

## Dissertation

# Resource Allocation in Wireless Communications under Time Constraints

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines  
Doktors der technischen Wissenschaften

unter der Leitung von  
Univ.-Prof. Dipl.-Ing. Dr. Norbert Görtz  
Institute of Telecommunications

eingereicht an der Technischen Universität Wien  
Fakultät für Elektrotechnik

von  
Dipl.-Ing. Johannes Gonter  
Johann Kaller Gasse 14  
A-2201 Gerasdorf bei Wien

Wien, im Januar 2016

---



Die Begutachtung dieser Arbeit erfolgte durch:

**1. Univ.-Prof. Dipl.-Ing. Dr. Norbert Görtz**

Institute of Telecommunications

Technische Universität Wien

**2. Univ.-Prof. Dipl.-Ing. Dr. Hermann Hellwagner**

Institute of Information Technology

Alpen-Adria-University Klagenfurt





—— To Irmí, Hannah, and Christoph ——



# Abstract

This thesis explores strategies for resource allocation in the wireless channel under time constraints. These are considered in two different manifestations, leading to a two-part structure of the thesis.

1) The first part explores delay-constrained rate requests in multiuser downlink scenarios. Although this topic has been explored repeatedly in the past, there is a lack of approaches that are practical from a computational complexity perspective, and flexible enough to adapt to a wide variety of channel conditions, rate constraints and delay limitations. The first contribution of this thesis is therefore the discussion of properties of widespread resource allocation strategies with respect to their suitability for delay-constrained transmission of data. Current implementations of scheduling algorithms generally lack information about rate demands and delay requirements of higher layers in the OSI model. As long as the channel quality does not change too quickly, applications typically adapt their demands to the current possibilities the communication system provides. This is, however, also a strategy to compensate for the inability of the scheduler to serve all requests in time due to the lack of commonly implemented information exchange between the application layer and the scheduler. Therefore, the very different limitations regarding required rate and acceptable delay for normal web-browsing or video streaming cannot be considered in resource allocation. This thesis therefore suggests a new concept - Power-Controlled Cross-Layer Scheduling (PCCLS) - that exploits exactly this information, opening up a range of possibilities that are not available in today's standard implementations. Compared to proportional fair (PF) scheduling, PCCLS adds a control loop that compensates for slow changes in channel attenuation, thus providing a constant long-term channel quality that makes the achievement of specific rate- and delay- requirements possible. The faster block-fading is not compensated for, i.e. it can be taken advantage of. In addition, PCCLS excludes certain users from scheduling as soon as the required rate has been achieved in a given delay-window. This is done even when the transmit buffer still provides data to be transmitted, freeing up resources for other users. Through numerical simulations, PCCLS is shown to achieve the given rate- and delay- constraints for a general scenario based on realistic requirements. Proportional fair scheduling, on the other hand, is demonstrated to spend too much energy on users with low rate requests, which leads to a failure regarding the requested rates of the more demanding users.

The second main contribution in this first part of the thesis derives the probability of an outage

for resource allocation schemes that are based on a memoryless selection of the user with the largest scheduling metric. Well-known members of this species of scheduling schemes are the proportional fair- and the opportunistic scheduler. The analysis does not constrain the number of users, their respective channel (fading) statistics or rate demands. The general derivation is subsequently specialized to the original implementations of proportional fair and opportunistic scheduling; simulations demonstrate the perfect match between analytic and numeric approaches. Finally, it is shown that accepting more relaxed delay requirements results in a huge reduction of the experienced outage probability.

2) While the first part of this thesis explores strategies for efficient use of resources when there are user-imposed time limitations in the form of rate- and delay- constraints, the second part considers time limitations that originate from the channel. These are experienced in the form of periods where a transmission is economically feasible, and intensified from an algorithmical point of view when the channel is not ergodic during these windows of opportunity. In practice, these scenarios can be observed in ad-hoc communications, specifically between vehicles on opposing highway lanes or in body-area communications. Sum-power constraints directly lead to a waterfilling solution, raising the question for the optimum transmit decision threshold (i.e., the “waterlevel”) when the future evolution of the channel is unknown. A new approach to directly calculate this decision threshold is derived, and its properties are explored analytically. Subsequently, the gathered insights are used to compare acausal, causal and causal-adaptive implementations regarding efficient use of the limited transmit power budget. The second part of this thesis concludes with the conceptual discussion of a strategy for recognizing transmit opportunities when the channel is available during recurring time windows.

# Zusammenfassung

Diese Dissertation betrachtet Strategien zur Nutzung von Ressourcen bei drahtloser Übertragung unter Berücksichtigung von Zeitbeschränkungen. Die Fokussierung auf zwei Arten der Zeitbeschränkung führt zur Struktur einer zweiteiligen Dissertation.

1) Der erste Teil diskutiert verzögerungsbegrenzte Übertragung an mehrere Nutzer unter der Bedingung garantierter Übertragungsraten, bzw. unter Berücksichtigung der Unterbrechungswahrscheinlichkeit. Obwohl diese Themen in der Vergangenheit bereits diskutiert wurden besteht ein Mangel an ökonomisch implementierbaren Strategien die flexibel genug sind, sich an unterschiedlichste Übertragungsbedingungen, Raten- und Verzögerungsbeschränkungen anzupassen. Der erste Beitrag dieser Dissertation besteht also in der Diskussion bekannter Ressourcenzuteilungsstrategien hinsichtlich ihrer Eignung für raten- und verzögerungsbeschränkte Übertragung an mehrere Teilnehmer. Gegenwärtige Implementierungen leiden an mangelndem Informationsaustausch zwischen der Applikationsschicht und der Steuerung der Ressourcenzuteilung. So lange sich die Übertragungsbedingungen nicht zu schnell ändern, passt typischerweise die Applikation ihre Raten- und Verzögerungsanforderungen an die gegenwärtigen Möglichkeiten an. Ein allgegenwärtiges Beispiel für eine derartige Strategie ist die Anpassung der Auflösung und des Zwischenspeichers bei Videoübertragung. Diese Strategie ist jedoch zum Teil nur notwendig, weil die Ressourcenzuteilung keine (genauen) Informationen über die momentanen Anforderungen der jeweiligen Applikation hat. In der Praxis können daher massive Unterschiede bei der Toleranz von Übertragungsgeschwindigkeit und Verzögerung bei so verschiedenen Anwendungen wie Videoübertragung und Internetsurfen nicht berücksichtigt werden. Dies motiviert die Einführung ebenjener Möglichkeit durch ein neues Ressourcenzuteilungskonzept “Power-Controlled Cross-Layer Scheduling” (PCCLS). Verglichen mit der in vielen Varianten eingesetzten “Proportional Fair” (PF) - Strategie wird bei PCCLS langsame Kanalvariation durch die Leistungsregelung kompensiert, während die schnellere, übertragungsblockweise Variation für einen ökonomischen Ressourceneinsatz bei der Übertragung genutzt wird. Zusätzlich gewinnt PCCLS im Vergleich zu PF an Effizienz da Teilnehmer, die ihre geforderte Übertragungsrate im betrachteten Zeitfenster erreicht haben, kurzzeitig von der Ressourcenzuweisung ausgenommen werden. Dies passiert im Gegensatz zu gegenwärtig implementierten Algorithmen auch dann, wenn der Übertragungszwischenspeicher des betreffenden Teilnehmers nicht leer ist. Mittels Simulationen wird gezeigt, dass PCCLS energie-

effizienter arbeitet und speziell Teilnehmer mit hohen Anforderungen an die Übertragungsrate deutlich besser bedient werden, da im Gegensatz zu PF keine Übertragungsenergie an Teilnehmer mit kleineren Anforderungen verschwendet wird.

Der zweite Beitrag im ersten Teil dieser Dissertation diskutiert die Frage nach der Unterbrechungswahrscheinlichkeit einer Übertragung mit einer gewissen Ratenforderung, wenn der Ressourcenzuteilungsalgorithmus die Teilnehmer auf Basis einer gedächtnislosen *max*-Entscheidung bedient. Diese Herangehensweise ist eine verallgemeinerte Formulierung für die Funktionsweise des PF- und des opportunistischen Zuteilungsschemas. Die mathematische Analyse beschränkt weder die Teilnehmeranzahl, deren Kanalstatistik (so lange sie ergodisch ist) oder deren Anforderung an Übertragungsrate und Verzögerung. Die verallgemeinerte Diskussion wird in weiterer Folge auf die beiden genannten Ressourcenzuteilungsstrategien angewendet, und der Vergleich mit Simulationen zeigt eine perfekte Übereinstimmung. Schlussendlich wird der Effekt mehr oder weniger entspannter Verzögerungsanforderungen veranschaulicht.

2) Während im ersten Teil dieser Dissertation auf Zeitbeschränkungen eingegangen wird die von den Teilnehmern vorgegeben werden, ist der zweite Teil jenen Beschränkungen gewidmet, die vom Kanal vorgegeben werden. Diese bestehen immer dann, wenn eine Übertragung nur in gewissen Zeiträumen ökonomisch sinnvoll oder überhaupt möglich ist. Eine zusätzliche algorithmische Herausforderung erwächst hier aus der Annahme, dass der Übertragungskanal in diesen Zeitfenstern nicht ergodisch ist. In der Praxis sind solche Szenarien vor allem bei ad-hoc Kommunikation zu beobachten, zum Beispiel bei Übertragungsszenarien auf Straßen zwischen Autos auf entgegengesetzten Fahrbahnen oder beim Datenaustausch körpernaher Kommunikationssysteme mit kurzer Reichweite. Eine Beschränkung der gesamten Übertragungsenergie führt hier zu schnell zu einer “Waterfilling”-Lösung, was wiederum die Frage nach einer optimalen Entscheidungsschwelle (“Waterlevel”) aufwirft, wenn die zukünftige Entwicklung des Kanals unbekannt ist. Zu diesem Zweck wird eine neue Lösung des Waterfilling-Problems entwickelt und analysiert, die diese Entscheidungsschwelle direkt ermittelt. In weiterer Folge werden die gewonnenen Erkenntnisse genutzt um akausale, kausale und adaptiv-kausale Strategien zur optimalen Nutzung des zeitlimitierten Übertragungskanals zu entwickeln und zu bewerten. Der zweite Teil dieser Dissertation schliesst mit der konzeptuellen Diskussion eines Algorithmus, der die besten Übertragungsgelegenheiten in wiederkehrenden “Zeitfenstern” erkennt und damit eine effiziente Nutzung der Ressource “Energie” ermöglicht.

# Acknowledgements

I would like to thank Prof. Norbert Görtz for his support and belief from the very beginning, countless hours of discussion and his patience.

I also want to express my thanks to Prof. Hermann Hellwagner for reviewing this thesis.

Furthermore, I am grateful to my colleagues at the institute: Ing. Wolfgang Gartner, BSc., Geetha Ramachandran, Dr. Gerhard Doblinger, Dr. Georg Kail, Dr. Hua Zhou and Dr. Mehdi Mortazawi-Molu for making the years we spent together truly unforgettable.

My gratitude belongs to my parents for their confident support.

Finally, I want to thank my wife Irmi, and my children Hannah and Christoph for their unconditional love and care - you mean everything to me.





# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Limited Resources in Wireless Multiuser Communications . . . . .	1
1.2. Outline of This Thesis . . . . .	4
1.3. Contributions . . . . .	5
1.4. Publications Associated with this Thesis . . . . .	6
 <b>I. Part I: Quality of Service in Infinite Lifetime Channels</b>	 <b>7</b>
<b>2. QoS in Infinite Lifetime Channels</b>	<b>9</b>
2.1. Requested Rates and Outage . . . . .	11
2.2. Considered Fundamental Preceding Literature . . . . .	12
 <b>3. A Scheduling Algorithm for Small Outage</b>	 <b>19</b>
3.1. System Model . . . . .	19
3.2. Information Theoretical Limits . . . . .	20
3.3. Popular Resource Allocation Schemes and their Suitability for low Outage . . . . .	21
3.3.1. Round Robin Scheduling and Modifications . . . . .	21
3.3.2. Channel-Adaptive Time-Division Access with Power Allocation by “Water-filling” (TD-W) . . . . .	24
3.3.3. TD-Waterfilling: Modification A . . . . .	27
3.3.4. Discussion of TD-Waterfilling Mod. A . . . . .	29
3.3.5. Modified Scheme (B) . . . . .	31
 <b>4. Power-Controlled Cross-Layer Scheduling</b>	 <b>37</b>
4.1. Scheduling Algorithm Details . . . . .	38
4.1.1. Power Scaling . . . . .	40
4.2. Comparison to Proportional Fair Scheduling . . . . .	41
4.3. Numerical Results . . . . .	42
4.4. Outage Analysis . . . . .	46

<b>5. Outage Analysis for MAX-Based Scheduling Schemes</b>	<b>53</b>
5.1. Channel Model . . . . .	54
5.2. Outage Probability of <i>max</i> -Based Scheduling . . . . .	55
5.3. Results for Specific Schedulers and Channels . . . . .	62
5.3.1. Opportunistic Scheduling . . . . .	62
5.3.2. Proportional Fair Scheduling . . . . .	64
<b>6. Conclusion of Part I</b>	<b>67</b>
 <b>II. Resource Allocation for Finite Lifetime Channels</b>	 <b>69</b>
<b>7. Limits on Information Transmission for the Finite Lifetime Channel</b>	<b>71</b>
7.1. Problem Statement . . . . .	71
<b>8. The Waterfilling Solution</b>	<b>75</b>
8.1. Normalisation of the Solution . . . . .	76
<b>9. Fixed-Point computation of <math>\lambda</math></b>	<b>79</b>
9.1. Structure of the Fixed-Point Function . . . . .	82
9.1.1. Derivation of $ \tilde{h}(j) ^2$ . . . . .	83
9.1.2. Derivation of $J(\lambda)$ . . . . .	84
9.1.3. The Fixed-Point Function for <i>large N</i> . . . . .	85
9.1.4. Value of the Fixed-Point Function for extreme $\lambda$ . . . . .	86
9.1.5. Relation between Fixed-Points and the First Derivative of $\tilde{\varphi}(\lambda)$ . . . . .	89
9.1.6. The Fixed-Point Function $\tilde{\varphi}(\lambda)$ in a Neighborhood of the Fixed-Point $\lambda_i$ . .	91
9.1.7. The Converse of Section 9.1.5 . . . . .	95
9.2. Proposition and Conclusion . . . . .	95
<b>10. Numerical Simulations</b>	<b>97</b>
10.1. Channel- and System-Model . . . . .	97
10.2. Convergence Behavior . . . . .	105
10.3. Performance Comparison with the “Classical Approach” (Direct Calculation of the Transmit Powers) . . . . .	108
10.3.1. Fixed-Point Calculation of the Waterlevel . . . . .	108
10.3.2. Direct Calculation of Optimum Transmit Powers . . . . .	109
10.3.3. Comparison of the Algorithms . . . . .	111
10.3.4. Conclusion . . . . .	113

<b>11. Power Allocation Strategies for the Finite Lifetime Channel</b>	<b>115</b>
11.1. Acausal Knowledge of the Channel Coefficients . . . . .	115
11.2. Causal Knowledge of the Channel Coefficients . . . . .	116
11.3. Simulation Results . . . . .	117
11.4. Conclusion . . . . .	124
<b>12. Use of the Fixed-Point Algorithm in Scenarios of Sporadically Available Channels</b>	<b>127</b>
12.1. EWMA-Triggered Waterfilling . . . . .	128
12.2. Estimation of the Implementation Complexity . . . . .	133
<b>13. Conclusion of Part II</b>	<b>135</b>
<b>Appendix A. Arithmetic Considerations around a Function Maximum</b>	<b>137</b>
<b>Appendix B. Derivation of <math>\tilde{\varphi}(\lambda)</math></b>	<b>139</b>
<b>Bibliography</b>	<b>141</b>



# 1. Introduction

This introductory chapter contains a general problem statement that is setting the stage for this thesis' objectives and research topics. It provides motivation through highlighting what is considered to be unresolved questions in the research field. The chapter then outlines the structure of the thesis, providing a brief overview of the main topics discussed. It is closing with an overview of main research contributions found in the thesis.

## 1.1. Limited Resources in Wireless Multiuser Communications

Wireless communications has, in recent years, become a widely used and therefore ubiquitously implemented technology - one that continues to penetrate into more and more applications of daily life. As such, it is no longer only an academically-discussed research topic but a living technology affecting life of many people. In 2013, the ICT Data and Statistics Division of the ITU [1] has published a forecast estimating the number of mobile-cellular subscriptions in the same year to surpass 6.8 billion, with a growth rate of approximately 100% since 2007. In May 2015, this forecast number has increased to over 7 billion cellular subscriptions according to the updated annual report by the ITU [2], which is a growth of nearly 9.5 times since the year 2000. In comparison, the whole human population is approximately 7.3 billion, although a reduction of the growth rate of mobile-cellular subscriptions should not be reasoned from a comparison of the two figures: with the increasing performance of cellular networks, wireless technologies will continue to replace wired connections even for throughput-demanding applications. This is backed by another figure just recently published by the ITU in the same document: in early 2015, the number of mobile broadband-subscriptions has climbed to nearly 3.5 billion, compared to about 807 million in 2010. These numbers reflect an increasing demand of customers for not only low-throughput services like voice telephony or SMS, but also high-throughput services like over-the-air video on demand. Mobile network providers offering and actively promoting such services therefore face an increasing pressure to optimally exploit their infrastructure in a market of reducing revenues and volatile customers, who are changing the provider if the advertised services are available with a better quality with a different provider.

## 1. Introduction

This change in demand has evolved quickly but gradually. Other developments, however, will happen much more quickly and be a consequence of new technologies that do not have a live counterpart at present. In wireless communications, this is certainly going to be the breakthrough of relaying and ad-hoc communications among (rapidly) moving users. In the very near term, vehicular communications (either vehicle-to-vehicle or vehicle-to-infrastructure) is going to enforce a new paradigm in wireless communications due to the demand for low latency and high reliability in an environment characterized by finite-lifetime channels. Traditional wireless networks cannot live up to these demands economically, due to delays introduced by the network infrastructure (e.g. routing, buffering, or scheduling), and the lack of guaranteed minimum Quality-of-Service (QoS), even if the communication partners are in close proximity of each other. This demand for implemented ad-hoc networking is going to emerge quickly, driven by car manufacturers who are advertising improved safety, and will soon thereafter become a must-have feature for all manufacturers who do not want to lose customers. In the wake of these developments, there will be new technologies introduced for pedestrian short-range communications as well, and open up a new mass-market for ad-hoc communication devices. In fact, there is a good reason for wireless network providers to build upon tested solutions (vehicle-to-vehicle communications) and relieve the demand on their main infrastructure by allowing ad-hoc connections among their customers. Although there are many open questions in this regard ranging from billing to security concerns, we are already seeing devices that would ultimately benefit from such an approach, for instance the experimentally-introduced Google Glass. This forerunner of a new class of wearable communication devices already showcases the demands that will have to be met for the three parties involved in driving this technology (the customer, the network provider and the equipment manufacturer):

- **New Content.** The customer as well as the equipment manufacturer will require new content for creating an augmented reality experience with real-time information about the near environment. This will cause new data traffic requirements: the exact location of the customer is suddenly of great importance. This can be derived either from satellite-based navigation systems which consume precious energy, or from ad-hoc connections to nearby communications partners who are generating and passing on information anyway. The challenge here is for network providers to incorporate ad-hoc communications into their business-models and technology.
- **Energy Efficiency.** Wearable devices will need to be very energy-efficient, given the small sizes and required unobtrusive looks. Especially head-wearable devices need to be as light as possible, yet they need to operate a whole day without intermediate charging. This requires energy-efficient transceiver designs and a network-infrastructure that supports energy-saving short-range communications with short periods of activity and long periods of power saving.

The challenge for the network providers is to keep the latency low despite the wide range of different devices that require different power-saving features for optimum energy efficiency.

- **Ad-Hoc Networking.** Augmented reality requires real-time information about the surrounding environment. The same applies to other wearable communication devices and vehicular communications. This challenges the development of intelligent ad-hoc communications that are exploiting the finite-lifetime characteristics of the typical wireless channel to improve energy efficiency and decrease interference with other devices. In augmented reality scenarios, it might not be necessary to exchange all information over the network's core infrastructure. Instead, for information about the near surrounding of the user, it might be more efficient to exchange information via ad-hoc short-range connections. This information would typically not change very frequently, nor would it be overly personalized. It might contain social media contact data, advertisements, updates regarding public transportation, or weather information.

In tomorrow's world of wireless communications, there is therefore growing pressure on equipment manufacturers as well as network operators to provide energy-efficient communications with high QoS, suited for real-time exchange of multimedia content. In a market situation where the customer expects a better price/performance ratio in each new product generation (be it network contracts or user equipment), the most efficient way to satisfy these demands is by optimized use of existing resources. Although this is already done on a network level - Self-Optimizing Networks are slowly emerging - there are other fields where conventional approaches are still widely taken. This thesis therefore aims to find answers concerning resource allocation strategies that meet demand in tomorrow's mobile communication scenarios.

Specifically, two main topics are addressed and subsequently dominate the two main parts of this thesis:

## 1.2. Outline of This Thesis

Part I: Outage as a measure of QoS in wireless communications: In modern communications systems, the ability to deliver a certain amount of data within a given time-span consistently is a crucial prerequisite for a satisfying consumption of multi-media content. Today's resource allocation schemes generally employ best-effort strategies that do not consider the actual demand of the bandwidth-requesting application and at the same time do not guarantee any minimum QoS. The first part of this thesis is therefore devoted to exploring QoS in popular scheduling-schemes and to introducing and discussing a new proportional-fair based scheduling scheme that takes into account the actual rate requirement.

Part II: Today's resource allocation algorithms assume continuous availability of the channel, thus denying time-limited access to communication partners as will be experienced in future ad-hoc communication scenarios. The second part of this thesis is therefore focused on the use of finite-lifetime channels, providing strategies to optimally exploit the limited resources for energy - efficient communications.



## 1.3. Contributions

The following main contributions to academic discussion have been published in peer-reviewed conference papers and are discussed in this thesis:

- Analytic outage analysis of max-based scheduling algorithms, without any assumptions that constrain the number of users, individual transmit powers for every single user to be served, or individual channel statistics for all users.
- A refined strategy for analytically and numerically assessing the average throughput for proportional fair scheduling.
- Specialization of the above analytics to opportunistic and proportional fair scheduling schemes.
- Discussion of a new proportional-fair based scheduling scheme without energy constraints. The scheme greatly reduces the probability of an outage and can therefore practically guarantee a minimum Quality-of-Service, something that existing schemes cannot provide.
- Strategies for optimal use of the finite-lifetime channel as experienced in vehicular or pedestrian ad-hoc communications.
- A new fixed-point algorithm for performing a maximization of transmitted information while being power constrained (waterfilling), and a mathematical analysis of its properties.
- An introduction of predictive waterfilling for use in finite-lifetime channel scenarios to reduce algorithmical complexity in resource allocation decisions.

## 1.4. Publications Associated with this Thesis

[3] J. Gonter, N. Görtz: *Power-Controlled Cross-Layer Scheduling*, in Proc. Proceedings of the IEEE International Conference on Communications (ICC 2013), Budapest, Hungary, June, 2013.

[4] J. Gonter, N. Görtz, A. Winkelbauer: *Analytical Outage Probability for MAX-Based Schedulers in Delay-Constrained Applications*, in Proc. Proceedings Wireless Days 2012, Dublin, Ireland, November, 2012.

[5] N. Görtz, J. Gonter: *Information Transmission over a Finite-Lifetime Channel under Energy-Constraints*, in Proc. Proceedings European Wireless 2011, Vienna, Austria, April, 2011.

[6] N. Görtz, J. Gonter: *Limits on Information Transmission in Vehicle-to-Vehicle Communication*, in Proc. Proceedings IEEE Vehicular Technology Conference Spring 2011, Budapest, Hungary, May, 2011.

[7] J. Gonter, N. Görtz: *An Algorithm for Highly Efficient Waterfilling with Guaranteed Convergence*, in Proc. Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on, Vienna, Austria, 2012.

[8] J. Gonter, N. Görtz, M. Rupp, W. Gartner: *EWMA-Triggered Waterfilling for Reduced-Complexity Resource Management in Ad-Hoc Connections*, in Proc. Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on, London, England, September, 2013.

Part I.

## Part I: Quality of Service in Infinite Lifetime Channels



## 2. QoS in Infinite Lifetime Channels

During the last decade, consumer communications has evolved away from speech and moved progressively towards the exchange of multimedia content. Together with more and more powerful user-equipment and emerging technologies like transmission of high-resolution 3D-video, this trend is and will be challenging even for the most elaborate infrastructure (including LTE advanced), particularly when high rates are requested with hard delay constraints, as e.g. in conversational video applications. There are hints that future contracts will be based on guaranteed data-rates rather than the absolute amount of data, which will lead to a much increased consumer-awareness with respect to Quality of Service (QoS) and Quality of Experience (QoE).

All of these developments call for new paradigms in wireless network management.

This first part of this thesis focuses on Quality of Service (QoS) in Infinite Lifetime Ergodic Channels (ILECs). Although all (wireless) communication is generally characterized by a start and an end and is therefore time-limited, there is a general consensus to consider wireless links *in static scenarios*<sup>1</sup> to be of infinite duration. Historically, this is motivated by the fundamentally necessary prerequisite of the formula for the channel-capacity to hold only for channels of time-unlimited availability. Although modern coding schemes like LDPC codes are indeed performing close to the channel capacity even for codewords comprised of some 1000 transmit symbols [9], ubiquitously implemented resource allocation strategies like opportunistic and proportional-fair scheduling have been designed to achieve certain goals in the long term. In opportunistic scheduling, only the user is served who experiences the most favorable conditions for a transmission. Typically, this will be the user with the shortest distance to the base station. This concept maximizes the cell-throughput, however, in its original form and intent it does not feature a functionality to consider the individual user's service requirements. Proportional fair scheduling, on the other hand, assigns resources to the user whose ratio of momentarily achievable rate and (recently) achieved average is highest. This is commonly referred to as a "fair" distribution of system resources, since the scheduling decision for each user will not take into account the absolute quality of the channel (where the users with less favorable channel conditions like for instance those at the cell-edge would hardly stand a chance). Instead, the relative quality of the channel, given the recent history, is compared

---

<sup>1</sup>In *static scenarios*, neither the transmitter nor the receiver or the environment is moving, so that the assumption of ergodicity is justified.

## 2. QoS in Infinite Lifetime Channels

among all competing users, allowing all of them to be assigned resources. These approaches will, however, allow no possibility to adapt the algorithm to the user’s needs concerning average rate and maximum delay in their original forms. Both of these resource allocation strategies are, therefore, ultimately best-effort schemes that optimize cell-wide parameters without providing the flexibility a modern wireless network requires. This ubiquitous implementation of these popular resource allocation strategies does, however, not mean that the disregard of user’s service requirements is an acceptable approach in today’s communication scenarios: in fact, quite the opposite is true. In today’s communication scenarios, the ability of a wireless network to deliver some service with a certain reliability is getting more and more attention. This is especially true for applications that require a certain data-rate to be delivered with a low probability of intermediate drops below a certain threshold, like video streaming or even speech transmission. It is these “multimedia”-applications that are conflicting with the originally intended purpose of most-widely implemented resource allocation strategies. The growing importance to deliver multimedia content with quality-of-service and therefore customer experience in mind has been illustrated by Intel in 2014, who published an infographic<sup>2</sup> to illustrate the average activity in one minute of today’s world wide web [10]. The following Table 2.1 summarizes the most important activities and tries to categorize these according to rate requirement and tolerance to delay:

Activity	Extent	Rate Requirement	Delay Tolerance
Youtube Views	1.3 Million	High	Low
Youtube Uploads	30h	High	Medium
Google Search Queries	+2 Million	Low	Medium
facebook Views	6 Million	High	Low
twitter New Tweets	100 Thousand	Low	High
flickr Photo Views	20 Million	High	Medium

Table 2.1.: Typical worldwide internet activity per minute as of 2014 according to Intel Corporation and a rough classification with respect to rate demand and delay tolerance for acceptable user experience.

As already pointed out in the introduction, these typical applications that are contributing a significant share of overall traffic (the same document states the overall volume of IP data per minute to be 639,800GB) will without any doubt continue to evolve towards sharing of multimedia content, increasing the demand even more.

---

<sup>2</sup><http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html>

## 2.1. Requested Rates and Outage

The ability of modern wireless communication systems to deliver some service with a certain reliability is often included in the notion of a “quality-of-service”. This is, however, a very flexibly used term that may mean entirely different things in different situations. The term “outage”, on the other hand, though very often seen as an aspect of QoS, describes the most crucial ability of a wireless multiuser system – the ability to deliver a certain amount of information within a given and usually limited period of time. Throughout this thesis, “outage” refers to the event that a required amount of information cannot be transmitted within a given amount of time, i.e. the actually achieved sum rate  $R_{sum}$ , which is a random number, is less than a determined required sum rate  $r_{sum}^{req}$ :

$$R_{sum} < r_{sum}^{req} . \quad (2.1)$$

The reason for an outage can generally be described as a lack of transmit power. If it is assumed that a channel will almost never completely block the transmission (the magnitude of the channel coefficient would have to be exactly 0), the only reason for a required amount of information not to be transmitted is clearly a power constraint. One can therefore approach outage from two different sides.

- Assuming unconstrained power, one can aim to find the use of resources that lead to an acceptably low outage. This thesis discusses a novel proportional-fair based resource allocation scheme that also takes into account rate requirements of the individual users, therefore called Power-Controlled Cross-Layer Scheduling (PCCLS).
- On the other side, assuming a power constraint in place, one can try to find the reliability of scheduling schemes. This thesis analytically derives a method to calculate the probability of an outage for max-based scheduling schemes such as opportunistic and proportional fair scheduling.

## 2.2. Considered Fundamental Preceding Literature

This section gives an overview on important publications on Quality-of-Service aspects in wireless scenarios. This list is by no means complete, but it contains the most widely cited corner-stone contributions that build the basis of today's state-of-knowledge. In order to facilitate insight into the evolution of the topic in scientific discussion, the contributions are discussed in the order of their appearance.

**[11]: McEliece, R.; Stark, W.E., "Channels with block interference," Jan 1984**

The authors contribute some insight into the information theoretic consequences of channels whose quality is not changing over time in an uncorrelated way, but can be described as being constant during  $m$  channel uses. Thus, the authors introduce (and also, for the first time, name) block interference channels. In their paper, the authors assume noise bursts as the reason for adverse transmit conditions, while also referring to *jamming* and *fading* as typical scenarios, in which the presented considerations will provide solutions to real-world problems. The authors investigate the information theoretic limits of block-interference channels for both receivers with and without CSI. It is shown that the capacity of a block-interference channel with side-information is the average capacity  $\bar{C}$  of the component channels, and that the capacity of the BI channel without side information at the receiver is less than  $\bar{C}$ , approaching  $\bar{C}$  for  $m \rightarrow \infty$ . Therefore, the authors conclude that  $C$  is indeed *independent* of  $m$  for channels with side information. This publication does consider the problems associated with the attempt to mitigate the effects of large  $m$  by the use of long convolutional codes, therefore briefly mentioning problematic coding delay. There is, however, no mention of delay-limited capacity or outage.

**[12]: Kaplan, G.; Shamai, S., "Error Exponents And Outage Probabilities For The Block-Fading Gaussian Channel," Sep 1991**

The authors investigate error exponents and outage probabilities for a channel that is slowly fading. The channel model is very similar to the one introduced in [11]. The authors are exploring information theoretic bounds and are, therefore, assuming perfect CSI at the receiver. The channel is exhibiting "block-fading" behavior, i.e. it is constant during  $K$  successive uses of the channel. The authors introduce a maximum codeword length  $N$  that represents the maximum tolerable delay and define  $N' = N/K$  as the number of uses of a "memoryless superchannel". Effectively  $N$  is therefore the number of fading states a codeword is spanning. An upper bound for the probability of an outage (that is defined as the instantaneous channel capacity being smaller than the transmission rate) is derived, and it is independent of  $K$ . The outage probability does, however, depend on  $N'$ , and it is shown that the maximum admissible rate for a given small outage probability of



1% is increased by a factor of 4 if  $N'$  is increased from 1 to 4.

**[13]: Ozarow, L.H.; Shamai, S.; Wyner, A.D., "Information theoretic considerations for cellular mobile radio," May 1994**

The authors consider wireless transmission to  $J$  users via a TDMA protocol. The message is transmitted in  $M$  blocks of duration  $T$  each. It is assumed that  $2WT \gg 1$  with  $W$  being the signal bandwidth, i.e. each block accommodates a large number of transmit symbols. The channel is assumed to be dominated by a two-ray propagation mechanism with  $B$  being the bandwidth of the independent Gaussian fading processes. It is assumed that  $BT \ll 1$ , i.e. the variability of the fading is small during each block and is correlated among blocks. Arguing with a “reasonable vehicle speed”,  $B$  is set to be  $50\text{Hz}$ . The receiver is supposed to have full CSI, whereas the transmitter does not. The message is spread across  $K$  blocks, which is defined by the delay constraints. The authors derive the probability distribution function of the mutual information for the single- and double-ray model, (as a function of the specific values of the channel parameters), as well as the average (expected) mutual information for both cases. The pdf of the mutual information in a transmission spanning  $K$  blocks is calculated by convolving the pdf of the mutual information  $K$  times with itself. The authors conclude that the achievable rates with an outage of less than 1% under stringent delay constraints ( $K = 1, 2$ ) differ strongly from average capacities (with  $K = \infty$ ) that are equal to the Shannon capacities, iff the channel satisfies the asymptotic mean stationarity property. This is the case if the two fading processes in the two-path transmission are ergodic. They show that, for fixed outage probability, the achievable rate is much better for  $K = 2$  than for  $K = 1$ , even when the fading coefficients are “sizeably correlated” in both transmitted blocks.

**[14]: Hanly, Steven V.; Tse, David N.C., "Multi-Access Fading Channels-Part II: Delay-Limited Capacities," Aug 1998**

In part two of their paper on optimal multiuser uplink resource allocation for fading channels, the authors refer to results by Gilhousen et al. [15], Hanly [16], and Yates [17] who used the term “delay-limited capacity” in conjunction with a minimum/constant mutual information that must be kept in all fading states. In their paper, the authors use the same approach, i.e. power control is used to enforce a constant mutual information, as well. They point out that in the multi-user case, this results in simple channel-inversion. This strategy is not successful for some fading distributions when a finite average transmit power is assumed - a prominent example being Rayleigh-fading. This paper is actually an extension of a previous publication by the same authors [18] where they explored the delay-limited capacity for the multiaccess channel in the symmetric case when all users have the same rate requirements and experience i.i.d. fading. According to the authors, this symmetric delay-limited capacity is the maximum common rate that can be achieved. The optimal

## 2. QoS in Infinite Lifetime Channels

power-control policy is CD with users being decoded in the order of *decreasing* channel quality. The power-allocation scheme assigns the least transmit power to the user with the weakest channel (i.e. who is furthest away), because after having decoded all the other users, this user does not suffer any interference by any other user. The authors point out that this strategy is in contrast to the CDMA-approach, where the user who is furthest away needs to transmit with the largest power, causing lots of interference in neighboring cells.

The capacity region of the two-user case can be calculated by varying the ratio of their respective power prices, i.e.  $\lambda = \frac{\lambda_1}{\lambda_2}$ , from 0 to  $\infty$ . There are extreme rate regions (resulting from a large difference in power prices) where all the power is allocated to either user 1 or user 2, and there is a rate region for intermediate  $\lambda$ , where the decoding order of the optimum CD scheme depends on the fading state. The authors suggest that in a fast fading environment (i.e. the coherence time of the channel changes during the duration of a block for which the power allocation is fixed), it is still optimal to assign power depending on the *initial* knowledge of the channel and “pretend” that the channel does not change. This is, according to the authors, a fundamental difference between the Shannon “throughput capacity” that considers the channel only statistically and can explicitly state the optimum CD decoding order and the “delay-limited capacity” where the optimum decoding order in a fast-fading channel is unknown. Additionally, the throughput capacity is generally characterized by average rates rather than fixed rates, i.e. a user may or may not be assigned a transmitted rate. This is in contrast to delay-limited capacity where each user transmits at the same rate in every block.

Considering the practical problem of finding the optimal power prices for a desired rate vector, the authors provide an iterative solution that exploits the polymatroid structure of the rate region: they introduce proportional, weighted fairness between the users, i.e. the assigned average powers for all users are balanced as much as possible. The authors consider the case when the channel conditions of the individual users do not allow for perfect balancing.

Finally, the authors consider the case when two types of fading, i.e. block- and fast fading are present. In this case, they assume that the fast fading be sufficiently fast to average out during each block. Then, an optimum resource allocation scheme will not require explicit knowledge about the (slow) fading statistic at all; instead, the power is adapted through the power prices to ensure the same rate at any fading state to all users.

**[19]: Caire, Guiseppe.; Taricco, Giorgio; Biglieri, Ezio, "Optimum power control over fading channels," Jul 1999**

The authors apply a block-fading AWGN channel model [11–13] and minimize the probability of an outage under the assumption of both perfect CSI at transmitter and the single receiver. Notably, this is also assumed for all *future* channel states, so that both transmitter and receiver can find an optimum transmission scheme. The block fading assumption leads to  $N$  transmit symbols experiencing the same channel state. Each codeword spans  $M$  blocks (and therefore contains  $MN$  symbols), so that  $M$  being finite defines the delay-constrained case. The minimum outage probability is achieved by transmitting a fixed code-book comprised of random gaussian symbols, followed by an optimum power controller. The authors adopt the notion of “delay-limited capacity” introduced by [14] and derive an explicit formula for the delay-limited capacity of the two-block BF AWGN channel with rayleigh i.i.d. blocks. Although the delay-limited capacity is shown to be zero when the transmission (and therefore the transmitted codeword) spans just one block ( $M=1$ ), it is shown to be only 5dB lower than capacity of the AWGN channel if the codeword spans 2 blocks ( $M=2$ ). In addition, the delay-limited capacity is only 2.5dB higher if the same rayleigh-fading channel is considered and the codeword spans an unlimited number of blocks ( $M \rightarrow \infty$ ).

**[20]: Li, Lifang.; Goldsmith, Andrea J., "Capacity and Optimal Resource Allocation for Fading Broadcast Channels - Part i: Ergodic Capacity," Mar 2001**

In part one of their landmark paper on optimal multiuser resource allocation , Goldsmith and Li base their work mainly on previous results by David N. Tse [21]. Instead of parallel gaussian broadcast channels used in [21], the authors consider (sub)optimum power allocation in the ergodic discrete-time  $M$ -user broadcast channel with  $M$  independent flat-fading subchannels with possibly different noise-power spectral densities. They assume perfect knowledge of CSI and noise by the transmitter and all receivers, so that they can perform perfect Code Division (CD) with successive decoding. The optimum, power-constrained CD allocation strategy is based upon the solution derived in [21], and it spans the ergodic capacity region. In the case of equally weighted individual user’s rates, it is optimal to assign power to the user least affected by noise only, and the allocated power is obtained by performing water filling that ensures the (total) average power constraint of the  $M$  users to be met. Again using a water-filling approach, the authors show that an optimal TD approach assigns power to only one user, for every fading state. That means that the capacity regions for CD and TD are actually the same if all users have the same rate-rewards - else CD performs better if the noise variances of the users are not similar (i.e. the average rates of the users are quite different). As the arguments for TD can be generalized to FD and CD without successive decoding, they all have the same ergodic capacity region. The results are illustrated with plotted capacity regions for the two-user case, for a AWGN, Rayleigh, and Rician channels.

