Uppsala University
Signals and Systems

RESOURCE ALLOCATION UNDER UNCERTAINTY

Applications in Mobile Communications

Mathias Johansson

UPPSALA UNIVERSITY 2004

Dissertation for the degree of Doctor of Philosophy
in Signal Processing at Uppsala University, 2004.

ABSTRACT
Johansson, M., 2004. Resource Allocation under Uncertainty – Applications in Mobile Communications, 221 pp. Uppsala. ISBN 91-506-1770-2.

This thesis is concerned with scheduling the use of resources, or allocating resources, so as to meet future demands for the entities produced by the resources. We consider applications in mobile communications such as scheduling users' transmissions so that the amount of transmitted information is maximized, and scenarios in the manufacturing industry where the task is to distribute work among production units so as to minimize the number of missed orders.
The allocation decisions are complicated by a lack of information concerning the future demand and possibly also about the capacities of the available resources. We therefore resort to using probability theory and the maximum entropy principle as a means for making rational decisions under uncertainty.
By using probabilities interpreted as a reasonable degree of belief, we find optimum decision rules for the manufacturing problem, bidding under uncertainty in a certain type of auctions, scheduling users in communications with uncertain channel qualities and uncertain arrival rates, quantization of channel information, partitioning bandwidth between interfering and non-interfering areas in cellular networks, hand-overs and admission control. Moreover, a new method for making optimum approximate Bayesian inference is introduced.
We further discuss reasonable optimization criteria for the mentioned applications, and provide an introduction to the topic of probability theory as an extension to two-valued logic. It is argued that this view unifies a wide range of resource-allocation problems, and we discuss various directions for further research.

Keywords: resource allocation, uncertainty, probability theory as logic, scheduling, multiuser diversity, Jaynes, maximum entropy, Bayesian probability theory.

Mathias Johansson, Signals and Systems, Uppsala University, PO Box 528, SE-751 20 Uppsala, Sweden. E-mail: mathias.johansson@signal.uu.se.

This thesis has been prepared using LaTeX.

ISBN 91-506-1770-2

Printed in Sweden by Elanders Infologistics Väst AB, Göteborg, September 2004.
Distributed by Signals and Systems, Department of Engineering Sciences,
Uppsala University, Uppsala, Sweden.

Contents

Mät aldrig bergets höjd
förrän du nått toppen.
Då ska du se
hur lågt det var.

Dag Hammarskjöld

Preface

As a basketball player, I was taught that everything was about mastering the fundamentals – how to move with and without the ball, how to position yourself in offense and defense, and how to handle the ball. Bill Walton, one of the all-time great basketball players, stressed that the difference between the professional players and the rest of us were how they practiced and focused on the fundamentals. There are players who can match the artistic moves of the top athletes in the game, but unless they command the fundamentals equally well, they would not stand a chance in a real competition. From solid fundamental skills, all aspects of the game follow. That is why the top players continue to practice the basic skills, these simple movements and techniques that constitute the foundation of the game. The difference between the best player and the second best lies in their fundamental skills.

As a student, I was not taught any similar fundamentals of science. In science, there seemed to exist only a vague picture of what the fundamentals were. In the beginning of my Ph.D. student days, all I could see was a vast number of different tools for various purposes, but no underlying unifying principle. Any new problem seemed to call for a new approach. The tools were seemingly picked at random. How would I ever be able to understand all these completely different ideas? I felt that the journey to a Ph.D. thesis was endless and perhaps simply too difficult for me.

By coincidence, while taking a course in information theory and browsing the Internet for some material, I came across an unfinished manuscript for a book entitled 'Probability Theory – The Logic of Science'. It was written by an American physicist, Edwin T. Jaynes. His style of writing was quite different from all other textbooks I had read. Writing in a friendly tone, he focused on the fundamentals of science, and showed that a few very simple rules were really all that was needed for conducting scientific inference. Starting from three simple 'desiderata' describing an ideal objective reasoner he constructed a theory for optimal reasoning under uncertainty. Although the theory used the same basic building blocks as conventional probability theory, the underpinnings were completely different and resulted in a

completely general method for inference. Gradually becoming more adept at the fundamentals that Jaynes stresses – much like a Bill Walton of science – has made it easier for me to understand the various tools that I had been confronted with as separate topics earlier. Nowadays, I find that these results are typically easily derived from the basic rules in just a few lines of maths. From then on, I have stuck to this view of science, and it has shaped my way of thinking about the world, not just in a scientific context.

Although his view unifies and simplifies science, most scientists have no acquaintance with Jaynes' approach. Compared to conventional probability theory, Jaynes' theory is a different paradigm altogether and many times I have found it difficult to communicate my work to others, as the word 'probability' has a very different meaning for them. In conventional probability theory, the technical term 'probability' does not correspond at all to what we mean colloquially by a probability as describing a reasonable degree of confidence in something. It is much like speaking different languages but using the same words. Therefore, in this thesis I take the opportunity to give a comprehensive introduction to probability theory as 'the logic of science'. I hope that this will at least reduce the number of misinterpretations concerning the later chapters.

Today, to my great pleasure, we are an increasing group who adopts the view of probability theory as logic at the Signals and Systems group. I hope that the group will continue its meetings and I am excited about the possibilities that the group have in forming a strong team in this emerging research area.

I have had the great luxury of having intellectually curious and adept supervisors in Professor Mikael Sternad and Professor Anders Ahlén, who have managed to give me constructive advice and criticisms even in an area where they had little prior experience. I thank you especially for allowing me to go on into these uncharted waters. Your emphasis on making *relevant* research and your high standards have served as a strong inspiration for me.

In addition to my two supervisors, a number of people has meant much for me during my time as a Ph.D. student at Magistern. The unofficial 'Thursday club' meetings spent at student nations include many memorable moments. My strongest memories from these evenings concern train movies, strict altruism, and the Cliff Barnes-versus-ice cream episode. A further special thanks goes to Mattias Wennström who has provided guidance during my Ph.D. endeavors and who is a good friend with a great sense of humor. More than eight years ago, I met Jonas Ruström. Since then, we have written a joint Master's thesis and have been fellow Ph.D. students for quite some years. Jonas deserves a warm acknowledgement for these years. It is a tribute to his positive attitude and easy-going nature that we have remained good friends over such a long joint venture which includes being part of founding a company. Thanks also to Erik Björnemo and Daniel Aronsson

who have contributed greatly to filling the gap left by Mattias Wennström when he left Magistern for industry. All the Ph.D. students at Magistern are greatly acknowledged for providing such a nice atmosphere.

This thesis work has been partly financed by PCC++ and Vinnova (The Swedish Agency for Innovation Systems), which is greatly appreciated.

Finally, and most importantly, my family – my mom and dad, my brother and my sister – has always supported me in everything. I owe a lot to you. Thankyou!

Mathias Johansson
Uppsala, August 2004.

Chapter 1

Introduction

IN this thesis we consider a number of problems with the common feature that they all require decisions on how to allocate resources among different tasks under uncertainty concerning the demand and potentially also the supply of resources.

We first study a model problem from the manufacturing industry in which a plant manager has a number of production units which are used to produce different sorts of widgets. The manager's aim is to meet the order intake, but the task is complicated by uncertainty concerning the future order intakes as well as possibly uncertain production capacities. We then consider the customer perspective in an auctioning situation. With only limited information concerning other customers' bids, what amount should an individual customer bid? The answer obviously depends on what the expected benefit of the customer will be from winning, and we therefore investigate a few different scenarios.

Based on the general ideas that we formulate in connection to these two problems, we then consider a number of specific problems which are of current interest in digital mobile cellular communications. The objective is to increase the resource efficiency, or to maximize the useful work performed by the resources, over a given time horizon and thereby achieve a more cost-efficient cellular network.

1.1 Probability Theory and Plausible Reasoning

A common problem in deciding on a satisfactory allocation of the available resources is that the actual future outcomes of our decisions are hard to predict in advance. At the time of the decision, the information at hand is too vague to uniquely determine a guaranteed best decision. Therefore, the fundamental tool that we will rely on throughout this thesis is probability calculus in its most gen-

eral formulation as a theory for optimal plausible reasoning. Our use of probability theory is quite different from the collection of methods taught at most schools and universities known as the frequentist interpretation and associated with the names of Venn, Fisher, Neyman, E. Pearson, and Feller, and instead follows in the steps of such names as Laplace, Jeffreys, Cox, and Jaynes. Our approach, championed by Jaynes (2003), is based on the interpretation that probabilities are the fundamental carriers of incomplete information, and describe a reasonable degree of belief that is, or should be, in the mind of an idealized objective and completely rational reasoner. It may come as a surprise to many that the ordinary rules of probability theory are uniquely determined as the only consistent rules for optimal information processing under uncertainty (deductive reasoning being a special case thereof), a result essentially due to Cox (1946) and further refined by Jaynes.

The frequency interpretation of probabilities maintains that a probability is a property of an idealized imagined 'random experiment', and is only a special case of the more general definition as a reasonable degree of belief. The interpretation of probability theory as an extension to logic dramatically affects the scientific method, and it can rightly be described as a new scientific paradigm in the sense of Kuhn (1970) (see also Chalmers, 1999). It must however be emphasized that under the Bayesian umbrella of interpretations, some maintain a quite different position than ours, claiming that probabilities are (1) still interpreted as frequencies in imagined random experiments, or (2) entirely subjective in the sense of arbitrariness. In the framework derived by Jaynes, probabilities are subjective in the sense that they depend on the information at hand and are not objectively verifiable in nature, as they are not properties of nature but of our actual information, and lack thereof, about nature. On the other hand, they are completely objective in the sense that for a given state of information, there is in principle only one correct probability assignment that corresponds to that information state. Failure from seeing this has resulted in a significant amount of misdirected criticisms towards all 'Bayesian' ideas (this is for instance manifest in the aforementioned work by Chalmers, 1999), and we therefore use the term 'probability theory as logic' rather than 'Bayesian probability theory' in this thesis in order to emphasize this distinction. In Chapter 2 we provide a comprehensive introduction to the subject intended for a reader with no previous acquaintance with Jaynes' ideas. Some of the lengthier mathematical derivations are left out but all major results and principles behind them are provided.

1.2 Some Basic Terminology in Cellular Communications

Here we present a minimum of terminology that may assist a reader unacquainted with mobile communications. Some additional more detailed techniques will be briefly explained in the next section, but the interested reader is referred to textbooks for more information.

Current large-area mobile radio networks are typically geographically split in a number of smaller areas called *cells*, each cell being served by one *base station* which all mobile terminals are directly connected to. Each mobile terminal thus transmits to and receives from a base station only, and the base station relays the message to destinations outside the cell via a *core network*. A cell, which often is depicted as a hexagon with the base station in the middle, can also be further divided into typically three or six *sectors* by the use of directional antennas. That way, more users can be supported in the cell area.

We distinguish between the transmission from base station to the mobile terminal and the transmission in the opposite direction, and denote the former by the *downlink channel* and the latter by the *uplink channel*.

Before connecting to the network, the base station performs *admission control*, deciding whether the user may connect or not based on the load of the network and on the propagation conditions as measured by the mobile terminal. When a mobile terminal moves from one cell to another, the network must further make a *hand-over* which means that a new base station takes over communication with that user.

1.3 Resource Allocation in Mobile Communications – Towards More Efficient Networks

Mobile radio networks, such as GSM and the third generation cellular system UMTS, are designed to provide coverage over large areas and for mobile terminals that may move at very high speeds. These two tasks are challenging engineering problems. Due to movements, reflections and multipath propagation, the received signal is a distorted and attenuated version of the transmitted signal. Several techniques are therefore required in order to decode the sent message. Already in the early history of wireless communications it was realized (Nyquist, 1928) that the signalling speed could be increased at the expense of increasing also the bandwidth of the channel, i.e. the width of the spectral contents of the transmitted signal. Shannon (1948) then established limits on the information rate for noise-free as well as noisy channels. He showed that even in the case of noisy channels, error-free reception is possible as long as the data rate does not surpass a certain

number, the capacity of the channel. The channel capacity of a channel limited in bandwidth and disturbed by additive thermal noise was seen to be proportional to the bandwidth of the channel and approximately proportional to the logarithm of the signal-to-noise ratio (SNR) at the receiver. Thus, two ways of increasing the performance of a communication link is to increase the power leveraged to the receiver and to increase the bandwidth of the channel. The latter is perhaps the simpler way, as bandwidth is in some sense an unlimited natural resource. In practice, however, increasing the bandwidth makes linear amplifier design a challenge, and moreover bandwidth usage is regulated by government agencies limiting the allowed spectrum usage. Similarly, increasing transmitter power or using more advanced antenna concepts such as beamforming to increase the amount of power delivered to the receiver increases the costs of the network. In addition to this, there are concerns that the power radiated in the microwave frequency bands used for mobile communications may have adverse effects on human health. For these reasons, increasing the transmitter power is not an attractive option.

Instead, it becomes important to increase the spectral efficiency and the power efficiency, i.e. transmitting more data per Hertz and Watt, and coming closer to Shannon's limit. Network design has become a problem of optimal resource allocation. For instance, how should the bandwidth be partitioned between users and areas to best utilize the limited spectrum available for the network? And how should we distribute power among users to maximize the system throughput?

It is not until recently that it has been realized that in order to maximize the system throughput it is not sufficient to simply use techniques which improve the performance of individual links between the transmitter and a receiver. A strategy which improves the single-user capacity of a link may actually decrease the capacity of the whole network. This makes the design task even more challenging. The designer must now consider the problem of jointly maximizing the throughput of all users. For instance, in a single-user scenario, the channel capacity increases when the number of transmit antennas increases when open-loop spatial diversity is employed, but in a multiuser scenario this strategy decreases the capacity (Jiang et al., 2004)!

In order not to over-generalize results such as these it is important to understand the assumptions imposed on the considered communications system in obtaining these results. The recent interest in multiuser optimizations was sparked by a result due to Knopp and Humblet (1995). They considered the Shannon capacity of a fading *Gaussian multiple-access channel*, a channel where several sources are sending independent information to one common receiver and where the transmitted message from each source may be attenuated by an individual factor which the receiver can measure perfectly. The received messages are also distorted by a common additive white Gaussian noise term. The received signal can thus be modelled

by

$$y = \sum_{u=1}^{U} \alpha_u x_u + n \tag{1.1}$$

where U is the number of sources (or users), x_u is the transmitted message from user u and n is the additive disturbance. Note that this model does not include multipath propagation where messages at different transmission times from each source would arrive simultaneously at the receiver. When the α_u are non-zero constants the capacity of the Gaussian multiple-access channel is (Cover and Thomas, 1991)

$$C = \frac{1}{2}\log_2\left(1 + \frac{\sum_{u=1}^{U} \mu_u(\underline{P})P_u}{N}\right), \tag{1.2}$$

where $\mu_u(\underline{P})$ is the normalized ($0 \leq \mu_u(\underline{P}) \leq 1$) allocated transmitter power for user u and N is the noise power at the receiver. $\underline{P_u}$ is the set of signal powers P_u for each user's message at the receiver would u transmit with full power. This capacity is often denoted the sum-of-rates capacity since it denotes the maximum achievable sum of rates from all users. Under the assumption that the channel attenuations α_u vary randomly over time according to some frequency distribution, the sum-of-rates capacity is obtained by averaging (1.2) over that distribution. This intends to model a mobile radio channel, where the received signal strength varies due to the changing environment. Knopp and Humblet showed that the sum-of-rates capacity averaged over any probability distribution for P_u is maximized by transmitting at any time only to the user with maximum instantaneous SNR when there is a constraint on each user's average transmit power. They further showed that the optimal power control law under the same average power constraint is to use a form of water-filling over time, i.e. increasing the transmit power when the SNR is high and decreasing when it is low. We should however keep in mind that the capacity-optimal strategy is highly dependent on the type of power constraint that is employed.

Following Knopp and Humblet (1995), Tse (1997) considered capacity-optimal power control over a set of parallel *Gaussian broadcast channels* under an average power constraint. A broadcast channel describes a situation where one transmitter wants to send independent information to several receivers. For instance, the considered scenario can model a downlink in a cellular system. The received signal at user u is given by

$$y_u = x + n_u \tag{1.3}$$

where x is the transmitted message and n_u is Gaussian receiver noise. Notice that the disturbances may have different power among users, and that if there is any attenuation of the transmitted signal each user is assumed to measure it perfectly.

Again the optimal power allocation turns out to consist of transmitting in each parallel channel only to the user experiencing the most favorable channel conditions using water-filling across the different channels.

As Knopp and Humblet pointed out, since the capacity-optimal strategy (under the average power constraint) is to transmit only to the user with the highest SNR at any time and that the capacity increases with that maximum instantaneous SNR, the capacity increases with the number of users experiencing channel variability. The stronger the variations (around some given mean) and the more the users, the higher the possible gain from scheduling. Therefore, since this effect is inherent in a multiuser environment, they coined the name *multiuser diversity*. They also noted in a numerical example where each user experienced Rayleigh fading independently of other users that utilizing multiuser diversity is equivalent in terms of average error probability to a system employing selection diversity[1] with equally many branches as there are users in the system. Thus, multiuser diversity can be considered as selection diversity on the transmitting end.

In order to actually realize the potential gains promised by these information-theoretic results, some technique for actually changing the rate according to channel quality must be employed in the network. The use of scheduling and *adaptive modulation* is one such method that has been shown to facilitate considerable throughput gains in the downlinks of cellular systems (see e.g. Chuang and Sollenberger, 2000, Li et al., 2002, Wang et al., 2003a). Adaptive modulation is carried out by predicting the channel conditions (typically the SNR) of the receiving user for the coming time slot, and choosing a modulation level[2] based on this which matches the bit-error rate (BER) requirements of the user. In many cases, a reasonable model for the rate-SNR-BER relation useful for most modulation formats is

$$r \propto \log_2\left(1 + \frac{\gamma}{\Gamma(\mathrm{BER})}\right) , \tag{1.4}$$

where r is the rate in bits per symbol, γ denotes the SNR at the receiver, and $\Gamma(\mathrm{BER})$ is the 'gap' – as a function of the BER – between the modulation (and any additional coding) technique and the Shannon capacity for a bandlimited Gaussian channel.

[1] Selection diversity is the technique where there are, say L, parallel channels to one user, each channel conveying the same message, and the receiver selects only the best channel in decoding the message.

[2] A modulation level [bits/transmitted symbol] determines the signalling speed of the communication. A higher modulation level implies a higher data rate.

1.4 Outline and Contributions of this Thesis

One of the aims of this thesis is to provide a comprehensive introduction to the subject of resource allocation under uncertainty using probability theory as logic. Consequently, we first give a concise but self-contained treatment of probability theory as logic in Chapter 2. This chapter should not be skipped if the reader is unfamiliar with the book by Jaynes (2003). In the following two chapters we then study two model problems in resource allocation, applying the framework in Chapter 2 and providing a methodology for resource allocation problems. Chapter 3 considers a manufacturing plant producing different types of widgets and investigates the problem of allocating production resources so as to meet future customer demands for the different widget types. In this chapter the basic reasoning format and many technical results are derived that form the basis for the remaining chapters. Chapter 4 discusses the customer perspective in a certain type of auctions and addresses the problem of optimal bidding under uncertainty.

Based on the results and ideas in Chapters 2-4 we then devote the remaining chapters to more specific resource allocation problems in mobile communications. We now give a short overview of the contributions of each chapter in the thesis.

Chapter 2

This chapter provides an introduction to probability theory as an extension to logic. We recapitulate the three underlying desiderata which yield the ordinary sum and product rules of probability theory as a uniquely determined consistent framework for plausible reasoning under uncertainty. We further emphasize two basic tools of probability theory, Bayes' rule and marginalization of nuisance parameters. We provide a thorough treatment of the maximum entropy principle as an essential rule for assigning probabilities and discuss its most important properties. In addition to this, we briefly discuss the Shannon capacity and the related concept of information. Before concluding the chapter with some comments on the history of the subject, we give an outline of decision theory from a Bayesian standpoint. The chapter is intended as an introduction rather than an overview and presents no new technical results but explains the most important conceptual and technical sides of the subject in some depth.

Chapter 3

Here, a problem facing a manager of a manufacturing plant is considered. The task is to assign different jobs to different production units so as to minimize the expected number of missed orders. Solutions are given for a number of scenar-

ios, differing in the information available to the manager. The chapter extends an early contribution by Jaynes (1963b) which was the first application of probability theory as logic to resource allocation problems. The chapter is based on presently unpublished material, but some derivations are found in the following works.

- M. Johansson and M. Sternad, "Resource allocation under uncertainty using the maximum entropy principle", submitted to IEEE Transactions on Information Theory, April 2002, revised December 2003.

- M. Johansson, "Benefits of multiuser diversity with limited feedback", SPAWC 2003 (IEEE Signal Processing Advances for Wireless Communications), Rome, June 2003.

Chapter 4

Chapter 4 considers the problem of determining the amount to bid in a certain type of auctions in which customers submit one sealed bid. Each customer has a carrying capacity (not necessarily equal among customers) denoting the amount of goods that the customer can currently receive. Only the bid with winning price-capacity product obtains any goods, and then obtains an amount equal to the carrying capacity of the customer. The auction is repeated many times, with only limited information concerning winning price-capacity products being announced to the customers. This situation is motivated in for example communication networks in which a possible way of obtaining a desired quality-of-service level is to use dynamic pricing in combination with competitive bidding. We derive optimum bidding rules for a few typical service requirements and show in simulations that the derived bidding strategies are successful. The material presented in this chapter has not yet been published.

Chapter 5

Here, we consider the problem of allocating bandwidth among different users in a downlink over a set of parallel channels. The objective is to maximize the expected system throughput over a given time interval while accounting for uncertain arrival rates and possibly uncertain channel predictions. Based on the framework developed in Chapter 3, we introduce the maximum entropy principle as a robust and powerful method to solve the notorious problem of modelling individual Internet data sources. This work generalizes the results of Knopp and Humblet (1995) and Tse (1997) concerning optimum scheduling policies for the case of infinitely much data to send and perfect channel knowledge with one time slot scheduling. Our

solutions provide maximum expected throughput for multiple parallel channels, finite queue lengths with uncertain arrival rates, arbitrary scheduling horizons, and include a general model for accounting for channel prediction inaccuracies.

We also comment on the notion of queue stability which has been taken as the primary criterion in some works on scheduling, and note some of its more serious deficiencies. Moreover, we discuss the use of logarithmic throughput criteria and argue that they may be more appropriate than previously thought.

The work presented in Chapter 5 is based on the following contributions, but includes new and previously unpublished material, mainly on queue-stabilizing schedulers, logarithmic criteria and competitive bidding as a mechanism for obtaining a desired level of service.

- M. Johansson and M. Sternad, "Resource allocation under uncertainty using the maximum entropy principle", submitted to IEEE Transactions on Information Theory, April 2002, revised December 2003.
- M. Johansson, "Benefits of multiuser diversity with limited feedback", SPAWC 2003 (IEEE Signal Processing Advances for Wireless Communications), Rome, June 2003.

Chapter 6

In Chapter 6 we discuss the implications of limited channel feedback for multiuser diversity. We study quantization of the channel information in a setting where adaptive modulation is used in combination with a pure multiuser-diversity strategy and propose to use common rate thresholds for all users. We derive an expression for determining such a quantization achieving maximum expected system throughput and also find an expression for the optimum amount of feedback taking both downlink throughput and feedback overhead into account. From this we find that the expected throughput does in theory not reduce at all as much as in traditional systems with fixed access schedules. It however turns out that unfairness increases with reduced channel feedback and that the promised theoretical throughput may reduce drastically in practice due to an inherent sensitivity to correctly chosen rate thresholds.

We propose two methods for achieving a high degree of multiuser-diversity gain with only 1-bit channel feedback. The first method adaptively changes a rate threshold based on usage statistics in a recent time interval. The second method combines individual rate thresholds – decreasing the sensitivity to correctly chosen levels – with a simple new scheduling strategy guaranteeing a fixed inter-access delay while still achieving a high multiuser-diversity gain.

The chapter is based on the work presented in

- M. Johansson, "On scheduling and adaptive modulation with limited channel feedback", submitted to IEEE Transactions on Communications, April 2004.

- M. Johansson, "Benefits of multiuser diversity with limited feedback", SPAWC 2003 (IEEE Signal Processing Advances for Wireless Communications), Rome, June 2003.

- M. Johansson, "Diversity-Enhanced Equal Access – Considerable throughput gains with 1-bit feedback", SPAWC 2004 (IEEE Signal Processing Advances for Wireless Communications), Lisbon, July 2004.

Chapter 7

As a related issue to that of scheduling users *within* a cell, we here investigate inter-cell scheduling, or reuse partitioning, i.e. partitioning bandwidth between interfering and non-interfering sub-sectors in a cellular network. The objective is to dynamically reallocate bandwidth to areas where it will be best utilized in a coming time period. The criteria which we develop, to maximize the total expected throughput in an area, extend the ideas presented in Chapters 3 and 5 and we show that the consequent framework can be used also for analyzing and making handovers and access control decisions.

The material covered in this chapter is based on the following contribution.

- M. Johansson, "Dynamic inter-cell scheduling based on local supply-demand fluctuations", submitted to IEEE Transactions on Vehicular Technology, April 2004.

Chapter 8

In many problems of resource allocation, the prior information and the computational power are limited, thus requiring some solution for adapting to unforeseen events at low complexity. In the final chapter of this thesis, we therefore introduce a method for conducting approximate Bayesian inference. The method is based on approximating a full Bayesian inference by adapting a simpler quantized distribution according to incoming data. We establish that the optimum approximation in the sense of maximizing the mutual information of the quantized and the unquantized distributions yields a quantized distribution with maximum entropy. The approximate pdf is then represented by a self-organizing histogram, where each bin is adjusted to attain equal probability mass. We show how this is accomplished in practice by using basic probability theory from Chapter 2.

The resulting algorithm provides a general-purpose approximation of Bayesian inference for arbitrary non-stationary distributions. It does however not take advantage of time dependencies. The resulting posterior distribution increases its resolution at regions where observations are frequent and decreases resolution in regions of low activity. It moreover provides easy assessment of expectations of arbitrary integrable (or summable, in the discrete case) functions of the uncertain quantity. The material in this chapter has not yet been submitted for publication.

1.5 Summary and Further Work

In this thesis we hope to show how using Jaynes' interpretation of probability theory as an extension to logic unifies a number of resource allocation problems. The applications discussed in the thesis range from manufacturing and bidding in auctions to diverse topics in mobile communications such as user scheduling, quantization of channel information, inter-cell scheduling, hand-overs and access control. Although the problems differ widely, the criteria, the models of uncertainty concerning the future outcomes, and the solutions will often be delusively similar. One reason is that the actual performance criteria of all these resource allocation problems are very similar; maximizing the resource efficiency, or the revenues of the manufacturer, customer, or network operator. Another even stronger reason is that we often find ourselves in situations where the only information we may have concerning future demand for widgets or data rates or whatever the goods we distribute, is quite the same irrespective of the actual application at hand. The similarity in information entails similar probability models when we regard probabilities as the fundamental carriers of information, and thus we often use similar probability distributions for very different entities. It is a subtle but essential insight that this is not equivalent to *assuming* that the different entities will behave in the same way. Using maximum-entropy distributions will lead us to always take precautious decisions that *avoid* assumptions (Roberts, 1971) concerning the actual long-run behavior of the entities. Instead, all actual information that we have is thoroughly distilled and utilized while the full extent of our uncertainty is frankly admitted. The resulting inferences, due to the property that the class of maximum entropy distributions is exactly the class of distributions with sufficient statistics (see Chapter 2), will only use those properties of the data that we actually used in assigning the probability and will not rely on any other properties of the data.

A main intention of this work has been to provide a readable introduction to probability theory as logic with an emphasis on examples in the fields of scheduling and resource allocation. Many problems are naturally posed as ones of improving the utility of some limited resource. Apart from the previously cited paper by

Jaynes (1963b), very little has yet been published on resource allocation under uncertainty from our present perspective on probability theory. Further work on bidding under uncertainty is an interesting topic; in Chapter 4 we only consider one specific type of auction, and there is clearly many other situations which call for other solutions. The format of the auction, the bidders' objectives, and especially the information announced to the bidders will have a strong impact on the resulting strategies.

In mobile communications, which is the main application area studied here, a number of outstanding issues require more research. For a number of years, the use of multiple antennas at the transmitting and receiving ends (known as multiple-input multiple-output (MIMO) channels) have been investigated in different scenarios. These techniques promise substantial performance gains in single-link scenarios, but as we noted earlier, some techniques may have adverse effects when employed in a multiuser setting. Multiple antennas are beginning to be used in some cellular systems, and are believed to emerge as a standard component in future base stations. If these extra investments are to be put to best use, it is imperative that a careful analysis of the joint scheduling-MIMO strategy is carried out. Although isolated results are beginning to appear, there is still a lack of practical high-performing strategies. Very recently, the capacity region and a capacity-optimal scheme for the Gaussian MIMO-broadcast channel has been found (Caire and Shamai Shitz, 2003, Jindal et al., 2004, Viswanath and Tse, 2003), but the results require full channel information at the transmitter. Moreover, the capacity-optimal scheme is extremely computationally demanding.

A key issue in realizing practical schemes is the balance between channel feedback and downlink system throughput. We see in Chapter 6 that some types of channel feedback can be substantially quantized without compromising the downlink throughput when multiuser diversity is taken into account. But the type of channel information that is required for many MIMO techniques, such as beamforming, does not have this desired property. The complex interplay between channel feedback, scheduling gain, and the choice of MIMO technique combines into one of the most important research topics for near-future cellular systems.

Chapter 2

Probability Theory As Logic

IN any area of science, or indeed daily life, we have to draw conclusions from limited knowledge. Only very rarely do we have enough information so as to draw conclusions with absolute certainty about any matter. For instance, will it rain today? Should I invest in the stock market or in bonds? Every day, decisions must be made in the face of uncertainty.

One should expect that in the evolution of animals, competition would favor those with a highly developed skill for making plausible inferences, i.e. making generalizations and logical decisions that turn out to serve the purpose of the decision maker[1] well and give that animal easier access to food, etc. In the most highly developed animals, therefore, we expect that some form of optimal reasoning under uncertainty should have evolved over time. We put forward this example as an objection to the opinion that has occasionally been advanced that a theory for induction is fundamentally impossible. The very fact that people and animals are making successful inferences based on limited knowledge should be sufficient reason to infer the possibility of such a theory.

In this chapter, we study the theory of plausible reasoning developed into its present form by Edwin T. Jaynes (2003). Although Jaynes should certainly be credited as the father of this scientific paradigm, his work is an extension to Jeffrey's (Jeffreys, 1939) and the seminal derivation due to Cox (1946) of the ordinary rules of probability theory as an extension to logic.

[1] Perhaps we should clarify for whom the decision should serve a purpose. Richard Dawkins (1976) makes it plausible that it is not the individual animal, or the species, that is the main beneficiary in evolution. Evolution is a process that critically affects which *genes* are transmitted to the next generation. An animal in itself is a collection of competing and cooperating genes, and is not the entity which evolution fundamentally involves.

The presentation here is of an introductory character, and requires no previous knowledge of probability theory. The chapter is intended as a comprehensive introduction with emphasis on the fundamental principles and rules. Since most readers probably have been exposed to some form of conventional probability theory, we will often point out differences between these two subjects, so as to better facilitate the understanding of the present theory. We hope that such a reader will be pleasantly surprised by the simplicity and generality of this theory.

2.1 Consistency and Common Sense – The Basic Desiderata

Our topic is optimal information processing, i.e. deductive reasoning whenever possible, and inductive reasoning when the information at hand is insufficient to render a conclusion with the status of absolute certainty. In order to derive a theory for this purpose, we will first state three desiderata (desired properties) that such a theory should reasonably satisfy. Having stated them explicitly, Jaynes (2003) shows that it is indeed possible to derive from these desiderata a unique set of rules for conducting inferences. We will simply state the resulting rules without derivation, as some of the mathematics is quite cumbersome. The interested reader is referred to (Jaynes, 2003) for the full derivation.

Our desiderata are:

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

(II) *Qualitative correspondence with common sense.*

(III) *Consistency requirements:*

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

(IIIb) *All evidence relevant to a question is always taken into account. No information is ever arbitrarily ignored.*

(IIIc) *Equivalent states of knowledge are always represented by equivalent probability assignments. That is, if in two problems, the reasoner's state of knowledge is the same (except perhaps for the labelling of propositions), then he or she must assign the same probabilities for both.*

As Jaynes remarks, desiderata (I), (II), and (IIIa) are the basic structural requirements on how plausibilities are processed internally, and (IIIb) and (IIIc) are 'interface' conditions which shows how probabilities relate to the outside world.

Recall now that our aim is to generalize deductive two-valued logic into inductive logic, thereby allowing us to reason consistently also under uncertainty. The basic building blocks are the same in both theories. The only difference is that we are no longer necessarily certain about the truth value (true or false) of some proposition of interest. Typically, a proposition is a combination of other more elementary propositions, and its truth value depends on whether other propositions are true or false. Consequently, in order to determine the plausibility for some event we first need to establish how it relates to other propositions and their truth values.

It is a fact from deductive logic (or boolean algebra) that an adequate set of operations for constructing any logical proposition[2] of statements is contained in the logic product (=conjunction, AND) and the negation (=NOT)

Logical product = conjunction, AND: $AB =$ both propositions A and B are true.

Negation = NOT: $\bar{A} = A$ is false.

By repeatedly applying these two operations it is possible to construct any arbitrary logical proposition. Apart from the logical product AB and the negation $\bar{A}$, two other operations are typically defined in deductive logic, with the following meanings:

Logical sum = disjunction, OR: $A + B =$ at least one of the propositions A and B are true.

Implication: $A \Rightarrow B = A$ implies B, i.e. if A is true, then B is also true, or equivalently, if B is false, then A is also false.

Note that the disjunction of A and B is equivalent to denying that both are false,

$$A + B = \overline{(\overline{A}\,\overline{B})}\,,$$

and that the implication $A \Rightarrow B$ is the same as denying that A and $\bar{B}$ are both true,

$$A \Rightarrow B = \overline{A\overline{B}}\,.$$

These two last operations are thus redundant in the sense that they are just combinations of AND and NOT.

It is important to emphasize the difference between logical implication and its meaning in colloquial usage. Logical implication makes no reference to time or

[2] A proposition here refers to any combination of statements that can, at least in principle, be deemed either true or false.

physical causation. This is brought out most clearly by an example. Let

$$\begin{aligned} A &= \text{Rain tonight} \\ B &= \text{Clouds tonight.} \end{aligned}$$

Then the correct logical relation is $A \Rightarrow B$, since if there is rain tonight, there is certainly clouds as well, and if there are no clouds, there can be no rain. Since we are accustomed to thinking in terms of physical causation rather than logical implication in everyday life, we sometimes tend to confuse these two distinct concepts and equate logical implication with physical causation. Then one is tempted to think, erroneously, that clouds implies rain, and not the other way around. Therefore, we stress this; Logical implication refers only to truth values and does not require or point to any causal effects. For instance, there is nothing illogical in a proposition implying a different proposition which makes statements about past events. This, seemingly trivial, remark becomes especially important in our extension of two-valued logic to a probability scale. If a probability for some event depends on some other event, it does not mean that the events are physically related in any way. For instance, the pear crop does not affect the apple crop, but knowing that this was a good year for apples, we probably have reason to believe that the pear crop will be good too. Or, if a probability for observing an electron in some state depends on the state of another electron (possibly separated from the former by a large distance), that does not imply that its state changes when measuring the state of the latter.

Returning now to our desiderata, some readers will note with dismay the seeming inexactness of our desideratum (II), common-sense correspondence. But note then from our preceding discussion on logical propositions that any arbitrary proposition can be constructed by only two operations, the logical product and the logical negation. So in order to determine the probability for any logical proposition fully, we only have to define rules for how the probability for the logical product of two propositions depend on the probabilities for the individual propositions, and how the probability for the logical negation of a proposition is written as a function of the probability for the proposition. Hence, it is in the formulation of these two basic rules that we require qualitative correspondence to common sense. Any reformulation of desideratum (II) would in the end need to have this correspondence to common sense or else it would be useless for our purposes.

Furthermore, desideratum (II) gives a 'sense of direction' for how probabilities change with information (not *how much* they change, but only in which direction).

This can be described in the form of three weak syllogisms,

$$\frac{\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ B \text{ is true}\end{array}}{\text{therefore, } A \text{ becomes more plausible}}$$

$$\frac{\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ A \text{ is false}\end{array}}{\text{therefore, } B \text{ becomes less plausible}}$$

$$\frac{\begin{array}{c}\text{if } A \text{ is true, then } B \text{ becomes more plausible}\\ B \text{ is true}\end{array}}{\text{therefore, } A \text{ becomes more plausible}}$$

These syllogisms may seem questionable at first sight, and the reader is urged to examine whether they are reasonable in some simple example. Try for instance the rain example above on the two first syllogisms,

$$\begin{aligned} A &= \text{Rain tonight} \\ B &= \text{Clouds tonight.} \end{aligned}$$

Neither of these syllogisms would be required in a theory that does not correspond qualitatively to common sense. Therefore, although desideratum (II) is necessarily reduced to a set of mathematical requirements on the probability rules in the derivations, we keep it in its current formulation because we think that exactness in the narrow mathematical sense may obscure understanding the over-all goal of the theory.

2.2 The Fundamental Rules

Jaynes shows the remarkable result that using only the three desiderata from Section 2.1, it is possible to construct a unique[3] theory for plausible reasoning. The resulting rules are the following:

[3]The rules are unique, but any one-to-one transformation of the rules is of course equivalent in content. As is the typical convention, we denote 0 for impossibility and 1 for certainty. It would also be possible to use an inverse scale where 1 denotes certainty and ∞ impossibility. The resulting theory would however look less familiar to us.

The product rule:

$$P(AB \mid C) = P(A \mid BC)P(B \mid C) = P(B \mid AC)P(A \mid C) \tag{2.3}$$

The sum rule:

$$P(A \mid B) + P(\bar{A} \mid B) = 1 \tag{2.4}$$

We have here introduced the notation $P(A \mid B)$ meaning the probability that A is true subject to knowledge that B is true, often abbreviated as A conditional on B, or simply A given B.

Note here an important difference to the random variable approach to probability theory: all probabilities in our theory are conditional on some other proposition. Since a probability is simply a statement about our degree of belief in a proposition, it cannot be determined without explicit statement of what information we base it on. It is therefore meaningless to speak of a true probability, as were it a physical property in nature, since a probability is only an artefact of our ignorance as to the true logical status of the proposition in consideration. A probability conditional on nothing is ill-defined; it has no 'logical context' from which it can be numerically determined.

As a convention, we will use the short-hand notation I to denote the prior information that is common to all probabilities in any particular case of inference.

From the product rule and the sum rule, we can derive a very useful rule determining the probability that at least one of propositions A and B are true, the extended sum rule:

$$P(A + B \mid I) = P(A \mid I) + P(B \mid I) - P(AB \mid I)\,. \tag{2.5}$$

Proof:

$$\begin{aligned}
P(A + B \mid I) &= P(\overline{\overline{A}\,\overline{B}} \mid I) = 1 - P(\overline{A}\,\overline{B} \mid I) \\
&= 1 - P(\overline{A} \mid \overline{B}I)P(\overline{B} \mid I) \\
&= 1 - \left[1 - P(A \mid \overline{B}I)\right] P(\overline{B} \mid I) \\
&= 1 - P(\overline{B} \mid I) + P(A\overline{B} \mid I) \\
&= P(B \mid I) + P(\overline{B} \mid AI)P(A \mid I) \\
&= P(B \mid I) + \left[1 - P(B \mid AI)\right] P(A \mid I) \\
&= P(B \mid I) + P(A \mid I) - P(AB \mid I)\,. \qquad (2.6)
\end{aligned}$$

■

If only one of A and B can be true, then the probability that both be true is $P(AB \mid I) = 0$, and the probability for A OR B reduces to the sum of the probabilities for the individual propositions.

2.3 Useful results: Bayes' Rule and Marginalization

By rearranging the factors in the product rule (2.3) we have (with $I = C$) that

$$P(A \mid BI) = P(A \mid I)\frac{P(B \mid AI)}{P(B \mid I)} \,. \tag{2.7}$$

This relation is often denoted Bayes' rule in memory of the British reverend and amateur mathematician Thomas Bayes who showed this relation in a specific case[4]. Its importance becomes clearer if instead of A and B we use the propositions

- $H =$ A hypothesis of interest
- $D =$ Observations of some data.

Then we obtain

$$P(H \mid DI) = P(H \mid I)\frac{P(D \mid HI)}{P(D \mid I)} \,, \tag{2.8}$$

which shows how our belief in a hypothesis H changes with the acquisition of new data D. Written in this form $P(H \mid DI)$ is often denoted the *posterior probability* for the hypothesis, $P(H \mid I)$ the *prior probability*, and $P(D \mid HI)$ the *likelihood*. Given the uniqueness of our rules subject to the constraints of desiderata (I)-(III), Bayes' rule can be seen to be a fundamental equation of optimal learning under uncertainty. We shall presently see by example how the rule corresponds to an ideal common sense. To prepare for this, we first derive another useful result from the two basic rules.

As mentioned in the ending notes of the preceding section, the extended sum rule (2.5) takes a particularly simple form when the individual propositions are *mutually exclusive*, i.e. when only one of the propositions can be true. If the propositions are also *exhaustive*, i.e. one of them must surely be true, then we obtain the simple sum rule. This can easily be extended by mathematical induction to an arbitrary number of propositions, resulting in

$$P(A_1 + \cdots + A_N \mid I) = \sum_{i=1}^{N} P(A_i \mid I) = 1 \,. \tag{2.9}$$

Now, note that the truth value (i.e. true or false) of a proposition A is the same as that of $A(B_1 + B_2 + \cdots B_N)$ if the propositions B_i are mutually exclusive and exhaustive (since the truth value of A is always the same as the truth value of A

[4]Laplace generalized the results of Bayes and should perhaps be attributed the discoverer of the rule rather than Bayes.

AND any proposition known to be true, i.e $A = A1$ always). This means that a probability for a proposition A can be resolved into

$$\begin{aligned} P(A \mid I) &= P\left(A \sum_{i=1}^{N} B_i \mid I\right) \\ &= P\left(\sum_{i=1}^{N} B_i \mid AI\right) P(A \mid I) \\ &= \sum_{i=1}^{N} P(B_i \mid AI) P(A \mid I) \\ &= \sum_{i=1}^{N} P(B_i A \mid I) \\ &= \sum_{i=1}^{N} P(A \mid B_i I) P(B_i \mid I) \,, \end{aligned} \tag{2.10}$$

given that the B_i, $i = 1 \ldots N$ are mutually exclusive and exhaustive. This, on first sight somewhat strange-looking, technique can be used to determine the probability that A is true regardless of which one of the possible B_i hypotheses is true. Called *marginalization*, this technique is a very useful way of making inferences about a particular hypothesis which depends on the value of some hypothesis or parameter, B_i, whose exact value is uncertain. Such a parameter, which enters the problem but is not the main object of interest, is often called a *nuisance parameter*. In the case of a continuous parameter θ, the sum is substituted into an integral

$$P(A \mid I) = \int P(A \mid \theta I) P(\theta \mid I) d\theta \,, \tag{2.11}$$

and we say that we integrate out the nuisance parameter.

2.3.1 Common-Sense Correspondence

When asked about whether an observation that was predicted by a certain theory H_1 confirms the theory, most people would answer a positive yes. Now, let us see what our rules say. Look at the individual factors in Bayes' rule on the form (2.8).

The observation D was predicted by the theory, so clearly $P(D \mid H_1 I)$ is large. But there are two factors left; the prior probability for the theory $P(H_1 \mid I)$ which could have been anything, and the normalizing factor $P(D \mid I)$. How do we

determine this last factor? We use marginalization,

$$P(D \mid I) = \sum_{i=1}^{N} P(D \mid H_i I) P(H_i \mid I) \,, \tag{2.12}$$

where the set of H_i contains all conceivable[5], mutually exclusive, hypotheses that may explain the observation D.

So we see that the ratio

$$\frac{P(D \mid H_1 I)}{\sum_{i=1}^{N} P(D \mid H_i I) P(H_i \mid I)} \tag{2.13}$$

determines whether the probability for H_1 increases, decreases or remains the same upon the observation. In order to give an answer to the question, we must therefore explicitly state all other alternative hypotheses and compare whether any of these alternatives may be more plausible on the observation. We can state this in a slightly different manner: the posterior probability for a hypothesis equals the ratio of the joint probability for the observation and the hypothesis to the sum of all joint probabilities of the observation and all possible hypotheses

$$P(H \mid DI) = \frac{P(DH \mid I)}{\sum_{i=1}^{N} P(DH_i \mid I)} \,. \tag{2.14}$$

This is obtained by inserting $P(DH \mid I) = P(D \mid HI) P(H \mid I)$ and (2.12) into (2.8). Perhaps, a better question is then: which of the hypotheses $H_1 \ldots H_N$ is more likely? That takes us into the area of hypothesis testing, where we simply compare the posterior probabilities for the individual hypotheses, and if prompted to bet only on one of them[6], select the one with the highest probability.

The main lesson to learn from this example is that we cannot say anything about the probability for a theory unless we clearly state alternative theories. We can only say how plausible a certain theory is in relation to other clearly stated theories. This brings out a useful feature of the present theory: the rules themselves tell us how to pose our questions. When confronted with the rules of probability theory, ill-posed questions are brutally exposed of their nature. The rules help us in determining what parts are missing to make a well-posed question.

[5]When we say 'all conceivable' hypotheses, we really mean 'all considered' hypotheses. We cannot hope to come up with all possible causes for some event, but we can always infer which one out of a set of considered alternatives that is best supported by the observations and our prior information.

[6]Note that according to probability theory, marginalization should always be used when there is uncertainty concerning which out of a number of alternatives is the true one. Thus, we should not select just the most likely theory and believe it blindly, but keep all the others in mind weighted by their posterior probabilities.

2.4 The Notion of Randomness

We have now derived the basic rules for manipulating probabilities. Given probabilities for individual statements, we can find probabilities for arbitrary propositions concerning these statements. We have so far not addressed the equally important question of how to determine the initial numerical values for probabilities. Before doing that, we must emphasize an essential feature of the present theory.

Nowhere in our desiderata, the consequent derivations, or the rules of probability theory, have we made any reference to randomness. This may be startling to some, as this is the starting point of the conventionally taught frequentist probability theory. There, a probability of an event[7] is typically defined as the limiting frequency with which a certain event occurs in a random experiment repeated under the same circumstances infinitely many times. It is taken as an axiom that probabilities can only refer to 'random variables', or 'stochastic processes', i.e. quantities that are fundamentally impossible to determine before the outcome is observed. The probability does not say anything about an individual outcome, but refers to an *ensemble* of all possible outcomes were the random experiment to be repeated an infinite number of times.

Now, if we are to apply the frequency definition of probability theory to realistic problems, then we must find or estimate the true probability of the event we are interested in. Let us for instance discuss the problem of estimating the impulse response of a mobile communications channel. The first question to ask is then: is the impulse response random? If we believe that it is determined by Maxwell's equations concerning electromagnetic waves or some more refined theory, i.e. if we believe that with knowledge of all initial conditions and some physical theory it is in principle possible to determine the impulse response, then, according to the frequentist definition, we must reject the use of probability theory. Still, we see probability theory used in the mobile communications literature. For instance, distributions such as Rayleigh, Nakagami-m, or Rice, are used for the envelope of the received signal. The reason must be that we still resort to a more relaxed definition of random variables: a random variable is taken to be a quantity that we have so little information about that we can hardly expect to determine it fully without actually knowing the outcome. But if that is the definition we adhere to, then we must frankly admit that the only reason for deciding that something is random is our own lack of information concerning the actual outcome. Then, how can we ask for a true probability distribution of the impulse response? The distribution we seek depends on how much we know about the impulse response, and to try to

[7] Note that in the frequentist definition, a probability is a property *of* the event, whereas in our theory a probability is determined *for* the event. It is a function not of the event, but of the information the inference is based on.

determine a true probability distribution by measuring frequencies would be like, as Jaynes aptly remarks, trying to assess a boy's love for his dog by performing measurements on the dog.

A main problem with the frequentist definition is that it does not even recognize such concepts as uncertainty or information which are central in conducting inferences. Indeed, in the standard reference of frequentist probability theory Feller (1968) remarks: 'There is no place in our system for speculations concerning the probability that the sun will rise tomorrow'. This seems to be precisely the type of problem that is of concern to an engineer. In constructing a bridge, he wants to be able to say with confidence something about the strain that this particular bridge will stand. Of course, his conclusion cannot take the status of absolute certainty, but he wants his statements made with a reasonable degree of belief attached to them. Since he has carefully chosen his materials and made his construction based on knowledge of the physics of elastic and rigid bodies, there is no random process involved and he must conclude that the frequentist theory cannot help in inferring properties of his bridge.

But still we think that probability theory could, and should, be used in both examples. There is no need to appeal to randomness, and if we instead of estimating 'true' probability distributions, as were they an actual physical property, shift our focus to making statements about our own uncertainty towards the object in question, we will realize that we can always find a probability distribution that adequately describes that uncertainty. Then, our theory is completely general, and can be applied to any problem of inference regardless of whether there is such a thing as true randomness involved.

In light of this, if we would still accept that there is a possibility that some things in nature are completely random, i.e. that even if we could fix all initial conditions of an experiment, the experiment would persist in giving different results on repeated trials, then we should at least have some objective procedure for choosing between the hypotheses $H_1 =$ *the outcome is fundamentally impossible to determine, i.e. the process is random, or stochastic*, and $H_2 =$ *the outcome is in principle possible to determine, but in order to do so information that we do not possess may be required.*

But how could we ever hope to determine that something is truly random? That would require complete knowledge of every aspect of nature's workings, since the determination of randomness requires rejecting all possible physical reasons that might explain the outcome. But that makes randomness impossible to prove, since it requires absolute evidence that all physical mechanisms are known.

Now, it seems that true randomness is in itself fundamentally unreasonable, because it requires that an outcome is impossible to determine, but the mere fact that something occurs should convince us of the opposite! However we twist the

explanations of randomness, we end up having to use it as if it is only an artefact of our own ignorance. Why, then, not take the simpler and more constructive route of admitting that uncertainty is the only 'cause' of apparent randomness?

We therefore stress that whatever one's outlook on the rationale of using probabilities is, in the end, when having to determine a probability distribution numerically, everyone actually uses our definition, although perhaps unwittingly, of probability as a description of uncertainty. Then, it should be clear that it would be a contradiction to ask for a true probability *of* an event.

2.5 Assigning Probabilities I – The Principle of Indifference

We now turn to the problem of assigning numerical values to probabilities. We will start with the perhaps most basic situation, in which we know very little about different outcomes. The situation can be formulated as a symmetric information condition, or a state of indifference. Consider two problems. In Problem 1, we have a set of mutually exclusive and exhaustive propositions, $\{A_1 \dots A_n\}$, and we wish to find the probabilities $P(A_i \mid I)_1$. In Problem 2, we face the same problem but here the set of propositions $\{A'_1 \dots A'_n\}$ is a permutation of the propositions in Problem 1. For instance, it might be that $A'_1 = A_3$, etc. We are thus in effect facing two identical problems, but the labels of the propositions have been changed. Suppose now that information I is indifferent between all propositions, i.e. if it says something about A_1 it says the same thing about A_2, A_3, and so on. Then, desideratum (IIIc), which says that equivalent states of knowledge must be represented by the same probability assignments, requires that

$$P(A_i \mid I)_1 = P(A'_j \mid I)_2 \qquad i, j = 1 \dots n \,. \tag{2.15}$$

Note that this holds whatever the exact information I is. The only requirement is that it says the same thing about all propositions A_i.

The *symmetry equations* (2.15) have only one solution. Since the n propositions are exhaustive and mutually exclusive,

$$P(A_i \mid I) = \frac{1}{n} \qquad i = 1 \dots n \,. \tag{2.16}$$

This rule, which says that probability assignments can be performed by breaking down propositions into more elementary propositions for which our background information I is indifferent and assign equal probability for these sub-propositions, is usually called the *principle of indifference*.

We can immediately see an extension to this rule in the following standard example from probability theory. There are n different balls, labelled $A_1 \dots A_n$, spread out in an urn which we are to make a blindfolded draw from. Out of the n balls, m of them, $\{A_1 \dots A_m\}$, are black. What is then the probability for drawing a black ball? Our background information is indifferent between different balls, and the probability for drawing ball A_i is thus $P(A_i \mid I) = 1/n$. Then, since the outcomes are mutually exclusive and exhaustive, the probability for black is

$$P(A_1 + A_2 + \cdots + A_m \mid I) = \sum_{i=1}^{m} P(A_i \mid I) = \frac{m}{n} \,. \tag{2.17}$$

This rule, which we here derived from our basic consistency requirement, desideratum IIIc, is a very common *definition* for probabilities, and was used by for instance Laplace. In this case, we find that the probability is equal to a frequency, not as an axiom, but as a consequence of the information indifference between different propositions. In other problems, this frequency correspondence does not occur. We shall come back to an example of this later in connection with the maximum entropy principle, and show that the usefulness of some probability assignments lie in making such frequencies irrelevant.

2.6 Assigning Probabilities II – Laplace's Rule of Succession

Let us now turn to another common scenario, in which our information concerning future outcomes consists of a record of the number of past occurrences for each possible outcome. Suppose that there are K distinct possible outcomes, and that outcome k has occurred m_k times out of the total record of M outcomes, i.e.

$$M = \sum_{k=1}^{K} m_k \,. \tag{2.18}$$

From these numbers, what can we say about the plausibility of recording r_k occurrences of k in a *future* period? If we solve this problem, the probability for each outcome k is then obtained by taking the expectation of the relative frequencies with which they occur. Assuming that the underlying causal mechanisms which determine the outcomes do not change significantly with time, it follows that the relative frequencies should remain constant as well. The problem of translating relative frequencies observed in a finite interval into predictive probabilities is not new, indeed it is one of the oldest in probability theory. The solution is found from a generalized form of Laplace's rule of succession (Jaynes (2003), ch. 18).

We seek to evaluate

$$\begin{aligned} P(f_1...f_K|m_1...m_K I) = \\ =\frac{P(m_1...m_K|f_1...f_K I)P(f_1...f_K|I)}{P(m_1...m_K|I)} \end{aligned} \tag{2.19}$$

where

$$f_k = \frac{r_k}{\sum_{j=1}^{K} r_j} \tag{2.20}$$

is the relative frequency with which outcome k *will* occur, and I contains only information about the past number of outcomes m_k.

We perform the derivations of (2.19) and $\langle f_k \rangle$ in Appendix 2.A but note the essential elements here.

The prior probability distribution for the relative frequencies f_k is defined by a distribution which is uniform over all combinations of K non-negative numbers that sum to unity (by the principle of indifference):

$$P(f_1...f_K|I) = C\delta(f_1 + ... + f_K - 1)\,, f_k \geq 0\,, \tag{2.21}$$

where $\delta(\cdot)$ is the Dirac delta ($\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ elsewhere). The likelihood term in (2.19), the probability for obtaining $m_1 \ldots m_K$ samples of each outcome $k = 1 \ldots K$ given that the frequencies $f_1 \ldots f_K$ are known, is a bit more complicated. We here interpret the frequencies as probabilities, in effect claiming that the causal mechanisms which determine the actual outcomes are so haphazard or complex that we cannot model them any better than simply assuming that the relative frequencies with which they occur will persist to be representative. The probability for obtaining a certain sequence of occurrences is according to the product rule given by the product of the individual probabilities for each outcome in the sequence, in this case $f_1^{m_1} \cdots f_K^{m_K}$. But since the given sample numbers can occur in several ways, depending on the order with which they occur in the sequence, the sum rule dictates that we must sum the probabilities for all possible sequences to obtain the probability for the given sample numbers regardless of their order. Since a sequence of length M with given sample numbers $m_1 \ldots m_K$ can arise in $\frac{M!}{m_1! \cdots m_K!}$ ways, the likelihood term in (2.19) is thus the following *multinomial* distribution,

$$\begin{aligned} P(m_1...m_K|f_1...f_K I) = \\ =\frac{M!}{m_1! \cdots m_K!} f_1^{m_1} \cdots f_K^{m_K}\,. \end{aligned} \tag{2.22}$$

Finally, the prior distribution $P(m_1...m_K|I)$ is obtained by averaging the joint distribution for m_k and f_k over all possible f_k.

As shown in Appendix 2.A, the probability for obtaining a certain outcome k is given by the expectation of the relative frequency with which that particular outcome occurs:

$$p_k \stackrel{\triangle}{=} P(k|m_1...m_K I) = \langle f_k \rangle = \frac{m_k + 1}{M + K} \,. \tag{2.23}$$

Note that when the number of observations, M, is very small compared to the number of possible outcomes, K, the distribution tends to a uniform distribution. This agrees with common sense; in order to obtain any sharp predictions, the number of observations must be relatively large in comparison to the number of hypotheses. If $M \gg K$, then the probability assigned to any outcome is practically independent of the number of possible outcomes, and depends only on the observed data. Note further that the probability assigned to any outcome will never be zero unless either K or M is infinite, which is never the case in reality. This can be understood from observing that (2.23) can be interpreted as using the observed frequencies as estimates of the predictive probabilities, but in addition using the fact that each of the outcomes actually *can* occur, corresponding to K additional observations, one for each outcome.

2.7 Assigning Probabilities III – The Maximum Entropy Principle

Suppose now that our information is of another, more informative, kind, consisting of mean values of functions of some variables. For example, suppose that a sales manager of an apple garden has information I that the average order size is 420.8 apples. How do we translate this into a probability statement, $P(\text{size of the next order} \mid I)$? Such a probability statement could then be used to guide decisions regarding whether more trees should be planted or not, or if the number of trees could be reduced.

The principle of indifference is not directly applicable, as it seems hard to partition order sizes in a way that make our information indifferent between different partitions. It is clear that, given I, some order sizes are more likely than others. We would certainly regard an order of size 400 more likely than one of size 100000. In some sense, we wish to assign a probability distribution which is as uniform as possible, so as to assume no more than necessary, but the uniformness will be constrained by the required mean value.

Is it possible to derive a measure of 'uniformness', or something that corresponds to the notion of *uncertainty*? Claude Shannon (1948) published his theory of communication in 1948. In that theory he derives a measure of uncertainty, which he denotes as *entropy*.

Shannon starts by considering a set of n possible events with the respective probabilities $p_1, p_2, \ldots, p_n$ (here, we use the shorthand notation $p_i = P(i \mid I)$). Then he asks, can we find a measure $H(p_1, \ldots, p_n)$ of how uncertain we are concerning which event will occur? As in the derivation of probability theory, a few desiderata are set up:

1. $H(p_1, \ldots, p_n)$ should be continuous in the p_i. Otherwise an arbitrarily small change in p_i would yield a large change in our uncertainty.

