Channel Estimation and Prediction from a Bayesian Perspective

Daniel Aronsson

May 2007

DEPARTMENT OF ENGINEERING SCIENCES
UPPSALA UNIVERSITY
UPPSALA, SWEDEN

Submitted to the Faculty of Science and Technology, Uppsala University in partial fulfillment of the requirements for the degree of Licentiate of Technology.

Printed in Sweden by Universitetstryckeriet, Uppsala University, 2007

Till mina tjejer

Abstract

Digital communications systems require the receiver to estimate the transmitted bit sequence from a noisy received signal. Estimation is therefore a crucial part in digital communications. Prediction of error rates, on the other hand, is not, but it enables capacity improving techniques in the form of fast link adaptation and opportunistic resource scheduling.

In this thesis, solutions to the estimation and prediction problems are proposed by inferring radio channels that vary rapidly due to the mobility of users. It is crucial not only to produce point estimates and predictions of the channels, but also to take the uncertainty of those estimates into account.

This thesis adopts the Bayesian probability interpretation, which regards probability theory an extension to logic. Orthodox statistics, which considers a probability to be a limiting frequency of an imagined experiment, will in many cases produce only point estimates, whereas the Bayesian method also always produces measures of uncertainty.

Linear state space models are designed for a number of system types, and the Kalman filter is used to infer the time-variant radio channels.

The proposed channel predictor is evaluated on a specific system proposal. It is found that control data aiding the channel estimation and prediction (so called pilot data) should be transmitted simultaneously by all users, and that the distribution pattern of pilot symbols should also be varied over time, in order to achieve a high prediction performance.

Two methods for predicting the bit error rate are proposed. It is shown that although the associated mathematical expressions are somewhat involved, the numerical complexity induced by those is negligible compared to the complexity of the channel predictor.

It is also suggested how the proposed algorithms may be used for evaluation and design of wireless multiuser systems.

Acknowledgments

A number of people deserves to be acknowledged:

First of all, my supervisors Prof. Mikael Sternad and Prof. Anders Ahlén, for encouragement and support,

the Bayesians at the Signals and Systems Group – my brothers-in-arms – Erik Björnemo, Mathias Johansson, and Tomas Olofsson, for endless intriguing discussions, and for introducing me to a field that shows what science really *is*,

all present and former colleagues at the Signals and Systems group, for good friendship and for providing a relaxed work atmosphere,

all the members of the Wireless IP project for insight and inspiration,

my parents and my sisters with families for their love and support,

and finally, and especially, Sara and our daughter Linn for their endless love and sunshine.

Contents

Chapter 1

Introduction

A world, in which the demand for mobility and accessibility is ever-increasing, requires improved technologies for enabling high data rates for mobile wireless users. The topology of *multiuser systems* takes a few forms of which the *cellular* system is the most popular; a few *base stations*, wired to one another and to other infrastructures via a core network, serve many *mobile wireless units* or *terminals* (or simply *mobile users* or *mobiles*). The system is so called because the geography is divided into *cells*, each cell being served by one base station. Wireless multiuser networks present many challenges mainly due to the fact that the radio channel qualities experienced by users vary as users move about. Efficient processing of information has always been called for to ensure good quality of received data, but in later years *adaptivity* has become a central concept in the development of wireless systems. An adaptive system continuously draws conclusions about the channel quality and may thereby use different transmission techniques depending on the current quality. In a multiuser system, the adaptivity is added to by introducing *scheduling*. A scheduler, centralised at the base station, then assigns resources to users based on needs and channel qualities, hence taking advantage of the channel quality diversity.

To allow for efficient scheduling, the scheduler needs to know the channel quality estimates of each mobile user in the system. Since these estimates are most conveniently produced at the terminals, they must be signalled back to the base station at regular intervals. Based on these estimates, the base station then take resource allocation decisions which it signals back to the terminals as special feedback information signals. This *control loop* takes a certain amount of time, typically on the scale of microseconds, which

requires the mobile terminals to *predict* the channel, so that the reported channel qualities will be valid at the time of transmission.

The channel quality diversity caused by the mobility of the users is what potentially makes adaptivity and scheduling beneficial. Rapidly varying channels however also makes prediction a difficult problem. Moving through a wave pattern with a wavelength on the order of a few decimeters means that one will experience fading of a few tens or a few hundred Hertz, depending on the exact wavelength and the exact velocity. The channel quality predictor, predicting a few milliseconds ahead in time, therefore needs to be able to predict the channel on the order of fractions of a wavelength, or at most one wavelength.

The focus of the present thesis will be on the estimation and prediction of the radio channel quality necessitated by the adaptivity functionality. The question of how to schedule resources in an optimal manner is outside the scope of this thesis. Instead I present algorithms for processing the information that necessarily forms a basis for such optimal scheduling decisions.

Probability theory offers optimal processing of information. The theory has over the years divided into two rivaling approaches. In this thesis I will take the so called *Bayesian* standpoint. Since it is the other approach, *frequentist* probability theory, that is prevalent in channel estimation and prediction theory, I will go through some lengths to explain the idea and method of Bayesian probability theory.

As the present work is intended both for those who have previous knowledge in channel estimation/prediction and want to see a Bayesian discourse on the subject, and for those who are already acquainted with Bayesian probability theory and would like to see a specific application of the theory, I presume no previous knowledge in either field. Relevant concepts and terminology will be introduced when needed.

1.1 Outline of the thesis

Chapter 2

Chapter 2 is entirely devoted to probability theory. The aim is to give the reader a thorough understanding of the Bayesian method, not only to make later chapters comprehensible, but also to provide the method for general scientific inference. I also provide some comparisons between Bayesian theory and frequentist theory, the latter of which is prevalent within the field of channel estimation and prediction. The comparisons focus mainly on actual results. Ideological differences are considered to be of subordinate impor-

tance.

The two main tools in Bayesianism – Bayes' theorem and marginalisation – are presented. A few methods for assigning so called prior probabilities are treated, as is the general method for producing point estimates.

Bayesianism does not equate probability with frequency. In this chapter the relation between the two concepts is investigated for certain circumstances. Understanding of the probability/frequency correspondence is especially important in the context of channel prediction, since a central measure of radio channel quality, the *bit error rate*, is a frequency.

Bayesian theory comes in many flavours (as does frequentist theory). The viewpoint taken in this thesis is that propounded by Edwin Thompson Jaynes in his book *Probability Theory – The Logic of Science*.

Chapter 3

The modelling of a radio channel usually amounts to modelling several individual *taps*. Chapter 3 is restricted to the modelling of one single tap. I distinguish between *single carrier systems* and *multi carrier systems* and describe how the respective systems are accurately modelled.

The estimation and prediction problems are formulated, and it is concluded that both are solved by inferring the channel taps. The construction of a linear model is outlined, which makes it possible in later chapters to conduct optimal inferences when known so called *pilot information* is transmitted over the radio link. To exemplify the modelling procedure, a specific model is designed step by step. This particular model is chosen so that it assigns equal probability to all frequencies of the fading.

The use of a linear model is suboptimal in the sense that it does not allow any uncertainty of the model parameters. Chapter 3 sketches the principle of the suboptimal linear estimator/predictor and compares it to the optimal Bayesian solution.

Chapter 4

The one tap model derived in Chapter 3 is here extended to a model representing several fading taps (impulse responses) at several simultaneous users. Many different types of system are studied; uplinks and downlinks of single carrier as well as multicarrier systems. It is shown how OFDM systems may be modelled in several different ways, which way to choose depending on resulting numerical complexity.

Correct scaling of the process noise covariance matrix so as to produce a

given process covariance is generally a very complex problem when setting up state spaces. Chapter 4 gives a closed-form expression for the process noise covariance matrix which applies to the special model structures studied here.

A linear model of time-varying channel taps is only motivated if model parameters such as velocity, fading descriptors, noise variance, and so on can be estimated with high accuracy. "Off-line" estimators, optimal or sub-optimal, that operates beside the linear model hence have to be used. Such estimators are briefly discussed at the end of Chapter 4.

Chapter 5

Here the models constructed in previous chapters are used to conduct near-optimal inferences about channel taps. The purpose of this is twofold. The first is actual applicability; by inferring channel taps, the estimation and prediction problems presented in Chapter 3 can be solved.

The prospect of the derived algorithms to be implemented in a real system is however restricted by the numerical complexity imposed on the hardware. The complexity is therefore investigated for a few different scenarios depending on which assumptions are made. It is shown that the use of the proposed algorithm is potentially feasible within a few years, if the number of data points measured at each sampling instant is kept low.

We study channel estimation and prediction performance in a specific system, and conclude that users should use overlapping pilot signals rather than taking turns in sending exclusive pilot signals. The benefits of using time-varying pilot patterns is also stressed. We finally conclude that prediction performance depend heavily on the fading statistics of the channel taps. Correct estimation of fading properties is therefore of ample importance.

In channel prediction research it is common to produce predictions of the squared channel magnitude. It is shown how optimal squared channel magnitude estimates are produced with the present algorithms, but it is also shown that such estimates correspond poorly with estimates of the bit error rate, which should be more suited to underlie the scheduling decisions.

Two approaches to bit error rate prediction are presented: expected bit error rate and probability of bit error rate. The question of which one to use is left open.

The second purpose for inferring channel taps is to show how the presented algorithms may be used for system evaluation and design. Probability theory describes how uncertainty propagates through a chain of related variables. Instead of examining one specific scenario, exactly defined by one particular channel realisation and one particular set of transmitted data (as is the case

when a simulation is carried out), we may therefore use probability theory to evaluate how a certain system will perform given that the radio channels experienced by users are accurately described by specific models. Chapter 5 sketches how such an evaluation process could be carried out.

Chapter 6

Chapter 6 discusses potential extensions to the present work.

1.2 Contributions

The material in this thesis has partly been published in

- D. Aronsson and M. Sternad, "OFDMA Uplink Channel Prediction to Enable Frequency-Adaptive Multiuser Scheduling", *European Signal Processing Conference*, Poznan, Poland, September 2007
- D. Aronsson and M. Sternad, "Kalman Predictor Design for Frequency-Selective Scheduling of FDD OFDMA Uplinks", Submitted to *International Symposium on Personal, Indoor and Mobile Radio Communications*, Athens, Greece, September 2007
- M. Sternad, S. Falahati, T. Svensson and D. Aronsson, "Adaptive TDMA/OFDMA for Wide-Area Coverage and Vehicular Velocities", *IST Summit*, Dresden, Germany, June 2005
- M. Sternad and D. Aronsson, "Channel Estimation and Prediction for Adaptive OFDMA/TDMA Uplinks based on Overlapping Pilots", *IEEE International Conference on Audio, Speech, and Signal Processing*, Philadelphia, USA, March 2005
- M. Sternad and D. Aronsson, "Channel Estimation and Prediction for Adaptive OFDM Downlinks", *IEEE Vehicular Technology Conference*, Orlando, USA, October 2003

Chapter 2

Bayesian inference

In this chapter I will introduce the reader to Bayesian probability theory. The aim is to give the reader a thorough understanding of the Bayesian framework, not only to make later chapters comprehensible, but also to provide the method for general scientific inference according to the Bayesian school. I provide some comparisons between Bayesian theory and orthodox (frequentist) theory, since the latter is prevalent within the field of channel estimation and prediction. The comparisons focus mainly on actual results.

The two main tools in Bayesianism – Bayes' theorem and marginalisation – are presented. A few methods of assigning so called prior probabilities are treated, as is the general method for producing point estimates.

Orthodox probability theory defines probabilities as observed limiting frequencies, but this is not the case in Bayesian theory, where a probability represents a state of knowledge. Under some circumstances there are however a strong correspondence between frequencies and probabilities as defined by Bayesianism. This is especially important to understand in the context of channel prediction, since a central measure of radio channel quality, the *bit error rate*, is a frequency.

There are many different kinds of Bayesianism. The specific viewpoint taken in this thesis is that advocated by Edwin Thompson Jaynes in his book *Probability Theory – The Logic of Science*[1].

2.1 The probability concept

The concept of probability is ubiquitous in all science. To some extent, measurements and observations are always subject to error, and so we have

to take uncertainty into account when we draw conclusions from data. In this thesis I will discuss the process of making inferences from noisy data about unknown radio channels and system performance in a digital communications system. How to formalise a procedure for inference, that in some way is optimal, has puzzled many great thinkers for at least the past two centuries. Their collected ideas have essentially condensed into two camps. To see what characterises the respective camp, we will take a look at how a few different authors in the field of probability theory have chosen to *define* probability.

William Feller [2, pp 4–5] exposes his view on what probability theory is about and what a probability really is:

> "[...] we are not concerned with the accidental circumstances of an actual experiment: the object of the theory is sequences (or arrangements) of symbols [...] Before speaking of [a probability] we should have to agree on an (idealized) model which would presumably run along the lines 'out of infinitely many worlds one is selected at random....' "

He continues with explicitly stating what he means when he talks about a probability:

> "[...] we shall not worry whether or not our conceptual experiment can be performed; we shall analyze abstract models. [...] We *imagine* the experiment performed a great many times. An event with probability 0.6 should be expected, in the long run, to occur sixty times out of a hundred."

Ronald A. Fisher [3, pp 34–35] is of roughly the same opinion when he attempts to clear up old misunderstandings. He exposes his view on probability hence:

> "Indeed, I believe that a rather simple semantic confusion may be indicated as relevant to the issues discussed, as soon as consideration is given to the meaning that the word probability must have to anyone so much practically interested as is a gambler, who, for example, stands to gain or lose money, in the event of an ace being thrown with a single die. To such a man the information supplied by a familiar mathematical statement such as: "If a aces are thrown in n trials, the probability that the difference in absolute value between a/n and $1/6$ shall exceed any positive value ε,

> however small, shall tend to zero as the number n is increased indefinitely", will seem not merely remote, but also incomplete and lacking in definiteness in its application to the particular throw in which he is interested. Indeed, by itself it says nothing about that throw. It is obvious, moreover, that many subsets of future throws, which may include his own, can be shown to give probabilities, in this sense, either greater or less than 1/6. Before the limiting ratio of the whole set can be accepted as applicable to a particular throw, a second condition must be satisfied, namely that before the die is cast no such subset can be *recognized*. This is a necessary and sufficient condition for the applicability of the limiting ratio of the entire aggregate of possible future throws as the probability of any one particular throw. On this condition we may think of a particular throw, or of a succession of throws, as a *random* sample from the aggregate, which is in this sense subjective homogeneous and without recognizable stratification."

Bruno de Finetti [4, pp 73–74] speaks of four different kinds of probability interpretations and declares himself in favour of what he calls the subjectivistic approach :

> "*The subjectivistic approach* [...] considers probability a measure of the degree of belief of a given subject in the occurrence of an event (proposition)."

De Finetti also refers to earlier proponents of the same approach to probability theory, among which Harrold Jeffreys was one of the stronger. Jeffreys [5, p. 20] writes:

> "We can now introduce the formal notation $\mathrm{P}(q|p)$ for the number associated with the probability of the proposition q on data p; it may be read 'the probability of q given p' provided that we remember that the number is not in fact the probability, but merely a representation of it in terms of a pair of conventions. The probability, strictly, is the reasonable degree of confidence and is not identical with the number used to express it."

That the general definition of probability is "reasonable degree of confidence" was also the opinion of George Pólya. In [6, p. 58], Pólya restricts himself to study experiments that allow themselves to be repeated under the same conditions over and over again – a case which he calls *random mass*

phenomena. The definition of probability he then uses seems very similar to that propounded by Feller and Fisher:

> "We have to consider the theoretical value of long range relative frequency and we shall call this theoretical value probability."

However, later ([7, pp 116–117]) when he expands the area of application and looks at general plausible reasoning, he writes:

> "We used the symbol $\Pr(A)$ to denote the probability of the event A, that is, the theoretical value of the long range relative frequency of the event A. In the present chapter, however, we have to deal with plausible reasoning. We consider some conjecture A, and we are concerned with the reliability of this conjecture A, the strength of the evidence in favor of A, our confidence in A, the degree of credence we should give to A, in short the *credibility of the conjecture* A. We shall take the symbol $\Pr(A)$ to denote the credibility of A."

Pólya then investigates whether the laws used in the case of random mass phenomena also apply in general plausible reasoning, or whether there is in fact an ambiguity in using the same symbol for denoting both probability and credibility (with Pólya's definitions). By carefully investigating its consequences, he finds that the same rules may in fact be applied in both cases.

From the above attempts to define probability, it is clear that Feller and Fisher belong to a camp that regards probabilities to be imagined limiting frequencies. We shall call this approach the *frequentist* school. Jeffreys, de Finetti and Pólya, on the other hand, look at probability as a general measure of belief. This approach will be called the *Bayesian* view. It is so called because one of its central theorems was introduced in 1763 by reverend Thomas Bayes as a solution to a problem concerning what has historically been called *inverse probability*. However, the idea of regarding probability a general degree of confidence should rightfully be attributed to Pierre-Simon Laplace who in 1814 introduced it in his *Essai philosophique sur les probabilités*. In this thesis I will take the Bayesian standpoint.

The progress of probability theory has since long been plagued by controversies between the two sides; Bayesianists accusing frequentists of producing erroneous results, and frequentists accusing Bayesianism for unsound ideology. Frankly, judging from the few examples given above, one can see why

the latter is; it may not seem very clear why there would exist rules to which something as seemingly ethereal as 'degree of belief' would have to conform.

Therefore a milestone was set in 1946 when Richard T. Cox [8] showed that such rules can be derived by just attaching a very few qualitative requirements on such a measure of belief. Cox's exposition helps clarify what the present definition of probability is, and the rules hence derived constitute the foundation for probability theory. We will look at the discoveries made by Cox below.

Bayesian probability theory has in later years been developed and refined to a great extent by Edwin Thompson Jaynes. A lifetime of thorough work in the Bayesian field is summarised in his excellent book *Probability Theory – The Logic of Science*[1]. The thoughts and ideas expressed in this book permeates the present thesis.

2.2 The Bayesian definition of probability

Probability theory operates on *propositions*, which is denoted by capital letters. Propositions can be of any general type, for example

A The measured voltage is between 1.5 and 2.0 Volts.

B The bridge will hold for normal stresses.

C You dress nicely.

are all valid propositions.

Cox [8] introduced the concept of *plausibility*[1]. Plausibility is a general measure of degree of belief. The plausibility of a specific proposition will vary depending on what other propositions we know to be true. For example, it is more plausible that it is freezing outside if we know that it is winter, than if we are ignorant to the time of year. In accordance with a notation introduced by John Maynard Keynes in 1921, I will denote plausibilities by

$$A|B, \tag{2.1}$$

meaning the plausibility of proposition A given that proposition B is true.

Cox aimed to show that rules for probability theory interpreted in the Bayesian sense – rules that by the time of Cox already had been employed

[1] To be precise, Cox really used the term *likelihood*, but today this notion is used for another purpose as we shall see shortly.

as axioms by several generations of workers in the field – could be derived from the axioms of classical logic[2] only by adding a few 'common sense' requirements.

These requirements were stated as functional relationships between plausibilities. Jaynes chose to reformulate these as verbal statements, making the exposition easier to follow. I will adopt his view here, although I leave out the actual derivation.

Jaynes uses three *desiderata*, starting with

(I) *Degrees of plausibility are represented by real numbers.*

I will adopt the convention that higher numbers correspond to higher plausibilities, without further specifying the exact relationship. Jaynes discusses other possibilities for constructing a theory for plausible reasoning, where this desideratum is not needed. Instead he replaces desideratum (I) with two more elementary ones,

(Ia) If $(A|X) \geq (B|X)$ and $(B|X) \geq (C|X)$ then $(A|X) \geq (C|X)$, and

(Ib) Given A, B, C, one of $(A|C) > (B|C)$, $(A|C) = (B|C)$, $(A|C) < (B|C)$ must hold,

and argues that any useful theory must be analogous to one that associates plausibility with real numbers, so that we might just as well accept desideratum (I).

The second desideratum is concerned with how plausibilities change when new data are obtained. If old information C is updated to new information C' so that the plausibility for A is increased,

$$(A|C') > (A|C), \tag{2.2}$$

while the plausibility for B stays the same,

$$(B|AC') = (B|AC), \tag{2.3}$$

then common sense says that

$$(AB|C') \geq (AB|C), \tag{2.4}$$

[2]Classical logic is usually formulated as six axioms expressing true/false relationships between propositions (see e.g. [8]). They may be compressed to fewer axioms, but at the expense of clarity.

and that

$$(\overline{A}|C') < (\overline{A}|C), \tag{2.5}$$

where $\overline{A}$ denotes the logical complement of A, that is the proposition that is always true when A is false and vice versa. The above 'common sense' requirements are expressed by desideratum II :

(II) *Qualitative correspondence with common sense.*

The third desideratum is divided into three statements, all having to do with the consistency of the theory :

(IIIa) *If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.*

(IIIb) *We must always take into account all of the evidence available that is relevant to the problem.*

(IIIc) *Equivalent states of knowledge must always be represented in the same way.*

Surprisingly, these three desiderata are all that is needed to derive a consistent theory for plausible reasoning. Although the plausibility measure is quite arbitrary, the derivation reveals that there must exist relationships between *functions* operating on plausibilities. One such example turns out to be

$$\mathrm{P}(AB|C) = \mathrm{P}(A|C)\mathrm{P}(B|AC) = \mathrm{P}(B|C)\mathrm{P}(A|BC) \quad \text{(The product rule)}$$
$$\mathrm{P}(A|B) + \mathrm{P}(\overline{A}|B) = 1 \quad \text{(The sum rule)}$$

The function $\mathrm{P}(A|B)$ is termed *the probability of A given B*. It has the additional property that $\mathrm{P}(\text{'truth'}) = 1$ and $\mathrm{P}(\text{'falsity'}) = 0$.

We now summarise what probability means in the present theory:

> **Plausibility** is a measure of belief isomorphic to the real numbers, so that the plausibility can be either increased, decreased or unaltered by new information.
> **Probability** is a monotonic increasing function of plausibility and obeys the product and sum rules, necessitated by Cox´s desiderata.

An important point that is often neglected in the literature is that the above definition of probability is not in any way more correct than using some one-to-one mapping $q(\cdot) = f \circ \mathrm{P}(\cdot)$. I elaborate on this in Appendix A. But the function P has qualities that make it preferable to other functions apart from the fact that its corresponding rules look simple. As we shall see in Section 2.10, there is an attractive correspondence between probabilities and frequencies in repeatable experiments. Also, if we know that there are, say, seven red and three white balls in an urn, then our choice of function P gives a probability 3/10 of a white ball being drawn. This certainly seems a sound property for a definition of probability. In fact, Laplace himself used this property as the definition of probability [9].

> The probability for an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally probable.

I shall therefore adopt the same definition and choose the function P as the representative for probability, *but always keeping in mind that specific probability values only possess relative meaning when connections to frequencies cannot be made*. Probability values can only tell whether a proposition is more or less plausible than some other proposition.

For it is important to note that the present theory makes no references to repeated experiments, observed frequencies or hypothetically observed frequencies. The definition of probability used here applies to all kinds of propositions. Probability according to Bayesianism is therefore an *extension of deductive reasoning* and has thus a wider range of application than the frequentist interpretation[3].

Another difference between the two schools is that Bayesianism always require a probability assignment to be specific about what information is given. This is not necessary in frequentist probability theory, where probabilities are regarded physical entities. In the following, whenever I talk about

[3]It should be mentioned that some controversy has been raised regarding whether probability theory as an extension to logic may be applied to all kinds of propositions. The derivation due to Cox requires the *excluded middle* axiom which says that any statement is always either true or false. The objection concerns, among others, statements that may be neither verifiable nor falsifiable. Many number theoretical statements belong to this group. The present thesis is mostly concerned with statements about physical entities to which, most would agree, the excluding middle axiom must apply. However, it is also common that we have to make inferences about parameters which describe certain properties of pdf:s, hence referring to probabilities that relate to other probabilities. For such parameters, application of the excluding middle axiom may seem controversial.

probability theory in general, I will condition on the information I, meaning 'whatever is known in advance'. When discussing a general "post-data" case I will condition on DI, meaning prior information I and data D.

2.2.1 Continuous variables

Usually, the number of propositions (that is, the size of the hypothesis space) presented to us in a given problem is very large. In fact, if we are to make inferences about a continuous variable, we require the number of propositions to be indefinite. Letting A=*The variable of interest has a value between* θ *and* $\theta + \mathrm{d}\theta$, we can write

$$\mathrm{P}(A|I) = p(\theta|I)\mathrm{d}\theta. \tag{2.6}$$

The function $p(\theta|I)$ is now a function representing the probability for a $\mathrm{d}\theta$-interval starting at θ. If we decide on a hypothesis space of mutually exclusive propositions so that, for example, $A_1 = \theta \in [0, \mathrm{d}\theta)$, $A_2 = \theta \in [\mathrm{d}\theta, 2\mathrm{d}\theta)$, and so forth, then the function $p(\theta|I)$ is not unique since it only matters to us what values it takes on the "sampling points" 0, $\mathrm{d}\theta$, $2\mathrm{d}\theta$, and so on. But if we allow the propositions to start at an arbitrary "phase", then $p(\theta|I)$ is uniquely defined at every point. The important special case arises when we let $\mathrm{d}\theta$ go to zero. $p(\theta|I)$ will then represent the probability *density* of the infinitude of propositions around point θ. It is therefore called the *probability density function* (pdf) of θ. A pdf has the properties

$$p(\theta|I) > 0, \text{ and}$$
$$\int_{-\infty}^{\infty} p(\theta|I)\mathrm{d}\theta = 1$$

Generally, the integral is taken over the entire domain of $p(\theta)$.

Now that the step is taken to represent probabilities for an infinite number of propositions by pdf:s, we may use pdf:s also to represent finite sets of propositions by using Dirac distributions. Hence

$$p(\theta|I) = \frac{1}{2}(\delta(\theta + 1) + \delta(\theta - 1)) \tag{2.7}$$

means that it is absolutely certain that θ is either -1 or 1, but that we are equally ignorant to which of these cases is true.

An unorthodox detail from a mathematical point of view is that θ is not only the free variabel in the function $p(\theta|I)$, but that it is also a part of the

function *name*. What if we want to evaluate $p(\theta|I)$ at the point $\theta = x$? How do we express it? $p(x|I)$ will not do, because that looks like a totally different function, namely the prior for the variabel x. $p(\text{`}\theta = x\text{'}|I)$ could work but it would not mean the same thing as $p(\text{`}x = \theta\text{'}|I)$, which seems strange. In this thesis, whenever I need to evaluate a probability function at a certain value or change variables in a probability function, I will assign a temporary function

$$f(\theta) = p(\theta|I). \tag{2.8}$$

This convention eliminates the need for complicating a notation that otherwise is very practical.

2.3 Tools in probability theory

The sum and product rules may now be used to derive more useful tools. Essentially only two tools are needed to conduct general inferential calculus. *Bayes' theorem* is used to switch places between parameters on the right and left sides of the conditioning bar in a probability assignment. *Marginalisation* is used to remove parameters from the left side of the conditioning bar. We investigate them below.

2.3.1 Bayes' theorem

By a simple rearrangement of the product rule, we get Bayes' theorem:

$$\mathrm{P}(X|DI) = \mathrm{P}(X|I)\frac{\mathrm{P}(D|XI)}{\mathrm{P}(D|I)} \tag{2.9}$$

Bayes' theorem describes how to update what we know about a certain proposition X when we get some data D that is related to X in some way. That is to say, we start off with a representation of our knowledge, $\mathrm{P}(X|I)$, and when we receive the information that proposition D is true, we may update this to $\mathrm{P}(X|DI)$. Usually, we will have some physical relationship between X and D which makes it easy to say what we know about D given that X is true ($\mathrm{P}(D|XI)$). Bayes' theorem is used to turn that knowledge around, so to speak.

Naturally, Bayes' theorem applies equally well to probability *density* functions as it does to proper probabilities.

2.3.2 Marginalisation

The sum rule allows us to calculate probabilities for 'compound' propositions:

$$\mathrm{P}(\text{`}a < x < b\text{'}|DI) = \int_a^b p(x|DI)\mathrm{d}x. \tag{2.10}$$

When we have two or more variables we have the important special case

$$\mathrm{P}(x|DI) = \mathrm{P}(x, \text{`}-\infty < y < \infty\text{'}|DI) = \int_{-\infty}^{\infty} p(xy|DI)\mathrm{d}y, \tag{2.11}$$

where the last statement comes from the fact that the proposition $-\infty < y < \infty$ is always true. What is accomplished here is that y on the left side of the conditioning bar is eliminated. Nearly every problem in probability theory requires us to use *marginalisation* in this way to remove so called *nuisance parameters*.

In the case of discrete hypothesis spaces, the above integrals are replaced by sums.

2.3.3 Change of variables

Changing variables is not an actual principle of probability theory, but purely one of mathematics. It is needed whenever we have a mathematical relationship between two parameters and need to calculate the pdf for one of them from the pdf of the other. For example, let us say that

$$p(x|DI) = 1, \qquad 0 < x < 1, \tag{2.12}$$

and that we want to know the pdf of $y = x^2$. Noting that

$$p(x)|\mathrm{d}x| = p(y)|\mathrm{d}y| \Rightarrow p(y) = p(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|\Bigg|_{x=\sqrt{y}}, \tag{2.13}$$

we get

$$p(y|DI) = \frac{1}{2\sqrt{y}}, \qquad 0 < y < 1. \tag{2.14}$$

Changing variables can become very complicated if the functions involved are multidimensional and/or the mapping from x to y is not one-to-one.

2.4 How to assign priors

Looking at the product and sum rules, it is evident that a pdf can only be deduced from other pdf:s. Hence there is the need for principles that can produce the initial distributions required to "get started". These initial pdf:s must be based on – and only on – whatever information we have beforehand. But how do we calculate such *prior probability densities* (or simply *priors*) when seemingly there is no prior information available? Below I will discuss three important principles.

2.4.1 The principle of indifference

Assume that two persons know n outcomes of some experiment to be possible, and that this is the only thing they know. Person A labels the propositions in a certain way, while person B uses some other labelling, unaware of the labelling method that A has used. But since A and B are in the same state of knowledge, this must mean that if we pick out a certain proposition, then both persons must assign the same probability to that proposition. Since one is unaware of the labelling of the other, the only scheme that guaranties that they make the same probability assignments is the one where they both assign the same probability $1/n$ to all events.

This is the *principle of indifference*. The name was first used by John Maynard Keynes in 1921, but the principle was used long before that by earlier probability theorists such as Jakob Bernoulli and Laplace, but then going under the name *the principle of insufficient reason*. It applies only to finite hypothesis spaces but is very useful when reasoning about everyday 'deck of cards' and 'game of dice' kinds of problems. It has however a number of pitfalls which usually creates heated debates when basic probability theory is discussed. I take the opportunity to mention one such pitfall which has confused myself for quite some time, although it is not relevant for the discussion following in this thesis.

Let us say that we have an urn filled with objects and that we are about to draw an object from this urn. Now we are given the information I_1=*the objects in the urn are either blue or red.* Let R=*'the next draw is a red object'* and B=*'the next draw is a blue object'*. The principle of indifference requires both persons to assign

$$\begin{aligned} \mathrm{P}(R|I_1) &= 1/2 \\ \mathrm{P}(B|I_1) &= 1/2, \end{aligned} \tag{2.15}$$

because the information I_1 only tells them that there are two possible outcomes and nothing more.

Now imagine that they are given the additional information I_2=*the blue object may be either spherical or cubical*, and let B_1 and B_2 be the new "sub-propositions". Should they interpret the joint information $I_1 I_2$ as that there now are three possible alternatives and go back and reassign

$$\begin{array}{ll} \mathrm{P}(R|I_1,I_2) & = 1/3 \\ \mathrm{P}(B|I_1,I_2) = \mathrm{P}(B_1|I_1,I_2) + \mathrm{P}(B_2|I_1,I_2) & = 2/3, \end{array} \tag{2.16}$$

or should they reiterate the principle of indifference over just the propositions B_1 and B_2, so that

$$\begin{array}{ll} \mathrm{P}(R|I_1,I_2) & = 1/2 \\ \mathrm{P}(B_1|I_1,I_2) & = 1/4 \\ \mathrm{P}(B_2|I_1,I_2) & = 1/4? \end{array} \tag{2.17}$$

The former alternative certainly does seem strange since it forces a reassignment of probabilities upon receiving I_2 despite the fact that I_2 does not say anything about a blue object being more probable than a red. And in fact it is a misapplication of the principle of indifference, because information I_2 *is not indifferent to the propositions* R *and* B; it says something about B but nothing about R.

In the latter, correct application, the principle of indifference was first applied to propositions R and B, and then to B_1 and B_2. The example shows that caution must be taken when using this principle, so that it is only applied to propositions to which the information at hand is indifferent[4].

When correctly applied, the principle of indifference is a very powerful tool for assigning priors. However, since the applicability of the principle of indifference is restricted to finite hypothesis spaces and the present work will be concerned with continuous hypothesis spaces, we have to turn to more general principles for assigning priors. Transformation group theory is one such principle. We will look at it next.

2.4.2 Transformation groups

The idea of using *transformation groups* to produce priors is to consider the change of parameters under which a persons state of knowledge doesn't change. Stated differently, if we are equally ignorant about a variabel θ as we are about some transformation $f(\theta)$, then we may use f to derive a prior for θ that will express total ignorance on our behalf. This is most easily explained with a few examples. First, let us consider a so called *scale parameter* α. We

[4]Thanks goes to Erik Björnemo for repeating this argumentation over and over again until I finally grasped it. It took me about three years.

want to produce a prior $f(\alpha) = \mathrm{P}(\alpha|I)$ from a few simple facts: we know α to be positive, but we have no perception about its scale. Generally we may say that

$$f(\alpha)\mathrm{d}\alpha = f(\alpha')\mathrm{d}\alpha', \tag{2.18}$$

and in this case $\alpha' = c\alpha$. Its prior must then obey the relation

$$f(\alpha) = cf(c\alpha), \tag{2.19}$$

which has the solution

$$f(\alpha) = \frac{1}{\alpha}. \tag{2.20}$$

The above prior, which is usually called *Jeffreys' prior*, is unnormalisable. This is seldom a problem, because most problems are well-behaved in the limit when the domain of f goes to the entire positive real axis, and the normalisation constant will usually cancel when Bayes' theorem is applied.

For another example, consider calculating the prior $f(\beta) = \mathrm{P}(\beta|I)$ for a *location parameter* β. We know it to be bounded between two values A and B, but except for that we have no perception about its value whatsoever. Hence we have

$$f(\beta) = f(\beta - r), \tag{2.21}$$

within the boundaries given. The solution is

$$f(\beta) = \left\{ \begin{array}{ll} \frac{1}{B-A} & A < \beta < B \\ 0 & \text{otherwise} \end{array} \right) \tag{2.22}$$

The method of transformation groups is more general than indicated above, since it allows taking into account several parameters simultaneously. It turns out that it doesn't suffice to merely say in which way we are ignorant to each parameter, but that also the order in which the transformations are carried out matters. Such subtleties are not relevant in the present work, so I will not discuss them further.

2.4.3 The maximum entropy principle

It is often the case that we have prior information that constrains the variabel of interest in some way or another. The *maximum entropy principle* allows such additional information to be incorporated into a prior.

In his seminal 1948 paper, Shannon [10] looked for a measure that would account for the 'amount of uncertainty' in a set of probabilities. Denoting such a function by $H(p_1, \ldots, p_n)$, he set up the following requirements:

- H should be continuous.
- $H(1/n, \ldots, 1/n)$ should be monotonic increasing in n.
- Different ways of calculating $H(p_1, \ldots, p_n)$ should give the same answer.

Shannon showed that the only function satisfying these requirements is

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log(p_i). \tag{2.23}$$

The choice of logarithm is arbitrary. $H(p_1, \ldots, p_n)$ is called the *entropy* of the distribution $\{p_i\}$. The principle of maximum entropy is that we should choose the prior that maximises the entropy while still adhering to known constraints. The prior thus produced is the least committal prior possible, since it is the distribution that has the largest possible degree of uncertainty among all distributions conforming to the constraints.

In the most common application of the maximum entropy principle, we assume that the expected values of a number of functions are known:

$$F_k = \sum_{i=1}^{n} p_i f_k(x_i), \qquad k = 1 \ldots m. \tag{2.24}$$

By using Lagrange multipliers we may now find the $\{p_i\}$ that maximises entropy without violating the m constraints.