**[22]: Li, Lifang.; Goldsmith, Andrea J., "Capacity and Optimal Resource Allocation for Fading Broadcast Channels - Part ii: Outage Capacity," Mar 2001**

Part two of the publication elaborates on the rate regions supported by flat-fading broadcast channels (with Nakagami  $m$ -fading as a - due to the scale parameter - rather flexible specialization). The two rate regions that are considered are the outage capacity region and the *zero* outage capacity region, the latter meaning the rate that can be maintained in *all* fading conditions. The authors cite [14] to remind that the single-user zero outage capacity region is actually zero for a Rayleigh fading channel. Two different power allocation strategies (as well as a combination of both of them) are explored: using the broadcast channel simultaneously for all users and declaring an outage if even a single user cannot maintain the required rate, and the possibility for user-individual outage. Following the structure of the first part of the publication, optimal power-allocation strategies are obtained for TD, FD, and CD with and without successive decoding. The optimum strategy is certainly CD with successive decoding, where the individually requested rates by the users must not be larger than the capacity of the AWGN channel. For CD without successive decoding, the capacity region is smaller due to the mutual interference when more than one user shall be served at a time. It should be noted that the average power is simply scaled to the value that will support the user's rate requests, i.e. the average sum-power that is distributed among the users is not the same for all strategies.

The authors show once again that the capacity regions of TD and FD strategies are identical for all considered cases (i.e., for zero-outage as well as for an allowed outage greater than zero, and for enforced simultaneous outage of all users as well as for user-individual outage). These strategies are also equivalent to CD without successive decoding when compared in a slow fading environment. The CD strategy without successive decoding is, however, inferior to F/TD in a fading environment since its capacity region is convex instead of concave for the other schemes. Still, CD with successive decoding achieves the largest capacity region in all cases. Also, compared to the zero-outage capacity region, the capacity region of CD without successive decoding shows a significantly smaller growth compared to the other strategies when a small outage is accepted.

**[23]: Xiangheng, Liu; Goldsmith, Andrea J., "Optimal Power Allocation over Fading Channels with Stringent Delay Constraints," 2002**

The authors explore optimal power allocation schemes for i.i.d. block-fading channels, considering a strict transmission delay constraint. They seek to maximize the total (i.e., not the user-individual!) throughput under a short-term power constraint. They distinguish the case that all channel-coefficients are known a priori from the case that transmitter and receiver(s) have full causal CSI, i.e. that the current channel state is known. When all channel-coefficients are known a priori, the

problem is equivalent to power allocation for parallel channels, with known optimal strategies and results. When the channel coefficients are not known a priori, and full CSI relates to the current and all previous channel states, the authors refer to an optimal solution suggested by Negi and Cioffi [24] who developed a dynamic programming algorithm performing an exhaustive search over all combinations of power prices. Along with the results by [24], it is shown for the single user case that in the high SNR regime, an optimization of the transmitted powers does not result in a large gain in transmitted information (compared to sub-optimal schemes), and transmitting constant power for all channel states is optimal. In contrast, when a low SNR is experienced, it is best to transmit all available power during one fading state only. Then, the threshold that determines if a transmission should take place needs to decrease towards the end of the delay-limit and is actually zero during the last transmit-opportunity (i.e. when the channel has been unfavorable until then). In the two-user case and under a low SNR, the user with the best combination of channel gain and scheduling priority will transmit, i.e. the optimal scheme applies a decision threshold which, as in the single-user case, decreases towards the end of the tolerable delay. In the high SNR regime however, it is shown that it is best to serve only the user with the highest priority will transmit at all times, if the channel gains are comparable.



### 3. A Scheduling Algorithm for Small Outage

In this section, available resource allocation schemes are explored for their ability to meet rate- and delay- demands in a multi user scenario. The discussion focuses on well known schemes and tries to find modifications that will eventually lead to the development of a new resource allocation scheme that is designed to ensure network-services for modern applications.

#### 3.1. System Model

Considering a multiuser scenario, it is assumed that  $J$  mobile users request data bit rates  $r_j$  from the base station for their applications with different hard delay constraints. Once the central network control has accepted to serve a user, the required resources (power in particular) are assumed to be available. This may be accomplished by measuring the path loss of the user requesting to be served. If the requested rate  $r_j$  is low enough to be supported by the system's free resources (with a certain margin  $Q > 1$ ), the user is served:

$$r_j \cdot Q \stackrel{!}{\leq} \log_2 \left( 1 + \frac{P_{available} \cdot |\bar{H}_j|^2}{\sigma^2} \right) \quad (3.1)$$

In (3.1),  $Q > 1$  is some margin that is necessary to avoid having to drop the user if its average channel quality degrades slightly, as well as to compensate for the impossibility of the coding and modulation scheme to achieve channel capacity. The bit rate requested by the mobile  $r_j$  user is a deterministic quantity and therefore written in lower-case. The symbols  $P_{available}$  and  $\bar{H}_j$  represent the random (i.e. not constant over time) available system-resource transmit power, and the random (i.e. dependent on the random recent history of the channel) recent average of the signal attenuation through the channel, respectively. The squared euclidean distance  $|H_j|^2$  represents the influence of the complex, random channel-coefficients  $H_j$  on the received signal power. Equation (3.1) therefore considers the received signal to noise ratio, with  $\sigma^2$  denoting the noise-power of additive, white Gaussian noise.

The current values of the channel power gains  $|H_j|^2$  in (3.1) are assumed to be known at the transmitter (this is referred to channel state information at the transmitter, or CSI-T). In the downlink, this can be practically realized by users estimating the channel power gains and returning them to

### 3. A Scheduling Algorithm for Small Outage

the base station. It is therefore assumed that this feedback information can be used as long as it is still valid - though this is not perfectly the case in practice, where the scheduling algorithm can usually only consider quantized channel coefficients that were measured a couple of milliseconds ago, it is a common assumption that represents a best-case scenario. The channel power gains include both the long-term path-loss (mainly depends on the distance to the base station) and the effects of block fading, e.g., due to scattering and interference.

A user-specific delay parameter  $\tau_j$  defines the maximum number of successive time slots over which one block of source data of a user can be spread out. More specifically, the hard delay constraints are met when the inequality

$$\bar{R}_j(k) \geq r_j \cdot t_s \quad \forall j, k \quad (3.2)$$

is fulfilled. In (3.2)  $t_s$  is the channel-symbol time in seconds,  $\bar{R}_j(k)$  is the average rate (in bits per channel-use) obtained by user  $j$  in time window  $k = n \cdot \tau_j - 1$ ,  $n = 1, 2, \dots$ ; it is defined by

$$\bar{R}_j(k) \doteq \frac{1}{\tau_j} \sum_{i=0}^{\tau_j-1} R_j(k-i), \quad (3.3)$$

and  $R_j(l)$  is the rate achieved (in bits per channel-use) by user  $j$  in time slot  $l$ .

## 3.2. Information Theoretical Limits

In the framework of Information Theory, the system model can be described as a block-fading Gaussian broadcast channel (e.g. [13]). In this scenario, the terms “block” and “time slot” will be used interchangeably: they refer to a sequence of transmit symbols that experience constant channel conditions. When user  $j$  is scheduled, Shannon’s well-known formula for the capacity of the Gaussian channel [25] is used to compute the achieved rate (bits/channel-use), i.e.,

$$R_j(k) = \log_2 \left( 1 + \frac{P_j(k) \cdot |H_j(k)|^2}{\sigma^2} \right), \quad (3.4)$$

with  $P_j(k)$  the transmit power allocated to user  $j$  in time slot  $k$  and  $H_j(k)$  the corresponding channel coefficient as in (3.1);  $\sigma^2$  is the variance of the Gaussian receiver noise which is assumed to be equal for all users (for simplicity). In a practical system, (3.4) cannot be achieved, but a power offset (that is also contained in  $Q$  in (3.1)) can be determined that is required by some practical coding and modulation scheme to achieve the rate  $R_j(k)$  at some (low) bit-error probability that is acceptable for the application.

Although there has been work on block fading channels with rate-averaging over more than one block (time slot), most of the results (e.g., [19]) only apply to single user channels. Results for



the fading broadcast channel are either restricted to single time slots (then, outage capacities are analyzed, [22]) or they apply only for averaging over an unlimited number of time slots (then, ergodic capacities can be stated, [20]). The dilemma with the practical setup considered in this thesis is that neither of the two cases really applies to the problem: the zero-outage (delay limited) capacity of the fading broadcast channel with averaging of the achieved rates over  $\tau$  time slots, with  $1 < \tau \ll \infty$ . Unfortunately, there are no analytical results available from literature. What is known, however, is that ergodic capacity ( $\tau \rightarrow \infty$ ) is a strict upper bound for zero-outage capacity with  $\tau \ll \infty$ , so the simulation results below are compared with ergodic capacity limits, assuming that the values of  $\tau$  are “close enough” to infinity.

### 3.3. Popular Resource Allocation Schemes and their Suitability for low Outage

In this section, well-known resource allocation schemes and explores modifications regarding the delivery of minimum quality of service are discussed. The gathered insights are going to be used for developing a new scheduling scheme that is improved in certain key aspects.

#### 3.3.1. Round Robin Scheduling and Modifications

In a round robin resource allocation scheme each user is scheduled for transmission according to a fixed pattern that is periodically repeated, and the scheduling decision does not depend on the channel quality. In the simplest case it is assumed that the scheduler has no knowledge of the applications either, which means the best, i.e. least discriminating approach is to allocate each user the same number of time-slots. A simple way to implement this is to schedule the user  $j_k^* \in \{1, 2, \dots, J\}$  in time-slot  $k$  according to

$$j_k^* = (k \bmod J) + 1 \quad (3.5)$$

with  $J$  the number of users in the system and “mod” denotes the remainder when dividing the integer  $k$  by the integer  $J$  (i.e.,  $k$  is reduced modulo  $J$ ).

A more general round-robin scheme would be given by

$$j_k^* = \mathcal{S}_{(k \bmod M_{\mathcal{S}})+1} \quad (3.6)$$

with  $\mathcal{S}$  the scheduling vector (“scheduling pattern”),  $\mathcal{S}_l$  the  $l$ -th component of the scheduling vector, and  $M_{\mathcal{S}}$  the scheduling period which equals the length of the pattern  $\mathcal{S}$ . The latter is given by

### 3. A Scheduling Algorithm for Small Outage

$$\mathcal{S} \doteq \{\underbrace{1, 1, \dots, 1}_{M_1 \text{ times}}, \underbrace{2, 2, \dots, 2}_{M_2 \text{ times}}, \dots, \underbrace{J, J, \dots, J}_{M_J \text{ times}}\} \quad (3.7)$$

with the condition

$$\sum_{j=1}^J M_j = M_{\mathcal{S}} . \quad (3.8)$$

The pattern  $\mathcal{S}$  allows for users to appear more than once in the periodic repetition of the pattern  $\mathcal{S}$  which, e.g., would allow applications with high rate requests to be allocated more time-slots than other users' applications. In practice it will often be reasonable to randomize the user-order within  $\mathcal{S}$  to avoid possibly correlated bad channel states. Moreover, note that (3.5) follows from (3.6) and (3.7) with  $M_j = 1 \forall j$  and  $M_{\mathcal{S}} = M$ .

#### **Variant: Channel-Inversion Round-Robin Scheduling (RR-I)**

The scheduled user has to be allocated power such that the requested rate is achieved. One way is to “invert” the current channel power-gain  $|H_j(k)|^2$ . As by (3.6) each user  $j$  is scheduled only  $M_j$  times in  $M_{\mathcal{S}} = \sum_1^J M_j$  time-slots, the rate that has to be achieved in every *scheduled* time-slot for this user is given by

$$R_{j,\text{slot}} = r_j \cdot t_s \cdot \frac{M_{\mathcal{S}}}{M_j} . \quad (3.9)$$

A consequence from the “channel inversion” approach is the constant rate that is achieved in every single scheduled time-slot. Applying Shannon's formula for the capacity of the Gaussian channel [25]

$$R_j(k) = \log_2 \left( 1 + \frac{P_j(k) \cdot |H_j(k)|^2}{\sigma^2} \right) \quad (3.10)$$

to calculate the power required to achieve a desired rate in time slot  $k$  leads to:

$$P_j(k) = (2^{R_{j,\text{slot}}} - 1) \frac{\sigma^2}{|H_j(k)|^2} . \quad (3.11)$$

Here, it is assumed that the channel power gain for user  $j$  at time-instant  $k$ ,  $|H_j(k)|^2$ , is known due to perfect channel-state information at the transmitter (i.e. CSI-T).

Of course it is well known that, as the support of the channel coefficient may contain the value “zero” (e.g. Rayleigh fading), this power allocation scheme generally has unlimited peak power. Hence, in an application the channel-inversion given by (3.11) has to be truncated at some point. However,  $P_j(k)$  will still have a huge dynamic range, which is very unfavorable from a practical

### 3.3. Popular Resource Allocation Schemes and their Suitability for low Outage

perspective (due to interference in neighboring cells, and non-linearity of RF-amplifiers). The problem of excessive noise amplification is mitigated by approaches that assign more power when the radio conditions are good, i.e. waterfilling. A more subtle issue is that (3.11), although the requested rates can be achieved exactly as *long-term* averages, the delay constraints as defined by (3.2) may still not be met *exactly*. The reason is that the delay limit  $\tau_j$  is unlikely to be an integer multiple of the length  $M_S$  of the scheduling pattern, which means the rate-average will be too small when the window of length  $\tau_j$  covers  $\lfloor \frac{\tau_j}{M_S} \rfloor$  scheduled slots and it will be too big when a window with the same length but different location on the time axis will cover  $\lfloor \frac{\tau_j}{M_S} \rfloor + 1$  scheduled slots. This effect can be easily compensated for by a (usually slight) increase of  $R_{j,\text{slot}}$  in (3.11) which ensures that the constraint on the delay-limited rate-average (3.2) is achieved, even for the location of the time-window of length  $\tau_j$  that produces the smallest value for the average. However, due to this effect a “ripple” in the average-rate curves is inevitable. The ripple will be the smaller the more time-slots the average is taken over and the better the spreading of each user in the scheduling pattern  $\mathcal{S}$  is: this explains the necessity of the randomization of the order of the user-indices in  $\mathcal{S}$ .

This approach therefore suffers from a non-constant transmit rate and unlimited peak power that has to be truncated and compensated for by requesting a higher rate  $R_{j,\text{slot}}$  in (3.11). This necessitates the use of an adaptive requested rate that is adjusted for the channel conditions.

### 3. A Scheduling Algorithm for Small Outage

#### 3.3.2. Channel-Adaptive Time-Division Access with Power Allocation by “Waterfilling” (TD-W)

This scheme is based on the suboptimal “TD-solution” from multiuser information theory for the ergodic capacity region of the fading Gaussian broadcast channel [20]. The idea is to exploit (as a scheduling and resource-allocation algorithm) the result of a mathematical optimization that can be used to obtain the boundary surface of the capacity region under a time-division constraint (i.e. at most one user per time-slot is scheduled).

##### Basic Scheme

According to [20, 26] the following optimization problem needs to be solved:

$$\begin{aligned} & \max \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^J \mu_j \cdot R_j[k] \\ & \text{subject to} \\ & \max \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^J P_j[k] = \bar{P} \\ & \sum_{j=1}^J \mu_j = 1 . \end{aligned} \tag{3.12}$$

i.e. the sum of the average weighted rates of all users is maximized under the constraint that the average power emitted by the “base station” is  $\bar{P}$ . In (3.12),  $K$  denotes the number of blocks across which the average power is computed, and  $\mu_j$  are the rate-weighting factors by which a desired point on the boundary surface of the capacity region is selected. The instantaneous rates  $R_j[k]$  in block  $k$  of the  $J$  individual users are calculated according to

$$R_j[k] = \log_2 \left( 1 + \frac{|h_j[k]|^2 P_j[k]}{\sigma^2} \right) . \tag{3.13}$$

Due to convexity of the problem (see [20]), the optimization can be written as (not taking into account the condition  $\sum_{j=1}^J \mu_j = 1$ )

$$\max_{P[k]} \sum_{k=1}^K \left( \sum_{j=1}^J \mu_j R_j[k] - \lambda \sum_{j=1}^J P_j[k] \right) . \tag{3.14}$$

If one user only is scheduled in every block, this reduces to

$$\sum_{k=1}^K \max_{P[k], j} (\mu_j R_j[k] - \lambda P[k]) \quad \text{with} \quad \frac{1}{K} \sum_{k=1}^K P[k] = \bar{P} . \tag{3.15}$$

### 3.3. Popular Resource Allocation Schemes and their Suitability for low Outage

Now, the maximization is carried out with respect to the transmit power for every block, as well as for finding the user to be scheduled in every block.

The **transmit power** is found by maximization of the transmitted rate under a power constraint:

$$\max_{P_j} \left( \mu_j \log_2 \left( 1 + \frac{|h_j[k]|^2 P_j[k]}{\sigma^2} \right) - \lambda P_j[k] \right) \quad (3.16)$$

Derivation of the *Lagrangian* expression that is to be maximized in (3.16) with respect to  $P_j[k]$  leads to

$$\mu_j \frac{|h_j[k]|^2}{\sigma^2 \log(2) \left( 1 + \frac{|h_j[k]|^2 P_j[k]}{\sigma^2} \right)} - \lambda \stackrel{!}{=} 0 . \quad (3.17)$$

The optimum transmit power  $P_j[k]$  for user  $j$  in block  $k$  is therefore:

$$P_j[k] = \max \left( \frac{\mu_j}{\lambda \cdot \log(2)} - \frac{\sigma^2}{|h_j[k]|^2}, 0 \right) = \left( \frac{\mu_j}{\lambda \cdot \log(2)} - \frac{\sigma^2}{|h_j[k]|^2} \right)^+ . \quad (3.18)$$

This is a standard waterfilling solution with the additional user-individual (but constant and therefore not time-dependent) rate-weighting factor  $\mu_j$ . In (3.18) a “max”-operation according to  $(x)^+ \doteq \max(0, x)$  ensures that the solutions for the powers do not take negative values; the Karush-Kuhn-Tucker [27] conditions guarantee that (3.18) is still an optimal solution to the problem.

To obtain exactly the same result as was proposed in [20], the “waterlevel”  $\lambda$  has to be redefined:

$$\tilde{\lambda} = \lambda \cdot \log(2) . \quad (3.19)$$

The sum in (3.15) is maximized if all addends are maximized individually. Together with the long-term power constraint from (3.12), this amounts to finding the optimum user to be served, considering the optimized transmit power (3.18).

The basic scheme can therefore be written in its algorithmic representation:

1. **Transmit Power:** According to (3.18), determine the optimum transmit powers  $P_j[k]$  for all users  $j$  in the current block  $k$ :

$$P_j[k] = \left( \frac{\mu_j}{\lambda \cdot \log(2)} - \frac{\sigma^2}{|h_j[k]|^2} \right)^+ . \quad (3.20)$$

### 3. A Scheduling Algorithm for Small Outage

2. **Weighted Rates:** With the optimum transmit powers  $P_j[k]$  and the rate-weighting factors  $\mu_j$ , calculate the weighted rates  $G_j \left( \frac{|h_j[k]|^2}{\sigma^2}, \mu_j \right)$  of all users  $j$  ( $G_j$  depends on the fading states of all users via the respective transmit powers),

$$G_j \left( \frac{|h_j[k]|^2}{\sigma^2}, \mu_j \right) = \mu_j R_j[k] - \lambda P_j[k] = \mu_j \log_2 \left( 1 + \frac{|h_j[k]|^2 P_j[k]}{\sigma^2} \right) - \lambda P_j[k]. \quad (3.21)$$

The set of users  $\mathcal{U}$  who satisfy the maximization (3.15) in a certain block (at time instant  $[k]$ ) and obtain the maximum weighted rate  $G$  for the current fading state is:

$$\mathcal{U} = \{ \forall j \in \{1, 2, \dots, J\} : G_j = G \}. \quad (3.22)$$

Here,  $G \left( \frac{|h[k]|^2}{\sigma^2}, \mu \right) = \max_{j \in \{1, \dots, J\}} G_j \left( \frac{|h_j[k]|^2}{\sigma^2}, \mu_j \right)$  denotes the maximum weighted rate that is reached by one or more users in the current block. While it is unlikely, that the set  $\mathcal{U}$  contains more than one user when continuous values for the weighted rates are considered, in practise these will be quantized, justifying the following “user selection”:

3. **User Selection:** From the set  $\mathcal{U}$ , user  $j^*$  that uses the smallest power is the only user that is really scheduled:

$$j^* = \arg \min_{j \in \mathcal{U}} P_j[k], \quad (3.23)$$

with the transmit power

$$P_{j^*}[k] = \max \left( \frac{\mu_{j^*}}{\lambda} - \frac{|h_{j^*}[k]|^2}{\sigma^2}, 0 \right). \quad (3.24)$$

The parameter  $\lambda$  in (3.18), (3.20), and (3.24) is chosen such that the average power used by the “base station” meets a given average power constraint  $\bar{P}$ , i.e.,

$$\bar{P} = \mathbb{E} (P_{j^*}[k]). \quad (3.25)$$

The average rate achieved for user  $j$  is given by the following expression:

$$\bar{R}_j = \mathbb{E} \left[ \log_2 \left( 1 + \frac{P_j[k] |h_j[k]|^2}{\sigma^2} \right) \right]. \quad (3.26)$$

The expectations of (3.25) and (3.26) can be evaluated by time-averaging over the realizations, as the channel-statistics are assumed to be stationary and ergodic.

### 3.3.3. TD-Waterfilling: Modification A

The scheduling scheme described above is not directly applicable to the problem of guaranteed rates for users with different rate demands: as a power constraint is given, the scheme provides a best-effort service in terms of the delivered rates. Therefore, the scheme has to be modified: instead of the power constraint, the rate-demands of the applications  $r_j$  (in bit/s) are used as constraints. Together with the symbol-time  $t_s$ , this provides the requested rate  $\bar{\bar{R}}_j$  in bit/channel use, for each user:

$$\bar{\bar{R}}_j = r_j \cdot t_s \forall j. \quad (3.27)$$

The power constraint is dropped, as now the powers required to meet the rate requests are subject to optimization (a similar problem is solved in [21]). Of course, this constraint ignores the delay limits as the rate-demands are met in a “long-term” average sense.

According to the algorithm in Section 3.3.2, the classic form of the waterfilling solution is retained by merging the weighting factors  $\mu_j$  for the rates with the Lagrange multiplier  $\lambda$  for the power constraint, i.e.,  $\tilde{\mu}_j = \mu_j/\lambda$ . Moreover, the time-slot index  $k$  is included to express that the fading states are given by different realizations of the channel coefficients at different times.

The Modified Scheme (A) for scheduling and resource allocation is given by

For every time-slot  $k$  perform 1) – 3):

1. **Transmit Powers:** If scheduled, the power that would be allocated to user  $j = 1, 2, \dots, J$  in time-slot  $k$  is given by

$$P_j(k) = \max \left( \tilde{\mu}_j - \frac{\sigma^2}{|h_j(k)|^2}, 0 \right). \quad (3.28)$$

### 3. A Scheduling Algorithm for Small Outage

2. **Weighted Rates:** At each time  $k$  the set  $\mathcal{U}(k)$  of users that obtain the maximum weighted rate  $G(k)$  is formed:

$$\mathcal{U}(k) = \left\{ \forall j : G_j(k) = G(k) \right\} \quad (3.29)$$

with

$$G(k) = \max_{j \in \{1, \dots, J\}} G_j(k) \quad (3.30)$$

and

$$G_j(k) = \tilde{\mu}_j \cdot \log_2 \left( 1 + \frac{P_j(k) \cdot |h_j(k)|^2}{\sigma^2} \right) - \frac{P_j(k) \cdot |h_j(k)|^2}{\sigma^2} . \quad (3.31)$$

Note that the value  $G(k)$  is different from the result of Section 3.3.2 as the weighting factors  $\tilde{\mu}_j$  are different. However, as the Lagrange multiplier  $\lambda$  in Section 3.3.2 is a constant (which is the same for all users), still the proper set of users with the largest weighted rates in (3.29) is found.

3. **User Selection:** From the set  $\mathcal{U}(k)$  the user  $j_k^*$  that uses the smallest power is selected; this is the only user scheduled:

$$j_k^* = \arg \min_{j \in \mathcal{U}(k)} P_j(k) \quad (3.32)$$

with the transmit power

$$P_{j_k^*} = \left( \tilde{\mu}_{j_k^*} - \frac{\sigma^2}{|h_{j_k^*}(k)|^2}, 0 \right)^+ . \quad (3.33)$$

The rate-weighting factors  $\tilde{\mu}_j$  must be calculated in advance and such that

$$\begin{aligned} \bar{\bar{R}}_j &= \mathbb{E} \left[ \log_2 \left( 1 + \frac{P_j(k) \cdot |h_j(k)|^2}{\sigma^2} \right) \right] \\ &\approx \frac{1}{L} \sum_{k=1}^L \log_2 \left( 1 + \frac{P_{j_k^*}(k) \cdot |h_{j_k^*}(k)|^2}{\sigma^2} \right) \cdot \gamma_0(j - j_k^*) , \end{aligned} \quad (3.34)$$

where



$$\gamma_0(x) \doteq \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.35)$$

and  $\bar{R}_j$  is the long-term average rate as requested in (3.27).

In the second line of (3.34) the expected value of the rate is estimated by time averaging over a large number  $L$  of channel realizations (ergodicity assumption) – this indicates how the selection of the  $\tilde{\mu}$ -factors can be carried out in practice: A simulation of  $L$  time-slots for an assumed vector of  $\tilde{\mu}$ -values is carried out and (3.34) is evaluated afterwards. If the rate constraint is not fulfilled with sufficient accuracy the  $\tilde{\mu}$ -values are corrected appropriately and a new simulation is run. The process is iteratively repeated until the rate constraints are achieved. Note that for this technique to work, the assumption of stationary and ergodic channel-statistics has to hold, which often will not be the case in practice.

### 3.3.4. Discussion of TD-Waterfilling Mod. A

TD-Waterfilling Mod.A from Section 3.3.3 has several disadvantages.

- As TD-Waterfilling Mod.A actually computes a point on the boundary surface of the ergodic capacity region, delay is completely ignored. Although the ergodic capacity region (due to lack of accurate capacity results for the given delay-limited problem; see Section 3.2) will provide a benchmark for the simulations, this does certainly not mean that the algorithm used to obtain the capacity region boundary is suitable as a scheduling and resource-allocation scheme for a delay-limited problem.
- Waterfilling requires, by definition, a huge dynamic range for the transmit power. In a practical application this will likely cause difficulties for RF amplifiers. Moreover, waterfilling may well allocate transmit powers significantly below the noise variance  $\sigma^2$ : although this is theoretically possible, it is very difficult in a real-world system considering synchronization and error rate issues.
- If in a time slot the channel coefficient is large, in principle this means that lots of information can be transmitted. Using high transmit power to benefit from the favorable channel condition even more, this will cause a huge variation in the transmit rate in the time slots. If the slots were infinitely long (as assumed by information theory), then capacity could be achieved within each slot by picking a channel code with the appropriate rate for the specific time slot: coding across multiple time slots is not necessary in this case. In the practical case, however, the time-slots will be too short for a code to perform close to Gaussian capacity and therefore coding across time slots is unavoidable. This, together with waterfilling which

### 3. A Scheduling Algorithm for Small Outage

amplifies the rate-differences between slots, means that *one* code that spans several slots has to cope with a huge variability of channel quality. As the problem is causal, it is not known at the time when channel encoding is conducted which parts of the codewords will in the future suffer from a bad channel. Therefore, the code design must be universal for many different types of channels which are a-priori unknown<sup>1</sup>. The result from information theory that capacity for fading channels can be achieved by both separate coding for fading states and cross-coding over all fading states [28] is not applicable here as this result applies only in the ergodic case.

- The delay constraint defined by (3.2) could be met by introducing a rate-offset for the requested rate: raising the rate request in order to *always* fulfill the minimum requirements will improve the situation by decreasing the probability of an outage. However, since the increased rate request will have to be adapted to the rare case when an unlikely combination of bad coefficients for all users is experienced, a much higher rate than requested will be achieved most of the time. As the transmitting application will provide data according to the originally requested rate, this will empty the transmit buffers quickly, and some of the allocated resources will not be used. A solution to this problem would be to switch off a user once the requested average rates has been achieved in a delay window (note that as the channel coefficients are known at the transmitter, this is possible). Both measures in combination will prove to be rather efficient.
- The scheme can be further improved when the delay-constraints are used to take scheduling decisions. All schemes discussed up to now do not involve delay limits whatsoever. However, if a delay window is coming to an end and the user still needs data, the scheduler metric should be scaled appropriately. Such a concept is well-known from Proportional Fair Scheduling [29], and it is included in the power-control feature of *Power-Controlled Cross-Layer Scheduling* in Chapter 4, see expression (4.5).
- The previous two items will be used in a modified algorithm for TD-Waterfilling. Although the above measures to stabilize the rate will provide good performance, there is no guarantee that the scheme will meet any rate requirement without any outages. When the channel conditions are bad, any waterfilling-based scheme will not consider the available channel at all.

This is a fundamental problem of waterfilling which, together with the practical difficulties mentioned above, will ultimately lead to the development of a less “aggressive” power allocation scheme.

---

<sup>1</sup>In practise, this problem is partly solved by interleaving, at the cost of additional en- and decoding delay.

### Contributions: TD-Waterfilling Mod. A

By exploring basic multi-user TD-waterfilling, the following contribution led to “Mod. A”:

- In order to guarantee a certain bit rate, the power constraint from basic TD-Waterfilling is dropped - the transmit powers are now subject to optimization with respect to the delivered long-term average information rate.

### 3.3.5. Modified Scheme (B)

For this scheme two ideas discussed above will be adopted:

1. Increase the rate demand such that the minimum of the average rate is above the rate request and
2. Temporarily exclude users from scheduling when their rate request have already been met before a delay-window has elapsed.

The Modified Scheme B based on waterfilling is given by

#### Initialize:

- Set the scheduling flag  $\mathcal{F}_j$  to “one” for all users:

$$\mathcal{F}_j = 1 \quad \forall j \quad (3.36)$$

- Initialize the counters  $\kappa_j$  for the time-slots passed within the delay windows:

$$\kappa_j = 0 \quad \forall j \quad (3.37)$$

For every time-slot  $k$  perform 1) – 4):

1. **Transmit Powers:** Compute the power that would be allocated to user  $j = 1, 2, \dots, J$  in time-slot  $k$ :

$$P_j(k) = \mathcal{F}_j \cdot \max \left( \tilde{\mu}_j - \frac{\sigma^2}{|h_j(k)|^2}, 0 \right) \quad (3.38)$$

### 3. A Scheduling Algorithm for Small Outage

2. **Weighted Rates:** At each time  $k$  form the set  $\mathcal{U}(k)$  of users that obtain the maximum weighted rate  $G(k)$ :

$$\mathcal{U}(k) = \left\{ \forall j : \tilde{\mu}_j \cdot \log_2 \left( 1 + \frac{P_j(k) \cdot |h_j(k)|^2}{\sigma^2} \right) - \frac{P_j(k) \cdot |h_j(k)|^2}{\sigma^2} = G(k) \right\} \quad (3.39)$$

with

$$G(k) = \max_{j \in \{1, \dots, J\}} G_j(k) \quad (3.40)$$

and

$$G_j(k) = \tilde{\mu}_j \cdot \log_2 \left( 1 + \frac{P_j(k) \cdot |h_j(k)|^2}{\sigma^2} \right) - \frac{P_j(k) \cdot |h_j(k)|^2}{\sigma^2} \quad (3.41)$$

3. **User Selection:** From the set  $\mathcal{U}(k)$  the user  $j_k^*$  that uses the smallest power is served (this is the only user scheduled):

$$j_k^* = \arg \min_{j \in \mathcal{U}(k)} P_j(k) \quad (3.42)$$

with the transmit power

$$P_{j_k^*} = \mathcal{F}_{j_k^*} \cdot \max \left( \tilde{\mu}_{j_k^*} - \frac{\sigma^2}{|h_{j_k^*}(k)|^2}, 0 \right). \quad (3.43)$$

4. **Scheduling Exclusion:** For  $j = 1, 2, \dots, J$ :

**If**  $\kappa_j = \tau_j - 1$ :

Set  $\kappa_j = 0$  and  $\mathcal{F}_j = 1$

**Else-If**

Compute the delay limited rate-average already achieved for user  $j$  in the current delay window:

$$\hat{R}_j(k, \kappa_j) = \frac{1}{\tau_j} \sum_{i=0}^{\kappa_j} C(k-i) \cdot \gamma_0(j - j_{k-i}^*) \quad (3.44)$$

with

$$C(l) \doteq \log_2 \left( 1 + \frac{P_{j_l^*}(l) \cdot |h_{j_l^*}(l)|^2}{\sigma^2} \right) \quad (3.45)$$

and  $\gamma_0(l)$  given by (3.35).

**If**  $\hat{R}_j(k, \kappa_j) \geq r_j \cdot t_s$ : Set  $\mathcal{F}_j = 0$

Increment the slot counter:  $\kappa_j := \kappa_j + 1$

**End-If**

**End-For**

5. **Goto 1)**

In order to fulfill the delay constraints, the rate-weighting factors  $\tilde{\mu}_j$  must be optimized in advance such that

$$\bar{R}_j(n \cdot \tau_j - 1) \geq r_j \cdot t_s \quad \forall n = 1, 2, \dots \quad (3.46)$$

with

$$\begin{aligned} \bar{R}_j(k) &= \frac{1}{\tau_j} \sum_{i=0}^{\tau_j-1} \left( 1 + \frac{P_j(l) \cdot |h_j(l)|^2}{\sigma^2} \right) \Big|_{l=k-i} \\ &= \frac{1}{\tau_j} \sum_{i=0}^{\tau_j-1} \left( 1 + \frac{P_{j_l^*}(l) \cdot |h_{j_l^*}(l)|^2}{\sigma^2} \right) \cdot \gamma_0(j - j_l^*) \Big|_{l=k-i} \end{aligned} \quad (3.47)$$

The second line of (3.47) merely expresses explicitly that the rate is zero when a user is not scheduled: this is implicitly contained in the upper line, as  $P_j(l) = 0$  when user  $j$  is *not* scheduled in time-slot  $l$ . The function  $\gamma_0(x)$  is therefore the unit impulse function or discrete-time delta function (the argument is always an integer):

The average-rate request (3.34) has now been replaced by a minimum request (3.46) for the *delay-limited* average rate (3.47). This will, however, increase the average rate (3.34), as (3.46) makes sure that the delay constraint is not violated even in the rare case that many consecutive channel coefficients occur that have low magnitudes.

Assuming stationary random processes again, the  $\tilde{\mu}$ -factors can be determined off-line by simulation before the system is actually used. Although the principle is similar to Algorithm 3.3.3 care must

### 3. A Scheduling Algorithm for Small Outage

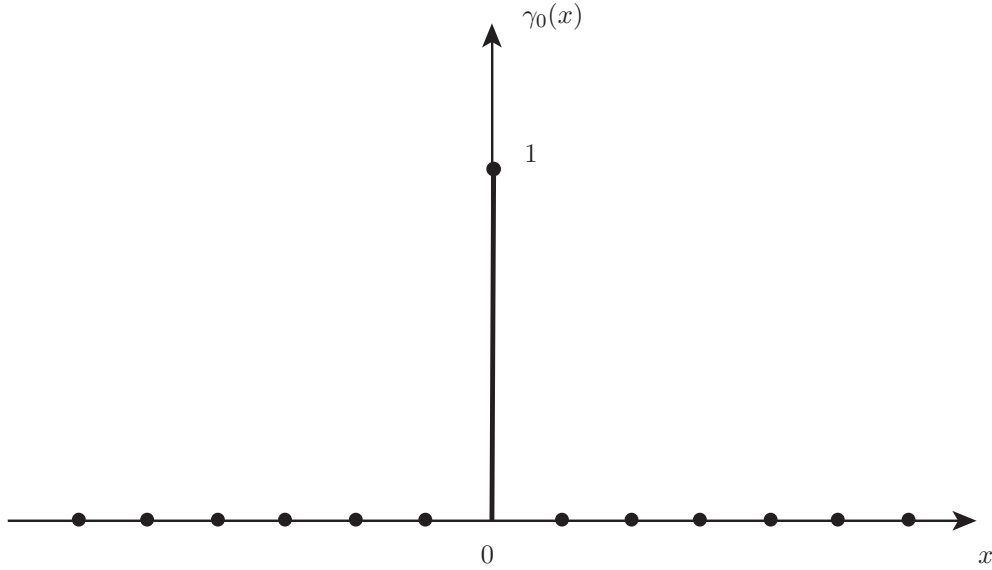


Figure 3.1.: Illustration of the discrete-time delta function  $\gamma_0(x)$ .

be taken: the optimization of the  $\tilde{\mu}$ -factors is potentially unstable if they are not jointly initialized with very small values. The reason is that the scheduling Algorithm 3.3.5 switches off users (Step d) who have achieved their rate requests in the current delay window already. If the  $\tilde{\mu}$ -factors are initialized with very large numbers, only the first few time-slots of each delay window would be used, as the scheduler would disable the users very quickly (by setting  $\mathcal{F}_j = 0$ ), as the rate requests have been achieved already. In this situation the delay constrained would be met immediately after initialization and no further optimization would be carried out for the  $\tilde{\mu}$ -factors: this solution would, of course, be bad as the peak power and also the average power used would be much higher<sup>2</sup> than with an optimized solution that starts from small initial values for the  $\tilde{\mu}$ -factors values.

It is therefore straightforward to adapt an opportunistic scheduling scheme for low outage: always assign transmit power to the user with the currently best channel, except if the respective rate request has already been met. In scenarios where all user's rate requests can be met, this scenario provides a much more efficient approach than the modified round-robin scheme (discussed in Section 3.3.1), since channel states are exploited efficiently: the scheme requires the use of much less transmit power. This approach will even make optimum use of system resources regarding the achievable cell throughput if the sum of the user's required rates exceeds the cell-capacity: then, the users with the weakest channels will hardly be served and subsequently be dropped, until all

---

<sup>2</sup>The power and rate are related to each other by a log-function. Hence, we need exponentially more power for the same average rate compared to a more frequent use of the channel.

### 3.3. Popular Resource Allocation Schemes and their Suitability for low Outage

remaining user's requests can be met. Therefore, this approach is unsatisfactory as well, since it is not a *fair* approach.





## 4. Power-Controlled Cross-Layer Scheduling

This chapter introduces a scheduling algorithm based on proportional fair scheduling that takes into account user's rate requests to minimize outage, while maintaining a fair distribution of system resources. The algorithm adapts resource allocation to changing channel conditions while still benefitting from multiuser diversity gains. According to these design criteria, this new approach is called *Power-Controlled Cross-Layer Scheduling* (PCCLS).

The major difference to most other scheduling schemes presented in literature is that there is no other widely known scheduler that guarantees (with very small outage) hard delay constraints for bit rates requested by the users and at the same time exploits multi-user diversity gains [30] with little loss compared to ergodic capacity limits [20]. The term “Cross-Layer Scheduling” characterizes the main property of the scheduler (which is classically [31, 32] based in the medium access layer) that requires both knowledge of the channel coefficients (known in the physical layer) for each user but also information about all users' applications (known in the application layer), in particular their delay constraints and rate requests, as is illustrated in Figure 4.1.

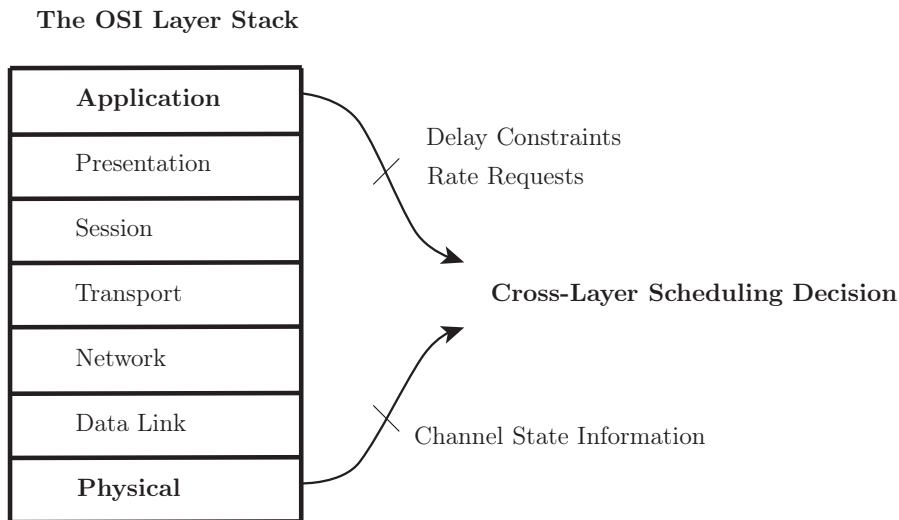


Figure 4.1.: Cross-Layer Scheduling takes into account both information from the application layer (delay constraints and rate requests) and from the physical layer (channel coefficients).

