsymbol{x}) = f(\boldsymbol{x}) - A\boldsymbol{x}$. Then $\|g(\boldsymbol{x})\| \leq K\|\boldsymbol{x}\|^2$. Furthermore, the nonlinear dynamics (6.1) can now be written as

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + g(\boldsymbol{x}) \tag{6.4}$$

Since the linearization is asymptotically stable, we have from Theorem 6.2.6 the existence of a positive definite P such that $A'P + PA = -I$. Now let us consider the Lyapunov function $V(\boldsymbol{x}) = \boldsymbol{x}'P\boldsymbol{x}$ and calculate its derivative along trajectories of (6.4). Then we obtain

$$\begin{aligned}
\dot{V}(\boldsymbol{x}) &= 2\boldsymbol{x}'P(A\boldsymbol{x} + g(\boldsymbol{x})) \\
&= \boldsymbol{x}'(A'P + PA)\boldsymbol{x} + 2\boldsymbol{x}'Pg(\boldsymbol{x}) \\
&\leq -\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{x}\|\|P\|\|g(\boldsymbol{x})\| \\
&\leq -\|\boldsymbol{x}\|^2 + 2K\|P\|\|\boldsymbol{x}\|^3 \\
&= -\|\boldsymbol{x}\|^2\big(1 - 2K\|P\|\|\boldsymbol{x}\|\big)
\end{aligned}$$

So for $\|\boldsymbol{x}\| \leq 1/(4K\|P\|)$ we have

$$\dot{V}(\boldsymbol{x}) \leq -\frac{1}{2}\|\boldsymbol{x}\|^2$$

which shows local asymptotic stability.

• Assume that the linearization is unstable, then a trajectory starting at $\boldsymbol{x}_0$ very small will move away from the origin. Thus the origin must be unstable.

□

6.4 Illustration: Stabilizing Control for Inherently Unstable Systems

We now show how to use the theory presented in the previous chapter and in this chapter to control an (A,B,C,D) system.

Consider the example of the Inverted Pendulum on a Cart, as presented in section 5.1. We repeat the resulting dynamics (5.9a), where we use the parameters $F = 1$, $M = 1$, $g = 10$, $L' = 1$:

$$\dot{\boldsymbol{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -10 & 0 & 10 & 0 \end{bmatrix} \boldsymbol{x}(t) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} u(t)$$

In order to stabilize this system, according to Theorem 6.2.8, we should makes sure that the eigenvalues of the closed-loop system are all in the negative half plane. So let us take as desired characteristic polynomial for the closed-loop system:

$$(s+2)^2(s+5)^2 = s^4 + 14s^3 + 69s^2 + 140s + 100. \tag{6.5}$$

Next, let $K_f = \begin{bmatrix} k_1 & k_2 & k_3 & k_4 \end{bmatrix}$. Applying the feedback $u(t) = -K_f \boldsymbol{x}(t)$ results in the closed-loop system

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -10 & 0 & 10 & 0 \end{bmatrix} \boldsymbol{x}(t) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} k_1 & k_2 & k_3 & k_4 \end{bmatrix} \boldsymbol{x}(t) \\ &= \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ -k_1 & -1-k_2 & -k_3 & -k_4 \\ 0 & 0 & 0 & 1 \\ -10 & 0 & 10 & 0 \end{bmatrix}}_{\tilde{A}} \boldsymbol{x}(t) \end{aligned}$$

Now we can determine the characteristic polynomial of the closed-loop matrix $\tilde{A}$:

$$s^4 + (1+k_2)s^3 + (k_1 - 10)s^2 - 10(1 + k_2 + k_4)s - 10(k_1 + k_3) \tag{6.6}$$

Next we can solve for k_1, k_2, k_3, and k_4 which make (6.5) and (6.6) equal. That is:

$$14 = 1 + k_2 \qquad 69 = k_1 - 10 \qquad 140 = -10(1 + k_2 + k_4) \qquad 100 = -10(k_1 + k_3)$$

resulting in

$$k_1 = 79 \qquad k_2 = 13 \qquad k_3 = -89 \qquad k_4 = -28.$$

From the above derivation we have a recipe the works in general for feedback design of linear systems. Given a desired characteristic polynomial, a matrix A and a vector $\boldsymbol{b}$ such that the pair $(A, \boldsymbol{b})$ is controllable, one can derive the feedback gain K_f simply by letting $K_f = \begin{bmatrix} k_1 & k_2 & \dots & k_n \end{bmatrix}$ and determine the characteristic polynomial for the resulting closed-loop system. One can prove that for single input systems the coefficients in the characteristic polynomial are all affine expressions in the k_i. By equating the coefficients of the desired characteristic polynomial and those of the characteristic polynomial of the resulting closed-loop system, one obtains a linear set of equations to solve for the k_i.

Unfortunately, this scheme does not work for the multi-input case. In the multi-input case the coefficients of the characteristic polynomial of the resulting closed-loop system will contain products of k_i. However, for the multi-input case there is a recipe that does work:

- Let a desired characteristic polynomial, a matrix A $(n \times n)$, and a matrix B $(n \times m)$ be given, where the pair (A, B) is controllable.
- Pick matrices N_1 $(m \times n)$ and N_2 $(m \times 1)$ such that the pair $(A - BN_1, BN_2)$ is controllable.
- Let $\bar{A} = A - BN_1$ and $\bar{\boldsymbol{b}} = BN_2$. The pair $(\bar{A}, \bar{\boldsymbol{b}})$ is controllable, so by letting $\bar{K} = \begin{bmatrix} \bar{k}_1 & \bar{k}_2 & \dots & \bar{k}_n \end{bmatrix}$ we can determine $\bar{K}$ such that the matrix $\bar{A} - \bar{\boldsymbol{b}}\bar{K}$ has the desired characteristic polynomial.
- By observing that $\bar{A} - \bar{\boldsymbol{b}}\bar{K} = (A - BN_1) - (BN_2)\bar{K} = A - B(N_1 + N_2\bar{K})$, we now take $K_f = N_1 + N_2\bar{K}$ and as a result the matrix $A - BK_f$ also has the desired characteristic polynomial.

A remaining question is: how to pick the matrices N_1 and N_2? By randomly picking the elements of these matrices, with probability 1 the resulting matrices are such that the pair $(\bar{A}, \bar{\boldsymbol{b}})$ is controllable. When working out an example by hand it will ease calculations if one puts many zeroes in. However, one has to assure that the pair $(\bar{A}, \bar{\boldsymbol{b}})$ is controllable.

Consider again the example of the Inverted Pendulum on a Cart, as presented in section 5.1, but this time we look at the observer design problem. That is, we consider the dynamics (5.9). Using the same parameters as before, we obtain:

$$\dot{\boldsymbol{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -10 & 0 & 10 & 0 \end{bmatrix} \boldsymbol{x}(t) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} u(t) \tag{6.7a}$$

$$\boldsymbol{y}(t) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \boldsymbol{x}(t) \tag{6.7b}$$

We can use the following Luenberger observer to reconstruct the state:

$$\dot{\hat{\boldsymbol{x}}}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -10 & 0 & 10 & 0 \end{bmatrix} \hat{\boldsymbol{x}}(t) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} u(t) - K_o\big(\hat{\boldsymbol{y}}(t) - \boldsymbol{y}(t)\big)$$
$$\hat{\boldsymbol{y}}(t) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \hat{\boldsymbol{x}}(t)$$

where K_o is still to be determined.

Typically one would like the state estimation error to converge quicker to zero than the system state. Therefore we take as a desired characteristic polynomial for the observer error dynamics:

$$(s+4)^2(s+10)^2 = s^4 + 28s^3 + 276s^2 + 1120s + 1600. \tag{6.8}$$

Due to duality between controllability and observability we have a recipe for determining K_o.

First we need to pick matrices N_1 (4×2) and N_2 (1×2), define $\bar{A} = A - N_1 C$ and $\bar{C} = N_2 C$ such that the pair (A, C) is observable.

If we take

$$N_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad N_2 = \begin{bmatrix} 0 & 1 \end{bmatrix},$$

then we obtain

$$\bar{A} = \begin{bmatrix} -1 & 1 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -10 & 0 & 10 & 0 \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \bar{C} \\ \bar{C}\bar{A} \\ \bar{C}\bar{A}^2 \\ \bar{C}\bar{A}^3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & -10 \end{bmatrix},$$

and the pair $(\bar{A}, \bar{C})$ is indeed observable.

Now we can take $\bar{K} = \begin{bmatrix} \bar{k}_1 & \bar{k}_2 & \bar{k}_4 & \bar{k}_4 \end{bmatrix}'$ and determine

$$\bar{A} - \bar{K}\bar{C} = \begin{bmatrix} -1 - \bar{k}_1 & 1 & -1 & 0 \\ -\bar{k}_2 & -1 & 0 & 0 \\ -\bar{k}_3 & 0 & 0 & 1 \\ -10 - \bar{k}_4 & 0 & 10 & 0 \end{bmatrix}$$

which has as characteristic polynomial

$$s^4 + (2+\bar{k}_1)s^3 + (\bar{k}_1 + \bar{k}_2 - \bar{k}_3 - 9)s^2 - (30 + 10\bar{k}_1 + 10\bar{k}_2 + \bar{k}_4)s - (20 + 10\bar{k}_1 + 10\bar{k}_2 + \bar{k}_4). \tag{6.9}$$

Next we can solve for $\bar{k}_1$, $\bar{k}_2$, $\bar{k}_3$, and $\bar{k}_4$ which make (6.8) and (6.9) equal. That is:

$$\begin{aligned} 28 &= 2 + \bar{k}_1 & 276 &= \bar{k}_1 + \bar{k}_2 - \bar{k}_3 - 9 \\ 1120 &= -(30 + 10\bar{k}_1 + 10\bar{k}_2 + \bar{k}_4) & 1600 &= -(20 + 10\bar{k}_1 + 10\bar{k}_2 + \bar{k}_4) \end{aligned}$$

resulting in

$$\bar{k}_1 = 26 \qquad \bar{k}_2 = -81 \qquad \bar{k}_3 = -340 \qquad \bar{k}_4 = -1070.$$

Now we can take $K_o = N_1 + \bar{K}N_2$, that is

$$K_o = \begin{bmatrix} 1 & 26 \\ 0 & -81 \\ 0 & -340 \\ 0 & -1070 \end{bmatrix}.$$

As a final step, we now obtain the following dynamic output feedback controller for stabilizing (6.7):

$$\begin{aligned} u(t) &= \begin{bmatrix} 79 & 13 & -89 & 28 \end{bmatrix} \hat{\boldsymbol{x}}(t) \\ \dot{\hat{\boldsymbol{x}}}(t) &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -10 & 0 & 10 & 0 \end{bmatrix} \hat{\boldsymbol{x}}(t) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} u(t) - \begin{bmatrix} 1 & 26 \\ 0 & -81 \\ 0 & -340 \\ 0 & -1070 \end{bmatrix} \big(\hat{\boldsymbol{y}}(t) - \boldsymbol{y}(t)\big), \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0, \\ \hat{\boldsymbol{y}}(t) &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \hat{\boldsymbol{x}}(t). \end{aligned}$$

6.5 Stability of Stochastic Systems

Definition 6.5.1. *A state i is said to be* recurrent *if, after starting from i, the process returns to i almost surely. In other words,*

$$\mathbf{P}[T_i < \infty \mid X_0 = i] = 1,$$

where $T_i = \inf\{n : X_n = i\}$ is the first return time to state i. Otherwise, the state i is said to be transient.

By definition, recurrence/transience is another class property: the states in the same class are either all recurrent or all transient. An irreducible Markov chain is said to be recurrent if its states are recurrent.

Is recurrence a sufficient condition for a Markov chain to have a stationary distribution? The long short answer is, no. (For the long story, consider a symmetric random walk on the integer line, where the system moves to ± 1 with probability $1/2$ each direction. This Markov chain is recurrent, but has no stationary distribution.) It turns out that we need to be more specific about the concept of recurrence.

Definition 6.5.2. *A recurrent state i is said to be* positive recurrent *if*

$$\mathbf{E}[T_i \mid X_0 = i] < \infty.$$

Otherwise, i is said to be null recurrent*.*

Positive/null recurrence is also a class property, so an irreducible Markov chain is either positive recurrent, or null recurrent, or transient. If Theorem **??** provides a sufficient condition for the existence and uniqueness of a stationary distribution, then the following theorem states a necessary and sufficient condition for the existence of a stationary distribution.

Theorem 6.5.3. *An irreducible Markov chain has a stationary distribution $\boldsymbol{\pi}$ if and only if it is positive recurrent.*

Together, theorems **??** and 6.5.3 imply that if an irreducible Markov chain is positive recurrent, then its stationary distribution $\boldsymbol{\pi}$ is unique and is also the unique limiting distribution.

...

Similarly, the definitions of positive recurrence, null recurrence and transience (Definitions 6.5.1 and 6.5.2) hold for CTMCs, and Theorem 6.5.3 applies: positive recurrence is a necessary and sufficient condition for an irreducible CTMC to have a unique stationary distribution.

TEXT PASTED UP TO HERE

The stochastic systems we have analyzed so far in this course are DTMCs and CTMCs, with a focus on structured cases. One can formulate several notions of stability for such Markov chains. One notion typically accepted notion for irreducible cases is that of having all states *positive recurrent*. This then implies that the stationary distribution, $\boldsymbol{\pi}$, exists and further,

$$\lim_{\ell \to \infty} \mathbb{P}\big(X(\ell) = i \;\big|\; X(0)\big) = \pi_i, \tag{6.10}$$

for any initial state $X(0)$, or distribution thereof (similarly for continuous time with t instead of ℓ). In the case of finite state space Markov chains (and assuming irreducibility), (6.10) always holds, yet if the state space is countably infinite, (6.10) is not guaranteed.

It is thus a matter of practical interest to see when a Markov Chain is positive recurrent (*stable*) and further characterize stability of structured Markov chains. Such characterizations appeared in Theorem 3.5.4 for birth-death processes as well as in Theorem 4.1.1 for level independent quasi-brith-death processes.

Up to this point, the only general tool introduced for establishing stability of Markov chains is the existence of a solution to $\boldsymbol{\pi}P = \boldsymbol{\pi}$ (DTMCs) or $\boldsymbol{\pi}Q = \mathbf{0}$ (CTMCs), such that $\boldsymbol{\pi}$ is a probability distribution. This may appear to be a promising method, as it not only verifies stability, but also gives further performance analysis input by actually yielding $\boldsymbol{\pi}$ (as in Theorem 3.5.4). Yet, the matter of the fact is, that in many cases

researchers and practionars have **not** been able to find an explicit solution, $\boldsymbol{\pi}$ and thus finding stability conditions has become a goal in its own right.

We now introduce a useful device for establishing stability of Markov Chains that follows similar lines to the Lyapunov function of Theorem 6.2.5 (that theorem is for deterministic dynamical systems):

Theorem 6.5.4. *Consider an irreducible DTMC, $\{X(\ell)\}$ on state space $\mathcal{S}$ with $|\mathcal{S}| = \infty$. Assume there exists a function,*

$$V : \mathcal{S} \to \mathbb{R}_+,$$

a subset $\mathcal{K} \subset \mathcal{S}$ with $|\mathcal{K}| < \infty$, and $\epsilon > 0$, such that,

$$(i) \qquad \mathbb{E}[V\big(X(1)\big) - V\big(X(0)\big)\big|X(0) = x] \qquad < \infty, \quad \forall x \in \mathcal{K},$$

and,

$$(ii) \qquad \mathbb{E}[V\big(X(1)\big) - V\big(X(0)\big)\big|X(0) = x] \qquad \leq -\epsilon, \quad \forall x \in \mathcal{S} \setminus \mathcal{K}.$$

The above theorem is called the *Foster-Lyapunov* condition for Markov chains. It first appeared in the 1950's in a paper by Foster. The resemblance of V in the theorem above, to the Lyapunov functions appearing in previous subsections is obvious: It is desired that outside of a compact set $\mathcal{K}$, the value of the Lyapunov function will have strictly negative drift. Note that it is required that ϵ be independent of x. Note that if the $-\epsilon$ term in the theorem is replaced by 0 then it is only guaranteed that the DTMC is recurrent (yet not necessarily positive-recurrent). There are many other variations and versions of the above theorem, here it is our goal simply to illustrate the general field.

While the above theorem is for DTMCs, it can be applied to CTMCs in certain cases. If for example the rates of the CTMC are not "too fast" nor "too slow", then we have the following:

Theorem 6.5.5. *Consider an irreducible CTMC with generator matrix Q. Assume that $|q_{i,i}| \in [a, b]$ for some, $a, b > 0$ and any $i \in \mathcal{S}$. Then if the associated embedded DTMC is positive recurrent then so is the CTMC.*

Other CTMCs that do not satisfy the preconditions of the above theorem can also be analyzed by means of the Foster-Lyapunov condition. We do not discuss the details further here.

Example 6.5.6. *Consider the M/M/1 queue as presented in Chapter 4. The embedded Markov chain on state space $\mathcal{S} = \{0, 1, 2, \ldots\}$, has transition probabilities,*

$$p_{0,1} = 1,$$

and,

$$p_{i,j} = \frac{\lambda}{\lambda + \mu}\mathbf{1}\{j = i + 1\} + \frac{\mu}{\lambda + \mu}\mathbf{1}\{j = i - 1\},$$

for $i = 1, 2, \ldots$ and any $j \in \mathcal{S}$. In this case, to show the DTMC is positive recurrent when $\lambda < \mu$, we can simply use $V(x) = x$ and $\mathcal{K} = \{0\}$.
With this choice, condition (i) of Theorem 6.5.4 holds trivially and condition (ii) follows since for any $x \in \{1, 2, \ldots\}$,

$$\mathbb{E}[V\big(X(1)\big) \mid X(0) = x] = \frac{\lambda}{\lambda+\mu}(x+1) + \frac{\mu}{\lambda+\mu}(x-1) = x + \frac{\lambda-\mu}{\lambda+\mu}.$$

And thus setting (for example),

$$\epsilon := \frac{\mu-\lambda}{\lambda+\mu} > 0,$$

we get,

$$\mathbb{E}[V\big(X(1)\big) - V\big(X(0)\big) \mid X(0) = x] = \frac{\lambda-\mu}{\lambda+\mu} = -\epsilon,$$

and satisfy the drift condition (ii).

Consider now level independent QBD processes. As stated in Theorem 4.1.1, such Markov chains are stable if,

$$\boldsymbol{\eta}\, A_1\, \mathbf{1} < \boldsymbol{\eta}\, A_{-1}\, \mathbf{1}, \tag{6.11}$$

where η is the stationary distribution of the phases independent of the level. I.e it is the stationary distribution associated with the transition probability matrix (alt. generator) $A = A_{-1} + A_0 + A_1$. As intuition for this result, assume that the level of the QBD is very high. In this case, it "acts as" a QBD process with no reflecting barrier and thus the phase process evolves independently of the level. Now $A_{-1}\mathbf{1}$ is a vector, whose i'th coordinate is the transition probability (or rate if CTMC) towards a downward level in case the current phase is i. Similarly $A_1\mathbf{1}$ is for upward phases. Thus, condition (6.11) makes sense because it implies that the average rate of downward transitions is greater than that of the average rate of upward transitions (this is intuitive because it relies on the assumption that the phase process has "mixed" to $\boldsymbol{\eta}$, but this will essentially happen for high levels).

Exercise 6.5.7. *(Non-trivial) Think of how to prove* (6.11) *using a Lyapunov function or a related method.*

6.6 Stability Criteria for QBD Processes (omitted)

This section is omitted from this version.

6.7 Congestion Network Stability via Fluid Limits (omitted)

This section is omitted from this version.

Bibliographic Remarks

Exercises

Consider the nonlinear dynamical system

$$y(t)^3 + \ddot{y}(t) + \sin y(t) = u(t)\big(u(t) - 1\big).$$

1. Determine the linearization around the equilibrium point $\bar{y} = 0$, $\bar{u} = 0$.
2. Determine a linear output feedback controller that locally stabilizes the equilibrium point $\bar{y} = 0$, $\bar{u} = 0$.
3. Determine the region of attraction of the system with this linear output feedback (or a non-empty subset of it). That is, determine a region around the origin where you can guarantee that if the system starts in that region it will stay in that region and furthermore that solutions will converge to the origin.

Chapter 7

Optimal Linear-Quadratic Control (3h)

This chapter gives a flavor of optimal control theory with specialization into optimal control of linear systems. It ends with an introduction of Model Predictive Control, a sub-optimal control method that has become popular in both theory in practice in the past 20 years.

7.1 Bellman's Optimality Principle

Consider a general nonlinear time-invariant control system in discrete time:

$$\boldsymbol{x}(\ell+1) = f\big(\boldsymbol{x}(\ell), \boldsymbol{u}(\ell)\big) \qquad\qquad \boldsymbol{x}(0) = \boldsymbol{x}_0. \tag{7.1}$$

We consider the problem of controlling this system optimally, that is, we want to find inputs $\boldsymbol{u}(0)$, $\boldsymbol{u}(1)$, $\ldots$, $\boldsymbol{u}(N-1)$ such that the following objective is minimized:

$$J = g_N\big(\boldsymbol{x}(N)\big) + \sum_{k=0}^{N-1} g\big(\boldsymbol{x}(k), \boldsymbol{u}(k)\big). \tag{7.2}$$

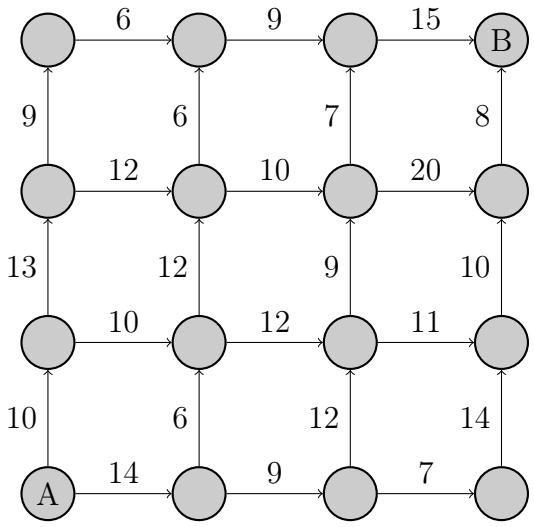
Richard Bellman (1920-1984) made a rather obvious observation, which became known as the Principle of Optimality:

Principle of Optimality [Bellman] *Let $u^* = \{\boldsymbol{u}(0),\ \boldsymbol{u}(1),\ \ldots,\ \boldsymbol{u}(N-1)\}$ be an optimal policy for minimizing* (7.2) *subject to the dynamics* (7.1)*, and assume that when using u^*, a given state $\boldsymbol{x}(i)$ occurs at time i. Consider now the subproblem where we are at $\boldsymbol{x}(i)$ at time i and wish to minimize the "cost-to-go" from time i to time N:*

$$g_N\big(\boldsymbol{x}(N)\big) + \sum_{k=i}^{N-1} g\big(\boldsymbol{x}(k), \boldsymbol{u}(k)\big).$$

Then the truncated policy $\{\boldsymbol{u}(i),\ \boldsymbol{u}(i+1),\ \ldots,\ \boldsymbol{u}(N-1)\}$ is optimal for this subproblem.

h



The intuitive idea behind this principle is very simple. If the truncated policy were not optimal, then we could reduce the cost for the original problem by switching to the better alternative once we reach $\boldsymbol{x}(i)$.

Consider for instance the problem of driving by car from Melbourne to Perth. The fastest route passes through Adelaide. Now the principle of optimality states the obvious fact that the Adelaide-Perth part of the route is also the fastest route from Adelaide to Perth.

However, this simple observations has far reaching consequences. Since in order to solve the optimal control problem (7.1), (7.2), we can first determine the optimal control for the subproblem starting at $i = N - 1$. Next, we can determine the optimal control for the subproblem starting at $i = N - 2$, etc., and proceed backwards until we determine the optimal control for the original problem.

This approach for solving the optimization problem (7.2), subject to (7.1), is also known as *Dynamic Programming (DP)*.

Exercise 7.1.1. *Consider the grid, shown in the figure below. Find the shortest path from A to B (moving only east or north) using Dynamic Programming.*

Exercise 7.1.2. *Consider again the grid of Exercise 7.1.1. This time, find the longest path from A to B (moving only east or north) using Dynamic Programming.*

From the above exercises the DP algorithm should have become clear:

Theorem 7.1.3. *For every initial state $\boldsymbol{x}_0$, the optimal cost $J^*(\boldsymbol{x}_0)$ resulting from minimizing* (7.2) *subject to* (7.1) *is equal to $J_0(\boldsymbol{x}_0)$, given by the last step of the following algorithm, which proceeds backwards in the time from period $N - 1$ to period* 0*:*

$$J_N\big(\boldsymbol{x}(N)\big) = g_N\big(\boldsymbol{x}(N)\big)$$

$$J_k\big(\boldsymbol{x}(k)\big) = \min_{\boldsymbol{u}(k)} g\big(\boldsymbol{x}(k), \boldsymbol{u}(k)\big) + J_{k+1}\Big(f\big(\boldsymbol{x}(k), \boldsymbol{u}(k)\big)\Big) \qquad k = 0, 1, \ldots, N-1 \qquad (7.3)$$

Furthermore, if $\boldsymbol{u}^*\big(\boldsymbol{x}(k)\big)$ *minimizes the right hand side of* (7.3) *for each* $\boldsymbol{x}(k)$ *and* k*, the policy* $u^* = \{\boldsymbol{u}^*\big(\boldsymbol{x}(0)\big), \boldsymbol{u}^*\big(\boldsymbol{x}(1)\big), \ldots, \boldsymbol{u}^*\big(\boldsymbol{x}(N-1)\big)\}$ *is optimal.*

The function $J_k\big(\boldsymbol{x}(k)\big)$ can be interpreted as the costs associated with the subproblem of starting in $\boldsymbol{x}(k)$ at time k, and therefore is often referred to as the *cost-to-go function* at time k.

7.2 The Linear Quadratic Regulator

In the previous section we considered optimal control in a more general setting. Now, we restrict ourselves to linear systems:

$$\boldsymbol{x}(\ell+1) = A\boldsymbol{x}(\ell) + B\boldsymbol{u}(\ell) \qquad\qquad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

and quadratic costs

$$J = \boldsymbol{x}(N)'Q_f\boldsymbol{x} + \sum_{k=0}^{N-1} \boldsymbol{x}(k)'Q\boldsymbol{x}(k) + \boldsymbol{u}(k)'R\boldsymbol{u}(k).$$

Engineers typically take Q and R to be diagonal matrices, where $Q_{ii} = 1/u_{i,\max}$ and $R_{ii} = 1/x_{i,\max}$. Here $u_{i,\max}$ and $x_{i,\max}$ denote the maximally allowed value for u_i and x_i respectively. Not that this choice for Q and R guarantees that these constraints will not be violated, but it is a rule of thumb, commonly used in practice. How to explicitly deal with constraints is subject of the next section.

For solving the optimal control problem we need to make the following assumptions:

- The pair (A, B) is controllable.
- $Q' = Q \geq 0$, i.e. positive semi-definite.
- $R' = R > 0$, i.e. positive definite.
- The pair $(A, \bar{C})$ is observable, where $\bar{C}$ solves $Q = \bar{C}'\bar{C}$

These conditions guarantee that in the remainder of this section inverses exist, limits exist, and that the resulting controller stabilizes the system at $\boldsymbol{x} = \boldsymbol{0}$.

We can solve this optimal control problem by means of dynamic programming.

As a first step, we obtain

$$J_N\big(\boldsymbol{x}(N)\big) = \boldsymbol{x}(N)'Q_f\boldsymbol{x}(N).$$

Writing (7.3) for $k = N-1$ results in

$$\begin{aligned}
J_k\big(\boldsymbol{x}(k)\big) &= \min_{\boldsymbol{u}(N-1)} \boldsymbol{x}(N-1)'Q\boldsymbol{x}(N-1) + \boldsymbol{u}(N-1)'R\boldsymbol{u}(N-1) + J_N\big(A\boldsymbol{x}(N-1) + B\boldsymbol{u}(N-1)\big) \\
&= \min_{\boldsymbol{u}(N-1)} \boldsymbol{x}(N-1)'Q\boldsymbol{x}(N-1) + \boldsymbol{u}(N-1)'R\boldsymbol{u}(N-1) + \\
&\qquad + \big(A\boldsymbol{x}(N-1) + B\boldsymbol{u}(N-1)\big)'Q_f\big(A\boldsymbol{x}(N-1) + B\boldsymbol{u}(N-1)\big) \\
&= \min_{\boldsymbol{u}(N-1)} \boldsymbol{u}(N-1)'\big(R + B'Q_fB\big)\boldsymbol{u}(N-1) + 2\boldsymbol{x}(N-1)'A'Q_fB\boldsymbol{u} + \boldsymbol{x}'\big(Q + A'Q_fA\big)\boldsymbol{x}
\end{aligned}$$

Now we differentiate the right hand term w.r.t. $\boldsymbol{u}(N-1)$, set the derivative to zero, and obtain

$$\big(R + B'Q_fB\big)\boldsymbol{u}(N-1) = -B'Q_fA\boldsymbol{x}(N-1)$$

and therefore

$$\boldsymbol{u}^*(N-1) = -\big(R + B'Q_fB\big)^{-1}B'Q_fA\boldsymbol{x}(N-1)$$

By substituting this back into the equation for $J_k\big(\boldsymbol{x}(k)\big)$ we obtain:

$$\begin{aligned}
J_k\big(\boldsymbol{x}(k)\big) &= \boldsymbol{x}(N-1)'A'Q_fB\big(R + B'Q_fB\big)^{-1}B'Q_fA\boldsymbol{x}(N-1) - \\
&\quad - 2\boldsymbol{x}(N-1)'A'Q_fB\big(R + B'Q_fB\big)^{-1}B'Q_fA\boldsymbol{x}(N-1) + \boldsymbol{x}'\big(Q + A'Q_fA\big)\boldsymbol{x} \\
&= \boldsymbol{x}(N-1)'\Big(\underbrace{A'Q_fA - \big(A'Q_fB\big)\big(R + B'Q_fB\big)^{-1}\big(B'Q_fA\big) + Q}_{P(N-1)}\Big)\boldsymbol{x}(N-1)
\end{aligned}$$

Proceeding similarly for the other steps we obtain

$$\boldsymbol{u}^*(k) = -\big(R + B'P(k)B\big)^{-1}B'P(k)A\boldsymbol{x}(k)$$

where $P(k)$ is given by the backward recursion

$$P(k-1) = A'P(k)A - \big(A'P(k)B\big)\big(R + B'P(k)B\big)^{-1}\big(B'P(k)A\big) + Q \quad P(N) = Q_f.$$

So for the finite time horizon optimal control problem, the resulting optimal controller is a linear (time-varying) feedback controller.

Now if we let $N \to \infty$, that is, we consider the infinite horizon optimal control problem, then it can be shown that $P(k)$ converges to a fixed matrix P which is the unique positive definite solution of the *discrete time algebraic Riccati equation*:

$$P = A'PA - (A'PB)(R + B'PB)^{-1}(B'PA) + Q$$

The corresponding optimal input now also becomes a static state feedback

$$\boldsymbol{u}^*(\ell) = -\underbrace{(R + B'PB)^{-1}B'PA}_{K_f}\boldsymbol{x}(\ell).$$

Exercise 7.2.1. *Solve the infinite horizon optimal control problem in discrete time for the scalar linear system*

$$x(\ell+1) = x(\ell) + 2u(\ell) \qquad\qquad x(0) = x_0$$

and cost function

$$J = \sum_{k=0}^{\infty} 2x(k)^2 + 6u(k)^2.$$

That is, determine the optimal steady state feedback $u(\ell) = -k_f x(\ell)$, as well as the cost-to-go $px(\ell)^2$.

Remark 7.2.2. *Similar results can be derived for continuous time. So consider the system*

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \qquad\qquad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

and cost function

$$J = \int_0^T \boldsymbol{x}(s)'Q\boldsymbol{x}(s) + \boldsymbol{u}(s)'R\boldsymbol{u}(s)ds$$

with the same assumptions on A, B, Q and R. Then the optimal input is given by the linear time-varying state feedback

$$\boldsymbol{u}^*(t) = -R^{-1}B'P(t)\boldsymbol{x}(t)$$

where $P(t)$ is given by the following differential equation (in backward time):

$$\dot{P}(t) = A'P(t) + P(t)A - P(t)BR^{-1}B'P(t) + Q \qquad\qquad P(T) = Q_f.$$

Again, if we let $T \to \infty$, we get a steady state feedback controller

$$\boldsymbol{u}^*(t) = -\underbrace{R^{-1}B'P}_{K_f}\boldsymbol{x}(t)$$

where P is the unique positive definite solution to the continuous time algebraic Riccati equation*:*

$$A'P + PA - PBR^{-1}B'P + Q = 0.$$

7.3 Riccati Equations

What the exercises at the end of the previous section show, is that in general MPC does not yield a stabilizing controller. However, by taking the prediction horizon p sufficiently large, and by adding terminal constraints, a stabilizing controller can be obtained.