2. Qualitative correspondence to common sense, in the sense that when there are many equally likely events, we are more uncertain of the outcome than when there are few. This means that if all the p_i are equal, $p_i = 1/n$, $H(p_1, \ldots, p_n)$ is a monotonic increasing function of n.

3. Additivity. If a choice be broken down in two successive choices, the original H should be the sum of the individual values of H weighted by the probability for each choice. For example, if we start with $p_1 = 1/2$, $p_2 = 1/3$, and $p_3 = 1/6$, and group events 2 and 3, then we can first determine the uncertainty in the choice between 1 and the disjunction $2+3$, $H(1/2, 1/2)$. Then, with probability $1/2$, there will be a remaining uncertainty $H(2/3, 1/3)$ to resolve concerning events 2 and 3. That is, $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + 1/2H(2/3, 1/3)$.

4. Consistency, in the sense that when there are several ways of calculating $H(p_1, \ldots, p_n)$ we must get the same answer for every possible way.

Shannon shows that there is only one function H that satisfies these requirements,

$$H(p_1, p_2, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log p_i \ , \tag{2.24}$$

where K is an arbitrary positive constant, and the logarithm is taken to any base. A similar proof is given by Jaynes (2003), Chapter 11. Typically, K is taken as unity, and the logarithm either in base 2 or the Napierian[8] (natural) base. Shannon gave H the name *entropy* because of the mathematical similarity with the thermodynamical definition of entropy.

The entropy H has a number of interesting properties. Shannon notes for example the following.

[8] John Napier (1550-1617) was a Scottish amateur mathematician who 'invented' the logarithmic function. His main work on logarithms appears in *Mirifici logarithmorum canonis descriptio* from 1614.

- $H = 0$ only when one $p_i = 1$, all others being zero. That means that we are certain of the outcome and thus there is no uncertainty. In all other cases, H is greater than zero.
- The maximum of H is $H_{max} = \log n$, which is reached when all the $p_i = 1/n$.
- The *joint entropy* for two variables x, y with the possible outcomes denoted by x_i and y_j respectively is

$$H(x, y) = -\sum_{i,j} P(x_i y_j \mid I) \log(P(x_i y_j \mid I)) \leq H(x) + H(y) \quad (2.25)$$

  with equality only if x and y are logically independent, i.e. if knowledge of one gives no information about the other.

Shannon goes on to define the conditional entropy for x given y as

$$H(x \mid y) = -\sum_{i,j} P(x_i y_j \mid I) \log(P(x_i \mid y_j I)) \,, \quad (2.26)$$

but here we shall part with Shannon's nomenclature. The problem with (2.26) is that it is not a measure of the uncertainty concerning x given knowledge of y. Because if we actually know y then our uncertainty concerning x is surely not dependent on other possible values of y that could, but in fact did not, occur. The true uncertainty concerning x given that y took the value $y = y_j$ is just the original entropy definition,

$$H(x \mid y = y_j) = -\sum_{i} P(x_i \mid y_j I) \log(P(x_i \mid y_j I)) \,, \quad (2.27)$$

and we see that a better name for (2.26) is the *average* conditional entropy, since it is equal to (2.27) averaged over y, as is easily seen:

$$\begin{aligned} & -\sum_{i,j} P(x_i y_j \mid I) \log(P(x_i \mid y_j I)) \\ = & -\sum_{i,j} P(x_i \mid y_j I) P(y_j \mid I) \log(P(x_i \mid y_j I)) \\ = & -\sum_{j} P(y_j \mid I) \sum_{i} P(x_i \mid y_j I) \log(P(x_i \mid y_j I)) \,. \end{aligned} \quad (2.28)$$

We will use the notation $H(x \mid y)$ for the *average conditional entropy* (2.26), and $H(x \mid y = y_j)$ for the *conditional entropy* (2.27) (which is consistent with

Kullback's definitions for mean conditional information and conditional information (Kullback, 1968)).

From their definitions (2.25) and (2.26) and the product rule (2.3), we find that the joint entropy and the average conditional entropy are related through the following formula,

$$H(x, y) = H(x) + H(y \mid x) = H(y) + H(x \mid y) \,, \tag{2.29}$$

similarly to the product rule (2.3). By the use of (2.25) we have

$$H(x) + H(y) \geq H(x, y) = H(x) + H(y \mid x) \,, \tag{2.30}$$

which leads to

$$H(y) \geq H(y \mid x) \,. \tag{2.31}$$

The uncertainty is thus on average reduced upon new knowledge. Only if x and y are logically independent is $H(y) = H(y \mid x)$.

Now, if we accept the interpretation of H as a measure of 'amount of uncertainty', then it follows that the most honest description of a state of knowledge should be represented by probabilities with maximum entropy subject to whatever knowledge is at hand. Then we have only accounted for information that we actually have and assume nothing further than that. This rule, the maximum entropy principle, was introduced by Jaynes (1957a,b) in two seminal papers in which he showed that all of conventional thermodynamics followed from interpreting probability theory as logic and using the maximum entropy principle in assigning probabilities. Thus he showed that the predictions from thermodynamics were not to be interpreted as physical laws, but rather as the best inferences that could be made given a particular state of knowledge.

Although the requirements that led to the entropy expression all seem reasonable, one would expect that the basic desiderata (I)-(III) of probability theory should be all that is required. Indeed, there are other derivations of the maximum entropy principle that suggest that the introduction of a measure of uncertainty is not really required. We show here an alternative derivation, referred to by Jaynes as the Wallis derivation after its inventor Graham Wallis, which may provide a more direct motivation for using the maximum entropy principle.

Consider a scenario where we are to distribute N little 'quanta' of probability among n alternatives. The quanta are scattered randomly among the alternatives, for instance by a proverbial team of monkeys tossing quanta into urns representing the different alternatives, so that each outcome is equally likely in any toss. If the resulting distribution conforms to our information (i.e. the expectations match the given mean values), then we will keep it. Otherwise, we reject it and restart the procedure. What distribution is most likely to result from this game?

Let each probability quantum have magnitude $1/N$. In an outcome where alternative i gets m_i quanta, we have constructed a discrete probability distribution

$$p_i = \frac{m_i}{N} \,, \qquad i = 1, 2, \ldots n. \tag{2.32}$$

The question is now in how many ways a particular such distribution can be obtained. The probability for obtaining the distribution (2.32) is

$$\frac{1}{n^N} \times \frac{N!}{m_1! \cdots m_n!} \,, \tag{2.33}$$

where the first factor is the probability for obtaining any of the n^N possible sequences, and the second factor is the number of ways in which a particular sequence can arise.

The most likely distribution is thus the one which maximizes (2.33), or equivalently, since n and N are fixed, maximizes

$$W = \frac{N!}{m_1! \cdots m_n!} \tag{2.34}$$

subject to the constraints that our information imposes.

Noting that we can equally well maximize the logarithm of the multiplicity factor W, we rewrite $\log(W)$ assuming N large. We use the Stirling approximation

$$\log(N!) = N \log(N) - N + \sqrt{2\pi N} + \frac{1}{12N} + O(1/N^2) \,. \tag{2.35}$$

Thus,

$$\begin{aligned} \log W &= N \log N - m_1 \log m_1 - \ldots - m_n \log m_n \\ &+ \sqrt{2\pi N} - \sum_{i=1}^{n} \sqrt{2\pi m_i} + \frac{1}{12N} - \sum_{i=1}^{n} \frac{1}{12 m_i} \qquad (2.36) \\ &+ O(1/N^2) - \sum_{i=1}^{n} O(1/m_i^2) \\ &= -\sum_{i=1}^{n} m_i \log \frac{m_i}{N} + \sqrt{2\pi N} - \sum_{i=1}^{n} \sqrt{2\pi m_i} \\ &+ \frac{1}{12N} - \sum_{i=1}^{n} \frac{1}{12 m_i} + O(1/N^2) - \sum_{i=1}^{n} O(1/m_i^2) \qquad (2.37) \end{aligned}$$

and as N and m_i go to infinity in such a way that $\frac{m_i}{N} \to p_i$ most of the terms in (2.37) tend to zero and we obtain

$$\frac{1}{N} \log(W) \to -\sum_{i=1}^{n} p_i \log p_i \,. \tag{2.38}$$

So, the distribution which is most likely to arise, or the one which can arise in the greatest number of ways W, is also the one which maximizes the entropy as defined in (2.24).

2.7.1 The general maximum entropy problem and its solution

Consider a problem where we have knowledge of mean values F_k of certain functions, $f_k(\cdot)$, of data. We are now to determine a probability distribution with expectations matching the measured mean values:

$$\sum_{i=1}^{n} p_i f_k(x_i) = F_k \ , \quad 1 \leq k \leq m \tag{2.39}$$

where p_i denotes the probability for each possible 'state of nature', x_i, indexed by $i \in \{1..n\}$.

We wish to find the probabilities p_i, for all possible i, which maximize the entropy

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{2.40}$$

subject to the constraints (2.39). This is a standard variational problem solvable by using Lagrange multipliers when $m < n$. In Appendix 2.B it is shown that using the partition function (Jaynes, 1957a)

$$Z(\lambda_1, \ldots, \lambda_m) \equiv \sum_{i=1}^{n} \exp\left[-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\right] \tag{2.41}$$

we have the formal solution

$$p_i = \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp\left[-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\right] , \tag{2.42}$$

where $\{\lambda_k\}$ are the Lagrange multipliers which are chosen so as to satisfy the constraints (2.39). This is the case when

$$F_k = -\frac{\partial}{\partial \lambda_k} \log Z, \quad 1 \leq k \leq m \ . \tag{2.43}$$

In (2.39) - (2.43) we have the general maximum entropy problem and its solution. It should be noted that the solution presented here automatically includes the constraint $\sum_{i=1}^{n} p_i = 1$ without need for an additional Lagrange multiplier.

The maximum entropy distribution (2.42) has the entropy

$$H_{\text{max}} = \lambda_0 + \sum_{j=1}^{m} \lambda_j F_j \ , \tag{2.44}$$

where $\lambda_0 = \log(Z(\lambda_1, \ldots, \lambda_m))$. Generally speaking, large values of the λ_i thus indicates large uncertainty.

It can further be shown (Jaynes, 2003) that the covariances between the different functions f_j, f_k obey the following relations,

$$\langle f_j f_k \rangle - \langle f_j \rangle \langle f_k \rangle = \frac{\partial^2 log Z}{\partial \lambda_j \partial \lambda_k} = -\frac{\partial \langle f_j \rangle}{\partial \lambda_k} = -\frac{\partial \langle f_k \rangle}{\partial \lambda_j} \ . \tag{2.45}$$

Here we give three common maximum entropy distributions obtained from different constraints (2.39).

EXAMPLE 2.1 No constraints

With no constraints except that the probability distribution should sum to unity, there are no Lagrange multipliers and the maximum entropy distribution is uniform over the space of all possible outcomes.

EXAMPLE 2.2 Mean and variance constraints

Using the following constraints

$$\mu = \int_{-\infty}^{\infty} p(x \mid I) x dx \tag{2.46}$$

$$\sigma^2 = \int_{-\infty}^{\infty} p(x \mid I)(x - \mu)^2 dx \ , \tag{2.47}$$

i.e. fixing the mean and the variance of a continuous distribution, the maximum entropy distribution is Gaussian

$$p(x \mid I) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{1}{2\sigma^2}(x - \mu)^2 \right\} . \tag{2.48}$$

In Figure 2.1 we plot this distribution for different standard deviations σ.

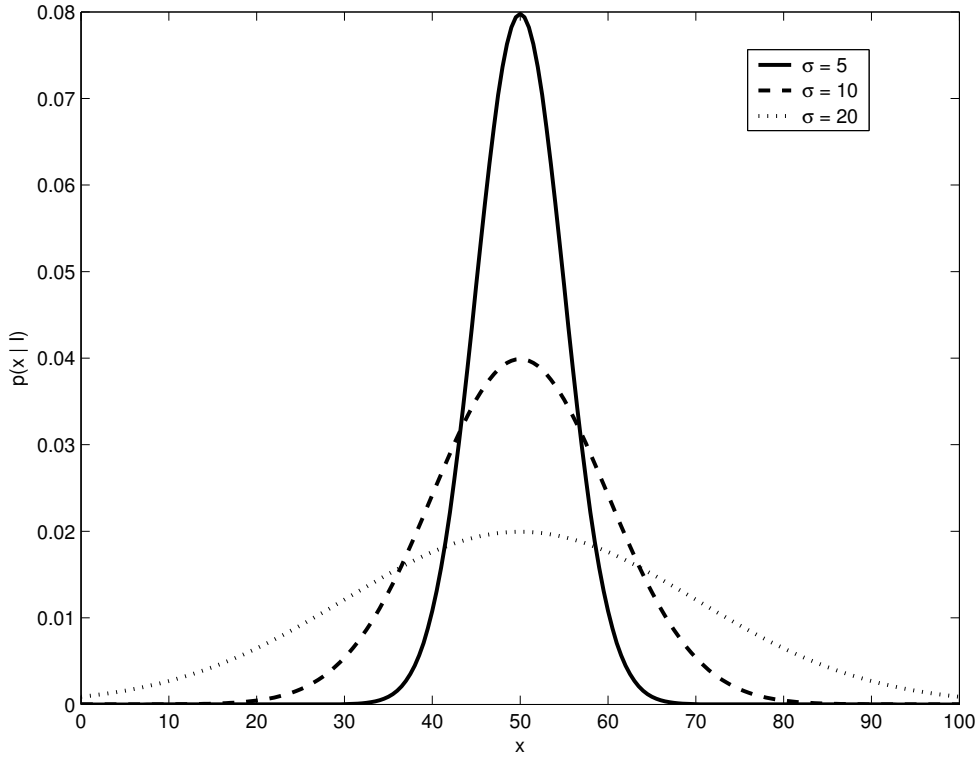

Figure 2.1: The maximum entropy probability distribution for a continuous variable x with known mean (here $\mu = 50$) and known standard deviation σ is Gaussian.

EXAMPLE 2.3 Mean and mean logarithm constraints

With knowledge of the mean and the mean of the logarithm of a non-negative variable,

$$F_1 = \int_0^\infty p(x \mid I) x dx \tag{2.49}$$

$$F_2 = \int_0^\infty p(x \mid I) \ln x dx \,, \tag{2.50}$$

the maximum entropy distribution becomes

$$\begin{aligned} p(x \mid I) &= \frac{1}{Z(\lambda_1, \lambda_2)} \exp\left(-\lambda_1 x - \lambda_2 \ln x\right) \\ &\propto x^{-\lambda_2} \exp\left(\lambda_1 x\right). \end{aligned} \tag{2.51}$$

This is on the same form as the Gamma distribution, and if we write $\gamma = 1/\lambda_1$ and $\alpha = 1 - \lambda_2$ we obtain the Gamma distribution in the conventional form

$$p(x \mid I) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\gamma^{\alpha}} \exp\left(\frac{-x}{\gamma}\right) , \tag{2.52}$$

where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the Gamma function. Special cases of the Gamma distribution include the exponential distribution and the chi-square distribution.

2.7.2 The entropy concentration theorem

For those who adhere to the frequency interpretation of probabilities, the rationale above hardly makes any sense. A valid motivation must refer to actual frequencies in random experiments. We here show that when the notion of a repeated experiment is valid, there is such a correspondence between maximum entropy distributions and frequencies. Then, having established such a correspondence, we will show a remarkable property of maximum entropy distributions which make such frequency correspondences irrelevant for the subsequent inference.

Consider an experiment that has been performed N times, each with n possible outcomes $x_1 \dots x_n$. Suppose that the only information we receive about the experimental outcomes are the sample mean values F_k of m ($m < n - 1$) functions of the observations,

$$F_k = \sum_{i=1}^{n} \frac{n_i}{N} f_k(x_i), \qquad k = 1 \dots m , \tag{2.53}$$

where n_i denotes the number of trials that yielded the outcome x_i.

The mean values F_k (2.53) are insufficient to determine the actual frequencies $g_i = n_i/N$ with which each outcome x_i occurred. But out of the n^N conceivable outcomes, how many would lead to any particular numbers n_i? The answer is again given by the multinomial coefficient

$$W = \frac{N!}{n_1! \cdots n_n!} = \frac{N!}{(Ng_1)! \cdots (Ng_n)!} . \tag{2.54}$$

The frequencies which can arise in the greatest number of ways maximizes W, or equivalently maximizes $\frac{1}{N}\log(W)$, which when n_i and N tend to infinity in such a way that $g_i = n_i/N \to p_i$ becomes

$$\frac{1}{N}\log(W) \to -\sum_{i=1}^{m} p_i \log(p_i) . \tag{2.55}$$

This result if obtained by using the same approximation as in the Wallis derivation.

If we compare the number of ways W that the frequency distribution maximizing (2.55) can arise in, with another distribution p' having entropy H' and multiplicity W', we see directly from (2.55) that the maximum at W becomes enormously sharp as N grows,

$$\frac{W}{W'} \to \exp\left\{N(H - H')\right\} \ . \tag{2.56}$$

We can now check how large the entropy deviation ΔH must be from the maximum to cover a certain percentage of the class of all possible outcomes constrained to satisfy (2.39). A certain fraction F of the class C of possible outcomes will yield an entropy in the range

$$H_{\text{max}} - \Delta H \leq H(p_1 \ldots p_n) \leq H_{\text{max}} \ . \tag{2.57}$$

Just how large must ΔH be to cover a certain fraction F? The following theorem gives the answer.

Theorem 2.1 (Jaynes (1982)) *Asymptotically, as $N \to \infty$, $2N\Delta H$ is chi-square distributed with $k = n - m - 1$ degrees of freedom according to*

$$2N\Delta H = \chi^2_k(1 - F) \ , \tag{2.58}$$

where $\chi^2_k(1 - F)$ denotes the critical chi-square value for k degrees of freedom at the $100(1 - F)$ percent significance level.

EXAMPLE 2.4 Entropy concentration for throwing dice

Suppose that we toss a die 1000 times, i.e. $n = 6$, $N = 1000$. If we have no information concerning the outcomes, then the maximum entropy distribution (2.42) is uniform (see Example 2.1) with $H_{\text{max}} = \ln 6 = 1.792$. From Theorem 2.1 we find that out of all distributions with $k = 6 - 1 = 5$, $100F = 99.5$ percent of them have an entropy in the range $2N\Delta H = 11.07$, or $1.783 \leq H \leq 1.792$.

Thus, if we would assign any distribution with entropy less than 1.783 we would reside in a tiny subset of all possible outcomes. In order to do so with a reason, we would certainly need strong evidence to support our choice.

Note that the entropy concentration theorem is only a combinatorial statement, expressing only a counting of the possibilities. It does not become a statement of probabilities unless we assign (by the principle of indifference) equal probability to each outcome in class C.

2.7.3 Frequency irrelevance and sufficiency

We noted that there are instances where there is a clear demonstrable frequency correspondence between frequencies in repeated experiments and probabilities assigned by the maximum entropy principle. We will now turn our attention to the question whether such frequency correspondence is required or even useful.

We first state a formal property of maximum entropy distributions that says that the class of distributions with *sufficient statistics* is exactly the class of maximum entropy distributions. The sufficient statistics of the resulting maximum entropy distribution are the same functions of data, whose mean values (2.39) we know and, thus, which constrain the entropy. Hence, the values of these functions are the only properties of the data that our inferences will depend on.

Definition 2.1 (Jaynes (2003), Kullback (1968)) *If the likelihood $P(D \mid \theta I)$ for the parameter θ factors in the form*

$$P(D \mid \theta I) = f(T(D) \mid \theta I)g(D) \tag{2.59}$$

where $T(D)$ is some function of the data, then $T(D)$ is called a sufficient statistic for the parameter θ.

Note that this means that any posterior inference about a parameter θ involving a sufficient statistic $T(D)$ depends only on the data through the function $T(D)$ since the posterior probability $P(\theta \mid DI)$ is a function of the data D only through the likelihood (the factor $P(D \mid I)$ being only a normalization constant). No other properties of the data will affect the inference. The definition generalizes immediately to the case where there are m jointly sufficient statistics $T_k(D)$ for some multi-dimensional parameter vector θ,

$$P(D \mid \theta I) = f(T_1(D), \ldots, T_m(D) \mid \theta I)g(D)\ . \tag{2.60}$$

Likewise, if there are two parameters θ_1, θ_2 and we can write the likelihood as

$$P(D \mid \theta_1\theta_2 I) = f(T_1(D), \mid \theta_1 I)h(T_2(D), \mid \theta_2 I)g(D)\ , \tag{2.61}$$

$T_1(D)$ is a sufficient statistic for θ_1 and $T_2(D)$ is a sufficient statistic for θ_2.

Theorem 2.2 (Kullback (1968)) *The class of all maximum entropy distributions (2.42) is exactly the class of all distributions with sufficient statistics. The sufficient statistics are given by the constraint functions (2.39),*

$$T_k(D) = f_k(D), \qquad k = 1 \ldots m \tag{2.62}$$

The implication of this result may not be immediately obvious. Interpreted in the framework of probability theory as logic it means that if we assign a probability distribution with maximum entropy subject to constraints on the expectations of some functions f_k, we are in effect demanding that our inferences shall depend only on these functions of the data. We can thus choose to base our inferences on any particular feature of the data that we can measure, and then make the least biased inference possible based on knowledge only of this. Of course, honesty requires that if the measured mean values are based only on a few data points, we must marginalize the resulting distribution with respect to the unknown true value.

This further implies that any long-run frequency correspondence is uncritical. Indeed, it is a subtle but important insight that *the actual long-run frequencies are made irrelevant by using maximum entropy distributions*. Since no other aspects (including frequencies with which different outcomes occur) in our data than the sufficient statistics will affect our inference, the actual frequencies will have no impact whatsoever on our resulting conclusions. This again shows that asking for a true probability distribution, or trying to 'estimate' it as were it a real property of nature, simply misses the point of what we are aiming for; to make the best, i.e. least biased, inferences from incomplete data.

If the reasoning above was not entirely clear, the following example given by Jaynes (2003) may help in understanding the role of sufficient statistics.

EXAMPLE 2.5 The success of Gaussian distributions – making frequencies irrelevant

---

Consider a problem where our observations y_i are modelled as

$$y_i = \mu_0 + e_i \,, \qquad i = 1 \ldots n \tag{2.63}$$

where μ_0 is an unknown location parameter and there is some unknown additive disturbance e_i which may vary from observation to observation. Our problem is to estimate μ_0 from the observations. Let us suppose that we *assign* an independent Gaussian distribution with mean zero and variance σ^2 for each of the disturbance terms e_i. Note that we are not assuming that the actual e_i are distributed in frequency according to our assignment. Rather, we will now investigate the consequences of our assignment, irrespective of how the frequencies of different values of e_i are actually distributed. Recall from Example 2.2 that the Gaussian distribution is the maximum entropy distribution subject to fixing the mean and the variance, and that according to Theorem 2.2 this means that the mean and the variance become sufficient statistics.

First, note that the actual estimation error $\Delta = \hat{\mu}_0 - \mu$ can only depend

on properties of the actually obtained data set $y_1 \dots y_n$. Frequencies, or other properties, in other data sets that could have, but in fact were not, observed can have no influence on the accuracy of our estimate. Thus, the long-run frequencies in an imagined ensemble of trials have no effect on our estimation accuracy.

We will use as our estimate the value $\hat{\mu}_0$ that maximizes the likelihood $P(y_1 \dots y_n \mid \mu I)$. If we assign a uniform prior for μ_0 this is identical to the maximum a posteriori estimate and thus seems to make sense (we shall however come back to the problem of choosing a reasonable course of action in a later section). The probability for obtaining a certain y_i given knowledge of μ_0 (i.e. the likelihood) is equal to the probability for obtaining a certain disturbance $e_i = y_i - \mu_0$, i.e.,

$$P(y_1 \dots y_n \mid \mu_0 I) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_0)^2\right\} , \tag{2.64}$$

and since we have that

$$\sum_{i=1}^{n}(y_i - \mu_0)^2 = n\left[(\mu_0 - \bar{y})^2 - s^2\right] \tag{2.65}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \mu_0 + \bar{e} , \tag{2.66}$$

$$s^2 = \overline{y^2} - \bar{y}^2 = \overline{e^2} - \bar{e}^2 \tag{2.67}$$

and

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i , \qquad \overline{e^2} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 , \tag{2.68}$$

the only properties of the data that can matter for our inference about μ_0, i.e. the sufficient statistics, are the first and second moments of the data.

The estimate that maximizes the likelihood is the arithmetic mean $\bar{y}$ of the observations. Using that estimate, the estimation error is

$$\Delta = \bar{y} - \mu_0 = \bar{e}, \tag{2.69}$$

the arithmetic mean of the actual disturbances in our data set. The total squared error is

$$\Delta^2 = (\bar{y} - \mu_0)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} e\right)^2 , \tag{2.70}$$

the expectation of which is identical to σ^2/n if we adopt our Gaussian assignment. The interesting result here is that the estimation error Δ is independent of the frequency distribution. Whether the actual errors are distributed accorded to a Gaussian histogram or not does not matter. The actual estimation error that we make is always exactly the arithmetic mean of the actual disturbances. Therefore, the true importance of the Gaussian probability assignment lies in the fact that it renders the actual frequencies irrelevant to the inference. Only the sufficient statistics have any effect on the estimate or its accuracy.

This example should make it clear that statements such as 'A Gaussian error distribution should not be used because we know that the actual errors are not Gaussian' are flawed in that they fail to realize which criteria are important in assigning a probability distribution. It is not frequency correspondence, but demonstrable information content, that is the valid criterion. Therefore, one should be careful in dismissing a probability distribution because the shape of the curve seems strange. One should instead assess the effects on the inference that the curve has.

2.7.4 A caveat – continuous variables

The above treatment of the maximum entropy principle was based on discrete variables. Shannon's derivation does not go through for continuous variables. Instead, Jaynes (1963a) derived the correct entropy expression for continuous variables by starting from the discrete expression for entropy and letting the points become more numerous. As the number of points increase, the density of points approaches a definite function $m(x)$ according to

$$\lim_{n\to\infty} \frac{1}{n}(\text{number of points in } a < x < b\ ) = \int_a^b m(x)dx\ . \tag{2.71}$$

The discrete probability distribution p_i tends to a continuous probability density $p(x \mid I)$ according to

$$p_i = p(x_i \mid I)(x_{i+1} - x_i)\ , \tag{2.72}$$

and supposing that the difference between any adjacent points will tend to zero in the manner

$$\lim_{n\to\infty} (n(x_{i+1} - x_i)) = (m(x_i))^{-1}, \tag{2.73}$$

the discrete probability distribution will tend into

$$p_i \to \frac{p(x_i \mid I)}{nm(x_i)}\ . \tag{2.74}$$

Hence, the discrete entropy (2.24) tends to the limiting expression

$$H \to H_c = -\int p(x \mid I) \log\left(\frac{p(x \mid I)}{nm(x)}\right) dx \,. \tag{2.75}$$

The $\log(n)$ term is a constant and can be subtracted. We then take the following expression as our continuous measure of uncertainty:

$$H_c = -\int p(x \mid I) \log\left(\frac{p(x \mid I)}{m(x)}\right) dx \,. \tag{2.76}$$

The continuous maximum entropy problem now becomes to find a probability density $p(x \mid I)$ that maximizes (2.76), constrained by information regarding the mean values

$$F_k = \int f_k(x) p(x \mid I) dx \,, \qquad k = 1 \ldots m \tag{2.77}$$

where the F_k are known numerical values. The solution obtained by maximizing (2.76) is

$$p(x \mid I) = \frac{m(x)}{Z(\lambda_1, \ldots, \lambda_m)} \exp\left[-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\right] \,, \tag{2.78}$$

where the partition function becomes

$$Z(\lambda_1, \ldots, \lambda_m) = \int m(x) \exp\left[-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\right] dx \tag{2.79}$$

and the Lagrange multipliers λ_k are given by the m equations

$$F_k = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k}, \quad 1 \leq k \leq m \,. \tag{2.80}$$

Apparently, if we have no constraints on the probability density there are no λ_k in (2.78), and the maximum entropy distribution is equal to

$$p(x \mid I) = \left(\int m(x) dx\right)^{-1} m(x) \,. \tag{2.81}$$

We are now left with the question of how to determine the density $m(x)$. Since it is the most uninformative density that we can obtain, the role of $m(x)$ is to define a completely ignorant distribution. We must therefore seek to answer the question: What is meant by complete ignorance concerning the variable x?

Jaynes offers the following solution. Complete ignorance concerning a certain parameter can very often be stated in terms of an invariance under some specific

parameter transformation. For example, consider the Gaussian density, containing two parameters: the expectation μ and the standard deviation σ. Suppose that we are to determine a probability assignment for μ expressing complete ignorance. It seems appropriate to express ignorance of μ by stating that the probability density function (pdf) for μ should be equal to that for a transformed variable μ' according to

$$\mu' = \mu + a \,, \tag{2.82}$$

i.e. we are saying that a shift of location does not change our state of knowledge. This is to say that shifting the origin does not change our pdf assignment. We are equally ignorant of μ. If that were not true, then we must have had some cogent information concerning the location, and thus we are not completely ignorant in this sense. If our ignorance concerning μ is thus expressed as translation invariance then we have that

$$p(\mu \mid I)d\mu = p(\mu + a \mid I)d(\mu + a) \tag{2.83}$$

and since a is constant $d(\mu + a) = d\mu$. Thus, the only pdf that satisfies (2.83) is

$$p(\mu \mid I) = \text{constant.} \tag{2.84}$$

A parameter for which this translation invariance property can be used to express ignorance about is appropriately described as a location parameter. In general, if we can write a pdf as

$$p(x \mid \mu\sigma I) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \tag{2.85}$$

then we call μ a *location parameter* and σ a *scale parameter*.

The standard deviation σ in our current problem is a scale parameter. A scale parameter refers to a size or a magnitude, something which describes the scale of something; for instance, the gain of a mobile radio channel, or the length of a molecule. Complete ignorance of a scale parameter, as Sivia (1996) vividly points out, must mean that in a plot our pdf should be invariant to any shrinking or stretching of the horizontal axis. The scale of the problem is unknown, it could equally well be centimeters as meters. Thus, this invariance can be expressed as

$$p(\sigma \mid I)d\sigma = p(b\sigma \mid I)d(b\sigma) \,, \tag{2.86}$$

where b is an arbitrary positive number. Since $d(b\sigma) = bd\sigma$, the only pdf which satisfies (2.86) is

$$p(\sigma \mid I) \propto \frac{1}{\sigma} \,. \tag{2.87}$$

This strange-looking pdf is known as Jeffreys' prior[9], and was used extensively by him (see for example Jeffreys, 1939). It may look less peculiar when we realize that (2.87) is equivalent to a uniform pdf on $\log(\sigma)$. In principle, in order to normalize this pdf, one should always confine σ to lie between a non-zero lower bound and finite upper bound.

2.8 Information Measures and the Shannon Capacity

In judging the merits of a communications system, it would be useful to be able to measure the amount of information sent over the link. Moreover, we would benefit from knowing whether there is a limit on how much information can be sent over a link, and in that case what the actual value of that limit is.

Claude Shannon (1948) considered these issues in his classic work on communications. We here derive the basic blocks of his theory from our present perspective that probabilities are the fundamental carriers of incomplete information.

Suppose that we are interested in knowing the value of some quantity X having N possible distinct outcomes, $x_1 \ldots x_N$. Now, if we instead of X are given the value of some other (hopefully related) quantity Y with M mutually exclusive possible outcomes $y_1 \ldots y_M$, how much information does that provide us about any specific outcome x_i? If the value of Y is y_j then an intuitive measure of the information in y_j about x_i for someone with prior information I is the change in probability upon receiving information y_j:

$$K_i \triangleq \frac{p(x_i \mid y_j I)}{p(x_i \mid I)} \; . \tag{2.88}$$

Notice that using Bayes' rule gives that K_i is symmetric,

$$K_i = \frac{p(x_i \mid y_j I)}{p(x_i \mid I)} = \frac{p(y_j \mid x_i I)}{p(y_j \mid I)} = K_j \tag{2.89}$$

so we can suppress the index, and simply denote the information by K. The ratio of the posterior probability and the prior probability can take any non-negative value, so we can equally well work with the more convenient choice of logarithms:

$$\log K = \log p(x_i \mid y_j I) - \log p(x_i \mid I) \; . \tag{2.90}$$

Suppose now that we wish to evaluate $\log K$ without knowing the actual outcome of X. We then take the expectation of $\log K$ as a reasonable guess and denote

[9] Although Jeffreys was its main advocate and first suggested its use, Haldane has been attributed (Howie, 2002) as providing an early motivation which reinforced Jeffreys' belief in it.

this as the prior information in $Y = y_j$ about an unknown outcome of X,

$$\sum_{i=1}^{N} p(x_i \mid I) \log K = \sum_{i=1}^{N} p(x_i \mid I) \log p(x_i \mid y_j I) - \sum_{i=1}^{N} p(x_i \mid I) \log p(x_i \mid I) \; . \tag{2.91}$$

Similarly, knowing the outcome $X = x_i$ but not the value of Y we take the posterior expectation of $\log K$,

$$\begin{aligned} \sum_{j=1}^{M} p(y_j \mid x_i I) \log K &= \sum_{j=1}^{M} p(y_j \mid x_i I) \log p(x_i \mid y_j I) \\ &\quad - \sum_{j=1}^{M} p(y_j \mid x_i I) \log p(x_i \mid I) \\ &= \sum_{j=1}^{M} p(y_j \mid x_i I) \log p(x_i \mid y_j I) - \log p(x_i \mid I) \; . \end{aligned} \tag{2.92}$$

This can be taken as the average information in an observation from Y about a particular x_i.

We shall finally take the average information $\mathcal{I}$ in an observation of Y about an observation of X as the joint average of $\log K$,

$$\begin{aligned} \mathcal{I} &\triangleq \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i y_j \mid I) \log p(x_i \mid y_j I) - \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i y_j \mid I) \log p(x_i \mid I) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i y_j \mid I) \log p(x_i \mid y_j I) - \sum_{i=1}^{N} p(x_i \mid I) \log p(x_i \mid I) \; . \end{aligned} \tag{2.93}$$

Another common name for the average information (2.93) is the *mutual information*.

Notice that the average information is thus the entropy for X (2.24) less the mean conditional entropy for X given Y (2.26),

$$\mathcal{I}(X, Y) = H(X) - H(X \mid Y) \; . \tag{2.94}$$

We can thus interpret the average information as the prior uncertainty minus the average posterior uncertainty.

The entropy expressions were here obtained from a more intuitive departure point given by our basic probability rules, while the original entropy derivation based on the desiderata in the previous section is perhaps more elegant and has a

more explicit motivation. Side by side, both derivations add to the understanding of the expressions.

If we agree to use (2.93) as a measure of information, then a natural communication-theoretic interpretation is that on receiving data Y the recipient is on average obtaining information corresponding to an amount $\mathcal{I}(X, Y)$ concerning the transmitted message X. On receiving a particular datum y_j however, the information (2.91) about the transmitted message can be larger or smaller than the average information.

A natural goal for a communication link is to maximize the average information transmitted[10]. That is carried out by making $H(X)$ as large as possible, i.e. coding messages so that they are as uniform as possible, while keeping $H(X \mid Y)$ as small as possible, i.e. making messages as easy to decode as possible given the received data. Clearly, minimizing $H(X \mid Y)$ can be carried out by simply repeating the message forever, but that does not constitute a good system design. We should instead maximize the obtained information per utilized resource, such as the *information rate*, i.e. the information received per second.

Shannon showed that if the entropy $H(X)$ for the transmitted message is less than a number C, the capacity of the channel, then reception at an arbitrarily small error rate is possible. The capacity of the channel defined by $p(X \mid Y, I)$ is defined as

$$C \stackrel{\triangle}{=} \max_{p(X|I)} \mathcal{I}(X, Y) \; . \tag{2.95}$$

If $H(X \mid Y) > C$ then error-free reception is not possible. (Shannon further gives explicit bounds on the mean conditional entropy $H(X \mid Y)$ for this case.) Shannon's theorem does however not tell us how to construct a system which achieves the channel capacity, and it does not guarantee that such a system does not impose infinite coding and decoding delays.

Thus, our questions in the beginning of this section were all answered by Shannon. It turned out that there is a strict upper limit on information transfer, which can be stated in an abstract mathematical language valid for arbitrary communication channels. As we have indicated here in our derivation of the information measures used by Shannon, the connection to probability theory is very strong. Indeed, we expected this based on Jaynes' derivation of probabilities as carriers of incomplete information. We think that this approach helps to understand the generality of Shannon's theorem above. It is applicable not only to man-made telephony systems; it is a fundamental constraint on any information exchange between any entities, and constrains interactions in biological and physical systems as well.

[10]It should be emphasized that it is the information averaged over Y and X that should be maximized, as we are typically designing communications systems which should be used to send not just a specific message, but any conceivable message.

Having said all this, we must finally point out a problem in the reasoning above. Maximizing average information is surely a good system-wide approach in communications, but for any individual receiver, the entropy $H(X)$ for the transmitted message simply *is what it is*! It cannot be adjusted by anything else than receiving information. So, for a receiver, the critical property is again only the posterior probability $p(X \mid Y = y_j I)$ as in all problems of inference. The average information is useless in inferring the actual message. The basic desiderata of this chapter ensure us that all information relevant to the question being asked is always taken into account fully and in the only consistent way possible if we use probability theory as logic. There is no need for any additional *ad hoc* rules in decoding received messages. Although Shannon's communication theory is an essential tool for analysis of all man-made and naturally occurring communications systems, it does not provide a rationale for making the optimal individual inferences and should not be construed as such. It provides important performance measures and shows certain critical limits of communications systems stated in terms of these performance measures. The importance lies in its analytical tools, rather than in providing constructive rules.

2.9 Decision Making in the Face of Uncertainty

In the sections preceding the previous one we have considered how our knowledge about an arbitrary uncertain event is updated with new information. When all is said and done, however, we typically have to take some definite decision based on all relevant information at hand. Nothing in our rules tells us how to do this. Probability theory is only concerned with describing a state of knowledge; it does not give any rules for which decision to make in a given situation. A moment's reflection shows that in order to make a rational decision we must consider what the effects of our decision will be given different possible outcomes. There is thus a certain amount of subjectivity involved in decision making, since it includes making value judgements. For example, what is more worth to me, choosing a more expensive apartment at a better location, or saving money but having to spend more time commuting? Of course, probability theory cannot determine that. But this line of thought implies a reasonable course of action: Determine a loss function $L(d_i, \theta_k)$ (or equivalently a utility function) describing the 'loss' incurred from making decision d_i should θ_k turn out to be the true 'state of nature'. Having determined numerically how large the loss for different decision-outcome combinations will be, the only remaining uncertainty resides in the outcomes θ_k. Thus, we must work out the probabilities for the respective states θ_k given the data D and any other relevant information I. A reasonable decision d_i then minimizes

the expected loss,

$$\langle L \rangle = \sum_k L(d_i, \theta_k) p(\theta_k \mid DI) \; , \tag{2.96}$$

which is a function of the decisions d_i. Of course, this generalizes in the obvious way to an integral over a pdf in the continuous case,

$$\langle L \rangle = \int L(d_i, \theta) p(\theta \mid DI) d\theta \; . \tag{2.97}$$

2.9.1 Parameter estimation

We now consider the problem of estimating a parameter, i.e. to guess the actual value of some parameter given whatever data and information that we might have. We can view this as a decision problem; we wish to make a decision as to the true parameter value which in some sense minimizes the bad effects (for instance the estimation error) of that choice. Estimating an unknown continuous-valued parameter α, the expected loss can be minimized by setting the derivative of (2.97) with respect to the estimate $\hat{\alpha}$ equal to zero,

$$\begin{aligned} \frac{\partial \langle L(\hat{\alpha}, \alpha) \rangle}{\partial \hat{\alpha}} &= \frac{\partial}{\partial \hat{\alpha}} \int L(\hat{\alpha}, \alpha) p(\alpha \mid DI) d\alpha \\ &= \int \frac{\partial L(\hat{\alpha}, \alpha)}{\partial \hat{\alpha}} p(\alpha \mid DI) d\alpha = 0 \; , \end{aligned} \tag{2.98}$$

where the order of differentiation and integration could be changed since the boundaries of the integral are independent of α (eq. 12.211 in Gradshteyn and Ryzhik, 2000).

According to (2.98), the expectation of

- a quadratic loss $L = (\hat{\alpha} - \alpha)^2$ is minimized if

$$\begin{aligned} & \int (\hat{\alpha} - \alpha) p(\alpha \mid DI) d\alpha = 0 \\ \Leftrightarrow \quad & \hat{\alpha} = \int \alpha p(\alpha \mid DI) d\alpha \end{aligned} \tag{2.99}$$

  which corresponds to using $\hat{\alpha} = \langle \alpha \rangle$, the expectation of the parameter over the posterior pdf.

- the absolute error $L = |\hat{\alpha} - \alpha|$ corresponds to using $\hat{\alpha} = \alpha_{\text{med}}$ where α_{med} is the median over the posterior pdf for α since the median α_{med} is defined by $\int_{-\infty}^{\alpha_{med}} p(\alpha \mid DI) d\alpha = \int_{\alpha_{med}}^{\infty} p(\alpha \mid DI) d\alpha = 0.5$. The median has

the interesting property that it is invariant under any monotonic transformation $f(\alpha)$. It is thus insensitive to the exact form of the posterior pdf, and consequently also to outliers.

- a loss function which only cares about being exactly right, represented by $L(\hat{\alpha}, \alpha) = 0$ if $\hat{\alpha} = \alpha$ and $L(\hat{\alpha}, \alpha) = 1$ otherwise, results in using the maximum of the posterior density for α as an estimate. Note that this common choice considers any error, regardless of size, to be equally bad. It can further be observed that this is equivalent to using a loss function which is $L = |\hat{\alpha} - \alpha|^k, k \to 0$.

2.9.2 Other approaches

Another criterion for decision making, and parameter estimation in particular, which is common in the random-variable approach to probability theory is to minimize a quantity R called the *risk*

$$R = \int \dots \int L(\hat{\alpha}, \alpha) p(x_1 \dots x_n \mid \alpha I) dx_1 \dots dx_n \tag{2.100}$$

where $x_1 \dots x_n$ denotes the observed data consisting of n samples. The loss is thus not averaged over the posterior pdf with respect to the parameter, but rather over the likelihood with respect to the data. This means that in general, the best estimate $\hat{\alpha}$ according to this criterion may depend on the actual unknown parameter value. Another severe problem with this approach is that the minimum of (2.100) cannot be found by variational methods (see e.g. Jaynes, 2003, Chapter 13), and thus we cannot in general find a truly best estimator by this criterion. Why then, would anyone still wish to use (2.100) as a criterion? Van Trees (1968) (p. 63) motivates it since 'in many cases it is unrealistic to treat the unknown parameter as a random variable'. Again, it is the fallacy to project ones own uncertainty onto nature, assuming that a probability for a parameter implies that the parameter must in fact be random by nature, that forbids the use of the expected posterior loss (2.97) as a valid criterion.

Since we cannot find a useful estimator based on (2.100), the conventional approach to estimating a 'non-random' parameter is to come up with a few candidate estimators and then compare them in terms of risk (and most often this last step is not even carried out). A common approach is to use the value of the parameter which maximizes the likelihood, i.e. the probability for the observed data, as the estimator. With a uniform prior for the parameter and a loss which does not care about the size of the error, this coincides with the Bayesian approach given above.

Nevertheless, from the definition of the risk (2.100) one can insert some commonly used loss function and then see what the risk becomes. With a quadratic

loss $L = (\hat{\alpha} - \alpha)^2$, we obtain

$$\begin{aligned} R &= \int \dots \int (\hat{\alpha} - \alpha)^2 p(x_1 \dots x_n \mid \alpha I) dx_1 \dots dx_n \\ &= \langle \hat{\alpha}^2 \rangle + \alpha^2 - 2\alpha \langle \hat{\alpha} \rangle = (\alpha - \langle \hat{\alpha} \rangle)^2 + \text{var}(\hat{\alpha}) \,, \end{aligned} \tag{2.101}$$

where $\text{var}(\hat{\alpha}) = \langle \hat{\alpha}^2 \rangle - \langle \hat{\alpha} \rangle^2$ is the variance of the likelihood for the estimator (remember that the estimator is just a function of the data, and we can therefore speak of an expectation of the estimator in this sense over the probability for the data given the parameter). A good estimator in the sense of low risk should thus have two properties:

1. $\langle \hat{\alpha} \rangle = \alpha$
2. minimum $\text{var}(\hat{\alpha})$.

An estimator satisfying the first condition is called *unbiased* in the random-variable literature, and an estimator with both properties (1) and (2) is called *efficient* or an *unbiased minimum variance* estimator. Of course, to tell whether an estimator is unbiased or not, we need to know the true parameter value, which seems like a rather bizarre condition. It is also important to remember that both the bias term and the variance term are equally important but hardly ever independent. An estimator which is made unbiased typically increases the variance at the same time and may lead to an overall larger mean squared error. The term 'unbiased' is thus misleading in that it may lead us into thinking that an unbiased estimate is more objectively valid in some sense; on the contrary, an unbiased estimate may perform worse than a biased one in the sense of increasing the risk.

Another approach would be to consider some other function of the loss, rather than its expectation. For instance, why not make the decision which minimizes the *maximum* loss? If some intelligent opponent foresees our decision and makes sure that the consequences of that decision will always be the least favorable possible, then this would be a reasonable criterion. This *mini-max* criterion is therefore not uncommon in game theory, but note that this criterion assumes that we face a player with perfect skills who always makes the best possible decisions. In reality this is over-pessimistic. Even if we have information that tells us the loss is always maximized subject to our decisions, then a probability distribution for the unknown outcome of our decision would reflect that; consequently the expected loss would equal the maximum loss. Thus, the expected loss criterion contains the mini-max criterion as a very special case. In most situations, however, they differ since it would be overly pessimistic to assume that whatever our decisions are, their consequences will invariably be the worst possible.

Similarly, an incurable optimist would make decisions that minimize the minimum loss; and again we see that this is a special case of the minimum expected loss criterion when our information gives us reason to believe that Nature is in its most helpful mood. It seems that whenever a criterion different from expected loss is suggested, it either reflects that the person who made the suggestion actually means that another loss function should be used, or, that that person does not allow probabilities to reflect information.

In contrast to the random-variable approach, the Bayesian approach to decision making under uncertainty is always the same, and, to summarize, consists of the following five steps:

1. Enumerate the possible states of nature and the possible decisions.
2. Determine the loss function for all combinations of decisions and outcomes.
3. Assign prior probabilities for the uncertain variables using the maximum entropy principle.
4. Digest any additional information or data by the use of Bayes' theorem.
5. Make the decision which minimizes the expected loss.

2.10 Comments

This chapter has given a brief introduction to probability theory from a perspective quite distinct from that conventionally taught at schools and universities. Historically, however, the early workers in probability theory seem to have held a view in line with that expounded here. Laplace, for instance, who made many of the most important early contributions remarked that 'probability theory is nothing but common sense reduced to calculation'. The great physicist Maxwell wrote the following in a letter in 1850.

> The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

It was first in the latter part of the 19th century that the frequentist interpretation became common, at a time when pure mathematicians started to dominate the subject. Their goal was quite different from that of physicists such as Laplace or Maxwell, and at this time focus started to shift from making inferences based on incomplete

data to proving limiting theorems of idealized 'random experiments'. But the subjective view, as it was called, still had its followers, and it was not until around the 1920s and 1930s that the frequentist theory monopolized the subject of probability theory. In hindsight this is not surprising; with most practitioners coming from fields such as agriculture or population ecology where the amount of data was massive and where background information was not easily assessed numerically, the Laplacean view did not offer much improvement over frequentist methods. At that time, there were very few followers of Laplace. Two notable exceptions were the famous economist Lord John Maynard Keynes and the geophysicist Sir Harold Jeffreys. Jeffreys and Sir Ronald A. Fisher, one of the main proponents on the frequentist, or 'objective', side debated these issues in publications and in private correspondence. Although they held very different views on the subject, both had the goal of finding practical tools for inference and they regarded each other highly. An interesting and impartial account of this important part of scientific history is given by Howie (2002) who also provides more details on the general history of probability theory. Howie also points out that the frequentist school is not really one common school, but a disparate collection of ideas. Fisher, for instance, probably had more in common with Jeffreys than with such mathematical statisticians as Neyman or Egon Pearson (not to be confused with his father Karl Pearson who invented the chi-squared test).

In the 1950s, however, the Bayesian movement experienced a revival and since then a slowly increasing attention has been paid to Bayesian ideas. Presently, the interest in the Bayesian paradigm is accelerating and although most workers still adopt a somewhat inconsistent interpretation of probabilities as frequencies while still using the Bayesian framework, we believe that with the publication of Jaynes' book probability theory as logic will become the standard theory in the future.

Jeffrey's book (Jeffreys, 1939) was the main inspiration for Jaynes and is the critical link between Laplace's work and today's understanding of the subject. Several other works also deserve mentioning; Zellner's book (Zellner, 1971) contains many technical results that cover basic problems in econometrics, signal processing, and control theory. Not as strong on emphasizing fundamentals as Jaynes' book, it however provides an invaluable technical reference; Sivia (1996) is an introductory text drawing strong inspiration from Jaynes' works and can serve as a clearly written complement to Jaynes's book. Although we have here stressed the contributions by Jaynes, it should be emphasized that Cox (1946) made the essential derivation of the product and the sum rules from elementary assumptions of consistency and common-sense correspondence.

In this chapter we have not given many examples, and for full understanding of the subject the reader is referred to Jaynes (2003). In the following chapters, however, we will use the results obtained here in solving a number of problems

related to allocating resources under uncertainty. Hopefully, these examples will stimulate understanding and inspire to new and improved solutions.

Appendix 2.A Derivation of Laplace's Rule of Succession

We here evaluate

$$\begin{aligned} P(f_1...f_K|m_1...m_K I) &= \\ &= \frac{P(m_1...m_K|f_1...f_K I)P(f_1...f_K|I)}{P(m_1...m_K|I)} \end{aligned} \tag{2.102}$$

where

$$f_k = \frac{r_k}{\sum_{j=1}^{K} r_j} \tag{2.103}$$

is the relative frequency with which outcome k *will* occur, m_k is the number of occurrences of outcome $k = 1 \ldots K$, and I is all our background information that is relevant to the problem.

The prior probability distribution for the relative frequencies f_k is uniform over all combinations of K non-negative numbers that sum to unity (by the principle of indifference):

$$P(f_1...f_K|I) = C\delta(f_1 + ... + f_K - 1)\,, f_k \geq 0, \tag{2.104}$$

where $\delta(\cdot)$ is the Dirac delta ($\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ elsewhere). The normalization constant C is obtained from

$$\int_0^\infty ... \int_0^\infty P(f_1...f_K|I) df_1...df_K = 1 \tag{2.105}$$

and defining

$$I(q) \stackrel{\triangle}{=} \int_0^\infty ... \int_0^\infty \delta(f_1 + ... + f_K - q) df_1...df_K \tag{2.106}$$

we can write (2.105) as

$$CI(1) = 1 \ . \tag{2.107}$$

In order to avoid difficulties in carrying out this integration due to the inter-dependency of the integration limits, we note that the Laplace transform of $I(q)$ is

$$\begin{aligned} &\int_0^\infty e^{-sq} I(q) dq = \\ &= \int_0^\infty ... \int_0^\infty e^{-s(f_1+...+f_K)} df_1...df_K = \\ &= \frac{1}{s^K} \ . \end{aligned} \tag{2.108}$$

But this is a standard formula and the inverse Laplace transform of (2.108) is

$$I(q) = \frac{q^{K-1}}{(K-1)!} \tag{2.109}$$

yielding the normalization constant[11]

$$C = \frac{1}{I(1)} = (K-1)! \,. \tag{2.110}$$

The likelihood term in (2.102) is the multinomial distribution

$$\begin{aligned} P(m_1...m_K|f_1...f_K I) = \\ = \frac{M!}{m_1!\cdots m_K!} f_1^{m_1}\cdots f_K^{m_K} \,, \end{aligned} \tag{2.111}$$

where we define $M = \sum_{k=1}^{K} m_k$.

The prior distribution $P(m_1...m_K|I)$ is obtained by averaging the joint distribution for m_k and f_k over all possible f_k. Since

$$\begin{aligned} P(m_1...m_K|I) &= \int\cdots\int P(m_1...m_K, f_1...f_K|I) df_1...df_K \\ &= \int\cdots\int P(m_1...m_K|f_1...f_K I) P(f_1...f_K|I) df_1...df_K \end{aligned} \tag{2.112}$$

the prior can be written as

$$\begin{aligned} P(m_1...m_K|I) = \frac{M!}{m_1!\cdots m_K!} \int ... \\ ... \int f_1^{m_1}\cdots f_K^{m_K} P(f_1...f_K|I) df_1...df_K = \\ = \frac{M!}{m_1!\cdots m_K!} \cdot C \cdot J(1) \end{aligned} \tag{2.113}$$

where C was obtained in (2.110) and

$$\begin{aligned} J(q) = \int_0^\infty ... \int_0^\infty f_1^{m_1}\cdots f_K^{m_K} \times \\ \times \delta(f_1 + ... + f_K - q) df_1...df_K \,. \end{aligned} \tag{2.114}$$

[11] One might casually expect that the normalization constant becomes K, not $(K-1)!$, since the different frequencies are equally likely . However, the constraint that the probabilities must sum to one in effect means that the normalization constant is obtained by counting the possible combinations that can arise while satisfying the sum constraint.

The Laplace transform of $J(q)$ is

$$\begin{aligned} & \int_0^\infty e^{-sq} J(q) dq \\ = & \int_0^\infty ... \int_0^\infty e^{-s(f_1+...+f_K)} f_1^{m_1} \cdots f_K^{m_K} df_1 ... df_K \\ = & \frac{m_1! \cdots m_K!}{s^{M+K}} , \end{aligned} \tag{2.115}$$

and taking the inverse Laplace transform yields

$$J(q) = \frac{m_1! \cdots m_K!}{(M+K-1)!} q^{M+K-1} . \tag{2.116}$$

Inserting this (with $q = 1$) into (2.113), we obtain

$$P(m_1 ... m_K | I) = \frac{M!(K-1)!}{(M+K-1)!} . \tag{2.117}$$

Combining (2.104), (2.111), and (2.117) into (2.102), we have

$$\begin{aligned} P(f_1 ... f_K | m_1 ... m_K I) = & \frac{(M+K-1)!}{m_1! \cdots m_K!} \times \\ & \times f_1^{m_1} \cdots f_K^{m_K} \delta(f_1 + ... + f_K - 1) . \end{aligned} \tag{2.118}$$

We set out to find the probability for obtaining a certain outcome k, which, due to the assumption of a fixed causal mechanism, would equal the future relative frequency f_k if it were known. Instead we take the probability for k as the expectation of the relative frequency with which that particular outcome occurs:

$$\begin{aligned} p_k & \triangleq P(k | m_1 ... m_K I) = \langle f_k \rangle \\ & = \int_0^\infty ... \int_0^\infty f_k P(f_1 ... f_K | m_1 ... m_K I) df_1 ... df_K \\ & = \frac{m_1 \cdots m_K}{(M+K)!} \cdot \frac{(M+K-1)!}{m_1! \cdots m_{k-1}! (m_k+1)! m_{k+1}! \cdots m_K!} \\ & = \frac{m_k + 1}{M+K} \end{aligned} \tag{2.119}$$

where we again use the Laplace transformation technique in exactly the same way as in deriving (2.117) to solve the integrals. As before, $M = \sum_{k=1}^K m_k$.

Appendix 2.B Derivation of the Discrete Maximum Entropy Distribution

The maximum entropy distribution is found using the Lagrange method. Using the constraints (2.39) we form the functional

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i + \sum_{k=1}^{m} \lambda_k \left(F_k - \sum_{i=1}^{n} p_i f_k(x_i) \right) \tag{2.120}$$

and differentiate with respect to p_i:

$$\frac{\partial H(p)}{\partial p_i} = -\log p_i - 1 - \sum_{k=1}^{m} \lambda_k f_k(x_i) \, . \tag{2.121}$$

Setting this equal to zero we have the general form of the entropy-maximizing probability mass:

$$p_i = \exp\big[-1 - \sum_{k=1}^{m} \lambda_k f_k(x_i)\big] \ . \tag{2.122}$$

However we have not yet included the constraint that $\sum_{i=1}^{n} p_i = 1$. This is just a normalization, and we obtain:

$$p_i = \frac{1}{\sum_{i=1}^{n} \exp\big[-\sum_{k=1}^{m} \lambda_k f_k(x_i)\big]} \exp\big[-\sum_{k=1}^{m} \lambda_k f_k(x_i)\big] \, . \tag{2.123}$$

The Lagrange multipliers λ_i are chosen so that the constraints (2.39) are satisfied.

This procedure is formulated in a compact form by introducing the partition function (2.41) and rewriting (2.123) as

$$p_i = \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp\big[-\sum_{k=1}^{m} \lambda_k f_k(x_i)\big] \, . \tag{2.124}$$

In order to find the Lagrange multipliers satisfying the constraints (2.39) we notice that differentiating $\log Z$ with respect to each λ_k gives:

$$\begin{aligned} \frac{\partial}{\partial \lambda_k} \log Z &= \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \sum_{i=1}^{n} \big(-f_k(x_i) \times \\ &\times \ \exp\big[-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\big]\big) \\ &= -\sum_{i=1}^{n} p_i f_k(x_i) \ , \end{aligned} \tag{2.125}$$

which is the formulation of the constraints (2.39).

Thus the constraints (2.39) are satisfied by choosing the Lagrange multipliers so that

$$F_k = -\frac{\partial}{\partial \lambda_k} \log Z \quad . \tag{2.126}$$

Chapter 3

Controlling Production Resources to Meet Customer Demands

IN the manufacturing industry a common class of resource allocation problems can be described as allocation of the production resources so as to meet future uncertain order intakes while minimizing production expenses, conditioned on satisfying a number of constraints on how the resources can be assigned.

In this chapter we formulate and solve a class of general resource allocation problems which can be stated in an abstract way as follows.

Consider the problem faced by the manager of a production plant. The plant manufactures a variety of widgets, and a number of production units (PU:s) are available for producing the widgets. The PU:s vary in quality; consequently each PU has a certain capacity, varying according to the type of widget to be produced. The capacity of a PU may also be time varying. If the PU is a machine, it may be time varying because at times it needs maintenance; if instead the PU consists of a team of workers the time variability is explained by simple facts as absence from work due to sickness or that the work force is decreased during nights and weekends. Evidently, the production capacities may be uncertain as well.