#### 4. Power-Controlled Cross-Layer Scheduling

PCCLS builds upon proportional fair scheduling (PFS) [29], which is also the baseline scheduler of the CDMA/HDR system [33]. The difference between the two scheduling approaches is, however, very fundamental: PFS aims to mitigate the “unfairness” of strictly opportunistic communication: users are scheduled whenever *their* instantaneous channel quality is good compared to *their* recent average channel quality. In principle, the number of time-slots across which the average is computed provides the possibility to integrate delay-constraints into the PFS-decision metric. However, the average rate a user will get still only depends on the channel quality and is, hence, a best-effort service - it was even shown in [34] that PFS is not stable at all, i.e. the scheme, though widely adopted, does not guarantee finite queue-lengths. Therefore, PFS would not be a suitable scheduler for applications with hard delay constraints (in the sense defined below), such as any real-time multimedia-application. The proposed scheduler adopts ideas from PFS and combines them with user-individual rate-monitoring and a “slow” power control. More specifically, the power control is designed to be slow compared to the fast fading, to ensure that the channel is not simply inverted, which has to be avoided because otherwise huge transmit power would be used for a temporarily bad channel.

##### 4.1. Scheduling Algorithm Details

The initial transmit powers for all users are determined by solving (3.10) for the transmit power, while considering the requested rates  $r_j$  and assuming all users are scheduled every  $J$  time slots:

$$P_j(1) = (2^{r_j \cdot (t_s \cdot J)} - 1) \cdot \frac{\sigma^2}{|\bar{H}_j|^2}. \quad (4.1)$$

By use of this initial power the *average* path-loss  $|\bar{H}_j|^2$  of each user is equalized and “symmetric” fading channels for all users in the first iteration of the scheduler are obtained. Simulations have shown that the long-term convergence of the power is not affected, as long as the initial assumptions are “reasonable” (as (4.1)).

Power-Controlled Cross-Layer Scheduling works as follows for every time slot  $k = 1, 2, \dots$ :

1. **Calculation of the rate**  $R_j(k)$  user  $j$  can get in time slot  $k$ :

$$R_j(k) = \log_2 \left( 1 + \frac{P_j(k) \cdot |H_j(k)|^2}{\sigma^2} \right) \cdot \mathcal{F}_j(k). \quad (4.2)$$

Here,  $\mathcal{F}_j(k)$  is a “scheduling flag”: only a user whose rate request hasn’t been satisfied already in the current delay-window will be considered for scheduling.

2. **Scheduling decision:** user  $j_k^* \in \{1, 2, \dots, J\}$  is scheduled in time slot  $k$  according to

$$j_k^* = \arg \max_{j=1, \dots, J} \frac{R_j(k)}{T_j(k)}. \quad (4.3)$$

$T_j(k)$  is the current rate-average for user  $j$ ; as an initialization,  $T_j(1) = \log_2 \left( 1 + (P_j(1) \cdot |\bar{H}_j|^2)/\sigma^2 \right)$  can be used.

3. **Update of the rate-average** for time-slot  $k+1$  for all users  $j = 1, 2, \dots, J$ :

$$T_j(k+1) = \begin{cases} \left(1 - \frac{1}{a_j}\right) T_j(k) + \frac{R_j(k)}{a_j}, & j = j_k^* \\ \left(1 - \frac{1}{a_j}\right) T_j(k), & j \neq j_k^* \end{cases}. \quad (4.4)$$

Only if user  $j$  is served in time slot  $k$  (that is,  $j = j_k^*$ ) the exponentially decaying average rate is increased by addition of  $R_j(k)/a_j$ , with  $R_j(k)$  the current possible rate on user  $j$ 's channel as determined in step 1.

The factor  $a_j$  defines the slope of the exponential decay in the computation of the rate average. In experiments  $a_j = \tau_j/10$  has been determined as a suitable choice. An exponentially decaying window is used for averaging rather than a rectangular window, since this smoothes the average and eliminates “panic-like” scheduling decisions when a time-slot with a “high-rate peak” moves out of the averaging window.

4. **Power control:** The power for user  $j$  in the next time slot  $k+1$  is calculated by

$$P_j(k+1) = \begin{cases} P_j(k) \cdot 2^{r_j \cdot t_s - T_j(k+1)}, & j = j_k^* \\ P_j(k), & j \neq j_k^* \end{cases}. \quad (4.5)$$

The *updated average* rate  $T_j(k+1)$  is used in (4.5) to check whether the rate request,  $r_j$ , was met and to correct the power if necessary. The *rate-average* is used with the aim to compensate for the *average long-term path-loss* but *not* for the fast fading, which is still to be exploited to obtain multi-user diversity gains [30].

5. **Increment slot counter**, i.e,  $k := k+1$ , and **go to step 1**.

#### 4.1.1. Power Scaling

While the above power scaling rule (4.5) seems intuitive, its derivation deserves to be explored. Its original form was developed from the simple comparison of the actually achieved rate  $R_1$  to some target rate  $R_2$ . The resulting original power scaling rule, however, showed a too aggressive behavior in simulations by scaling the transmit power by too large quantities.

Rate  $R_1$  is calculated according to the transmit power  $P_1$  with some unknown channel coefficient  $|H|^2$  on a complex Gaussian channel with a noise power of  $N_0/(2t_s)$  in each real sub-channel,

$$R_1 = \log_2 \left( 1 + \frac{|\bar{H}|^2 \cdot P_1}{\sigma^2} \right). \quad (4.6)$$

This formula is an approximation as the channel coefficient  $|\bar{H}|^2$  in (4.6) is actually an *average* coefficient because of the use of the rate *average* for power correction in (4.5). This approach is chosen since the scheduler is designed to equalize the pathloss, while still exploiting block fading, thus exploiting the positive effects of multiuser gain.

If  $R_1$  is the rate which was delivered and  $R_2$  is the target rate that should be achieved, one possibility is to scale the power  $P_1$  to get the requested rate  $R_2$ . Therefore, (4.6) is solved for the power which results in

$$P_1 = (2^{R_1} - 1) \cdot \frac{\sigma^2}{|\bar{H}|^2}, \quad |\bar{H}|^2 \neq 0. \quad (4.7)$$

The ratio of the powers  $P_2$  for the rate  $R_2$  and  $P_1$  that achieved rate  $R_1$  is then:

$$\frac{P_2}{P_1} = \frac{(2^{R_2} - 1)}{(2^{R_1} - 1)} \doteq f(R_1, R_2). \quad (4.8)$$

As a rule for power scaling, the function  $f$  in (4.8) leads to an erratic behavior of the power adaptation in (4.5). A less aggressive scaling function can be found by (optimistically) assuming that the rate difference  $\Delta R \doteq R_2 - R_1$  is small (it is the job of the scheduler to ensure exactly this). Introducing the rate difference in (4.8) obtains

$$\frac{P_2}{P_1} = \frac{(2^{R_1 + \Delta R} - 1)}{(2^{R_1} - 1)} = 2^{\Delta R} \frac{(2^{R_1} - 2^{-\Delta R})}{(2^{R_1} - 1)}. \quad (4.9)$$

Assuming that  $\Delta R \approx 0$ , a Taylor series expansion based on  $2^{-\Delta R} \approx 1 - \Delta R \cdot \log(2)$  leads to

$$\frac{P_2}{P_1} = 2^{\Delta R} \frac{(2^{R_1} - 1 + \Delta R \cdot \log(2))}{(2^{R_1} - 1)}, \quad (4.10)$$

and by neglecting the additive term  $\Delta R \cdot \log(2)$  in the numerator the simple scaling rule is found:

$$P_2 = P_1 \cdot 2^{R_2 - R_1}. \quad (4.11)$$

## 4.2. Comparison to Proportional Fair Scheduling

The major differences between PFS and PCCLS are the introduction of the outer power-control loop and the user-individual scheduling exclusion, as well as user-individual adaptation of the transmit power to satisfy the requested maximum delay. It is the task of the outer power-control loop to compensate for the slowly changing long-term path-loss so that every user can be served such that their rate- and delay-requests are met and not only on a best-effort basis as in PFS. The goal is *not* to combat fading but rather to exploit it, i.e., to schedule a user when the fading channel's fast fluctuations temporarily lead to a "good" channel. By introduction of the "schedule Flag"  $\mathcal{F}_j(k)$ , the system resources can be used much more efficiently compared to PFS, since PCCLS aims to use the minimum transmit power for every user to guarantee requested rate and delay requirement. PCCLS does not have any queue-stability problems: it was reported [34] that PFS does not ensure queue stability, i.e., there is no guarantee that the length of a user's queue remains limited. This problem does not exist with PCCLS, as the rate requests are to be met within a specified time window: PCCLS will make sure that all arriving data will be delivered in time – otherwise, data packets would be useless due to violated delay constraints and could be dropped. This does, however, only hold if the required transmit power is available so that the user's rate request is satisfiable. To ensure exactly this is the task of the scheduling admission decision (3.1). The same argument does, however, not apply to PFS, since this scheme does - in its basic implementation - not provide the possibility to selectively exclude users from competing for resources. In practice, a user is being considered for scheduling for as long as the corresponding transmit buffer contains data, rather than being scheduled only if the requested rate by the application has not been satisfied. This functionality is a distinguishing factor for cross-layer scheduling schemes and therefore not available in traditional resource allocation algorithms.

The power control in a real-world system would have to be adapted, as, in the basic concept proposed, the power required for some data rate is derived from the "Shannon-Capacity" that can not be achieved in practice. Hence, a power margin that depends on the specific coding and

#### 4. Power-Controlled Cross-Layer Scheduling

User $j$	Rate Request $r_j$ in kbits/s	Delay Limit $\tau_j$ slots (Time/ms)	Average Channel Pow. Gain $G_j$ /dB
1	50.0	50 (20)	-27.0
2	100.0	186	-26.0
3	150.0	321	-25.0
4	200.0	457 (183)	-24.0
5	250.0	593	-23.0
6	300.0	729	-22.0
7	350.0	864	-21.0
8	400.0	1000 (400)	-20.0

Table 4.1.: User profiles for simulations; the channel symbol time is set to  $0.4\mu\text{s}$ .

User $j$	CD (optimal) $SNR_j$ in dB	TD $SNR_j$ in dB
1	17.21 (3.17)	17.71 (3.05)
2	17.57 (5.90)	18.00 (5.77)
3	17.53 (7.15)	17.90 (7.05)
4	17.31 (7.78)	17.66 (7.71)
5	16.96 (8.06)	17.32 (8.05)
6	16.50 (8.15)	16.93 (8.19)
7	15.93 (8.07)	16.48 (8.16)
8	15.22 (7.87)	16.02 (8.09)
Average	16.30	16.31

Table 4.2.: Theoretical limits: users' channel SNRs required to achieve ergodic capacities according to their rate requests and average channel power gains given in Table 4.1: general case (Code-Division, CD) [20] and with Time-Division constraint (TD). SNRs within brackets include “zero power” when users are not scheduled.

modulation schemes in use would have to be introduced. The scheduling decisions as such (i.e., which user is allowed to access the channel) would, however, still be justified.

### 4.3. Numerical Results

The numerical results presented are based on a generic system example summarized in Table 4.1 with  $J = 8$  users with different rate requests and delay constraints. The system is assumed to operate with a channel symbol time of  $0.4\mu\text{s}$  (parameters similar to CDMA2000) and a time-slot duration of 0.4 ms (1000 channel symbols per slot).

The profile of User 1 is that of a mobile phone call (hard delay constraint of  $\tau_j = 50$  time slots (i.e., 20 ms) but low rate request) while user 8 runs video streaming (relaxed delay constraints of

$\tau_j = 1000$  time slots (i.e., 400 ms) but higher rate demand). For the average channel power gains (that are to account for path loss) values were chosen suitable for nice visualization, since they don't affect the relative performance of the scheduling scheme 4.1. Apart from the fixed path loss, the transmission is subject to Rayleigh fading; the fading coefficients are assumed to be constant within each time slot but they change randomly from one time slot to another.

Table 4.2 shows the theoretical limits for the scheduling problem described by Table 4.1: it lists the channel SNRs (ratio of the transmit power and the receiver noise in dB) required to achieve ergodic capacities that equal the rate requests of the users given in Table 4.1; of course those SNRs are controlled by the allocated transmit power with the channel coefficients known at the scheduler. The general “code division” solution is optimal but it requires code superposition and successive interference cancellation at the decoder, which is impossible to implement in practice, as delay constraints would be violated. In this case, it is necessary to receive complete sets (i.e., packets) of decodable information from all users so that the interfering transmissions meant for the other users can be reconstructed from the decoded signals and be subtracted from the received signal. This strategy, while strictly optimal, is not compatible with a broad spectrum of delay constraints because the packet-length needs to be synchronized among all users: those with large rate requests and relaxed delay constraints will generally require large packet lengths. On the other hand, having to wait for the completed transmission of large packets will hurt those with very tight delay requirements. Furthermore, the complexity will grow in an unknown manner with the number of users who shall be decoded, since for interference cancellation it is necessary to reconstruct and subtract the signals of the stronger interferers. It is unknown however, how many interferers a specific user needs to decode and subtract before the desired signal becomes “available”. Therefore, the “TD” strategy is a much more desirable approach; the respective column in Table 4.2 contains the theoretical limits that apply with a time-division constraint (at most one user is scheduled per time slot). As obvious from the table, there is hardly any difference between the “CD” and the “TD” solution with respect to the average power used by the base station, so in practice one would certainly opt for the TD solution as it is much more simple to implement (even if there was no problem with delay constraints). The theoretical limits for the SNRs in Table 4.2 are lower bounds for the scheduling and power-allocation problem described by Table 4.1 because the rate requests are seen as ergodic capacities that have to be achieved and ergodic theory ignores delay constraints as all rates are *long-term* averages. In contrast to that the users' rate requests need to be achieved within each successive time window of length  $\tau_j$  (number of time slots), i.e. the rate average (3.3) within each time-window needs to fulfill (3.2): that is what defines hard delay constraints in the given context. Applying a scheduling scheme derived from the information theoretical solution for the time-division constraint will utterly fail in terms of the delay constraints but it will provide a lower limit of 16.31 dB for the average power used (see Fig. 4.2, Table 4.2).

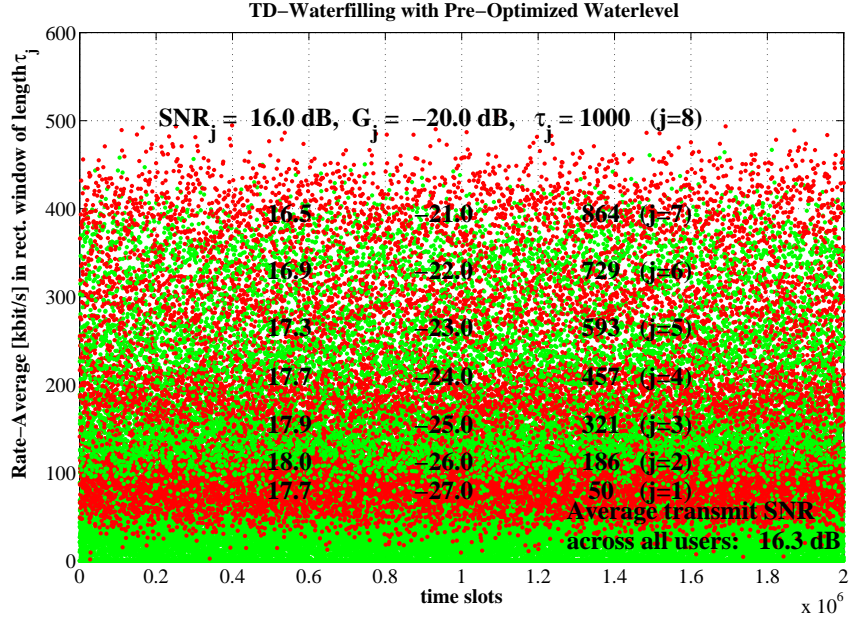


Figure 4.2.: Simulation results for the information-theoretical solution “TD-Waterfilling” according to [20]. The average rates (3.3) are plotted on the y-axis; on the x-axis the time slot indices  $k = n\tau_j - 1$  are given with  $n = 0, 1, 2, \dots$ .

As described in Section 3.3.2, the TD-solution maximizes the sum of the average weighted rates of all users under a power constraint. The desired point at the boundary surface of the capacity region is selected by tuning the rate-weighting factors  $\mu_j$  (3.12) so that the long-term average rates of the users are equal to their respective requested rates - this was done in advance of the simulation for this specific scenario. This scheme does however not provide the possibility to satisfy any delay constraints as only a long-term average rate can be achieved. The histogram of used channel states for each user can be seen in Fig. 4.3.

An alternative resource allocation scheme, “TD-Waterfilling Mod. B” has been discussed in Section 3.3.5. It improves upon the information-theoretical approach in several ways: the rate-demand of each user is increased so that the requested rate is satisfied (the power constraint is dropped so that this is actually possible), and users can temporarily be excluded from scheduling if their rate-request has been met in the current delay-window. These modifications do make a huge difference according to Fig. 4.4, and introduce an additional need for only 1.2 dB of transmit power. However, there is a small residual probability for an outage (i.e. the requested rate has not been met in a delay-window) still visible.

A simple, practical approach for guaranteeing the satisfaction of the user’s rate-demands is Round-Robin scheduling where every user is scheduled in a fixed repetitive pattern, and the delay con-



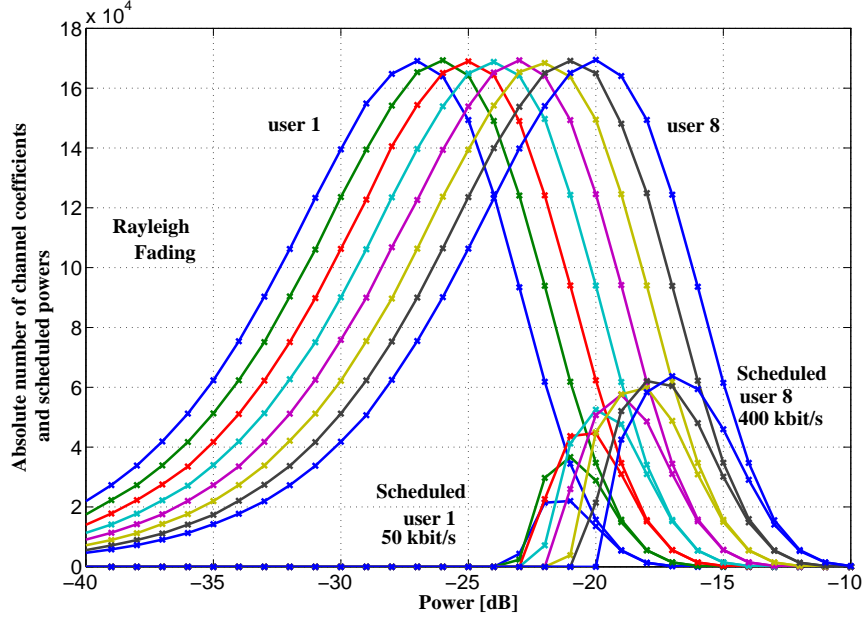


Figure 4.3.: Simulation results for the information-theoretical solution “TD-Waterfilling” according to [20]. The illustration shows that this scheme selects only very good channel states for scheduling. This behavior leads to extremely high total throughput, but it prevents any delay constraints to be met.

straints are fulfilled by equalizing the channel in each slot  $k$  with suitable transmit power (i.e., channel inversion). Such a scheme will consume 32.5 dB average transmit power (for the problem described by Table 4.1), so more than 16 dB more power is needed compared to the ergodic limits in Table 4.2; perhaps even worse, the peak power is unlimited for certain fading statistics, like e.g. Rayleigh. This is illustrated in Fig. 4.5. Obviously, this scheme does meet the required demands of all users at all times, but in the most uneconomical way, as can be observed in Fig. 4.6.

Proportional fair scheduling (PFS) can not be used in the given setup as PFS is a best-effort scheme that will serve the users with rates according to their channel qualities. Hence, it is impossible to give any rate guarantees or fulfill delay constraints.

Fig. 4.7 shows the results for the PCCLS-scheme: The achieved average rates (3.3) within time windows of size  $\tau_j$  are depicted. While the delay constraints according to (3.2) are nicely met for the high-rate user ( $j = 8$ , 400 kbit/s), the achieved average rates show some fluctuations for user  $j = 1$  with hard delay constraints (rate request of 50 kbit/s). But even in the latter case, the rate requests are fulfilled, as the value of (3.3) does not drop below the requested rate. The most significant result is that PCCLS fulfills the delay constraints and at the same time the average power used leads to a channel SNR of only 18.2 dB compared to 16.31 dB for the theoretical

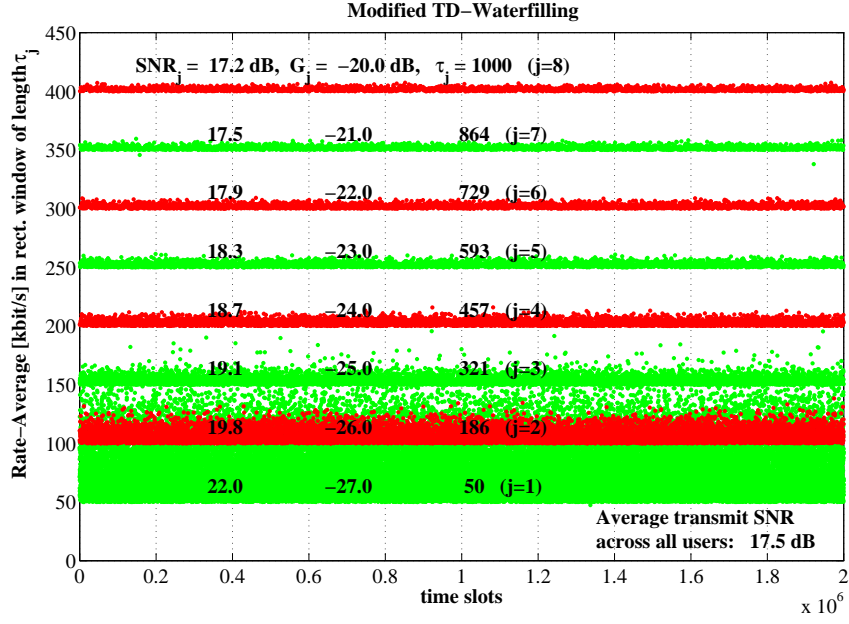


Figure 4.4.: Simulation results for “Modified TD-Waterfilling, Mod. B”. The average rates (3.3) are plotted on the y-axis; on the x-axis the time slot indices  $k = n\tau_j - 1$  are given with  $n = 0, 1, 2, \dots$

limits. Hence, compared to Round-Robin scheduling with channel-inversion, a *practical* power gain of more than 13 dB is achieved. A further comparison to Round-Robin scheduling reveals another practical advantage: the required power - although not constant - requires much smaller dynamic range.

Fig. 4.8 compares the histogram (in this case, the absolute amount) of the scheduled channel coefficients to the scheduled powers. It can clearly be seen (by comparing the graphs that correspond to the same users) that PCCLS tends to schedule a user when experiencing a favorable channel attenuation. This functionality is not available to Round Robin scheduling which is therefore much less efficient with respect to the transmitted information.

#### 4.4. Outage Analysis

An outage is defined as the event that the actually achieved average rate for user  $j$  is smaller than the requested rate  $r_j$ , within the delay window of length  $\tau_j$ . The probability of such an event is the outage probability. Based on the findings in the second part of this thesis, and on a corresponding publication that derives an analytic expression to calculate the outage probability for *max*-based scheduling schemes [4], the outage-probability of PCCLS and PFS is compared for the above

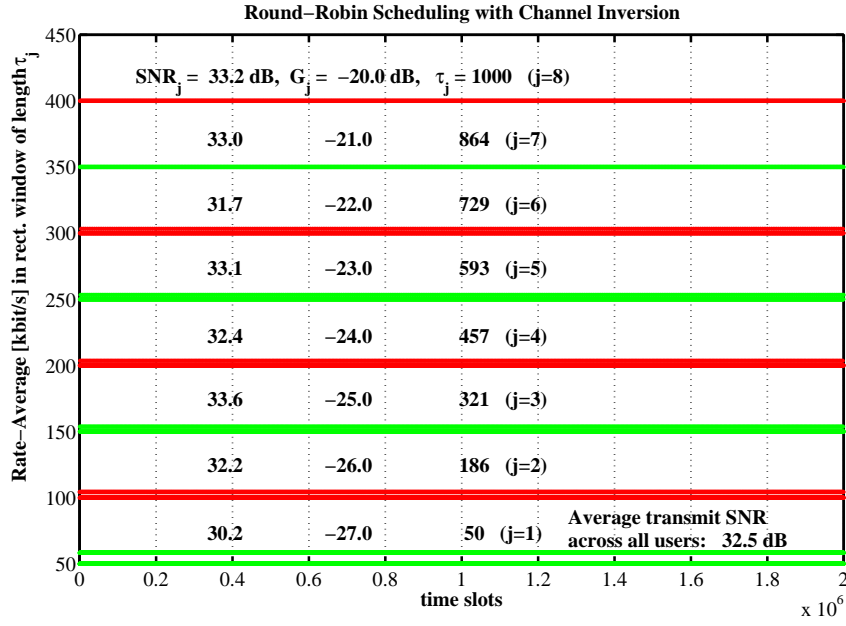


Figure 4.5.: Simulation results for Round-Robin scheduling with channel inversion. The average rates (3.3) are plotted on the y-axis; on the x-axis the time slot indices  $k = n\tau_j - 1$  are given with  $n = 0, 1, 2, \dots$

system parameters. The optimum outage probability characteristic for a scheduler guaranteeing delay constraints as well as a certain rate is a step-function. Ideally, the desired outage probability is zero for the requested rate and one for any bigger rate. PCCLS is designed to achieve this goal by optimizing the transmit power for every user and excluding users from the scheduling decision if their rate requests have been satisfied in the current delay window. Therefore, PCCLS can be seen to spend as much power as needed, but not more energy than required on every user. The following Fig. 4.9 illustrates the benefits of this behavior.

The curves were obtained by simulation, where first PCCLS was run on  $10^6$  blocks, and subsequently PFS was performed on the *same* channel coefficients, with the transmit power previously optimized in the PCCLS process. In addition, the filter used for the calculation of the average rate that PFS bases its scheduling decision on (an EWMA / exponentially-weighted, moving-average filter) was initialized with an optimized seed, so that the whole sequence of  $10^6$  blocks could be used for the outage-assessment. The advantage of this approach is that the PFS average does not require time to overcome a “settling time” - thus, there is no need to exclude this time at the beginning of the simulation from statistical analysis. For “User 1”, the characteristic behavior of PCCLS, as compared to PFS, is already visible: PCCLS optimizes the transmit-power, initialized as in (4.1), such that no outage occurs for the duration of the simulation (every simulation was run

#### 4. Power-Controlled Cross-Layer Scheduling

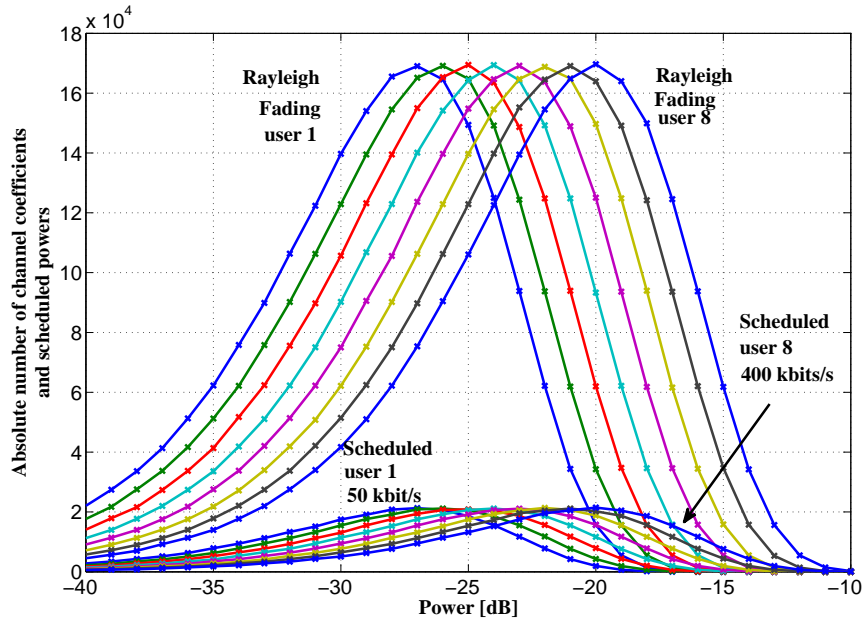


Figure 4.6.: Simulation results for Round-Robin scheduling with channel inversion. This scheme uses all channel states, regardless of any economical considerations.

two times so that the PCCLS scheduler could start with an already optimized transmit power). Due to users no longer being scheduled as soon as their respective rate-requests are met within the current delay-window, the probability of an outage increases rapidly above the requested rate, which is 50 kbit/s in case of “User 1”. PFS, on the other hand, is a best-effort service with a good performance for “User 1”, however, energy is wasted for rates above the required rate. This is hurting users with high rate demands, such as “User 8”: with a rate requirement of 400 kbit/s, this user suffers from energy being spent on users whose rate-requests have already been satisfied. In Fig. 4.9, this less than ideal behavior can be observed by comparing the cdfs of the outage probability of PFS (solid lines) to the target rates on the horizontal axis. For “User 1”, who demands a rate of exactly 50 kbit/s, the corresponding solid red cdf suggests that the requested bit rate is almost certainly surpassed within any given delay window. While this may not seem problematic at first glance (the requested rate is provided with high probability and PFS is a successful approach for this user), it also means that resources are being wasted on the service of “User 1”, when they should actually be spent on serving users with more demanding rate requests. In case of the 400 kbit/s requesting “User 8” for example, it can be seen that the the rate is almost never achieved within any delay window. “User 7”, on the other hand, is served a much higher rate than requested - in fact, the achieved rate would actually satisfy “User 8”. This is a result from using the user-individual powers that were optimized with the PCCLS scheduler for PFS-scheduling. PC-

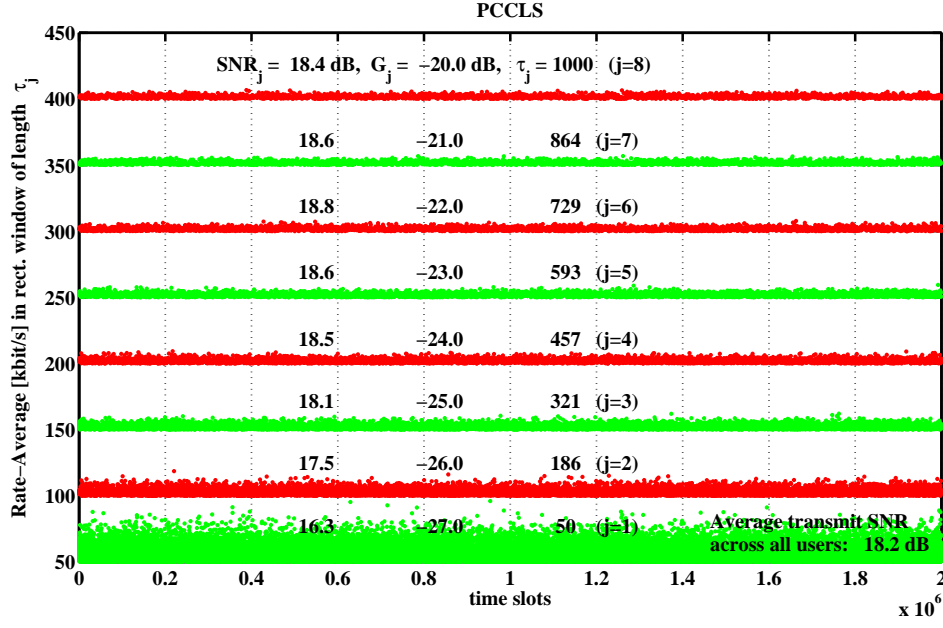


Figure 4.7.: Simulation results for Power-Controlled Cross-Layer Scheduling. The average rates (3.3) are plotted on the y-axis; on the x-axis the time slot indices  $k = n\tau_j - 1$  are given with  $n = 0, 1, 2, \dots$ .

CLS will, on the other hand, exclude any user from scheduling whose rate request has been met in a delay window and free up resources to serve users with higher demands. The distinct sign of this strategy is a range of outage probability cdfs that resemble step-functions: the outage is (close to) 0 for the requested rate, but almost instantly 100 % for any higher rate. For small requested rates, this is not exactly the case (i.e., the cdf is not a pure step-function for any greater rate than the requested one), since the transmission is not just stopped whenever the requested rate is reached *during* a delay-window, causing the scheme to transmit more than minimally required in the last scheduled block. PCCLS can therefore, while reducing the energy being spent to a minimum (even no users may be served temporarily), provide service with close-to-zero outage-probability, where PFS causes a waste of energy and will almost certainly fail to provide the requested rates to demanding users.

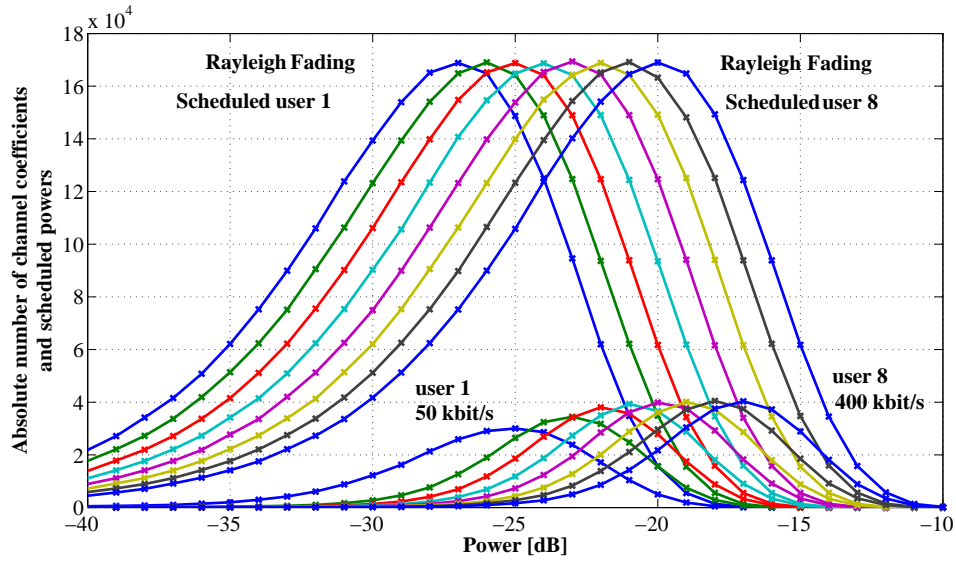


Figure 4.8.: Simulation results for Power-Controlled Cross-Layer Scheduling. The absolute number of occurrences is plotted over either the power of the channel coefficients (graphs with larger values) or the scheduled powers, in dB. It is interesting to observe the position of the maximum of the used channel coefficients histograms with respect to the maximum of the scheduled powers. Obviously, the PCCLS-scheme finds the “good” time-slots and avoids those that would lead to a rather inefficient use of transmit power.

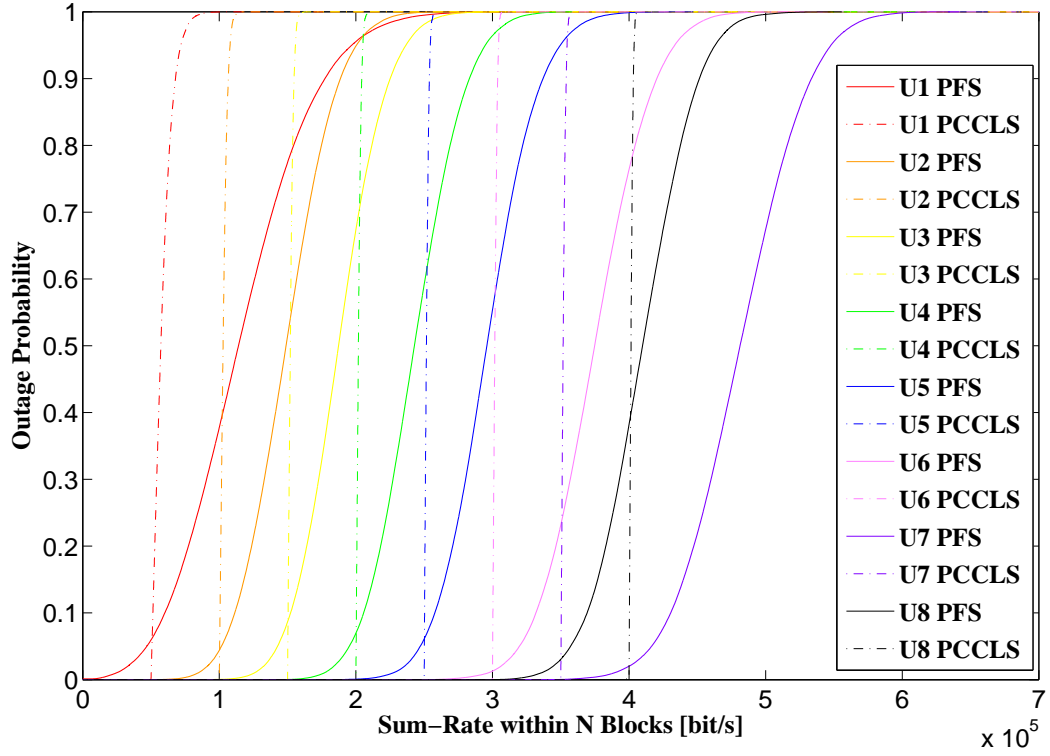


Figure 4.9.: Outage Probability comparison between Power-Controlled Cross-Layer Scheduling and Proportional Fair Scheduling, based on the same transmit powers for both schemes. The dashed lines show the outage probabilities for PCCLS, the solid lines for PFS.





## 5. Outage Analysis for *MAX*-Based Scheduling Schemes

The previous chapter has explored the design of a resource allocation strategy for minimizing outage in a delay-limited system. It was shown that two main strategies greatly improve system efficiency, these are user-individual scheduling exclusion when a user's requested rate has been met, user-individual scaling of the transmit power to meet the rate demands, and user-individual number of blocks over which the recently achieved rate is calculated. It was pointed out that outage is not a design criterion in classic resource allocation algorithms because of the classic ergodic approach and analysis that puts emphasis on asymptotic results rather than short-term behavior.

This chapter contributes a fully analytic approach to the probability of an outage, given a range of system parameters, for *max*-based scheduling algorithms. In the presented analysis, these are constant-power algorithms that base resource allocation decisions purely on a memoryless *max*-operation. Therefore, the following analysis is not suitable for PCCLS with its variable power allocation. This family of schedulers can be characterized by assigning a *figure of merit* (*FOM*) to each user, selecting the user with the maximum *FOM* for scheduling – two prominent scheduling schemes based on this approach, and therefore within the scope of the present analysis, are opportunistic [30] and proportional fair scheduling [29].

While proportional fair scheduling has, due to its popularity and use in modern wireless networks, drawn a lot of attention, there seems to be a lack of analysis concerning its “short-term” behavior with respect to *hard*<sup>1</sup> delay constraints. While proportional fairness has been investigated for fairness in delay-constrained environments [14,35], most publications address cell throughput analyses [35,36] rather than throughput for individual users. Also, asymptotic analyses [37,38] have usually been performed rather than analyses that take care of practically important short-term availability of throughput. Furthermore, perhaps for mathematical tractability, many analyses specialize on Rayleigh fading assumptions [38,39].

