Towards Energy Aware Wireless Network Operation: Analysis and Algorithms for Load Aware Network Topology Control

vorgelegt von
Diplom-Ingenieur
Emmanuel Pollakis
geboren in Überlingen

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Adam Wolisz, TU Berlin
Gutachter:	Prof. Dr.-Ing. Slawomir Stanczak, TU Berlin
	Prof. Dr.-Ing. Eduard A. Jorswieck, TU Dresden
	Dr. Gabor Fodor, KTH und Ericsson Research

Tag der wissenschaftlichen Aussprache: 06.07.2017

Berlin 2017

Abstract

Todays cellular communication networks have come a long way from voice only capabilities to wireless networks that serve millions of users and machine type devices with high data rates and low latencies. Typically, network designers overprovision the capacity of those networks in order to allow for some increase in demand over the years. This approach together with a largely static network operation leads to the consumption of a huge amount of energy which contributes to the constantly increasing CO_2 footprint of the ICT-sector. Against this background, this thesis addresses the energy consumption of cellular communication networks with a timely varying traffic load as it is observable in todays networks. We focus on the problem of implementing an optimal network topology control that is concerned with the energy consumption of all active network elements.

We start by giving an insight in the growth behavior of wireless communication systems when the number of users tends to infinity. A scaling law analysis of throughput as a function of the number of users having some QoS requirement is provided and we extend the results into the direction of embodied network energy consumption. In our analysis we specifically include the energy consumed by hardware for, e.g., cooling or processing, which is typically neglected in theoretical studies from information theory. The identified scaling behavior is further investigated by simulations that allow to verify the scaling results predicted by theory and to assess the speed of convergence of the asymptotic results. The main conclusions of this scaling analysis highlight the tradeoff between throughput and energy efficiency scaling, and provide valuable insights into the question how the number of infrastructure nodes, such as base stations, needs to scale with the number of wireless devices so as to achieve high energy efficiency with practical communication schemes.

Motivated by these results we develop an optimization framework targeting the minimization of the overall energy consumption in the downlink channel of mobile cellular networks including the energy consumed by hardware and auxiliary equipment. We follow an approach that aims to selecting the set of network elements consuming the least amount of energy while satisfying the QoS requirements of all users in the system. The underlying problem is of combinatorial nature and thus it is in general hard to solve. In order to obtain solutions in a computationally efficient manner we apply relaxation techniques to approximate the objective function by a concave function and relax all non-convex constraints which leads to a problem that is amenable to majorization-minimization techniques. With these steps we are able to propose an optimization framework that can find good solutions to the combinatorial problem in relatively short time. A major advantage of the framework is

its ability to cope with a variety of different network elements and energy consumption models. We provide extensive simulation results to show the framework's performance. With the core optimization framework at hand we include several extensions to account for novel aspects of networks such as CoMP transmission, the desire for a distributed implementation or to account for buffering capabilities at users' devices. The latter is used for an extension towards anticipatory scheduling, where the decisions on resource allocation and user-cell assignments are based not only on present channel state information but also on information about future propagation conditions. Thereby, we are able to exploit the knowledge about users' mobility and path loss to proactively build user-cell assignment and resource allocation schedules that greatly support energy savings in cellular communication systems.

Finally, we turn our attention towards the general framework of interference calculus to obtain computationally efficient algorithms for a more accurate estimation of the load at network elements. The knowledge about cell load is highly beneficial to develop tools that allow us to identify feasible network settings for given user-cell assignments. These tools are then used to facilitate the network topology control optimization framework for increased energy savings. We conclude the thesis with the presentation of a framework which allows us to identify feasible cell configurations (e.g., antenna tilts) supporting the energy savings algorithm that deactivate redundant network elements. In other words, with the help of heuristic arguments we combine the improved algorithm for topology control with the optimization of cell configurations to show how the optimization of down tilts can help to save energy.

Zusammenfassung

Die Entwicklung im Mobilfunk hat in den letzten Jahrzehnten die Leistungsfähigkeit zellularer Kommunikationsnetze beachtlich gesteigert. Netze, die ursprünglich der reinen Sprachübertragung dienten, sind zu drahtlosen Kommunikationsnetzen geworden, die Millionen von Nutzer und Geräte mit hohen Datenraten bei niedrigen Latenzen versorgen. Üblicherweise planen Netzdesigner die Kapazität dieser Netze großzügig, um einen gewissen Anstieg der Nachfrage über die Jahre hinweg abfedern zu können. In Verbindung mit einem weitgehend statischen Netzbetrieb führt dieser Ansatz zu einem immensen Energieverbrauch, der einen steigenden Anteil der IKT am CO_2-Ausstoß fördert. Vor diesem Hintergrund liegt der Fokus dieser Arbeit auf dem Energieverbrauch von zellularen Kommunikationsnetzen bei einer zeitlich schwankenden Verkehrslast, wie sie in heutigen Netzen zu beobachten ist.

Im ersten Schritt wird ein Überblick über das Skalierungsverhalten von drahtlosen Kommunikationssystemen gegeben, in denen die Anzahl der Nutzer gegen unendlich strebt. Dabei wird die Skalierung von Durchsatz als Funktion der Anzahl von Nutzern, die eine bestimmte QoS-Anforderung haben analysiert und in Richtung ganzheitlichem Netzwerkenergieverbrauch erweitert. Unsere Analyse berücksichtigt explizit den Energieverbrauch von Hardware, wie z.B. Kühlung und Datenverarbeitungseinheit, die typischerweise in theoretischen Studien der Informationstheorie vernachlässigt wird. Mithilfe von Simulationen, die es ermöglichen, die aus der Theorie bekannten Skalierungsergebnisse und die Konvergenzgeschwindigkeit der asymptotischen Ergebnisse zu bestimmen, wird weiterhin das identifizierte Skalierungsverhalten untersucht. Die wichtigsten Ergebnisse dieser Skalierungsanalyse unterstreichen den Abtausch der Skalierung von Datendurchsatz und Energieeffizienz. Daraus ergeben sich geeignete Designrichtlinien für die Skalierung von Infrastrukturknoten, z. B. Basisstationen, die zur Einsparung von Energie in praktischen Kommunikationssystemen beitragen.

In Anbetracht dieser Ergebnisse wird ein Optimierungs-Framework entwickelt, das auf die Minimierung des gesamten Energieverbrauchs im Downlink-Kanal drahtloser zellularer Netze abzielt und dabei den Energieverbrauch der Hardware berücksichtigt. Der Ansatz dieser Arbeit fokussiert die Menge der Netzelemente, die die geringste Energie verbrauchen, während sie die QoS-Anforderungen aller Nutzer im System erfüllen. Das zugrunde liegende Problem ist kombinatorischer Natur und daher im Grundsatz schwer zu lösen. Um rechnerisch effiziente Lösungen zu erhalten, werden Relaxationstechniken angewandt, die die Zielfunktion durch eine konkave Funktion annähern und alle nichtkonvexen Randbedingungen relaxieren. Dadurch wird ein verwandtes Problem konstruiert,

das durch Majorisierungs-Minimierungstechnik gelöst werden kann. Schritt für Schritt wird so ein Optimierungs-Framework vorgeschlagen, das in relativ kurzer Zeit geeignete Lösungen für das kombinatorische Problem findet. Ein großer Vorteil des hier beschriebenen Lösungsansatzes gegenüber bestehender Ansätze aus der Literatur ist die Fähigkeit, mit einer Vielzahl von verschiedenen Netzwerk-Elementen und Energieverbrauch-Modellen umzugehen. Mittels umfangreicher Simulationen wird die Leistungsfähigkeit des Frameworks dargelegt. Auf Basis des entwickelten Optimierungs-Frameworks werden Erweiterungen abgeleitet, um neue Aspekte wie die CoMP-Übertragung, eine verteilte Implementierung oder die Zwischenspeicherung von Daten bei Nutzern zu berücksichtigen. Letzteres zeigt die Erweiterung in Richtung der antizipatorischen Ressourcenplannung. Die Entscheidungen über die Ressourcenzuteilung und die Nutzer-zellen-zuweisungen basieren allerdings nicht nur auf gegenwärtigen Kanalzustandsinformationen, sondern auch auf Informationen über zukünftige Ausbreitungsbedingungen. Auf diese Weise wird das Wissen über die Mobilität von Nutzern und dem Pfadverlust verwenden, um proaktiv Nutzer und Ressourcen zuzuweisen. Ziel ist es, die Energieeinsparung in zellularen Kommunikationssystemen sukzessive zu maximieren.

Schließlich bedienen wir uns der allgemeinen Theorie des Interferenz Calculus, um rechnerisch effiziente Algorithmen zu entwickeln, mit deren Hilfe die Last an Netzwerkelementen genauer abzuschätzen ist. Bei der Entwicklung von Optimierungsmethoden ist die Kenntnis über die exakte Zellauslastung gewinnbringend, da gültige Netwerkkonfigurationen für gegebene Nutzer-zellen-zuweisungen identifiziert werden können. Diese Ergebnisse werden wiederum verwendet, um das entwickelte Framework zur Netzwerktopologie-Optimierung weiterzuentwickeln. Dadurch soll die Energieeinsparung weiter erhöht werden. Abschließend zeigen wir auf, wie das Optimierungsframework zu erweitern ist, um Konfigurationen der Netzelemente (z.B. Antennenabstrahlwinkel) zu ermitteln, die den Algorithmus zur optimierten Netztopologie in seiner Arbeit unterstützen. Genau gesagt wird der optimierte Algorithmus zur Netzwerk-Topologiekonfiguration mittels heuristischer Argumente mit der Optimierung von Zellkonfigurationen kombiniert. Hierdurch wird aufgezeigt, wie die Optimierung von Antennenabstrahlwinkeln zur Energieeinsparung beitragen kann.

Acknowledgement

This thesis would not have been possible without the support I received from a number of people in the course of my PhD studies.

First, I want to express my gratitude to Prof. Dr.-Ing. Sławomir Stańczak for giving me the opportunity to do research under his supervision. His passion for research was a source of inspiration to me. His constant support and encouragement helped me develop the skills necessary to be a good researcher.

My gratitude extends to my colleague Dr. Renato Luis Garrido Cavalcante. Renato is a phenomenal researcher and mentor with a remarkable fascination for research. I am very fortunate to have worked with him throughout my PhD studies and appreciate all the knowledge and advice he shared with me.

I also want to thank Prof. Dr.-Ing. Eduard A. Jorswieck and Dr. Gabor Fodor for dedicating their time and energy to act as a reviewer of my thesis. I also thank Prof. Dr.-Ing. Adam Wolisz for acting as chairman on the examination board.

A special thank you to all my colleagues at Fraunhofer Heinrich Hertz Institute for such a friendly and pleasant work environment. The discussions during lunch and coffee breaks helped me to develop not only on the scientific side but also in many other parts of life that matter.

Above all, my gratitude goes to my wife Jacqueline Pollakis. She was always there for me when I needed her and lifted my spirit in bad times. I am so grateful that she cut back her own interest many times and stood by my side to give me her full support.

Finally, I want to thank my parents for their unconditional love and support throughout my whole life.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

3G	Wireless communication system of the 3rd generation
3GPP	3rd Generation Partnership Project
4G	Wireless communication system of the 4th generation
5G	Wireless communication system of the 5th generation
APSM	Adaptive Projected Sub-gradient Method
AWGN	Additive White Gaussian Noise
BMWi	German Federal Ministry of Economics and Technology
CAGR	Compound Annual Growth Rate
cCZ	Centralized Cell Zooming Algorithm
CoMP	Coordinated Multi-Point transmission
D2D	Device to Device
DVB	Digital Video Broadcast
EM	Expectation Maximization
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HD	High Definition
HTP	Hotspot Test Point
ICT	Information and Communications Technology
IoT	Internet of Things
ITU	International Telecommunication Union
LP	Linear Programming problem
LTE	Long Term Evolution
M2M	Machine-to-Machine communication
MCS	Modulation and Coding Scheme
METIS	Mobile and wireless communications Enablers for the Twenty-twenty (2020) Information Society (project)
MILP	Mixed Integer Linear Program
MINLP	Mixed Integer Non-Linear Program
MIMO	Multiple Input Multiple Output
MM	Majorization Minimization algorithm
MNO	Mobile Network Operator
MTC	Machine Type Communication
NP-hard	Non-deterministic Polynomial-time hard
OFDM	Orthogonal Frequency Division Multiplexing

PC	Personal Computer
PRAUA	Proactive Resource Allocation and User Assignment
QoS	Quality of Service
RAM	Random Access Memory
RAN	Radio Access Network
RAT	Radio Access Technology
REA	Research Executive Agency
RRM	Radio Resource Management
SIC	Successive Interference Cancellation
SINR	Signal to Interference plus Noise Ratio
sMM	Sparsity-supporting MM algorithm
SNR	Signal to Noise Ratio
STP	Standard Test Point
TC	Test Case
TDMA	Time Division Multiple Access
TP	Test Point
UMTS	Universal Mobile Telecommunications System
VoLTE	Voice over LTE
WCDMA	Wideband Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access

Notation

In this work, sets are presented by calligraphic capital letters (such as $\mathcal{X}$) and the cardinality of a set is denoted by the same letter of the set in italic writing $X := |\mathcal{X}|$. The set $\mathbb{R}_+$ denotes the set of non-negative real numbers, while $\mathbb{R}_{++} := \mathbb{R}_+ \backslash \{0\}$ is the set of positive real numbers. We further define $0/0 := 0$ and $c/0 := \infty$ for $c > 0$.