The manager of the plant makes manufacturing decisions (allocating PU 1 to produce x A-widgets, y B-widgets, and so on) so as to meet as many future order intakes for the different types of widgets as possible. The job is then to decide how many widgets of each type each PU is to produce over a specified time horizon. Depending on the type of widgets produced, the time horizon may vary from days, or even hours, to several years. The job is complicated by the fact that the managing unit does not know exactly how many orders will be be placed for each type of widget over the specified time horizon. There may also be certain constraints on

how the PU:s may be utilized. Some PU:s may only be able to produce a certain type of widgets, while some other PU:s only make parts of widgets, etc.

A further complication may be that some customers are considered more important than others, there may be different expenses associated with changing the production patterns, and there may be different costs related to manufacturing different widgets.

Clearly, the problem faced by the manager can be treated to some extent by applying existing works in the field of operations research. In the operations research literature, methods for scheduling and resource allocation are studied mainly with the objective of minimizing the *make-span*, i.e. the greatest completion time for a number of pre-specified tasks, or similar delay-related criteria (see e.g. Brucker et al., 1999, Hillier and Lieberman, 1990, Negenman, 2001). Typical constraints are formulated as precedence requirements, i.e. certain tasks must be completed before certain other tasks can begin. The majority of operations research publications concern different instances of this type of problem with deterministic parameters. The most common ways of handling uncertainty, according to Penz et al. (2001), are sensitivity analysis, robust design or the use of a stabilization process. Stochastic problems are studied to a lesser extent, and in general by assuming certain fault frequency distributions, etc. (Sox et al., 1999).

Instead of incurring assumptions concerning the true order sizes, the manager needs a way to make decisions that use whatever knowledge he may have *without* introducing unwarranted assumptions. He needs a mathematical description of his state of uncertainty which takes into account all the possibilities not ruled out by his knowledge. We have already seen in Chapter 2 how this objective can be achieved by using the maximum entropy principle and probability theory. The first use of probability theory as extended logic in problems of resource allocation was given by Jaynes (1963b). In that work, Jaynes considered a similar problem as ours. Our treatment generalizes that model to a more flexible manufacturing plant, and extends the framework in several directions as will be explained later. The main extensions are the following:

- The model of the manufacturing plant includes an arbitrary number of widgets and production units, and an arbitrary scheduling horizon.
- The optimization criterion accounts for widget-type specific pricing and arbitrary costs for resource utilization, which for instance may include increased production costs when the work assignments for production units are changed.
- Solutions are presented which take into account uncertain production capacities, either based on a Gaussian prediction or based on previously recorded

capacity fluctuations.

- Whereas Jaynes only presented maximum entropy solutions for accounting for uncertain order sizes, we further introduce a Bayesian solution which is able to extract information regarding patterns observed in previous order intakes.
- We show how the given problem formulation translates into a flow-control problem which forms the basis for an application in mobile communications presented in Chapter 5.

Our work is a continuation of Jaynes', and to facilitate reading both works simultaneously we are using the same notation as in his work.

In Section 3.1 we formulate the problem. In Section 3.2 solutions are derived for a number of basic cases. Numerical examples are given in Section 3.3, whereas Section 3.4 gives some extensions and modifications to the problem before we conclude the chapter in Section 3.5.

3.1 Minimizing the Expected Number of Missed Orders

Consider the problem outlined in the introduction. We are to schedule the use of our production resources over a time horizon divided into T time slots. The plant produces U different types of widgets and there are R production units, where the rth production unit has the capacity to produce c_{urt} widgets of type u during time slot t. Suppose that the resource allocation decisions ρ_{urt}, $t = 1 \dots T$, are all made at $t = 0$.

During the period $t = 1 \dots T$ the plant receives orders for $n_u = \sum_{t=1}^{T} n_{ut}$ type u widgets, and if we fail to meet an order for such a widget our cost, or lost income, (measured in some appropriate unit/currency) is denoted by v_u.

The object is now to assign a fraction ρ_{urt} of production unit r to produce type u widgets during time slot t, so that the future order intakes are met while the manufacturing costs are minimized.

Before stating the complete problem in which future order intakes are uncertain, we here assume knowledge of all incoming order sizes during the time horizon $t = 1 \dots T$. The problem is then to minimize the loss function L:

$$L = \sum_{u=1}^{U} \left(v_u \cdot g(n_u - S_u - \sum_{r=1}^{R} \sum_{t=1}^{T} \rho_{urt} c_{urt}) + h(\rho_{\mathbf{u}}) \right) \tag{3.1}$$

where S_u is the number of type u widgets already in stock, $g(x) = x$ if $x > 0$, $g(x) = 0$ otherwise, and $h(\cdot)$ is a function describing the cost (in the same unit/currency as v_u) for the utilization of the production units.

The first term in (3.1), $v_u \cdot g(n_u - S_u - \sum_{r=1}^{R}\sum_{t=1}^{T}\rho_{urt}c_{urt})$, represents failed incomes due to orders that can not be met by the stock S_u or the production $\sum_{r=1}^{R}\sum_{t=1}^{T}\rho_{urt}c_{urt}$ under the coming interval $t = 1 \ldots T$. In the second term of (3.1), $h(\rho_{\mathbf{u}})$, we define $\rho_{\mathbf{u}}$ as the vector of all assignments to produce type u widgets, i.e.

$$\rho_{\mathbf{u}} \equiv \{\rho_{u11}, \rho_{u12}, \ldots, \rho_{u1T}, \rho_{u21}, \ldots, \rho_{u2T}, \ldots, \rho_{uRT}\ ,\} \tag{3.2}$$

which can be further generalized to depend on previous resource allocations (i.e. for $t < 1$). The function $h(\cdot)$ should be defined according to actual production costs and varies from problem to problem. The use of $h(\cdot)$ is a simple way to include costs for transferring production of one widget to another PU, etc.

There may also be various constraints on resource utilization. The basic constraints on ρ_{urt} are

$$\sum_u \rho_{urt} \leq 1 \quad \forall r, t \tag{3.3}$$

$$0 \leq \rho_{urt} \leq 1 \quad \forall u, r, t\ , \tag{3.4}$$

but in general we may have an additional number of matrix equalities and inequalities representing constraints on different resources. For instance, we may require a certain minimum number, φ_u, of widgets produced, i.e.

$$\sum_{t=1}^{T}\sum_{r=1}^{R} c_{urt}\rho_{urt} \ \geq \ \varphi_u \ . \tag{3.5}$$

Another common restriction may be that production units are constrained to have only one assignment in each time slot, i.e. ρ_{urt} must belong to the set $\{0, 1\}$. In Section 3.4 other types of modifications are described which transform the problem to a variety of common problem scenarios.

The problem of minimizing (3.1) presents no conceptual difficulties, but is in fact of little use for the manager of the plant. The main problem facing him is that the incoming orders are highly uncertain. Typically, there is some limited information available. For instance, he may have at his disposal records from the previous period in which he can gather the average order sizes for different widgets.

What is needed is thus a probability distribution describing his uncertainty as to the true order sizes. Having such a distribution, we can determine the expectation $\langle L \rangle$ of (3.1) as

$$\langle L \rangle = \sum_{u=1}^{U}\sum_{n_u=1}^{\infty} P(n_u|I)\left(v_u \cdot g(n_u - S_u - \sum_{r=1}^{R}\sum_{t=1}^{T}\rho_{urt}c_{urt}) + h(\rho_{\mathbf{u}})\right) \tag{3.6}$$

Table 3.1: Definitions of the main variables in this chapter.

L	The loss function, representing the total cost for production and unfilled orders
U	The number of widget types
R	The number of production units
T	The number of time slots a resource allocation is optimized over
S_u	The number of widgets of type u in stock
n_u	The total order size for widgets of type u over the T time slots
c_{urt}	The production capacity [number of widgets] at production unit r, time slot t for type-u widgets
ρ_{urt}	The fraction ($0 \leq \rho_{urt} \leq 1$) of production unit r that is used for producing type-u widgets at time t. Adjusted so that $\langle L \rangle$ is minimized
x_u	The total production of type-u widgets over the T time slots ($x_u = \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt}\rho_{urt}$)
v_u	The per-widget cost for failing to meet orders for type-u widgets
$\rho_{\mathbf{u}}$	A vector of all past and present allocations ρ_{urt} used to determine production costs for type-u widgets
$h(\rho_{\mathbf{u}})$	The production cost for type-u widgets given past and present allocations

where $P(n_u|I)$ denotes the probability distribution for the order sizes n_u conditioned on whatever information I the manager may have.

The most reasonable course of action is now to make the decisions ρ_{urt} which minimize the expected loss (3.6), while agreeing with the constraints (3.3), (3.4) and other relevant constraints as mentioned above.

Furthermore, the production capacities c_{urt} may be uncertain due to the possibility of machine failures, etc. In that case, we should also determine a probability distribution for the capacities and average (3.6) over that distribution too. That way, our resource allocation decisions take into account the possibility of lower or higher capacities than expected.

In the next section we derive the expected loss expressions for different states of information concerning the order intakes and production capacities. For reference throughout the chapter, Table 3.1 lists the definitions of the main variables in this chapter.

3.2 Solutions for Uncertain Order Intakes and Uncertain Production Capacities

In any problem of inference, we use whatever information we have in order to narrow our list of possible outcomes. We assign different degrees of plausibility to different outcomes corresponding to that information. As the process of inference turns into a decision problem, we find that a rational decision should take into account all possibilities that have not been ruled out by our information.

Here, we attempt to accomplish this objective by using probability assignments

in (3.6) which have maximal entropy subject to constraints given by our information I. As we stressed in Chapter 2, it corresponds to the aim of avoiding gratuitous assumptions (Roberts, 1971). In one specific formulation, given in Section 3.2.2, we will further use Bayes rule to take advantage of unforeseen patterns in the order intakes.

The final expression for the expected loss depends critically on the information we use to assign probabilities. We here investigate four basic cases, each representing a typical scenario that may arise in practice.

3.2.1 Knowledge of expected order intakes

A common type of information available for this type of problem consists of expected order intakes for the coming period. This may be based on sales records from the previous period.

Here we derive the maximum-entropy probability distribution for future order sizes under the condition of knowing the expected order sizes for each widget type, and then proceed to give the expected loss for this scenario.

The distribution for future order sizes

We are to assign a prior probability distribution for non-negative integer quantities, n_u, $u = 1 \dots U$, having known means $\langle n_u \rangle$. Denoting this information by I, we now turn to find the $P(n_u|I)$ which maximizes the entropy, c.f. (2.24),

$$H = -\sum_{n_u} P(n_u|I) \log P(n_u|I) \tag{3.7}$$

under the constraints

$$\langle n_u \rangle \;\; = \;\; \sum_{n_u=0}^{\infty} n_u P(n_u|I) \;\; , \;\; u = 1 \dots U \;\; . \tag{3.8}$$

Notice that the summation index reflects that the integer n_u is non-negative. In order to find the $P(n_u|I)$ with maximum entropy we follow the steps in Section 2.7.1. The partition function (2.41) becomes

$$\begin{aligned} Z(\lambda_1, \ldots, \lambda_U) &= \sum_{n_1=0}^{\infty} \ldots \sum_{n_U=0}^{\infty} \exp(-\lambda_1 n_1 - \ldots - \lambda_U n_U) \\ &= \sum_{n_1=0}^{\infty} \left(\ldots \left(\sum_{n_U=0}^{\infty} \exp(-\lambda_U n_U) \right) \ldots \right) \exp(-\lambda_1 n_1) \\ &= \prod_{u=1}^{U} \frac{1}{1 - e^{-\lambda_u}} \ , \end{aligned} \tag{3.9}$$

where we first rewrote the expression according to $x^{a+b} = x^a x^b$ and then used the closed form expression for the geometric series. The Lagrange multipliers are now determined from (2.43):

$$\langle n_u \rangle = -\frac{\partial}{\partial \lambda_u} \log Z = \frac{1}{e^{\lambda_u} - 1} \ . \tag{3.10}$$

Independence between different probabilities yields higher entropy than dependencies, and consequently the maximum-entropy probability assignments $P(n_u|I)$ factor:

$$P(n_1, \ldots, n_U|I) = P(n_1|I) \ldots P(n_U|I) \ . \tag{3.11}$$

Inserting (3.9) into the expression for the general maximum-entropy distribution (2.42) and using (3.11) we obtain

$$\begin{aligned} P(n_u|I) &= \frac{1}{Z(\lambda_u)} e^{-\lambda_u n_u} \qquad n_u = 0 \ldots \infty \\ &= (1 - e^{-\lambda_u}) e^{-\lambda_u n_u} \ . \end{aligned} \tag{3.12}$$

From (3.10) we see that

$$e^{-\lambda_u} = \frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \ , \tag{3.13}$$

and consequently

$$\begin{aligned} P(n_u|I) &= (1 - e^{-\lambda_u}) e^{-\lambda_u n_u} \\ &= \frac{1}{\langle n_u \rangle + 1} \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{n_u} \end{aligned} \tag{3.14}$$

is the distribution of highest entropy subject to the constraints (3.8) and $\sum P(n_u|I) = 1$.

The maximum-entropy derivation of the exponential distribution above can also be found in (Jaynes, 1963b). In Figure 3.1 the distribution is plotted for different mean values. The skewness of the curve arises because n_u is only defined

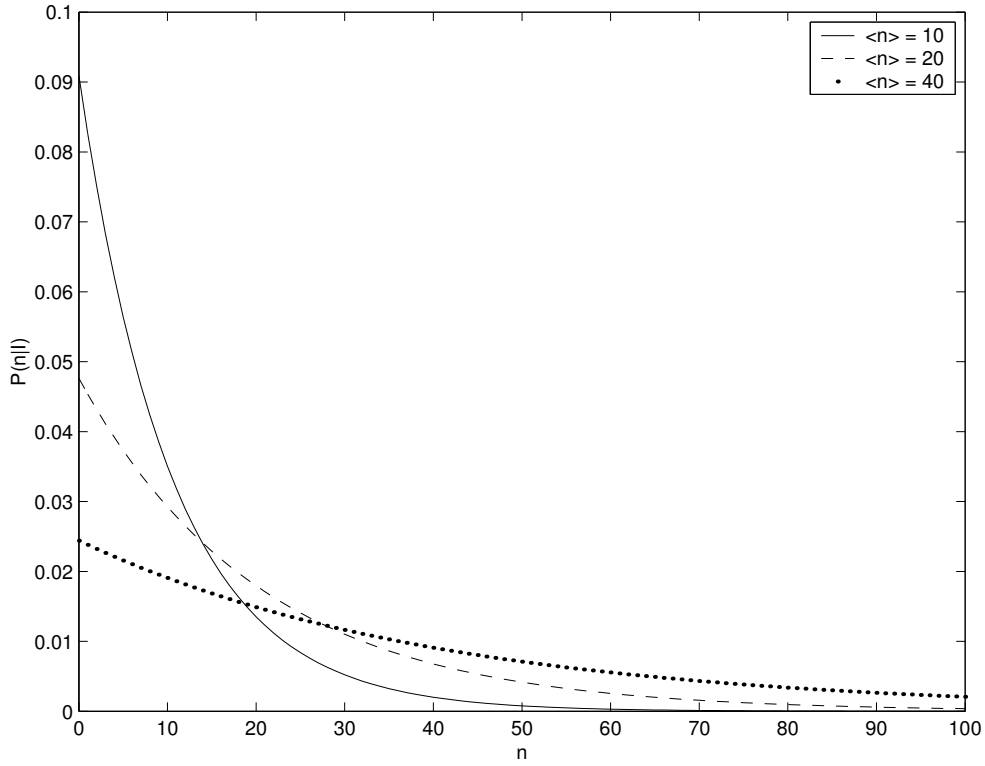

Figure 3.1: The maximum entropy probability distribution for a non-negative integer quantity n with known mean $\langle n \rangle$.

for non-negative values. Hence, for a larger mean value the curve tends more and more towards a uniform distribution. The distribution would be different if n_u had a known upper bound. For instance, if the n_u represent the number of dots on the face of a die, we must include that $1 \leq n_u \leq 6$ in our probability derivation. This yields a distribution which is skewed differently depending on the given mean values.

The expected loss

For brevity, we introduce

$$x_u = \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt} \rho_{urt} \quad , \tag{3.15}$$

describing the total number of type u widgets produced during the scheduled time horizon $t = 1 \dots T$. With $P(n_u|I)$ given by (3.14) the expected loss (3.6) becomes:

$$\langle L \rangle = \sum_{u=1}^{U} \sum_{n_u=0}^{\infty} P(n_u|I) \left(v_u \cdot g(n_u - S_u - x_u) + h_u(\rho_{\mathbf{u}}) \right) \tag{3.16}$$

$$= \sum_{u=1}^{U} \left(v_u \langle L_u \rangle + h_u(\rho_{\mathbf{u}}) \right) \ . \tag{3.17}$$

It is shown in Appendix 3.A that $\langle L_u \rangle = \sum_{n_u=0}^{\infty} P(n_u|I) g(n_u - S_u - x_u)$ is equal to

$$\langle L_u \rangle = \begin{cases} \langle n_u \rangle \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{x_u + S_u} & , \ x_u + S_u > 0 \\ \langle n_u \rangle - S_u - x_u & , \ x_u + S_u \leq 0 \ . \end{cases} \tag{3.18}$$

The situation that $x_u + S_u \leq 0$ in (3.18) may seem like an impossible circumstance, but although x_u is certainly positive, S_u may be negative if we have a number of outstanding orders left from previous scheduling rounds or if there is a number of known orders in the coming period. The case $x_u + S_u > 0$ is however more likely in the type of application we consider in this chapter. The expression (3.17) is to be minimized by adjusting ρ_{urt} under the system utilization constraints (3.3) and (3.4). This is a constrained nonlinear optimization problem which can be solved using nonlinear programming methods.

A practical complication – unknown expectations

In practice, the probability assignment $P(n_u \mid I)$ has a substantial shortcoming; it requires exact knowledge of the expected order sizes. This information may in reality be highly uncertain, or even non-existing in cases where for instance a new type of widget is to be released. Here, we will briefly describe how to integrate out the uncertain parameter $\langle n_u \rangle$ based on knowledge of a finite record of previous order sizes. In Section 3.2.2 we then treat this problem more fully, and update the entire probability distribution according to Bayes' rule based on past records of order sizes. Although the latter approach takes better advantage of the patterns of past order sizes, the former may be used in situations where we believe that the only operative constraint on the entropy is actually the mean value constraint. In effect, we then assume less structure consistent over time in the order patterns than in the latter approach.

Denoting the probability assignment (3.14) by $P(n_u \mid \mu, I)$ (where $\mu = \langle n_u \rangle$ for compact notation) to make explicit the dependence on the expectation, we now

wish to find the marginal probability for n_u conditioned on a short number of previous order sizes.

Let $\{y\} = \{y_1 \dots y_N\}$ be known order sizes for widgets of type u over N (possibly non-consecutive) previous periods of length T (the same length as the scheduling horizon). We then wish to derive the probability $P(n_u \mid \{y\}, I)$ for obtaining orders for n_u widgets over the next T periods given past order sizes $\{y\}$. According to the sum rule we find this distribution by integrating over all possible values of μ,

$$P(n_u \mid \{y\}, I) = \int_0^\infty P(n_u \mid \mu, \{y\}, I) P(\mu \mid \{y\}, I) d\mu \ . \tag{3.19}$$

If we have no information of correlations between order sizes at different time periods we are better off leaving them out (due to the higher entropy), and thus $P(n_u \mid \mu, \{y\}, I) = P(n_u \mid \mu, I)$. In order to determine $P(\mu \mid \{y\}, I)$ we use Bayes' theorem (2.8) to obtain

$$P(\mu \mid \{y\}, I) \propto P(\{y\} \mid \mu, I) P(\mu \mid I) \ . \tag{3.20}$$

Thus, in order to fully specify the marginal distribution $P(n_u \mid \{y\}, I)$ given past data, the only new element consists of the prior $P(\mu \mid I)$.

Letting $Y = y_1 + \ldots + y_N$ be the sum of all past order sizes for a particular widget type u and using a Jeffrey's prior, we show in Appendix 3.B that

$$P(n_u \mid \{y\}, I) = N \frac{(n_u + Y - 1)!(N + Y - 1)!}{(Y - 1)!(N + n_u + Y)!} = N \frac{\prod_{i=0}^{N-1}(Y + i)}{\prod_{i=0}^{N}(Y + n_u + i)} \ , \tag{3.21}$$

and that an excellent approximation to (3.21) is given by the exponential distribution (3.14) with $\langle n_u \rangle = \frac{Y}{N-1}$.

Consequently, when we have only a very short data record for the determination of $\langle n_u \rangle$ we can use the expected loss in (3.18) using $\langle n_u \rangle = \frac{Y}{N-1}$.

3.2.2 A predictive distribution based on logarithmic histograms

In the previous section we only used the record of past order sizes to marginalize the expectation μ. A disadvantage of that approach is that even if there are very obvious patterns in the past data (for instance, suppose that in a record of 1000 orders, half of them were of size 1 and the other half of size 10) they will be dogmatically ignored. In effect we would not follow our Chapter 2 desiderata in doing so, as we would arbitrarily throw away possibly relevant information. No matter what the individual order sizes could tell us, we would flat out reject using this information and stick to our exponential probability assignment and only average

over the uncertainty pertaining to μ. This would be reasonable if all information we obtained from the records were the number of orders and the sum size. However, if we are actually given the entire sales record, we should, as always on receiving new information, invoke Bayes' rule.

We would like to calculate the posterior probability for n_u given the past order sizes $\{y\}$

$$P(n_u \mid DI) = \frac{P(D \mid n_u I)P(n_u \mid I)}{P(D \mid I)} \,, \tag{3.22}$$

where $D = \{y\}$ is the past sales record of size N. If we would only have a small set of possible order sizes, a natural procedure would be to use Laplace's rule of succession (see Section 2.6) to determine the probability for future orders of the different sizes. Such a probability assignment would not assume any temporal correlations, but would express an expectation that the underlying market mechanisms do not change appreciably. We have however not restricted the order sizes to a small set, but should rather be prepared for any positive integer size. Based only on a finite order record, the resulting distribution would be very close to uniform. Calculation of the expected loss would also require a numerical summation over an infinite number of terms, which clearly precludes this approach from further consideration.

Nonetheless, it is possible to resort to a similar approach where we first partition the order sizes in a discrete set covering a closed interval on the non-negative real line, and then use the rule of succession on the 'bins' that constitute this interval. We can thus find a reasonably informative posterior probability for receiving future orders within a certain size interval given by the number of bins to use and the minimum and maximum order sizes. An important question is then the issue of partitioning the order sizes. We could use a linear partitioning consisting of using equal sizes for all bins and spreading them uniformly from 0 to some upper limit. There are a number of problems with this idea however; first of all, the upper limit must be very large, and that means that the bins become too numerous; secondly, order sizes are perhaps not naturally partitioned in a linear fashion. A quick look in the product catalogue of any large vendor of, say, electronic chips reveals that packages are typically priced in ranges of $1-10$, $10-100$, $100-1000$, $1000-$, etc. This is typical of industrial products, where the *order* of the order sizes is a more natural partitioning rule than the absolute order sizes. For instance, a company would first buy a few samples of the widgets to try them in their products, and then the company might decide to pursue the use of the widgets in their products. The number of companies that buy individual sample widgets may be quite large. The companies that continue using the widgets may be small-scale, mid-scale, or large-scale companies, and might for instance require an amount on the order of

100, 1000, or 10000 widgets depending on their customer base. For each such logarithmic interval, we might expect an approximately equal number of orders with no further information from previous sales. Thus, a logarithmic partitioning of the order sizes may be appropriate, indicating a Jeffrey's prior (which is uniform over the logarithm) for the order sizes.

Based on the reasoning above, we propose to partition the non-negative n_u-line in K *bins*, or sub-intervals, spread uniformly over the logarithm of n_u and use the rule of succession to update the probability for the number of orders in the size interval corresponding to a given bin. We will define a lower limit and an upper limit for the uniformly distributed $\log n_u$ bins, and use one bin for all sizes below the lower limit and one bin for the sizes above the upper limit.

We partition the logarithm of the order sizes for widget u in $K-2$ bins of equal width

$$w = \frac{\log n_{max} - \log n_{min}}{K-2} \tag{3.23}$$

between two numbers $\log n_{min}$ and $\log n_{max}$. The additional two bins refer to log-order sizes below $\log n_{min}$ and above $\log n_{max}$. We then count the number m_{uk} of orders for widget u of log-size corresponding to each bin k. The posterior probability for the number of orders in each bin k for widget u is then obtained by Laplace's rule of succession

$$P(n_{uk}|m_{u1}...m_{uK}I) = \frac{m_{uk}+1}{M_u+K} \tag{3.24}$$

where $M_u = \sum_{k=1}^{K} m_{uk}$. Note that this reflects a prior which is uniform over the bins, but that two of these bins are not of equal log-size as the others. We thus assume that the chance for receiving an order in any bin is equal.

Now, having determined the probability for obtaining orders for n_u type-u widgets, we still need the probability for receiving a particular order size within the bin interval. Otherwise we cannot determine the expected loss. There are two possible choices, either a uniform distribution or a Jeffrey's distribution over the bin interval. We have already employed Jeffrey's distribution in motivating the rule of succession for the logarithmic intervals, but here we are concerned with distributing probability over a closed (non-logarithmic) interval and the uniform distribution is then more appropriate according to the principle of indifference. We shall thus take

$$P(n_u \mid n_u \in k) = \frac{1}{b_k - a_k} \tag{3.25}$$

where b_k denotes the lower limit of the closest bin to the right and a_k is the lower limit of the current bin k. Letting

$$\alpha_k \stackrel{\triangle}{=} \max(S_u + x_u, a_k) \tag{3.26}$$

and

$$\beta_k \stackrel{\triangle}{=} \max(S_u + x_u + 1, b_k) \tag{3.27}$$

we thereupon obtain the expected loss contribution given that the order size is within bin k

$$\begin{aligned}\langle L_u \mid n_u \in k\rangle &= \sum_{n_u=a_k}^{b_k-1} \frac{1}{b_k - a_k} g(n_u - S_u - x_u)\,, \quad k = 1 \ldots K-1 \\ &= \frac{1}{2}\frac{\beta_k^2 - \beta_k - (\alpha_k^2 - \alpha_k)}{b_k - a_k} - \frac{\beta_k - \alpha_k}{b_k - a_k}(S_u + x_u)\,. \end{aligned} \tag{3.28}$$

The derivation is given in Appendix 3.C. Note that if we allow order sizes of infinite size, the Kth bin would range over an interval which is open to the right giving an infinite β and consequently an infinite expected loss. The solution is to use Jeffrey's distribution over a bounded interval for the probability for obtaining a particular order size given that the order size lies in bin K. For $k = K$ we thus have

$$\begin{aligned}\langle L_u \mid n_u \in K\rangle &\approx \sum_{a_K}^{b_K-1} \frac{1}{\log(b_K/a_K)\, n_u} g(n_u - S_u - x_u) \\ &\approx \frac{\beta_K - \alpha_K - \log(\beta_K/\alpha_K)\,(S_u + x_u)}{\log(b_K/a_K)}\,, \end{aligned} \tag{3.29}$$

where we approximated the normalization of the discrete Jeffrey's distribution with the normalization for a continuous Jeffrey's distribution and the sum with the corresponding integral. The continuous and the discrete results are however almost perfectly identical and in practice we can use them interchangeably.

Accordingly, we obtain the following expression for the expected loss contribution:

$$\langle L_u \rangle = \sum_{k=1}^{K} \frac{m_{uk} + 1}{M_u + K} \langle L_u \mid n_u \in k\rangle\,. \tag{3.30}$$

Minimization of this expression inserted in (3.17) typically improves on the performance obtained from using only the average values $\langle n_u \rangle$, given that K is not too small. The reason is that we here take advantage of patterns in the sales records that are not accounted for by only using the expected order sizes regardless of what our data actually tell us.

We should further observe that if we would let the bin widths adapt according to the incoming order sizes instead of using fixed logarithmic widths, we would be able to obtain even better performance. In Chapter 8 we investigate this problem of optimal approximate Bayesian inference.

3.2.3 Uncertain production capacities

Here, we investigate the expected loss for the case when the production capacities c_{urt} are uncertain. We build on the case developed in Section 3.2.1 presupposing knowledge of expected order sizes $\langle n_u \rangle$.

Consider a problem where we have predicted the capacity c_{urt} of every production unit r for producing type u widgets for each time slot t, with some known accuracy.

Our task is now to calculate the expected loss (3.6) with respect to the joint distribution $P(n_u c_{urt} \mid I)$. In Appendix 3.D we show that since n_u and c_{urt} are logically independent, the expected loss contribution from type-u widgets becomes

$$\begin{aligned} \langle L_u \rangle &= \int_{-S_u}^{\infty} P(x_u \mid I) \langle L_{1u,P(n|I)} \rangle dx_u \\ &+ \int_{-\infty}^{-S_u} P(x_u \mid I) \langle L_{2u,P(n|I)} \rangle dx_u \ , \end{aligned} \tag{3.31}$$

where $\langle L_{1u,P(n|I)} \rangle$ and $\langle L_{2u,P(n|I)} \rangle$ denotes the expected per-widget-type loss from Section 3.2.1 for $x_u + S_u > 0$ and $x_u + S_u \leq 0$ respectively (recall that $x_u = \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt} \rho_{urt}$). This new notation is used to make a distinction between the expected loss with respect to $P(n_u \mid I)$ in (3.17) and the one currently under investigation. $\langle L_{1u} \rangle$ and $\langle L_{2u} \rangle$ will now be used to describe the latter. The total loss is obtained by inserting (3.31) into

$$\sum_{u=1}^{U} \left(v_u \langle L_u \rangle + h_u(\rho_{\mathbf{u}}) \right) \ . \tag{3.32}$$

The determination of $P(c_{urt} \mid I)$ (which in turn gives $P(x_u \mid I)$ since $x_u = \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt} \rho_{urt}$) depends on what information we have concerning the production capacities. We will study two cases which are useful in different situations. The first is based on having a prediction of each production capacity along with a measure of the prediction accuracy. This may be reasonable in certain flow control applications where the capacity may turn out to be higher or lower than the predicted value. This is similar to the wireless communication set-up considered in Chapter 5, but there the capacity can never increase beyond the transmission rate chosen by the transmitter. We will not study that situation here, but using the results in Chapter 5 the corresponding solution can easily be worked out also in this application.

In the second scenario, we consider a problem where the production capacity can only take a small set of values, and we have a record of how many times each of the different possible capacities have been used in a previous time interval of

some known length. This case is appropriate when the production units are of a static nature, but may have different quality at different times. For instance, this may be the case if the production units are some form of transport vehicles, some of which have a certain size, others having other sizes, and it is unknown which vehicle will actually be used.

Prediction with known accuracy

Let us assume that the accuracy of the prediction of a particular c_{urt} is represented by a known variance, σ_{urt}^2, and the prediction itself is the expected value of the capacity, $\langle c_{urt} \rangle$. In the case of a nonnegative integer quantity such as the effective capacity, finding the maximum-entropy distribution for known expectation and variance is analytically intractable. However, it is well-known (Shannon, 1948) that the Gaussian distribution has the highest entropy for a given mean and variance if the quantity of interest is defined over the entire real axis. Negative capacities are not possible and we should therefore calculate the expected loss using a Gaussian distribution truncated at 0. As that solution turns out to be somewhat analytically inelegant we instead derive the expected loss using a Gaussian distribution defined over the entire real axis as a simpler solution, valid when the expectation $\langle c_{urt} \rangle$ is large in comparison to the standard deviation σ_{urt} (so that the probability mass for negative values are negligible). We thus make the probability assignment

$$P(c_{urt} \mid I) = \frac{1}{\sqrt{2\pi\sigma_{urt}^2}} \exp\left\{ -\frac{1}{2\sigma_{urt}^2}(c_{urt} - \langle c_{urt} \rangle)^2 \right\} . \tag{3.33}$$

In order to determine

$$\langle L_{1u} \rangle = \int_{-S_u}^{\infty} P(x_u \mid I) \langle L_{1u,P(n|I)} \rangle dx_u \tag{3.34}$$

we first need to find the probability for $x_u = \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt}\rho_{urt}$ conditioned on I. Since x_u is a sum of scaled independent Gaussian variables, x_u is also Gaussian according to

$$x_u \sim \mathcal{N}(\sum_{r=1}^{R} \sum_{t=1}^{T} \rho_{urt} \langle c_{urt} \rangle, \sum_{r=1}^{R} \sum_{t=1}^{T} \rho_{urt}^2 \sigma_{urt}^2) . \tag{3.35}$$

Inserting (3.35) and $\langle L_{1u,P(n|I)} \rangle$ from (3.18) into (3.34) we have (the integral is equivalent to (A.4) in Appendix A with the solution (A.9)) the resulting expected

loss contribution for $x_u + S_u > 0$,

$$\begin{aligned}\langle L_{1u}\rangle &= \langle n_u\rangle\Big(\frac{\langle n_u\rangle}{\langle n_u\rangle+1}\Big)^{\sum_t^T\sum_r^R \rho_{urt}\gamma_{urt}+S_u} \\ &\times \left(\frac{1}{2}+\frac{1}{2}\text{erf}\left(\frac{S_u+\langle x_u\rangle+\delta_u^2\log\left(\frac{\langle n_u\rangle}{\langle n_u\rangle+1}\right)}{\sqrt{2\delta_u^2}}\right)\right) , \end{aligned} \tag{3.36}$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the error function, and

$$\gamma_{urt} = \frac{1}{2}\rho_{urt}\sigma_{urt}^2\log\left(\frac{\langle n_u\rangle}{\langle n_u\rangle+1}\right)+\langle c_{urt}\rangle , \tag{3.37}$$

$$\langle x_u\rangle = \sum_{r=1}^{R}\sum_{t=1}^{T}\rho_{urt}\langle c_{urt}\rangle , \tag{3.38}$$

$$\delta_u^2 = \sum_{r=1}^{R}\sum_{t=1}^{T}\sigma_{urt}^2\rho_{urt}^2 . \tag{3.39}$$

Observe that as the variance σ_{urt}^2 goes to zero the erf$(\cdot)$ expression tends to 1 and we obtain the familiar solution (3.17) with known capacities. Note also that the average loss (3.36) for uncertain production capacities modelled by a Gaussian distribution is equal to that of an exactly known value $\tilde{c}_{urt} = \gamma_{urt} \leq \langle c_{urt}\rangle$ when the argument to the erf$(\cdot)$ expression is large. Hence, our uncertainty concerning c_{urt} has the effect that it decreases the value of the predicted capacity by an amount which is proportional to σ_{urt}^2. Peculiarly, the proportionality constant depends on how much we utilize the resource and also to a lesser extent on the expected order size for the widgets produced by production unit u.

We may expect that as the variance increases, approximating (3.36) by (3.17) will gradually lead to worse performance. But just how high variance is needed for it to be worthwhile using the more complex model (3.36)?

In the expression for γ_{urt} it is readily seen that the term $\frac{1}{2}\rho_{urt}\sigma_{urt}^2\log(\frac{\langle n_u\rangle}{\langle n_u\rangle+1})$ will be negligible compared to $\langle c_{urt}\rangle$ unless $\sigma_{urt}^2 > \langle c_{urt}\rangle$ or $\langle n_u\rangle$ is very small, say in the range $1-5$. Thus, when the variance σ_{urt}^2 is small compared to the expected value $\langle c_{urt}\rangle$ and $\langle n_u\rangle$ is not too small we can safely ignore the effects of the term $\frac{1}{2}\rho_{urt}\sigma_{urt}^2\log(\frac{\langle n_u\rangle}{\langle n_u\rangle+1})$ and use the simpler scheduler minimizing (3.17). The intuition for this is quite simple. When the variance is small compared to the expectation, the Gaussian distribution will be approximately a Dirac delta in comparison to the wider pdf:s $P(n_u \mid I)$ for the inflows (remember from Figure 3.1 that $P(n_u \mid I)$ grows wider with larger $\langle n_u\rangle$ explaining the fact that the Gaussian

pdf will be more Dirac-like in comparison to $P(n_u \mid I)$ when $\langle n_u \rangle$ is large), and the simpler solution from (3.17) will give equally valid scheduling decisions.

As for the calculation of $\langle L_{2u} \rangle$, the part of the loss when $S_u + x_u \leq 0$, we obtain (following the procedure in Appendix A)

$$\begin{aligned} \langle L_{2u} \rangle &= \int_{-\infty}^{-S_u} P(x_u \mid I) \langle L_{2u,P(n|I)} \rangle dx_u \\ &= \frac{1}{2}\left[(\langle n_u \rangle - S_u - \langle x_u \rangle) \left(1 - \text{erf}\left(\frac{S_u + \langle x_u \rangle}{\sqrt{2\delta_u^2}} \right) \right) \right] \\ &+ \sqrt{\frac{\delta_u^2}{2\pi}} \exp\left(-\frac{1}{2\delta_u^2} (S_u + \langle x_u \rangle)^2 \right) . \end{aligned} \tag{3.40}$$

Notice that when the uncertainties δ_u concerning x_u becomes large, then depending on the sign of $S_u + x_u$, $\langle L_{2u} \rangle$ either vanishes (when $S_u + x_u \leq 0$) or becomes equal to $\langle n_u \rangle - S_u - \langle x_u \rangle$, as the erf$(\cdot)$ expression becomes equal to 1 or -1. This corresponds beautifully to the behavior we wish for, and even though the desiderata of Chapter 2 should guarantee this type of results, the fact that simple mathematical rules can yield such complex and at the same time intuitive behavior is nevertheless remarkable.

In summary, the expected loss with predicted production capacities $\langle c_{urt} \rangle$, known prediction accuracies σ_{urt} and known expected demands $\langle n_u \rangle$ is

$$\langle L \rangle = \sum_{u=1}^{U} \left(h(\rho_{\mathbf{u}}) + v_u \langle L_{1u} \rangle + v_u \langle L_{2u} \rangle \right) , \tag{3.41}$$

with $\langle L_{1u} \rangle$, $\langle L_{2u} \rangle$ given by (3.36), (3.40), respectively. Note that when $S_u \geq 0$ (implying that $x_u + S_u > 0$ for sure) the expected loss simplifies to

$$\langle L \rangle = \sum_{u=1}^{U} \left(h(\rho_{\mathbf{u}}) + v_u \langle L_{1u} \rangle \right) \qquad S_u \geq 0 . \tag{3.42}$$

Small set of capacities with known number of past occurrences

Let us now turn to the problem of finding the probability distribution for the production capacities c_{urt} of each production unit when the capacity in each time slot can assume only a limited set of values, $c_{urt} = c_{ur,1} ... c_{ur,K_{ur}}$.

The production manager monitors and keeps a record of the relative frequencies with which the different $c_{ur,k}$ are used. Assume that in its past history, the r:th production unit could produce $c_{ur,k}$ type-u widgets in $m_{ur,k}$ time slots out of the

total number of monitored time slots this unit was in production. The total number M_{ur} of monitored slots that unit r produced widgets of type u is

$$M_{ur} = \sum_{k=1}^{K_{ur}} m_{ur,k} \ .$$

From these numbers, what can we say about the plausibility for achieving capacity $c_{ur,k}$ in each of the time slots that constitute the coming period $t = 1 \dots T$? We shall assume that the frequencies with which different $c_{ur,k}$ occur are stationary over time, and take the expectation of the relative frequencies with which they occur as the probability for each $c_{ur,k}$ in all time slots. Assuming that the underlying physical mechanisms which determine the capacities do not change significantly with time, it follows that the relative frequencies should remain constant as well. The problem of translating relative frequencies observed in a finite interval into predictive probabilities is given by Laplace's rule of succession, derived and commented on in Section 2.6.

We seek to evaluate

$$\begin{aligned} &P(f_{ur,1}...f_{ur,K_{ur}}|m_{ur,1}...m_{ur,K_{ur}}I) = \\ &\quad = \frac{P(m_{ur,1}...m_{ur,K_{ur}}|f_{ur,1}...f_{ur,K_{ur}}I)P(f_{ur,1}...f_{ur,K_{ur}}|I)}{P(m_{ur,1}...m_{ur,K_{ur}}|I)} \end{aligned} \tag{3.43}$$

where $f_{ur,k}$ is the relative frequency with which $c_{ur,k}$ will be used, and I is all our background information that is relevant to the problem.

In the following, we will require that the production capacities for all widget types are known for the monitored M_{ur} time slots. In some cases however, it may be that the production capacities can only be recorded for the widget type that was actually produced in a given time slot. The other $u - 1$ capacities for that time slot would then be unknown. This is an instance of a *missing data* problem (also known as *censored*, or *gapped* data in the vast literature on this matter). It presents no new conceptual problems to us; we just apply our Chapter 2 rules and average the result we obtain below over the pdf for the unknown records. In the random-variable approach to probability theory, however, this is a problem which causes major concerns. The most usual *ad hoc* approach is to use estimates of the unknown data and treat them as if they were real. Obviously, the accuracy of the resulting inferences are then overestimated. Although an interesting topic in itself, we leave the missing data problem and treat only the case where we know the m_{ur} for all widget types u.

The probability for having production capacity $c_{ur,k}$ in an arbitrary time slot during the next scheduled T slots is then given by

$$p_{c_{ur,k}} \stackrel{\triangle}{=} P(c_{ur,k}|m_{ur,1}...m_{ur,K_{ur}}I) = \langle f_{ur,k} \rangle = \frac{m_{ur,k} + 1}{M_{ur} + K_{ur}} \ . \tag{3.44}$$

As in the case with predicted capacities and known prediction accuracy (3.41), the expected loss with a small set of possible capacity levels and known past frequencies is made up of the known cost $h(\rho_{\mathbf{u}})$ and the contributions $v_u\langle L_{1u}\rangle$ and $v_u\langle L_{2u}\rangle$ for the cases $x_u + S_u > 0$ and $x_u + S_u \leq 0$ respectively,

$$\langle L\rangle = \sum_{u=1}^{U} \left(h(\rho_{\mathbf{u}}) + v_u\langle L_{1u}\rangle + v_u\langle L_{2u}\rangle\right) \ , \tag{3.45}$$

but now with

$$\langle L_{1u}\rangle = \langle n_u\rangle \Big(\frac{\langle n_u\rangle}{\langle n_u\rangle + 1}\Big)^{S_u} \prod_{t=1}^{T}\prod_{r=1}^{R} \sum_{k \text{ s.t. } x_u > -S_u} p_{c_{ur,k}} \Big(\frac{\langle n_u\rangle}{\langle n_u\rangle + 1}\Big)^{c_{ur,k}\rho_{urt}} \ , \tag{3.46}$$

(where we rewrote the expected loss for known c_{urt} (3.18) according to the algebraic relation $x^{a+b} = x^a x^b$ and then averaged over $p_{c_{ur,k}}$) with $p_{c_{ur,k}}$ given by (3.44). Similarly,

$$\langle L_{2u}\rangle = \sum_{t=1}^{T}\sum_{r=1}^{R} \sum_{k \text{ s.t. } x_u \leq -S_u} p_{c_{ur,k}} \left(\langle n_u\rangle - S_u - \rho_{urt}c_{urt}\right) \ . \tag{3.47}$$

Computing the exact expected loss becomes difficult due to the summation over k s.t. $x_u \leq -S_u$ and k s.t. $x_u > -S_u$. In cases where $S_u \geq 0$, which typically would be the case, then of course $x_u + S_u \geq 0$ and the expected loss reduces to

$$\langle L\rangle = \sum_{u=1}^{U} \left(h(\rho_{\mathbf{u}}) + v_u\langle L_{1u}\rangle\right) \tag{3.48}$$

where the summation is considerably simplified,

$$\langle L_{1u}\rangle = \langle n_u\rangle \Big(\frac{\langle n_u\rangle}{\langle n_u\rangle + 1}\Big)^{S_u} \prod_{t=1}^{T}\prod_{r=1}^{R}\sum_{k=1}^{K_{ur}} p_{c_{ur,k}} \Big(\frac{\langle n_u\rangle}{\langle n_u\rangle + 1}\Big)^{c_{ur,k}\rho_{urt}} \qquad S_u > 0 \ . \tag{3.49}$$

3.3 Numerical Examples

In the following examples, we will concentrate on how uncertainty affects the resource allocation decisions. We assume (c.f. the loss expression (3.1)) that the known cost is equal to zero, $h(\rho_{\mathbf{u}}) = 0$, and $v_u = 1$ for all widget types u. This means that we concentrate only on the cost associated with unfulfilled orders and

set the cost per unit equal to one for all types of widgets. We will also assume in these examples that if an order is not met within the scheduled time horizon, that order does not expire, but simply reduces the number of widgets in stock to a negative number. In effect, there is no deadline for meeting the orders. In many cases, orders may expire unless met within a given time frame. In these cases, that deadline sets the natural scheduling horizon T. Note that the average loss expressions are exactly the same in both cases, but we choose to focus on the case without deadlines in the following simulations.

3.3.1 Comparison with a simple *ad hoc* approach

Assuming that we know exactly all production capacities but only the expected order sizes, as in Section 3.2.1, what could we do without using probability theory as extended logic?

Most people would presumably make an estimate of the loss

$$\hat{L} = \sum_{u=1}^{U} \left(v_u \cdot g(\langle n_u \rangle - S_u - \sum_{r=1}^{R} \sum_{t=1}^{T} \rho_{urt} c_{urt}) + h(\rho_{\mathbf{u}}) \right) , \tag{3.50}$$

using the expected demand $\langle n_u \rangle$ in lieu of the true future demand. Now, this is a violation of the sum rule which behooves us to take into account all possible future demands by summing together all such loss contributions weighed by their respective probabilities. Using (3.50) is the same thing as dogmatically denying that any other value can occur. We will now look at what the effects of not admitting the full extent of our uncertainty may lead to in a specific scenario.

In the considered situation there are three widget types, $U = 3$, the average order sizes are

$$\langle n_1 \rangle = 120 , \qquad \langle n_2 \rangle = 130 , \qquad \langle n_3 \rangle = 90 , \tag{3.51}$$

and the optimization horizon is $T = 3$ weeks (we here drop the generic expression 'time slots' which seems inappropriate). There are $R = 2$ production units, having independently varying manufacturing capacities but the same average capacities. The capacities are produced by a random-number generator mimicking a Rayleigh frequency distribution[1] with the average capacities

$$\overline{c_{1r}} = 150 , \qquad \overline{c_{2r}} = 138 , \qquad \overline{c_{3r}} = 81 , \qquad r = 1, 2.$$

Over a total time of 60 weeks, two factories are simulated; one (A) relying on the expected loss expression (3.17), and another one (B) using instead the loss

[1] This choice is arbitrary and only amounts to simulating manufacturing capacities with some variation. The Rayleigh frequency distribution does not reflect any typical real situation.

estimator (3.50). Identical orders and manufacturing capacities are generated for the two factories, and at the end of the 60-week period the number of widgets that have been ordered but not yet produced (i.e. the number of outstanding orders given that one order is always for exactly one widget) are reported. The order sizes are generated from a Poisson[2] random-number generator with the average sizes given by (3.51).

Running the simulation and averaging the result over 50 simulations, the number of widgets left (or if negative, the number of remaining unfilled orders) in stock after the 60 weeks are:

$$\begin{aligned} &\text{Factory A:} \quad \overline{S_1} = 1\,, \quad \overline{S_2} = -61\,, \quad \overline{S_3} = -255 \\ &\text{Factory B:} \quad \overline{S_1} = -7\,, \quad \overline{S_2} = -19\,, \quad \overline{S_3} = -762\,. \end{aligned}$$

At the end of the 60-week period factory A has $255 + 61 = 316$ unfilled orders (and an extra widget of type 1 in stock), whereas factory B has 788 unfilled orders, approximately two and a half times as many as does factory A. The difference in absolute numbers is large as well. Whatever the value of each widget, multiply that number by 472 and you obtain the resulting loss that factory B makes because it uses an uncertain estimate $\langle n_u \rangle$ as were it indeed the true value instead of assigning probabilities for different possible outcomes of n_u.

3.3.2 The behavior of the expected loss as a function of widgets in stock

In Figure 3.2 the expected loss (3.18) with known mean order size $\langle n_1 \rangle = 20$ and exactly known capacity is plotted for one widget, one production unit and $T = 1$ as a function of $x_1 + S_1$ (negative and positive values, the former indicating outstanding known orders). It is seen that the expected loss is equal to $\langle n_1 \rangle$ at $x_1 + S_1 = 0$, and it then decays very slowly towards zero as the stock size increases. This reflects that even very large order sizes cannot be ruled out on the information at hand. Only with more definitive knowledge or order sizes, for instance in the form on known upper bounds, can we hope to achieve a faster decay to zero expected loss.

3.3.3 The effects of increasing capacity uncertainty

When the production capacity cannot be predicted with absolute certainty, but we instead can use a Gaussian probability distribution for the c_{urt} we would expect

[2] Again, this is an arbitrary choice. A better test would rely on real data from some manufacturing plant. Regrettably, we do not have access to such records.

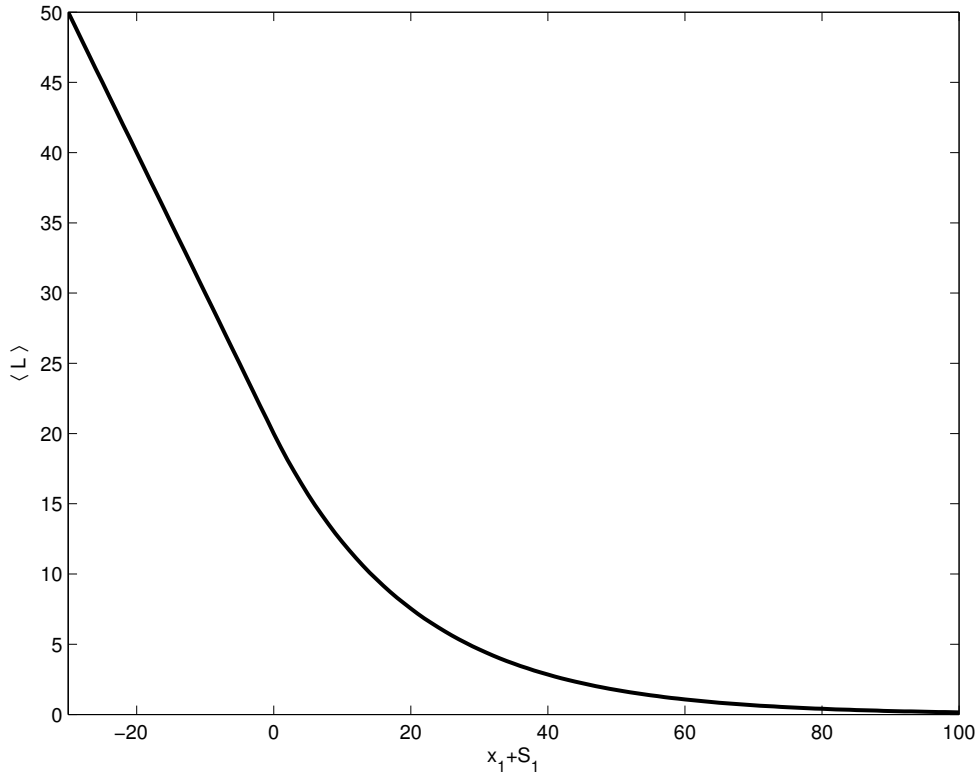

Figure 3.2: The expected loss (3.18) with known average demand $\langle n_1 \rangle = 20$ as a function of the production capacity plus the widgets in stock $x_1 + S_1$.

that the resource allocation policy will be less inclined to use exclusive allocations as the risk of obtaining zero capacity ought to be larger than when distributing the workload over several production units. We should however note that when we have a very large number of unfilled orders, so that $x_u + S_u < 0$ almost surely, then the expected loss for widget u with prediction uncertainty becomes $\langle L_u \rangle = \langle n_u \rangle - \langle x_u \rangle - S_u$, which is independent of the prediction uncertainty σ_{urt} in the Gaussian model. This may be somewhat surprising, as in this situation we might think that a rational decision in the choice between two production units is to use one with slightly lower expected capacity if that unit has much lower uncertainty than the other, or at least that we should spread the risk and split the work between both units.

To see why our resource allocation algorithm always picks only the production unit with the largest expected capacity, irrespective of the prediction uncertainty,

we have to understand our choice of criterion and our uncertainty model. First, the loss criterion in the case when we know $x_u + S_u < 0$ says that everything we produce will be sold. There is no risk of producing widgets that will not be sold. Second, the Gaussian distribution is perfectly symmetrical, meaning that even if there is a risk for lower capacity than expected, there is an equal chance of *larger* capacity.

Let us think about a simpler, but similar, situation in which there is only one widget type. Suppose that there are three possible capacities, $c = 1$, $c = 3$ or $c = 5$ with probability 0.25, 0.5 and 0.25, respectively for production unit 1, while unit 2 has capacity $c = 3$ with certainty. With our current loss function and $S \ll 0$ any allocation will amount to the same expected loss.

How would you decide? If your choice disagrees with that of our algorithm, the reason must be that you use a different loss criterion. One who prefers to use unit 2 exclusively would in effect have a mental loss function which does not grow as fast as our function $L = c$. The symmetry would be broken. For instance a logarithmic function $L = \log(c)$ would give that use of unit 2 is slightly better than using unit 1. Indeed, a logarithmic loss is often a very adequate description of a 'rational' decision metric, since it does place equal weight to an increase of a factor x as to a decrease of a factor x. Many times, a doubling has the same positive effect as a halving of some quantity has a negative effect. Further, for someone with a yearly income of 30000 Euros a salary increase of 30000 Euros would presumably mean much more to him than to someone with a yearly income of 30000000 Euros. With a linear loss function both persons would benefit just as much from the 30000 Euro increase. A more sensible loss function would be the logarithm of the new salary relative to the old salary, giving a clear preference for the less well-situated fellow. Daniel Bernoulli (1738) described in a very clear and well-written memoir – which is still, almost 300 years later, well worth reading – how the logarithmic loss function, or equivalently utility, mostly correspond to how people tend to reason in practice. We will return to discussing logarithmic loss functions in Chapter 5 in connection with scheduling users in cellular communications systems. In the current application, however, we think that the linear loss (for positive values) is suitable in that factories typically are expected to maximize absolute incomes[3].

But how does prediction uncertainty affect decisions for the case when $x_u + S_u$ is or may be larger than zero? In such situations, there is a risk that overproduction occurs if the capacity becomes larger than expected. Thus, we should expect a tendency to spread the production over several widget types when capacity uncertainty increases. Figure 3.3 shows the expected loss (3.41) for different prediction

[3]But the ideas of Daniel Bernoulli might still make better sense also in companies. We leave this as an issue for further contemplation.

uncertainties in a scenario with one production unit, $R = 1$, two widget types $U = 2$ and $T = 1$, as a function of ρ_1 (note that $\rho_2 = 1 - \rho_1$). The expected demand, the number of widgets in stock of each type, and the predicted capacities are respectively

$$\begin{aligned} \langle n_1 \rangle &= 20 \,, & \langle n_2 \rangle &= 10 \,, \\ S_1 &= 10 \,, & S_2 &= 10 \,, \\ \langle c_1 \rangle &= 10 \,, & \langle c_2 \rangle &= 10. \end{aligned}$$

As expected, when uncertainty increases the optimum resource allocation becomes less inclined to concentrate all resources on producing only one type of widget.

3.4 Extensions and Modifications

We have used a problem formulation (3.1) where the demand adds to the loss function, and the known supply S_u and the production capacity subtracts from the loss. A mathematically very similar problem is to instead consider an additive demand consisting of an unknown component n_u, and a known component S_u where there is no previous 'supply' which subtracts from the loss, but only a subtractive future component c_{urt}. This would be the case in flow control problems, where the allocation decisions consist of turning on or off (partly or fully) 'knobs' which control the magnitude c_{urt} of a flow. The demand n_u is then to be thought of as the number of 'packets' that are to be sent to some destination u. These packets, if left in the outgoing buffers, add to the loss just as unfilled orders do in the manufacturing plant. The only difference is that there is no 'stock' of capacity to build up in advance. The capacities cannot be saved for later, but must be used or wasted. The formulation is then

$$L = \sum_{u=1}^{U} v_u g(n_u + S_u - \sum_{r=1}^{R} \sum_{t=1}^{T} \rho_{urt} c_{urt}) + h(\rho_{\mathbf{u}}) \,, \tag{3.53}$$

where n_u is the number of incoming packets, S_u is the number of packets already in stock, and c_{urt} is the capacity in terms of the number of packets that a 'channel' between the allocation central and the destination u can send at time t. This formulation is clearly analogous to the cases we have covered in this chapter, but applies in partly different problems where capacities cannot be stored for later use. This formulation will be our starting point in scheduling users in mobile communications, Chapter 5.

In the loss formulation (3.1), there are two components that we have not commented much on. The role of v_u is simply to associate a cost with different widget

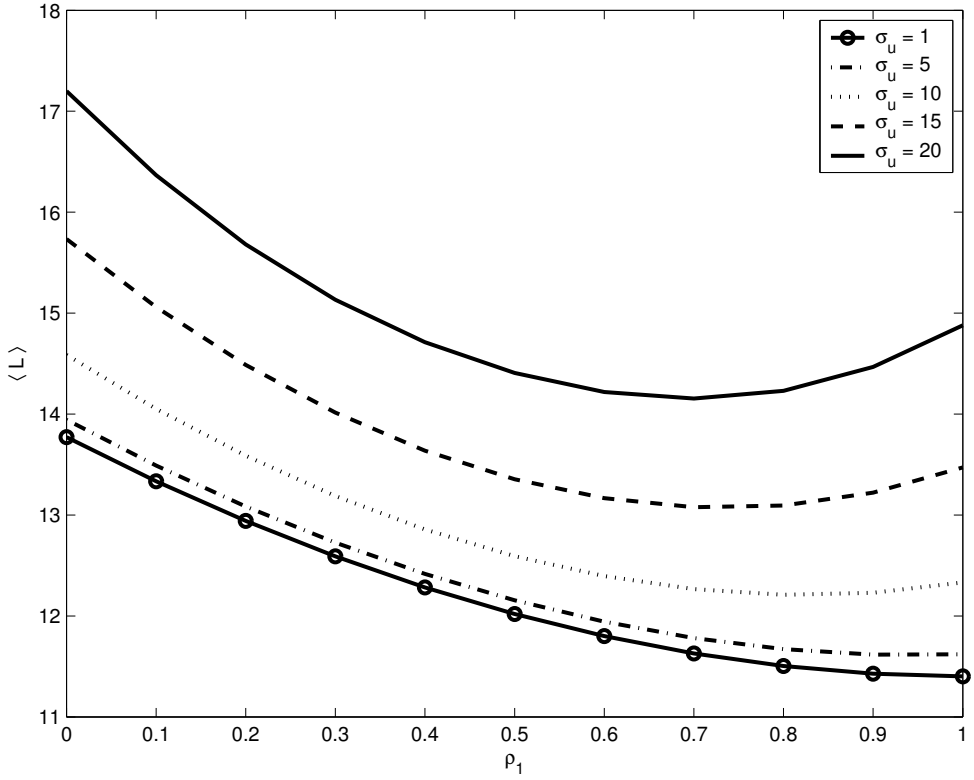

Figure 3.3: The expected loss $\langle L \rangle$ in (3.41) for different production uncertainties σ_u as a function of the allocation $\rho_1 = 1 - \rho_2$. When the uncertainty increases, the optimum allocation is to spread out production due to the risk of overproduction. Note however that there is an additional effect which comes into play when the uncertainty σ_u becomes large; the Gaussian distribution's tail is then non-negligible for $c_{urt} < 0$ which means that our approximation of the truncated distribution by the entire distribution in (3.33) looses accuracy.

types. A further refinement would be to have a cost factor v_{ur} for each production unit, thus setting

$$x_u = \sum_r \sum_t v_{ur} \rho_{urt} c_{urt} \ . \tag{3.54}$$

This would reflect that some production units are more costly to use than others.