The approach discussed in Section 5.2 calculates the outage probability for a specific set of parameters, which include hard delay constraints, arbitrary channel statistics for an arbitrary number

---

<sup>1</sup>The definition of a *hard* delay constraint is detailed below.

## 5. Outage Analysis for MAX-Based Scheduling Schemes

of users, a constant transmit-power of every user, and a receiver noise-level for every user. The present analysis can even be applied to users with an arbitrary number of antennas (as long as selection diversity is applied). In this case, the statistical properties of the channel have to be adapted - this does not have any consequences for the analytic approach.

### 5.1. Channel Model

The following analysis elaborates on the downlink case with perfect Channel State Information (CSI) at the transmitting base-station serving  $U$  users; CSI is required for the scheduler to take optimum scheduling decisions. The following analysis applies a block-fading AWGN channel model for which the “Gaussian Shannon-capacity” is assumed to be achieved within each block. However, the results of the analysis can be extended to non-ideal systems by assuming a power margin. An outage is defined as the event that the sum rate  $R_{\text{sum}}$  (a random variable) transmitted within a window of  $N$  consecutive blocks is below a required rate  $r_{\text{sum}}^{\text{req}}$  (a given number, typically determined by the service constraints of an application). This definition also clarifies the notion of a “hard” delay constraint: there is no possibility of carrying over “excess-information” transmitted within one window of  $N$  consecutive blocks to the next window during which a required rate has to be achieved. This is a realistic assumption, e.g., in wireless multimedia transmission where a source signal is divided into blocks of samples (e.g. speech/audio samples or pixels) and each block is separately source-encoded and transmitted. At the receiving end, all source-bits that belong to a block of samples must be received “in time”; otherwise the source signal can not be reconstructed<sup>2</sup> rendering the block of received data useless. The analysis does not distinguish the cases where these  $N$ -block windows overlap or not: the probability of an outage is strictly defined as the probability that  $R_{\text{sum}} < r_{\text{sum}}^{\text{req}}$  in *any* window of  $N$  consecutive blocks. This definition does therefore not allow for any “delay”, as it was used in the previous chapter 3: there, a certain rate needed to be provided in non-overlapping windows on  $N$  consecutive blocks, leading to a scenario where a user needed to wait for the passing of a delay-window of  $N$  blocks, before the required rate was available. In this analysis however, the user may require a minimum rate during  $N$  consecutive blocks *at any time* - this was previously labeled “short-term availability” of throughput.

---

<sup>2</sup>Although embedded source coding schemes (or multiple descriptions) may be used that produce a “basic” quality from a subset of the transmitted source bits, those techniques are not widely used, because embedded multimedia coding does often not produce a satisfactory quality trade-off compared with the quality-level of non-embedded coding at various rate-levels transmitted in parallel (multi-cast).

For the statistical analysis of the outage probability to hold, uncorrelated channel-coefficients (both between users and individually for each user) are assumed. Without loss of generality, the probability of an outage is calculated for one distinct user ( $u_1$ , representing user 1), given the statistical channel knowledge of all users in the system.

## 5.2. Outage Probability of *max*-Based Scheduling

As pointed out above, the fundamental approach does not distinguish different scheduling schemes, as long as the core-decision is memoryless and based solely on a *max*-operation to identify the user with the *currently* best *FOM*. This *FOM* may be the channel-power coefficients  $|h|^2$  in case of opportunistic scheduling (which is fully equivalent to taking the rate  $R$  achievable in the current block into account), or it may be the normalized current rate  $\bar{R} = R/T$  in case of proportional-fair scheduling, with  $T$  denoting the long-term average of rates achieved in the past. To preserve most general validity, the following analysis will therefore not deal with *rates* or *channel-coefficients* per se, but will consider *FOMs* exclusively, so that the same analysis can directly be applied to both opportunistic and proportional-fair scheduling. Specific results are subsequently presented in Section 5.3.<sup>3</sup>

Based on the above presented naming conventions, the probability of an outage for user 1 is the following sum of probabilities of mutually exclusive events:

$$P_{1,\text{out}}(x_{1,\text{sum}}^{\text{req}}) = \sum_{k=0}^N P\{X_{1,\text{sum}} < x_{1,\text{sum}}^{\text{req}} | S_1 = k\} P\{S_1 = k\}. \quad (5.2)$$

Here,  $X_{1,\text{sum}}$  denotes the sum of the *FOMs* in individual blocks of user  $u_1$  that is achieved within a window of  $N$  consecutive blocks, i.e.  $N$  expresses a hard delay constraint. The quantity  $x_{1,\text{sum}}^{\text{req}}$  is the sum of *FOMs* that is requested<sup>4</sup> to be achieved within a window of  $N$  blocks; if it is *not* achieved an outage event occurred, and the goal in what follows is to analytically compute this outage probability.  $S_1$  denotes the number of times user 1 is scheduled out of  $N$  possible times within a window of length  $N$  blocks:  $S_1$  is, therefore, an integer-valued random variable ranging from 0 to  $N$ . Throughout, the algorithms are developed to compute the probability of an outage

---

<sup>3</sup>Please note that the analysis presented in this thesis is, in fact, not limited to strictly *max*-based scheduling. With the same approach, also scheduling schemes performing *min*-based scheduling (such as minimization of interference) can be assessed, since

$$\arg \min\{x_1, x_2, \dots, x_U\} = \arg \max\{-x_1, -x_2, \dots, -x_U\}. \quad (5.1)$$

The presented considerations are therefore applicable to all scheduling schemes based on the search for extremums, however, all examples refer to *max*-based schemes.

<sup>4</sup> $x_{1,\text{sum}}^{\text{req}}$  is a number such as a bit rate requested by an application

## 5. Outage Analysis for MAX-Based Scheduling Schemes

for user 1. If the probability of an outage shall be calculated for a different user, this can be achieved by reordering of the user indices; hence, no generality is lost.

The first addend of (5.2) can be simplified to  $P\{S_1 = 0\}$ , since an outage *will* occur when the user is not scheduled at all, i.e.,

$$P\{X_{1,\text{sum}} < x_{1,\text{sum}}^{\text{req}} | S_1 = 0\} = 1. \quad (5.3)$$

Analysis of (5.2) reveals two different structures: the expressions  $P\{X_{1,\text{sum}} < x_{1,\text{sum}}^{\text{req}} | S_1 = k\}$  on the one hand, and the probabilities,  $P\{S_1 = k\}$ , for user 1 to be scheduled  $k$  times on the other hand. Following this structure, the mathematical derivation of the outage-calculation for *max*-based schedulers is organized accordingly: In Section 5.2, the expression  $P\{S_1 = k\}$  is analyzed, and in Section 5.2 the expressions of the form  $P\{X_{1,\text{sum}} < x_{1,\text{sum}}^{\text{req}} | S_1 = k\}$  are explored.

### Probability of User 1 being Scheduled $k$ Times.

Given that user 1 can only be scheduled once *or* twice *or* three times, etc (where “*or*” is exclusive) the event “user 1 is scheduled  $a$  times within  $N$  blocks” is distinct from the event “user 1 is scheduled  $b$  times within  $N$  blocks” for  $a \neq b$ . Therefore,  $P\{S_1 = k\}$  can be calculated by drawing a probability tree of depth  $N$ , with  $p_1$  being the probability that user 1 is scheduled in a block, and  $1 - p_1$  meaning the opposite. Therefore,

$$P\{S_1 = k\} = \binom{N}{k} \cdot p_1^k \cdot (1 - p_1)^{(N-k)}. \quad (5.4)$$

The actual calculation of an outage probability requires  $p_1$  to be known, with  $p_1$  being the probability that user 1 is scheduled in a particular block. From a *max*-based scheduling perspective this is equivalent to the *current FOM*  $X_1$  of user 1 being larger than the *current FOMs*  $X_2, \dots, X_U$  of all other users  $u_2, \dots, u_U$ . These random variables (or their realizations) could be rates in the case of opportunistic scheduling or normalized rates in case of proportional-fair scheduling or something entirely different. It is only important for this analysis that these *FOMs* are compared, and the user associated with the maximum *FOM* is chosen to be scheduled in the current block (out of  $N$  blocks, over which the sum of *FOMs* is calculated to determine the outage probability by (5.2)). The “current” *FOMs* are going to be called  $X_u$ , with  $X_1$  the *FOM* of user 1 (the user of interest). Since all competing users (i.e. the users associated with *FOMs*  $X_2, \dots, X_U$  are compared to user 1 in a *memoryless max*-operation, the first step is to simplify the problem by only considering the largest *FOM* among all competing users, and this *FOM* will be denoted by  $Y$ . This approach does not ignore any information available, since the statistics of all users are implicitly taken into account. With those assumptions, the following theorem is stated:

## 5.2. Outage Probability of max-Based Scheduling

**Theorem 5.2.1.** *The probability  $p_1$  that the FOM  $X_1$  of user 1 ( $u_1$ ) is larger than the FOMs  $X_2, X_3, \dots, X_U$  of all competing users  $u_2 \dots u_U$  is given by*

$$p_1 \doteq P\{X_1 > Y\} = 1 - \sum_{u=2}^U \int_{-\infty}^{\infty} \left( \frac{d}{dy} F_{X_u}(y) \right) \prod_{\substack{k=1 \\ k \neq u}}^U F_{X_k}(y) dy \quad (5.5)$$

$$= \int_{-\infty}^{\infty} \left( \frac{d}{dy} F_{X_1}(y) \right) \prod_{k=2}^U F_{X_k}(y) dy \quad (5.6)$$

where  $Y \doteq \max\{X_2, X_3, \dots, X_U\}$  and the cumulated density functions (cdfs) of the FOMs of the users are defined by  $F_{X_u}(y) \doteq P\{X_u < y\}$ .

*Proof.* (Theorem 5.2.1)

$$P\{X_1 > Y\} = \int_{-\infty}^{\infty} P\{X_1 > Y | Y = y\} f_Y(y) dy \quad (5.7)$$

$$= \int_{-\infty}^{\infty} (1 - F_{X_1}(y)) f_Y(y) dy \quad (5.8)$$

$$= 1 - \int_{-\infty}^{\infty} F_{X_1}(y) f_Y(y) dy \quad (5.9)$$

This leaves  $F_Y(y)$ , the cdf of the largest competing FOM, to be defined. As  $F_Y(y)$  is the probability that all other users at the same time have a FOM smaller than  $y$ , statistical independence of all FOMs leads to the simple “product probability”

$$F_Y(y) = P\{Y \leq y\} = \prod_{u=2}^U P\{X_u \leq y\} = \prod_{u=2}^U F_{X_u}(y), \quad (5.10)$$

and the corresponding pdf  $f_Y(y)$  is, therefore,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \prod_{u=2}^U F_{X_u}(y) \quad (5.11)$$

$$= \sum_{u=2}^U \left[ \left( \frac{d}{dy} F_{X_u}(y) \right) \prod_{\substack{k=2 \\ k \neq u}}^U F_{X_k}(y) \right]. \quad (5.12)$$

Hence, the probability of the FOM of user 1 being larger than those of the competing users can

## 5. Outage Analysis for MAX-Based Scheduling Schemes

be expressed as:

$$P\{X_1 > Y\} = 1 - \int_{-\infty}^{\infty} F_{X_1}(y) \sum_{u=2}^U \left( \frac{d}{dy} F_{X_u}(y) \right) \prod_{\substack{k=2 \\ k \neq u}}^U F_{X_k}(y) dy \quad (5.13)$$

$$= 1 - \sum_{u=2}^U \int_{-\infty}^{\infty} F_{X_1}(y) \left( \frac{d}{dy} F_{X_u}(y) \right) \prod_{\substack{k=2 \\ k \neq u}}^U F_{X_k}(y) dy \quad (5.14)$$

$$= 1 - \sum_{u=2}^U \int_{-\infty}^{\infty} \left( \frac{d}{dy} F_{X_u}(y) \right) \prod_{\substack{k=1 \\ k \neq u}}^U F_{X_k}(y) dy \quad (5.15)$$

Now, the integral in (5.15) will be evaluated by integration by parts ( $\int_a^b u'v = [uv] \big|_a^b - \int_a^b uv'$ ):

$$p_u \doteq \int_{-\infty}^{\infty} \left( \frac{d}{dy} F_{X_u}(y) \right) \prod_{\substack{k=1 \\ k \neq u}}^U F_{X_k}(y) dy \quad (5.16)$$

$$= F_{X_u}(y) \prod_{\substack{k=1 \\ k \neq u}}^U F_{X_k}(y) \Big|_{-\infty}^{\infty} \quad (5.17)$$

$$- \int_{-\infty}^{\infty} F_{X_u}(y) \sum_{\substack{k=1 \\ k \neq u}}^U \left( \frac{d}{dy} F_{X_k}(y) \right) \prod_{\substack{n=1 \\ n \neq k \\ n \neq u}}^U F_{X_n}(y) dy$$

$$= 1 - \sum_{\substack{k=1 \\ k \neq u}}^U \int_{-\infty}^{\infty} F_{X_u}(y) \left( \frac{d}{dy} F_{X_k}(y) \right) \prod_{\substack{n=1 \\ n \neq k \\ n \neq u}}^U F_{X_n}(y) dy \quad (5.18)$$

$$= 1 - \sum_{\substack{k=1 \\ k \neq u}}^U \int_{-\infty}^{\infty} \left( \frac{d}{dy} F_{X_k}(y) \right) \prod_{\substack{n=1 \\ n \neq k}}^U F_{X_n}(y) dy \quad (5.19)$$

$$= 1 - \sum_{\substack{k=1 \\ k \neq u}}^U p_k = p_u \quad (5.20)$$

with (5.19) proving (5.5), and due to (5.20),

$$\sum_{k=1}^U p_k = 1. \quad (5.21)$$

Finally, (5.6) is obtained from (5.19) according to

$$P\{X_1 > Y\} = 1 - \sum_{u=2}^U p_u = p_1. \quad (5.22)$$

The normalization property in (5.21) leads to the simplification in the calculation of  $P\{X_1 > Y\}$  in (5.22). □

### Outage Probability, Conditioned on $S_1 = k$ .

The pdf of the “system-*FOM*” (as opposed to user-individual *FOMs*) that can be achieved system-wide, i.e. across all users, in the current transmit block (out of  $N$  blocks we consider), can be found applying the maximum-order statistic theorems for i.n.i.d (independent, *non*-identically distributed) random variables [40]. However, observing the *FOMs* that are assigned to a *specific* user *in case of being scheduled* reveals that these *FOMs* are not distributed according to the maximum-order statistic. Instead, it will be shown that the weighted pdfs of the *FOMs* assigned to the users will sum up to the maximum-order statistic of the system-*FOM*. The weights can be found by calculating the probabilities of the users being scheduled in a certain block  $n$ .

With

$$Z \doteq \max\{X_1, X_2, \dots, X_U\}, \quad (5.23)$$

the cdf  $F_Z(z)$  of the maximum of the system-*FOMs* (i.e. of the current *FOMs* of all users) can, for i.n.i.d. *FOMs*, be calculated as [40]:

$$F_Z(z) = P\{Z \leq z\} = \prod_{u=1}^U P\{X_u \leq z\} = \prod_{u=1}^U F_{X_u}(z). \quad (5.24)$$

Calculating the pdf  $f_Z(z) = \frac{d}{dz}F_Z(z)$  of (5.24) reveals the following structure:

$$f_Z(z) = \sum_{u=1}^U \left[ \left( \frac{d}{dz} F_{X_u}(z) \right) \prod_{\substack{k=1 \\ k \neq u}}^U F_{X_k}(z) \right] \quad (5.25)$$

A comparison of (5.16) and (5.25) reveals structural similarity and suggests that the maximum order statistics of statistically independent (but not necessarily identically distributed) random variables is a linear combination of the statistics of the instances of the individual random variables which contribute to the maximum order statistics, weighted with the probability  $p_u$  of each individual random variable to contribute the maximum. The probability  $p_1$  was, however, computed above in (5.5) and (5.6) with the general form for any user  $u$  given in (5.16). Hence, (5.25) is equivalently written as

$$f_Z(z) = \sum_{u=1}^U \left[ \tilde{f}_{X_u}(z) \cdot p_u \right]. \quad (5.26)$$

## 5. Outage Analysis for MAX-Based Scheduling Schemes

Figure 5.1 illustrates this elegant relation (5.26, compare 5.2) between the individual user's *FOMs* and the *system-FOM*, in this particular case for Proportional Fair Scheduling. The statistics for the users' Rayleigh-fading channels were chosen in a way that make the weighted summation clearly visible. When the channel statistics are similar, the stair-like appearance is lost and replaced by something like a "smooth" cdf, as can be observed in Figure 5.2. Specifically, Fig. 5.1 depicts the weighted cdfs of the outage probabilities for user 1 being scheduled  $k = 1 \dots N$  times. The weights are the probabilities of user 1 being scheduled exactly  $k$  times. The sum of these weighted cdfs results in the total outage probability of user 1.

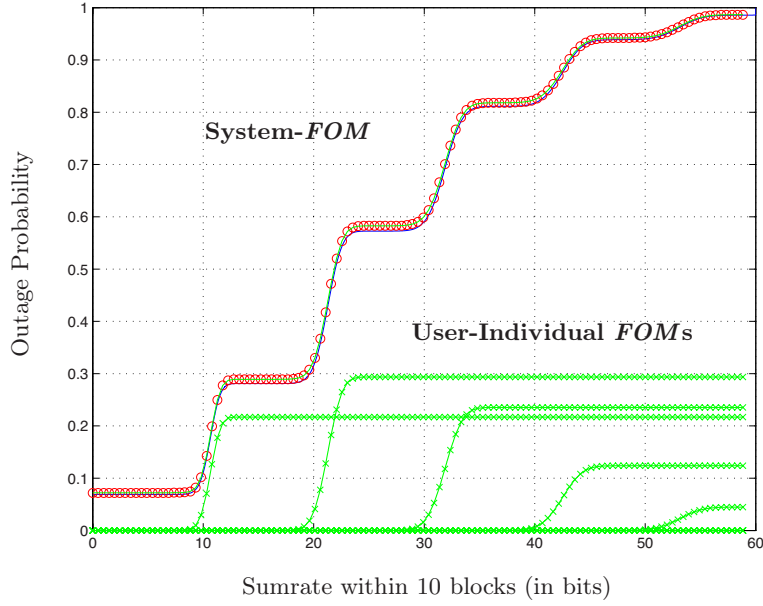


Figure 5.1.: Illustration of Equation (5.2). The plot shows an example for Proportional Fair Scheduling and different channel statistics for all users (i.e., no two *FOM*-cdfs are the same or even similar, allowing for a nice separation and characteristic stair-like appearance of the *system-FOM*). The plot illustrates the suitability of the algorithmical approach by comparing simulated values (solid lines) to calculated values (red circles).

With

$$\tilde{f}_{X_u}(z) = \frac{(\frac{d}{dz}F_{X_u}(z)) \prod_{\substack{k=1 \\ k \neq u}}^U F_{X_k}(z)}{p_u}, \quad (5.27)$$

the pdf associated to the *FOM* of user  $u$  in case this user is scheduled for transmission. The



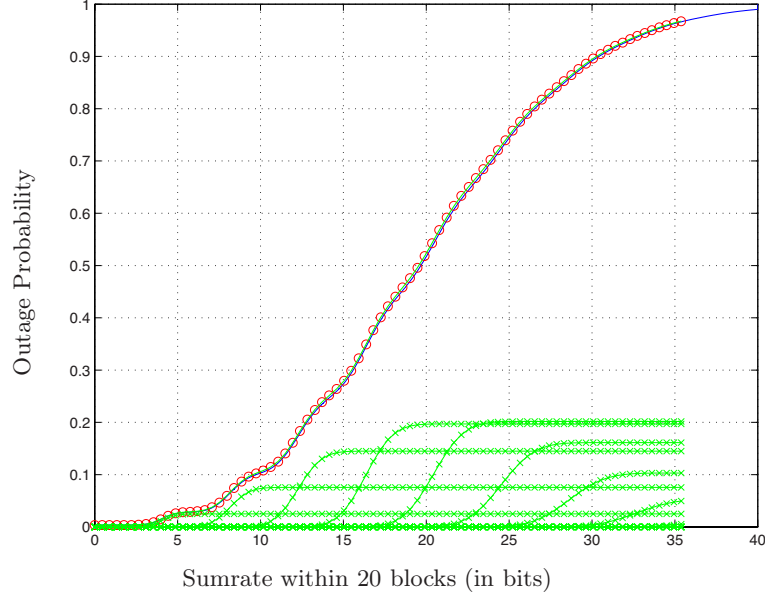


Figure 5.2.: In contrast to Fig. 5.1, this illustration shows the effect of a less well-defined separation between the individual user's *FOMs* cdfs.

corresponding pdf of the *FOM* of the scheduled user 1 can therefore be written as

$$\tilde{f}_{X_1}(z) = \frac{f_{X_1}(z) \cdot F_{X_2}(z) \cdot F_{X_3}(z) \cdots F_{X_U}(z)}{p_1} \quad (5.28)$$

and be used to compute the outage probability

$$P\{X_{1,\text{sum}} < x_{1,\text{sum}}^{\text{req}} | S_1 = 1\} = \int_0^{x_{1,\text{sum}}^{\text{req}}} \tilde{f}_{X_1}(x) dx \quad (5.29)$$

for the case that user 1 is scheduled once ( $S_1 = 1$ ) within  $N$  transmit blocks. For  $S_1 = 2$  (i.e., user 1 being scheduled twice within  $N$  blocks) the result reads

$$P\{X_{1,\text{sum}} < x_{1,\text{sum}}^{\text{req}} | S_1 = 2\} = \int_0^{x_{1,\text{sum}}^{\text{req}}} \int_0^\infty \tilde{f}_{X_1}(\rho) \cdot \tilde{f}_{X_1}(x - \rho) d\rho dx \quad (5.30)$$

as the pdf of a sum of two independent pdfs follows from a convolution of the pdfs. Consequently, for  $S_1 = k$ ,

$$P\{X_{1,\text{sum}} < x_{1,\text{sum}}^{\text{req}} | S_1 = k\} = \int_0^{x_{1,\text{sum}}^{\text{req}}} \underbrace{\left( \tilde{f}_{X_1} * \cdots * \tilde{f}_{X_1} \right)}_{k \text{ times}}(x) dx, \quad (5.31)$$

where “ $*$ ” denotes convolution.

### 5.3. Results for Specific Schedulers and Channels

In the following, the pdfs of the *FOMs* of opportunistic and proportional-fair scheduling are derived for the special case of user-individual Rayleigh-fading channels. The reason to pick Rayleigh-fading is that it allows for compact analytical expressions for the statistics of the *FOMs*, i.e.,  $f_{X_1}(z)$  and  $F_{X_u}(z)$ , that are used in (5.6) and (5.28): using the results from there, the outage probabilities are calculated by (5.29)–(5.31) and finally in (5.4) and (5.2). For the last two steps a numerical approach (incl. convolutions) is required anyway (as analytical solutions are intractable) so the whole scheme can as well be run with measured pdfs  $f_{X_1}(z)$  and cdfs  $F_{X_u}(z)$  that are used in the numerical implementations, e.g. of the integrations in (5.29)–(5.31). This is to say that, although the analytical results from Section 5.2 are demonstrated for simple Rayleigh fading, they could equally be applied for any measured channel statistics, so the method as such is not restricted to any specific simple channel model.

#### 5.3.1. Opportunistic Scheduling

If a channel-coefficient  $h$  is Rayleigh-distributed with scale-parameter  $\lambda$ , the pdf can be expressed as:

$$f_H(h) = \frac{h}{\lambda^2} e^{-\frac{h^2}{2\lambda^2}}, \quad h > 0. \quad (5.32)$$

According to elementary probability calculus (e.g. [41, p. 130]) the magnitude-square  $|H|^2$  has the exponential distribution

$$f_{|H|^2}(|h|^2) = \frac{1}{2\lambda^2} e^{-\frac{|h|^2}{2\lambda^2}}. \quad (5.33)$$

If the scheduler applies a classic opportunistic scheme (i.e. decides in favor of the user with the highest achievable rate), it applies a *max*-operation on the rates for all users. Therefore, the pdf of the rate  $r$  needs to be determined. With

$$r = g(|h|^2) = \log_2(1 + \frac{P}{2\sigma^2}|h|^2) \quad (5.34)$$

(Gaussian channel capacity for a fixed channel coefficient  $h$ ) and the derivative

$$g'(|h|^2) = \frac{1}{\log(2)} \cdot \frac{\frac{P}{2\sigma^2}}{|h|^2 \frac{P}{2\sigma^2} + 1}, \quad (5.35)$$

the standard pdf-transformation (e.g. [41])

$$f_R(r) = f_{|h|^2}(|h|^2) / g'(|h|^2) \Big|_{|h|^2=g^{-1}(r)} \quad (5.36)$$

(note that  $g(|h|^2)$  is monotonic) results in

$$f_R(|h|^2) = \log(2) \cdot \frac{|h|^2 \frac{P}{2\sigma^2} + 1}{\frac{P}{2\sigma^2}} \cdot \frac{1}{2\lambda^2} \cdot e^{-\frac{|h|^2}{2\lambda^2}}, \quad (5.37)$$

and substituting  $|h|^2 = \frac{2\sigma^2}{P}(2^r - 1)$  leads to the rate-pdf

$$f_R(r) = \log(2)\beta \cdot 2^r \cdot e^{-\beta(2^r-1)} . \quad (5.38)$$

with

$$\beta = \frac{\sigma^2}{\lambda^2 P} . \quad (5.39)$$

The cdf follows by direct integration of (5.38); it reads

$$F_R(r) = 1 - e^{-\beta(2^r-1)} . \quad (5.40)$$

The pdf (5.38) and the cdf (5.40) are to be used in place of  $f_{X_1}(z)$  and  $F_{X_u}(z)$  in (5.6) and (5.28) to compute  $p_1$  and  $\tilde{f}_{X_1}(z)$  respectively. The pdf and the cdf only depend on the parameter  $\beta$  in (5.39), which contains the transmit power  $P$ , the parameter  $\lambda$  of the fading process and noise variance  $\sigma^2$  at the receiver. All those parameters can (and will in practice) be different for every user.

It is interesting, however, to study the case when  $\beta$  is the same for all users, as then compact analytical expressions can be found for  $p_1$  and  $\tilde{f}_{X_1}(z)$ . For the probability of user  $u_1$  to be scheduled,

$$p_1 = \frac{1}{U} , \quad (5.41)$$

with  $U$  the number of users, as there is no preference for any user as they all share the same statistics. The same would, of course, result from evaluating (5.6) with (5.38) and (5.40).

Using a Matlab implementation, Figure 5.3 illustrates the outage probability for a system with 5 users, a window-length of 10 and 20 blocks and transmit SNRs ( $= 10 \log_{10} \frac{P}{2\sigma^2}$ ) of 0, 10 and 20dB. The fading statistics of all users are assumed to be Rayleigh-distributed with scale-parameters  $\lambda = [2, 2, 3, 4, 5]$  for the 5 users, respectively. The x-axis is the sum-rate across all blocks in a window of size  $N$  ( $= 10, 20$ ) that is achieved by user 1. This sum rate adds the rates in “bits per channel-use” according to (5.34) from the individual blocks (all of which are assumed to be infinitely long by the common block fading model) within a window of size  $N$  blocks: this approach has been chosen to separate the curves in the plots (and, hence improve readability). In practice one could normalise the sum-rate axis by the window length  $N$  to obtain a normalized average rate in bits per channel use. The probability of outage was calculated for user 1 by the method described in Section 5.2 for transmit SNRs of (0, 10, 20 dB from left to right in the curves). The solid curves stem from a Monte-Carlo simulation approach, and the individually marked points were calculated according to the analytical approach stated above.

As to be expected, the figure shows that higher SNR and larger window-size  $N$  always cause a lower outage probability. However, the outage probability for the proportional fair scheduling scheme is about 0.6 *at best*, so that the approach is rather unusable for any practical purposes.

## 5. Outage Analysis for MAX-Based Scheduling Schemes

The simulation parameters were chosen to be like this in order to show the achievable gain with proportional-fair scheduling that is shown with identical simulation parameters in the next Section 5.3.2.

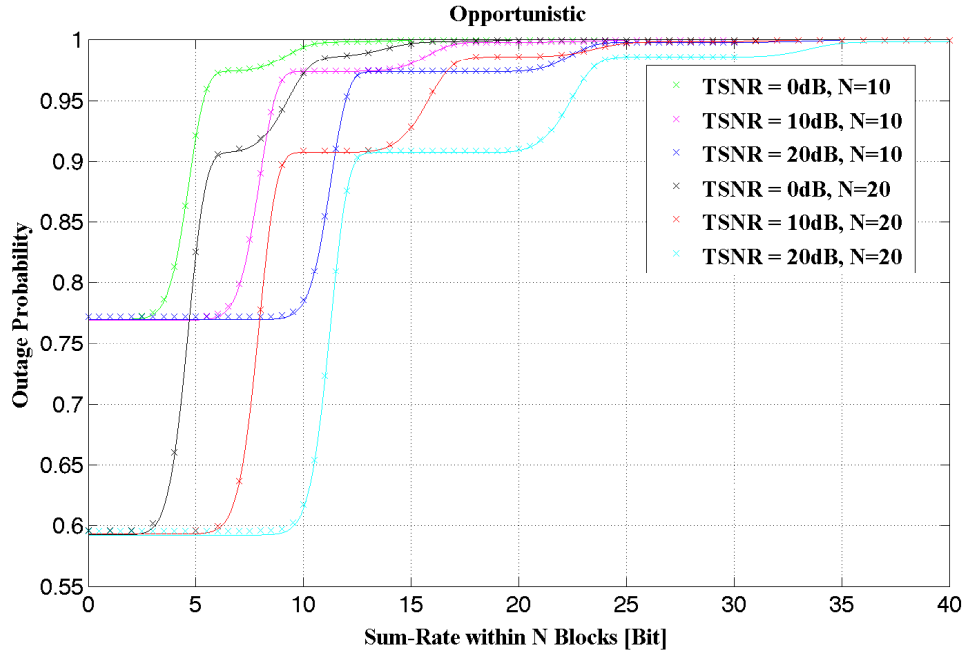


Figure 5.3.: Outage probability of user 1 for 5-user opportunistic scheduling in a Rayleigh-fading environment with transmit-SNRs 0dB, 10dB, and 20dB and window-lengths of 10 and 20 blocks. Numerical results based on  $10^6$  simulated windows of length  $N$  (solid lines) are compared with outage probabilities computed analytically (individually marked points).

The figure also demonstrates that the analytical approach (with numerical computations of integrals and convolutions) is rather accurate as the marked points computed are sitting almost exactly on the curves found by extensive Monte-Carlo simulations.

### 5.3.2. Proportional Fair Scheduling

Opportunistic scheduling and proportional fair scheduling are similar schemes in the sense that both select the user with the highest  $FOM$ . For opportunistic scheduling, the  $FOM$  is the achievable rate in the current block; for proportional-fair scheduling, the  $FOM$  is the rate that can be achieved in the current block, divided by the “recently achieved” rate: a user is scheduled, if the ratio is larger than for any other user. This scheme is modeled by using a long-term average for the normalization

of the current rate. Since there is currently no universally-valid closed-form expression for the long-term average achieved rate known for proportional-fair scheduling, the iterative solution suggested in [37] and refined in [42] is used:

$$E\{T_1\} = \int_0^\infty r \cdot f_{R_1}(r) \prod_{u=2}^U F_{R_u} \left( r \cdot \frac{E\{T_u\}}{E\{T_1\}} \right) dr . \quad (5.42)$$

This expression has been shown by [43] to converge to the measured mean throughput of  $u_1$ , if the instantaneous rate follows a continuous distribution  $f_{R_u}(r)$ .

Therefore,  $u_1$ 's figure of merit for PF-scheduling is

$$\bar{R}_1 = \frac{R_1}{E\{T_1\}} , \quad (5.43)$$

which is also the decision rule for PF-scheduling (the decision is based on this figure of merit / FOM by applying a simple memory-less maximization across all users).

Then, the pdf of user 1's *FOM* can be written as (compare (5.38)):

$$f_{\bar{R}_1}(\bar{r}_1) = E\{T_1\} \cdot \log(2)\beta_1 \cdot 2^{E\{T_1\}\bar{r}_1} \cdot e^{-\beta_1(2^{E\{T_1\}\bar{r}_1}-1)} \quad (5.44)$$

Again, this pdf is to be used in (5.29), in place of  $f_{X_1}(z)$ . The following Figure 5.4 illustrates the precision of the approach. The simulation parameters chosen are the same as in the opportunistic scheduling case. Please note that a proportional-fair scheduler updates the average-achieved rates  $T_u$  of the served users by applying an exponentially-weighted moving-average (EWMA) filter. Since the expected value for the average-achieved rates  $T_u$  was used, the EWMA-filter in the simulations was assigned a long memory, i.e., a weighting-factor of  $\alpha = 10^5$ .

The presented results for opportunistic and proportional-fair scheduling demonstrate the accuracy of the approach. It should be noted that, although a Rayleigh-fading channel was assumed for all users (albeit with different parameters for each user), the presented approach is by no means limited to a Rayleigh-fading environment. This was only chosen for mathematical tractability and ease of analysis. The presented principles can be applied to any scheduler which assigns resources based on a *max*-operation. In that, the approach is highly general, but at the same time, as demonstrated, also practically useful. There are, however, limitations to the generality of the approach: it is only suitable for resource allocation schemes that are based on a *memoryless max-operation*. PCCLS can therefore not be analyzed, since it includes a user-individual scheduling-exclusion which allows the scheduler to assign resources to users with less favorable channel conditions. In this case, the maximization is not performed on the *FOM* exclusively, but also the availability of the user for resource allocation is considered.

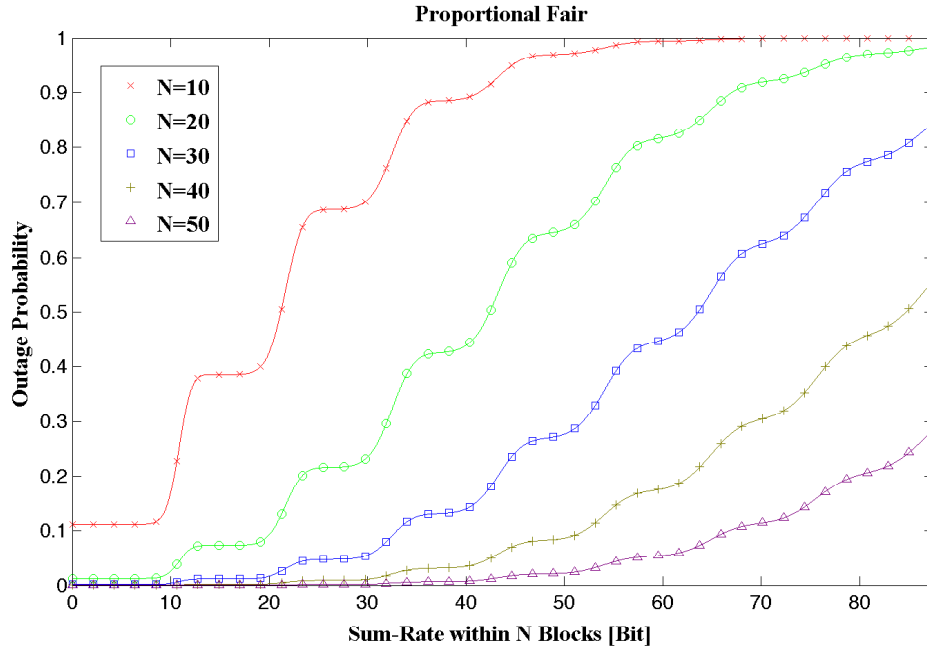


Figure 5.4.: Outage probability of user 1 for 5-user Proportional-Fair scheduling in a Rayleigh-fading environment with a transmit-SNR of 10dB and window-lengths of 10, 20, 30, 40, and 50 blocks. Numerical results based on  $10^7$  simulated blocks and an EWMA weighting-factor of  $10^5$  (solid lines) are compared to outage probabilities computed analytically (individually marked points).

## 6. Conclusion of Part I

The first part of this thesis has explored time-constrained rate-requests in multiuser downlink scenarios. Although this topic has been explored repeatedly in the past, there is a lack of approaches that are practical from a computational complexity perspective, and flexible enough to adapt to a wide variety of channel conditions, rate constraints and delay limitations. The first contribution of this thesis is therefore the discussion of properties of widespread resource allocation strategies with respect to their suitability for delay-constrained transmission of data. Current implementations of scheduling algorithms generally lack information about rate demands and delay requirements of higher layers in the OSI model. As long as the channel quality does not change too quickly, applications typically adapt their demands to the current possibilities the communication system provides. Video streaming is an especially good example for this strategy: if the offered rate and its reliability decreases, the video quality will be reduced and the buffering will increase. This is, however, also a strategy to compensate for the inability of the scheduler to serve all requests in time - there is simply no commonly implemented information exchange between the application layer and the scheduler. In practice, this lack of information at the scheduler is often - from a user-perspective - substituted by the TCP “slow start” and “congestion avoidance” mechanisms. These tend to keep the scheduling buffer rather empty if the requested data is being provided by many different servers at the same time (as is the case with many websites), freeing up resources for more demanding (e.g., streaming) applications sharing the same scheduler resources. With increasing performance of wireless and backbone technology however, this “natural” TCP prioritization of demanding services loses its effect. The presented considerations have shown that in powerful systems that are not limited by data delivery, resources can potentially be assigned much more efficiently, especially for users with high rate demands. This thesis therefore suggests a concept - Power-Controlled Cross-Layer Scheduling (PCCLS) - that exploits information from the application layer, opening up a range of possibilities that are not available in today’s standard-implementations. Compared to proportional fair scheduling, PCCLS adds a control loop that compensates for slow changes in channel attenuation, thus providing a constant long-term channel quality that makes the achievement of specific rate- and delay- requirements possible. The faster block-fading is not compensated for, i.e. it can be exploited. In addition, PCCLS excludes certain users from scheduling as soon as the required rate has been achieved in a given delay-window. This

## 6. Conclusion of Part I

is done even when the transmit buffer still provides data to be transmitted, freeing up resources for other users. Through numerical simulations, PCCLS is shown to achieve the given rate- and delay-constraints for a general scenario based on realistic requirements. Proportional fair scheduling, on the other hand, is demonstrated to spend too much energy on users with low rate requests, which leads to a failure regarding the requested rates of the more demanding users.

The second contribution in this first part of the thesis derives the probability of an outage for resource allocation schemes that are based on a memoryless selection of the user with the largest scheduling metric. Well-known members of this species of scheduling schemes are the proportional fair- and the opportunistic scheduler. The analysis does not constrain the number of users, their respective channel (fading) statistics or rate demands. The general derivation is subsequently specialized to the classic (i.e., unweighted) implementations of proportional fair and opportunistic scheduling; simulations demonstrate the perfect match between analytic and numeric approaches. Further discussion of the results reveals that the outage probability of the individual users can be derived from the statistics of the individual user's scheduling metrics, under the constraint that they are scheduled. This leads to some very elegant properties of the probability that a certain rate cannot be supported in a specific delay window. Finally, it is shown that accepting more relaxed delay requirements leads to a huge reduction of the experienced outage probability.



Part II.

## Resource Allocation for Finite Lifetime Channels



## 7. Limits on Information Transmission for the Finite Lifetime Channel

The second part of this thesis focuses on a very different yet closely related topic compared to the first part: While discussion has so far evolved around the possibility to provide a minimum rate while satisfying delay requirements, the second part is now going to explore the possibility of exploiting a channel of finite lifetime in an optimum way - regardless of the application requests.

### 7.1. Problem Statement

One main property of the radio channel between any two communication partners moving relative to each other is its non-existence that applies most of the time.

The concept of finite lifetime of the radio channel implies a number of boundary-conditions that are setting the stage for this second part of the thesis. First and foremost, the channel will not allow for any considerable data-transmission at all, almost all of the time. Only when the communication partners (this shall also include a scenario with only one moving transceiver and one stationary access point, as in cellular communication along a highway) approach each other, there will be a short opportunity to transmit data.