In this section we show that by adding a terminal constraint, a specific choice for the terminal costs, and carefully selecting the prediction horizon (given the current state), the resulting MPC controller solves the infinite horizon LQR-problem with constraints. This guarantees that the resulting controller stabilizes the system. Furthermore, we obtain the optimal controller for the infinite horizon LQR-problem with constraints, where we only have to solve a finite optimization problem.

To be precise, we consider the system

$$\boldsymbol{x}(\ell+1) = A\boldsymbol{x}(\ell) + B\boldsymbol{u}(\ell) \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{7.4}$$

and want to minimize te objective

$$J = \sum_{k=0}^{\infty} \boldsymbol{x}(k)'Q\boldsymbol{x}(k) + \boldsymbol{u}(k)'R\boldsymbol{u}(k)$$

subject to the constraints

$$E\boldsymbol{x} + F\boldsymbol{u} \leq \boldsymbol{g}. \tag{7.5}$$

In addition to the assumptions we made in section 7.2, we require that the elements of the vector $\boldsymbol{g}$ satisfy $g_i > 0$. Furthermore we assume that the initial condition $\boldsymbol{x}_0$ is in the set of states for which the optimal control problem is feasible. In case the matrix A is such that the system $\boldsymbol{x}(\ell+1) = A\boldsymbol{x}(\ell)$ is marginally stable, we can take for this set $\mathbb{R}^n$.

In the remainder of this section we outline how to solve this problem. Since the problem is feasible, i.e., a solution exists which results in finite costs, we also know, due to the fact that $g_i > 0$, that after a finite amount of time the constraints (7.5) will not be active anymore. From Bellmans Optimality Principle we know that from then on, the solution to the unconstrained infinite horizon optimal control problem as presented in section 7.2 is followed.

So the first step is to solve the Discrete Time Algebraic Riccati equation

$$P = A'PA - (A'PB)(R + B'PB)^{-1}(B'PA) + Q.$$

From section 7.2 we know that the associated optimal input is given by $\boldsymbol{u}(\ell) = -K_f\boldsymbol{x}(\ell)$ where $K_f = R^{-1}B'P$. Furthermore, the cost to go is given by $x(\ell)'Px(\ell)$

The second step is to determine the *maximally output admissible set*. That is, the largest set Z of $\boldsymbol{x}$ satisfying $(E - FK_f)\boldsymbol{x} \leq \boldsymbol{g}$ such that $(A - BK_f)\boldsymbol{x}$ is also contained in that set.

As a third step, we consider the MPC problem to minimize

$$J = \boldsymbol{x}(\ell+p)'P\boldsymbol{x}(\ell+p) + \sum_{k=0}^{p-1} \boldsymbol{x}(\ell+k|\ell)'Q\boldsymbol{x}(\ell+k|\ell) + \boldsymbol{u}(\ell+k|\ell)'R\boldsymbol{u}(\ell+k|\ell).$$

subject to the dynamics (7.4), the constraints (7.5), and the terminal constraints $\boldsymbol{x}(\ell + p) \in Z$. Here we should take p large enough such that this MPC problem is feasible.

7.4 Model-based Predictive Control (omitted)

This section is omitted from this version.

Bibliographic Remarks

Exercises

Chapter 8

Fluid Buffer Models

8.1 Deterministic Fluid Buffer Models for Switching Servers

<<< Put here type of stuff Erjen has been doing on switching systems >>>.

8.2 Anick, Mitra and Sondhi's Model

In 1982 Anick, Mitra and Sondhi developed a model for a 'Data-Handling System with Multiple Sources'. To my knowledge this was the first paper that proposed a model of the class which have become known as *Stochastic Fluid Models*. We shall start by describing Anick, Mitra and Sondhi's model and then move on to a discussion of stochastic fluid models in general.

Consider a switch which has input lines from N sources, which independently and asynchronously alternate between 'on' and 'off' states. The 'on' periods are exponentially distributed with a parameter which we can take to be one, and the off periods are exponentially distributed with a parameter λ. When a source is 'on', it transmits a continuous flow of data at a rate of one per unit time, so when r sources are 'on' data arrives to the switch at a rate of r per unit time. The switch can process data at a rate of c per unit time, so the net rate of data arriving at the switch is $r - c$, which might be positive, zero, or negative. The net amount of data buffered at the switch changes by this amount per unit time, unless $r - c$ is negative and the level of buffered data is zero, in which case the level remains at zero. We assume that $c \in (1, N)$ so that there are some states where the buffer fills and some others where it empties and, to avoid complications (which can be dealt with), that c is not an integer.

Anick, Mitra and Sondhi didn't do this, but we shall denote the amount of data by $M(t)$, and the number of transmitting sources by $\varphi(t)$. They were interested in the stationary

distribution of the amount of data buffered at the switch,

$$F_i(x) \equiv P(M(t) \leq x, \varphi(t) = i) \tag{8.1}$$

under stationary conditions. Letting $\boldsymbol{F}(x)$ be a column with entries $F_i(x)$, Anick, Mitra and Sondhi showed that $\boldsymbol{F}(x)$ satisfies the equation

$$D\frac{d\boldsymbol{F}(x)}{dx} = M\boldsymbol{F}(x) \tag{8.2}$$

where $D = \text{diag}(-c, 1-c, \ldots, N-c)$ and

$$M = \begin{bmatrix} -N\lambda & 1 & & & & & & \\ N\lambda & -[(N-1)\lambda+1] & 2 & & & & & \\ & (N-1)\lambda & -[(N-2)\lambda+2] & 3 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & & 2\lambda & -[\lambda+N-1] & N & \\ & & & & & \lambda & -N & \end{bmatrix}. \tag{8.3}$$

Note that the matrix M is the transpose of the generator matrix of the continuous-time birth and death process that governs the number of transmitting sources. A modern stochastic modeller would consider $\boldsymbol{F}(x)$ to be a row and write

$$\frac{d\boldsymbol{F}(x)}{dx}D = \boldsymbol{F}(x)T \tag{8.4}$$

where $T = M'$.

The assumption that c is not an integer means that D is non-singular, and we can write (8.5) as

$$\frac{d\boldsymbol{F}(x)}{dx} = \boldsymbol{F}(x)TD^{-1}. \tag{8.5}$$

This is a linear first-order differential equation of the type that you have been dealing with throughout this subject. However, we do not, at this stage, know the initial conditions the derivation of which depends on some more in-depth analysis. A further feature is that TD^{-1} has some eigenvalues with positive real parts, so the system is not stable in the sense that we have talked about in this subject. This had rendered attempts to solve (8.5) numerically unsuccessful. Anick, Mitra and Sondhi used arguments about the nature of the eigenvalues and eigenvectors of TD^{-1} to define the initial conditions and came up with a nice solution. The paper is worth a look if you are interested.

8.3 A General Stochastic Fluid Buffer Model

Our interest here is in a general class of models, the stochastic fluid models, that have a similar form to Anick, Mitra and Sondhi's model. They have been discussed by a number

of authors. Notable amongst these are Rogers [**?**], Asmussen [**?**], Ramaswami [**?**], Da Silva Soares and Latouche [**?**] and Bean,O'Reilly and Taylor [**?**].

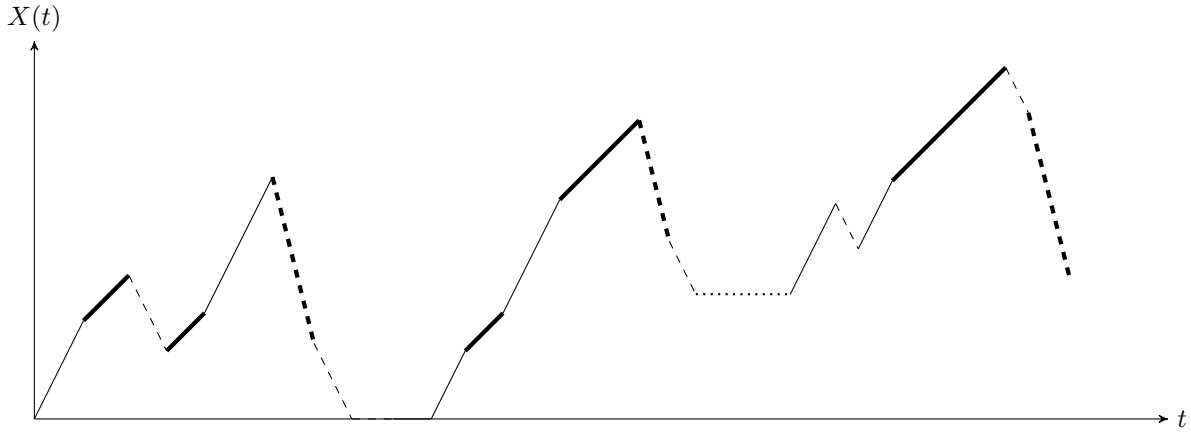
A stochastic fluid model is a two-dimensional continuous-time process $\{(M(t), \varphi(t)), t \in \mathbb{R}^+\}$ where

- $M(t) \in \mathbb{R}^+$ is the *level*, that is the content of the fluid buffer
- $\varphi(t) \in \mathcal{S}$ is the *phase*, that is the state of the underlying Markov process, with its state space S assumed to be finite and its generator T assumed to be irreducible.

When $\varphi(t) = i$,

$$\frac{dM(t)}{dt} = \begin{cases} r_i & \text{if } X(t) > 0 \\ \max(0, r_i) & \text{if } X(t) = 0 \end{cases}$$

In the above diagram, black and red correspond to controlling phases with $r_i > 0$, blue



and green correspond to controlling phases with $r_i < 0$, while purple corresponds to a controlling phase with $r_i = 0$.

Following a similar argument to that used by Anick, Mitra and Sondhi, we can show that

$$\frac{\partial}{\partial t}\pi_j(x,t) + \frac{\partial}{\partial x}\pi_j(x,t)\, r_j = \sum_{i \in \mathcal{S}} \pi_i(x,t) T_{ij}$$

$$\pi_i(x,t) = \frac{\partial}{\partial x} P[M(t) \le x, \varphi(t) = i],$$

is the density (in x) of the fluid level and state at time t. Provided it exists, the stationary density $\pi_i(x) = \lim_{t\to\infty} \pi_i(x,t)$ satisfies the equation

$$\frac{d}{dx}\pi_j(x)\, r_j = \sum_{i \in \mathcal{S}} \pi_i(x) T_{ij}$$

which we can write in matrix form as

$$\frac{d}{dx}\boldsymbol{\pi}(x)R = \boldsymbol{\pi}(x)T, \tag{8.6}$$

with $R = \text{diag}(r_i)$.

Let $\boldsymbol{\xi}$ be the stationary distribution of the continuous-time Markov chain with generator T, which satisfies

$$\begin{array}{rcl} \boldsymbol{\xi} T & = & \mathbf{0} \\ \boldsymbol{\xi}\,\mathbf{1} & = & 1. \end{array}.$$

Then the stationary density vector $\boldsymbol{\pi}(x)$ exists if and only if

$$\boldsymbol{\xi}\,\boldsymbol{r} < 0,$$

where $\boldsymbol{r} = (r_i)$. we can think of $\boldsymbol{\xi}\,\boldsymbol{r}$ as the *mean stationary drift*.

As with Anick, Mitra and Sondhi, if there are any $r_i = 0$, we can define an equivalent model in which all r_i are non-zero. Thus, we can assume without loss of generality that $r_i \neq 0$ for all $i \in \mathcal{S}$. R is then invertible.

Many authors have seen the problem mainly in terms of solving the ODE (8.6), with suitable boundary conditions. However the ODE approach frequently leads to unstable numerical procedures arising from the positive eigenvalues, and it would be good to have another approach. Rogers [**?**] and Asmussen [**?**] used Wiener-Hopf factorization. Ramaswami [**?**], who was a former PhD student of the founder of the field of matrix-analytic methods, Marcel Neuts, suggested a matrix-analytic approach, which we shall present here.

First, we can show that we do not lose generality in restricting the analysis to fluid queues with net input rates equal to $r_i = +1$ or $r_i = -1$ only and we assume this here. Partition $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ where $r_i = 1$ for $i \in \mathcal{S}_1$ and $r_i = -1$ for $i \in \mathcal{S}_2$. Corresponding to this we can partition the generator T into four blocks so that

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

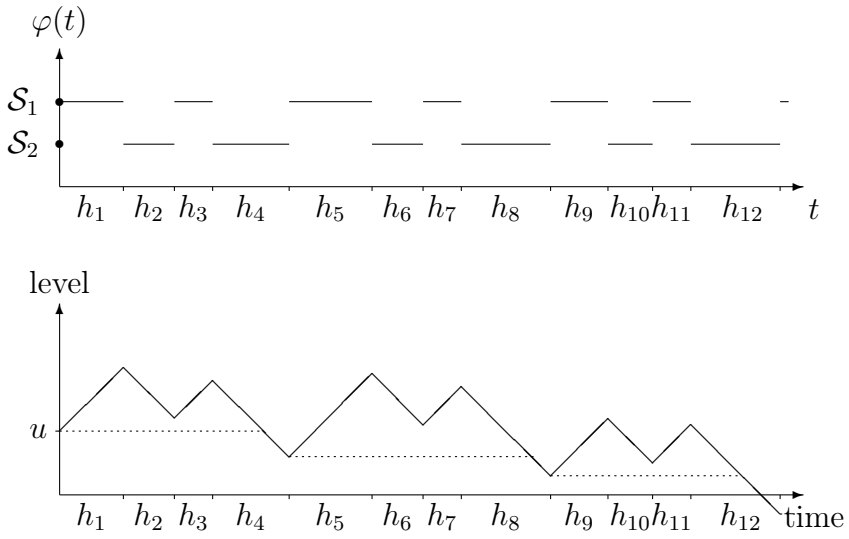
and the stationary density $\boldsymbol{\pi}(x)$ into two blocks so that $\boldsymbol{\pi}(x) = [\boldsymbol{\pi}_1(x), \boldsymbol{\pi}_2(x)]$. The sample paths of this process look like

When the state is in $\mathcal{S}_1$ the fluid level will immediately move away from level 0, so the stationary probability that the fluid level is equal to zero is

$$\lim_{t\to\infty} P[M(t) = 0, \varphi(t) = i] = 0,$$

when $i \in \mathcal{S}_1$. However, because the fluid level can stay at zero while it is in $\mathcal{S}_2$, there is a positive stationary probability mass associated with the states where $M(t) = 0$ and $i \in \mathcal{S}_2$. So we write

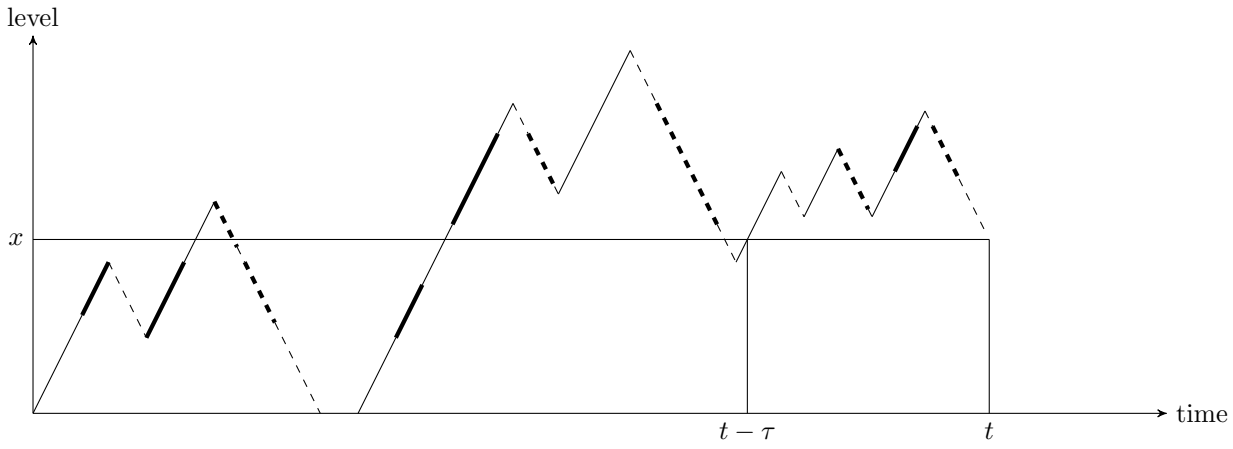
$$\lim_{t\to\infty} P[M(t) = 0, \varphi(t) = i] = \beta_i, \quad i \in \mathcal{S}_2.$$



and

$$\lim_{t\to\infty} P[M(t) = 0, \varphi(t)] = [0, \boldsymbol{\beta}_2].$$

Taking $M(0) = 0$, we condition on the last visit to level x so that we can show that $\pi_j(x,t)$



satisfies a Markov renewal equation of the form

$$\pi_j(x,t) = \sum_{i\in\mathcal{S}_1} \int_0^t \pi_i(x, t-\tau)\phi_{ij}(d\tau)$$

where $\phi_{ij}(d\tau)$ is the probability that the return to the same level occurs in the time interval $(\tau, \tau + d\tau)$ *and* in phase $j \in \mathcal{S}_2$. Using a theorem about Markov renewal

equations, we can write

$$\pi_j(x,t) = \sum_{i\in\mathcal{S}_1}\int_0^t \pi_i(x,t-\tau)\phi_{ij}(d\tau)$$
$$\underset{t\to\infty}{\longrightarrow} \sum_{i\in\mathcal{S}_1}\pi_i(x)$$

and so

$$\int_0^\infty \phi_{ij}(d\tau)\pi_j(x) = \sum_{i\in\mathcal{S}_1}\pi_i(x)\Psi_{ij} \tag{8.7}$$

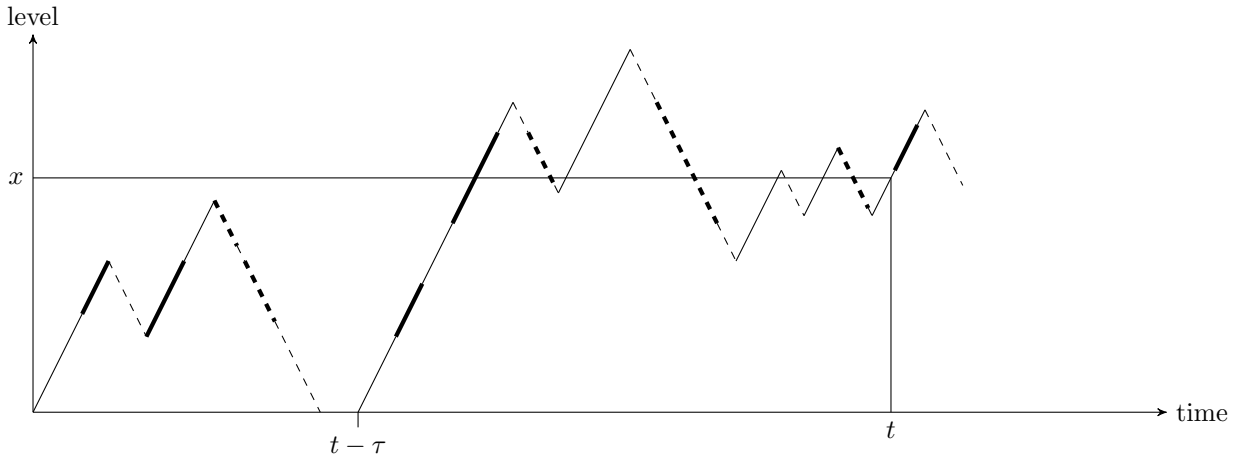
where $\Psi_{ij} = \int_0^\infty \phi_{ij}(d\tau)$ is the probability of returning to the same level in finite time, in j. In matrix notation this tells us that

$$\boldsymbol{\pi}_2(x) = \boldsymbol{\pi}_1(x)\Psi.$$

So we now have

$$\begin{aligned}[\boldsymbol{\pi}_1(x) \quad \boldsymbol{\pi}_2(x)] &= [\boldsymbol{\pi}_1(x) \quad \boldsymbol{\pi}_1(x)\Psi] \\ &= \boldsymbol{\pi}_1(x)[I \quad \Psi]\end{aligned}$$

We still need to compute $\boldsymbol{\pi}_1(x)$. To get this, condition on the last time that the process was in level 0, so that



$$\pi_j(x,t) = \sum_{i\in\mathcal{S}_2}\sum_{k\in\mathcal{S}_1}\int_0^t \beta_i(t-\tau)T_{ik}\gamma_{kj}(x,\tau)d\tau$$

where $\beta_i(t)$ is the probability of being in $(0,i)$ at time t, and $\gamma_{kj}(x,\tau)$ is the probability (that is expected value of the indicator function) of crossing level x at time interval τ in

phase j avoiding level 0, starting from level 0 in phase k. Hence

$$\pi_j(x,t) = \sum_{i\in\mathcal{S}_2}\sum_{k\in\mathcal{S}_1}\int_0^t \beta_i(t-\tau)T_{ik}\gamma_{kj}(x,\tau)d\tau \tag{8.8}$$

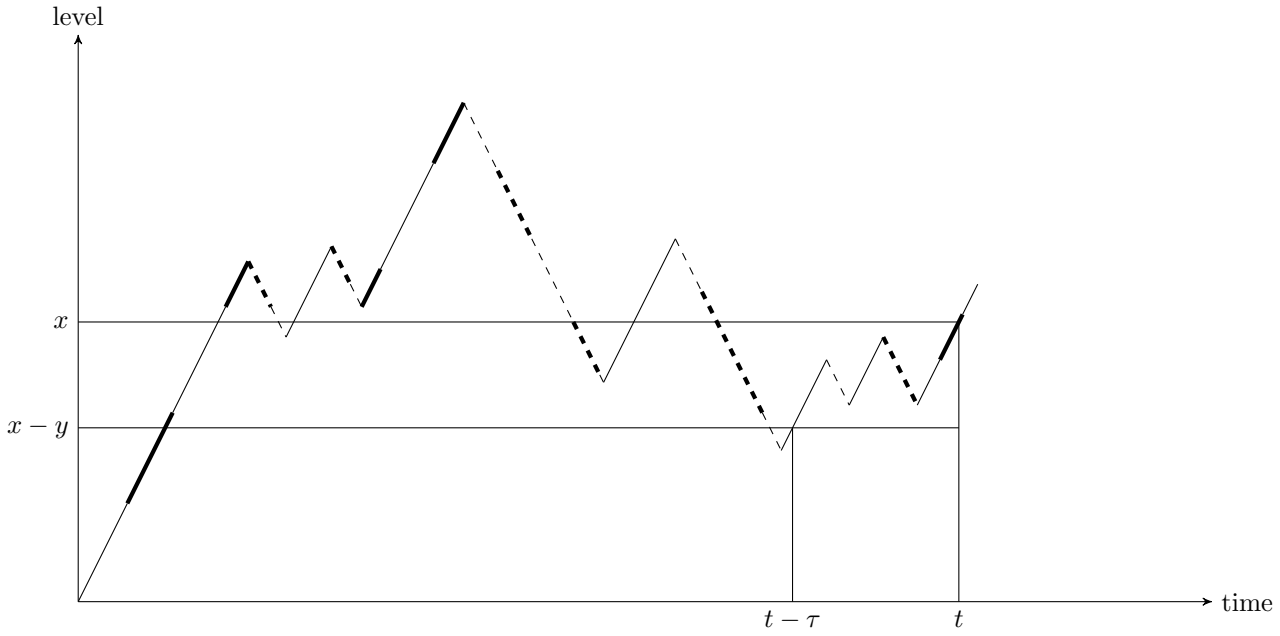
Taking limits as $t\to\infty$, we see that

$$\begin{aligned}\pi_j(x) &= \lim_{t\to\infty}\sum_{i\in\mathcal{S}_2}\sum_{k\in\mathcal{S}_1}\beta_i T_{ik}\int_0^\infty \gamma_{kj}(x,\tau)d\tau \\ &= \sum_{i\in\mathcal{S}_2}\sum_{k\in\mathcal{S}_1}\beta_i T_{ik}\Gamma_{kj}(x),\end{aligned}$$

and so

$$\boldsymbol{\pi}_1(x) = \beta_2 T_{21}\Gamma(x)$$

where $\Gamma(x)$ is the expected number of visits to level x, starting from level 0, under taboo of level 0.



Taking $i,j\in\mathcal{S}_1$, and conditioning on last visit to level $x-y$, we can derive the form of $\Gamma(x)$ by observing that

$$\gamma_{ij}(x,t) = \sum_{k\in\mathcal{S}_1}\int_0^t \gamma_{ik}(x-y,t-\tau)\gamma_{kj}(y,\tau)d\tau.$$

Integrating from 0 to ∞, we see that

$$\Gamma_{ij}(x) = \sum_{k \in \mathcal{S}_1} \Gamma_{ik}(x-y)\Gamma_{kj}(y)$$

which tells us that

$$\Gamma(x) = \Gamma(x-y)\Gamma(y), \qquad \text{all } 0 \leq y \leq x$$

and we can conclude that $\Gamma(x)$ must have an exponential form $\Gamma(x) = e^{Kx}$ for some K. So we now have concluded that

$$[\boldsymbol{\pi}_1(x) \;\; \boldsymbol{\pi}_2(x)] = \beta_2 T_{21} e^{Kx} [I \quad \Psi]$$

We still need to determine K, β_2 and Ψ. We get β_2 by considering the censored process in level 0, obtained by observing the stochastic fluid model only when it is in level 0. This is a continuous-time Markov chain on the phases. We can derive its generator by noting the fact that a transition from $i \in \mathcal{S}_2$ to $j \in \mathcal{S}_2$ can occur in one of two ways:

- by a direct jump from i to j, with rate T_{ij}, or
- by jump from i to some k in $\mathcal{S}_1$ followed by a return to level 0 in j, which occurs with rate $\sum_{k \in \mathcal{S}_1} T_{ik}\Psi_{kj}$.

The generator of the censored process is thus given by

$$U = T_{22} + T_{21}\Psi,$$

and so we know that

$$\beta_2 U = 0.$$

To get the matrix K, we look at the equations for the probability density that the the process is at level x under taboo of level zero. We use $\gamma(x,t)$ for the matrix whose entries are $\gamma_{ij}(x,t)$ and partition this into blocks, so that

$$\gamma(x,t) = \begin{bmatrix} \gamma_{11}(x,t) & \gamma_{12}(x,t) \\ \gamma_{21}(x,t) & \gamma_{22}(x,t) \end{bmatrix}.$$

Then

$$\frac{\partial}{\partial t}\gamma_{11}(x,t) + \frac{\partial}{\partial x}\gamma_{11}(x,t) = \gamma_{11}(x,t)T_{11} + \gamma_{12}(x,t)T_{21}.$$

Taking the stationary regime and integrating both sides from $t = 0$ to ∞, we derive

$$\frac{\partial}{\partial x}\Gamma_{11}(x) = \Gamma_{11}(x)T_{11} + \Gamma_{12}(x)T_{21}$$

with $\Gamma_{11}(x) = e^{Kx}$ and $\Gamma_{12}(x) = e^{Kx}\,\Psi$. We obtain

$$Ke^{Kx} = e^{Kx}T_{11} + e^{Kx}\Psi T_{21}$$

and, substituting $x = 0$, we find that

$$K = T_{11} + \Psi T_{21}.$$

We see that

$$[\boldsymbol{\pi}_1(x) \;\; \boldsymbol{\pi}_2(x)] = \boldsymbol{\beta}_2 T_{21} e^{Kx} [I \quad \Psi]$$

where

$$K = T_{11} + \Psi T_{21}$$

and

$$\beta_2 U = 0,$$

with

$$U = T_{22} + T_{21}\Psi.$$

Finally, we need to work out the correct normalisation of $\boldsymbol{\beta}_2$. We get this by observing that

$$\boldsymbol{\beta}_2 \mathbf{1} + \int_0^\infty [\boldsymbol{\pi}_1(x) \;\; \boldsymbol{\pi}_2(x)] dx = 1.$$

This reduces to

$$\boldsymbol{\beta}_2(\mathbf{1} + 2T_{21}(-K^{-1})\mathbf{1}) = 1.$$

The only object that we still need to determine is the matrix Ψ that contains the first passage probabilities from phases $i \in (0, \mathcal{S}_1)$ back to $j \in (0, \mathcal{S}_2)$. Let $G_{ij}(y)$ be the probability that, starting from (y, i) at time 0, the fluid goes down to 0 in a finite time *and* the phase at that time is j. Conditioning on the first time t at which the fluid stops increasing, we can write

$$\Psi = \int_0^\infty e^{T_{11} y} \, T_{12} \, G(y) dy$$

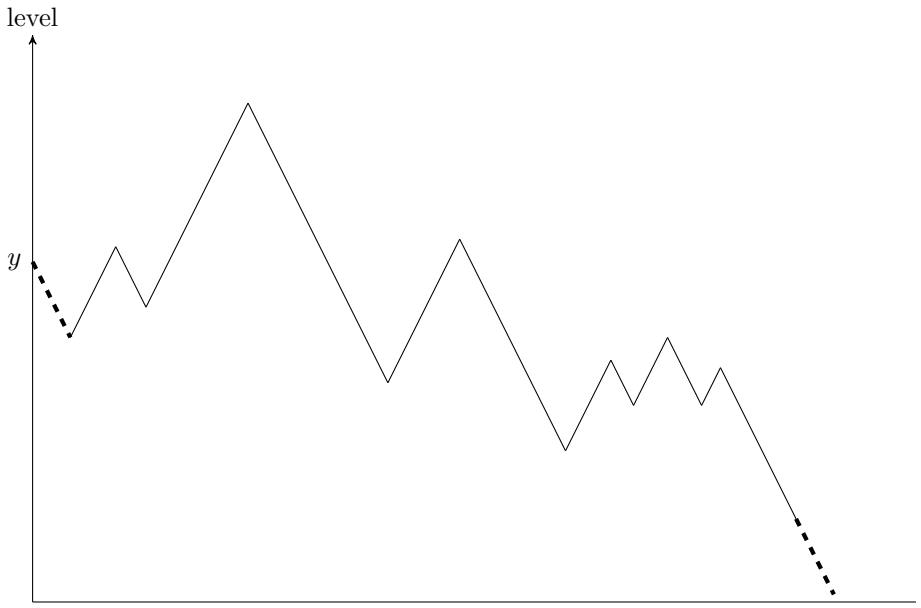
Notice that since the input rates are ± 1, we can conclude that $t = y$. Now look at the downward records process The Markov process of successive minima has transition matrix

$$V = T_{22} + T_{21},$$

and we can write

$$G(y) = e^{Vy} \qquad \text{with} \qquad V = T_{22} + T_{21}\Psi$$

$$\begin{aligned} \Psi &= \int_0^\infty e^{T_{11} y} \, T_{12} \, G(y) dy \\ &= \int_0^\infty e^{T_{11} y} \, T_{12} \, e^{Vy} dy \end{aligned}$$



Now we integrate by parts, observing that $T_1 1$ is nonsingular (because it has at least one negative row sum in each communicating class)

$$\begin{aligned} T_{11}\Psi &= \left[e^{T_{11}y}\, T_{12}\, e^{Vy}\right]_0^\infty - \int_0^\infty e^{T_{11}y}\, T_{12}\, e^{Vy} V\, dy \\ &= -\, T_{12} - \Psi V \end{aligned}$$

Substituting for V, we see that Ψ is the minimal nonnegative solution of the *Riccati equation*,

$$T_{12} + T_{11}\Psi + \Psi T_{22} + \Psi T_{21}\Psi = 0$$

Bean, O'Reilly and Taylor [**?**] proposed fourteen algorithms to solve the equation

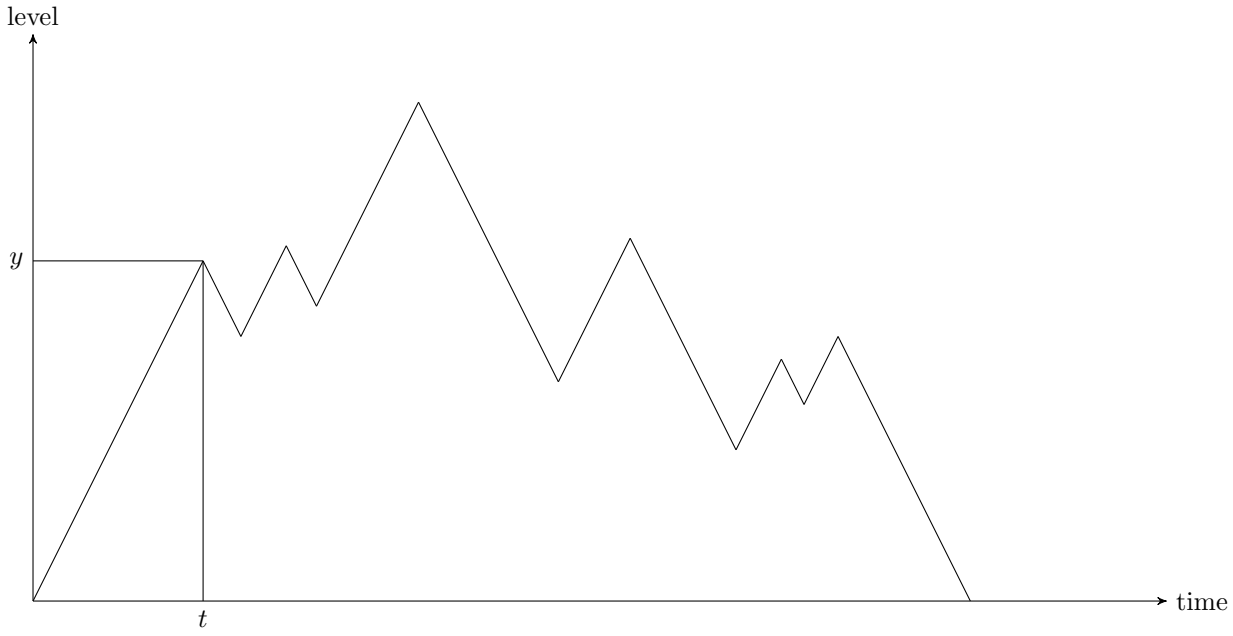
$$T_{12} + T_{11}\Psi + \Psi T_{22} + \Psi T_{21}\Psi = 0.$$

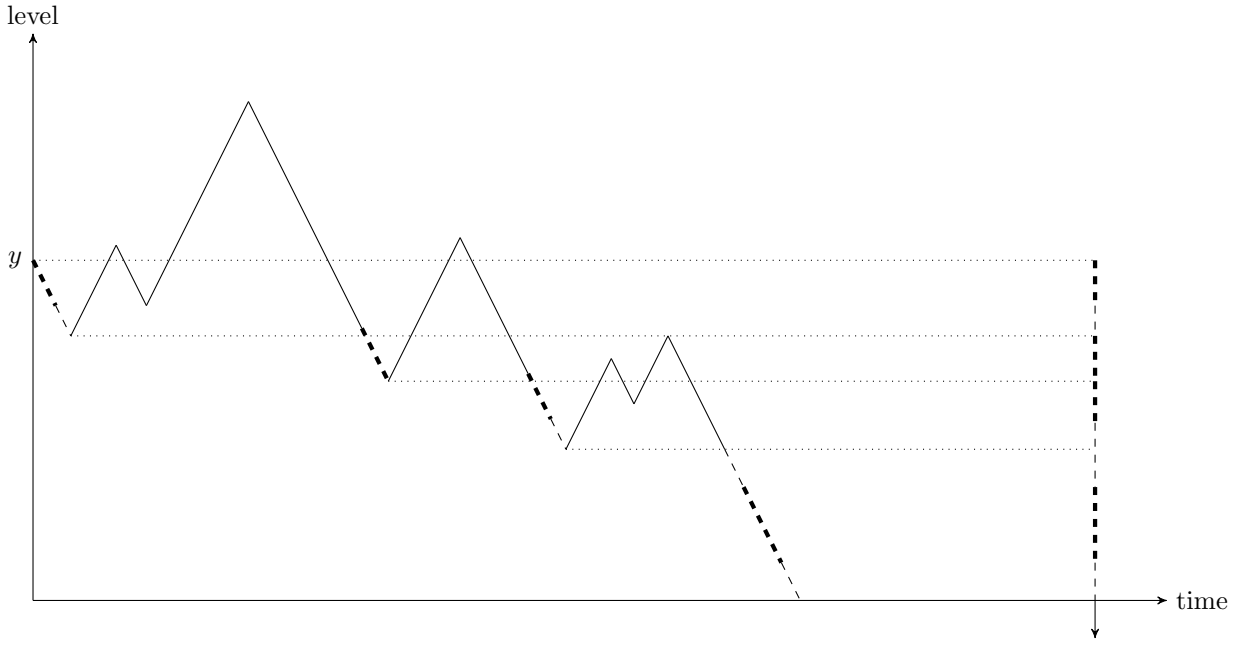
The best is Newton's method, which involves from starting from $\Psi_0 = 0$ and iteratively solving the Lyapunov equation

$$T_{12} + (T_{11} + \Psi_n T_{21})\Psi_{n+1} + \Psi_{n+1}(T_{22} + T_{21}\Psi_n) - \Psi_n T_{21}\Psi_n = 0.$$

Bibliographic Remarks

Exercises


level
y
t
time


level
y
time

Chapter 9

Deterministic Models with Additive Noise

We have spent plenty of time in the book dealing with systems of the form:

$$\begin{array}{rcl} \dot{x}(t) &=& Ax(t)+Bu(t) \\ y(t) &=& Cx(t)+Du(t) \end{array} \quad \text{and} \quad \begin{array}{rcl} x(n+1) &=& Ax(n)+Bu(n) \\ y(n) &=& Cx(n)+Du(n) \end{array} . \tag{9.1}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. The focus was mostly on the continuous time version $\big(u(t), x(t), y(t)\big)$. In unit 4 we saw how to design a state feedback controller and an observer and in later units we dealt with optimal control of such systems.