The additive known cost $h(\rho_{\mathbf{u}})$ can be used to express costs for transferring production from one unit to another or to put other costs on the detailed structure of the allocation matrix $\rho_{\mathbf{u}}$. Typically, changes in production patterns may induce certain costs, and thus any $\rho_{\mathbf{u}}$ that differs from that of the previous scheduling

period may be penalized.

None of these two parameters however allow us to separate customers and make priorities among them. In order to do that we could try to generalize the model further, but that does not seem to be a straightforward route to take. Instead, we could set hard constraints on the resource allocation matrix if we know in advance how many orders are placed for different widgets by different customers. We can also prioritize some customers by our choice of how we distribute the produced widgets.

3.5 Conclusions

In this chapter we extended the 'widget problem' described by Jaynes (1963b) to encompass a slightly more general problem scenario and discussed how uncertainty regarding supply and demand affects optimum resource allocation decisions. We saw that acknowledging uncertainty generally results in 'hedging the bets' and spreading production over several production units if the loss function or the probability distribution is asymmetric. If we neglect uncertainty and treat estimates as if they were true values substantial economic loss can result.

The problem we treated in this chapter is a basic building block for the remaining problems that we consider in this thesis. The three probability distributions that we found useful in solving this problem – the exponential distribution, the Gaussian, and the general form of Laplace's rule of succession – will come to play a central part in the chapters to come. We stress that their importance and their frequent occurrence in many practical problems lie not in any imagined frequency correspondences with real phenomena but rather that they represent certain basic states of knowledge. They are uniquely determined as the valid models for exactly those states of knowledge. To use any other distribution in such a situation would require additional information. In many cases, such extra information must be rather precise and limit the entropy of the reasoner significantly for it to be worthwhile to use that more complex model. This should be emphasized and is a topic that requires more research – how large must the entropy difference between two probability distributions be for it to have a significant impact on the resulting inferences? A general answer may be difficult to give, as there is a clear dependence on the specific loss function. We however suspect that for some rather general class of loss functions, say differentiable symmetric functions, a precise answer may be within reach. With such a result, we could determine what type of information to look for in order to get the highest performance improvement, and how to make a proper balance between computational costs and the quality of the corresponding decision making. We leave these questions for future research.

Appendix 3.A Derivation of Expected Loss given Expected Order Sizes

In Section 3.2.1, in the derivation of the expected loss assuming knowledge of expected order sizes, we need to evaluate the summation over n_u in (3.16). Depending on the sign of $x_u + S_u$ we obtain different solutions[4]. First assume $x_u + S_u$ is positive. Then, using the probability assignment (3.14) for the order sizes and neglecting the additive term $h_u(\rho_{\mathbf{u}})$ (which is known and independent of the probability assignments), we obtain:

[4]Notice that in the case of known capacities the term $x_u + S_u$ is always known.

$$
\begin{aligned}
& \sum_{u=1}^{U} \sum_{n_u=0}^{\infty} P(n_u \mid I) g(n_u - S_u - x_u) = \\
& \sum_{u=1}^{U} \Big(\sum_{n_u=0}^{\infty} P(n_u \mid I)(n_u - S_u - x_u) \\
& - \sum_{n_u=0}^{x_u+S_u} P(n_u \mid I)(n_u - S_u - x_u) \Big) \\
= & \sum_{u=1}^{U} \Big(\sum_{n_u=0}^{\infty} \frac{1}{\langle n_u \rangle + 1} \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{n_u} (n_u - S_u - x_u) \\
& - \sum_{n_u=0}^{x_u+S_u} \frac{1}{\langle n_u \rangle + 1} \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{n_u} (n_u - S_u - x_u) \Big) \\
= & \sum_{u=1}^{U} \Big(\sum_{n_u=0}^{\infty} \frac{1}{\langle n_u \rangle + 1} \Big[\Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{n_u} n_u && (3.55) \\
& - \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{n_u} (S_u + x_u) \Big] && (3.56) \\
& - \sum_{n_u=0}^{x_u+S_u} \frac{1}{\langle n_u \rangle + 1} \Big[\Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{n_u} n_u && (3.57) \\
& - \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{n_u} (S_u + x_u) \Big] && (3.58) \\
= & \sum_{u=1}^{U} \Big[\langle n_u \rangle - S_u - x_u && (3.59) \\
& + (S_u + x_u) \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{x_u+S_u+1} && (3.60) \\
& - \langle n_u \rangle \Big(1 - \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{x_u+S_u} \Big) && (3.61) \\
& + (S_u + x_u) \Big(1 - \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{x_u+S_u+1} \Big) \Big] && (3.62) \\
= & \sum_{u=1}^{U} \langle n_u \rangle \Big(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \Big)^{x_u+S_u} . && (3.63)
\end{aligned}
$$

The infinite progression in lines (3.55) and (3.56) are standard sums which can be found in (Gradshteyn and Ryzhik, 2000) (eqns. 0.231.2 and 0.231.1). They correspond to the solution (3.59). The finite sum in lines (3.57) and (3.58) can also be found in (Gradshteyn and Ryzhik, 2000) (eqns. 0.113 and 0.112). The arithmetico-

geometric progression (3.57) corresponds to the solution spanning lines (3.60) and (3.61), while the geometric series (3.58) corresponds to the solution (3.62).

Now, if $x_u + S_u$ is zero or negative[5] the sum on line (3.55) vanishes. We then obtain:

$$\sum_{u=1}^{U} \sum_{n_u=0}^{\infty} P(n_u \mid I) g(n_u - S_u - x_u) = \sum_{u=1}^{U} \langle n_u \rangle - S_u - x_u \; . \tag{3.64}$$

In summary, the solution is to minimize (3.63) if $x_u + S_u > 0$, and (3.64) otherwise.

Appendix 3.B Derivation of Expected Loss given Past Orders

We here determine the expected loss when marginalizing the maximum-entropy distribution (3.14) over its expectation $\mu = \langle n_u \rangle$ based on a short data record $\{y\} = \{y_1 \ldots y_N\}$ of past order sizes.

Since all that is known about the expected order size is that it is non-negative, a reasonable prior $P(\mu \mid I)$ for the expected order size is Jeffrey's prior:

$$P(\mu \mid I) \propto \frac{1}{\mu} \; . \tag{3.65}$$

Note that a uniform prior would be inappropriate since for any given point on the μ axis, the ratio of the probability for obtaining a larger value and the probability for obtaining a smaller value would always be infinite. A motivation for using Jeffrey's prior lies in the observation that it gives equal probability to the parameter being larger than any given value, as to it being smaller. Note that this is a different motivation than the one we used in determining that Jeffrey's prior is uninformative for the case of scale parameters, c.f. Section 2.7.4.

Let $Y = y_1 + \ldots + y_N$ be the sum of all past order sizes for a particular widget.

[5] x_u is of course non-negative, but S_u may be negative, corresponding to a number of outstanding orders remaining from previous periods or new known orders.

Using Jeffrey's prior (3.65) and then inserting (3.20) into (3.19) we obtain:

$$\begin{aligned} P(n \mid \{y\}, I) &= \int_0^\infty P(n \mid \mu, I) P(\{y\} \mid \mu, I) P(\mu \mid I) d\mu \\ &\propto \int_0^\infty \left(\frac{1}{\mu+1}\right)\left(\frac{\mu}{\mu+1}\right)^n \left(\frac{1}{\mu+1}\right)\left(\frac{\mu}{\mu+1}\right)^{y_1} \cdots \\ &\cdots \left(\frac{1}{\mu+1}\right)\left(\frac{\mu}{\mu+1}\right)^{y_N} \mu^{-1} d\mu \\ &= \int_0^\infty \left(\frac{1}{\mu+1}\right)^{N+1+n+Y} \mu^{n+Y-1} d\mu \qquad (3.66) \\ &= \frac{N!(n+Y-1)!}{(N+n+Y)!} \ , \qquad (3.67) \end{aligned}$$

where the solution to the last integral is given by equation 3.194.3 in Gradshteyn and Ryzhik (2000). The normalizing constant (to make the probability sum to unity) is found by summing (3.67) over all n. In summary, we obtain

$$P(n \mid \{y\}, I) = N \frac{(n+Y-1)!(N+Y-1)!}{(Y-1)!(N+n+Y)!} = N \frac{\prod_{i=0}^{N-1}(Y+i)}{\prod_{i=0}^{N}(Y+n+i)} \ . \quad (3.68)$$

Interestingly, the expected value turns out not to be the arithmetic mean, but

$$\langle n \rangle = \sum_{n=0}^{\infty} n P(n \mid \{y\}, I) = \frac{Y}{N-1} \ , \qquad (3.69)$$

which reflects the fact that the interval is open to the right side while bounded from the left. This means that the best estimate as to the next value of an independent non-negative sequence is slightly larger than the arithmetic mean. This is an estimate that unguided intuition would not conjecture. After giving it some thought however, we find that it is a very reasonable estimate, indeed more plausible than the arithmetic mean. The reason, as mentioned, being that there are always infinitely many more larger values than there are smaller ones compared to any single number. It can be noted that the use of a uniform prior would make the estimate even larger[6] as it gives much higher initial probability to a large value. Of course, in the limit as $N \to \infty$, both estimates converge to the arithmetic mean.

Now, turning to our actual problem, we use our probability distribution (3.21) conditioned only on knowledge of past order sizes (and I of course) and determine the expected loss (3.6). Writing the expected loss $\langle L \rangle$ as in (3.17):

$$\sum_{u=1}^{U} h(\rho_{\mathbf{u}}) + v_u \langle L_u \rangle \ , \qquad (3.70)$$

[6] A straightforward calculation gives the result $\langle n \rangle = \frac{Y}{N-2}$ for a uniform prior.

we have for $x_u + S_u > 0$

$$\begin{aligned}\langle L_u \rangle &= \sum_{n_u=0}^{\infty} N \frac{(n_u + Y - 1)!(N + Y - 1)!}{(Y-1)!(N + n_u + Y)!}(n_u - S_u - x_u) \\ &= \frac{Y}{N-1} - S_u - x_u \ . \end{aligned} \tag{3.71}$$

For $x_u + S_u \leq 0$ we have

$$\begin{aligned}\langle L_u \rangle &= \sum_{n_u=0}^{x_u+S_u} N \frac{(n_u + Y - 1)!(N + Y - 1)!}{(Y-1)!(N + n_u + Y)!}(n_u - S_u - x_u) \\ &= \frac{Y}{N-1} - x_u - S_u + \Big[(Y + x_u + S_u + 1 + N)(Y + x_u + S_u)! \\ &\times \ (Y + N - 1)!((N+1)S_u - Y - (S_u + x_u + 1)N)\Big] \\ &\Big/ \ \Big[(N-1)(Y-1)!(Y + x_u + S_u + N + 1)!\Big] \\ &= \frac{Y}{N-1} - S_u - x_u - \frac{(Y + x_u + S_u)!(Y + N - 1)!(x_u + S_u + Y + N)}{(N-1)(Y-1)!(Y + x_u + S_u + N)!} \\ &= \frac{Y}{N-1} - S_u - x_u - \frac{(Y + x_u + S_u)!(Y + N - 1)!}{(N-1)(Y-1)!(Y + x_u + S_u + N - 1)!} \ . \end{aligned} \tag{3.72}$$

Taking the difference between (3.71) and (3.72) we obtain

$$\begin{aligned}&\frac{(Y + x_u + S_u)!(Y + N - 1)!}{(N-1)(Y-1)!(Y + x_u + S_u + N - 1)!} \\ &= \frac{1}{N-1} \frac{\prod_{i=0}^{N-1}(Y+i)}{\prod_{i=1}^{N-1}(Y + x_u + S_u + i)} \\ &= \frac{Y}{N-1} \prod_{i=1}^{N-1} \frac{Y+i}{x_u + S_u + Y + i} \\ &= \frac{Y}{N-1} \prod_{i=1}^{N-1} \frac{1}{1 + \frac{x_u + S_u}{Y+i}} \ , \end{aligned} \tag{3.73}$$

which yields the sum solution

$$\begin{aligned}\langle L_u \rangle &= \sum_{n_u=0}^{\infty} P(n_u \mid \{y_u\}, I)\,(g(n_u - S_u - x_u)) \\ &= \begin{cases} \frac{Y}{N-1} \prod_{i=1}^{N-1} \frac{1}{1+\frac{x_u+S_u}{Y+i}} & , \ x_u + S_u > 0 \\ \frac{Y}{N-1} - S_u - x_u & , \ x_u + S_u \leq 0 \ . \end{cases} \end{aligned} \tag{3.74}$$

The result (3.74) for $x_u + S_u > 0$ is an elegant but non-trivial weighting of the expected order size $\mu = \frac{Y}{N-1}$, the widgets in stock, S_u, and the number of widgets to be produced, x_u. The actual calculation of (3.74) for $x_u + S_u > 0$ is however a bit cumbersome when N_u becomes large. The following result establishes that an excellent approximation of (3.74) is obtained by using (3.18) with $\langle n_u \rangle = \frac{Y}{N-1}$.

Result 3.1 *Let Y and $x_u + S_u$ be positive real numbers, and $N > 2$ a positive integer. Then,*

$$\frac{Y}{N-1} \prod_{i=1}^{N-1} \frac{1}{1 + \frac{x_u + S_u}{Y+i}} \geq \mu \left(\frac{\mu}{\mu + 1} \right)^{x_u + S_u} . \tag{3.75}$$

where $\mu = \frac{Y}{N-1}$. The inequality tends to equality as $N \to \infty$.

Proof: Recognizing that the right-hand side of (3.75) can be rewritten as

$$\mu \left(\frac{\mu}{\mu + 1} \right)^{x_u + S_u} = \frac{Y}{N-1} \left(\frac{1}{1 + \frac{N-1}{Y}} \right)^{x_u + S_u} \tag{3.76}$$

the inequality (3.75) is simplified, and we rewrite the relation in a more compact form,

$$\prod_{i=1}^{M} \left(1 + \frac{\alpha}{y+i} \right) \leq \left(1 + \frac{M}{y} \right)^{\alpha} , \tag{3.77}$$

with $\alpha > 0$, $y > 0$ and $M > 0$.

Taking the logarithm of the left side we obtain

$$\sum_{i=1}^{M} \log \left(1 + \frac{\alpha}{y+i} \right) \leq \sum_{i=1}^{M} \frac{\alpha}{y+i} \tag{3.78}$$

with equality for $\frac{\alpha}{y+i} = 0$. Inserting this into (3.77) we further find that

$$\exp \left(\alpha \sum_{i=1}^{M} \frac{1}{y+i} \right) \leq \exp \left[\alpha \left(\ln(y+M) - \ln(y+1) \right) \right] \tag{3.79}$$

$$= \left(\frac{y+M}{y+1} \right)^{\alpha} = \left(\frac{1 + M/y}{1 + 1/y} \right)^{\alpha} \leq (1 + M/y)^{\alpha} \tag{3.80}$$

where the inequality in (3.79) tends to equality as M grows to infinity. This concludes the proof, and it is seen that the approximation (3.75) gains in accuracy when the number of observations, N, is large and the sum of the past order sizes, Y, is large. It has been verified by simulation that the approximation is excellent even for small values of N and Y.

■

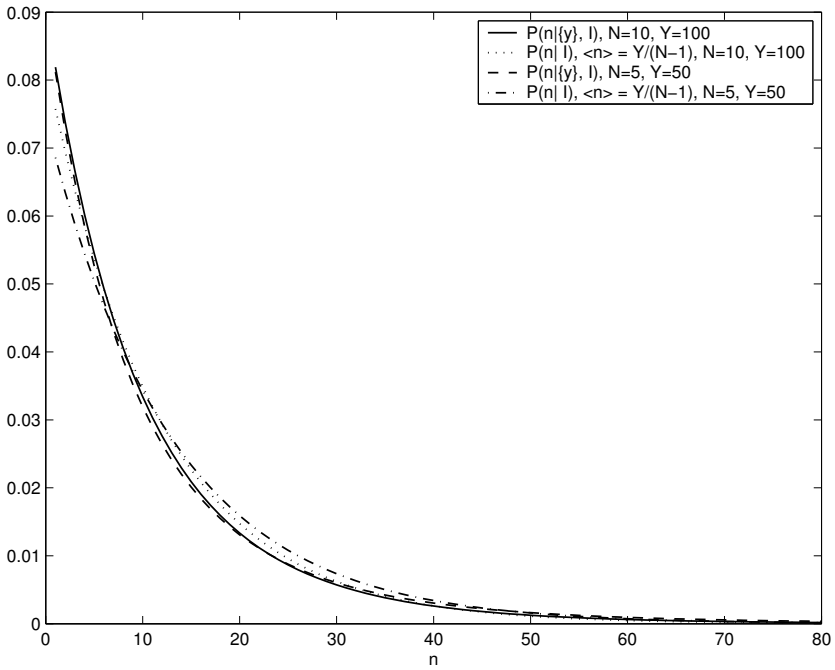

Figure 3.4: The exact marginal probability distribution (3.21) for n with knowledge of past data, and the approximation (3.14) with $\langle n_u \rangle = \frac{Y}{N-1}$.

In Figure 3.4, we plot the exact probability distribution (3.21) for future order sizes given past data and, in comparison, the approximate distribution using (3.14) with $\langle n_u \rangle = \frac{Y}{N-1}$ for different values of N. We see that the approximation is indeed very near the exact curve.

Appendix 3.C Derivation of Expected Loss for Partitioned Intervals

When the n_u axis has been partitioned into K intervals, for example logarithmically spaced such as in Section 3.2.2, the expected loss contribution given that the order size is within bin k (ranging over $a_k \leq n_u \leq b_k - 1$) becomes

$$\langle L_u \mid n_u \in k \rangle = \sum_{n_u=a_k}^{b_k-1} \frac{1}{b_k - a_k} g(n_u - S_u - x_u) \,, \qquad k = 1 \ldots K-1. \quad (3.81)$$

Recalling that $g(n_u - S_u - x_u) = 0$ if $n_u \leq S_u + x_u$ and $g(n_u - S_u - x_u) = n_u - S_u - x_u$ otherwise, we rewrite the expected loss as

$$\langle L_u \mid n_u \in k \rangle = \sum_{n_u=\alpha_k}^{\beta_k-1} \frac{n_u - S_u - x_u}{b_k - a_k} \tag{3.82}$$

where

$$\alpha_k \stackrel{\triangle}{=} \max(S_u + x_u, a_k) \tag{3.83}$$

and

$$\beta_k \stackrel{\triangle}{=} \max(S_u + x_u + 1, b_k) \; . \tag{3.84}$$

The first part of the sum in (3.82) is

$$\begin{aligned} & \frac{1}{b_k - a_k} \sum_{n_u=\alpha_k}^{\beta_k-1} n_u \\ = \; & \frac{1}{2} \frac{\beta_k^2 - \beta_k - (\alpha_k^2 - \alpha_k)}{b_k - a_k} \; . \end{aligned} \tag{3.85}$$

To see this, note that

$$\sum_{n_u=\alpha_k}^{\beta_k-1} n_u = \sum_{n_u=0}^{\beta_k-1} n_u - \sum_{n_u=0}^{\alpha_k-1} n_u \tag{3.86}$$

and that a sum of the type $\sum_{n_u=0}^{x-1} n_u$ describes an area of a large triangle with sides of length $x-1$, i.e. having area $\frac{1}{2}(x-1)^2$, plus the area of $x-1$ small triangles, each of area $1/2$ (see Figure 3.5). The total area described by such an area is thus

$$\frac{1}{2}(x-1)^2 + \frac{1}{2}(x-1) = \frac{1}{2}(x^2 - x) \; . \tag{3.87}$$

The second part of the sum in (3.82) is

$$\begin{aligned} & \frac{1}{b_k - a_k} \sum_{n_u=\alpha_k}^{\beta_k-1} (-S_u - x_u) \\ = \; & -\frac{\beta_k - \alpha_k}{b_k - a_k}(S_u + x_u) \; . \end{aligned} \tag{3.88}$$

Combining these results we obtain

$$\langle L_u \mid n_u \in k \rangle = \frac{1}{2} \frac{\beta_k^2 - \beta_k - (\alpha_k^2 - \alpha_k)}{b_k - a_k} - \frac{\beta_k - \alpha_k}{b_k - a_k}(S_u + x_u) \; . \tag{3.89}$$

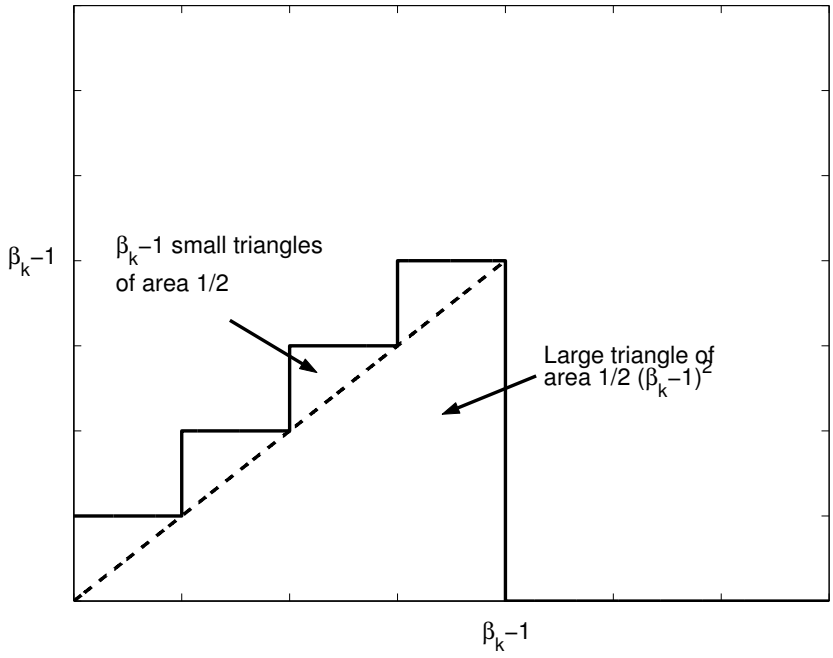

Figure 3.5: The sum $\sum_{n_u=0}^{\beta_k-1} n_u$ describes the area under the curve, the sum of a large triangle and a number of smaller triangles. Here, $\beta_k - 1 = 4$.

Appendix 3.D Derivation of Expected Loss given Uncertain Production Capacities

We here derive a general expression for the expected loss when production capacities and order sizes are uncertain.

Recall that the expected loss (3.17) was either one of two possibilities, depending on the sign of $x_u + S_u$. This did not present any problem since $x_u = \sum_{t=1}^{T}\sum_{r=1}^{R} c_{urt}\rho_{urt}$ and S_u were known. Here, however, c_{urt} is uncertain, and consequently so is $x_u + S_u$. We must therefore also consider the probability q_u for the event $x_u + S_u > 0$. The expected loss contribution for widget type u is thus the expectation of L over the joint probability $P(n_u x_u \mid \{x_u + S_u > 0\}I)$ with probability q_u, and with probability $1 - q_u$ it is the expectation of L over $P(n_u x_u \mid \{x_u + S_u \leq 0\}I)$. Assuming that knowledge of n_u does not give any information about c_{urt} the joint probabilities factor into two independent factors

$$\begin{aligned} P(n_u x_u \mid \{x_u + S_u > 0\}I) &= P(n_u \mid \{x_u + S_u > 0\}I) \\ &\times \; P(x_u \mid \{x_u + S_u > 0\}I) \end{aligned} \tag{3.90}$$

(and likewise for the case when $x_u + S_u \leq 0$). This means that we can first determine the expectation of the loss based on the uncertainty concerning the orders n_u (this was derived in the previous section with the result (3.18)) and then average this expression over the uncertainty concerning x_u to arrive at the expected loss expression for the joint uncertainty about n_u and x_u.

Thus, according to the sum and product rules of probability theory, the expected loss contribution for type u widgets (c.f. (3.17)) becomes

$$\begin{aligned} \langle L_u \rangle &= q_u \int P(x_u \mid \{x_u + S_u > 0\} I) \langle L_{1u,P(n|I)} \rangle dx_u \\ &+ (1 - q_u) \int P(x_u \mid \{x_u + S_u \leq 0\} I) \langle L_{2u,P(n|I)} \rangle dx_u \end{aligned} \quad (3.91)$$

where $q_u = P(x_u + S_u > 0 \mid I)$, and $\langle L_{1u,P(n|I)} \rangle$ and $\langle L_{2u,P(n|I)} \rangle$ denotes the expected per-widget-type loss from Section 3.2.1 for $x_u + S_u > 0$ and $x_u + S_u \leq 0$ respectively. This new notation is used to make a distinction between the expected loss with respect to $P(n_u \mid I)$ in (3.17) and the one currently under investigation. $\langle L_{1u} \rangle$ and $\langle L_{2u} \rangle$ will now be used to describe the latter.

Now, in order to determine (3.91) we must first compute the probability distribution $P(x_u \mid \{x_u + S_u > 0\} I)$. According to the product rule (2.3),

$$P(x_u \mid \{x_u + S_u > 0\} I) = \frac{P(x_u \{x_u + S_u > 0\} \mid I)}{P(\{x_u + S_u > 0\} \mid I)} , \quad (3.92)$$

where

$$P(\{x_u + S_u > 0\} \mid I) = \int_{-S_u}^{\infty} P(x_u \mid I) dx_u \quad (3.93)$$

is a normalizing constant and

$$P(x_u \{x_u + S_u > 0\} \mid I) = P(x_u \mid I) \quad x_u > -S_u . \quad (3.94)$$

Thus, conditioning on $\{x_u + S_u > 0\} \Leftrightarrow \{x_u > -S_u\}$ simply truncates $P(x_u \mid I)$, putting a limit on the range of values that x_u can take, and yields

$$P(x_u \mid \{x_u + S_u > 0\} I) = \frac{P(x_u \mid I)}{\int_{-S_u}^{\infty} P(x_u \mid I) dx_u} \quad x_u > -S_u . \quad (3.95)$$

Note that the normalizing constant in this expression is equal to q_u and consequently the expected loss contribution for $x_u + S_u > 0$ becomes

$$\begin{aligned} \langle L_{1u} \rangle &= q_u \int P(x_u \mid \{x_u + S_u > 0\} I) \langle L_{1u,P(n|I)} \rangle dx_u \\ &= \int_{-S_u}^{\infty} P(x_u \mid I) \langle L_{1u,P(n|I)} \rangle dx_u . \end{aligned} \quad (3.96)$$

The case for $x_u + S_u \leq 0$ is entirely analogous.

We thus find that (3.91) becomes

$$\begin{aligned} \langle L_u \rangle &= \int_{-S_u}^{\infty} P(x_u \mid I) \langle L_{1u,P(n|I)} \rangle dx_u \\ &+ \int_{-\infty}^{-S_u} P(x_u \mid I) \langle L_{2u,P(n|I)} \rangle dx_u \; . \end{aligned} \tag{3.97}$$

Chapter 4

Bidding under Uncertainty in a Certain Type of Auctions

WE here consider a bidding situation in which customers compete for a resource which can only be used by one customer at a time. To each competitor the resource carries a certain utility, the *carrying capacity* of the resource, which varies over time. For instance, the carrying capacity may in a mobile telecommunications network be the time-varying data rate over the communications channel.

At an auction each competitor submits one sealed bid, and after all bids have been collected a winner is announced who gets access to the resource for a certain time period. For the next period, a new auction is carried out again under similar circumstances. A customer may come and go at any time, but in the presently considered applications a customer normally stays for a large number of auctions.

If the winning bid was q and the carrying capacity of the winning customer was c, the winning customer pays qc monetary units, i.e. q is the price per unit utility.

The auctioneer's income for each auction is thus determined by qc, and the winning customer is hence the one with maximum price-capacity product qc.

Our problem set-up is the following:

- Different bidders u may have different carrying capacities c_u
- Each bidder u reports its own carrying capacity c_u to the auctioneer along with its bid q_u. Both values are hidden for other customers.
- Although all information reported to the auctioneer is sealed, a bidder obtains some implicit information regarding other bidders' carrying capacities and bids from how many times the bidder wins the auction. The bidder does

however not know who wins an auction that is not won by the bidder, nor, in that case, the winning price-capacity product.

- The auctioneer knows all bidders' carrying capacities and bids.

The question we seek to answer is then: What is the best bid that a customer can make? Clearly, the answer depends on the customer's need for carrying capacity, and – having established a loss function describing this – any information at hand that can assist in reaching a decision. This type of problem was considered by Friedman (1956), and a similar strategy as the one we will use here was suggested. Friedman considers the objective of bidding for maximum expected profit in a scenario where a government agency invites a large number of companies in the same industry to bid for contracts. Friedman notes that 'the difficulty in determining the expected profit lies in determining ... the probability of winning as a function of the amount bid'. He suggests the use of histograms of bids from old auctions, assuming that all previous bids are made public after an auction. In our scenario, we do not assume knowledge of all previous bids. In many auctions only the winning bids are announced, and then Friedman's method would fail to determine a probability distribution for the other customers' bids. From our present understanding of probability theory as logic, however, the solution is straightforward. As always, a probability distribution should not reflect old frequencies but carry all information, and lack thereof, that we actually have concerning the unknown event. In our specific scenario, the information we assume to be in possession of will lead to a maximum entropy problem. In general, additional information should be processed through Bayes rule.

Before turning to the actual formalization of the problem, let us first examine a model scenario in mobile communications.

EXAMPLE 4.1 Bidding for quality-of-service

In a cellular mobile telecommunications network customers compete for access to the communications channel. In traditional networks, the users are all treated equally with respect to the number of times a user gets access. This however implies suboptimal resource utilization, and consequently if users pay a certain amount per transmitted bit, the network operator fails to maximize its revenues. Instead, the operator should transmit to the user with maximum capacity if all users pay the same amount per unit throughput, or, if users pay different prices, to the user with maximum price-capacity product. This corresponds to maximizing the revenues over a short time horizon.

In Section 5.6 we discuss such a scenario in which users are allowed to dy-

namically change the prices that they are willing to pay for transmission. In order to realize such a system, each user should report to the network its transmission capacity (the number of bits that can be transmitted over the channel at some desired bit-error or packet-error rate) for the coming time slot and the price that the user will pay per unit throughput.

The user can then bid according to its needs for transmission capacity. The price per unit throughput thus becomes lower for a user near the base station, and higher for a user with worse channel conditions. This may be construed as unfair, but consider then a user with severe channel conditions who gets access without having to compensate for that by paying a higher price. In an overloaded network, prioritizing a user with a bad average channel results in rejecting perhaps two or three users having better channels.

Thus, in order to be fair to one user with a bad channel, we find ourselves being unfair to several other users! At the same time, we are also loosing revenues. The same resources could have been awarded to these other users and thereby more than one paying customer could have been given satisfactory service.

We argue that the least unfair policy is the one where the winning user is the one that has the largest price-capacity product, but note that fairness is a somewhat elusive concept, which has not yet been given any satisfactory mathematical definition[1].

4.1 The Basic Reasoning of Bidding under Uncertainty

Consider a customer, Mr A, who desires access to a certain resource, the level of desire being characterized by a utility or a loss function $L(d, \theta)$. The loss function determines the loss suffered by the customer upon making decision d should θ turn out to be the true state of nature. In our problem, d is the bid q_A that Mr A makes, and θ is the throughput that he is awarded. Here, θ is either 0 or c_A, Mr A's carrying capacity associated with the resource.

Mr A should make a bid q_A so that the chance of winning the auction is such that his loss is minimized. Now, assume that he has no information concerning the

[1]There are a number of more or less *ad hoc* mathematical definitions of fairness, such as min-max fairness and proportional fairness (see e.g. Boudec, 2003), but there is no single one that is generally agreed upon. The entropy (relative to a 'fair distribution' describing the relative requirements of different users) of the instantaneous resource distribution or its average could be a reasonable measure of fairness, but it does not seem to help in forming constructive criteria for resource allocation decisions.

outcomes or winning bids of previous auctions, nor knowledge of other customers' utility functions or channel conditions. Clearly, at this stage Mr A is at a loss, and has too little information to be able to give any well-grounded bid. Depending on his loss function he would either bid very little, or make a very generous offer. The former case would correspond to Mr A being a man concerned about his expenses, whereas in the latter case Mr A's loss function would reflect a less constrained budget. In any case, the information at hand is insufficient for Mr A to feel comfortable that he has made a sound decision.

In terms of probability theory, Mr A's probability distribution for θ has too high entropy to confidently rule out any specific course of action. Mr A would be happy for any information that could reduce this entropy and single out a specific bid.

From the auctioneer's viewpoint, having uninformed customers is of no advantage. The customers have no way of obtaining a given service level with any degree of confidence, since there is no information to guide their decisions. Consequently, the auctioneer would soon be out of business.

Consider a more reasonable auctioneer, who after every nth auction announces the average winning price-capacity product for that period, $\mu_w = \langle q_w c_w \rangle$, and some measure of the variability of the same quantity, say the variance $\sigma_w^2 = \langle (q_w c_w - \langle q_w c_w \rangle)^2 \rangle$. With this information, Mr A, knowing his own carrying capacity c_A, can compare his price-capacity product $q_A c_A$ to the average winning one and, accounting for the variance, make a bid with some confidence of minimizing his loss.

4.2 The Bidding Policy

It is clear that Mr A should make the decision which minimizes his expected loss,

$$\langle L(q_A, c_A) \rangle = \int p(c_A|I) L(q_A, c_A) dc_A \tag{4.1}$$

where $p(c_A|I)$ represents the probability that Mr A receives the resource and thereby obtains c_A units of carrying capacity, conditioned on all information I available to Mr A in making the bid q_A.

The bid will depend just as much on the choice of loss function as on the prior information. The onus is therefore on Mr A to formulate a loss function which matches his values. Of course, different customers have different needs for the resource being put up for sale, and thus different users will in general reach different conclusions as to the best bid even though their information is equivalent. There is nothing irrational in this; on the contrary it reflects a great deal of rationality as it portrays the differing requirements of each user.

In the following section we present a number of loss functions reflecting different optimization objectives. We then derive the probability distribution $p(c_A|I)$ for the case where the auctioneer after each nth auction announces the average winning price-capacity product and its variance over the preceding n auctions.

4.2.1 Typical loss functions

Different customers may have different service demands. We here propose a number of loss functions that are intended to reflect typical requirements. The loss functions would moreover often be supplemented by a constraint on the maximum allowed bid.

Constant demand

A customer u wishing to obtain a certain amount ϕ_u of goods over the coming N time slots should use

$$L(q_u, x_u(q_u)) = |x_u(q_u) - \phi_u| \,, \tag{4.2}$$

where $x_u(q_u)$ is the actual amount of goods that the user will obtain for q_u monetary units.

Price-performance ratio

A customer u may wish to increase his bid if that bid would result in a significantly increased amount of delivered goods. In some sense, the price-performance ratio should be optimized. A possible formalization is the following: A price increase of 1 unit is acceptable given that the amount of goods obtained then increases by at least a factor a. Then the following loss function should be used.

$$L(q_u, x_u(q_u)) = \frac{a^{q_u}}{\max(x_u(q_u), b)} \,, \tag{4.3}$$

where $x_u(q_u)$ is the actual amount of goods that the customer will obtain for q_u monetary units. If $x_u(q_u) > b$ then an increased bid, $q_u \to q_u + 1$ will result in a lower loss if and only if $x_u(q_u + 1) > a x_u(q_u)$, because then we obtain

$$L(q_u + 1, x_u(q_u + 1)) = \frac{a^{q_u+1}}{x_u(q_u + 1)} < \frac{a^{q_n}}{x_u(q_u)} = L(q_u, x_u(q_u)) \,. \tag{4.4}$$

The formulation (4.3) also includes a minimum acceptable delivery size; if the user is to pay more than 0 monetary units per bit then the throughput must satisfy $x_u(q_u)/a^{q_u} > b$.

For example, if the customer requires at least an amount of 50 units per time slot, and if a price raise of 1 unit is acceptable only if the obtained goods then double, the loss function is $2^{q_u}/\max(x_u(q_u), 50)$.

4.2.2 The basic probability distribution

The following distribution is fundamental for the bidding problem, because it shows in general how to calculate the probability for obtaining a given service level. The procedure follows the same pattern for other states of knowledge as well.

Let the probability that a certain customer u will have the largest price-capacity product of all customers be denoted by $P(u \mid I)$. Then $P(u \mid I)$ is equal to the probability that the customer v with the largest price-capacity product of all other customers has a lower price-capacity product than customer u. Letting $y \equiv q_v c_v$ denote the largest price-capacity product among all customers except u, we can thus find the probability that u wins by marginalization: first determine the probability that $y < c_u q_u$ assuming knowledge of c_u, i.e. $\int_0^{c_u q_u} P(y \mid c_u I) dy$, then multiply this with the probability distribution for c_u given I to obtain the joint probability for c_u and $y < c_u q_u$, and integrate the result over all possible outcomes of c_u. In summary, we have

$$P(u \mid I) = \int P(c_u \mid I) \int_0^{c_u q_u} P(y \mid c_u I) dy dc_u \,. \tag{4.5}$$

In order to determine this probability distribution we must first find the probability distribution for c_u and that for y. We will consider a general case in which the carrying capacities c_u may be unknown in advance, as that is often the case in mobile communications. If the capacity is already known the solution simplifies straightforwardly.

Assume that there are K different possible capacities c_k. We suppose further that each customer stores the number of time slots that each capacity c_k could be used during a recent time window. If nothing else than these numbers are known the probability that the customer's carrying capacity will be c_k is then the expected frequency with which that capacity will be used. According to Laplace's rule of succession (see Section 2.6), the probability for having the carrying capacity c_k is

$$P(c_k \mid I) = \frac{n_k + 1}{N + K} \,, \tag{4.6}$$

where n_k is the number of time slots over the last N records that capacity c_k (but not higher) could be attained.

Now, the distribution $P(y \mid I)$ of the other customers' best price-capacity product depends heavily on the information I that customer u possesses. Several alternatives are possible. For instance, if the auctioneer does not give any information

about the most recent winning price-capacity products, then each customer has very vague information about the other customers. Based only on the observed number of time slots in which the customer has received goods, a resulting inference would be very uncertain.

A more reasonable approach would be for the auctioneer to periodically broadcast the expected winning price-capacity product for the coming period along with a measure of the prediction uncertainty. The simplest such scheme would consist of recording the average of the most recent winning price-capacity products and its variance. More advanced schemes include determining a model for the time evolution of price-capacity products. Here, we will assume that an expectation is available along with a variance for the prediction. These two quantities are broadcast to all users at regular intervals.

With no other knowledge than the mean and the variance of a variable, the least biased probability distribution according to the maximum entropy principle is Gaussian (see Section 2.7.1). Thus, we shall take

$$P(y \mid I) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left\{-\frac{1}{2\sigma_y^2}(y-\mu_y)^2\right\} , \tag{4.7}$$

with μ_y denoting the expectation of y, and σ_y^2 the variance of the distribution. Here, by not truncating the distribution at zero we have assumed that the variance of the distribution is not too large compared to the mean, so that the tail of the distribution below $y = 0$ is negligible. It should also be pointed out that we are told the mean and the variance of *all* winning price-capacity products, which includes those times when customer u won. However, we should actually determine a distribution for the winning price-capacity products of all customers *except* u. Below, we discuss how to adjust the mean and the variance to subtract out the contributions from customer u. However, it is not clear in general that this distribution, having excluded one of the components, should also be Gaussian. We have good reason to use a Gaussian distribution if there are many bidders with independently and symmetrically varying price-capacity products around some mean. Now, the bids are not logically independent since all customers base their decisions on partly the same information. However, the capacity variations will often, for instance in the mobile communications scenario described in Example 4.1, be independent among customers, which to some extent will have a 'randomizing' effect on the price-capacity products. Nonetheless, we may argue that a correlated distribution might be a better model. We will leave this alternative as a topic for future research, and here continue to work with the Gaussian model.

Inserting (4.6) and (4.7) into (4.5) (replacing the integral over c_u with a sum,

reflecting that c_u is discrete) we obtain

$$\begin{aligned} P(u \mid I) &= \sum_{k=1}^{K} \frac{n_k+1}{N+K} \\ &\times \int_{-\infty}^{q_u c_k} \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left\{-\frac{1}{2\sigma_y^2}(y-\mu_y)^2\right\} dy \\ &= \sum_{k=1}^{K} \frac{n_k+1}{N+K} \times \frac{1}{2}\text{erfc}\left(\frac{\mu_y - q_u c_k}{\sqrt{2}\sigma_y}\right) , \end{aligned} \tag{4.8}$$

where $\text{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt$ is the complementary error function (see Appendix A for the evaluation of the integral in (4.8)).

4.2.3 Making the decision – expectations and computations

The expected throughput $\langle x_u(q_u)\rangle$ per time slot as a function of the bid q_u is

$$\langle x_u(q_u)\rangle = \sum_{k=1}^{K} c_k \times \frac{n_k+1}{N+K} \times \frac{1}{2}\text{erfc}\left(\frac{\mu_y - q_u c_k}{\sqrt{2}\sigma_y}\right) . \tag{4.9}$$

Similarly, the expected loss using the loss function (4.2) is

$$\langle L(q_u)\rangle = \sum_{k=1}^{K} |c_k - \phi_u| \times \frac{n_k+1}{N+K} \times \frac{1}{2}\text{erfc}\left(\frac{\mu_y - q_u c_k}{\sqrt{2}\sigma_y}\right) . \tag{4.10}$$

The expected loss using (4.3) involves determining the expectation of $1/x_u$ for the Gaussian-distributed uncetainty of x_u, an expectation which is not available in closed form. We shall instead use the expected value of x_u directly in (4.3), thus obtaining a suboptimal solution that does not fully account for our actual uncertainty in making the bid. The estimated loss $\hat{L}(q_u)$ is then

$$\hat{L}(q_u) = \frac{a^{q_u}}{\max(\langle x_u(q_u)\rangle, b)} , \tag{4.11}$$

where $\langle x_u(q_u)\rangle$ is defined in (4.9).

It is important to see that in the preceding derivations, y is the winning price-capacity product of all customers *except* customer u. In calculating the best bid, a customer must therefore adjust the variance and the mean of the distribution for the best price-capacity product since these quantities are broadcast and based on all customers. These adjustments are quite difficult to carry out for a customer who

has been awarded all or almost all resources over the last period. Usually, however, we would expect that there are many different customers who obtain at least some goods, and then the following adjustments may be used.

The average μ_y is estimated from the broadcast value μ_w (the average of the winning bids) by

$$\mu_y = \frac{l\mu_w - q_u(t-1)x_u(t-1)}{l - l_u} \tag{4.12}$$

where l is the number of time slots between consecutive price updates, l_u is the number of time slots that customer u won, and $q_u(t-1)x_u(t-1)$ is the sum of customer u's price-capacity products for the l_u time slots that were won by customer u in the previous period of l slots.

Similarly, the variance is estimated by

$$\sigma_y^2 = \frac{l\sigma_w^2 - l_u\sigma_u^2(t-1)}{l - l_u} \tag{4.13}$$

where σ_u^2 is the sample variance for the price-capacity product of customer u in the slots that this customer won.

In order to compute the minimum of either of the two expected loss expressions (4.10) and (4.11) a numerical one-dimensional search is carried out using e.g. the Nelder-Mead simplex algorithm (Nelder and Mead, 1965).

4.3 Examples

We now consider the performance of the scheme outlined in this chapter based on simulations of the mobile communications scenario described in Example 4.1. Assume one transmitting base station and $U = 4$ users in the cell. With a periodicity of $n = 20$ time slots, each mobile user updates its bid and submits it to the base station. Each user is unaware of the other users' bids and the feedback channel is assumed to be error-free. An upper limit on the bid, $q_u \leq 5$ is also assumed. There are $K = 4$ different transmission rates, and each user determines and tells the base station the rate that can be used in the next time slot based on SNR measurements and bit-error rate requirements. The base station then transmits exclusively in each time slot to the user with the highest price-capacity product. All users have similar channel statistics, the unquantized rates being generated by independent Gaussian number generators. On average, 80 bits per time slot is supported, and the standard deviation is 20 bits. The rate is then quantized to the nearest level below the unquantized value. The quantized levels are determined from maximizing the expected system throughput for 4 users employing multiuser diversity as described

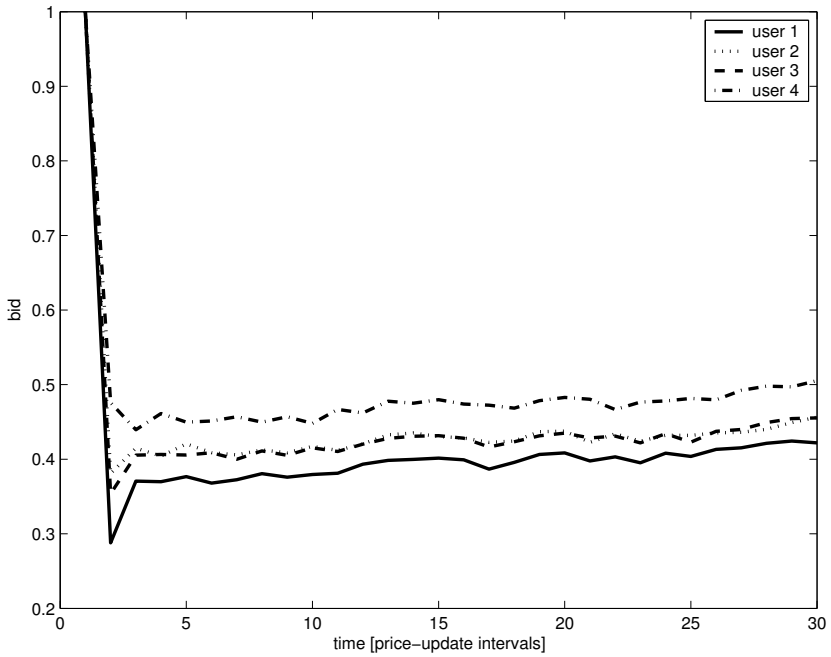

Figure 4.1: The evolution of the bids for the four users with desired rates 15, 20, 20 *and* 30 *respectively.*

in Chapter 6 with the result (in bits per time slot)

$$c_1 = 0 \quad c_2 = 74 \quad c_3 = 92 \quad c_4 = 106. \tag{4.14}$$

The rate probabilities (4.6) are updated continuously as more data becomes available.

4.3.1 Maintaining a desired throughput

We now consider a case where all four users have a desired rate per time slot according to

$$\phi_1 = 15 \quad \phi_2 = 20 \quad \phi_3 = 20 \quad \phi_4 = 30 \tag{4.15}$$

and attempt to minimize (4.10). Figures 4.1 and 4.2 show the resulting bids and obtained throughput per time slot from this test in a simulation lasting for 600 time slots (i.e. 30 price-update intervals). The plotted results are averages from 25 simulations.

It can be observed that there are quite substantial variations around the mean desired rate, but on average the obtained throughput matches the desired rate well.

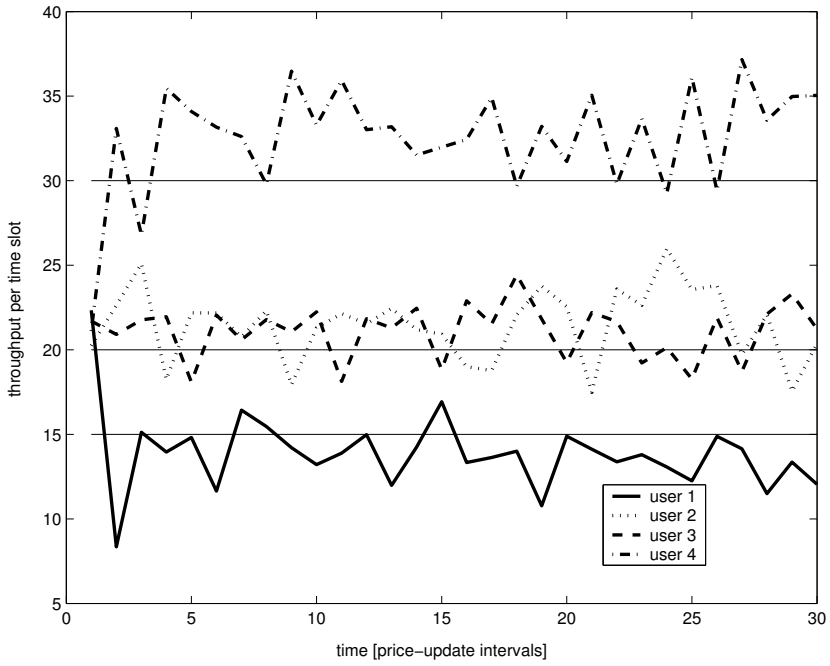

Figure 4.2: The obtained throughput per time slot for the four users with desired rates 15, 20, 20 *and* 30 *respectively.*

The average obtained rates over the entire simulated period were found to be

$$\bar{x}_1 = 14 \quad \bar{x}_2 = 21 \quad \bar{x}_3 = 21 \quad \bar{x}_4 = 33. \tag{4.16}$$

Under otherwise similar circumstances, Figures 4.3 and 4.4 show the bids and the obtained throughput when the desired rate of user 3 was increased to 25 bits per time slot, yielding a more competitive setting. Here, we see that the prices tend to increase because the users have trouble obtaining the desired quality of service. The average obtained throughput per time slot over the entire simulated period now becomes

$$\bar{x}_1 = 13 \quad \bar{x}_2 = 19 \quad \bar{x}_3 = 26 \quad \bar{x}_4 = 31. \tag{4.17}$$

4.3.2 Buying when the price is low and the performance high

In a similar setting as the previous one, we now let user 1 minimize the approximate expectation (4.11) of the price-performance-related loss

$$\frac{a^{q_u}}{\max(x_u(q_u), b)} \tag{4.18}$$

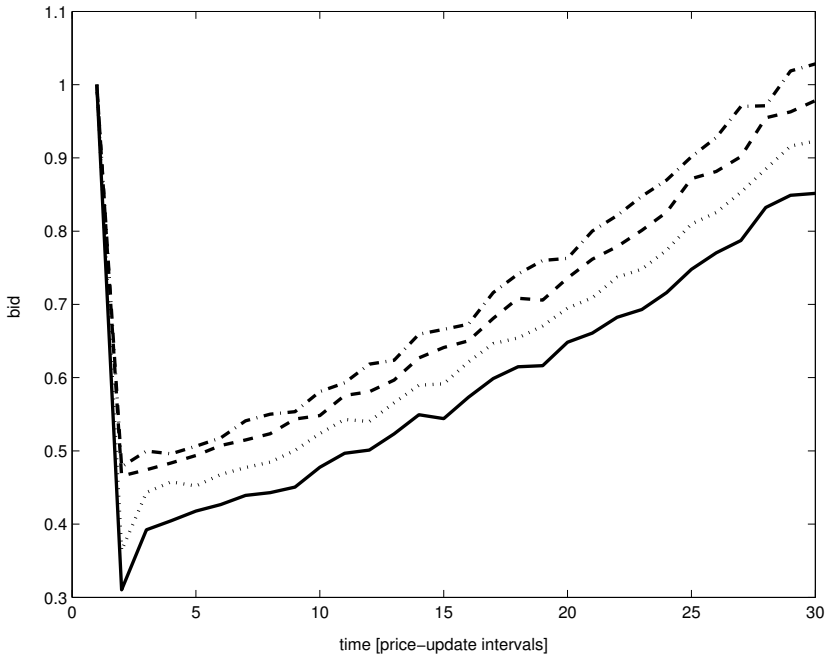

Figure 4.3: The evolution of the bids for the four users with desired rates 15, 20, 25 *and* 30 *respectively.*

with $a = 2$ and $b = 8$. Recall that use of this loss means that a 1-unit price increase is acceptable only if it leads to more than a doubling of the obtained throughput. Only if the throughput becomes more than $2^{q_u} \times 8$ bits is a non-zero bid q_u preferable. Users $2 - 4$ continue to minimize the expected loss (4.10) for a desired rate per time slot of

$$\phi_2 = 10 \quad \phi_3 = 20 \quad \phi_4 = 20. \tag{4.19}$$

In Figures 4.5, 4.6 and 4.7 the bids, obtained throughput and the price-to-obtained-throughput ratio (PTR) q_u/x_u are plotted as a function of time. The results are averages from running a simulation consisting of 1800 time slots 25 times. The average obtained throughput per time slot in this case becomes

$$\bar{x}_1 = 34 \quad \bar{x}_2 = 11 \quad \bar{x}_3 = 21 \quad \bar{x}_4 = 21, \tag{4.20}$$

where we see that users $2 - 4$ obtain rates corresponding well to their preferences. From Figure 4.7 we see that user 1 achieves the lowest PTR while the user with the lowest rate requirement has the worst PTR.

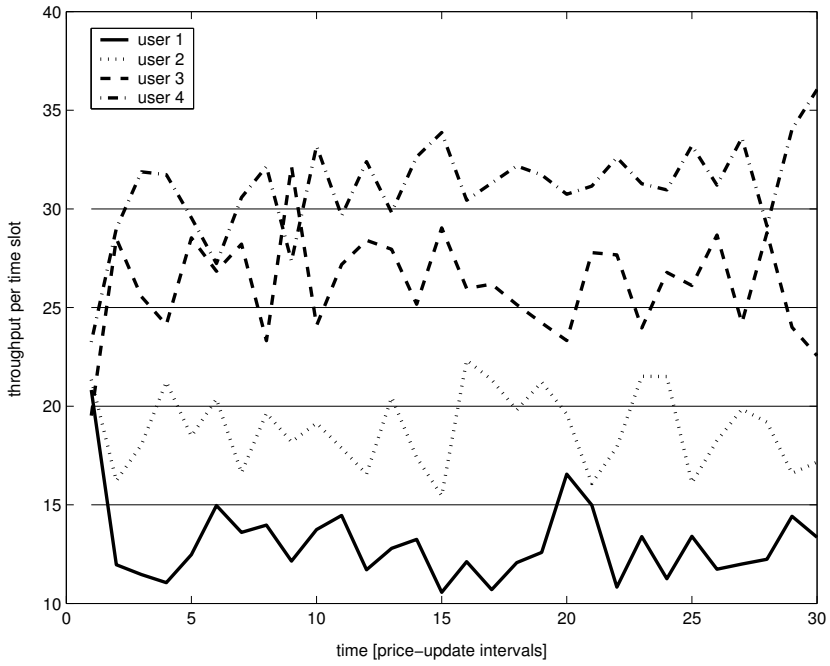

Figure 4.4: The obtained throughput per time slot for the four users with desired rates 15, 20, 25 *and* 30 *respectively.*

4.4 Comments

We have seen in this chapter how to make competitive bids in a repetitive auction with limited information. We considered specifically an auctioneer who after a set of auctions announces the average winning price-capacity product along with its sample variance for the preceding period. The considered auction format sells exclusively to the customer with the highest price-capacity product in order to maximize profits in a short perspective. We should however keep in mind that optimization over a short time period may be far from optimal in the long run. Taking other long-term effects, such as customer reactions to this type of procedure and its inherent unpredictability, into account is a vastly more difficult issue.

The performance examples show that the bidding strategies seem to perform well, but it should be noted that a full analysis of the behavior of the bidding policies is extremely complex and has not been carried out here. The individual bidder, in trying to make a reasonable bid in terms of his/her loss function, bases his/her decision on information which is different for different customers (because the estimates of the other users' best price-capacity products become different for different

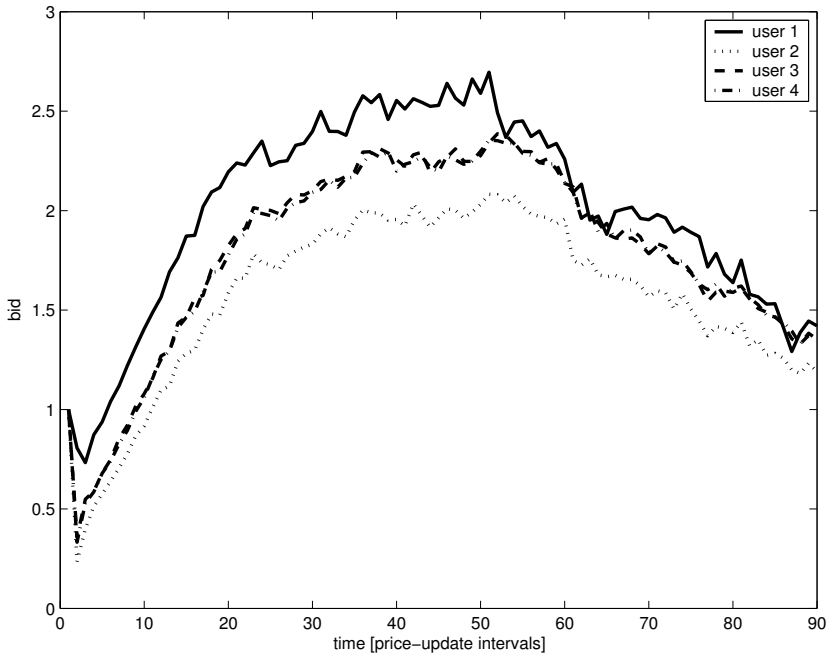

Figure 4.5: The evolution of the bids for the four users with user 1 *minimizing the price-performance-related estimated loss (4.11) and the other users employing (4.10) with desired rates* 10, 20 *and* 20 *respectively.*

users depending on the number of wins for that customer). Therefore, the behavior becomes very complex and hard to predict. A more general analysis must probably be based on some form of theoretical analysis rather than rely on simulations. This however is a quite complicated task and to the knowledge of this author there are no tools from the field of game theory that are immediately suited for analysis of this type of situation.