The communication partners will, during this finite time-window, experience adverse conditions for data exchange due to two inevitable phenomena: multipath propagation due to a scattering environment and Doppler shifts due to relative movement of the communication partners. This temporal and frequency dispersion are causing selectivity of the radio conditions both in frequency and time, leading to channel fading. The assumption of a finite lifetime channel moreover implies a time-variant long-term path loss that is slow compared to the fading but causes the random process describing the channel coefficients to be non-ergodic. This prevents time-continuous transmission, causing the classic concept of a channel capacity that is defined by the average transmitted information over a very long time-span to lead to zero.

Literature provides more general definitions than the classic Shannon capacity to evaluate the capability of a channel to transmit information, though:

## 7. Limits on Information Transmission for the Finite Lifetime Channel

For instance, Dobrushin [44] has proved that for information stable channels, i.e. channels whose input that maximizes mutual information and its corresponding output behave ergodically, the capacity is given as

$$C = \lim_{n \rightarrow \infty} \sup \frac{1}{n} I(X^n; Y^n), \quad (7.1)$$

where  $n$  is the encoded block length. In 1994, Verdú and Han [45] derived a formula for channel capacity which does not require any assumption such as memorylessness, information stability, stationarity, causality, etc:

$$C = \sup_{\mathbf{X}} \underline{I}(\mathbf{X}; \mathbf{Y}), \quad (7.2)$$

where  $\underline{I}(\mathbf{X}; \mathbf{Y})$  is the limes inferior in probability of the normalized information density between input sequence  $\mathbf{X}$  and output sequence  $\mathbf{Y}$ .

These expressions, however, are not applicable to a channel of finite life-time, since a vector of infinite length needs to be encoded (see the limit  $n \rightarrow \infty$  in 7.1 above)<sup>1</sup>.

An analysis of information transmission in time-constrained scenarios requires a clear concept of the physical boundary-conditions setting the stage. The common assumption about the time-variability of the channel widely adopted in literature is the applicability of a block fading model [46]. Already assumed in the first part of this thesis, the reason to apply the same fundamental model again is mathematical tractability. The assumption of the classical channel capacity being a close upper bound for the information that can be transmitted when the channel can be considered coherent does constrain generality: It is not applicable to a finite lifetime of the channel in vehicular communications as well as in ad-hoc connections between body-area networks. Here, the channel capacity will certainly be a close upper bound if the communication partners remain stationary. When they move and cause the channel to die within a relatively short amount of time, the assumption of the channel capacity to be a *close* upper bound is not justified in general. There is, however, an exception to this obvious truth: If the coherence time of the channel in a certain time-window can be assumed to be long enough for block-fading to be applicable, the upper bound of information transmission derived from the classical channel capacity will indeed be acceptable for vehicular communications and body-area networks. These scenarios can be distinguished mainly by the speed of the moving communication partner (at least one of them has to be moving relative to the other one in order to cause a time-constrained communication scenario). In case of vehicular communications, the speed of the vehicular-based transceiver together with

---

<sup>1</sup>Strictly speaking, all practically relevant information channels have a finite lifetime: this may be due to the fundamental problem that the message should be used at the receiving side at some point and there is no use for an efficient scheme that causes not only infinite transmit time but also decoding time. It might also be other, less mundane problems like finite battery lifetime or communication partners who will stay in each others range for only a short time - as discussed in this thesis.

the inevitable presence of other cars and reflecting roadside structures guarantees the existence of multiple propagation paths (and therefore delays) for the transmitted signal. The combination of scattering and movement will, at a first glance, lead to extremely fast fading. In an extensive measurement campaign [47–52], it was shown that inter-vehicular communication along a highway can be characterized by a dominant line-of-sight propagation path, when the vehicles are close to each other. While other cars and metallic buildings at the roadside contribute some significant multipath components, they get relatively weak compared to the line-of-sight component the closer the communication partners approach each other. Since it will be shown that efficient vehicular communications necessarily exploits the best channel coefficients that the finite lifetime channel provides (these occur typically when the line-of-sight component is dominating the channel), it is therefore assumed that a block fading model is applicable to this scenario.

A pedestrian scenario on the other hand, such as the ad-hoc connection between close body-area networks, is characterized by slow movement. Although rich scattering - and therefore fast fading - should be assumed in urban surroundings at a first glance, it is exactly the slow movement and expected high attenuation of the signal around the human body that introduce a directivity to the wave-propagation that promotes, again, a dominant line-of-sight component through absorbing body tissue and devices that are worn closely to the body. This intuitively justifies the assumption of block fading.

Although the following discussion is restricted to single-input single-output (SISO) channels, the results will also apply for the multiple-input multiple-output (MIMO) case when the degrees of freedom are used to increase the reliability of system (multiplexing gain is “1”). In this case orthogonal space-time block codes can be used that will convert the MIMO channel into a SISO Gaussian channel with an equivalent channel gain derived from the MIMO channel matrix [53]. Of course, the use of MIMO will also reduce the experienced fading, further justifying the approach of block fading.

As mentioned above, the finite-lifetime channel is characterized by a limited number of channel states that can be used for transmission, i.e.  $|h(i)|^2 > 0$  only for a limited subset

$$\mathcal{S} \doteq \{i : |h(i)|^2 > 0\} \text{ with } |\mathcal{S}| < \infty . \quad (7.3)$$

This definition does not necessarily mean the set  $\mathcal{S}$  is compact (i.e., without gaps), so the “lifetime” may also consist of several periods, with a “broken connection” in between.

In each block  $i$  the receive power  $P_r(i)$  will equal the transmit power  $P(i)$  scaled by the magnitude square  $|h(i)|^2$  of the channel coefficient  $h(i)$ , which will subsequently be called the channel-power coefficient, i.e.,

$$P_r(i) = |h(i)|^2 \cdot P(i) . \quad (7.4)$$

## 7. Limits on Information Transmission for the Finite Lifetime Channel

“Channel-power coefficients”  $|h(i)|^2$  are used rather than “channel coefficients”  $h(i)$  because coherent detection is assumed, i.e. perfect channel state information (CSI) is available at both the transmitter and the receiver. This is in order to properly focus on the development of an efficient scheme for data transmission. In real-world systems, the channel state is usually available with a delay of a few fading blocks due to a feedback or channel analysis that takes some time.

The capacity  $C(i)$  (in bits per channel-use) within block  $i$  is given by

$$C(i) = \log_2 \left( 1 + \frac{|h(i)|^2 \cdot P(i)}{2\sigma^2} \right) \quad (7.5)$$

where  $\sigma^2$  is the variance of the Gaussian receiver noise in each real component of the transmit channel (inphase and quadrature component); by definition of (7.5) a carrier-modulated system is implicitly assumed, i.e., complex modulation. The total power  $P(i)$  is spread equally across the two real channel parts. For symmetric channels, this is necessary (but not sufficient) to achieve capacity (e.g., [54]).

Under the assumption that each block  $i$  contains  $M$  uses of the channel (with  $M$  very large), at most

$$N(i) = C(i) \cdot M \quad (7.6)$$

bits of information can be transmitted in block  $i$ .

Hence, the total number  $N$  of bits of information that can be transmitted through the channel equals

$$N = M \sum_{i=-\infty}^{\infty} C(i) = M \sum_{i=-\infty}^{\infty} \log_2 \left( 1 + \frac{|h(i)|^2 \cdot P(i)}{2\sigma^2} \right). \quad (7.7)$$

Here,  $N$  will be bounded as  $|h(i)|^2 > 0$  holds only for a limited number  $|\mathcal{S}| < \infty$  of blocks  $i$ . Hence, the following upper bound holds:

$$N = M \sum_{i \in \mathcal{S}} \log_2 \left( 1 + \frac{|h(i)|^2 \cdot P(i)}{2\sigma^2} \right). \quad (7.8)$$

Although more general definitions for the channel capacity have been stated in the literature (e.g., [13, 45, 55–57]) those results do not apply in a straightforward way to the problem at hand. In the following, this problem is circumvented by first considering the absolute amount of information that can be transmitted during the limited lifetime of the channel, and a suitable “rate normalization” is introduced in a second step to achieve a result that can be expressed in “bits per channel use”.

## 8. The Waterfilling Solution

The following considerations aim to maximize the number  $N$  of bits in (7.8), given  $|h(i)|^2$  for all  $i \in \mathcal{S}$ . Since the transmitted information  $N$  grows without bounds for growing transmit power, there is need to introduce a constraint. This approach is common in information theory, where an “average-power” constraint is frequently used. This is, however, not applicable in case of a channel with finite lifetime due to the non-ergodicity of the channel. Restricting the average transmit power would still allow for a very large amount of energy to be used for a short period (the channel lifetime) and zero everywhere else. Restricting the average power transmitted in any number of consecutive blocks (as long as this number is smaller than infinity) does not make sense either, since it leads to a sub-optimum transmission if the channel lifetime is not contained in exactly one interval over which the average power constraint should hold. This consideration directly leads to the only constraint that is applicable to the finite-lifetime channel: an energy constraint. The total energy used for the transmission is therefore introduced as

$$E = \sum_{\forall i \in \mathcal{S}} MP(i) \cdot T_s = M \cdot T_s \sum_{\forall i \in \mathcal{S}} P(i) . \quad (8.1)$$

In order to simplify notation, and exploiting the discrete-time property of the system, the “time-scaling” factor  $T_s$  is omitted in the following derivations. Reconsideration of (8.1) therefore leads to

$$E = \sum_{\forall i \in \mathcal{S}} MP(i) = M \sum_{\forall i \in \mathcal{S}} P(i) . \quad (8.2)$$

Limiting the value of the total energy according to

$$0 \leq E = M \sum_{\forall i \in \mathcal{S}} P(i) \leq E_0 , \quad (8.3)$$

a constrained optimization problem can be derived. Convex optimization technique allows for setting up the functional  $L$  with the Lagrange multiplier  $\lambda > 0$  [58]:

$$L \doteq \frac{M}{\log(2)} \sum_{\forall i \in \mathcal{S}} \log \left( 1 + \frac{|h(i)|^2 \cdot P(i)}{2\sigma^2} \right) + \lambda (E_0 - M \sum_{\forall i' \in \mathcal{S}} P(i')) . \quad (8.4)$$

## 8. The Waterfilling Solution

Calculating the derivative<sup>1</sup> for the unknown transmit power  $P_j$  in block  $j \in \mathcal{S}$  finds an expression for the slope of the convex problem:

$$\frac{\partial L}{\partial P(j)} = \frac{M}{\log(2)} \frac{1}{1 + \frac{|h(j)|^2 P(j)}{2\sigma^2}} \cdot \frac{|h(j)|^2}{2\sigma^2} - \lambda \cdot M. \quad (8.5)$$

Setting the derivative to zero, cancelling  $M \neq 0$ , setting  $\tilde{\lambda} \doteq \lambda \log(2)$  obtains

$$P^*(j) = \left( \frac{1}{\tilde{\lambda}} - \frac{2\sigma^2}{|h(j)|^2} \right)^+ \quad (8.6)$$

which is a “waterfilling” solution (e.g. [28]). In (8.6) a “max”-operation according to  $(x)^+ \doteq \max(0, x)$  ensures that the solutions for the powers do not take negative values; the Karush-Kuhn-Tucker [27] conditions guarantee that (8.6) is still an optimal solution to the problem.

The value of  $\tilde{\lambda}$  must be chosen such that (8.3) is fulfilled. The maximum number of information bits  $N^*$  is given by substituting (8.6) into (7.7) with  $P(i) = P^*(i)$ .

For a measured sequence  $\{|h(i)|^2\}$  of channel-power coefficients that is known in advance  $\tilde{\lambda}$  can be found numerically by trying a value for it, solving (8.3) and checking if  $E$  is close to the pre-specified value of  $E_0$ . If  $E > E_0$ , then  $\tilde{\lambda}$  has to be increased to get closer to  $E_0$ : this process can be iteratively repeated to get the optimal solution for  $\tilde{\lambda}$  for this particular sequence of channel coefficients.

### 8.1. Normalisation of the Solution

The constrained search for  $\tilde{\lambda}$  in the previous section is impractical since it has to be carried out iteratively to find the optimum overall transmit energy  $E$ . The problem with this kind of search is usually the necessity to decrease the adjustments to  $\tilde{\lambda}$  the closer the iterations approach the optimum. Therefore, this *step-size* depends on the specific realization of  $\{|h(i)|^2\}$  and requires knowledge about the properties of the function input. This section therefore develops an alternative approach to solving this optimization problem that is not dependent on the specific sequence of channel-power coefficients.

The problem and its solution can be rewritten equivalently by dividing (8.3) by  $M$  and  $2\sigma^2$ , obtaining

$$\frac{E/M}{2\sigma^2} = \sum_{\forall i \in \mathcal{S}} \frac{P(i)}{2\sigma^2} \leq \frac{E_0/M}{2\sigma^2}. \quad (8.7)$$

The fraction  $\frac{P(i)}{2\sigma^2} \doteq \text{SNR}(i)$  can be interpreted as a “transmit” signal-to-noise ratio (as it does not contain the channel fading factor  $h(i)$ ) for block  $i$ . The ratio  $E/M$  on the left-hand side the

---

<sup>1</sup>To avoid confusion with the sum index  $i$ , we change the index to  $j$  for the power for which we take the derivative.



represents the average transmit power for each use of an equivalent substitute channel that would transmit the same amount of information but use only one block (also of size  $M$ ) instead of  $|\mathcal{S}|$  blocks of size  $M$  that are used in the channel. Hence, an alternative form of (8.7) is obtained according to

$$\frac{E/M}{2\sigma^2} \doteq SNR = \sum_{\forall i \in \mathcal{S}} SNR(i) \leq \frac{E_0/M}{2\sigma^2}. \quad (8.8)$$

The solution for the optimal transmit  $SNR^*(i)$  then follows directly from (8.6) by dividing through the non-negative (total) noise variance  $2\sigma^2$ . The solution reads<sup>2</sup>

$$SNR^*(i) = \left( \frac{1}{\lambda} - \frac{1}{|h(i)|^2} \right)^+, \quad \lambda > 0 \quad (8.9)$$

with the upper limit for the largest possible number  $N^*$  of information bits given by

$$N^* = M \sum_{\forall i \in \mathcal{S}} \log_2 (1 + |h(i)|^2 SNR^*(i)), \quad (8.10)$$

and  $\lambda$  in (8.9) has to be chosen (typically by a numerical approach) such that with  $SNR(i) = SNR^*(i)$  (8.8) is fulfilled.

Instead of considering the absolute number  $N^*$  of information bits it is more convenient to introduce the relative quantity

$$C^* \doteq \frac{N^*}{M} = \sum_{\forall i \in \mathcal{S}} \log_2 (1 + |h(i)|^2 SNR^*(i)) \quad (8.11)$$

that can be interpreted as a “capacity” measured in *information bits per use of the equivalent substitute channel* that transmits a single block of size  $M$  only. Note that as long as the lifetime of the channel is limited (i.e.,  $|\mathcal{S}| < \infty$ ), the sum in (8.11) will always converge and it is strictly non-zero if there is at least one  $j$  for which  $|h(j)|^2 > 0$ . This assumption is necessary for a meaningful analysis of information transmission in a finite-lifetime scenario.

From (8.11) the *receive* signal-to-noise ratio  $SNR_r^*$  of the one-block substitute channel can also be expressed: from introducing

$$C^* = \log_2 (1 + SNR_r^*) = \sum_{\forall i \in \mathcal{S}} \log_2 (1 + |h(i)|^2 SNR^*(i)), \quad (8.12)$$

the following expression for the receive SNR is obtained:

---

<sup>2</sup>Although the Lagrange multiplier  $\lambda$  in (8.9) is different from that in (8.4), the same variable-name is used as there is no risk of confusion.

## 8. The Waterfilling Solution

$$SNR_r^* = \prod_{\forall i \in \mathcal{S}} (1 + |h(i)|^2 SNR^*(i)) - 1; \quad (8.13)$$

when (8.9) is substituted for  $SNR^*(i)$ , the receive SNR becomes

$$SNR_r^* = \prod_{\forall i \in \mathcal{S}^*} \frac{|h(i)|^2}{\lambda} - 1 \quad (8.14)$$

with

$$\mathcal{S}^* \doteq \{i : |h(i)|^2 > \lambda\}. \quad (8.15)$$

The effective receive signal-to-noise ratio  $SNR_r^*$  is maximized by the waterfilling power allocation in (8.9), now implicitly contained in (8.14).

## 9. Fixed-Point computation of $\lambda$

The previously gathered insights are now developed into an efficient algorithm that finds the optimum  $\lambda$ , considering the sum-Energy constraint (8.3). Using (8.9) in (8.8) leads to

$$\sum_{\forall i \in \mathcal{S}} \left( \frac{1}{\lambda} - \frac{1}{|h(i)|^2} \right)^+ \leq \frac{E_0/M}{2\sigma^2}. \quad (9.1)$$

With the definition (8.15) this is equivalent to

$$\frac{E_0/M}{2\sigma^2} \stackrel{!}{=} \sum_{\forall i \in \mathcal{S}^*} \left( \frac{1}{\lambda} - \frac{1}{|h(i)|^2} \right) = \frac{|\mathcal{S}^*|}{\lambda} - \sum_{\forall i \in \mathcal{S}^*} \frac{1}{|h(i)|^2}. \quad (9.2)$$

Please note that here, implicitly, equality ( $\stackrel{!}{=}$ ) for the energy constraint was assumed. (9.2) can be solved for  $\lambda$ , resulting in

$$\lambda = \frac{|\mathcal{S}^*|}{\frac{E_0/M}{2\sigma^2} + \sum_{\forall i \in \mathcal{S}^*} \frac{1}{|h(i)|^2}}. \quad (9.3)$$

Here, (9.3) is only an implicit characterisation of  $\lambda$  as the right-hand side also depends on  $\lambda$  (via (8.15)). The result for  $\lambda$  does, however, *not* depend on the *order* with which the summation in the denominator of (9.3) is calculated. Hence, in order to compute  $\lambda$ , the set  $\mathcal{H} \doteq \{h(i), i \in \mathcal{S}\}$  of non-zero channel coefficients is sorted in descending order such that an ordered set  $\tilde{\mathcal{H}} \doteq \{\tilde{h}(j), j = 1, 2, \dots, |\mathcal{S}|\}$  is obtained, with  $\tilde{h}(j) \geq \tilde{h}(j')$  if  $j < j'$ .

According to (8.15), the set

$$\tilde{\mathcal{S}}^* \doteq \{j : |\tilde{h}(j)|^2 > \lambda\}. \quad (9.4)$$

of all index-numbers  $j$  (of the sorted channel coefficients) for which the inequality  $|\tilde{h}(j)|^2 > \lambda$  is fulfilled is introduced. The *largest* of those indices is denoted by  $J(\lambda)$ , i.e.,

$$J(\lambda) = \max_{j: |\tilde{h}(j)|^2 > \lambda} j. \quad (9.5)$$

## 9. Fixed-Point computation of $\lambda$

With this definition of  $J(\lambda)$ , (9.3) is rewritten according to

$$\lambda = \frac{J(\lambda)}{\frac{E_0/M}{2\sigma^2} + \sum_{j=1}^{J(\lambda)} \frac{1}{|\tilde{h}(j)|^2}} . \quad (9.6)$$

For what follows, the right side of (9.6) will be called the fixed-point function  $\varphi(\lambda)$ :

$$\varphi(\lambda) = \frac{J(\lambda)}{\frac{E_0/M}{2\sigma^2} + \sum_{j=1}^{J(\lambda)} \frac{1}{|\tilde{h}(j)|^2}} . \quad (9.7)$$

Obviously, (9.6) calls for the use of a fixed-point iteration to find the optimal value for  $\lambda$ <sup>1</sup>. In this section, the convergence of (9.6) is shown for very general assumptions concerning the statistical properties of the channel-power coefficients  $|h(i)|^2$ .

A sensible starting point for the fixed-point iteration can be found as follows: it is safe to assume that  $J(\lambda)$  will never be zero, as, even for the worst channel, the transmit energy budget will be used for at least one (the largest) channel coefficient even if almost no information can be transmitted (the case that no transmission is possible in any transmit block does obviously not need to be considered, since then the channel can not be considered a finite-lifetime channel). Moreover,  $J(\lambda)$  as given by (9.5) will lead to the same result for some *interval* of  $\lambda$ -values, i.e., once the optimal value for  $J(\lambda)$  is known, the optimal exact value for  $\lambda$  can be computed from (9.6), and this value will reproduce  $J(\lambda)$  by use of (9.5). Based on those considerations the following procedure to compute  $\lambda$  is proposed:

1. **Start-value for  $J(\lambda)$ .** From the above considerations it is known that  $J(\lambda) \in \{1, 2, \dots, |\mathcal{S}|\}$ . Simulations have shown that of the  $|\mathcal{S}|$  non-zero channel coefficients in most cases just a few are really used for data transmission. This follows from the fact that, for a channel model where two communication partners approach, pass, and leave each other, most of the time the relative speed is constant. With a channel exponent of approximately 2.5 for inter-vehicular communication on a highway or more for pedestrian scenarios, this leads to an at least quadratic behavior of the channel quality and thus a rather small portion of the measurable channel coefficients being used for actual transmission. Therefore the start-value is set according to  $J(\lambda) = 1$ . This implicitly also defines the start-value for  $\lambda$  as it follows from the initial choice of  $J(\lambda)$ .

---

<sup>1</sup>A non-iterative algorithm for finding the optimum transmit powers for all channel-coefficients is compared in Section 10.3.2

2. With the current value of  $J(\lambda)$ ,  $\lambda$  is computed by (9.6).
3. Then,  $\lambda$  as obtained in Step 2) is used in (9.5) to compute a new value for  $J(\lambda)$ .
4. If the new value for  $J(\lambda)$  and the old one are the same or within a predefined range, the iterations terminate and the most recent value of  $\lambda$  is the solution. Otherwise, the algorithm proceeds with Step 2).

The convergence of this fixed-point function remains to be shown for very general assumptions concerning the statistical properties of the channel-power coefficients  $|h(i)|^2$ . This proof can be split into the proof of two statements, the combination of which will prove the convergence of the fixed-point iteration. In order to facilitate the readability of this proof, some more technical parts can be found in the Appendix A. Here, just the results of these considerations will be used. The significance of the proof stems from the fact that some remarkable properties of the fixed-point function are found that will lead to more efficient approach in the actual implementation.

1. In a first step, the structure of the fixed-point function  $\varphi(\lambda)$  is explored by considering a large number of observed channel-power coefficients  $|h|^2$ . This leads to a representation equivalent to  $\varphi(\lambda)$  employing the cdf of the channel-power coefficients. This approach finds the *expected* fixed-point function, facilitating, on the one hand, the mathematical analysis while maintaining the principal properties that can still be observed for a small set of observed channel-power coefficients  $|h|^2$ . Due to the random nature of  $|h|^2$ , the actual fixed-point  $\lambda$  will differ for any observed ensemble of channel-power coefficients.
2. The fixed-point function  $\varphi(\lambda)$  is extended to a continuously differentiable generalized function  $\tilde{\varphi}(\lambda)$  with the same properties as  $\varphi(\lambda)$  (most notably the same fixed-point), only adding differentiability.
3. The behavior of  $\tilde{\varphi}(\lambda)$  is explored at  $\lambda = 0$  and  $\lambda = 1$ , showing that  $\tilde{\varphi}(0) > 0$ ,  $\frac{d}{d\lambda}\tilde{\varphi}(0) \geq 0$ ,  $\tilde{\varphi}(1) = 0$ , and  $\frac{d}{d\lambda}\tilde{\varphi}(1) = 0$ . It is concluded that  $\tilde{\varphi}(\lambda)$  has *at least one* fixed-point.
4. The first derivative of  $\tilde{\varphi}(\lambda)$  is calculated, and it is shown that every  $\lambda_i$ , for which  $\frac{d}{d\lambda}\tilde{\varphi}(\lambda_i) = 0$  is a fixed-point.
5. It is shown that every fixed-point  $\lambda_i$  lies in a local maximum of  $\tilde{\varphi}(\lambda)$ .
6. Based on the previous findings, it is shown that the converse of 4) holds: every fixed-point  $\lambda_i$  has the property  $\frac{d}{d\lambda}\tilde{\varphi}(\lambda_i) = 0$ .
7. It is concluded that, based on 6), the function  $\tilde{\varphi}(\lambda)$  has *at most one* fixed-point. Together with 3), it is therefore true that  $\tilde{\varphi}(\lambda)$  has exactly one fixed-point, which is also the maximum of  $\tilde{\varphi}(\lambda)$ .

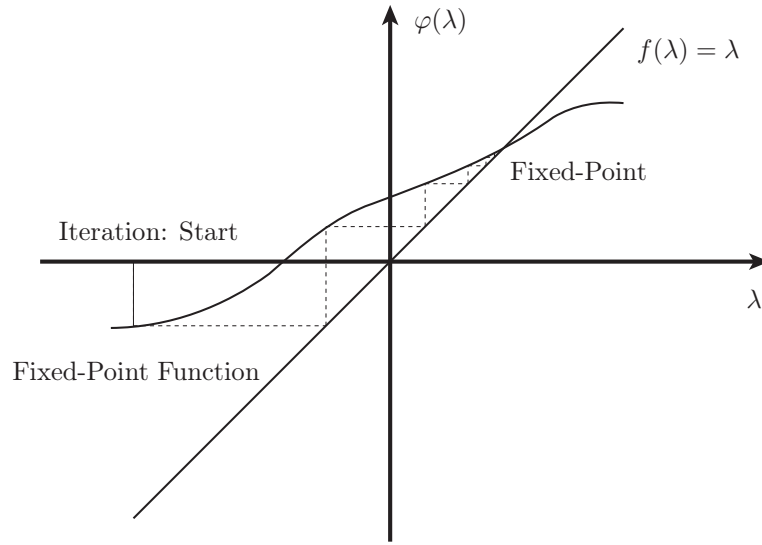


Figure 9.1.: Repeated application of a function with an attractive fixed-point on itself will converge efficiently.

## 9.1. Structure of the Fixed-Point Function

Literature knows several theorems about the existence of fixed-points under various conditions. In this section, the mere proof of the existence of a unique solution, however, is not the main scope of the analysis, since it does not provide deep insights into the structure of the problem at hand. Instead, the focus lies on a brief characterization of a fixed-point function in general, and, more importantly, the analysis of the fixed-point function  $f(\lambda) = \lambda$  based on a probabilistic approach. In general, a fixed-point is any intersection of a function with the line  $f(x) = x$ . In this point, the function therefore returns the function argument itself. Obviously, not every function has a fixed-point, while those that do may have more than one. The practical relevance of the fixed-point lies in the existence of *attractive fixed-points*. These are obtained by repeated application of a function on itself (which will subsequently be referred to as fixed-point iterations):

$$x_0 = f(f(f(\dots(x)))) \quad (9.8)$$

If the iteration was started sufficiently near an attractive fixed-point, this will converge efficiently. Considering the problem at hand, the existence of a fixed-point is easily inferred from the nature of the waterfilling solution derived in (9.7), which is the solution to a constrained convex optimization problem with a single optimum point.

Application of these insights to the fixed-point function (9.7):

$$\varphi(\lambda) = \frac{J(\lambda)}{\frac{E_0/M}{2\sigma^2} + \sum_{j=1}^{J(\lambda)} \frac{1}{|\tilde{h}(j)|^2}} \quad (9.9)$$

with

$$J(\lambda) = \max_{j: |\tilde{h}(j)|^2 > \lambda} j, \quad (9.10)$$

that returns the number of channel-power coefficients which are bigger than  $\lambda$ , it can be readily seen that  $\varphi(\lambda)$  is necessarily a piecewise-constant function. Therefore,  $\varphi(\lambda)$  is not differentiable. The fixed-point function that will have to be solved in a realistic scenario is constructed from an ordering of the random channel-power coefficients. Its shape is therefore random itself. In order to analyze the general properties of  $\varphi(\lambda)$ , the expected function is considered, that can be constructed either by ordering a *very large* number of channel-power coefficients, or by taking the average over many realizations. This approach is justified for nearly all channels, except for degenerated realizations (all coefficients are zero, and all are exactly equal). The analysis finds an equivalent expression to the empirical fixed-point function  $\varphi(\lambda)$  which depends on the *cdf* of the channel-power coefficients, and a variant of  $\varphi(\lambda)$  is introduced which is differentiable. This differentiable function will be called  $\tilde{\varphi}(\lambda)$ .

In the following, despite the fact that the wireless channel is inherently non-ergodic in a finite lifetime scenario, the channel-power coefficients will be characterized by cumulative density functions (cdfs) and probability density functions (pdfs) that are not describing i.i.d. random processes. This is not a contradiction, however: measuring the finite lifetime channel several times with the same physical boundary conditions will allow for the construction of a statistic of the whole measured channel when the time-dependency is deliberately discarded.

### 9.1.1. Derivation of $|\tilde{h}(j)|^2$

Sorting all channel-power coefficients, that can potentially be used for a transmission of more than 0 bits according to their absolute values from large to small, generates the function  $|\tilde{h}(j)|^2$ . Therefore, while  $j$  increases from 1 to  $N$  (i.e. the number of channel coefficients being used for transmission),  $|\tilde{h}(j)|^2$  decreases from (at most) 1 to a value near, but larger than, 0.

Given that the statistics of the channel power (during its whole lifetime) are assumed to be known,  $|\tilde{h}(j)|^2$  can be expressed in the following way:

Let  $F_{|\tilde{H}|^2}(|\tilde{h}|^2)$  denote the cumulative distribution function of the channel-power coefficients. Let's find an expression for the  $j^{th}$  largest channel-power coefficient. Analytically, for a very large

### 9. Fixed-Point computation of $\lambda$

amount of channel-power coefficients, this amounts to determining the expected value of the  $j^{th}$  largest channel-power coefficient.

Suppose the  $j^{th}$  largest channel-power coefficient is asked, out of a large number of  $N$  measured coefficients. The empirical probability that a channel-power coefficient which is drawn randomly from the set of all measured coefficients is larger than  $\lambda$ , is

$$P\{|h|^2 > |\tilde{h}(j)|^2\} = \frac{j}{N}. \quad (9.11)$$

Reformulating this expression so that well-known statistical items can be used, this is equivalent to:

$$1 - P\{|h|^2 \leq |\tilde{h}(j)|^2\} = \frac{j}{N}, \quad (9.12)$$

and therefore:

$$P\{|h|^2 \leq |\tilde{h}(j)|^2\} = 1 - \frac{j}{N}. \quad (9.13)$$

This can be reformulated to be the cdf of the channel-power coefficients  $F_{|H|^2}(|h|^2)$ :

$$F_{|H|^2}(|\tilde{h}(j)|^2) = 1 - \frac{j}{N}, \quad (9.14)$$

and using the inverse function of the cdf,  $|\tilde{h}(j)|^2$  can finally be expressed analytically as:

$$|\tilde{h}(j)|^2 = F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right). \quad (9.15)$$

The inverse function of the cdf,  $F_{|H|^2}^{-1}(|h|^2)$  exists if no two channel-power coefficients are exactly the same, which is herewith assumed.

#### 9.1.2. Derivation of $J(\lambda)$

The function  $J(\lambda)$ , which forms the numerator on the right side of the fixed-point equation  $\varphi(\lambda) = \lambda$  (9.6) returns the number of channel-power coefficients, measured during the whole lifetime of the finite lifetime channel, which are bigger than  $\lambda$ . Obviously, since this return-value can only be a natural number from the interval  $\{0...N\}$ , but the argument of the function,  $\lambda$ , can take any value from the interval  $\{0...1\}$  (assuming that the channel can only attenuate the signal or let it pass unattenuated, to include an upper bound), the function  $J(\lambda)$  is necessarily a step-function.

In the following, and as in the previous subsection, an analytical expression for  $J(\lambda)$  shall be derived. Again, for a large number of samples  $N$  taken from the whole lifetime of the channel, the empirical probability that a randomly-drawn channel-power coefficient is larger than  $\lambda$ , can be expressed similarly as before:



$$P\{|h|^2 > \lambda\} = \frac{J(\lambda)}{N} \quad (9.16)$$

$$1 - P\{|h|^2 \leq \lambda\} = \frac{J(\lambda)}{N} \quad (9.17)$$

and therefore:

$$1 - F_{|H|^2}(\lambda) = \frac{J(\lambda)}{N}. \quad (9.18)$$

Taking into account that  $J(\lambda)$  can only yield a natural number, a *floor*-operation has to be included, which leads to the final analytical expression for  $J(\lambda)$ :

$$J(\lambda) = \lfloor N \cdot (1 - F_{|H|^2}(\lambda)) \rfloor. \quad (9.19)$$

### 9.1.3. The Fixed-Point Function for *large* $N$

The fixed-point equation (9.6) has two major drawbacks. First, the numerator-function  $J(\lambda)$  is not a continuous nor a differentiable function, and second the upper bound of the sum in the denominator depends on the numerator. In order to facilitate working with this function in a mathematically meaningful way, according to the above arguments, it is supposed that a *large* number  $N$  of channel-power coefficients  $|h(j)|^2$  has been observed (see previous section). Then, the fixed-point equation can be rewritten as:

$$\lambda = \frac{\lfloor N \cdot (1 - F_{|H|^2}(\lambda)) \rfloor}{\frac{E_0/M}{2\sigma^2} + \sum_0^{\lfloor N \cdot (1 - F_{|H|^2}(\lambda)) \rfloor} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)}}. \quad (9.20)$$

For the sake of mathematical tractability, let's now assume that the function  $J(\lambda)$  is not restricted to the field of natural, but real numbers. Then, the function  $\tilde{\varphi}(\lambda)$  *smoothes* the original  $\varphi(\lambda)$  (they coincide in all points for which  $N \cdot (1 - F_{|H|^2}(\lambda)) \in \mathbb{N}$ : here, the *floor*-function, applied to the set of *real* numbers coincides with the set of *natural* numbers), and makes analytical analysis feasible.

$$\tilde{\varphi}(\lambda) = \frac{N \cdot (1 - F_{|H|^2}(\lambda))}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1 - F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj}. \quad (9.21)$$

## 9. Fixed-Point computation of $\lambda$

### 9.1.4. Value of the Fixed-Point Function for extreme $\lambda$

Considering  $\tilde{\varphi}(\lambda)$  at the two extreme values of  $\lambda$ , i.e.  $\lambda = 0$  and  $\lambda = 1$ , allows for a characterization at the “boundaries”. In order to do that, a few basic properties of the pdf  $f_{|H|^2}(|h|^2)$  and the cdf  $F_{|H|^2}(|h|^2)$  need to be stated:

#### Properties of $f_{|H|^2}(|h|^2)$ and $F_{|H|^2}(|h|^2)$

For  $|h|^2 = 0$ ,  $f_{|H|^2}(0) = 0$ . This is due to the fact that in a Finite Lifetime channel scenario, only channel coefficients are taken into account which amount to possible transmission of information, and therefore the pdf  $f_{|H|^2}(|h|^2)$  equals zero for  $|h|^2 = 0$ . In other words, the pdf of the finite-lifetime channel is defined such that the probability of the channel-power coefficient  $|h|^2$  to be zero is zero. For the cdf  $F_{|H|^2}(|h|^2)$ , this means that it can only be equal to zero at  $|h|^2 = 0$ .

For the first derivative of the pdf,  $f'_{|H|^2}(|h|^2)$ , evaluated at  $|h|^2 = 0$ , this means that it can only be bigger or equal to zero, since the pdf is strictly positive or equal to zero. Therefore, the following holds:

$$\begin{aligned} F_{|H|^2}(0) &= 0 \\ f_{|H|^2}(0) &= 0 \\ f'_{|H|^2}(0) &\geq 0. \end{aligned} \tag{9.22}$$

For  $|h|^2 = 1$ ,  $f_{|H|^2}(1) = 0$ , since it was assumed that *no* channel will conduct signals *absolutely unattenuated*. Together with the basic property of the cdf to equal 1 for  $|h|^2 = 1$ , this amounts to:

$$\begin{aligned} F_{|H|^2}(1) &= 1 \\ f_{|H|^2}(1) &= 0. \end{aligned} \tag{9.23}$$

For the fixed-point function, these properties lead to the following function values at the boundaries of the interval over which the function is defined:

#### Analysis of $\varphi(\lambda)$ at $\lambda = 0$

For  $\lambda = 0$ ,  $\tilde{\varphi}(\lambda)$  can be simplified to be (considering especially that  $F_{|H|^2}(0) = 0$ ):

$$\frac{N \cdot (1 - 0)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-0)} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} > 0 \tag{9.24}$$

Note that this is *strictly bigger* than 0.

The first derivative  $\frac{\partial \tilde{\varphi}}{\partial \lambda}$  can be evaluated, at  $|h|^2 = 0$ , to be (taking into account that  $f_{|H|^2}(0) = 0$  and  $F_{|H|^2}(0) = 0$ ):

$$\left. \frac{\partial \tilde{\varphi}}{\partial \lambda} \right|_{|h|^2=0} = - \frac{N \cdot f_{|H|^2}(0)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(0))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.25)$$

$$+ \frac{N \cdot (1 - F_{|H|^2}(0)) \cdot \frac{N}{0} f_{|H|^2}(0)}{\left[ \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(0))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj \right]^2} \quad (9.26)$$

And substitution of the previously derived (9.23) numerical values leads to:

$$\left. \frac{\partial \tilde{\varphi}}{\partial \lambda} \right|_{|h|^2=0} = - \frac{0}{\frac{E_0/M}{2\sigma^2} + \int_0^N \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.27)$$

$$+ \frac{N \cdot \frac{N}{0} \cdot 0}{\left[ \frac{E_0/M}{2\sigma^2} + \int_0^N \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj \right]^2}$$

This expression, however, contains an undetermined form, which can be resolved using L'Hôpital's rule:

$$\left. \frac{N}{\lambda} f_{|H|^2}(\lambda) \right|_{\lambda=0} = \left. \frac{N}{1} f'_{|H|^2}(\lambda) \right|_{\lambda=0} \quad (9.28)$$

$$= N \cdot " \geq 0 " \quad (9.29)$$

$$\geq 0. \quad (9.30)$$

And, therefore, the first derivative of  $\tilde{\varphi}(\lambda)$ , evaluated at  $\lambda = 0$ , is shown to be bigger or equal to zero:

$$\left. \frac{\partial \tilde{\varphi}}{\partial \lambda} \right|_{\lambda=0} = \frac{\geq 0}{\left[ \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-0)} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj \right]^2} \quad (9.31)$$

$$\geq 0. \quad (9.32)$$

### 9. Fixed-Point computation of $\lambda$

To summarize, at  $\lambda = 0$ ,  $\tilde{\varphi}(\lambda)$  is bigger than zero, and its first derivative is bigger or equal to zero.

#### Analysis of $\varphi(\lambda)$ at $\lambda = 1$

For  $\lambda = 1$ , and keeping in mind that  $F_{|H|^2}(1) = 1$ ,  $\tilde{\varphi}(\lambda)$  can easily be evaluated to be:

$$\frac{N \cdot (1 - 1)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-1)} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} = 0. \quad (9.33)$$

And the first derivative can as easily be simplified to yield (keep in mind that  $f_{|H|^2}(1) = 0$  and  $F_{|H|^2}(1) = 1$ ):

$$- \frac{N \cdot f_{|H|^2}(1)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(1))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.34)$$

$$+ \frac{N \cdot (1 - F_{|H|^2}(1)) \cdot \frac{N}{1} f_{|H|^2}(1)}{\left[ \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(1))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj \right]^2} \quad (9.35)$$

$$= 0. \quad (9.36)$$

### Illustration of the Findings

Figure 9.2 illustrates that according to the considerations in the Sections 9.1.4 and 9.1.4, the following properties of  $\tilde{\varphi}(\lambda)$  hold in general:

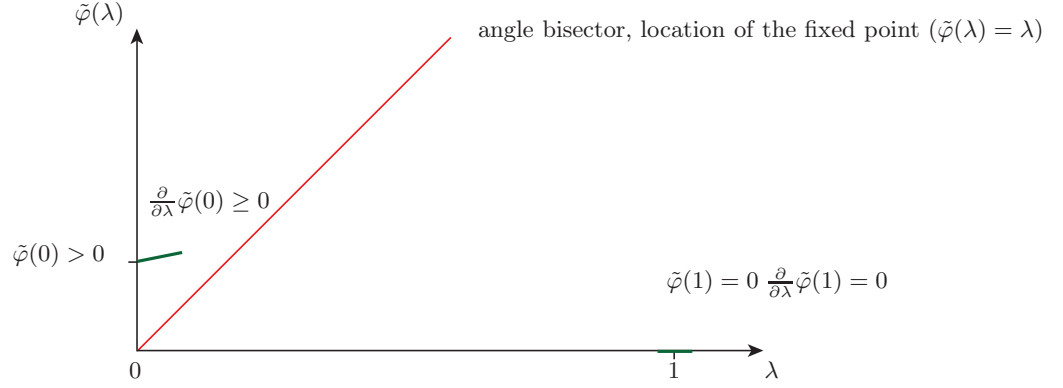


Figure 9.2.: The first part of the analysis focuses on the properties of  $\tilde{\varphi}(\lambda)$  for extreme values of  $\lambda$ .