The entropy principle can also be extended to the continuous case. The continuous entropy is defined by

$$H(\mathrm{P}(x|I)) = -\int \mathrm{P}(x|I) \log\left[\frac{\mathrm{P}(x|I)}{m(x)}\right] \mathrm{d}x. \tag{2.25}$$

The function $m(x)$ describes how 'dense' the points x_i become at different locations when we let n go to infinity. It is calculated through use of transformation groups.

A typical maximum entropy example is that where we take $f_1(x) = x$ and $f_2(x) = x^2$, and we know that $F_1 = \mu$ and $F_2 = \sigma^2 + \mu^2$. The distribution $\mathrm{P}(x|\mu\sigma I)$ that then maximises entropy is

$$\mathrm{P}(x|\mu\sigma I) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2.26}$$

that is, the gaussian distribution.

We now have a few very powerful and general tools for producing priors. Note that, since the maximum entropy principle requires the principle of transformation invariance, we may regard the maximum entropy the sole principle for assigning priors, with the other two as special cases; when there are no constraints available, then maximum entropy will give the answer $\mathrm{P}(x|I) = m(x)$, so then we have the transformation groups solution. When we have a discrete distribution without constraints, then maximum entropy gives $p_i = 1/n$ – the principle of indifference.

On the other hand we can also choose to look upon transformation groups as the only principle for assigning priors, and use MaxEnt as a means of *updating* the pdf when our state of knowledge changes from complete ignorance to one where a set of constraints on the pdf is given.

My personal objection against the maximum entropy principle is that we cannot know the mean values F_k, since they aren't physical entities but properties of our state of knowledge. No data can ever force a specific mean value upon a distribution that represent my state of knowledge. It may certainly seem sober to think that a sample of ten million measurements with arithmetic mean value μ should imply a distribution with mean value μ, but such conclusions should be made with utter care. What entitles us to keep the sample mean and variance in a data series while disposing of all other information (as is often the case)? There might be reasons for doing so, but then we would need to adhere to principles outside the theory so far presented, and such principles must always be stated explicitly.

In fact, it is doubtful whether the information-theoretical rationale for the MaxEnt principle given here is needed at all. As we will see later, there is also a combinatorial rationale. Motivated from combinatorial principles, MaxEnt becomes a purely mathematical method, and its range of application is restricted to fields in which the situation under study can be described to be in one of a large number of "states". However, such fields (for example statistical mechanics and image reconstruction) are exactly those where MaxEnt has proven successful.

Prior assignment for noise measurements is often motivated through the use of maximum entropy; knowing the noise power and nothing more forces a gaussian prior upon the noise. However, when we speak of thermal noise (as is the case in the present work), the gaussian prior is also easily motivated by the central limit theorem and the fact that thermal noise is generated by adding contributions from a large number of similar physical processes. The effect is that the impact of moments higher than the second moment vanishes so that the noise is characterised by its variance alone. See Appendix D. In fact, the motivation of assigning a gaussian prior with independence

between samples is made stronger if the central limit theorem is used, than if a maximum entropy argument is employed. We shall see in Section 2.10 that careless use of the maximum entropy principle carries some risks with it.

2.4.4 The near-irrelevance of priors

It should be mentioned that, however important from a philosophical viewpoint, the choice of prior rarely plays an important part in the final result. While the present theory maintains that there is a *unique* correct prior for each parameter, for all practical purposes we may not need to be so strict when assigning priors, since the impact of the prior in many cases become negligible even when only a few data points are available. As a rule of thumb one may say that the prior is worth one data point, although it is of course possible to construct examples where the prior is of higher importance. The reader is encouraged to test different priors on a particular problem to see what significance it has on the final result.

2.5 How to make a decision

The assignment of priors lets us specify a 'starting point' of an inference problem. Then, through use of Bayes' theorem and marginalisation, we may produce a 'post-data' (posterior) pdf for the parameter that we are interested in. But what do we do with this pdf? Parameter point estimation problems, which is a very common type of problem in statistical inference, require us to answer the question 'What value do you think the parameter will take?'. Value judgment inevitably enters the theory at this point; our guess will depend on what we are prepared to lose if we happen to make the wrong guess. The associated theory is called *decision theory* and it a large field on its own. Looking for the estimate $\hat{\theta}$ of a parameter θ, we construct a *loss function* $L(\theta, \hat{\theta})$ that describes the loss associated with making the wrong guess. We will then choose the estimator $\hat{\theta}$ that minimises

$$\int L(\theta, \hat{\theta}) p(\theta|DI) \mathrm{d}\theta. \tag{2.27}$$

In this thesis we will be concerned with conditions that repeat over and over again. We therefore want to choose a loss function that produces good performance over the course of many trials. The *quadratic* loss function is by far the most commonly employed criterion and is reasonably the optimal

criterion to use in repeated scenarios. Minimising the quadratic error conforms to choosing the mean value of a pdf as estimate for that parameter, and so we will always choose

$$\hat{\theta} = \int \theta p(\theta|DI)\mathrm{d}\theta, \tag{2.28}$$

which minimises the expected value of the error $(\theta - \hat{\theta})^2$. As we will see later, when consecutive samples are independent, the arithmetic mean will eventually approach $\hat{\theta}$.

2.6 Model selection

Up to this point I have only considered the problem of how to assign prior pdf:s, manipulating pdf:s by means of marginalisation and Bayes' theorem, and taking a decision from a given pdf. In most problems of inference however, we have a set of data from which we wish to draw conclusions. It is then necessary to have a model which relates the measured data D to the parameter of interest, which will be called θ:

$$D = f(\theta, \xi) \tag{2.29}$$

Additional nuisance parameters ξ are also in general included in the model. In the above model, all parameters – D, θ, and ξ – are generally vector-valued. It is now possible to construct the pdf $p(\theta|D, I)$ by using the tools of probability theory and the principles for assigning priors. The to-do list of Bayesian inference now looks like this :

1. Construct a model.
2. Assign priors.
3. Use probability theory to derive the pdf for θ given whatever information is available.
4. Invoke decision theory to produce an answer to whatever question was asked.

But then, how do we know which model to choose in the first place? The somewhat disheartening answer is that there exists no formal procedure for model selection. That is to say, once we have established a particular *set* of models, then probability theory provides us with the means of telling which

one is best suited given a set of measurements (as we shall see shortly), but we are still left in the dark when it comes to choosing the original set. However optimal the theory for making inferences, the output result will still be bad if the model is bad. It may very well be that many scientific areas are still waiting for the discovery of "good" models.

Imagine then that we do have a number of models, and that we want to evaluate them against one another based on a measurement $D = \{d_0, d_1, \ldots\}$. Denote the models M_k. Using Bayes' theorem, we have

$$p(M_k|D, I) = p(M_k|I)\frac{p(D|M_k, I)}{p(D|I)}. \tag{2.30}$$

Since we only want to compare the probabilities of the different models we do not need to calculate the denominator $p(D|I)$. Also, in most cases we would assign equal values to all prior probabilities $p(M_k|I)$. Thus we have

$$p(M_k|D, I) \propto p(D|M_k, I). \tag{2.31}$$

The right-hand-side is usually called the *likelihood* of M_k and is denoted $L(M_k)$. Although is may require marginalisation over a few nuisance parameters, $L(M_k)$ is usually much easier to evaluate than the probability $p(M_k|D, I)$.

Hence we see that the model selection process consists of first choosing a set of models, and then evaluating the likelihood for each model given a set of data. Taken that all models are assigned the same prior probability, we choose the model that gives the highest likelihood.

2.7 The methods of frequentism and Bayesianism

Since the conventional attitude towards probability theory in physical channel estimation and prediction theory is the one propounded by frequentists, it may be a good idea to look at the differences between Bayesian and frequentist methods.

Problems of inference[5] come in many different forms of which the most important ones in the present context are sampling theory, hypothesis testing, and parameter estimation.

[5] 'Inference theory' in normally distinguished from 'decision theory'; In Bayesianism, the inference part of a problem is to produce the posterior of the parameter under consideration, whereas the decision part amounts to determining the course of action from the posterior. For an engineer it is of little interest to produce a pdf without any concrete suggestion of course of action coming out of the calculations, and so here I will be careless about the terminology and talk about 'inference' when I mean the joint process of inference and decision making.

Sampling theory is the theory of determining the probabilities for outcomes (samples) in data series, discrete or continuous. Problems in sampling theory are often analogous to problems regarding balls being drawn from an urn. Sampling theory is not actually directly relevant to the present problem in digital communications, and as a Bayesian, I will have no need for it in this thesis. The reason I mention it as an important topic is that it is central to orthodox statistics; in order to find the most likely values of some model parameters given some data, we may use sampling theory to calculate the probability for the data *that was actually received*, as a function of the model parameters, and then search for the values of those parameters which maximise the probability. This principle is called *the maximum likelihood principle*. When no cogent prior information is available, and when we do not care about the size of the error in the final guess, the maximum likelihood solution coincides with the Bayesian solution.

Hypothesis testing is the procedure of deciding which model that best describes a given set of data. Model selection is an alternative name for the same thing, although model selection often concerns the model structure, while hypothesis testing usually concerns values of fixed parameters in the model.

Parameter estimation is, as the name strongly indicates, the estimation of one or many parameters from given data. Its output can either be specific values, in which case one talks about *point estimation*, or intervals, which is called *interval estimation*.

All problems of inference begin with a model relating the parameters of interest with the data and possibly additional nuisance parameters,

$$D = f(\theta, \xi), \tag{2.32}$$

where the measured data D, the parameters of interest θ, and the nuisance parameters ξ are generally vector valued.

The general method for obtaining information about θ varies widely depending on whether one confesses to the frequentist or the Bayesian school. In frequentist theory, which method to use depends on the type of problem, whereas in Bayesianism the procedure is more or less the same regardless of the problem statement. This means that the Bayesian method can be listed as a 'do this–then that' procedure, whereas frequentist methods must be expressed as 'do this–*or* that'. Below is a summary of the methods used in the respective camps.

Bayesianism:

1. Determine the prior distributions.

2. Use the tools of probability theory to derive the posterior distribution. In sampling theory the posterior is often given directly by the model and the prior. Estimation and hypothesis testing often requires application of Bayes' theorem and marginalisation.

3. Investigate the posterior to produce the sought-after result. *Point estimation* is performed through the use of a loss function integrated over the posterior (see Section 2.5). *Interval estimation* amounts to finding an interval of the posterior (usually the shortest interval possible) having a certain area. *Hypothesis testing* can be seen as a special case of parameter estimation. In hypothesis testing, it is mostly common to choose the hypothesis that has the highest probability. This is the same as using a maximum criterion as loss function.

Frequentism:

- **Point estimation** of a parameter θ is commonly performed by using intuition to invent an *unbiased estimator* $\theta^*(D)$ which is a function of data D. It is chosen so that its mean value *over the sampling distribution*, $\int f(D)p(D|\theta I)\mathrm{d}D$, equals the parameter value θ. Note that this is quite different from the Bayesian least mean squares estimate, $\int \theta p(\theta|DI)\mathrm{d}\theta$. Other orthodox parameter estimation methods are the maximum likelihood method and the least squares method.

- **Interval estimation** also starts by inventing an estimator $\theta^*(D)$. The sampling distribution $p(\theta^*)$ is then calculated. Technically, this is done by a change of variables, $p(\theta^*)\mathrm{d}\theta^* = p(D|\theta)\mathrm{d}D$. Last, one finds the least interval having an area of 0.9 or so over this distribution.

- **Hypothesis testing** relies on a number of *significance tests*, among which are the commonly employed χ^2-test. Maximum likelihood is another method, which – if no cogent prior information is available and the loss function prescribes the same value to any error, regardless of its size – produces the same answer as the Bayesian approach.

Frequentist theory requires the division of model parameters into *random variables* and *deterministic but unknown parameters*. Only random variables are allowed to have pdf:s. There exists no formal procedure for determining which parameter belongs to which group. Usually, measured data is taken to be random variables while all other parameters are considered deterministic. A rule of thumb is therefore to let that which is known (the data) be random variables and let all other parameters – those that are the subject of the inference in most cases – be deterministic.

The concept of random variables vs. deterministic parameters bear no meaning in Bayesianism. Examples of other concepts which are relevant in frequentism but lack meaning in Bayesianism are *sufficiency* (which is not the same as *sufficient statistics*), *ancillarity*, and *optimal stop rules*.

2.8 Bayesianism/frequentism comparisons

In this section I show a few examples where Bayesian methods and frequentist methods are utilised on the same problems. We begin with a simple example of a point estimation problem:

EXAMPLE 2.1 **PARAMETER POINT ESTIMATION**

Let us assume that we have used some principle to assign a gaussian pdf to data $D = \{d_1, \ldots, d_N\}$:

$$p(D|\mu\sigma I) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i - \mu)^2\right). \tag{2.33}$$

Given data D, what is our estimate of μ (let us pretend that we already know the variance σ^2)? The frequentist method of unbiased estimators begins with guessing an estimator, which is then adjusted so that it conforms to a property called *unbiasedness*. The arithmetic mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{N} d_i \tag{2.34}$$

is known to be such an estimate.

Moving on to the Bayesian solution, we write down Bayes' theorem:

$$p(\mu|D\sigma I) = p(\mu|\sigma I)\frac{p(D|\mu\sigma I)}{p(D|\sigma I)}. \tag{2.35}$$

We are completely ignorant to the value of the location parameter μ, so its appropriate prior is uniform on an interval that we will make infinitely wide by going to the limit. The corresponding normalisation constant will be present also in the denominator, which will basically make the prior $\mathrm{P}(\mu|I)$ disappear. But then we see that

$$p(\mu|D\sigma I) \propto p(D|\mu\sigma I), \tag{2.36}$$

which peaks at $\mu = \sum d_i/N$. Whether we use the maximum value, the mean, or the median as estimate does not matter since the distribution is gaussian; they all yield the arithmetic mean. We see that the Bayesian and the frequentist method yield the same answer in this case.

When we have gaussian distributions and no cogent prior information, there is a symmetry that produces the same answer regardless of whether we use Bayesian och frequentist methods. This is relevant in the present context, because these conditions often apply in digital communications.

But let us also shortly look at the estimation of the variance σ^2 (now μ is assumed to be known). For convenience, let $\rho^2 = \sum(d_i - \mu)^2$. The conventional frequentist (unbiased) estimate is

$$\frac{\rho^2}{N-1}, \tag{2.37}$$

but what about the Bayesian solution? Using Jeffreys' prior and the expectancy as estimate, we have

$$\int \sigma^2 \mathrm{P}(\sigma|I) \frac{\mathrm{P}(D|\mu\sigma I)}{\mathrm{P}(D|\mu I)} \mathrm{d}\sigma = \frac{\pi^{-N/2}\rho^{2-N}\Gamma\left(\frac{N-2}{2}\right)/4}{\pi^{-N/2}\rho^{-N}\Gamma\left(\frac{N}{2}\right)/2} = \frac{\rho^2}{N-2}. \tag{2.38}$$

We see that the frequentist and Bayesian estimate are virtually the same already for moderately low values of n. Indeed, it would seem strange if they would deviate considerably in such a common case as estimating the variance of a gaussian distribution.

Next we will take a look at an interval estimation problem. The methods used in this example are vastly different between the two schools, and – as we shall see shortly – so are the results. The example is taken from [11].

EXAMPLE 2.2 **INTERVAL ESTIMATION**

We now consider an *interval estimation problem*. A certain type of devices will operate without failure for a time period θ, after which they start to break down following an exponential law. The probability that x number of failures occur between time t and $t + \mathrm{d}t$ is hence $p(x|\theta I)\mathrm{d}x$ for small values of $\mathrm{d}x$, where

$$p(x|\theta I) = \begin{cases} \exp(\theta - x), & x > \theta \\ 0, & x < \theta \end{cases} \tag{2.39}$$

We are interested in determining the 'life expectancy' parameter θ to some level of confidence, that is we want to establish boundaries within which θ is likely to lie. The data that we have are N measurements of failure times $x_1, \ldots, x_N$.

The frequentist method of *confidence intervals* starts by inventing an estimator of θ. Again, we use an unbiased estimator

$$\theta^* = \frac{1}{N}\sum(x_i - 1), \tag{2.40}$$

which is unbiased because the expectancy for θ (over the sampling distribution) is $\theta + 1$. We must then find the sampling distribution for θ^*. The method of characteristic functions reveals that it is proportional to $(\theta^* - \theta + 1)^{N-1}\exp(-N(\theta^* - \theta + 1))$. Here we will look at the specific case $N = 3$ with data $\{x_1, x_2, x_3\} = \{12, 14, 16\}$ and look for the shortest 90% confidence interval. By integrating the sampling distribution and numerically looking for the shortest interval, we find

$$12.1471 < \theta < 13.8264. \tag{2.41}$$

Next we look at the bayesian solution. It demands that we assign a prior to θ. Since θ is a location parameter, our ignorance about its value is properly represented by a constant prior. By applying Bayes' rule we then find

$$p(\theta|x_1, x_2, x_3, I) = \begin{cases} N\exp N(\theta - x_1), & \theta < x_1 \\ 0, & \theta > x_1 \end{cases} \tag{2.42}$$

Since the posterior is decaying, the Bayesian interval is clearly shorter the lower the values we choose. Hence, the interval has upper limit x_1, and its span is easily found to be $-N^{-1}\log(1 - 0.90)$. We therefore have the Bayesian solution

$$11.23 < \theta < 12.0. \tag{2.43}$$

Comparing the two approaches, we see that both the methods used and the results are quite different. The reason for the frequentist solution yielding an absurd answer (we know from the data that $\theta < 12$) is that the estimator θ^* was poorly chosen. How to choose a 'good' estimator is a nontrivial problem and there exists no formal procedure for how to do it.

Finally we look at a hypothesis testing problem. It is relevant in channel estimation theory in connection with a type of channel estimators called

blind estimators, although this kind is not used in the present thesis. The below example is borrowed from [1, p. 300].

EXAMPLE 2.3 **HYPOTHESIS TESTING**

The English one pound coin is sufficiently thick that it might stand on edge after a toss. An experiment of 29 tosses is performed, with the outcome $n_1 = 14$ heads, $n_2 = 14$ tails, and $n_3 = 1$ standing on edge. Person A is familiar with the coin in question, and so A assigns probabilities $p_1 = p_2 = 0.499$ and $p_3 = 0.002$. Person B, however, is ignorant to the fact that a coin has been tossed and is only aware of that there are three possible outcomes and that the different trials (tosses) are logically independent. B must therefore use the principle of indifference and assigns $p_1 = p_2 = p_3 = 1/3$.

Now we want to evaluate the hypotheses of A and B against one another[6]. Which one is most plausible? The Bayesian solution is simple:

$$\frac{\mathrm{P}(H_A|D)}{\mathrm{P}(H_B|D)} = \frac{\mathrm{P}(H_A)}{\mathrm{P}(H_B)} \frac{\mathrm{P}(D|H_A)}{\mathrm{P}(D|H_B)} = \frac{\mathrm{P}(D|H_A)}{\mathrm{P}(D|H_B)}. \tag{2.44}$$

Here, $D = \{n_1, n_2, n_3\}$ and the probabilities for the respective hypotheses are considered to be the same. Because of the independence we have $\mathrm{P}(D|H) = p_1^{n_1} p_2^{n_2} p_3^{n_3}$. The ratio (2.44) is readily found to be about 483.5, supporting person A's hypothesis.

Note that in the above, we just invoked the usual Bayes' theorem. However, since we are only interested in calculating the *ratio* between the two probabilities, we save ourselves the trouble of calculating the prior for the data.

Orthodox statistics based on a frequentist approach has no tools for calculating probabilities for hypotheses; hypotheses cannot be regarded as random variables and so even talking about probabilities for hypotheses is considered an abomination. Instead, orthodox statisticians have invented a range of *significance tests* for comparing hypotheses. The most common of these is the so called χ^2-test. We calculate the factor

$$\chi^2 = \sum_{k=1}^{n} \frac{(n_k - np_k)^2}{np_k} \tag{2.45}$$

[6]What *are* the hypotheses? We shall later see that in independent experiments, there is an exact correspondence between probabilities and observed long run frequencies. The hypotheses here is therefore that the frequencies of outcomes that will be observed over a long time are the respective p_k

for both cases and see which one gives the *lowest* value:

$$\chi^2_A = 2\frac{(14 - 29 \times 0.499)^2}{29 \times 0.499} + \frac{(1 - 29 \times 0.002)^2}{29 \times 0.002} = 15.33,$$
$$\chi^2_B = 2\frac{(14 - 29 \times 0.333)^2}{29 \times 0.333} + \frac{(1 - 29 \times 0.333)^2}{29 \times 0.333} = 11.66,$$

which supports person B's hypothesis, that each outcome should account for one third of the tosses in the long run. We see that Bayesianism and frequentist reasoning express different opinion about which hypothesis is the most likely. The reason is that the χ^2-test is very sensitive to unexpected data. In this case, even given the hypothesis of person A it was a bit unexpected to see the coin stand on edge once in just 29 tosses (On performing more "batches" of 29 tosses, we would only expect the coin to stand on edge in about one of 18 such batches). For example, if the data had been $n_1 = 14, n_2 = 15, n_3 = 0$, the result of the two approaches would have been nearly the same.

We have seen that the two perspectives – Bayesianism and frequentism – take quite different attitudes towards the probability concept. Frequentism has an *ontological* perspective and takes probabilities to be real physical entities. Bayesianism, on the other hand, takes an *epistemological* stand, claiming that probability expresses the state of knowledge of an individual. Jaynes often warns against what he refers to as the *mind projection fallacy*, that a person's private thoughts are mistaken for actual existing physical reality. Quoting Jaynes [1, p. 75]:

> "Anyone who believes that he is proving things about the real world, is a victim of the mind projection fallacy."

Having decided to adhere to the Bayesian view, we must then pose an inevitable question. The frequentist's statement about a physical parameter can certainly be validated by careful measurements, but how do we validate the Bayesian's statement that is *not* about the real world, but merely one about a person's state of knowledge?

2.9 Validation

Cox's derivation of the rules of inference calculus originated from the axioms of classical logic. To these he added a plausibility measure along with a

few rules to which such a measure must conform. Hence were conceived the sum and product rules. These rules had up until then been employed by both frequentists and Bayesians simply because they seemed reasonable, but only by Cox's derivations were they derived as the only consistent rules for conducting plausible inference. Since these rules are the only ones that are used in Bayesian probability theory – unlike frequentist theory which consists of a vast number of *ad hoc* methods – we know with certainty that the result produced by Bayesianism is always *logically consistent* with the information that was put in, that is the data, the mathematical model relating the parameter(s) of interest with the data, and the prior(s)[7].

Additionally, if the posterior pdf for the parameter of interest as derived by Bayesian theory would turn out to be incredibly sharply peaked with all its mass collected at one single value, then not only would it be logically consistent with whatever information was offered to the calculations, but it would also make a strong statement about the actual physical reality *if we have full confidence in the accuracy of the model.* This presents us with two distinctive ways of validating the result of a Bayesian probability calculation:

Logical consistency
Since all results sprung from Bayesian calculations are logically consistent with the information presented to the calculations, they need not be validated against measured data. If we believe in the information put into the equations – the data, the model, the priors – strongly enough that virtually no evidence would make us lose confidence in them, then we know also the result to be correct. Thanks to the rules of probability theory, we know with certainty that any result derived from strict application of these rules will conform to any extreme test of logical conformity we can put it to.

Physical prediction
In very special cases it might be that some pdf involved in the calculations turn out extremely narrow. We have then in effect made a prediction without uncertainty about the *real world.* But this can only occur if we also put certainty into the equations, in the form of prior distributions.

It is clear that if we use Dirac functions for prior distributions for some parameters, then that expresses absolute certainty about those parameters. The inference process might then collapse into pure deductive logic which lacks all elements of uncertainty. But there is another case, relevant to the present problem in digital communications, in which there is a risk that we

[7] Whenever frequentism yields an answer that differs from the Bayesian result, we know for a fact that it is *not* consistent with the information put in. The result then contradicts the desiderata in one way or another.

inadvertently put in certainty. It has to do with observed long run frequencies. We study it next.

2.10 Probability/frequency correspondence

We saw earlier that the particular choice of the function $p(\cdot)$ as a representative for the probability concept is arbitrary in the sense that any other monotonic mapping $f \circ p(\cdot)$ will act as an equally valid representative, as long as we remember also to adjust the sum and product rules. But in this section we will see that the function p implies an interesting correspondence between probability and observed frequency. We will examine which implications the assignment of independent pdf:s has on a person's state of knowledge about long run frequencies. I start with investigating the consequences, then I comment on the relevance.

Assume that our hypothesis space for a certain experiment consists of K propositions which we denote X_k. For one reason or another – it is not relevant in this context – we have made the prior assignment

$$\mathrm{P}(X_k|I) = p_k, \quad k = 1, \ldots K \tag{2.46}$$

We assume that the circumstances allow us to perform an indefinite number of experiments, and that the outcome of one experiment does not provide us with any relevant information about the outcome of the next. This means that we should assign independent distributions to all experiments. Let us investigate what this means for the prior information we have about observed frequencies as we perform more and more experiments. That is, what probability do we implicitly assign to the event that, say, X_1 comes up 15 percent of the cases when we carry out a large number of experiments?

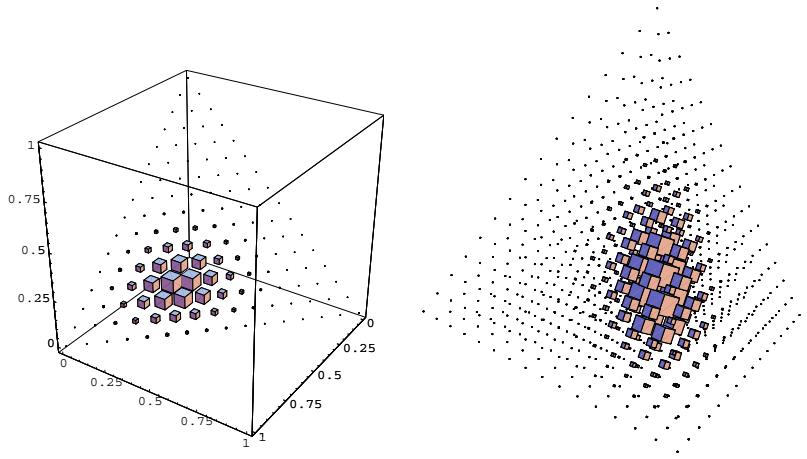
Let us say we perform a total of N trials. The number of ways in which we can combine $f_1 N$ apples, $f_2 N$ oranges, $f_3 N$ pears and so on is given by the multinomial coefficients. f_k is the fraction of times that proposition X_k takes place. The probability for a certain combination of frequencies is therefore given by

$$p(f_1 \ldots f_K|I) = \frac{N!}{(f_1 N)! \ldots (f_K N)!} p_1^{f_1 N} \ldots p_K^{f_K N} \tag{2.47}$$

Using the Stirling approximation $\log x! \approx x \log x$ we get[8]

$$\begin{aligned} \log p(f_1 \dots f_K|I) &\approx N \log N + N \sum_k f_k [\log \frac{p_k}{f_k} - \log N] \\ &= N \sum_k \log \frac{p_k}{f_k}. \end{aligned} \tag{2.48}$$

But the inequality $\log x < x - 1$ shows that the above sum is less or equal to zero, with equality only when $f_k = p_k$ (Gibbs' inequality). Hence, in the limit when $N \to \infty$, the probability for observed frequencies $f_1 \dots f_K$ is unity at the point $p_1 \dots p_K$ and zero otherwise.



(a) $K = 3$ (b) $K = 4$

Figure 2.1: The pdf for observed frequencies when there are three and four different possible outcomes, respectively. The pdf:s have support in the plane $\sum f_k = 1$. In the case $K = 4$, the tetrahedron with corners at the tips of the 4-D unit vectors has been projected onto a 3-dimensional space. The size of each cube indicates the value of the pdf at that point. As the number of trials N gets larger, the mass of the pdf concentrates to the point $f_1 = p_1, \dots, f_K = p_K$. In the left figure, $p = [0.5, 0.3, 0.2]$. In the right figure, $p = [0.1, 0.1, 0.3, 0.5]$. Here, $N = 15$.

This shows that the assignment of independent priors results in a concentration of mass in the pdf for observed long-run frequencies. The mass gathers at a point which represents the prior for one single experiment. This

[8] Actually, the Stirling approximation contains a few more terms which are needed to fully appreciate the limit : $\log x! \approx x \log x - x + \log \sqrt{2\pi x} + \mathcal{O}(1/x)$. It is easily seen that the additional terms tend to zero in this case.

is illustrated in Figure 2.1. Already after just 15 independent trials, the pdf:s for observed frequencies are sharply peaked around the mean value for a single trial. As the number of trials increases, the pdf:s will rapidly approach a dirac distribution, which we have to interpret as 'absolute certainty'. But this suggests that the assignment of independent trials is not – as one might expect – an expression of ignorance, but contrary one of enlightenment. The following examples will illustrate this.

EXAMPLE 2.4 **PROBABILITY/FREQUENCY MISMATCH**

Person A assembles an infinite sequence of ones and zeros by block-wise mixing 750 zeros and 250 ones. This means that the first 1000 'bits' consist of exactly three quarters ones and one quarter zeros, and the same goes for the next 1000 bits, and the next and so on.

Now person A gives these bits, one by one, to person B for as long as B desires. Person B is not aware of the scheme that person A has used to construct the sequence, just that it will consist of just ones and zeros and that it will go on indefinitely. We call this 'information I'.

What probability does B assign to the event *the next bit will be a zero*? B wants to be as conservative as possible, since virtually no information exists, so he sets $p(\text{'next comes a zero'}|I) = 0.5$, and the same with 'next comes a one'. He states that the probability is one half for any future bit being zero, given outcomes up to the present. He lets different bits be independent, since he has no reason to believe otherwise.

This might seem very reasonable and conservative. B could for example calculate the probability that bits number ten and eleven from now will both be zeros, and find the answer to be one fourth. This certainly expresses a lot of uncertainty, which should be the case.

But what happens if person B looks at the probability for the ratio of zeros in the next, say, N bits? The expression is easily found to be

$$\mathrm{P}(\text{'ratio of zeros is } k/N\text{'}|I) = 2^{-N}\binom{N}{k}. \tag{2.49}$$

Studying the ratio of zeros in just a few bits, we see that the uncertainty is quite large, centred at 0.5. But as the number of bits increases, we see that the curve claims that we should become increasingly certain that the ratio is exactly 0.5. It rapidly approaches something that we have to interpret as

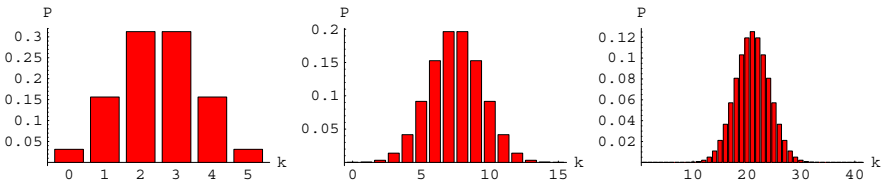


Figure 2.2: The probability of observing k zeros in a sequence of N bits, as given by the model $p(\text{'zero'}|I) = 0.5$. The cases $N = 5$, $N = 15$, and $N = 40$ are shown.

'absolute certainty'. Mathematically, this is easily seen if we use the normal approximation of the binomial function[9],

$$2^{-N}\binom{N}{fN} \approx \sqrt{\frac{2}{\pi N}}\exp\left(-2N(f-\frac{1}{2})^2\right), \tag{2.50}$$

where f is the frequency k/N.

But how could this be? If we study the bits that person B has already got, then, as he gets more and more, we will discover that the ratio of zeros will be exactly three quarters when we have studied an integer multiple of 1000 bits (remember that person A mixed 750 zeros with 250 ones). In between whole blocks, the ratio will vary a little around 3/4, but as B collects more and more it will settle down to that steady value. So the 'certain' result that B gets from his probability calculations certainly does not represent what really will happen!

We see now that person B's attempt to represent very little knowledge backfired and turned out to express much more knowledge than he had!

The message is that the assignment of independent priors is a hazardous undertaking. Assigning independent priors to an indefinite number of parameters always implies claiming full knowledge about the long run frequency distribution (which is a physical parameter). As mentioned in Section 2.4.3, the gaussian representation of noise is often motivated by using maximum entropy and claiming knowledge of the first two moments (mean value and variance). Then, using independent pdf:s between samples is motivated by

[9]Note that these are proper probabilities and not probability densities. As $N \to \infty$, the probability for any f goes to zero. However, the probability *density*, calculated by multiplying the expression with N, goes to infinity for $f = 1/2$ in the limit.

lack of further information about the noise. But we have seen in this section that independence implies full knowledge about long-run frequencies, and this can never be defended by an *absence* of knowledge. In the case of thermal white noise, independent samples *are* well motivated since it is known that the underlying physical mechanism consists of a lot of similar physical processes, but in other cases we have to be careful when using independent pdf:s.

The specific relevance in the present context, apart from the representation of thermal noise, is when linear models are used. Such models claim certainty about long run frequencies. Before investigating this phenomenon I will make a more formal introduction of the gaussian distribution than hitherto has been made.

2.11 The gaussian distribution

The gaussian distribution will be used extensively in the forthcoming chapters. It is therefore convenient to introduce a compact notation for it. For a complex parameter x with mean value μ and variance σ^2 equally divided between the real and imaginary parts, I will use the abbreviation

$$\mathcal{CN}(x;\mu,\sigma^2) \triangleq \frac{1}{\pi\sigma^2}\exp(-|x-\mu|^2/\sigma^2), \tag{2.51}$$

$\mathcal{CN}$ denoting *complex normal*. I then assume independence between the real and imaginary parts. To allow for correlation between them, we have to write $x = x_r + jx_i$ and use a vector-valued real distribution:

$$p(x|\mu,\Sigma,I) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^*\Sigma^{-1}(x-\mu)\right), \tag{2.52}$$

where

$$x = \begin{bmatrix} x_r \\ x_i \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_r \\ \mu_i \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_r^2 & \sigma_{ri}^* \\ \sigma_{ri} & \sigma_i^2 \end{bmatrix}, \tag{2.53}$$

are the real-valued vector representations of the complex parameters. Note that one gets (2.51) by setting $\Sigma = \sigma^2/2 \cdot I$ in (2.52).

The multivariate complex gaussian distribution is written

$$\mathcal{CN}(x;\mu,\Sigma) \triangleq \pi^{-n}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^*\Sigma^{-1}(x-\mu)\right), \tag{2.54}$$

where x and μ are now vectors of length n. Again, the real- and imaginary parts are assigned to be independent and the variance is divided equally between them.

2.12 Linear models

In the forthcoming chapters I will use *state space* models to mimic the behaviour of time-varying radio channels. The states (to be introduced later) are modelled by

$$x_{t+1} = Fx_t + Gu_t \tag{2.55}$$

where the (generally vector-valued) so called *process noise* u_t is assigned a gaussian distribution, independent between samples, so that

$$p(u_t|I) = \mathcal{CN}(u_t; 0, Q) \tag{2.56}$$

and

$$p(u_t u_\tau|I) = p(u_t|I)p(u_\tau|I), \quad t \neq \tau. \tag{2.57}$$

The model matrices F and G may be time-varying but I am here considering a time-invariant model.

Imagine now that we believe in the "correctness" of the linear model (2.55) indefinitely. Since the process noise u_t is white, so that we hence claim to know its frequency distribution with certainty, this implicates that we also claim knowledge about the long-term frequency distribution of $X_T = \{x_0, x_1, \ldots, x_T\}$. For example, we will find that

$$\lim_{T\to\infty} \sum X/T = 0, \tag{2.58}$$

and that

$$\lim_{T\to\infty} \sum XX^*/T = \bar{\Pi}, \tag{2.59}$$

where $\bar{\Pi}$ solves the *Lyapunov equation*

$$\bar{\Pi} = F\bar{\Pi}F^* + GQG^*. \tag{2.60}$$

It may not have been the intention of the person who constructed the model to build in that kind of cogent information. Again we have to conclude that the assignment of independent pdf:s must be undertaken with utter care.

Chapter 3

The modelling of a single tap

In the forthcoming chapters I will apply the Bayesian probability theory treated in Chapter 2 to radio channel estimation and prediction, the motivation of which is given below. The modelling of a radio channel usually amounts to modelling several scalar complex-valued radio channel descriptors called *taps*. The present chapter is restricted to the modelling of one single tap. A linear model will be used to represent the fading statistics of the tap. This approach requires the assumption that a number of parameters are known to exact precision. Here I will therefore assume that estimates of high quality of such parameters are provided by stand-alone estimators.