We however have reasonably strong confidence in that the individual policies put forward here will continue to work well also in other cases than the ones tested in the previous section. Our belief is founded on the desiderata of probability theory, which should convince us that if the information used in our policies is valid and adequate, the resulting inferences will indeed always be reasonable. Since our proposed policies are based on fundamental principles of optimal reasoning, our worries instead concern whether the information broadcast by the auctioneer is sufficiently informative, and whether the chosen loss functions actually represent what a customer desires. This is an easier problem which readily lends itself to analysis based on for the one part customer polls and for the other part com-

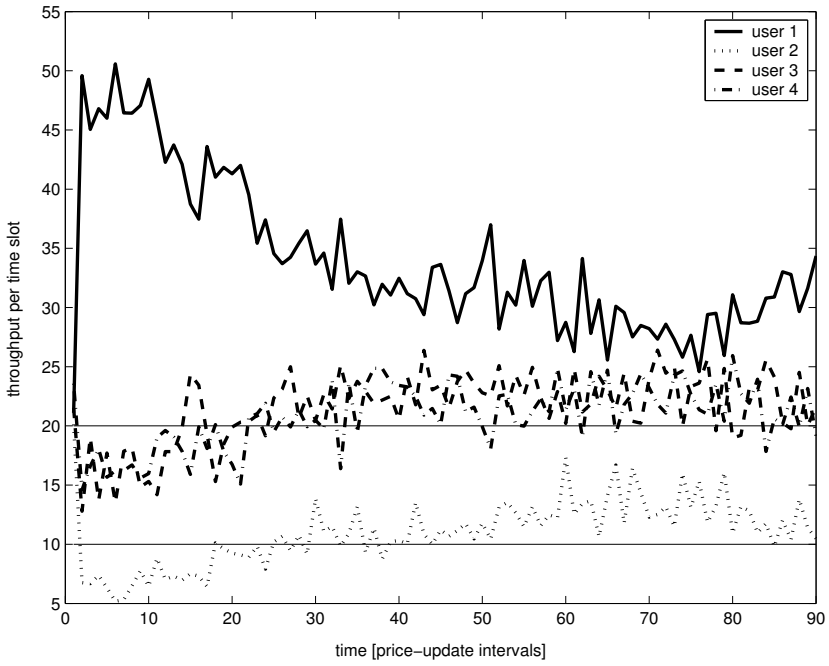

Figure 4.6: The obtained throughput per time slot for the four users with user 1 *minimizing the price-performance-related estimated loss (4.11) and the other users employing (4.10) with desired rates* 10, 20 *and* 20 *respectively.*

puter simulations such as those carried out in the previous section. We will come back in Chapter 5 to discussing the issue of whether to use competitive bidding in mobile communication networks makes sense from a technical and a commercial perspective.

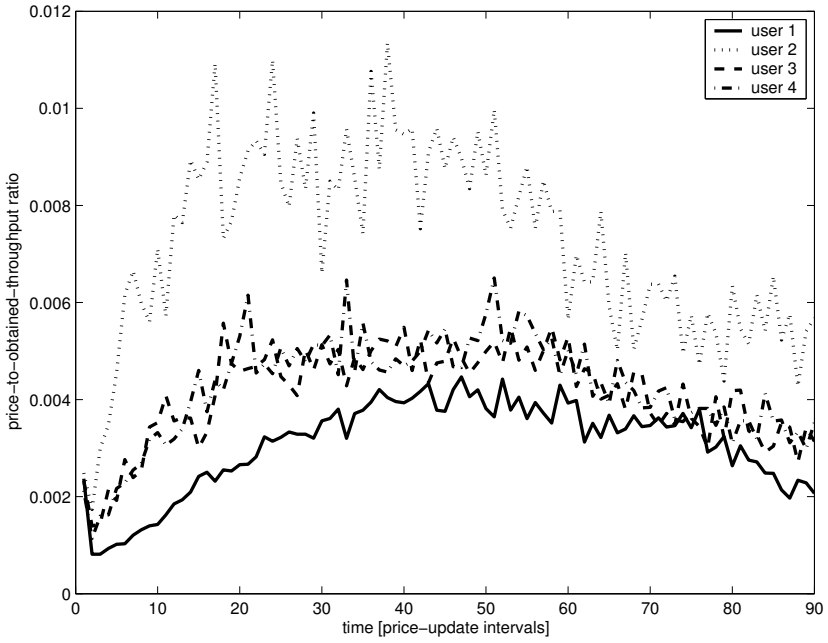

Figure 4.7: The evolution of the price-to-performance ratio (the bid divided by the obtained throughput) for the four users with user 1 *minimizing the price-performance-related expected loss (4.11) and the other users employing (4.10) with desired rates* 10, 20 *and* 20 *respectively.*

Chapter 5

Scheduling for Maximum Throughput under Uncertainty

IN this chapter we consider a problem of scheduling transmissions from a base station to a set of users in a cellular communications system. The problem consists of distributing bandwidth among users who share a number of channels. A number U of sources are producing bits at unknown rates. The bits from each source are to be transmitted to one of U users (or receivers). The sources share a number, R, of transmitters (or resources, or channels) which may be used to send the produced bits to the receivers.

The problem is a reformulation of the 'widget problem' studied in Chapter 3, with some differences due to the nature of communication links.

In our present problem each transmitter-receiver pair has a time-varying number associated with it, denoting the number of bits that can be sent over the link at a prescribed bit error rate (BER), given that the transmitter is used exclusively for transmitting to that specific receiver. We will henceforth denote this number as the *effective capacity*[1] of that link.

Bits produced by the sources are stored in buffers monitored by a transmission controller. The transmission controller aims to distribute the bits over the transmitters so that the number of bits in the buffers is minimized, or equivalently so that the system throughput is maximized. The question that we address is then:

[1]The term capacity is here used in a non-traditional way and should not be confused with any of the usual information theoretic capacity definitions. The effective capacity denotes the transmission rate for a given BER requirement that a user obtains if no other users transmit simultaneously on the channel. The actual transmission rate becomes less than that if the channel is shared among several users.

given only limited knowledge of the actual source rates and effective capacities, how should the controller distribute the resources?

The main information-theoretic motivation for using scheduling in mobile communications comes from the observation by Knopp and Humblet (1995) that the sum-of-rates capacity increases with the number of users and that it is maximized by transmitting exclusively to the user with highest signal-to-noise ratio (SNR) at the receiver. This phenomenon, denoted multiuser diversity by Knopp (1997), suggests that independent channel fluctuations between different users should be taken advantage of instead of being combatted. The result of Knopp and Humblet (1995) however assumes perfect channel knowledge, additive Gaussian disturbances only, and that transmission buffers cannot be emptied (there is always data to send).

Following the publication of Knopp and Humblet (1995), scheduling in wireless communications has received an increasing amount of attention, but the focus has been on assuming that there is always data to send (buffers are never emptied) and that the scheduler has perfect channel knowledge.

In high-level schedulers, stochastic channels are sometimes introduced by two-state models (error-free or random errors) (see e.g. Cao and Li, 2001), which might be considered too coarse. Casimiro Ericsson et al. (2000) suggest a framework for scheduling several time-slots ahead which takes known buffer sizes into account but requires perfect channel prediction (see also Casimiro Ericsson, 2001, for a more detailed account). Another rule, the proportional fair scheduler (Viswanath et al., 2002), gives exclusive access to the user who currently has the highest effective capacity normalized by its average allocated throughput, thus striking a balance between fairness and performance, but again requiring complete knowledge of the effective capacities. A similar result to that of Knopp and Humblet (1995) is obtained by Tse (1997) for a set of parallel broadcast channels corrupted only by additive white Gaussian noise. Another line of work (Tassiulas and Ephremides, 1991, 1992), which has been used for multi-hop networks and on-off types of links with constant effective capacity, considers queue stability as the main criterion. An interesting application of this criterion which also shows a relation to the proportional fair scheduler is reported by Andrews et al. (2001), where queue stabilizing schedulers are adapted to support quality-of-service (QoS) constraints.

Except for base station assignments in the uplink with the objective of minimizing allocated mobile powers (Rashid-Farrokhi et al., 1998) and a similar downlink problem (Bengtsson, 2001), little has been published concerning allocation of multiple shared transmitters. Scheduling transmissions under uncertain channel conditions and uncertain source rates with the objective of maximizing total throughput under quality-of-service constraints has hitherto not been investigated in any detail. The aim of this chapter is to provide such a study.

In summary, this work extends the current literature by providing means for

resource allocation with uncertain source rates (the traffic entering the buffers), taking buffer levels into account, and scheduling with multiple transmitters over arbitrary time periods. Furthermore, the scheduling framework is extended to take into account inaccurate channel predictions.

In two seminal papers, Jaynes (1957a, 1957b) introduced the maximum entropy principle as a consistent method for determining probability distributions under constraints on mean values of functions of data. The principle is applicable to inference problems with well-defined hypothesis spaces but incomplete data. We noted in Chapter 2 that the maximum entropy distribution can be realized in overwhelmingly more ways than any other distribution. It can thus be considered as the least biased solution for determining prior probabilities under the given constraints. It has been successfully applied to a variety of problems, the reference list providing a sample of examples from image reconstruction (Daniell and Gull, 1980, Gull and Daniell, 1978), spectrum estimation (Burg, 1975), finance (Buchen and Kelly, 1996), language modelling (Rosenfeld, 1996), and physics (Gruver et al., 1994, Jr., 1980). We here propose that the maximum entropy principle be used for modelling uncertain data flows in mobile communications systems.

The chapter is organized as follows: in Section 5.1 we present the problem formulation, whereas in Section 5.2 we explain how the maximum-entropy principle can be used to model the uncertain source flows. Following this, Section 5.3 presents the solutions for different states of knowledge concerning source rates and effective capacities. In Section 5.4 some observations are made concerning the behavior of the scheduler for different degrees of uncertainty. The performance is also compared to that obtained by the proportional fair scheduler from Viswanath et al. (2002). Before concluding the chapter, we discuss other approaches to scheduling in Section 5.5.

5.1 Distributing Bandwidth among Users Sharing a Set of Channels

The problem we shall investigate is how to allocate transmission resources with possibly uncertain effective capacities to sources with uncertain bit rates. A motivating application has been the problem of link-level predictive scheduling of a broadband downlink radio resource to mobile users with independently varying channel capacities due to fast fading (see e.g. Casimiro Ericsson, 2001, Wang et al., 2003a). Here we consider a slightly generalized problem.

In Figure 5.1 an overview of the system is given. There are U users, and equally many buffers. We will schedule the use of the channels for T time slots. During the scheduling horizon T, each buffer is filled with n_u bits, u denoting the

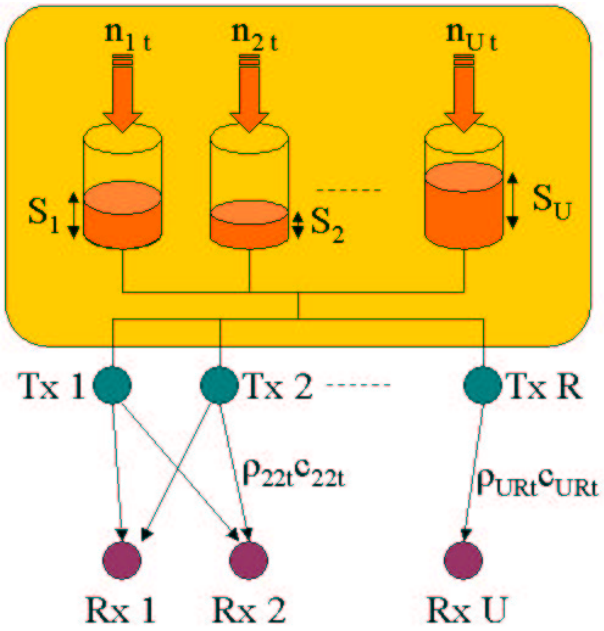

Figure 5.1: The system consists of U buffers, one for each receiver. R transmission resources are available and user u receives $\rho_{urt}c_{urt}$ bits at time t from transmitter r.

user index. A buffer may also have a number, S_u, of bits remaining in stock from previous scheduling rounds. The objective of interest will be to minimize the buffer contents at the end of the scheduled time horizon. In a situation where all influxes and effective capacities are known, this amounts to minimizing the loss function

$$L = \sum_{u=1}^{U} g\big(S_u + n_u - \sum_{t=1}^{T}\sum_{r=1}^{R} c_{urt}\rho_{urt}\big) \ , \tag{5.1}$$

where $g(x) = x$ if $x > 0$, else $g(x) = 0$. The time-varying effective capacity for the link between transmitter r and user u is denoted by the integer c_{urt}, while ρ_{urt} is the fraction ($0 \leq \rho_{urt} \leq 1$) of the bandwidth of the rth transmitter that we allocate to user u at time t. For instance, if $\rho_{urt} = 1$, then user u uses the rth channel exclusively at time t. The total channel usage $\sum_u \rho_{urt}$ for a given channel r at a time t must satisfy $\sum_u \rho_{urt} \leq 1$. The minimization of (5.1) is performed by adjusting ρ_{urt} under whatever constraints the specific system poses on ρ_{urt}.

The total number of incoming bits, n_u, in the time interval T is the sum of the

Table 5.1: Definitions of the main variables in this chapter.

L	The loss function, representing the sum of all users' buffer contents after T time slots, each term weighed by a user-specific cost per bit $\pi(u, \{\theta_u\})$
U	The number of users
R	The number of transmitters
T	The number of time slots a resource allocation is optimized over
S_u	The number of bits in stock for user u
n_u	The influx into user us transmission buffer summed over the T time slots
c_{urt}	The effective capacity (the number of bits that the channel supports at some desired BER) for the channel from transmitter r to user u at time slot t
ρ_{urt}	The fraction ($0 \leq \rho_{urt} \leq 1$) of the bandwidth that is used to transmit from transmitter r to user u at time slot t. Adjusted so that $\langle L \rangle$ is minimized
x_u	The total amount of data sent to user u over the T time slots ($x_u = \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt}\rho_{urt}$)
$\pi(u, \{\theta_u\})$	The cost per each bit intended for user u remaining in stock after the scheduled horizon
$\{\theta_u\}$	A set of known parameters that determine the user-specific cost per remaining bit

influxes at each time slot t:

$$n_u = \sum_{t=1}^{T} n_{ut} \quad . \tag{5.2}$$

In cases where we have knowledge of time variations, we will use this more detailed notation. In general, as a notational convention, for any quantity a, we will use at most three indices: a_{urt}, where u ($1 \leq u \leq U$) denotes user index, r ($1 \leq r \leq R$) transmitter resource index, and t ($1 \leq t \leq T$) time slot index. In this chapter, whenever any of these three indices are omitted the quantity represents the *sum over all values* of the omitted index. For reference throughout the chapter, Table 5.1 provides a list of definitions of the main variables that we use in this chapter.

In general, complete knowledge of the effective capacities or the number of incoming bits at any specific future time is unavailable. Therefore we cannot directly minimize L but must resort to assigning probability densities for the influx n_u and the effective capacities c_{urt} and minimize the expected loss. Assuming that knowledge of effective channel capacities gives no information of incoming bit rates [2],

[2] Although certain communication protocols actually change their transmission rates due to channel variations, these protocols, eg. TCP (Transmission Control Protocol), react on slower time scales than would normally be used in scheduling decisions at the link layer.

and vice versa, we can factor the joint probability[3]

$$\begin{aligned} P(n_u c_{urt}|I) &= P(n_u|c_{urt}, I)P(c_{urt}|I) = \\ &= P(n_u|I)P(c_{urt}|I) \end{aligned} \tag{5.3}$$

and the expected loss becomes

$$\langle L \rangle = \sum_{u=1}^{U} \sum_{c_{urt}=0}^{\infty} \sum_{n_u=0}^{\infty} P(n_u|I)P(c_{urt}|I) g\left(S_u + n_u - \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt}\rho_{urt} \right) \ . \tag{5.4}$$

Throughout the rest of the chapter we will find it convenient to use the notation $\langle L_u \rangle$ for the expected loss contribution corresponding to user u, with the total expected loss being the sum of all user contributions:

$$\langle L \rangle = \sum_{u=1}^{U} \langle L_u \rangle \ . \tag{5.5}$$

The scheduling framework we propose relies on minimizing (5.4) subject to various constraints. The rest of the chapter is concerned with investigating the expected loss contributions $\langle L_u \rangle$ for a few typical cases in mobile communications and the consequences of using buffer level minimization as a scheduling criterion. It should be emphasized that the cases differ only in what knowledge the scheduler uses.

Finding the minimum of (5.4) will in general turn out to require non-linear programming. The basic constraints on ρ_{urt} are:

$$\textstyle\sum_u \rho_{urt} \leq 1 \quad \forall r, t \tag{5.6}$$

$$0 \leq \rho_{urt} \leq 1 \quad \forall u, r, t \ , \tag{5.7}$$

but in general we may have an additional number of matrix equalities and inequalities representing constraints imposed by the specific system architecture on different resources. Examples of such constraints include

- a limited set Ω of rate levels, implying that the transmission rate $\rho_{urt}c_{urt}$ must belong to the set Ω,
- in a time division system, ρ_{urt} can only be 0 or 1,

[3]To indicate that the probability expressions will change according to the information at hand, all probabilities are conditioned on I, which denotes any available information relevant for inferring n_u or c_{urt}.

- some channels may not be accessible to all users, i.e. $\exists r, \exists u$, $\rho_{urt} = 0$,

- in a network guaranteeing some minimum level of service quality, constraints may take the form of user-specific minimum channel access levels, $\rho_{urt} \geq \eta_{urt}$, or minimum transmission rates $\sum_r \sum_t \rho_{urt} c_{urt} \geq \varphi_u$.

These types of constraints are readily treated by available software for solving non-linear programming problems and present no conceptual difficulties. The general problem can thus be transformed to different specialized settings, all represented by the same average loss function but with different optima due to the restrictions on ρ_{urt}.

Minimizing the number of bits remaining in stock is equivalent to maximizing the sum of the users' bit rates. With this criterion, user specific priorities can be introduced as multipliers to each user's loss contribution in (5.5). This can be interpreted as a user-specific cost per bit, expressed as a function $\pi(u, \{\theta_u\})$ of any set $\{\theta_u\}$ of known parameters (such as time, delay, buffer levels, average effective capacities, average influxes, bit prices, etc.). The generalized criterion is then to minimize

$$\langle L \rangle = \sum_{u=1}^{U} \pi(u, \{\theta_u\}) \langle L_u \rangle \ . \tag{5.8}$$

For instance, if $\pi(u, \{\theta_u\})$ is defined as the reciprocal of user u's average throughput and we restrict ourselves to exclusive allocations, then we obtain a generalized version of the proportional fair scheduler (Viswanath et al., 2002). We will not consider fairness in any detail in this chapter; note that any fairness requirement or user priority that can be formulated as a deterministic function describing an equivalent user-specific cost per bit is compatible with the given formulation. In Chapter 6 we will come back to the issue of fairness in connection with a study of how limitations in channel feedback affects the performance of systems using multiuser diversity.

Another possible approach could be to minimize the sum of the squared buffer contents in order to prioritize large buffers and consequently aim at reducing the risk of buffer overflow. A disadvantage of using a quadratic criterion here is that the scheduler would no longer maximize the sum of the users' bit rates, hence capacity would be wasted. Another problem is that if priorities are introduced as multiplicative factors for each user's contribution to the total loss, the priorities will loose their intuitive meaning as incurring a certain cost per bit to the network.

In the next sections we derive the expected loss contribution for each user u, $\langle L_u \rangle$ for different states of prior information by the use of the maximum entropy principle. Solutions are given for the following cases:

- Section 5.3.1 assumes knowledge of *average* source rates and *exactly* known capacities.
- In Section 5.3.2 we relax the requirement of perfect channel knowledge and instead assume *capacity predictions of varying accuracy*.
- In Section 5.3.3 source flows are subdivided into packets and the scheduler requires knowledge of the *average number of packets produced for each packet size* and the exact effective capacities.
- Finally, Section 5.3.4 provides a solution which takes account of arrival rate patterns by the use of Laplace's rule of succession applied on logarithmically spaced intervals. Perfect channel knowledge is assumed.

5.2 The Maximum Entropy Approach to Source Flow Modelling

The source flows in the current problem are not assumed to be known in detail. A common assumption concerning near-future networks is that traffic to a large extent will consist of Internet flows. Modelling an individual Internet data source is however a notoriously difficult problem (see the discussion in Floyd and Paxson, 2001). Various distributions have been proposed, the most commonly used consists of assuming that the number of packets per time unit is Poisson distributed. This distribution has some justification when the incoming packet streams stem from a large number of independent sources, but not in the case of a single-user source flow. Another approach would be to record individual histograms for each user in the transmitter and use them as approximate probability distributions. That is however not realistic; the amount of data that has to be collected would typically be larger than that obtainable during a user's connection. A possible way around this problem is however briefly investigated in Section 5.3.4.

Instead, we propose to use the maximum entropy approach. We shall use the maximum entropy principle to model the source rates n_u subject to knowledge of the average source rate $\langle n_u \rangle$ for each user[4]. Using the results from Chapter 3 we thus obtain

$$P(n_u|I) = \frac{1}{\langle n_u \rangle + 1} \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{n_u} \tag{5.9}$$

as the distribution of highest entropy subject to knowledge of $\langle n_u \rangle$.

[4]The average source rate can be determined at the transmitter based on the incoming data. An initial estimate can be obtained by using the average of all users' data streams. With only a short data record, the expectation (3.69) conditioned on the data record should be used in (5.9).

Note that, as mentioned in Chapter 3, the distribution would be different if n_u had a known upper bound. For the case of data flows, there is an upper bound which is determined by the bandwidth of the fixed network preceding the buffers. This limit is however neglected here because it is usually much larger than the expected source flow of each user.

5.3 Expected Loss Expressions for the General Resource Allocation Problem

5.3.1 Knowledge of average source rates and exact capacities

Here we will work out the expected loss contribution of user u, $\langle L_u \rangle$ (cf. (5.5)), for the scheduling problem when the average number of incoming bits during the interval T, $\langle n_u \rangle$, in each buffer is known and the effective capacities c_{urt} of the transmitters are exactly known. For clarity, we use

$$x_u = \sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt} \rho_{urt} \ , \tag{5.10}$$

describing the total number of bits sent from buffer u over the scheduled time horizon T. With $P(n_u|I)$ given by (5.9) the expected loss contribution with known c_{urt} becomes:

$$\langle L_u \rangle \ = \sum_{n_u=0}^{\infty} P(n_u|I) g(S_u + n_u - x_u) \tag{5.11}$$

$$= \ \begin{cases} \langle n_u \rangle \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{x_u - S_u} & , \ x_u > S_u \\ S_u + \langle n_u \rangle - x_u & , \ x_u \leq S_u \ . \end{cases} \tag{5.12}$$

The summation over n_u in (5.11) is equivalent to the derivation carried out in Appendix 3.A.

In certain problems the expected values of the influxes at time t defined in (5.2), n_{ut}, may vary over time, i.e. we have knowledge of $\langle n_{ut} \rangle$ for specified times t. For instance, a traffic predictor may be employed which gives the expectation of the traffic flow at different times. In Appendix 5.A the solution for this case is derived. The resulting loss contribution for time-varying expectations of incoming bit rates is:

$$\langle L_u \rangle \ = \begin{cases} \langle n_{u1} \rangle \left(\frac{\langle n_{u1} \rangle}{\langle n_{u1} \rangle + 1} \right)^{x_u - S_u} \times \\ \quad \times \prod_{k=2}^{T} \frac{1}{\langle n_{uk} \rangle + 1} \times \frac{1}{1 - \frac{\langle n_{uk} \rangle}{\langle n_{uk} \rangle + 1} \frac{\langle n_{u1} \rangle + 1}{\langle n_{u1} \rangle}} & , \ x_u > S_u \\ S_u + \langle n_u \rangle - x_u & , \ x_u \leq S_u \ , \end{cases} \tag{5.13}$$

where, for the case $x_u > S_u$, the averages are no longer ordered chronologically, but have been reordered by decreasing size, with the index k, to ensure convergence of the geometrical series. Notice also that the product over all averages which are smaller than $\langle n_{u1} \rangle$ does not depend on x_u, and consequently not on ρ_{urt}. Therefore, if the minimum loss is calculated iteratively, the constant factor need not be recalculated at each iteration.

5.3.2 Knowledge of average source rates and accuracy of capacity predictions

In this section we turn to a case which is of particular interest in applications for mobile communications. Here, a transmitter may predict future channel conditions with some known accuracy based on measured fading patterns at the receivers (see e.g. Ekman, 2002, Ekman et al., 2002). Adaptive modulation is then used to adjust the transmission rate based on the predicted channel quality.

We must now consider three different effective capacities: the *predicted* one $\hat{c}_{urt}$, the *potential* one $\bar{c}_{urt}$, and the eventually *obtained* one c_{urt}. The potential effective capacity $\bar{c}_{urt}$ is the number of bits that could be sent over the channel at time t with a prescribed error rate if we knew the channel and thus could choose the optimal modulation level. With inaccurate channel knowledge however, if the potential effective capacity is lower than predicted, then the modulation level may be set too high leading to a performance degradation due to increasing bit error rates. If on the other hand the predicted capacity is lower than the potential capacity, then the modulation level is set lower than the optimum and the *obtained* effective capacity will equal the predicted capacity (i.e. the obtained capacity will again be lower than the *potential* capacity). Thus, the probability for the outcome of the prediction (in the sense of being larger than, smaller than, or equal to the potential capacity) will determine the probability for obtaining a given effective capacity.

We assume that the accuracy of prediction is represented by a known variance, σ^2_{urt}, and that the prediction itself $\hat{c}_{urt}$ is the expected value of the potential (but unknown) effective capacity, $\bar{c}_{urt}$. As an example of how the prediction can be obtained, Ekman (2002), Ekman et al. (2002) derive a quadratic channel power predictor, based on which it is possible to derive a pdf for the channel power (Ekman, 2002, ch. 7-8). Using that pdf one can determine the corresponding pdf for the effective capacity given a certain BER requirement by a change of variables. This can for instance be carried out by using the approximate BER expressions from Chung and Goldsmith (2001). We will however not use this particular pdf as it is would only be valid for that predictor. Using instead the predicted value and the standard deviation for the predictive pdf, we have a more general model,

although of slightly higher entropy (thus disregarding some information).

In the case of a nonnegative integer quantity such as the potential effective capacity, finding the maximum-entropy distribution for known expectation and variance is analytically intractable. However, it is well-known (Shannon, 1948) that the Gaussian distribution has the highest entropy for a given mean and variance if the quantity of interest is defined over the entire real axis. If the expectation of a Gaussian distribution is positive and large compared to its standard deviation, then it has negligible probability mass for negative numbers. Therefore, for reasonably accurate predictions of $\bar{c}_{urt}$ we may safely assign a Gaussian distribution as an accurate description of our state of knowledge.

However, as mentioned, the obtained capacity depends on the prediction error $\hat{c}_{urt} - \bar{c}_{urt}$. There are three possible cases:

1. $\underline{\hat{c}_{urt} \leq \bar{c}_{urt}}$. In this case the obtained effective capacity will equal the predicted one, $c_{urt} = \hat{c}_{urt}$.

2. $\underline{\bar{c}_{urt} \leq \hat{c}_{urt} \leq c^*_{urt}}$. If the predicted value is higher than the potential effective capacity, then the modulation level will be set too high and thus the obtained effective capacity will decrease. Here, c_{urt} is given by a function $f(\hat{c}_{urt})$ which depends on coding and other system-specific parameters. A reasonable approximation is to assume that the obtained effective capacity decreases linearly with the predicted value, reaching zero at a point $c^*_{urt} = v\bar{c}_{urt}$. We comment further on this model choice and the determination of v in the end of this section.

3. $\underline{\hat{c}_{urt} \geq c^*_{urt}}$. In this interval, the obtained capacity is zero.

In summary we obtain an effective capacity curve as described by Figure 5.2.

In Appendix 5.B the probability for the obtained effective capacity c_{urt} given the predicted value is derived as the sum of the contributions from each of the three cases. It is shown that the probability for the obtained capacity is

$$P(c_{urt}|I) = P_1(c_{urt}|I) + P_2(c_{urt}|I) + P_3(c_{urt}|I) \tag{5.14}$$

where

$$P_1(c_{urt}|I) = \frac{1}{2}\delta(c_{urt} - \hat{c}_{urt}) \tag{5.15}$$

$$P_2(c_{urt}|I) = \frac{v-1}{\sqrt{2\pi}\sigma_{urt}v}\exp\left[-\left(\frac{v-1}{\sqrt{2}\sigma_{urt}v}\right)^2(c_{urt}-\hat{c}_{urt})^2\right] \times (H(c_{urt}) - H(c_{urt} - \hat{c}_{urt})) \tag{5.16}$$

$$P_3(c_{urt}|I) = \delta(c_{urt})\left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{(v-1)\hat{c}_{urt}}{v\sigma_{urt}\sqrt{2}}\right)\right) \tag{5.17}$$

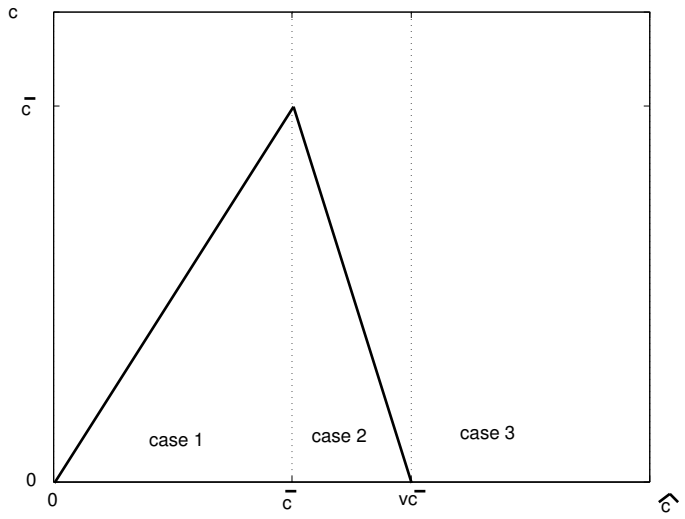

Figure 5.2: The obtained capacity as a function of the predicted capacity with linear decline for too large predictions.

where $H(x)$ denotes the Heaviside step function and $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. The probability distribution (5.14) for the obtained capacity is plotted for $\hat{c}_{urt} = 40$ and for different values of σ_{urt} and v in Figure 5.3.

We will now calculate each user's contribution $\langle L_u \rangle$ to the expected loss (5.4) with respect to $P(n_u|I)$ and $P(c_{urt}|I)$. The two probability distributions are logically independent, and hence we average the results obtained in the last section over c_{urt}. The expected loss contribution will consist of a sum of two components, one for $x_u > S_u$ and another for $x_u \leq S_u$, weighted by their respective probabilities $P(x_u > S_u|I)$ and $1 - P(x_u > S_u|I)$:

$$\langle L_u \rangle = P(x_u > S_u|I)\langle L_{u1} \rangle + (1 - P(x_u > S_u|I))\langle L_{u2} \rangle \ . \tag{5.18}$$

It is however reasonable to assume that $P(x_u > S_u|I)$ is approximately 1 or 0, eg. when the standard deviation for the prediction is not extremely large. Hence we use the simpler rule

$$\langle L_u \rangle \approx \begin{cases} \langle L_{u1} \rangle & , \langle x_u \rangle > S_u \\ \langle L_{u2} \rangle & , \langle x_u \rangle \leq S_u \ , \end{cases} \tag{5.19}$$

where $\langle L_{u1} \rangle$ and $\langle L_{u2} \rangle$ are derived below with the results (5.31) and (5.32), and

$$\langle x_u \rangle = \sum_{r=1}^{R} \sum_{t=1}^{T} \rho_{urt} \langle c_{urt} \rangle \tag{5.20}$$

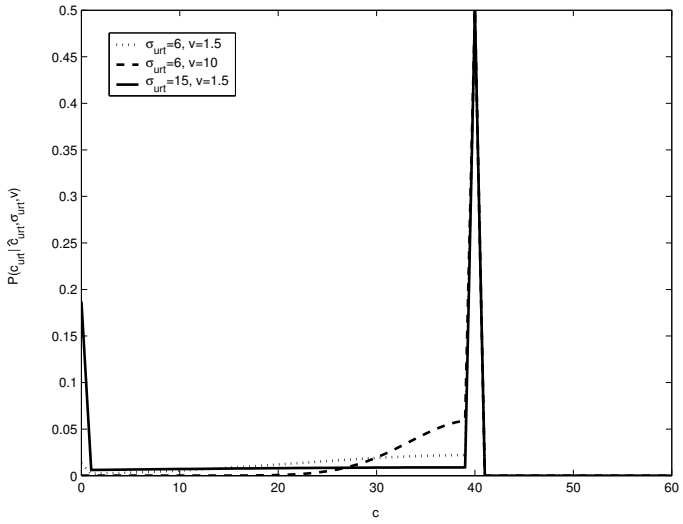

Figure 5.3: The probability distribution for the obtained capacity c_{urt} given the prediction $\hat{c}_{urt} = 40$. The spike at zero corresponds to setting the transmission rate too high, leading to unacceptable bit-error rates; the spike at $c_{urt} = 40$ corresponds to the predicted capacity being less than the potential one (the Gaussian distribution is symmetric, giving probability $1/2$ for this event); the intermediate range covers the case when the predicted capacity is higher than the potential one, increasing bit-errors but not so much as to render the data completely useless.

where, inserting (5.15), (5.16) and (5.17) into (5.14),

$$\langle c_{urt} \rangle = \int_0^{\hat{c}_{urt}} c_{urt} P(c_{urt}|I) dc_{urt} \tag{5.21}$$

$$= \frac{1}{2}\Bigg\{ \hat{c}_{urt} + \hat{c}_{urt}\text{erf}\left(\alpha_{urt}\hat{c}_{urt}\right) + \frac{1}{\sqrt{\pi}\alpha_{urt}}\left[\exp\left(-\alpha_{urt}^2\hat{c}_{urt}^2\right) - 1\right]\Bigg\} \tag{5.22}$$

with

$$\alpha_{urt} = \frac{v-1}{\sqrt{2}\sigma_{urt}v} \quad . \tag{5.23}$$

The integral in (5.21) consists of three additive terms; the Dirac contributions (5.15), (5.17) at zero and $\hat{c}_{urt}$, respectively, simply extracts the loss at those values; the integral of the truncated Gaussian part (5.16) follows from the procedure in Appendix A. Adding them together yields the second equality in (5.22).

Consider the calculation of $\langle L_{u1} \rangle$ which is the expectation with respect to $P(c_{urt}|I)$ of the corresponding case in (5.12). To distinguish between the expected loss with respect to $P(n_u|I)$ from (5.12) and the one currently under investigation we here assign the notation $\langle L_{u1} \rangle_{P(n_u|I)}$ for the former one.

We rewrite the expression for $x_u > S_u$ in (5.12) using the algebraic relation $x^{a+b} = x^a x^b$, and obtain

$$\begin{aligned} \langle L_{u1} \rangle_{P(n_u|I)} &= \langle n_u \rangle \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{\sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt} \rho_{urt} - S_u} \\ &= \langle n_u \rangle \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{-S_u} \prod_{t=1}^{T} \prod_{r=1}^{R} \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{c_{urt} \rho_{urt}} . \end{aligned}$$

Averaging over $P(c_{urt}|I)$ gives the expected loss contribution with respect to both $P(n_u|I)$ and $P(c_{urt}|I)$:

$$\begin{aligned} \langle L_{u1} \rangle &= \langle n_u \rangle \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{-S_u} \\ &\times \prod_{t=1}^{T} \prod_{r=1}^{R} \int_{-\infty}^{\infty} P(c_{urt}|I) \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{c_{urt} \rho_{urt}} dc_{urt} . \end{aligned} \quad (5.24)$$

Inserting (5.14) into (5.24), the integral over c_{urt} contains three mutually exclusive intervals. We label the corresponding integrals I_1, I_2, and I_3. The first integral I_1 corresponding to the point $c_{urt} = \hat{c}_{urt}$ is simply

$$I_1 = \frac{1}{2} \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{\hat{c}_{urt} \rho_{urt}} . \quad (5.25)$$

The second integral, I_2, ranges from 0 to $\hat{c}_{urt}$. Using (5.16) and following the procedure in Appendix A we obtain

$$I_2 = \int_0^{\hat{c}_{urt}} P_2(c_{urt}|I) \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right)^{c_{urt} \rho_{urt}} dc_{urt} \quad (5.26)$$

$$\begin{aligned} &= \frac{1}{2} \exp \left(\rho_{urt} \hat{c}_{urt} \log \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right) + \rho_{urt}^2 \gamma_{urt}^2 \right) \times \\ &\times \left(\text{erf} \left(\frac{(v-1)\hat{c}_{urt}}{v \sigma_{urt} \sqrt{2}} + \rho_{urt} \gamma_{urt} \right) - \text{erf} \left(\rho_{urt} \gamma_{urt} \right) \right) , \end{aligned} \quad (5.27)$$

where

$$\gamma_{urt} = \frac{\sigma_{urt} v}{(v-1)\sqrt{2}} \log \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1} \right) . \quad (5.28)$$

Finally, the third integral, I_3, represents the single point $c_{urt} = 0$ and using (5.17) we have

$$I_3 = \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{0\rho_{urt}} \left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{(v-1)\hat{c}_{urt}}{v\sigma_{urt}\sqrt{2}}\right)\right) \tag{5.29}$$

$$= \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{(v-1)\hat{c}_{urt}}{v\sigma_{urt}\sqrt{2}}\right) . \tag{5.30}$$

Using I_1 from (5.25), I_2 from (5.27), and I_3 from (5.30) in (5.24) the expected loss contribution of user u with predicted capacities is, if $P(x_u > S_u \mid I) = 1$,

$$\langle L_{u1} \rangle = \langle n_u \rangle \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{-S_u} \prod_{t=1}^{T} \prod_{r=1}^{R} (I_1 + I_2 + I_3) . \tag{5.31}$$

The second case in the expected loss contribution from user u (5.19) assuming that $P(x_u \leq S_u \mid I) = 1$ is, using (5.12) for $x_u \leq S_u$ and the definitions of x_u (5.10) and $\langle x_u \rangle$ (5.20),

$$\langle L_{u2} \rangle = \int P(c_{urt}|I)(S_u + \langle n_u \rangle - x_u) dc_{urt} = S_u + \langle n_u \rangle - \langle x_u \rangle \tag{5.32}$$

The loss contributions in (5.31) and (5.32) are valid when predicted capacities can be modelled by a Gaussian distribution with known variance and expected value $\langle \hat{c}_{urt} \rangle = \bar{c}_{urt}$. They also require that the obtained capacity decreases linearly when the predicted capacity $\hat{c}_{urt}$ is larger than the potential capacity $\bar{c}_{urt}$. It should however be emphasized that the linear decrease and the actual choice of v is a subjective choice, and not a property of the channel. The value of v depends on how sensitive the application is to departures from the desired BER. For low BER requirements, even a small prediction error leads to a substantial departure from the desired BER. For example, with Gray-coded M-QAM modulation[5], increasing from 4 bits to 5 bits per symbol at an SNR of 20 dB increases the BER by a factor of more than 200. Typically, in order to determine v we find the BER increase which means that the data must be retransmitted. We then determine the corresponding rate increase that would cause this BER discrepancy. If for instance M-QAM is used with a desired BER of 10^{-4}, and if a BER increase of a factor 100 would require that the data be retransmitted, then it can be found that $v \approx 1.5$ will be a good model. If a BER increase of a factor 10 would require retransmission, then $v \approx 1.2$. Typical values of v are thus in the range $1 < v < 2$. The linear decrease in c_{urt} for predictions larger than the potential capacity can be questioned, but clearly it satisfies the obvious requirement that the curve should be monotonic

[5] Approximate BER formulas from Proakis (1995) are used in these calculations.

decreasing. Other alternatives would be to use either some concave of some convex decreasing function, but that could hardly make any substantial difference for the actual expected loss value unless the magnitude of the function's derivative would be very nearly zero for one interval and large for the remaining part. These cases will not be considered here, as they would rarely be encountered in practice.

The final expression for $\langle x_u \rangle > S_u$ (5.31) is rather complex and in the simulations of Section 5.4.4 we investigate whether the basic scheduler assuming perfect channel knowledge can be used with predicted values as an alternative to the more computationally burdensome minimization of (5.19).

5.3.3 Knowledge of average rates for each packet size

We now consider the case where the possible sizes of incoming packets of each size are known to the scheduler. If the number of possible packet sizes is small, then having knowledge of the possible sizes significantly reduces the possible influx sizes and thus we should be able to make better scheduling decisions. We further assume that the expected number of incoming packets of each size in the time interval T is known to the scheduler. Similarly, the effective capacities c_{urt} are also assumed known.

Let the packet sizes in the uth buffer, cf. Figure 5.1, belong to a set $\{k_u\}$ with K_u elements. Let m_{uk} be the number of packets of size k which are received in the uth buffer during the scheduling horizon T, with $\langle m_{uk} \rangle$ assumed known. In order to find a closed-form expression for the expected loss, we make a logic partitioning of each buffer u into K_u buffers. Hence, each user's buffer is split so that each packet size gets its own buffer. The remaining number of bits from the previous round, S_u, is also split into K_u partitions $S_u = \sum_k kS_{uk}$. Note however that this is only a logical separation for mathematical convenience.

Our new loss function is

$$L = \sum_{u=1}^{U} \sum_{k \in \{k_u\}} g\Big(km_{uk} + kS_{uk} - \frac{\sum_{t=1}^{T} \sum_{r=1}^{R} c_{urt}\rho_{urt}}{K_u} \Big), \tag{5.33}$$

where km_{uk} is the size (in bits) of the packet multiplied by the number of packets received by that size. It should be noted that the packet-enumerated loss function (5.33) is perfectly equivalent to the bit-enumerated (5.1). With the new loss function it is however easier to model knowledge of packet-rates than when using (5.1).

For each user u we assign a probability distribution describing our knowledge of the future influxes m_{uk} corresponding to packets of size k. The probability

assignment is analogous to (5.9):

$$P(m_{uk}|I) = \frac{1}{\langle m_{uk}\rangle + 1}\left(\frac{\langle m_{uk}\rangle}{\langle m_{uk}\rangle + 1}\right)^{m_{uk}} \quad , \tag{5.34}$$

and the resulting expected loss contribution of user u is

$$\langle L_u\rangle = \sum_{k\in\{k_u\}}\sum_{m_{uk}=0}^{\infty} P(m_{uk}|I)g\Big(km_{uk} + kS_{uk} - \frac{x_u}{K_u}\Big) \ . \tag{5.35}$$

For each $k \in \{k_u\}$ we must separate between two possible cases, $\frac{x_u}{kK_u} > S_{uk}$ and $\frac{x_u}{kK_u} \leq S_{uk}$, which leads to different expressions. The derivation follows the procedure in Appendix 3.A. Consequently the total user contribution consists of the sum

$$\langle L_u\rangle = \sum_{k\in\{k_u\}} \langle L_{uk}\rangle \tag{5.36}$$

where

$$\langle L_{uk}\rangle = \begin{cases} k\langle m_{uk}\rangle\Big(\frac{\langle m_{uk}\rangle}{\langle m_{uk}\rangle+1}\Big)^{\frac{x_u}{kK_u}-S_{uk}} & , \ \frac{x_u}{kK_u} > S_{uk} \\ k\langle m_{uk}\rangle + kS_{uk} - \frac{x_u}{K_u} & , \ \frac{x_u}{kK_u} \leq S_u \ . \end{cases} \tag{5.37}$$

It should be noted that if there is a wide variety of packet sizes, i.e. if K_u is large, then the expression above would consist of too many terms for it to be tractable in actual calculations. We should then instead assign a probability density for n_u, the number of incoming bits in each buffer. This is possible (see Jaynes, 1963b, for a similar derivation) and results in a Gaussian approximation.

5.3.4 Knowledge of past order sizes

If we have the possibility of collecting histograms of past source influxes for each user, then we could use Laplace's rule of succession to obtain better performance than using just the mean influx of each stream. We would then assume that a connection always carries similar traffic throughout its lifetime, and that there are no temporal correlations that we can infer from the data records. Again, just as in Section 3.2.2 the arrival rates may vary over a very large interval, say from bits per second to tens of megabits per second, and the resulting histograms would be very uninformative if we use one bin for each integer rate. Instead, we distribute a given number K of bins logarithmically over the non-negative integers below a certain upper bound. Using a logarithmic bin scaling[6], we in effect consider the order of

[6]In Chapter 8 we extend the ideas formulated in Section 5.3.4 to adapt the bin sizes according to data instead of using a fixed logarithmic partition.

the influxes to be unknown below some upper limit. Then, we construct histograms over these bins for each user and use the rule of succession (c.f. Section 2.6)

$$P(n_{uk}|m_{u1}...m_{uK}I) = \frac{m_{uk}+1}{M_u+K} \tag{5.38}$$

where m_{uk} is now the number of time slots with influx of size within bin interval k and $M_u = \sum_{k=1}^{K} m_{uk}$. In accordance with the derivation in Section 3.2.2 we then arrive at the expected loss contribution

$$\langle L_u \rangle = \sum_{k=1}^{K} \frac{m_{uk}+1}{M_u+K} \langle L_u \mid n_u \in k \rangle \ , \tag{5.39}$$

where

$$\begin{aligned} \langle L_u \mid n_u \in k \rangle &= \sum_{n_u=a_k}^{b_k-1} \frac{1}{b_k-a_k} g(S_u+n_u-x_u) \ , \qquad k = 1 \ldots K-1 \\ &= \frac{1}{2}\frac{\beta_k^2-\beta_k-(\alpha_k^2-\alpha_k)}{b_k-a_k} + \frac{\beta_k-\alpha_k}{b_k-a_k}(S_u-x_u) \ , \end{aligned} \tag{5.40}$$

and

$$\alpha_k \stackrel{\triangle}{=} \max(x_u - S_u, a_k) \tag{5.41}$$

$$\beta_k \stackrel{\triangle}{=} \max(x_u - S_u + 1, b_k) \ . \tag{5.42}$$

Finally, for $k = K$ we have

$$\begin{aligned} \langle L_u \mid n_u \in K \rangle &\approx \sum_{a_K}^{b_K-1} \frac{1}{\log(b_K/a_K)\, n_u} g(S_u+n_u-x_u) \\ &\approx \frac{\beta_K-\alpha_K+\log(\beta_K/\alpha_K)\,(S_u-x_u)}{\log(b_K/a_K)} \ . \end{aligned} \tag{5.43}$$

5.4 Comments and Simulations

By using prior probability distributions with maximum entropy subject to our information constraints, we avoid assumptions concerning the 'underlying' long-run behavior of the sources. The use of the maximum entropy distribution is motivated because it is the distribution which can arise in the greatest number of ways when the outcomes are constrained to agree with the given information (see the Entropy Concentration Theorem, Theorem 2.1).

Other reasonable approaches to modelling the influxes include using more information in the initial probability assignments, and adapting the distributions according to incoming data using Bayes' theorem. For instance, if we have knowledge of correlations over time or among different user streams, then we can use this information in the maximum entropy formalism to obtain prior distributions of lower entropy than using the mean values only. If such correlations are known to exist but their absolute values are unknown *a priori*, then the initial probability distribution should be updated recursively according to Bayes' theorem as observations of the data streams become available. Another approach, where each radio connection is assumed to operate under stationary conditions but without any correlations, was given in the previous section (with a further generalization to adaptive bin sizes given in Chapter 8) and could be used to improve the performance of the maximum-entropy solutions given earlier. We will however not study the performance of that approach here, as its merits relative to the maximum entropy approach cannot be judged without having access to real traffic. Using a simple simulation set-up based on random-number generators as is done here cannot determine which approach is better in real networks. Our simulation examples will rather be confined to studying the effects of uncertainty concerning arrival rates and effective capacities; therefore, we will here rely on the maximum entropy approach for modelling uncertain source flows.

5.4.1 On the optimality of time division multiple access (TDMA)

Previous work (Bedekar et al., 1999) claims that time division is an optimal scheduling policy in CDMA on the grounds that it minimizes the received power levels from other users. However, in CDMA systems, the bad effects of interference are alleviated by well-designed codes. The interfering users' signal levels are not necessarily harmful to the detection performance of the desired user and thus we cannot conclude that it is always appropriate to use time division.

In spite of this one might conjecture that, would the buffers never be emptied, it might be optimal to use time division also when interference does not affect receiver performance. This conjecture was proven to be true in the deterministic case in the sense of maximizing the sum-of-rates capacity of an uplink in a multiuser single-cell scenario by Knopp and Humblet (1995) when the time-varying fading channels were perfectly tracked and known at the transmitters. In general, however, neither source rates nor channels are perfectly known and buffers may be emptied. Hence, time division is not always the best choice. To see this, consider the problem of scheduling one transmitter one time slot at a time, ie. $R = 1, T = 1$. It can be observed from the expected loss expression (5.12) that if the buffer contents of the user with the highest effective capacity c_u satisfies $S_u \geq c_u$, then the minimum

loss is obtained by transmitting exclusively to that user. If this condition is not met, then we cannot conclude that exclusive transmission is optimal in the sense of maximizing expected throughput.

EXAMPLE 5.1 Sub-optimality of TDMA

Consider the problem of assigning bandwidth across two users using one transmitter and one time slot, i.e. $U = 2$, $R = 1$, $T = 1$. Assume that the users have $S_1 = S_2 = 10$ bits in stock and their expected influx for the next time slot is $\langle n_1 \rangle = \langle n_2 \rangle = 10$. Assume knowledge of the effective capacities, $c_1 = 17$ and $c_2 = 20$.

Figure 5.4 plots the total expected buffer contents using (5.12) as a function of $\rho_1 = 1 - \rho_2$. The optimum assignment is to split the bandwidth almost equally among the users. Even though the user with the highest capacity seems to have a large probability for being able to transmit 20 bits (since $S_2 + \langle n_2 \rangle = 20$) the uncertainty is still considerable and the best decision is to refrain from exclusive transmission. The probability that $n_2 = 0$ is large, and we can only be certain about transmitting 10 bits (the number of bits already in stock) to user 2. Therefore, it would be unnecessarily risky to let user 2 obtain all bandwidth when we know for certain that it can be used to reduce the buffer levels of user 1.

If the scheduler uses a longer time horizon, the minimum loss is obtained with exclusive allocations for each time slot if for every time slot the user with maximum capacity at that time fulfills the criterion $S_u \geq c_{ut}$. If there at any time slot is some user with maximum effective capacity having less data to send than the channel allows, no general conclusion about the optimality of exclusive transmission at any time slot can be drawn. We may conjecture that the scheduler will indeed use exclusive assignments also in many cases that are not covered by the general conditions for optimality; the loss expression does however not give any simple criterion for this to be the optimal choice in general.

Further, for the conjecture to be true, the transmission resources (consisting of antennas, codes, modulation format, etc.) must be such that there is no additional advantage of letting two users transmit at the same time. For instance, some resources might not be mutually exclusive, i.e. two users may utilize them fully at the same time. The model used here does not consider such resources.

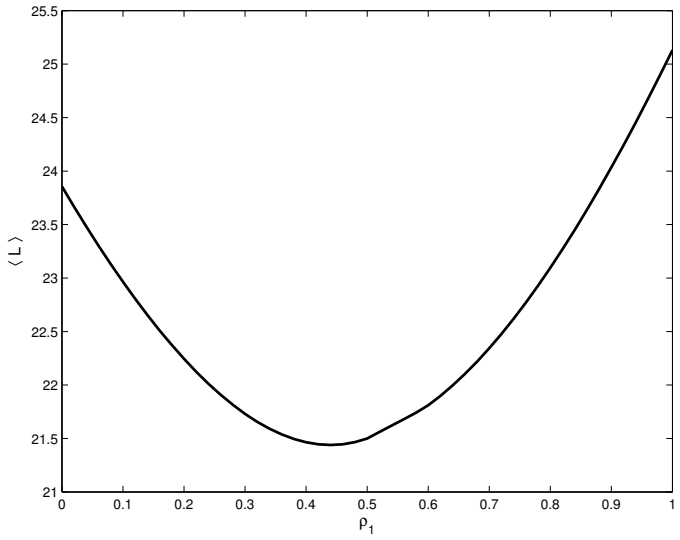

Figure 5.4: The expected loss using (5.12) as a function of $\rho_1 = 1 - \rho_2$ for the scenario in Example 5.1.

5.4.2 Multiuser diversity gain

In this section we investigate how the capacity of a system increases with the number of users when utilizing multiuser diversity.

In Figure 5.5 the sum throughput is plotted as a function of the number of users in a simulated system. The results were obtained using the basic scheduler with perfect channel knowledge using (5.12) in a scenario with two access points. Each user experienced independent Rayleigh fading on the time scale of slots, and the effective capacity was modelled as the integer nearest below the Shannon capacity for a band-limited channel disturbed by additive white Gaussian noise only[7],

$$c_{urt} = \log_2 (1 + \gamma_{urt}) \ , \quad \text{[bits/second/Hertz]} \tag{5.44}$$

where γ_{urt} denotes the SNR at the receiver. Assuming one-tap Rayleigh fading, γ_{urt} is exponentially distributed. The average SNR was set to 10 dB, and the source rates were set so that the transmission buffers were never emptied.

Define the multiuser diversity gain, or scheduling gain, α, as the ratio between the actually obtained total throughput, x, over some given period of time, and the throughput that would have been obtained by simple round-robin scheduling,

[7]The model used here would in reality require perfect channel adaptation and a continuum of modulation levels and coding rates.

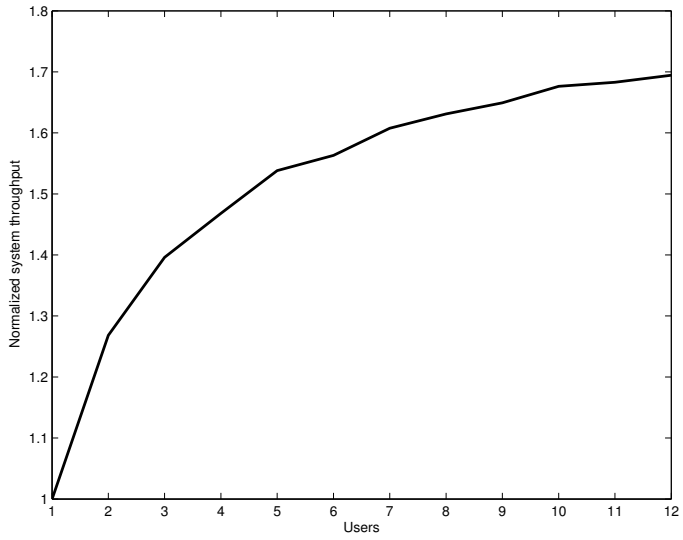

Figure 5.5: The total downlink throughput obtained in a system employing the basic scheduler increases with the number of users. Each user experienced independent Rayleigh fading on the time scale of slots, with an average SNR of 10 *dB.*

$x^{(RR)}$, over the same period,

$$\alpha = \frac{x}{x^{(RR)}} \ . \tag{5.45}$$

Figure 5.5 then describes the scheduling gain of the simulated scenario, since round-robin scheduling gives a sum throughput equal to the average effective capacity for any one of the users.

Remember that the effective capacity increases logarithmically with SNR (c.f. (5.44)). Consequently, SNR fluctuations around a small average SNR causes rate fluctuations of the same order as the SNR fluctuations, while for a larger average channel gain, SNR fluctuations give smaller effects on the rate scale. It follows that multiuser diversity gains are more pronounced at low SNR averages. Consequently, if we weigh each user's loss contribution with the reciprocal of that user's average effective capacity, users with low average SNRs will be favored compared to high-SNR users if their SNR fluctuations are of the same magnitude. This results in reduced system throughput, and puts users with good channel conditions at an unexpected disadvantage. Non-obvious effects like this may follow for other forms of suggested fairness constraints as well. Compensating one set of users often puts other users at an unforeseen disadvantage.

5.4.3 Comparison with proportional fair scheduling

Viswanath et al. (2002) considered a diversity scheme consisting of a scheduler and randomized beamforming and compared it to the two-antenna space-time block coding scheme by Alamouti (1998) and coherent beamforming without scheduling. It was found that scheduling is not only a viable and economic alternative requiring little feedback; in a multiuser setting with enough users the proposed scheme also strictly outperformed space-time coding. With many users and few antennas, scheduling with randomized beamforming also approaches the performance of coherent beamforming while requiring significantly less feedback.

There is however an obvious problem with randomized beamforming. In typical settings, the merits of scheduling depends on channel predictions; this is effectively contradicted by randomized channels.

In a new set of simulations we compared the proportional fair scheduler of Viswanath et al. (2002) and the basic scheduler from Section 5.3.1 with knowledge of effective capacities (using (5.12)). Both these schedulers use knowledge of the channel to guide their decisions. The proportional fair scheduler does however not consider the effects of source rates and hence the possibility of empty buffers. Implicitly it assumes that there is always data to send.

The proportional fair scheduler works as follows. The data rates that the users can receive at (given some BER requirement) at each time slot t (the effective capacity, c_{urt}) is known to the scheduler. The scheduler then keeps track of the average throughput $T_u(r,t)$ of each user u in a past window of length t_c. At each base station r and time slot t, the scheduler transmits exclusively to the user with the largest $\frac{c_{urt}}{T_u(r,t)}$. The parameter t_c is used as a forgetting factor in the calculation of the windowed average throughput. It is used as a means of obtaining fairness, by giving a user access to a channel when its effective capacity is high relative to its own average throughput over the time scale t_c. Viswanath et al. (2002) considered a single base station only. Here, we adapt the proportional fair scheduler to multiple base stations/antennas simply by treating an additional base station as more time slots. In other words, if we are to assign two base stations and three time slots, the scheduler works exactly as if it were to schedule one base station and six time slots. After each single assignment, the average throughput $T_u(i)$ (where i indexes assignments regardless of whether it describes time slot or base station) is recalculated according to (Tse, 2001)

$$T_u(i) = (1 - \frac{1}{t_c})T_u(i-1) + \frac{1}{t_c}c_{u,i-1}\delta(u - u^*) \ , \tag{5.46}$$

where $\delta(u - u^*) = 1$ if user u was the transmitting user u^* in the most recent assignment, otherwise, $\delta(u - u^*) = 0$.

The schedulers were run on the same data sets, with source rates n_{ut} drawn from a Poisson random number generator[8], and effective capacities generated from the rate expression (5.44) using an exponential pdf for the SNR. The parameters used are listed in Table 5.2. The forgetting factor for the proportional fair scheduler was set to $t_c = 7$.

The simulated scenario consisted of two base stations and three users. The scheduling horizon was $T = 3$ time slots, and the schedulers were run for a total of 60 time slots. The results listed in Table 5.3 are averages from 100 realizations. The table reports average throughput and average buffer levels after the 60 time slots (the averages being over the 100 realizations).

The results show that in this scenario the total throughput increases by approximately 30% using (5.12) compared with using the proportional fair scheduler. In particular, the throughput of user 2 is severely degraded when buffer contents are neglected. In terms of buffer levels it is clear that the second user's buffer would overflow, causing further throughput degradation and increasing delays due to the invoking of higher-layer mechanisms such as decreasing transmission rates or retransmissions.