Vectors are presented by bold lower case letters (such as $\mathbf{x}$), whereas matrices are written in bold upper case letters (such as $\mathbf{X}$). For a vector $\mathbf{x} \in \mathbb{R}^N$, its ith component is $x_i \in \mathbb{R}$. Similarly, the (i,j)-th component of matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ is $x_{i,j} \in \mathbb{R}$. Inequalities involving vectors, such as $\boldsymbol{x}_1 \geq \boldsymbol{x}_2$, are to be understood as component-wise inequalities.

Given a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$, we use $\tilde{\mathbf{x}} := \text{vec}(\mathbf{X}) \in \mathbb{R}^{MN}$ to denote the vector obtained by stacking the columns of $\mathbf{X}$.

Definition 1 (l_p-norm)**.** *For $p \geq 1$ the l_p-norm of a vector $\mathbf{x} \in \mathbb{R}^N$ is defined by*

$$||\mathbf{x}||_p := \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}.$$

In Chapter 3 we make use of the the cardinality function obtained from the l_p-norm definition (Definition 1) with $p = 0$. Although this l_0-norm is not a norm, since it does not satisfy all properties of a norm (it is not homogenous), we use the term "norm" as it is common practice in literature.

Definition 2 (l_0-norm)**.** *For any vector $\mathbf{x} \in \mathbb{R}^N$ and matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$, their l_0-norms $|\boldsymbol{x}|_0$ and $|\boldsymbol{X}|_0$ are equal to the number of nonzero elements of $\boldsymbol{x}$ and $\mathbf{X}$, respectively. For a scalar $x \in \mathbb{R}$, $|x|_0 := 1$ if $x \neq 0$ and $|x|_0 := 0$ otherwise.*

Further notation

In addition the following notation is used:

$\lvert \cdot \rvert$	Absolute value of a scalar or cardinality of a set
argmin	Argument of the minima
$\exists$	At least one exists
$:=$	Equal by definition

$d(X_i, X_j)$	Euclidean distance between X_i and X_j
$\forall$	For all
∇f	Gradient of a (multivariate) function f
∞	Infinity
$\lim f$	Limit of a function f
log	Logarithmic function
max	Maximum
min	Minimum
ln	Natural logarithm
$\|\| \cdot \|\|_p$	p-norm
$P[\mathcal{E}]$	Probability of event $\mathcal{E}$
$\mathbf{A}^T$	Transpose of a matrix $\mathbf{A}$

1 Introduction

1.1 Motivation and background

Since the introduction of the first cellular communication system, the wireless communication landscape has experienced numerous evolutions of technology. Especially during the last decade, novel techniques have revolutionized the way we communicate. The early systems allowed for voice services in only a few isolated regions. Over the last decades, coverage has been improved to make mobile communication services available for the general public. Data communication services were added and network capabilities were gradually improved with every new generation of mobile networks. From the early stages of GSM (Global System for Mobile Communications) to communication systems of the third (3G), fourth (4G) and the upcoming fifth (5G) generation, we have seen researchers and network designers putting large efforts into pushing the limits of wireless communication network. Thereby, the constantly increasing popularity of mobile communications lead to interesting dynamics in the counter-play of service demand and network performance. The system throughput, end-to-end latency and the supported number of users were greatly improved and novel techniques such as Orthogonal Frequency-Division Multiplexing (OFDM), multiple-input and multiple-output (MIMO) systems or beamforming helped to make a more efficient use of the wireless spectrum. The improved network performance is welcomed with open arms by the users and new services are implemented that exploit the improved Quality of Service (QoS). This effect results in the situation that the capabilities of the network are used almost to its full extend quickly. The chase for higher network performance and the demand for better QoS became somewhat like a self-fulfilling prophecy. The moment the network provides a better QoS, the demand for better QoS almost instantly rises. A popular example is video streaming over wireless networks. In the late 1990s, when it was a novelty to have internet access on the go, the network was able to provide only small downlink throughputs and video streaming was hardly possible. Later, with the improved downlink data rates of 3G, videos could be streamed over the network but the supported video resolution was low. Today, with long term evolution (LTE) deployed, we have a wireless network that supports downlink peak data rates of up to 300 Mbits-per-second enabling video streaming in high definition. This continuous increase is still ongoing with mobile network connection speeds growing 20% in 2014 [Cis15]. At the same time the growth in consumption of data over wireless networks is even more pronounced. According to [Cis15], the global mobile data traffic in 2014 was nearly 30 times the size of the entire global (wired) Internet in the year 2000 and this trend is predicted to continue over the

next years. Mobile communication has become part of many areas in our daily lives. The possibility to communicate with each other without being tied to a fixed place (by a wire) made us more mobile than ever and our society adopted the new way to communicate quickly. The study in [Qur15] highlights the continuation of this trend and predicts that the number of mobile subscribers will increase from 7.1 billion in 2014 to 9.2 billion in 2020 consuming over 30.000 PetaByte of data per month, whereby the lion's share is video data. These numbers are supported by [Cis15] predicting the key driver of mobile data growth to be video streaming (or on-demand video) with a 66% annual increase through 2019.

Besides these benefits for people connecting to wireless networks, mobile communications technology pushes open doors to a new horizon fuelling progress in every sector of our economy (industry and private life). Wireless communications is increasingly linking the digital and the physical world forming a cornerstone of the global economy. The huge potential of mobile communications has been started to be exploited by new sectors foreseeing huge changes in the years to come. Under the umbrella of the Internet of Things (IoT) it is envisioned that all devices and things are connected, thereby integrating billions of sensors and actuators into devices and communicating with them wirelessly. Such devices are found in a broad spectrum of application areas including wearable health and fitness devices, smart power readers, smart lamp poles on the street and many other sensors from manufacturing and maintenance processes. Device connection density will hit $200,000$ connections per km^2 according to [AL15] introducing never seen requirements to the communication network. Especially requirements for reduced latency, improved reliability, long battery life for devices and more consistent user bit rates become more and more important. Even in the more conservative industry sectors like production, wireless communications is emerging as a hot topic, which is usually referred to as the fourth industrial revolution (industry 4.0). Smart cities will help us in every thinkable way in our daily lives through improved self-awareness and novel services made possible by ubiquitous connectivity. In these settings high importance will be given to Machine-to-Machine (M2M) and Device-to-Device (D2D) connections whose growth will come from a range of significant verticals (smart cities, eHealth, industry 4.0 and automotive). The net result of this trend towards more devices with a huge variety of service applications is an even more drastic increase in the mobile data traffic which needs to be accounted for.

A straightforward and easy to implement method to provide better QoS to the increasing number of mobile-connected devices typically involves to deploy more cells in the service areas [Int15]. However, these ultra dense network deployments result in a massive increase in the cost for installed hardware and network operation which in turn has a negative effect on the mobile network operator's revenue. In fact, contrary to the increasing demand, the study in [Int15] has identified a slowdown in industry revenue growth with a compound annual growth rate (CAGR) of 4% in the period 2008-2014. And it is expected to decrease even further to 3.1% per annum through to 2020. It has been mentioned before, that the major drivers for this effect are the large costs involved in deploying and powering wireless

communications networks. Thus, the densification of cellular communication networks challenges the energy efficiency in such networks in a way that the drastic increase of infrastructure used to match the increasing demand is accompanied with a potentially unacceptable level of total energy consumption.

Therefore, in this thesis, we give some answers and guidelines for how to operate todays and future mobile communications networks in a more energy efficient manner. We address questions from network planning about the optimal number of cells in the network and provide insight in how to operate such networks to save energy. Thereby, we go beyond most works in literature, which usually accounts only for the energy radiated by the antennas, and we explicitly include energy consumed by hardware (e.g., coolers and circuit energy consumption). The later source of energy consumption usually accounts for a much bigger part of the total energy budget [CYZ+11].

1.2 Outline and main contributions

In **Chapter 2** we present fundamental results related to the growth behavior of wireless communication systems when the number of users grows without bounds. We start from three main characteristics of today's networks: nodes/users are assumed to be stationary, interference is treated as noise and non-dense infrastructure. The first assumption does not mean that users are not moving but rather it is meant that mobility of users is not (yet) exploited to improve the network performance. The second assumption, that interference is treated as noise in the system, relates to the fact that in current system deployments the exploitation of the interference by advanced techniques such as successive interference cancelation (SIC) or network coding is not standard. The last characteristic highlighted is that the network infrastructure is not dense but rather sparse compared to the number of users served by one infrastructure node. For networks of such kind, we provide a scaling law analysis for throughput and energy consumption of typical communication networks as a function of the number of users. Such a scaling analysis gives valuable guidelines for network designers to choose the order of the number of infrastructure nodes needed to satisfy the service demand without wasting energy. Our analysis is done for a hybrid communication network, where nodes can communicate in an ad-hoc manner as well as classical cellular communication via infrastructure nodes (base stations or cells). Using two different communication models (physical and protocol model) to capture interference between transmissions, we summarize the throughput scaling in such networks. We show that the throughput per user in classical ad hoc networks is inevitably vanishing with increasing number of users in the system and that even advanced techniques such as MIMO etc. can not change this effect in the asymptotic regime. Note, that one should not be inclined to think that those techniques are not beneficial in practice. They are able to improve the situation in the non-asymptotic regime where constants matter. One key finding is that when dropping one of the three assumptions it is possible to achieve stable throughput when accepting some trade-off. These scaling laws are evaluated by simulations and compared

to the theoretical results. Building upon the throughput scaling results, we establish a connection to the scaling of the energy required to transmit one bit of information in such a network. The novelty of our approach is to incorporate the static energy consumption that is consumed by hardware in addition to the transmit energy consumption. This energy consumption is typically neglected in information theory but has significant impact in practice. The chapter is concluded by showing that the energy spent to communicate one bit of information grows without bounds in networks where the number of users tends to infinity. *The results presented in Chapter 2 are partially published in [BPS13, CSS+14].*

Motivated by the results on the scaling of energy in wireless communication networks, we focus in **Chapter 3** on dense infrastructure deployments where the number of infrastructure nodes such as base stations and cells is large and present results for load-aware topology control. In contrast to energy efficiency studies, where one seeks to maximize the ratio of throughput and energy consumed, the goal in our study is to minimize the energy consumption for a fixed throughput. The starting point for our analysis is the observation that the peak demand is always increasing and new services and more powerful user devices with larger screens introduce higher requirements for mobile communication systems. Additionally, new types of users such as from the IoT and Machine Type Communications (MTC) have a largely varying user pattern which typically results in a mismatch of requirements and provided service for certain time frames (over-provisioning). This highly fluctuating traffic pattern of network subscribers can be exploited in order to save energy. Therefore, we turn to the question of how to better match the actual traffic demand with an already deployed cellular network. We present methods how to select network components and deactivate parts of the network that are redundant and not needed to provide the required QoS. Thus, we save energy by deactivating parts of the network. The identification of the set of network components consuming the least amount of energy while providing exactly the needed QoS to the users in the network turns out to be a nontrivial problem. We develop efficient algorithms that are able to find good solutions in reasonable time making them amenable to near-realtime implementation. In this chapter we further provide variants of our algorithm that can be applied to different types of networks. Starting with single radio access technology (single-RAT LTE) networks we also show the application to a multi radio access technology (multi-RAT) network consisting of UMTS and LTE cells of different size with overlapping coverage. Whereas these algorithms are optimizing snapshots of a network and their results are valid for a certain time period after which the algorithm will have to produce a new valid solution, we also take into account the capability of user devices to buffer data for applications like video streaming. This work stream uses ideas from anticipatory scheduling to find a suitable resource allocation schedule over time that allows for the use of a set of network elements consuming the least amount of energy over time. The proposed algorithms have a theoretical justification for their good performance which is outlined and empirically shown by simulations. Thereby, we provide insight to the performance of our algorithms in several scenarios and network deployments. It is shown that a significant amount of energy can be saved, when our algorithms are

applied in cases where there is a significant mismatch of provided and required QoS. *The results presented in Chapter 3 are partially published in [PCS12, PCS13, CPS+13b, PCS16, PS16].*

In **Chapter 4** the framework of standard interference functions is introduced to compute more accurately the cell loads in wireless communication systems. We start by showing how the framework of standard interference functions can be used to answer questions of network feasibility. When reducing the active set of network components in an uncoordinated way a network can become infeasible in the sense that some network element does not have enough resources to serve all users associated with it. The framework of standard interference functions provides a set of tools that serve as the basis for our algorithms to identify such situations in a computationally efficient manner. In addition, the developed algorithms provide indications for network planners which network elements are overloaded in a given infeasible network setting. Thus, these network elements need to be equipped with more resources in order to arrive at a stable and feasible network state. Results from literature have shown that it is optimal with respect to the transmit power consumption if a network is operated at a point where all network elements are fully loaded. Motivated by these results, we develop an algorithm for transmit power planning targeting a feasible network state with maximum load at network elements. Being based on the framework of standard interference functions, our developed algorithm is of low computational complexity and shows good performance in a realistic simulation scenarios. With the tools for accurate cell load calculation at hand, we enhance our topology control algorithms for energy savings from Chapter 3. This combination resolves some shortcomings of our topology control framework leading to conservative results that did not exploit the full energy savings potential. The combination of interference calculus with the topology control optimization leads to efficient heuristics that enable feasible network topologies saving even more energy. This chapter is concluded with an extension of the studied approach towards the identification of beneficial network element configurations. Based on intuitive observations we derive an alternating algorithm that selects a configuration for all network elements in the network which support the developed energy saving algorithms. *The discussed results of this chapter were in part published in [CPS13a, CPS14].*

We summarize the main findings and conclusions in **Chapter 5**.