Taps rotate with time in the complex plane (a phenomenon which is called *fading*). Tap behaviour is characterised by the frequencies with which the fading can occur. To exemplify the tap modelling procedure, I describe in this chapter how to construct one particular model. The model described is very cautious and has only the restrictions imposed by vehicular velocity built into it. All other fading frequencies are set to be equiprobable so that no kind of behaviour is favoured before any other. It is later shown that such a cautious stand is very costly for system performance. Considerable effort should therefore be put into estimation of the fading behaviour.

The present chapter also serves as a very brief introduction to digital communication theory for the uninitiated reader. I examine both *single carrier systems* and *multi carrier systems* and establish that optimal detection of digitally transmitted information essentially is a problem of geometric character.

3.1 Adaptive multiuser systems

Most of present day wireless multiuser systems assign radio resources among users regardless of their current instantaneous radio channel conditions[1]. The systems instead perform averaging over the unknown conditions by utilising diversity in various dimensions (time, frequency, polarisation, and space). However, there is a lot to be gained in taking an opportunistic approach, in which a user gets assigned resources when its channels conditions are good, an rests idle to the benefit of other users when its channel conditions are bad.

In such systems, users would compete for the resources by signalling their respective temporal channel quality to a central scheduler. This scheduler would then distribute the resources among the users based on these qualities, and possibly also on some kind of fairness criteria. However, since there would inevitably be some delay involved in the signalling and the scheduling process, the users would have to *predict* the channel quality. The prediction horizon would typically be a few microseconds long. This is the *prediction* problem. The *estimation* problem is that of estimating the most probable transmitted bit sequence at the time instant when data arrives to a specific user, so that payload data may be recovered as well as possible.

To tackle the estimation/prediction problem, we need a model for the situation under study. In the present chapter I lay out the first part of the modelling process, which ultimately is about describing the fading behaviour of time variant radio channels. The next chapter presents the second part, in which an entire system is modelled.

3.2 Digital communication systems

Schematically, a communication system is constructed as depicted in Figure 3.1. The principle is the same for most types of digital communications systems; information units – bits – are grouped together and mapped onto *symbols*. The amount of symbols available – the size of the symbol *constellation* or the symbol *alphabet* – is determined by the size of these groups. For example, if bits are grouped three by three, then there are eight symbols in the constellation.

The constellation also has a certain dimensionality, so that each symbol

[1] There are exceptions. For example, the standards EDGE (Enhanced Data rates for GSM Evolution) and HSDPA (High Speed Downlink Packet Data Access) are adaptive to some extent.

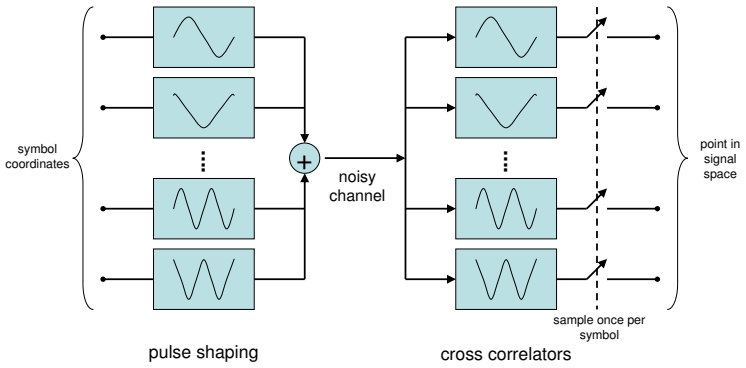


Figure 3.1: Schematic description of a digital communication system.

is represented by combinations of one, two, or several unit vectors. The number of available unit vectors determines the dimensionality of the *signal space*.

Each of the unit vectors is associated with a waveform from a set of orthonormal signals, so that the transmitted waveform associated with a symbol is simply a weighted sum – a superposition – of signals from this set. The orthogonality is meant with respect to a certain time interval. This interval need not be the same as the symbol *duration*, that is the time interval between each new symbol. The important thing is that the signal set is chosen so that orthogonality holds not only between different members of the signal set, but also between any one member and a time-delayed copy of itself, if the delay is an integer multiplicative of the sampling period.

The model described here is valid for *linear* modulation formats. There are also *nonlinear* modulation formats, such as *Continuous Phase Frequency Shift Keying* (CPFSK), but they will not be considered in this thesis.

One symbol duration apart, a new symbol is released from the transmitter. On its way to the receiver, the signal is distorted by nonlinear amplifiers and by noise and multipath propagation. At the receiver, the signal is passed through a bank of filters, each filter corresponding to a unit vector in the signal space. These filters are called *cross correlators* or simply *correlators*. If the impulse responses of the correlators are the same as the respective waveforms – and I shall assume this in this thesis – then we say that they are *matched filters*. A matched filter maximises the signal-to-noise ratio at

the output of the filter.

The output of a cross correlator is sampled at a certain rate. This rate is usually the symbol rate but may in some cases be higher depending on the characteristics of the channel. If the channel distorts the signal only by introducing additive (white) thermal noise, we say that we have an *additive white gaussian noise* (AWGN) channel.

In the AWGN case, we would then receive at the output of each cross correlator a signal $r(t)$ described by

$$r(t) = \sum_k a_k p(t - kT_s) + n(t). \tag{3.1}$$

Here, $n(t)$ is the filtered noise, the $\{a_k\}$ are the symbol amplitudes (the coordinate) for this particular dimension, T_s is the symbol duration, and $p(t)$ is the filtered waveform or the *pulse shape*. The pulse shape may be chosen so that time-adjacent symbols do not interfere with one another. This means that at the sampling instant, $p(t - kT_s) = 0$ for all k but $k = 0$.

A natural choice for $p(t)$ is a rectangular pulse of width T_s. However, such a pulse produces high spectral side lobes because of the sharp flanks. Instead, one may choose from a class of pulses of long time duration called *Raised Cosine* (RC) pulses. These share with the rectangular pulse the property of zero crossings at the sampling instant, but have better spectral properties. To ensure that the pulse shape *after* the cross correlator is an RC pulse, one would use *Square Root Raised Cosine* filters for pulse shaping at the transmitter and for the cross correlators, so that their total impulse response is an RC pulse.

As a short side note, we may set this into relation with the relevance of prior information. Assume a noiseless situation so that $n(t) = 0$. Then if the pulse shape $p(t)$ is known, we may reconstruct the continuous signal $r(t)$ perfectly only from knowledge of the samples $p(kT_s) = a_k/p(0)$. But according to the Nyquist theorem, a signal can only be reconstructed from samples if its single side bandwidth is strictly less than the Nyquist frequency $1/2T_s$. In our case, the bandwidth of $r(t)$ is given by the power spectral density (psd) of $p(t)$ (if the $\{a_k\}$ are white), and this psd may be arbitrarily wide. And still we reconstruct the signal perfectly from samples taken with period T_s. This illustrates that if cogent prior information is available, signal reconstruction may be possible even if the requirements for the Nyquist theorem are not met.

The sampled output of each cross correlator represents how much of a certain unit vector is "contained" in the received signal. The collected samples from all correlators mark a point in the signal space. Ideally, that point

would be one of the points in the symbol constellation, but the noise will disturb the sampled outputs, causing a displacement to occur. It is the responsibility of the *detector* to decide which symbol is likely to be the correct one, and hence which bit pattern the sender had in mind.

When the distortion is restricted to AWGN alone and all symbols are equiprobable, the optimal detection scheme reduces to simply deciding which symbol representation in the signal space is closest to the received point. We may think of each point corresponding to a transmitted symbol as being surrounded by a cloud of gaussian density. The density at a certain point correlates to the probability of the received symbol to end up at that particular point.

Distortions of a more complex character than AWGN introduces dependencies between samples. So called *equalisation* is used to combat this phenomenon, but we may also have to sample the signal at a higher rate than the symbol rate in order to achieve a high detection performance. The optimal detection algorithm may then become extremely complicated, forcing the system designer to use suboptimal schemes.

3.2.1 Coding and other functionality

It is often the case that we want to make our information bits resistant to a reasonable percentage of transmission bit errors. This is achieved by introducing redundancy into the bit sequence. For example, we may map k information bits onto n *coded* bits, where n is larger than k (we say that we apply an n/k block code). By applying clever algorithms at the receiving side, such an n-block may then be decoded to the correct k-block, even though one or a few of its bits have been subjected to errors.

The optimal decision strategy can be understood in many ways. I will here give a "geometric" explanation. To study a specific case, let us say that we use a 7/4 block code and that our symbol alphabet has eight symbols and two dimensions. Then 12 information bits are mapped onto 21 coded bits which in turn are mapped onto 7 symbols. This means that we can think of the encoder as directly mapping the 12 bits onto one of $2^{12} = 4096$ symbols in an "expanded" signal space. We may call the symbols residing in this space *supersymbols*. We then detect the received supersymbol as usual, by picking the nearest proper supersymbol. The expanded signal space would in this case have 14 dimensions, since one supersymbol corresponds to seven regular symbols, each having two dimensions. This example is more closely investigated in Appendix B.

Supersymbols are constructed is such a way that coding and other func-

tionality only affect bits and symbols within the same supersymbol. Another way of expressing it is to say that one supersymbol corresponds to one codeword. By construct, a supersymbol is therefore always received in AWGN. Needless to say, this way of looking upon a communications system quickly becomes complicated when we continue to introduce more and more functionality so that the size of the codeword grows. For example, utilising a so called convolutional encoder instead of a block code, in which every coded bit is related to the other in endless succession, would force us to regard the *whole* information bit sequence as one single supersymbol, since no natural subdivisions in the information bit stream exist. Needless to say, this abstraction then becomes useless.

The natural and common thing to do is of course to regard the encoder / decoder mechanism as a separate pair of blocks inserted before the modulator and after the demodulator, respectively (and accordingly for other types of functionality such as encryption, spectral spreading etc). Treating them separately may however yield suboptimal solutions, as Appendix B illustrates.

3.3 Multicarrier systems

Communications systems can roughly be divided into two groups: multicarrier systems and single carrier systems, both of which will be studied presently. We start with multicarrier systems.

The wave propagation through the air can do little to our signal than to add noise and to "smear" it out, that is to split it up into an infinitude of echoes of different time lags and attenuations. Hence the channel – by which I mean the total system consisting of the pulse shaping filters at the transmitter, the wave propagation, and the filters (correlators) at the receiver – constitutes a linear system, though however it is generally a *time varying* linear system for mobile users[2].

A linear system has eigenfunctions on the form $e^{j\omega t}$. Complex exponentials are therefore an excellent choice for the orthogonal set of waveforms. To guarantee orthogonality, all waveforms need to have an integer number of periods in the symbol time (we here take the symbol time to be the time over which orthogonality should hold). Hence we choose the set of N complex

[2]This requires us to omit certain phenomena that might be present in reality but which we hope will be negligible. Apart from nonlinear behaviour of amplifiers, metal oxidation such as rusty fences might cause nonlinear scattering of radio signals.

waveforms

$$e^{j(\omega_c + n\Delta\omega)t}, \quad 0 \leq n \leq N-1, 0 \leq t < Ts \tag{3.2}$$

where ω_c is the system's centre angular frequency. $\Delta\omega = 1/2T_s$, and T_s is the symbol duration. Of course, we cannot use complex signals, so we have to resort to using pairwise signals on the form $\sin(\omega t)$ and $\cos(\omega t)$ instead.

The 'smearing interval", that is the duration of the channel's impulse response, will however impose a problem. For time-adjacent symbols not to overlap, we need to introduce a *guard time* between each symbol. We may refrain from transmitting anything in the guard time, but this will mean that not all energy will be captured from those symbols that arrive late, which destroys orthogonality.

Orthogonality between the complex exponentials regardless of time of arrival can instead be preserved by letting the waveforms continue to sound also in the guard time. This is efficiently accomplished by first constructing the symbol in the usual way, by adding together waveforms of a symbol duration, and then copying a small part of its end (of the guard time's duration) and adding it to the beginning of the symbol.

By choosing the above set of orthogonal signals and adding this *cyclic prefix*, we have in effect constructed an *orthogonal frequency division multiplex* (OFDM) system. The OFDM system is an example of a *multicarrier system*. There are other kinds, but the OFDM system is the one that most efficiently utilises the radio spectrum.

Since the signal cannot cross over between different complex correlators even when the channel introduces multipath propagation, it is not necessary to regard the entire bandwidth as one single channel. Instead, we may look upon each frequency in the set of, say, N eigenfunctions as a separate *subcarrier*, over which symbols from a 2D signal space travel independently of symbols sent over other subcarriers.

The independence of the subcarriers makes OFDM an excellent system design choice. Firstly, it opens up possibilities for adaptation because different modulation formats may be used on different subcarriers depending on individual subcarrier quality. Secondly, it makes OFDM well suited to be used in a multiuser environment, since different users may be assigned different subcarriers, and this allocation may also change over time. Hence, users in a multiuser OFDM system share both time and bandwidth.

Observe that each sampled symbol in OFDM marks a point in a $2N$-dimensional signal space, where N is the number of subcarriers in the system. Twodimensional cross sections of this space correspond to the individual subcarriers. Hence we observe that OFDM is a modulation format of very high

dimensionality. The symbol rate, on the other hand, is low in OFDM systems; to avoid wasting most of the transmitted energy on overhead, the cyclic prefix needs to be short compared to the symbol period. The symbol period is therefore considerably longer than the maximum length of the channel impulse response. OFDM thus stands in sharp contrast to many other kinds of modulation methods, where signals from a low-dimensional signal space are transmitted at a high rate. Such systems will be studied in the next section.

Having to split up the received signal and sending it through a bank of $2N$ cross correlators may seem unfeasible, given that the number N of subcarriers could very well be several thousands, and it would indeed be unfeasible if we had to do it. Luckily, we don't; sampling the output of N parallel complex correlators once every symbol, is equivalent to sampling the signal at N times the symbol rate and taking the *Fast Fourier Transform* (FFT) of N samples at a time. This means that we only need *one* receiving filter which operates on the whole system bandwidth, from whose output we can produce all $2N$ metrics in $\mathcal{O}(N \log N)$ operations.

Briefly described, an OFDM detector splits the received signal in blocks, cuts away the cyclic prefix, and performs an FFT on the remaining sequence, thus producing the N complex outputs.

The phase shift/attenuation for a certain subcarrier n for a certain symbol t will be denoted the *time-frequency tap* $h_{n,t}$. Hence we have the relationship

$$r_{n,t} = h_{n,t} s_{n,t} + v_{n,t} \tag{3.3}$$

between the transmitted symbol $s_{n,t}$ and the received symbol $r_{n,t}$, where $v_{n,t}$ denotes the noise contribution at symbol time index t and subcarrier n. The complex number $h_{n,t}$ is hence associated with the channel perceived by exactly one *time-frequency symbol*.

A few assumptions are needed for (3.3) to be valid. Firstly, I have assumed that the cyclic prefix is indeed longer than the channel impulse response, and that time syncronisation between transmitter and receiver is perfect. Secondly, is is implicitly assumed that also the carrier synchronisation between transmitter and receiver is perfect. If this is not the case, energy will leak over into adjacent subcarriers.

We now turn to study systems in which one single carrier bears all information.

3.4 Single carrier systems

In the OFDM system, we gave the transmitted symbols some margin of arrival through the cyclic prefix, ensuring that an OFDM symbol delayed in time due to the influence from the channel will not leak over into the next OFDM symbol period. If we instead choose to design the system so that the symbol duration is short compared to the length of the channel's impulse response, then consecutive symbols will overlap, causing *intersymbol interference* (ISI) to occur.

ISI may seem very troublesome, especially when the delay of a particular "echo" of the signal is not a multiple integer of the symbol duration. Then, during one symbol interval, the receiver will perceive that echo not as one isolated symbol, but as two joined fragments of adjacent symbols. We now investigate what implications this has on the received signal.

Let us say that the signal that we transmit is limited to a (baseband) bandwidth $W/2$. Denote the transmitted signal by $s(t)$ and its Fourier transform by $S(f)$. The signal is sent over a channel with impulse response $c(t)$ and frequency response $C(f)$. Following [12], we then have

$$s(t) = \sum_{l=-\infty}^{\infty} s(l/W)\mathrm{sinc}(Wt - l) \tag{3.4}$$

and

$$S(f) = 1/W \sum_{l=-\infty}^{\infty} s(l/W)e^{-j2\pi fl/W}, \quad \text{if } |f| \leq W/2 \text{ and zero otherwise} \tag{3.5}$$

We here only look at baseband signals, but the theory holds also for bandpass signals. The received signal $r(t)$ becomes

$$\begin{aligned} r(t) &= \int_{-\infty}^{\infty} C(f)S(f)e^{j2\pi ft}\mathrm{d}f \\ &= 1/W \sum_{l=-\infty}^{\infty} s(l/W) \int_{-W/2}^{W/2} C(f)e^{j2\pi(t-n/W)}\mathrm{d}f \\ &= 1/W \sum_{l=-\infty}^{\infty} s(l/W) \int_{-\infty}^{\infty} C'(f)e^{j2\pi(t-n/W)}\mathrm{d}f \\ &= \sum_{l=-\infty}^{\infty} s(l/W)c'(t - l/W), \end{aligned} \tag{3.6}$$

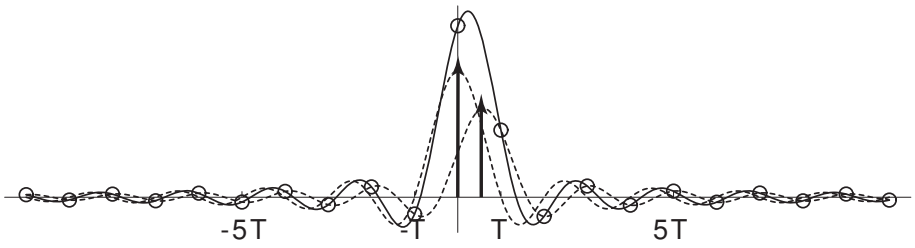


Figure 3.2: The impulse response of an echoing channel, $c(t) = \delta(t) + 0.7\delta(t - 0.6)$, and the corresponding tapped delay line model when the receiver filter is matched to the signal bandwidth $W = 1/2T$. The coefficients of the resulting tapped delay line model stretch far beyond the echo at $0.6T$. Ideal Nyquist (sinc) pulses where used, which allows symbols to be transmitted with period T. Other pulse shapes reduce the length of the sampled impulse response, but also forces a reduction of the symbol rate.

where $c'(t)$ is the impulse response of a channel with the same frequency response as $c(t)$ within the frequency support region $|f| \leq W/2$ and zero otherwise. The result could also be produced by directly convolving (3.4) with the channel impulse response $c(t)$.

We see that the received (noiseless) signal can be constructed from the transmitted signal sampled with frequency W. The resulting impulse response $c'(t)$ may however be long and the ISI severe, depending on the choice of pulse shape. Still, the fact that some echoes enter the correlator "out of synch" does not cause any strange effect other than possibly a long discrete impulse response.

Figure 3.2 illustrates the filter coefficients of the *tapped delay line model* resulting from a direct path and one echo. An important conclusion is that the filter taps stretch far beyond the echo, which occurs fairly close in time to the primary ray, at $0.6T$. The filter model illustrated in Figure 3.2 is non-causal. In practice, one would use a pulse shape that is truncated and delayed, producing a causal tapped delay line model.

The equality (3.6) holds for any bandlimited signal $s(t)$. If $s(t)$ is a modulated information signal on the form

$$s(t) = \sum_{k=-\infty}^{\infty} \alpha_k p(t - kT_s), \tag{3.7}$$

we would rather represent $r(t)$ in terms of the $\{\alpha_k\}$ than in terms of samples taken at a frequency W. To write the received signal as a weighted sum of

the transmitted symbols $\{\alpha_k\}$, we go back and rewrite Equation (3.4) as

$$s(t) = \sum_{l=-\infty}^{\infty} s(lT)p(t - lT_s), \tag{3.8}$$

where $p(t)$ is the pulse shape. Following the previous derivation, we find

$$r(t) = \sum_{l=-\infty}^{\infty} s(lT_s)c \star p(t - lT_s) \tag{3.9}$$

If ideal Nyquist pulses are used, then $W = 1/T$ and the two results (3.6) and (3.9) become identical.

The derivations in this section show that a tapped delay line (single carrier) channel can be modelled by

$$r_t = h_{0,t}s_t + h_{1,t}s_{t-1} + \ldots + h_{N_{CP}-1,t}s_{t-N_{CP}+1} + v_t \tag{3.10}$$

where

$$h_{m,t} = c \star p(t - mT - t_{\text{delay}}) \tag{3.11}$$

$$\tag{3.12}$$

with a sufficient delay t_{delay} to prevent non-causality. Here, N_{CP} is the (maximum) length of the channel impulse response.

Note that, in this section I use $h_{m,t}$ to denote taps (the mth tap in the channel impulse response at time t), whereas in the previous section I used $h_{n,t}$ to denote time-frequency taps (the nth subcarrier at time t). There will be no need to use different notations for the two; the meaning will always be clear from the context.

This concludes the short introduction to digital communications. The rest of this chapter is devoted to the modelling of one tap.

3.5 Making inferences from observed data

The objective of a communications system is to infer on the receiving side the bits transmitted from the sender. In the adaptive multiuser systems considered here, we also want to schedule resources opportunistically to the user that is temporarily able to utilise them the best.

The general problem facing us is as follows. The measurement equation

$$y_t = \varphi_t^* h_t + v_t \tag{3.13}$$

expresses how symbols φ_t distorted by the channel h_t are measured in gaussian noise v_t. In a multiuser system the measurements will comprise several carriers at each time instant. The measurement y_t will then be a vector and φ_t will be a diagonal matrix holding the transmitted symbols. In the single carrier scenario the transmitted symbols will be stacked in a vector φ_t, rendering the measurements scalar. The symbol vector/matrix φ_t is assumed to constitute an entire supersymbol, so that the $\{\varphi_t\}$ are mutually independent. Given that the encoding used is a block code, this can always be achieved by making the dimensionality of the measurements sufficiently large.

Considering the estimation problem, we seek the most probable transmitted symbols given the measurements, which is accomplished by finding the maximum a posteriori value (the peak) of the distribution

$$p(\varphi_t|Y_tI), \tag{3.14}$$

where Y_t represents all passed measurements y_t, y_{t-1}, y_{t-2} and so on. Applying the rules of probability theory, we have

$$\begin{aligned} p(\varphi_t|Y_tI) &= \int p(\varphi_t|Y_th_tI)p(h_t|Y_tI)dh_t \\ &\propto p(\varphi_t|Y_{t-1}h_tI)\int p(y_t|\varphi_th_tY_{t-1}I)p(h_t|Y_tI)dh_t \\ &= p(\varphi_t|I)\int p(y_t|\varphi_th_tI)p(h_t|Y_tI)dh_t \end{aligned} \tag{3.15}$$

Since φ_t corresponds to a codeword, the distribution $p(\varphi_t|Y_{t-1}h_tI)$ reduces to the prior $p(\varphi_t|I)$. The pdf for h_t given passed measurements is very difficult to compute since it will require marginalisation over all unknown transmitted symbols. Instead, by approximating $p(h_t|Y_tI)$ with a pdf $p(h_t|\bar{Y}_tI)$ based on a subset $\bar{Y}_t$ of all measurements Y_t, the problem is transformed into feasible form. $\bar{Y}_t$ is based exclusively on known transmitted symbols which are commonly termed *pilot* symbols. As will be demonstrated in Chapter 5, $p(h_t|\bar{Y}_tI)$ will then turn out to be gaussian with some mean value $\hat{h}_{t|t}$ and some covariance matrix A. The pdf that we seek is now written

$$\begin{aligned} p(\varphi_t|Y_tI) &\propto p(\varphi_t|I)\int p(y_t|\varphi_th_tI)p(h_t|Y_tI)dh_t \\ &\approx p(\varphi_t|I)\int p(y_t|\varphi_th_tI)p(h_t|\bar{Y}_tI)dh_t \\ &= p(\varphi_t|I)\int \mathcal{CN}(y_t;\varphi_t^*h_t,R)\times\mathcal{CN}(h_t;\hat{h}_{t|t},A)dh_t \end{aligned} \tag{3.16}$$

One may assume that the uncertainty in the tap estimation as expressed by A is very low at the moment of detection. Letting A go to zero, the second gauss distribution in 3.16 will approach a Dirac distribution and hence we have that $p(\varphi_t|Y_tI)$ is approximately proportional to

$$p(\varphi_t|I) \times \mathcal{CN}(y_t; \varphi_t^* \hat{h}_{t|t}, R) \propto p(\varphi_t|I) \times \exp\left(-\frac{1}{2}(y_t - \varphi_t^* \hat{h}_{t|t})^* R^{-1} (y_t - \varphi_t^* \hat{h}_{t|t})\right) \quad (3.17)$$

How do we use this to find the most probable transmitted sequence φ_t? First, let us look at an OFDM system where the measurements are vector valued and where each symbol is affected by one time-frequency tap only. The symbol φ_t^* then represents a diagonal matrix with the transmitted symbols along the diagonal.

To proceed we rewrite the vector $\varphi_t^* \hat{h}_{t|t}$. Swapping vector and matrix structures, we may express this vector as $\bar{h}_{t|t}\bar{\varphi}_t$, where the bars indicate the changes in structure. $\bar{h}_{t|t}$ is here a matrix with the elements of $\hat{h}_{t|t}$ along its diagonal, and $\bar{\varphi}_t$ is a column vector holding the transmitted symbols (note that $\bar{\varphi}_t$ and φ_t^* contain the same values; the conjugate operator is closely associated with transposing and is therefore unsuited for use on column vectors).

We now have

$$\begin{aligned} p(\varphi_t|I) &\times \exp\left(-\frac{1}{2}(y_t - \varphi_t^* h_t)^* R^{-1} (y_t - \varphi_t^* h_t)\right) \\ &= p(\varphi_t|I) \times \exp\left(-\frac{1}{2}(y_t - \bar{h}_{t|t}\bar{\varphi}_t)^* R^{-1} (y_t - \bar{h}_{t|t}\bar{\varphi}_t)\right) \\ &= p(\varphi_t|I) \times \exp\left(-\frac{1}{2}(\bar{h}_{t|t}^{-1} y_t - \bar{\varphi}_t)^* \bar{h}_{t|t}^* R^{-1} \bar{h}_{t|t} (\bar{h}_{t|t}^{-1} y_t - \bar{\varphi}_t)\right) \end{aligned} \quad (3.18)$$

The estimation (detection) problem is to find the peak of $p(\varphi_t|Y_tI)$ which is now trivial; we choose the sequence $\bar{\varphi}$ in the set of all possible sequences, expressed by the prior $p(\varphi_t|I)$, that is closest to the mean $\bar{h}_{t|t}^{-1} y_t$ if the gauss distribution. If the noise covariance R is diagonal and $p(\varphi_t|I)$ allows any combination of symbols, then, since the matrix $\bar{h}_{t|t}^* R^{-1} \bar{h}_{t|t}$ is also diagonal, we get the following algorithm for choosing the most probable transmitted sequence:

$$\arg\max p(\varphi_t|Y_tI) = \text{elementwise hard decisions on } \bar{h}_{t|t}^{-1} y_t \quad (3.19)$$

That is, all we have to do is to "derotate" the received signal with the estimated channel, and then round to the nearest proper symbol sequence.

Consider now instead a TDMA system where φ_t is a vector. Looking at (3.17), one might think that φ_t should be straighforward to find also in this case. Unfortunately, intuition defies fact at this point, and the most likely sequence can only be found by thoroughly searching through a large set of candidates. If the noise is uncorrelated between different taps so that R is diagonal, this search need not be exhaustive. One may instead use the *Viterbi* algorithm to find the correct sequence, which has exponential complexity in the length of the sequence.

We turn now to the prediction problem. Some measure of error, such as packet error, symbol error, bit error, is calculated and used as basis for resource scheduling decisions made by a centralised scheduler. Exactly how this is done depends among other things on how much feedback information one can tolerate on the radio link. In this thesis I imagine yes-or-no answers for whether the expected bit error rate is below some predefined limit being signaled to the base station. This way, only one bit of information needs to be transmitted per user and candidate modulation/coding format[3].

For the cases considered here, error rates are always functions of the magnitude of the channel taps; the better the channel, the lower the error rate. Hence we have some function of, say, the bit error rate

$$P_b(|h|) \tag{3.20}$$

that decreases with increasing $|h|$. For single carrier systems where the channel impulse response has many taps, all taps have to be taken into account in $P_b(|h|)$. For example, it is much easier to get a good error rate performance if the impulse response energy is concentrated to a few taps than if it is spread out over the whole impulse response.

Marginalising over the unknown tap(s), we have

$$p(P_b|YI) = \int P_b(|h|)p(h|YI)\mathrm{d}h \tag{3.21}$$

Again, we see that we need the distribution $p(h|YI)$, which we also here approximate with $p(h|\bar{Y}I)$. However, we will here be concerned with *predicting* the error rate (and hence the channel) a few steps ahead, to compensate for

[3]The size of the feedback information is actually less than that. In a system where each user can choose from a set of eight modulation/coding formats, only the best of these that can be used needs to be signalled back to the base station. This means that the feedback information constitutes three bits, not eight.

the delay introduced by the control loop. Luckily, as we will see in Chapter 5, channel prediction and estimation fit into the same framework.

The above discussion shows that both the estimation and the prediction problems may be solved by inferring the channel taps. The large part of the remainder of this thesis will therefore be concerned with channel estimation and prediction.

3.5.1 Tap modelling

Inference begins with constructing a model, and so I therefore commence by constructing a model for one single channel tap. In Chapter 4 I will expand this model to account for multiple taps in multiuser systems. The actual inference is then considered in Chapter 5.

What does an adequate model for a fading tap look like? We could attempt to build the model *ab initio*, directly using the laws of physics (Maxwell's equations mainly) and the uncertainty we have about whatever parameters are included. A model built from first principles is very powerful in the respect that it will always yield reasonable answers, heedless of whatever extreme questions we may ask. There is no risk of getting a results that is caused by modelling errors. Its weakness is that it is rarely possible to construct such a model at all, since a fully correct physical description of a situation almost always is to complicated to design; this is indeed the situation that we face here.

At the other end of the scale is *phenomenological* modelling. Phenomenological modelling is done by first observing (measuring) the process of interest, followed by the construction of a model that mimics the process' behaviour.

In practice, we use something in between the two methods of modelling: phenomenological modelling supported by physical reasoning. For radio channel modelling, it is common to let the chain of reasoning leading to a model start at physical principles, and every now and then along the route make simplifications and abstractions based on observations of the actual process (the radio channel). See e.g. [13], [14].

For illustration, I will here begin closer to the phenomenological end of the scale. Figure 3.3 shows the complex impulse response of a radio channel measurement in suburban Stockholm. The channel, sampled at approximately 9.1 kHz, clearly contains many taps, although it is dominated by only a few. The taps rotate around the origin with an angular frequency depending on the angle of arrival of the respective rays. The strongest tap moves about 90° in 16 samples and thus makes one revolution in about 7 ms. One revo-

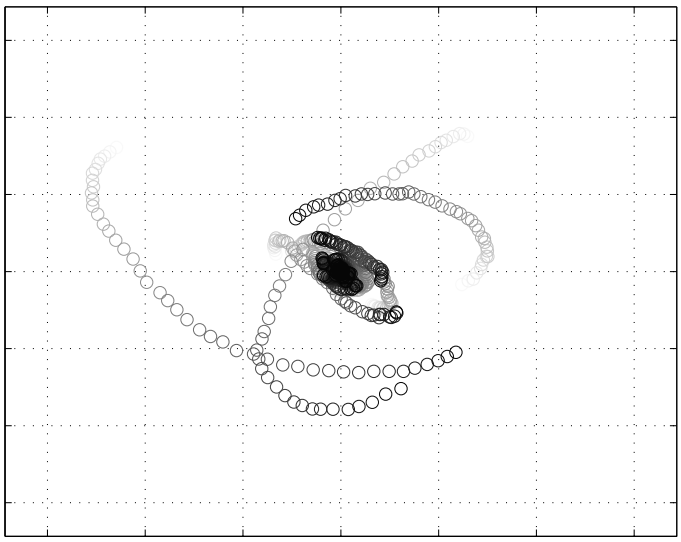


Figure 3.3: Impulse response measurements of a 1800 MHz radio channel in suburban Stockholm. The received signal was correlated with the known transmitted sequence. The figure has no particular scaling, but both axes span the same range.

lution around the origin means moving through one wavelength, which here is $3 \cdot 10^8/1.8 \cdot 10^9$ m. This corresponds to about 85 km per hour, which is probably the velocity of the vehicle taking the measurements. Details regarding the measurements can be found in [14]. Although this particular measurement concerns a single carrier system, the general behaviour of each individual tap doesn't deviate from those in an OFDM system.

The most obvious feature of the time varying impulse response is the tendency of the taps to rotate about the origin. This is quite expected; when the receiver moves one wavelength in the direction of the propagation path corresponding to a certain tap, then, accordingly, that tap will make one revolution. A first attempt to model a tap h_t could then be

$$h_{t+1} = \alpha_t e^{j\Delta\omega_t} h_t, \tag{3.22}$$

where $\Delta\omega_t$ would be a function of the mobile unit's velocity (including direction), and the attenuation α_t would usually be slowly varying, determined in large by the velocity with which the receiver approaches or moves away from nearby reflectors. Uncertainty about $\Delta\omega_t$ and α_t needs to be taken into consideration according to the Bayesian methodology, which makes the model (3.22) nonlinear.

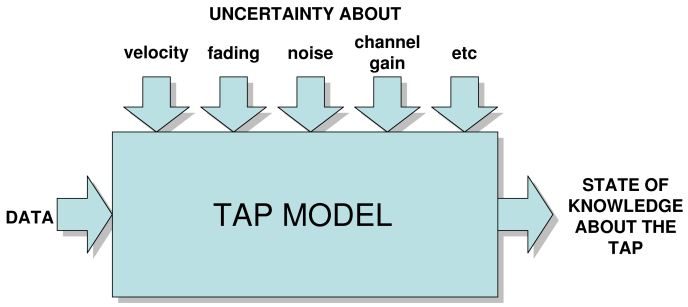


Figure 3.4: Schematic illustration of the structure of an optimal, Bayesian tap estimator/predictor.

The tap model (3.22) is, directly or indirectly, dependent on parameters such as the velocity v, the tap power σ_h^2, and suitable fading statistics descriptors. A correct application of Bayesian theory would take into account all prior knowledge about these parameters to produce optimal inferences of h_t from measurements of some sort. This is illustrated in Figure 3.4.

However, as indicated by Equation (3.22), such parameters enter the model in a nonlinear way. This makes the inference problem hard, because each pdf update as given by Bayes' theorem would generally yield a new type of pdf. The number of parameters required to describe each new pdf would grow without limit. Using some kind of intermediate approximation/simplification step would be crucial for the utility of such a model.

Had the model instead been linear, then, quite remarkably, Bayes' propagation would yield gaussian distributions all the time, keeping the degrees of freedom constant. I describe this in detail in Chapter 5.

We can "force" the model to become linear by neglecting to propagate the uncertainties about the "nonlinear" parameters, and instead separate optimal or suboptimal estimators for the nonlinear parameters from the linear tap model. These estimators will then occasionally update the linear model with *point estimates* of the nonlinear parameters. This is illustrated in Figure 3.5.

How big is the difference between the suboptimal model illustrated in Figure 3.5 and the optimal scheme depicted in Figure 3.4? If the parameter estimates are very accurate (of low uncertainty), then the suboptimal model will agree extremely well with the optimal model. In this thesis I will assume that such accurate point estimates are available.

As is shown in [14], a fading radio channel may be accurately modelled by an *autoregressive moving average* (ARMA) model. For the remainder of this chapter, I will restrict the model used to a general *autoregressive* (AR)

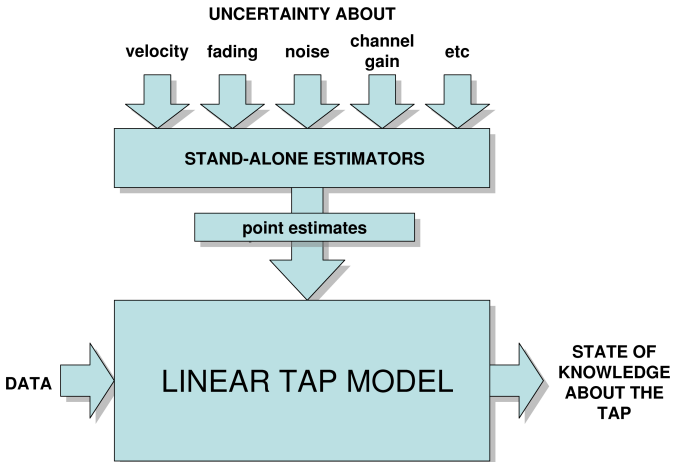


Figure 3.5: Schematic illustration of the structure of the sub-optimal tap estimator/predictor used in the present thesis.

model, which may approximate an ARMA model arbitrary well by making the model order sufficiently large. However, the modelling principle to be described is valid for general ARMA models.