We now augment our system models by adding *noise* components. To the first equation we shall add *disturbance noise* (ξ_x) and to the second equation we shall add *measurement noise* (ξ_y). This yields:

$$\begin{array}{rcl} \dot{x}(t) &=& Ax(t)+Bu(t)+\xi_x(t) \\ y(t) &=& Cx(t)+Du(t)+\xi_y(t) \end{array} \quad \text{or} \quad \begin{array}{rcl} x(n+1) &=& Ax(n)+Bu(n)+\xi_x(n) \\ y(n) &=& Cx(n)+Du(n)+\xi_y(n) \end{array} .$$

One way of modeling the noise is by assuming that $\xi(\cdot)$ is from some function class and assuming that in controlling the system we have no knowledge of what specific $\xi(\cdot)$ from that class is used. This is the method of *robust control*. Alternatively, we can think of $\xi(\cdot)$ as a *random process(es)* by associating a probability space with the model. We shall focus on the latter approach.

The level of mathematical care that is needed for handling the continuous time case is beyond our scope as it requires some basic understanding of *stochastic calculus* (e.g. Brownian motion, Stochastic Differential Equations, Ito's formula etc...). We shall thus focus on the discrete time case which is simpler in that the random processes (es) are discrete time sequences of *random variables*. Luckily, the methods that we shall survey (Kalman filtering and Linear Quadratic Gaussian (LQG) optimal control) are often

applied in practice in discrete time on digital computers. So understanding the discrete time case is both pedagogically simpler and often of greater practical interest.

In treating $\xi_x(n)$ and $\xi_y(n)$ as discrete time random processes we shall assume they are each i.i.d. (independent and identically distributed) sequences of zero-mean Gaussian (normal) random vectors with covariance matrices Σ_x and Σ_y respectively (we review this below). In many physical situations this is actually a practical assumption:

- Having the noise of one time slot independent of the disturbance at other time slots is the practical situation (especially for short lived disturbances). (This is the first 'i' of i.i.d.).
- Having noise of a constant statistical law makes sense for time invariant systems. (This is the second 'i' of i.i.d.).
- Having noise that have a mean of 0 implies there is no general direction of the disturbance.
- Having noise that follows the Gaussian distribution is sensible if the noise is a summation of many small factors. In this case the *central limit theorem* implies that the noise distribution is Gaussian.

Note 1: We are not restricting individual coordinates of $\xi(n)$ (at any time n) to be independent.

Note 2: Note that even though the noise terms are i.i.d., $x(\cdot)$ is no longer an i.i.d. process (it will be in the pathological case in which $A = 0$ and $B = 0$).

Note 3: In many situations the variance (covariance matrix) of ξ can be modeled from "first principles" just as the (A, B, C, D) model is. This is the case of noise is due to well understood electromagnetic effects as well as due to rounding errors appearing in digital control.

What will we do with the stochastic model?

1. **State estimation (Kalman filtering):** For the deterministic system, we saw the Luenberger observer:

   $$\hat{x}(n+1) = A\hat{x}(n) + Bu(n) + K\big(y(n) - \hat{y}(n)\big).$$

   The Kalman filter is used to do essentially the same thing, yet now taking into control the fact that now $x(\cdot)$ is a random process.

2. **Optimal control (LQG):** For the deterministic system we saw how to design a state feedback control such that,

   $$\sum_{k=0}^{\infty} x'(k)Qx(k) + u'(k)Ru(k),$$

is minimized (if $Q \geq 0$ and $R > 0$). Now with random noise, $x(\cdot)$ is a random process. Further if we use a state feedback control then $u(\cdot)$ is random process. We are thus interested in finding a control law that minimizes,

$$\mathbb{E}\Big[\sum_{k=0}^{\infty} x'(k)Qx(k) + u'(k)Ru(k)\Big]$$

We will have time to touch LQG only briefly and focus mostly on the Kalman filter.

A practical note: The celebrated Kalman filter is implemented in a variety of engineering applications dealing with tracking, positioning and sensing. It is a good thing to know about outside the scope of control also.

9.1 Minimum Mean Square Estimation

Consider now the general situation in which you observe the value of a random vector $X_b = x_b$ and would like to use it to estimate the value of X_a. Here we model X_a and X_b as two random vectors (measurable functions) on the same probability space and hope that they are somewhat dependent (i.e. knowledge of X_b can give us some information on X_a). We are thus looking for a function $f(\cdot)$ such that $f(X_b)$ is a "good" estimate on X_a. There are all kinds of definitions of "good" – here is perhaps the most popular one:

$$\min_h \mathbb{E}\Big[||X_a - h(X_b)||^2\Big], \tag{9.2}$$

where $||\cdot||$ is the Euclidean norm and the minimization is over all $h(\cdot)$ in some function class whose definition we leave vague for the purpose of this informal discussion. Note that the expectation is with respect to both X_a and X_b. Does this criterion make sense? Yes, of book! Further, it turns out to be very tractable in certain cases since it turns out that the $h(\cdot)$ that minimizes (9.2) is:

$$h^*(x_b) = \mathbb{E}[X_a \mid X_b = x_b]. \tag{9.3}$$

The above is read as the "conditional expectation of the random vector X_a, given the observed value x_b". Does the best estimator $h^*(\cdot)$ make sense? Yes of book!

Brief reminder: If two random vectors X_a and X_b are distributed say with a density $f_{ab}(x_a, x_b)$, then the conditional density of X_a given $X_b = x_b$ is:

$$f_{a|b}(x_a|x_b) = \frac{f_{ab}(x_a, x_b)}{f_b(x_b)},$$

where the denominator is the marginal density of X_b, namely (assuming X_a is k-dimensional):

$$f_b(x_b) = \int_{x_a \in \mathbb{R}^k} f_{ab}(x_a, x_b) dx_a.$$

I.e. to get the marginal density of X_b you need to "integrate out" all of the values that X_a may get. And to get the conditional distribution of X_a given the information that X_b takes a specific values x_b, you need to "rescale" the joint density by the marginal of X_b. Try to draw this in two dimensions.

Now the conditional expectation (for a given value of X_b) that appears in (9.3) is simply evaluated as follows:

$$\mathbb{E}[X_a \mid X_b = x_b] = \int_{x_a \in \mathbb{R}^k} x_a \; f_{a|b}(x_a|x_b) dx_a.$$

Further note that the expression $\mathbb{E}[X_a \mid X_b]$ (where we do not specify a given values for X_b) is actually a random variable that is a function of the random variable X_b, where the function is:

$$g(X_b) = \int_{x_a \in \mathbb{R}^k} x_a \; f_{a|b}(x_a|X_b) dx_a.$$

Hence the conditional expectation $\mathbb{E}[X_a \mid X_b]$ is actually a random variable in itself. And we may thus attempt to take its expectation. It turns out that in this case:

$$\mathbb{E}\big[g(X_b)\big] = \mathbb{E}\big[\mathbb{E}[X_a \mid X_b]\big] = \mathbb{E}[X_a]. \tag{9.4}$$

Note: The above "brief reminder" about conditional expecation is very informal as technical details are missing. Yet this is enough for our needs.

Here is now (an informal) proof of (9.3):

Proof. First use the conditional expectation formula similar to (9.4):

$$\mathbb{E}\big[||X_a - h(X_b)||^2\big] = \mathbb{E}\Big[\mathbb{E}\big[||X_a - h(X_b)||^2 \mid X_b\big]\Big] = \int \mathbb{E}\big[||X_a - h(X_b)||^2 \mid X_b(\omega)\big] dP_{X_b}(\omega). \tag{9.5}$$

The last expression represents the outer expectation as a Lebesgue integral with respect to the probability measure associated with the random variable X_b. This is not needed to understand the proof, but is here for additional clarity on the meaning of expectation.

Note that the internal conditional expectation (conditional on X_b) is a function, $\tilde{g}(\cdot)$ of the random variable X_b. Let's investigate this function in a bit greater detail. Assume that the estimator $h(X_b)$ takes on the value z (i.e. assume that in the probability sample space associated with the random variable X_b, we get and ω such that $h(X_b(\omega)) = z$). Then,

$$\mathbb{E}\big[||X_a - z||^2 \mid X_b\big] = \mathbb{E}\big[||X_a||^2 \mid X_b\big] - 2z'\mathbb{E}\big[X_a \mid X_b\big] + ||z||^2$$

Taking derivative with respect to z (note that z is generally a vector) and equating to 0 implies that the above is minimized by $z = \mathbb{E}[X_a|X_b]$. I.e. the integrand in (9.5) is minimized by setting,

$$h(X_b) = \mathbb{E}[X_a|X_b].$$

Thus the integral (the outer expectation) is also minimized by this choice of $h(\cdot)$ and thus the (9.2) is minimized by (9.3). □

Evaluating (9.3) for arbitrarily distributed X_a and X_b can be a complicated (not explicitly solvable) task. Yet for Gaussian random vectors we are blessed with a clean result. Indeed as we saw in the case of Gaussian random vectors that this conditional expectation has the closed (linear) form. So if you believe (9.3), in the case of Gaussian random vectors,

$$h^*(x_b) = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(x_b - \mu_b).$$

We thus see that for Gaussian random vectors, the optimal estimator $h^*(\cdot)$ is an linear (affine to be precise) function of x_b. It is thus tempting to restrict the function class of $h^*(\cdot)$ in (9.2) to,

$$h(x_b) = Gx_b + g,$$

where G and g are a matrix and a vector of the appropriate dimension. The pair (G, g) that minimizes (9.2) is sometimes called the LMMSE estimator (Linear Minimum Mean Square Error estimator).

Exercise 9.1.1. *What are G and g in the case of Gaussian random variables?*

Exercise 9.1.2. *Prove the following proposition by taking derivatives w.r.t. to G and g.*

Proposition 9.1.3. *Let (X_a, X_b) be random vectors with means μ_a and μ_b respectively and with a covariance matrix (of $(X_a, X_b)'$) being:*

$$\begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma'_{ab} & \Sigma_b \end{bmatrix}.$$

Then LMMSE estimator of X_a given $X_b = x_b$ is:

$$h^*(x_b) = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(x_b - \mu_b).$$

Further the covariance matrix of the error vector $X_a - h^*(X_b)$ is given by:

$$\mathbb{E}\big[\big(X_a - h^*(X_b)\big)\big(X_a - h^*(X_b)\big)'\big] = \Sigma_a - \Sigma_{a,b}\Sigma_b^{-1}\Sigma'_{a,b}.$$

In the case of non-Gaussian random variables, restricting to an affine estimator based on G and g is often a compromise:

Exercise 9.1.4. *Let X_b have a uniform distribution on the interval $[-1, 1]$ and let $X_a = X_b^2$. Find the best affine estimator of X_a in terms of X_b and compare its performance (using the objective (9.2)) to the best estimator (9.3).*
Repeat for the case of,

$$f_{a,b}(x_a, x_b) = \begin{cases} 2e^{-(x_a+x_b)} & 0 \le x_b \le x_a < \infty, \\ 0 & \textit{elsewhere.} \end{cases}$$

9.2 The Kalman Filtering Problem "Solved" by LMMSE

Our goal is to have a state estimate, $\hat{x}(n)$ for a given (A, B, C, D) + noise system:

$$\begin{array}{rcl} x(n+1) & = & Ax(n) + Bu(n) + \xi_x(n) \\ y(n) & = & Cx(n) + Du(n) + \xi_y(n) \end{array}.$$

More specifically we assume we have controlled this system over times $k = 0, \ldots, N-1$ by setting inputs $u(0), \ldots, u(N-1)$ (which we know) and have measured outputs $y(0), \ldots, y(N-1)$. Note that we treat $x(0)$ as a random variable also where we assume we know its mean and covariance.

We will now show that his problem can be posed as estimating X_a based on measurement of X_b (as presented in the previous section) where,

$$X_a = (x(0)', x(1)', \ldots, x(N)')', \qquad X_b = (y(0)', y(1)', \ldots, y(N)')',$$

and the inputs $u(0), \ldots, u(N-1)$ are known values.

By iterating the system, we get:

$$\begin{array}{rcl} x(1) & = & Ax(0) + Bu(0) + \xi_x(0), \\ x(2) & = & A^2x(0) + ABu(0) + Bu(1) + A\xi_x(0) + \xi_x(1), \\ x(3) & = & A^3x(0) + A^2Bu(0) + ABu(1) + Bu(2) + A^2\xi_x(0) + A\xi_x(1) + \xi_x(2), \\ & \vdots & \\ x(N) & = & A^Nx(0) + \sum_{k=0}^{N-1} A^{N-1-k}Bu(k) + \sum_{k=0}^{N-1} A^{N-1-k}\xi_x(k). \end{array}$$

Plugging the above in the output equations, we get,

$$\begin{array}{rcl} y(0) & = & Cx(0) + Du(0) + \xi_y(0), \\ y(1) & = & CAx(0) + CBu(0) + C\xi_x(0) + Du(1) + \xi_y(1) \\ y(2) & = & CA^2x(0) + CABu(0) + CBu(1) + CA\xi_x(0) + C\xi_x(1) + Du(2) + \xi_y(2) \\ & \vdots & \\ y(N) & = & CA^Nx(0) + \sum_{k=0}^{N-1}(CA^{N-1-k}B)u(k) + Du(N) + \sum_{k=0}^{N-1} CA^{N-1-k}\xi_x(k) + \xi_y(N) \end{array}$$

It is thus a simple matter to write out constant matrices $\tilde{A}, \tilde{C}$ and well as functions of

the known input, $\tilde{b}(u), \tilde{d}(u)$, such that:

$$X_a = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N) \end{bmatrix} = \tilde{A} \begin{bmatrix} x(0) \\ \xi_x(0) \\ \vdots \\ \xi_x(N-1) \end{bmatrix} + \tilde{b}\big(u(0), \ldots, u(N-1)\big),$$

$$X_b = \begin{bmatrix} y(0) \\ \vdots \\ y(N) \end{bmatrix} = \tilde{C} \begin{bmatrix} x(0) \\ \xi_x(0) \\ \vdots \\ \xi_x(N-1) \end{bmatrix} + \begin{bmatrix} \xi_y(0) \\ \vdots \\ \xi_y(N) \end{bmatrix} + \tilde{d}\big(u(0), \ldots, u(N)\big)$$

Exercise 9.2.1. *Specify $\tilde{A}, \tilde{C}$ as well as $\tilde{b}(u), \tilde{d}(u)$ explicitly.*

It is now useful to consider the combined random vector,

$$\zeta = \begin{bmatrix} x(0) \\ \xi_x(1) \\ \vdots \\ \xi_x(N-1) \\ \xi_y(0) \\ \vdots \\ \xi_y(N) \end{bmatrix}.$$

We may now rewrite the equations for X_a and X_b as follows:

$$\begin{bmatrix} X_a \\ X_b \end{bmatrix} = \tilde{F}\zeta + f\big(u(0), \ldots, u(N)\big).$$

Exercise 9.2.2. *Specify $\tilde{F}$ as well as $\tilde{f}(u)$ explicitly.*

We further have,

$$\Sigma_\zeta := Cov(\zeta) = \begin{bmatrix} \Sigma_{x(0)} & 0 & 0 & 0 \\ 0 & \begin{bmatrix} \Sigma_x & & 0 \\ & \ddots & \\ 0 & & \Sigma_x \end{bmatrix} & \begin{bmatrix} \Sigma_{xy} & & 0 \\ & \ddots & \\ 0 & & \Sigma_{xy} \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} \Sigma'_{xy} & & 0 \\ & \ddots & \\ 0 & & \Sigma'_{xy} \end{bmatrix} & \begin{bmatrix} \Sigma_y & & 0 \\ & \ddots & \\ 0 & & \Sigma_y \end{bmatrix} & 0 \\ 0 & 0 & 0 & \Sigma_y \end{bmatrix}.$$

Here the $\Sigma_{x(0)}$ is an assumed covariance matrix for $x(0)$. The other Σ elements are the covariances of the noise vectors: Σ_x is the covariance matrix of the disturbance. Σ_y is the

covariance matrix of the measurement noise. And $\Sigma_{x,y}$ is the cross-covariance between disturbance and measurements (this is often assumed 0).

Thus,

$$Cov\Big(\begin{bmatrix} X_a \\ X_b \end{bmatrix}\Big) = \tilde{F}\Sigma_\zeta \tilde{F}' =: \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma'_{ab} & \Sigma_b \end{bmatrix}.$$

Observe also that,

$$\begin{aligned} \mu_a = \mathbb{E}[X_a] &= [\mathbb{E}[x(0)]' \;\; 0' \;\; \dots \;\; 0']' + \tilde{b}\big(u(0), \dots, u(N-1)\big), \\ \mu_b = \mathbb{E}[X_b] &= \tilde{d}\big(u(0), \dots, u(N)\big). \end{aligned}$$

We now have all of the needed ingredients of Proposition 9.1.3 to calculate the LMMSE of X_a based on X_a. I.e. take,

$$h^*(x_b) = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(x_b - \mu_b).$$

and then the predictor at for the state at time n is:

$$\hat{x}(n) = \Big[h^*(x_b)\Big]_{(nN+1,\dots,nN+n)}.$$

While this is very nice, it is not efficient from a control theory perspective since getting an estimate for X_a requires computation of the order of $O((nN)^3)$. It would be much better to have some sort of recursive solution that yields $\hat{x}(N)$ at each step. This is the celebrated Kalman filtering algorithm which we present in the next section.

Exercise 9.2.3. *Consider the scalar system:*

$$\begin{aligned} x(n+1) &= 2x(n) + u(n) + \xi_x(n) \\ y(n) &= x(n) + \xi_y(n) \end{aligned}$$

Where $\xi_x(n)$ and $\xi_y(n)$ are both of unit variance and assumed uncorrelated.
Assume $x(0)$ is such that $\mathbb{E}[x(0)] = 0$ and $Var\big(x(0)\big) = 0$. Assume a control input of $u(n) = 1$ was applied to the system over the times $n = 0, 1, 2$. And the measured output was, $\big(y(0), y(1), y(2), y(3)\big) = (y_0, y_1, y_2, y_3)$.
Use the derived LMMSE in this section to obtain an estimator for $x(3)$ in terms of (y_0, y_1, y_2, y_3).

9.3 The Kalman Filtering Algorithm

For simplicity in this section, we assume $B = 0$ and $D = 0$ and thus our system is

$$\begin{aligned} x(n+1) &= Ax(n) + \xi_x(n) \\ y(n) &= Cx(n) + \xi_y(n) \end{aligned}.$$

The more general case (with inputs) easily follows and is left as an exercise. We shall also assume for simplicity that $\Sigma_{xy} = 0$. This assumption can also be relaxed.

In general the Kalman filtering algorithm is based on (deterministic) sequence $K(0), K(1), \ldots$ that is used as follows:

$$\hat{x}(n+1) = A\hat{x}(n) + K(n)\big(y(n+1) - CA\hat{x}(n)\big). \tag{9.6}$$

In this sense it is like a Luenberger observer yet where the matrices K generally depend on time (even in the case presented here where A and C are constant). As an aid for calculating $K(n)$ we have,

$$S(n) := Cov\big(x(n+1) - \hat{x}(n+1) \;\big|\; x(n), x(n-1), \ldots, x(0)\big),$$

with $S(n)$ following the following recursion:

$$S(n+1) = A\Big(S(n) - S(n)C'\big(CS(n)C' + \Sigma_y\big)^{-1}CS(n)\Big)A' + \Sigma_x.$$

Now $S(n)$ is used to obtain $K(n)$ as follows:

$$K(n) = S(n)C'\big(CS(n)C' + \Sigma_y\big)^{-1}.$$

Note that in many applications we may also use the *steady state Kalman filter* in which we take $S(n)$ as the fixed unique positive definite S solving equation:

$$S = A\Big(S - SC'\big(CSC' + \Sigma_y\big)^{-1}CS\Big)A' + \Sigma_x.$$

This then yields a constant K in (9.6).

It is obvious that the Kalman filter and (even more) the steady state Kalman filter are computationally efficient compared to the method described in the previous section.

Exercise 9.3.1. *Consider the scalar system,*

$$\begin{aligned} x(n+1) &= \frac{4}{5}x(n) + \xi_x(n), \\ y(n) &= x(n) + \xi_y(n). \end{aligned}$$

Take, $Var\big(\xi_x(n)\big) = 9/25$ *and* $Var\big(\xi_y(n)\big) = 1$. *Find the form of the predictor* $\hat{x}_n(y)$. *Find the steady state predictor.*

We have the following:

Theorem 9.3.2. *The sequence defined in* (9.6) *is the LMMSE estimator of* $x(n)$.

Note that the proof below is based on the the fact the noise terms are Gaussian. In this case the LMMSE is also the optimal MSE estimator. A more general proof based on the *orthogonality principle*, based on the representation of square integrable random vectors as elements of a Hilbert space is also known but is not discussed here. In that case Gaussian assumptions are not required and (9.6) is still the LMMSE (yet not necessarily the best MSE estimator).

Proof. Denote $Y(n) = \Big(y(0), y(1), \ldots, y(n)\Big)$ and set,

$$\hat{x}^-(n) := \mathbb{E}[x(n)|Y(n-1)], \qquad \hat{x}(n) := \mathbb{E}[x(n)|Y(n)].$$

Observe by (9.3) that $\hat{x}(n)$ is the optimal MSE estimator of $x(n)$ and thus also the LMMSE estimator since $x(\cdot)$ is Gaussian. Denote the respective conditional covariance matrices:

$$\begin{aligned} P(n) &:= \mathbb{E}\big[\big(x(n)-\hat{x}(n)\big)\big(x(n)-\hat{x}(n)\big)' \mid Y(n)\big], \\ P^-(n) &:= \mathbb{E}\big[\big(x(n)-\hat{x}^-(n)\big)\big(x(n)-\hat{x}^-(n)\big)' \mid Y(n-1)\big]. \end{aligned}$$

Further for $n=0$ set, $P^-(0) := \Sigma_{x(0)}$ and $\hat{x}^-(0) := \mathbb{E}[x(0)]$. Observe that in addition to $y(\cdot)$ and $x(\cdot)$, the sequences $\hat{x}^-(\cdot)$ and $\hat{x}(\cdot)$ are also jointly Gaussian since they are all generated by linear combinations of the "primitives" of the process, $\xi_x(\cdot), \xi_y(\cdot)$ and $x(0)$ and also since $\hat{x}^-(\cdot)$ and $\hat{x}(\cdot)$ follow from the formula for the conditional expecation in (D.7).

The key step is to observe that when conditioning on $Y(n-1)$, the distribution of $[x(n)', y(n)']'$ is,

$$\mathcal{N}\Big(\begin{bmatrix} \hat{x}^-(n) \\ C\hat{x}^-(n) \end{bmatrix}, \begin{bmatrix} P^-(n) & P^-(n)C' \\ CP^-(n) & CP^-(n)C' + \Sigma_y \end{bmatrix} \Big). \tag{9.7}$$

Noting that,

$$\hat{x}(n) = \mathbb{E}[x(n) \mid Y(n)] = \mathbb{E}[x(n) \mid y(n), Y(n-1)],$$

we apply the mean and covariance formulas of(D.7) based on (9.7) with everything preconditioned on $Y(n-1)$ to get:

$$\hat{x}(n) = \hat{x}^-(n) + P^-(n)C'\big(CP^-(n)C' + \Sigma_y\big)^{-1}\big(y(n) - C\hat{x}^-(n)\big), \tag{9.8}$$

$$P(n) = P^-(n) - P^-(n)C'\big(CP^-(n)C' + \Sigma_y\big)^{-1}CP^-(n). \tag{9.9}$$

Now observe that,

$$\hat{x}^-(n+1) = \mathbb{E}[x(n+1)|Y(n)] = \mathbb{E}[Ax(n) + \xi_x(n)|Y(n)] = A\mathbb{E}[x(n)|Y(n)] = A\hat{x}(n),$$

and thus substitution in (9.8) for time $n+1$ yields,

$$\hat{x}(n+1) = A\hat{x}(n) + P^-(n+1)C'\big(CP^-(n+1)C' + \Sigma_y\big)^{-1}\big(y(n+1) - CA\hat{x}(n)\big).$$

Further,

$$P^-(n+1) = Cov\big(x(n+1) \mid Y(n)\big) = Cov\big(Ax(n) + \xi_x(n) \mid Y(n)\big) = AP(n)A' + \Sigma_x.$$

Substitution of (9.9) in the above yields

$$P^-(n+1) = A\Big(P^-(n) - P^-(n)C'\big(CP^-(n)C' + \Sigma_y\big)^{-1}CP^-(n)\Big)A' + \Sigma_x.$$

Now denote $S(n) := P^-(n+1)$ to obtain the desired equations:

$$\begin{aligned}\hat{x}(n+1) &= A\hat{x}(n) + K(n)\big(y(n+1) - CA\hat{x}(n)\big)\\ K(n) &= S(n)C'\big(CS(n)C' + \Sigma_y\big)^{-1}\\ S(n+1) &= A\Big(S(n) - S(n)C'\big(CS(n)C' + \Sigma_y\big)^{-1}CS(n)\Big)A' + \Sigma_x.\end{aligned}$$

□

Exercise 9.3.3. *What is the Kalman filter for the case of $B \neq 0$ and $D \neq 0$. Describe any needed changes in the proof above.*

9.4 LQR Revisited: LQG

We only touch LQG briefly and informally. Consider the system,

$$\begin{aligned}x(n+1) &= Ax(n) + Bu(n) + \xi_x(n)\\ y(n) &= Cx(n) + Du(n) + \xi_y(n),\end{aligned}$$

and assume our goal is to find an optimal output feedback law: $u^*(y)$, such that the following is minimized:

$$\mathbb{E}\big[\sum_{k=0}^{N} x(n)'Qx(n) + u(n)'Ru(n)\big],$$

with N either finite or infinite and $Q \geq 0$, $R > 0$. Assume further that (A, B) is controllable and (A, C) is observable.

This generalization of the linear quadratic regulator (LQR) problem studied in previous units, is often refereed to as the LQG problem (Linear quadratic Gaussian). Note that the LQR formulation that we studied ignored the output y and assumed state-feedback.

It turns out that solution of the LQG problem by means of dynamic programming (yet with a stochastic element) is essentially equivalent to dynamic programming solution of LQR. The basic ingredient is once again *Bellman's principle of optimality*, yet this time presented in a stochastic (Markovian) setting:

In somewhat greater generality, consider systems of the form:

$$x(n+1) = f\Big(x(n), u\big(x(n)\big), \xi(n)\Big), \qquad n = 0, 1, \ldots, N-1,$$

where $f(\cdot)$ is some function and ξ is an i.i.d. sequence. For any prescribed $u(\cdot)$ such a system is a Markov chain (informally a stochastic process whose next step only depends on the current state and some noise component and not on the past). The basic setting of *stochastic dynamic programming* (a.k.a. Markov decision processes) is to find a

$u_n^*(x), n = 0, 1, \ldots, N-1$ such that,

$$\mathbb{E}\Big[g_N\big(x(N)\big) + \sum_{k=0}^{N-1} g_k\big(x(k), u_k\big(x(k)\big), \xi(k)\big)\Big],$$

is minimized. Here $g_k(\cdot), k = 1, \ldots, N-1$ is the cost per stage and $g_N(\cdot)$ is the terminal cost. Note also the slight change of notation, where we put the time index as a subscript of u.

Principle of optimality (stochastic version): Let $u^* = \big(u_0^*(\cdot), \ldots, u_{N-1}^*(\cdot)\big)$ be an optimal policy. Assume that in the stochastic process resulting from $u^*(\cdot)$ it is possible to reach a given state at time n . Consider now the *subproblem* whereby the process is in state $x(n)$ at time n and wish to minimize:

$$\mathbb{E}\Big[g_N\big(x(N)\big) + \sum_{k=n}^{N-1} g_k\big(x(k), u_k\big(x(k)\big), \xi(k)\big)\Big],$$

then the truncated policy $\big(u_n^*(\cdot), u_{n+1}^*(\cdot), \ldots, u_{N-1}^*(\cdot)\big)$ is optimal for this subproblem. □

By application of the principle of optimality in similar spirit to as is done for the solution of discrete time LQR, we get a solution to the LQG problem that parallels that of the LQR problem, yet takes the noise into account in the following beautiful manner:

1. The Kalman filtering solution yields an estimator of $\hat{x}(\cdot)$.
2. The deterministic LQR solution (assuming known x) is applied to $\hat{x}$.

In view of the brevity of this section, we omit details, yet mention that this is a stochastic manifestation of the *separation principle* presented in Unit 4, where the observer and feedback control law can be designed separately and then combined. Non-linear (deterministic and stochastic) systems usually do not exhibit this clean property – and are a current active area of research.