Further results that are not included in this thesis

During the course of my research I came across a variety of different fields. For the reason of consistency of this thesis some material has not been included. Nevertheless for completeness following is a brief description of the publications that are not included in this thesis.

- Our publication [KPS11] addresses the problem of multicast transmit beamforming

for the multiple antenna multicast fading channel. We propose to use coded transmission at the application layer over a number of channel realizations and design a communication scheme that achieves good scaling of the achievable rate for the increasing number of users. Whereas to other multicast schemes without coding do have an achievable rate that does decrease to zero, the proposed scheme has a non-vanishing achievable rate. Therefore, we can improve the rate compared to so-called max-min transmit beamforming schemes. The publication presents algorithms to solve the involved beamforming problem and shows its performance by means of empirical evaluation.

- In [KPS12] we address beamforming problems for multi-group multicast communication. In such communication scenarios the performance is usually limited by the weakest link within a multicast group and by the interference from other transmissions. Therefore, we propose a communication scheme, in which application layer coding over a number of channel realizations is used. We formalize a transmit beamforming problem with the goal of maximizing the weighted sum of rates achieved in each group. It is shown that the optimal transmit strategy depends only on the current channel realization. Furthermore, we show that the utility-based power control framework can be generalized to the case of multi-group multicast, which allows for the application of iterative beamforming algorithms. Building upon this framework, we propose iterative beamforming algorithms which can be applied in scenarios both with and without additional coding at the application layer.
- In [PCSP12] we address the problem of reconstructing interference patterns in UMTS and LTE networks. A novel cognitive interference identification technique is presented which uses a priori system knowledge, limited user information and sparse pathloss and interference measurements for kernel-based machine learning techniques. It is shown that the obtained information can be used as an input to multi-RAT optimization procedures aiming at energy efficient network configurations.

Copyright information

Acknowledgements

I would like to acknowledge some projects that were relevant and helped me with viable insights into the broad field of wireless communications. The discussions with project partners helped me sharpen the results and make them more suitable for practical implementation. I appreciate to have gotten an insight in practical problems in industry.

- GreenNets: The research leading to these results has received funding from the European Union's Seventh Framework Programme managed by REA - Research Executive Agency (FP7/2007-2013) under grant no 286822.
- ComGreen: This work has been partly supported by the framework of the research project ComGreen under the grant-number 01ME11010, which is funded by the German Federal Ministry of Economics and Technology (BMWi).
- METIS: Part of this work has been performed in the framework of the FP7 project ICT-317669 METIS, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in METIS, although the views expressed are those of the authors and do not necessarily represent the project.

2 Scaling of Infrastructure for Large Wireless Networks

In this chapter, we give an insight in the growth behavior of wireless communication systems when the number of users tends to infinity. For this purpose we make a scaling law analysis of various performance metrics (such as throughput and energy consumption) as a function of the number of users having some QoS requirements. Our analysis concerns *hybrid wireless networks* where ad hoc communication is supported by infrastructure nodes[1]. Such hybrid networks include communication protocols in the sense of classical cellular communication, i.e., when the used communication protocol restricts to cellular only, as well as future communication protocols of 5G where cellular communication is for example combined with direct D2D communication in a multi-hop fashion.

In the context of network planning, such a scaling analysis provides valuable guidelines for, among others, the following questions:

- How fast should the number of infrastructure nodes grow with the number of users so that the provision of the desired performance to every user is feasible?
- How will the network energy consumption scale for a certain growth in the number of infrastructure nodes?

To address these questions, we present an information-theoretic framework that assesses the relationship between network growth and energy consumption in infrastructure-based wireless communication networks. This framework will help to shed light on the question of how the energy that is needed to operate an infrastructure-based wireless communication system given a certain traffic demand is scaling when scaling up the network. The scaling law analysis addresses the problem of the wireless infrastructure growth rate: As the number of users and traffic demand increase, the question is how fast the number of infrastructure nodes needs to grow (with the number of users) so that the provision of the desired performance to every user is feasible.

The scaling law analysis does not address the question of how many infrastructure nodes are necessary (and sufficient) for achieving an optimal tradeoff between throughput and energy consumed per transmitted information bit. This is a longstanding and fundamental problem in information theory that is far from being solved. In fact, even the information-theoretic

[1] In this chapter we refer to infrastructure nodes as nodes that are interconnected via a high capacity (wired) backbone network and are no source of data; thus they only act as relays. In classical cellular networks such infrastructure nodes are called base stations.

capacity of infrastructure-based wireless communication networks, such as cellular wireless networks, has been an open problem for more than fifty years. This is despite the fact that information theory makes a lot of assumptions such as neglecting dynamicity, energy consumption of hardware, bursty traffic and signaling overhead. It is currently not possible to exactly estimate the minimum average energy that is required for reliable wireless communication. More precisely, for an established network topology with a fixed number of infrastructure nodes (e.g. base stations) and wireless users attempting to transmit their data to the infrastructure nodes at some given rates over a common bandwidth, the minimum average energy for transmission of one information bit at some given probability of error is not known.

Scaling law analysis is usually carried out in *random wireless networks* to arrive at generally valid statements that are not limited to a particular network deployment. In such networks, each node has a randomly chosen destination. Moreover, positions of the users over some area in the plane are random variables drawn from a predefined (two-dimensional) stochastic process. The performance measure of interest is therefore a random variable whose probability distribution depends on this stochastic process. An immediate consequence of this is that the scaling laws of random wireless networks must be analyzed using tools from probability theory. The scaling law becomes increasingly accurate when the number of users grows and is asymptotically exact; therefore the results presented in this chapter are of *asymptotic nature*.

We proceed with a summary of existing results in the area of scaling laws for throughput capacity in wireless networks. After reviewing fundamental results about the throughput capacity of random networks and the tradeoffs between throughput capacity, mobility, and delay, we continue presenting our main contribution to the topic of hybrid wireless networks.

2.1 Background and contribution

2.1.1 Notation

We use the Big-O notation to classify the behavior of functions in the asymptotic regime.

Given two functions $f, g : \mathbb{N} \to \mathbb{R}$ we say that

- f is *asymptotically upper bounded* by g for

$$f \in \mathcal{O}(g) \Leftrightarrow \lim_{n\to\infty} \frac{f(n)}{g(n)} < \infty.$$

- f is *asymptotically lower bounded* by g for

$$f \in \Omega(g) \Leftrightarrow 0 < \liminf_{n\to\infty} \frac{f(n)}{g(n)} \leq \infty.$$

- g is a *strict asymptotic upper bounded* for f for

$$f \in o(g) \Leftrightarrow 0 < \limsup_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

- g is a *strict asymptotic lower bound* for f for

$$f \in \omega(g) \Leftrightarrow \lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty.$$

- g is an *asymptotically tight bound* for f for

$$f \in \Theta(g) \Leftrightarrow f \in \mathcal{O}(g) \text{ and } g \in \mathcal{O}(f).$$

2.1.2 Related work

The landmark paper [GK00] analyzes the information-theoretic capacity in wireless ad hoc networks (infrastructure-less networks) in the asymptotic regime (i.e. large-scale wireless networks). The main result of the study is that the throughput capacity of an *arbitrary network* with n wireless nodes[2] communicating in an ad-hoc fashion is *at most* $\Theta\left(\frac{c}{\sqrt{n}}\right)$, where $c > 0$ is a constant that depends on system parameters like area, bandwidth, power constraints and the number of antenna elements. In arbitrary networks nodes are optimally placed and the communication has optimal characteristics (traffic pattern and transmission range). For random networks the order of the throughput capacity provided to each node is only $\Theta\left(\frac{c}{\sqrt{n \log n}}\right)$. These results give rise to the following fact

Fact 1 (Vanishing Throughput)**.** *The throughput per node in a wireless ad-hoc network scales at most as* $\Theta\left(\frac{c}{\sqrt{n}}\right)$ *for some positive constant* $c \in \mathbb{R}_{++}$*, and thus vanishes as the number of nodes* n *increases to infinity.*

The analysis of [GK00] together with subsequent works (e.g. see [GK03]) show that the main limiting factor in the asymptotic regime is interference which is the result of three underlaying assumptions made by [GK00].

Assumption 1. *The considered wireless ad-hoc network is static and wireless nodes are stationary (do not move).*

Assumption 2. *All wireless nodes are equipped with standard air interfaces that treat residual interference as noise.*

Assumption 3. *Communication is performed in a wireless multi-hop manner using only other wireless nodes to relay the traffic. In other words, there is no infrastructure support.*

With the goal to stabilize the per node throughput (i.e. $\Theta(1)$) in the asymptotic regime, the subsequent work has loosened Assumption 1-3.

[2] Users are usually termed wireless nodes. We adopt this notion.

The effect of having mobile users in a wireless ad hoc network, and therefore lifting Assumption 1, is analyzed in [GT02, NM05]. The work in [GT02] analyzes an infrastructure-less wireless network with mobile nodes and interference treated as noise. It is shown that a stable per node throughput can be achieved (i.e. $\Theta(1)$) when exploiting the nodes mobility. A two-hop relaying scheme that allows transmissions only when nodes are close to each other is presented that achieves $\Theta(1)$ for an i.i.d mobility model of each node. However, the improvement in throughput comes at the cost of potential excessive/unbounded delay. The throughput-delay tradeoff is further studied in [NM05], where a capacity-achieving scheme is presented that reduces the delay by sending redundant packets along multiple paths to the destination. However, it was also shown that when exploiting mobility there is a necessary delay-rate tradeoff: $\frac{\text{delay}}{\text{rate}} \geq \mathcal{O}(n)$. We conclude that exploiting mobility of nodes may be a viable method to improve the throughput but unfortunately many wireless applications have strict delay constraints.

By dropping Assumption 2, we can make nodes exploit, shape, or reject interference through advanced multi-user transmission and reception techniques. [GK03] presents a scheme achieving $\Theta(1)$ by employing zero-forcing beamforming under perfect channel knowledge at all nodes. Using a minimum separation distance and coherent multistage relaying with interference subtraction is proposed in [XK04] to improve the throughput scaling. [OLT07] also shows the benefits of hierarchical cooperation and other multi-user communication strategies motivated by information-theoretic results on relay, multiple-access and broadcast channels. The benefit of interference-shaping techniques such as interference alignment to support non-vanishing throughputs is demonstrated in [CJ08]. Such schemes usually assume global channel knowledge which is very restrictive for today's communication networks. We can summarize, that results have only been obtained for special networks with optimal conditioning (arbitrary networks) or they are not easy to be implemented due to obstacles in real systems, like high signaling and coordination requirements or the lack of perfect synchronization. It is worth noting that even if such interference treatment schemes cannot further improve the per-node throughput in the asymptotic regime, they can significantly improve the throughput for a finite number of users (see Figure 2.1). Even if such physical limits do exist and sophisticated strategies like the hierarchical cooperation cannot further improve the per-node throughput in the scaling limit sense, these strategies generally could be considerably beneficial in networks of any finite size.