#### 9.1.5. Relation between Fixed-Points and the First Derivative of $\tilde{\varphi}(\lambda)$

In this section, it is shown that each point of the fixed-point function, except for  $\lambda = 0$  and  $\lambda = 1$ , in which  $\frac{d}{d\lambda} \tilde{\varphi}(\lambda) = 0$ , is also a fixed-point.

In order to find the first derivative of  $\tilde{\varphi}(\lambda)$  (9.21) with respect to  $\lambda$ , the quotient rule needs to be applied (details concerning the derivation of  $\tilde{\varphi}(\lambda)$  are contained in Appendix A:

$$\left(\frac{u}{v}\right)' = \frac{vu' - uv'}{v^2}, \text{ where } v \neq 0. \quad (9.37)$$

$$\frac{u'}{v} - \frac{uv'}{v^2} \doteq - \frac{N \cdot f_{|H|^2}(\lambda)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.38)$$

$$+ \frac{N \cdot (1 - F_{|H|^2}(\lambda)) \cdot \frac{N}{\lambda} f_{|H|^2}(\lambda)}{\left[ \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj \right]^2} \quad (9.39)$$

## 9. Fixed-Point computation of $\lambda$

Setting this zero, it is obtained accordingly:

$$\frac{N \cdot f_{|H|^2}(\lambda)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.40)$$

$$= \frac{N \cdot (1 - F_{|H|^2}(\lambda)) \cdot \frac{N}{\lambda} f_{|H|^2}(\lambda)}{\left[ \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj \right]^2} \quad (9.41)$$

and multiplying with the denominator of the left side leaves:

$$N \cdot f_{|H|^2}(\lambda) = \frac{N \cdot (1 - F_{|H|^2}(\lambda)) \cdot \frac{N}{\lambda} f_{|H|^2}(\lambda)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.42)$$

furthermore, canceling the left side of the equation (which is bigger than zero by definition of the lifetime of the channel!) leaves

$$1 = \frac{\frac{N}{\lambda} \cdot (1 - F_{|H|^2}(\lambda))}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.43)$$

which is exactly equal to (9.21), which is the original function:

$$\lambda = \frac{N \cdot (1 - F_{|H|^2}(\lambda))}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.44)$$

That shows now implicitly, that *each* point of  $\tilde{\varphi}(\lambda)$  in which the first derivative equals zero is also a fixed-point  $\lambda_i$  ( $\frac{d}{d\lambda} \tilde{\varphi}(\lambda_i) = 0$ ).

### Illustration of the Findings

Figure 9.3 illustrates that according to the considerations in Section 9.1.5, the following properties of  $\tilde{\varphi}(\lambda)$  hold in every fixed-point:

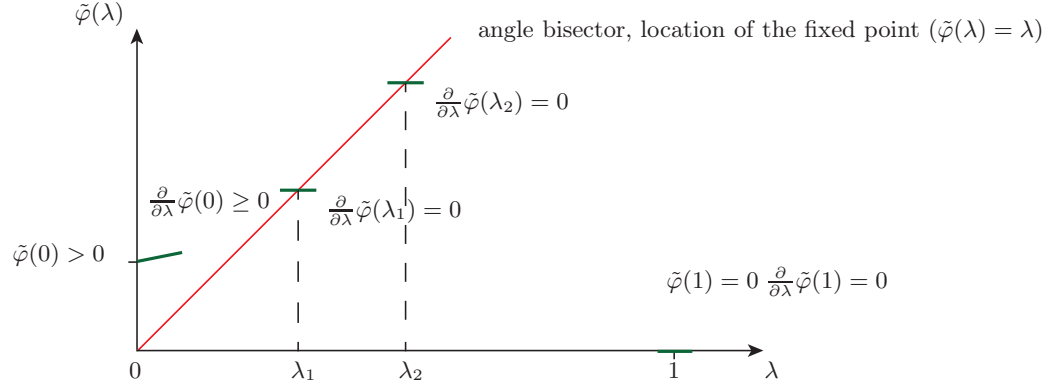


Figure 9.3.: The second part of the analysis shows that each point  $\lambda_i$  in which  $\frac{\partial}{\partial \lambda} \tilde{\varphi}(\lambda_i) = 0$  is also a fixed-point.

#### 9.1.6. The Fixed-Point Function $\tilde{\varphi}(\lambda)$ in a Neighborhood of the Fixed-Point $\lambda_i$

Now, it remains to be shown that every fixed-point  $\lambda_i$  is indeed a local maximum of the fixed-point function  $\tilde{\varphi}(\lambda)$ . In order to do so, it is necessary to show the following.

Suppose a fraction  $\frac{a}{b}$ . Now, add a value, let's say  $c$ , to the numerator, and another value,  $d$ , to the denominator. All the variables  $a, b, c, d \in \mathbb{R} > 0$ . For this fraction  $\frac{a+c}{b+d}$ , the following statement holds (for the proof, see Appendix A):

$$\frac{a+c}{b+d} < \frac{a}{b} \text{ if and only if } \frac{c}{d} < \frac{a}{b} \quad (9.45)$$

For subtracting terms in both the numerator and the denominator of the original fraction  $\frac{a}{b}$ , as in  $\frac{a-c}{b-d}$ , this rule cannot be applied. Instead, the following holds (for the proof, see Appendix A):

$$\frac{a-c}{b-d} < \frac{a}{b} \text{ if and only if } \frac{c}{d} > \frac{a}{b} \quad (9.46)$$

Using these relations, it shall now be shown that each fixed-point  $\lambda_i$  is indeed a (local) maximum of  $\tilde{\varphi}(\lambda)$ .

Considering  $\tilde{\varphi}(\lambda)$  (9.21), it shall be shown that replacing  $\lambda$  by  $\lambda + \varepsilon$  (for  $\varepsilon > 0$  small) yields a smaller value.

## 9. Fixed-Point computation of $\lambda$

Therefore, the following substitutions are made:

$$N \cdot (1 - F_{|H|^2}(\lambda)) \quad (9.47)$$

$$\rightarrow N \cdot (1 - F_{|H|^2}(\lambda + \varepsilon)) \quad (9.48)$$

$$\xrightarrow{\text{for small } \varepsilon} N \cdot (1 - F_{|H|^2}(\lambda) + \varepsilon \cdot f_{|H|^2}(\lambda)) \quad (9.49)$$

And  $\tilde{\varphi}(\lambda)$  can be written as:

$$\frac{N \cdot (1 - F_{|H|^2}(\lambda) + \varepsilon \cdot f_{|H|^2}(\lambda))}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1 - F_{|H|^2}(\lambda) + \varepsilon \cdot f_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.50)$$

Since the substitution  $\lambda \rightarrow \lambda + \varepsilon$  amounts to an additional summation-term both in the numerator as well as (due to the integral) the denominator of  $\tilde{\varphi}(\lambda)$ , the above relation (9.45) is applicable. Therefore, if

$$\frac{N \cdot \varepsilon \cdot f_{|H|^2}(\lambda)}{\int_{N \cdot (1 - F_{|H|^2}(\lambda))}^{N \cdot (1 - F_{|H|^2}(\lambda) + \varepsilon \cdot f_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.51)$$

$$< \frac{N \cdot (1 - F_{|H|^2}(\lambda))}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1 - F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj}, \quad (9.52)$$

then the function decreases for values of  $\lambda > \lambda_i$ . By noting that  $\tilde{\varphi}(\lambda)$  is considered at an actual fixed-point, the above relation can be simplified to (a fixed-point is always an intersection of  $\tilde{\varphi}(\lambda)$  with the function  $f(\lambda) = \lambda$ ):

$$\frac{N \cdot \varepsilon \cdot f_{|H|^2}(\lambda)}{\int_{N \cdot (1 - F_{|H|^2}(\lambda))}^{N \cdot (1 - F_{|H|^2}(\lambda) + \varepsilon \cdot f_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} < \lambda_i. \quad (9.53)$$

This relation is now to be shown. In other words, the left side of (9.53) needs to be upper-bounded. Before (9.53) can be shown to hold, first the monotonicity of the function in the integral,  $\frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)}$ , needs to be explored. Under the assumption that the pdf of the channel-power coefficients not to have zeros or to be non-differentiable, the cdf  $F_{|H|^2}(|h|^2)$  is of course strictly monotonically increasing. Therefore, the inverse  $F_{|H|^2}^{-1}(|h|^2)$  must be strictly monotonically increasing as well.



For the function  $\frac{1}{F_{|H|^2}^{-1}(|h|^2)}$ , strictly monotonuous decreasing follows suit, and therefore,

$$\frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} \quad (9.54)$$

is, due to the sign in the argument, strictly monotonically increasing. An upper bound for the left side of (9.53) therefore amounts to finding a lower bound (in this case the lower bound cannot be reached, due to strict monotonicity) for the denominator.

A lower bound for the integral is indeed easy to find, and the fraction (9.53) can be upper bounded as (remember the strictly monotonuous increasing property of the function (9.54)):

$$\frac{N \cdot \varepsilon \cdot f_{|H|^2}(\lambda)}{\int_{N \cdot (1 - F_{|H|^2}(\lambda))}^{N \cdot (1 - F_{|H|^2}(\lambda) + \varepsilon \cdot f_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.55)$$

$$< \frac{N \cdot \varepsilon \cdot f_{|H|^2}(\lambda)}{\frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{N \cdot (1 - F_{|H|^2}(\lambda))}{N}\right)} \cdot N \cdot \varepsilon \cdot f_{|H|^2}(\lambda)} \quad (9.56)$$

$$= F_{|H|^2}^{-1}(F_{|H|^2}(\lambda)) \quad (9.57)$$

$$= \lambda, \quad (9.58)$$

which proves the decreasing of  $\tilde{\varphi}(\lambda)$  towards larger  $\lambda$ , i.e.  $\lambda > \lambda_i$ .

On the other hand, towards smaller  $\lambda$ , a similar argument can be used, however with a different condition on the decrease of the fixed-point function (see condition (9.46)).

First,  $\lambda$  is substituted by  $\lambda - \varepsilon$ :

$$\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1 - F_{|H|^2}(\lambda - \varepsilon))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj \quad (9.59)$$

$$= \frac{N \cdot (1 - F(\lambda)) - N\varepsilon \cdot f(\lambda)}{N \cdot (1 - F(\lambda))} \cdot \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1 - F(\lambda))} \frac{1}{F^{-1}\left(1 - \frac{j}{N}\right)} dj - \int_{N \cdot (1 - F(\lambda - \varepsilon))}^{N \cdot (1 - F(\lambda))} \frac{1}{F^{-1}\left(1 - \frac{j}{N}\right)} dj \quad (9.60)$$

(Notice the difference in the lower bound of the integral compared to the previous case! Due to limited space, the subscripts in (9.60) have been omitted).

### 9. Fixed-Point computation of $\lambda$

And, as above, the fraction made up of the additional summation-terms in the nominator and numerator can be (loosely, i.e. the bound can not be reached) lower bounded as:

$$\frac{-N\varepsilon \cdot f_{|H|^2}(\lambda))}{-\int_{N \cdot (1-F_{|H|^2}(\lambda-\varepsilon))}^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} \quad (9.61)$$

$$> \frac{N \cdot \varepsilon \cdot f_{|H|^2}(\lambda)}{\frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{N \cdot (1-F_{|H|^2}(\lambda))}{N}\right)} \cdot N \cdot \varepsilon \cdot f_{|H|^2}(\lambda))} \quad (9.62)$$

$$= \lambda. \quad (9.63)$$

This shows that the left neighborhood of every fixed-point  $\lambda_i$ , as well, lies below the value of  $\tilde{\varphi}(\lambda_i)$ . This shows that every fixed-point is in a local maximum.

### Illustration of the Findings

Figure 9.4 illustrates that according to the considerations in Section 9.1.6, every fixed-point of  $\tilde{\varphi}(\lambda)$  lies in a local maximum:

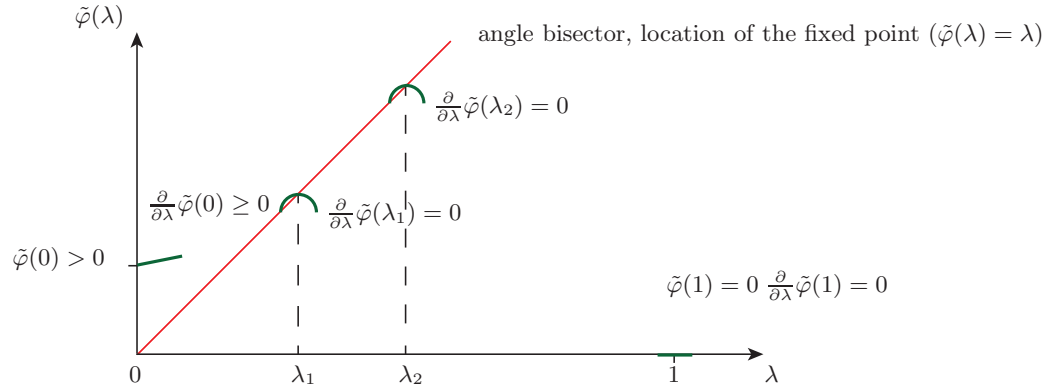


Figure 9.4.: The third part of the analysis shows that each fixed-point  $\lambda_i$  is also a local maximum.

### 9.1.7. The Converse of Section 9.1.5

In Section 9.1.5 it has been shown that each point of the fixed-point function, except for  $\lambda = 0$  and  $\lambda = 1$ , in which the first derivative  $\frac{d}{d\lambda}\tilde{\varphi}(\lambda) = 0$ , is also a fixed-point.

Section 9.1.6 contains proof that every fixed-point  $\lambda_i$  is indeed a local maximum of the fixed-point function  $\tilde{\varphi}(\lambda)$ .

These two insights, in turn, can be used to show the converse of Section 9.1.5. Since every fixed-point is a local maximum, and every point of  $\tilde{\varphi}(\lambda)$  for which  $\frac{d}{d\lambda}\tilde{\varphi}(\lambda) = 0$  is a fixed-point, This means that the first derivative of the fixed-point function is zero for every fixed-point.

## 9.2. Proposition and Conclusion

In Chapter 9.1.7, it has been shown that  $\tilde{\varphi}(\lambda)$  satisfies  $\frac{d}{d\lambda}\tilde{\varphi}(\lambda_i) = 0$  for every fixed-point  $\lambda_i$ . In Section 9.1.6, it has been shown that every fixed-point  $\lambda_i$  lies in a local maximum. The conclusion from these two findings is that there can be at most *one* fixed-point  $\lambda_0$ , since between any two local maxima there must lie a local minimum. Since the first derivative is zero in a local minimum, this must also be a fixed-point. However, since every fixed-point must also lie in a local maximum, there can only be *one* fixed-point *at most*.

Section 9.1.4, it was shown that the fixed-point function  $\tilde{\varphi}(\lambda)$  is greater than zero for  $\lambda = 0$ , and that it is equal to zero for  $\lambda = 1$ , which means that it must intersect the linear function  $f(\lambda) = \lambda$  at some point, since it is positive or equal to zero in the interval  $\lambda = [0 \dots 1]$ .

Since every intersection between the functions  $\tilde{\varphi}(\lambda)$  and  $f(\lambda) = \lambda$  is a fixed-point, there must be *at least one* fixed-point.

This insight leads to the corollary that the function  $\tilde{\varphi}(\lambda)$  has *exactly one* fixed-point, which is also the maximum of the function  $\tilde{\varphi}(\lambda)$ . This in turn means that the fixed-point is the unique solution to the problem at hand. While this follows from the approach of formulating and solving a convex optimization problem, the presented analysis also provides some deep insight into the well-behaved structure of the fixed-point function itself.

### Illustration of the Conclusion

Figure 9.5 illustrates that since every fixed-point of  $\tilde{\varphi}(\lambda)$  lies in a local maximum and every point of  $\tilde{\varphi}(\lambda)$  with a first derivative being zero ( $\frac{d}{d\lambda}\tilde{\varphi}(\lambda) = 0$ ) is a fixed-point, there can only be one fixed-point. Furthermore, due to the “boundary-conditions” for  $\tilde{\varphi}(\lambda)$  for  $\lambda = 0$  and  $\lambda = 1$ , there must be at least one fixed-point - hence, there is exactly one fixed-point, which represents the unique optimum solution to the problem:

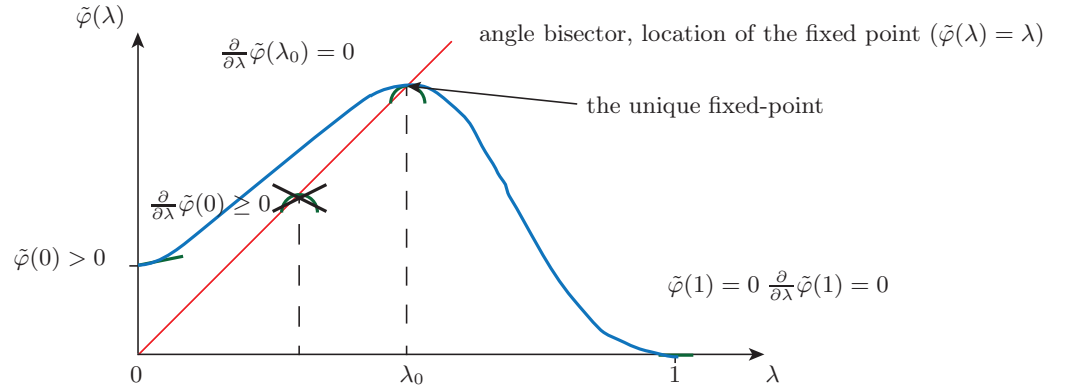


Figure 9.5.: The conclusion of the analysis shows that there must be exactly one fixed-point  $\lambda_i$ .

# 10. Numerical Simulations

## 10.1. Channel- and System-Model

For a numerical comparison of the power allocation schemes, the channel model depicted in Fig. 10.1 is used. The zero-mean Gaussian receiver noise is assumed to have a variance of  $\sigma^2$  in each real channel component. Due to coherent detection, there are no phase rotations at the receiving end and only an amplitude scaling that is equal in both the inphase and the quadrature component of the channel occurs.

The channel coefficients in the model are composed of a long-term path loss  $a_i$  and a factor  $b_i$  that is to model block fading, due to inter symbol interference and multipath propagation. Both factors  $a_i, b_i$  are assumed to be fixed during each block  $i$ , with the path-loss  $a_i$  changing between blocks but much more slowly than the fading factor  $b_i$ , which is assumed with a mean power of, i.e.  $E\{b_i^2(i)\} = 1$ . Therefore, the block fading does not change the power of the received signal:  $E\{X'^2(i)\} = E\{X''^2(i)\}$  (see Figure 10.1).

The above analysis is exact in an asymptotic way: If either the number of measured channel coefficients is very large or the non-ergodic channel can be averaged over many realizations, then the analysis matches the measured results perfectly, as shown below. A channel of finite lifetime is, however, characterized by a relatively small number of actually usable channel realizations. Therefore, it remains to be shown that the insights gathered in the above sections can still be observed in realistic scenarios.

The channel model is derived from a Vehicle to Vehicle (V2V) centric measurement campaign by

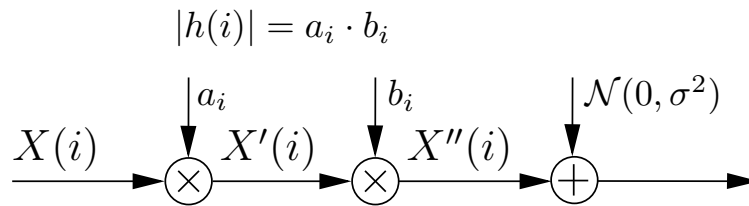


Figure 10.1.: Channel model used for simulations; depicted is one of two independent, statistically identical parallel transmit channels (“I” or “Q”-path).

## 10. Numerical Simulations

Paier et al. (e.g. [47, 48, 59]). The measurement results were obtained by this group in a series of drive tests with two vehicles on opposed highway-lanes in Lund, Sweden.

The measurement setup can be best illustrated with the following information, taken from [59]:

Table 10.1.: Measurement Setup according to [59]

Center Frequency	5.2GHz
Measurement Bandwidth	240MHz
Delay Resolution	4.17ns
Transmit Power	27dBm
Number of Tx Antennas	4
Number of Rx Antennas	4
Snapshot Time	102.4 $\mu$ s
Snapshot Time Interval	307.2 $\mu$ s
Number of Snapshots	32,500
Traffic Intensity	approx. 1 Vehicle per Second
Individual Speed of the two Vehicles	90km/h
Relative Speed of the two Vehicles	approx. 180km/h
Tx Antenna Height	2.4m
Rx Antenna Height	2.4m

From the channel measurements, the authors concluded that the channel is stationary during the time it took the two cars to approach each other by twenty wavelengths. At the given snapshot repetition rate, this corresponded to 75 snapshots or a stationarity time of 23ms.

For this thesis' numerical simulations, a channel model was developed to mimic the measurements by Paier et al. Considering a "worst case scenario", the following transceiver parameters were chosen:

Table 10.2.: Simulation Assumptions

System Bandwidth	1MHz
Channel Stationarity Time	10ms
Ambient Temperature	300K
Channel Uses per Block	10 <sup>4</sup>

From these assumptions, the following derived parameters can be calculated:

- From the Channel Stationarity of 10ms, the block duration of equally 10ms was assumed.
- Together with the system bandwidth of 1MHz, this leads to 10,000 channel uses per block under the simplifying assumption that the transmit energy is totally contained within the rectangular transmit spectrum.
- The lifetime of the channel is assumed to be 10s, i.e. 5s before and after the closest approach between transmitter and receiver. This means that the channel lifetime is 1000 blocks.
- For the calculations, the vehicles are going at 90km/h (i.e. 25m/s) each, resulting in a relative speed between close to 50m/s down to 0m/s during the lifetime of the channel. Doppler shifts were, however, not taken into account in the simulations. This is certainly a problematic approach, given the previous assumption of a perfectly rectangular transmit spectrum that measures exactly the system bandwidth. Moreover, the doppler shift is not constant but time-varying between

$$f_r = \frac{c_0 - v_r}{c_0 + v_t} \cdot f_t , \quad (10.1)$$

where  $f_t$  is the transmitted frequency,  $f_r$  is the received frequency and  $v_t$  and  $v_r$  are the transmitter and receiver speed along a straight line connecting the two. Therefore, assuming that the transmit bandwidth of 1MHz is centered at 5.2GHz (i.e. the transmit spectrum ranges from 5.1995GHz to 5.2005GHz), the usable bandwidth is actually 996,533kHz, if the rectangular spectrum of the transmit signal must not leave its dedicated range. This is, however, a difference that is well below 1% of the simplified version of calculating with 1MHz and is therefore neglected.

- The waterfilling approach is a direct consequence of a constraint regarding the sum energy that is available (and should be spent) for transmission (8.3). Of course, this value is completely arbitrary in that it does not influence the algorithm's efficiency or speed: however it has to be chosen reasonably so that the application of the algorithm on a set of realistic channel coefficients will lead to results that can actually be achieved in reality. Adopting the channel attenuation from [59] to be -70dB when the vehicles are close to each other, and a transmit bandwidth of 1MHz, the noise power and the transmit power need to be defined for an arbitrarily chosen channel throughput of 1Mbit/s.

## 10. Numerical Simulations

**The noise power** follows directly from the above defined ambient temperature of 300K and the bandwidth of 1MHz. Together with

$$2\sigma^2 = 2kTB \quad (10.2)$$

with  $k$  being the Boltzmann constant ( $k \approx 1.38 \cdot 10^{-23} \frac{m^2 kg}{s^2 K}$ ), the received noise power is  $2\sigma^2 = 10^{-14.1} W$ .

**The transmit power** is then found by solving Shannon's channel capacity formula for the transmit power:

$$C = B \cdot \log_2 \left( 1 + \frac{|h|^2 \cdot P}{2\sigma^2} \right) \quad (10.3)$$

$$10^6 = 10^6 \cdot \log_2 \left( 1 + \frac{10^{-7} \cdot P}{10^{-14.1}} \right), \quad (10.4)$$

where it is important to note that the channel power  $|h|^2$  is taken from the measurements by Paier et al. [47, 48, 59]. Together with the arbitrarily chosen value for the target capacity, this leads to a transmit power of

$$P = 10^{-7.1} Watt. \quad (10.5)$$

This is a remarkably low transmit power due to the extremely low target capacity of just 1Mbit/s. This choice is justified by the intent to select parameters resulting in simulation results that are in a practically realistic range.

The energy constraint follows from (8.3):

$$E_0 = M \cdot \sum_i P(i) \quad (10.6)$$

$$= 10^4 \cdot 50 \cdot 10^{-7.1} = 5 \cdot 10^{-2.1}. \quad (10.7)$$

Here, the "discrete time" definition of  $E_0$  from (8.3) has been used, and the number of used channel coefficients has been arbitrarily set to 50.

With these estimates, the spectral efficiency is  $\log_2 \left( 1 + \frac{10^{-7} \cdot P}{10^{-14.1}} \right) = 1$  Bit/s/Hz (which is certainly low), and the total amount of information to be transmitted should be around

$$\text{Number of bits } D = 50 \cdot 10000 \cdot 1 = 500 \cdot 10^3. \quad (10.8)$$



With these assumptions, the geometric channel model needs to be defined. Based upon the measurements by Paier et al. [47, 48, 59], a geometric model describing the relative distance of the two vehicles to each other has to take into account the following parameters:

- Speed  $v$
- Lifetime  $2 \cdot \Delta$
- Min. Side Distance  $\delta$

And therefore, the function describing the distance as a function of time can be derived according to these parameters in their geometric interpretation.

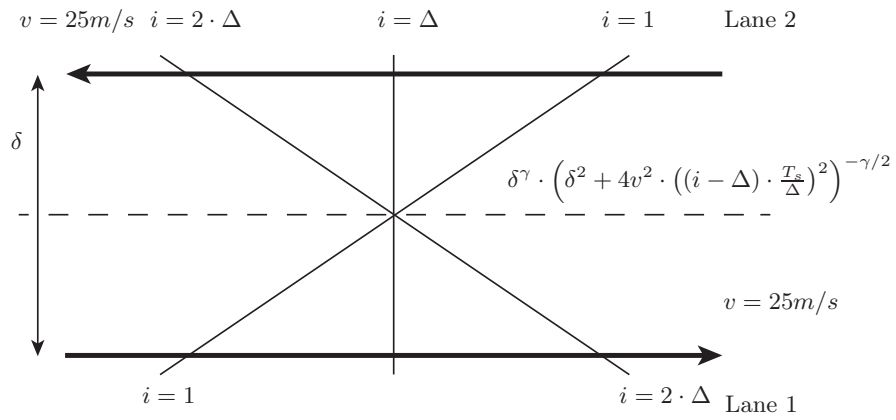


Figure 10.2.: The geometric channel model is based on the relative distance between the two vehicles.

The geometric model for the pathloss  $a_i$  (with the index  $i$  representing the discrete-time index) can therefore be formulated as:

$$a_i = \begin{cases} 0 & , \quad i < 1 \\ \alpha \cdot \delta^\gamma \cdot \left( \delta^2 + 4v^2 \cdot \left( (i - \Delta) \cdot \frac{T_s}{\Delta} \right)^2 \right)^{-\gamma/2} & , \quad 1 \leq i \leq 2\Delta \\ 0 & , \quad i > 2\Delta \end{cases} \quad (10.9)$$

Based on the assumptions that the lanes are parallel and straight with a minimal side-distance of  $\delta$ , the distance can be calculated by applying the pythagorean theorem. The meaning of the other parameters in the model is explained in the following list:

- $(2v(i - \Delta) \cdot T_s/\Delta)^2$ : this describes the evolution of the pythagorean triangle over time, it is therefore the *dynamic part* in the movement model. When  $i = \Delta$  the vehicles are closest.

## 10. Numerical Simulations

When  $i \in [12\Delta]$ , the vehicles are just entering or leaving the range when a transmission is potentially possible. This range is therefore limited by the total channel lifetime  $T_s$ . The discrete-time index  $i$  determines the positions of the vehicles relative to each other at the various time-slots, and  $2\Delta$  is the number of time-slots that are potentially available for transmission. The factor 2 merely describes the fact that both vehicles travel at velocity  $v$ .

- $-\gamma/2$ : the pathloss exponent and the square-root required for the pythagorean expression.
- $\alpha \cdot \delta^\gamma$ : this term normalizes the pathloss to 1 for the nearest side-distance and introduces the desired channel attenuation at this point as  $\alpha$  to match the measured results.

Therefore, for the simulations the geometric parameters

- Minimum side-distance  $\delta$ ,
- Pathloss Exponent  $\gamma$ , and
- Attenuation  $\alpha$

were optimized by minimization of the geometric model's Euclidean distance of (10.9) to the measured sequence by Paier et al. ([59]):

Please observe the close match of the measured channel (solid red) versus the geometric model with optimized parameters (dashed blue). The geometric model also leads to a better match compared to the model originally applied by the authors (printed in black), especially for small attenuation (short distance between the cars).

The geometric model used for the simulations therefore applies the following parameters:

- Minimum side-distance  $\delta = 50m$ ,
- Pathloss Exponent  $\gamma = 2.7$ , and
- Attenuation  $\alpha = 10^{-7}$ .

These parameters are used for all simulations in this thesis. They are not in any way influencing the algorithmic performance of the presented algorithms, but they will ensure numeric results that are plausible under (information-theoretically) optimum circumstances.

The green line in Fig. 10.3 represents the channel fading that is acting "on top" of the long-term path loss. It is obviously much larger than the measured result. The sequence was measured by a 4x4 MIMO system and averaged during the channel stationarity time of 23ms (or 20 wavelengths at the given speed). Furthermore, all 769 delay bins were summed up to effectively combat the small scale fading. This approach, however, is not likely going to be implemented in small and cheap

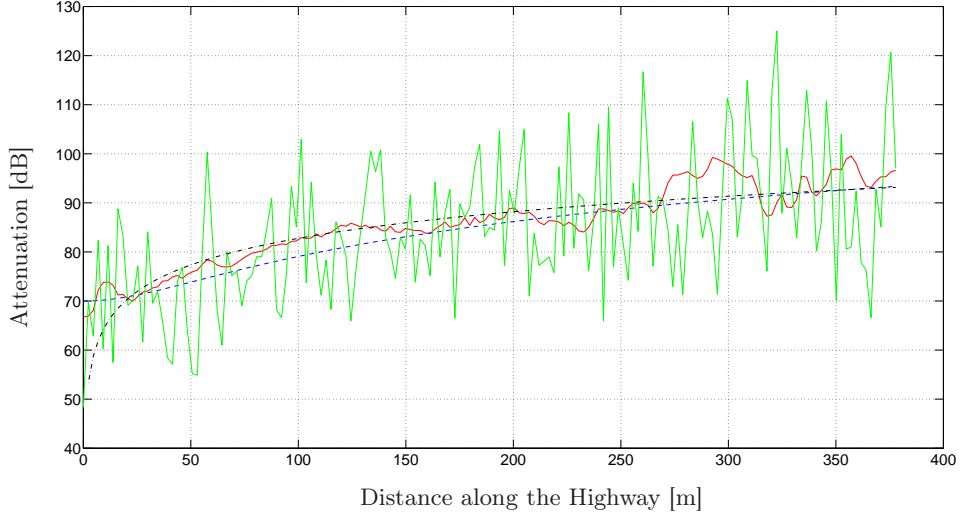


Figure 10.3.: The geometric channel model compared to the measurements by [59]. Solid red: measured channel. Dashed blue: channel according to the geometric model. Dashed black: Channel model originally suggested by [59]. Green: channel model with additional fading used for simulations.

devices that should furthermore provide a long battery life. The channel coefficients are subject to Rayleigh fading with  $E\{\tilde{b}_i^2\} = 1$ , i.e. a mean power of 1 (this is granted by using  $\sigma_b^2 = 1/2$ ). Therefore, the fading does not change the mean power of the channel:

$$f_{\tilde{b}_i}(\tilde{b}_i) = \frac{\tilde{b}_i}{\sigma_b^2} e^{-\tilde{b}_i^2/2\sigma_b^2} \Big|_{\sigma_b^2=1/2} = 2\tilde{b}_i e^{-\tilde{b}_i^2}. \quad (10.10)$$

To bring down the fading a bit and to prevent unrealistically large spikes when the channel power is actually still very low, the actual fading power is assumed to be obtained by averaging over 50 fading-power coefficients:

$$\bar{b}_i^2 = \frac{1}{50} \sum_{b=1}^{50} b^2. \quad (10.11)$$

Since the fading-power coefficients are obtained by averaging over 50 realizations of the fading power, the expected value of the fading power is still 1. However, the probability of unrealistically large channel coefficients a long time before the channel is actually *usable* in the measurements is reduced dramatically, as can be observed from Fig. 10.4.

This distribution is also very similar to a typical Rice-distribution that can be observed when there is a dominant line-of-sight propagation path between transmitter and receiver. This justifies

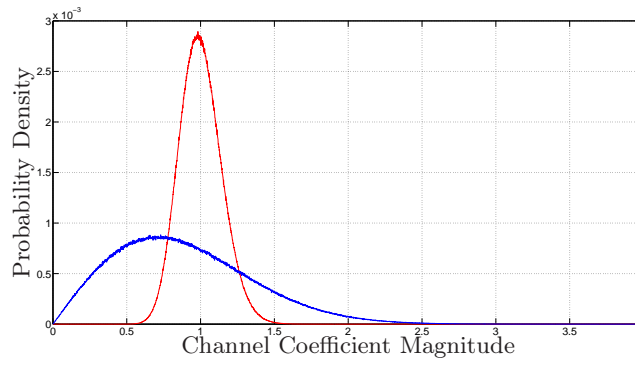


Figure 10.4.: The fading histogram that is obtained by averaging over 50 Rayleigh-Fading Power Coefficients (red) is very similar to the one that can be obtained by applying Rician fading. However, the probability of using unrealistically large channel-power coefficients in the simulations is almost negligible compared to the use of Rayleigh-distributed channel-power coefficients (plotted in blue) with the same mean power of 1.

the approach further, since measurements have shown (e.g. [49]), that there is indeed a dominant line-of-sight propagation path to be expected when the vehicles are at a close distance. Numerical results in Chapter 11 will show that, for a limited amount of transmit energy, most if not all transmission of information typically happens when the communication partners are very close - therefore, the statistical properties of the simulated channel model should ideally reflect the conditions at this point. While all that might call for the use of Rice-distributed fading coefficients, the presented approach was chosen for its simplicity and scalability (the standard deviation can be easily varied by averaging over more or less fading-power coefficients). Finally, the exact choice of the fading distribution is not critical, since it does not affect the performance of the presented algorithms and their comparison.

## 10.2. Convergence Behavior

In Chapter 9 it was shown theoretically that there will always be one optimum solution to the optimization problem at hand. It was proved that this solution can be characterized as the unique maximum of the fixed-point function, assuming that there is a large number of channel coefficients. Based upon the channel characteristics of the numerical simulations in Chapter 11, it shall be demonstrated that this will hold in practice as well. The following Fig. 10.5 depicts the expected and extreme results for the fixed-point function for  $10^6$  different realizations of the channel, being comprised of 1000 channel coefficients.

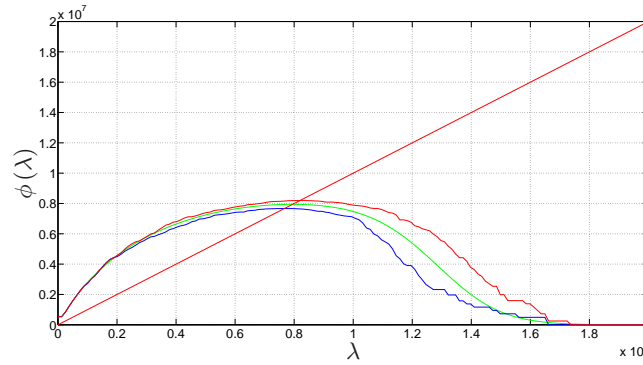


Figure 10.5.: The numerically obtained fixed-point functions reflect the properties derived analytically. The plots are explained in detail below.

As expected, the optimum *waterlevel*  $\lambda$  depends on the specific channel coefficients and is therefore a random value. Since  $\lambda$  is the result of a fixed-point iteration, it is the intersection of the function  $f(x) = x$  (i.e. the angle bisector) with the fixed-point function. For simplicity, there are only three fixed-point functions drawn in the plot:

**Green** The average of all  $10^6$  different realizations of fixed-point functions represents the *asymptotic result* that was derived in (9.20). The smooth characteristic can be achieved in two ways: either by averaging over many realizations of fixed-point functions for relatively few channel coefficients, or for a very long sequence of channel coefficients of the same statistical properties. This case could be experienced in case of a periodically available channel.

## 10. Numerical Simulations

**Blue** The lower curve represents the fixed-point function with the smallest euclidean norm, i.e.

$$Min_{\phi} \left( \sqrt{\sum_i \phi_i^2} \right) \quad (10.12)$$

where  $\phi$  represents the numerically calculated fixed-point function vector containing all the discrete vector elements  $\phi_i$ .

**Red** The upper curve represents, in turn, the fixed-point function with the largest euclidean norm, i.e.

$$Max_{\phi} \left( \sqrt{\sum_i \phi_i^2} \right) \quad (10.13)$$

While the extreme realizations are not as smooth as the averaged representation of the fixed-point function, they certainly display the same properties and most notably the same behavior around their respective maximum, where the fixed-point lies.

The optimum waterfilling parameter  $\lambda$  is a random value depending on the random realization of the sequence of channel coefficients (the other parameters it depends on, like total amount of energy to be spent or the number of channel uses per block is deterministic and does not contribute to the randomness of  $\lambda$ ). The exact value of the optimum  $\lambda$  does, however, not influence the efficiency of the algorithm.

The following considerations shall therefore provide some insight into a practical measure to increase the efficiency of the algorithm by choosing a sensible starting-point for the iteration.

By reconsidering the principal way that fixed-point iterations are finding their optimum, a very simple yet effective rule can be introduced to decrease the number of iterations needed by one: Illustration 10.6 schematically plots two convergence traces that are typical for the type of fixed-point function that occurs for channels of finite life-time.

As can be observed, choosing the starting-point to the left of the eventual optimum will reduce the number of iterations needed for convergence. This is a consequence of the fixed-point function to be (almost) zero for large  $\lambda$ . When the initial value of the iteration leads the fixed-point function to return zero (i.e. there are no channel coefficients larger than the initial value), then the first iteration is *wasted* and the next one is initialized with 0 - that could have been the choice from the very beginning without any loss in this case.

Fig. 10.7 shows the number of iterations for the same set of 1000 channel coefficients, but with 1000 different starting points that are evenly distributed between 0 and 1.1 times the largest channel-power coefficient. The worst-case number of 7 iterations to approach the optimum fixed-point by less than 1% is only required if the initial value for the fixed-point iterations is chosen too high.