Bibliographic Remarks

Exercises

Appendix A

Basics

A.1 Sets

A *set* is a collection of objects, e.g. $\mathcal{A} = \{1, -3, 8, a\}$. Sets are not regarded as ordered and can have a finite or infinite number of objects. $x \in \mathcal{A}$ is read as "x is an element of $\mathcal{A}$". Similarly $x \notin \mathcal{A}$. E.g. for the set above we have $1 \in \mathcal{A}$ and $4 \notin \mathcal{A}$.

We say $\mathcal{A}$ is a *subset* of $\mathcal{B}$ (denoted by $\mathcal{A} \subset \mathcal{B}$) if whenever $x \in \mathcal{A}$ we also have $x \in \mathcal{B}$. We say two sets $\mathcal{A}$ and $\mathcal{B}$ are equal (denoted $\mathcal{A} = \mathcal{B}$) if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{A}$. The empty set, denoted $\emptyset$ has no elements ($\emptyset = \{\}$). It is a subset of any other set.

We often have a *universal set* (in probability theory it is often denoted Ω). Having such a set allows us to define the *complement* of any subset of Ω: $\mathcal{A}^c$. This is the set of all elements that are not in $\mathcal{A}$ but in Ω. This can also be written as,

$$\mathcal{A}^c = \{x \in \Omega : x \notin \mathcal{A}\}.$$

Note that $(\mathcal{A}^c)^c = \mathcal{A}$. Also, $\Omega^c = \emptyset$.

The *union* of two sets $\mathcal{A}$ and $\mathcal{B}$, denoted $\mathcal{A} \cup \mathcal{B}$, is the set that contains all elements that are in either $\mathcal{A}$, $\mathcal{B}$ or both. E.g. $\{-2, 0, 3\} \cup \{0, 1\} = \{0, -2, 3, 1\}$. Note that $\mathcal{A} \cup \mathcal{A}^c = \Omega$.
The *intersection* of two sets $\mathcal{A}$ and $\mathcal{B}$, denoted $\mathcal{A} \cap \mathcal{B}$, is the set of all elements that are in both $\mathcal{A}$ and $\mathcal{B}$. E.g. $\{-2, 0, 3\} \cap \{0, 1\} = \{0\}$. Note that $\mathcal{A} \cap \mathcal{A}^c = \emptyset$.

Exercise A.1.1. *Prove the following:*

1. $\mathcal{A} \cap \mathcal{B} \subset \mathcal{A} \cup \mathcal{B}$.
2. *Commutative properties:* $\mathcal{A} \cup \mathcal{B} = \mathcal{B} \cup \mathcal{A}$ *and* $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$.
3. *Associative properties:* $\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cup \mathcal{C}$ *and* $\mathcal{A} \cap (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C}$.
4. *Distributive properties:* $\mathcal{A} \cup (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C})$ *and* $\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C})$.

5. *DeMorgan's rules:* $(\mathcal{A} \cup \mathcal{B})^c = \mathcal{A}^c \cap \mathcal{B}^c$, $(\mathcal{A} \cap \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}^c$.

Two sets $\mathcal{A}$ and $\mathcal{B}$ are said to be *disjoint* if $\mathcal{A} \cap \mathcal{B} = \emptyset$. The *difference* of $\mathcal{A}$ and $\mathcal{B}$, denoted $\mathcal{A} \backslash \mathcal{B}$ is the set of elements that are in $\mathcal{A}$ and not in $\mathcal{B}$. Note that $\mathcal{A} \backslash \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$.
We can use the following notation for unions: $\bigcup_{\gamma \in \Gamma} \mathcal{A}_\gamma$, or similarly for intersections $\bigcap_{\gamma \in \Gamma} \mathcal{A}_\gamma$. This means taking the union (or intersection) of $\mathcal{A}_\gamma$ for all γ in Γ. E.g. if $\Gamma = \{1, 2\}$ it implies $\mathcal{A}_1 \cup \mathcal{A}_2$ (or similarly for intersection).

Exercise A.1.2. *Prove DeMorgan's rules for arbitrary collections:*

$$\Big(\bigcup_{\gamma \in \Gamma} A_\gamma\Big)^c = \bigcap_{\gamma \in \Gamma} A_\gamma^c, \qquad \textit{and} \qquad \Big(\bigcap_{\gamma \in \Gamma} A_\gamma\Big)^c = \bigcup_{\gamma \in \Gamma} A_\gamma^c.$$

The *power set* of a set $\mathcal{A}$, denoted $2^{\mathcal{A}}$ is the set of all subsets of $\mathcal{A}$, e.g.,

$$2^{\{a,b\}} = \{\emptyset,\ \{a\},\ \{b\},\ \{a, b\}\}.$$

A.2 Functions

A *function*, f, is an object denoted by $f : \mathcal{X} \to \mathcal{Y}$, where the set $\mathcal{X}$ is called the *domain* and the set $\mathcal{Y}$ is called the *codomain* and for every $x \in \mathcal{X}$ there is a unique $y \in \mathcal{Y}$ denoted $y = f(x)$.
The subset of the codomain, $\mathcal{Y}$ which is actually hit by the function is called the *image* of the function, denoted,

$$f(\mathcal{X}) := \{y \in \mathcal{Y}\ :\ \exists x \in \mathcal{X},\ f(x) = y\}.$$

The function is called a *surjection* (or *surjective function* or *onto*) if $f(\mathcal{X}) = \mathcal{Y}$. The function is called an *injection* (or *injective function* or *one-to-one*) if does not map distinct elements of its domain to the same element of the codomain.
The function is called a *bijection* (or *bijective function* or *one-to-one correspondence*) if it is both a surjection and an injection. Bijections are useful since they allow imply that there is an inverse function, denoted $f^{-1} : \mathcal{Y} \to \mathcal{X}$ that is also a bijection.

A.3 Counting

For a finite set $\mathcal{A}$, $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$. E.g. $|\{a, b, c\}| = 3$. A *k-tuple* is simply an ordered list with values $(x_1, \ldots, x_k)$. The multiplication principle: The number of distinct ordered k-tuples $(x_1, \ldots, x_k)$ with components $x_i \in \mathcal{A}_i$ is $|\mathcal{A}_1| \cdot |\mathcal{A}_2| \cdot \ldots \cdot |\mathcal{A}_k|$.

Exercise A.3.1. *Show that for $\mathcal{A}$ finite,*

$$|2^{\mathcal{A}}| = 2^{|\mathcal{A}|}.$$

The number of ways to choose k objects from a finite set $\mathcal{A}$ with $|\mathcal{A}| = n$, not requiring the objects to be distinct is: n^k. This is sometimes called sampling with replacement and with ordering. Note that this also corresponds to the number of ways of distributing k distinct balls in n bins where there is no limit on the number of balls that can fit in a bin.

The number of ways to choose k distinct objects from a finite set $\mathcal{A}$ of size n where order matters is

$$n \cdot (n-1) \cdot \ldots \cdot (n-k+1).$$

I.e. this is the number of k-tuples with distinct elements selected from $\mathcal{A}$. This is number also corresponds the number of ways of distributing k distinct balls in n bins where there is a limit of at most one ball per bin. Note that if $k = n$ this number is $n!$ (e.g. $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$). Each ordering of a finite set of size n is called a *permutation*. Thus the number of permutations is $n!$. Note Stirling's formula:

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}.$$

The "similar sign" $\sim$ indicates that the ratio of the left hand side and right hand side converges to 1 as $n \to \infty$. Note: We often use $\sim$ to indicate the distribution of a random variable - something completely different.

The number of ways of choosing k distinct objects from a finite set $\mathcal{A}$ where order does not matter is similar to the case where order matters but should be corrected by a factor of $k!$. This number is sometimes called the binomial coefficient:

$$\left(\begin{array}{c} n \\ k \end{array} \right) := \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}.$$

I.e. this is the number of subsets of size k of a set of size n. It also corresponds to the number of ways of distributing k indistinguishable balls in a n bins with room for at most one ball per bin.

Exercise A.3.2. *Prove each of these properties both algebraiclly and using counting arguments:*

1. $$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right).$$

2. $$\left(\begin{array}{c} n \\ 0 \end{array} \right) = \left(\begin{array}{c} n \\ n \end{array} \right) = 1.$$

3.

$$\binom{n}{1} = \binom{n}{n-1} = n.$$

4.

$$\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}.$$

5. The binomial theorem:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

6.

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

A.4 Number Systems

The set of *natural numbers*, denoted $\mathbb{N}$ is $\{1, 2, 3, \ldots\}$. A set, $\mathcal{S}$ is said to be *countable* if it is either finite, or it is infinite and there exists a one-to-one mapping between $\mathcal{S}$ and $\mathbb{N}$, in the latter case, it is sometimes refered to as *countably infinite*.
The set of *integers*, denoted $\mathbb{Z}$ is $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. The non-negative integers are denoted $\mathbb{Z}_+ := \{0\} \cup \mathbb{N}$. The set of *rational numbers*, denoted $\mathbb{Q}$ are all numbers that can be represented in the form m/n with $m, n \in \mathbb{Z}$.

Exercise A.4.1. *Show $\mathbb{Z}$, $\mathbb{Z}_+$ and $\mathbb{Q}$ are countably infinite sets.*

The set of *reals* or *real numbers*, denoted $\mathbb{R}$ contains $\mathbb{Q}$ as well as all limits of sequences of elements in $\mathbb{Q}$. A useful subset of the reals is the interval $[0, 1] := \{x : 0 \leq x \leq 1\}$. Any element of $[0, 1]$ can be represented by an infinite sequence of binary digits such as,

$$0010100111010011111010101011010 1 \ldots,$$

by this representation it can be shown that $[0, 1]$ and hence $\mathbb{R}$ is not a countable set.

Theorem A.4.2. *The set $\mathbb{R}$ is not countable.*

The above theorem is proved by assuming that $[0, 1]$ is countable and thus its elements can be ordered. Then showing that the number represented by flipping the i'th digit of the i'th element of the ordered sequence does not equal any of the ordered numbers, yet is an element of $[0, 1]$.

A.4.1 Complex Numbers

A complex number is an ordered pair (u, v) with $u, v \in \mathbb{R}$ that can also be represented as $u + iv$. The *real part* of $z = (u, v) = u + iv$ is denoted, $\Re(z) = u$ and the *imaginary part* is, $\Im(z) = v$. The set of complex numbers is,

$$\mathbb{C} = \{u + iv : u, v \in \mathbb{R}\}.$$

Addition of two complex numbers (u_1, v_1) and (u_2, v_2) is defined as though they are elements of the vector space (see below) $\mathbb{R}^2$:

$$(u_1 + iv_1) \ + \ (u_2 + iv_2) := (u_1 + v_1) + i(v_1 + v_2).$$

Further, multiplication of the two numbers is defined as follows:

$$(u_1 + iv_1)\ (u_2 + iv_2) = (u_1u_2 - v_1v_2) + i(u_1v_2 + u_2v_1).$$

Note that vector spaces (and $\mathbb{R}^2$) do not have a multiplication operation of two vectors associated with them. Hence complex numbers generalize $\mathbb{R}^2$ by allowing elements to be multiplied.

Exercise A.4.3. *Verify the following:*

1. $i^2 = -1$.
2. commutativity*:* $wz = zw$ *for all* $w, z \in \mathbb{C}$.
3. associativity*:* $(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$ *for* $z_1, z_2, z_3 \in \mathbb{C}$.
4. $z + 0 = z$ *and* $z1 = z$ *for all* $z \in \mathbb{C}$ *where* $0 \in \mathbb{C}$ *is defined as* $(0, 0)$ *and* $1 \in \mathbb{C}$ *is defined as* $(1, 0)$.
5. *For every* $z \in \mathbb{C}$ *there is a unique* $w \in \mathbb{C}$ *such that* $z + w = 0$.
6. *For every* $z \in \mathbb{C}$ *with* $z \neq 0$ *there is a unique* $w \in \mathbb{C}$ *such that* $zw = 1$.
7. distributivity*:* $z_1(z_2 + z_3) = z_1z_2 + z_1z_3$ *for all* $z_1, z_2, z_3 \in \mathbb{C}$.

Think of i as $\sqrt{-1}$: There is no real number x such that $x^2 = -1$, hence there is no real solution to the equation, $x^2 + 1 = 0$, thus the imaginary number i introduced. Note that some text used by electrical engineers sometimes use the notation j for i, reserving the latter for current.

The *conjugate* of a complex number $z = u + iv$ is $\bar{z} = u - iv$. The *absolute value* of z is the non-negative real number,

$$|z| = \sqrt{z\bar{z}} = \sqrt{u^2 + v^2}.$$

Thus multiplying a complex number by it's conjugate yields a real number $(u^2 + v^2)$, this is useful for dividing. Assume $w \neq 0$, then,

$$\frac{z}{w} = \frac{z\bar{w}}{c^2 + d^2} = \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2} i.$$

Exercise A.4.4. *Show the following:*

1. $\overline{z \pm w} = \bar{z} \pm \bar{w}$.
2. $\overline{zw} = \bar{z}\bar{w}$.
3. $\overline{z/w} = \bar{z}/\bar{w}$.
4. $\frac{1}{z} = \frac{\bar{z}}{|z|^2}$.

Polar form representation of complex numbers is very useful for multiplication/division. For $z = a + bi$, the magnitude (called modulus) of z is $r = |z|$ and the argument of z, φ is the angle between the x-axis and z expressed in radians. For positive, a this is simply $\arctan(b/a)$, observe that this is a value in the range $(-\frac{\pi}{2}, \frac{\pi}{2})$. For other values, more care is needed.

We can now express,

$$z = a + bi = re^{\varphi i}.$$

The nice thing is that rules of exponentials work, so for example,

$$zw = r_z e^{\varphi_z i} r_w e^{\varphi_w i} = r_z r_w e^{(\varphi_z + \varphi_w)i)}.$$

A.5 Polynomials

We deal here only with polynomials having real coefficients. Given $a_0, a_1, \ldots, a_m \in \mathbb{R}$ with $a_m \neq 0$, which we refer to as *coefficients*, a function, $p : \mathbb{C} \to \mathbb{C}$ such that,

$$p(z) = a_0 + a_1 z + a_2 z^2 + \ldots + a_m z^m,$$

is called a *polynomial* of the m'th *degree*. A number $\lambda \in \mathbb{C}$ is called a *root* of the polynomial if $p(\lambda) = 0$.

Theorem A.5.1. *Let $p(\cdot)$ be a polynomial of degree $m \geq 1$ and let $\lambda \in \mathbb{C}$. The λ is a root of $p(\cdot)$ if and only if there is a polynomial $q(\cdot)$ of degree $m - 1$ such that for all $z \in \mathbb{C}$,*

$$p(z) = (z - \lambda)q(z).$$

Further, any polynomial $p(\cdot)$ of degree $m \geq 0$ has at most m distinct roots.

Proof. □

Exercise A.5.2. *Take $a_0, a_1, \ldots, a_m \in \mathbb{C}$. Show that if for all $z \in \mathbb{C}$*

$$a_0 + a_1 z + a_2 z^2 + \ldots + a_m z^m = 0,$$

then,

$$a_0 = \ldots = a_m = 0.$$

The following is the *fundamental theorem of algebra* which we state without proof:

Theorem A.5.3. *Every non-constant polynomial $p(\cdot)$ has a root and may further be uniquely factorized (up to the order of the factors), as follows:*

$$p(z) = c(z - \lambda_1)^{k_1} \cdot \ldots \cdot (z - \lambda_n)^{k_n},$$

where $c \in \mathbb{R}$, $\lambda_1, \ldots, \lambda_n$ are distinct and $k_1, \ldots, k_n$ are positive integers.

Note: The above holds for polynomials with complex coefficients also. In that case $c \in \mathbb{C}$. In the case of real coefficients we further have:

Exercise A.5.4. *Show that if λ is a root of $p(z) = \sum_{i=0}^{m} a_i z^i$ then so is the conjugate $\bar{\lambda}$ taking the conjugate of both sides of $p(\lambda) = 0$.*

Theorem A.5.5. *Every non-constant polynomial $p(c)$ has q unique factorization (up to the order of the factors), as follows:*

$$p(x) = c \prod_{i=1}^{n_r} (x - \lambda_i)^{k_i} \prod_{i=1}^{n_c} (x^2 + \alpha_i x + \beta_i)^{K_i},$$

where $c \in \mathbb{R}$, $\lambda_1, \ldots, \lambda_{n_r} \in \mathbb{R}$ are distinct, $(\alpha_1, \beta_1), \ldots, (\alpha_{n_c}, \beta_{n_c}) \in \mathbb{R}^2$ are distinct, and $\alpha_i^2 - 4\beta_i < 0$ for $i = 1, \ldots, n_c$.

Proof. □

A.6 Vectors

A.6.1 Vectors as Tuples

Given a set $\mathcal{S}$, an *n-tuple* is an ordered sequence of elements $(\alpha_1, \ldots, \alpha_n)$ where $\alpha_i \in \mathcal{S}$. The set of all possible n-tuples is denoted $\mathcal{S}^n$, short for $\mathcal{S} \times \ldots \times \mathcal{S}$, n times. Some important examples are $\mathcal{S} = \mathbb{Z}$ or $\mathcal{S} = \mathbb{R}$, in those cases we call $\mathbb{Z}^n$ or $\mathbb{R}^n$ n-dimensional *vectors*. Observe that $\mathbb{Z}^n \subset \mathbb{R}^n$. We denote such vectors in **bold face**. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, we denote the i'th element of the vector by x_i, $i = 1, \ldots, n$.

A.6.2 Vector Operations

There are two basic operations on elements of $\mathbb{R}^n$: *vector addition* and *multiplication by scalars*. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$:

- **Vector addition**: $\boldsymbol{x} + \boldsymbol{y}$ is again an element of $\mathbb{R}^n$ with $(x + y)_i = x_i + y_i$. I.e. the i'th element of the sum is the sum of the i'th elements of a and b.
- **Multiplication by scalar**: $\alpha\boldsymbol{x}$ is again an element of $\mathbb{R}^n$ with $(\alpha x)_i = \alpha x_i$. I.e. the i 'th element of the multiplication by the scalar is the i'th element of a multiply by the scalar α.

Exercise A.6.1. *Refresh the geometric illustration of addition and multiplication of vectors in* $\mathbb{R}^2$.

A.6.3 More General Vectors

.

A.6.4 Eucleadiean Inner Products, Norms, Orthogonallity and Projections

Given two vectors, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the *inner product* of $\boldsymbol{x}$ and $\boldsymbol{y}$, denoted $\boldsymbol{x}'\boldsymbol{y}$ is,

$$< \boldsymbol{x}, \boldsymbol{y} >= \boldsymbol{x}'\boldsymbol{y} := \sum_{i=1}^{n} x_i y_i.$$

Exercise A.6.2. *Show that* $\boldsymbol{x}'\boldsymbol{y}$ *satisfies the following properties:*

1. $< \alpha\boldsymbol{x}, \boldsymbol{y} >= \alpha < \boldsymbol{x}, \boldsymbol{y} >$.
2. *QQQQ*

The (Euclidean) *norm* of $\boldsymbol{x}$, denoted $||\boldsymbol{x}||$, is,

$$||\boldsymbol{x}|| := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}'\boldsymbol{x}}.$$

Theorem A.6.3. *For* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$|\boldsymbol{x}'\boldsymbol{y}| \leq ||\boldsymbol{x}||\ ||\boldsymbol{y}||.$$

Proof. QQQQ. □

Exercise A.6.4. *Draw a vector $\boldsymbol{x} \in \mathbb{R}^2$ and illustrate Pythagoras theorem,*

$$||\boldsymbol{x}||^2 = x_1^2 + x_2^2.$$

Exercise A.6.5. *Show that $||\boldsymbol{x}||$ satisfies the following properties:*

1. *Homogeneity: $||\alpha\boldsymbol{x}|| = |\alpha|\,||\boldsymbol{x}||$.*
2. *Triangle inequality: $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$*
3. *Non-negativity: $||\boldsymbol{x}|| \geq 0$*
4. *Definiteness: $||\boldsymbol{x}|| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.*

Exercise A.6.6. *Draw two non-zero vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^2$. Let θ be the angle between the two vectors. Let θ_x be the angle between the horizontal axis and $\boldsymbol{x}$ and let θ_y be the angle between the horizontal axis and $\boldsymbol{y}$. Show:*

1. *$\sin\theta_x = x_2/||\boldsymbol{x}||$ and $\cos\theta_x = x_1/||\boldsymbol{x}||$*
2. *Use $\cos(\beta - \alpha) = \cos\beta\cos\alpha + \sin\beta\sin\alpha$ to show*

$$\cos\theta = \frac{\boldsymbol{x}'\boldsymbol{y}}{||\boldsymbol{x}||\;||\boldsymbol{y}||}.$$

3. *Use the above to show that,*

$$|\boldsymbol{x}'\boldsymbol{y}| \leq ||\boldsymbol{x}||\;||\boldsymbol{y}||.$$

The third item of the exercise above is called the Cauchy-Schwarz inequality (in that case for $\mathbb{R}^2$). Here is the more general case,

Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be *orthogonal* if $\boldsymbol{x}'\boldsymbol{y} = 0$. In geometrical terms this implies that they are perpendicular.

Exercise A.6.7.

1. *Show that all vectors in $\mathbb{R}^n$ are orthogonal to the zero vector, $\boldsymbol{0} \in \mathbb{R}^n$.*
2. *Describe (and try to draw) the set of vectors in $\mathbb{R}^3$ orthogonal to $(1,1,1)'$. Is this same set of vectors orthogonal to $(-2,-2,-2)'$.*

Two sets of vectors $\mathcal{V}, \mathcal{W} \subset \mathbb{R}^n$ are said to be *orthogonal* if every vector in $\mathcal{V}$ is orthogonal to every vector in $\mathcal{W}$. Given a set of vectors $\mathcal{V} \subset \mathbb{R}^n$, the *orthogonal complement* of $\mathcal{V}$, denoted $\mathcal{V}^\perp$ is the set of vectors in $\mathbb{R}^n$ orthogonal to $\mathcal{V}$:

$$\mathcal{V}^\perp := \{\boldsymbol{x} \in \mathbb{R}^n \;:\; \boldsymbol{x}'\boldsymbol{v} = 0,\ \forall \boldsymbol{v} \in \mathcal{V}\}.$$

We now come to the concept of *projecting* a vector $\boldsymbol{x} \in \mathbb{R}^n$ onto a vector $\boldsymbol{y} \in \mathbb{R}^n$. This *projection* is given by,

$$\boldsymbol{p} = \frac{\boldsymbol{y}'\boldsymbol{x}}{\boldsymbol{y}'\boldsymbol{y}}\boldsymbol{y} = \frac{\boldsymbol{y}'\boldsymbol{x}}{||\boldsymbol{y}||^2}\boldsymbol{y}.$$

Exercise A.6.8. *Draw the projection of some $\boldsymbol{x} \in \mathbb{R}^2$ onto some $\boldsymbol{y} \in \mathbb{R}^2$.*

A.7 Matrices

An *m by n matrix* is a rectangular array with m rows and n columns appearing as such,

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}.$$

We typically write matrices using upper case letters, and treat the i, j'th element with the corresponding lower case subscripted by i and j. The set of all $m \times n$ matrices with real elements (typically our case) is $\mathbb{R}^{m \times n}$. An important class of matrices is that of *square matrices*, $\mathbb{R}^{n \times n}$.

It is often extremely useful to *block-partition* matrices. For example for an $m \times n$ matrix A, we can write,

$$A = \left[\begin{array}{c|c} A_{1,1} & A_{1,2} \\ \hline A_{2,1} & A_{2,2} \end{array}\right],$$

where the dimensions of the sub-matrices are as follows: $A_{1,1}$ is $p \times q$, $A_{1,2}$ is $p \times (n-q)$, $A_{2,1}$ is $(m-p) \times q$ and $A_{2,2}$ is $(m-p) \times (n-q)$. We can take $p \in \{0, \ldots, m\}$ and $q \in \{0, \ldots, n\}$ so that if $p = 0$ or $p = m$ then two of the sub-matrices are empty and similarly if $q = 0$ or $q = n$.

Block partitioning can also be done with more than two blocks.

We can sometimes treat vectors as special cases of matrices of dimension $n \times 1$ in which case we call the vector a *column vector* or dimension $1 \times n$ in which case we call the vector a *row vector*. Unless otherwise specified, vectors are to be treated as column vectors by default. Some important useful vectors/matrices are:

- The *identity matrix*, denoted I, or I_n if the dimension is not clear from context and we wish to specify it is an element of $\mathbb{R}^{n \times n}$. The elements of this matrix are $\delta_{i,j}$ where $\delta_{i,j}$ is the *Kronecker delta*, equaling 1 if $i = j$ and 0 if $i \neq j$.

- The *zero matrix/vector*. All elements of this matrix/vector are 0 scalars. Again, if the dimension is not clear from context we can specify it by 0_n or $0_{n \times m}$.

- The *identity vector*. This vector, denoted by $\mathbf{1}$ has all elements equal to the scalar 1 (some texts denote $\mathbf{1}$ by $\mathbf{e}$ – we don't).

- The canonical basis vectors, e_i, $i = 1, \ldots, n$. These values (coordinates) of the vector e_i are $\delta_{i,j}$. Again here the dimension should be understood from context. So for example e_2 in the context of $\mathbb{R}^2$ is not e_2 in the context of $\mathbb{R}^3$

A.7.1 Operations on Matrices

The operations of vector addition and multiplication by scalars that we defined for vectors, carry over directly to matrices. Just as we can only add two vectors of the same dimension, we can only *add two matrices* of the same dimension.

Exercise A.7.1. *Write out the matrix* $2A - 3I$*, where*

$$A := \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}.$$

It turns out to be extremely useful to define a further operation on matrices, *matrix multiplication*. Specifically, if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ then $C = AB$ is a matrix element of $\mathbb{R}^{m\times p}$ with,

$$C_{i,j} = \sum_{k=1}^{n} A_{i,k}B_{k,j}.$$

Exercise A.7.2. *In addition to* A *defined in the previous exercise, define,*

$$B := \begin{bmatrix} 1 & 2 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

Is the product BA *defined? Is the product* AB *defined? If so, what is it?*

Exercise A.7.3. *For any* $A \in \mathbb{R}^{n\times m}$ *show that,* $I_n A = A$ *and* $AI_m = A$*.*

The *transpose* of a matrix $A \in \mathbb{R}^{n\times m}$ is a matrix $B \in \mathbb{R}^{m\times n}$ such that $B_{i,j} = A_{j,i}$. The transpose is denoted A'.

Exercise A.7.4. *Prove that*

1. $I' = I$.
2. $(A + B)' = A' + B'$.
3. $(AB)' = B'A'$.

Of the many uses of transpose, one important one is converting column vectors into row vectors and visa versa. Thus in text we may often define $a = (a_1, \ldots, a_n)'$ implying that a is a column vector since it is a transpose of the row $(a_1, \ldots, a_n)$. Further, while multiplication of two vectors in $\mathbb{R}^n$ is not defined, if we treat vectors as matrices the multiplication is defined. Specifically take two (treated as column) vectors $a, b \in \mathbb{R}^n$. Then,

$$a'b = \sum_{i=1}^{n} a_i b_i \in \mathbb{R}^1,$$

is called the *inner product* of a and b. Further,

$$a\,b' = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{bmatrix} \in \mathbb{R}^{n\times n},$$

is called the *outer product* of a and b.

Exercise A.7.5. *1. Show that $a'b = b'a$.*

2. *Express the operation of matrix multiplication in terms of inner products of rows and columns.*

3. *Is it true that $ab' = ba'$?*

4. *We define $e_{i,j} \in \mathbb{R}^{n\times n}$ as the matrix e_ie_j', write the elements of this matrix in terms of the Kronecker delta.*

Of further use is a matrix of the form $G = A'A$, where A is some arbitrary $n \times m$ matrix. This is called the *Gram Matrix* (or *Gramian*).

Exercise A.7.6. *Deduce the following simple properties of $G = A'A$:*

1. *It is symmetric.*

2. *The entries $G_{i,j}$ of the Gram matrix are inner products.*

Matrix multiplication can have several meanings that are useful to consider:

Exercise A.7.7. *Consider and describe in words the following interpretations of matrix multiplication for $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times m}$ and $x \in \mathbb{R}^n$:*

1. *The vector $\mathbf{y} = A\mathbf{x}$ can be represented as $y = \sum_{j=1}^{n} x_j\mathbf{a}_j$ where $\mathbf{a}_j$ is the j'th column of A.*

2. *The elements of $C = AB$ are the inner products, $c_{i,j} = \tilde{\mathbf{a}}_i'\mathbf{b}_j$, where $\tilde{\mathbf{a}}_i'$ is the i'th row of A and $\mathbf{b}_j$ is the j'th column of B.*

3. *The elements of $\mathbf{y} = C\mathbf{x}$ are the inner products $\mathbf{y}_i = \tilde{\mathbf{a}}_i'\mathbf{x}$.*

4. *If we use the matrices to define functions $f_A(\mathbf{u}) = A\mathbf{u}$ and $f_B(\mathbf{u}) = B\mathbf{u}$, where $f_A : \mathbb{R}^n \to \mathbb{R}^m$ and $f_B : \mathbb{R}^m \to \mathbb{R}^p$, then the composed function, $g(\mathbf{x}) = f_A\big(f_B(\mathbf{x})\big)$ can be represented as, $g(\mathbf{x}) = C\mathbf{x}$ where $C = AB$.*

5. *The columns $\mathbf{c}_1, \ldots, \mathbf{c}_p$ of the product $C = AB$ are, $\mathbf{c}_j = A\mathbf{b}_j$.*

6. *Similarly, the rows of C, $\tilde{\mathbf{c}}'_1, \ldots, \tilde{\mathbf{c}}'_m$ are $\tilde{\mathbf{c}}'_i = \tilde{\mathbf{a}}'_i B$.*

A matrix $A \in \mathbb{R}^{n\times n}$ is *invertible* or *non-singular* if there exists a matrix $B \in \mathbb{R}^{n\times n}$ such that $AB = I$. In this case, B is unique and is called the inverse of A and denoted A^{-1}. Further $BA = I$.

Exercise A.7.8. *Prove that:*

1. *A^{-1} is unique.*
2. *The inverse of A^{-1} is A.*
3. *In general, the inverse of $A + C$ is not $A^{-1} + C^{-1}$.*
4. *$(AB)^{-1} = B^{-1}A^{-1}$.*

Exercise A.7.9. *A diagonal matrix $A \in \mathbb{R}^{n\times n}$ is a matrix with $A_{i,j} = 0$ if $i \neq j$. Show that diagonal matrixes are invertible if and only if $a_{i,i} \neq 0$ for all i, and find A^{-1} in such cases.*

The *transpose* of a matrix $A \in \mathbb{R}^{n\times m}$ is a matrix $B \in \mathbb{R}^{m\times n}$ such that $B_{i,j} = A_{j,i}$. The transpose is denoted A'.

Exercise A.7.10. *Prove that,*

1. *$(AB)' = B'A'$*
2. *For square matrices $(A')^{-1} = (A^{-1})'$*

For a matrix $A \in \mathbb{R}^{n\times n}$, and a non-negative integer k, the *matrix power*, A^k is defined as the product $AA \cdot \ldots \cdot A$, k times. If $k = 0$ it is the identity matrix.

A.7.2 Kronecker Products and Sums

Given a matrix $A \in \mathbb{R}^{m\times n}$ and a matrix $B \in \mathbb{R}^{p\times q}$ the *Kronecker product* of A and B, denoted $A \otimes B$ is a matrix of dimension $mp \times nq$ written in block form as follows:

$$A \otimes B = \begin{bmatrix} a_{1,1}B & a_{1,2}B & \cdots & a_{1,n}B \\ a_{2,1}B & a_{2,2}B & \cdots & a_{2,n}B \\ \vdots & \vdots & & \vdots \\ a_{m,1}B & a_{m,2}B & \cdots & a_{m,n}B \end{bmatrix}.$$

Exercise A.7.11. 1. *If B is a scalar show that $A \otimes B = B\,A$.*

2. *If $A = I$ what is the form of $A \otimes B$?*
3. *If $B = I$ what is the form of $A \otimes B$?*

A.8 Complex Vectors and Matrices

In most of this book we treat vectors and matrices as being defined over the field of real numbers. Nevertheless, it is sometimes useful to consider vectors and matrices over complex numbers. Namely $\boldsymbol{u} \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n \times n}$ respectively. Now denote $\overline{\boldsymbol{u}}$ and $\overline{A}$ as the complex conjugate vector and matrix respectively. That is, in these objects the elements are complex conjugates of $\boldsymbol{u}$ and $\boldsymbol{A}$.

Exercise A.8.1. *Show that:*

1. $\overline{A\,\boldsymbol{u}} = \overline{A}\,\overline{\boldsymbol{u}}$.
2. $\overline{A}' = \overline{A'}$.

A.9 Derivatives and Continuity

We now present some basic results from real (multi-variate) analysis. We skip the basic definitions of continuity and derivatives, these can be obtained from any calculus source. But we present some further definitions and properties.

Consider a function $f : \mathbb{R} \to \mathbb{R}^n$.