The remaining pillar to overcome the problem of vanishing throughputs is to use an underlaying infrastructure, i.e. base stations (Assumption 3). In fact, it has been shown in [KT03, LLT03, AK04, ZdV05] that in hybrid networks, where $m(n)$ infrastructure nodes support the ad hoc communication of n wireless nodes, the throughput capacity can be significantly improved. In particular, for random hybrid networks it was shown in [KT03] that in a setting where the number of wireless nodes per infrastructure node is bounded above and all nodes use a fixed transmission range, the throughput capacity can be improved to $\Theta\left(\frac{1}{\log n}\right)$ from $\Theta\left(\frac{1}{\sqrt{n \log n}}\right)$ for pure ad-hoc networks [GK00]. Furthermore, it was proven that in such scenarios, where no power control is employed a per node

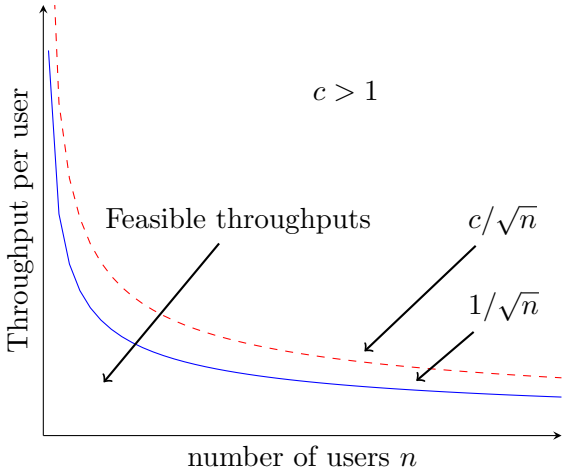


Figure 2.1: The throughput capacity in wireless networks with standard air interfaces scales with $c\sqrt{1/n}$, where c is a constant depending on system parameters and n denotes the number of nodes. Relaying and employing multiple antenna systems cannot improve the (asymptotic) throughput order but they can improve the constant c to shift the throughput capacity curve upward (dashed line).

throughput of $\Theta(1)$ cannot be obtained. In turn, [AK04] shows that the throughput of a constant fraction of wireless nodes in a random hybrid network can be improved from $\Theta\left(\frac{1}{\log n}\right)$ to $\Theta(1)$ with high probability when allowing for power control and using a slotted operation. [LLT03] considers arbitrary hybrid networks where wireless nodes are randomly placed but infrastructure nodes are arbitrarily placed according to a predefined pattern. It is shown for two routing strategies that if the number of infrastructure nodes (base stations) scales at least as $\Omega(\sqrt{n})$, then adding new base stations results in a linear throughput increase compared with the pure ad-hoc communication. Even more interestingly, for an infrastructure node scaling of $\Theta(n)$ a constant per node throughput of $\Theta(1)$ for all nodes is achievable.

Therefore, the following sufficient condition for non-vanishing throughput in arbitrary hybrid networks can be derived.

Fact 2 (Stable Throughput)**.** *Non-vanishing throughput for all wireless nodes is possible provided that the number of infrastructure nodes scales linearly with the number of wireless nodes* $m \in \Theta(n) \rightarrow \lambda(m,n) \in \Theta(1)$*. In such a case, the throughput capacity per wireless node is said to be of order* $\Theta(1)$*, which means that the per-user throughput does not vanish asymptotically because it is of the same order as any constant function.*

In other words, for the per node throughput to be non-vanishing the number of infrastructure nodes should be as high as possible; optimally the same number of infrastructure nodes as wireless nodes. The number of base stations needs to grow linearly with the number of users (personal base stations) and power control is needed.

A concise overview of scaling regimes according to some infrastructure node scaling in

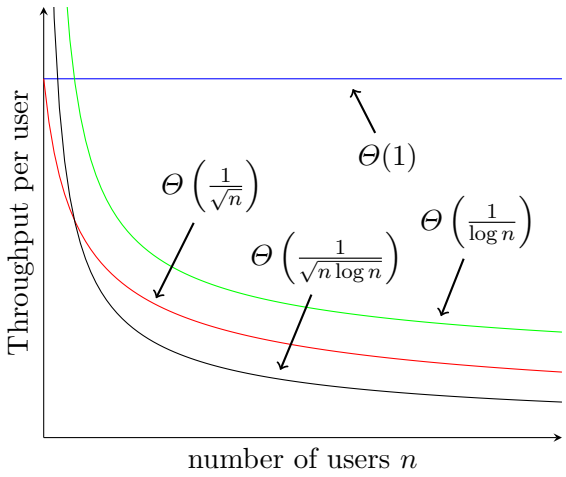


Figure 2.2: Any constant function belongs to $\Theta(1)$, while the remaining functions tend to zero at different rates and are therefore of different orders. In particular, we have $o\left(\frac{1}{\sqrt{n \log n}}\right) \subset o\left(\frac{1}{\sqrt{n}}\right) \subset o\left(\frac{1}{\log n}\right)$.

arbitrary hybrid networks is presented in [ZdV05]. The linear scaling of throughput above some threshold of infrastructure node scaling is confirmed. In addition, an upper bound for the throughput scaling is presented originating from the used scheme to form source destination pairs. In more detail, it is shown, that with some positive probability there exists at least one node which is selected as an intended receiver by $\Theta(\log n)$ wireless nodes leading to an upper bound in throughput of $\Theta(\frac{1}{\log n})$. In [ZdV05] it was shown that without power control the asymptotic throughput capacity can be separated in three different scaling *regimes* depending on the scaling of infrastructure nodes $m(n)$. More precise, for an infrastructure node scaling of $m \in \mathcal{O}\left(\sqrt{\frac{n}{\log n}}\right)$ the per user throughput is $\lambda \in \Theta\left(\frac{1}{\sqrt{n \log n}}\right)$, for $m \in \omega\left(\sqrt{\frac{n}{\log n}}\right)$ and $m \in \mathcal{O}\left(\frac{n}{\log n}\right)$ it is $\lambda \in \Theta\left(\frac{m}{n}\right)$ and for $m \in \omega\left(\frac{n}{\log n}\right)$ it is $\lambda \in \Theta\left(\frac{1}{\log n}\right)$. The different scaling regimes are illustrated in Figure 2.2. The idea of densely deployed infrastructure nodes has been adopted by research activities for future wireless networks to enhance the throughput and energy efficiency of cellular networks at relatively low operational costs [HKD11]. It is envisioned that small and low cost base stations form small cells to provide access by using short-distance links [OBB+14].

Accompanied with the benefits of a high density of future wireless networks is the increase in capital and operational costs. It is known that a significant part of the costs for operating a wireless communication network lies in the energy costs for transmission energy and operation of infrastructure [CZB+10]. Therefore, an important question is the following: *How dense do base stations (and other infrastructure nodes having access to a wired backhaul network) need to be deployed in order to ensure the desired throughput-cost tradeoff?*

We are interested in the energy efficiency in such wireless communication systems with infrastructure support. Classically, the energy efficiency of communication systems with

finite energy constraints is understood to be the amount of energy needed to transmit a finite number of bits under a given error probability. In literature there are two different approaches towards investigating the energy efficiency. The work [Gal87] and also [GW02, p.15] are interested in the maximum rate per energy and refer to the energy efficiency as the capacity per unit energy (bits per second per Joule). For information-theoretic studies a more suitable notion has been established, namely the energy per one bit. It is defined as the energy required to reliably transmit one bit of information at some rate [Ver02]. Using this notion of energy efficiency the minimum energy per bit for the two terminal additive white Gaussian noise (AWGN) channel with a relay is still an open problem. A lower bound on the transmit energy per bit required in the broadcasting case for dense and extended large networks is established in [JKV11]. [KNG11] studies the capacity per unit energy of large wireless networks. For Gaussian fading channels with pathloss exponent α it is shown that in dense networks the capacity per unit area scales as $\Theta(n)$ and $\Theta(n)$ for $2 \leq \alpha \leq 3$ and $\alpha \geq 3$ respectively.

The main point which distinguishes our work from other studies in literature is that we not only consider the energy radiated by the transmitters but also incorporate the energy consumption of hardware (e.g. for cooling, power amplifiers or signal processing). With this compound notion of energy consumption we derive the energy required to reliably transmit one bit of information in a hybrid ad hoc network. Along those lines our work will contribute to a better understanding of the connection between approximate throughput scaling and energy efficiency of large wireless networks employing infrastructure nodes. Moreover, it will enable us to give justified suggestions for large wireless communication networks that will lead to increased energy efficiency. Therefore, the energy required to operate a scheme which achieves the throughput scaling presented in [ZdV05] is investigated.

2.1.3 Main contribution

Our main contribution to the field of scaling laws for wireless communication networks is twofold.

1. First, we use existing results from literature considering infrastructure nodes (allowing to model cellular networks) [KT03, LLT03, AK04, ZdV05] and evaluate their throughput scaling laws. We provide a simulation environment that allows to verify the scaling results predicted by theory and to assess the speed of convergence of the asymptotic results.

2. The second cornerstone of this chapter is the derivation of growth rates for energy-per-bit for practical communication schemes achieving the throughput scaling laws from literature. The novelty of our study is the incorporation of static energy consumption that is consumed by hardware in addition to the transmit energy consumption usually included as only source of energy consumption in other studies.

In more detail, our first observation concerns the average per connection throughput in wireless communication networks with standard air interfaces where no advanced techniques

such as cooperative systems, massive MIMO and interference cancellation techniques are used, and the residual interference is treated as noise. Such networks have been identified to have vanishing throughput in the asymptotic regime. Our analysis and simulation results point out the following fact.

Result 1. *The results of vanishing throughput in random hybrid wireless communication systems for the asymptotic regime are already observable for a finite number of users. Thus the problem of diminishing per node throughput is a problem to be addressed even for smaller number of nodes.*

When looking at energy consumption we are not only accounting for transmit energy but also for static energy. Usually information theory neglects static energy consumption only focusing on transmit energy and sometimes processing energy. We extend existing scaling results by taking into account the static energy consumed by hardware. We try to give an answer to the question of how the total energy-per bit spent on both transmission and hardware scales with the number of wireless nodes for different infrastructure growth rates. The energy consumption and energy per bit results are not of general applicability, but we conjecture that they provide scaling laws for specific communication schemes currently used in cellular networks such as GSM, UMTS, LTE and future 5G. In particular, we concentrate on a particular communication scheme presented in [ZdV05] that is justified by its close relation to practical systems. It consists of a cellular network model where users communicate with their closest infrastructure node and a high capacity backhaul network is used interconnecting infrastructure nodes. For such networks we have our second main result.

Result 2. *The energy-per-bit $E_b(m,n)$ of random hybrid networks with n wireless nodes and $m(n)$ infrastructure nodes operating under a practical communication scheme as outlined in [ZdV05] satisfies*

$$E_b(m,n) \in \begin{cases} \Omega\left(\sqrt{n \log n}\ h_M\right) & m(n) \in o\left(\sqrt{\frac{n}{\log n}}\right), \\ \Omega\left(\sqrt{n \log n}\ h_M + h_{BS}\right) & m(n) \in \Theta\left(\sqrt{\frac{n}{\log n}}\right), \\ \Omega\left(\frac{n}{m} h_M + h_{BS}\right) & m(n) \in \omega\left(\sqrt{\frac{n}{\log n}}\right), \mathcal{O}\left(\frac{n}{\log n}\right), \\ \Omega\left(\log n\ h_M + \log n\ \frac{m}{n} h_{BS}\right) & m(n) \in \omega\left(\frac{n}{\log n}\right), \end{cases} \tag{2.1}$$

with h_M and h_{BS} the static energy consumption of wireless nodes and infrastructure nodes, respectively.

Results from literature not taking into account the hardware energy consumption of a network have shown that the transmission energy in a hybrid network operating under a cellular communication scheme has to decrease in order to implement an acceptable level of interference. When incorporating the energy consumption of hardware, the total energy consumption increases whereas the transmit throughput is at most stable resulting in a net increase of the energy-per-bit scaling.

Result 3. *In the asymptotic regime (wireless nodes n tend to infinity), the energy-per-bit scaling of a random hybrid network operating under a cellular communication scheme increases without bounds.*

2.2 System model

In the following, we detail on the system model that underlies our scaling analysis. We first describe the general setup and later detail on the communication scheme that achieves the throughput scaling.

We consider a *hybrid network* where ad-hoc communication is supported by infrastructure nodes. The communication area is a disc of area size A. In our study we scale space and thus consider the area size to be fixed. Without loss of generality we use $A = 1\text{m}^2$. Such a network is referred to as a *dense network* as opposed to an extended network where the area A scales with the number of wireless nodes n (representing user). We place n wireless nodes uniformly at random in the area and refer to the k-th node as X_k, $k = 1, \ldots, n$. Each wireless node is both source and destination resulting in n communication pairs. Without loss of generality, we assume that X_k transmits to X_{k+1} for $k \leq n-1$ and X_n transmits to X_1. Each node transmits or receives data at an arbitrary rate of W bits per second. Additionally, there are $m(n)$ infrastructure nodes (representing base stations) placed *arbitrarily*[3] in the area, that support the communication of the wireless nodes. These infrastructure nodes are interconnected by a wired backhaul network.

Assumption 4 (Backhaul Network)**.** *The backhaul network consists of wired links with high capacity and zero delay.*

Infrastructure nodes use the same channel to communicate with wireless nodes as the wireless nodes itself. For a concise naming we refer to the k-th infrastructure node as X_k with $k = n+1, \ldots, n+m$. In contrast to wireless nodes, infrastructure nodes are neither source nor destination of data and thus act only as relays for wireless nodes.