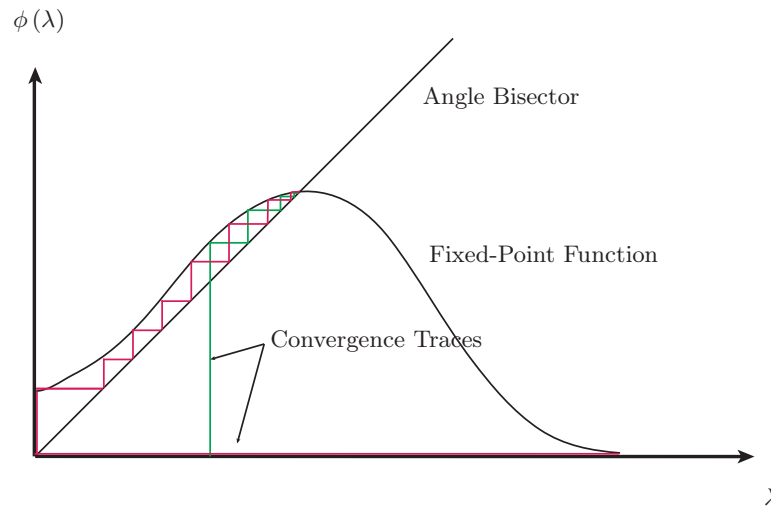


Figure 10.6.: The starting-point of the fixed-point iterations influence the number of iterations to find the optimum.

Fig. 10.7 furthermore suggests that avoiding the an initial value for the fixed-point iterations for which the fixed-point function is zero is not a sufficient condition to avoid the maximum number of iterations that is also dependent on the shape of the fixed-point function. It can, however, safely be stated that avoiding an initialization with large values will exclude the possibility to spend more time calculating than absolutely required.

## 10. Numerical Simulations

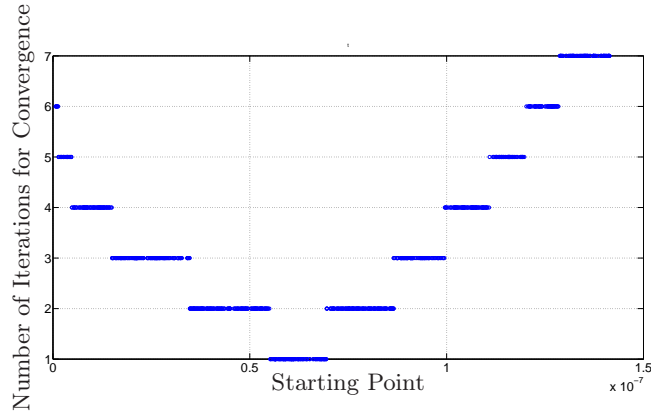


Figure 10.7.: For a large initial value of the fixed-point iterations, one iteration is wasted.

### 10.3. Performance Comparison with the “Classical Approach” (Direct Calculation of the Transmit Powers)

In the following, the fixed-point algorithm is compared to a classical, efficient, non-iterative approach to point out the differences between the two approaches. In order to provide a fair comparison, both algorithms are developed into a pseudo-code with a strong reminiscence of the well-known Matlab-syntax, and runtimes are measured for various numbers of input channel coefficients.

#### 10.3.1. Fixed-Point Calculation of the Waterlevel

The previously developed fixed-point equation to find the waterlevel  $\lambda$  reads

$$\lambda = \frac{J(\lambda)}{\frac{E_0/M}{2\sigma^2} + \sum_{j=1}^{J(\lambda)} \frac{1}{|\hat{h}(j)|^2}} . \quad (10.14)$$

In (10.14), the function  $J(\lambda)$  in the numerator returns the number of channel-power coefficients which are larger than the waterlevel  $\lambda$ :

$$J(\lambda) = \max_{j: |\hat{h}(j)|^2 > \lambda} j . \quad (10.15)$$



### 10.3. Performance Comparison with the “Classical Approach” (Direct Calculation of the Transmit Powers)

The following implementation similar to Matlab-Syntax illustrates the approach (boldface characters denote vectors):

**Initialise:**

```
|h|sorted2 = sort(|h|2, 'descend');  
λ = 0.3; λold = 0.5;  $\frac{E_0/M}{2\sigma^2} = 10$ ;
```

**Fixed-Point Iterations:**

```
while |λ - λold| > 0.00001  
    λold = λ  
    J = length(find((|h|sorted2 - λold)>0));  
    λ = J / ( $\frac{E_0/M}{2\sigma^2} + \text{sum}(1./\text{|h|sorted2(1:J))$ );  
end
```

**Normalize Result:**

```
λ =  $\frac{\lambda}{2\sigma^2}$ 
```

The fraction  $\frac{E_0/M}{2\sigma^2}$  provides the possibility to adjust the algorithm for sum energy and noise-power spectral density. It is a calculation parameter that does not affect algorithm performance. The fixed-point algorithm can easily be implemented recursively (see [7]), however this does not produce any performance advantage in Matlab and might even be problematic for integration in small real-time systems.

#### 10.3.2. Direct Calculation of Optimum Transmit Powers

An alternative, non-iterating (in the sense that the result does not have to be refined after running the algorithm once) algorithm, presented in [60], is used for efficiently and precisely calculating the transmit powers associated to every channel coefficient:

$$I = M \sum_{n=1}^N \log_2 \left( 1 + \frac{|h(n)|^2 \cdot P(n)}{2\sigma^2} \right). \quad (10.16)$$

Please note that here,  $N$  is the number of channel coefficients that will finally be used for transmission according to the algorithm.  $N$  is still an unknown at this time, but the discussion of an implementation pseudo-code will show that  $N$  does not need to be known for the calculation of the transmit powers for the respective channel coefficients. Instead,  $N$  is a by-product of the stopping criterion.

## 10. Numerical Simulations

Limitation of the total energy according to

$$\sum_{n=1}^N P(n) = E_0, \quad (10.17)$$

defines a constrained optimization problem. The corresponding functional  $L$  with the Lagrange multiplier  $\lambda > 0$  is:

$$L \doteq M \sum_{n=1}^N \log_2 \left( 1 + \frac{|h(n)|^2 \cdot P(n)}{2\sigma^2} \right) + \lambda (E_0 - M \sum_{n=1}^N P(n)). \quad (10.18)$$

Therefore, the waterfilling solution determining the transmit powers for the individual channel coefficients reads

$$P^*(n) = \left( \frac{1}{\lambda \log(2)} - \frac{2\sigma^2}{|h(n)|^2} \right)^+. \quad (10.19)$$

which is a “waterfilling” solution (e.g. [28]). Equation (10.19) introduced a “max”-operation according to  $(x)^+ \doteq \max(0, x)$  to ensure that the solutions for the powers do not take negative values; the Karush-Kuhn-Tucker [27] conditions guarantee that (10.19) is still an optimal solution to the problem.

Together with (10.17), the following statement is therefore true:

$$E_0 = \frac{N}{\lambda} - \sum_{n=1}^N \frac{2\sigma^2}{|h(n)|^2}, \quad (10.20)$$

where it is important to note that only those channel-coefficients are taken into account that are contributing a transmission power  $P^*(n)$  in (10.19) that is greater than zero. Reformulating (10.20), an expression for the waterlevel can be found:

$$\lambda = \frac{N}{E_0 + \sum_{n=1}^N \frac{2\sigma^2}{|h(n)|^2}}. \quad (10.21)$$

Therefore, the transmit powers are, according to (10.19),

$$P^*(n) = \frac{E_0 + \sum_{n=1}^N \frac{2\sigma^2}{|h(n)|^2}}{N} - \frac{2\sigma^2}{|h(n)|^2}. \quad (10.22)$$

This algorithm, however, is not practical in this specific form, since, as stated above, there will be some channel coefficients contributing positive, and some will be contributing negative transmit power.

There is a simple solution: In a waterfilling solution, and assuming that at least one channel coefficient is greater than zero, there will always be energy assigned to at least one - the biggest -

### 10.3. Performance Comparison with the “Classical Approach” (Direct Calculation of the Transmit Powers)

channel coefficient. Furthermore, there will be a smallest channel coefficient that will be used for transmission, and all the coefficients that are still smaller will be contributing “negative transmit Power” according to (10.22). Therefore, the channel coefficients (or channel SNRs, if the noise power is not constant) will be sorted from large to small, and the stopping criterion for the algorithm is reached as soon as (10.22) is negative.

The following schematic illustrates the implementation of the algorithm. Again, boldface characters denote vectors.

#### **Initialise:**

```
|h|sorted2 = sort(|h|2, 'descend');  
CSNR = 1./(|h|sorted2./2σ2);  
k = 1;  
SUM = CSNR(k);
```

#### **Determine Stopping Criterion and Calculate SUM**

```
while (E0+SUM+CSNR(k+1))/k - CSNR(k+1) > 0  
    SUM = SUM + CSNR(k+1);  
    k=k+1;  
end
```

#### **Calculate Transmit Powers:**

```
for i = 1:k  
P(i) = (E0+SUM)/k - CSNR(i);  
end
```

### 10.3.3. Comparison of the Algorithms

Comparison of the algorithms shows some fundamental differences. Above all, the fixed-point algorithm is of the iterative kind with a notably fast convergence. In Fig. 10.7 it was shown that the algorithm converges within at most 7 iterations for a set of 1000 typical channel coefficients with an accuracy that does not affect the total number of transmitted bits in the Matlab simulations. On the other hand, the second “benchmark” - algorithm is non-iterative and does therefore not require refinement of the result.

## 10. Numerical Simulations

The following Fig. 10.8 illustrates the runtime differences between the two algorithms: the average runtime in Matlab (as an average of 1000 runs) is plotted over the number of generic (i.e. not necessarily realistic) channel coefficients. In Matlab, the relative performance of the two algorithms changes as the number of channel coefficients increases: for a large number of channel coefficients, the ratio of the runtimes decreases roughly exponentially (Fig. 10.9).

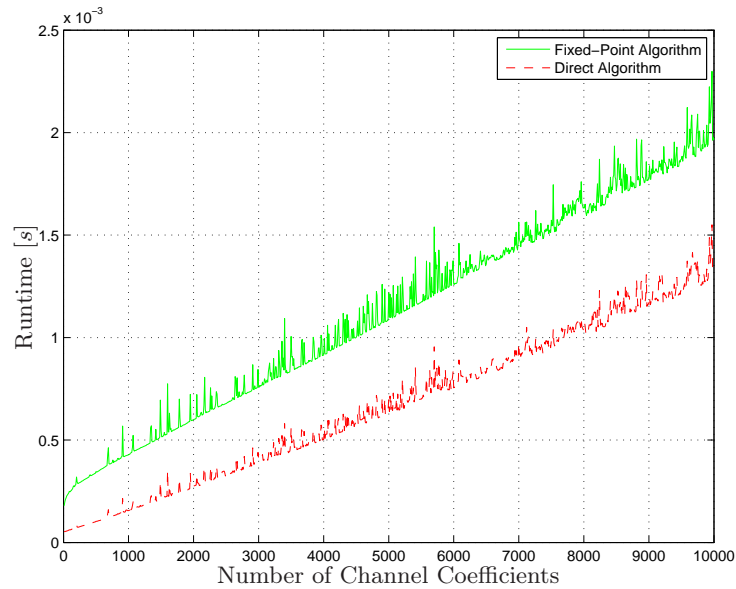


Figure 10.8.: Comparison of Algorithm Runtime in Matlab for Direct Calculation vs. Fixed-Point Algorithms. Lines smoothed by taking the average of 1000 algorithm runs for each data-point.

### 10.3. Performance Comparison with the “Classical Approach” (Direct Calculation of the Transmit Powers)

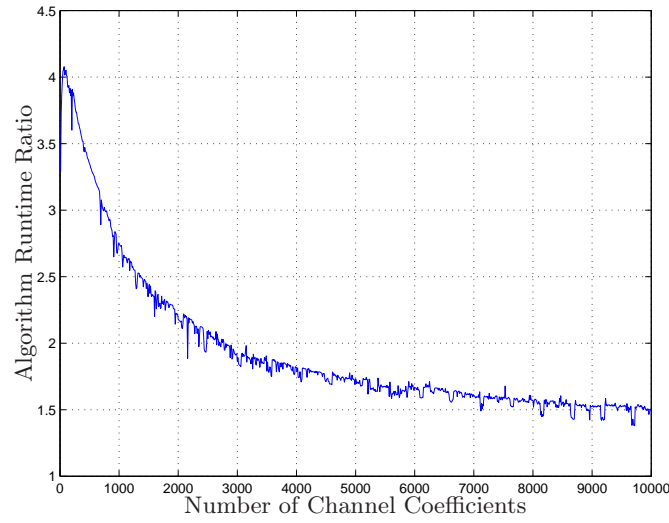


Figure 10.9.: Algorithm Runtime Ratio in Matlab for Direct Calculation vs. Fixed-Point Algorithms. The Fixed-Point Algorithm is slower. Lines smoothed by taking the average of 1000 algorithm runs for each data-point.

#### 10.3.4. Conclusion

The considerations in this Section have shown that from a performance point of view, the fixed-point approach is inferior to the classical approach of directly calculating the transmit powers from a set of known channel coefficients. On the other hand of course, in their presented, basic incarnations, the two algorithms are solving two different problems: the fixed-point approach aims to efficiently answer the question for the calculation of the “waterlevel”  $\lambda$ , while the classical approach is designed to return the channel powers for the respective channel coefficients only. There is, of course, the option to answer both questions using either of the two algorithms, at the cost of additional complexity and a degrading performance. In practice, the choice of the algorithms does depend on the specific requirements of the applications.



# 11. Power Allocation Strategies for the Finite Lifetime Channel

In this section, the previously developed and analyzed fixed-point function is used to develop an algorithm for optimum resource allocation in finite-lifetime channels. The various approaches are compared among each other to find an optimum scheme that is feasible and efficient. The properties of the finite-lifetime channel have been developed in Chapter 10.1 and are only important for comparison. Interpreting the numerical results, it is important to keep in mind that while the channel-coefficients and the evolution of the channel over time are based on actual channel measurements in a V2V scenario, the actual amount of information that is transmitted in the scenarios is an information-theoretical upper bound. The approach is still valid however, since the numbers are used for comparison, and should not be considered as actually achievable, although they are certainly in the range of what could be expected. Furthermore, any reason why realistically-achievable amount of information should be lower than stated applies to all discussed schemes equally and does therefore not distort the comparison: The relative loss of performance, e.g. due to the use of non-ideal modulation and coding schemes affects all schemes equally.

## 11.1. Acausal Knowledge of the Channel Coefficients

For causal channel knowledge at the transmitter, i.e. in a practical system, it seems impossible to solve *in advance* for the value of  $\lambda$  that, via power allocation will satisfy (8.8) (with  $SNR(i) = SNR^*(i)$ ), because the future channel coefficients  $h_{i+1}, h_{i+2}, \dots$  are unknown at time  $i$ .

It is, however, possible to determine the performance of a hypothetical acausal scheme that can take into account the *future* channel coefficients. Realistic schemes can then be compared to this optimum performance.

## 11.2. Causal Knowledge of the Channel Coefficients

**Non-Adaptive Scheme** If the channel was not only available during a finite amount of time, and if it were ergodic, finding a threshold for *good* channel coefficients would be easy by measuring over an extended period and keeping the optimized threshold for all times. Alternatively, a model of the channel's statistical properties could be estimated based on a number of observed coefficients. This channel model could be used to simulate a “fake” sequence of channel coefficients with the same statistical properties as the ones that are actually experienced, calculating the threshold based on this “fake sequence” and applying it to the real channel. This approach is also applicable to the channel of finite lifetime. Here, the “threshold”  $\lambda$  would be calculated off-line, and the transmit power would be determined by the currently available channel coefficient. Due to finite lifetime and non-ergodicity, however,  $\lambda$  will be impossible to determine exactly off-line. If a good physical model describing the long-term path loss is available and the life-time of the channel is large enough, the method might provide a good approximation to the actually experienced scenario. Due to the randomness of the channel coefficients and the finite lifetime, however,  $\lambda$  will never be exactly correct. This leads to the risk of assigning the wrong amount of energy: In case of  $\lambda$  being estimated too low, too much power would be used in the scheme - this can be corrected easily by just stopping power allocation in simulation when the energy budget is exhausted. The opposite case is much more difficult to assess: Overestimating  $\lambda$  will lead to the situation that the total energy budget can not be assigned. There is no way of combating this situation: by the time it becomes clear that the energy budget cannot be fully assigned, the “peak”-values of the channel coefficients have passed. An optimum waterfilling-based power-allocation is not possible any more, and assigning the remaining rest of the energy budget to weak channel coefficients will not help a lot. Nevertheless, this *non-adaptive* power allocation scheme that is based on a fixed “threshold”  $\lambda$  is included in the simulations, since it provides valuable insights into the performance penalty of non-adaptive schemes.

**Adaptive Scheme** A more elaborate, *adaptive* strategy to solve the problem is to re-calculate  $\lambda$  for every block, using the available current channel coefficient  $h(i)$  and append a fake-sequence  $\{h'(j), j = i + 1, \dots, |\mathcal{S}|\}$  of *future* channel coefficients. With every new known channel coefficient  $h(i)$ , the corresponding value in the fake sequence is replaced and the new sequence of channel coefficients is used for power allocation in the current block. After taking the decision what part of the available energy is to be used in the current block, the remaining energy budget that is left for the future use has to be updated. If  $|\mathcal{S}|$  – the lifetime of the channel – is sufficiently large, the error introduced by this procedure will be small and the available energy-budget will be used efficiently. The price to pay is the complexity for an update of  $\lambda$  for every transmit block.



### 11.3. Simulation Results

The following simulation results in Fig. 11.1-11.8 showcase the different approaches of the algorithms. The transmitted amount of information is what appears typical, the simulation results are, however, a way of illustrating complex improvements and setbacks when comparing the proposed approaches. To maintain strict comparability, the simulation parameters are kept the same for all illustrations - the channel model has been explained in Section 10.1. This includes the same initialization of the Matlab random number generator, so that the different algorithms operate on *the same* set of channel coefficients. The sum energy has been derived in (8.3) to be  $5 \cdot 10^{-2.1}W$ , and the noise power has been estimated to be  $10^{-14.1}W$  in (10.2). For all the simulations, the channel lifetime is  $2\Delta = 1000$  Blocks.

In the following illustrations, the acausal power allocation scheme (Chapter 8) in Fig. 11.1 is compared with the two causal schemes (Figs 11.2, 11.4) described in Section 11.1. In Fig. 11.2 the results for power allocation using  $\lambda$  (for use in 8.9) that are calculated in advance from a “fake” coefficient sequence of length  $2\Delta$  are shown. In Fig. 11.4 the results for block-adaptive  $\lambda$  as also described in Section 11.2 are exemplarily illustrated to show the typical behavior of the adaptive version of the waterfilling algorithm when the channel statistics are known.

For all illustrations, a channel coefficient is used for transmission if the magnitude square  $|h(i)|^2$  of the coefficient is larger than  $\lambda$  and this is indicated by “circles” in the upper plots of all figures. The optimal acausal scheme in Fig. 11.1 is the benchmark as it indicates the best possible theoretical performance for the given parameters. The SNR can be seen to increase and decrease over time according to the geometrical model of two vehicles driving on opposing lanes along a highway, as described in Section 10.1. When the two vehicles approach each other, the short-term average SNR grows until it reaches its maximum when the vehicles are closest. It then decreases again - far beyond the point when it is economically usable. Fig. 11.1 illustrates nicely that the notion of the *Finite Lifetime Channel* is correct only if the *economically feasible* transmission is sought, since there is some measurable signal power available for a long time. It furthermore shows that the movement model postulates a symmetric increase and decrease of the channel quality. Based on the given set of channel coefficients (that is the same for all simulations), the optimal acausal scheme assigns a total of approx. 552211 bits during the lifetime of the channel. Notably, despite the higher variance of the block fading in simulations than compared to the measurement campaign by e.g. [49], still all of the used channel coefficients fall into the period where a line-of sight connection can be expected. The achieved rates for the individual channel uses (the number of bits transmitted in each block were divided by the constant number of channel uses  $M$  in the block) are shown in the lower part of Fig. 11.1: those numbers are not higher than approximately 1 bit per channel use and, hence, reasonable for a practical implementation. Naturally, as the

## 11. Power Allocation Strategies for the Finite Lifetime Channel

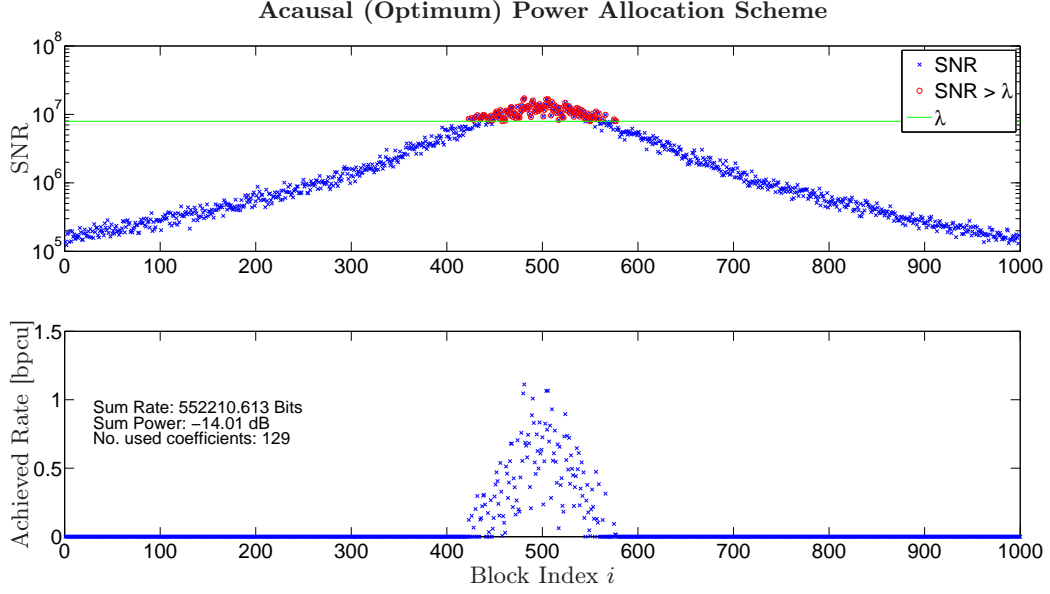


Figure 11.1.: Simulation of a vehicle-to-vehicle scenario with optimal acausal power allocation; upper plot: channel coefficients with indication which coefficients are used for transmission, lower plot: achieved rates in bits per channel-use. The channel lifetime is 1000 blocks, the sum transmit power is limited to  $5 \cdot 10^{-2.1} W$ .

scheme operates acausally on the complete set of actually experienced channel coefficients, the sum of the used powers is exactly as defined in (8.3), i.e.  $5 \cdot 10^{-2.1} W$ . The green line symbolizes the decision threshold  $\lambda$  - if the SNR is above the green line, the respective channel coefficient is assigned transmit power. The decision threshold is constant.

The problem of the required acausally known channel coefficients can be mitigated, as already pointed out in the introduction of the chapter, by using a “fake” sequence from a random generator that has exactly the same statistics as the original channel (non-adaptive approach from Section 11.1) for calculating  $\lambda$ . The results in Fig. 11.2 show that the power constraint is perfectly met (the assigned power would have been larger, had the algorithm not stopped to assign power to further channel coefficients). This can be seen by noting that there are a few channel coefficients that are above the decision threshold, but have not been used for transmission (as they are not marked with a red circle). As pointed out previously, the case of too many channel coefficients qualifying for transmission can easily be handled by stopping the assignment of further transmit power. The degradation in performance has to be accepted if the power constraint should be met for comparison purposes. If, on the other hand, the fake sequence of channel coefficients is such that the threshold  $\lambda$  is too high, not all of the available power can be assigned and the performance degrades some more.

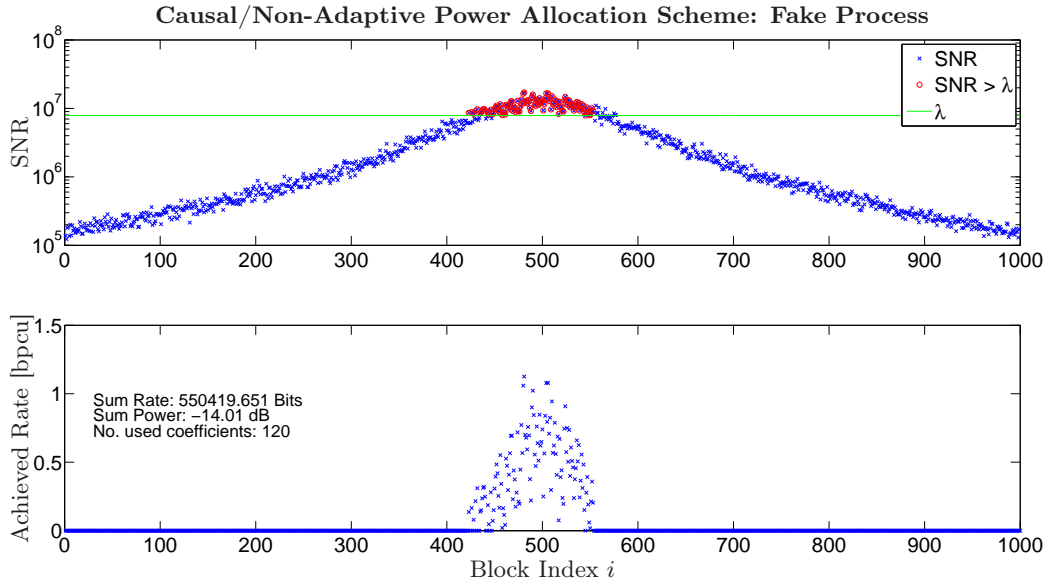


Figure 11.2.: Simulation of a vehicle-to-vehicle scenario with causal power allocation using a fake random process (with correct statistics); upper plot: channel coefficients with indication which coefficients are used for transmission, lower plot: achieved rates in bits per channel-use. The channel lifetime is 1000 blocks, the sum transmit power is limited to  $5 \cdot 10^{-2.1} W$ . Note that there are some channel coefficients that are above the green line but that are not assigned any transmit power: this is due to the exhausted power budget.

Fig. 11.3 illustrates the further degradation in performance if the artificially calculated threshold  $\lambda$  is too large. Compared to the opposite case when  $\lambda$  is too small, the performance degrades further in that there is some power left in the budget that is not used. Therefore, strictly speaking the performance (i.e. the number of transmitted bits) cannot be compared on a strict absolute basis, since there are fewer bits transmitted *and* less power used. However, as this is an expected and typical problem of this approach, it is still considered in the comparisons in Fig. 11.8.

The algorithm illustrated in Fig. 11.2 and 11.3 can be improved by considering the currently available channel coefficient and appending the fake sequence with correct statistics to “fill in” the rest of the channel life-time. Fig. 11.4 illustrates this approach. This real-time capable version of the algorithm is similar in performance to the previously presented algorithm that works purely off-line.

## 11. Power Allocation Strategies for the Finite Lifetime Channel

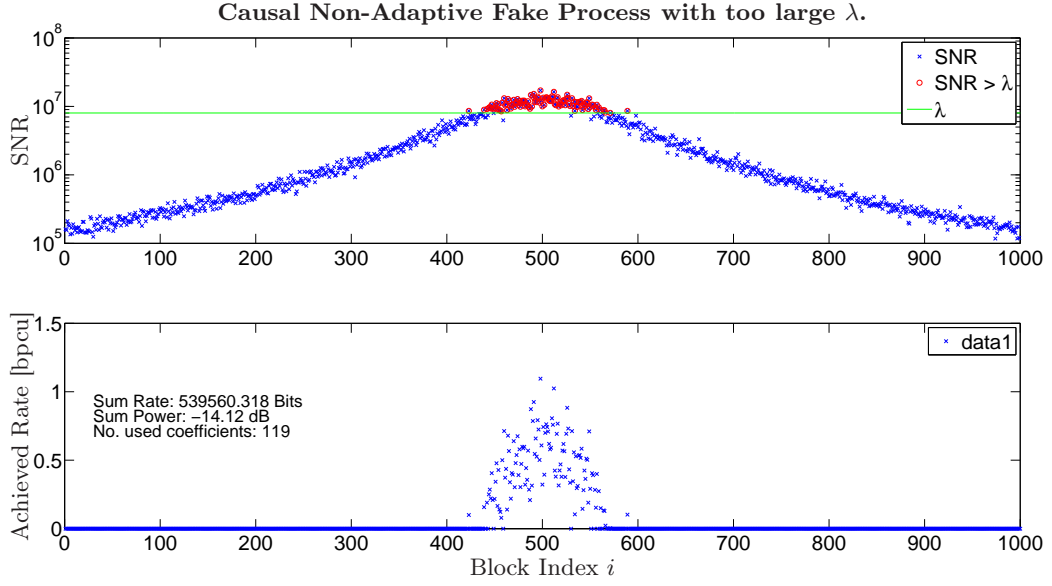


Figure 11.3.: Simulation of a vehicle-to-vehicle scenario with causal power allocation using a fake random process (with correct statistics but  $\lambda$  too large); upper plot: channel coefficients with indication which coefficients are used for transmission, lower plot: achieved rates in bits per channel-use. The channel lifetime is 1000 blocks, the sum transmit power is limited to  $5 \cdot 10^{-2.1}W$  but not fully used up.

In Fig. 11.4 an exemplary run of the causal scheme with adaptation of  $\lambda$  in every block  $i$  as described in Section 11.1 is shown. This time, and compared to the previous algorithms depicted in Fig. 11.1, 11.2, and 11.3, the threshold  $\lambda$  is not kept constant. Instead, the algorithm considers the presently available channel coefficient and appends a fake sequence with correct statistics to “fill in” the rest of the channel life-time. Here, the absolute life-time is not critical, i.e. if the long-term change in path-loss is known, it is irrelevant what exactly is defined as the channel life-time, as long as the coefficients that are eventually used for transmission are lying within the lifetime. This follows from the way the algorithm only considers the largest coefficients for resource allocation - the exact number or magnitude of the very small coefficients does not play a role in the scheduling decision. The remaining power is updated every time a channel coefficient was used for transmission, so that this improved version is guaranteed to assign the total available power. If all the power has been used up during the channel life-time, the algorithm stops assigning any more channel coefficients and the threshold  $\lambda$  is not drawn any further. This is depicted in Figure 11.4. Conversely, in cases when the previous, non-adaptive algorithm had failed to allocate all the available power budget, this time the algorithm decreases  $\lambda$  until *all* the power is used up.

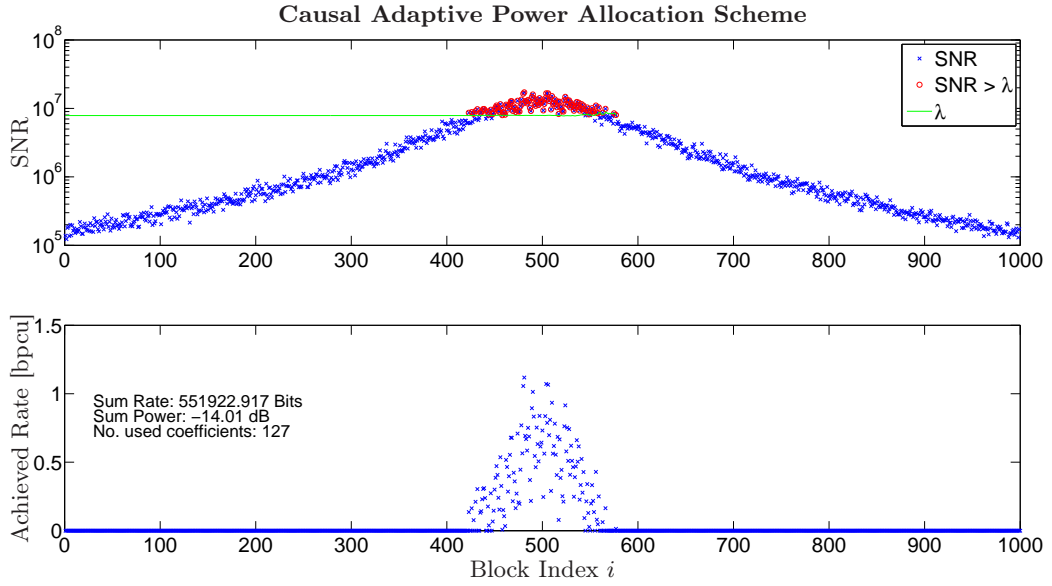


Figure 11.4.: Simulation of a vehicle-to-vehicle scenario with causal power allocation using a fake random process (with correct statistics) with adaptive  $\lambda$ ; upper plot: channel coefficients with indication which coefficients are used for transmission, lower plot: achieved rates in bits per channel-use. The channel lifetime is 1000 blocks, the sum transmit power is limited to  $5 \cdot 10^{-2.1}W$ .

This does not need to be illustrated, since only minor adaptations to  $\lambda$  that are barely visible in a plot will cause an allocation of the total available power budget.

The adaptive scheme also makes use of a fake sequence, and if the statistics of the fake sequence are significantly different from the true statistics of the channel, a loss in performance is inevitable. This is demonstrated by Fig. 11.5, where a path loss was assumed to generate the fake sequence that is 10dB lower than in the real channel.

When the estimated long-term path loss is 10dB too low compared to the actually experienced channel,  $\lambda$  is initialized with a value that is too high. In the present example in Fig. 11.5, this leads to a threshold that is actually above the best experienced channel coefficients - the non-adaptive scheme would not allocate *any* channel-coefficients any power. The adaptive scheme improves a lot in this regard: As soon as the channel coefficients are recognized to produce no data transmission at all, the decision threshold is continuously adapted. The resulting transmitted amount of information is still low with approx. 70,303Bits: This is the consequence of the *best* channel coefficients being assigned at a time when  $\lambda$  is still high (the coefficients are just barely above the threshold). Fig. 11.6 illustrates the opposite situation: The long-term path loss estimate

## 11. Power Allocation Strategies for the Finite Lifetime Channel

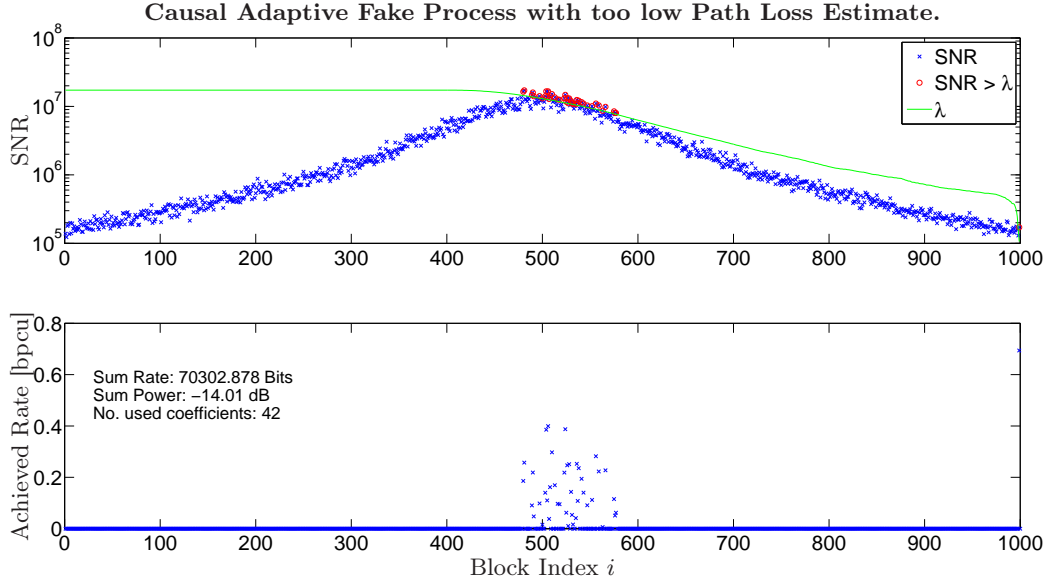


Figure 11.5.: Simulation of a vehicle-to-vehicle scenario with causal power allocation using a fake random process (with correct statistics but  $\lambda$  too large); upper plot: channel coefficients with indication which coefficients are used for transmission, lower plot: achieved rates in bits per channel-use. The channel lifetime is 1000 blocks, the sum transmit power is limited to  $5 \cdot 10^{-2.1} \text{W}$ . The channel statistics used at the transmitter are incorrect: the path loss is assumed to be 10dB too low compared to the actually experienced channel: therefore, the estimate for  $\lambda$  is initialized too high.

is too high by 10dB. This leads to an initial  $\lambda$  that is too low by a significant margin. In this case, the decision threshold is corrected upwards gradually but the power budget is used up before the *best* channel coefficients are actually experienced. In this case, the total amount of information that is transmitted is approx. 247,837Bits, and this is significantly higher than the previous case when  $\lambda$  is initialized too high. The non-adaptive algorithm with the same wrong assumption about the path loss performs worse, as can be seen in Fig. 11.7. This is the result of more power being assigned in better transmit conditions by the adaptive scheme as opposed to the case when there is hardly any power assigned to the very best coefficients due to too high  $\lambda$ .

Of course the simulation results discussed above are only snapshots for specific realizations of the channel coefficient sequences. They are useful to understand the operation of the schemes and to explain effects observed but for a fair performance comparison averaging over many channel realizations is required. The results are given in Fig. 11.8: To obtain the histograms for the number of bits that were transmitted, the above simulations were repeated  $10^6$  times.

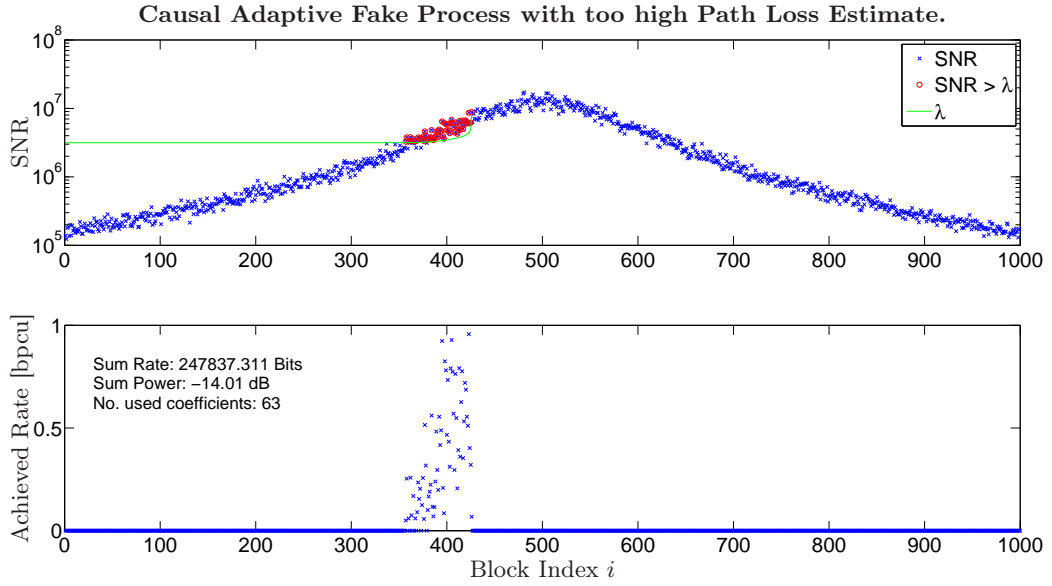


Figure 11.6.: Simulation of a vehicle-to-vehicle scenario with causal power allocation using a fake random process (with correct statistics but  $\lambda$  too small); upper plot: channel coefficients with indication which coefficients are used for transmission, lower plot: achieved rates in bits per channel-use. The channel lifetime is 1000 blocks, the sum transmit power is limited to  $5 \cdot 10^{-2.1} \text{W}$ . The channel statistics used at the transmitter are incorrect: the path loss is assumed to be 10dB too high compared to the actually experienced channel: therefore, the estimate for  $\lambda$  is initialized too low.