The function is said to be *Lipschitz continuous* if there exists a $K \geq 0$ such that for any x, $y \in \mathbb{R}$,

$$||f(x) - f(y)|| \leq K\,|x - y|.$$

Bibliographic Remarks

Exercises

Appendix B

Linear Algebra Basics

B.1 Vector Spaces in $\mathbb{R}^n$ and Their Bases

B.1.1 General Vector Spaces

A subset $\mathcal{V}$ of $\mathbb{R}^n$ is a *vector space in* $\mathbb{R}^n$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$ and $\alpha \in \mathbb{R}$,

$$\boldsymbol{x} + \boldsymbol{y} \in \mathcal{V}, \qquad \alpha \boldsymbol{x} \in \mathcal{V}.$$

Exercise B.1.1. *Show the following:*

1. *If $\mathcal{V}$ is a vector space in $\mathbb{R}^n$ then the vector $\mathbf{0}_n \in \mathcal{V}$.*
2. *The vector spaces in $\mathbb{R}^1$ are either $\{0\}$ or $\mathbb{R}$.*
3. *The vector spaces in $\mathbb{R}^2$ are either $\{0\}$, lines passing through the origin, or $\mathbb{R}^2$.*
4. *The vector spaces in $\mathbb{R}^3$ are either $\{0\}$, lines (in 3D space) passing through the origin, planes passing through the origin, or $\mathbb{R}^3$.*

A *subspace* of a vector space $\mathcal{V}$ is a set, $\tilde{\mathcal{V}} \subset \mathcal{V}$ that is also a vector space.

In general, *Linear Algebra* is the study of vector spaces, related sets and their properties. This is not just for vector spaces in $\mathbb{R}^n$ as defined above but also for general vector spaces either over F^n (where F is some *field* – a set with together with two operations such as addition and multiplication and with a zero element such as the real 0 and a one element such as the real 1), or where elements of the vector space $\mathcal{V}$ are other types of objects (e.g. functions). We do not take this more general abstract approach here, nevertheless we comment that for $\mathcal{V}$ to be a vector spaces there needs to be an operation of *addition* defined over elements of $\mathcal{V}$ as well as an operation of *scalar multiplication*. With these two operations at hand, the set $\mathcal{V}$ is said to be a *vector space* with an associated *scalar field*, F, if the following properties hold:

- *commutativity*: $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$
- *associativity*: $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z})$
- *existence of additive identity*: There exists, $\tilde{0} \in \mathcal{V}$ such that $\boldsymbol{x} + \tilde{0} = \boldsymbol{x}$.
- *existence of additive inverse*: There exists, $\tilde{\boldsymbol{x}} \in \mathcal{V}$ such that $\boldsymbol{x} + \tilde{\boldsymbol{x}} = \tilde{0}$.
- *scalar multiplicative identity*: The scalar 1 (from the associated scalar field) has the property: $1\boldsymbol{x} = \boldsymbol{x}$.
- *scalar distributive properties*: For a scalars α, β, $\alpha(\boldsymbol{x} + \boldsymbol{y}) = \alpha\boldsymbol{x} + \alpha\boldsymbol{y}$, $(\alpha + \beta)\boldsymbol{x} = \alpha\boldsymbol{x} + \beta\boldsymbol{x}$.

Except for the exercise below, we will not be concerned with general vector spaces (only with vectors spaces in $\mathbb{R}^n$):

Exercise B.1.2. *Consider the set of polynomials as a vector space.*

1. *How would you define vector addition?*
2. *How would you define scalar multiplication?*
3. *Show that the set of of polynomials is a vector space based on your definitions in 1 and 2 above.*

There are many standard examples of vector spaces and we do not spend too much time on those. But here is one: Consider,

$$\mathcal{V} = \{x : \mathbb{R}_+ \to \mathbb{R}^n \mid x \text{ is differentiable}\},$$

where the vector sum is the sum of functions,

$$(x + z)(t) = x(t) + z(t),$$

and the scalar multiplication is defined by,

$$(\alpha x)(t) = \alpha\, x(t).$$

Exercise B.1.3. *Show that, $\tilde{\mathcal{V}} = \{x \in \mathcal{V} \mid \dot{x} = Ax\}$ is a subspace. That is x are solutions of a linear autonomous system.*

B.1.2 Linear Combinations and Span

A *linear combination* of a set of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^n$ and a set of scalars $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, is,

$$\alpha_1 \boldsymbol{x}_1 + \ldots \alpha_m \boldsymbol{x}_m.$$

Exercise B.1.4. *This exercise deals with the several ways of looking at matrix multiplication in terms of linear combinations.*

1. *Take $A \in \mathbb{R}^{n\times m}$ and $a \in \mathbb{R}^m$. Express the elements of A a in terms of linear combinations of columns of A.*
2. *Take $A \in \mathbb{R}^{n\times m}$ and $a \in \mathbb{R}^n$ treated now as a row vector. Express the elements of aA in terms of linear combinations of rows of A.*
3. *Take now...(do it for matrix multiplication).*

The *span* of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^n$ is the set of all linear combinations:

$$\text{span}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) := \{\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i \ : \ \alpha_1, \ldots, \alpha_m \in \mathbb{R}\}.$$

The vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are said to be *linearly independent* is the only choice of $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ that satisfies,

$$\alpha_1 \boldsymbol{x}_1 + \ldots \alpha_m \boldsymbol{x}_m = 0_n,$$

is $\alpha_1 = \ldots = \alpha_m = 0$, otherwise we say the vectors are *linearly dependent.*

Theorem B.1.5. *The span of the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$ is a vector space (subspace) in $\mathbb{R}^n$*

Proof. □

Exercise B.1.6. *Prove that if $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are linearly independent, then for every $\tilde{\boldsymbol{x}} \in$ span$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ there is a unique representation,*

$$\alpha_1 \boldsymbol{x}_1 + \ldots + \alpha_m \boldsymbol{x}_m = \tilde{\boldsymbol{x}},$$

i.e. the scalars $\alpha_1, \ldots, \alpha_m$ are unique.

Theorem B.1.7. *If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ is linearly independent in V and $\boldsymbol{x}_1 \neq 0$ then there exists $j \in \{2, \ldots, m\}$ such that $a_j \in$ span$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1})$. Further if the j'th vector is removed from $_1, \ldots, a_m$ the span of the remaining $m-1$ vectors equals span$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$.*

Theorem B.1.8. *Take a vector space $\mathcal{V}$ in $\mathbb{R}^n$, with $\mathcal{V} =$ span$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$. Assume $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \in \mathbb{R}^n$ are linearly independent, with,*

$$\alpha_1 \boldsymbol{y} + \ldots + \alpha_k \boldsymbol{y} \in \mathcal{V}.$$

Then $k \leq m \leq n$.

B.1.3 Basis and Dimension

A vector space $\mathcal{V}$ is said to be *finite dimensional* if it equals the span of some finite set of vectors.

Exercise B.1.9. *Prove that all vector spaces in $\mathbb{R}^n$ are finite-dimensional. E.g. use the canonical vectors $e_1, \ldots, e_n \in \mathbb{R}^n$.*

If a set of vector $\mathcal{B} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is a linearly independent set of vectors and $\text{span}(\mathcal{B}) = \mathcal{V}$ then $\mathcal{B}$ is said to be a *basis* of $\mathcal{V}$.

Take a vector space $\mathcal{V}$ and a basis of the vector space, $\mathcal{B}$. The number of elements in $\mathcal{B}$ is refereed to as $\dim \mathcal{V}$ – the *dimensions* of the vector space. If there is no such (finite) basis $\mathcal{B}$ we denote, $\dim \mathcal{V} = \infty$. We have the following

Theorem B.1.10. *All bases of the vector space have the same number of elements hence $\dim \mathcal{V}$ is well defined.*

B.2 Linear Transformations and Systems of Equations

Let $\mathcal{V}$ and $\mathcal{W}$ be two vector spaces with, $L : \mathcal{V} \to \mathcal{W}$ a function with domain $\mathcal{V}$ and range $\mathcal{W}$. If for every $v_1, v_2 \in \mathcal{V}$ and a scalar c we have,

$$L(v_1 + v_2) = L(v_1) + L(v_2) \quad \text{and} \quad L(cv_1) = cL(v_2),$$

then $L(\cdot)$ is a called a *linear transformation* (this is sometimes called a *linear map*).

Exercise B.2.1.

1. *Show that in the case $\mathcal{V}, \mathcal{W} = \mathbb{R}$ the only linear transformations are the functions $L(x) = kx$, $k \in \mathbb{R}$.*
2. *Show that in the case $\mathcal{V} = \mathbb{R}^n$ and $\mathcal{W} = \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, then $L(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$ is a linear transformation if and only if $\boldsymbol{b} = \boldsymbol{0}$.*

B.2.1 The Matrix of a Linear Transformation

Let $L : \mathcal{V} \to \mathcal{W}$ be a linear transformation between finite dimensional vector spaces. Let $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ be a basis for $\mathcal{V}$ and $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$ be a basis for $\mathcal{W}$. Given these bases, we have the following.

Theorem B.2.2. *The linear transformation $L(\cdot)$ operates on $\boldsymbol{v} \in \mathcal{V}$ by the matrix multiplication $M_L\boldsymbol{v}$, where given the bases $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ and $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)$ for $\mathcal{V}$ and $\mathcal{W}$ respectively, the $m \times n$ matrix $M_L = [a_{i,j}]$ has unique $a_{i,j}$ defined by*

$$L(\boldsymbol{v}_j) = \sum_{i=1}^{m} a_{i,j}\boldsymbol{w}_i.$$

Proof. □

Exercise B.2.3. *Given two linear transformations L_1 and L_2 with the same domain and range and given a scalar c, show the following:*

1. $M_{L_1+L_2} = M_{L_1} + M_{L_2}$
2. $M_{cL_1} = cM_{L_1}$
3. $M_{L_1L_2} = M_{L_1}M_{L_2}$

B.2.2 Null Spaces and Ranges

For a linear transformation, $L : \mathcal{V} \to \mathcal{W}$ (or alternately for a matrix $A \in \mathbb{R}^{n\times m}$ with $\mathcal{V} = \mathbb{R}^n$ and $\mathcal{W} = \mathbb{R}^m$), there are two basic sets (which happen to be subspaces of $\mathcal{V}$ and $\mathcal{W}$) associated with L. These are called the *null space* and the *range*.
The *null space* of L, denoted null L is the set of vectors of $\mathcal{V}$ that are mapped to $0 \in \mathcal{W}$:

$$\text{null } L := \{v \in \mathcal{V} \ : \ L(v) = 0\} \quad \text{or,} \quad \text{null } A := \{x \in \mathbb{R}^n \ : \ Ax = 0_m\}.$$

That is, null A is the set of vectors orthogonal to all rows of A. The following exercise shows that null A gives the "ambiguity" in the (transformation) A:

Exercise B.2.4. *Show the following an explain why it implies that null A implies the ambiguity in A:*

1. $0 \in$ *null* A.
2. *If* $y = Ax$ *and* $z \in$ *null* A *then* $y = A(x + z)$.
3. *If* $y = Ax$ *and* $y = A\tilde{x}$ *then* $\tilde{x} = x + z$ *for some* $z \in$ *null* A.

Exercise B.2.5. *Assume that null* $A = \{0\}$. *Show that,*

1. x *can always be uniquely determined from* $y = Ax$ *(i.e. the mapping* $f(x) = Ax$ *is one-to-one.*
2. *The columns of* A *are independent.*
3. A *has a left inverse...QQ*
4. $\det(A'A) \neq 0$.

The *range* of L, denoted range L is the image of L in $\mathcal{W}$. That is,

range $L := \{w \in \mathcal{W} \ : \ \exists v \in \mathcal{V},\ L(v) = w\}$.or, $\quad$ range $A := \{y \in \mathbb{R}^m \ : \ \exists x \in \mathbb{R}^n,\ Ax = y\}$.

Theorem B.2.6. *For a linear transformation, $L : \mathcal{V} \to \mathcal{W}$, we have that null L is a subspace of $\mathcal{V}$ and range L is a subspace of $\mathcal{W}$.*

Proof. QQQQ. □

Theorem B.2.7. *A linear transformation, L, is injective*[1] *if and only if null $L = \{0\}$.*

Proof. QQQQ. □

A key result is the following:

Theorem B.2.8. *If $\mathcal{V}$ is finite dimensional and $L : \mathcal{V} \to \mathcal{W}$ is a linear transformation then range L is finite dimensional and,*

$$dim\ \mathcal{V} = dim\ null\ L\ +\ dim\ range\ L.$$

Proof. QQQQ. □

The following is an important corollaries:

Corollary B.2.9. *Assume $\mathcal{V}$ and $\mathcal{W}$ are finite-dimensional. Then,*

(i) If dim $\mathcal{V}$ > dim $\mathcal{W}$ then no linear map from $\mathcal{V}$ to $\mathcal{W}$ is injective.

(ii) If dim $\mathcal{V}$ < dim $\mathcal{W}$ then no linear map from $\mathcal{V}$ to $\mathcal{W}$ is surjective[2]*.*

B.2.3 Invertibility

A linear transformation $L : \mathcal{V} \to \mathcal{W}$ is called *invertible* if there exists a linear transformation $L^{-1} : \mathcal{W} \to \mathcal{V}$ such that LL^{-1} is the identity map on $\mathcal{V}$.

B.2.4 The Four Fundamental Subspaces of a Matrix

For $A \in \mathbb{R}^{n \times m}$ we describe four subspaces associated with A. These are (1) The *row space of A* which is range A', a subspace of $\mathbb{R}^m$. (2) The *nullspace of A* which is null A, a subspace of $\mathbb{R}^n$. (3) The *column space of A* which is range A, a subspace of $\mathbb{R}^n$. (4) The *null space of A'* which is null A', a subspace of $\mathbb{R}^m$. These four subspaces are related through what is sometimes called *the fundamental theorem of linear algebra*:

[1] A function, $L : \mathcal{V} \to \mathcal{W}$ is *injective* (also known as *one-to-one*) if for every $u, v \in \mathcal{V}$ such that $L(u) = L(v)$, we have that $u = v$

[2] QQQQ

Theorem B.2.10. *Given that dim range $A' = r$ (the dimension of the row space of A is r) then:*

$$\textit{dim null } A = n - r, \qquad \textit{dim range } A = r, \qquad \textit{dim null } A' = m - r.$$

and further,

$$\textit{null } A = \big(\textit{range } A'\big)^{\perp}, \qquad \textit{range } A' = \big(\textit{null } A\big)^{\perp},$$

and,

$$\textit{null } A' = \big(\textit{range } A\big)^{\perp}, \qquad \textit{range } A = \big(\textit{null } A'\big)^{\perp}.$$

Proof. QQQQ. □

B.2.5 Left and Right Inverses

B.2.6 Linear Equations

Consider now the *homogeneous linear system of equations*,

$$A\boldsymbol{x} = \mathbf{0},$$

where $A \in \mathbb{R}^{m\times n}$ and thus $\boldsymbol{x} \in \mathbb{R}^n$ and $\mathbf{0} \in \mathbb{R}^m$. Here we think of the elements, $a_{i,j}$ of A as known and of the elements of $\boldsymbol{x}$ as unknown.

With this system we can associate the linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$:

$$L(x_1, \ldots, x_n) = \big(\sum_{k=1}^{n} a_{1,k}x_k, \ldots, \sum_{k=1}^{n} a_{m,k}x_k\big),$$

where $a_{i,j}$ are elements of A.

Exercise B.2.11. *Show that L is a linear transformation.*

Obviously $(x_1, \ldots, x_n)' = (0, \ldots, 0)'$ is a solution. Are there other solutions? This question is the same as asking if null L is bigger than $\{0\}$. From Theorem **??** this happens precisely when L is not injective. Then from Corollary **??** (i) this happens if $n > m$. Hence if there are more unknowns than equations there must be non-zero solutions.

Move now to the *non-homogeneous linear system of equations*,

$$A\boldsymbol{x} = \boldsymbol{a},$$

where $\boldsymbol{a} \in \mathbb{R}^m$ is considered a known vector.

Exercise B.2.12. *Show that if $n < m$ (there are more equations than unknowns) then there is no solution for some choice of $\boldsymbol{a}$.*

B.2.7 Orthogonallity

Two subspaces $\mathcal{V}$ and $\mathcal{W}$ of $\mathbb{R}^n$ are called *orthogonal* if every vector $v \in \mathcal{V}$ is orthogonal to every vector $w \in \mathcal{W}$.

B.3 Determinants

Determinants play a central role in linear algebra. Here is an axiomatic definition of the determinant of a matrix: A function $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ is said to be a determinant if:

1. It is linear as a function of each of the rows of the matrix.
2. It gets the value 0 for any matrix having two equal rows.
3. It assigns 1 to the identity matrix.

We state the following without proof:

Theorem B.3.1. *Each matrix, $A \in \mathbb{R}^{n\times n}$ has a unique determinant,* $\det(A)$ *computed as follows:*

$$\det(A) = \sum_{p\in perm(1,\ldots,n)} sign(p) A_{1,p(1)} \cdot A_{2,p(2)} \cdot \ldots \cdot A_{n,p(n)} = \sum_{p\in perm(1,\ldots,n)} sign(p) A_{p(1),1} \cdot A_{p(2),2} \cdot \ldots \cdot A_{p(n),n},$$

where $perm(1,\ldots,n)$ is the set of all permutations of $\{1,\ldots,n\}$ where for a permutation, p, $p(i)$ is the i'th element of the permeation and $sign(p)$ is $+1$ if the number of transpositions between p and the identity permeation $(1,\ldots,n)$ is even and -1 if that number is odd.

Consider for example the 2×2 case,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Here $perm(1,2) = \{(1,2),(2,1)\}$ so,

$$\det(A) = sign(1,2) A_{1,1} A_{2,2} + sign(2,1) A_{1,2} A_{2,1} = ad - bc.$$

Alternatively (using the second sum of Theorem B.3.1):

$$\det(A) = sign(1,2) A_{1,1} A_{2,2} + sign(2,1) A_{2,1} A_{1,2} = ad - bc.$$

Exercise B.3.2. *Show that*

$$\det\left(\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \right) = \begin{array}{l} a_{1,1}a_{2,2}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2} \\ -a_{1,3}a_{2,2}a_{3,1} - a_{1,2}a_{2,1}a_{3,3} - a_{1,1}a_{2,3}a_{3,2}. \end{array}$$

Here are some elementary properties of the determinant (for conveinece we also include the defining properties (mentioned above):

Theorem B.3.3. *The following hold:*

1. *Linear function of the rows.*
2. *The determinant changes sign when rows are exchanged.*
3. *The determinant is* 0 *if there are two equal rows.*
4. $\det(I) = 1$.
5. *Subtraction a multiple of one row from another leaves the determinant unchanged.*
6. *If there is a* 0 *row then the determinant is* 0.
7. *For a triangular matrix the determinant is the product of the diagonal entries.*
8. *For* $A, B \in \mathbb{R}^{n\times n}$, $\det(AB) = \det(A)\,\det(B)$.
9. $\det(A') = \det(A)$.

Proof. Prove some of the properties leaving others as exercises. □

Our key use of determinants is in the following.

Theorem B.3.4. *For* $A \in \mathbb{R}^{n\times n}$ *we have* $rank(A) = n$ *if and only if* $\det(A) \neq 0$. *That is* A^{-1} *exists if and only if* $\det(A) \neq 0$.

This is particularly useful when looking at families of matrices, say parameterized by a complex value λ: $\{\tilde{A}(\lambda) \ : \ \lambda \in \mathbb{C}\}$. In this case the solutions of the equation (in λ):

$$det\big(\tilde{A}(\lambda)\big) = 0,$$

are exactly the $\lambda \in \mathbb{C}$ for which $A(\lambda)$ is singular.

Proof. Prove that $det = 0$ iff not full rank. □

B.3.1 Minors, Cofactors, Adjugate Matrix and Cramer's Rule

Given $A \in \mathbb{R}^{n\times n}$, the sub matrix $M_{i,j} \in \mathbb{R}^{(n-1)\times(n-1)}$ called the *minor* is formed by throwing away the i'th row and j'th column of A. Then the (i, j) *cofactor* is the determinant of of the minor with a possible sign adjustment:

$$C(i, j) = (-1)^{i+j} \det(M_{i,j}).$$

Theorem B.3.5. *An alternative calculation of* $\det(A)$*,*

$$\det(A) = A_{i,1}C(i,1) + A_{i,2}C(i,2) + \ldots + A_{i,n}C(i,n),$$

where i *is some row* $i \in \{1, \ldots, n\}$*.*

The *adjugate matrix* of the matrix A is the matrix whose (i,j) is the (j,i)'th cofactor of A:

$$\mathrm{adj}A = \begin{bmatrix} C(1,1) & C(2,1) & \cdots & C(n,1) \\ C(1,2) & C(2,2) & \cdots & C(n,2) \\ \vdots & \vdots & & \vdots \\ C(1,n) & C(2,n) & \cdots & C(n,n) \end{bmatrix}.$$

We now have,

Theorem B.3.6. *For any* $A \in \mathbb{R}^{n\times n}$*,*

$$A\ adjA = \det(A)I,$$

and thus if $\det(A) \neq 0$*,*

$$A^{-1} := \frac{1}{\det(A)}\, adjA.$$

Further consider the system of equations $Ax = b$ *and assume* $\det(A) \neq 0$*. Then the* j*'th component of* $x = A^{-1}b$ *is,*

$$x_j = \frac{1}{\det(A)} \det\Big(\begin{bmatrix} A_{1,1} & \cdots & A_{1,j-1} & b_1 & A_{1,j+1} & \cdots & A_{1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ A_{1,1} & \cdots & A_{1,j-1} & b_n & A_{1,j+1} & \cdots & A_{1,n} \end{bmatrix} \Big).$$

The last result is called *Cramer's rule.*

Proof. QQQQ □

B.4 The Characteristic Polynomials

Consider the expression determinant, $\det\big(A - \lambda I\big)$.

Exercise B.4.1. *Show that* $\det\big(A - sI\big)$ *can be represented as the polynomial,*

$$p(s) = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \ldots + \alpha_n s^n.$$

The expression $p(s)$ is called the *characteristic polynomial* of the matrix A, and the equation (in λ),

$$p(s) = 0,$$

is called the *characteristic equation.* Note that we may take this equation with s being a complex valued scalar and then the right hand side is the scalar 0. Similarly, we can take it with s being a square matrix in which case the right hand side is the the 0 matrix.

Exercise B.4.2. *Consider the 2 by 2 matrix and write it's characteristic equation.*

B.5 Eigenvalues, Eigenvectors and Characteristic Polynomials

Given a matrix $A \in \mathbb{R}^{n\times n}$, a scalar $\lambda \in \mathbb{C}$ and a vector $\boldsymbol{v} \neq \mathbf{0}$, we say that λ is an *eigenvalue* with corresponding *eigenvector* v if,

$$Av = \lambda \boldsymbol{v},$$

or alternatively,

$$(A - \lambda I)\boldsymbol{v} = \mathbf{0}. \tag{B.1}$$

Since λ is an eigenvalue only if there is some corresponding eigenvector $\boldsymbol{v}$, the matrix, $A - \lambda I$ must be singular in order for λ to be an eigenvalue. To see this, assume that it is non-singular, in this case its null-space contains only $\mathbf{0}$. I.e. the only solution to (B.1) is $\boldsymbol{v} = \mathbf{0}$.
This reasoning also equips us with a way of finding all of the eigenvalues of A and for each one describing its eigenvectors:

- To find the eigenvalues of A, solve,

  $$\det\big(A - \lambda I\big) = 0,$$

  It is exactly for these values, λ that $A - \lambda I$ is singular and has a non-trivial null-space

- Given some eigenvalue λ, all of its corresponding eigenvectors are the vectors in the null-space of $A - \lambda I$ excluding $\mathbf{0}$.

Exercise B.5.1. *Show that if $\boldsymbol{v}$ is an eigenvector then so is $\alpha\boldsymbol{v}$ for any scalar $\alpha \neq 0$.*

Theorem B.5.2. *Every $A \in \mathbb{R}^{n\times n}$ has an eigenvalue.*

Theorem B.5.3. *Eigenvectors corresponding to distinct eigevalues are linearly independent.*

Exercise B.5.4. *Prove that the number of distinct eigenvalues of $A \in \mathbb{R}^{n\times n}$ is at most n.*

B.6 Some Eigenvalue Propeties

The *trace* of a square matrix A is the sum of its diagonal elements, denoted, $\mathrm{Tr}(A)$. This is a key property.

Exercise B.6.1. *Prove that for $A, B \in \mathbb{R}^{n\times n}$, $Tr(AB) = Tr(BA)$.*

Theorem B.6.2. *Let $A \in \mathbb{R}^{n\times n}$ have eigenvalues $\lambda_1, \ldots, \lambda_n$ then the following properties hold:*

1. $\sum_{i=1}^{n} \lambda_i = Tr(A)$.
2. $\prod_{i=1}^{n} \lambda_i = Det(A)$.
3. *If A^{-1} exists then it's eigenvalues are $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$.*
4. *The eigenvalues of A' are $\lambda_1, \ldots, \lambda_n$.*
5. *For $\alpha, \beta \in \mathbb{R}$ and $k \in \mathbb{Z}_+$, the eigenvalues of $(\alpha A + \beta I)^k$ are $(\alpha\lambda_1 + \beta)^k, \ldots, (\alpha\lambda_n + \beta)^k$.*

Bibliographic Remarks

Exercises

Appendix C

Further Linear Algebra

C.1 Properties of Symmetric Matrices

A matrix is symmetric if $A' = A$.

Proposition C.1.1. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix then:*

1. *The eigenvalues of A are real.*
2. *Eigenvectors of A corresponding to different eigenvectors are orthogonal.*

Proof. Let $A \in \mathbb{R}^{n\times n}$. We have $\lambda u = Au$ where λ is an eigenvalue and $u \neq 0$. Now premultiply with $\overline{u}'$ to get,

$$\lambda \overline{u}'u = \overline{u}'Au = (A'\overline{u})'u = (A\overline{u})'u = (\overline{A}\overline{u})'u = (\overline{\lambda}\overline{u})'u = \overline{\lambda}\overline{u}'u.$$

Hence $(\lambda - \overline{\lambda})\overline{u}'u = 0$. But since $u \neq 0$, we have that $\overline{u}'u > 0$ hence $\lambda = \overline{\lambda}$ so λ is real. Now let $Au_1 = \lambda_1 u_1$ and $Au_2 = \lambda_2 u_2$ with $\lambda_1 \neq \lambda_2$. Hence,

$$\lambda_1 u_2' u_1 = u_2' A u_1 = (A'u_2)'u_1 = (Au_2)'u_1 = \lambda_2 u_2' u_1.$$

Hence, $(\lambda_1 - \lambda_2)u_2'u_1 = 0$ so we must have $u_2'u_1 = 0$. □

Exercise C.1.2. *Show that the product of two symmetric matrices is not necessarily symmetric.*

Example C.1.3.

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

In this case the characteristic polynomial is $(1-\lambda)^2 + x^2 = 0$. If $x = 0$ the matrix is symmetric and the eignevalues are both 1. If $x = 1$ ....

C.2 Cayley–Hamilton Theorem and Implications

Theorem C.2.1. *(Cayley-Hamilton) Every square matrix satisfies its characteristic equation. That is for $A \in \mathbb{R}^{n\times n}$, $p(A) = 0$.*

Before we see a proof, it is good to see a "non-proof" (a faulty proof):

$$p(A) = \det(A - A\,I) = \det(0) = 0.$$

Exercise C.2.2. *Describe why the above is a faulty proof.*

Exercise C.2.3. *Show the validity of the Cayley-Hamilton theorem for the 2×2 matrix,*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

That is, show that,

$$A^2 - (a+d)A + (ad - bc)I = 0_{2\times 2}.$$

C.3 Quadratic Forms, Positive Definiteness and Convexity

Given a vector of variables $x \in \mathbb{R}^n$, a *quadratic form*, is a function,

$$q : \mathbb{R}^n \to \mathbb{R},$$

of the form,

$$q(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{a}_{i,j} x_i x_j.$$

The above summation can be represented by means of a matrix $\tilde{A} \in \mathbb{R}^{n\times n}$ composed of entries $\tilde{a}_{i,j}$, as follows:

$$q(x) = x'\tilde{A}x.$$

Since for any $i, j \in \{1, \ldots, n\}$ with $i \neq j$, the coefficient of $x_i x_j$ in the quadratic form, is $\tilde{a}_{i,j} + \tilde{a}_{j,i}$. We can also represent the quadratic form in terms of a **symmetric** matrix $A \in \mathbb{R}^{n\times n}$:

$$q(x) = x'Ax.$$

where the elements of the matrix A, $a_{i,j}$ are,

$$a_{i,j} = \frac{\tilde{a}_{i,j} + \tilde{a}_{j,i}}{2}.$$

Thus to every matrix $\tilde{A}$, there corresponds a quadratic form that can be represented by the *symmetrized* matrix:

$$A = \frac{1}{2}\big(\tilde{A} + \tilde{A}'\big).$$

Theorem C.3.1. *Consider a symmetric matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ (they are all real by Theorem QQQQ). Then the following are equivalent:*

1. *For any $x \in \mathbf{R}^n$, $x'Ax > 0$.*
2. *The function $q(x) = x'Ax$ is strictly convex.*
3. *$\lambda_i > 0$, $i = 1, \ldots, n$.*
4. *Determinants...*
5. *Pivots...*

Proof. □

We call a matrix that satisfies (1)–(5) of Theorem **??** *positive definite* and denote it,

$$A > 0.$$

C.4 Linear Matrix Inequalities

A linear matrix inequality (LMI) has the form

$$F(x) = F_0 + \sum_{i=1}^{m} x_i F_i > 0$$

where $x \in R^m$ is the variable and the symmetric matrices $F_i = F_i^T \in R^{n\times n}$, $i = 0, \ldots, m$, are given.

An example is the Lyapunov inequality for stability:

$$A^T P + PA < 0$$

where A is given and $P = P^T > 0$ is the variable.

Let $P_1, \ldots P_m$ be a basis for symmetric $n \times n$ matrices ($m = n(n+1)/2$). Take $F_0 = 0$ and $F_i = -A^T P_i - P_i A$.

An other example is the quadratic matric inequality

$$A^T P + PA + PBR^{-1}B^T P + Q < 0$$

where A, B, $Q = Q^T$, $R = R^T > 0$ are given matrices and $P = P^T > 0$ is the variable. This *quadratic* matric inequality in the variable P can be expressed as the *linear* matrix inequality

$$\begin{bmatrix} -A^T P - PA - Q & PB \\ B^T P & R \end{bmatrix} > 0$$

C.5 Perron Frobenious

In this section, we introduce the Perron-Frobenius theory for matrices with non-negative elements, which is not only among the most elegant theories in mathematics, but it is also among the most useful.

We shall deal with square matrices $T = (t_{ij})_{ij=1,\ldots,n}$ such that $t_{ij} \geq 0$ for all i, j, which we summarize as $T \geq 0$. The material of this section is taken from Chapter 1 in Seneta (1981).

A square non-negative matrix T is said to be *primitive* if there exists a positive integer k such that $T^k > 0$. The strongest version of the *Perron-Frobenius Theorem* holds for primitive matrices.

Theorem C.5.1 (Perron-Frobenius Theorem for primitive matrices)**.** *Suppose T is an $n \times n$ non-negative primitive matrix. Then, there exists an eigenvalue r such that*

(a) r is real and strictly positive;

(b) with r can be associated strictly positive left and right eigenvectors;

(c) $r > |\lambda|$ for any eigenvalue $\lambda \neq r$;

(d) the eigenvectors associated with r are unique to constant multiples.

(e) If $0 \leq B \leq T$ and β is an eigenvalue of B, then $|\beta| \leq r$. Moreover, $|\beta| = r$ implies $B = T$.

(f) r is a simple root of the characteristic polynomial of T.

Note that assertion (d) of Theorem C.5.1 states that the *geometric multiplicity* of r is one, whereas (f) states that its *algebraic multiplicity* is one.

Corollary C.5.2.

$$\min_i \sum_{j=1}^{n} t_{ij} \leq r \leq \max_i \sum_{j=1}^{n} t_{ij}$$

with equality on either side implying equality throughout (i.e. r can only be equal to the maximal or minimal row sum if all row sums are equal).

A similar statement holds for column sums.

Let λ_2 be the second largest eigenvalue of T (in terms of absolute value) and let m_2 be its algebraic multiplicity (if there exists λ_i such that $|\lambda_2| = |\lambda_i|$, then we assume $m_2 \geq m_i$).

Theorem C.5.3. *For a primitive matrix T:*

(a) if $\lambda_2 \neq 0$, then as $k \to \infty$,

$$T^k = r^k \, \boldsymbol{w} \, \boldsymbol{v}' + O(k^s |\lambda_2|^k)$$

elementwise, where $s = m_2 - 1$;

(b) if $\lambda_2 = 0$, then for $k \geq n - 1$

$$T^k = r^k \, \boldsymbol{w} \, \boldsymbol{v}'.$$

In both cases $\boldsymbol{w}$ and $\boldsymbol{v}'$ are any positive right and left eigenvectors corresponding to r guaranteed by Theorem C.5.1, providing only they are normed so that $\boldsymbol{v}'\boldsymbol{w} = 1$.

Theorem C.5.1 can be adapted to *irreducible* matrices. We say that an $n \times n$ matrix is *irreducible* if for every pair i, j of its index set, there exists a positive integer $m = m(i, j)$ such that $(T^m)_{ij} > 0$.