Comparing the results for users 2 and 3, having equal channel statistics, we see that the throughput ratio of the two users is identical to the ratio of their average inflows when using maximum entropy scheduling. If the inflows are taken to reflect each user's service requirements, then this means that fairness is obtained without any explicit fairness constraint on the policy. On the other hand, a user with very low average SNR and small channel variability would obviously risk starvation with the proposed scheduler.

It can be noted that the proportional fair scheduler could approach the performance of the maximum entropy scheduler were the transmission buffers constantly flooded with data. A more important observation is that this case is normally prevented from occurring in a real system due to rate-control mechanisms such as provided by TCP. Schedulers should therefore always take buffer contents into account. The additional use of *source rate diversity* further increases the performance of the maximum entropy scheduler.

Another interesting result from this simulation can be observed by studying the throughput obtained for the second user, 326 bits. Instead of trying to use multiuser diversity to our advantage we could split the available bandwidth into three equal parts, and always transmit to all users. Instead of 326 bits, user 2 would then obtain a total throughput of $\frac{3.7}{3} \times 2 \times 60 = 148$ bits. Thus, the individual throughput increases by 120% when using the fluctuating channel as a source of diversity.

[8]This choice is admittedly somewhat arbitrary. For a discussion of the problems involved in modelling and simulating individual Internet sources, see Floyd and Paxson (2001).

Table 5.2: Parameters for the comparison of proportional fair scheduling with the maximum entropy scheduler for known channels. Average inflows per time slot, $\frac{\langle n_i \rangle}{T}$, average SNR (dB) at the receiver, γ_u, and the corresponding average effective channel capacity (number of bits per time slot), $\langle c_{urt} \rangle$.

	$\frac{\langle n_i \rangle}{T}$	$\gamma_u(dB)$	$\langle c_{urt} \rangle$
User 1	2	10	2.9
User 2	6	13	3.7
User 3	1.5	13	3.7

Table 5.3: Results for the comparison of proportional fair scheduling with the maximum entropy scheduler for known channels. The average number of bits remaining in the buffers after 60 time slots are listed in columns 1 and 2 for the proportional fair scheduler (PF) and the scheduler with known c_{urt} proposed here (ME). The last two columns display average total throughput in bits.

	S_{60}(PF)	S_{60}(ME)	Tp(PF)	Tp(ME)
User 1	2	11	117	108
User 2	170	35	191	326
User 3	0	4	92	88
Total	172 bits	50 bits	400 bits	522 bits

The proportional fair scheduler only achieves an increase of 29% since it does not take the varying source rates into account. Evidently, there are substantial benefits associated with taking advantage of the fact that, on average, the other users' source rates are lower than their effective capacities. Neglecting this source of diversity results in decreased individual and total throughput.

5.4.4 Results for different amounts of channel uncertainty

Having established that taking channel information and source rates into account are critical issues, two questions naturally arise:

1. How does the accuracy of channel predictions affect individual and total throughput?

2. Do we need to use the more complex scheduler when using inaccurate channel predictions or can we equally well use the simpler one, assuming perfect channel knowledge?

To answer the first question, we study the throughput degradation of a user as a function of increasing prediction inaccuracy. The simulation setup consists of scheduling six users according to (5.19), with two transmitters, $R = 2$, and a scheduling horizon of $T = 3$ time slots. All users have an average SNR of 10 dB, and the Rayleigh fading model from Section 5.4.2 is used with the effective capacity described by (5.44). (The average potential effective capacity is thus approximately 2.9 bits.) The buffer influxes are large compared to the effective capacities. All users except the first one have nearly perfect prediction, $\sigma_{urt} = 0.1$. During a simulation run for 60 time slots, user one's prediction accuracy was held at a constant value. The simulation was then repeated for a range of increasing prediction inaccuracies $\sigma_{1rt} = 0.1 \ldots 3.5$. Figure 5.6 shows the throughput of user one for two different BER sensitivities, $v = 1.3$ and $v = 1.1$. We see that the throughput degrades very quickly for decreasing prediction accuracy. Already at $\sigma_{1rt} = 0.15$ the throughput has degraded to roughly 60% of what a user with $\sigma_{1rt} = 0.1$ obtains. The reason is that there is almost always another user with equally high predicted capacity, but with higher accuracy, thereby leaving user one at a disadvantage since a larger uncertainty σ_{urt} results in lower expected effective capacity (5.22).

In terms of an individual user's performance, therefore, an important property of the predictor is that its accuracy should be comparable to that of the other users. On the level of system throughput, since the expected throughput $\langle x_{urt} \rangle$ decreases with prediction inaccuracy, the total throughput necessarily decreases too if the accuracy is equal among users. But if the accuracy varies independently among users, it is likely that there is at least one user with both high SNR and high accuracy. In this sense, prediction accuracy should preferably vary substantially around some average, rather than be constant at that average. Furthermore, prediction accuracy in the high-SNR region is more important than for low SNR, since a user will only be scheduled for transmission in the former case.

Addressing the second question, the same simulation setup was also run with the basic scheduler using (5.12) but employing the predicted values of the effective capacity, $\hat{c}_{urt}$, instead of the true values. The sum throughput using (5.19) relative to the throughput corresponding to using (5.12) is given in Figure 5.7. It can be seen that there is a significant performance difference between the two schedulers[9] when there is a considerable prediction uncertainty for some users (in this case only one) while other users have high prediction accuracy. This implies that the

[9]Notice that if all users would have had the same prediction accuracy (this is unlikely, since different users move at different velocities and at higher velocities the channel changes faster than for a stationary user), then there would not have been any performance difference between the two schedulers, since using (5.19) would merely reduce all users' expected capacity by a nearly equal amount.

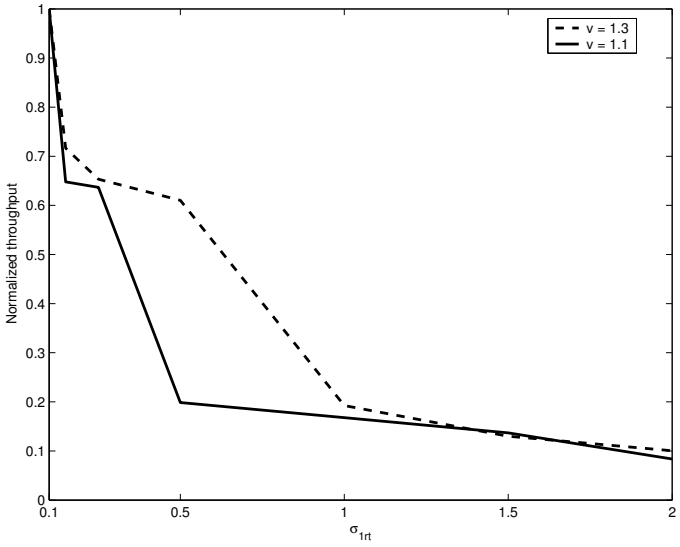

Figure 5.6: The normalized throughput (1 corresponding to the throughput of user one if $\sigma_{1rt} = 0.1$) for user one as a function of σ_{1rt}. All users had the same average source rates and potential effective capacities ($\langle \bar{c} \rangle \approx 2.9$) (cf. Section 5.4.4). The two curves correspond to different values of the BER sensitivity v.

more complex scheduler should be used in situations where different users have different prediction accuracies, for instance due to different user velocities (which affect how fast the channel varies and thus how predictable it is). There is however an intermediate solution which offers better performance than just using the estimate $\hat{c}_{urt}$ and also lower complexity than using (5.19); note that if we replace c_{urt} by $\langle c_{urt} \rangle$ from (5.22) as an estimate, since increased prediction uncertainty leads to a decreased estimate of the effective capacity, we will come closer to the performance of the more complex expected loss (5.19). The two approaches coincide when $\langle x_u \rangle \leq S_u$; when $\langle x_u \rangle > S_u$ it shows a qualitatively similar behavior to that of (5.19) since the conditional-mean estimate $\langle c_{urt} \rangle$ takes account of the risk for lower-than-predicted capacity.

5.4.5 Scheduling one time slot at a time using exclusive allocations

Despite the fact that exclusive allocations are generally suboptimal, one may in practice use them anyway. Certain architectures only allow exclusive transmissions, and there is often a computational advantage as well. In most works on scheduling in wireless communications, due to the optimality result of Knopp and

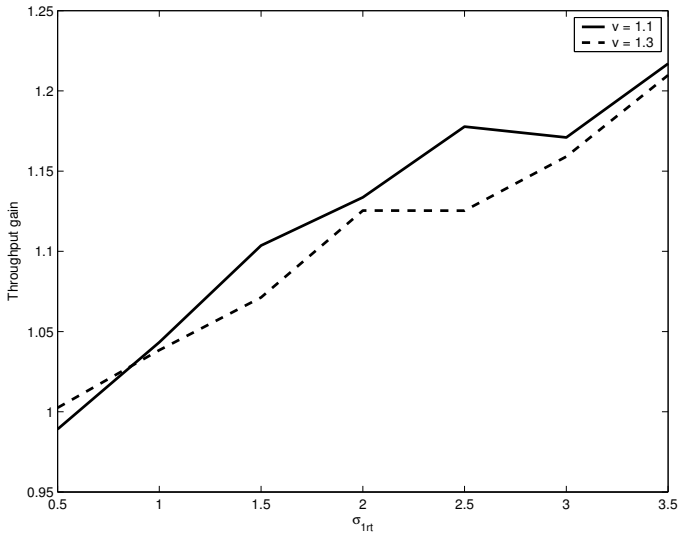

Figure 5.7: The relative throughput gain of the scheduler using knowledge of prediction accuracy as compared to the basic scheduler using the predictions $\hat{c}_{urt}$. All users had the same average source rates and average potential effective capacities ($\langle \bar{c} \rangle \approx 2.9$) (cf. Section 5.4.4). The two curves correspond to different values of the BER sensitivity v.

Humblet (1995) and perhaps for reasons of practical constraints, exclusive allocations are the only alternative considered. Similarly, almost without exception, scheduling over more than one time slot is not discussed.

We have seen that exclusive allocations are in general not the optimal choice in the sense of maximizing expected throughput. Moreover, the length T of the scheduling horizon should be chosen with some care. If possible, a longer horizon should be used since the flexibility of the scheduler increases with T. This is particularly important when the allocations are constrained so as to meet quality-of-service demands. There is however a trade-off involved in the choice of T; the channel prediction accuracy decreases with the prediction horizon, and the uncertainty concerning the source rates increases similarly. In the current literature, it is typically assumed that there is no possibility of obtaining accurate channel predictions for more than one time slot ahead; hence the choice $T = 1$. On the other hand, Ekman (2002) has shown that it is indeed possible to obtain reliable channel prediction for longer horizons[10], and we could therefore choose a slightly larger T.

[10]The prediction performance for a given horizon (measured in distance) depends on the physics of the surrounding environment and the speed of the mobile terminal in relation to the carrier frequency.

As it is many times supposed that Bayesian solutions are prohibitively complex, we would like to point out that this is often a misconception. If we consider the same case that the current literature focuses on, i.e. exclusive allocations and $T = 1$, then the decision which maximizes the expected throughput is to choose the user u with maximum

$$\langle L_u^{diff} \rangle \stackrel{\triangle}{=} \langle L_u(\rho_{ut} = 0) \rangle - \langle L_u(\rho_{ut} = 1) \rangle \tag{5.47}$$

with $\langle L_u(\rho_{ut}) \rangle$ given by (5.12), or, for uncertain capacities, (5.19). To see this, note that the best decision is to choose the user which reduces the total expected loss $\langle L \rangle = \sum_{u=1}^{U} \langle L_u \rangle$ the most of all users. If user u is chosen, only the term $\langle L_u \rangle$ corresponding to that user is affected, and the reduction is the expected number of bits that can be sent to user u, that is $\langle L_u(\rho_{ut} = 0) \rangle - \langle L_u(\rho_{ut} = 1) \rangle$. This rule, to select the user maximizing (5.47), involves no dynamic programming or numerical optimization, and is computationally approximately equivalent to other proposed schedulers.

5.5 Other Approaches to Scheduling in Mobile Communications

5.5.1 Queue stability

We noted in the introductory section of this chapter that there is a line of work that takes queue stability as the most important property of a scheduler. The idea is that the transmission buffers should not unnecessarily overflow, which intuitively seems like a desirable quality. Before considering how to determine a scheduler with this property, we must decide what we mean by the term 'stability'.

In the papers on queue-stable schedulers 'stability' of a scheduler is taken to mean that all data are transmitted in a bounded amount of time, i.e., the queues are bounded over time. Now, obviously no scheduler can guarantee stability for any amount of traffic or any channel quality. Instead, the system of data arrivals and transmission capacities is said to be 'admissible' if there exists some schedule which can maintain queue stability for the particular system. A scheduler is then said to be stable if it keeps the queues bounded whenever the system is admissible.

Andrews et al. (2000) announced a mathematical theorem which says that if and only if the average arrival rate for each user is less than that user's average effective capacity, then the system is admissible. Obviously then, stability may appear to be possible over a certain time scale, and then if the effective capacities

For an analysis of attainable prediction horizons under the Jakes Rayleigh fading model see Ekman et al. (2002)

or data rates change among users, it may turn out that the system is not admissible any more. The same reference also proves, for the case $R = 1$ and $T = 1$, that transmitting exclusively to the user u with maximum

$$\psi_u c_u S_u^\beta \,, \tag{5.48}$$

where ψ_u and β are arbitrary non-negative constants, is a stable scheduling rule. The stability result holds also when replacing S_u by the maximum time d_u that any bit in buffer u has spent in that buffer. Putting $\beta = 1$ and $\psi_u = 1$, we see that this queue stabilizing scheduler (denoted Modified Largest Weighted Delay First, M-LWDF) chooses users with high transmission rates who have not been served in a long time.

Although queue stability may seem desirable at first glance, it would be interesting to see some examples of criteria that lead to queue stabilizing schedulers. For instance, does our criterion, to maximize the expected throughput, guarantee stability if possible at all? Certainly so, since the buffer levels are always kept to its minimum allowed value. But our criterion is more explicit; it always maximizes throughput, and continues to do so even when stability cannot be guaranteed. Further, we should ask ourselves whether stability is that important after all. In order to guarantee that all queues be bounded at *all* times when possible, we effectively put equal weight to the most unlikely, but possible, events as we do to the most typical ones. We maintain that stability may perhaps be an indicator that a scheduling discipline is useful in certain cases, but it rarely, if ever, corresponds to the actual goal we have set for our communications system. Such a goal should be stated clearly in a loss function, so that we can then minimize the *expected* loss placing the optimal weight (i.e. the posterior probability) on the different possible outcomes.

Although scheduling the user that maximizes (5.48) apparently has the property that it keeps queues bounded whenever possible, that in itself does not tell us what loss function the M-LWDF scheduler actually corresponds to. It is an *ad hoc* scheduler, but can we find a loss function that actually leads to the M-LWDF discipline? Several motivations may lead to the same decision in the end, and here we provide one possible such motivation. Casimiro Ericsson (2004) notes that the M-LWDF method can be derived as an approximation to a loss which sums the squares of the buffer levels. Writing the loss as

$$L_u = (\overline{S_u})^2, \tag{5.49}$$

where

$$\overline{S_u} = S_u - \delta S_u \tag{5.50}$$

is the buffer level after the next scheduled time slot(s) with $\delta S_u = n_u - x_u$, we

have

$$(\overline{S_u})^2 = (S_u + \delta S_u)^2 = S_u^2 + \delta S_u^2 + 2\delta S_u S_u \quad (5.51)$$
$$= S_u^2 + \delta S_u(\delta S_u + 2S_u)\ . \quad (5.52)$$

The first term is not affected by δS_u and thus the decision which minimizes the sum-quadratic loss (5.49) minimizes $\delta S_u(\delta S_u + 2S_u)$. Assuming that δS_u is much smaller than S_u, an approximation to minimizing (5.49) is to minimize $2\delta S_u S_u$, or equivalently, since n_u is fixed, to maximize

$$x_u S_u\ . \quad (5.53)$$

This is the M-LWDF method (5.48) with $\beta = 1$ and $\psi_u = 1$.

In the literature on queue stability, scheduling algorithms that maintain stability whenever that is possible are called 'throughput optimal'. This, however, is severely misleading. In the standard scenario considered in these references (see e.g. Andrews et al., 2001) there is one time slot and one resource to schedule under no uncertainty. The decision that maximizes the throughput in simply to choose the user with maximum c_u (if that user has at least that much data to send). The M-LWDF scheduler certainly does not follow this rule, and we can see no reason why this scheduler, or any other that does not attempt to achieve maximum throughput, should be defined as 'throughput optimal'. The repeated misuse of this term leads to a false impression of the far from optimal results that these *ad hoc* schedulers achieve and may mislead unwary workers and reviewers in comparing different approaches.

5.5.2 Proportional fairness vis-à-vis logarithmic loss

We have already discussed the proportional fair scheduler in some detail in Section 5.4.3. Here we note a feature that has not been emphasized in the current literature, and which we think provide an important rationale for its use.

Consider the problem of scheduling U users over one time slot and one transmitter. Suppose now that only exclusive allocations are possible and that we know the effective capacity c_u and the expected influx $\langle n_u \rangle$. Assuming further that the number of bits in stock for each user is larger than that user's effective capacity, the optimal allocation is to transmit to the user with maximum c_u. Now, consider two users, Mr A and Mr B. Mr A has an average throughput of 10 bits per time slot, while Mr B on average receives 1000 bits per time slot. If Mr A has $c_A = 20$, and Mr B has $c_B = 21$, our scheduler will award Mr B channel access. But Mr B is not likely to even notice that he gets the extra 10 bits, since this amount is extremely small compared to his average throughput. If he is downloading a large

file, the time it takes will hardly be affected by this extra throughput, while on the other hand Mr A would have noticed a most dramatic performance increase, receiving twice the amount of data that he is accustomed to, would he instead have been given access. As we noted in Chapter 3, Daniel Bernoulli (1738) observed that the latter decision in general seems a more rational course of action to most people. Indeed, we see the reason very clearly in our example. A doubling of the average rate implies halving the download time. It seems that whether the amount of time we halve is a minute or an hour, the utility for the user is the same. This is reminiscent of the scale invariance argument for priors which led to a uniform distribution for the logarithm of the parameter, and similarly, Bernoulli concluded that the 'utility resulting from any small increase in wealth will be inversely proportionate to the quantity of goods previously possessed'. From this he finds that the corresponding utility y for someone already in possession of an amount α is

$$y = b \log \frac{x}{\alpha} \tag{5.54}$$

when increasing his possessions to the total amount x. The constant b is arbitrary. Notice that $x = \alpha + \Delta$ where Δ is the new amount that the person received. In our problem, we may thus use the individual loss

$$L_u = -\log\left(\frac{\min(x_u, S_u + n_u)}{\bar{x}_u}\right) , \tag{5.55}$$

where $\bar{x}_u$ denotes the mean allocated throughput that user u has actually obtained, and $\min(x_u, S_u + n_u)$ is the number of bits transmitted over the link, the $\min()$ accounting for the case when the buffer levels are lower than the effective capacity. Under the assumption that $S_u + n_u \geq x_u$ at all times, we see that for $T = 1$ and $R = 1$ the optimal decision according to this rule is to choose the user with maximum $\frac{x_u}{\bar{x}_u}$, the same decision that the proportional fair scheduler makes. It is well-known that the proportional fair scheduler can be derived from the logarithmic loss, but in the literature this loss is motivated because of its fairness property. We stress the converse relation. The logarithmic rule is useful because it is a more natural measure of the actual 'moral' value of a given allocation than the absolute value of the throughput. It further has the desirable property of achieving fairness in the sense that users with relatively low channel quality are not completely shut-off from transmission, but this is not the primary reason for using it. Fairness, in some sense of this elusive concept, can be achieved in many ways, and if fairness in itself is the ultimate goal then we should explicitly state in a loss function how we define it. If instead – as is done here – we take the stance that the value of a communication link for a user lies in receiving data, then the throughput is more fundamentally important than fairness, and we should define in exactly what sense

a rate allocation is useful for the user. In this chapter we have taken the absolute throughput as our main criterion, but perhaps, as our discussion here indicates, we should indeed use the logarithmic measure.

With the logarithmic loss (5.55), we have for the case of uncertain source rates but perfectly known capacities

$$\begin{aligned}\langle L_u \rangle &= P(n_u + S_u \geq x_u | I) \log\left(\frac{x_u}{\bar{x}_u}\right) \\ &+ P(n_u + S_u < x_u | I) \left\langle \log\left(\frac{S_u + n_u}{\bar{x}_u}\right) \middle| n_u + S_u < x_u \right\rangle \end{aligned} \quad (5.56)$$

where we use $\langle A|B\rangle$ to denoted the expectation of A given knowledge of B. If $S_u > x_u$ then the resulting loss is simply

$$L_u = \log\left(\frac{x_u}{\bar{x}_u}\right) \qquad S_u > x_u \,, \quad (5.57)$$

where there is no longer any uncertainty to average over. This corresponds to the traditional proportional fair scheduler when $R = 1$ and $T = 1$. If however $S_u \leq x_u$, then the uncertainty as to the outcome remains, and we must consider the second term as well. With $P(n_u|I)$ based on knowledge of $\langle n_u \rangle$, i.e. using (5.9), we cannot obtain a closed-form expression for the expectation of $\log(\frac{S_u+n_u}{\bar{x}_u})$. If we instead use $n'_u = \langle n_u | n_u + S_u < x_u \rangle$ directly in the logarithm instead of carrying out the correct sum, then, using (5.56), we have an estimate $\hat{L}_u$

$$\begin{aligned}\hat{L}_u &= \sum_{n=x_u-S_u}^{\infty} \frac{1}{\langle n_u \rangle + 1} \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{n_u} \log\left(\frac{x_u}{\bar{x}_u}\right) \\ &+ \sum_{n=0}^{x_u-S_u-1} \frac{1}{\langle n_u \rangle + 1} \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{n_u} \log\left(\frac{S_u + n'_u}{\bar{x}_u}\right) \\ &= \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{x_u - S_u} \log\left(\frac{x_u}{\bar{x}_u}\right) \\ &+ \left(1 - \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{x_u - S_u}\right) \log\left(\frac{S_u + n'_u}{\bar{x}_u}\right) \qquad S_u \leq x_u \end{aligned} \quad (5.58)$$

where the adjusted expectation[11] of n_u is

$$n'_u = \langle n_u \rangle + \frac{(S_u - x_u)\left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{x_u - S_u}}{1 - \left(\frac{\langle n_u \rangle}{\langle n_u \rangle + 1}\right)^{x_u - S_u}} \ . \tag{5.60}$$

In this expression we have a straightforward (although approximate) generalization of the proportional fair scheduler taking uncertainty of the arrival rates into account and extending it to an arbitrary scheduling horizon T as well as an arbitrary number of transmitters R. For an extension to the case of uncertain effective capacities, we would have to carry out the steps in Section 5.3.2 for the new loss expression. We leave this as an open door for further development.

5.6 Competitive Bidding – A Possible Solution to the Quality-of-Service Dilemma?

Recall the generalized throughput criterion (5.8) where each user's buffer contents are weighed according to

$$\langle L \rangle = \sum_{u=1}^{U} \pi(u, \{\theta_u\}) \langle L_u \rangle \ . \tag{5.61}$$

Now consider setting

$$\pi(u, \{\theta_u\}) = q_u \tag{5.62}$$

where q_u is the price that user u pays per transmitted bit. It then follows that by allowing users to adjust their prices on-line different quality-of-service targets can be achieved. There are at least two possible ways of setting the dynamic prices. One approach is to let the network operator set the prices according to each user's demand, but the lack of transparency of such a solution is clearly undesirable. A user would have to trust the operator in not increasing prices without reason.

Instead it may be a better idea to let each user set his/her own prices in a procedure such as that considered in Chapter 4. There, the base station was supposed to

[11] The expectation of n_u when $S_u \leq x_u$ is obtained by evaluating

$$\frac{\sum_{n_u=0}^{x_u - S_u - 1} n_u P(n_u|I)}{\sum_{n_u=0}^{x_u - S_u - 1} P(n_u|I)} \ . \tag{5.59}$$

These arithmetic-geometric and geometric series are solved in the derivations of the expected loss (5.12).

generate reports in regular time intervals consisting of the average winning price-capacity product and its sample variance. It was found that such a solution allows users to differentiate their prices according to service demands. Simulations showed that rate requirements were indeed satisfied with a reasonable degree of confidence. The advantages of such a solution include that the bit prices would actually reflect the current demand-supply situation, thereby yielding a true market-economic mechanism, and that with individual price adjustments at the mobile terminals, a complicated U-dimensional dynamic-programming problem would be avoided at the base station. Potential disadvantages include that very rich customers could starve all other users, and that the feedback information sent to and from the base station increases. The latter problem can however be alleviated by allowing only a discrete set of possible price changes. Using for instance 2 bits of feedback for price updates ($+/-$ 1 or 2 units) at regular but infrequent intervals should be sufficient to be able to maintain a desired service level. Another disadvantage is that a customer can never acquire a guaranteed service level by use of this scheme. There is an element of uncertainty concerning the future service level which may be unacceptable for applications with real-time service requirements. For other types of traffic the gain in flexibility and the probable over-all reduction in prices from using competitive bidding may be compelling reasons to adopt the considered scheme. A more serious concern, however, stems from the customers' uncertainty regarding the prices in this type of scheme. Customers may not appreciate price uncertainty, but rather prefer a fixed price regardless of the amount of data downloaded. An operator adopting a fixed price regardless of the amount of throughput must however set a price which reflects the operator's uncertainty about the customers' bandwidth requirements, yielding higher prices than with the auctioning procedure, or introduce flow control mechanisms that put hard regulations on the customers' data flows. Most likely, a combination of higher prices and flow regulations would result from this type of fixed pricing. In the end, whether the customers actually prefer the auctioning procedure or the fixed-price scheme would depend on the typical price reduction a customer obtains with the auctioning procedure and the typical service-level variations.

In summary, for networks where the major revenue streams come from data traffic such as Internet browsing, dynamic pricing using competitive bidding could constitute an attractive compromise between resource utilization efficiency, quality of service and low costs for the end user, but the uncertainty regarding prices and service-levels may potentially outweigh these advantages.

5.7 Conclusions

In this chapter a problem of optimizing channel assignments in the presence of uncertainty was considered for applications in mobile communications. The problem was formulated as a minimization of the expected total buffer contents, given by the general expression (5.4), a sum of contributions from each user. It was noted that the framework is compatible with user priorities represented by known functions describing an equivalent cost per bit. In Chapter 4 we noted that introducing competitive bidding in combination with maximum-throughput scheduling as a means for acquiring a desired quality of service was a feasible solution, although the additional signaling over-head and the potential problems from unpredictability may limit its usefulness for real-time traffic.

Each user's contribution to the total expected loss was calculated for four different cases, each representing a typical state of knowledge at the scheduler. With knowledge of effective capacities and of average influxes, the expected loss contribution was found in (5.12). Using knowledge of the accuracy of capacity predictions, a Gaussian distribution was assigned for the predicted capacities. It was noted that the obtained capacity is a function of the prediction, and the resulting probability distribution for the effective capacities was derived for the case when too large predictions result in a linear decrease of obtained capacity. The consequent expected loss contribution was found in (5.19). In a packet data system with knowledge of packet sizes, effective capacities, and average influxes for each packet size, the resulting expected loss contribution was described by (5.35). Finally, with knowledge of past influxes, an expected loss expression based on the rule of succession applied to a logarithmic partitioning of the influx sizes was given by (5.39).

A substantial increase in throughput due to multiuser diversity gain from maximum entropy scheduling was demonstrated in simulations. A comparison of maximum entropy scheduling with the proportional fair scheduler showed that the maximum entropy scheduler achieved higher throughput by also utilizing source rate diversity. Further simulations demonstrated that in order to obtain high throughput the scheduler needs to have accurate channel knowledge. Degradation of channel prediction accuracy for one user inevitably led to reduced throughput for that user as described by Figure 5.6. Including knowledge of prediction accuracy into the criterion resulted in improved system performance compared to using the basic criterion with predicted capacities instead of the true values. The performance difference was a consequence of exploiting the variations in prediction accuracy. The larger the variations in channel prediction accuracy and the more users in the system, the larger the resulting gain of using the full Bayesian solution (5.19). With small or no variations of prediction accuracy among the users there was no

performance difference (cf. Figures 5.6 and 5.7).

The Bayesian solution thus prioritizes users with well-determined high-rate channels, and with data to send. In the limit, as the number of users tends to infinity and the prediction accuracies vary independently over the users, the full Bayesian solution would approach the throughput of the scheduler with perfect channel knowledge.

Observe also that any of the proposed expected loss expressions could be used in other types of schedulers as well. For instance, with strict delay requirements, a simple and effective scheme for exclusive one-slot scheduling would be to transmit to the user u who yields the largest total loss decrease, $\langle L(\rho_{ur} = 0)\rangle - \langle L(\rho_{ur} = 1)\rangle$ (which is the best exclusive scheduling policy in the sense of minimizing expected loss). Then at the next time slot, the remaining $U - 1$ users would compete similarly. For each time slot, the set of competing users is reduced, and after U time slots, the process repeats. The maximum delay for any user would then be $2U - 1$ time slots. We will have reason to come back to this simple scheduler in Chapter 6 when we consider the implications of limited feedback channels on system performance.

In conclusion it should be pointed out that, although the framework was formulated in a communication theoretic setting, the rationale can be employed in other forms of flow optimization problems where the demand, n_u, is incompletely known. The case of incompletely known supply, c_{urt} corresponding to the solution laid out in Section 5.3.2, would however require a different supply distribution than here. This is in principle straightforward; given any testable information regarding the actual supply mechanisms, find the $P(c_{urt}|I)$ that maximizes the corresponding entropy. Given that model, the solution that maximizes the number of satisfied orders is again given by (5.4).

Appendix 5.A Derivation of Expected Loss given Time-Varying Influx Averages

Here we derive the expected loss contribution for known time-varying influx averages, assuming perfect knowledge of the effective capacities. The probabilities for n_{ut} for different times t factor according to the maximum entropy principle and thus we can rewrite the expected loss contribution as a product of independent terms. As in (5.12) we need to separate between the cases $x_u > S_u$ and $x_u \leq S_u$. It follows immediately from the derivation of (5.12) in Appendix 3.A that for $x_u \leq S_u$ the loss contribution for user u is

$$\begin{aligned} \langle L_u \rangle &= S_u + \textstyle\sum_{t=1}^{T} \langle n_{ut} \rangle - x_u \\ &= S_u + \langle n_u \rangle - x_u \qquad , \; x_u \leq S_u \; . \end{aligned}$$

Consider the calculation of $\langle L_u \rangle$ in the case $x_u > S_u$. For reasons we shall come back to in the derivation we need to reorder the $\langle n_{ut} \rangle$ by decreasing size. Thus, we replace the time indexes t by size indexes k, where larger k corresponds to smaller size. We start by deriving the average loss with respect to $P(n_{u1}|I)$, for given smaller influxes $n_{u2}, n_{u3}, \ldots$, which we denote by $\langle L_u \rangle_{P(n_{u1}|I)}$. By substituting $S_u + \sum_{k=2}^{T} n_{uk}$ for S_u in the derivation of (5.12) in Appendix 3.A it follows directly that:

$$\begin{aligned} \langle L_u \rangle_{P(n_{u1}|I)} &= \langle n_{u1} \rangle \Big(\frac{\langle n_{u1} \rangle}{\langle n_{u1} \rangle + 1} \Big)^{x_u - S_u - \sum_{k=2}^{T} n_{uk}} \\ &= \langle n_{u1} \rangle \Big(\frac{\langle n_{u1} \rangle}{\langle n_{u1} \rangle + 1} \Big)^{x_u - S_u} \prod_{k=2}^{T} \Big(\frac{\langle n_{u1} \rangle}{\langle n_{u1} \rangle + 1} \Big)^{-n_{uk}} \; . \end{aligned} \tag{5.64}$$

This means that the expected loss averaged over the influxes at the remaining times, $n_{u2}, \ldots,$ becomes:

$$\langle L_u \rangle = \langle n_{u1} \rangle \Big(\frac{\langle n_{u1} \rangle}{\langle n_{u1} \rangle + 1} \Big)^{x_u - S_u} \prod_{k=2}^{T} \sum_{n_{uk}=0}^{\infty} P(n_{uk}|I) \Big(\frac{\langle n_{u1} \rangle}{\langle n_{u1} \rangle + 1} \Big)^{-n_{uk}} \tag{5.65}$$

The sum over n_{uk} in (5.65) is, by using (5.9), given by

$$\begin{aligned} & \sum_{n_{uk}=0}^{\infty} \frac{1}{\langle n_{uk}\rangle + 1} \Big(\frac{\langle n_{uk}\rangle}{\langle n_{uk}\rangle + 1}\Big)^{n_{uk}} \Big(\frac{\langle n_{u1}\rangle}{\langle n_{u1}\rangle + 1}\Big)^{-n_{uk}} \\ = & \sum_{n_{uk}=0}^{\infty} \frac{1}{\langle n_{uk}\rangle + 1} \left(\frac{\langle n_{uk}\rangle}{\langle n_{uk}\rangle + 1} \frac{\langle n_{u1}\rangle + 1}{\langle n_{u1}\rangle}\right)^{n_{uk}} \qquad (5.66) \\ = & \frac{1}{\langle n_{uk}\rangle + 1} \left(\frac{1}{1 - \frac{\langle n_{uk}\rangle}{\langle n_{uk}\rangle + 1} \frac{\langle n_{u1}\rangle + 1}{\langle n_{u1}\rangle}}\right) . \qquad (5.67) \end{aligned}$$

In the last equality the reordering of $\langle n_{uk}\rangle$ by decreasing size is needed to ensure convergence of the geometric series (5.66) (eqn. 0.231.1 in (Gradshteyn and Ryzhik, 2000)) , which requires $\frac{\langle n_{uk}\rangle}{\langle n_{uk}\rangle + 1} \frac{\langle n_{u1}\rangle + 1}{\langle n_{u1}\rangle} < 1$. The average loss is then:

$$\langle L_u\rangle = \langle n_{u1}\rangle \left(\frac{\langle n_{u1}\rangle}{\langle n_{u1}\rangle + 1}\right)^{x_u - S_u} \prod_{k=2}^{T} \frac{1}{\langle n_{uk}\rangle + 1} \left(\frac{1}{1 - \frac{\langle n_{uk}\rangle}{\langle n_{uk}\rangle + 1} \frac{\langle n_{u1}\rangle + 1}{\langle n_{u1}\rangle}}\right) \qquad (5.68)$$

Appendix 5.B Derivation of Channel PDF given Prediction and Variance

In Section 5.3.2 the probability for the obtained effective capacity c_{urt} given a prediction is needed in order to calculate the expected loss. We derive the probability for each of the three cases (cf. Figure 5.2) and then add the resulting distributions to obtain the total probability distribution.

1. When $\hat{c}_{urt} \leq \bar{c}_{urt}$ the obtained capacity is $c_{urt} = \hat{c}_{urt}$. Because the distribution for the predicted capacity is symmetric and centered at the potential capacity $\bar{c}_{urt}$ we have

$$P_1(c_{urt}|I) = \frac{1}{2}\delta(c_{urt} - \hat{c}_{urt}) \qquad (5.69)$$

   where δ is the Dirac delta.

2. In the second interval, $\bar{c}_{urt} \leq \hat{c}_{urt} \leq c^*_{urt}$, we use the aforementioned linearly decreasing function in describing the obtained capacity:

$$c_{urt} = -\frac{1}{v-1}\hat{c}_{urt} + \frac{v}{v-1}\bar{c}_{urt} . \qquad (5.70)$$

Leaning on previous remarks we model the potential capacity as a Gaussian distribution according to $\bar{c}_{urt} \sim \mathcal{N}(\hat{c}_{urt}, \sigma^2_{urt})$. Using the result

$$x \sim \mathcal{N}(m\ ,\ \sigma^2) \Rightarrow ax + b \sim \mathcal{N}(am + b\ ,\ a^2\sigma^2) \tag{5.71}$$

and the relation (5.70) it is concluded that

$$\begin{aligned} c_{urt} &\sim \mathcal{N}\left(-\frac{1}{v-1}\hat{c}_{urt} + \frac{v}{v-1}\hat{c}_{urt}\ ,\ \left(\frac{v\sigma_{urt}}{v-1}\right)^2\right) &(5.72)\\ &= \mathcal{N}\left(\hat{c}_{urt}\ ,\ \left(\frac{v\sigma_{urt}}{v-1}\right)^2\right)\ . &(5.73)\end{aligned}$$

Notice that this distribution is attained only for the interval $0 \leq c_{urt} \leq \hat{c}_{urt}$.

3. In the third interval, $\hat{c}_{urt} \geq v\bar{c}_{urt}$ or equivalently $-\infty \leq \bar{c}_{urt} \leq \hat{c}_{urt}/v$, the obtained capacity is zero. The probability for this is

$$\begin{aligned} P_3(c_{urt}|I) &= \delta(c_{urt}) \int_{-\infty}^{\hat{c}_{urt}/v} P(\bar{c}_{urt}|I) d\bar{c}_{urt} = \\ &= \delta(c_{urt}) \int_{-\infty}^{\hat{c}_{urt}/v} \frac{1}{\sqrt{2\pi\sigma^2_{urt}}} \exp\left[-\frac{1}{2\sigma^2_{urt}}(\bar{c}_{urt} - \hat{c}_{urt})^2\right] d\bar{c}_{urt} \\ &= \delta(c_{urt}) \left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{(v-1)\hat{c}_{urt}}{v\sigma_{urt}\sqrt{2}}\right)\right)\ , &(5.74)\end{aligned}$$

where erf(x) is the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt\ . \tag{5.75}$$

The integral in (5.74) is solved by following the procedure in Appendix A.

Chapter 6

Implications of Limited Feedback for Scheduling and Adaptive Modulation – Throughput, Sensitivity, Fairness and A Way Out

WE have seen in the previous chapter that the combined use of scheduling and adaptive modulation promises substantial throughput gains in the downlinks of cellular communication systems.

Remember that the scheduling policy that maximizes system throughput is to transmit exclusively to the user that can receive at the highest rate at any particular time, provided that this user has at least as much data to send as his channel can support (Knopp and Humblet, 1995). In order to realize the potential throughput increase, we consider a system using adaptive modulation to set the transmission rate based on the signal-to-noise ratio (SNR) at the receiver and the required bit-error rate (BER). The receiver thus predicts its SNR for the next time slot to be scheduled, and determines the corresponding rate with which it can receive data. This rate is then quantized and fed back to the base station.

With adaptive modulation on each sub-carrier in an OFDM system, or on several antennas, the required amount of channel feedback may severely degrade the spectral efficiency of the total system. The gain in spectral efficiency from channel adaptation may even be less than the degradation due to the extra feedback information. In this chapter, therefore, we investigate the implications of quantizing the feedback information so as to maximize the expected downlink throughput in a cell where scheduling and adaptive modulation is employed. We study the performance

degradation, the sensitivity to quantization errors, and how fairness is affected due to reduced feedback.

It was shown by Johansson (2003), Florén et al. (2003), Gesbert and Alouini (2003) that the multiuser-diversity gain is not considerably reduced when channel feedback is limited, provided that accurate knowledge of the individual channel statistics of every user is at hand. Such aspects as how to realize these gains in practice, the sensitivity to sub-optimum quantizations, and the effects on fairness, were however not addressed in these references or in other works.

In Section 6.1 we discuss the case where the individual users' channel pdf's are known in detail and find the optimum number of bits to use for feedback as well as the corresponding quantization. The section ends with a discussion of the performance implications, which leads to the conclusion that although optimum performance would in theory be very high, an extreme sensitivity to correct quantizations may in practice lead to drastic throughput losses. Then, in Section 6.2 we discuss on-line adaptation of the quantizations as the channel conditions and the number of users vary. We show how rate levels can be optimized adaptively based on the relative frequencies with which the prior levels have been used. In Section 6.3, in the light of our findings in Section 6.1, we investigate an alternative scheduling and quantization procedure based on a simple modification of fixed access which we briefly mentioned in the concluding section of Chapter 5. The proposed scheduler guarantees a minimum inter-access time, and is therefore well suited for real-time services such as speech. It further generalizes straightforwardly to systems using multiple orthogonal channels, such as OFDM. In comparison to traditional Round Robin scheduling, the proposal is seen to yield substantial throughput gains without affecting fairness. Simulations verify that the proposed scheme overcomes the shortcomings of pure multiuser diversity with only a small throughput degradation.

6.1 Quantization for Maximum Expected Throughput

We consider adaptation of downlink transmission over a fading channel. A quantization scheme is used in which the mobile terminals predict their SNR, determine the corresponding attainable transmission rate, and send a quantized value of the rate to the base station.

In adaptive modulation, the problem of determining SNR thresholds where to switch from one modulation level to another under bit error rate constraints has been investigated in many works under different assumptions and with different optimization criteria (see e.g. Alamouti and Kallel, 1994, Chung and Goldsmith, 2001, Falahati et al., 2004, 2003, Wang et al., 2003b). In the present work we assume that the receiver calculates the appropriate rate (modulation level) accord-

ingly based on the predicted SNR and the desired BER.

The quantization scheme then works as follows. Let $\log_2(M+1)$ be the number of bits per time slot used for feedback, where M is the number of non-zero levels. Each bit pattern corresponds to one of $M+1$ allowed modulation and coding levels [payload bits/symbol] $q_0, q_1, \ldots, q_M$ (the q_m are in general rational numbers). If a user can receive q_m bits per symbol but not q_{m+1} bits per symbol (where we assume that $0 = q_0 < q_1 < q_2 \ldots < q_M$), the user sends the bit pattern corresponding to q_m to the base station. The base station will then transmit to the user who signalled the highest quantized rate, using that rate, in the next time slot, here assumed to consist of l symbols. We here assume that the thresholds $q_0 \ldots q_M$ are common to all users. The option to use individually adjusted thresholds will be discussed in Section 6.3.

We now consider how the M non-zero rate levels are chosen so that the expected throughput in the cell is maximized.

Let A_m denote the proposition that at least one user can receive at a rate r_u such that $r_u \geq q_m$. Let B_m denote the proposition that at least one user can transmit at a rate r_u such that $q_m \leq r_u \leq q_{m+1}$. Assuming that the transmitter sends exclusively to the user with the highest instantaneous transmission rate, the expected throughput $\langle x \rangle$ per transmitted symbol[1] can then be written as a function of the modulation and coding rates $\{q_m\} \equiv q_0 \ldots q_m$,

$$\langle x(\{q_m\}) \rangle = \sum_{m=1}^{M} q_m P(B_m \bar{A}_{m+1} | I) \tag{6.1}$$

$$= \sum_{m=1}^{M} q_m P(B_m | \bar{A}_{m+1} I) P(\bar{A}_{m+1} | I) \tag{6.2}$$

where $\bar{(\cdot)}$ means logical complement and I denotes any background information we might have that is relevant to the determination of the joint probability for B_m and $\bar{A}_{m+1}$. Note that $P(B_m \bar{A}_{m+1} | I)$ is the probability that there is at least one user which can receive at rate q_m but no user that can receive at rate q_{m+1} or higher.

If the number M of non-zero rates is fixed, then the optimal rates are obtained by maximizing (6.2) by adjusting the q_m, $m = 1 \ldots M$. If we also want to decide on the optimal number of rates, then we should maximize the expected throughput minus the number of bits required for feedback

$$J(\{q_m\}, M) = l\langle x \rangle - U \log_2(M+1) \tag{6.3}$$

[1] Throughput here is not defined as the number of correctly received bits, but as the number of received bits at the desired BER.

which describes the net expected throughput gained from using M non-zero rate levels (and thus $\log_2(M+1)$ bits for feedback per user), where U denotes the number of users, and l is the number of symbols that make up a time slot. The optimization of J is now over both q_m and M.

For notational convenience, we first derive an expression for $\langle x \rangle$ in the case where all users' rates are modelled by identical probability distributions. We then state the general result where users have different rate distributions.

Note that

$$P(\bar{A}_{m+1}|I) = \left(\int_0^{q_{m+1}} P(r_u|I) dr_u \right)^U \tag{6.4}$$

and

$$\begin{aligned} P(B_m|\bar{A}_{m+1}I) &= 1 - P(r_u < q_m | r_u < q_{m+1}, I)^U \\ &= 1 - (1 - P(r_u \geq q_m | r_u < q_{m+1}, I))^U \\ &= 1 - \left(1 - \frac{\int_{q_m}^{q_{m+1}} P(r_u|I) dr_u}{\int_0^{q_{m+1}} P(r_u|I) dr_u} \right)^U . \end{aligned} \tag{6.5}$$

In (6.5), the term within the parentheses describe the probability that a user has a rate lower than q_m conditional on the statement that no user (in particular, user u) has a higher rate than q_{m+1}[2].

Multiplying (6.4) and (6.5) we obtain the joint pdf

$$\begin{aligned} P(B_m \bar{A}_{m+1}|I) = & \left[1 - \left(1 - \frac{\int_{q_m}^{q_{m+1}} P(r_u|I) dr_u}{\int_0^{q_{m+1}} P(r_u|I) dr_u} \right)^U \right] \\ & \times \left(\int_0^{q_{m+1}} P(r_u|I) dr_u \right)^U \\ = & \left(\int_0^{q_{m+1}} P(r_u|I) dr_u \right)^U \\ & - \left(\int_0^{q_{m+1}} P(r_u|I) dr_u - \int_{q_m}^{q_{m+1}} P(r_u|I) dr_u \right)^U \\ = & \left(\int_0^{q_{m+1}} P(r_u|I) dr_u \right)^U - \left(\int_0^{q_m} P(r_u|I) dr_u \right)^U . \end{aligned} \tag{6.6}$$

[2]Note that the conditioning on $\bar{A}_{m+1}$ limits the possible outcomes to below q_{m+1} and leads to a re-normalization ensuring that the sum probability becomes unity within the range $0 \ldots q_{m+1}$.

From this it is easily seen that the joint pdf with non-identical distributions is

$$\begin{aligned} P(B_m\bar{A}_{m+1}|I) &= \prod_{u=1}^{U}\int_0^{q_{m+1}} P(r_u|I)dr_u \\ &- \prod_{u=1}^{U}\int_0^{q_m} P(r_u|I)dr_u \ , \end{aligned} \tag{6.7}$$

and the expected throughput can be written as

$$\begin{aligned} \langle x(\{q_m\})\rangle &= \sum_{m=1}^{M} q_m\left(\prod_{u=1}^{U}\int_0^{q_{m+1}} P(r_u|I)dr_u\right. \\ &- \left.\prod_{u=1}^{U}\int_0^{q_m} P(r_u|I)dr_u\right) . \end{aligned} \tag{6.8}$$

Thus, the optimal rates $\{q_m\}$ for a fixed M can be found by maximizing (6.8).

Maximizing $J(\{q_m\}, M)$ in (6.3) by adjusting M and $\{q_m\}$ simultaneously yields the optimal expected total net throughput increase that can be obtained by multiuser diversity and rate adaptation taking the feedback rate into account. The maximization generally requires numerical methods. Note that $J(\{q_m\}, M)$ is valid for all $M > 0$, which covers all practical cases since for $M = 0$ the receiver cannot even tell the transmitter that it has access to a channel.

Note further that the method presented here can also be used to analyze a given quantization by calculating the ratio of the expected throughput obtained with the given quantization and the optimum expected throughput with the same number of feedback bits for a certain number of users and channels. For any arbitrarily chosen quantization $\{q_m\}$, we define this ratio as the quantization efficiency, $\kappa(U, M, \{P(r_u \mid I)\})$,

$$\kappa(U, M, \{P(r_u \mid I)\}) = \frac{\langle x(\{q_m\})\rangle}{\langle x^*\rangle} \tag{6.9}$$

where $\langle x^*\rangle$ denotes the expected throughput with optimum thresholds. We take κ as a measure of how efficient a given quantization is.

6.1.1 Implications

Consider the case of a 1-bit quantization under the assumption that all users have identical but independent rate distributions. In this case the expected throughput (6.8) simplifies to

$$\langle x\rangle = q\left(1 - P(r_u < q \mid I)^U\right) \ , \tag{6.10}$$

where q is the single non-zero modulation and coding rate.

We can draw some interesting conclusions about the behavior of a throughput-maximizing policy already from (6.10). The probability that there is at least one user who can receive with an arbitrary rate q is

$$P(q \mid I) = 1 - P(r_u < q \mid I)^U \ . \tag{6.11}$$

Now, assume that a user's transmission rate r_u can be modelled by the relation

$$r_u = \log_2 \left(1 + \frac{\text{SNR}_u}{\Gamma_u} \right) \ , \tag{6.12}$$

where r_u is the transmission rate[3] of the uth user that attains the prescribed BER at SNR $=$ SNR$_u$, SNR$_u$ is the predicted SNR at the receiver of user u, and Γ_u is a system-specific value which depends on the desired BER and the type of modulation and coding used. For instance, (6.12) is a good approximation of the attainable rate using Gray-coded M-QAM modulation (Chung and Goldsmith, 2001) with

$$\Gamma_u = -\frac{\ln(5\text{BER}_u)}{1.6} \ . \tag{6.13}$$

Under the assumption that the SNR pdf for each user is exponential (corresponding to the case of a Rayleigh fading channel) with known mean $\langle \text{SNR}_u \rangle$,

$$P(\text{SNR}_u \mid I) = \frac{1}{\langle \text{SNR}_u \rangle} \exp \left\{ -\frac{\text{SNR}_u}{\langle \text{SNR}_u \rangle} \right\} \ , \tag{6.14}$$

and that the relation between SNR and rate is given by (6.12), the rate pdf $P(r_u \mid I)$ for each user is obtained by a variable transformation:

$$\begin{aligned} P(r_u \mid I) &= P(\text{SNR}_u \mid I) \left| \frac{d\text{SNR}_u}{dr_u} \right| \\ &= P(\text{SNR}_u \mid I) \Gamma_u 2^{r_u} \ln 2 \\ &= \frac{\Gamma_u 2^{r_u} \ln 2}{\langle \text{SNR}_u \rangle} \exp \left\{ -\frac{\Gamma_u \left(2^{r_u} - 1\right)}{\langle \text{SNR}_u \rangle} \right\} \ . \end{aligned} \tag{6.15}$$

From (6.12) we have that $\text{SNR}_u = \Gamma_u(2^{r_u} - 1)$ and consequently $\frac{d\text{SNR}_u}{dr_u} = \Gamma_u 2^{r_u} \ln 2$. The probability that a user can receive at a rate in the interval $q_1 <$

[3]Here we treat r_u as a continuous variable; in practice it should be rounded off to the nearest smaller (rational) number specified by the modulation-coding scheme.

$r_u < q_2$ is then

$$\begin{aligned} P(q_1 < r_u < q_2 \mid I) &= \int_{q_1}^{q_2} P(r_u \mid I) dr_u \\ &= \int_{q_1}^{q_2} \frac{\Gamma_u 2^{r_u} \ln 2}{\langle \text{SNR}_u \rangle} \exp\left\{ -\frac{\Gamma_u \left(2^{r_u} - 1\right)}{\langle \text{SNR}_u \rangle} \right\} dr_u \\ &= \exp\left\{ -\frac{\Gamma_u \left(2^{q_1} - 1\right)}{\langle \text{SNR}_u \rangle} \right\} - \exp\left\{ -\frac{\Gamma_u \left(2^{q_2} - 1\right)}{\langle \text{SNR}_u \rangle} \right\} . \end{aligned} \tag{6.16}$$

With $q_1 = 0$ as in (6.11), (6.16) becomes

$$P(r_u < q \mid I) = 1 - \exp\left\{ -\frac{\Gamma_u \left(2^{q} - 1\right)}{\langle \text{SNR}_u \rangle} \right\} . \tag{6.17}$$

We can easily find the throughput-maximizing value of q, by inserting (6.17) in (6.10) and finding the integer q which maximizes (6.10) for a given number of users U. For $U = 30$, with mean individual SNR $\langle \text{SNR}_u \rangle = 13$ dB and Gray-coded M-QAM with a desired BER of 10^{-3}, i.e. Γ_u determined by (6.13), we find the optimum to be $q = 4$, yielding an expected throughput of $\langle x \rangle = 3.71$ bits per symbol[4]. With perfect channel information at the transmitter (i.e. without quantization) and adaptive modulation supporting any integer positive rate, the expected throughput

$$\langle x \rangle = \sum_{k=0}^{\infty} k \left(\left(\int_0^{k+1} P(r_u|I) dr_u \right)^U - \left(\int_0^{k} P(r_u|I) dr_u \right)^U \right)$$

becomes 4.09 bits per symbol. The performance drop by going from unlimited resolution to a 1-bit quantization is thus only 10%!

Compare this to the case of using a traditional fixed-access scheme, in which users transmit in the same order regardless of channel quality. Then multiuser diversity is completely lost, and, under the same assumptions as above, the expected throughput with perfect channel knowledge becomes $\langle x \rangle = \langle r_u \rangle = \int r_u P(r_u \mid I) dr_u \approx 2.35$ bits per symbol. With a 1-bit quantization, the optimally adjusted q for maximum expected throughput is determined from (6.10) with $U = 1$. The result is $q = 2$, yielding an expected throughput of $\langle x \rangle = 1.22$. Evidently, with fixed access the expected throughput is approximately halved, from 2.35 bits per symbol to 1.22 bits per symbol, by a 1-bit quantization as compared with perfect channel knowledge. Hence, with regard to optimum throughput, it is clear that multiuser diversity-driven systems do not suffer at all as badly from reduced feedback as does the traditional fixed-access scheme.

[4]Remember that each user who is allowed to transmit will use 4 bits/symbol; even if the corresponding channel could support more than this, there is no way for the receiver to inform the transmitter about that.

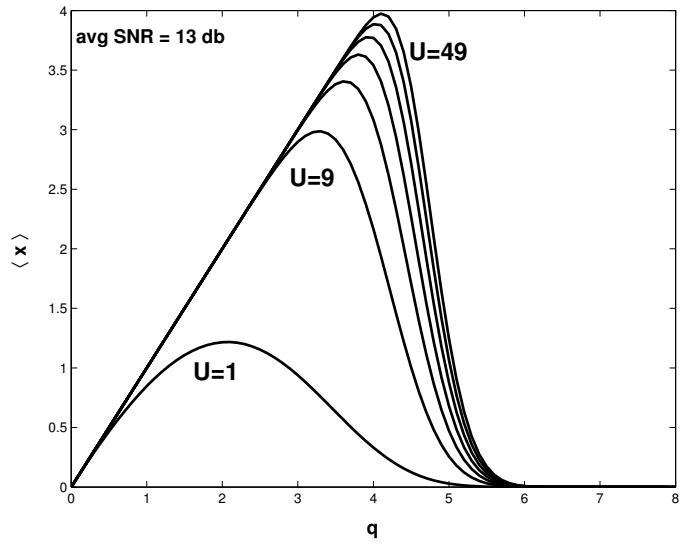

Figure 6.1: The expected throughput [bits/symbol] as a function of the used modulation level q for different number of users. Each curve corresponds to an increase of 8 users from the curve below. The average SNR of each user was 13 dB and Gray-coded M-QAM was used with a desired bit-error rate of 10^{-3}.

Let us now discuss the sensitivity to erroneously set rate thresholds q. Consider again a system employing pure multiuser diversity; at each time slot the user with the highest current rate is served. With a large number of users, the probability distribution for the rate that will be used may become extremely sharp[5]; up until a certain level there will be almost probability 1 that someone can receive at that rate, but then it suddenly drops down to zero. This drop will be extremely steep, as illustrated in Figure 6.1 where the expected throughput is plotted as a function of the chosen level q for different number of users. For instance, consider the same scenario as in the preceding paragraph. Then the expected throughput with $q = 4$ is 3.71 bits per symbol. Increasing the threshold to $q = 5$ however yields an expected throughput of only 0.81 bits per symbol, a most dramatic performance decrease! The probability for being able to transmit at a particular rate is almost certainty; just adding one bit to that rate leads to a probability for transmission of only 16%. The expected throughput decreases by a factor of 4.56 if the selected threshold changes by a factor of only $1/4$. The throughput degrades to below what can be expected from using fixed access!

[5]In particular, this happens when all users have the same mean SNR, e.g. due to slow power control.

In practice, the base station has very little information regarding individual channels and is therefore in the unenviable position of realizing the risk for potential performance breakdown (to a level well below that of ordinary fixed access) but having no information as to ensure its avoidance.

Moreover, since a correctly chosen threshold q will rely heavily on the upper tails of the individual rate distributions, there is a large risk that the throughput-maximizing q will be set so high that only a very small number of users will ever be able to receive at that rate. Consider for example a case in which the mean SNRs of different users range from, say, $6 - 30$ dB according to distance from the base station. The optimum rate threshold will depend almost entirely on the channels representative for the users near the base station, while the border users will be completely shut off. Typically, the upper tail of the distribution for attainable rates is dominated by just one or a few users. With more than 1-bit feedback, some thresholds would be set rather low as to always guarantee some throughput, but with only one bit for feedback, the threshold q will be set much higher. The problem with unfairness will consequently become more pronounced as the amount of feedback is reduced.

In summary, there are apparent risks associated with using a pure multiuser-diversity strategy, but on the other hand the system throughput may become very large if the situations which cause the extreme sensitivity are unlikely to occur in practice.

There are evidently two ways of tackling these problems. One way is to find a robust mechanism for determining the optimal rate thresholds adaptively as the channels and the number of users vary. Another way is to modify the scheduling policy in some way as to ensure a larger degree of fairness and/or a smaller sensitivity to quantization errors. In the next section we study the former alternative, and in Section 6.3 we investigate the latter.

6.2 Feedback Adaptation

We now assume that there is only one non-zero rate threshold q, i.e. that we use a 1-bit quantization. In Section 6.4.3 we show that in a single-channel system, using 1-bit feedback often results in a larger net throughput gain (taking into account the required feedback channel's bandwidth) than using several bits.

Assume that we have the possibility of changing the rate threshold q on-line at certain intervals. This requires that the transmitter has the possibility to broadcast updated rate levels to the receivers, thus incurring some extra signalling in the downlink. The transmitter can then tune the rate levels based on how often the different current levels are used. Further, assume that the transmitter can transmit

at any integer rate (or at any rate from a discrete set of rational numbers) below or at the maximum of r_{max} bits per symbol.

The average throughput per symbol, $\bar{x}$, over an arbitrary time interval given the current modulation level q, $0 \leq q \leq r_{max}$, can be modelled as

$$\bar{x} = \rho(q)q + e \tag{6.19}$$

where $\rho(q)$ is an unknown decreasing function of q defined on the interval $0 \leq q \leq r_{max}$, and e is any outstanding variation not explained by $\rho(q)$. The function $\rho(q)$ must further satisfy the evident property that the average throughput is non-negative and not larger than the used modulation level, i.e.

$$0 \leq \rho(q) \leq 1 \ , \ \forall q \ . \tag{6.20}$$

The true non-linear relation between the used modulation level and the corresponding throughput varies with the number of users and the properties of the individual users' channels. We shall thus have to content ourselves with choosing a function $\rho(q)$ containing adjustable parameters that allow us to adapt the function to the data at hand in any given situation. The function should be flexible enough to fit different data sets, and have as few parameters as possible. Two one-parameter functions suggest themselves: a straight line parameterized by its slope, and an exponential parameterized by its exponent. The former alternative is clearly inadequate; to be useful it would need an adjustable intercept, and even with one it could not model the typical behavior with a relatively flat region with $\rho(q)$ near 1 followed by a sharp knee at some critical value of q where the throughput suddenly drops and then remains nearly 0 (cf. the examples in Section 6.1.1). The latter alternative is not much better; an exponential can clearly not model the first flat region, and would need to be augmented with some modification in this region.