2.2.1 Interference model

The wireless channel is shared between all transmissions from or to wireless nodes. The transmission from one infrastructure node to another is realized in a wired fashion such that it is subject to no capacity and delay constraint. We adopt the *protocol model* and the *physical model* from [GK00].

The *Protocol Model* employs a common *transmission radius* $r(m, n)$ and uses some guard distance $\Delta > 0$. Then node X_i can successfully transmit to node X_j if and only if the following two conditions hold:

$$d(X_i, X_j) \leq r(m, n) \tag{2.2}$$

[3]The position of infrastructure nodes could, e.g. be a pre-defined pattern, independent of the realization of the wireless node placement.

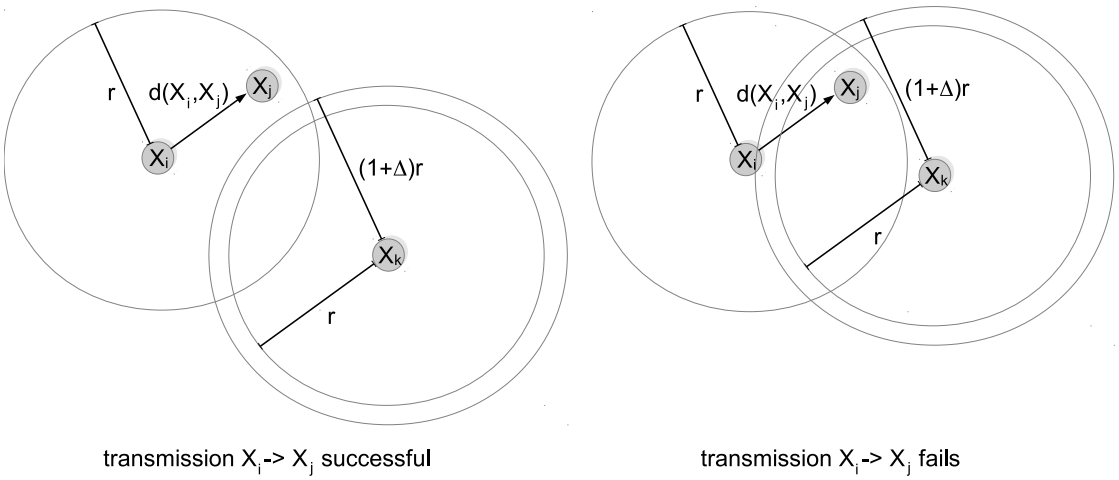


Figure 2.3: Illustration of the definition of the Protocol Model.

and

$$d(X_k, X_j) > (1 + \Delta)r(m, n), \tag{2.3}$$

where X_k is any other node simultaneously transmitting as X_i; $d(X_i, X_j)$ denotes the Euclidean distance between X_i and X_j (cf. Figure 2.3).

In the *Physical Model* we suppose that all nodes choose a common power level[4] P and the channel follows a distance dependent path loss model with path loss exponent $\alpha > 2$; i.e. the received power at distance d is given by

$$P^{\mathrm{RX}}(d) = \frac{P}{d^{\alpha}}. \tag{2.4}$$

Then the transmission from node X_i to node X_j is successful if and only if

$$\frac{\frac{P}{d(X_i,X_j)^{\alpha}}}{N_0 + \sum_{k \in \mathcal{T}} \frac{P}{d(X_k,X_j)^{\alpha}}} \geq \beta, \tag{2.5}$$

where $\mathcal{T}$ is the set of nodes transmitting at the same time, N_0 is noise power and $\beta > 0$ is some constant specifying the SINR requirement at each node (cf. Figure 2.4).

The models are interchangeable under certain conditions. More precisely, the results for the protocol model can be carried over to the physical model by choosing Δ to satisfy (cf. [GK00])

$$(1 + \Delta)^2 > \left(2\left(c\beta\left(9 + \frac{3}{\alpha - 1} + \frac{6}{\alpha - 2}\right)\right)^{\frac{1}{\alpha}} - 1\right)^2. \tag{2.6}$$

for sufficiently large P and a positive constant c not depending on n or m. If a transmission is successful, the source node can transmit W bits per second to its corresponding receiving node. For simplicity we use $W = 1$. The data may be transmitted directly to its destination

[4]This assumption will simplify equations in the following but is no limitation to the approach.

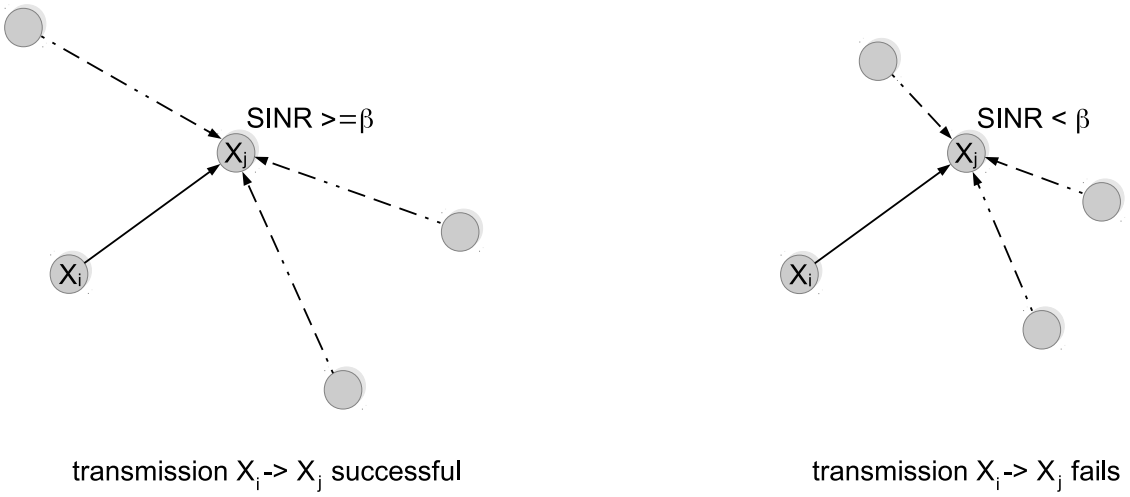


Figure 2.4: Illustration of the definition of the Physical Model.

node or in a multi-hop fashion, routing the data message through several wireless or infrastructure nodes.

2.2.2 Definition of throughput and energy-per-bit

For any communication network, including ours, there exists a feasible throughput. The following definitions have also been used by [GK00].

Definition 3 (Feasible Throughput)**.** *Let n denote the number of wireless nodes in the system. A per-node throughput of $\lambda(n)$ is said to be feasible if there exists a communication scheme that allows, for each node, to transport $\lambda(n)$ bits per second on average to its destination node.*

The scaling analysis carried out in this study is done for a class of random networks where wireless nodes are randomly located according to some given distribution; i.e., independently and uniformly at random. Source-destination communication pairs are as well chosen from a given random distribution. Hence, probabilistic tools are used to analyze such networks. In particular we introduce the notation of the throughput capacity scaling.

Definition 4 (Throughput capacity scaling [GK00])**.** *The throughput capacity is of order $\Theta(f)$ (also written as $\lambda \in \Theta(f)$) if there are (deterministic) constants $c > 0$ and $c' < \infty$ such that*

$$\lim_{n\to\infty} P[\lambda(n) = cf(n) \text{ is feasible}] = 1 \tag{2.7}$$

$$\liminf_{n\to\infty} P[\lambda(n) = c'f(n) \text{ is feasible}] < 1. \tag{2.8}$$

So, in words, we have throughput capacity scaling of order $\lambda \in \Theta(f)$ if $\lambda(n)$ is equal to $f(n)$ up to a multiplicative constant with high probability provided that the number of users is sufficiently large.

In Section 2.2.1 we will elaborate on our communication scheme that is used to achieve a particular throughput scaling and derive the corresponding energy-per-bit scaling. This quantity measures the amount of energy that is required to reliably communicate one bit of information at a certain transmission rate. In contrast to most work in literature, in addition to the transmit energy radiated by antennas we account for the energy consumption of hardware that is consumed as long as the node is turned on. For this purpose we adopt a simple, but commonly used energy consumption model [ABG+10]. The energy consumed by hardware components at infrastructure nodes is denoted by $h_{\mathrm{BS}} \in \mathbb{R}_+$ whereas wireless nodes have a static hardware energy consumption of $h_{\mathrm{M}} \in \mathbb{R}_+$. The chosen transmit power at node X_k is denoted by $P_k^{\mathrm{TX}}(m,n) \in \mathbb{R}_{++}$. This gives us the average total energy consumption of the network.

Definition 5 (Aggregate Network Energy Consumption)**.** *Given a particular communication scheme, the aggregate total energy consumption $E_\Sigma(m,n)$ is given by*

$$E_\Sigma(m,n) = \sum_{k=1}^{n} \left(\tau_k P_k^{TX}(m,n) + h_M\right) + \sum_{k=n+1}^{m+n} \left(\tau_k P_k^{TX}(m,n) + h_{BS}\right), \tag{2.9}$$

where τ_k is the average fraction of time that node k is transmitting.

Now, for a particular communication scheme achieving a total throughput $T(m,n)$ for n wireless nodes with the aggregate energy consumption $E_\Sigma(m,n)$, we say that the energy-per-bit scaling of

$$E_b(m,n) = \frac{E_\Sigma(m,n)}{T(m,n)} \tag{2.10}$$

is feasible.

Definition 6 (Achievable Energy-per-Bit)**.** *An energy-per-bit scaling of order $E_b(m,n) \in \Theta(f(m,n))$ is achievable if there is a (deterministic) constant $c > 0$ such that*

$$\lim_{n\to\infty} P[E_b(m,n) = cf(m,n) \text{ is feasible}] = 1. \tag{2.11}$$

Definition 7 (Energy-per-Bit scaling)**.** *The energy-per-bit scaling is of order $E_b(m,n) \in \Theta(f(m,n))$ if there are (deterministic) constants $c < \infty$ and $c' > 0$ such that*

$$\lim_{n\to\infty} P[E_b(m,n) = cf(n) \text{ is feasible}] = 1 \tag{2.12}$$

$$\liminf_{n\to\infty} P[E_b(m,n) = c'f(n) \text{ is feasible}] < 1. \tag{2.13}$$

2.3 Throughput scaling

In the following we review the results of [ZdV05] on the throughput order capacity under the protocol model which will serve as a basis for the derivation of the energy-per-bit scaling behavior.

The throughput capacity per node in a pure ad hoc network under a random traffic pattern exhibiting no spatial locality scales with $\Theta\left(\frac{1}{\sqrt{n \log n}}\right)$ [GK00]. It has been lined out that placing infrastructure nodes can be beneficial for the throughput order capacity. To this end, three different regimes for the infrastructure node scaling have been identified, i.e. the number of infrastructure nodes $m(n)$ deployed depending on the number of wireless nodes n. In order to refer to the three different scaling regimes we use the following definition.

Definition 8. *Infrastructure node scaling. The number of infrastructure nodes m scales with the number of wireless nodes n according to*

$$m(n) = \Theta\left(\left(\frac{n}{\log n}\right)^b\right), \tag{2.14}$$

where b is a constant determining the regime according to

$$b = \begin{cases} \left(-\infty, \frac{1}{2}\right] & \textit{regime i)} \\ \left(\frac{1}{2}, 1\right] & \textit{regime ii)} \\ (1, \infty) & \textit{regime iii).} \end{cases} \tag{2.15}$$

Now the following corollary is an immediate consequence of [ZdV05, Theorem 2].