We call the *period* $d(i)$ of the index i the greatest common dividor of those k for which

$$(T^k)_{ii} > 0.$$

Note that if $T_{ii} > 0$, then $d(i) = 1$.

In an irreducible matrix, all indices have the same period. An irreducible matrix is said to be *cyclic* (or *periodic*) with period d if the period of any one (and so each one) of its indices satisfies $d > 1$, and is said to be *acyclic* (or *aperiodic*) if $d = 1$.

Theorem C.5.4. *An irreducible acyclic matrix T is primitive and conversely.*

We now state the Perron-Frobenius Theorem for irreducible matrices.

Theorem C.5.5 (Perron-Frobenius Theorem for irreducible matrices)**.** *Suppose T is an $n \times n$ irreducible non-negative matrix. Then all of the assertions (a)–(f) of Theorem C.5.1 hold, except that (c) is replaced by the weaker statement: $r \geq |\lambda|$ for any eigenvalue λ of T. Corollary C.5.2 holds also.*

We shall therefore call r the *Perron-Frobenius eigenvalue* of an irreducible matrix T, and its corresponding left and right positive eigenvectors the *Perron-Frobenius eigenvectors.*

Theorem C.5.6. *For a cyclic matrix T with period $d > 1$, there are exactly d distinct eigenvalues λ with $|\lambda| = r$, where r is the Perron-Frobenius eigenvalue of T. These eigenvalues are: $r \exp(i2\pi k/d), k = 0, 1, \ldots, d - 1$ (that is, the roots of the equation $\lambda^d - r^d = 0$).*

Corollary C.5.7. *If $\lambda \neq 0$ is an eigenvalue of T, then the numbers $\lambda exp(i2\pi k/d), k = 0, 1, \ldots, d-1$ are eigenvalues also. (Thus, rotation of the complex plane about the origin through angles of $2\pi/d$ carries the set of eigenvalues into itself.)*

Bibliographic Remarks

Exercises

Appendix D

Probability

D.1 The Probability Triple

The basic thing to start with is $\mathbb{P}(A)$. What is this? Read this as the *probability* of the event A. Probability is a number in the interval $[0, 1]$ indicating the chance of the event A occurring. If $\mathbb{P}(A) = 0$ then A will not occur. If $\mathbb{P}(A) = 1$, occurrence is certain. If $\mathbb{P}(A) = 0.78$ then we can read this as a chance of 78% for the event. It can also be read as that if we repeat the experiment that we are talking about many times, the proportion of times of which we will observe the event A occurring is 78%. The higher the probability the more likely the event will occur.

But $\mathbb{P}(A)$ doesn't live by itself. Sometimes people ask me: "You are a researcher in the field of probability, so what is the probability of finding another life form on a different planet?". My response often follows the lines: "Sorry, guys, I need a probability model. For example, you can ask me what is the chance of getting a double when tossing a pair of dice. Then my probability model will tell you this is 1/6. But for finding life forms on a different planet, I don't have a model that I can use. Sorry... But we do have some friendly astrophysicists here at UQ so go ask them!".

So what is a probability model? Well the basic way to handle this is through a *probability triple*, $\left(\Omega, \mathcal{F}, \mathbb{P}\right)$. The basic idea is that of an *experiment*. Think of every dynamic situation as an experiment. By this I mean every situation in which there can be one of several possible outcomes. The set of possible outcomes to this experiment is Ω. For example in the case of tossing a pair of dice Ω can be represented by,

$$\Omega = \{(i, j) \ : \ i, j = 1, 2, \ldots, 6\}.$$

I.e. when you roll a pair of dice you can get $(3, 4)$ indicating that the first die was 3 and the second was 4 and you can get any other combination. The set Ω is called the *sample space*. Caution: don't confuse this with "sample" as used by statisticians; In general, you shouldn't confuse the (applied) mathematical field of probability with statistics! Do you know the difference? If not, give it some thought.

Back to the probability triple: How about events? Well an *event* is a subset of Ω and we denote the set of these by $\mathcal{F}$. In complicated experiments not all subsets of Ω are in $\mathcal{F}$, but in elementary examples such as the rolling of a pair of dice we can take $\mathcal{F}$ to be composed of all possible subsets. Specifically this is the case when Ω is finite. In our specific case there are 2^{36} possible outcomes! Also for our specific example, the event $A \subset \Omega$ which indicates "getting a double" is:

$$A = \{(1,1),\, (2,2),\, (3,3),\, (4,4),\, (5,5),\, (6,6)\}.$$

One of the events in $\mathcal{F}$ is $\emptyset$. This is called the null-event. Another event is Ω itself. So basically, events are sets (subsets of Ω and elements of $\mathcal{F}$). The appendix to these notes can help you, if you are not an ace on basic set notation and operations and similarly if you have some gaps of knowledge with respect to basic counting (combinatorics).

Now $\mathbb{P}(\cdot)$ is the *probability measure* (sometimes loosely called the *probability function*). It is a function taking elements of $\mathcal{F}$ (events) and mapping them to $[0,1]$. The basic (and most sensible) model for rolling a pair of dice is to believe that each outcome (i,j) is equally likely. In this case (this is often called a *symmetric probability space*) the probability measure is obvious:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

So for the event we discussed before, $\mathbb{P}(A) = 6/36 = 1/6$. But in other examples we may have a different type of $\mathbb{P}(\cdot)$ that does not give the same chance for all outcomes.

What properties do we expect Ω, $\mathcal{F}$ and $\mathbb{P}$ to obey? Well, $\mathcal{F}$ needs to be a *sigma-Algebra* (also called *sigma-field*). This is a regularity property on the set (family) of events that ensures that the mathematics end up being well defined. Basically we need:

1. $\emptyset \in \mathcal{F}$.
2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. The set A^c is the complement with respect to Ω. I.e. $A^c = \Omega \setminus A$.
3. If $A_1, A_2, \ldots \subset \mathcal{F}$ then, $\cup_i A_i \in \mathcal{F}$. The number of sets in the union can be finite or countably infinite.

Some properties follow quite easily:

Exercise D.1.1. *Show that:*

1. $\Omega \in \mathcal{F}$.
2. *If* $A_1, A_2, \ldots \subset \mathcal{F}$ *then,* $\cap_i A_i \in \mathcal{F}$.

Having defined the (boring) machinery of $\mathcal{F}$ let's move to the key ingredient of any probability model: $\mathbb{P}(\cdot)$. The probability measure must satisfy:

1. For any $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. For any countable sequence of disjoint events, $A_1, A_2, \ldots$:
$$\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i).$$

Key in (3) is the fact that the events are disjoint. I.e. for any A_i and A_j with $i \neq j$ we have $A_i \cap A_j = \emptyset$. The above *probability axioms* imply the following:

Exercise D.1.2. *Show that:*

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. $\mathbb{P}(\emptyset) = 0$.
3. $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$ *(this is called the inclusion-exclusion principle).*

D.2 Independence

Two events A and B are said to independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. A typical example is an experiment where you do two things and they don't affect each other. For the rolling of the dice experiment, this is typically the case: One die does not affect the other. And indeed consider for $i \in \{1, \ldots, 6\}$, the events:

$$A_i := \{(i,1), (i,2), (i,3), (i,4), (i,5), (i,6)\},$$
$$B_i := \{(1,i), (2,i), (3,i), (4,i), (5,i), (6,i)\}.$$

The event A_i implies "The first die yielded i". The event B_i implies "The second die yielded i". What is the event $A_i \cap B_j$? It is read as "The first yield i and the second yielded j." Indeed,

$$A_i \cap B_j = \{(i,j)\},$$

and thus,

$$\mathbb{P}(A_i \cap B_j) = \frac{|A_i \cap B_j|}{|\Omega|} = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \frac{|A_i|}{|\Omega|}\frac{|B_j|}{|\Omega|} = \mathbb{P}(A_i)\,\mathbb{P}(B_j).$$

So the events are independent.

This example is almost too trivial to be interesting. But the concept of independence goes a long way in probability. This will become more evident when random variables and conditional probability come into play.

Students starting with probability often get confused between "two events being disjoint" and "two events being independent". After all, both terms specify that the events are non-related in some way. But in fact, these concepts are very different.

Exercise D.2.1. *Consider the experiment of tossing a fair coin (yielding 'H' or 'T') and spinning a wheel divided into three parts (yielding '1', '2' or '3'). Assume the underlying probability space is symmetric. Write* Ω, $\mathcal{F} = 2^\Omega$ *and specify* $\mathbb{P}(A)$ *for all* $A \in \mathcal{F}$ *(you'll have 64 events!). Fish out which events are disjoint and which events are independent. See that if two (non-null) events are disjoint they are not independent. And conversely if two (non-null) events are independent, they are not disjoint.*

Independence goes further than just two events. the events $A_1, \ldots, A_n$ are said to be *pair-wise independent* if for each $i \neq j$, A_i and A_j are independent. This set of events is said to be *independent* (without the "pair-wise prefix") if for any set of indexes, $1 \leq i_1 < i_2 \ldots < i_k \leq n$:

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}(A_{i_k}).$$

D.3 Conditional Probability

Given two events, $A, B \subset \Omega$, with $\mathbb{P}(B) > 0$, the *conditional probability* of A given B, denoted $\mathbb{P}(A \,|\, B)$ is defined as:

$$\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{D.1}$$

Exercise D.3.1. *Assume* $\mathbb{P}(A) > 0$ *and* $\mathbb{P}(B) > 0$. *Show that* A *and* B *are independent if and only if* $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$.

Mmmm... So if A and B are independent then the chance of A happening is not influenced by B. But if there is some dependence, then $\mathbb{P}(A \,|\, B) \neq \mathbb{P}(A)$.

Exercise D.3.2. *Suppose you roll a die. I tell you that the result is an even number. So now what is the chance that the result is* 6*?*

There are mathematical subtleties in defining conditional probability, but we won't touch these. From our perspective, we can consider the conditional probability $\mathbb{P}(\cdot \,|\, B)$, (D.1), as a new probability measure in a new probability triple, $\big(B, \tilde{\mathcal{F}}, \mathbb{P}(\cdot \,|\, B)\big)$. It is as though the sample space was reduced from Ω to B and all probabilities were simply normalised. This means that all the properties of $\mathbb{P}(\cdot)$ from the previous section carry over. For e.g.,

$$\mathbb{P}(A \,|\, B) = 1 - \mathbb{P}(B \setminus A \,|\, B).$$

Below are three useful basic results that follow immediately from the definition in (D.1). Let $A, B_1, B_2, B_3, \ldots \subset \Omega$ with $\{B_i\}$ mutually disjoint sets such that $\cup_i B_i = \Omega$:

1. *The multiplication rule:* Assume $\mathbb{P}(B) > 0$, then $\mathbb{P}(A \cap B) = \mathbb{P}(B)\,\mathbb{P}(A \,|\, B)$.

2. *The law of total probability:* $\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i) = \sum_i \mathbb{P}(A \cap B_i)$.

3. *Bayes' rule:* $\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)}$.

Note that in certain cases, the law of total probability and the celebrated Bayes' rule hold also when there is an non-countable family of events $\{B_t\}$. In that case, replace the summations over i by integrals over t.

Exercise D.3.3. *Prove (1)–(3) above.*

Have you heard of Bayesian statistics? The underlying mechanism is Bayes' rule.

An example that surprises many people is the following: Suppose you are in a television gameshow where you need to choose one of three boxes, one of which has a prize, and the others are empty. The game-show host knows where the prize is. You point at one of the boxes and say with a hesitant voice: "this is my box". At that point, the flashy gameshow host follows the producer's protocol and reveals another box, showing you that the prize is not in that one. Now you know that either your first choice was the correct box, or perhaps the prize is in the third box. The gameshow continues to follow protocol and says: "So, do you want to stay with your box, or change (to the third box)?". What do you do?

The immediate intuitive answer would be to say: "It doesn't matter, there is a 50% chance for having the prize in either the current box or the other option." But let's look more closely.

Denote the boxes by $1, 2, 3$ and assume without loss of generality that you choose box 1 at first. Denote the event that the prize is in box i by A_i. Clearly,

$$\mathbb{P}(A_i) = \frac{1}{3}, \quad i = 1, 2, 3.$$

Now the host will never reveal a box with a prize. If you initially guessed the correct box, the host will have an option between two boxes to reveal. But if you initially guessed the wrong box, the host only has one option of what to reveal. Denote by B the event that the host reveals box 2 after your choice. I.e. B^c is the event that the host reveals box 3. So:

$$\mathbb{P}(B \mid A_1) = \frac{1}{2}, \qquad \mathbb{P}(B^c \mid A_1) = \frac{1}{2}.$$

and,

$$\mathbb{P}(B \mid A_2) = 0, \qquad \mathbb{P}(B^c \mid A_2) = 1,$$

and similarly,

$$\mathbb{P}(B \mid A_3) = 1, \qquad \mathbb{P}(B^c \mid A_3) = 0.$$

Now using the law of total probability,

$$\mathbb{P}(B) = \mathbb{P}(B \mid A_1)\,\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\,\mathbb{P}(A_2) + \mathbb{P}(B \mid A_3)\,\mathbb{P}(A_3) = \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{1}{2}.$$

So (not surprisingly) there is a 50% chance that the host reveals box 2.

Now let's put you back in that situation. You are on TV! You just made a choice (box 1), and the gameshow guy (or flashy gal if you wish) just revealed box 2. So you observed the event B. Now you want to compare,

$$\mathbb{P}(A_1 \,|\, B), \qquad \text{v.s.} \qquad \mathbb{P}(A_3 \,|\, B),$$

and choose the box which maximises this probability. Using Bayes' Rule

$$\mathbb{P}(A_1 \,|\, B) = \frac{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1)}{\mathbb{P}(B)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3},$$

and the complement,

$$\mathbb{P}(A_3 \,|\, B) = \frac{\mathbb{P}(B \,|\, A_3)\mathbb{P}(A_3)}{\mathbb{P}(B)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

So you are better of changing boxes!!! Go for it.

I must admit that this is one of the most confusing elementary examples of conditional probability (and Bayes' rule) that are out there. But it is also one of the more shocking examples; hence it is interesting. I was recently reminded of it by a curious pool-safety-person[1], and I couldn't explain it to him without resorting to formalism. Maybe you can...

Exercise D.3.4. *Think about this example. Try to describe (in lay-person terms) why you are better off changing boxes.*

D.4 Discrete Random Variables and their Probability Distributions

So now you know what a probability triple is and you even know about independence and basic conditional probability. What next? Well typically we work with random quantities. And by "quantity" I mean something that is easier to handle and manipulate in comparison to arbitrary sets (events). By this I mean real numbers, integers, complex numbers, vectors, matrices etc... But let's just think of random quantities that are either real valued (continuous) or integer valued (discrete). Our focus is in fact on discrete (basically integer) quantities..

A *random variable*, X (also referred to sometimes as a *measurable function*), is a mapping from Ω to $\mathbb{R}$ or $\mathbb{N}$ or some other sensible set (vectors, complex numbers etc...). Think for now about integer random variables so, $X : \Omega \to \mathbb{Z}$. Now the idea is that since the

[1]Dan Adelman (Finishing Touch Pool Safety Inspections and Compliance Repairs) – highly recommended for pool safety certificates as well as for a long chat about probability once the job is done.

$\omega \in \Omega$ is a random outcome of an experiment, then so is $X(\omega)$. Formally, the way to handle this is to define for sensible subsets of $B \subset \mathbb{Z}$, an inverse image set,

$$A = \{\omega \in \Omega \,:\, X(\omega) \in B\}.$$

Think of A as an event; B should not be thought of as an event. It is rather a set of values that the random variable may take.

Now if everything is well defined meaning that $\mathcal{F}$ is rich-enough and that $X(\cdot)$ and B are not too crazy, then $A \in \mathcal{F}$ and hence it is a proper event which we can stick in $\mathbb{P}(\cdot)$. Often instead of the event A we often just write "$X \in B$" instead. So we can calculate probabilities of the form, $\mathbb{P}(X \in B)$. Of course if the set B contains just one point, say b, then we can try and evaluate $\mathbb{P}(X = b)$ or if B is say an interval $[a, b]$ (with possibility one or two of the endpoints being $-\infty$ or ∞, then we can try and evaluate $\mathbb{P}(a \leq X \leq b)$, etc.. etc... The point is that random variables quantify the outcome of the experiment. And for some possible set of outcomes, B, we are asking for the probability of $X \in B$.

Now consider sets B of the form, $(-\infty, b]$. For such sets we have,

$$\mathbb{P}(A) = \mathbb{P}(X \in B) = \mathbb{P}(X \leq b).$$

Such subsets, B are useful because if we know the value of $\mathbb{P}(X \leq b)$ for all b then we can use this to calculate $\mathbb{P}(X \in B)$ for any sensible B. This motivates us to define the *distribution function*:

$$F_X(b) = \mathbb{P}(X \leq b).$$

The subscript X is just part of the notation of the function - it reminds us that this is the distribution of the random variable X. This function is also (less ambiguously) called: the *cumulative distribution function* (CDF). Some prefer to work with the *complementary cumulative distribution function* (CCDF):

$$\overline{F}_X(b) := 1 - F_X(b) = \mathbb{P}(X > b).$$

Some call the above the *survival function* - but these guys are typically wearing suits and don't smile too much because they work in insurance companies or are reliability engineers. The CDF or CCDF are alternative descriptions of the *distribution* of X. There are other descriptions which are sometimes useful also (probability mass function, probability density function, moment generating function, probability generating function, characteristic function, Laplace transform, hazard rate, renewal measure,...). What I'm trying to say is that there are many ways to describe the *distribution of a random variable*, each useful in its own way. But let's get back to CDF:

Exercise D.4.1. *Show that,*

1. $\lim_{x \to -\infty} F_X(x) = 0$.
2. $\lim_{x \to \infty} F_X(x) = 1$.

3. $F_X(\cdot)$ is non-decreasing.

The above three properties are often taken to be defining properties of CDFs. For any function satisfying the above, we can actually find a probability triple supporting a random variable X with the desired CDF.

In these notes we focus mostly on random variables whose values fall within a discrete set such as $\{0, \ldots, n\}$ for some finite n or $\mathbb{N}$ or $\mathbb{Z}$ etc. These are sometimes called *discrete random variables*. We call the set of values which the random variable may take, the *support*. If (for e.g.) the support does not have negative values then we say the random variable is *non-negative*.

Exercise D.4.2. *Consider the first example of these notes (tossing of two dice). Let the random variable be the sum of the dice. Illustrate the graph $F_X(x)$. At points of discontinuity, make sure to note open and closed indications on the graph.*

For discrete random variables an alternative (and sometimes easier to handle) representation of the distribution is the *probability mass function* (PMF):

$$p_X(k) := \mathbb{P}(X = k).$$

Assuming that the support is some subset of $\mathbb{Z}$ then,

$$F_X(k) := \sum_{i=-\infty}^{k} p_X(i) \quad \text{and} \quad p_X(k) = F_X(k) - F_X(k - \epsilon),$$

where ϵ is any value in the range $(0, 1]$. For k that are not in the support we simply have $p_X(k) = 0$. Keep this in mind, because when we write things such as,

$$\sum_{k=-\infty}^{\infty} p_X(k),$$

this is equivalent to,

$$\sum_{k \,\in\, \text{support of } X} p_X(k).$$

Exercise D.4.3. *Draw the PMF associated for the previous exercise. Place your illustration under the CDF. Exhibit the relationship between the CDF and the PMF.*

Some people call refer to PMF as "density". I respect these people, some of them are even my friends, but I'm not one of them. I keep the word density for functions $f_X(x)$ that describe the CDF of continuous random variables through:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

But more on this later (when we briefly touch continuous distributions). Also I should state that in the continuation of these notes, I won't use the notation $p_X(\cdot)$ much, even though PMFs will appear everywhere.

D.5 Expectation, Mean, Variance, Moments

The *mean* of a (discrete) random variable, denoted $\mathbb{E}[X]$ is:

$$\mathbb{E}[X] = \sum_{k=-\infty}^{\infty} k\, p_X(k).$$

An alternative name for the mean is the *expectation* or *expected value*. The expected value describes the "center of mass" of the probability distribution. Another meaning follows from the law of large numbers described in the sequel: If we observe many random variables having this distribution and calculate their average, it will be near the mean. Note that in the summation above, it is enough to sum over the support of the random variable since for other values of k, $p_X(k) = 0$.

Observe that the mean of an integer valued random variable does not have to be an integer.

Exercise D.5.1. *What is the mean value for the sum of two dice? Use the probability model and random variable that appeared in previous exercises.*

Exercise D.5.2. *Show that for a non-negative random variable,*

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \overline{F}_X(k).$$

Take now $h : \mathbb{R} \to \mathbb{R}$ then $h(X(\omega))$ is some new random variable. We can calculate the mean of this new random variable simply as follows:

$$\mathbb{E}[h(X)] = \sum_{k=-\infty}^{\infty} h(k)\, p_X(k). \tag{D.2}$$

I.e. the expectation functional, $\mathbb{E}[\cdot]$ takes as input a random variable and returns a number. When $h(x) = x^n$ then $\mathbb{E}[h(X)]$ is called the n'th *moment*. I.e. the first moment is the mean. Another important case $h(x) = (x - \mathbb{E}[X])^2$ then $\mathbb{E}[h(X)]$ is called the *variance* and denoted, $\mathbb{V}\mathrm{ar}\,(X)$. Note that it is non-negative. The square root of the variance is called the *standard deviation*. Both the variance and the standard deviation are measures of the spread of the distribution (each one of these is useful in its own way). You can see that:

$$\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}\big[X^2 - 2\mathbb{E}[X]\, X + (\mathbb{E}[X])^2\big]. \tag{D.3}$$

Note that inside the expectation operator we are doing algebra involving both the random variable X and the constant values, 2 and $\mathbb{E}[X]$.

Exercise D.5.3. *Show that,*

1. *If c is a constant (non-random quantity), then $\mathbb{E}[c\,X] = c\,\mathbb{E}[X]$.*

2. *For any two random variables, X and Y,*

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

(Illustrate this through the meaning of a random variable – a function of ω).

Now with these basic properties of the expectation, you are ready to proceed with (D.3) to show that,

$$\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

This implies that for "zero-mean" random variables, the variance is simply the second moment.

Exercise D.5.4. *Let, c_1, c_2 be some constants. What is $\mathbb{V}\mathrm{ar}\,(c_1X + c_2)$ in terms of $\mathbb{V}\mathrm{ar}\,(X)$?*

Exercise D.5.5. *Show that if $\mathbb{V}\mathrm{ar}\,(X) = 0$ then the support of X contains a single value (i.e. there is some k_0 such that $p_X(k_0) = 1$).*

Another very important $h(\cdot)$ is obtained by setting some $B \subset \mathbb{R}$ and then $h(x) = \mathbf{1}_B(x) := \mathbf{1}\{x \in B\}$ (and indicator function returning 1 if $x \in B$ and 0 otherwise). In this case $\mathbb{E}[h(X)] = \mathbb{P}(X \in B)$. Nice, no?

D.6 Bernoulli Trials

We now consider probability spaces where Ω is the set of binary sequences,

$$\Omega = \{(b_1, b_2, b_3, \ldots),\ b_i \in \{0,1\}\}$$

and where $\mathbb{P}(\cdot)$ is such that the events $\{b_i = 1\}$ are independent. We further assume that $\mathbb{P}(\{b_i = 1\}) = p$ for all i. I.e. this probability space describes experiments involving a sequence of independent "coin flips", each with having the same probability of success: p.

There are now many random variables associated with this probability space. We say X follows a *Bernoulli distribution*, with probability p if,

$$\mathbb{P}(X = 0) = (1-p), \qquad \text{and} \qquad \mathbb{P}(X = 1) = p.$$

We say that X follows a *binomial distribution* with parameters n and p if,

$$\mathbb{P}(X = k) = \binom{n}{k} p^k\,(1-p)^{n-k}, \qquad k = 0, \ldots, n. \tag{D.4}$$

Here n is any integer ≥ 1 and $p \in [0,1]$.

Exercise D.6.1. *Show that with respect to the Bernoulli Trials probability space,*

$$X(\omega) = \sum_{i=1}^{n} \mathbf{1}\{b_i^\omega = 1\},$$

where b_i^ω is the i'th element of ω. That is, derive the right hand side (D.4).

Exercise D.6.2. *Verify for the binomial distribution of* (D.4)*, that*

$$\sum_{i=0}^{n} \mathbb{P}(X = i) = 1.$$

Exercise D.6.3. 1. *Show that the mean of a binomial distribution is $n\,p$.*

2. *Let X be binomially distributed with n and p. What is the distribution of $Z = n - X$?*

Exercise D.6.4. *Assume you are guessing answers on a multiple choose test that has 20 questions, and each can be answered (a), (b), (c), or (d). What is the chance of getting* 10 *or more answers correct?*

Consider now, $X(\omega) = \inf\{k \in \{1, 2, 3, \ldots\} \mid b_k^\omega = 1\}$. I.e. This is the index of the trial with the first success. Such a random variable is said to follow a *geometric distribution* with success probability p.

Exercise D.6.5. *Show that,*

$$\mathbb{P}(X = k) = (1 - p)^{k-1}\, p, \qquad k = 1, 2, \ldots.$$

Verify that (as needed),

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = 1.$$

Exercise D.6.6. *The chance of getting a flat-tire on a bicycle ride is* 0.01*. What is the chance of having* 20 *consecutive bicycle rides without a flat tire?*

A related random variable (also referred to as "geometric"), counts the "number of failures until success" as opposed to the "number of trials until success".

Exercise D.6.7. *What is the support and distribution of this version of the geometric?*

A generalisation of the geometric distribution is the *negative binomial distribution.* Here X counts the number of trials till m successes:

$$X(\omega) = \inf\{k \in \{1, 2, 3, \ldots\} \mid \sum_{i=1}^{k} b_i^\omega = m\}.$$

The support of this distribution is $\{m, m + 1, m + 2, \ldots\}$.

Exercise D.6.8. *Develop the pmf of the negative binomial distribution with parameters $p \in [0,1]$ and $m \geq 1$ from first principles. Do the same for a modification (as was for the geometric) which counts the number of failures till m successes. The support here is $\{0,1,2,\ldots\}$.*

D.7 Other Common Discrete Distributions

You can think of the binomial distribution as follows: You are fishing in a lake where there are M brown fish and N gold fish. You are fishing out n fish, one by one, and whenever you catch a fish you return it to the lake. So assuming your chance of catching a fish of a given type is exactly its proportion, and further assuming that fishing attempts don't interact, the number of gold fish that you get is binomially distributed with n and $p = N/(N+M)$. The thing here is that by catching a fish, you didn't alter the possible future catches.

But what if you (weren't a vegetarian like me), and as you catch a fish, you bop it in the head, fry eat and eat it. Then with every fish you are catching, you are altering the population of fish, and then the binomial description no longer holds. In this case X, the number of gold fish that you catch follows a *hyper-geometric distribution.*

$$\mathbb{P}(X = k) = \frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}}.$$

Exercise D.7.1. *The hyper-geometric distribution is constructed by basic counting arguments on a symmetric probability space. Carry out these arguments. Further, what is the support of this distribution?*

Exercise D.7.2. *When $N + M \to \infty$ (i.e. big lakes) such that $N/(N+M) \to p$, you would expect that it doesn't matter if you return the fish to the lake or not. This can be formalised by showing the the pmf of the hyper-geometric distribution converges to the binomial distribution. Find this some place in the literature and carry out the computations, describing the steps. Or if you have already had several courses of probability, maybe try to do it without looking elsewhere.*

Another useful discrete distribution is the *Poisson distribution* (incidentally "poisson" means fish in French – but we are now done with fish). The random variable X is distributed Poisson with parameter λ if,

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0,1,2,\ldots.$$

Exercise D.7.3. *Show that the mean and variance are both λ.*

The Poisson distribution is useful for describing the number of events in a time-interval. Especially when events occur in a "completely random manner". That is, it may be a

good model for the number of shooting stars that you observe while looking at a moonless desert sky for an hour. To see this, consider the hour and divide it input n intervals, each interval being quite small. Then it is sensible that within each such interval there is a probability of p_n for seeing a shooting star. Here the subscript indicates the dependence on n. The bigger the n the smaller the p. In fact, how about setting $\lambda = n\,p_n$ (this is the mean number of shooting stars during that hour). Now if we increase $n \to \infty$ then $p_n \to 0$ in such a way that their product remains λ. For any finite n, the number of stars is distributed Binomial(n, p_n). But as $n \to \infty$ this converges to Poisson.

Exercise D.7.4. *Show that for every k,*

$$\lim_{n\to\infty} \binom{n}{k} \big(\frac{\lambda}{n}\big)^k \big(1 - \frac{\lambda}{n}\big)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}.$$

As a final example of a discrete distribution, consider,

$$\mathbb{P}(X = k) = \frac{1}{k(k+1)}, \quad k = 1, 2, \ldots.$$

Indeed by writing

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1},$$

we get a telescopic sum and see that,

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = 1,$$

as desired. This distribution is an example of a power-law, since the tails of it decay to 0 like a power law. Such distributions are sometimes called *heavy tailed* and indeed the following distribution does not have a finite mean.

Exercise D.7.5. *Show that the mean is infinite.*

Note that while the mean is infinite it is well defined. I.e. this series diverges to infinity:

$$\sum_{k=1}^{\infty} k\,\mathbb{P}(X = k) = \infty.$$

But in other cases, the mean is not even defined. For e.g. consider this distribution:

$$\mathbb{P}(X = k) = \frac{1}{2|k|(|k|+1)}, \quad k = \ldots, -3, -2, -1, 1, 2, 3 \ldots.$$

D.8 Vector Valued Random Variables

A vector valued random variable doesn't differ much from the scalar (uni-variate) cases described above. We'll present things for a vector of two random variables, X and Y. The generalisation to n random variables is straight forward.

The basic object is the *joint probability mass function*:

$$p_{X,Y}(k,\ell) = \mathbb{P}(X = k,\, Y = \ell).$$

The requirement is that,

$$\sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} p_{X,Y}(k,\ell) = 1.$$

This is sometimes called the *joint distribution* of X and Y. Knowing this joint distribution, how can we recover the distributions of the individual random variables, X and Y? To get the distribution of X, we sum over all possible Y:

$$p_X(k) = \sum_{\ell=-\infty}^{\infty} p_{X,Y}(k,\ell).$$

Similarly to get the distribution of Y we can sum over all possible X.

Exercise D.8.1. *Derive the above using the law of total probability.*

We know about independence of events, but what is independence of random variables? The random variables X and Y are said to be *independent* if,

$$p_{X,Y}(k,\ell) = p_X(k)\, p_Y(\ell).$$

When the random variables are independent, the knowledge of X yields no information about Y and visa-versa.

Given some function, $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we can compute the expectation of the random variable $h(X,Y)$ as follows:

$$\mathbb{E}[h(X,Y)] = \sum_{k} \sum_{\ell} h(k,\ell)\, p_{X,Y}(k,\ell).$$

The *covariance* of X and Y, denoted $\mathbb{C}\mathrm{ov}\,(X,Y)$ is computed in this way using

$$h(x,y) = (x - \mathbb{E}[X])(y - \mathbb{E}[Y]).$$

Exercise D.8.2. *Show that* $\mathbb{C}\mathrm{ov}\,(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

Exercise D.8.3. *Show that if X and Y are independent then $\mathbb{E}[X\,Y] = \mathbb{E}[X]\,\mathbb{E}[Y]$ and hence, $\mathbb{C}\mathrm{ov}\,(X,Y) = 0$.*

Exercise D.8.4. *Take a case where the support of X is $\{1,2,3\}$ and the support of Y is $\{1,2\}$.*

1. *Find $p_{X,Y}(x,y)$ such that $\mathbb{C}\mathrm{ov}\,(X,Y) \neq 0$.*
2. *Find $p_{X,Y}(x,y)$ such that X and Y are not independent but $\mathbb{C}\mathrm{ov}\,(X,Y) = 0$.*

D.9 Conditioning and Random Variables

Now that you know about multiple random variables living together in the same probability space, you can start seeing how they interact. Consider first the conditional probability:

$$\mathbb{P}(X = k \,|\, Y = \ell).$$

Since you can read "$X = k$" and "$Y = \ell$" as events then $\mathbb{P}(X = k \,|\, Y = \ell)$ is well defined (well, as long as $Y = \ell$ can occur with a positive probability). Continuing this, define the function, $p_{X|Y=\ell}(\cdot, \cdot)$, as:

$$p_{X|Y=\ell}(k, \ell) := \mathbb{P}(X = k \,|\, Y = \ell) = \frac{\mathbb{P}(X = k,\, Y = \ell)}{\mathbb{P}(Y = \ell)} = \frac{p_{X,Y}(k, \ell)}{p_Y(\ell)}.$$

The function $p_{X|Y=\ell}(\cdot, \ell)$ specifies the *conditional distribution* of X given that $Y = \ell$.