A linear AR tap model of order K can be described on transfer function form:

$$h_t = -a_{1,t}h_{t-1} - a_{2,t}h_{t-2} - \ldots - a_{K,t}h_{t-K} + u_{t-1} \tag{3.23}$$

The *process noise* u_{t-1} is zero mean gaussian and independent between time samples (white). Its variance implicitly determines the variance of the tap h_t. For practical reasons that will be evident in Section 3.5.3, I lag the process noise one time sample compared to the modelled tap. For the behaviour of the process h_t, this is of no importance since u_t is white.

It should also be pointed out that the time variability of the coefficients $\{\alpha_{k,t}\}$ will be much slower than that of the process h_t, which makes the task easier for the stand-alone estimators that provide the coefficients.

The model (3.23) is actually well capable of representing the prior knowledge that we wish to put into the model. We would have to be in possession of quite specific prior information for it to be inadequate. The gaussian property of the plant noise will cause also h_t to be gaussian. This is well motivated by the fact that the tap can be seen as the sum of infinitely many radio propagation paths. Although we do not know which pdf correctly represents our state of knowledge about one single path, we know from the central limit theorem that what we know about the *sum* of paths is correctly

represented by a gaussian pdf.

How do we choose the $\{a_{k,t}\}$ so that it accounts for our prior knowledge about the channel conditions? Assuming a static surrounding and that we know the velocity v, the standing wave pattern formed around the mobile unit is composed of a set of wavelengths having the carrier wavelength as upper limit. This causes the perceived channel taps to fade with a frequency no greater than $f_D = f_c \cdot v/c_0$, where f_c is the carrier frequency and c_0 is the speed of light.

The frequencies with which a model allows a tap to fade are given by the model's *doppler spectrum*. Taking the z-transform of model (3.23) (skipping the time index),

$$H(z) = \frac{z^{-1}}{1 - a_1 z^{-1} + \ldots + a_K z^{-K}} = \frac{z^{-1}}{(1 - p_0 z^{-1})(1 - p_1 z^{-1}) \ldots (1 - z_{K-1} z^{-1})}, \quad (3.24)$$

allows us to place the poles $\{p_k\}$ so as to correctly represent our prior knowledge about the channel behaviour (The doppler spectrum for normalised frequencies f, so that $-1 < f < 1$, is then given by $S(f) = |H(j2\pi f)|^2$).

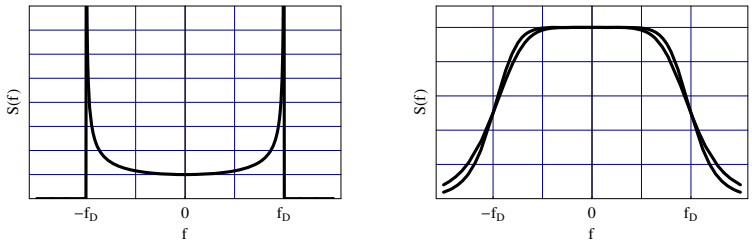
The off-line estimators may now be designed to provide estimates of the coefficients $\{a_{k,t}\}$ or the poles $\{p_{k,t}\}$ directly, or they may produce estimates of physical parameters, which implicitly gives the values of $\{a_{k,t}\}$. In the example that will be given here, I use the latter approach and employ a model that is determined by the vehicular velocity only.

A commonly used fading model is the so called *Jakes model*. It is derived from the assumption that radio wave scatterers are distributed on the circumference of a circle around the mobile unit. This produces the Jakes doppler spectrum[4]:

$$S_{\text{Jakes}}(f) = \begin{cases} \frac{1}{\sqrt{1-(f/f_D)^2}} & |f| < f_D \\ 0 & \text{otherwise} \end{cases}, \quad (3.25)$$

which strongly favours frequencies near the maximum doppler frequency f_D. See figure 3.6(a).

[4]Normalisation is not important here, since the integral of the doppler spectrum depends on the total "power" of the signal (if we consider the fading channel tap to be a signal) which varies from case to case. Commonly, the doppler spectrum is normalised so that the integral evaluates to unity. One then has to add a normalising constant with value $1/\pi f_D$.



(a) The Jakes doppler spectrum.

(b) Doppler spectra for Butterworth filters of order 3 and 4.

Figure 3.6: Doppler spectra for different models.

Environments such as urban surroundings dominated by streets, in which sideways scatterers are closer than scatterers located in the direction of movement, will produce a doppler spectrum that is more flat than the Jakes' spectrum. Estimation and prediction on such models is necessarily more difficult than on the Jakes' model, since no single frequency dominates over any other. As this model represents a kind of "worst case", I will use it as a case study.

3.5.2 A flat doppler spectrum

A perfectly flat model, as well as the Jakes' model, would have a doppler spectrum with a discontinuity at $f = \pm f_D$, but this cannot be described by a linear model such as (3.23). We necessarily need to resort to approximations. A good candidate is the *Butterworth* filter. The power frequency response for the Butterworth filter (which corresponds to the doppler spectrum in channel modelling) is given by

$$S_{\text{Butterworth}}(f) = \frac{1}{1 + (f/f_D)^{2K}}. \tag{3.26}$$

Of all linear filters, it is the one that has least amount of ripple in the passband.

A (continuous-time) Butterworth filter has poles distributed on a semicircle around the origin with a radius given by the normalised cut-off frequency:

$$p_{c,k} = 2\pi f_D \exp\left(j\frac{(K + 2k + 1)\pi}{2K}\right), \quad k = 0..K-1 \tag{3.27}$$

Note that the cutoff frequency in the present context is the same thing as the doppler frequency.

Discretising a continuous-time filter to a model with sampling time T_s inevitably involves approximations. We may calculate the time discrete model poles $\{p_k\}$ by directly taking $p_k = \exp(p_{c,k}T_s)$(this is called the *matched z transform*), but this introduces aliasing by repeatedly wrapping the frequency axis around the unit circle. Alternatively, we use the *bilinear transformation*

$$p_k \approx \frac{1 + p_{c,k}T_s/2}{1 - p_{c,k}T_s/2}, \tag{3.28}$$

which preserves features but distorts frequencies. Here I will use the latter.

Since the bilinear transform causes frequency distortion in the conversion from the time continuous domain to the time discrete domain, one usually "prewarps" essential frequencies so that they come out correctly after the transform. A frequency f is prewarped through

$$f_{\text{p.w.}} = \frac{2}{T_s}\tan(\pi f T_s) \tag{3.29}$$

Hence we find the poles of a model that represents a flat doppler spectrum by calculating the doppler frequency $f_D = f_c v/c_0$ from the velocity v, prewarping f_D according to (3.29), calculating the continuous-time poles $p_{c,k}$ from (3.27), and finally applying the bilinear transformation (3.28) to calculate the discrete-time model poles $\{p_k\}$.

The filter order K determines the rolloff of the filter (which can be shown to be $-6K$ dB per octave). See Figure 3.6(b). It is however important not to make the filter flanks too steep. The reason is that the strict absence of frequencies outside $-f_D < f < f_D$ is true only in a static environment. In reality, moving objects which surround the mobile unit will cause the doppler spectrum to spread outside the $\pm f_D$ limits. For examples of doppler measurements that illustrate this fact, see [14]. The linear model's inability to represent a box-shaped doppler spectrum therefore actually turns to our advantage, since it constitutes a better representation of our prior knowledge. Note that the difference between using a third-order filter and a fourth-order filter is small, and that difference gets even smaller as the model order is increased. As we shall see later, the filter order affects the computational complexity quite severely. A filter of order three or four should be adequate for our purposes.

3.5.3 State space model

The model (3.23) is easily translated to state-space representation by use of e.g. the controllable canonical form

$$\begin{aligned} x_{t+1} &= \underbrace{\begin{pmatrix} -a_{1,t} & -a_{2,t} & \dots & -a_{K,t} \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{pmatrix}}_{F} x_t + \underbrace{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{G} u_t \\ h_t &= \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}}_{H} x_t \end{aligned} \tag{3.30}$$

The reason for earlier lagging the process noise u_t one time sample is now evident; doing so allows us to construct a state space representation of lowest possible order.

For several reasons (complexity, stability, ease of finding process noise covariance) it is convenient to have a diagonal F matrix. We construct a *"diagonal form"* by diagonalising F by means of eigenvalue decomposition. Skipping time indices to ease the notation, we have

$$F = V \Lambda V^{-1}, \tag{3.31}$$

where V are the eigenvectors and Λ is a diagonal matrix consisting of the eigenvalues of V. By constructing new states through $x_{new} = V^{-1} x_{old}$ and multiplying (3.30) with V^{-1} from the left, we have

$$\begin{aligned} x_{t+1,new} &= \Lambda x_{t,new} + V^{-1} G e_t \\ h_t &= H V x_{t,new} \end{aligned} \tag{3.32}$$

This means that our new diagonalised form is given by $F_{new} = \Lambda$, $G_{new} = V^{-1} G_{old}$, and $H_{new} = V H_{old}$.

When we now have the model poles p_k, we have only to expand the denominator of (3.24) and insert the resulting $\{a_k\}$ into the state space form (3.32). It is desirable to avoid having to perform the eigenvalue decomposition (3.31) every time the model is updated. I will therefore present closed form expressions for the matrices in the diagonal state space form.

The eigenvalue decomposition on the companion form turns out to be

$$\Lambda = \begin{pmatrix} p_0 & & & \\ & p_1 & & \\ & & \ddots & \\ & & & p_{K-1} \end{pmatrix}, \quad V = \begin{pmatrix} rp_0^{K-1} & p_1^{K-1} & \dots & p_{K-1}^{K-1} \\ \vdots & \vdots & & \\ p_0 & p_1 & \dots & p_{K-1} \\ 1 & 1 & \dots & 1 \end{pmatrix}, \tag{3.33}$$

where V is a so called *Vandermonde* matrix.

The inverse V^{-1} is somewhat more involved. Equation (3.34) shows V^{-1} for the particular case $K = 4$.

$$V^{-1} = \begin{pmatrix} \frac{1}{(p_0-p_1)(p_0-p_2)(p_0-p_3)} & -\frac{p_1+p_2+p_3}{(p_0-p_1)(p_0-p_2)(p_0-p_3)} \\ \frac{1}{(p_1-p_0)(p_1-p_2)(p_1-p_3)} & -\frac{p_0+p_2+p_3}{(p_1-p_0)(p_1-p_2)(p_1-p_3)} \\ \frac{1}{(p_2-p_0)(p_2-p_1)(p_2-p_3)} & -\frac{p_0+p_1+p_3}{(p_2-p_0)(p_2-p_1)(p_2-p_3)} \\ \frac{1}{(p_3-p_0)(p_3-p_1)(p_3-p_2)} & -\frac{p_0+p_1+p_2}{(p_3-p_0)(p_3-p_1)(p_3-p_2)} \end{pmatrix}$$
$$\begin{pmatrix} \frac{p_1p_2+p_3p_2+p_1p_3}{(p_0-p_1)(p_0-p_2)(p_0-p_3)} & -\frac{p_1p_2p_3}{(p_0-p_1)(p_0-p_2)(p_0-p_3)} \\ \frac{p_0p_2+p_3p_2+p_0p_3}{(p_1-p_0)(p_1-p_2)(p_1-p_3)} & -\frac{p_0p_2p_3}{(p_1-p_0)(p_1-p_2)(p_1-p_3)} \\ \frac{p_0p_1+p_3p_1+p_0p_3}{(p_2-p_0)(p_2-p_1)(p_2-p_3)} & -\frac{p_0p_1p_3}{(p_2-p_0)(p_2-p_1)(p_2-p_3)} \\ \frac{p_0p_1+p_2p_1+p_0p_2}{(p_3-p_0)(p_3-p_1)(p_3-p_2)} & -\frac{p_0p_1p_2}{(p_3-p_0)(p_3-p_1)(p_3-p_2)} \end{pmatrix} \tag{3.34}$$

For general matrix sizes $0 \leq i, j \leq K-1$, the individual elements of the respective matrices are

$$\begin{aligned} \Lambda[i,j] &= p_i\delta_{ij} \\ V[i,j] &= p_j^{K-1-i} \\ V^{-1}[i,j] &= (-1)^j \frac{\sum(\prod \text{All combinations of } (\{p_k\} \backslash r_i))}{\prod_{k \neq i}(p_i - p_k)} \end{aligned} \tag{3.35}$$

Now we just have to carry out the appropriate matrix multiplications to formulate the diagonal form (3.32). This gives

$$F = \begin{pmatrix} p_0 & & & \\ & p_1 & & \\ & & \ddots & \\ & & & p_{K-1} \end{pmatrix}, G = \begin{pmatrix} \frac{1}{(p_0-p_1)(p_0-p_1)\cdots(p_0-p_{K-1})} \\ \frac{1}{(p_1-p_0)(p_1-p_2)\cdots(p_0-p_{K-1})} \\ \vdots \\ \frac{1}{(p_{K-1}-p_0)(p_{K-1}-p_1)\cdots(p_{K-1}-p_{K-2})} \end{pmatrix}$$

$$H = \begin{pmatrix} p_0^{K-1} & p_1^{K-1} & \dots & p_{K-1}^{K-1} \end{pmatrix} \tag{3.36}$$

The diagonal $K \times K$-matrix F and the vectors G and H now constitute the single tap model that we sought for. To complete the model it is also necessary to compute the variance for the process noise u_t, so that the model generates the variance of h_t that we desire (just as with the velocity and the fading statistics, this variance is assumed to be given by estimators external to the linear model). This problem will however have to be considered in a larger context, where the covariance between different taps are taken into account. I therefore postpone the issue of determining the process noise variance to the next chapter.

Finally, it is worth repeating that the above derivation of the model for one (time-frequency) tap requires us to know the values of a few parameters exactly. These are the variance of the tap, the velocity of the mobile unit, and suitable descriptors of the fading statistics (here represented as the poles $\{p_k\}$). Needless to say, we cannot know them exactly. The correct Bayesian way would be to take into account also the uncertainties that we have about the above parameters. Unfortunately they would enter the model equations nonlinearly (for the case of the velocity, this is indicated by Eq. (3.22)), which would effectively terminate our attempts to infer the channel taps.

It is therefore crucial that the algorithms that provide our model with estimates of above parameters, operates with such an accuracy that the uncertainties are virtually zero.

3.6 Summary

The present chapter discusses the construction of a state space model for one single fading tap. It is assumed that off-line estimators provide a linear tap model with highly accurate estimates of the model parameters. These parameters may be the coefficients or poles and zeros of the linear model, in which case the state space representation is constructed according to (3.36) (the example given here is for an AR model, but the structure generalises straightforwardly also to ARMA models).

The estimated parameters may also be of a more direct physical character. In this chapter a very cautious model is suggested, which assigns approximately the same probability to all frequencies of fading up to the maximum doppler frequency. The algorithm for finding the poles for this model is given by the following steps:

- Achieve an estimate of high accuracy of the velocity v.
- Calculate the doppler frequency : $f_D = f_c v/c_0$, where f_c=carrier fre-

quency and c_0=speed of light.

- Calculate the discrete time (normalised) doppler frequency : $\Omega_D = 2\pi f_D T_s$.
- Calculate the prewarped time continuous doppler frequency : $f_{D,\text{p.w.}} = 2/T_s \tan(\pi f_D T_s)$.
- Calculate the Butterworth poles : $p_{c,k} = 2\pi f_{D,\text{p.w.}} \exp j\theta_k$ whit arguments $\theta_k = \pi(2k + 1 + K)/2K$, $k = 0 \ldots K - 1$.
- Make a bilinear mapping of the poles to the discrete time domain : $p_k = (2/T_s + p_{c,k})/(2/T_s - p_{c,k})$.

Chapter 4

The modelling of frequency-selective channels for many users

Chapter 3 was devoted to the construction of a model for a single fading tap, either subcarrier channel $h_{n,t}$ in a multicarrier system, or an impulse response tap $h_{m,t}$ in a single carrier system. A linear model for a tap h,

$$\begin{aligned} x_{t+1} &= Fx_t + Gu_t, \\ h_t &= Hx_t, \end{aligned} \tag{4.1}$$

was conceived under the assumption that optimal or suboptimal estimators of parameters such as the mobile unit velocity, the tap power, and suitable descriptors of the fading environment (for example in the form of a doppler spectrum), are available and ready to update the linear model when conditions change.

In this chapter I generalise the single tap model to represent general multiuser single carrier and multicarrier systems. As before, a linear model is feed with certain parameter estimates from "external" estimators. It is important to realise that this model is suboptimal and cannot lead to optimal, Bayesian inferences unless these parameter estimates are extremely accurate. The actual algorithm used for conducting the inferences – the Kalman filter – is introduced in Chapter 5.

To allow for efficient inference in later chapters, state space models are used here. A problem that usually faces the state space model constructor is the determination of the process noise covariance, given that the actual process covariance is known. For the model structures considered here, I give an explicit formula for determining this covariance.

Before I delve into the details of modelling the channels for multiple users, a few words on the notation is appropriate. The matrices used in this chapter will be of block diagonal type. There will be a hierarchical structure built into the models that we are about to construct, the levels being, in ascending order, the tap/subcarrier level, the user level, and the time level. A matrix representing some part of our state of knowledge about a specific tap therefore needs to be identified by indices for all three of these levels, whereas the corresponding "entire system" matrix only needs the time index. However, since all matrices in a model refer to the same time instant, for most cases the time index will be dropped to reduce notational clutter. The superscript (t) will be used to indicate the tap/subcarrier level, and (u) will indicate the user level. Hence I will use the notation

$X_{n,u}^{(\mathrm{t})}$	Matrix referring to the time-frequency tap n for user u
$X_{m,u}^{(\mathrm{t})}$	Matrix referring to the impulse response tap m for user u
$X_{u}^{(\mathrm{u})}$	Matrix referring to all taps for user u
X	Matrix referring to the entire system

Also, when we look at downlinks (that is, the link from the base station to the mobile unit), the inference will be carried out on the mobile unit side. We may therefore drop also the u index in the downlink case, since there is only one user to consider.

Throughout this thesis, I assume the use of only one antenna at both the transmitter and receiver. The formalism is however equally valid for multiple antennas, since each antenna may be look upon as an individual "user", competing for resources against antennas at other users but also against other antennas at the same user.

4.1 General system model

Thus far I have only described how to correctly represent the knowledge that we have about the behaviour of one single tap. In a communication system we will experience the influence from many taps at the same time, how many and in which way depending on which kind of system we are examining. Representing several taps is generally accomplished by constructing a *block-diagonal* state space. To exemplify, assume a one-user system (so that the u index may be dropped). A convenient state space representation of all its

taps would be

$$x_{t+1} = \underbrace{\begin{pmatrix} F_1^{(\mathrm{t})} & & & \\ & F_2^{(\mathrm{t})} & & \\ & & F_3^{(\mathrm{t})} & \\ & & & \ddots \end{pmatrix}}_{F} x_t + \underbrace{\begin{pmatrix} G_1^{(\mathrm{t})} & & & \\ & G_2^{(\mathrm{t})} & & \\ & & G_3^{(\mathrm{t})} & \\ & & & \ddots \end{pmatrix}}_{G} u_t$$

$$h_t = \underbrace{\begin{pmatrix} H_1^{(\mathrm{t})} & & & \\ & H_2^{(\mathrm{t})} & & \\ & & H_3^{(\mathrm{t})} & \\ & & & \ddots \end{pmatrix}}_{H} x_t \tag{4.2}$$

Each triplet of blocks $F^{(\mathrm{t})}$, $G^{(\mathrm{t})}$, and $H^{(\mathrm{t})}$ is a model constructed in accordance with Chapter 3, and h_t is now a *vector* holding all modelled taps.

The block-diagonal forms introduced above will be so commonly employed that we need a compact notation. In the case where all blocks are the same we could use the Kronecker product, here indicated by $\otimes$, in combination with the identity matrix. Hence would a matrix consisting of five identical blocks be denoted by $\mathbf{I}_5 \otimes X$. Our needs are more general, so I will use the notation

$$\mathrm{diag}(X_l, l = 1..L) \triangleq \begin{pmatrix} X_1 & & & \\ & X_2 & & \\ & & \ddots & \\ & & & X_L \end{pmatrix} \tag{4.3}$$

to indicate a general L-block matrix. In many cases the number of blocks will be obvious, so it will suffice to use the short notation

$$\mathrm{diag}(X_l) \triangleq \mathrm{diag}(X_l, l = 1..L) \tag{4.4}$$

It is now a simple matter to construct a model for the fading taps of U users having channels of M fading taps each:

$$\begin{aligned} x_{t+1} &= \mathrm{diag}(\mathrm{diag}(F_{m,u}^{(\mathrm{t})}, m = 1..M), u = 1..U)x_t \\ &\quad + \mathrm{diag}(\mathrm{diag}(G_{m,u}^{(\mathrm{t})}, m = 1..M), u = 1..U)u_t \\ &= Fx_t + Gu_t \\ h_t &= \mathrm{diag}(\mathrm{diag}(H_{m,u}^{(\mathrm{t})}, m = 1..M), u = 1..U)x_t \\ &= Hx_t \end{aligned} \tag{4.5}$$

To be able to make inferences about the taps h, we need some kind of measurements, but before introducing these, we need to make sure that the elements of the vector h are scaled correctly.

4.1.1 The process noise covariance

Recall that we require estimators of a number of parameters to operate beside the linear model, feeding it with new estimates whenever conditions change. Among those are the tap variances (or tap powers) $\{\sigma_m^2\}$. Actually, what we need is the entire *covariance matrix* R_h for the taps. However, often one assumes that the taps fade independently, and so will R_h be diagonal with the $\{\sigma_m^2\}$ along its diagonal. In any case, henceforth I will assume R_h to be known.

The model (4.5) is "driven" by the white (generally vector valued) process noise u. The covariance Q for the process noise need to be adjusted so as to produce the correct R_h. For a general model, this is an extremely difficult problem. However, the special case considered here possesses special properties that yields a direct solution. We use the fact that

- the vectors h and u have the same dimensions,
- the matrices G and H are block diagonal, and
- the matrix F is diagonal.

Calculating the covariance matrices of h and x from (4.5), we have

$$R_h = H\overline{\Pi}H^*, \tag{4.6}$$

where $\overline{\Pi}$ is the covariance of the states x solving the Lyapunov equation

$$\overline{\Pi} = F\overline{\Pi}F^* + GQG^* \tag{4.7}$$

Since all matrices are block diagonal the Lyapunov equation can be written on block form:

$$\begin{pmatrix} \overline{\Pi}_{1,1} & \overline{\Pi}_{1,2} & \cdots \\ \overline{\Pi}_{2,1} & \overline{\Pi}_{2,2} & \\ \vdots & & \ddots \end{pmatrix} = \begin{pmatrix} F_1\overline{\Pi}_{1,1}F_1^* & F_1\overline{\Pi}_{1,2}F_2^* & \cdots \\ F_2\overline{\Pi}_{2,1}F_1^* & F_2\overline{\Pi}_{2,2}F_2^* & \\ \vdots & & \ddots \end{pmatrix} + \begin{pmatrix} q_{1,1}G_1G_1^* & q_{1,2}G_1G_2^* & \cdots \\ q_{2,1}G_2G_1^* & q_{2,2}G_2G_2^* & \\ \vdots & & \ddots \end{pmatrix}, \tag{4.8}$$

where the $\{q_{m,n}\}$ are the elements of Q.

Since F is diagonal, we may rewrite the product $F\overline{\Pi}F^*$ as

$$F\overline{\Pi}F^* = \overline{\Pi} \otimes ff^* \tag{4.9}$$

Here, f is a vector holding the diagonal elements of F, and $\otimes$ is the elementwise multiplication operator. We may now solve the Lyapunov equation:

$$\overline{\Pi} = GQG^* \oslash (\mathbf{1} - ff^*), \tag{4.10}$$

where $\oslash$ denotes elementwise division. The symbol $\mathbf{1}$ stands for a matrix in which all elements are ones.

Finally, using (4.6) and (4.10) and the fact that also H is block diagonal, we have the element (m, n) of R_h:

$$[R_h]_{m,n} = q_{m,n}H_m(G_mG_n^* \oslash (\mathbf{1} - f_mf_n^*))H_n^*, \tag{4.11}$$

and so we may calculate Q in one fell swoop:

$$Q = Rh \oslash H(G\mathbf{1}G^* \oslash (\mathbf{1} - ff^*))H^* \tag{4.12}$$

We will study several different types of systems, but the process noise covariance can always be calculated according to (4.12).

4.1.2 The measurements

We now add the measurement equation, which explicitly expresses the form in which we receive new information. Generally, the measurements consist of transmitted symbols or some transformation of transmitted symbols, distorted by the channel (which is the centre of attention here). The measurements also contain an additive element of noise. Hence we have a *measurement equation* that typically is

$$y_t = \varphi_t Dh_t + v_t = \underbrace{\varphi_t DH_t}_{C_t} x_t + v_t = C_t x_t + v_t, \tag{4.13}$$

although the exact form varies between systems as we shall see shortly. The measurement y_t may be a scalar or a vector, depending on system design. Due to pulse shaping and/or a sparse impulse response, the energy contribution from the taps may be distributed over several samples. The matrix D is a mapping from the vector of taps, h, to the actual impulse response caused by the tap delays and the pulse. The structure of the regressor matrix φ_t depends on which kind of system we are looking at. We will study

several different structures below. However, in order for us to make inference from the equations, φ_t needs to be known (otherwise we have to carry out a marginalisation which will take us into the nonlinear domain). We will therefore make the restriction that the filter will only operate on *pilot symbols*. Pilot symbols are symbols known to both sender and receiver (hence they do not carry any information). It is a delicate problem to balance the ratio

$$\frac{\text{pilot symbol rate}}{\text{payload (information) symbol rate}}$$

for an optimal trade-off between system throughput and channel prediction performance.

To aid the filter, we will thus scatter pilot symbols over the radio resource. The measurement signal will be designed in such a way that it only "sees" pilot symbols.

The measurement noise is caused by thermal noise and by inference from other systems or from adjacent cells in the same system. In most cases it will be proper to assign independence between the noise contributions on the different components of the measurement vector. In these cases we let the *noise covariance matrix* R be

$$R = \sigma_n^2 \mathbf{I}, \tag{4.14}$$

with the noise power σ_n^2 being given by an estimator operating outside of the linear model. In some cases a nondiagonal noise covariance matrix may be more appropriate. The off-line estimator then needs to estimate the entire matrix. This is discussed in Section 4.6.

We now have the general tools for constructing specific system models. To summarise, we want to construct a model

$$\begin{aligned} x_{t+1} &= Fx_t + Gu_t, & \operatorname{var}(u_t|I) &= \mathrm{E}(u_t u_t^*|I) = Q \\ h_t &= Hx_t \\ y_t &= \varphi_t DHx_t + v_t = Cx_t + v_t, & \operatorname{var}(v_t|I) &= \mathrm{E}(v_t v_t^*|I) = R \end{aligned} \tag{4.15}$$

The matrices F, G, and H are built from blocks constructed according to Chapter 3. The noise covariance R was discussed above and the covariance matrix Q is set according to (4.12). It remains to determine the tap covariance matrix R_h, the pilot information φ_t, and the energy distribution matrix D. They are all system specific, and we will therefore commence to study specific systems. For later convenience, we would also like to summarise all matrices between the states x and the measurements y into one single output matrix C so that

$$y_t = C_t x_t + v_t. \tag{4.16}$$

Different systems will require the construction of matrices of different sizes and structures, as we will see presently. All matrix sizes and structures are summarised in Tables 4.1 and 4.2.

4.2 The TDMA downlink

We begin by looking at single carrier systems. Having just one carrier means that any one user utilises the whole systems bandwidth during transmission. The bandwidth allocation need not be exclusive. For example, *Code Division Multiple Access* (CDMA) is a technique where many users share the same bandwidth at the same time. In this thesis I will however assume that single carrier systems use exclusive allocation. Hence we regard time as the only resource and slot the signal into time frames. The purpose of prediction is then to provide the centralised scheduler with foundations for taking the right decisions. A time-slotted multiuser system is called a *Time Division Multiple Access* (TDMA) system. First we look at the *downlink*, which is the access link from the base station to the mobile unit. The estimator/predictor then resides on the mobile terminal, and from its perspective we have a single user system.

As always, we construct a block diagonal model:

$$\begin{aligned} x_{t+1} &= \operatorname{diag}(F_m^{(\mathrm{t})})x_t + \operatorname{diag}(G_m^{(\mathrm{t})}), \\ h_t &= \operatorname{diag}(H_m^{(\mathrm{t})})x_t, \\ y_t &= \varphi_t D \operatorname{diag}(H_m^{(\mathrm{t})})x_t + v_t, \end{aligned} \tag{4.17}$$

where we have fixed the number of taps in the model to M.

Estimators residing outside the linear model provide the terminal velocity (maximum doppler frequency) and doppler spectrum needed to construct the blocks $\{F_m^{(\mathrm{t})}, G_m^{(\mathrm{t})}, H_m^{(\mathrm{t})}\}$ (see Chapter 3).

We have also assumed that estimates of the tap variances $\{\sigma_m^2\}$ are available. Assuming that the taps fade independently, we set

$$R_h = \operatorname{diag}(\sigma_m^2, m = 1..M), \tag{4.18}$$

and determine Q from (4.12). The noise covariance matrix R is set according to Section 4.1.2.

Along with estimates of the tap powers we will also have estimates of the delays $\{\delta_m\}$ of the respective taps. We also know the pulse shape $p(t)$ used by the system. Hence we can create a mapping from the M-vector of tap

values, to the actual impulse response:

$$D = \begin{pmatrix} p(0-\delta_0) & p(0-\delta_1) & \cdots \\ p(T_s-\delta_0) & \vdots & \\ p(2T_s-\delta_0) & & \\ \vdots & & \\ p((N-1)T_s-\delta_0) & p((N-1)T_s-\delta_1) & \end{pmatrix} \tag{4.19}$$

Here, T_s is the symbol time. Apart from influencing the matrix D, the symbol time T_s also affects the tap model as described in Chapter 3. The symbol time is generally much shorter in a TDMA system than in an OFDM system. Note that the pulse $p(t)$ is time-shifted so that $p(t) = 0$ for $t < 0$. The number N must be chosen so that the total impulse response duration NT_s becomes longer than the duration of the pulse $p(t)$ plus the delay of the last tap in the model (δ_{M-1}).

We now have

$$\text{Channel impulse response} = Dh_t,$$

in which are included both transmitter and receiver filters. The impulse response should then be convoluted with the transmitted pilots to produce the received signal y_t (save the noise).

The measurement y_t does not have to be scalar; we may also choose to collect data blockwise. To be general, let us say that we collect W data points at a time. Then $N+W-1$ consecutive pilots need to be transmitted in the corresponding time slot in order for the regressor matrix φ_t to be known (see Figure 4.1). Denoting the pilot symbols by $\{s\}$, we set

$$\varphi_t = \begin{pmatrix} s_0 & s_{-1} & \cdots & & s_{2-N} & s_{1-N} \\ s_1 & s_0 & s_{-1} & \cdots & & s_{2-N} \\ s_2 & s_1 & s_0 & s_{-1} & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ s_{W-1} & & & & & \end{pmatrix}, \tag{4.20}$$

which is Toeplitz and therefore performs convolution.

The TDMA downlink model is now complete, and the output matrix in (4.16) is

$$C = \varphi_t D H, \tag{4.21}$$

with dimension $W \times KM$, where K is the tap model order, and $H = \text{diag}(H_m^t)$ in accordance with (4.17).

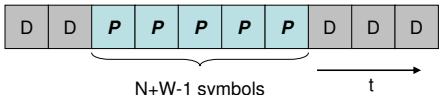


Figure 4.1: Example of pilot distribution in a single carrier system.

4.3 The TDMA uplink

Modelling the uplink – the link from the mobile terminal to the base station – may seem a harder problem than modelling the downlink. It is however easily accomplished by means of our block-diagonal structures. Extending the downlink model by one level, we get

$$\begin{aligned} F &= \mathrm{diag}(\mathrm{diag}(F^{(\mathrm{t})}_{m,u}, m = 1..M), u = 1 \dots U), \\ G &= \mathrm{diag}(\mathrm{diag}(G^{(\mathrm{t})}_{m,u}, m = 1..M), u = 1 \dots U), \\ H &= \mathrm{diag}(\mathrm{diag}(H^{(\mathrm{t})}_{m,u}, m = 1..M), u = 1 \dots U) \end{aligned} \tag{4.22}$$

The covariances Q and R are set as before, with the diagonal tap covariance

$$Rh = \mathrm{diag}(\mathrm{diag}(\sigma^2_{m,u}, m = 1..M), u = 1..U), \tag{4.23}$$

if the impulse response taps fade independently, and the energy distribution matrix is block-diagonal:

$$D = \mathrm{diag}(D_u, u = 1..U), \tag{4.24}$$

where D_u is the impulse response matrix for user u, chosen as (4.19).

It is however not clear how the pilots should be distributed. One could certainly choose to transmit pilots exclusively, so that users take turns in sending pilots. This would impede heavily on the system performance though, since only one user would send pilots at any one time, while the other users would have to be quiet.

Instead I will here suggest the use of *overlapping* pilots, meaning that all users transmit pilots at the same time (the same time-frequencies in OFDM systems). Designing the measurement equation to account for overlapping pilots is straightforward:

$$\varphi_t = \left[\varphi_{1,t}\, \varphi_{2,t}\, \dots\, \varphi_{U,t}\right], \tag{4.25}$$

which is a $W \times KNU$ matrix. Here, $\varphi_{u,t}$ is the pilot matrix for user u, designed according to (4.20). Note that the measurement equation still has

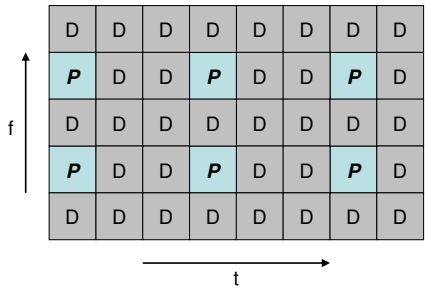


Figure 4.2: Example of pilot distribution in an OFDM system.

dimension W, since the contribution from all users add together. The total $W \times KMU$ output matrix in (4.16) for the TDMA uplink is

$$C = \varphi_t DH, \tag{4.26}$$

with φ_t, D, and H given by (4.25), (4.24), and (4.22), respectively.

4.4 The OFDM downlink

Turning to multicarrier multiuser systems, we now study the OFDM downlink. Contrary to the TDMA systems, we here consider both the time domain and the frequency domain as resources to be distributed among users. The total radio resource can be viewed as a grid, each element holding one time-frequency symbol, where time runs horisontally (divided in *OFDM symbols*), and frequency runs vertically (divided into *subcarriers*). Pilot symbols are distributed over this grid according to some predefined pattern. This is illustrated in Figure 4.2. We will make sure that the measurement equation will only take into consideration those time-frequency slots that hold pilots.

In an OFDM system, the received baseband signal is collected in blocks after which the *cyclic prefix* – which is at least as long as the channel impulse response – is cut away. The remaining part of length N is fourier transformed. As we saw in Section 3.3, this is equivalent to sampling the output of $2N$ cross correlators. It should be noted that, whereas N in the TDMA systems corresponded to the length of the channel impulse response, in the OFDM systems N is the block length. It should be set to a value corresponding to a considerably longer period than the maximum impulse response length to minimise the relative overhead caused by the cyclic prefix.