Both causal schemes (adaptive and non-adaptive) operate close to the theoretical performance limits as long as the channel statistics are known exactly at the transmitter. Moreover, for all simulations a relatively long channel lifetime of 1000 blocks is assumed, so a realization (sequence of coefficients) will reveal most of the statistical properties of the random process. This explains why both causal schemes perform very well, with the non-adaptive algorithm being slightly inferior when the power budget cannot be fully exploited - this problem does not exist in the adaptive version. The picture changes, however, when the statistical properties are not known exactly (as in the illustrated simulation results above, the path-loss was wrongly estimated to be 10dB lower and higher than the actually experienced channel). This is demonstrated by the two lower curves in Fig. 11.8 that reflect the performance of the adaptive scheme to cope with too high or too low initial thresholds  $\lambda$ . There are no histograms for the non-adaptive scheme since overestimating  $\lambda$  like in the adaptive case will not allow for any transmission at all, while a low initial threshold will produce results that are similar to the adaptive scheme. This all-or-nothing approach is highly

## 11. Power Allocation Strategies for the Finite Lifetime Channel

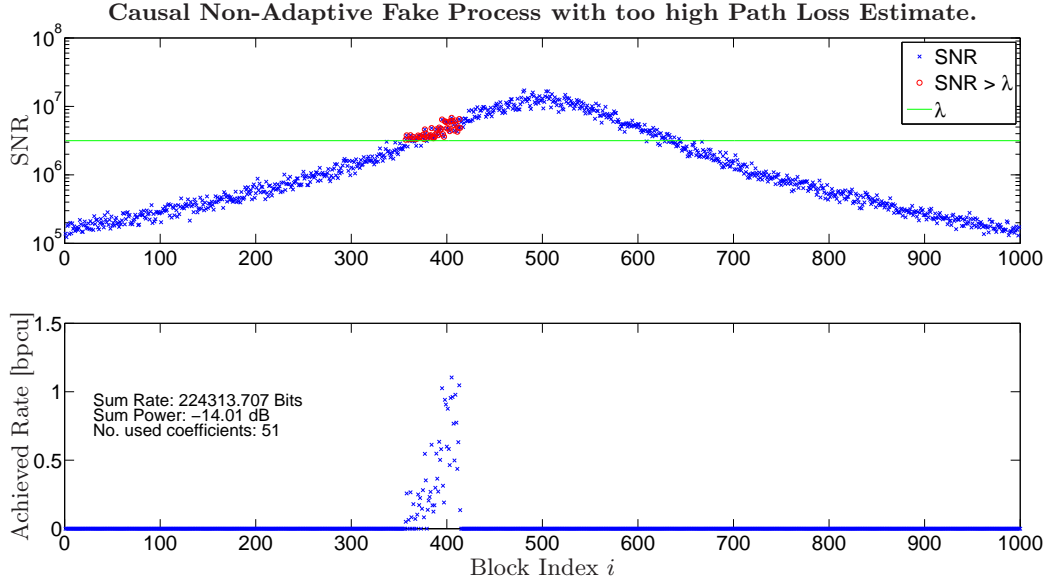


Figure 11.7.: Simulation of a vehicle-to-vehicle scenario with causal power allocation using a fake random process (with correct statistics but  $\lambda$  too large); upper plot: channel coefficients with indication which coefficients are used for transmission, lower plot: achieved rates in bits per channel-use. The channel lifetime is 1000 blocks, the sum transmit power is limited to  $5 \cdot 10^{-2.1} \text{W}$ . The channel statistics used at the transmitter are incorrect: the path loss is assumed to be 10dB too high compared to the actually experienced channel: therefore, the estimate for  $\lambda$  is initialized too low.

impractical and therefore not included in the illustration.

### 11.4. Conclusion

Based on the previously - developed Fixed-Point approach (Chapter 8), this Chapter explores two approaches for efficient use of the Finite Lifetime Channel. These are compared to an acausal (and therefore not feasible) benchmark solution that takes into account *all* channel coefficients that are available during the whole lifetime of the channel.

The two approaches that are discussed are causal, strictly taking into account only already - experienced channel measurements. The simplest possibility to assign power in a Finite Lifetime Channel environment is by estimation of the channel model either by early measurements or by learning from previous realizations and subsequently using a fixed, non-adaptive decision threshold  $\lambda$  for power allocation. This approach will suffer from a non-ideal choice of  $\lambda$ : If  $\lambda$  is initially set



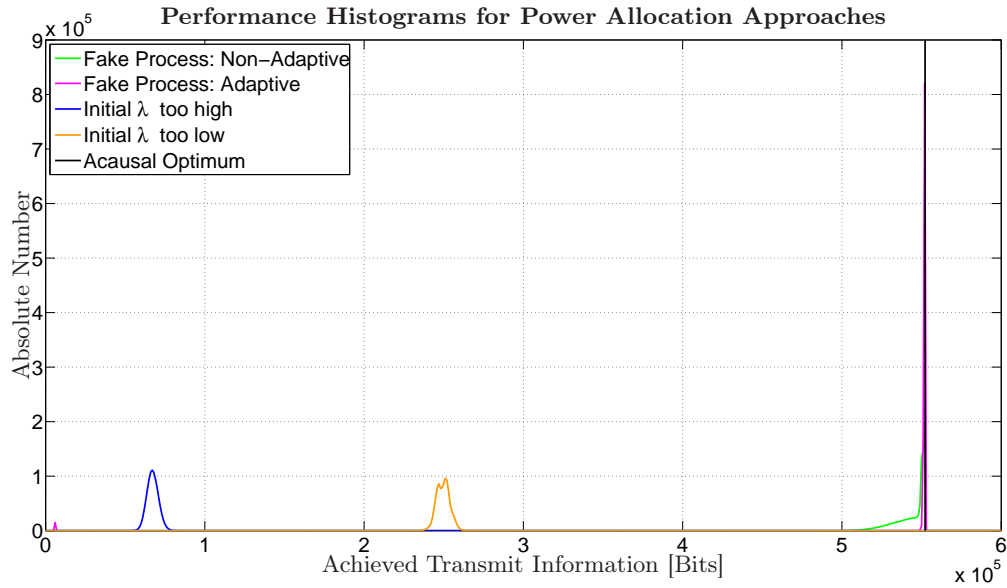


Figure 11.8.: Simulating the proposed schemes  $10^6$  times produces the histogram with absolute numbers of occurrence (and a bin width of  $10^3$ ). This makes the comparison of the discussed algorithms meaningful.

too high, the transmit energy budget will not be fully used (it may actually not be used at all if  $\lambda$  is larger than the highest experienced SNR). If, on the other hand,  $\lambda$  is set too low, the energy budget may be used up before the channel conditions are best, leading to a poor performance as well.

The obvious solution is an adaptive approach that constructs the future evolution of the channel from already-experienced channel coefficients and a fake sequence or previously experienced realizations. The latter possibility is much more practical since it does not require the estimation of a complicated channel model. In the simulations, a *fake sequence* is used instead of a previously measured sequence, because it allows tuning the sequence much easier. From the simulations it becomes clear that the optimum approach is to apply the adaptive scheme with a slightly low  $\lambda$  to start with (if the choice is only to initialize  $\lambda$  too high or too low). This will prevent the risk of missing the best coefficients for transmission. With an initial threshold near the optimum, the performance of the adaptive resource allocation approach is close to the acausal benchmark.



## 12. Use of the Fixed-Point Algorithm in Scenarios of Sporadically Available Channels

The classic and optimum (in an absolute, global sense) approach for maximizing the transmitted information under a sum-energy constraint is waterfilling, as illustrated in 12.1.

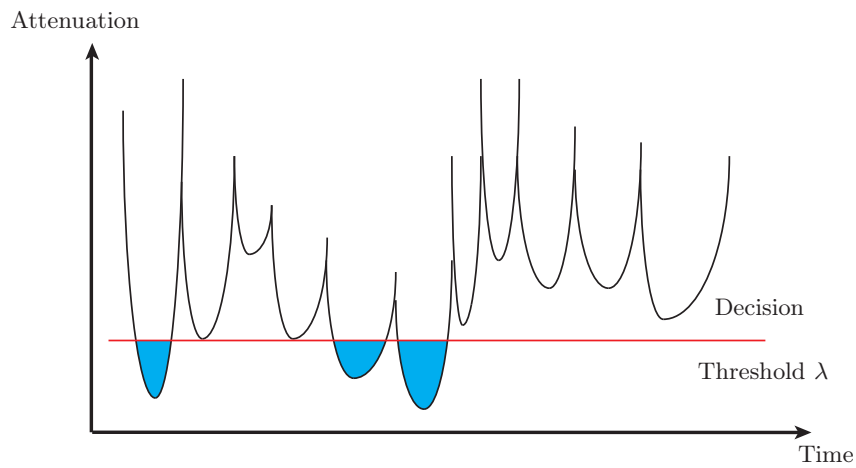


Figure 12.1.: The classic waterfilling approach in a scenario of recurring transmit opportunities of different quality. The threshold  $\lambda$  is not adaptive.

In practical scenarios like e.g. body-area networks or vehicular communications, classical waterfilling suffers one big drawback that is its *global*, i.e. long-term optimality - this is depicted in Fig. 12.1. From a connectivity-point of view, however, it may well be much more desirable to have an adaptive scheme that adjusts the decision threshold to the recently experienced channel quality, as illustrated in Fig. 12.2. Here,  $\lambda$  follows the channel attenuation just enough to trigger transmission regularly, without spending energy on every single attenuation notch. The advantage of such an approach is the potentially increased number of communication partners that will increase the degree of interworking among the devices. On the other hand, of course, keeping the long-term energy constraint on the same level that is used for the globally optimum approach, the amount

## 12. Use of the Fixed-Point Algorithm in Scenarios of Sporadically Available Channels

of transmitted information will decrease. This stems from the strategy of assigning some of the sum energy to weak channels that will not allow for comparatively efficient transmission.

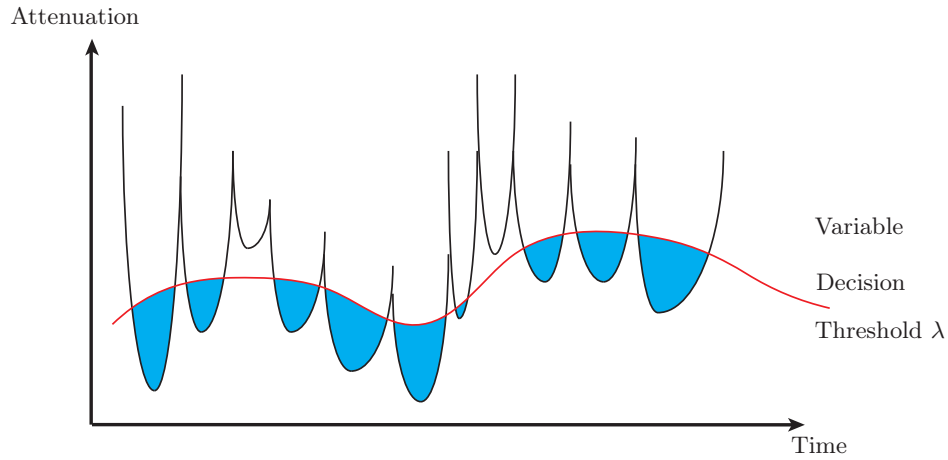


Figure 12.2.: Modified waterfilling with an adaptive  $\lambda$ : Transmit opportunities are found much more frequently, at the cost of a lower total amount of information being transmitted.

Compared to the classical approach and under the same long-term energy constraint, the total amount of information that can be transmitted with the variable scheme is naturally less, depending on the variability of the channel and the freedom to wait longer for good transmit opportunities at the expense of connected time. In future ad-hoc networking, the absolute amount of information might, on the other hand, not be the only critical performance aspect for power-allocating algorithms. Instead, the amount of relays might be a crucial factor - in this regard, variable-threshold waterfilling certainly outperforms the classical variant. This chapter sketches an approach to adapt the decision threshold  $\lambda$  to the recently-experienced slow fading without spending too much computational time on time- and energy-consuming waterfilling calculations.

### 12.1. EWMA-Triggered Waterfilling

In a scenario of recurring transmit opportunities, such as between different cars along a highway, or, very similarly, for ad-hoc-communications of personal devices which may exchange data (special case: act as a relay) in a crowded place (such as a shopping mall), it is an unsolved problem how to decide if a wireless link which is detected is good enough to establish a connection, and, if that is the case, how to optimally exploit the channel and at the same time save resources and minimize interference with others. One straight-forward idea might be to exploit *every* chance to transmit information (within practical limitations), or to introduce thresholds for different

kinds of recently experienced channels that are used to decide if a transmission should take place. The optimum strategy for assigning transmit power to channel states is waterfilling, and the problem with the approach is that the performance is highly dependent on the decision threshold  $\lambda$ . Broadly categorizing the recently experienced channel into distinct classes with individual transmit thresholds will be suboptimal.

Section 10.1 has introduced the channel model as one which takes into account two different types of fading: slow fading  $a_i(t) = \sin(2 \cdot \pi \cdot 1 \cdot t) \cdot 0.1 + 0.4$  and fast Rayleigh fading with scale parameter  $1/\sqrt{2}$  for the average power of the fast fading process to be 1. Such a channel behavior may occur along the highway in vehicular communications scenarios, or, physically very similarly, in pedestrian scenarios.

The proposed fixed-point algorithm intended for finite-lifetime channels tries to forecast the channel quality and adapt the decision threshold accordingly. However, this approach - while efficient - does not solve the problem of the fundamental decision if a connection shall be established at all, given all transmit opportunities to other devices within a certain timespan.

This section therefore suggests an alternative approach: if the waterlevel has been calculated based on the most recent  $N$  channel coefficients, and assuming that this time-span is long enough to contain, due to physical proximity of communication partners, a few “good” channel states, it will automatically provide a decision threshold when to establish a connection or to let the opportunity pass. In order to reduce accumulation of data (channel states), it is possible to sub-sample the channel state and consider only e.g. each  $10^{th}$  channel coefficient for the calculation of the waterlevel. Traditional waterfilling approaches typically calculate the transmit power in every current transmit block. For comparison, it is assumed that they do that for the same most recent  $N$  channel coefficients and with the same sub-sampling of the channel states. This frequent recalculation will, however, produce a steady processing demand based on a set of channel coefficients that is largely (except for the oldest and the newest) identical to the ones just analyzed.

The main idea is to re-calculate the decision threshold  $\lambda$  only when the mean attenuation of the  $N$  sub-sampled channel-coefficients has changed more than a certain threshold compared to the previous calculation. This can easily be implemented using an exponentially weighted moving average (EWMA) filter:

$$EWMA_{n+1} = EWMA_n \cdot \left(1 - \frac{1}{N}\right) + |h(n)|^2 \cdot \frac{1}{N} \text{ and } EWMA_1 = |h(1)|^2. \quad (12.1)$$

This filter returns the expected value for the channel powers  $|h(n)|^2$  they are stationary: For  $n = 1$ , the expected output of the filter is  $E(|h(1)|^2) = \mu_{h^2}$ , i.e. the mean channel power. For  $n = 2$ , it is  $\mu_{h^2} \cdot (1 - \frac{1}{N}) + E(|h(2)|^2) = \mu_{h^2}$ .

## 12. Use of the Fixed-Point Algorithm in Scenarios of Sporadically Available Channels

Continuing this argument, it can be seen that the expected filter output is always the mean channel power, albeit with an emphasis on the latest channel coefficients.

The following pseudo-code illustrates the procedure:

**while** 1

**Transmit Decision:**

**if**  $P^*(n) = \left( \frac{1}{\lambda \log(2)} - \frac{2\sigma^2}{|h(n)|^2} \right) > 0$

            transmit with power  $P^*(n)$ ;

**end**

**Update EWMA Filter:**

        EWMA = EWMA ·  $\left(1 - \frac{1}{N}\right)$  +  $|h_n|^2 \cdot \left(\frac{1}{N}\right)$ ;

**Check for Slow Fading:**

**if**  $\text{abs}(\text{EWMA} - \text{EWMA\_old}) > \Delta \cdot \text{EWMA\_old}$                       re-calculate  $\lambda$ ;

            EWMA\_old = EWMA;

**end**

**end**

1. Derive the current channel coefficient  $|h_n|^2$  and compare it to the recently-calculated water-level to decide on the use of the current block. This decision is very fast.
2. Use  $|h_n|^2$  for updating an Exponentially-Weighted Moving-Average (EWMA) filter.
3. Compare the current EWMA filter-output to some previously stored reference output, generated when the waterlevel was updated last.
4. If the EWMA filter output has changed more than e.g. 5% compared to the reference, update the waterlevel and make the current filter output the new reference.

There are two important parameters in this algorithm:

- The relative change of the EWMA filter output compared to the reference that triggers the re-calculation of the waterlevel. If the allowed change is too small, the performance of the algorithm will be equal to an algorithm directly calculating the transmit power in each block, therefore consuming more resources. If it is too large, the algorithm will not (or slowly) adapt to a change of the average link quality. Thresholds in the range of 5 to 10% have produced excellent results.

- The number of recent channel-coefficients  $N$ , which is also the forgetting factor of the EWMA-filter. If this value is too small, the algorithm does not recognize good connection opportunities, and it increases the computational load since the moving average will change a lot. If  $N$  is too large, the re-calculation of the waterlevel will take a long time and might not have finished when another re-calculation is already required.

Table 12.1 summarizes results that have been obtained with the above simulation parameters and  $10^7$  channel coefficients. The performance of the EWMA-triggered approach is compared to the previously discussed algorithm with direct calculation of the transmit powers (Section 10.3.2). Of course, fixed-point waterfilling is not the only algorithm that lends itself to an EWMA-triggered approach. The computational gain demonstrated in the following table can be realized with any waterfilling scheme that allows for calculation of the waterlevel. This does include the “direct” algorithm, as the waterlevel can be derived from the transmit powers and the channel coefficients in a second step without too much computational cost.

Comparing the presented EWMA-Approach to the previously discussed direct calculation of channel coefficients discussed in Section 10.3.2 reveals a large reduction in computational load. For comparison, the “direct” calculation of the transmit powers is carried out every time, i.e. there is no EWMA-filter that helps decide when a re-calculation of  $\lambda$  is needed because the direct algorithm does not compute  $\lambda$ . While the direct calculation of transmit powers is superior to the EWMA-triggered approach with  $N = 10$ , the latter algorithm is superior with respect to runtime in all the other cases. Interestingly the EWMA-triggered approach suffers only very little in terms of throughput, and this seems to hold for all values of  $N$ . This may result from the rather low re-calculation threshold of just 5% regarding the EWMA filter output. The gain with respect to Run Time is, on the other hand, clearly visible with growing  $N$ : with  $N = 10^2$ , the Run Time is already reduced by 71%,  $N = 10^3$  brings a reduction of 89% and for  $N = 10^4$ , the reduction in processing time is even 95%. This trend is also visible if the “Idle Time” is observed: this is the average number of channel coefficients between two updates of the waterlevel. Based on these observations, it seems practical to consider values of  $N$  in the range of  $10^3$ . Of course, this depends mainly on the application and hardware-restrictions.

12. Use of the Fixed-Point Algorithm in Scenarios of Sporadically Available Channels

Update Threshold: 5%		
	Direct Algorithm	EWMA-Triggered
<b>N = 10</b>		
Transmitted Bits	6.826e+07	3.853e+07
Energy used	2.294e+06	2.046e+06
Run Time	4.855e+02	9.708e+02
Idle Time	1	1.58
<b>N = 10<sup>2</sup></b>		
Transmitted Bits	4.487e+07	4.521e+07
Energy used	1.049e+06	1.065e+06
Run Time	5.467e+02	1.603e+02
Idle Time	1	24.75
<b>N = 10<sup>3</sup></b>		
Transmitted Bits	4.744e+07	4.733e+07
Energy used	1.000e+06	9.971e+05
Run Time	1.463e+03	1.634e+02
Idle Time	1	219.81
<b>N = 10<sup>4</sup></b>		
Transmitted Bits	4.746e+07	4.695e+07
Energy used	1.000e+06	9.856e+05
Run Time	1.473e+04	7.551e+02
Idle Time	1	7.752e+04

Table 12.1.: The table summarizes the results from calculating the decision threshold every time a scheduling decision has to be made (“Direct Algorithm”) and from updating the decision threshold only if the recent average changes more than 5% compared to the latest update (“EWMA-Triggered”). While the amount of transmitted information as well as the energy used is similar for the two strategies, the run-time (measured as Matlab simulation run-time) is much shorter in the “EWMA-Triggered” case if the scheduling decision is based on a long history ( $N > 10^2$  channel coefficients).



## 12.2. Estimation of the Implementation Complexity

The question of practicability in terms of implementational complexity and demand for resources is still unanswered. Based on the simulation results for  $N = 10^3$ , and considering a block duration of  $1ms$ , the following estimations can be made:

If every  $10^{th}$  channel coefficient is sampled and stored into a cyclic buffer of length  $N = 10^3$ , the waterfilling algorithm has to be run approximately every 2.2 seconds. If all the channel coefficients are stored with a resolution of 16 Bit, this will cause a memory requirement of 2 kB, which seems absolutely reasonable. In the case of an EWMA-triggered update of the waterlevel, the processor would copy the whole cyclic buffer to another memory and perform the waterfilling on this set of channel coefficients. There is no need to finish the calculation within  $1ms$  - the assumed minimum channel coherence time - the transmit decisions can be made on the old waterlevel, until the new one is available.



## 13. Conclusion of Part II

While the first part of this thesis has explored strategies for efficient use of resources when there are user-imposed time limitations in the form of rate- and delay- constraints, the second part considers time limitations that originate from the channel. These are experienced in the form of periods where a transmission is economically feasible, and intensified from an algorithmical point of view when the channel is not ergodic during these windows of opportunity. Sum-power constraints directly lead to a waterfilling solution which raises the question for the optimum transmit decision threshold (i.e., the “waterlevel”) when the future evolution of the channel is unknown. A new approach to directly calculate this decision threshold is derived, and its properties are explored analytically. Subsequently, the gathered insights are used to compare acausal, causal and causal-adaptive implementations regarding efficient use of the limited transmit power budget. The second part of this thesis concludes with the conceptual discussion of a strategy for recognizing transmit opportunities when the channel is available during recurring time windows.



## A. Arithmetic Considerations around a Function Maximum

All the variables used in this section represent real, positive numbers.

Suppose a fraction  $\frac{a}{b}$ . Now, add a value  $c$ , to the numerator, and another value,  $d$ , to the denominator. We have now the fraction  $\frac{a+c}{b+d}$ , and the following statement holds:

$$\frac{a+c}{b+d} < \frac{a}{b} \text{ if and only if } \frac{c}{d} < \frac{a}{b} \quad (\text{A.1})$$

*Proof.* Instead of writing  $\frac{c}{d} < \frac{a}{b}$ , we can also write  $\delta \cdot \frac{a}{b} < \frac{a}{b}$ , and it remains to be shown that  $\delta < 1$ .

So, if  $\frac{a+\delta \cdot a}{b+b} \stackrel{!}{<} \frac{a}{b}$ , then

$$\frac{a(1+\delta)}{2b} \stackrel{!}{<} \frac{a}{b} \quad (\text{A.2})$$

$$\delta \stackrel{!}{<} 1 \quad (\text{A.3})$$

Which proves (9.45). □

For subtracting terms in both the numerator and the denominator of the original fraction  $\frac{a}{b}$ , as in  $\frac{a-c}{b-d}$ , this rule cannot be applied. Instead, the following holds:

$$\frac{a-c}{b-d} < \frac{a}{b} \text{ if and only if } \frac{c}{d} > \frac{a}{b} \quad (\text{A.4})$$

### A. Arithmetic Considerations around a Function Maximum

*Proof.* Let  $0 < z < 1$  and  $0 < n < 1$ , and find  $z$ , and  $n$  such that:

$$\frac{a - z \cdot a}{b - n \cdot b} < \frac{a}{b}. \quad (\text{A.5})$$

$$\frac{a - z \cdot a}{b - n \cdot b} \stackrel{!}{<} \frac{a}{b} \quad (\text{A.6})$$

$$\frac{a(1 - z)}{b(1 - n)} \stackrel{!}{<} \frac{a}{b} \quad (\text{A.7})$$

$$\frac{1 - z}{1 - n} \stackrel{!}{<} 1 \quad (\text{A.8})$$

$$1 - z \stackrel{!}{<} 1 - n \quad (\text{A.9})$$

$$\text{and therefore } z \stackrel{!}{>} n. \quad (\text{A.10})$$

That shows that subtracting a number  $c$  from the numerator  $a$  of a fraction and subtracting  $d$  from the denominator  $b$  of a fraction will yield a *smaller* fraction if and only if a fraction, comprised of  $c$  in the numerator and  $d$  in the denominator, is bigger than the original fraction  $\frac{a}{b}$ .

This proves (A.4). □

## B. Derivation of $\tilde{\varphi}(\lambda)$

In order to derive  $\tilde{\varphi}(\lambda)$  with respect to  $\lambda$ , the quotient rule is applied:

$$\frac{u}{v} \doteq \tilde{\varphi}(\lambda) = \frac{N \cdot (1 - F_{|H|^2}(\lambda))}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1 - F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj} . \quad (\text{B.1})$$

The derivation of the numerator with respect to  $\lambda$  can easily be carried out to be:

$$u' = \frac{d}{d\lambda} N \cdot (1 - F_{|H|^2}(\lambda)) = -N \cdot f_{|H|^2}(\lambda) . \quad (\text{B.2})$$

And for the denominator, we set (compare B.1):

$$v(\lambda) = \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1 - F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1}\left(1 - \frac{j}{N}\right)} dj. \quad (\text{B.3})$$

In order to derive this function with respect to  $\lambda$ , with the derivation acting on the upper bound of the integral, the following method is chosen. Reconsidering  $v(\lambda)$  such that

$$v(\lambda) = \text{const.} + \int_0^{a(\lambda)} g(y) dy = v(a(\lambda)) \quad (\text{B.4})$$

enables us to apply the chain-rule

$$v'(a(\lambda)) = \frac{d}{da} v \cdot \frac{d}{d\lambda} a . \quad (\text{B.5})$$

According to the fundamental theorem of calculus,

$$v(a(\lambda)) = \text{const.} + \int_0^{a(\lambda)} g(y) dy \quad (\text{B.6})$$

$$= \text{const.} + G(a(\lambda)) - G(0) \quad (\text{B.7})$$

### B. Derivation of $\tilde{\varphi}(\lambda)$

the derivative  $v'$  can be expressed as:

$$\frac{d}{d\lambda} \left( \text{const.} + \int_0^{a(\lambda)} g(y) dy \right) \quad (\text{B.8})$$

$$= \frac{d}{d\lambda} (\text{const.} + G(a(\lambda)) - G(0)) \quad (\text{B.9})$$

$$= \frac{d}{d\lambda} G(a(\lambda)) - \frac{d}{d\lambda} G(0) \quad (\text{B.10})$$

$$= g(a(\lambda)) \cdot a'(\lambda) . \quad (\text{B.11})$$

Plugging (B.3) back into this general relation leads to the following derivation:

$$g(a(\lambda)) \cdot a'(\lambda) \quad (\text{B.12})$$

$$\doteq \frac{1}{F_{|H|^2}^{-1} \left( 1 - \frac{N(1-F_{|H|^2}(\lambda))}{N} \right)} \cdot (-N \cdot f_{|H|^2}(\lambda)) \quad (\text{B.13})$$

$$= \frac{1}{F_{|H|^2}^{-1} (1 - 1 + F_{|H|^2}(\lambda))} \cdot (-N \cdot f_{|H|^2}(\lambda)) \quad (\text{B.14})$$

$$= \frac{1}{F_{|H|^2}^{-1} (F_{|H|^2}(\lambda))} \cdot (-N \cdot f_{|H|^2}(\lambda)) \quad (\text{B.15})$$

$$= -\frac{N}{\lambda} \cdot f_{|H|^2}(\lambda) \quad (\text{B.16})$$

Now, returning to the quotient rule:

$$\left( \frac{u}{v} \right)' = \frac{vu' - uv'}{v^2} = \frac{vu'}{v^2} - \frac{uv'}{v^2} = \frac{u'}{v} - \frac{uv'}{v^2} , \quad (\text{B.17})$$

the expressions (B.16) and (B.2) can be resubstituted to give the first derivative of the fixed-point function  $\tilde{\varphi}(\lambda)$  (B.3):

$$\frac{u'}{v} - \frac{uv'}{v^2} \quad (\text{B.18})$$

$$\doteq - \frac{N \cdot f_{|H|^2}(\lambda)}{\frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1} \left( 1 - \frac{j}{N} \right)} dj} \quad (\text{B.19})$$

$$+ \frac{N \cdot (1 - F_{|H|^2}(\lambda)) \cdot \frac{N}{\lambda} f_{|H|^2}(\lambda)}{\left[ \frac{E_0/M}{2\sigma^2} + \int_0^{N \cdot (1-F_{|H|^2}(\lambda))} \frac{1}{F_{|H|^2}^{-1} \left( 1 - \frac{j}{N} \right)} dj \right]^2} . \quad (\text{B.20})$$



# Bibliography

- [1] “The World in 2013: ICT Facts and Figures,” ITU ICT Data and Statistics Division, February 2013. [Online]. Available: [www.itu.int/ict](http://www.itu.int/ict)
- [2] “The World in 2015: ICT Facts and Figures,” ITU ICT Data and Statistics Division, May 2015. [Online]. Available: [www.itu.int/ict](http://www.itu.int/ict)
- [3] J. Gonter and N. Goertz, “Power-Controlled Cross-Layer Scheduling,” in *Proceedings 2013 IEEE International Conference on Communications (ICC)*, June 2013, pp. 5463–5467.
- [4] J. Gonter, N. Goertz, and A. Winkelbauer, “Analytical Outage Probability for max-based Schedulers in Delay-Constrained Applications,” in *Proceedings 2012 IFIP Wireless Days Conference*, Nov. 2012.
- [5] N. Görtz and J. Gonter, “Information Transmission over a Finite-Lifetime Channel under Energy-Constraints,” in *Proceedings European Wireless 2011*, Vienna, April 2011.
- [6] —, “Limits on Information Transmission in Vehicle-to-Vehicle Communication,” in *Proceedings 2011 IEEE Vehicular Technology Conference Spring (ITC)*, Budapest, Hungary, May 2011.
- [7] J. Gonter and N. Görtz, “An Algorithm for Highly Efficient Waterfilling with Guaranteed Convergence,” in *Proceedings 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2012, pp. 126–129.
- [8] J. Gonter, N. Goertz, M. Rupp, and W. Gartner, “EWMA-Triggered Waterfilling for Reduced-Complexity Resource Management in ad-hoc Connections,” in *Proceedings 2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept 2013, pp. 2555–2559.
- [9] J. Costello, D.J. and J. Forney, G.D., “Channel coding: The Road to Channel Capacity,” *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1150–1177, June 2007.
- [10] “What happens in an Internet Minute?” Intel Corporation, 2014. [Online]. Available: <http://www.intel.de/content/www/de/de/communications/internet-minute-infographic.html>

## Bibliography

- [11] R. McEliece and W. Stark, "Channels with Block Interference," *IEEE Transactions on Information Theory*, vol. 30, no. 1, pp. 44–53, 1984.
- [12] G. Kaplan and S. Shamai, "Error Exponents And Outage Probabilities For The Block-Fading Gaussian Channel," in *Proceedings 1991 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 1991, pp. 329–334.
- [13] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information Theoretic Considerations for Cellular Mobile Radio," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 2, pp. 359–378, May 1994.
- [14] S. Hanly and D. Tse, "Multi-Access Fading Channels-Part II: Delay-Limited Capacities," in *Proceedings 1998 IEEE International Symposium on Information Theory (ISIT)*, Aug. 1998, p. 397.
- [15] K. Gilhousen, I. Jacobs, R. Padovani, A. Viterbi, J. Weaver, L.A., and I. Wheatley, C.E., "On the Capacity of a Cellular CDMA System," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 2, pp. 303–312, May 1991.
- [16] S. Hanly, "An Algorithm for Combined Cell-Site Selection and Power Control to Maximize Cellular Spread Spectrum Capacity," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1332–1340, Sep 1995.
- [17] R. Yates, "A Framework for Uplink Power Control in Cellular Radio Systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, Sep 1995.
- [18] S. Hanly and D. Tse, "Multi-Access Fading Channels: Shannon and Delay-Limited Capacities," in *Proceedings 33rd Allerton Conference*, Oct. 1995.
- [19] G. Caire, G. Taricco, and E. Biglieri, "Optimum Power Control Over Fading Channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1468–1489, Jul. 1999.
- [20] L. Li and A. Goldsmith, "Capacity and Optimal Resource Allocation for Fading Broadcast Channels .I. Ergodic Capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1083–1102, 2001.
- [21] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," [www.eecs.berkeley.edu/~dtse/broadcast2.pdf](http://www.eecs.berkeley.edu/~dtse/broadcast2.pdf), unpublished.
- [22] L. Li and A. Goldsmith, "Capacity and Optimal Resource Allocation for Fading Broadcast Channels .II. Outage Capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1103–1127, 2001.

- [23] X. Liu and A. Goldsmith, "Optimal Power Allocation over Fading Channels with Stringent Delay Constraints," in *Proceedings 2002 IEEE International Conference on Communications (ICC)*, vol. 3, 2002, pp. 1413–1418 vol.3.
- [24] R. Negi and J. Cioffi, "Delay-Constrained Capacity with Causal Feedback," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2478–2494, Sep 2002.
- [25] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423/623–656, July/October 1948 (reprint available at <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>).
- [26] M. Shaqfeh, "Resource Allocation and Flexible Scheduling in Wireless Networks," Ph.D. dissertation, 2009. [Online]. Available: <http://books.google.at/books?id=UVjpcQAACAAJ>
- [27] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability (available online at <http://projecteuclid.org>)*, Berkeley, CA, USA, July/August 1950, pp. 481–492.
- [28] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge University Press, 2005.
- [29] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic Beamforming Using Dumb Antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.
- [30] R. Knopp and P. Humblet, "Information Capacity and Power Control in Single-Cell Multiuser Communications," in *Proceedings 1995 IEEE International Conference on Communications (ICC)*, Seattle, WA, USA, Jun. 1995, pp. 331–335.
- [31] H. Zimmermann, "OSI Reference Model – The ISO Model of Architecture for Open Systems Interconnection," *IEEE Transactions on Communications*, vol. COM-28, no. 4, pp. 425–432, Apr. 1980.
- [32] ISO/IEC, "Information Technology – Open Systems Interconnection – Basic Reference Model: The Basic Model," *ISO/IEC 7498-1:1994(E)*, 1994.
- [33] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "A Bandwidth Efficient High Speed Data Service for Nomadic Users," *IEEE Communications Magazine*, Jul. 2000.
- [34] M. Andrews, "Instability of the Proportional Fair Scheduling Algorithm for HDR," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1422–1426, Sep. 2004.

- [35] R. K. G. Caire, R.R. Müller, “Hard Fairness versus Proportional Fairness in Wireless Communications: The Single Cell Case,” *IEEE Transactions on Information Theory*, vol. 53, no. 4, pp. 1366–1385, Apr. 2007.
- [36] J.-G. Choi and S. Bahk, “Cell-Throughput Analysis of the Proportional Fair Scheduler in the Single-Cell Environment,” *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 766–778, Mar. 2007.
- [37] J. Holtzman, “Asymptotic Analysis of Proportional Fair Algorithm,” in *Proceedings 2001 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 2, Sept./Oct. 2001, pp. F-33 – F-37 vol. 2.
- [38] E. Liu, Q. Zhang, and K. Leung, “Asymptotic Analysis of Proportionally Fair Scheduling in Rayleigh Fading,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 6, pp. 1764–1775, Jun. 2011.
- [39] E. Liu and K. Leung, “Proportional Fair Scheduling: Analytical Insight under Rayleigh Fading Environment,” in *Proceedings 2008 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2008, pp. 1883–1888.
- [40] H. David and H. Nagaraja, *Order Statistics, 3rd Edition*, 3rd ed. Wiley & Sons, 2003.
- [41] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002, vol. Fourth Edition.
- [42] G. Horvath and C. Vulkan, “Throughput Analysis of the Proportional Fair Scheduler in HSDPA,” in *Proceedings 2008 14th European Wireless Conference (EW)*, June 2008.
- [43] H. Kushner and P. Whiting, “Convergence of Proportional-Fair Sharing Algorithms under General Conditions,” *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1250 – 1259, July 2004.
- [44] R. I. Dobrushin, “General Statement of Shannon’s main Theorem in the Information Theory,” *Doklady Akademii Nauk SSSR*, vol. 126, no. 3, pp. 474–477, 1959.
- [45] S. Verdú and T. S. Han, “A General Formula for Channel Capacity,” *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
- [46] D. Tse and S. Hanly, “Multiaccess Fading Channels. I. Polymatroid Structure, Optimal Resource Allocation and Throughput Capacities,” *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2796–2815, Nov 1998.

- [47] A. Paier, J. Karedal, N. Czink, H. Hofstetter, C. Dumard, T. Zemen, F. Tufvesson, C. Mecklenbrauker, and A. Molisch, "First Results from Car-to-Car and Car-to-Infrastructure Radio Channel Measurements at 5.2GHz," in *Proceedings 2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sept 2007, pp. 1–5.
- [48] A. Paier, J. Karedal, N. Czink, H. Hofstetter, C. Dumard, T. Zemen, F. Tufvesson, A. Molisch, and C. Mecklenbrauker, "Car-to-Car Radio Channel Measurements at 5GHz: Pathloss, Power-Delay Profile, and Delay-Doppler Spectrum," in *Proceedings 2007 4th International Symposium on Wireless Communication Systems (ISWCS)*, Oct 2007, pp. 224–228.
- [49] A. Paier, T. Zemen, L. Bernado, G. Matz, J. Karedal, N. Czink, C. Dumard, F. Tufvesson, A. Molisch, and C. Mecklenbrauker, "Non-WSSUS Vehicular Channel Characterization in Highway and Urban Scenarios at 5.2GHz using the Local Scattering Function," in *Proceedings 2008 International ITG Workshop on Smart Antennas (WSA)*, Feb 2008, pp. 9–15.
- [50] J. Karedal, F. Tufvesson, N. Czink, A. Paier, C. Dumard, T. Zemen, C. Mecklenbrauker, and A. Molisch, "A Geometry-Based Stochastic MIMO Model for Vehicle-to-Vehicle Communications," *IEEE Transactions on Wireless Communications*, vol. 8, no. 7, pp. 3646–3657, July 2009.
- [51] —, "Measurement-Based Modeling of Vehicle-to-Vehicle MIMO Channels," in *Proceedings 2009 IEEE International Conference on Communications (ICC)*, June 2009, pp. 1–6.
- [52] A. Alonso, C. Mecklenbrauker, A. Paier, T. Zemen, N. Czink, and F. Tufvesson, "Temporal Evolution of Channel Capacity in Vehicular MIMO Channels in the 5GHz Band," in *Proceedings 2010 URSI International Symposium on Electromagnetic Theory (EMTS)*, Aug 2010, pp. 938–941.
- [53] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-Time Block Codes from Orthogonal Designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.
- [54] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.
- [55] M. Effros, A. Goldsmith, and Y. Liang, "Generalizing Capacity: New Definitions and Capacity Theorems for Composite Channels," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3069–3087, Jul. 2010.
- [56] E. Biglieri, J. Proakis, and S. S. (Shitz), "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

## Bibliography

- [57] J. C. Kieffer, “ $\epsilon$ -Capacity of Binary Symmetric Averaged Channels,” *IEEE Transactions on Information Theory*, vol. 53, no. 1, pp. 288–303, Jan. 2007.
- [58] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.
- [59] A. Paier, J. Karedal, N. Czink, C. Dumard, T. Zemen, F. Tufvesson, A. Molisch, and C. Mecklenbrauker, “Characterization of Vehicle-to-Vehicle Radio Channels from Measurements at 5.2GHz,” 2008.
- [60] P. Palomar and J. R. Fonollosa, “Practical Algorithms for a Family of Waterfilling Solutions,” *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 686–695, Feb. 2005.