Exercise D.9.1. *Show that $p_{X|Y=\ell}(\cdot, \ell)$ is a valid probability mass function (in the first variable) for any ℓ such that $\mathbb{P}(Y = \ell) > 0$.*

Exercise D.9.2. *Show that if X and Y are independent random variables, then $p_{X|Y=\ell}(\cdot, \ell) = p_X(\cdot)$.*

Exercise D.9.3. *For your example used as solution of Exercise D.8.4 calculate, $p_{X|Y=\ell}(\cdot, \cdot)$ and $p_{Y|X=k}(\cdot, \cdot)$ for all possible values. I.e. specify 6 distributions.*

The geometric distribution is said to be *memoryless* due to this property:

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s).$$

Exercise D.9.4.

1. *Show that the memoryless holds for geometric random variables.*
2. *Comment on why this property makes sense (considering the sequence of Bernoulli trials).*
3. *Find another discrete distribution which does* **not** *satisfy the memoryless property.*

Now that you know about conditional distributions, you can talk about the *conditional expectation*, variance, etc... Simply define:

$$\mathbb{E}[h(X) \,|\, Y = \ell] = \sum_k h(k)\, p_{X|Y=\ell}(k, \ell).$$

Exercise D.9.5. *Calculate the conditional means of the 6 distributions of the previous example. Compare these means to the two (unconditional) means of X and Y.*

Observe that you can think of $\mathbb{E}[h(X) \,|\, Y = \ell]$ as a function of ℓ. So what if you left ℓ unspecified and let it simply be the result of the random variable Y? In this case, you get (also called conditional expectation) the random variable: $\mathbb{E}[h(X) \,|\, Y]$. The conditional expectation is a random variable because it is a function of the random variable on which we are conditioning.

Exercise D.9.6. *Show that,*

$$\mathbb{E}\Big[\mathbb{E}[h(X) \,|\, Y]\Big] = \mathbb{E}[h(X)]. \tag{D.5}$$

Note that the outer expectation is with respect to the random variable Y.

The formula (D.5) is sometimes called the smoothing formula. It is sometimes super-useful because, evaluation of $\mathbb{E}[h(X)]$ in its own may be tough, but if we condition on another random variable Y, things get much easier. This is a classic example: Let $X_1, X_2, \ldots$ be a sequence of i.i.d. (*independent and identically distributed*) random variables independent of some discrete random variable N. Denote,

$$S := \sum_{i=1}^{N} X_i.$$

The new random variable S is sometimes called a random sum. For example, N may be the number of insurance claims a company has during a month, and each insurance claim is assumed to be distributed as X_1. What is $\mathbb{E}[S]$? Intuition may tell you that, $\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[X_1]$. This is for example the case if N equals some fixed value with probability 1 (the linearity of expectation). But how can you show (prove) this? Well, condition on N:

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{N} X_i] &= \mathbb{E}[\,\mathbb{E}[\sum_{i=1}^{N} X_i \,|N]] \\
&= \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i \,|\, N]] \\
&= \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i]] \\
&= \mathbb{E}[N\,\mathbb{E}[X_1]] \\
&= \mathbb{E}[X_1]\,\mathbb{E}[N].
\end{aligned}$$

Exercise D.9.7. *Detail (in words) what is happening in each step of the above derivation.*

D.10 A Bit on Continuous Distributions

The random variables discussed up to now were discrete. Their support is finite or countably infinite. For our purposes, these are indeed the critical cases to master. Nevertheless, we now briefly touch on *continuous random variables*. In the continuous case, the support is some non-countable subset of $\mathbb{R}$: E.g. $[a, b]$ or $[0, \infty)$ or all of $\mathbb{R}$. For such random variables, $\mathbb{P}(X = x) = 0$ for any specific x, but for intervals of strictly positive length, the probability can be non-zero. Such random variables are best described by a *density function*: $f_X(x) : \mathbb{R} \to \mathbb{R}_+$. The best way to think of the density is that it is a function satisfies the following:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

Exercise D.10.1. *1. What is $\int_{-\infty}^{\infty} f_X(x)\, dx$?*

2. *Given a density, $f_X(\cdot)$, what is the CDF? Is the CDF a continuous function? Or only if the density is continuous?*

3. *Given any integrable, non-negative function $\tilde{f}(x)$, describe how to make a density $f_X(\cdot)$ such that $f_X(x) = K\tilde{f}(x)$ for some constant K.*

For statisticians, the typical way of thinking about a distribution is through the density. If you think about it, indeed a PMF and a density are not so different. You should also know that random variables don't need to be continuous or discrete, you can get mixtures of the two or even more exotic objects. But for an elementary and introductory treatment such as ours, this dichotomy is fine.

The mean, moments and variance of continuous random variables are defined in an analogous way to the discrete case. The basic definitions is:

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)\, f_X(x)\, dx. \tag{D.6}$$

Once you realise that $\mathbb{P}\big(X \in [x, x + dx)\big) \approx f_X(x)\, dx$, the above should make perfect sense. I.e. compare (D.6) with (D.2). As with discrete random variables, make sure that you know what is the support of the random variable. For x's not in the support, $f_X(x) = 0$. So the region of integration in (D.6) may be limited to the support.

There are many types (parametrised families) of continuous probability distributions and manipulation of these encompasses a good part of a full course of probability. Here we shall outline three key types:

The *uniform distribution* on the range $[a, b]$ has density,

$$f_X(x) = \frac{1}{b - a}, \qquad x \in [a, b].$$

Exercise D.10.2. *Calculate the mean and variance of the uniform distribution. The mean should make "perfect sense" – explain it. The variance: not intuitive.*

Exercise D.10.3. *Write out the CDF of the uniform distribution. Make sure to specify it for the three regions, $x \leq a$, $x \in [a, b]$ and $x > b$.*

But come on! The uniform density is a bit boring. This one is much more exciting:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

This is the *normal* (also known as Gaussian) density with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$.

Exercise D.10.4. *Show that the mean is μ and that the variance is σ^2.*

Gaussian random variables are everywhere. I said everywhere!!! In the sequel when we discuss the central limit theorem there is some evidence for that.

Exercise D.10.5. *Do you believe me that for the Gaussian case, $f_X(\cdot)$ is a density? Carry out numerical integration (for some selected μ and σ) to check that,*

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

The final example that we briefly describe is the exponential distribution with parameter $\lambda > 0$.

$$f_X(x) = \lambda\, e^{-\lambda x}, \quad x > 0.$$

Exercise D.10.6.

1. *Verify that $\int_0^{\infty} f_X(x)\, dx = 1$.*
2. *Calculate the mean.*
3. *Calculate the variance.*

You can get a discrete distribution by transforming a continuous one. Here is one such example:

Exercise D.10.7. *Let X be distributed exponential(λ). Let $Y = \lfloor X \rfloor$. What is the distribution of Y?*

Exercise D.10.8. *Show that (as for geometric random variables), exponential random variables also satisfy the memoryless property.*

D.11 Limiting Behaviour of Averages

Much of modern probability deals with limiting results associated with sequences of random variables and stochastic processes. Here we only discuss the two fundamental classic results:

The first result states that the sample mean converges to the mean:

Theorem D.11.1 (The Strong Law of Large Numbers (SLLN)). *Let $X_1, X_2, \ldots$ be and i.i.d. sequence of random variables with finite mean μ. Then with probability 1:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

Exercise D.11.2. *Let $q = \mathbb{P}(X_i > \alpha)$. Use the SLLN to show that with probability 1:*

$$\lim_{n\to\infty} \frac{\#_n\{X_i > \alpha\}}{n} = q,$$

where $\#_n\{A_i\}$ is the number of times out of the first n during which the event A_i occurs.

The next result is called the *central limit theorem*. It is the reason for the universality of the normal distribution. It shows that normalised sums of random variables converge in distribution to the normal distribution.

Theorem D.11.3 (The Central Limit Theorem (CLT)). *Let $X_1, X_2, \ldots$ be and i.i.d. sequence of random variables with mean μ and finite variance $\sigma^2 > 0$. Then,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \le x\Big) = \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}\, du, \quad \forall x \in \mathbb{R}.$$

Exercise D.11.4. *Another version (often more popular with statisticians) of the CLT deals with the asymptotic distribution of the sample mean, $\frac{1}{n}\sum_{i=1}^{n} X_i$:*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - a_n}{b_n} \le x\Big) = \Phi(x) \quad \forall x \in \mathbb{R}.$$

Here a_n is the mean of the sample mean and b_n is it's standard deviation. What are a_n and b_n?

Exercise D.11.5. *Let X_1, X_2 and X_3 be i.i.d. uniform$(0,1)$ random variables. Using either a convolution (analytically – if you know how to do that) or via simulation (overviewed in the next section), plot the density of $S_n = \sum_{i=1}^{n} X_i$ for $n = 2$ and 3. What is the relation of this exercise to the CLT?*

D.12 Computer Simulation of Random Variables

When you invoke the **rand**() function in matlab (or similar functions in similar software packages) you get a *pseudo-random* number in the range $[0, 1]$. This number is an element in a deterministic (non-random) sequence initialised by a *seed*. A good pseudorandom sequence has statistical properties similar to an i.i.d. sequence of uniform$(0, 1)$ random variables.

What if you want to use a computer to generate (simulate) random variables from a different distribution? In certain cases, it should be obvious how to do this:

Exercise D.12.1. *Generate on a computer,* $10,000$ *Bernoulli random variables with success probability* $p = 0.25$*. Calculate the sample mean and sample variance. How far are these values from the theoretical values?*

So you figured out how to generate Bernoulli random variables. But what about other types of random variables? Below is a general method.

Proposition D.12.2 (Inverse probability transform)**.** *Let* $U \sim$ *uniform*$(0, 1)$ *and Let* $F(\cdot)$ *be a CDF with inverse function,*

$$F^{-1}(u) := \inf\{x \mid F(x) = u\}.$$

Then the random variable $X = F^{-1}(U)$ *is distributed with CDF* $F(\cdot)$*.*

Proof.

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

□

So if you want to generate from distribution X, you need to find out $F^{-1}(\cdot)$ and apply this function to a pseudorandom uniform. For continuous random variables, this is often very easy.

Exercise D.12.3. *Generate* $10,000$ *exponential*$(1/2)$ *values. Plot their histogram. Calculate their sample mean and sample variance. Compare this to the theoretical values.*

You will often need to generate from a discrete distribution with probability masses given by some vector $\boldsymbol{p}$. Proposition D.12.2 can be used for that.

Exercise D.12.4. *Write a function that takes as input* $\boldsymbol{p}$ *of some arbitrary finite length and generates a random variable distributed according to this vector. Try this on the vector,*

$$\boldsymbol{p} = [0.35,\ 0.25,\ 0.1,\ 0.3].$$

Generate $10,000$ *values distributed according to* $\boldsymbol{p}$ *and compare their empirical frequencies to* $\boldsymbol{p}$*.*

This section recalls some basics of probability theory.

D.13 Gaussian Random Vectors

In this section, we briefly summarize Gaussian random vectors. We begin with Gaussian scalars: a random variable, X is said to have a Gaussian (normal) distribution with mean μ and variance $\sigma^2 > 0$, denoted, $X \sim N(\mu, \sigma^2)$ if,

$$\mathbb{P}\Big(a \leq X \leq b\Big) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx.$$

We have,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = \mu.$$

Further,

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = \sigma^2.$$

Let us now consider random vectors. We say that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, denoted $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ if,

$$\mathbb{P}\Big(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n\Big) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} \phi(x_1, \ldots, x_n) dx_1 \ldots dx_n,$$

with the density function being

$$\phi(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

It can be calculated that in this case,

$$\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}, \qquad Cov(\boldsymbol{X}) = \Sigma.$$

Further, the marginal distribution of each of the X_i's is normal.

Note that

(i) Distributions of Gaussian random vectors are characterized by their mean vector and covariance matrix.

(ii) If two coordinates are non-correlated (covariance 0) then they are independent.

(iii) Linear transformations of Gaussian random vectors yield random vectors that still follow the Gaussian distribution with mean and covariance as given by Exercise **??**.

The final property that we shall overview for Gaussian random vectors deals with *conditional distributions*. Partition $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ into $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ and have,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}. \qquad \Sigma = \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma'_{ab} & \Sigma_b \end{bmatrix}.$$

We have that the distribution of $\boldsymbol{X}_a$ conditional on $\boldsymbol{X}_b = \boldsymbol{x}_b$ is

$$\mathcal{N}\Big(\boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_b^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b), \quad \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ab}'\Big). \tag{D.7}$$

This is useful for estimating $\boldsymbol{X}_a$ based on *measurements* of $\boldsymbol{X}_b$. A sensible estimate in this case is, $\boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_b^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$. As a "sanity check" of this formula observe that if $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ are independent then $\Sigma_{ab} = 0$ and thus the estimate is simply $\boldsymbol{\mu}_a$.

D.14 Stochastic Processes

A collection of random variables $\{X_t,\ t \in T\}$ (or $\{X(t), t \in T\}$) on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *stochastic process*. The index variable t is often called 'time'.

- If $T = \{1, 2, \cdots\}$ or $\{\cdots, -2, -1, 0, 1, 2, \cdots\}$, the process is a *discrete time process*.
- If $T = \mathbb{R}$ or $[0, \infty)$, the process is a *continuous time process*.
- If $T = \mathbb{R}^d$, then the process is a *spatial process*, for example temperature at $t \in T \subset \mathbb{R}^2$, which could be, say, the University campus.

In the present context, we can think of the sample space Ω as consisting of the set of sample paths or realizations ω of the stochastic process $\{X_t : t \in T\}$, that is a set of sequences if T is discrete or a set of functions if T is continuous. Each $\omega \in \Omega$ has a value at each time point $t \in T$. With this interpretation,

- For a fixed ω, we can think of t as a variable, and $X_\omega(t)$ as a deterministic function (realization, trajectory, sample path) of the process.
- If we allow ω to vary, we get a collection of trajectories.
- For fixed t, with ω varying, we see that $X_t(\omega)$ is a random variable.
- If both ω and t are fixed, then $X_t(\omega)$ is a real number.

We give below a few examples of stochastic processess.

- If X_t is the number of sales of an item up to time t, then the stochastic process $\{X_t, t \geq 0\}$ is called a *counting process*. If X_t is a counting process, then
  - For fixed ω, $X_t(\omega)$ is a non-decreasing step function of t.
  - For fixed t, $X_t(\omega)$ is a non-negative integer-valued random variable.
  - For $s < t$, $X_t - X_s$ is the number of events that have occurred in the interval $(s, t]$.

- If X_t is the number of people in a queue at time t, then $\{X_t : t \geq 0\}$ is a stochastic process where, for each t, $X_t(\omega)$ is a non-negative integer-valued random variable, but it is NOT a counting process because, for fixed ω, $X_t(\omega)$ can decrease.

- If $X_t \sim N(0,1)$ for all t, then X_t is a *Gaussian* process.

  Different processes can be modelled by making different assumptions about the dependence between the X_t for different t.

- The *Standard Brownian Motion* is a Gaussian process where $X(t_1) - X(s_1)$ and $X(t_2) - X(s_2)$ are independent for all disjoint intervals $[s_1, t_1]$ and $[s_2, t_2]$. We also have $V(X(t_1) - X(s_1)) = t_1 - s_1$ for all $s_1 < t_1$.

Remark D.14.1. *Knowing just the one-dimensional (individual) distributions of X_t for all t is not enough to describe a stochastic process.*
To specify the complete distribution of a stochastic process $\{X_t,\ t \in T\}$, we need to know the finite-dimensional distributions that is the family of joint distribution functions $F_{t_1,t_2,\cdots,t_k}(x_1, \cdots, x_k)$ of $X_{t_1}, \cdots, X_{t_k}$ for all $k \geq 1$ and $t_1, \cdots, t_k \in T$.

Bibliographic Remarks

Exercises

Appendix E

Further Markov Chain Results

E.1 Communication and State Classification

Definition E.1.1. *A state j is said to be* accessible *from state i if, given that the system has started at i, there is a positive probability that it will eventually be in j, that is,*

$$\mathbf{P}[\bigcup_{n=0}^{\infty}\{X_n = j\} \mid X_0 = i] > 0. \tag{E.1}$$

Equivalently,

$$\sum_{n=0}^{\infty}[P^n]_{i,j} = \sum_{n=0}^{\infty}\mathbf{P}[X_n = j \mid X_0 = i] > 0. \tag{E.2}$$

Definition E.1.2. *A state i is said to* communicate with *state j if i is accessible from j and j is accessible from i.*

Definition E.1.3. *A Markov chain is said to be* irreducible *if all pairs i and j, for $i, j \in \mathcal{S}$, communicate. In other words, it is possible to go from any state i to any other state j. Otherwise, the chain is said to be* reducible.

One can partition the state space $\mathcal{S}$ into subsets of communicating states. Then each subset is called a *communicating class*, or *class*, and communicating classes of a Markov chain are mutually exclusive and exhaustive. Every irreducible Markov chain has exactly one communicating class, and every reducible Markov chain has zero, two or more communicating classes.

Example 2. No one is communicating with anyone else, including with herself! To illustrate a discrete-time Markov chain with no communicating class, consider

the infinite state space $\mathcal{S} = \mathbb{N}$ and the probability transition matrix P with the structure

$$P = \begin{bmatrix} 0 & p_{01} & & & \\ & 0 & p_{12} & & \\ & & 0 & p_{23} & \\ & & & \ddots & \ddots \end{bmatrix}, \tag{E.3}$$

that is, for $i \in \mathbb{N}$, $p_{i,j} > 0$ for $j = i + 1$ and $p_{i,j} = 0$ otherwise. □

Definition E.1.4. *The* period d_i *of state* i *is defined as* $d_i = \gcd\{n : [P^n]_{i,i} > 0\}$*, where* gcd *denotes the greatest common divisor.*

If two states i and j communicate, then $d_i = d_j$. Thus, periodicity is a *class property*: all states in the same communicating class have the same period. In the case of an irreducible Markov chain, all states in the state space share a common period, $d = d_i$ for all $i \in \mathcal{S}$.

Definition E.1.5. *We say that an irreducible Markov chain is periodic if* $d > 1$*, and is aperiodic otherwise.*

Here are some definitions.

- State k is accessible from state j, denoted by $j \to k$, if there exists an $n \geq 1$ such that $p_{jk}^{(n)} > 0$. That is, there exists a path $j = i_0, i_1, i_2, \cdots, i_n = k$ such that $p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} > 0$.
- If $j \to k$ and $k \to j$, then states j and k communicate, denoted by $j \leftrightarrow k$.
- State j is called non-essential if there exists a state k such that $j \to k$ but $k \not\to j$.
- State j is called essential if $j \to k$ implies that $k \to j$.
- A state j is an absorbing state if $p_{jj} = 1$. An absorbing state is essential but essential states do not have to be absorbing.

Exercise E.1.6. *Draw a transition diagram and then classify the states of a DTMC with transition matrix*

$$P = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

A state j which is such that $j \not\leftrightarrow j$ is called ephemeral. Ephemeral states usually do not add anything to a DTMC model and we are going to assume that there are no such states.

With this assumption, the communication relation $\leftrightarrow$ has the properties

- $j \leftrightarrow j$ (reflexivity),
- $j \leftrightarrow k$ if and only if $k \leftrightarrow j$ (symmetry), and
- if $j \leftrightarrow k$ and $k \leftrightarrow i$, then $j \leftrightarrow i$ (transitivity).

A relation that satisfies these properties is known as an equivalence relation.

Consider a set S whose elements can be related to each other via any equivalence relation $\Leftrightarrow$. Then S can be partitioned into a collection of disjoint subsets $S_1, S_2, S_3, \ldots S_M$ (where M might be infinite) such that $j, k \in S_m$ implies that $j \Leftrightarrow k$.

So the state space of a DTMC is partitioned into communicating classes by the communication relation $\leftrightarrow$.

An essential state cannot be in the same communicating class as a non-essential state. This means that we can divide the sets in the partition $S_1, S_2, S_3, \ldots S_M$ into a collection of $S_1^n, S_2^n, S_3^n, \ldots S_{M_n}^n$ of non-essential communicating classes and a collection of $S_1^e, S_2^e, S_3^e, \ldots S_{M_e}^e$ of essential communicating classes.

If the DTMC starts in a state from a non-essential communicating class S_m^n then once it leaves, it can never return. On the other hand, if the DTMC starts in a state from a essential communicating class S_m^e then it can never leave it.

If a DTMC has only one communicating class, that is all states communicate with each other, then it is called an *irreducible* DTMC.

Example E.1.7. *Classify the states of the DTMC with*

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0.25 & 0.15 & 0.45 & 0.15 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Exercise E.1.8. *Classify the states of the DTMC with*

$$P = \begin{pmatrix} 0 & 0 & + & 0 & 0 & 0 & + \\ 0 & + & 0 & + & 0 & 0 & + \\ + & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & + & 0 & 0 & 0 \\ 0 & + & 0 & 0 & 0 & 0 & 0 \\ 0 & + & 0 & 0 & + & + & 0 \\ 0 & 0 & + & 0 & 0 & 0 & + \end{pmatrix}$$

We say state j is *periodic* with *period* $d > 1$ if $\{n : p_{jj}^{(n)} > 0\}$ is non-empty and has greatest common divisor d.

If state j has period 1, then we say that it is aperiodic.

Exercise E.1.9.

- *What is the period of the DTMC with* $P = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$?
- *Find the period for the DTMC with* $P = \begin{pmatrix} 0 & 0 & 0.5 & 0.5 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$.

E.1.1 Solidarity Properties

The arguments above bring us to the following theorem, which discusses some solidarity properties of states in the same communicating class.

Theorem E.1.10. *In any communicating class* S_r *of a DTMC with state space* S*, the states are*

- *either all recurrent or all transient, and*
- *either all aperiodic or all periodic with a common period* $d > 1$*.*
- *If states from* S_r *are periodic with period* $d > 1$*, then* $S_r = S_r^{(1)} + S_r^{(2)} + \cdots + S_r^{(d)}$ *where the DTMC passes from the subclass* $S_r^{(i)}$ *to* $S_r^{(i+1)}$ *with probability one at a transition.*

Exercise E.1.11. *Find the classes and properties of the DTMC:*

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Theorem E.1.12. *If an DTMC has finitely-many states, then there must be at least one recurrent state.*

If on the other hand $\mathcal{S}$ is countably infinite then it can be that there are no-recurrent states (instability).

E.2 Poisson's Equation and Generalized Inverses

So you know a few things about Markov Chains on finite state spaces. For example you know that transition probabilities evolve by multiplication by the transition probability matrix, and you know the meaning of powers of that matrix, you know about class

structure, irreducibility, transience, recurrence and periodicity. You also know about the stationary distribution. You understand the basic stuff. You've reached some level of maturity in life. Good for you!

The other thing to really know about is about the behaviour of Markov Reward Processes (MRP). These are processes of the form $\big(X_t,\, r(X_t)\big)$ where X_t is a Markov chain and $r(\cdot)$ is some function of the state space. Reward is the accumulated and possibly it's time average is taken. When dealing with *policy evaluation* for Markov Decision Processes (MDP), understanding how to analyse MRP is critical. The theory for the case of finite state spaces is closed and well understood. This is the subject of these notes.

The main use of these notes is to be an aid for following Chapter 8 of [Put94], "Average Reward and Related Criteria". That chapter relies on Appendix A of the book, which gives a concise treatment of the subject. Specifically the Drazin Inverse and the Deviation Matrix. One complication in the book and it's appendix is that the treatment is general in the sense that it supports Markov chains with several classes (not irreducible) and also periodic Markov chains. This is all nice and good, but for a first reading one may want to assume irreducibly and non-periodicity so as to understand the key concepts without complication. This is what we do in these notes. I.e. the notes summarise the results of Appendix A of [Put94] together with Section 8.2, "Markov Reward Processes and Evaluation Equations", but assume throughout an irreducible and aperiodic Markov chain.

The notes contain exercises. Some of these exercises take you through steps of proofs. As you do that, make sure you also understand the assumptions in the exercises.

E.3 Basics

This material was covered in the lecture notes: "Basic Probability and Markov Chains" (and is covered in many other places also). Here we simply put down the key concepts for the purpose of notation (which differs slightly from the aforementioned notes).

E.3.1 Basic Definitions and Properties

Let $\{X_t,\, t = 0, 1, 2, \ldots\}$ be a sequence of random variables taking values in $\mathcal{S} = \{1, \ldots, N\}$. We say this sequence is a *Time Homogenous Finite State Space Discrete Time Markov Chain* if,

$$\mathbb{P}(X_t = j \,|\, X_{t-1} = s,\, X_{t-2} = j_{t-2}, \ldots, X_0 = j_0) = \mathbb{P}(X_1 = j \,|\, X_0 = s) := p(j\,|s).$$

The matrix, P with s, j'th entry being $p(j \,|\, s)$ is called the *transition probability matrix*. We have that P^m (m'th matrix power) is a matrix with entries, s, j being,

$$p^{(m)}(j \,|\, s) := \mathbb{P}(X_m = j \,|\, X_0 = s).$$

We say that the Markov chain is irreducible if for each pair of states j and s there exists an $m > 0$ such that $p^{(m)}(j \,|\, s) > 0$. In a finite-state space irreducible Markov chain all states are *positive recurrent*. This is typically defined to be the property that the expected return to each state (from itself) is finite.

If the greatest common divisor of $\{m : p^{(m)}(s \,|\, s) > 0\}$ equals 1 for state s then the state is said to have period 1. In an irreducible Markov chain, if this holds for one state, then it holds for all states. In this case we say the Markov chain is *aperiodic*.

In the remainder of these notes we deal with **finite state space** Markov chains that are **irreducible** and **aperiodic**. All statements are based on this assumption. The more general case is covered in [Put94].

E.3.2 The Limiting Matrix

Define the limiting matrix P^* to be with elements,

$$p^*(j \mid s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} p^{(t)}(j \,|\, s) = \lim_{T\to\infty} p^{(T)}(j \,|\, s). \tag{E.4}$$

The first limit is called a *Cesaro limit* it equals the second limit only in a case where the second limit exists (this is the case if the Markov chain is aperiodic). We assume this is the case. The matrix with s, j'th entry being $p^*(j \,|\, s)$ is denoted P^* and is called the *limiting matrix*.

Exercise E.3.1. *Show that P^* satisfies the following equalities:*

$$PP^* = P^*P = P^*P^* = P^*.$$

Exercise E.3.2. *Show that*

1. $(I - P^*)^2 = (I - P^*)$.
2. $P^*(I - P^*) = 0$.
3. P^* *is a stochastic matrix.*

The following theorem describes the *stationary distribution.*

Theorem E.3.3. *The system of equations,*

$$\mathbf{q}'P = \mathbf{q}',$$
$$\mathbf{q}'\mathbf{1} = 1,$$

has a unique positive solution.

Since $P^*P = P^*$, we can write,

$$P^* = \mathbf{1}\mathbf{q}'.$$

That is P^* is an outer product of $\mathbf{1}$ and $\mathbf{q}$. It is a rank one matrix, with all rows equal the stationary distribution $\mathbf{q}'$.

Exercise E.3.4. *Exercises* (E.3.1) *and* (E.3.2) *did not assume* $P^* = \mathbf{1}\mathbf{q}'$. *Carry out the computations in these exercises again under this structure of* P^*.

E.4 The Generalized Inverses

An inverse of a matrix, A is a matrix A^{-1} such that $A\,A^{-1} = I$. For non-square matrices and for singular square matrices, such an inverse does not exist. But once can define the *generalized inverse* (in several ways). This is a big subject in linear algebra. In this chapter, we discuss certain generalized inverses associated with Markov chains.

E.4.1 The Underlying Linear Algebra

The following theorem summarizes a basic property of a stochastic, irreducible, aperiodic matrix (Markov Chain), P. It encompass what is called as the "Perron-Frobenius" theorem. There are also more general versions for irreducible and not necessarily aperiodic Markov chains. Denote by $\sigma(A)$ the spectral radius of the matrix A, this is the maximal modulus of all eigenvalues of A.

Theorem E.4.1. *Under the finite state space, irreducible, aperiodic assumptions, the following hold:*

1. *The value* 1 *is an eigenvalue of* P *with algebraic and geometric multiplicity one and and a single linearly independent eigenvector.*
2. *There exists a non-singular matrix* W *for which,*

$$P = W^{-1} \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} W, \tag{E.5}$$

   where Q *is an* $(N-1) \times (N-1)$ *matrix with the following properties:*

   (a) *It holds that* $\sigma(Q) < 1$ *(so* 1 *is not an eigenvalue of* Q*).*
   (b) *The inverse* $(I-Q)^{-1}$ *exists.*
   (c) $\sigma(I-Q) = \sigma(I-P)$.

3. *The matrix* P^* *is unique and may be represented by,*

$$P^* = W^{-1} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W.$$

One use Theorem E.4.1 is to prove that the limit in (E.4) holds. We follow this proof now through a series of three straight forward exercises:

Exercise E.4.2. *Show that,*

$$P^n = W^{-1} \begin{bmatrix} Q^n & 0 \\ 0 & 1 \end{bmatrix} W,$$

and hence,

$$\frac{1}{T}\sum_{t=0}^{T-1} P^t = W^{-1}\begin{bmatrix} \frac{1}{T}\sum_{t=0}^{T-1} Q^t & 0 \\ 0 & 1 \end{bmatrix} W.$$

Next,

Exercise E.4.3. *Show that since $I - Q$ is non-singular,*

$$\sum_{t=0}^{T-1} Q^t = (I - Q^T)(I - Q)^{-1},$$

and hence since $\sigma(Q) < 1$, Q^T is bounded (in T) so that,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} Q^t = 0.$$

Finally,

Exercise E.4.4. *Show now that,*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} P^t = W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W = P^*.$$

E.4.2 The Drazin Inverse

Take a matrix B that has the representation,

$$B = W^{-1}\begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix} W,$$

where C and W are nonsingular. Then the *Drazin inverse* (or *group inverse*), denoted $B^{\#}$ is defined as,

$$B^{\#} = W^{-1}\begin{bmatrix} C^{-1} & 0 \\ 0 & 0 \end{bmatrix} W.$$

Exercise E.4.5. *Show the following:*

1. $B^{\#}\,B\,B^{\#} = B^{\#}$.
2. $B\,B^{\#} = B^{\#}\,B$.
3. $B\,B^{\#}\,B = B$.

The Drazin inverse is a particular generalised inverse of B (we do not cover more definitions related to generalised inverses here).
We now study the Drazin inverse $(I - P)^{\#}$.

Theorem E.4.6. *The following holds:*

1. *The matrix $(I - P + P^*)$ is non-singular with Z_P denoting it's inverse, i.e.,*
$$Z_P \equiv (I - P + P^*)^{-1}.$$
2. *The Drazin inverse of $(I - P)$ denoted by H_P satisfies,*
$$H_P \equiv (I - P)^{\#} = (I - P + P^*)^{-1}(I - P^*) = Z_P(I - P^*).$$
3. *It holds that,*
$$H_P = \sum_{t=0}^{\infty}(P^t - P^*).$$

We now illustrate the proof of (1) and (2) through through exercises (we skip the proof of (3)):

Exercise E.4.7. *Prove (1) by showing the representation,*
$$I - P + P^* = W^{-1} \begin{bmatrix} I - Q & 0 \\ 0 & 1 \end{bmatrix} W.$$

Exercise E.4.8. *Show that definition of the Drazin inverse implies,*
$$\begin{aligned}(I - P)^{\#} &= W^{-1} \begin{bmatrix} (I - Q)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W \\ &= W^{-1} \begin{bmatrix} (I - Q)^{-1} & 0 \\ 0 & 1 \end{bmatrix} W - W^{-1} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W\end{aligned}$$
and from this (2) follows.

The matrix H_P is referred to as the *deviation matrix*. The matrix Z_P is referred to as the *fundamental matrix*. They are both vaguely referred to as generalized inverses associated with P.

Exercise E.4.9. *Derive the following:*

1. $(I - P)H_P = H_P(I - P) = I - P^*$.
2. $H_P P^* = P^* H_P = 0$.
3. $H_P = Z_P - P^*$.
4. $Z_P P^* = P^*$.
5. $P^* = I - (I - P)(I - P)^{\#}$.

E.5 The Laurent Series

For $\rho > 1$ define the *resolvent* of $P - I$ denoted by R^ρ by,

$$R^\rho = \big(\rho I + (I - P)\big)^{-1}.$$

Letting $\lambda = (1+\rho)^{-1}$, we get,

$$(I - \lambda P) = (1+\rho)^{-1}\big(\rho I + (I - P)\big).$$

When $\lambda \in [0, 1)$, $\sigma(\lambda P) < 1$ so $(I - \lambda P)^{-1}$ exists. Hence for $\rho > 1$, R^ρ exists.

Exercise E.5.1. *Show the following:*

1. $(I - \lambda P)^{-1} = (1+\rho)R^\rho$.
2. $R^\rho = \lambda(I - \lambda P)^{-1}$.

This is the Laurent series expansion for the resolvent:

Theorem E.5.2. *For* $\rho \in \big(0, \sigma(I - P)\big)$,

$$R^\rho = \rho^{-1}P^* + \sum_{n=0}^{\infty}(-\rho)^n H_P^{n+1}.$$

Once again we supply the proof through a series of exercises.