In the light of these two examples, it is seen that a two-parameter function would be more suitable. A model which captures the typical behavior of the throughput – a flat region with $\rho(q)$ nearly 1, then a knee, followed by a new flat region with $\rho(q)$ nearly 0 – while satisfying the quantitative requirements is

$$\rho(q) = \frac{1}{2}\text{erfc}\left(\frac{q-\mu}{\sigma}\right) \tag{6.21}$$

where μ and σ are adjustable parameters, determining the location and the sharpness of the knee respectively, and $\text{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}}\int_0^x \exp\left(-t^2\right) dt$ is the complementary error function. A plot of the model function $\rho(q)$ is given in Figure 6.2.

With knowledge of previously used modulation levels q and their corresponding average throughput per symbol $\bar{x}$ (obtained from knowledge of the number of

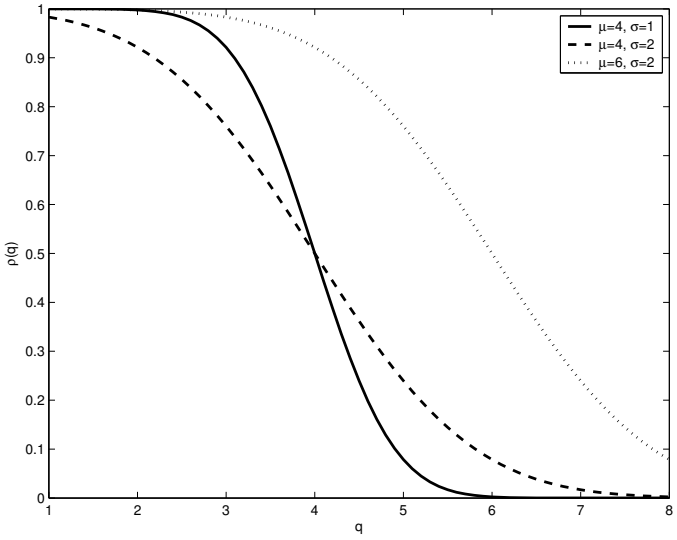

Figure 6.2: The model function $\rho(q) = \frac{1}{2}\text{erfc}\left(\frac{q-\mu}{\sigma}\right)$ used in the non-linear regression (6.19) describing the relation between the used modulation level q and the corresponding average throughput.

times the modulation level could be used in the previous time interval), we can estimate the parameters of the non-linear regression, μ and σ.

The joint posterior distribution for μ and σ is

$$P(\mu, \sigma \mid D, I) \propto P(D \mid \mu, \sigma, I) P(\mu, \sigma \mid I) \,, \tag{6.22}$$

where D denotes the observed input-output pairs, ($\bar{x}$, q), under the M most recent updating intervals. Using a Gaussian model for the error term e in (6.19), and assuming that the parameters in $\rho(q)$ has not changed significantly during the N most recent updating intervals, the likelihood at time t becomes

$$P(D \mid \mu, \sigma, I) \propto exp\left\{ -\frac{1}{2\delta^2} \sum_{k=t-N+1}^{t} (\rho(q_k) q_k - \bar{x}_k)^2 \right\} \,, \tag{6.23}$$

with δ^2 denoting a constant variance for the e distribution. We shall take the priors for μ and σ to be independent and uniform in small intervals, $1 \leq \mu \leq \mu_{max}$, and $0.5 \leq \sigma \leq \sigma_{max}$.

The parameters are thus found by maximizing the likelihood in the constrained parameter space, $1 \leq \mu \leq \mu_{max}$, and $0.5 \leq \sigma \leq \sigma_{max}$, or equivalently minimizing

the corresponding log likelihood,

$$(\mu, \sigma) = \arg\min_{\mu,\sigma} \sum_{k=t-N+1}^{t} \left(\rho(q_k)q_k - \bar{x}_k\right)^2 . \tag{6.24}$$

The minimization is carried out in a numerical search, e.g. using the Nelder-Mead simplex algorithm (Nelder and Mead, 1965).

Having estimated μ and σ, we shall use the modulation level q which maximizes (6.19) using the estimated parameter values. This is not exactly equivalent to maximizing the expected throughput, which would be obtained by averaging (6.19) over the joint posterior probability distribution for μ and σ. If μ and σ are reasonably well-determined the difference is however small.

We should also observe another important point; maximizing the (approximate) expected throughput for the next time interval may result in a succession of choices of the same modulation level q if the size of the user population remains approximately constant. In the worst case we would try to estimate μ and σ based only on one value of q. Obviously, in such a case, the joint likelihood for the parameters becomes flat, and the accuracy of the estimate is poor. A better strategy would be to maximize the expected throughput over a longer time horizon, which results in a balance between short-run performance and information gathering (see e.g. Zellner, 1971). Such a policy is however not analytically tractable. Instead, a simple constraint on the size of consecutive changes will be used: never allow the modulation level to increase or decrease by more than one level at a time. Then, in cases where the same level has been used for quite a while and the uncertainty concerning μ and σ becomes large, we are still guaranteed that a reasonable choice is made. The mere fact that the same level has been used for a long while indicates that a large sudden change is unlikely to be correct.

6.3 Diversity-Enhanced Equal Access – Rate Quantization and Scheduling with Fairness

We have seen that a disadvantage of using scheduling with the sole objective of maximizing throughput is that it may lead to an uneven distribution of transmissions. Some users may be completely shut off from transmission for long time periods. In systems where all users are guaranteed a certain time of access to the network, this should be avoided. One way of reducing the risk of uneven resource distributions is to use scheduling with other objectives than pure throughput maximization, e.g. by including user-specific priorities that depend on past channel accesses, bit rate requirements, payment options, etc.

We here propose the use of a simple method which attains both some multiuser-diversity gain and a fixed maximum inter-access delay. The method, which was briefly introduced in the concluding section of Chapter 5, consists of transmitting at each time slot to the user who can transmit the largest number of bits in that time slot. In the next time slot the procedure is repeated, but now only the remaining users are allowed to compete for channel access. After U time slots, all users have thus accessed the channel, and the process is restarted. This method guarantees a maximum inter-access time of $2U - 1$ time slots.

At the first time slot, the proposed policy employs a pure multiuser-diversity strategy for U users; in the second slot it does so again but only among $U - 1$ users, and so on. Thus, over a period of U time slots the policy can be interpreted as taking full advantage of multiuser diversity among a number of users that is decreasing by one for every time slot. We would then expect that in terms of throughput the policy would on average achieve full multiuser diversity gain for a system of approximately $U/2$ users[6]. This is the price that is paid by guaranteeing equal access. It can however be observed that the multiuser diversity gain increases more slowly the larger U becomes (c.f. Figure 5.5). Thus, with many users in the system, the gain obtained with the proposed policy will not be far from that of the maximum throughput strategy.

We now consider the problem of determining a good quantization for this modification of round-robin scheduling. We again restrict our study to the 1-bit feedback case, a choice which is further discussed in Section 6.4.3. The generalization to several bits is possible; it follows from modifying (6.29) analogously to the general expression (6.8) for strict multiuser diversity .

The scheme consists of determining an individual quantization q_u for each user. Each receiver calculates its quantization based on channel measurements over the last N time slots and updates the base station every Sth time slot with a new quantization. (Thus a small extra amount of feedback is used in addition to the bit transmitted each slot. The updates will however be made infrequently and the number of bits required for feeding back the new threshold can be made small, e.g. 1 or 2 bits.) The base station keeps a table with a record for each user containing the user-specific transmission rates q_u, and the record gets updated with a period of S time slots. (S and N does not have to be equal.) Then, for each time slot, each user u sends 1-bit feedback to the base station indicating whether it supports the rate q_u or not. When a user u sends a 1 to the base station, it means that it can receive at rate q_u in the next time slot.

[6]It should be observed that characterizing multiuser-diversity gain only as a function of the number of users requires that all users have identical and independent rate distributions. In general, the diversity gain should be characterized as the ratio of the expected throughput of the scheduling policy and that of round-robin scheduling.

An advantage of this scheme compared to the ones studied in Section 6.1 and Section 6.2 is that the individual thresholds will be determined locally by the mobile terminals, where more channel information is available for the calculation.

Assume that the system can transmit at any of K different non-zero rates, $r_1 \dots r_K$. Let them be ordered by increasing size, so that $0 < r_1 < r_2 \dots < r_K$. Over a period of the N most recent time slots, each receiver measures the SNR every time slot and keeps a record of the number of slots, n_i, that it was possible to receive at the rate r_i but not at the higher rate r_{i+1}. The probability that the channel can support rate r_i on a future time slot is then calculated based on the N most recent measurements. Assuming that the channel is 'stationary'[7] during the last N plus the next S time slots, the probability, p_i, that the channel supports rate r_i at a future time slot is

$$p_i = \frac{n_i + 1}{N + K} \,, \tag{6.25}$$

which is the same general version of Laplace's rule of succession with K possible outcomes that we derived and used in Section 2.6 (the same derivation with many interesting historical comments can be found in (Jaynes, 2003), Ch. 18).

Consider the determination of the rate threshold q_u for a particular user u. A simple approach would be to maximize

$$q_u P(r_u > q_u \mid I) \,, \tag{6.26}$$

but note that this expression does not take into account the fact that a user on average competes for access over more than one time slot. In effect, the expression does not take full advantage of the multiuser diversity that is utilized by the proposed scheduling policy. If the user would know the number of slots, n_u ($1 \leq n_u \leq U$), that this user has the highest rate of all users, then he should use the q_u that maximizes his expected throughput per received symbol in that time slot that u obtains access,

$$\langle x_u \rangle = q_u \left(1 - P(r_u < q_u \mid I)^{n_u}\right) \,, \tag{6.27}$$

where $1 - P(r_u < q_u \mid I)^{n_u}$ is the probability that the rate r_u is larger than q_u at least one out of n_u time slots[8].

In practice however, n_u is unknown and we must assign a probability for n_u which represents our uncertainty concerning its actual value. The expected

[7] When we say that a channel is stationary over a certain time we mean simply that the causal processes underlying the main channel variations (i.e. the geography and the velocity of the receiver) do not change significantly over that time period.

[8] A useful analogy is to consider the probability for obtaining at least one 5, say, or higher when throwing a regular die a number of times. As the number of trials increase, the probability increases correspondingly.

throughput is then obtained by multiplying (6.27) by $P(n_u \mid I)$ and then integrating out n_u as a nuisance parameter.

As no value of n_u within the range $1 \ldots U$ is more likely than any other the principle of indifference applies, and we assign a uniform probability distribution to n_u:

$$P(n_u \mid I) = \frac{1}{U} \,. \tag{6.28}$$

The expected throughput with unknown n_u thus becomes

$$\begin{aligned} \langle x_u \rangle &= q_u \frac{1}{U} \sum_{n_u=1}^{U} \left(1 - P(r_u < q_u \mid I)^{n_u}\right) \\ &= q_u \left(1 - \frac{1}{U} \sum_{n_u=1}^{U} P(r_u < q_u \mid I)^{n_u}\right) \\ &= q_u \left(1 - \frac{P(r_u < q_u \mid I)^{U+1} - P(r_u < q_u \mid I)}{U\left(P(r_u < q_u \mid I) - 1\right)}\right) \,, \end{aligned} \tag{6.29}$$

where the sum on the second line is a geometric progression (eqn. 0.112 in Gradshteyn and Ryzhik, 2000) which yields the final equality. Note that $P(r_u < q_u \mid I)$ is determined from

$$P(r_u < q_u \mid I) = \sum_{r_u < q_u} P(r_u \mid I) \,, \tag{6.30}$$

where $P(r_u \mid I)$ is the probability distribution (6.25) for the individual rates r_u.

Each mobile terminal thus selects the q_u which maximizes its expected throughput per received symbol (6.29). The maximum is found by a one-dimensional numerical search over K integers with very low computational demands.

Notice that the rate probabilities and the rate thresholds are based on the N most recent channel measurements. The number of time slots to use for channel measurements, N, is consequently of importance. Typically, N and S would be chosen as the same number of slots, and the number should be large enough to cover a number of fading dips and highs, i.e. N should be on the time scale of shadow fading rather than on that of fast fading.

The proposed scheduling and quantization policy can straightforwardly be used in a system with multiple orthogonal channels, such as OFDM. Now, the service guarantee requires that each user obtain one channel access on each channel over a time span of U time slots. The scheduler is then run in parallel on each channel, and each user has a single rate threshold that is used on all channels.

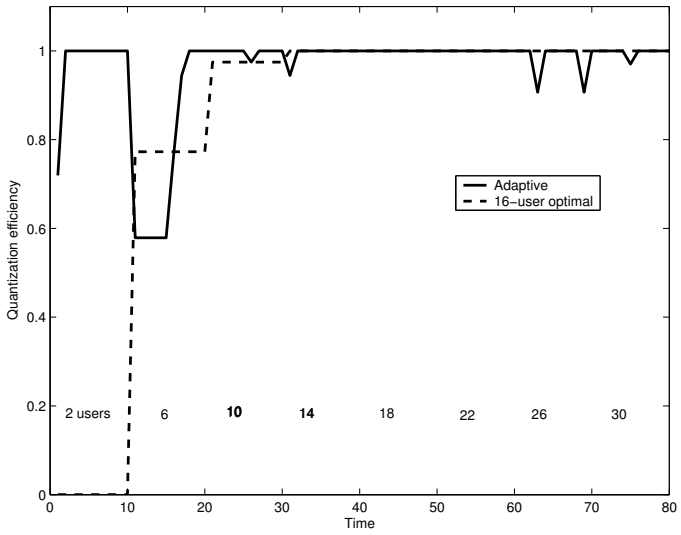

Figure 6.3: The quantization efficiency of a 1-bit adaptation using the procedure in Section 6.2 in a system where all users have exponentially distributed SNR with uniformly distributed mean SNR. The dashed line corresponds to a quantization optimized for the 16*-user case with perfect knowledge of channel parameters.*

6.4 Examples and Simulations

6.4.1 On-line adaptation

It is clear that the on-line adjustment procedure given in Section 6.2 will not in general give as good performance as the optimal procedure in Section 6.1 with detailed knowledge of the number of users and their individual channel pdf's. We therefore need to analyze the performance in a controlled experiment where one can compare the evolution of the adaptive solution and see whether it converges to the better informed solution.

Letting the number of users increase from $U = 2$ to $U = 30$ by additions of 4 users every 10th time the rate level was updated, we tested the adaptive quantizer on a population in which each user had exponentially distributed SNR with average SNR generated from a uniform distribution between 0 and 11.76, i.e below 15 dB. The Rayleigh distributions for the channel gains were independent among users with no correlation between adjacent time slots. Only integer rate levels in the range $[1...8]$ were allowed. The estimation of μ and σ was based on the $M = 5$ most recent pairs of q, and $\bar{x}$ was obtained from $\bar{x} = fq$ where f is the proportion of time slots that a receiver signaled the possibility of receiving at that rate.

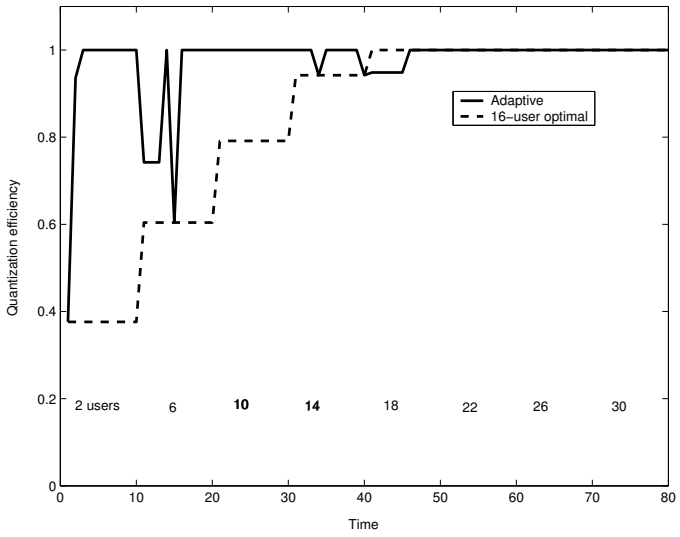

Figure 6.4: The quantization efficiency of a 1-bit adaptation using the procedure in Section 6.2 in a system where all users have exponentially distributed SNR with the same mean SNR. The dashed line corresponds to a 16*-user optimal quantization with perfect knowledge of channel parameters.*

The test was run by choosing an initial quantization level $q_{init} = r_{max}/2 = 4$ and generating the frequencies with which rate q_{init} would be used according to (6.6) and the rate distribution (6.15) derived from the exponential SNR distribution. The gap factor in the rate-SNR relation was set to $\Gamma = 2$. The rate level updates were assumed to take place with long intervals under stationary conditions, so that the usage frequencies f were equal to the rate probabilities used by the random number generator in the simulation of the next time slot. In order to be able to observe any oscillations or slow convergence in the adaptation law, 10 rate level updates were carried out in succession before increasing the number of users.

The quantization efficiency (6.9), i.e. the ratio of the expected throughput with the best q possible and the expected throughput with the level determined from adaptation, is shown in Figure 6.3. For comparison, the efficiency of the optimal 16-user quantization is also presented. This gives the highest attainable performance for 16 users, but is an unrealistic ideal as it assumes knowledge of all channel statistics. We however show this to see how close the adaptive solution comes to the optimum, and also to see how the 16-user optimum performs for other population sizes. In this simulation, the first two users had relatively low mean SNR, which results in zero throughput for the 16-user optimal case. The adaptive scheme, on the other hand, generally achieves a high quantization efficiency. For

the 6-user case we see that the convergence is somewhat slow. The reason is that the two first users happened to experience bad channels; when increasing from 2 to 6 users, the optimal q increased from 3 to 6 bits per symbol.

In Section 6.1.1 we noted that the pure multiuser-diversity strategy may sometimes lead to drastic performance drops when the users have identical independent rate distributions and the details of the distributions are unknown. The proposed adaptation law is bound to suffer from this risk. In order to quantify what may happen in an extreme case, a simulation was set up in which all users had exponentially distributed SNR with mean SNR 13 dB, yielding the rate distribution (6.15) with Γ given by (6.13) and desired BER 10^{-3}. Figure 6.4 shows the quantization efficiency (6.9) of the adaptation and a 16-user optimal quantization as the number of users was gradually increased. It can be seen that the adaptive quantizer actually avoids using a too large q when U is large (as otherwise the quantization efficiency would be extremely low). In the case of few users, there are occasional mistakes, but the overall performance is very high. Obviously, in some cases the proposed adaptive quantizer will occasionally try a too large q with an inevitable performance loss. This difficulty is inherent in multiuser diversity due to its extreme sensitivity for too large thresholds. Any adaptation mechanism must try to explore possible improvements from increasing q and thus balance this with the risk of performance loss.

6.4.2 Diversity-Enhanced Equal Access

In this section we aim to investigate to what extent the scheme proposed in Section 6.3 does indeed overcome the problems of fixed access and those associated with the pure multiuser-diversity policy.

A set of simulations was carried out in which 16 users were spread out uniformly over the cell radius[9], and where each individual user's SNR was exponentially distributed with a fixed mean proportional to d^{-2} where d is the distance to the base station. The proportionality constant was chosen so that the mean SNR of the 16 users ranged from 30 dB down to 6 dB. The rate-SNR relation (6.12) was used with $\Gamma_u = 2$.

In order to test the scheduling and quantization policy under the circumstances that it was designed for, the rate distributions were assumed to be stationary and the system assumed to have been started in an infinite past (ensuring that the probabilities $P(r_u \mid I)$ were set 'correctly' for all users). The simulation was run for 1600 time slots, each consisting of only one symbol, and the reported results are aver-

[9]Note that this set-up is not equivalent to a uniform user distribution over the cell area, but was chosen for simplicity. The results are however representative also for other user distributions, as briefly mentioned in the end of the section.

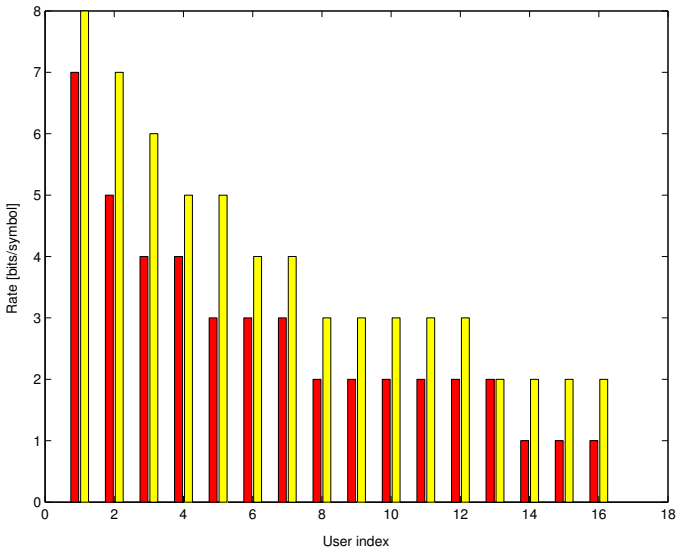

Figure 6.5: The optimized rate thresholds for 16 *users having exponentially distributed SNR with mean SNR ranging from* $30 - 6$ *dB. The users are ordered by decreasing mean SNR. The dark color refers to the optimum fixed-access thresholds, while the light color refers to the optimum thresholds using (6.29).*

ages from 100 simulation runs. In order to make a fair comparison, the throughput was set to zero in time slots when none of the remaining users could transmit at their rate threshold. In reality, one would obviously choose to transmit to another user who has already received service in such cases[10].

Figure 6.5 shows the rate thresholds obtained from maximizing (6.29) and the thresholds obtained from maximizing (6.26), i.e. the optimum quantization for a fixed access scheme that does not utilize multiuser diversity. The general tendency in using (6.29) is, as expected, to set the levels somewhat higher since a user typically competes for more than one time slot, thereby increasing his chances for obtaining a higher rate at least once in the U slots.

In Figure 6.6, the total individual throughput obtained from using the proposed scheduling and quantization policy is plotted and compared to the throughput obtained by using the same scheduling policy but with the rate thresholds obtained from (6.26). It can be seen that almost every user obtains increased throughput by choosing the more aggressive quantization strategy. The total throughput summed over all users increases by approximately 27% by using the higher rate thresholds.

[10] With such a mechanism, the proposed scheme would have an even bigger performance advantage than the present simulations suggest.

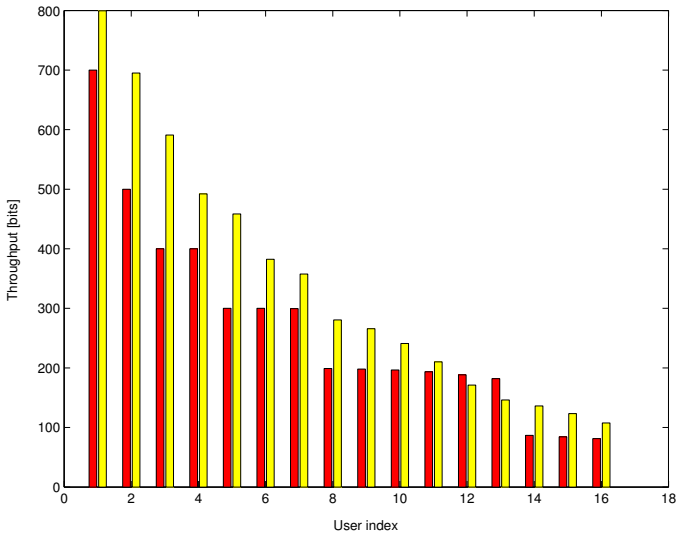

Figure 6.6: The obtained individual throughput for 16 *users using rate thresholds from Figure 6.5. The users are ordered by decreasing mean SNR. The dark color refers to the optimum fixed-access thresholds, while the light color refers to the optimum thresholds using (6.29).*

Figure 6.7 shows the individual throughput for each user using the proposed 1-bit quantization compared to the maximum attainable when having unquantized channel knowledge and a continuum of possible rates. We see that the performance drop is larger for the users with low average SNR. The total throughput increase of the unquantized case is 24%.

In order to see how the use of individual thresholds affect the performance, we also tested using a common quantization level optimized for the median user. With individually optimized thresholds using (6.29), the throughput increase was approximately 80% compared to this case.

The multiuser-diversity gain was quantified by comparing the obtained throughput to a fixed-access schedule with a common rate threshold optimized for the median user. The throughput increase was now 168%. In comparison to a fixed-access scheme with individually and for fixed-access optimally adjusted rate thresholds, the throughput increase was 90%.

Under somewhat different channel assumptions, with $U = 16$ users having identical but independent rate distributions (6.15) with mean SNR 15 dB and $\Gamma_u = 2$, the performance gain of using the proposed scheduling and quantization policy was about 25% as compared to using the same scheduling policy but with the

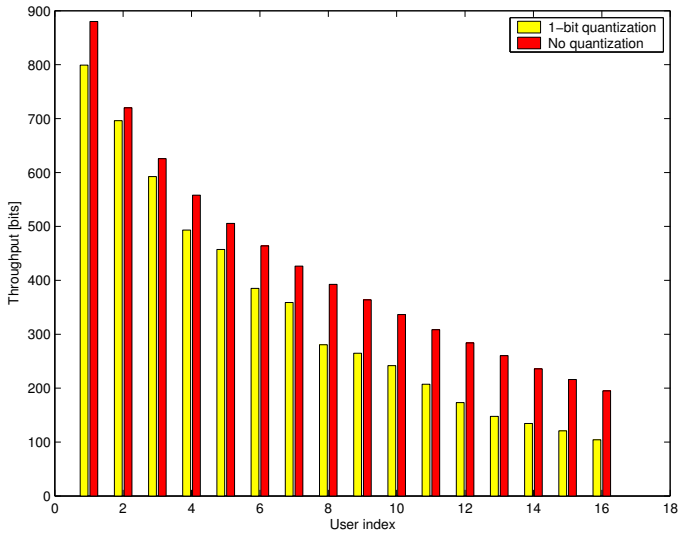

Figure 6.7: The obtained individual throughput for 16 *users using rate thresholds from Figure 6.5 in light color, and in dark color that of using unquantized and untruncated rates (i.e. assuming a continuum of available rates) in the same setting. The users are ordered by decreasing mean SNR.*

rate thresholds obtained from (6.26). The optimum rate thresholds were found to be $q_u = 4$ for all users. In this scenario, it is possible to determine how much throughput is lost by using the proposed scheme in comparison to using a pure multiuser-diversity strategy. A numerical search found the optimum common rate threshold[11] for pure multiuser-diversity to be $q = 5$. In order to carry out a fair comparison between the two approaches we let our proposed policy be augmented by a mechanism for avoiding transmitting zero bits in the time slots when none of the remaining users can reach their rate threshold. In such time slots, the policy instead transmits to an arbitrarily chosen user with non-zero rate. The throughput increase from using the pure multiuser-diversity strategy with the optimum quantization is then just below 25%, as expected.

In the previous section we conjectured that the proposed scheduling and quantization policy would be roughly equivalent to a pure multiuser-diversity strategy with $U/2$ users. With 8 users, the optimum q for pure multiuser diversity in the current simulation scenario is $q = 4$, which is also the individual optimum for the proposed policy for 16 users. As predicted, there is no throughput difference.

[11] Note that in this case, since all users have identical independent rate distributions, nothing would be gained by having individual rate thresholds. This applies to both strategies.

6.4.3 The number of feedback bits

After introducing the M-level selection problem, we have focused on the case with 1-bit feedback. Here, we show that this is indeed a proper choice in many cases when strict multiuser diversity is used. The optimum number of levels is obtained by maximizing the net throughput gain (6.3) with respect to M and $\{q_m\}$.

Denote the expected downlink throughput per symbol for an arbitrary choice of $\log_2(M+1)$ $\langle x \rangle_{\log_2(M+1)}$. An increase from 1 bit of feedback to 2 bits is then worthwhile according to (6.3) only if

$$\begin{aligned} l\langle x\rangle_2 - 2U &> l\langle x\rangle_1 - U \\ \Leftrightarrow \langle x\rangle_2 - \langle x\rangle_1 &> \frac{U}{l}\,, \end{aligned} \tag{6.31}$$

i.e. if the expected downlink throughput increases by at least U/l bits per symbol (where l is the number of symbols per time slot). Typically, l is chosen as the number of symbols that the channel is expected to be approximately constant, which e.g. depends on expected vehicle speeds and the channel bandwidth. If $l \approx U$, then it would be worthwhile to use 2 bits instead of 1 only if the expected downlink throughput increases by at least 1 bit per symbol. But by utilizing multiuser diversity we have seen that the throughput decrease by using only 1-bit feedback may be about 10% for large U as compared to unlimited resolution. Thus, in order to use 2 feedback bits instead of 1, we would at the very least require the expected downlink throughput for a 1-bit quantization to be a remarkable 10 bits per symbol.

Note that this however assumes the use of strict multiuser diversity and that the number of users is approximately the same as the slot length in symbols. It may however be advantageous to use more than 1 bit for feedback when the slot length l is very large compared to the number of users or when the throughput gain from increasing to 2 feedback bits is higher than the 10% assumed above. For instance, using a modified scheduler, such as the one proposed in Section 6.3, the simulations in Section 6.4.2 suggest that there is a larger gain from increasing to 2 bits as compared to strict multiuser diversity.

6.5 Conclusions

We have seen that in order to achieve a certain fraction of the potential performance gain from using adaptive modulation and coding, taking advantage of multiuser diversity leads to lower feedback requirements than using a fixed schedule. In addition, reducing the number of feedback bits does not affect throughput nearly as much as for the traditional single-user perspective. This illustrates why traditional

adaptive modulation with many modulation levels substantially increases the performance of non-scheduling based systems. With only one non-zero transmission rate (i.e. no adaptive modulation), the actual bit rate reduces roughly to half of what could be obtained without quantization. With scheduling based on channel quality, we have the advantage of higher possible throughput as evidenced by the scheduling gain of the unquantized case, and just as importantly, less degradation from the unquantized theoretical throughput due to limited amounts of feedback.

However, the theoretical advantages of multiuser diversity were seen to suffer from two distinct difficulties. First, unfairness generally increases when the number of feedback bits is reduced and users have different rate distributions. Second, the theoretical throughput advantage has a critical proviso; the rate level must not be set too high. It was seen that if the level is chosen just one bit over the optimal value, in cases where users have identical and independent rate distributions, the throughput may drop to below that of fixed access. The risk of sudden drastic performance drops is inevitable in practice, as rate levels must be adjusted without complete channel information.

In cases where this risk is considered small and unfairness is acceptable, a practical scheme for threshold selection is required. In Section 6.2 we developed an adaptive scheme which was seen to result in high quantization efficiency in simulations.

For systems where unfairness and the potential performance drops of pure multiuser diversity are unacceptable, we proposed a multiuser diversity-enhanced version of fixed access, guaranteeing that all users get equal channel access in a time span of U slots, thereby facilitating real-time services. The scheme was seen to yield a multiuser-diversity gain that approximately equals that obtained by a strict multiuser-diversity strategy with $U/2$ users. Furthermore, the proposed strategy avoids the quantization sensitivity of multiuser diversity by decentralizing the determination of rate thresholds.

We also saw that with strict multiuser diversity, unless the number of symbols that constitute a time slot is much larger than the number of active users in the cell, using 2 bits for feedback results in a net throughput loss in comparison to using just 1 bit.

It should finally be pointed out that both the proposed schemes could be used simultaneously in a cellular system using two traffic classes; one with guaranteed channel access and another providing best-effort service with pure multiuser diversity. Such a system would be a hybrid between today's wide-area coverage cellular networks and the hot-spot info-station scenario as suggested e.g. by Frenkiel et al. (2000).

Chapter 7

Inter-Cell Scheduling, Access Control, and Hand-Overs

A CRITICAL aspect in realizing a cost-efficient mobile communications network is to utilize the spectral resources as efficiently as possible. Anticipating that a substantial part of the traffic in current and coming mobile networks will stem from data applications, the traffic load of each user will fluctuate much more strongly than for traditional voice services. Accordingly, as the aggregate demand for transmission capacity in an area becomes more unpredictable, it becomes increasingly important to allow dynamic reallocation of the supplies of transmission resources to areas with currently high demands.

At the same time, each user experiences shadow fading, fast fading and distance-related attenuation of the transmitted signal. Thus, both supply and demand for transmission capacity is subject to a high degree of local variability. From a general standpoint of optimal resource utilization, variations in demand and supply are the driving forces which make dynamic optimization advantageous. In contrast, if we fix the resource partitioning for all time, the variations are a nuisance which degrades the resource efficiency.

In this chapter we will use the mentioned sources of variability as a means to optimize spectral efficiency in the specific case of partitioning down-link transmission channels among interfering and non-interfering sectors in a cellular network. The object is to maximize the expected total throughput in the considered area, while using probability theory to explicitly take the inherent uncertainty concerning individual users' channels and traffic loads into account. The formulation, detailed in Section 7.1, is also intended to serve as a unifying basis for a variety of resource management problems with a common aim to optimize capacity usage.

We outline how this objective can be met for hand-overs and admission control in Section 7.4.

It can be observed that the topics of this chapter are related to that of scheduling users *within* a sector according to channel quality and traffic requirements. This subject is discussed in detail in Chapter 5 and Chapter 6. In discussing practical aspects of the framework derived in this chapter, we will often assume that multiuser diversity is exploited within each sector. The derivations, on the other hand, do not presume so.

There is an extensive literature in the related areas of dynamic spectrum partitioning, hand-overs, and admission control for mobile communications. As indicated by Katzela and Naghshineh (1996) and Verdone and Zanella (2002), for the most part the solutions, either explicitly or implicitly, assume voice traffic, but more recently (Chuang and Sollenberger, 1998, Li et al., 2002, Qiu et al., 2001, Zhang et al., 2002) attempts have been made to meet the anticipated requirements of data traffic. Burstiness, the size of fluctuations, and its unpredictability make resource management for data traffic a challenging problem. Critical aspects that have not been sufficiently investigated in previous studies include uncertain traffic and uncertain transmission capacities.

Further, allocation policies which maximize the aggregate throughput within a group of sectors and take transmission buffers into account have not been reported previously. Our study does not place a lot of weight on fairness and quality of service, although we briefly discuss these issues in connection with admission control. Instead, we set out to find a solution which tells us how to optimally partition a finite set of transmission resources under realistic levels of uncertainty. Analyzing its behavior could then help in designing algorithms aimed at providing certain quality-of-service levels without sacrificing too much capacity.

In the following we will assume (without loss of generality) that the considered network uses OFDM with each frequency bin being slotted in time. The set of transmission resources to be partitioned then consists of time-frequency slots according to Figure 7.1.

7.1 Partitioning Bandwidth for Maximum Expected Throughput

Let us first investigate the problem of partitioning bandwidth dynamically between two sectors which cause high interference in the border zone between the two sectors. Following this solution, we will see how to extend the discussion to multiple sectors.

Consider the problem of distributing N time-frequency slots among two down-

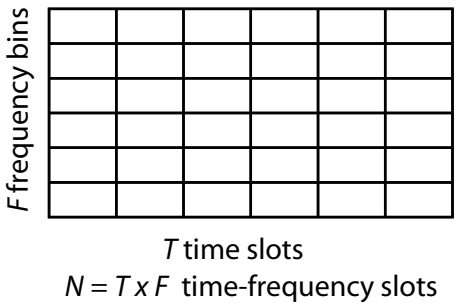

Figure 7.1: The set of transmission resources consists of $N = T \times F$ time-frequency slots.

link sectors and a subarea within which a user experiences unacceptably high interference from the sector which is not transmitting to the specific user. A situation like this is depicted in Figure 7.2, where the similar case of three interfering sectors is also shown. The same situation arises on the border between two sectors lying side-by-side and belonging to the same base station. In all positions apart from the high-interference area, the interference from the other sector is assumed acceptable. Acceptable interference is here taken to mean that the system capacity becomes higher if the same channels are used simultaneously in the low-interference areas than it would be if the total set of channels are divided into two mutually exclusive subsets of channels, one for use in one area, one for use in the other. This means that the geographical partitioning will typically remain more or less the same irrespective of the exact bandwidth partitioning.

N_3 time frequency slots are allocated to the high-interference zone (the black area in Figure 7.2 (a) which we henceforth denote by zone 3), and the remaining $N_1 = N_2 = N - N_3$ slots are used simultaneously by base stations 1 and 2 respectively in the non-disturbed (shaded) areas (which we denote by zone 1 and 2).

In the high-interference zone, a number of transmitter options are possible. The simplest options are exclusive transmission by the nearest base station or joint transmission using macro diversity from all base stations. In the present study we do not consider the macro diversity approach any further, but note that the following problem definition is compatible with any choice of transmission strategy in the high-interference zone.

The aim in this chapter is to find a resource partitioning which maximizes the system throughput, which we shall define as the capacity, within the considered area. The global optimum over the entire network would in principle involve a global coordination, which is not tractable, but an approximation to this end can be

obtained by using a succession of nearest-neighbor partitionings.

The partitioning is likely to be carried out at regular intervals over which the user population in each area does not change significantly. Over the coming period for which the partitioning is to be optimized, the traffic generated by the totality of the respective user populations is incompletely known, as is the exact transmission capacity. Hence, we must first assign a loss function $L(N_3, \theta_j)$ describing the 'loss' incurred to the system on making decision N_3 should θ_j turn out to be the true 'state of nature' in terms of supply and demand for transmission capacity. Then, having decided on a loss function, we must find probability distributions for the remaining uncertainty, which in this case resides in the actual supply and demand for transmission capacity. The optimal partition shall in this work be taken as the solution found by adjusting N_3 so that the expected loss, which we denote by $\langle L \rangle$, is minimized. The loss function describes the amount of data remaining in the transmission buffers. As has been mentioned in previous chapters, minimizing the buffer levels is equivalent to maximizing the throughput in the considered area.

The criterion to maximize the expected capacity may be subject to scrutiny in some applications. Depending e.g. on the network operator's business model, certain events that lie far out in the tails of the probability distributions may in some cases be very costly. In those cases, another criterion should be developed, e.g. one which is more sensitive to such extreme events, i.e. a loss which is more sharply curved than the absolute value of the queue sizes. Note however that the main contribution of this work is not the actual partitioning strategies, but rather the resulting probability distributions and expectations, which are of a more general interest, and equally valid for uses requiring other criteria.

Let N_i denote the number of time-frequency slots allocated to each zone i as defined above, and remember that $N_1 = N_2 = N - N_3$, reflecting that the same slots can be reused in the non-disturbed zones. In the following sections, we will use the term *frame* to describe a set of time-frequency slots that are allocated to a zone. The entire scheduling frame is then the N time-frequency slots that are being partitioned.

Let S_i denote the current number of bits in the transmission buffers corresponding to zone i, and let c_i represent the effective transmission rate per time-frequency slot in the i:th zone. Notice that we use the term *effective* rate to emphasize that c_i represents the transmission rate that is *actually used*, which in a system using multiuser diversity may be significantly larger than the average of all users' individual transmission rates (see e.g. Chapter 5).

Further, let n_i denote the number of bits that will enter the i:th buffer[1] over the

[1]Formally, we here consider one buffer per zone containing the data for all users in that zone, but in practice this buffer is typically made up of individual buffers for each user, just as in Chapter 5.

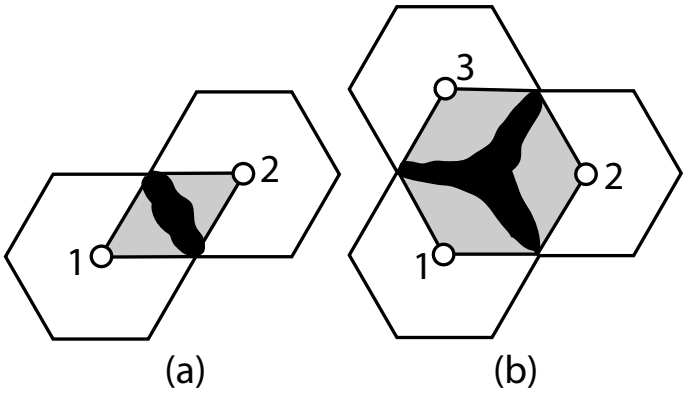

Figure 7.2: The black areas denote the high-interference area where N_3 time frequency slots are allocated. The remaining slots are used simultaneously in the shaded areas, where the interference is acceptably low. Figure (a) shows two interfering 60 degree sectors, and (b) three-sector coordination using 120 degree sectors.

coming scheduled time interval of T time slots.

Maximizing the throughput is equivalent to minimizing the total amounts of data remaining in the transmission buffers for each of the three areas after T time slots. With the given definitions, we then formulate the corresponding loss function as

$$\begin{aligned} L(N_3, \{c_i, n_i\}) = g\left(S_1 + n_1 - (N - N_3)c_1\right) + \\ +g\left(S_2 + n_2 - (N - N_3)c_2\right) + g\left(S_3 + n_3 - N_3 c_3\right) \ , \end{aligned} \tag{7.1}$$

where $g(x) = x$ if $x > 0$, else $g(x) = 0$. Each of the three terms in the loss function describes the number of bits remaining in the transmission buffers for the respective zones, i.e. the sum of the data in stock, S_i, and the influx, n_i, over the coming period, minus the number of bits to be transmitted, $N_i c_i$. We take each c_i to be fluctuating according to different probability distributions for each c_i. Notice that the transmission rate c_i is here assumed to be fixed within each frame of scheduled slots [2], which may seem to be a severe restriction. However, even if the transmission rates vary within a frame, the resulting expression will still be entirely correct provided that the partition allocates bandwidth such that each zone has more data in its buffers than that zone's available transmission rate.

[2]Otherwise, we would need to replace the single c_i with N terms representing individual time-frequency slots, as well as a decision variable for each slot. The corresponding optimal allocation would require calculation of the probability for each possible frame of transmission rates.

The reason is that then the non-linearities due to $g(\cdot)$ disappear and the expectation calculated from the aggregate c_i becomes equal to that of the sum of sub-divided c_i. That would normally be the case. In all other cases, however, the partition may be suboptimal.

In the following section we determine the probability distributions for n_i, the incoming amounts of data, and c_i, the effective transmission rate, and then in Section 7.3 we determine the expectation of the loss (7.1) and find the solution which maximizes the expected capacity. Following this, in Section 7.4 we extend the solution to several sectors, and show that the derived expected loss unifies a number of resource allocation problems, where we emphasize hand-overs and admission control.

7.2 Derivations of Supply and Demand Distributions

7.2.1 The demand distribution

The distribution for the total transmission capacity demand in each zone is denoted by $P(n_i|I)$ given information I. The background information I includes that the total demand in the area in terms of bits per T time slots, the scheduled horizon, is a sum of the influxes into each user's transmission buffer for each time slot, i.e.

$$n_i = \sum_{u=1}^{U_i} \sum_{t=1}^{T} n_{ut}$$

where U_i is the number of users in the i:th zone. If we regard the data streams as originating from some type of best-effort data service such as the Internet, each n_{ut} can be regarded as an independent unknown variable which taken together with the fact that $U_i \times T$ is a large number (most likely > 100), makes the resulting distribution tend into a Gaussian shape by a central limit theorem argument. In Chapter 5 each individual user's influx was modelled by a negative exponential distribution according to the maximum entropy principle subject to known average influxes. A sum of such variables can be shown in computer simulations to converge to a Gaussian distribution with reasonable accuracy even for a small (< 10) number of terms, giving another justification for the choice of a Gaussian model.

In summary, we model the total transmission capacity demand in each zone i in terms of number of bits, n_i, required over the scheduling horizon as

$$P(n_i|I) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{1}{2\sigma_i^2} \left(n_i - \langle n_i \rangle\right)^2 \right) , \tag{7.2}$$

with $\langle n_i \rangle$ and σ_i^2 denoting the mean and the variance, respectively, as determined by the base station serving zone i.

7.2.2 The supply distribution

We now determine the probability distribution for the effective transmission rates c_i of each zone i. Suppose that the transmission rate for each slot can assume only a limited set of values, $c_i = c_{i,1}...c_{i,K}$ and that the base station monitors and stores the relative frequencies with which the different $c_{i,k}$ are used in each zone. Recall from Chapter 6 that in a system employing multiuser diversity, the distribution of relative frequencies with which the $c_{i,k}$ are used depend on the number of users currently in the area[3]. Therefore, the relative frequencies for the different $c_{i,k}$ should be monitored and stored as a function of population size.

Assume that according to these records, the i:th zone has until now served $m_{i,k}$ time-frequency slots at the transmission rate $c_{i,k}$. The total number M_i of monitored slots can then be written as

$$M_i = \sum_{k=1}^{K} m_{i,k} \ ,$$

where K is the number of rate levels supported by the base station.

We are now interested in determining the probability for serving $r_{i,k}$ time-frequency slots at rate $c_{i,k}$ in the *next* frame. Assuming that the underlying causal mechanisms which determine the transmission rates do not change significantly with time, it follows that the relative frequencies should remain constant as well, and we take the probability for each $c_{i,k}$ as the expectation of the relative frequencies with which it occurs.

We seek to evaluate

$$\begin{aligned} &P(f_{i,1}...f_{i,K}|m_{i,1}...m_{i,K}I) = \\ &\quad = \frac{P(m_{i,1}...m_{i,K}|f_{i,1}...f_{i,K}I)P(f_{i,1}...f_{i,K}|I)}{P(m_{i,1}...m_{i,K}|I)} \end{aligned} \tag{7.4}$$

where

$$f_{i,k} = \frac{r_{i,k}}{\sum_{j=1}^{K} r_{i,j}} \tag{7.5}$$

is the relative frequency with which $c_{i,k}$ will be used, and I is the background information stated above. This problem was solved in Section 2.6, where we under similar circumstances derived the probability for the occurrence of an event

[3]The probability that there is at least one user who can transmit at rate $c_{i,k}$ but no user that can transmit at the nearest larger rate $c_{i,k+1}$ is, according to (6.7),

$$\prod_{u=1}^{U_i} \int_0^{c_{i,k+1}} P(r_u|I)dr_u - \prod_{u=1}^{U_i} \int_0^{c_{i,k}} P(r_u|I)dr_u \ , \tag{7.3}$$

where $P(r_u|I)$ is the probability distribution for user u:s rate.

given only a record of its previous number of occurrences. The solution was the generalized rule of succession due to Laplace.

The probability for transmitting at a certain rate $c_{i,k}$ in an 'average' time-frequency slot during the next scheduled frame is then given by

$$p_{c_{i,k}} \stackrel{\triangle}{=} P(c_{i,k}|m_{i,1}...m_{i,K}I) = \frac{m_{i,k}+1}{M_i+K} \,. \tag{7.6}$$

For an interpretation and a discussion on common-sense correspondences for this probability assignment, see Section 2.6.

7.3 Solution to the Resource Partitioning Problem

Having derived the probability distributions for the supply and demand in each area, we now determine the expectation of the loss (7.1). Under the condition that the influxes n_i and the effective transmission rates $c_{i,k}$ are logically independent, we have

$$\begin{aligned}
\langle L \rangle &= \sum_{k=1}^{K} p_{c_{1,k}} \int_{-\infty}^{\infty} P(n_1|I) g(S_1 + n_1 - (N - N_3)c_{1,k}) dn_1 \\
&+ \sum_{k=1}^{K} p_{c_{2,k}} \int_{-\infty}^{\infty} P(n_2|I) g(S_2 + n_2 - (N - N_3)c_{2,k}) dn_2 \\
&+ \sum_{k=1}^{K} p_{c_{3,k}} \int_{-\infty}^{\infty} P(n_3|I) g(S_3 + n_3 - N_3 c_{3,k}) dn_3 \,.
\end{aligned} \tag{7.7}$$

Here we have used the more compact notation $p_{c_{i,k}} = P(c_{i,k}|m_{i,1}...m_{i,K}I)$ introduced in (7.6). Integrals of the type in (7.7) are evaluated in Appendix A. Adjusting the lower integration limit due to $g(\cdot)$, we find that

$$\begin{aligned}
&\int_{-\infty}^{\infty} P(n_i|I) g(S_i + n_i - N_i c_{i,k}) dn_i = \\
=&\int_{N_i c_{i,k} - S_i}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{1}{2\sigma_i^2}\left(n_i - \langle n_i \rangle\right)^2\right) \times \\
&\times (S_i + n_i - N_i c_{i,k}) dn_i \\
=&\frac{1}{2}\left[\sqrt{\frac{2}{\pi}}\sigma_i \exp\left(-\frac{\alpha_i^2}{2\sigma_i^2}\right) + \alpha_i \left(\text{erf}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right) - 1\right)\right]
\end{aligned} \tag{7.8}$$

where

$$\alpha_i = N_i c_{i,k} - S_i - \langle n_i \rangle \,. \tag{7.9}$$

The resulting expected loss is

$$\langle L \rangle = \sum_{i=1}^{3} \sum_{k=1}^{K} p_{c_{i,k}} \frac{1}{2} \left[\sqrt{\frac{2}{\pi}} \sigma_i \exp\left(-\frac{\alpha_i^2}{2\sigma_i^2} \right) + \alpha_i \left(\text{erf}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i} \right) - 1 \right) \right], \quad (7.10)$$

with $p_{c_{i,k}}$ defined in (7.6) and α_i defined in (7.9).

In Appendix 7.A we prove the following theorem which gives the optimum partition between the zones when the N_i are allowed to be continuous. We shall take the discrete solution to be the integer N_i closest to the continuous optimum.

Theorem 7.1 *The partition N_3 which minimizes the expected buffer levels (7.10) is obtained by solving the equation*

$$\sum_{k=1}^{K} \left(p_{c_{3,k}} c_{3,k} \mathit{erfc}\left(\frac{\alpha_3}{\sqrt{2}\sigma_3} \right) - \sum_{i=1}^{2} p_{c_{i,k}} c_{i,k} \mathit{erfc}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i} \right) \right) = 0 \quad (7.11)$$

with

$$\alpha_i = N_i c_{i,k} - S_i - \langle n_i \rangle \quad (7.12)$$

where it should be remembered that $N_1 = N_2 = N - N_3$.

The term erfc $\left(\frac{\alpha_i}{\sqrt{2}\sigma_i} \right)$ in (7.11) is twice the probability that n_i is larger than $N_i c_{i,k} - S_i$, i.e. it is proportional to the probability that there is a non-zero loss contribution from zone i. Assuming that the transmission rates c_i are known, the optimum partition (7.11) thus balances the transmission rate in an average time-frequency slot multiplied by the probability for a non-zero loss contribution from the high-interference zone with the sum of the corresponding quantity for the two low-interference zones. Likewise, when the c_i are uncertain, the optimum is obtained by balancing the expectation over p_{c_i} of these quantities.

The balance equation (7.11) does not admit a general solution in closed form but can be solved numerically. The left hand side of (7.11) is either monotonically increasing or monotonically decreasing as a function of $N_1 = N_2$, and the optimum can be found in a few iterations. The computational complexity should therefore not limit the potential use of this scheduler.

7.4 Extensions

7.4.1 Several sectors

From the balance equation to be solved for optimal local partitioning (7.11) the generalization to l sectors with one common zone of high interference is immediate:

$$\sum_{k=1}^{K}\left(p_{c_{3,k}}c_{3,k}\text{erfc}\left(\frac{\alpha_3}{\sqrt{2}\sigma_3}\right)\right. \\ \left. -\sum_{i=1}^{l}p_{c_{i,k}}c_{i,k}\text{erfc}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right)\right)=0\,. \tag{7.13}$$

A global optimization for all sectors in a network in the general case of interference between several sectors is not tractable due to the interdependence of all partitions. But if we assume that the sectors are mainly disturbed by the three neighboring sectors (the one standing opposite to it, and the nearest sectors to the left and to the right) then we can make a sequential partitioning with one neighbor at a time.

7.4.2 Hand-overs

When a user requests a hand-over from one zone or sector to another, the maximum capacity criterion translates into recalculating the partitioning according to (7.11) with the user transferred to the zone requested. If the optimal partition yields a higher expected loss (7.10) than the optimal partition with the user remaining in the current zone, then the hand-over request is rejected, otherwise it is granted. One may also use a less throughput-oriented scheme by allowing a hand-over request if the optimal new partition gives an expected loss that is below a given bound. Such a bound may be calculated by weighing the cost of decreased throughput with the cost of lost connections. Using (7.10) it is possible to explicitly calculate the performance loss from service guarantees and decide on acceptable bounds .

There are two important factors when recalculating the partitioning with the user changing zones. First, the aggregate buffer contents, influx expectations, and influx variances must be adjusted in each zone by adding/subtracting the respective quantities of that user in the new/old zone.

Secondly, the transmission rate distributions $p_{c_{i,k}}$ must also change accordingly. If the network takes advantage of multiuser diversity, the average transmission rate increases with the number of users, and particularly so when the user population is small (see Chapter 5). This should be taken into account by keeping separate records of the relative transmission rate frequencies according to the

number of users in the zone. This implies assuming that the relative frequencies of transmission rates are constant over time for each population size, but that they vary with the population size. When the number of users in a zone is large this assumption is valid, but if the number of users is very small, the effect from multiuser diversity is lost and the specific locations and mobility of the few users take over as the rate-determining factor. But on the other hand, the relative frequencies in the case of few users will be almost uniform as a consequence of the mentioned effect; the resulting rate distributions will thus reflect the inherent uncertainty and lead us to take a precautious decision. Improved tracking of the actual capacity supply could only be obtained from detailed channel predictions for each user, which is not realistic on the considered time scales.

7.4.3 Admission control

In systems employing a strict capacity-optimal regimen, admission control may be neglected altogether since the system then assigns each time-frequency slot to the user that has the highest transmission rate. In this respect, guaranteeing certain levels of service quality is simply suboptimal and thus any user is allowed to enter the system, which however does not imply that the user actually gets any service.

In a less extreme network, however, where all connected users are given at least some minimum level of service, admission control is an important issue. The decision to admit or reject a requesting user can clearly be put in the framework we build upon here. If the system promises some minimum service level to its users, this means that the expected loss (7.10) cannot be allowed to grow too large.

Let the service guarantee consist of a commitment to transmit to each user u at a rate such that the expected buffer level of that user after the completion of a scheduling frame does not surpass a certain amount q_u. For this to be a meaningful guarantee, the expected influx $\langle n_u \rangle$ of the user must be known to the network.

The fulfillment of the guarantee for users admitted to the network must be carried out partly on the level of spectrum partitioning between zones, but mainly on the level of user scheduling within each zone. This level of scheduling is not studied here, but the approach presented in Chapter 5 could be used with constraints on average allocated rates.

The decision to admit or reject a new user under the described service guarantee resembles the hand-over solution from Section 7.4.2. First, calculate the optimal expected loss (7.10) with the user having gained access using the same adjustments as for hand-over decisions. Then, if the sum of all users' (including the new user) service guarantees $\sum_u q_u$ within the total two-sector area is lower than the expected loss[4] $\langle L \rangle$, the new user is admitted. In this case, optimal spectrum

[4]Remember that the expected loss is equal to the expected amount of remaining data in the buffers

Table 7.1: Standard parameters for performance tests for the three zones, $i = 1...3$. The parameter c_i is the effective transmission rate, S_i is the current number of bits in stock, and $\langle n_i \rangle$ and σ_i is the expectation and the standard deviation, respectively, for the number of incoming bits over a scheduling interval. The total number of scheduled slots is $N = 500$

i	c_i	S_i	$\langle n_i \rangle$	σ_i
1	15	500	2500	200
2	15	500	2500	200
3	10	500	2500	200

efficiency is obtained simultaneously with guaranteed service quality.

An alternative is to allow suboptimal partitions and instead find N_3 under the criterion that the expected loss (7.10) is less than $\sum_u q_u$. This strategy leads to reduced throughput but admits more users.

7.5 Performance Examples

As an illustration of how the proposed scheduling framework performs, we here investigate a few different scenarios with varying uncertainty and traffic load. We study the basic partitioning problem for two sectors with one area of high mutual interference (cf. Figure 7.2), where the solution is obtained by solving (7.11) for N_3. The other issues studied in the chapter – hand-overs, sequential nearest-neighbor partitioning, and call-admission control – all use the same unifying framework and their behavior thus follow a similar pattern. In all tests, if not otherwise stated, the parameters in Table 7.1 are used, and the total number of scheduled slots is $N = 500$.

7.5.1 Known transmission rates

Assuming that the effective transmission rate per time-frequency slot in each zone is fixed and known[5], (7.11) simplifies to

$$c_3 \text{erfc}\left(\frac{\alpha_3}{\sqrt{2}\sigma_3}\right) - \sum_{i=1}^{2} c_i \text{erfc}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right) = 0\,. \tag{7.14}$$

after the completion of the scheduled period

[5]This corresponds to a situation in which rate adaptation is not used, but instead power control is employed to give all users in a zone the same c_i

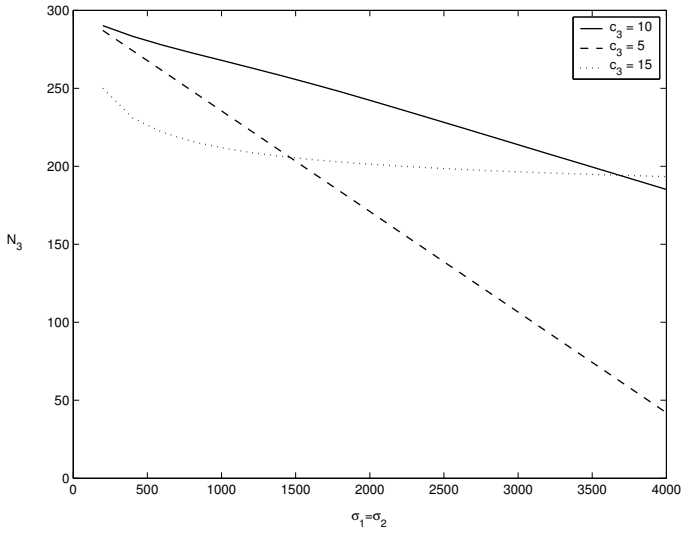

Figure 7.3: The optimal N_3 for fixed σ_3 and varying σ_1 and σ_2 for known and fixed transmission rates. Expected traffic loads etc. are shown in Table 7.1. It should be observed that as σ_i becomes very large the probability mass for negative values of n_i becomes non-negligible with the Gaussian demand distribution, a fact which may affect the accuracy at large values of $\sigma_1 = \sigma_2$.