Corollary 1. *The aggregate system throughput in a random hybrid network under the protocol model satisfies*

$$T(m,n) \in \begin{cases} \mathcal{O}\left(\sqrt{\frac{n}{\log n}}\right) & b \leq \frac{1}{2} \\ \mathcal{O}\left(\left(\frac{n}{\log n}\right)^b\right) & b \in \left(\frac{1}{2}, 1\right] \\ \mathcal{O}\left(\frac{n}{\log n}\right) & b > 1. \end{cases} \tag{2.16}$$

Proof. The per node throughput of a random hybrid network is of order

$$\lambda(m,n) \in \begin{cases} \mathcal{O}\left(\frac{1}{\sqrt{n \log n}}\right) & \text{regime i)} \\ \mathcal{O}\left(\frac{m}{n}\right) & \text{regime ii)} \\ \mathcal{O}\left(\frac{1}{\log n}\right) & \text{regime iii)} \end{cases} \tag{2.17}$$

in [ZdV05, Theorem 2]. The result now follows with Definition 8 for the number of

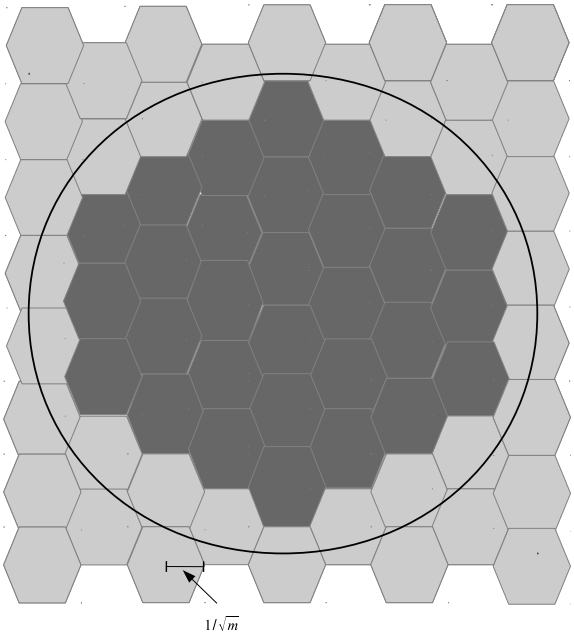


Figure 2.5: System area and regular base station placement with radius $\frac{1}{\sqrt{m}}$.

infrastructure nodes m and having n wireless nodes in the system with a per node throughput of (2.17). □

In the following, we briefly sketch how the above throughput scaling laws can be achieved in the different regimes. The schemes presented below are just examples and it is not clear whether there are other schemes achieving the same throughput capacity scaling. Note that these capacity scaling laws are for a certain class of wireless networks and not of general nature.

In *regime i)* the per node throughput is equivalent to the throughput order capacity of pure ad-hoc networks proven by [GK00] and hence a communication scheme using ad-hoc communication only can achieve $\lambda(m,n) \in \mathcal{O}\left(\frac{1}{\sqrt{n \log n}}\right)$.

A scheme which achieves the capacity scaling in *regime ii)* is as follows. The communication area is divided by a hexagonal tessellation with a cell radius of $\rho(m,n) = \frac{1}{\sqrt{m}}$ (see Figure 2.5). Each cell has an infrastructure node in its center and each wireless node is associated with the infrastructure node closest to it. Transmission time slots are divided into *uplink* phase and *downlink* phase. The uplink phase is reserved for wireless nodes transmitting to associated infrastructure nodes. In the downlink phase infrastructure nodes transmit to wireless nodes in corresponding cells. In line with the protocol model, a guard interval Δ is chosen which translates to a further division of the cell activity. This can be referred to as the *cell activation pattern* and a practical implementation is a conventional frequency reuse pattern (note that the activation pattern can equivalently be implemented by using different sub-channels)[5]. See Figure 2.6 for a reuse 3 example which is also used in our

[5] More generally, as described in [GK00], the activation pattern corresponds to a *coloring* of the graph modeling the neighboring cell structure using a constant number of colors.

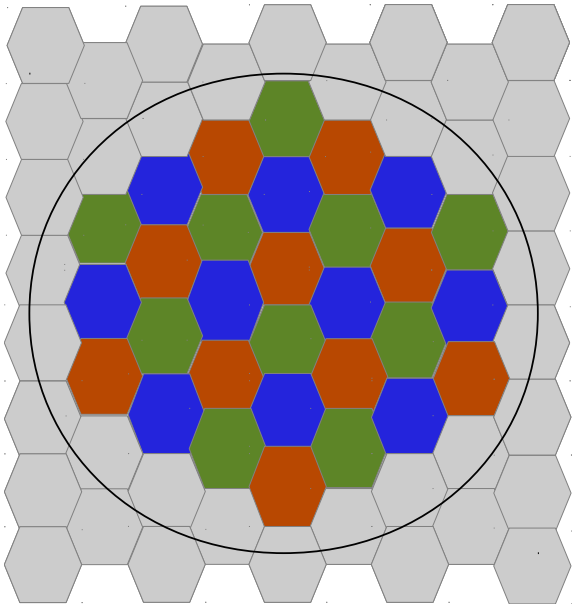


Figure 2.6: Cell activation (or frequency reuse) pattern using three different activation slots (or frequencies).

simulations and which implements a guard distance of $\Delta = 1$. Additionally, infrastructure nodes implement Round Robin to serve their wireless nodes in both phases. Messages are routed with two wireless hops, i.e. wireless nodes transmit to the closest infrastructure node that forward the message to the infrastructure node closest to the destination node (using its wired connection). The infrastructure node sends the message to its final destination node wirelessly. To ensure connectivity of all wireless nodes a transmission radius r is used for the protocol model which satisfies (cf. [ZdV05, Proposition 1])

$$r(m,n) \in \begin{cases} \Omega\left(\sqrt{\frac{\log n}{n}}\right) & \text{if } m \in o\left(\frac{n}{\log n}\right) \\ \Omega\left(\frac{1}{\sqrt{m}}\right) & \text{if } m \in \Omega\left(\frac{n}{\log n}\right). \end{cases} \tag{2.18}$$

In order to assure that each wireless node is able to communicate directly with its closest infrastructure node we use a transmission radius of (cf. [ZdV05, Fact 2])

$$r \sim \frac{1}{\sqrt{m}}. \tag{2.19}$$

Along the same line we have to specify a transmission power P^{TX} for the physical model. It is chosen as

$$P^{\mathrm{TX}} \sim \frac{1}{\sqrt{m}^{\alpha}} \tag{2.20}$$

at each node. The proportionality factor depends on the SINR requirement β and on the frequency reuse pattern that is employed, which in turn specifies the guard distance parameter Δ. Further details on the exact value are provided in Section 2.6.

For an infrastructure node scaling as in regime iii) we observe that the throughput order

capacity scales as for an infrastructure node growth of $\Theta\left(\frac{n}{\log n}\right)$ even if the number of infrastructure nodes scales faster. One implication is that the infrastructure nodes deployed additionally in regime iii) to the ones already deployed in regime ii) are not beneficial for the throughput order capacity. Hence, the scheme which achieves the throughput order capacity of $\lambda(m,n) \in \mathcal{O}\left(\frac{1}{\log n}\right)$ is the same as for regime ii) with only $m'(n) = \frac{n}{\log n}$ infrastructure nodes used out of the $m(n)$ available ones.

2.4 Scaling of energy-per-bit

To study the scaling of energy-per-bit, we show first how the aggregated energy consumption scales with increasing number of nodes. For this purpose we use the physical model from Section 2.2.1 which gives us the energy used for transmission and use (2.6) in order to carry over the results from the protocol model to the physical model. To ensure a minimum received power level in every cell a simple consequence from (2.4) is

$$P^{\mathrm{TX}}(m,n) \sim r(m,n)^{\alpha}, \tag{2.21}$$

which is used for all nodes. More precisely, we use the minimum connectivity range (2.19).

With the achievable schemes for the three different regimes described in Section 2.3 and the notion of energy-per-bit (2.10) we are now able to derive the main result of this chapter: the achievable energy-per-bit scaling for the three infrastructure scaling regimes.

Theorem 1. *The energy-per-bit growth rate $E_b(m,n)$ of a random hybrid network with $m(n) \in \mathcal{O}\left(\sqrt{\frac{n}{\log n}}\right)$ infrastructure nodes operating under the protocol model and the ad-hoc communication scheme described in Section 2.3 satisfies in the asymptotic regime*

$$E_b(m,n) \in \begin{cases} \Omega\left(\sqrt{n \log n}\ h_M\right) & b < \frac{1}{2} \\ \Omega\left(\sqrt{n \log n}\ h_M + h_{BS}\right) & b = \frac{1}{2}. \end{cases} \tag{2.22}$$

Proof. By definition of regime i) there are $m(n) = \left(\frac{n}{\log n}\right)^b$, $b \leq \frac{1}{2}$ infrastructure nodes deployed consuming h_{BS} amount of static energy each. Wireless nodes consume h_{M} of static energy in addition to the energy spent when transmitting with power level of $P^{\mathrm{TX}} = \sqrt{\frac{\log n}{n}}^{\alpha}$. Definition 5 gives the aggregate total energy consumption as

$$E_{\Sigma}(m,n) = n\left(\tau\sqrt{\frac{\log n}{n}}^{\alpha} + h_{\mathrm{M}}\right) + \left(\frac{n}{\log n}\right)^b h_{\mathrm{BS}}. \tag{2.23}$$

Note, that infrastructure nodes spend no energy on transmission since they do not participate in the communication of regime i). However, since they are deployed, they consume

some amount of static energy. Using (2.23) with $\tau = 1\text{s}$ and (2.16) for $b \leq \frac{1}{2}$ we obtain the energy-per-bit scaling of

$$E_b(m,n) \in \frac{n\left(\sqrt{\frac{\log n}{n}}^{\alpha} + h_{\mathrm{M}}\right) + \left(\frac{n}{\log n}\right)^{b} h_{\mathrm{BS}}}{\mathcal{O}\left(\sqrt{\frac{n}{\log n}}\right)} \tag{2.24}$$

$$= \Omega\left(n\left(\frac{\log n}{n}\right)^{\frac{1+\alpha}{2}} + h_{\mathrm{M}}\sqrt{n\log n} + h_{\mathrm{BS}}\left(\frac{\log n}{n}\right)^{\frac{1}{2}-b}\right). \tag{2.25}$$

To find the dominant factor we analyze each summand individually. By applying l'Hopital's rule and noting that $\alpha > 2$ we have

$$\lim_{n\to\infty} n\left(\frac{\log n}{n}\right)^{\frac{1+\alpha}{2}} = 0. \tag{2.26}$$

Furthermore, we have

$$\lim_{n\to\infty}\left(\frac{\log n}{n}\right)^{\frac{1}{2}-b} = \begin{cases} 0 & b < \frac{1}{2} \\ 1 & b = \frac{1}{2}, \end{cases} \tag{2.27}$$

since $\frac{1}{2} - b \geq 0$. Thus, the term for hardware energy consumption of wireless nodes dominates the scaling behavior in the asymptotic regime and we obtain the scaling stated in the theorem. □

The following corollary is an immediate consequence from Theorem 1.

Corollary 2. *Without the consideration of the static energy consumption of mobile nodes h_M and infrastructure nodes h_{BS}, the energy-per-bit scaling tends to zero in the asymptotic regime.*

Proof. With no static energy consumption $h_{\mathrm{M}} = h_{\mathrm{BS}} = 0$ the energy-per-bit scaling in (2.25) changes to

$$E_b(m,n) \in \Omega\left(n\left(\frac{\log n}{n}\right)^{\frac{1+\alpha}{2}}\right), \tag{2.28}$$

which tends to zero for $n \to \infty$ shown in (2.26). □

Theorem 2. *The energy-per-bit growth rate $E_b(m,n)$ of a random hybrid network with infrastructure node scaling according to (2.14) in regime ii) and iii), that operates under the protocol model and the cellular communication schemes in Section 2.3 satisfies in the asymptotic regime*

$$E_b(m,n) \in \begin{cases} \Omega\left(n\left(\frac{\log n}{n}\right)^{b} h_M + h_{BS}\right) & \frac{1}{2} < b \leq 1 \\ \Omega\left(\log n\ h_M + \left(\frac{n}{\log n}\right)^{b-1} h_{BS}\right) & b > 1. \end{cases} \tag{2.29}$$

Proof. Recall that $m(n) = \left(\frac{n}{\log n}\right)^b$ with $b > \frac{1}{2}$. The cellular communication scheme requires a successful end-to-end communication to consist of exactly two wireless hops with transmit power $P^{\mathrm{TX}} = \frac{1}{\sqrt{m}^\alpha} = \left(\frac{\log n}{n}\right)^{\frac{b\alpha}{2}}$; one from source node to its closest infrastructure node and one from the destination's closest infrastructure node to the destination. With n wireless nodes and m infrastructure nodes we have the aggregate energy consumption for the cellular scheme as

$$E_\Sigma(m,n) = n\left(2\frac{1}{\sqrt{m}^\alpha}\tau + h_{\mathrm{M}}\right) + m\ h_{\mathrm{BS}} \tag{2.30}$$

$$= n\left(2\left(\frac{\log n}{n}\right)^{\frac{b\alpha}{2}}\tau + h_{\mathrm{M}}\right) + \left(\frac{n}{\log n}\right)^b h_{\mathrm{BS}}. \tag{2.31}$$

According to Corollary 1 the aggregate throughput scaling is different depending on b resulting in different scaling regimes. We first show the result for $\frac{1}{2} < b \leq 1$. Using this and (2.30) with $\tau = 1\mathrm{s}$ in (2.10) we obtain the energy-per-bit scaling as

$$E_b(m,n) \in \frac{n\left(2\frac{1}{\sqrt{m}^\alpha} + h_{\mathrm{M}}\right) + m\ h_{\mathrm{BS}}}{\mathcal{O}(m)} \tag{2.32}$$

$$= \Omega\left(\frac{n}{m}\left(2\frac{1}{\sqrt{m}^\alpha} + h_{\mathrm{M}}\right) + h_{\mathrm{BS}}\right) \tag{2.33}$$

$$= \Omega\left(2n\left(\frac{\log n}{n}\right)^{\frac{b\alpha}{2}+b} + h_{\mathrm{M}}\ n\left(\frac{\log n}{n}\right)^b + h_{\mathrm{BS}}\right). \tag{2.34}$$

Analyzing each summand we have $\lim_{n\to\infty} n\left(\frac{\log n}{n}\right)^b = \lim_{n\to\infty} n^{1-b}(\log n)^b = \infty$ since $1 - b \geq 0$ for $\frac{1}{2} < b \leq 1$ and $\lim_{n\to\infty} 2n\left(\frac{\log n}{n}\right)^{\frac{b\alpha}{2}+b} = 0$ bearing in mind that $\alpha > 2$ and $\frac{1}{2} < b \leq 1$. This gives the result of the theorem for $\frac{1}{2} < b \leq 1$.