Define the $N \times N$ *fourier matrix*

$$\mathcal{F} = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & & \omega^{(N-1)(N-1)} \end{bmatrix}, \tag{4.27}$$

where $\omega = e^{-2\pi i/N}$. The inverse fourier transform is simply given by its conjugate transpose, so that

$$\mathcal{F}^{-1} = \mathcal{F}^*. \tag{4.28}$$

Now define as the *partial fourier matrix* a number W of rows from the fourier matrix. For example, we could choose

$$\mathcal{F}_W = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & \omega^{105} & \omega^{210} & \cdots & \omega^{105(N-1)} \\ 1 & \omega^{110} & \omega^{220} & \cdots & \omega^{110(N-1)} \\ \vdots & \vdots & \vdots & & \\ 1 & \omega^{100+5W} & \omega^{(100+5W)\times 2} & \cdots & \omega^{(100+5W)(N-1)} \end{bmatrix}, \tag{4.29}$$

where I have made the specific choice to start at row 105 and then pick every fifth row. In the way the partial fourier matrix will be used, this corresponds to taking the first measurement at the 105th subcarrier in a system where pilots are located at every fifth subcarrier. To better assess the properties of the partial fourier transform, note that $\mathcal{F}_W \mathcal{F}_W^* = \mathbf{I}$, and that $\mathcal{F}_W^* \mathcal{F}_W$ is the linear operation that cancels out all frequencies in an N-vector, except those frequencies represented by the rows of $\mathcal{F}_W$.

OFDM systems may be modelled in a few different ways. Since we will use near-Bayesian algorithms for conducting the inferences, system performance will be independent of this choice. The only significant reason for choosing one model over another is numerical complexity, an issue that will be addressed in the next chapter.

4.4.1 Model in time, measure in frequency

Identically to the TDMA downlink case, set

$$\begin{aligned} x_{t+1} &= Fx_t + Gu_t = \mathrm{diag}(F_m^{(\mathrm{t})})x_t + \mathrm{diag}(G_m^{(\mathrm{t})})u_t, \\ h_t &= Hx_t = \mathrm{diag}(H_m^{(\mathrm{t})})x_t, \end{aligned} \tag{4.30}$$

and set Q according to (4.12) (the tap covariance is still diagonal as given by Equation (4.18)). Note that h_t is still the same vector as in (4.17), holding *time-domain* impulse response taps.

The energy distribution matrix D is chosen as (4.19). If we assume, idealistically, that the pulses $p(t)$ have zero crossings at all integer multiples of the symbol time T_s except for its peak, and that the duration $\{\delta_m\}$ of the impulse response taps are multiple integers of the symbol time, then all the energy from any given tap in the impulse response arrive at one single sample and we may set

$$D = \begin{pmatrix} 1 & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & 1 & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & 1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}, \tag{4.31}$$

where the symbol $\cdot$ indicates a zero and the position of the ones is determined by the tap delays $\{\delta_m\}$, analogous to (4.19). Note that D is dimensioned so that it accounts for the entire block of length N. Since the actual impulse response will be much shorter than this, a large portion of the lower rows of D will all be null.

It should also be noted that it is common to assume that rectangular pulses are used in OFDM. This however refers to the pulses used in each separate subcarrier. These pulses are rectangular in the sense that the harmonic waveform transmitted over a subcarrier starts suddenly at $t = 0$ and ends suddenly at $(N + N_{CP}T_s)$, where N_{CP} is the number of samples in the cyclic prefix. This generates spectral sidelobes, but the bandwidth of each subcarrier is so small compared to the total system bandwidth that this does not matter. The "system" pulse shape $p(t)$, corresponding to a sampling rate that is $N + N_{CP}$ times higher than T_s, will however determine the total system bandwidth and should therefore be selected carefully.

In OFDM, the frequency-domain measurements y_t consist of W parallel subcarriers influenced by pilots. How to distribute the pilots is a matter for the system designer to decide. Once this is done, the partial fourier matrix $\mathcal{F}_W$ is set according to (4.29). Since we have the channel impulse response Dh_t, the time-frequency taps on these subcarriers take the value $\mathcal{F}_W Dh_t$.

Multiplying elementwise with the frequency-domain pilots

$$\varphi_t = \begin{pmatrix} s_0 & & \\ & \ddots & \\ & & s_{W-1} \end{pmatrix} \tag{4.32}$$

gives the measurement equation

$$y_t = \varphi_t \mathcal{F}_W D h_t + v_t, \tag{4.33}$$

so that the output matrix C in (4.16) for the OFDM downlink modelled in time and measured in frequency is

$$C = \varphi_t \mathcal{F}_W D H. \tag{4.34}$$

An important special case

All matrices involved in the ongoing discussion are time variant. The pilot matrix φ_t is however varying on a much faster time scale than the rest of the matrices; whereas the latter change when the fading statistics or the velocity changes – which happens relatively seldom compared to the sampling rate of the measurements – φ_t will generally change at every sample (this is indicated by the subscript t in (4.32)). Keeping φ_t constant for long periods would cause spectral spikes, which would make it difficult for amplifiers to process the signal without introducing distortion.

Still, a time invariant model (over a period spanning several measurements) is to prefer for complexity reasons, as we shall see in the forthcoming chapter. It is therefore interesting to note that we may take advantage of the fact that φ_t is quadratic. If we let all pilots have modulus one, for example we could let them be QPSK symbols, then $\varphi_t^{-1} = \varphi_t^*$. If we now multiply the measurement equation with φ_t^* from the left, we get

$$\varphi_t^* y_t = \mathcal{F}_W D H x_t + v_t \tag{4.35}$$

Thus we may replace the measurement y_t with the "adjusted measurement" $\varphi_t^* y_t$. The output matrix then becomes (nearly) time invariant:

$$C = \mathcal{F}_W D H \tag{4.36}$$

We also need to adjust he noise covariance matrix:

$$R \mapsto \varphi_t^* R \varphi_t \tag{4.37}$$

4.4.2 Model in frequency, measure in frequency

Instead of modelling the M taps in the channel impulse response, we may directly model the W time-frequency taps that we are actually interested in:

$$\begin{aligned} x_{t+1} &= \mathrm{diag}(F_w^{(\mathrm{t})}, w = 1..W)x_t + \mathrm{diag}(G_w^{(\mathrm{t})}, w = 1..W)u_t, \\ h_t &= \mathrm{diag}(H_w^{(\mathrm{t})}, w = 1..W)x_t, \end{aligned} \tag{4.38}$$

where h_t now represents W taps at the OFDM symbol with time index t. The individual taps are still modelled as described in Chapter 3, but we have to reconsider the value of the tap covariance matrix R_h. Since the mapping from the time taps to the time-frequency taps is given by $\mathcal{F}_W D$, we have

$$R_h = \mathcal{F}_W D \times \mathrm{diag}(\sigma_m^2, m = 1..M) \times D^* \mathcal{F}_W^* \tag{4.39}$$

The process noise covariance Q is then calculated from (4.12) and (4.39).

The measurement equation is straightforward since we now have the time-frequency taps straight from the model:

$$y_t = \varphi_t h_t + v_t = \varphi_t H x_t + v_t \tag{4.40}$$

Evidently, the output matrix in (4.16) for the OFDM downlink modelled and measured in frequency is

$$C = \varphi_t H, \tag{4.41}$$

which has dimensions $W \times KW$, where K is the tap model order.

As in the previous section, we may use the fact that φ_t is square to make the output matrix time-invariant. We then have

$$C = H. \tag{4.42}$$

4.4.3 Model in time, measure in time

If the pilots are distributed in such a way that *every* time-frequency symbol in a particular OFDM symbol is a pilot, then that entire OFDM symbol is known to the receiver. We may then omit the fourier transform, and instead regard the received block as a known time-domain signal of N samples, and then model the channel as we did for the TDMA downlink. We denote this known time domain signal

$$[s_0 \cdots s_{N-1}] \tag{4.43}$$

(note that $\{s\}$ are the *transformed* time-frequency pilots) and construct the pilot matrix φ_t:

$$\varphi_t = \begin{pmatrix} s_0 & s_{N-1} & \cdots & s_{N-N_{CP}} & \text{payload symbols} & \\ s_1 & s_0 & s_{N-1} & \cdots & s_{N-N_{CP}} & \text{payload symbols} \\ \vdots & \ddots & \ddots & \ddots & & \\ s_{N-1} & s_{N-2} & & & \cdots & s_0 \end{pmatrix}, \tag{4.44}$$

where N_{CP} is the number of samples in the cyclic prefix. It may seem troublesome that unknown payload data enter the pilot matrix. This is however not a problem since there is at most N_{CP} non-zero elements in the impulse response, so that the last $N - N_{CP}$ rows of the energy distribution matrix D are null. The payload symbols will therefore not influence the measurements.

Hence we may either cut away the last $N - N_{CP}$ columns of φ_t and the last $N - N_{CP}$ rows of D, or make it simple and set

$$\varphi_t = \text{Circulant}([s_0,\, s_{N-1},\, \ldots,\, s_1]), \tag{4.45}$$

where Circulant(x) is a circulant square matrix whose first row is x. I will assume the latter, to keep consistency in the matrix dimensions, although it will introduce some redundant numerical overhead.

In (4.44) I have chosen to measure the whole OFDM symbol as one block. This means that we keep the sampling frequency of the original OFDM system (sample once per OFDM symbol), but that the dimensionality is huge (N). This will inevitably generate a huge numerical complexity. But just as in the TDMA downlink, we are perfectly free to increase the sampling frequency and at the same time decrease the block size. At the extreme end, we would have a scalar measurement sampled at the full system sampling rate. Such an approach may very well reduce the total numerical complexity of the inferential process considerably. Numerical complexity is studied in Chapter 5.

The state space model and the measurement equation are now constructed exactly as for the TDMA downlink. Again, we have

$$C = \varphi_t D H. \tag{4.46}$$

Sparse pilots

In the above it was assumed that the measured signal during an interval of one OFDM symbol's length consists of only known pilots. In many system

one would prefer to interleave pilots with payload data, so that, say, every fourth subcarrier is a pilot. Seemingly, one would then have to perform the following steps in order to remove the (unknown) influence from the payload symbols:

- Transform the received time domain signal to the frequency domain (FFT).
- Null out all symbols but the pilot symbols.
- Transform back to the time domain (IFFT).

Fourier transforms are expensive, so this may discourage anyone from using OFDM time modelling and measuring. A closer examination will however reveal that the above steps are easier to carry out than one might expect.

Remember that the partial fourier matrix (4.29) may be used to filter out certain frequencies by forming the product $\mathcal{F}_W^*\mathcal{F}_W$. Now we construct a particular kind of a partial fourier matrix, where the spacing between the frequencies are N/W, so that its row covers all the N frequency bands. I denote this matrix by $\bar{\mathcal{F}_W}$. Examining the product $\bar{\mathcal{F}}_W^*\bar{\mathcal{F}}_W$, we find

$$\bar{\mathcal{F}}_W^*\bar{\mathcal{F}}_W = \frac{W}{N}\begin{pmatrix} \overbrace{\begin{matrix} 1 & \cdot & \cdots & \cdot \\ \cdot & 1 & \ddots & \cdot \\ \vdots & \ddots & \ddots & \cdot \\ \cdot & \cdot & \cdot & 1 \end{matrix}}^{W} & \begin{matrix} 1 & \cdot & \cdots & \cdot \\ \cdot & 1 & \ddots & \cdot \\ \vdots & \ddots & \ddots & \cdot \\ \cdot & \cdot & \cdot & 1 \end{matrix} & \cdots \\ \begin{matrix} 1 & \cdot & \cdots & \cdot \\ \cdot & 1 & \ddots & \cdot \\ \vdots & \ddots & \ddots & \cdot \\ \cdot & \cdot & \cdot & 1 \end{matrix} & \begin{matrix} 1 & \cdot & \cdots & \cdot \\ \cdot & 1 & \ddots & \cdot \\ \vdots & \ddots & \ddots & \cdot \\ \cdot & \cdot & \cdot & 1 \end{matrix} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}, \tag{4.47}$$

which I formalise with the following theorem:

Theorem 1 *$\bar{\mathcal{F}}^*\bar{\mathcal{F}}$ is a band matrix where the bands are separated by W elements and the band elements all have the value W/N.*

Proof. Let $\bar{\mathcal{F}_W}$ be the partial fourier matrix with spacing N/W between each row, so that its elements are given by

$$\bar{\mathcal{F}_W}[m,n] = \omega^{mnN/W}/\sqrt{N}, \quad 0 \le m < W, 0 \le n < N, \tag{4.48}$$

and $\omega = \exp(-2\pi i/N)$. Note that

$$\sum_{w=0}^{W-1} \omega^{wcN/W} = \sum_{w=0}^{W-1} \exp(-2\pi iwc/W) = \frac{1-\exp(-2\pi ic)}{1-\exp(-2\pi ic/W)} = 0, \quad c \in \mathbb{Z} \text{ but } c/W \notin \mathbb{Z} \tag{4.49}$$

If $c/W \in \mathbb{Z}$, then

$$\sum_{w=0}^{W-1} \omega^{wcN/W} = W \tag{4.50}$$

The element (m,n) of $\bar{\mathcal{F}}^*\bar{\mathcal{F}}$ now evaluates to

$$(\mathcal{F}^*\mathcal{F})[m,n] = \frac{1}{N}\sum_{w=0}^{W-1} \omega^{wmN/N - wnN/W} = \sum_{w=0}^{W-1} \omega^{w(m-n)N/W} = \begin{cases} W/N & \text{if } (m-n)/W \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \tag{4.51}$$

What does this mean? It means that in a system in which the pilot frequency spacing is N/W, all we have to do to rid the received N-block y_t of the influence from payload, is to form the product

$$\mathcal{F}_W^*\mathcal{F}_W y_t, \tag{4.52}$$

and this simply amounts to form a new time-domain signal where each element is a sum of every W:th element in the originally received block. Since we then remove every influence from the payload data, we may use this new signal as measurement signal.

4.5 The OFDM uplink

Extending the OFDM downlink model to a representation of the uplink, where the channels from U users are simultaneously tracked, is done by following the same convention as we did for the TDMA systems. By adding a level, we get

$$\begin{aligned} F &= \text{diag}(F_u^{(\text{u})}, u = 1 \dots U), \\ G &= \text{diag}(G_u^{(\text{u})}, u = 1 \dots U), \\ H &= \text{diag}(H_u^{(\text{u})}, u = 1 \dots U), \end{aligned} \tag{4.53}$$

where the models $F_u^{(u)}$, $G_u^{(u)}$, and $H_u^{(u)}$ for the individual users are given by Section 4.4.

The partial fourier matrix, when needed, is constructed in the same straightforward way:

$$\mathcal{F}_W = \mathbf{I}_U \otimes \mathcal{F}_{W,\text{signle user}}, \tag{4.54}$$

where $\otimes$ here represents the Kronecker product. This special case can be used instead of $\text{diag}(\cdots)$, since there is no reason for the individual blocks to differ. Note that the dimension of $\mathcal{F}_W$ is now $WU \times NU$. The subscript W thus indicates the property of each block, not the dimension of the matrix.

The covariances Q and R are set as before, with the diagonal tap covariance

$$R_h = \text{diag}(R_{h,u}, u = 1..U), \tag{4.55}$$

and the impulse response matrix is block-diagonal:

$$D = \text{diag}(D_u, u = 1..U), \tag{4.56}$$

where $R_{h,u}$ and D_u are the respective matrices for user u.

The question of how to distribute the pilots for the different users is more or less open-ended. We could allocate exclusive time-frequency locations for each user, in which only one user transmits symbols and everyone else is quiet. As previously noted, this scheme is however very costly in the respect that a lot of resources are spent on pilots. Nevertheless, if we choose to take that approach, each user will in effect experience a one-user system which means that the system representation for the uplink will be the same as that for the downlink.

Another approach is to let all users transmit pilots at the *same* time-frequency locations (overlapping pilots). The pilot matrix is then constructed from stacking the pilot matrices from the individual users:

$$\varphi_t = \left[\varphi_{1,t}\, \varphi_{2,t}\, \cdots\, \varphi_{U,t}\right], \tag{4.57}$$

As before, the output matrix C is equal to $\varphi_t \mathcal{F}_W DH$ if we model in time and measure in frequency, $\varphi_t H$ if we model in frequency and measure in frequency, or $\varphi_t DH$ if we model in time and measure in time.

4.6 Parameter estimation

I close this chapter with briefly discussing how to produce the "external" parameter estimates needed by the linear model.

4.6.1 Frequency offset

Throughout this thesis I assume that the time and frequency synchronisation is perfect. While timing is more or less straightforward to recover based on pilot sequences, frequency synchronisation requires more effort. A Bayesian approach to frequency synchronisation can be found in [15].

4.6.2 The mobile unit velocity

Since there is a direct relationship between the velocity and the maximum doppler frequency, velocity estimation amounts to estimating the doppler spread/doppler shift of a pilot sequence. This is most simply accomplished if the pilot sequence is a steady tone. The estimation algorithm can be primitive; a simple FFT-based estimator will probably suffice to produce a very accurate estimate. A component of the frequency shift may be due to drifting oscillators, but the "virtual velocity" induced by this phenomenon should be considered just as seriously as the real velocity. There is no need to discriminate between the two.

4.6.3 The noise power and the noise covariance matrix

Assuming additive white gaussian noise, the noise power σ_n^2 is most easily estimated by measuring the channel during a short period of radio silence. Such periods will naturally impede on the total system performance, but would on the other hand appear rarely, since measurement noise statistics is expected to vary slowly.

Taking N measurements $D = \{y_t\}$ over a silent period, we would then have

$$\mathrm{E}(\sigma_n^2|DI) = \int_0^\infty \sigma_n^2 \mathrm{P}(\sigma_n|I)\frac{\mathrm{P}(D|\sigma_n I)}{\mathrm{P}(D|I)}\mathrm{d}\sigma_n = \frac{\int_0^\infty \sigma_n^2\sigma_n^{-1}(2\pi\sigma_n^2)^{-N/2}\exp(-N\overline{y^2}/2\sigma_n^2)\mathrm{d}\sigma_n}{\int_0^\infty \sigma_n^{-1}(2\pi\sigma_n^2)^{-N/2}\exp(-N\overline{y^2}/2\sigma_n^2)\mathrm{d}\sigma_n} = \frac{N\overline{y^2}\Gamma\left(\frac{N}{2}-1\right)}{2\Gamma\left(\frac{N}{2}\right)}, \quad (4.58)$$

where I have used Jeffrey's prior and assigned $N\overline{y^2} = \sum y_t^2$. Also, $N > 2$ is needed for the integral to converge. If $N/2 \in \mathbb{N}$ we have

$$\mathrm{E}(\sigma^2|DI) = \frac{N\overline{y^2}(N/2-2)!}{2(N/2-1)!} = \frac{N}{N-2}\overline{y^2} \approx \overline{y^2}, \quad (4.59)$$

which should be quite expected.

As noted in Section 4.1.2, there may be circumstances under which the entire noise covariance matrix will have to be estimated. This is not all that trivial, especially since one then needs to consider the assignment of a prior distribution on the set of all non-negative hermitian matrices. For a comprehensive treatise on this subject, see [16].

We may however note that we here require the estimate to be extremely accurate. This means that the prior does not play a crucial part in the estimation process; the data needs to produce a sharp likelihood, which renders any uninformative prior irrelevant. We therefore seek the maximum-likelihood of the covariance matrix in a zero-mean, generally multidimensional gaussian distribution. It can be shown (see e.g. [17]) that this estimate is given by

$$\frac{1}{N}\sum_{n=1}^{N} y_n y_n^*, \tag{4.60}$$

where N is the number of (vector-valued) samples y_n taken over a silent period.

4.6.4 The power delay profile

The power delay profile is easily and near-optimally estimated from a known transmitted sequence; correlating the received signal with the pilot sequence produces an estimate of the channel impulse response. Repeating the measurement several times over a short time period will naturally improve the estimate.

	R	φ_t	$\mathcal{F}_W$	D	H	F	G	Q/R_h
TDMA downlink	$W\times W$	$W\times N$	-	$N\times M$	$M\times KM$	$KM\times KM$	$KM\times M$	$M\times M$
OFDM downlink, t-model, t-meas.	$W\times W$	$W\times N$	-	$N\times M$	$M\times KM$	$KM\times KM$	$KM\times M$	$M\times M$
OFDM downlink, t-model, f-meas.	$W\times W$	$W\times W$	$W\times N$	$N\times M$	$M\times KM$	$KM\times KM$	$KM\times M$	$M\times M$
OFDM downlink, f-model, f-meas.	$W\times W$	$W\times W$	$(W\times N)$	$(N\times M)$	$W\times KW$	$KW\times KW$	$KW\times W$	$W\times W$
TDMA uplink	$W\times W$	$W\times NU$	-	$NU\times MU$	$MU\times KMU$	$KMU\times KMU$	$KMU\times MU$	$MU\times MU$
OFDM uplink, t-model, t-meas.	$W\times W$	$W\times NU$	-	$NU\times MU$	$MU\times KMU$	$KMU\times KMU$	$KMU\times MU$	$MU\times MU$
OFDM uplink, t-model, f-meas.	$W\times W$	$W\times WU$	$WU\times NU$	$NU\times MU$	$MU\times KMU$	$KMU\times KMU$	$KMU\times MU$	$MU\times MU$
OFDM uplink, f-model, f-meas.	$W\times W$	$W\times WU$	$(WU\times NU)$	$(NU\times MU)$	$WU\times KWU$	$KWU\times KWU$	$KWU\times WU$	$WU\times WU$

Table 4.1: Summary of matrix dimensions used in different systems. The matrices are listed according to causality, with matrices associated with the process noise to the right, and matrices associated with the measurements to the left. Adjacent matrices therefore have matching dimensions. Parenthesised matrices are associated with the process noise and should therefore be thought of as being located to the far right.

	R	φ_t	$\mathcal{F}_W$	D	H	F	G	Q/R_h
TDMA downlink	diag.	full (Toeplitz)	-	nearly full or sparse	bl.diag.	diag.	bl.diag.	diag.
OFDM downlink, t-model, t-meas.	"	"	-	"	"	"	"	"
OFDM downlink, t-model, f-meas.	"	diag.	full	"	"	"	"	"
OFDM downlink, f-model, f-meas.	"	"	"	"	"	"	"	full (hermit.)
TDMA uplink	"	full	-	bl.diag.	"	"	"	diag.
OFDM uplink, t-model, t-meas.	"	"	-	"	"	"	"	"
OFDM uplink, t-model, f-meas.	"	sparse	bl.diag.	"	"	"	"	"
OFDM uplink, f-model, f-meas.	"	"	"	"	"	"	"	full (hermit.)

Table 4.2: Summary of matrix structures used in different systems.

Chapter 5

Inference

Thus far we have modelled all fading taps in a multiuser environment as well as measurement of these taps through pilot symbols:

$$\begin{aligned} x_{t+1} &= Fx_t + Gu_t, \quad \text{var}(u_t|I) = \text{E}(u_t u_t^*|I) = Q \\ h_t &= Hx_t \\ y_t &= Cx_t + v_t, \qquad \text{var}(v_t|I) = \text{E}(v_t v_t^*|I) = R \end{aligned} \tag{5.1}$$

The focus of our interest here are the channel taps h_t, which we want to infer with aid of the measurements y_t. For the estimation problem, we want to calculate $p(h_t|y_t, y_{t-1}, \ldots, I)$. For the prediction problem, we are interested in calculating $p(h_{t+L}|y_t, y_{t-1}, \ldots, I)$, where L is the prediction horizon in the current system.

As explained in Chapter 4, the dimensionality of h_t, that is the number of taps that we model, depends on which kind of system we are looking at. In TDMA systems for U users we model M (time-domain) taps for each user, which means that the vector h_t has length MU. The time-domain taps, constituting the impulse response, are also what we are interested in in TDMA systems, because the bit error rate will here be determined by the performance of the channel equaliser, which in turn will be decided by the properties of the impulse response and the signal-to-noise ratio.

In OFDM systems we may choose to model the M time-domain taps in the impulse response model (producing a vector h_t of length MU), or to model W parallel time-frequency taps in the competition band that we are currently looking at (h_t then has WU elements). Bear in mind, though, that it is always the time-frequency taps that are ultimately determining the performance in an OFDM system; each subchannel is subjected to flat fading

only (meaning that the radio channel for that subchannel is characterised by one tap only), so equalisation is trivially carried out by "derotating" the received time-frequency symbol with the estimated time-frequency tap.

Let us now finally turn to the question of how to actually carry out the inference.

5.1 Kalman's great discovery

Let us say that we have constructed a linear model of a generally vector-valued process x_t that we here shall be calling the *states*:

$$x_{t+1} = F_t x_t + G_t u_t, \quad p(u_t|I) = \mathcal{CN}(u_t; 0, Q_t), \tag{5.2}$$

where we call the white noise u the *process noise*. u_t is the "driving force" behind the changes of $\{x_t\}$.

Further, the states $\{x_t\}$ are measured in AWGN and we model the measurements by

$$y_t = H_t x_t + v_t, \quad p(v_t|I) = \mathcal{CN}(v_t; 0, R_t) \tag{5.3}$$

Following [18], we denote all measurements up to time t by

$$Y_t \triangleq y_t, y_{t-1}, y_{t-2}, \ldots \tag{5.4}$$

Let us say that at some point t we know that $p(x_t|Y_t I) = \mathcal{CN}(x_t; \hat{x}_{t|t}, P_{t|t})$. To calculate what we know about the next x when we get a new measurement, we apply Bayes' theorem and marginalise over nuisance parameters (as always). Conditioning everything on $Y_t I$, we get

$$p(x_{t+1}|Y_{t+1}I) = p(x_{t+1}|y_{t+1}Y_t I) = p(x_{t+1}|Y_t I)\frac{p(y_{t+1}|x_{t+1}Y_t I)}{p(y_{t+1}|Y_t I)} = \frac{\int p(y_{t+1}|x_{t+1}I)p(x_{t+1}|x_t I)p(x_t|Y_t I)dx_t}{\iint p(y_{t+1}|x_{t+1}I)p(x_{t+1}|x_t I)p(x_t|Y_t I)dx_t dx_{t+1}} \tag{5.5}$$

and from the model (5.2) and (5.3) we see, after some consideration, that

$$\begin{aligned} p(y_{t+1}|x_{t+1}I) &= \mathcal{CN}(y_{t+1}; H_{t+1}x_{t+1}, R_{t+1}) \\ p(x_{t+1}|x_t I) &= \mathcal{CN}(x_{t+1}; F_t x_t, G_t Q_t G_t^*) \end{aligned} \tag{5.6}$$

and we knew already that

$$p(x_t|Y_t I) = \mathcal{CN}(x_t; \hat{x}_{t|t}, P_{t|t}) \tag{5.7}$$

But the gaussian function $\mathcal{CN}(\cdot;\cdot,\cdot)$ has remarkable properties. Firstly, the product of two (generally multidimensional) gaussians is a gaussian. Secondly, integrating a multidimensional gaussian over one or more of its free variables also yields a gaussian.

These properties makes it possible to solve Equation (5.5) analytically. Since we now know that the left side of the equation (the posterior) is also gaussian, we may directly express its two (generally multidimensional) degrees of freedom – the mean $\hat{x}_{t+1|t+1}$ and the covariance matrix $P_{t+1|t+1}$ – in terms of the right hand side parameters. The expressions turn out to be somewhat involved so using a few intermediate expressions simplifies matters. An exercise in algebra reveals (see e.g. [18], [19])

$$\begin{aligned}
\hat{x}_{t+1|t} &= F_t \hat{x}_{t|t} \\
P_{t+1|t} &= F_t P_{t|t} P_t^* + G_t Q_t G_t^* \\
R_{e,t} &= R_t + H_t P_{t|t-1} H_t^* \\
K_{f,t} &= P_{t|t-1} H_t^* R_{e,t}^{-1} \\
\hat{x}_{t+1|t+1} &= \hat{x}_{t+1|t} + K_{f,t+1}(y_{t+1} - H_{t+1}\hat{x}_{t+1|t}) \\
P_{t+1|t+1} &= (I - K_{f,t+1} H_{t+1}) P_{t+1|t}
\end{aligned} \tag{5.8}$$

The "intermediate" matrix R_e is the covariance matrix for the measurement at time t, given the measurements up to time $t-1$. The matrix K_f is called the *Kalman gain*. It expresses to which extent one should take into account new data, and to which extent to extrapolate older estimates.

This is what Kalman discovered in his seminal 1960 paper, that when the underlying model is linear and all parameters are gaussian (given the information at hand), then there are closed form expressions for Bayes propagation. It is interesting to note that the most elegant formulation of the pdf $p(x_{t+1}|Y_{t+1}I)$ involves calculating the pdf $p(x_{t+1}|Y_t I)$ with mean $\hat{x}_{t+1|t}$ and variance $P_{t+1|t}$ as intermediate steps. In fact, the one-step prediction $p(x_{t+1}|Y_t I)$ is so common in linear filter theory that we henceforth will use the short-hand notations

$$\begin{aligned}
\hat{x}_{t+1} &\triangleq \hat{x}_{t+1|t} \\
P_{t+1} &\triangleq P_{t+1|t}
\end{aligned} \tag{5.9}$$

Rearranging the order of (5.8), we produce the equations for propagating

the one-step predictions:

$$\begin{aligned} R_{e,t} &= R_t + H_t P_t H_t^* \\ K_{f,t} &= P_t H_t^* R_{e,t}^{-1} \\ \hat{x}_{t|t} &= \hat{x}_t + K_{f,t}(y_t - H_t \hat{x}_t) \\ P_{t|t} &= (I - K_{f,t} H_t) P_t \\ \hat{x}_{t+1} &= F_t \hat{x}_{t|t} \\ P_{t+1} &= F_t P_{t|t} F_t^* + G_t Q_t G_t^* \end{aligned} \tag{5.10}$$

By repeated application of the last two equations we get a recursive formula for the pdf of the many-steps prediction $p(x_{t+L}|Y_t I)$:

$$\begin{aligned} \hat{x}_{t+L|t} &= F_{t+L-1} \hat{x}_{t+L-1|t} \\ P_{t+L|t} &= F_{t+L-1} P_{t+L-1|t} F_{t+L-1}^* + G_{t+L-1} Q_{t+L-1} G_{t+L-1}^* \end{aligned} \tag{5.11}$$

Explicitly, this means the the MMSE prediction $\hat{x}_{t+L|t}$ is calculated through

$$\hat{x}_{t+L|t} = F_{t+L-1} F_{t+L-2} \dots F_t \hat{x}_{t|t} \tag{5.12}$$

and simply

$$\hat{x}_{t+L|t} = F^L \hat{x}_{t|t} \tag{5.13}$$

if the model is static. The error covariance $P_{t+L|t}$ however has to be updated iteratively.

5.2 Channel estimation and prediction

5.2.1 Estimation

The quest for optimal *channel* estimation and prediction has now come to and end. From the mean value (expectancy) and covariance matrix of the states x, as given by the Kalman recursions, we can now calculate the full joint pdf of the taps h:

$$p(h|YI) = \mathcal{CN}(h; H\hat{x}, HPH^*) \tag{5.14}$$

I have deliberately left out all subscripts; if we are interested in, say, $h_{t+10|t}$, then we apply (5.10) and (5.11) accordingly to produce $\hat{x}_{t+10|t}$ and $P_{t+10|t}$. These are then inserted into (5.14) which yields the pdf that we sought for.

Equation (5.14) illuminates a useful property of linear models: Once we have the pdf of the states, which is fully described by ($\hat{x}$,P), changing

variables to another variable $z = Ax$ simply yields the pdf $p(z|YI) = \mathcal{CN}(z; A\hat{x}, APA^*)$.

As far as channel *estimation* is concerned, where the objective is to recover payload data as well as possible, one should always use as much data as possible. This means that the *filter estimate* $h_{t|t}$ should be used. If one can accept a certain delay in the bit stream recovery, one can also use *smoothing*, where all data up to time t is used to estimate older taps (for example $h_{t-4|t}$). The recursions needed to produce smoothing estimates are not presented in this thesis. The interested reader is recommended [19].

We may now summarise the procedure for inferring the transmitted bit stream:

- Use the Kalman recursions and the measurements to infer the states x of time-frequency positions or time instants containing payload data.
- Calculate the tap estimates $\hat{h} = H\hat{x}$.
- Equalise the channel based on the tap estimates.
- Detect the bits. The decisions are usually based on decision regions.

This is admittedly somewhat simplified; separating equalisation and detection is only optimal if the tap estimates are highly accurate (of low uncertainty), and if the received symbol sequence does not constitute an entire supersymbol (so that it corresponds to only a fraction of a codeword), then one may have to do joint equalisation and detecting to get a good performance even if the tap estimates are of high quality.

Equalisation is much easier in an OFDM system than in a TDMA system. In the former, all we have to do is "derotate" the received symbol on a subcarrier with the estimated channel tap for that particular subcarrier. In the latter, we will have to regard *all* taps in the impulse response.

I should also make clear the following. The measurement equation as constructed in Chapter 4 will only observe those symbols that are known pilot symbols. Naturally, these are not the symbols that we want to infer, since they are already known. Instead it is the payload symbols located in between the pilots whose value we seek. It is certainly possible to use the Kalman formulation to produce optimal estimates of these from the states estimates of the pilot locations, but a simple interpolation of the tap values at the pilot locations works almost as well if the channel variability between the pilots positions is small.

For example, let us say that we have an OFDM system with pilots on every fifth subcarrier and every second OFDM symbol. If we then look at a "frame"

of 20 subcarriers by ten OFDM symbols, the measurement equation will be constructed so that it samples the frame at five distinctive times, each time taking four parallel subcarriers into account. The $200 - 20 = 180$ payload symbols in the frame are then detected by interpolation of the 5×4 taps at the pilot locations.

5.2.2 Prediction

What, then, about channel prediction? As previously concluded, the Kalman recursions are sufficiently general that (5.14) can be used for prediction just as well as for estimation.

But the prediction problem doesn't end here, because the relationship between predicted bit error rate and predicted channel quality is intricate. In the channel estimation case, one can assume that the estimation accuracy is high (by using smoothing if necessary). The values of estimated taps then bear a direct correspondence to the success of the bit detection. When channel prediction is considered, it is necessary to take into consideration both the estimated values and their accuracy.

We shall study the changing of variables from the channel taps to the bit error rate presently. However, research in the field of channel prediction often stops at this point – as indeed its name indicates – and I shall acknowledge this convention and pause for a few results.

5.2.3 Simulation versus analysis

Needless to say, it is highly important to assess the performance of the predictor algorithm. Prior to setting up the state space model, one has to consider the problem of how actual performance results are to be produced. In academic research we rarely have a real system at our disposal, into which we can implement the channel predictor and then take it for a test drive in a real fading environment. The road ahead forks off into two distinctive directions: simulation and analysis.

Simulation is a very powerful technique in that it allows any algorithm to be evaluated. By that I mean that the algorithm to be scrutinised need not be subjected to any kind of adaptation to fit the evaluation process, but may be processed "as is". It is generally relatively straightforward to carry out a simulation; the algorithm is constructed just the way it would be in the real application, and it is then run on artificial data.

A simulation does not allow any element of uncertainty. What one in effect

calculates by running a simulation is

$$p(\text{'performance measure'}|\text{tap values}, \text{noise samples}, \ldots, I), \tag{5.15}$$

which always yields an absolute certain result (that is, we may see it as that it outputs a pdf in the form of a Dirac distribution) since all parameters (including data and pseudo-random noise sequences) are known. A simulation takes a "snapshot" in the space of all possible outcomes of an experiment. It may be difficult to know whether that snapshot in some sense is representative of the overall behaviour of the algorithm. In any case, it will not tell *how* representative it is.

As a contrast to this, one could attempt to undertake a full Bayesian analysis of the algorithm. Such an analysis would establish exactly how any prior uncertainty about governing parameters would propagate through every step of the algorithm. We are free to choose which parameters are uncertain and which are given, so that we could calculate

$$p(\text{'performance measure'}|\text{tap values}, I), \text{ or } p(\text{'performance measure'}|I), \tag{5.16}$$

or put any combination of parameters on the right side of the conditioning bar (which means that a simulation in a sense is a "degenerate" case of a Bayesian analysis). The drawback is that a full analysis is hardly ever possible to carry out for most algorithms. The mathematics simply becomes too involved. When it is possible to conduct, though, it will give a complete description of the situation.

'Precision' is the keyword here. To fully describe the details about the result of a simulation, one has to account for the precise value of every tap value and every noise sample used throughout the simulation. On the other hand, a Bayesian analysis conditioned on only, say, the channel power delay profiles, fading statistics, velocities, and signal-to-noise ratios, can be described precisely by presenting only those values.