Exercise E.5.3. *Let Q be defined through* (E.5) *and set $B = I - Q$. Then show how to use,*

$$\rho I + I - P = W^{-1}\begin{bmatrix} \rho I + B & 0 \\ 0 & \rho \end{bmatrix} W,$$

to obtain,

$$R^\rho = \rho^{-1}W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W + W^{-1}\begin{bmatrix} (\rho I + B)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W. \tag{E.6}$$

Exercise E.5.4. *Show that the first term of* (E.6) *equals $\rho^{-1}P^*$.*

Exercise E.5.5. *Show that the second term of* (E.6) *equals,*

$$\sum_{n=0}^{\infty}(-\rho)^n W^{-1}\begin{bmatrix} (I-Q)^{-(n+1)} & 0 \\ 0 & 0 \end{bmatrix} W = \sum_{n=0}^{\infty}(-\rho)^n H_P^{n+1}.$$

Do this by first showing that,

$$(\rho I + B)^{-1} = (I + \rho B^{-1})^{-1}B^{-1}$$

E.6 Evaluation of Accumulated/Discounted/Average Reward

We now show how to use some of the tools from above for evaluating the reward in a Markov Chain. While we do not mention a Markov Decision Process (in these notes), think of this as the reward obtained in an MDP with some given decision rule.

E.6.1 The Gain and Bias

Consider now some reward function: $r : \mathcal{S} \to \mathbf{R}$. Since we take $\mathcal{S} = \{1, \ldots, N\}$ it is convenient to denote the vector of values $r(1), \ldots, r(N)$ by $\mathbf{r}$. We are now interested in the *gain* (infinite horizon average cost):

$$g(s) = \lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}_s\Big[\sum_{t=1}^{T} r(X_t)\Big] = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [P^{t-1} r](s) = [P^* r](s) = \mathbf{q}'\,\mathbf{r}.$$

The fact $g(s)$ does not depend on the initial state s is because of the irreducibility assumption (holding throughout these notes). Denote the vector of constant values g, by $\mathbf{g}$

The *bias* (vector) is defined to be,

$$\mathbf{h} := H_P\, \mathbf{r},$$

where H_P is the fundamental matrix.

Exercise E.6.1. *Show that,*

$$\mathbf{h} = \sum_{t=0}^{\infty} P^t(\mathbf{r} - \mathbf{g}),$$

and explain why,

$$h(s) = \mathbb{E}_s\Big[\sum_{t=1}^{\infty} \big(r(X_t) - g(X_t)\big)\Big].$$

The *total reward* (vector) in N time units is,

$$\mathbf{v}_{T+1} = \sum_{t=1}^{T} P^{t-1}\mathbf{r},$$

i.e. $\mathbf{v}_{T+1}(s)$ is the total reward when starting in state s.

Exercise E.6.2. *Show that,*

$$\mathbf{v}_{T+1} = T\mathbf{g} + \mathbf{h} + o(1),$$

where $o(1)$ is a vector with components that approach 0 as $T \to \infty$.

E.6.2 Using the Laurent Series Expansion

We define,

$$\mathbf{v}_\lambda = (I - \lambda P)^{-1}\mathbf{r}.$$

where $\lambda \in [0, 1]$. In fact, $\mathbf{v}_\lambda$ is the expected discounted cost with discount factor λ (this was shown in earlier when studying MDP).
Setting $\rho = (1 - \lambda)\lambda^{-1}$ or alternatively $\lambda = (1 + \rho)^{-1}$, we have

$$\mathbf{v}_\lambda = (1 + \rho)\big(\rho I + [P - I]\big)^{-1}\mathbf{r}.$$

Exercise E.6.3. *Show the following:*
Let ν denote the nonzero eigenvalue of $I-P$ with smallest modulus. Then for $0 < \rho < |\nu|$ (ρ "sufficiently small"),

$$\mathbf{v}_\lambda = (1 + \rho)\Big[\rho^{-1}y_{-1} + \sum_{n=0}^{\infty} \rho^n y_n\Big],$$

where,

$$\begin{aligned} y_{-1} &= P^*\mathbf{r}, \\ y_0 &= \mathbf{g}, \\ y_n &= (-1)^n\, H_P^{n+1}\mathbf{r}, \qquad n = 1, 2, \ldots \end{aligned}$$

As a consequence the following holds:

Exercise E.6.4. *Establish,*

$$\mathbf{v}_\lambda = \frac{1}{1 - \lambda}\mathbf{g} + \mathbf{h} + o(1 - \lambda),$$

where $o(1 - \lambda)$ is a vector that converges to 0 as $\lambda \uparrow 1$.

As a consequence, the following relation between the gain and the discounted reward holds:

Exercise E.6.5. *Establish,*

$$\mathbf{g} = \lim_{\lambda \uparrow 1}(1 - \lambda)\mathbf{v}_\lambda.$$

E.6.3 Evaluation Equations

Computing $\mathbf{g}$ and $\mathbf{h}$ through direct evaluation of P^* and H_P can be done in very specific cases, but is otherwise inefficient. An alternative is using a system of equations (sometimes refereed to as *Poisson's equation* – although not in [Put94]).

Theorem E.6.6. *The following holds,*

$$\mathbf{g} + (I - P)\mathbf{h} = \mathbf{r}. \tag{E.7}$$

Further, if $\mathbf{g}$ *and* $\mathbf{h}$ *are some vectors that satisfy* (E.7) *then* $P^*\mathbf{h} = \mathbf{0}$ *and* $\mathbf{h} = H_P\,\mathbf{r} + \gamma\mathbf{1}$ *for some arbitrary scalar* γ*.*

We establish (E.7) in this exercise.

Exercise E.6.7. *Use now* $P^* + (I - P)H_P = I$ *to establish* (E.7)*.*

A consequence is that (E.7) uniquely determines h up to an element of the null space of $I - P$. This (in the irreducible case) is a space of dimension 1. Thus we we can find the *relative values* $h(j) - h(k)$ by setting any component of $\mathbf{h}$ to 0 and solving (E.7).

Bibliographic Remarks

Exercises

Appendix F

Transforms, Convolutions and Generalized Functions

F.1 Convolutions

F.1.1 Definitions and Applications

Let $f(\cdot)$, $g(\cdot)$ be two functions. The convolution of f and g is the function $\big(f * g\big)(\cdot)$:

$$\big(f * g\big)(t) := \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau.$$

If the functions are of positive support ($= 0$ for $t < 0$) the range of integration in the convolution integral reduces to $\tau \in [0, t]$.

For a probabilist, the convolution is the basic tool of finding the distribution of the sum of two independent random variables X and Y, say with densities $f_X(\cdot)$ and $f_Y(\cdot)$:

$$F_{X+Y}(t) := \mathbb{P}(X + Y \le t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} \mathbb{P}\big((X, Y) \in [x, x+dx) \times [y, y+dy)\big) dy\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f_X(x) f_Y(y) dy\, dx = \int_{-\infty}^{\infty} f_X(x) \Big(\int_{-\infty}^{t-x} f_Y(y) dy \Big)\, dx.$$

So for the density, $f_{X+Y}(t) := \frac{d}{dt} F_{X+Y}(t)$, we have

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x) \Big(\frac{d}{dt} \int_{-\infty}^{t-x} f_Y(x) dy \Big)\, dx = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x) dx = \big(f_X * f_Y\big)(t).$$

Convolution is also defined for discrete time functions (in probability theory this often corresponds to the probability mass function of the sum of two independent discrete random variables):

$$P_{X+Y}(n) = \mathbb{P}\big(X + Y = n\big) = \sum_{k=-\infty}^{\infty} \mathbb{P}\big(X = k\big) \mathbb{P}\big(Y = n - k\big) = \big(P_X * P_Y\big)(n).$$

Note again that if P_X and P_Y are of positive support ($= 0$ for $t < 0$) then the range of summation in the convolution sum reduces to $k \in \{0, \ldots, n\}$.

Another way to view discrete convolutions is as a representation of the coefficients of polynomial products. Denote,

$$A(x) = \sum_{j=0}^{n-1} a_j x^j, \quad B(x) = \sum_{j=0}^{n-1} b_j x^j, \quad C(x) = A(x)B(x) = \sum_{j=0}^{2n-2} c_j x^j.$$

Exercise F.1.1. *Show that* $c_j = \sum_{k=0}^{j} a_k b_{j-k}$.
Note: Assume $a_i, b_i = 0$ *for* $i \notin \{0, \ldots, n-1\}$.

F.1.2 Algebraic Properties

- Commutativity:

$$\big(f * g\big)(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau = \int_{-\infty}^{\infty} f(t-\tau) g(\tau) d\tau = \big(g * f\big)(t)$$

- Associativity:

$$\big(f * g\big) * h = f * \big(g * h\big)$$

- Distributivity:

$$f * (g + h) = f * g + f * h.$$

- Scalar multiplication:

$$\alpha(g * h) = (\alpha g) * h = g * (\alpha h).$$

- Shift/Differentiation:

$$D(g * h) = (Dg) * h = g * (Dh),$$

  where D is either the "delay by one" operator for discrete time or the differentiation operator for continuous time.

Exercise F.1.2. *Show the shift/differentiation property. Do both shift (discrete time) and differentiation (continuous time).*

Sometimes the notation f^{*m} is used for $f * f * \ldots * f$, m times.

If f is a probability density with mean μ and finite variance σ^2, the central limit theorem (CLT) in probability says that as $m \to \infty$, $\frac{f^{*m}(t) - m\mu}{\sqrt{m}\sigma}$ converges to the normal (Gaussian) density:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

F.1.3 Sufficient conditions for existence of the convolution

The *support* of a function f is the (closure of the) set of values for which $f(t) \neq 0$. We often talk about *positive support* if the the support does not contain negative values, and also about *bounded support* if the support is a bounded set..
A continuous time function, f is *locally integrable* if $\int_a^b |f(t)|dt$ exists and is finite for every a, b.

Theorem F.1.3. *The convolution $f_1 * f_2$ in continuous time exists if both signals are locally integrable and if one of the following holds*

1. *Both signals have bounded support.*
2. *Both signals have positive support.*
3. *$||f_1||_2$ and $||f_2||_2$ are both finite.*

Theorem F.1.4. *The theorem above holds for discrete time signals without the locally integrable requirement. In that case the $\mathcal{L}_2$ norms above are taken as ℓ_2 norms.*

F.2 Generalized Functions

Engineering (and mathematics) practice of continuous time signals is often greatly simplified by use of *generalized functions*. The archetypal such function is the *delta*-function denoted by $\delta(t)$, also called *impulse*. This "weird" mathematical object has the following two basic properties:

1. $\delta(t) = 0$ for $t \neq 0$.
2. $\int_{-\infty}^{\infty} \delta(t)dt = 1$.

Now obviously there is no such function $\delta : \mathbb{R} \to \mathbb{R}$, that obeys these two properties if the integral is taken in the normal sense (e.g. Reiman integral). The rigorous description of delta functions is part of the theory of distributions (not to be confused with probability distributions). We shall overview it below informally and then survey a few useful properties of the delta function. First, one should be motivated by the fact that in practice the delta function can model the following:

1. The signal representing the energy transfer from a hammer to a nail.
2. The "derivative" of the unit step function,
3. A Gaussian (normal) density of variance 0.

A more formal (yet not fully rigorous) way to define delta functions is "under the integral sign". It can be thought of as an "entity" that obeys,

$$\int_{-\infty}^{\infty} \delta(t)\phi(t)dt = \phi(0), \tag{F.1}$$

for every (regular) function ϕ that is continuous at 0 and has bounded support (equals 0 outside of a set containing the origin). Entities such as $\delta(t)$ are not regular functions - we will never talk about the "value" of $\delta(t)$ for some t, but rather always consider values of integrals involving $\delta(t)$. Yet from a practical perspective they may often be treated as such.

F.2.1 Convolutions with Delta Functions

The delta function gives a way to represent any signal $u(t)$. Consider the convolution, $\delta * u$:

$$\int_{-\infty}^{\infty} \delta(\tau)u(t-\tau)d\tau = u(t-0) = u(t). \tag{F.2}$$

Thus we see that the δ function is the identity "function" with respect to convolutions:

$$\delta * u = u.$$

In this case a discrete parallel of (F.2) is,

$$\big(\delta * u\big)(\ell) = \sum_{k=-\infty}^{\infty} \delta[k]u(\ell - k) = u(\ell). \tag{F.3}$$

Here $\delta[\ell]$ is the *discrete delta function* (observe the square brackets), a much simpler object than $\delta(t)$ since it is defined as,

$$\delta[\ell] := \begin{cases} 1, & \ell = 0, \\ 0, & \ell \neq 0. \end{cases}$$

Observe that the all elements in the summation of (F.3) are 0 except for possibily the element corresponding to $k = 0$. Thus we have again that $\delta * u = u$. Note that part of the motivation for introducing for the continuous time delta function is to be able to mimic the representation (F.3).

F.2.2 Working with Generalized Functions

We shall soon present other generalized functions related to the delta function. Since such functions are "defined under the integral" sign, two functions $\eta_1(t)$ and $\eta_2(t)$ are equal if,

$$\int_{-\infty}^{\infty} \eta_1(t)\phi(t)dt = \int_{-\infty}^{\infty} \eta_2(t)\phi(t)dt,$$

for a "rich enough class" of functions, $\phi(\cdot)$.

For generalized signals $\eta_1(\cdot)$ and $\eta_2(\cdot)$ and for scalars α_1, α_2, we define the function of the linear combination as,

$$\int_{-\infty}^{\infty} \big(\alpha_1\eta_1 + \alpha_2\eta_2\big)(t)\phi(t)dt = \alpha_1 \int_{-\infty}^{\infty} \eta_1(t)\phi(t)dt + \alpha_2 \int_{-\infty}^{\infty} \eta_2(t)\phi(t)dt.$$

Exercise F.2.1. *Prove that:* $\alpha_1\delta + \alpha_2\delta = (\alpha_1 + \alpha_2)\delta$.

For regular functions $f(\cdot)$ and $\alpha \neq 0$ we have (by a simple change of variables) that,

$$\int_{-\infty}^{\infty} f(\alpha t)\phi(t)dt = \frac{1}{|\alpha|} \int_{-\infty}^{\infty} f(\tau)\phi(\frac{\tau}{\alpha})d\tau.$$

For generalised signals this is taken as the definition of time scaling:

$$\int_{-\infty}^{\infty} \delta(\alpha t)\phi(t)dt = \frac{1}{|\alpha|} \int_{-\infty}^{\infty} \delta(\tau)\phi(\frac{\tau}{\alpha})d\tau = \frac{1}{|\alpha|}\phi(0) = \frac{1}{|\alpha|} \int_{-\infty}^{\infty} \delta(t)\phi(t)dt.$$

Here the first equality is a definition. and the second and third equalities come from the defining equation (F.1). This then implies that

$$\delta(\alpha t) = \frac{1}{|\alpha|}\delta(t).$$

Consider now translation. Take some time shift θ:

$$\int_{-\infty}^{\infty} \delta(t-\theta)\phi(t)dt = \int_{-\infty}^{\infty} \delta(\tau)\phi(\tau+\theta)d\tau = \phi(0+\theta) = \phi(\theta).$$

Hence delta functions translated by θ, denoted $\delta(t-\theta)$ are defined by

$$\int_{-\infty}^{\infty} \delta(t-\theta)\phi(t)dt = \phi(\theta).$$

Consider now what happens when $\delta(t)$ is multiplied by a function $f(t)$ continuous at 0. If $\delta(t)$ was a regular function then,

$$\int_{-\infty}^{\infty} \Big(f(t)\delta(t)\Big)\phi(t)dt = \int_{-\infty}^{\infty} \delta(t)\Big(f(t)\phi(t)\Big)dt = f(0)\phi(0)$$

It is then sensible to define the generalized function, $f(t)\delta(t)$ (for any regular function $f(\cdot)$) as satisfying:

$$\int_{-\infty}^{\infty} \Big(f(t)\delta(t)\Big)\phi(t) = f(0)\phi(0)$$

Hence we have that,

$$f(t)\delta(t) = f(0)\delta(t).$$

This again follows from (F.1).

Exercise F.2.2. *Take τ as fixed and $t \in \mathbb{R}$. Show that,*

$$f(t)\delta(t-\tau) = f(\tau)\delta(t-\tau).$$

Example F.2.3. *A useful generalized function is the so-called "Dirac Comb", also known as "impulse train":*

$$\Delta_T(t) = \sum_{k=-\infty}^{\infty} \delta(t-kT).$$

Here of course one needs to justify the existence of the series (of generalized functions !!!) etc, but this is not our interest.

Impulse trains are very useful for representing the operation of sampling a continuous time (analog) signal. This is done by taking the signal $u(t)$ and multiplying by $\Delta_T(t)$. The resulting signal has values $u(t)$ for $t = kT$, $k \in N$ and 0 elsewhere.

The derivation of the famous Nyquist-Shannon sampling theorem is greatly aided by the impulse train. That theorem says that a "band limited" analog signal $u(t)$ can be perfectly reconstructed if sampled at a rate that is equal or greater than twice its highest frequency.

Related to the delta function is the *unit step function*,

$$\mathbf{1}(t) = \begin{cases} 0, & t < 0, \\ 1, & 0 \le t. \end{cases} \tag{F.4}$$

This is sometimes called the "Heaviside unit function". Other standard notation for it is $u(t)$, but in control theory we typically reserve $u(t)$ for other purpuses (i.e. the input to a system). While it is a function in the regular sense, it can also be defined as a generalized function:

$$\int_{-\infty}^{\infty} \mathbf{1}(t)\phi(t)dt = \int_{0}^{\infty} \phi(t)dt, \tag{F.5}$$

where $\phi(t)$ is any integrable function.

Exercise F.2.4. *Derive (F.4) from (F.5).*

Given a generalized function $\eta(t)$, we define it's *generalized derivate*, $\eta'(t)$ (again a generalized function) by:

$$\int_{-\infty}^{\infty} \eta'(t)\phi(t)dt = -\int_{-\infty}^{\infty} \eta(t)\phi'(t)dt.$$

The above definition applied to $\mathbf{1}(t)$ yields,

$$\int_{-\infty}^{\infty} \mathbf{1}'(t)\phi(t)dt = -\int_{-\infty}^{\infty} \mathbf{1}(t)\phi'(t)dt = -\int_{0}^{\infty} \phi'(t)dt = -\big(\phi(\infty) - \phi(0)\big) = \phi(0).$$

We have just shown that $\mathbf{1}' = \delta$.

Exercise F.2.5. *Show that* $\mathbf{1}'(t-\theta) = \delta(t-\theta)$.

We can also look at the derivative of the delta function:

$$\int_{-\infty}^{\infty} \delta'(t)\phi(t)dt = -\int_{-\infty}^{\infty} \delta(t)\phi'(t)dt = -\phi'(0).$$

This generalized function is sometimes called a *doublet*. Higher order derivatives of a generalized function η are defined by,

$$\int_{-\infty}^{\infty} \eta^{(n)}(t)\phi(t)dt = (-1)^n \int_{-\infty}^{\infty} \eta(t)\phi^{(n)}(t)dt,$$

here $\phi(t)$ needs to be any function from a "suitable" set of test functions. We will not discuss generalized functions in any more depth than covered here. Students interested in functional analysis and related fields can study more about Schwartz's theory of distributions independently.

F.3 Integral and Series Transforms

F.3.1 Laplace Transforms

Let s be a complex number, the Laplace transform of a continuous time function $f(t)$ at the "*frequency*" $f(\cdot)$ is,

$$\mathcal{L}\{f(\cdot)\}(s) = \int_{0^-}^{\infty} e^{-st} f(t)\, dt. \tag{F.6}$$

We shall often denote $\mathcal{L}\{f(\cdot)\}$ by $\hat{f}$. Observe the lower limit to be 0^- and read that as,

$$\lim_{\epsilon \to 0^-} \int_{\epsilon}^{\infty} e^{-st} f(t)\, dt.$$

This is typical "engineering notation" as the function $f(\cdot)$ may sometimes have "peculiarities" at 0. For example may have a generalized function component. In applied probability and other more rigorous mathematical contexts, the Laplace-Stiltijes Transform is often used,

$$\int_{0}^{\infty} e^{-st} dF(t),$$

where the above is a Stiltijes integral. Our Laplace transform, (F.6) is sometimes referred to as the *one-sided Laplace transform*. Whereas,

$$\mathcal{L}_B\{f(\cdot)\}(s) = \int_{-\infty}^{\infty} e^{-st} f(t)dt,$$

is the *bilateral Laplace transform*. The latter is not as useful and important as the former for control purposes. An exception is the case of $s = i\omega$ (pure imaginary) in which case,

$$\hat{f}(\omega) = \mathcal{L}_B\{f(\cdot)\}(i\omega),$$

is the (up to a constant) Fourier transform of f (here we slightly abuse notation by using the "hat" for both Laplace and Fourier transforms). Note that in most engineering text the symbol $i = \sqrt{-1}$ is actually denoted by j.

In probability, the Laplace transform of a density, $f_X(\cdot)$ of a continuous random variable has the meaning, $\mathbb{E}[e^{-sX}]$.

F.3.2 Existence, Convergence and ROC

A function $f(t)$ is said to be of *exponential order* as $t \to \infty$ if there is a real σ and positive real M, T such that for all $t > T$,

$$|f(t)| < Me^{\sigma t}. \tag{F.7}$$

The function e^{t^2} is not of exponential order but most signals used in control theory are.

Exercise F.3.1. *Show that the following is an alternative definition to* exponential order*: There exists a real $\tilde{\sigma}$ such that,*

$$\lim_{t\to\infty} \left|f(t)e^{-\tilde{\sigma}t}\right| = 0.$$

Exercise F.3.2. *Show that any rational function is of exponential order.*

For a function of exponential order, the *abscissa of convergence*, σ_c, is the greatest lower bound (infimum) of all possible values σ in (F.7). Hence for polynomials, $\sigma_c = 0$ while for functions of the form $e^{t\alpha}$ with $\alpha > 0$, $\sigma_c = \alpha$.

Exercise F.3.3. *What is the abscissa of convergence of a rational function $f(t) = \frac{a(t)}{b(t)}$ (here $a(t)$ and $b(t)$ are polynomials and $a(t)$ is of lower degree)?*

Theorem F.3.4. *Functions $f(t)$ that are locally integrable and are of exponential order with σ_c have a Laplace transform that is finite for all $Re(s) > \sigma_c$.*

The regieon in the complex *s-plane*: $\{s : Re(s) > \sigma_c\}$ is denoted the *regieon of convergence* (ROC) of the Laplace transform.

Proof.

$$|\hat{f}(s)| = \left|\int_{0^-}^{\infty} e^{-st} f(t)dt\right| \leq \int_{0^-}^{\infty} \left|e^{-st}\right|\left|f(t)\right|dt.$$

Writing $s = \sigma + i\omega$ we have $|e^{-st}| = e^{-\sigma t}$, so for all $\sigma' > \sigma_c$

$$|\hat{f}(s)| \leq M\int_{0^-}^{\infty} e^{-\sigma t}e^{\sigma' t}dt = M\int_{0^-}^{\infty} e^{-(\sigma-\sigma')t}dt.$$

This integral is finite whenever $\sigma = Re(s) > \sigma'$. Now since σ' can be chosen arbitrarily close such that $\sigma' > \sigma_c$ we conclude that the transform exists whenever $\sigma > \sigma_c$. □

F.3.3 Uniqueness

Laplace transforms uniquely map to their original "time-functions". In fact, this is the inversion formula:

$$f(t) = \lim_{M\to\infty} \frac{1}{2\pi i} \int_{\sigma-iM}^{\sigma+iM} e^{st}\hat{f}(s)ds,$$

for any $\sigma > \sigma_c$. The integration is in the complex plane and is typically not the default method.

Exercise F.3.5. *Optional (only for those that have taken a complex analysis course). Apply the inversion formula to show that,*

$$\mathcal{L}^{-1}\Big(\frac{1}{(s+a)^2}\Big) = te^{-at}.$$

F.3.4 Basic Examples

Example F.3.6. *The Laplace transform of $f(t) = c$:*

$$\mathcal{L}(c) = \int_0^\infty e^{-st}c\,dt = \lim_{T\to\infty}\int_0^T e^{-st}c\,dt = \lim_{T\to\infty}\Big[-\frac{c}{s}e^{-st}\Big]_0^T = \frac{c}{s}\big(1 - \lim_{T\to\infty}\ e^{-sT}\big).$$

When does the limit converge to a finite value? Take $s = \sigma + i\omega$,

$$\lim_{T\to\infty} e^{-sT} = \lim_{T\to\infty}\ e^{-\sigma T}(\cos\omega T + i\sin\omega T).$$

So we need $\sigma > 0$ to get $\lim_{T\to\infty}\ e^{-sT} = 0$, hence,

$$\hat{f}(s) = \frac{c}{s}, \quad Re(s) > 0.$$

Exercise F.3.7. *Show that the transform of $f(t) = e^{\alpha t}$ is,*

$$\hat{f}(s) = \frac{1}{s-\alpha}, \quad Re(s) > Re(\alpha).$$

Exercise F.3.8. *Derive the Laplace transform (and find ROC) of*

$$f(t) = e^{-at}\cos(bt).$$

For other examples see a *Laplace transform table.*

F.3.5 Basic Properties

You should derive these.

- Linearity:
$$\mathcal{L}\big(\alpha_1 f_1(t) + \alpha_2 f_2(t)\big) = \alpha_1 \hat{f}_1(t) + \alpha_2 \hat{f}_2(t).$$
- Time shift:
$$\mathcal{L}\big(f(t-\theta)\big) = \int_{0^-}^{\infty} f(t-\theta)e^{-st}dt = \int_{0^-}^{\infty} f(t)e^{-s(t+\theta)}dt = e^{-s\theta}\hat{f}(t).$$
- Frequency shift:
$$\mathcal{L}\big(e^{-at}f(t)\big) = \hat{f}(s+a).$$
- Time Scaling:
$$\mathcal{L}\big(f(at)\big) = \frac{1}{|a|}\hat{f}\big(\frac{s}{a}\big).$$
- Differentiation:
$$\mathcal{L}\big(f'(t)\big) = \int_{0^-}^{\infty} f'(t)e^{-st}dt = f(t)e^{-st}\big|_0^{\infty} + s\int_{0^-}^{\infty} f(t)e^{-st}dt = -f(0^-) + s\hat{f}(s).$$
- Integration:
$$\mathcal{L}\big(\int_0^t f(x)dx\big) = \frac{1}{s}\hat{f}(s).$$

More basic properties are in one of tens of hundreds of tables available in books or on the web.

F.3.6 Relation To Differential Equations

The differentiation formula allows to transform differential equations into algebraic equations for s. Then the equations may be solved in the *s-plane* and transformed back to obtain the solutions.

Exercise F.3.9. *Solve using the Laplace transform:*

$$\ddot{x}(t) + 6x(t) = \cos\big(\frac{t}{2}\big),$$

with $x(0) = 0$, $\dot{x}(0) = 0$.

F.3.7 Relation To Convolution

This property is very important:

$$\mathcal{L}\big(f_1(t) * f_2(t)\big) = \hat{f}_1(s)\hat{f}_2(s).$$

Exercise F.3.10. *Prove it.*

F.3.8 Rational Laplace Transforms and Partial Fraction Expansion

Often Laplace (as well as Fourier and Z) transforms are of the rational form,

$$\hat{f}(s) = \frac{p(s)}{q(s)} = \frac{p_m s^m + \ldots + p_1 s + p_0}{q_n s_+^n \ldots\ldots\ldots + q_1 s + a_0},$$

with p_i, q_i either real or complex coefficients (we mostly care about real coefficients) such that, $p_m, q_n \neq 0$. The function $\hat{f}(\cdot)$ is called *proper* if $m \leq n$, *strictly proper* if $m < n$ and *improper* if $m > n$.
If $\hat{f}(s)$ is not strictly proper, then by performing *long division* it may be expressed in the form,

$$r(s) + \frac{v(s)}{q(s)},$$

where $r(s)$ is a polynomial of degree $m - n$ and $v(s)$ is a polynomial of degree $< n$.

Exercise F.3.11. *Carry long division out for,*

$$\hat{f}(s) = \frac{s^4 + 2s^3 + s + 2}{s^2 + 1},$$

to express it in the form above.

The action of performing *partial fraction expansion* is the action of finding the coefficients A_{ik} such that a strictly proper $\hat{f}(\cdot)$ in the form,

$$\hat{f}(s) = \sum_{i=1}^{K} \Big(\sum_{k=1}^{m_i} \frac{A_{ik}}{(s - s_i)^k} \Big),$$

where $s_1, \ldots, s_K$ are the distinct real or complex roots of $q(s)$, and the multiplicity of root s_i is m_i.
After carrying out long division (if needed) and partial fraction expansion, $\hat{f}(s)$ may be easily inverted, term by term.

Example F.3.12. *Consider,*

$$\hat{f}(s) = \frac{1}{s^2+3s+2} = \frac{1}{(s+1)(s+2)}.$$

We want the form,

$$\hat{f}(s) = \frac{A_{11}}{s+1} + \frac{A_{21}}{s+2}.$$

This to equation,

$$1 = A_{11}(s+2) + A_{21}(s+1). \tag{F.8}$$

or,

$$1 = (A_{11}+A_{21})s + (2A_{11}+A_{21}). \tag{F.9}$$

Now "identity coefficients of terms with like powers of s" to get a set of linear equations:

$$\begin{aligned} A_{11}+A_{21} &= 0 \\ 2A_{11}+A_{21} &= 1 \end{aligned}$$

to get $A_{11}=1$ and $A_{21}=-1$.

Example F.3.13. *Consider,*

$$\hat{f}(s) = \frac{s-1}{s^3-3s-2} = \frac{s-1}{(s+1)^2(s-2)}.$$

We want the form,

$$\hat{f}(s) = \frac{A_{11}}{s+1} + \frac{A_{12}}{(s+1)^2} + \frac{A_{21}}{s-2}.$$

Similar to before, we may get a system of equations for the A_{ik}.

Exercise F.3.14. *Complete the partial fraction expansion of the above example.*

When the coefficients of $q(\cdot)$ are real, the roots are complex conjugate pairs (say with multiplicity m_i). In this case we may write for any pair of roots, s_i and $\overline{s_i}$,

$$(s-s_i)(s-\overline{s_i}) = s^2 + a_i s + b_i,$$

where a_i and b_i are real coefficients. In this case, the partial fraction expansion is of the form,

$$\hat{f}(s) = \ldots\ldots + \frac{B_{i1}s+A_{i1}}{s^2+a_is+b_i} + \frac{B_{i2}s+A_{i2}}{(s^2+a_is+b_i)^2} + \ldots + \frac{B_{im_i}s+A_{im_i}}{(s^2+a_is+b_i)^{m_i}} + \ldots\ldots.$$

A similar technique may be used to find the B's and A's.

Exercise F.3.15. *Carry out a partial fraction expansion for,*

$$\hat{f}(s) = \frac{s+3}{(s^2+2s+5)(s+1)}.$$

F.3.9 The Fourier Transform in Brief

The Fourier transform of $f(t)$ is:

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt.$$

The inverse fourier transform is,

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(\omega)e^{i\omega t}dw.$$

Exercise F.3.16. *Find the Fourier transform of* $f(t) = \frac{\sin(t)}{t}$.

F.3.10 Conditions for convergence:

Theorem F.3.17. *A sufficient condition for convergence of the Fourier integral is that* $f(\cdot)$ *satisfies the following:*

- $\int_{-\infty}^{\infty} \big|f(t)\big|dt < \infty$.
- $f(\cdot)$ *has a finite number of maxima and minima in any finite interval.*
- $f(\cdot)$ *has a finite number of discontinuities within any finite interval. Furthermore each of these discontinuities must be finite.*

By means of generalized functions, the Fourier transform may also be defined (and convergences) for periodic functions that are not absolutely integrable.

F.3.11 Basic Properties

Many properties are very similar to the Laplace transform (the Fourier transform is a special case of the bilateral Laplace transform).
Some further important properties are:

- The transform of the product $f_1(t)f_2(t)$ is $\big(\hat{f}_1 * \hat{f}_2\big)(\cdot)$. This has far reaching implications in signal processing and communications.
- Parseval's Relation (energy over time = energy over spectrum):

$$\int_{-\infty}^{\infty} \big|f(t)\big|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \big|\hat{f}(\omega)\big|^2 d\omega.$$

F.3.12 Graphical Representaions

Plots of $\left|\hat{f}(\omega)\right|$ and $\angle\hat{f}(\omega)$ are referred to by engineers as *Bode plots*. It is typical to stretch the axis of the plots so that the horizontal axis is $\log_{10}(\omega)$ and the vertical axis are $20\log_{10}\left|\hat{f}(\omega)\right|$ and $\angle\hat{f}(\omega)$. There is a big tradition in engineering to generate approximate bode plots by hand based on first and second order system approximations. An alternative plot is the *Nyquist plot*

Exercise F.3.18. *Generate a Bode and a Nyquist plot of a system with transfer function,*

$$H(s) = \frac{1}{s^2 + s + 2}.$$

F.3.13 The Z Transform in Brief

This is the analog of the Laplace transform for discrete time functions, $f(\ell)$. The *Z-transform* is defined as follows,

$$\hat{f}(\ell) = \sum_{k=-\infty}^{\infty} f(\ell) z^{-\ell}.$$

Many of the things we do for continuous time using the Laplace transform may be done for discrete time using the Z-transform. We will not add further details in this unit, but rather touch discrete time systems when we talk about general (MIMO) linear systems.

Bibliographic Remarks

Exercises