In this case, if the traffic load in all zones exceeds the transmission capacity and the traffic uncertainty σ_i is low, then the minimum required effective transmission rate c_3 for zone 3 to obtain any time-frequency slots is (assuming $c_1 = c_2$) $c_3 \geq 2c_1$. This follows directly from the definition of the loss function (7.1). But when the system is less heavily trafficked[6] the scheduler will allocate resources to all zones according to their respective demands and effective transmission capacities.

Let us first see how the system reacts to varying amounts of uncertainty concerning the capacity demands. We use the parameters listed in Table 7.1, and vary the standard deviation of the traffic generated in zones 1 and 2 while keeping σ_3 fixed. The resulting optimum N_3 for three cases of effective transmission rates in zone 3 are displayed in Figure 7.3.

We see that for higher uncertainties σ_1 and σ_2, the general tendency of the scheduler is to lower N_3 and thus increase the number of time-frequency slots for zones 1 and 2. The optimal partition N_3 is very nearly a linear function of σ_1 and σ_2 for $c_3 = 5$ and $c_3 = 10$. But when the effective transmission rate of zone 3

[6] A well-dimensioned system should for the most part operate below the congestion level, or else it needs to increase its transmission capacities by either adding more base stations or increasing the bandwidth.

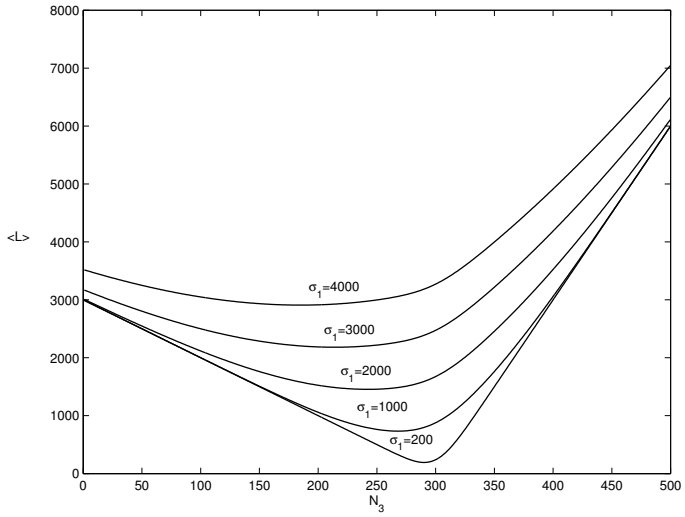

Figure 7.4: The expected loss $\langle L \rangle$ as a function of N_3 for varying $\sigma_1 = \sigma_2$ with $c_3 = 10$. Note how the optimum becomes sharper with decreasing uncertainty. Expected traffic loads etc. are shown in Table 7.1.

equals that of the other zones, the slope decreases for increasing uncertainty. This rather complex behavior can be understood from the observation that for increasing σ_1 and σ_2 the expected loss contributions of these two zones also increase, while the contribution from zone 3 remains the same. Thus, the relative advantage of giving more time-frequency slots to zones 1 and 2 increases with σ_1 and σ_2, explaining the sign of the slope of N_3, but it decreases with c_3, which explains the difference in magnitude of the slopes. For $c_3 = 15$ the magnitude of the slope actually decreases with uncertainty; here, the scheduler strikes a balance between the potentially higher loss contributions from zones 1 and 2, and the high utilization which is certain to result from spectrum usage in zone 3.

The expected loss as a function of N_3 is displayed in Figure 7.4 for $c_3 = 10$ and for different values of σ_1 and σ_2. From this plot it can be seen that, as expected, lower uncertainty translates into a sharper and lower optimum.

Fixing $\sigma_1 = \sigma_2 = 200$ and instead varying σ_3, the optimal N_3 varies according to Figure 7.5. The variations for $c_3 = 5$ and $c_3 = 10$ are now small, and N_3 decreases slightly as the uncertainty increases. The high-interference zone simply obtains the time-frequency slots that are left when the other zones with higher transmission rates and better known traffic loads have filled their needs. But when $c_3 = 15$, the fact that the expected loss contribution from zone 3 increases with the added uncertainty takes over as the determining factor, and the optimal N_3

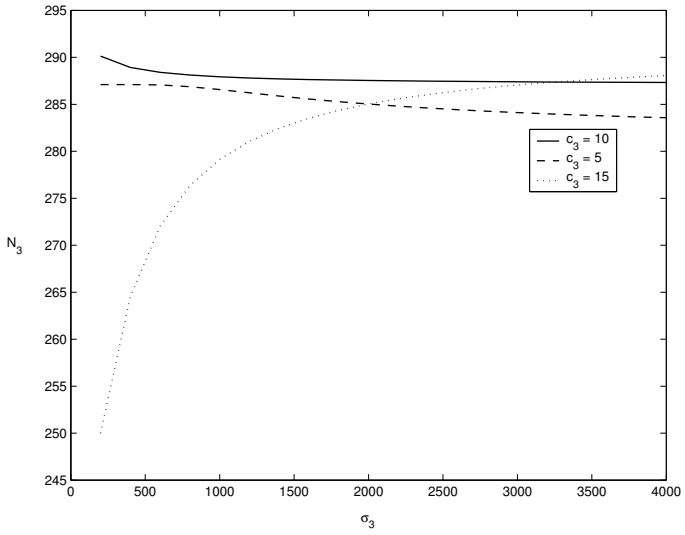

Figure 7.5: The optimal N_3 for varying σ_3 and fixed σ_1 and σ_2 with known and fixed transmission rates. Expected traffic loads etc. are shown in Table 7.1.

Table 7.2: Transmission rates $c_{i,k}$ ($K = 4$) and corresponding probabilities $p_{c_{i,k}}$.

k	1	2	3	4
$c_{i,k} \forall i$	5	10	15	20
$p_{c_{1,k}}$	0.15	0.25	0.35	0.25
$p_{c_{2,k}}$	0.15	0.25	0.35	0.25
$p_{c_{3,k}}$	0.25	0.35	0.25	0.15

consequently increases with σ_3.

In Figure 7.6 the optimal N_3 is plotted as a function of the expected traffic in zones 1 and 2, $\langle n_1 \rangle = \langle n_2 \rangle$. In this test, the standard deviations were fixed at $\sigma_i = 200$. The three curves correspond to $c_3 = 5, 10, 15$. The curves contain no surprises, for small traffic loads in the low-interference zones, the optimal partition is loss-free, and thus the majority of the slots are awarded to zone 3. When the traffic in zones 1 and 2 reaches a critical level however, N_3 decreases, reflecting the higher spectral efficiency that follows when these zones can use the available resources.

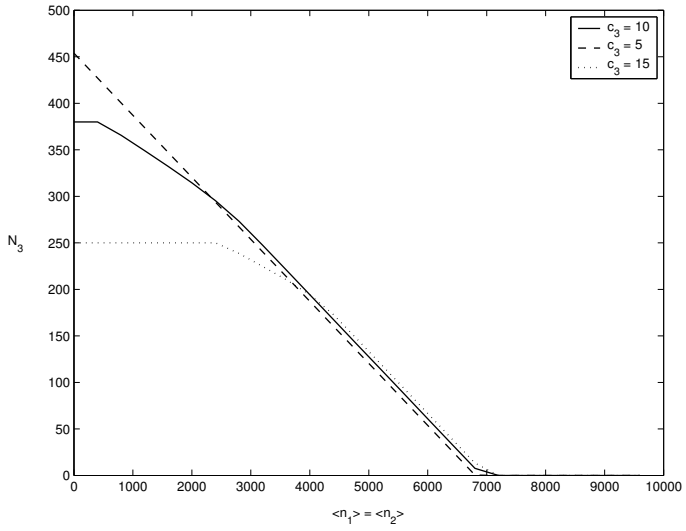

Figure 7.6: The optimal N_3 for varying $\langle n_1 \rangle = \langle n_2 \rangle$ and fixed standard deviations with known and fixed transmission rates.

7.5.2 Uncertain transmission rates

With uncertain effective rates c_i according to Table 7.2, the resulting optimal N_3 as a function of the expected traffic in zones 1 and 2 are given in Figure 7.7. Apart from the parameters just mentioned, the conditions are the same as in the equivalent test in the case of known and fixed rates. As a comparison, the figure shows both the true optimum obtained from solving (7.11) for N_3 (solid line), and the N_3 obtained by simply plugging in the average effective rates $c_i = \sum_k c_{i,k} p_{c_{i,k}}$ in (7.14)[7] (dashed line). The difference is not insignificant, and shows a surprising behavior. The true optimum is at first higher than the 'estimate plug-in' solution, then for an intermediate range of traffic intensity lower, and then for high loads once again higher. For the lowest traffic loads the estimate plug-in solution has a wide interval of N_3 which reaches the same estimate of the loss and that interval actually includes the true optimum from (7.11). However, with the use of (7.11) there is a single sharp optimum singling out a more conservative solution, while the suboptimal scheduler does not see any difference between a range of N_3 as wide as 100 time-frequency slots. Investigating the range of values around $\langle n_1 \rangle = \langle n_2 \rangle = 3000$, the discrepancy is no longer due to the same effect; here both

[7] It should be noted that this corresponds to using a loss function without the $g(\cdot)$ function. The decision may then become to allocate more slots than can actually be used to some zone (while others could in fact use it) since over-allocation decreases such a loss function.

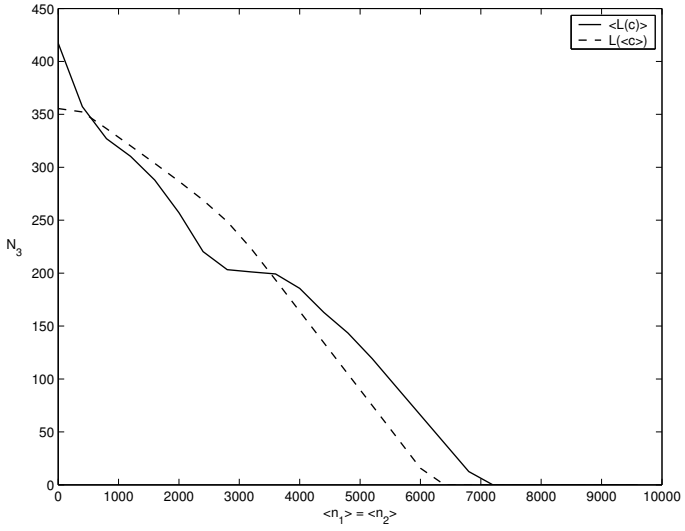

Figure 7.7: The optimal N_3 for varying $\langle n_1 \rangle = \langle n_2 \rangle$ and fixed standard deviations with uncertain c_i according to Table 7.2. The solid line is the true optimum obtained from solving (7.11), the dashed line shows the decision when using the average transmission rate in (7.14).

schedulers see one distinct optimum but the correct scheduler, aware of the actual uncertainty concerning the transmission rate, makes a more conservative decision which at this traffic load results in a lower value of N_3. A similar situation holds for the higher traffic intensities as well, but here a more precautious decision is to give more time-frequency slots to zone 3 than would be obtained with the estimate plug-in scheduler. This can be understood from studying the extreme case when $\langle n_1 \rangle = \langle n_2 \rangle \geq 6000$. At that traffic load, the estimate plug-in solution, confident of the fact that c_1 and c_2 are fixed at the average 13.5, sees that when the buffer loads corresponding to these two zones are larger than $13.5 \times 500 = 6750$, all slots can be used by these two zones without any risk of emptying the buffers. Compare this to the true optimum including knowledge of the rate uncertainty. Now there is a definite chance that the transmission rates are higher than 13.5 and thus a few slots should be left for zone 3 where it is certain that these slots can be used. These remarks are given further confirmation from Figure 7.8 which shows the same scenario as above but with uniform rate distributions for all three zones. We see that the difference becomes larger in this state of larger uncertainty, particularly for higher traffic intensities. For example, at $\langle n_1 \rangle = \langle n_2 \rangle = 5500$ the difference in N_3 for the two schedulers is almost 100 slots. In terms of expected total throughput the difference is however not very large; for $\langle n_1 \rangle = \langle n_2 \rangle = 5500$, the true expected

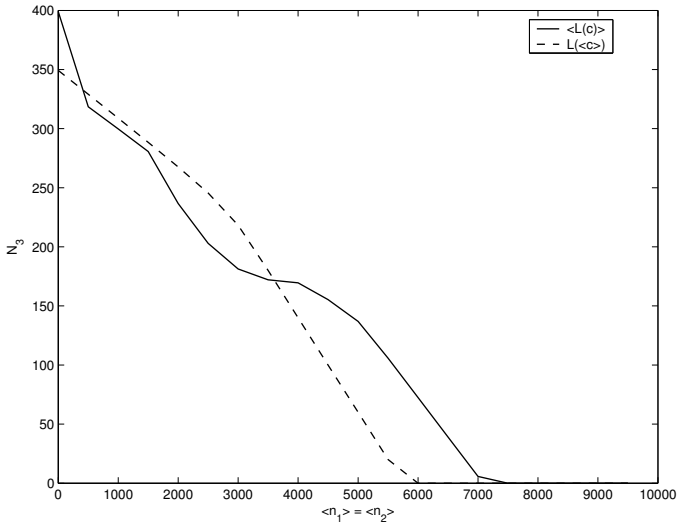

Figure 7.8: The optimal N_3 for varying $\langle n_1 \rangle = \langle n_2 \rangle$ and fixed standard deviations with uniform probability distributions for all c_i. The possible rates c_i are the same as in Table 7.2. The solid line is the true optimum obtained from solving (7.11), the dashed line shows the decision when using the average transmission rate in (7.14).

loss becomes $\langle L \rangle = 5150$ bits for the estimate plug-in solution, and $\langle L \rangle = 4786$ bits for the true optimum. The relative performance difference is thus less than 10%.

7.6 Conclusions

We have presented a method for dynamic partitioning of transmission channels among interfering sectors resulting in maximum expected throughput within the total area. As the main case of interest in this work, we investigated two sectors with one zone characterized by high mutual interference. Maximal expected throughput for this case is obtained by solving (7.11) for N_3, the number of channels allocated to the high-interference zone.

In Section 7.4 a natural extension to several interfering sectors was given. It was further shown that the introduced framework can also be used for hand-overs and admission control with quality-of-service constraints in terms of buffer levels.

The behavior of the channel partitioning solution was investigated in Section 7.5. The results showed that the optimal partition is highly dependent on

the amount of uncertainty concerning both traffic loads and transmission rates. It was observed that if transmission rate uncertainty is neglected by using estimates instead of averaging over the loss function, the resulting partitions become slightly more hazardous. In contrast, by the correct procedure, as dictated by probability theory, the partitions are more precautious, yielding solutions better in line with what common sense would suggest. Even though the relative differences in Figure 7.7 are only about 10% one should keep in mind the comments made in Section 7.5.2; the estimate plug-in solution does not see any difference in the incurred loss in intervals as wide as 100 slots. Therefore, the actual performance difference may become quite large depending on which of these 100 values the optimization program happens to choose. Further, Figure 7.8 shows that for large rate uncertainties the differences increase.

In calculating the expected loss (7.10), we derived probability distributions based on two assumptions. The critical assumption for both supply and demand distributions is that of an approximately constant number of users within each area. This should not be restrictive, but merely place an upper limit on the length of the scheduling intervals. It was emphasized that in networks employing multiuser diversity, the transmission rate distributions depend on the number of users which implies that these distributions should be calculated and stored for a number of typical population sizes.

The schemes considered here does not rely on measurements carried out by the receivers, which is a common problem with dynamic channel assignments. This is both a strength and a weakness of our proposal. The problem resides in the fact that the method completely neglects the SNR at the receiver, and therefore it treats users very far from the base station exactly in the same way as users near the base station. In practice therefore, in a decision whether to accept a new user or not, there must first be a pre-access control decision on whether to consider the user at all or not, preferably on the basis of the user's distance to the base station. If the mobile terminals are equipped with a positioning technology, then our approach is very reasonable, and then there is no need to rely on SNR measurements at the receiver. The main problem of relying on such receiver measurements in access control decisions is that the SNR varies strongly both over fast and slow time scales. As a connection typically lasts for several minutes, during which the user may move quite far, the initial SNR measurements are unrepresentative for the mean SNR during the connection. We believe that a better alternative is to make pre-access decisions to consider the user for access or not based on the position of the user, and then rely on the statistical channel properties of the cell for the final access decision.

In conclusion, it should be pointed out that the probability distributions for supply and demand were derived from particular information which is possible to

collect by the base stations in today's networks. Thus, the partitioning proposed here should be possible to deploy in current or near-future systems.

Appendix 7.A Derivation of the Optimum Partition

An N_3 which minimizes (7.10) can be found using Lagrange multipliers with the constraints $N_1 + N_3 = N$ and $N_2 = N_1$. There may not exist a point where the derivative of the loss function is actually zero. In that case the solution is simply $N_3 = 0$ or $N_3 = N$ according to whether the sign of the derivative of (7.10) is negative or positive.

We form (remembering that $N_1 = N_2$)

$$J(N_1, N_3, \lambda) = \langle L \rangle - \lambda(N - N_1 - N_3) \tag{7.15}$$

and differentiate with respect to N_1, N_3, and λ, respectively,

$$\begin{aligned}
\tfrac{\partial J}{\partial N_1} &= \lambda \\
&+ \sum_{i=1}^{2}\sum_{k=1}^{K} p_{c_{i,k}} \frac{1}{2}\left[\sqrt{\frac{2}{\pi}}\sigma_1 \frac{\partial \exp\left(-\frac{\alpha_i^2}{2\sigma_i^2}\right)}{\partial N_1} + \frac{\partial\left(\alpha_i\left(\mathrm{erf}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right) - 1\right)\right)}{\partial N_1}\right] \\
&= \lambda + \sum_{i=1}^{2}\sum_{k=1}^{K} p_{c_{i,k}} \frac{1}{2}\left[\sqrt{\frac{2}{\pi}}\sigma_1 \frac{\partial \exp\left(-\frac{\alpha_i^2}{2\sigma_i^2}\right)}{\partial \alpha_i^2}\frac{\partial \alpha_i^2}{\partial N_1}\right. \\
&+ \left.\frac{\partial \alpha_i}{\partial N_1}\left(\mathrm{erf}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right) - 1\right) + \alpha_i \frac{\partial\left(\mathrm{erf}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right) - 1\right)}{\partial \alpha_i}\frac{\partial \alpha_i}{\partial N_1}\right] \\
&= \lambda + \sum_{i=1}^{2}\sum_{k=1}^{K} p_{c_{i,k}} \times \frac{1}{2}\left[\sqrt{\frac{2}{\pi}}\sigma_i\left(-\frac{\alpha_i c_{i,k}}{\sigma_i^2}\exp\left(-\frac{\alpha_i^2}{2\sigma_i^2}\right)\right)\right. \\
&+ \left. c_{i,k}\left(\mathrm{erf}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right) - 1\right) + \frac{\sqrt{2}\alpha_i c_{i,k}}{\sqrt{\pi}\sigma_i}\exp\left(-\frac{\alpha_i^2}{2\sigma_i^2}\right)\right] = 0
\end{aligned} \tag{7.16}$$

where the exponential terms cancel and the result is

$$\frac{\partial J}{\partial N_1} = \lambda - \sum_{i=1}^{2}\sum_{k=1}^{K} p_{c_{i,k}} c_{i,k} \mathrm{erfc}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right) = 0\,. \tag{7.17}$$

In the same way, the derivative with respect to N_3 is

$$\frac{\partial J}{\partial N_3} = \lambda - \sum_{k=1}^{K} p_{c_{3,k}} c_{3,k} \mathrm{erfc}\left(\frac{\alpha_3}{\sqrt{2}\sigma_3}\right) = 0\,, \tag{7.18}$$

and the derivative with respect to the Lagrange multiplier is

$$\frac{\partial J}{\partial \lambda} = N_3 + N_1 - N = 0 \Leftrightarrow N_1 = N - N_3\,. \tag{7.19}$$

Noting that (7.17) and (7.18) are both equal to zero, we have

$$\sum_{k=1}^{K}\left(p_{c_{3,k}}c_{3,k}\text{erfc}\left(\frac{\alpha_3}{\sqrt{2}\sigma_3}\right) - \sum_{i=1}^{2}p_{c_{i,k}}c_{i,k}\text{erfc}\left(\frac{\alpha_i}{\sqrt{2}\sigma_i}\right)\right) = 0 \tag{7.20}$$

with, as before,

$$\alpha_i = N_i c_{i,k} - S_i - \langle n_i \rangle \, . \tag{7.21}$$

Chapter 8

A New Method for Adaptive Approximation of Non-Stationary Posterior Distributions and Expectations

HERE we introduce a simple and practical method for making approximate Bayesian inference. An approximate discretized posterior probability distribution is computed on block-wise data. The method is valid for arbitrary probability distributions including those that vary between blocks, but any information regarding time-dependencies is neglected. If information of time-dependent behavior is available then the method does not provide an optimal approximation.

The method relies on approximating an optimal inference by using a probability distribution for quantized intervals of the unknown quantity, and by adapting the quantization so as to obtain higher resolution in regions of higher probability. The probability distribution is partitioned into K bins. After a block of data is observed, the posterior probability for each bin is computed by the use of Laplace's rule of succession. The total probability in each bin is then spread out uniformly over the individual values within the bin. Based on this posterior probability distribution, the widths of the K bins are adjusted so as to maximize the mutual information between the quantized distribution[1] and the unquantized distribution. As we shall

[1] In this chapter, whenever we speak of a quantized distribution we really mean a continuous-valued distribution over discrete intervals of the variable of interest. It is not the probabilities that are quantized, but rather the variables for which the probability is calculated.

see, this approach is equivalent to maximizing the entropy of the quantized distribution, and we provide a low-complexity algorithm for approximately attaining equal probability mass within each bin. The resulting quantized distribution can be regarded as a histogram with K bars of equal area, but in general of unequal width. Using this strategy, the posterior quantized distribution will increase the resolution in regions of high probability and decrease it in low-intensity regions.

The method can be used to provide adaptive quantization of arbitrary data sequences, or to approximate the posterior expectation of for instance some loss function by summing over K terms. A useful feature is that the method adapts to incoming data and takes optimal advantage of any patterns by Bayes' theorem.

In the following example taken from mobile communications we provide a motivating application for the method.

EXAMPLE 8.1 Adaptive inference on data streams

Consider the problem studied in Chapter 5 of scheduling transmissions to users in a mobile communications system. A controller wishes to schedule the use of the channel for T time slots ahead, but then faces the problem that the channel quality and the arrival rates into each buffer is unknown. Focusing here only on the arrival rates, a possible approach to handling the uncertainty regarding the number of bits entering the buffer would be to assign a probability distribution based only on the maximum entropy principle, as was discussed in Chapter 5. This is a valid approach if the controller has information about for instance the average arrival rate in each buffer. However, as time evolves the controller can monitor the arrival rates and thus learn any patterns in the arrival rates by the use of Bayes' rule. Assuming that the statistics of the arrival rates do not change considerably during a certain period, we could use Laplace's rule of succession to obtain the probability p_k for an influx of size k bits,

$$p_k = \frac{n_k + 1}{N + K} \,, \tag{8.1}$$

where n_k is the number of times over the N most recent observations that the influx consisted of k bits, and K is the number of possible influx sizes. But if the possible data rates vary over a large interval, say from 0 bits/second to 1 megabit/second, K would be so large that the posterior distribution p_k would be uniform[2] for all practical purposes (since the observations N would then typi-

[2]By uniform, we here refer to the fact that the majority of all possible outcomes will be equally likely, although the distribution will have occasional peaks. When we say that a distribution is close to uniform, we mean this in the sense that the entropy of the distribution is close to that of a uniform one (i.e. $\log K$).

cally be much smaller than K).

Instead, it could prove useful to partition the interval of possible influxes into a smaller set of regions, or bins, and apply the rule of succession on this smaller set of possibilities. For improved performance we should let the bin widths be adapted based on incoming data. Then the bins should spread out and become wide in regions where little activity is observed, and become denser in the rate interval of frequent observations. Thus, high fidelity is attained where it is suggested by the data, and less attention is paid to atypical rate regions. Within each bin, the probability for individual values is assigned by the principle of indifference. The expectation of any function of the arrival rates can then be obtained by a simple summation over the quantized posterior distribution and the function.

Related Work

The problem of approximating a pdf by a simpler one is certainly not new. Indeed, since a solution to this problem has the potential of strongly simplifying Bayesian inferences by replacing complicated integrals over nuisance parameters by simpler integrals or sums, it is of obvious interest to a large audience. In pattern recognition and adaptive quantization, a problem known as 'non-parametric density estimation' is closely related. Here, the problem is to 'estimate' a probability for obtaining a certain value x based on a number of observations. The resulting pdf should resemble the true, but unknown, distribution as closely as possible. Of course, we would state the problem somewhat differently as we regard probabilities as information carriers rather than properties of nature. Interpreting the density estimation problem as one of approximating a given pdf which may otherwise be difficult to use, then we see that this problem is indeed similar to ours.

There are two standard techniques, see e.g. Fukunaga (1990), used for non-parametric density estimation, or density *approximation* as we would phrase it. The first approach, the *k-nearest-neighbor* approach, finds the probability at the point x by defining a region consisting of the k nearest observations around x. The probability for x is then taken as

$$\hat{p}(x) = \frac{k}{Nw} \tag{8.2}$$

where N is the total number of observations and w is the width of the region[3]. The problem of determining which k to use can be solved in the sense of mini-

[3]Sometimes $\frac{k-1}{Nw}$ is used instead, as this provides an unbiased estimate.

mum mean squared error, but the solution depends on the true distribution $p(x)$. A problem with the k-nearest-neighbor approach is that it does not result in a proper probability density, as it does not integrate to unity, c.f. Bishop (1995). It is however mostly used for classifying observations into different classes in which case it yields a simple rule regardless of this.

The second approach, the *kernel-based* or *Parzen-window* approach, computes histograms using constant bin sizes and smoothes the obtained histograms with some windowing function.

A disadvantage of both methods is that they require all samples to be retained (increasing storage requirements), not just in which bin a sample occurred. The kernel-based approach moreover requires quite intensive computational work.

8.1 Maximizing the Mutual Information Between an Approximate and an Exact Distribution

We here show that maximizing the mutual information between a quantized posterior distribution and an exact posterior is equivalent to maximizing the entropy of the quantized distribution. Let K be the number of bins to use in the approximation, and $i_{min} \leq i < i_{max}$ be the lower and upper bounds on the unquantized variable i between which we want to approximate $p(i \mid DI)$ (where D is the observed data and I our omnipresent background information). Denoting the mutual information[4] between the quantized and the exact distributions $\mathcal{I}(k, i)$ and writing $p(k) = p(k \mid DI)$ for the posterior probability for obtaining an observation in bin k, and $p(i) = p(i \mid DI)$ for the posterior probability for obtaining the exact value i, we now prove the following theorem.

Theorem 8.1 *The optimum approximation to an exact distribution $p(i)$ for a quantity i, in terms of maximum mutual information between $p(i)$ and an approximate distribution $p(k)$ for quantized intervals (bins) k of the same underlying variable, is obtained when the bin widths of the latter distribution are adjusted so that the resulting distribution for k has maximum entropy.*

Proof: The mutual information between the distribution for the quantized variable k and the distribution for the unquantized variable i is given by (c.f. (2.94))

[4] We here assume i to be integer-valued, but the argument goes through also for continuous quantities.

$$
\begin{aligned}
\mathcal{I}(k,i) &= H(k) - H(k \mid i) && (8.3)\\
&= \sum_{k=1}^{K} \sum_{i=i_{min}}^{i_{max}} p(ik) \log p(k \mid i) - \sum_{k=1}^{K} p(k) \log p(k) && (8.4)\\
&= \sum_{k=1}^{K} \sum_{i=i_{min}}^{i_{max}} p(ik) \log p(k \mid i) \\
&- \sum_{k=1}^{K} \sum_{i=i_{min}}^{i_{max}} p(i \mid k) p(k) \log p(k) && (8.5)\\
&= \sum_{k=1}^{K} \sum_{i=i_{min}}^{i_{max}} p(ik) \log \frac{p(k \mid i)}{p(k)} && (8.6)\\
&= -\sum_{k=1}^{K} \sum_{i \in \text{bin } k} p(i \mid k) p(k) \log p(k) && (8.7)\\
&= -\sum_{k=1}^{K} p(k) \log p(k) \ , && (8.8)
\end{aligned}
$$

where (8.5) follows from (8.4) by using the fact that $\sum_{i=i_{min}}^{i_{max}} p(i \mid k) = 1$ We obtain (8.7) from (8.6) by noting that given knowledge of i we know in which bin k the observation lies, i.e. $p(k \mid i) = 1$ or $p(k \mid i) = 0$ depending on whether i is in bin k or not. Since $p(i \mid k)$ sums to unity we finally have (8.8) from (8.7). The theorem can also be obtained directly from (8.3) by proving that $H(k \mid i) = 0$. (Given i, there is no uncertainty concerning which is the corresponding bin k.) ■

Thus, in order to obtain a quantized distribution which is as similar in information content to the unquantized distribution as possible, we should adjust the bin sizes to obtain equal probability mass in each bin (c.f. Example 2.1).

8.2 Maximizing the Entropy of the Approximate Distribution

Assume that we observe N samples of data before updating the bin widths. Within bin k we obtain n_k observations, and we have K bins in total. Assuming that the underlying causal mechanisms which determine the outcomes are stationary over the N observations and the coming period of N observations, and taking no

account of possible time-dependencies, the probability for a future observation in bin k is

$$p_k = \frac{n_k + 1}{N + K} \tag{8.9}$$

according to Laplace's rule of succession (see Section 2.6).

Now, in order to adjust the bin widths so as to obtain equal probability for all bins (and thus maximum entropy of the approximate distribution), we need to determine the probability for an individual value within an arbitrary bin k. Assume that the width of bin k is w_k, i.e. the bin covers exactly w_k values of the underlying quantity i. Then our task reduces to distributing the probability p_k over w_k values. In order to assume anything else than a uniform distribution within the bin we would require some information which is not indifferent between the different w_k values. Here, we shall keep our solution general and therefore assume information indifference between the different values. Then the principle of indifference (see Section 2.5) behooves us to distribute the probability as

$$p_i = p_k / w_k \qquad i \in \text{bin } k \,. \tag{8.10}$$

An argument can be made for assigning a Jeffrey's distribution summing to p_k in the bin with upper limit i_{max} if that maximum is taken to be very large in comparison with typical values. Similarly, if i can take on negative values, the bin with lower limit i_{min} could also be assigned a Jeffrey's prior with the absolute value of i as argument (so as to reverse the slope). We will henceforth assume a uniform distribution in all bins, but the algorithm below does not change if we instead use a Jeffrey's distribution in the edge bins.

Since we then have the probability for all values i between i_{min} and i_{max}, we can now simply redistribute the bins so that each bin contains approximately probability $p_k = 1/K$. We here suggest a simple method which distributes the bins so as to approximately attain the maximum entropy distribution by a single sweep of i. The emphasis is on low complexity rather than on performance, and several other methods could easily be devised.

(1) Set $k := 1$, $P' := 1$, $x_a := i_{min}$, $x_b := i_{min} + 1$ and $J := 1$.

(2) **If**

$$|P'/(K - k + 1) - \sum_{i=x_a}^{x_b - 1} p_i| > J \qquad \text{OR} \qquad i_{max} - (K - k) < x_b \tag{8.11}$$

then end bin k at $x_b - 1$ (i.e. bin k is the interval $x_a \leq i \leq x_b - 1$)
else set $J = |P'/(K - k + 1) - \sum_{i=x_a}^{x_b - 1} p_i|$, $x_b := x_b + 1$ and go to (2).

(3) If

$$k < K \tag{8.12}$$

then set $k := k+1$, $P' = 1 - \sum_{i=x_a}^{x_b-1} p_i$, $x_a := x_b$, $x_b := x_b + 1$ and $J := 1$ and go to (2)
else end (since the upper limit of bin K is always i_{max}).

The algorithm starts at i_{min} and then step-wise increases[5] the bin width until the total bin probability is close to $1/K$. Specifically, it adjusts the bin end-point so as to have probability as close to $P'/(K-k+1)$ as possible, where P' is the total probability mass remaining to be partitioned and $K-k+1$ is the number of remaining bins (including the one currently under adjustment). Notice that this is achieved by comparing the current probability mass in the bin to J, the deviation from the desired value at the previous candidate end-point of the bin. It is important to adjust to $P'/(K-k+1)$ instead of $1/K$ since a narrow bin with many observations may have much larger probability than $p_k = 1/K$, and if the next bin then tries to cover an interval of probability $1/K$ the remaining bins may have much less than probability $1/K$ to share. The second stop condition in step (2) makes sure that in the end there are not more bins to allocate than the remaining values of i.

After each block of N data, the procedure is repeated taking into account the new data and the previous bin sizes (which to some extent is a reflection of previously observed data). The distribution can thus adapt to changing statistics and produce optimal approximate learning (or, to be exact, the given algorithm provides an approximately optimal approximation to optimal learning).

It should be noted that the number of bins K should be chosen based on N. When N is small there is no point in using a large K, because then the rule of succession will caution us by assigning an almost uniform distribution since the number of observations must be significantly larger than the number of hypotheses if we are to draw any detailed conclusions about the plausibility for the different hypotheses. This suggests that K could be optimized as a function of N, but we leave that as a topic for further research.

There are also two variants of the approach described here:

- We could update the bins based on *all* previous observations, not just those in the most recently obtained block, if we have reason to believe that the probability distribution will remain stationary for all times.
- If the probability distribution is known to be stationary for a certain period, we should set N according to the length of that period.

[5] If the range of i is very large, the step-wise increase of x_b should be made larger than 1 to decrease complexity further.

In order to track changes as quickly as possible, we should adapt the bin widths as often as possible, i.e. as soon as we have obtained any new data. But if we perform updates after each new observation based on a sliding window of the N latest samples (instead of updating the bins after every Nth observation based on these N samples), a disadvantage is that each exact sample value must be stored, and not just in which bin it occurred (since the bins have changed during the data gathering interval.) Moreover, the computational complexity is proportional to how often updates are carried out. Therefore, in the following we only consider the basic case where the bins are updated after every Nth sample according to the N most recent observations.

8.3 Computing Approximate Posterior Expectations

Given the approximate posterior distribution p_k, what is the expectation of some function $f(\cdot)$ of the unquantized variable? The expectation of i given the N most recent data is obtained *before* repartitioning the bins (because the statistics were collected based on the previous partition, not on the new one) as

$$\langle i \rangle = \sum_{k=1}^{K} p_k \frac{i_{k-1} + i_k - 1}{2} \tag{8.13}$$

where we define i_k as the upper limit of bin k, i.e. bin k includes all values[6] from i_{k-1} up to $i_k - 1$, and where we define $i_0 = i_{min}$. Similarly, the posterior expectation for an arbitrary function $f(i)$ is given by

$$\begin{aligned} \langle f(i) \rangle &= \sum_{k=1}^{K} p_k \sum_{i=i_{k-1}}^{i_k-1} p(i \mid k) f(i) \\ &= \sum_{k=1}^{K} \frac{p_k}{w_k} \sum_{i=i_{k-1}}^{i_k-1} f(i) \,, \end{aligned} \tag{8.14}$$

where the second equality was obtained by noting that $p(i \mid k) = 1/w_k$. (Given which bin we are in, each value within the bin is equally likely and has a probability equal to the inverse of the bin width.) If i is instead a continuous variable, which we denote by x to separate the two cases, the expectation is

$$\langle f(x) \rangle = \sum_{k=1}^{K} \frac{p_k}{w_k} \int_{x_{k-1}}^{x_k} f(x) dx \tag{8.15}$$

[6] If i is continuous then the upper limit for values of i within bin k is defined as $i < i_k$ instead of $i \leq i_k - 1$.

Table 8.1: The bin limits after each of the five first blocks of data were observed.

Block	Bin limits						
1	0	3	6	9	12	14	100
2	0	1	2	6	7	8	100
3	0	1	2	7	8	54	100
4	0	1	2	7	8	54	100
5	0	1	2	7	8	54	100

where bin k covers the continuous range $x_{k-1} \leq x < x_k$ and w_k is the bin width $w_k = x_k - x_{k-1}$. In case a Jeffrey's distribution is used in the Kth bin, the Kth term in the expectation (8.15) is replaced by

$$\frac{p_K}{\log(x_K/x_{K-1})} \int_{x_{K-1}}^{x_K} \frac{f(x)}{x} dx \tag{8.16}$$

where $\log(\cdot)$ represents the Napierian, or natural, logarithm, and $\frac{1}{\log(x_K/x_{K-1})}$ normalizes the Jeffrey's distribution to unity within the bin interval.

8.4 Examples

8.4.1 Convergence for a two-valued alternating sequence

We here study the performance of the proposed adaptive approximate inference for a case with $N = 100$ samples per block of data. The data were generated so that each data block consists of 50 samples taking the value $i = 1$ and 50 samples of value $i = 7$, i.e. there are only two values and they occur with equal frequency. An approximate inference is carried out on the interval of integers between 0 and 100. Using $K = 6$ bins, and an initial uniform partition over the integer interval $0...100$, we let the partitioning be updated based on the relative frequencies for the bins according to the algorithm laid out in Section 8.2. Figure 8.1 shows the probabilities for each bin after each of the first five updates and Table 8.1 lists the resulting repartitioning of the bins. The bins quickly concentrate around $i = 1$ and $i = 7$, the only bins where any activity is registered, leaving larger implausible values nearly unattended. After the first update the expectation of i becomes 9.9, after the second and the later updates the expectation is between 4 and 5, near the arithmetic mean $(7 + 1)/2 = 4$ of the sequence.

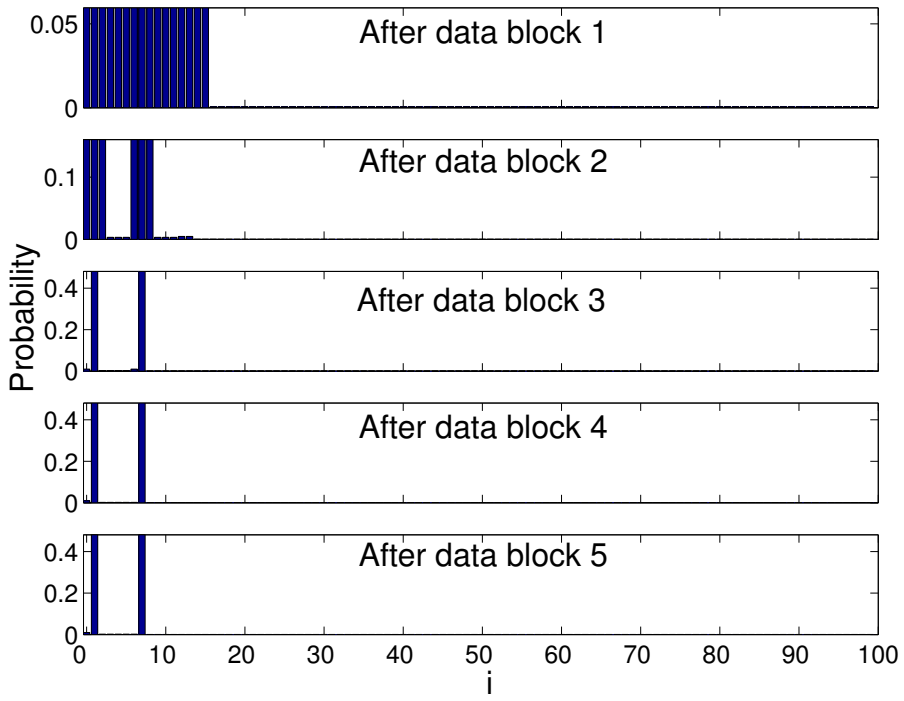

Figure 8.1: The evolution of the probabilities in each bin based on a quantized probability distribution in an example where each block of $N = 100$ samples contained only two values, $i = 1$ and $i = 7$, occurring with exactly the same frequency. The convergence is quick and nearly all attention is focused around the two observed values.

8.4.2 Approximating a Rayleigh distribution

Using $K = 4$ bins, the approximate inference is here tested on samples generated from a Rayleigh random-number generator with parameter $\gamma = 10$, yielding an expected value of 12.53. Each observed data block consists $N = 100$ samples, and the approximate inference is carried out on a range of integers between 0 and 50. Running the simulation repeatedly, we have found that the expectation obtained from the approximate inference after having observed 3 blocks of data ranges between 11 and 17 (depending on the particular number sequence generated). Figure 8.2 plots the probabilities in each bin and the new bin partition after observation of 3 blocks for one particular simulation. In this case the expectation obtained from the approximation was 12.27.

Moreover, varying the number of bins K, we obtained almost exactly the same performance for all $K > 2$. Changing the block length to $N = 10$, the performance

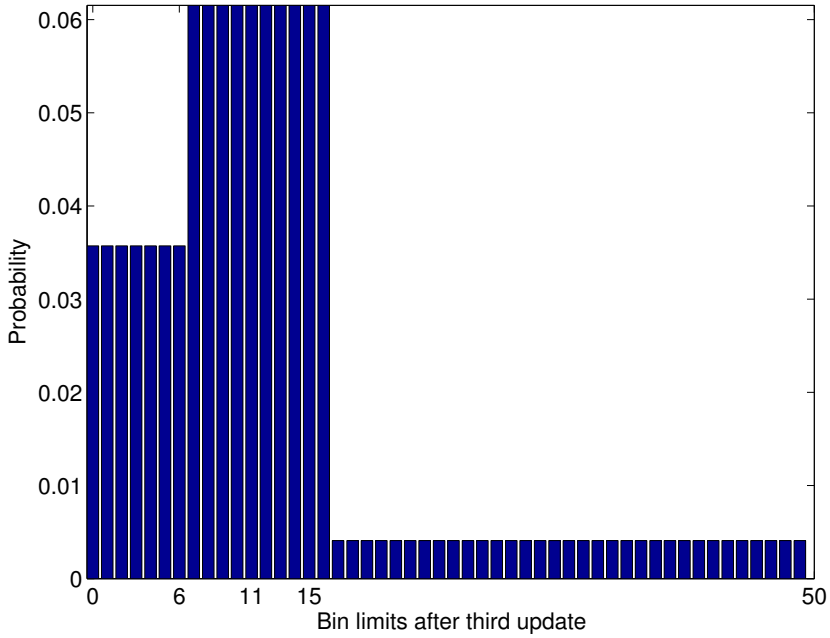

Figure 8.2: The bin probabilities and the new bin limits ($K = 4$ bins) after the third update in a scenario where the approximate inference was run on data blocks of size $N = 100$ produced by a Rayleigh random-number generator.

was nearly the same. Slightly higher variability of the approximate expectations could be detected due to the small number of samples, but the difference was very small.

8.5 Comments

We have so far only discussed the one-dimensional case. The criterion to distribute bins so that all bins have as equal probability mass as possible generalizes straightforwardly to the multi-variable case. The problem however lies in constructing a simple and effective algorithm for repartitioning the bins after a block of observations. The simplest approach would be to use the algorithm given above independently on each variable with a constant number of bins for each dimension. We would however expect to attain much higher approximation accuracy if we repartition bins more flexibly to take advantage of dependencies between different dimensions. On the other hand, a more flexible reallocation would generally have higher computational requirements as well. A challenge for future research

is to find an algorithm with a bin geometry constraint which is flexible enough to provide high accuracy for general dependencies and yet allows for low-complexity implementation. We suspect that solutions to this problem may already be accessible in the general mathematical literature, but have so far not found good candidate algorithms.

Another important direction for future research is finding means for taking time dependencies into account. In many cases, a quantity of interest evolves continuously over time under constraints on velocity and acceleration. It would greatly generalize the method suggested here if we could include simple time-dependent behavior into the model. A natural starting-point would be to include block-wise correlations and use the observed data to find a probability distribution for the possible correlations. By marginalizing over this distribution, we should be able to make better inferences when there is some dynamical process with constant parameters that generates our observations.

Appendix A

Some Integrals Related to the Gaussian Distribution

In many chapters in this thesis, we require a solution to an integral of the sort

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} g(x-c)dx \tag{A.1}$$

where c is a constant and

$$g(x) \quad = \begin{cases} x & , \ x > 0 \\ 0 & , \ x \leq 0 \ , \end{cases} \tag{A.2}$$

Using the definition of $g(\cdot)$, we rewrite the integral (A.1) as

$$I = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-c)dx \tag{A.3}$$

which is the difference between two terms, $I = I_2 - I_1$, with

$$I_1 = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} cdx \ . \tag{A.4}$$

and

$$I_2 = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} xdx \tag{A.5}$$

Let us now evaluate the first integral. Rewriting I_1 as

$$I_1 = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) c \int_{c}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x^2 - 2x\mu)\right\} dx \tag{A.6}$$

and using the standard formula[1] (eqn. 3.322.1 in Gradshteyn and Ryzhik, 2000)

$$\int_u^\infty \exp\left(-\frac{x^2}{4\beta} - \gamma x\right) dx = \sqrt{\pi\beta}\exp(\beta\gamma^2)\left[1 - \text{erf}\left(\gamma\sqrt{\beta} + \frac{u}{2\sqrt{\beta}}\right)\right] \quad [\text{Re}\beta > 0,\ u > 0]\ , \tag{A.7}$$

where

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt \tag{A.8}$$

is the error function, we find that

$$I_1 = \frac{c}{2}\left(1 + \text{erf}\left(\frac{\mu - c}{\sqrt{2}\sigma}\right)\right)\ . \tag{A.9}$$

The second part of (A.1), I_2, is obtained by integrating by parts. Defining

$$\begin{aligned} F(x) &= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\ &= \frac{1}{2}\text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\ , \end{aligned} \tag{A.10}$$

where the second equality is obtained directly from the definition of the error function (A.8), we have

$$I_2 = [xF(x)]_c^\infty - \int_c^\infty F(x)dx\ . \tag{A.11}$$

Using the relation (eqn. 5.41 Gradshteyn and Ryzhik, 2000)

$$\int \text{erf}(ax)dx = x\text{erf}(ax) + \frac{1}{a\sqrt{\pi}}e^{-a^2x^2} \tag{A.12}$$

we obtain

$$\int_c^\infty F(x)dx = \left[(x-\mu)F(x) + \frac{\sigma}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\right]_c^\infty\ . \tag{A.13}$$

Inserting this result into (A.11) gives

$$\begin{aligned} I_2 &= \left[\mu F(x) - \frac{\sigma}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\right]_c^\infty \\ &= \frac{\mu}{2}\left(1 - \text{erf}\left(\frac{c-\mu}{\sqrt{2}\sigma}\right)\right) + \frac{\sigma}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(c-\mu)^2\right\}\ . \end{aligned} \tag{A.14}$$

[1] There is an unfortunate double definition of a function $\Phi(x)$ in Gradshteyn and Ryzhik (2000) which may easily mislead the reader. In equation 3.321 it is first defined as $\Phi(x) = \frac{\sqrt{\pi}}{2}\text{erf}(x)$ while everywhere else in the book, including the equations immediately following 3.321, it is defined as (see Section 8.25) $\Phi(x) = \text{erf}(x)$. The latter definition is the correct one in our case. This error does not appear in earlier editions of the book.

Finally, we obtain

$$\begin{aligned} I &= I_2 - I_1 \\ &= \frac{\mu - c}{2}\left(1 - \text{erf}\left(\frac{c-\mu}{\sqrt{2}\sigma}\right)\right) + \frac{\sigma}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}\left(c-\mu\right)^2\right\} \end{aligned} \quad \text{(A.15)}$$

as the solution to the integral (A.1).

Bibliography

S. M. Alamouti. A simple transmitter diversity scheme for wireless communications. *IEEE J. Selected Areas in Communications*, 16:1451–1458, October 1998.

S. M. Alamouti and S. Kallel. Adaptive trellis-coded multiple-phase-shift keying for rayleigh fading channels. *IEEE Transactions on Communications*, 42(6): 2305–2314, June 1994.

M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. CDMA data QoS scheduling on the forward link with variable channel conditions. Technical report, Bell Labs Tech. Memo, April 2000.

M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting. Providing quality of service over a shared wireless link. *IEEE Communications Magazine*, 39(2):150–154, February 2001.

A. Bedekar, S. Borst, K. Ramanan, P. Whiting, and E. Yeh. Downlink scheduling in CDMA data networks. In *IEEE Globecom'99*, December 1999.

M. Bengtsson. Jointly optimal downlink beamforming and base station assignment. In *ICASSP - 2001*, May 2001.

D. Bernoulli. Specimen theoriae novae de mensura sortis. In *Commentarii Academiae Scientiarum Imperialis Petropolitanae, Tomus V*, pages 175–192. 1738. Translated to English by L. Sommer, published in *Econometrica* vol. 22, Issue 1 (Jan., 1954), pp 23-36.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

J.-Y. Le Boudec. Rate adaptation, congestion control and fairness: A tutorial. Technical report, Ecole Polytechnique Fédérale de Lausanne (EPFL), October 2003.

P. Brucker, A. Drexl, R. Möhring, K. Neumann, and E. Pesch. Resource-constrained project scheduling: Notation, classification, models, and methods. *European Journal of Operational Research*, (112):3–41, 1999.

P. W. Buchen and M. Kelly. The maximum entropy distribution of an asset inferred from option prices. *The Journal of Financial and Quantitative Analysis*, 31(1): 143–159, March 1996.

J. P. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Stanford University, 1975. Proc. 37th Meet. Soc. Exploration Geophysicists, 1967.

G. Caire and S. Shamai Shitz. On the achievable throughput of a multiantenna Gaussian broadcast channel. *IEEE Transactions on Information Theory*, 49(7): 1691–1706, July 2003.

Y. Cao and V. O. K. Li. Scheduling algorithms in broad-band wireless networks. *Proceedings of the IEEE*, 89(1):76–87, January 2001.

N. Casimiro Ericsson. On scheduling and adaptive modulation in wireless communications. Licentiate Thesis, Signals & Systems Group, Uppsala University, June 2001.

N. Casimiro Ericsson. *Revenue Maximization as a Criterion for Resource Allocation in Wireless Communications*. PhD thesis, Uppsala University, Signals and Systems, October 2004. under preparation.

N. Casimiro Ericsson, S. Falahati, A. Ahlén, and A. Svensson. Hybrid type-II ARQ/AMS supported by channel predictive scheduling in a multi-user scenario. In *IEEE VTC Fall 2000*, September 2000.

A. F. Chalmers. *What is this thing called Science?* Hackett Publishing Company, third edition, 1999.

J. Chuang and N. Sollenberger. Beyond 3G: Wideband wireless data access based on OFDM and dynamic packet assignment. *IEEE Communications Magazine*, 38(7):78–87, July 2000.

J. C.-I. Chuang and N. Sollenberger. Spectrum resource allocation for wireless packet access with application to advanced cellular internet service. *IEEE Journal on Selected Areas in Communications*, 16(6):820–829, August 1998.

S. T. Chung and A. J. Goldsmith. Degrees of freedom in adaptive modulation: A unified view. *IEEE Transactions on Communications*, 49(9):1561–1571, September 2001.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, January-February 1946.

G. J. Daniell and S. F. Gull. The maximum entropy algorithm applied to image enhancement. *Proceedings of the IEEE*, 5(127):170, 1980.

R. Dawkins. *The Selfish Gene*. Oxford University Press, 1976. second edition 1989.

T. Ekman. *Prediction of Mobile Radio Channels – Modeling and Design*. PhD thesis, Uppsala University, Signals and Systems, October 2002.

T. Ekman, M. Sternad, and A. Ahlén. Unbiased power prediction on broadband channels. In *IEEE VTC Fall 2002*, September 2002.

S. Falahati, A. Svensson, T. Ekman, and M. Sternad. Adaptive modulation systems for predicted wireless channels. *IEEE Transactions on Communications*, To Appear 2004.

S. Falahati, A. Svensson, M. Sternad, and H. Mei. Adaptive trellis-coded modulation over predicted flat fading channels. In *VTC 2003 Fall*, 2003.

W. Feller. *An Introduction to Probability Theory and Its Applications, Volume I*. John Wiley & Sons, third edition, 1968.

F. Florén, O. Edfors, and B.A. Molin. The effect of feedback quantization on the throughput of a multiuser diversity scheme. In *IEEE Globecomm 03*, December 2003.

S. Floyd and V. Paxson. Difficulties in simulating the internet. *IEEE/ACM Transactions on Networking*, 9(4):392–403, August 2001.

R.H. Frenkiel, B.R. Badrinath, J. Borras, and R. Yates. The infostations challenge: Balancing cost and ubiquity in delivering wireless data. *IEEE Personal Communications Magazine*, 7(2):66–71, April 2000.

L. Friedman. A competitive-bidding strategy. *Operations Research*, 4(1):104–112, February 1956.

K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Morgan Kaufmann, Academic Press, second edition, 1990.

D. Gesbert and M.S. Alouini. Selective multi-user diversity. In *IEEE ISSPIT 03*, December 2003.

I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, sixth edition, 2000.

J. L. Gruver, J. Aliaga, H. A. Cerdeira, and A. N. Proto. Nontrivial dynamics induced by a Jaynes-Cummings Hamiltonian. *Physics Letters A*, 190(5-6):363–369, August 1994.

S. F. Gull and G. J. Daniell. Image reconstruction from incomplete and noisy data. *Nature*, 272:686, 1978.

F. S. Hillier and G. J. Lieberman. *Introduction to Operations Research*. McGraw-Hill, fifth edition, 1990.

D. Howie. *Interpreting Probability: Controversies and Developments in the Early Twentieth Century*. Cambridge University Press, 2002.

E. T. Jaynes. Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630, May 1957a.

E. T. Jaynes. Information theory and statistical mechanics II. *The Physical Review*, 108(2):171–190, October 1957b.

E. T. Jaynes. Information theory and statistical mechanics. In K. W. Ford, editor, *Statistical Physics*, pages 181–218. W. A. Benjamin, 1963a.

E. T. Jaynes. New engineering applications of information theory. In Bogdanoff and Kozin, editors, *Engineering Uses of Random Function Theory and Probability*, pages 163–203. Wiley, 1963b.

E. T. Jaynes. On the rationale of of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, September 1982.

E. T. Jaynes. *Probability Theory – The Logic of Science*. Cambridge University Press, March 2003.

H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford University Press, first edition, 1939. (later editions 1948, 1961, 1967, 1988, 1998).

J. Jiang, R. M. Buehrer, and W. H. Tranter. Antenna diversity in multiuser data networks. *IEEE Transactions on Communications*, 52(3):490–497, March 2004.

N. Jindal, S. Vishwanath, and A. Goldsmith. On the duality of Gaussian multiple-access and broadcast channels. *IEEE Transactions on Information Theory*, 50 (5):768–783, May 2004.

M. Johansson. Benefits of multiuser diversity with limited feedback. In *IEEE SPAWC 03*, June 2003.

W. T. Grandy Jr. Principle of maximum entropy and irreversible processes. *Physics Reports*, 62(3):175–266, July 1980.

I. Katzela and M. Naghshineh. Channel assignment schemes for cellular mobile telecommunications systems: a comprehensive survey. *IEEE Personal Communications*, 3(3):10–31, June 1996.

R. Knopp. *Coding and Multiple-Access over Fading Channels*. PhD thesis, Swiss Federal Institute of Technology (Lausanne), Dept. of Electrical Engineering, 1997.

R. Knopp and P.A. Humblet. Information capacity and power control in single-cell multiuser communications. In *IEEE ICC 95*, June 1995.

T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1970.

S. Kullback. *Information Theory and Statistics*. Dover Publications, second edition, 1968.

J. Li, Y. Lee, H. Kim, and Y. Kim. Adaptive resource allocations based broadband wireless OFDMA systems with macro transmit diversity for downlink in cellular communications. In *7th WWRF meeting*, December 2002.

E. G. Negenman. Local search algorithms for the multiprocessor flow shop scheduling problem. *European Journal of Operational Research*, (128):147–158, 2001.

J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7(4):308–313, January 1965.

H. Nyquist. Certain topics in telegraph transmission theory. *Transactions of the A. I. E. E.*, pages 617–644, February 1928. reprinted in Proceedings of The IEEE, vol. 90, no. 2, Feb. 2002.

B. Penz, C. Rapine, and D. Trystram. Sensitivity analysis of scheduling algorithms. *European Journal of Operational Research*, (134):606–615, 2001.

J. G. Proakis. *Digital Communications*. McGraw-Hill, third edition, 1995.

X. Qiu, K. Chawla, J. C.-I. Chuang, and N. Sollenberger. Network-assisted resource management for wireless data networks. *IEEE Journal on Selected Areas in Communications*, 19(7):1222–1234, July 2001.

F. Rashid-Farrokhi, L. Tassiulas, and K. J. R. Liu. Joint optimal power control and beamforming in wireless networks using antenna arrays. *IEEE Transactions on Communications*, 46(10):1313–1324, October 1998.

L. H. Roberts. A discipline for the avoidance of unnecessary assumptions. *ASTIN Bulletin*, 5(3):205–217, 1971.

R. Rosenfeld. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech and Language*, 10(3):187–228, 1996.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

D. S. Sivia. *Data Analysis – A Bayesian Tutorial*. Clarendon Press, 1996.

C. R. Sox, P. L. Jackson, A. Bowman, and J. A. Muckstadt. A review of the stochastic lot scheduling problem. *International Journal of Production Economics*, (62):181–200, 1999.

L. Tassiulas and A. Ephremides. Allocation of a single server to a set of parallel queues with time dependent demands. In *IEEE ISIT*, June 1991.

L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, December 1992.

D. N. Tse. Optimal power allocation over parallel Gaussian broadcast channels. In *IEEE ISIT*, June 1997. An unpublished more detailed version is available at http://degas.eecs.berkeley.edu/∼dtse/pub.html.

D. N. Tse. Multiuser diversity in wireless networks, April 2001. Presentation at Stanford University.

H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley & Sons, 1968.

R. Verdone and A. Zanella. Performance of received power and traffic-driven handover algorithms in urban cellular networks. *IEEE Wireless Communications*, 9 (1):60–70, February 2002.

P. Viswanath and D. N. C. Tse. Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. *IEEE Transactions on Information Theory*, 49(8):1912–1921, August 2003.

P. Viswanath, D. N. C. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, 48(6):1277–1294, June 2002.

W. Wang, T. Ottosson, M. Sternad, A. Ahlén, and A. Svensson. Impact of multiuser diversity and channel variability on adaptive OFDM. In *VTC Fall 2003*, October 2003a.

W. Wang, T. Ottosson, M. Sternad, A. Ahlén, and A. Svensson. Impact of multiuser diversity and channel variability on adaptive OFDM. In *VTC 2003 Fall*, 2003b.

A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, 1971.

J. Zhang, M. Hu, and N. B. Shroff. Bursty data over CDMA: MAI self similarity, rate control and admission control. In *IEEE Infocom*, 2002.