It so happens that the algorithm considered here allows itself to be analysed for unknown noise and unknown tap values. If we set up a model according to Chapters 3 and 4, and assume that conditions do not change over the course of many filter samples, then the state error covariance P_t will be iterated by

$$P_{t+1} = FP_tF^* + GQG^* - FP_tH^*(HP_tH^* + R)^{-1}HP_tF^*, \tag{5.17}$$

as seen by studying (5.10). The sequence of covariance matrices $\{P_t\}$ will eventually – usually quickly – settle down on a steady value $\bar{P}$ (the *Discrete Algebraic Riccati Equation* (DARE)). The same is true for the state covariance, here denoted Π_t, which settles down to the solution of the *Lyapunov*

equation

$$\bar{\Pi} = F\bar{\Pi}F^* + GQG^*. \tag{5.18}$$

But, as we saw at the end of Chapter 2, since the model is static, we are now injecting absolute certainty into it by claiming exact knowledge about the *frequency* distributions of the process noise u_t and the measurement noise v_t. This certainty propagates through the model so that we know with absolute certainty all frequency distributions in the model. So, for example, will we know that the frequency distributions of the states x and the errors $\tilde{x} = x - \hat{x} - \hat{x}$ being the predictions – are $\bar{\Pi}$ and $\bar{P}$ respectively. From these we may easily calculate the frequency distribution of the taps h, the frequency distribution of the prediction errors, $\tilde{h}$, and the frequency distributions of the predictions, $\hat{h}$:

$$\begin{aligned} fr(h) &= \mathcal{CN}(h; 0, H\bar{\Pi}H^*), \\ fr(\tilde{h}) &= \mathcal{CN}(h; 0, H\bar{P}H^*), \\ fr(\hat{h}) &= \mathcal{CN}(h; 0, H\bar{\Sigma}H^*), \end{aligned} \tag{5.19}$$

where $\bar{\Sigma} = \bar{\Pi} - \bar{P}$ is the covariance for the predicted states.

Note that I am no longer talking about probability densities; the symbol $fr(\cdot)$ is to be regarded as a physical entity, just like mass or length, whose value we may assess with more or less certainty. Here, no uncertainty exists, and just as we by

$$p(d|DI) = \delta[d, 123] \tag{5.20}$$

would express that some data D and some piece of information I give us reason to be absolutely sure that a distance d is 123 units of length, so could we here write

$$p(fr(\tilde{h})|I) = \delta[fr(\tilde{h}), \mathcal{CN}(h; 0, H\bar{P}H^*)], \tag{5.21}$$

and so on, to express the fact that we are certain about frequency distributions. This notation however has a look of uncalled-for complexity, and I will refrain from using it.

Note also that the statement about the frequency distribution of $\tilde{h}$ was made without saying anything about whether data (taps, noise) were available or not. It is a property of the Kalman filter that the error covariances are independent of the measurements. This is quite amazing, because it allows us to calculate the frequency distribution, and hence the prediction performance, without the need of marginalising over large data sets.

Hence the scheme for producing channel prediction results look like this:

- Set up the model, given pilot patterns, power delay profiles, fading statistics, velocities, and noise power.
- Calculate the stabilising solution to the DARE (5.17).
- Calculate the variances of the prediction errors of the respective taps. These now appear on the diagonal of $H\bar{P}H^*$. They are usually normalised with their respective tap variances $\{\sigma_m^2\}$ (which are known).

What the result then says is exactly this: Over the course of a long time, these are the error variances (or the normalised error variances) that one will experience, given that the parameter values were correct.

The procedure will allow us to perceive details in the results, which if they were given by a simulation, we would not now whether they were caused by the particular data series used.

5.3 Case study: The WINNER system

Although up to this point I have made considerable efforts to treat both single carrier and multicarrier systems, I will in the examples below look at OFDM systems exclusively. The reason is that frequency-adaptive transmission in OFDMA (*Orthogonal Frequency Division Multiple Access*) downlinks is of high interest in the research community, for example in the ongoing 3GPP long-term evolution (LTE) standardization effort [20], for WiMAX and in the European beyond-3G WINNER project [21].

Recall that we model a multiuser OFDM system by

$$\begin{aligned} x_{t+1} &= Fx_t + Gu_t, \quad \mathrm{var}(u_t|I) = \mathrm{E}(u_t u_t^*|I) = Q \\ h_t &= Hx_t \\ y_t &= Cx_t + v_t, \qquad \mathrm{var}(v_t|I) = \mathrm{E}(v_t v_t^*|I) = R, \end{aligned} \tag{5.22}$$

with block-diagonal model matrices F, G, and H as described in Chapter 4. Choosing to model in frequency and measure in frequency according to Section 4.4.2, we set the output matrix $C = \varphi_t H$, with

$$\varphi_t = \mathrm{stack}_U[\varphi_{u,t}] = [\varphi_{1,t}\, \varphi_{2,t}\, \ldots\, \varphi_{U,t}]\,, \tag{5.23}$$

where stack$[\cdot]$ is the horizontal stacking operator and $\varphi_{u,t}$ is a diagonal matrix with the (generally time-varying) pilots for user u along its diagonal.

The prediction performance will be evaluated with respect to the baseline system design of the WINNER FDD mode [22]. This design has a system sampling period of 12.5 ns, giving a FFT bandwidth of 80 MHz. The signal

bands are 45 MHz in both uplinks and downlinks. Each OFDM symbol is 2048 samples plus an additional 256 samples for the cyclic prefix. The subcarrier width is 39.06 kHz and the OFDM symbol + guard duration is 28.8 μs. The centre (carrier) frequency is 3.7 GHz.

The time-frequency radio resource is divided into "frames" of 8 subcarriers (312.5 kHz) by 12 OFDM symbols (345.6 μs). A frame duration is denoted a *slot*. These frames constitute the unit for frequency-adaptive resource allocation. As in any OFDM system, the frame size is selected to make the channel moderately flat within frames. Uplink pilot symbols known to the receiver facilitate the prediction. They are here assumed located on one of the 12 OFDM symbols. This entire OFDM symbol is allotted to pilots so that no payload data is transmitted here. We assume a full-duplex FDD uplink, so uplink pilots will be transmitted within each slot.

To prepare for frequency adaptive uplink transmission, the terminal is allocated a competition band and begins to send pilots in that band. Estimators of the noise variance, the velocity, fading descriptors, and channel power delay profile aid in the setup of a model of the fading taps as well as a model of the channel measurements through pilots.

Channel predictions are then produced for this users channel. When a packet for uplink transmission arrives, the terminal sends a transmission request during slot j. The scheduler may grant the request and sends the allocation information over a downlink control channel during slot $j+1$. The transmission then commences over the uplink in slot $j+2$. The required prediction horizon is two slots, or 0.7 ms, or $L = 2$ channel samples. This tight control loop requires the update of the channel prediction from the last measurement in slot j, the scheduling and the downlink control transmission to be executed within less than 1.5 slot durations (0.5 ms).

The results in this section are evaluated on two channel models: A flat (frequency non-selective) channel, and a frequency selective non-line-of sight channel for urban environments (WINNER C2 channel) with power delay profile

Delay[ns]	Power[dB]
0, 5, 135, 160, 215, 260, 385, 400, 530, 540, 650, 670, 720, 750, 800, 945, 1035, 1185, 1390, 1470	-0.5, 0.0, -3.4, -2.8, -4.6, -0.9, -6.7, -4.5, -9.0, -7.8, -7.4, -8.4, -11.0, -9.0, -5.1, -6.7, -12.1, -13.2, -13.7, -19.8

When not explicitly stated otherwise, we set the velocity of the terminals to 50 km/h, the average signal-to-noise ratio E_s/N_0 to 12 dB, and the filter width W to 8 subcarriers (one chunk width). The estimation horizon is set to

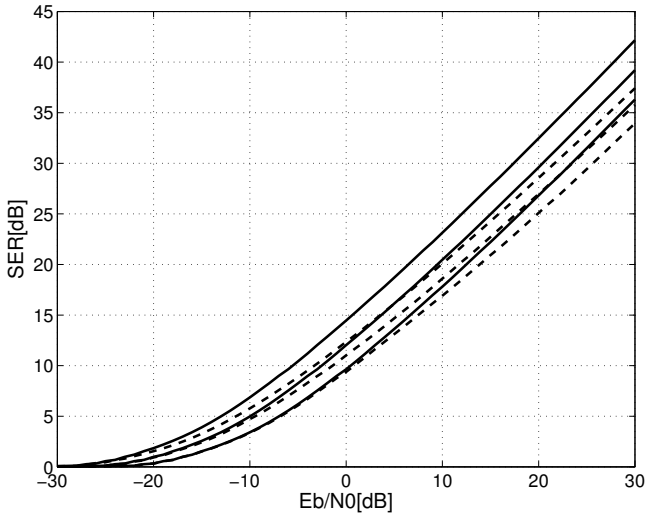


Figure 5.1: Filter performance versus filter width for widths 4, 8, and 16. Solid lines show the performance on a flat fading channel. Dashed lines for the frequency selective WINNER C2 non-line-of sight channel.

two steps (slots) and the fading statistics is set to the flat doppler spectrum described in Chapter 3. Performance is expressed either in terms of the mean value over all $W \times U$ channel taps of the signal-to-estimation error power ratio (SER), or by the normalised mean square error NMSE = $(\text{SER})^{-1}$.

5.3.1 The impact of filter width

One may adjust the dimensionality W of y (the filter width), i.e. the number of simultaneous subcarriers to be tracked, depending on the performance/complexity tradeoff. A competition band that comprises c predicted subcarriers will then require the use of int$[c/W]$ Kalman predictors run in parallel.

An increased filter width W should increase the performance. For a flat fading channel with noise without frequency correlation ($R = \sigma_n^2 \mathrm{I}$), the filter performance as measured by the signal-to-error ratio (SER) will increase by 3 dB when the filter width is doubled. The improvement is not as pronounced when the channel is frequency selective. We illustrate this in Figure 5.1.

5.3.2 The choice of pilots

The pilots $\{\phi_t\}$ are also design parameters. Should the pilot symbols transmitted by each user be placed on all W subcarriers that are tracked, hence making the pilots from the different users overlap? Or should one instead use dedicated pilots, so that each user concentrates its pilot energy to one single subcarrier, not transmitting anything on the remaining $W-1$ subcarriers?

Assuming that the number of users U in the competition band is less or equal to the number of subcarriers W, we may represent the pilots by a $W \times W$-matrix Φ, where each column contains the complex-valued time-frequency pilots for one user. The diagonal of the diagonal pilot matrices $\{\phi_u\}$ are then constructed from the columns of Φ.

The choice of pilots is crucial to achieve a high performance. We here evaluate two pilot schemes:

- *Dedicated* pilots, where each user puts pilots only on one subcarrier out of the W subcarriers tracked by one filter, with zero energy on the subcarriers used by other users.
- The use of *overlapped* pilots, where all users place pilots on all W subcarriers. We here use Walsh sequences to ensure that these pilots are orthogonal as long as the number of users U is less than or equal to W.

Dedicated pilots are simply obtained through $\Phi = \sqrt{W}\mathbf{I}_W$, where $\mathbf{I}$ denotes the identity matrix. Overlapping pilots are constructed through $\Phi = \text{hadamard}(W)$.

Although complex hadamard matrices are possible to find, they have no advantage over real matrices. We will here use Sylvester's construction which yields pilot symbols of either -1 or 1, that is BPSK symbols. Hence we construct a $2^n \times 2^n$ Hadamard matrix by setting $H_0 = 1$ and iterating

$$H_{n+1} = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix} \tag{5.24}$$

It is not possible to construct more than W real or complex-valued orthogonal pilot sequences. If the number of users is greater than the filter width W, we therefore need to construct additional non-orthogonal pilot sequences from the orthogonal set Φ. There is no general scheme for how to do this optimally. In this experiment we construct new pilots by pairwise combining pilots from the original set and multiplying the sum with $1/\sqrt{2}$ to preserve energy. The matrix used here that maps 8 orthogonal pilots onto

16 non-orthogonal pilots is

$$\begin{bmatrix} & \alpha & \alpha & \cdot & \alpha & \cdot & \cdot & \alpha & \cdot \\ & \alpha & \cdot & \alpha & \cdot & \alpha & \cdot & \cdot & \alpha \\ & \cdot & \alpha & \alpha & \cdot & \cdot & \alpha & \cdot & \cdot \\ \mathbf{I}_8 & \cdot & \cdot & \cdot & \alpha & \alpha & \alpha & \cdot & \cdot \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \alpha & \alpha \\ & & & & 3{\times}8 \text{ zeros} & & & & \end{bmatrix} \tag{5.25}$$

where $\mathbf{I}_8$ is the 8×8 identity matrix, and $\alpha = 1/\sqrt{2}$.

We compare the performance of overlapped pilots against the performance of dedicated pilots. Figure 5.2 shows SER versus number of users U when the channels experienced by the users are flat and the filter width is set to $W = 8$. Here we turn our attention only to the case $U \leq W$. All subcarriers fade in unison and the pilots for users 1 through 8 are completely orthogonal. In the noise-free case, the W measurements provided at time t would then provide a solvable linear system of equations with respect to the $U \leq W$ different channel coefficients. This holds regardless of whether we use overlapped (black, solid line) or dedicated pilots (grey, dashed line), which may be seen by letting

$$h_t = \text{diag}_U(\mathbf{I}_W)\bar{h}_t, \tag{5.26}$$

where $\bar{h}_t = (\bar{h}_t^{(1)}, \ldots \bar{h}_t^{(U)})^T$, where $\bar{h}_t^{(j)}$ is the flat-fading scalar channel coefficient for user j and $\mathbf{1}_W$ is a column vector of W ones. In the noise-free case, we then have

$$y_t = \text{stack}_U(\phi_{j,t})h_t = \text{stack}_U(\phi_{j,t})\text{diag}_U(\mathbf{I}_W)\bar{h}_t = \Phi_{W\times U}\bar{h}_t \quad ,$$

where $\Phi_{W\times U}$ equals the first U columns of the matrix Φ which, by construction, will have full rank U. When the channels are flat fading we therefore have the result that

- the choice of pilots is irrelevant as long as the pilots are orthogonal, and
- the performance does not degrade with an increasing number of users U as long as $U \leq W$.

The situation is vastly different when the channels are frequency selective. The importance of measuring over the entire filter bandwidth is evident when we study Figure 5.3 for users 1–8. For the particular working point $E_s/N_0 = 12$ dB studied here, the gain is about 3 dB for one user, and decreases when the number of simultaneous users increases.

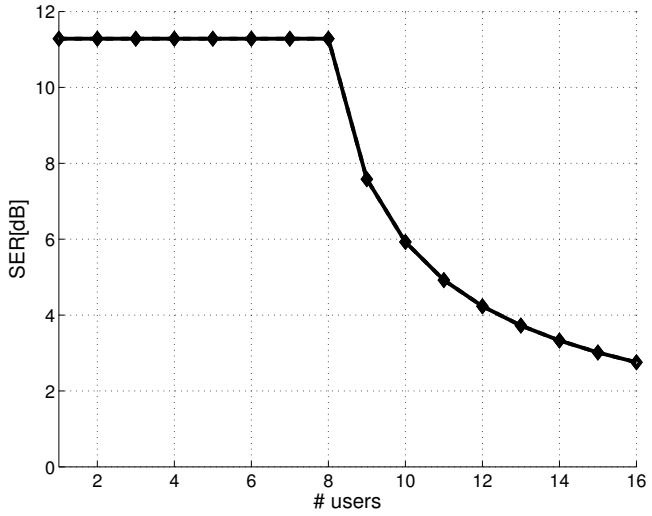


Figure 5.2: Prediction performance versus number of users for dedicated (dashed line) and overlapped (dolid line) pilots on flat fading channels. The lines overlap. Average $E_s/N_0 = 12$ dB. The prediction horizon is 2 steps.

The curves merge at the point $U = 8$, indicating that the choice of pilots is unimportant when the orthogonal set has been filled. This conclusion should however be drawn with care, because in the dedicated pilots case, the SER will vary considerably, from high (on the carrier over which pilots are transmitted), to low (on carriers far from the pilot carrier). Since modulation format is selected per frame, one would prefer a more even distribution of the SER, such as is produced by the overlapping pilots.

The reason for the performances for dedicated pilots (grey, dashed) actually *increasing* with U is due to the way the pilot subcarriers have been allocated to users in this experiment. The first user here puts pilot energy on the first subcarrier, which is on the border of the filter bandwidth, while users 4 and 5 invest their pilots in the middle of the bandwidth. The latter is the better tactic when we rate performance based on the mean value of the SER over all subcarriers. This is the reason for the performance increase when users 2, 3 and so on are added to the system.

This illustrates that if dedicated (and time static) pilots are to be used, then one should assign one of the middle subcarriers to the first user to enter the system, and only assign border subcarriers when necessary.

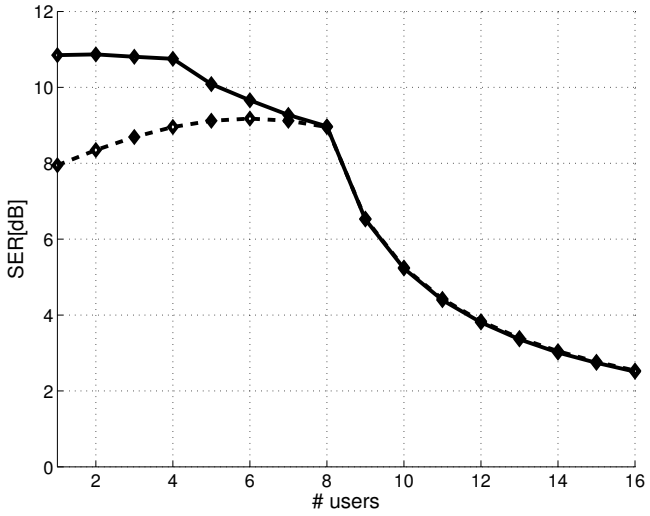


Figure 5.3: Prediction performance versus number of users for dedicated (dashed line) and overlapped (dolid line) pilots on frequency selective channels. WINNER C2 channel model and time-invariant pilot patterns. Kalman estimators track $W = 8$ adjacent subcarriers. Average $E_s/N_0 = 12$ dB, velocity 50 km/h, 3.7 GHz carrier. The prediction horizon is 2 steps.

5.3.3 Time varying pilot patterns

When the number of users U to share a certain bandwidth is larger than the corresponding filter width W, it is not possible to find a set of U orthogonal pilots. As presented earlier, we then construct new pilots from the original set of W orthogonal pilots by weighing together them two by two. Restudying Figures 5.2 and 5.3, we note two facts. One is that the performance drop when we go from orthogonal to non-orthogonal pilots ($U = 8$ to $U = 9$) is considerable. The other fact is that the performance is unaffected by the choice of dedicated versus overlapping pilots.

The performance can be improved by providing the filter with more information about the time variability of the channels. We have seen that the filtering performance increases if we spread out the pilot energy and let the pilots vary over the different frequencies in the frequency band. In the same manner we may design the pilots to make optimal use of previous channel samples. In the case of noiseless, *frequency-selective but time-invariant* channels (i.e. immobile terminals), we would obtain a linear system of equations

$$Y = Ah \tag{5.27}$$

where

$$\begin{aligned} A &= [\text{stack}_U(\phi_{j,t})^T \dots \text{stack}_U(\phi_{j,t+W-1})^T]^T, \\ Y &= (y_t^T, \dots, y_{t+M-1}^T)^T, \text{ and} \\ h_t &= h. \end{aligned} \tag{5.28}$$

If a set of orthogonal pilots is cycled over time so that A obtains full rank WU, the system (5.27) becomes solvable. That should improve the estimation also for time-varying channels and noisy measurements.

For *dedicated* pilots, this property is obtained for $M = 8$ by simply rotating the original $\Phi_{t=0} = \sqrt{8}\mathbf{I}_8$ one step left every time step, hence producing all eight time steps.

Time varying *overlapping* pilots are constructed as follows. We here study the specific case $U \leq 8$. For the first time step we use the same Hadamard matrix as used for the static pilots:

$$\begin{aligned} \Phi_{t=0} &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \\ &= \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \Phi_4 & \Phi_5 & \Phi_6 & \Phi_7 & \Phi_8 \end{bmatrix}, \end{aligned} \tag{5.29}$$

where Φ_i is the pilot pattern used by user i at time $t = 0$. This Hadamard matrix is then used a second time to construct all time steps. In the resulting matrix below, each row correspond to one time step $t = 0, 1, \dots 7$.

$$\begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \Phi_4 & \Phi_5 & \Phi_6 & \Phi_7 & \Phi_8 \\ \Phi_1 & -\Phi_2 & \Phi_3 & -\Phi_4 & \Phi_5 & -\Phi_6 & \Phi_7 & -\Phi_8 \\ \Phi_1 & \Phi_2 & -\Phi_3 & -\Phi_4 & \Phi_5 & \Phi_6 & -\Phi_7 & -\Phi_8 \\ \Phi_1 & -\Phi_2 & -\Phi_3 & \Phi_4 & \Phi_5 & -\Phi_6 & -\Phi_7 & \Phi_8 \\ \Phi_1 & \Phi_2 & \Phi_3 & \Phi_4 & -\Phi_5 & -\Phi_6 & -\Phi_7 & -\Phi_8 \\ \Phi_1 & -\Phi_2 & \Phi_3 & -\Phi_4 & -\Phi_5 & \Phi_6 & -\Phi_7 & \Phi_8 \\ \Phi_1 & \Phi_2 & -\Phi_3 & -\Phi_4 & -\Phi_5 & -\Phi_6 & \Phi_7 & \Phi_8 \\ \Phi_1 & -\Phi_2 & -\Phi_3 & \Phi_4 & -\Phi_5 & \Phi_6 & \Phi_7 & -\Phi_8 \end{bmatrix} \tag{5.30}$$

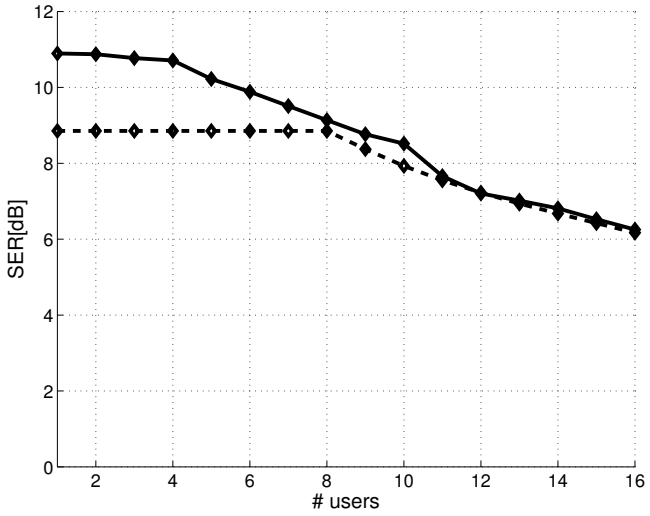


Figure 5.4: Prediction performance versus number of users for dedicated (dashed line) and overlapped (dolid line) pilots on frequency selective channels. Conditions as in Figure 5.3, but here the pilot patterns of each user cycle over time with period 8. The prediction horizon is 2 steps.

The impact of using cyclic pilots is studied in Figure 5.4. When the number of simultaneous users U is less or equal to eight, we see an improvement of about 1 dB for the dedicated pilots and somewhat less for the overlapped pilots, as compared to the case when static pilots were used (Figure 5.3).

When $U \geq 9$ the improvement is more dramatic. The steep performance drop at $U = 9$ is now gone. We conclude that the use of cyclic pilots is highly important to maintain a high estimation performance when the number of users competing for a frequency band is larger than the bandwidth W.

5.3.4 The impact of fading statistics

The Doppler spectrum, caused by the angular distribution of local scatterers around each terminal, relative to its direction of travel, has a crucial impact on the channel predictability. So far we have used a fading model with Doppler spectrum that is almost flat for frequencies less than the maximum Doppler frequency f_D. This corresponds to a situation where scatterers are placed mainly sideways relative to the direction of travel, e.g. due to buildings along streets. We evaluate the prediction performance for different signal-to-noise ratios (E_s/N_0) over a wide range of prediction horizons. The

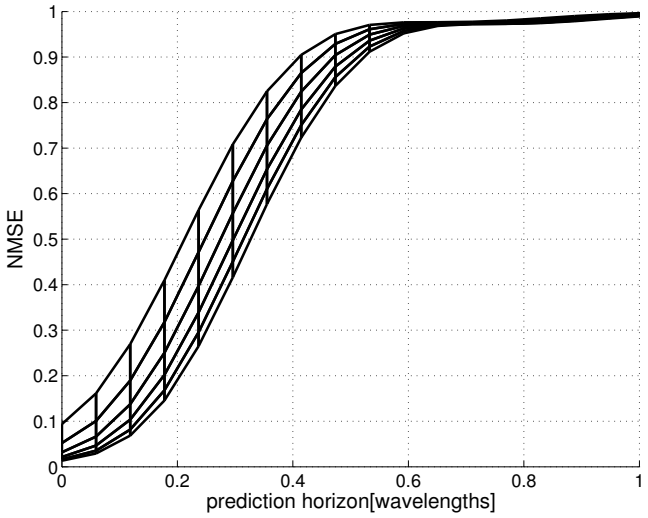


Figure 5.5: The predictor performance for flat Doppler spectrum measured by the NMSE, versus prediction horizon measured in wavelengths. The signal-to-noise ratio Es/N_0 goes from 0 dB (upper curve) to 25 dB (lower curve) in steps of 5 dB. $U = 8$ users with overlapping pilots, $W = 8$, WINNER C2 channel model.

result for a flat Doppler spectrum is presented in Figure 5.5.

An opposite extreme case is illustrated for the same situation as in Figure 5.5, but for a Doppler spectrum modelled by four poles close to the unit circle. This corresponds to a situation dominated by reflectors in front of or behind the terminal. This results in very good predictability. See Figure 5.6.

Originally, my own stand on the matter of how to choose fading model, was that one should be cautious and always use the flat Doppler spectrum derived in Chapter 3. From the above results it is evident that one cannot afford to do this; the fading statistics has far too great an impact on the system performance.

5.3.5 Complexity

To evaluate the feasibility of using optimal filters as proposed here, we need to assess the complexity of the Kalman filter (KF). The complexity is mainly determined by the number of states n, which is the product of the fading model order K, the number of modelled taps X, and the number of users U. The value of X depends on how the modelling is carried out: If time-

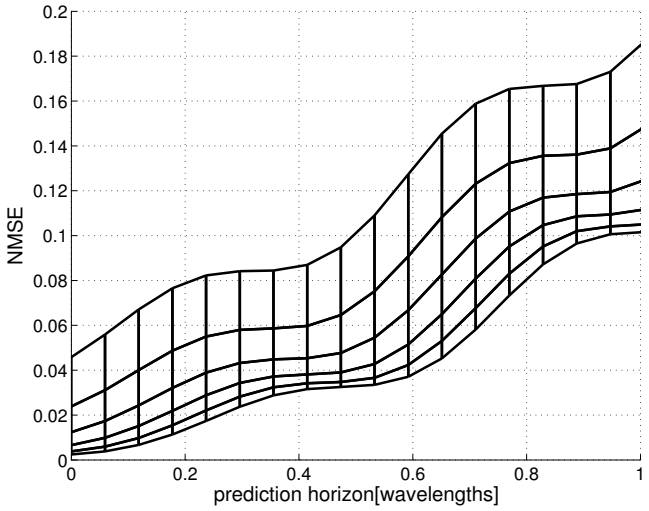


Figure 5.6: The predictor performance for oscillative AR4 Doppler spectrum measured in NMSE, versus prediction horizon measured in wavelengths. The signal-to-noise ratio Es/N_0 goes from 0 dB (upper curve) to 25 dB (lower curve) in steps of 5 dB. $U = 8$ users with overlapping pilots. $W = 8$, WINNER C2 channel model.

domain prediction of M taps is used, we set $X = M$. Otherwise we use frequency-domain prediction and choose $X = W$. Both cases are covered below.

The KF has to produce channel tap estimates (update the state estimates $\hat{x}$), predictions of the channel taps, as well as updates and predictions of the state error covariance matrix P. Due to the block diagonal structures used in this paper, the complexity is reduced considerably as compared to the general KF. It can be shown (see Appendix G) that the number of complex operations required for one KF update is

$(3/2 + W/2)n^2 + (KW + W^2)n$ $+KW^2/2 + W^3/6$	P update
$3n^2/2$	P prediction
$(W + 1)n + KW$	$\hat{x}$ update
n	$\hat{x}$ prediction

To cover a competition band that contains c predicted subcarriers, $\text{int}[c/W]$ KFs are run in parallel. The solid lines in Figure 5.7 display the number of real operations required per update versus number of users for $C = 160$

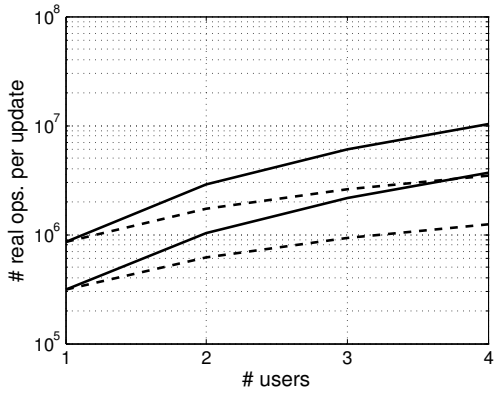


Figure 5.7: Total numerical complexity vs. number of users k for predicting a competition band containing 160 predicted subcarriers, using either 40 KFs of filter width $W = 4$ (lower) or 20 KFs with width $W = 8$ (upper). Solid lines represent a general choice of pilots. Dashed lines represent the use of dedicated pilots and k decoupled KFs for each set of W subcarriers.

predicted subcarriers and $K = 4$ uplink users per competition band, for designs with $W = 4$ or $W = 8$. An operation represents one multiplication *and* one addition. Furthermore we assume that one complex operation has the same complexity as four real operations.

The above calculations hold for general choices of pilot symbols, but the complexity may be decreased further by considering only dedicated pilots (as opposed to overlapping pilots). The measurement equation is then completely decoupled between different users, which makes the state error covariance matrix P block diagonal. This, in turn, means that we may run a separate KF for each user without losing performance, which means that the complexity increases only linearly with the number of users. In that case, the complexity is easily computed by setting $n = KX$ instead of $n = KXU$ in the above expressions, and then multiplying the final result by U. The dashed lines of Figure 5.7 show the number of real operations required for one update for filter widths $W = 4$ and $W = 8$ when these decoupled KFs can be used.

The WINNER baseline system would require a new prediction for each slot of duration 0.34 ms for vehicular users. To assess the feasibility of the required computational complexity, we here investigate the consequences of setting 10^{10} real operations per second as a target for feasibility for uplink

predictors realised in the base station.[1] This would correspond to a limit to $3.4 \cdot 10^6$ operations per update (0.345 ms). As is evident from Figure 5.7 (solid), using eight parallel subcarriers would then be infeasible, while four parallel subcarriers touches upon the limit. For the decoupled case (dashed), the total complexity of filters of width 4 falls well within our boundary while $W = 8$ is on the limit for $k = 4$. The use of Kalman-based uplink prediction seems feasible under these assumptions.

The numerical load imposed by the KF is dominated by the covariance matrix update. However, P usually converges very quickly to a stationary value (the solution to the Discrete Algebraic Riccati Equation) when the model matrices are kept constant. This holds also when cyclic pilots are used, in the sense that P will then approach a cyclostationary state in which the same value reoccurs with the same period as the cyclic pilots. In the experiment performed in this paper, P converged to a fixed value[2] in a few tens of iterations, and for most cases the iteration count was below 20.

Therefore, Kalman iterations only need to be performed burst-wise when fading models or number of users change, due to the fast convergence of P. This reduces energy consumption relative to the case of continuous updates.

5.4 Channel gain prediction

It is commonly suggested that the squared magnitude of the channel taps (that is, the channel *gain* or the channel *power*) should function as basis for scheduling decisions. See e.g. [23], [24],[25].

Producing the optimal estimate of the channel gain is straightforward. Let $z = |h|^2$ for a tap h. Then, using well-known results[3], a change of variables reveals

$$p(h|YI) = \mathcal{CN}(x; \hat{h}, \sigma^2) \Rightarrow p(|h|^2|YI) = \chi^2(|h|^2; |\hat{h}|^2 + \sigma^2, 2|\hat{h}|^2 + 4\sigma^2)), \tag{5.31}$$

where $\chi^2(\cdot)$ is the non-central χ^2-distribution with mean value and variance in the second and third argument. The optimal mean value tap power estimate is hence the squared magnitude of the complex tap estimate plus the error variance:

$$\mathrm{E}(|h|^2|YI) = |\hat{h}|^2 + \sigma^2 \tag{5.32}$$

[1] Lower targets would be realistic for predictors located in terminals.

[2] We consider P to have converged when the maximum element-wise relative change in magnitude between iterations is below one percent.

[3] See Appendix C for a proof.

This result was obtained with considerably more effort in [24] and [14], without the appreciation of its optimality.

When h is a vector we simply have

$$\mathrm{E}(|h|^2|YI) = \text{diagonal elements of } H(\hat{x}\hat{x}^* + P)H^*. \tag{5.33}$$

However, it turns out that the channel gain alone does not serve very well as a basis for scheduling decisions. Before making this clear, we have to consider how to actually predict the bit error rate.

5.5 Bit error rate prediction

The bit error rate, here denoted P_b, is conveniently measured versus the ratio *bit energy* (E_b) to *noise spectral density* (N_0). That way, different modulation formats/encoding schemes can be justly compared to one another[4].

The bit error energy E_b is related to the symbol energy E_s through $E_b \log_2 M = E_s$. Here, M is the number of symbols in the symbol constellation. In an OFDM system, I will denote the energy of the total OFDM symbol by $\bar{E}_s$. The symbol energy on each subchannel, E_s, relates to $\bar{E}_s$ through $NE_s = \bar{E}_s$, where N is the number of subcarriers in the system. The received OFDM symbol energy is $\bar{E}_s = S|h|^2$, where S is the transmitted signal energy per OFDM symbol, and the tap power $|h|^2$ should be thought of as the mean value over all subcarriers in this context.

The AWGN has a spectral power density of $N_0/2$ over all frequencies (naturally, this model is invalid for very high frequencies). When the cross correlators at the receiver are matched to the transmitted pulses, then we have the relation $\sigma_n^2 = N_0/2$ between the variance σ_n^2 of the noise samples and the noise power spectral density. This relation holds for each cross correlator, so if each symbol constellation has two dimensions – as is the case here – then the total variance for the sampled noise is $\sigma_n^2 = N_0$.

In total, we have the following relation between the tap power $|h|^2$ for a subcarrier and the E_b/N_0:

$$\frac{E_b}{N_0} = |h|^2 \frac{S}{N\sigma_n^2 \log_2 M} \tag{5.34}$$

Note that $|h|^2$ is regarded the only unknown parameter since σ_n^2 is given by the noise power estimator.

[4] This is a simplification. E_b/N_0 in combination with P_b does not show the whole picture. For example, bandwidth and algorithmic complexity are properties that are not expressed.

Henceforth I will denote

$$\gamma_0 \triangleq \frac{E_b}{N_0} \tag{5.35}$$

for convenience.

As noted in Appendix B, there is a functional relationship between P_b and γ_0. For simple cases it is possible to derive this relationship analytically, whereas one may need to use simulation for more complicated problems. Once the function $P_b(\gamma_0)$ has been retrieved, it can be approximated by

$$P_b(\gamma_0) \approx \sum c_k \exp\{-\alpha'_k \gamma_0\} = \sum c_k \exp\{-\alpha_k |h|^2\}, \tag{5.36}$$

for suitable values on $\{c_k\}$ and $\{\alpha_k\}$. Note that the $\{\alpha_k\}$ incorporate the scaling factor in Equation (5.34). Focusing on the taps rather than on E_b/N_0, we thus can write

$$P_b(|h|^2) = \sum c_k \exp\{-\alpha_k |h|^2\} \tag{5.37}$$

and I deliberately sacrifice rigour ($P_b(\gamma_0)$ and $P_b(|h|^2)$ are different functions, not the same function with different arguments) to the benefit of readability.

In (5.37) it was assumed that the bit error rate depends on one tap alone. This is only true in uncoded OFDM systems, where each time-frequency symbol is the same as a supersymbol. In all other types of system, one would have to extend the expression to take many taps into account:

$$P_b(|h_1|^2, \ldots, |h_X|^2) = \sum_x \sum_k c_{k,x} \exp\{-\alpha_{k,x} |h_x|^2\}, \tag{5.38}$$

where X is the number of taps affecting the supersymbol. In practice, however, this will not be necessary for OFDM systems, since they are dimensioned in such a way that the channel is moderately flat within a received block, so that all time-frequency taps have approximately the same value. If coding is introduced only within the received block itself, then (5.37) can be used. Single-carrier systems will however have to use (5.38).