Now, if $b > 1$ the aggregate throughput is $T(m,n) \in \mathcal{O}\left(\frac{n}{\log n}\right)$ and thus

$$E_b(m,n) \in \frac{n\left(2\frac{1}{\sqrt{m}^\alpha} + h_{\mathrm{M}}\right) + m\ h_{\mathrm{BS}}}{\mathcal{O}\left(\frac{n}{\log n}\right)} \tag{2.35}$$

$$= \Omega\left(\log n\left(2\left(\frac{\log n}{n}\right)^{\frac{b\alpha}{2}} + h_{\mathrm{M}}\right) + \left(\frac{n}{\log n}\right)^{b-1} h_{\mathrm{BS}}\right). \tag{2.36}$$

We can see that $\lim_{n\to\infty} \log n\left(\frac{\log n}{n}\right)^{\frac{b\alpha}{2}} = 0$ which yields the result of the theorem for $b > 1$. □

The question remains, for which infrastructure node scaling we obtain the best energy-per-bit scaling. In order to answer this question we search for the value of b yielding the energy-per-bit scaling with the slowest growth rate.

Proposition 1. *In a random hybrid network under the protocol model and using the ad-hoc*

and the cellular communication scheme for $b \leq \frac{1}{2}$ and $b > \frac{1}{2}$, respectively, the energy-per-bit scaling has the slowest growth rate for $b' = 1$.

Proof. With the energy-per bit scaling from Theorem 1 and Theorem 2 we have

$$b' = \underset{b \in \mathbb{R}}{\mathrm{argmin}} \begin{cases} \sqrt{n \log n}\ h_{\mathrm{M}} & \text{for } b < \frac{1}{2} \\ \sqrt{n \log n}\ h_{\mathrm{M}} + h_{\mathrm{BS}} & \text{for } b = \frac{1}{2} \\ n \left(\frac{\log n}{n}\right)^{b} h_{\mathrm{M}} + h_{\mathrm{BS}} & \text{for } \frac{1}{2} < b \leq 1 \\ \log n\ h_{\mathrm{M}} + \left(\frac{n}{\log n}\right)^{b-1} h_{\mathrm{BS}} & \text{for } b > 1. \end{cases} \tag{2.37}$$

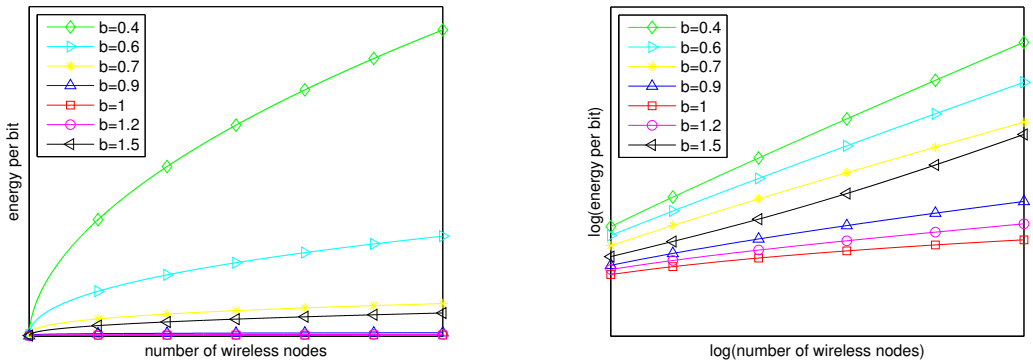
First we observe that $n \left(\frac{\log n}{n}\right)^{b} h_{\mathrm{M}} + h_{\mathrm{BS}}$ is monotonically decreasing for fixed n with increasing parameter b resulting in faster decrease in the scaling behavior for larger values of b. Similarly, we see that $\log n\ h_{\mathrm{M}} + \left(\frac{n}{\log n}\right)^{b-1} h_{\mathrm{BS}}$ is monotonically increasing for fixed n with increasing b which in turn leads to a faster increase of larger values of b. Now we can show that $\lim_{b \to 1} \log n\ h_{\mathrm{M}} + \left(\frac{n}{\log n}\right)^{b-1} h_{\mathrm{BS}} = \lim_{b \to 1} n \left(\frac{\log n}{n}\right)^{b} h_{\mathrm{M}} + h_{\mathrm{BS}} = \log n h_{\mathrm{M}} + h_{\mathrm{BS}}$. We also have $\lim_{b \to \frac{1}{2}} n \left(\frac{\log n}{n}\right)^{b} h_{\mathrm{M}} + h_{\mathrm{BS}} = \sqrt{n \log n}\ h_{\mathrm{M}} + h_{\mathrm{BS}}$ which is the scaling function for $b = \frac{1}{2}$. We complete the proof by noting that $\sqrt{n \log n}\ h_{\mathrm{M}} + h_{\mathrm{BS}} \geq n \left(\frac{\log n}{n}\right)^{b} h_{\mathrm{M}} + h_{\mathrm{BS}}$ for $b \geq \frac{1}{2}$ in the asymptotic regime (large n). □

2.5 Discussion of the results

We have derived the energy-per-bit scaling for two communication schemes achieving the throughput order capacity of (2.16) depending on the infrastructure growth rate. In all cases the terms accounting for transmit energy consumption tend to zero for large n. The densification of infrastructure nodes results in a reduced transmit radius, which in turn, allows for a decreased transmit energy. This decrease is faster than the decrease in throughput and hence gives the observed result.

In contrast, the static energy consumption of hardware is not decreasing with the increase in infrastructure nodes. The derived energy-per-bit scaling for the schemes used in the three regimes is depicted in Figure 2.7. For better illustration we have provided a plot with linear axes in Figure 2.7(a) and a second plot with logarithmic axes in Figure 2.7(b).

Unfortunately, the energy-per-bit scaling tends to infinity with increasing number of wireless nodes n in all three regimes, but for a finite number of nodes there are significant differences between them. The benefits of deploying infrastructure nodes is evident from the comparison of the energy-per-bit scaling for an increasing infrastructure node growth rate. When the number of infrastructure nodes relative to the number of wireless nodes is increased the energy-per-bit growth rate is decreased. In fact, the best energy-per-bit scaling is achieved



(a) Linear scale.

(b) log-log scale.

Figure 2.7: Illustration of the Energy-per-Bit scaling $E_b(m, n)$ for different infrastructure growth rates.

for $b = 1$ where the infrastructure growth rate is $m(n) = \frac{n}{\log n}$. When the infrastructure nodes scale faster ($b > 1$) the energy-per-bit scaling can not be improved, because the additional infrastructure nodes are not needed in the used communication scheme and thus, do not contribute to the communication[6]. From an energy-per-bit perspective it is hence suggested to implement an infrastructure node scaling with $b = 1$ resulting in $m(n) = \frac{n}{\log n}$.

Not only in terms of energy-per-bit scaling but also for a good throughput scaling using $b = 1$ seems beneficial. Figure 2.8 illustrates the throughput scaling behavior for various infrastructure growth rates. In all cases the throughput tends to zero with increasing n. However, it can be observed that the decrease is slower, the more infrastructure nodes there are deployed until $m(n) = \frac{n}{\log n}$ ($b = 1$) is reached where the throughput scaling can't be improved any more. Thus, for a finite number of nodes the deployment of infrastructure nodes according to $m(n) = \frac{n}{\log n}$ will result in the best throughput performance for the outlined schemes.

From this discussion, we conclude that the scaling behavior of the (natural cellular communication) scheme we used for deriving the achievable energy-per-bit scaling is optimal for a choice of $b = 1$ both in terms of throughput and energy-per-bit values. In other words, this suggests to use $m(n) = \frac{n}{\log n}$ infrastructure nodes. We remark that of course, there might be other communication schemes that are more efficient in terms of energy-per-bit scaling. Here, we provide achievable scaling laws and derived the optimal operating point for a large class of schemes.

[6]The reason for this behavior is the upper bound in throughput identified by [ZdV05]. However, since this result is due to the chosen way how source-destination pairs are selected, in other cases this upper bound might be lifted.

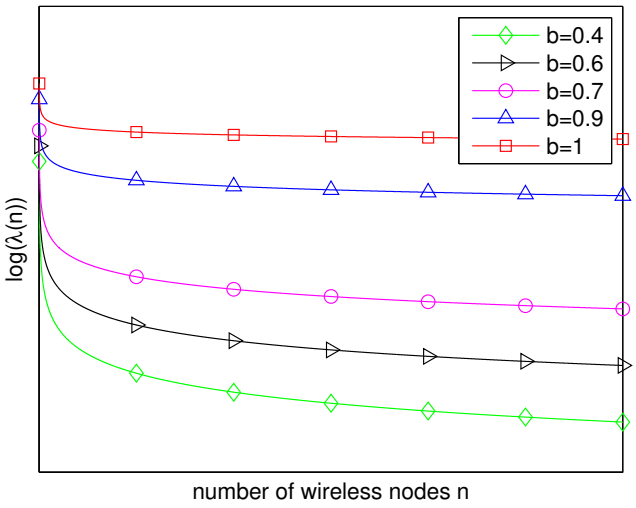


Figure 2.8: Illustration of the throughput per node scaling for different infrastructure node growth rates. The y-axis is in logarithmic scale.

2.6 Numerical evaluation

In order to evaluate the theoretical statements from Section 2.3 and 2.4, we perform simulations for several random realizations of the network. In our simulations, we focus on regime ii), where infrastructure nodes are used. The simulation setup and the results will be discussed in detail in the subsequent section. Our simulation scenario resembles the cellular communication scheme presented in Section 2.3. For the purpose of varying the infrastructure node growth rate we use (2.14) with different values of b. The resulting infrastructure growth rate for selected values of b is illustrated in Figure 2.9. Our simulations are performed for values of $\frac{1}{2} \leq b \leq 1$. By doing so, we obtain networks which exploit the throughput gains from placement of infrastructure nodes as identified by [ZdV05]. For completeness, we summarize the construction of the simulation scenario for a fixed number of users n.

The underlying communication area is a circular area with unit area. We deploy hexagonal cells with equal sides of $\frac{1}{\sqrt{m}}$. The infrastructure nodes are placed in the center of each hexagonal cell that is fully contained within the unit area disc.

Remark 1. *Here we differ from the scheme in [ZdV05] in the sense that we do not apply a Voronoi tessellation to increase the service area of the outmost infrastructure nodes. We just decrease the total simulation area to the size which is actually covered by the remaining hexagons. This approach is reasonable because for increasing number of wireless nodes and hence increasing number of infrastructure nodes within the unit area disc, the area which will be removed tends to zero.*

The obtained layout has been depicted in Figure 2.5. Wireless nodes are placed according to

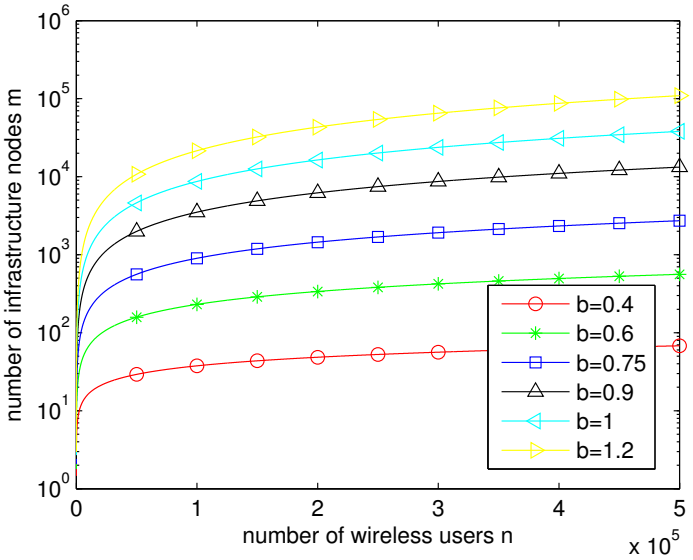


Figure 2.9: Illustration of different infrastructure growth rates $m(n)$.