In most systems one can find a bit error rate TBER that is the highest tolerable that still allows the system to function properly. I will make this assumption in what follows. Consequently, TBER is regarded as a given system design parameter.

To solve the bit error rate prediction problem, two approaches present themselves: We may, given the pdf for a tap[5], calculate the *expected value*

[5] Taken that the unit for scheduling decisions is a "frame" of some size (time in TDMA, time-frequency in OFDM), this tap should be the "worst" tap within that frame.

of P_b for each of the candidate modulation formats. The highest modulation format whose expected P_b conforms to the bit error rate requirement is then signalled to the base station.

The other approach is to instead calculate the *probability* of P_b being lower than TBER for each modulation format. The highest format whose probability is higher than, say, 90%, is signalled to the base station.

Based on requested modulation formats from each user and possibly fairness criteria, the base station will then distribute resources among the users. The decisions are signalled back to the terminals, either on control symbols intertwined with the pilots and the payload symbols, or on a separate control channel.

Next, the two bit error prediction methods are studied more closely.

5.5.1 Expected bit error rate

We here look at a particular time-frequency tap in an OFDM system. Looking at the time-frequency taps one by one mean that we are in effect considering uncoded systems, since coding introduces dependencies between different taps which therefore have to be considered jointly. The exposition following here will be restricted to the uncoded case.

The mean value $\hat{h}$ of its gaussian distribution is taken from the vector $H\hat{x}$. The variance σ^2 is taken from the corresponding diagonal element in HPH^*. The expected value of P_b can now be calculated:

$$\begin{aligned} \mathrm{E}(P_b|YI) &= \sum c_k \int_{\mathbb{C}} \exp\{-\alpha_k|h|^2\}\mathcal{CN}(h;\hat{h},\sigma^2)\mathrm{d}h \\ &= \sum c_k \int_{\mathbb{C}} \frac{1}{2\pi\sigma^2}\exp\{-\frac{|h-\hat{h}|^2}{\sigma^2} - \alpha_k|h|^2\}\mathrm{d}h \\ &= \sum c_k \frac{\exp(-\alpha_k|\hat{h}|^2/(1+\alpha_k\sigma^2))}{2(1+\alpha_k\sigma^2)} \end{aligned} \tag{5.39}$$

The expected value (5.39) is calculated for each modulation format (each having a unique set of values $\{c_k\}, \{\alpha_k\}$). The highest format for which $\mathrm{E}(P_b) <$ TBER is signalled to the base station.

For TDMA systems, the situation is slightly more complicated. The bit error rate P_b is now a function of more than one tap. It will not suffice then to take into account the most recent tap h_t, but one must also regard the pdfs of h_{t-1}, h_{t-2} and so on. I will not treat this issue here.

5.5.2 Probability of bit error rate

We want to calculate

$$\mathrm{P}(P_b < \mathrm{TBER}|YI) \tag{5.40}$$

for all modulation formats and signal back to the base station those who have a higher probability than, say, 90%. Again, we consider only the uncoded case. Since P_b is a decreasing function of b, where $b = |h|$, we might just as well calculate

$$\mathrm{P}(b > f(\mathrm{TBER})|YI) = \int_{f(\mathrm{TBER})}^{\infty} p(b|YI)\mathrm{d}b, \tag{5.41}$$

where $f(\mathrm{TBER})$ is found by numerical inversion of (5.37). It is well-known that if the pdf of h is gaussian with mean $\hat{h}$ and variance σ^2, then the pdf of b is a Rice distribution:

$$p(b|YI) = \frac{b}{\sigma^2}\exp^{-(b^2+\hat{h}^2)/2\sigma^2} I_0(b\hat{h}/\sigma^2) \tag{5.42}$$

(To see this, make a straightforward change of variables in the non-central χ^2-distribution derived in Appendix C). The expression (5.42) may seem complicated, but the integral (5.41) is conveniently expressed by means of the *Marcum Q function*. Below, I define it and show an efficient algorithm for calculating it.

The Marcum Q function

The *Marcum Q function* is defined

$$Q_1(a,b) = \int_b^{\infty} xe^{-(x^2+a^2)/2} I_0(ax)\mathrm{d}x \tag{5.43}$$

An efficient approximation of $Q_1(a,b)$ is given in [26]. I reproduce it here for convenience:

$$Q_1(a,b) = \begin{cases} \frac{\alpha_N}{2\beta_N} e^{-(a-b)^2/2} & \text{if } a < b \\ 1 - \frac{\alpha_N}{2\beta_N} e^{-(a-b)^2/2} & \text{if } a \geq b \end{cases} \tag{5.44}$$

in which

$$\begin{aligned} \alpha_n &= d_n + \frac{2n}{ab}\alpha_{n-1} + \alpha_{n-2} \\ \beta_n &= 1 + \frac{2n}{ab}\beta_{n-1} + \beta_{n-2} \end{aligned} \tag{5.45}$$

are iterated until $N = 5(1 + \sqrt{ab})$, using

$$d_{n+1} = d_n d_1, \quad \alpha_{-1} = 0, \quad \beta_{-1} = 0, \quad \beta_0 = 0.5, \tag{5.46}$$

and

$$\begin{array}{llll} \alpha_0 = 1, & d_1 = a/b & & \text{if } a < b \\ \alpha_0 = 0, & d_1 = b/a & & \text{if } a \geq b \end{array} \tag{5.47}$$

Hence we have

$$\mathrm{P}(P_b < \mathrm{TBER}|YI) = \frac{1}{\sigma} Q_1\left(\frac{\hat{h}}{\sigma}, \frac{b}{\sigma}\right), \tag{5.48}$$

which completes the solution of the prediction problem.

Note that the fact that (5.42) requires only a few hundred floating point operations, means that the total numerical complexity is completely dominated by the Kalman filter.

Whether we should use the criterion $\mathrm{E}(P_b) < \mathrm{TBER}$ or the criterion $\mathrm{P}(P_b < \mathrm{TBER}) > 90\%$ to decide which users and modulation formats are candidates for resource allocation depends on which properties one wishes for the system. The expectancy criterion guaranties that the total bit error rate is TBER, whereas the second criterion makes sure that the bit error rate over 90% of the time is lower than TBER. Which one to prefer is left as an open question in this thesis.

Let us finally study the relation between predicted bit error rate and predicted gain. We will do so by studying a specific example. It will show why I think it is a bad idea to use predicted channel gain as foundation for scheduling decisions.

EXAMPLE 5.1 **CHANNEL GAIN VERSUS BIT ERROR RATE**

We here calculate the predicted expectations of P_b for the particular modulation format *Differential Phase Shift Keying* (DPSK) and compare it to predictions of the channel gain. DPSK has the nice property of fitting the format (5.37) precisely:

$$P_b = \frac{1}{2} e^{-E_b/N_0}, \tag{5.49}$$

meaning that

$$c_0 = \frac{1}{2} \quad \text{and} \quad \alpha_0 = \frac{S}{N\sigma_n^2 \log_2 M} = \frac{S}{N\sigma_n^2}, \tag{5.50}$$

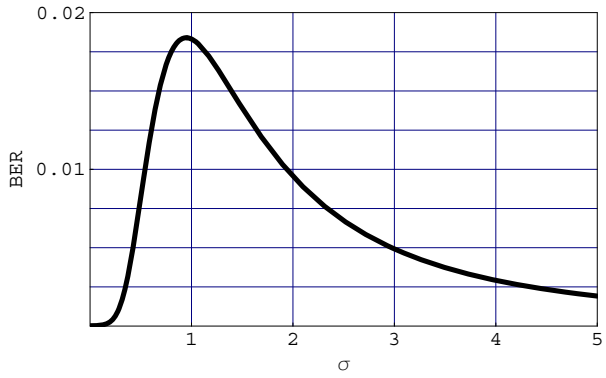


Figure 5.8: The expected bit error rate versus predictor uncertainty for DPSK.

since $\log_2 M = 1$ for binary modulation formats. Inserting this into (5.39), we get

$$\begin{aligned} \mathrm{E}(P_b) &= c_0 \frac{\exp(-\alpha_0 |\hat{h}|^2/(1+\alpha_0\sigma^2))}{1+\alpha_0\sigma^2} \\ &= \frac{1}{2} \frac{\exp(-\alpha_0 |\hat{h}|^2/(1+\alpha_0|\hat{h}|^2 \times \sigma^2/|\hat{h}|^2))}{1+\alpha_0|\hat{h}|^2 \times \sigma^2/|\hat{h}|^2}, \end{aligned} \tag{5.51}$$

where $\alpha_0|\hat{h}|^2 = \gamma_0$ and $\sigma^2/|\hat{h}|^2$ is the normalised prediction variance. A typical value for the former would be between 0 dB and 20 dB. The latter should stay below 1 but may be higher under difficult circumstances.

We now investigate how $\mathrm{E}(P_b)$ varies for varying σ^2 but static $|\hat{h}|^2$. Setting $\alpha_0 = 10$ and $|\hat{h}|^2 = 1$, we plot $\mathrm{E}(P_b)$ versus σ. See Figure 5.8.

It is evident that good bit error rates will be attained for both high and low values of σ, as long as it stays away from 1 (a high σ in combination with a low $|\hat{h}|$ usually means that the data is unable to support a decision, but that the prior knowledge is that the channel has an overall high quality).

As a contrast to this we study a predictor that predicts the channel gain. Such a predictor will always favour a high σ before a low, since $\mathrm{E}(|h|^2) = |\hat{h}|^2 + \sigma^2$ as we saw earlier. Using the predicted channel gain alone is therefore inadequate, since it does not tell how the energy is distributed among the squared tap prediction $|\hat{h}|^2$ and the tap prediction uncertainty σ^2. A certain fixed tap gain prediction can therefore correspond to many different predicted error rates.

5.6 System design

The details of how to apply the present algorithm have now been described, both regarding estimation and prediction. But there are more to investigate. Just as we examined the channel prediction performance in Section 5.3, we would want to do the same thing for bit error rate prediction.

In Section 5.3 we investigated the *channel* prediction performance for a specific system. However, we saw later that both complex channel tap predictions and predictions of the squared channel magnitude generally correspond poorly with bit error rate predictions. We would therefore like to develop a tool for analysing the *bit error rate* prediction performance of a system, or rather, given the predictors developed in this thesis, we would like to assess the resulting bit error rates for the respective users in a specific system.

However, this is a work in progress, and I will only here be able to outline a few vague ideas on how to pursue this goal.

Let me first reiterate what the objective is here. We want to examine a system – the system being characterised by number of users, fading statistics, pilot patterns, and so on – under *static* conditions. In a real scenario such parameters as signal-to-noise ratios and fading statistics will change over the course of time. These parameters will be updated by "off-line" estimators that feed the linear channel model with high-quality estimates. The transition between one set of parameters and another will however usually be swift due to the fast convergence of the DARE. When the off-line estimators update the linear model, the uncertainty P of the states x will then quickly settle down to a steady value. The system will spend most of its time in such a steady state which is why the steady state scenario is what we are most eager to investigate.

Section 5.2.3 showed that a steady state induces certainty about frequency distributions. If, for example, we would use a prediction horizon of one time step and hence use the one-step predictions, we would have the frequency distributions

$$\begin{aligned} fr(h) &= \mathcal{CN}(h; 0, H\bar{\Pi}H^*), & \bar{\Pi} &= F\bar{\Pi}F^* + GQG^* \\ fr(\tilde{h}) &= \mathcal{CN}(h; 0, H\bar{P}H^*), & \bar{P} &= F\bar{P}F^* + GQG^* - \bar{K}_p\bar{R}_e^{-1}\bar{K}_p^* \\ fr(\hat{h}) &= \mathcal{CN}(h; 0, H\bar{\Sigma}H^*), & \bar{\Sigma} &= F\bar{\Sigma}F^* + \bar{K}_p\bar{R}_e^{-1}\bar{K}_p^*, \end{aligned} \tag{5.52}$$

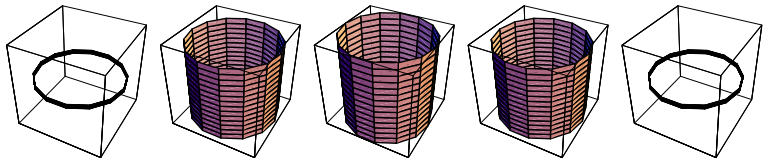


Figure 5.9: An attempt to illustrate a polydisc in two complex dimensions, defined by $|h_1| < r_1, |h_2| < r_2$. Since the number of real dimensions is four, this has to be done by using cross-sections in one dimension. The leftmost block has (nearly) $Re(h_2) = -r_2$, and the rightmost block has (nearly) $Re(h_2) = r_2$. The cylinders should be solid.

where h is the vector of the actual taps, $\hat{h}$ is the predictions, and $\tilde{h}$ is the errors in the predictions so that $h = \hat{h} + \tilde{h}$. For longer horizons we simply apply (5.11) to $\bar{P}$ and use $\bar{\Pi} = \bar{\Sigma} + \bar{P}$. Note that $\bar{\Pi}$ describes the actual fading of the taps and hence does not change with a changing horizon.

As means for deciding whether a user should compete for a resource or not, I have proposed two methods. The expected bit error rate criterion ensures a certain bit error rate over a long time period. The bit error rate probability criterion makes sure that the bit error rate over a specified fraction of time stays below some limit.

A steady state scenario has a fixed P matrix, which means that the variance of the pdf of each tap is fixed. The predicted bit error rate will then be a decreasing function of the squared magnitude of the tap *prediction*, regardless of which method is used. It is therefore possible to find "decision boundaries"

$$\Omega_{r_1 r_2 \ldots r_U} = \{c_{1,r_1} \leq |\hat{h}_1|^2, \ldots, c_{U,r_U} \leq |\hat{h}_U|^2\}, \tag{5.53}$$

which each defines the set of tap prediction vectors for which a certain combination of modulation formats will be claimed to be useful. A number k written in base $M+1$, so that each of its digits r_u has a value between 0 and M, will therefore be a convenient way to define each such set; Ω_k is the manifold in which a user u claims to be able to use *at least* modulation format r_u.

The complement to Ω_k is called a *polydisc*. A polydisc is a Cartesian product of discs. Think of constructing a polydisc in n complex dimensions as first constructing a circle in (real) dimensions 1 and 2 and filling it, then letting the centre of the disc move in a new circle in dimensions 3 and 4 and successively filling the cylinders between the discs, then having the centre of this four-dimensional body move in a circle in dimensions 5 and 6 and so on. Figure 5.9 attempts to display a polydisc in two complex dimensions.

We may also define the set

$$\Lambda_k = \Omega_k - \Omega_{k+11\ldots1}, \tag{5.54}$$

where $k = r_1 r_2 \ldots r_U$. Λ_k is then the manifold in which a user u claims to be able to use modulation format r_u, but not format $r_u + 1$.

The number k is conveniently called the *state* of the system. We now have that the fraction of time that the system spends in state k is given by

$$t_k = \int_{\Lambda_k} fr(\hat{h})\mathrm{d}h, \tag{5.55}$$

where h is a vector holding the U taps, and U is the number of users (in systems other than uncoded OFDM systems, one has to consider more than one tap per user).

Using the model (5.37), we may also calculate the average bit error rate in state k:

$$\begin{aligned} P_{b,k} &= \sum c_k \int_{\Lambda_k} \exp\{-\alpha_k |h|^2\} \times fr(h)\mathrm{d}h \\ &= \sum c_k \int_{\Lambda_k} \int_{\mathbb{C}^U} \exp\{-\alpha_k |\tilde{h}|^2 - \alpha_k |\hat{h}|^2 - \alpha_k \tilde{h}\hat{h}^* - \alpha_k \tilde{h}\hat{h}^*\} \\ &\quad \times fr(\tilde{h}) \times fr(\hat{h})\mathrm{d}\tilde{h}\mathrm{d}\hat{h} \end{aligned} \tag{5.56}$$

Note that both (5.55) and (5.56) are integrals over gaussian distributions.

By integrating multidimensional gaussian functions over the volumes Λ_k, which is ultimately the same as integrating over polydics, we may therefore calculate the average bit error rate for each user in the system. Moreover, we will see how much time is spent in each state, so that we will know how "balanced" the system is with respect to different modulation formats.

How easy is it to integrate zero-mean gaussian functions over polydiscs? Unfortunately, no closed form expression exists. There *do* exist expressions for approximations of integrals over balls, although they are complicated and expressed on series form. It may be possible to "fill" the volumes $\{\Lambda_k\}$ with balls and calculate the associated integrals that way.

More feasibly, one may attempt to do the corresponding thing with "boxes"; oriented along the eigenvectors of the gaussian function, the volume of a box is simply given as a product of error functions, or as a product of differences between error functions when the box is not centred at the origin.

The question of how to efficiently tessellate the polydiscs so as to fill them efficiently with as few boxes as possible, I leave open for the time being.

Another issue also needs to be addressed. To make a diagnosis of a system, we seemingly need to calculate integrals over all volumes $\{\Lambda_k\}$. But there is an awful lot of them. In fact, they may easily be counted in trillions! I do not think this is a big problem, though, because for most of them, the corresponding times t_k will be negligibly small.

One could therefore start by calculating the integrals for a state in which the system should spend a lot of its time. Say, for example, that the system in question has five modulation formats and four users. Then one could begin by calculating t_{3333} and $P_{b,3333}$. According to some algorithm, one would then continue to examining the states "around" $k = 3333$, for example $k = 4333$, $k = 4433$, $k = 4443$ and so on. My guess is that, much like the "mass concentration" in Section 2.10, one would quickly cover almost the entire gaussian distributions. Monitoring the accumulated sum of the times t_k, it is possible that one would find $\sum t_k = 0.999$ only after a few hundred or a few thousand integrations, which would mean that further integration would be pointless.

Of course, these intuitive guesses may be wrong, and in any case it might be quicker to run a pseudo-random simulation to assess the system performance. Whether this is true or not depends on how many iterations a simulation would require to give a good numerical resolution, and how easily and swiftly the above integrations can be carried out. These questions cannot be answered before the matter has been further investigated.

Chapter 6

Future work

During the work leading up to this thesis, many suggestions for future work have presented themselves.

Chapter 2

Probability theory as an extension to logic has been used in this thesis. It is however evident that frequentist theory, whose range of application is smaller that that of Bayesianism, often works equally well as Bayesianism in channel estimation and prediction theory, provided that the Kalman filter is allowed to be used.

It is therefore much more interesting, and rewarding, to apply Bayesian theory to other fields. I expect that the use of Bayesian theory will increase especially in such fields as medicine and economy in a short future.

Chapter 3

In Chapter 3, a cautious auto-regressive (AR) one-tap model was designed, that assigns equal probability to all frequencies of the fading. The design procedure may be extended to include also autoregressive moving average (ARMA) models, but some thought is required to correctly include model zeros into the state space model.

Chapter 4

Chapter 4 describes how multi-tap channels in multiuser systems are modelled. It is assumed that model parameters are fixed (that is, given without

uncertainty) by stand-alone estimators. The channel predictor performance is later evaluated given that these model parameters are correct, but the question of how sensitive prediction performance is to the correctness of these parameters is left unanswered.

It would be interesting to conduct a sensitivity study by maintaining the idea of avoiding simulations, and instead try to assess error rates from a model structure that takes *two* state space models into account: one that describes the actual system, and one that describes the system as given by the model parameter estimators.

There is no hope of using optimal model parameter estimators. However, given a set of different estimators and some actual measurements, they may be evaluated against one another by performing model selection of the different state space models that they yield (see Section 2.6). The algorithm for how to do this is described in Appendix F.

The block structures used in Chapter 4 may easily be extended to include multiantenna (MIMO) systems.

Chapter 5

The system evaluation and design procedure is in its infancy and requires a great deal of research and work.

Appendix A

The non-uniqueness of the probability function

The function $\mathrm{P}(\cdot)$, with its sum and product rule, is usually used as definition of probability. Obviously, one might just as well use some other function $\mathrm{Q}(\cdot) = f \circ \mathrm{P}(\cdot)$, where f is a one-to-one mapping. The sum and product rules,

$$\begin{aligned} \mathrm{P}(A|B) + \mathrm{P}(\overline{A}|B) &= 1, \\ \mathrm{P}(A|BI)\mathrm{P}(B|I) &= \mathrm{P}(B|AI)\mathrm{P}(A|I), \end{aligned} \tag{A.1}$$

will then transform into other forms. In this appendix we will investigate how the rules of probability theory may turn out if we use a function $\mathrm{Q}(\cdot)$ rather than $\mathrm{P}(\cdot)$ for measuring what we then might call *qrobability*.

EXAMPLE A.1 **ALTERNATIVE PROBABILITY DEFINITION**

Let $\mathrm{P}(\cdot)$ be the usual function that represent probabilities. We measure the constant x, which we happen to know is either -1 or 1, in some noise e, so that the measured value y is

$$y = x + e. \tag{A.2}$$

To allow for a simple analysis we will let the noise e have a discrete and somewhat constructed density function. It is most easily explained illustrated – see Figure A.1. If we let the proposition E_1='e takes the value $-3/2$', then $\mathrm{P}(E_1|I) = 1/10$ and so on.

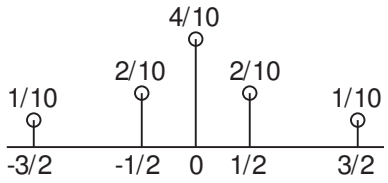


Figure A.1: The probability distribution of the noise e.

The only thing we know about x is that it is either -1 or 1, so the principle of indifference demands us to set

$$\mathrm{P}(X_1|I) = \mathrm{P}(X_2|I) = 1/2, \tag{A.3}$$

where X_1='x is -1' and X_2='x is 1'. The question is: How is our knowledge of x updated when we get the measurement y? Let us say that we measure the specific value 1/2, so that the proposition Y is 'y equals 1/2'. If we choose to concentrate on the proposition X_1 (i.e. that x=1), we may easily update our state of knowledge from $\mathrm{P}(X_1|I)$ to $\mathrm{P}(X_1|YI)$ through the use of Bayes' theorem :

$$\mathrm{P}(X_1|YI) = \mathrm{P}(X_1|I)\frac{\mathrm{P}(Y|X_1I)}{\mathrm{P}(Y|I)}, \tag{A.4}$$

where the factors on the left side may be readily calculated :

$$\begin{aligned} &\mathrm{P}(X_1|I) = 1/2 \\ &\mathrm{P}(Y|X_1I) = \mathrm{P}(\text{`}e\text{=-1/2'}|I) = 2/10 \\ &\mathrm{P}(Y|I) = \mathrm{P}(Y|X_1I)\mathrm{P}(X_1|I) + \mathrm{P}(Y|X_2I)\mathrm{P}(X_2|I) \\ &\qquad = 1/2 \cdot (2/10 + 1/10) = 3/20 \end{aligned} \tag{A.5}$$

Inserting this into Bayes' theorem, we get

$$\mathrm{P}(X_1|YI) = 2/3. \tag{A.6}$$

That is the probability of X_1 being true increases slightly from 1/2 to 2/3 when we get the measurement y=1/2.

What happens if we choose to transform $\mathrm{P}(\cdot)$ with a mapping f so that with instead look at the *qrobability* $\mathrm{Q}(\cdot) = f \circ \mathrm{P}(\cdot)$? Let us make the specific choice

$$\mathrm{Q}(x) = \log \frac{\mathrm{P}(x)}{1 - \mathrm{P}(x)}. \tag{A.7}$$

To simplify notation, we'll use the short-hand forms

$$\begin{array}{ll} \mathrm{P}_n = \mathrm{P}(X_n|I), & \mathrm{P}_{n,y} = \mathrm{P}(X_n|YI) \\ \mathrm{P}_{y,n} = \mathrm{P}(Y|X_nI), & \mathrm{P}_y = \mathrm{P}(Y|I) \end{array} , \tag{A.8}$$

and accordingly, $\mathrm{Q}_n = \log p_n/(1-p_n)$. What we have to do next is to transform the rules. Starting with the sum rule, we get

$$\begin{aligned} (N-1)e^{\mathrm{Q}_1+\ldots+\mathrm{Q}_N} + (N-2)(e^{\mathrm{Q}_2+\ldots+\mathrm{Q}_N} + e^{\mathrm{Q}_1+\mathrm{Q}_3+\ldots+\mathrm{Q}_N} + \ldots) \\ + (N-3)(e^{\mathrm{Q}_3+\ldots+\mathrm{Q}_N} + e^{\mathrm{Q}_2+\mathrm{Q}_4+\ldots+\mathrm{Q}_N} + \ldots) + \ldots + \\ + (e^{\mathrm{Q}_1+\mathrm{Q}_2} + e^{\mathrm{Q}_1+\mathrm{Q}_3} + \ldots) = 1. \end{aligned} \tag{A.9}$$

(The terms with coefficient $(N-k)$ have $(N-k+1)$ terms in the exponent). Evidently, the sum rule becomes quite complicated when we use the transformation. Note however that for the particular case $N=2$, the sum rule looks very attractive :

$$\mathrm{Q}_1 + \mathrm{Q}_2 = 0. \tag{A.10}$$

Although already the general form of the sum rule is fairly involved, things are getting worse. The transformed Bayes' theorem is

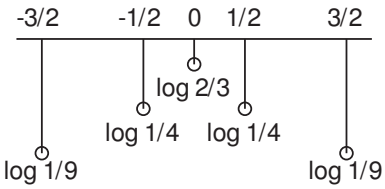
$$\mathrm{Q}_{n,y} = \log\left(\frac{e^{\mathrm{Q}_y}+1}{e^{\mathrm{Q}_y}(e^{\mathrm{Q}_n}+e^{\mathrm{Q}_{y,n}}+1) - e^{\mathrm{Q}_n+\mathrm{Q}_{y,n}}}\right) + \mathrm{Q}_{y,n} + \mathrm{Q}_n, \tag{A.11}$$

and for general number of propositions the transformed marginalisation formula is so complicated that we will restrict ourselves to the case $N=2$:

$$\begin{aligned} \mathrm{Q}_y = \\ \log\Big(\frac{e^{\mathrm{Q}_1+\mathrm{Q}_{y,1}}(1+e^{\mathrm{Q}_2}+e^{\mathrm{Q}_{y,2}})+e^{\mathrm{Q}_2+\mathrm{Q}_{y,2}}(1+e^{\mathrm{Q}_1}+e^{\mathrm{Q}_{y,1}})+2\alpha}{e^{\mathrm{Q}_1}+e^{\mathrm{Q}_{y,1}}+e^{\mathrm{Q}_2}+e^{\mathrm{Q}_{y,2}}+e^{\mathrm{Q}_1+\mathrm{Q}_2}+e^{\mathrm{Q}_1+\mathrm{Q}_{y,2}}+e^{\mathrm{Q}_2+\mathrm{Q}_{y,1}}+e^{\mathrm{Q}_{y,1}+\mathrm{Q}_{y,2}}+1-\alpha}\Big), \\ \text{with } \alpha = e^{\mathrm{Q}_1+\mathrm{Q}_{y,1}+\mathrm{Q}_2+\mathrm{Q}_{y,2}}. \end{aligned}$$

Now we have only to transform our prior knowledge about e and apply the transformed rules:

$$\begin{aligned} &\mathrm{Q}(X_1|I) = \mathrm{Q}(X_2|I) = 0 \\ &\mathrm{Q}(Y|X_1I) = \mathrm{Q}_{y,1} = \log 1/4, \qquad \mathrm{Q}(Y|X_2|I) = \mathrm{Q}_{y,2} = \log 1/9 \\ &\mathrm{Q}(Y|I) = \mathrm{Q}_y = \log 3/17, \quad \text{(by transformed marginalisation)} \\ &\Rightarrow \mathrm{Q}(X_1|YI) = \mathrm{Q}_{1,y} = \log 2 \end{aligned} \tag{A.12}$$



(Of course we could just as well have used the regular rules (A.1) and then transformed the result with (A.7), but that would miss the point). The result is in agreement with the result obtained with the traditional probability measure, since $\log(\frac{2/3}{1-2/3}) = \log 2$. We see that when we measure our state of knowledge using this qrobability, the measurement increases from $\mathrm{Q}(X_1|I) = 0$ to $\mathrm{Q}(X_1|YI) = \log 2$ upon receiving y=1/2. Whether this is little or much is hard to say; the proportional increase is indeed infinite.

There are two points to be made in this example. Firstly, it is evident that the function $\mathrm{P}(\cdot)$ is preferable to $\mathrm{Q}(\cdot)$ from an algebraic point of view. Secondly, it is important to understand that there is nothing fundamentally more *correct* about P as compared to Q. The only thing that Cox´s theorem states is that any measure of belief is *isomorphic* to a probability measure[1]. To me, this is reminiscent of Euclids five postulates of geometry; four of them seem very basic and it is obvious to anyone that they are needed to axiomatise geometry. The fifth postulate – the parallel postulate – however seem unnecessarily complicated and for two and a half millennia people tried to derive it from the other four. It was not until the 19th century that it was proven that the parallel postulate is indeed needed to define standard geometry, and that digressions from it yield new types of geometries.

In just the same way we need the three desiderata of Chapter 2 to find the definition of a measure of belief, but to make it unique we also need a fourth "fifth postulate":

(IV) *When applicable, probabilities and frequencies should take the same value,*

[1] In mathematics, a *probability measure* is in effect a function obeying the rules of probability (Eq. A.1), although it is defined in a somewhat more rigid manner.

or,

(IV) *The rules of inference should be as simple as possible,*

however one would formalise that last requirement. Without it, we do not know whether to use $P(\cdot)$ or any of the infinite possibilities of $Q(\cdot)$. Unlike the case of Euclid's postulates, digression from it will not unveil wonderful new theories which Dutch artists may turn into intricate images of interlocking lizards. Since all rules will transform along with any transformation of $P(\cdot)$, decisions will necessarily be independent on which theory is used. The lesson taught is that caution must be taken when interpreting absolute values of probability.

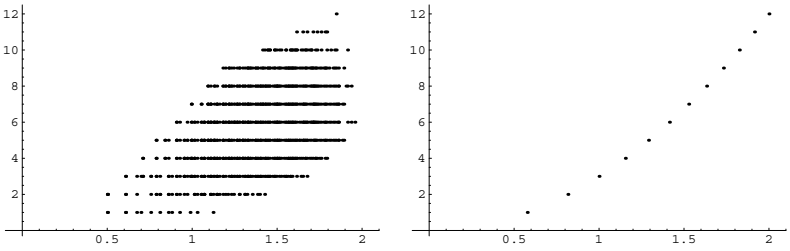
Appendix B

Supersymbols

Here we exemplify the concept of "supersymbols" by examining the use of a $7,4$ block code and a 2D modulation format of eight symbols and studying it from the geometric viewpoint proposed in Chapter 3. A good choice for the code is a so called (7,4) Hamming code. It is "optimal" in the sense that its 16 (2^4) 7-bit patterns (called *codewords*) are spread out evenly, so that the difference in bit pattern between each pair of codewords – the *Hamming distance* – is always 3 or 4. Hamming codes exist for all block sizes ($2^m - 1, 2^m - 1 - m$) where $m \geq 2$. A 2D 8-ary modulation format is gray coded 8PSK. All eight symbols have the same energy and are evenly distributed around a circle in the two-dimensional signal space. The gray coding ensures that adjacent symbols only differ in one bit, which also makes it "optimal" in a sense.

If we view this code/modulation combination as an "extended" modulation of "supersymbols" in 14D-space, then how are the supersymbols distributed? It's difficult to depict a 14-dimensional cluster of points on a two-dimensional sheet of paper, but plotting the Hamming distance versus the signal distance for each signal pair will tell a great deal about the properties of transmission method; the performance of a communication system is largely dependent on how well separated the symbols are and how many bits differ between close symbols. See Figure B.1(a).

Could we come up with a better alternative? We have 4096 symbols to distribute in a 14D space. An obvious choice would be to place the symbols on the corners of a 12D cube (and leave the two remaining dimensions unused). We plot the corresponding curve for this "hypercube" modulation and compare it to the Hamming+8PSK method. See Figure B.1(b). One



(a) The symbol distribution in a system that uses a Hamming (7,4) code in combination with 8PSK modulation. The asymmetry degrades the performance compared to the "hypercube" modulation.

(b) The distribution of symbols when each symbol is located on the corner of a 12D hypercube. The symmetry gives a good performance.

Figure B.1: Hamming distance versus signal distance for Hamming (7,4) encoded 8PSK and the "hypercube" modulation. The hamming distance is measured in bits. The signal distance is measured in units of the supersymbol energy.

might expect – and one would expect correctly – that the hypercube alternative yields better performance, despite that it seems considerably more simple and straightforward when we view the communication system from the geometric level discussed in Section 3.1. This shows how it might be beneficial to abstractly remove elements from the system when evaluating its performance.

Further we may compare the two methods by deriving the bit error rate expression for the hypercube modulation.

First we note that the bit error rate (P_b) versus bit energy/noise power spectral density ratio (E_b/N_0) for Hamming (7,4)+8PSK can be approximated by (See e.g. [12])

$$P_{b,8PSKcoded} \approx \sum_{i=2}^{7} \frac{i+1}{n} \binom{n}{i} p^i (1-p)^{7-i}, \quad p = \frac{2}{3} Q\left(\sqrt{\frac{24}{7}\frac{E_b}{N_0}} \sin\frac{\pi}{8}\right). \tag{B.1}$$

Here, $Q(\cdot)$ is a common function i digital communications and is defined

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt. \tag{B.2}$$

We need it every time our signal is subjected to AWGN, which is more or less always, and we want to evaluate the probability that the output from a correlator is misinterpreted as the wrong symbol, that is the risk that a

received symbol is displaced beyond a *decision boundary* x_b that separates the respective *decision regions* of two symbols. We then need to calculate the area of the tail of a gaussian distribution, which through a simple change of variables can be expressed in terms of the Q-function :

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{x_b}^{\infty} e^{-\frac{(x-x_0)^2}{2\sigma^2}}dx = Q\left(\frac{x_b - x_0}{\sigma}\right). \tag{B.3}$$

The noise variance is commonly described in terms of the single-sided noise spectral density N_0 : $\sigma^2 = N_0/2$. Moreover, if we by d denote twice the signal distance between the undistorted symbol and the decision boundary so that $x_b - x_0 = d/2$, we get the simple formula

$$Q\left(\frac{d}{\sqrt{2N_0}}\right) = \begin{array}{l}\text{The probability that the output of a}\\ \text{correlator is displaced beyond a de-}\\ \text{cision boundary.}\end{array} \tag{B.4}$$

We can use this result to derive the performance of the hypercube modulation. The most sound way to associate bit patterns with symbols is to let symbols that are close to each other represent bit patterns of little difference. Hence we let bit patterns of symbols that share all but one dimension differ in only one bit, and those symbols that differ in two dimension have bit patterns with two differing bits and so on. With respect to any given symbol, there are $\binom{N}{k}$ other symbols that differ in k dimensions. This gives us the bit error rate for the hypercube modulation. The symbol energy is the squared distance from the centre of the cube to a corner. The distance d between two adjacent symbols constitutes the length of one side, so we have

$$\sqrt{E_s} = \sqrt{NE_b} = \frac{d}{2}\sqrt{N} \quad \Rightarrow \quad d = 2\sqrt{E_b}. \tag{B.5}$$

Now we can calculate the bit error rate (note that a detection error where k dimensions come out erroneously generates k/N bit errors). Let $Q = Q(d/\sqrt{2N_0}) = Q(\sqrt{2E_b/N_0})$:

$$\begin{aligned} P_{b,Hypercube} &= \sum_{k=0}^{N}\frac{k}{N}\binom{N}{k}Q^k(1-Q)^{N-k} \\ &= \sum_{k=1}^{N}\frac{k}{N}\binom{N}{k}Q^k(1-Q)^{N-k} \\ &= Q\sum_{k=1}^{N}\binom{N-1}{k-1}Q^{k-1}(1-Q)^{(N-1)-(k-1)} \\ &= Q(\sqrt{2E_b/N_0}) \end{aligned} \tag{B.6}$$

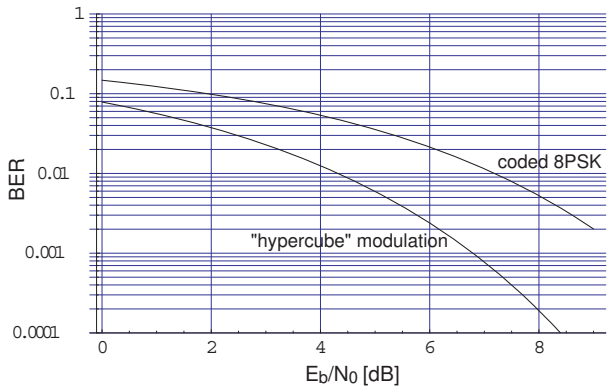


Figure B.2: The bit error rate of the two communication schemes versus E_b/N_0. The "hypercube" modulation, which turns out to be nothing more than a generalisation of regular BPSK or QPSK, outperforms coded 8PSK.

It might seem strange that the bit error rate is the same regardless of the dimensionality N. For those already familiar with digital communications this is however of little surprise; What we have done here is to send our 12 bits, two by two, in consecutive time slots. But this is the same as sending bits in blocks of two using *quadrature phase shift keying* (QPSK), a modulation type for which the bit error performance is the well known result (B.6). We compare the bit error performances of the two modulation types studied here in Figure B.2.

Normally there are more things than merely the bit error probability to consider when one chooses modulation/coding format, such as complexity and bandwidth requirements. The hypercube format suggested in this example will actually require *less* bandwidth than the 8PSK+Hamming combination, but the aim here is only to illustrate the merits of looking at the communications system from a geometrical perspective.

Appendix C

Posterior distribution for the channel power

I here show that the pdf of the squared magnitude of a tap h having a gaussian pdf is the non-central χ^2-distribution.

Denoting the real and imaginary part of a tap by h_r and h_i, respectively, we seek the pdf for the channel power $z = h_r^2 + h_i^2$, conditioned on that we have the mean value $\mu = \mu_r + j\mu_i$ and the variance σ^2 of $h_r + jh_i$. This is easily done by calculating $P(t < z|\mu\sigma I)$ and then taking the derivative with respect to z. Assume that h_r and h_i are uncorrelated and that the variance is split equally between them. Let $f(z) = p(z|\mu\sigma I)$. Then

$$\int_{t<z} f(t)dt = \int_{h_r^2+h_i^2<z} p(h_r|\mu\sigma I)p(h_i|\mu\sigma I)dh_r dh_i$$
$$= \int_{h_r^2+h_i^2<z} \frac{1}{\pi\sigma^2} e^{(-(h_r-\mu_r)^2-(h_i-\mu_i)^2)/\sigma^2} dh_r dh_i.$$

This can be rewritten by use of the law of cosines, which says that the sides in a triangle relates as $a^2 = b^2 + c^2 - 2bc\cos\theta$, where θ is the angle between sides b and c. Changing to polar coordinates, we now have

$$\int_{t<z} f(t)dt = \int_0^{\sqrt{z}} \int_0^{2\pi} \frac{1}{\pi\sigma^2} e^{(-|\mu|^2-r^2+2|\mu|r\cos\theta)/\sigma^2} d\theta r dr.$$

Using $I_0(z) = \frac{1}{\pi}\int_0^\pi \exp(z\cos\theta)d\theta$, where $I_0(z)$ is the modified Bessel func-

tion of the first kind, the above evaluates to

$$\int_{t<z} f(t)dt = \int_0^{\sqrt{z}} e^{-(|\mu|^2+r^2)/\sigma^2} 2I_0(2|\mu|r/\sigma^2)/\sigma^2 r dr.$$

Finally, taking the derivative with respect to z and noting that $d/dz \int^{\sqrt{z}} f(t)dt = f(\sqrt{z})/2\sqrt{z}$, we have

$$p(z|\mu\sigma I) = e^{-(|\mu|^2+z)/\sigma^2} I_0(2|\mu|\sqrt{z}/\sigma^2)/\sigma^2.$$

This is the non-central χ^2-distribution with mean value $|\mu|^2+\sigma^2$ and variance $2|\mu|^2 + 4\sigma^2$. Note that μ and σ are the mean value and standard deviation of the complex gaussian distribution $p(h_r + jh_i|\sigma\mu I)$.

Appendix D

The central limit theorem

Here it is shown that the pdf of the sum of variables, each having pdf $\mathrm{P}(x)$, approaches a gaussian distribution as the number of terms grow large.

Let $x = \sum_{n=1}^{N} x_n/N$. The proof it found by taking the fourier transform:

$$
\begin{aligned}
\mathcal{F}[\mathrm{P}(x)](\omega) &= \int_{-\infty}^{\infty} \mathrm{P}(x) e^{-j\omega x} \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \mathrm{P}(x) \sum_{k=0}^{\infty} \frac{(-j\omega x)^k}{k!} \mathrm{d}x \\
&= \sum_{k=0}^{\infty} \frac{(-j\omega)^k}{k!} \int_{-\infty}^{\infty} x^k \mathrm{P}(x) \mathrm{d}x \\
&= \sum_{k=0}^{\infty} \frac{(-j\omega)^k}{k!} \int_{-\infty}^{\infty} N^{-k} (x_1 + \ldots + x_N)^k \mathrm{P}(x_1) \ldots \mathrm{P}(x_N) \mathrm{d}x_1 \ldots \mathrm{d}x_N \\
&= \int_{-\infty}^{\infty} \sum_{k=0}^{\infty} \left(\frac{-j\omega(x_1 \ldots x_N)}{N} \right)^k \frac{1}{k!} \mathrm{P}(x_1) \ldots \mathrm{P}(x_N) \mathrm{d}x_1 \ldots \mathrm{d}x_N \\
&= \int_{-\infty}^{\infty} e^{-j\omega(x_1 \ldots x_N)/N} \mathrm{P}(x_1) \ldots \mathrm{P}(x_N) \mathrm{d}x_1 \ldots \mathrm{d}x_N \\
&= \left(\int_{-\infty}^{\infty} e^{-j\omega x_1/N} \mathrm{P}(x_1) \mathrm{d}x_1 \right) \times \ldots \times \left(\int_{-\infty}^{\infty} e^{-j\omega x_N/N} \mathrm{P}(x_N) \mathrm{d}x_N \right) \\
&= \left(\int_{-\infty}^{\infty} e^{-j\omega x_1/N} \mathrm{P}(x_1) \mathrm{d}x_1 \right)^N
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\int_{-\infty}^{\infty} \left[1 + \left(\frac{-j\omega}{N} \right) x_1 + \frac{1}{2} \left(\frac{-j\omega}{N} \right)^2 x_1^2 + \ldots \right] \mathrm{P}(x_1) \mathrm{d}x_1 \right)^N \\
&= \left(1 + \frac{-j\omega}{N} \langle x_1 \rangle + \frac{(-j\omega)^2}{2N^2} \langle x_1^2 \rangle + \mathcal{O}(N^{-3}) \right)^N \\
&= \exp \left(N \log \left(1 + \frac{-j\omega}{N} \langle x_1 \rangle + \frac{(-j\omega)^2}{2N^2} \langle x_1^2 \rangle + \mathcal{O}(N^{-3}) \right) \right) \\
&= \exp \left(N \left(\frac{-j\omega}{N} \langle x_1 \rangle + \frac{(-j\omega)^2}{2N^2} \langle x_1^2 \rangle - \frac{(-j\omega)^2}{2N^2} \langle x_1 \rangle^2 + \mathcal{O}(N^{-3}) \right) \right) \\
&= \exp - \left(j\omega \langle x_1 \rangle + \omega^2 \frac{\langle x_1^2 \rangle - \langle x_1 \rangle^2}{2N} + \mathcal{O}(N^{-2}) \right) \\
&= \exp - \left(j\omega\mu + \omega^2 \frac{\sigma_x^2}{2N} + \mathcal{O}(N^{-2}) \right)
\end{aligned}
\tag{D.1}
$$

Neglecting higher order terms and taking the inverse Fourier transform, we arrive at the answer

$$
\begin{aligned}
\Rightarrow \mathrm{P}(x) &\approx \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left(j\omega(x - \mu) + \omega^2 \frac{\sigma_x^2}{2N} \right) \mathrm{d}\omega \\
&= \frac{1}{\sqrt{2\pi\sigma_x^2/N}} \exp \left(-\frac{(x - \mu)^2}{2\sigma_x^2/N} \right)
\end{aligned}
\tag{D.2}
$$

Appendix E

Bayesians, frequentists, and pragmatists

The notations and expressions used in Bayesianism can differ quite a bit from those used in frequentist theory, which may cause confusion when scholars in the two fields try to share research and experience.

But the problems do not stop there, because although the bayesian solution may coinside with a frequentist method, there are many frequentist methods yielding different answers. Moreover, it is often the case that researchers take a pragmatic stand towards a problem and invent new methods unheard of in either bayesian or frequentist litterature.

We will study such a case here. The example is taken from [14]. A fading complex channel tap h_t is modelled with a time invariant linear model. The tap is measured in noise over n samples,

$$\varphi(t) = [h_t + n_t, h_{t-1} + n_{t-1}, \ldots, h_{t-n+1} + n_{t-n+1}]^T. \tag{E.1}$$

A time invariant model is conveniently expressed as the covariance function $r_h(t)$ for the tap, the covariance function $r_\varphi(t)$ for the data, and the cross-covariance $r_{h\varphi}(t)$ between the tap and the measurements. The objective here is to predict the squared magnitude of a future tap, $|h_{t+L|t}|^2$.

It is known that the optimal procedure for making inference for parameters from a linear gaussian model is given by the Kalman recursions. When the model is time invariant, the solution converges towards the Wiener infinite impulse response (Wiener IIR) solution as more and more data is collected. The convergence is in general rapid and the approximation will usually be good after only a few data points. If the Wiener IIR solution yields a short

impulse response, then the Wiener finite impulse response (Wiener FIR) solution will also be close to optimal. Ekman [14] makes this assumption and so he uses the minimum mean square (MMSE) estimate

$$\hat{h}_{t+L|t} = \beta\varphi_t, \tag{E.2}$$

where β is a row vector containing the optimal weights given by the Wiener FIR solution. But we know more than this, because filter theory gives us the entire pdf:

$$p(h_{t+L}|\varphi_t I) = \mathcal{N}(h_{t+L}; \beta\varphi_t, r_h(0) - r_{h\varphi}R_\varphi r_{\varphi h}). \tag{E.3}$$

The solution now follows readily from a change of variables; any statistical textbook tells us that if a parameter x is gaussian with non-zero mean, then $|x|^2$ has a non-central χ^2-distribution. More precisely,

$$p(x|DI) = \mathcal{CN}(x; \mu, \sigma^2) \Rightarrow p(|x|^2|DI) = \chi^2(|x|^2; |\mu|^2 + \sigma^2, 2|\mu|^2 + 4\sigma^2), \tag{E.4}$$

where the first and second arguments are the mean and variance, respectively (a derivation is given in Appendix C). The MMSE estimate of the squared magnitude is simply given by the mean of the χ^2-distribution:

$$\mathrm{E}(|h_{t+L}|^2|\varphi_t I) \approx |\beta\varphi_t|^2 + r_h(0) - r_{h\varphi}R_\varphi r_{\varphi h}. \tag{E.5}$$

The relation is not exact since the Wiener FIR estimate is not optimal. Note that the above solution is not bound to Bayesianism. The process $\{h_t\}$ is considered to be stochastic and so the solution is good also for frequentists.

But frequentism has many solutions to the same problem, and in general they all give different answers. One alternative is to use a so called unbiased estimate. Imagining that we search for an estimate of a parameter θ, we pick an estimator $\hat{\theta}(D)$, more or less out of thin air. Then we make sure that its mean value is equal to the true value θ. For θ to appear at all in our expressions, we must take the mean value over the *sampling distribution* $p(D|\theta I)$. Hence we act as though we already knew the value of θ and make sure that

$$\mathrm{E}(\hat{\theta}(D)|\theta I) = \theta. \tag{E.6}$$

In the present example, one might attempt to simply square the estimate of the tap and use that as a first rough estimator.

$$\hat{h}_{t+L|t,\mathrm{biased}}(\varphi_t) = |\beta\varphi_t|^2 \tag{E.7}$$

The next step is to adjust this so that it becomes unbiased. This is most easily done by simply adding the appropriate constant. To find this constant, we must calculate the expectancy

$$\mathrm{E}(|\beta\varphi_t|^2||h_{t+L}|^2 I), \tag{E.8}$$

and the best way to do this is probably to start by seeking the pdf

$$p(\varphi_t||h_{t+L}|^2 I), \tag{E.9}$$

and then calculating the pdf for the squared linear combination by changing variables. But already calculating (E.9) is exceedingly difficult and a research project on its own. The procedure would be to start by calculating

$$p(\varphi_t|h_{t+L} I), \tag{E.10}$$

which would have to be done by running a Kalman filter backwards in time from the point $t+L$ and setting the initial uncertainty P_0 to zero. The exact procedure can easily be derived by generalising the procedure outlined in Appendix F. Once this is done, we need to marginalise over the unknown phase of h_{t+L}, which probably cannot be done analytically. Finding $\mathrm{E}(|\beta\varphi_t|^2||h_{t+L}|^2 I)$ is indeed a difficult problem.

In any case, the method of unbiased estimators was recognised as flawed a long time ago, the reason for which can be seen by algebraically manipulating the variance of the error over the sampling distribution:

$$\begin{aligned}
&\int (\hat{\theta}(D)-\theta)^2 p(D|\theta I)\mathrm{d}D = \\
&\int \hat{\theta}(D)^2 p(D|\theta I)\mathrm{d}D + \theta^2 - 2\theta\mathrm{E}(\hat{\theta}(D)|DI) = \\
&\Big[\mathrm{E}(\hat{\theta}(D)|DI)\Big]^2 + \theta^2 - 2\theta\mathrm{E}(\hat{\theta}(D)|DI) + \\
&\int \hat{\theta}(D)^2 p(D|\theta I)\mathrm{d}D + \Big[\mathrm{E}(\hat{\theta}(D)|DI)\Big]^2 - 2\Big[\mathrm{E}(\hat{\theta}(D)|DI)\Big]^2 = \\
&\Big(\mathrm{E}(\hat{\theta}(D)|DI)-\theta\Big)^2 + \int \Big[\hat{\theta}(D)-\mathrm{E}(\theta|DI)\Big]^2 p(D|\theta I)\mathrm{d}D = \\
&\Big(\mathrm{E}(\hat{\theta}(D)|DI)-\theta\Big)^2 + \mathrm{var}(\hat{\theta}(D)|DI)
\end{aligned} \tag{E.11}$$

The original motivation for using unbiased estimators was to minimise the variance of the error, but from the above expression we see that although

$\Big(\mathrm{E}(\hat{\theta}(D)|DI) - \theta\Big)^2$ cancels by choosing $\mathrm{E}(\hat{\theta}(D)|DI) = \theta$, it is not evident how $\mathrm{var}(\hat{\theta}(D)|DI)$ is affected by this choice.

Ekman attempts to calculate the unbiased estimate of the squared tap magnitude. However, frequentists unually do not specify the conditions when evaluating mean values, and so Ekman tries to evaluate the "bias compensation" over the prior distribution instead of over the sampling distribution. He finds that

$$\mathrm{E}(|\beta\varphi_t|^2|I) = \beta\mathrm{E}(\varphi_t\varphi_t^*|I)\beta^* = \beta R_\varphi \beta^*. \tag{E.12}$$

To complete the compensation, he calculates the true squared magnitude, again over the prior distribution:

$$\mathrm{E}(|h_t + L|^2|I) = r_h(0) \tag{E.13}$$

The final "unbiased" estimator is now

$$\hat{h}_{t+L|t} = |\beta\varphi_t|^2 + r_h(0) - \beta R_\varphi \beta^*, \tag{E.14}$$

which is equal to the expression (E.5).

How is it possible for a faulty application of a flawed method to give the correct answer? Well, this is in many ways typical for simple problems where all parameters are gaussian and relationships are linear.

Optimal estimators often evaluate to sums of means and variances of model parameters, and so there are not many combinations to choose from if one would pick a blind guess. It is therefore a fairly good chance that suboptimal methods accidentally stumble upon the correct solution.

However, we may easily study cases where ad hoc procedures do not work out that well. One such case is constructed if we simply remove the squaring in the example above, so that we seek to estimate $|h_{t+L}|$ from φ_t.

Analogous to before, a change of variables takes us from the gaussian distribution to a Rice distribution (I drop the subindex $t + L$):

$$p(|h||\varphi I) = \frac{|h|}{\sigma^2}\exp\{-(|h^2| - |\beta\varphi_t|^2)/2\sigma^2\}I_0(|h||\beta\varphi_t|/\sigma^2) \tag{E.15}$$

But the expected value now look considerably more complicated and contains a Laguerre polynomial ($L_{1/2}$ below):

$$\mathrm{E}(|h||\varphi I) = \sigma\sqrt{\pi/2}L_{1/2}(-|\beta\varphi|^2/2\sigma^2), \tag{E.16}$$

This result can be derived as easily as (E.5) straight from schoolbook results, both by Bayesians and frequentists, but hardly through the use of ad-hoc methods.

Appendix F

Model selection for linear models

Given a set of data $Y = y_0, y_1 \ldots y_{N-1}$, model selection amounts to deciding which model M_k from a predefined set of models yields the highest probability. Assuming that all models are assigned the same prior probability, we have

$$p(M_k|Y, I) \propto p(Y|M_k, I), \tag{F.1}$$

and the problem reduces to calculating the likelihood $L(M_k) = p(Y|M_k, I)$.

An important class of models is the linear model structure in which all parameters are gaussian. Such a model can always be expressed in state space form :

$$\begin{aligned} x_{i+1} &= F_i x_i + G_i u_i, \\ y_i &= H_i x_i + v_i. \end{aligned} \tag{F.2}$$

The covariances and mean values are assumed given as prior information:

$$\mathrm{E}(\begin{bmatrix} u_i \\ v_i \\ x_0 \end{bmatrix}, \begin{bmatrix} u_j^* & v_j^* & x_0^* & 1 \end{bmatrix} | I) = \begin{bmatrix} Q_i\delta_{ij} & 0 & 0 & 0 \\ 0 & R_i\delta_{ij} & 0 & 0 \\ 0 & 0 & \Pi_0 & 0 \end{bmatrix}. \tag{F.3}$$

In order to calculate the likelihood, we write

$$\begin{aligned} p(Y|M_k, I) &= p(y_{N-1}, \ldots, y_0|M_k, I) \\ &= p(y_{N-1}|y_{N-2}, \ldots, y_0, M_k, I) \times p(y_{N-2}, \ldots, y_0|M_k, I) \\ &= p(y_{N-1}|y_{N-2}, \ldots, y_0, M_k, I) \times p(y_{N-2}|y_{N-3}, \ldots, y_0, M_k, I) \\ &\quad \times \ldots \times p(y_1|y_0, M_k, I) \times p(y_0|M_k, I), \end{aligned} \tag{F.4}$$

noting that the above is a product of the pdf:s of the one-step predictions for the process y. These pdf:s are given by the Kalman recursions :

$$\begin{aligned} R_{e,i} &= R_i + H_i P_i H_i^* \\ K_{f,i} &= P_i H_i^* R_{e,i}^{-1} \\ e_i &= y_i - H_i \hat{x}_i \\ \hat{x}_{i|i} &= \hat{x}_i + K_{f,i} e_i \\ P_{i|i} &= (I - K_{f,i} H_i) P_i \\ \hat{x}_{i+1} &= F_t \hat{x}_{i|i} \\ P_{i+1} &= F_i P_{i|i} F_i^* + G_i Q_i G_i^* \end{aligned} \tag{F.5}$$

The pdf for the one-step prediction for y is given by

$$p(y_i|y_{i-1}, \ldots, y_0, M_k, I) = \mathcal{CN}(y_i; H_i \hat{x}_i, R_{e,i}). \tag{F.6}$$

Considering the expression for a gaussian distribution, we may hence update the likelihood $L(M_k)$ alongside the Kalman recursions through

$$\log L_{i+1}(M_k) = \log L_i(M_k) - p \log \pi - \frac{1}{2} \log |R_{e,i}| - \frac{1}{2} e_i^* R_{e,i}^{-1} e_i, \tag{F.7}$$

where p is the dimensionality of the measurements y. Comparing different models, one would then choose the model that yields the highest likelihood.

Note that the algorithm requires – apart from the usual recursions – a determinant and an inverse to be calculated at each time update. If the measurement y has a high dimensionality it might be a better idea to update $R_{e,i}^{-1/2}$ instead of $R_{e,i}$. $R_{e,i}^{-1/2}$ here refers a so-called square-root factor[1]. It is non-unique but can be made unique by for example our requiring it to be lower triangular. Hence we define the square root $A^{1/2}$ of a matrix A to be a lower triangular matrix satisfying

$$A = A^{1/2} A^{*/2}, \tag{F.8}$$

where the last factor refers to the conjugate transpose of $A^{1/2}$. The restriction that the square root should be triangular is particularly suited for the present purpose, since its determinant can then be calculated simply by multiplying its diagonal elements. A recursion for updating $R_{e,i}^{-1/2}$ can be constructed through use of an *array algorithm* as follows (it is here assumed

[1] The terminology is a historical quirk. A more correct term would be *Choleski factor*.

that F_i and R_i are invertible). Construct a particular matrix (the left hand side below) and use QR decomposition to form (see [19, Sec. 12.8.5])

$$\begin{pmatrix} R_i^{-*/2} & 0 & 0 \\ -F_i^{-*}H_i^*R_i^{-*/2} & F_i^{-*}P_i^{-*/2} & 0 \\ Q_i^{*/2}G_i^*F_i^{-*}H_i^*R_i^{-*/2} & -Q_i^{*/2}G_i^*F_i^{-*}P_i^{-*/2} & I \\ -y_i^*R_i^{-*/2} & \hat{x}_i^*P_i^{-*/2} & 0 \end{pmatrix}$$

$$= \Theta \begin{pmatrix} R_{e,i}^{-*/2} & -K_{p,i}^*P_{i+1}^{-*/2} & (*) \\ 0 & P_{i+1}^{-*/2} & (*) \\ 0 & 0 & (*) \\ -e_i^*R_{e,i}^{-*/2} & \hat{x}_{i+1}^*P_{i+1}^{-*/2} & (*) \end{pmatrix},$$

where Θ is a unitary matrix satisfying $\Theta\Theta^* = I$. The $(*)$ are elements whose values need not be calculated. Evidently, $R_{e,i}^{-1/2}$ is extracted as the upper left $p \times p$-block of the QR decomposition, and $e_i^*R_{e,i}^{-*/2}$ constitutes the lower left $1 \times p$ vector. Note also that $P_{i+1}^{-*/2}$ is used in the next iteration. In combination with the above QR decomposition, we can now iterate the likelihood:

$$\log L_{i+1}(M_k) = \log L_i(M_k) - p\log\pi + \mathrm{tr}(\log R_{e,i}^{-1/2}) - \frac{1}{2}e_i^*R_{e,i}^{-*/2}R_{e,i}^{-1/2}e_i. \tag{F.9}$$

Keep in mind that QR decomposition is numerically quite cumbersome. Which approach to use is therefore not obvious. However, the array algorithm has an advantage over the usual recursions in terms of numerical robustness.

Appendix G

Numerical complexity

The Kalman filter (KF) recursions requires significant numerical effort. The computations are either dominated by the multiplications of the large matrices F_i and P_i, or by the matrix inversion $R_{e,i}^{-1}$, depending on how difficult the latter is to implement on the available hardware architecture. Below follows a thorough analysis of the number of arithmetic operations needed to iterate the KF.

G.1 Matrix multiplications

The multiplication of two matrices, one of dimension $m \times n$ and the other of dimension $n \times p$, takes n multiplications and $n-1$ additions per element, hence requiring a total of about mnp arithmetic operations (one arithmetic operation being equal to one multiplication *and* one addition). The left-over addition means that in an expression such as $A+BCD$, we need not consider the matrix addition, but need only count the matrix multiplications. In the KF expressions it is also common that a hermitian (conjugate symmetric) matrix is multiplied from the left with a certain matrix, and from the right with the same matrix conjugated, so that we know the total result to also be hermitian. Then we need only carry out half of the arithmetic operations in the last multiplication. Which of the two multiplications to regard as the last naturally depends on which one saves us the most operations.

G.2 Matrix inversions

To most people it is surprising to see that a general matrix inversion only takes n^3 operations, which is the same as it takes to square the same matrix[27, p. 33]. Moreover, inverting a hermitian matrix takes only $n^3/6$. There is however a big difference between matrix multiplications and matrix inversions. Whereas matrix multiplications are tailored for implementation on most DSP:s, which effectively pipeline alternating additions and multiplications, the numerical effort taken by a matrix inversion depends very much on the specific architecture of the system on which the algorithm is to be implemented. In the below numerical complexity calculations, I have therefore chosen to express the complexity of an inversion as a function call, $\text{minv}(n^3)$. I use the argument n^3 instead of n to indicate that the operation is generally an $\mathcal{O}(n^3)$ operation. Assuming the most efficient implementation, one should equate $\text{minv}(n^3)$ with $n^3/6$.

G.3 Complexity of the Kalman filter

Assume a general state space model :

$$\begin{aligned} x_{i+1} &= F_i x_i + G_i u_i, \\ y_i &= H_i x_i + v_i. \end{aligned} \tag{G.1}$$

The covariances and mean values are given as prior information:

$$\text{E}\left(\begin{bmatrix} u_i \\ v_i \\ x_0 \end{bmatrix}, \begin{bmatrix} u_j^* & v_j^* & x_0^* & 1 \end{bmatrix} | I \right) = \begin{bmatrix} Q_i\delta_{ij} & 0 & 0 & 0 \\ 0 & R_i\delta_{ij} & 0 & 0 \\ 0 & 0 & \Pi_0 & 0 \end{bmatrix}. \tag{G.2}$$

Let n be the dimensionality of the states x, let m be the dimensionality of the process noise u, and let p be the dimensionality of the measurements y. We then have the following matrix dimensions:

$$F, P : n \times n, \quad G : n \times m, \quad H : p \times n, \quad Q : m \times m, \quad R, R_e : p \times p, \quad K_f : n \times p \tag{G.3}$$

The KF recursions, needed for updating the filtered pdf $\{\hat{x}_{i|i}, P_{i|i}\}$ or the one-step prediction pdf $\{\hat{x}_i, P_i\}$, and their corresponding numerical complexity, are given in Table G.5. The total number of arithmetic operations, one operation being one addition and one multiplication, sums to

$$3n^3/2 + (3p/2 + m/2)n^2 + (3p^2/2 + m^2)n + \text{minv}(p^3), \tag{G.4}$$

in agreement with [19, p. 475] and close agreement with [28][1]. To be

[1] Verhaegen *et al*[28] has an extra $3pn^2/2$ term.

consisted with the references, I have in the above expression neglected the "quadratic" terms in the $\{\hat{x}_{i|i}, \hat{x}_{i+1}\}$ update.

The model matrices presented in this thesis are block structured so that only K elements per row/column are non-zero. This gives rise to a tremendous relaxation of the numerical complexity, which is also presented in Table G.5. The total count is $(3K/2 + p/2)n^2 + (pK + p^2)n + p^2K/2 + \text{minv}(p^3)$ arithmetic operations. I have then also assumed that the matrix product $G_i Q_i G_i^*$ has been precomputed.

G.4 Complexity of alternativ KF formulations

Apart from the block-structures of the model matrices, the channel estimation problem considered in this thesis has other structural properties that may reduce the numerical complexity. So, for example, will we observe that the fading statistics induced from a mobile unit's moving about in an urban or suburban environment, will be constant for long time periods, hence imposing static model matrices for large periods of time.

By restricting the matrices to be constant, the KF can be formulated in a way which reduces its complexity considerably. This formulation is called the CKMS algorithm after its developers Chandrasekhar, Kailath, Morf, and Sidhu. Instead of propagating $P_{i|i}$ and P_i directly, the CKMS formulation propagates four smaller matrices K_i, L_i, $R_{e,i}$, and $R_{r,i}$. The matrices L_i and $R_{r,i}$ are initiated at $i = 0$ by a spectral factorisation

$$-L_0 R_{r,0}^{-1} L_0^* = F\Pi_0 F^* + GQG^* - K_0 R_{e,0}^{-1} K_0^* - \Pi_0, \tag{G.5}$$

with $K_0 = F\Pi_0 H^*$ and $R_{e,0} = R + H\Pi_0 H^*$. The spectral factorisation yields a matrix L_0 of size $n \times \alpha$, and a matrix $R_{r,0}$ of size $\alpha \times \alpha$. The lower the α, the more beneficial the use of CKMS as compared to the standard KF formulation. The exact CKMS recursions and their corresponding numerical complexity is presented in Table G.5. Unfortunately, as is clear from the table, the CKMS recursions do not benefit very much from the block-structures of the model matrices. The reason is that the iterated matrices immediately become full which means that no savings can be made in the multiplications. Therefore, in the specific application considered here, the CKMS recursions are actually *more* computationally demanding than the original KF formulation.

The Kalman filter can also be implemented as an *array algorithm*. One then updates so called *square-root factors* of $P_{i|i}$ or P_i. Array algorithms can be implemented in such a way that they take virtually the same number of

arithmetic operation to iterate as the standard KF. However, they too suffer from not being able to exploit the block-structures of the matrices presented in this thesis. One array algorithm is briefly presented in Appendix F. For more information on array algoritms, see [19].

G.5 KF complexity vs. number of taps and users

So far I have only considered the updating of the P matrix. In order to study the complexity of a real filter we must also consider the updating of the mean value $\hat{x}$ as well as the complexity of the many-steps predictions needed for the channel state feedback. The complexity of the latter is determined by the number of steps to predict (the prediction horizon). It is carried out by simply iterating the recursions for $\hat{x}_{i+1}$ and P_{i+1}. Taking into account the block-structured model (4.2) where K represents the block size, the complexity budget required for one KF iteration hence looks like this:

$$\begin{array}{ll} (3/2+p/2)n^2+(pK+p^2)n+p^2K/2+\mathrm{minv}(p^3) & P \text{ update} \\ (p+K)n+pK & \hat{x} \text{ update} \\ 3n^2/2\times \#\text{ prediction steps} & P \text{ prediktion} \\ n\times \#\text{ prediction steps} & \hat{x} \text{ prediction} \end{array} \tag{G.6}$$

This deviates immensely from the effort needed to iterate the general Kalman filter. Due to the difficulties in exploiting the block structures in alternative KF formulations, it is not very likely that this complexity can be reduced further to any considerable extent.

Table G.1: Numerical complexity of the general Kalman filter and the corresponding relaxations emerging from block-structured matrices. The term unchanged refers to the fact that the expression in question is the same as for the general case.

Operation	# arithmetic ops. (general case)	# arithmetic ops. (block structures)
$A_i = H_i P_i$	pn^2	pnK
$R_{e,i} = R_i + H_i P_i H_i^* = R_i + A_i H_i^*$	$\frac{1}{2}p^2 n$	$\frac{1}{2}p^2 K$
$K_{f,i} = P_i H_i^* R_{e,i}^{-1} = A_i^* R_{e,i}^{-1}$	$p^2 n + \text{minv}(p^3)$	*unchanged*
$P_{i\|i} = (I - K_{f,i} H_i) P_i = P_i - A_i R_{e,i}^{-1} A_i^* = P_i - K_{f,i} A_i^*$	$\frac{1}{2}pn^2$	*unchanged*
$P_{i+1} = F_i P_{i\|i} F_i^* + G_i Q_i G_i^*$	$\frac{3}{2}n^3 + \frac{1}{2}mn^2 + m^2 n$	$\frac{3}{2}n^2 K$ or $\frac{3}{2}n^2$
$\hat{x}_{i\|i} = \hat{x}_i + K_{f,i}(y_i - H_i \hat{x}_i)$	$pn + pn$	$pK + pn$
$\hat{x}_{i+1} = F_i \hat{x}_{i\|i}$	n^2	nK or n

Table G.2: Numerical complexity of the general CKMS algorithm and the corresponding relaxations emerging from block-structured matrices. The term unchanged refers to the fact that the expression in question is the same as for the general case.

Operation	# arithmetic ops. (general case)	# arithmetic ops. (block structures)
$U_i = FL_i$	αn^2	αnK
$V_i = HL_i$	$p\alpha n$	$p\alpha K$
$K_{p,i} = K_i R_{e,i}^{-1}$	$p^2 n + \text{minv}(p^3)$	*unchanged*
$K_{i+1} = K_i - U_i R_{e,i}^{-1} V_i^*$	$p^2 n + \frac{1}{2} p n^2$	*unchanged*
$L_{i+1} = U_i - K_{p,i} V_i$	$p\alpha n$	*unchanged*
$R_{e,i+1} = R_{e,i} - V_i^* R_{r,i}^{-1} V_i^*$	$p\alpha + \frac{1}{2} p^2 \alpha + \text{minv}(\alpha^3)$	*unchanged*
$R_{r,i+1} = R_{r,i} - V_i^* R_{e,i}^{-1} V_i$	$p\alpha^2 + \frac{1}{2} p^2 \alpha$	*unchanged*
$P_{i+1} = P_i - L_i R_{r,i}^{-1} L_i^*$	$\frac{1}{2}\alpha n^2 + \alpha^2 n$	*unchanged*

Bibliography

[1] E. T. Jaynes, *Probability Theory: the Logic of Science*. Cambridge University Press, 2003.

[2] W. Feller, *An Introduction to Probability Theory and its Applications*. John Wiley and Sons, 3 ed., 1968.

[3] R. A. Fisher, *Statistical Methods and Scientific Inference*. Hafner Press, 3 ed., 1973.

[4] B. de Finetti, *Probability,Induction and Statistics*. John Wiley and Sons, 1972.

[5] H. Jeffreys, *Theory of Probability*. Oxford University Press, 3 ed., 1961.

[6] G. Pólya, *Mathematics and Plausible Reasoning*, vol. 2. Princeton university press, 2 ed., 1968.

[7] G. Pólya, *Mathematics and Plausible Reasoning*, vol. 2. Princeton university press, 2 ed., 1968.

[8] R. T. Cox, "Probability, frequency, and reasonable expectation," *American Journal of Physics*, 1946.

[9] P. S. Laplace, *Théorie Analytique des Probabilités*. Courcier Imprimeur, 1812.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, 1948.

[11] E. T. Jaynes, "Confidence intervals vs. bayesian intervals," *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, 1976.

[12] J. G. Proakis, *Digital Communications*. McGraw-Hill, 4 ed., 2001.

[13] R. Vaughan and J. B. Andersen, *Channels, Propagation and Antanneas for Mobile Communications*. The Institution of Electrical Engineers, 2003.

[14] T. Ekman, *Prediction of Mobile Radio Channels – Modeling and Design*. PhD thesis, Uppsala University, 2002.

[15] E. Björnemo, "Frequency synchronisation in ofdm – a bayesian analysis," in *IEEE Vehicular Technology Conference 2005-Spring, Stockholm, Sweden*, 2005.

[16] L. Svensson, *Bayesian Inference with Unknown Noise Covariance*. PhD thesis, Chalmers University of Technology, 2004.

[17] Wikipedia, "Estimation of covariance matrices — wikipedia, the free encyclopedia," 2007. [Online; accessed 20-May-2007].

[18] T. Soderstrom, *Discrete-time Stochastic Systems*. Springer, 2 ed., 2002.

[19] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[20] H. Ekström, A. Furuskär, J. K. ans M. Meyer, S. Parkvall, J. Torsner, and M. Wahlquist, "Technical solutions for the 3g long-term evolution,"

[21] IST WINNER and WINNER II projects, partly funded by the European Commission. Available on https://ist-winner.org.

[22] IST-4-027756 WINNER II, "D6.13.7: Test Scenarios and Calibration Cases; Issue 2," Dec. 2006. Available on https://ist-winner.org.

[23] M. Sternad, T. Ekman, and A. Ahlen, "Power prediction on broadband channels," in *IEEE Vehicular Technology Conference 2001-Spring, Rhodes, Greece*, 2001.

[24] T. Ekman, M. Sternad, and A. Ahlen, "Unbiased power prediction of rayleigh fading channels," in *IEEE Vehicular Technology Conference 2002-Fall, Vancouver, Canada*, 2002.

[25] A. Duel-Hallen, S. Hu, and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Processing Magazine*, vol. 17, pp. 62–75, May 2000.

[26] S. Parl, "A new method for calculation the generalized q function," *IEEE Transactions on Information Theory*, vol. 26, pp. 121–124, 1980.

[27] G. Strang, *Linear Algebra and Its Applications*. Academic Press, Inc., 2 ed., 1976.

[28] M. Verhaegen and P. V. Dooren, "Numerical aspects of different kalman filter implementations," *IEEE Transactions on Automatic Control*, 1986.

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 1991.

[30] S. Falahati, A. Svensson, T. Ekman, and M. Sternad, "Adaptive modulation systems for predicted wireless channels," *IEEE Transactions on Communications*, 2004.

[31] R. Penrose, *The Road to Reality*. Jonathan Cape, 2004.