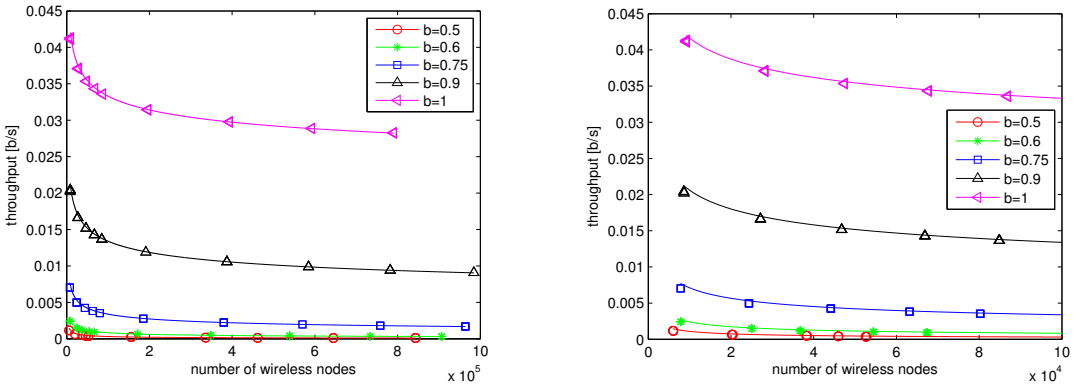
a random uniform distribution on the unit area disc. Recall that the actual communication area is slightly decreased since we discard cells which are not fully contained within the unit area disc. Thus, we also discard all wireless nodes placed outside these hexagonal cells. This determines the actual number of wireless nodes and infrastructure nodes for our simulation.

According to the *achievable scheme* we have to choose a frequency re-use pattern which ensures the achievability of the minimum SINR β at each node. The SINR for our communication scenario using the physical model is lower-bounded by [GK00]

$$\beta > \frac{\frac{P(m,n)}{N_0}}{r^{\alpha}(n) + \frac{1}{\left(1+\frac{\Delta}{2}\right)^{\alpha}} \frac{P(m,n)}{N_0} \left(9 + \frac{3}{\alpha-1} + \frac{6}{\alpha-2}\right)}. \tag{2.38}$$

Together with equation (2.6) the transmit power $P(m,n)$ can be chosen such that the results from the protocol model carry over to the physical model. The simulated communication scheme is the physical model from Section 2.3 and the noise power is calculated according to $N_0 = k \cdot T \cdot B_S \approx 8 \cdot 10^{-16}\text{W} = -120.98\text{dBm}$ with system bandwidth $B_S = 200$, Boltzman constant $k = 1.3807 \cdot 10^{-23}$ and temperature $T_0 = 290\text{K}$. The guard transmit power and the corresponding guard interval Δ in the protocol model is chosen according to Equation (2.6). This choice corresponds to a frequency re-use pattern of three (as shown in Figure 2.6) which in turn results in $\Delta = 1$.

Remark 2. *Note, that the chosen parameter values strongly depend on the considered scenario and will have influence on absolute values. Though the scaling behavior and trends will not be affected by the choice as long as they are chosen according to the limitations outlined above.*



(a) Large number of wireless nodes.

(b) Small number of wireless nodes.

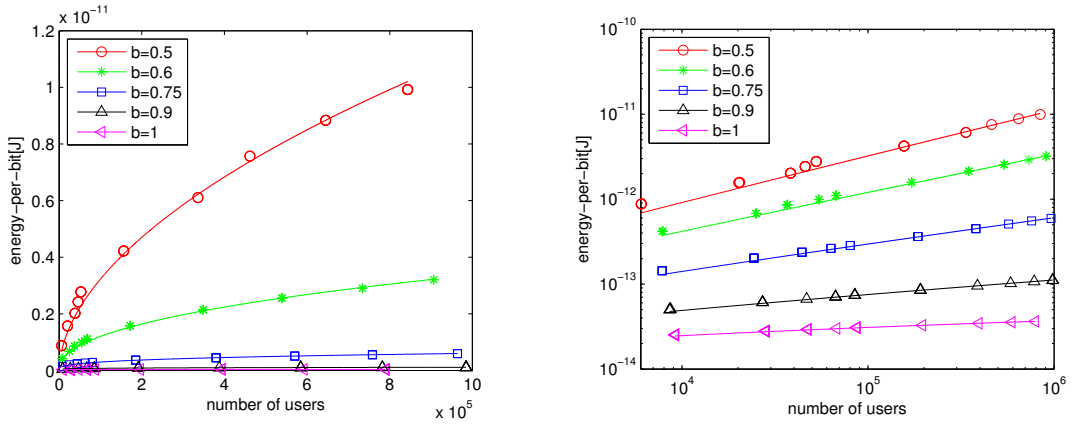
Figure 2.10: Throughput vs. number of users for scaling of number of base stations $m(n)$. The solid line is the theoretical scaling function and the markers are the actual data points obtained by simulation.

2.6.1 Throughput scaling

The scaling behavior of the throughput achieved by the cellular communication scheme is shown for a number of wireless nodes $n \in \left[500, 10^6\right]$. We simulate for different infrastructure growth rates achieved by selecting $b \in [0.5, 1]$ in Equation (2.14). The simulation results are illustrated in Figure 2.10. The decrease in throughput with increasing number of wireless nodes is shown in Figure 2.10(a). The resulting throughput values from our simulations are represented by the respective markers, whereas the solid lines represent fitted curves of the theoretical results from Equation (2.29). For an increase in the infrastructure growth rate we can observe an improvement in the throughput in the finite regime as expected by the theory. Furthermore, we can see that in the finite regime the throughput decrease matches closely the theoretical evolution. Even for a small number of wireless nodes the theory is matched with acceptable margins which can be seen in the detailed view in Figure 2.10(b).

2.6.2 Energy-per-bit scaling

The energy consumption in our simulation setup follows the model defined in Definition 5. The energy consumption parameters are chosen to be in the order of magnitude of the chosen transmit powers for the different links and cell sizes. In particular, we choose $h_{MS} = h_{BS} = -120$ dBm for the transmit power obtained with $\Delta = 1$ and $c = 3.1$ for our simulations. As for the throughput scaling we apply a fitting procedure to compare the simulation results with the theoretical findings from Section 2.4. The resulting plot is shown in Figure 2.11. The simulation results support the theoretical findings that the energy-per-bit growth rate increases for an increasing number of wireless nodes n and thus, an increase in the infrastructure nodes $m(n)$. We observe that the simulation results closely follow the theoretical scaling curves approving the theory. We also see that for a



(a) Linear scale.

(b) Log-log scale.

Figure 2.11: Energy-per-bit in J for different regimes and number of users. Theoretical scaling behavior shown by solid line and simulations as markers.

faster infrastructure node growth rate we can achieve better energy-per-bit scaling rates with the slowest energy-per-bit scaling for $b = 1$. However, we can also identify that for all infrastructure growth rates the energy-per-bit tends to infinity for $n \to \infty$. Thus, it is suggested to operate the network under the proposed cellular communication scheme with an infrastructure growth rate of $b = 1 \to m(n) = \frac{n}{\log n}$. The scaling behavior matches more closely the theory for increasing number of wireless nodes n.

2.7 Conclusion

In this chapter we have presented an information-theoretic analysis for estimating the scaling behavior of throughput and energy-per-bit in hybrid wireless communications networks. Of particular interest is the asymptotic regime of large-scale random wireless networks with a reasonable communication scheme in which the number of users randomly placed over some area in the plane tends to infinity. Results from literature have shown that the throughput per user in such networks will vanish in the asymptotic regime with the increasing traffic demand, unless cooperation, mobility or the placement of additional infrastructure nodes is exploited. We have highlighted the cause for those discouraging results of vanishing throughputs. The fundamental limit is that wireless resources shared by every node in the network which constricts the capacity. This fact causes that nodes transmitting concurrently over the same wireless channels create mutual interference and thereby limiting its capacity.

In order to overcome the vanishing throughput in the case of finite wireless nodes, we showed the benefits of deploying infrastructure nodes. We argue for the infeasibility of a linear infrastructure node scaling ($m(n) = n$) and highlight good alternative growth rates for which the throughput per user moderately tends to zero. By doing so, the throughput

regime	used scheme	energy-per-bit scaling $E_b(m,n)$	per node throughput scaling $\lambda(m,n)$
$b < \frac{1}{2}$	multi-hop ad-hoc	$\Omega\left(\sqrt{n \log n}\ h_{\mathrm{M}}\right)$	$\Theta\left(\frac{1}{\sqrt{n \log n}}\right)$
$b = \frac{1}{2}$	multi-hop ad-hoc	$\Omega\left(\sqrt{n \log n}\ h_{\mathrm{M}} + h_{\mathrm{BS}}\right)$	$\Theta\left(\frac{1}{\sqrt{n \log n}}\right)$
$\frac{1}{2} < b \leq 1$	cellular communication	$\Omega\left(n \left(\frac{\log n}{n}\right)^{b} h_{\mathrm{M}} + h_{\mathrm{BS}}\right)$	$\Theta\left(\frac{1}{n}\left(\frac{n}{\log n}\right)^{b}\right)$
$b > 1$	cellular communication	$\Omega\left(\log n\ h_{\mathrm{M}} + \left(\frac{n}{\log n}\right)^{b-1} h_{\mathrm{BS}}\right)$	$\Theta\left(\frac{1}{\log n}\right)$

Table 2.1: Energy-per-bit scaling $E(m,n)$ and per node throughput scaling $\lambda(m,n)$ for the three different regions. To arrive at different infrastructure node growth rates we use $m(n) = \left(\frac{n}{\log n}\right)^{b}$ with $b \in \mathbb{R}$.

performance is traded in for higher energy efficiency expressed in terms of scaling laws in the asymptotic regime. Most of the existing scaling law results are applicable to infrastructure-less and cellular wireless network, where each user (transmitter-receiver pair) is equipped with a standard air interface that treats interference from other users as noise; this includes radio access technologies of all existing telecommunications standards such as GSM, UMTS, WiMAX and LTE. The main distinguishing point of our study to existing studies is that we take into account the energy consumption for hardware components of all wireless nodes including the infrastructure nodes. The analysis provides a valuable insight into the rate at which the cellular infrastructure needs to grow to provide a good compromise between throughput and energy efficiency. It becomes evident that the energy consumption of hardware can not be neglect. In fact, we have shown that the energy-per-bit scaling is governed by the static energy consumption of hardware. The obtained results for energy-per-bit scaling $E_b(m,n)$ and throughput scaling $\lambda(m,n)$ depending on the infrastructure growth rate are summarized in Table 2.1.

A main point of scaling laws is that they are effective in the asymptotic regime, i.e. for a large number of wireless nodes n. Therefore, a highly relevant question is how large the number of wireless nodes n needs to be for the scaling laws to be effective. This question is addressed in Section 2.6.1 and Section 2.6.2. We are not aware of any publication that provides simulations to evaluate the scaling behavior in order to see how large n should be for the network to behave as predicted by the scaling laws. By means of simulation we can observe, that the theoretical scaling behavior of throughput and energy-per-bit for the asymptotic regime ($n \to \infty$) is already valid for a relatively small number of wireless nodes in the finite regime. The scaling behavior matches the theory more closely for an increasing number of wireless nodes n.

We see that the static energy consumption of the hardware plays a key role when investigating the energy efficiency of a cellular communication network in terms of energy-per-bit.

Therefore, future wireless communication networks will have to reduce the energy consumption of hardware by i) making hardware itself more energy efficient and ii) designing communication schemes that minimize the active time of hardware.

From these observations we can draw conclusions for large-scale networks (large values of n)

- For a practical communication scheme the infrastructure nodes need to scale accordingly to have the best energy per bit scaling.
- Under our considered communication scheme, a too large infrastructure scaling is not beneficial in terms of energy per bit.
- More advanced radio access technologies are needed to fully exploit the potential of cellular networks with very dense infrastructure and to achieve a better scaling tradeoff between throughput and energy efficiency.

3 Load-aware Network Topology Control

This chapter addresses the energy consumption of cellular communication networks for longer time periods in the order of several minutes or hours. We propose novel techniques to save energy in such networks by exploiting redundancies and finding network configurations that better match the capacity of the network to the actual demand over longer time periods.

Networks deployed today are typically designed and operated to provide the best QoS during peak demand. Global network parameters and the network topology are largely static, although, as pointed out in many studies (see for instance [WMBW09, CH08, OKLN11, AL08]), the traffic load fluctuates significantly over time and space. In a static network setup such spatio-temporal load fluctuations lead to huge capacity surpluses at times of low traffic demand; a generic example is depicted in Figure 3.1. These surpluses are amenable for energy savings when adapting the provided service to the actual demand. In order to exploit the fluctuations for energy savings, it is essential to reduce the energy consumed by hardware and auxiliary equipment (e.g. coolers). Studies have shown that, in typical networks with current technology, the energy consumption of base stations exceeds 50% of the total network energy budget [HHA$^+$11]. The largest portion of energy is consumed by powering base station hardware and auxiliary equipment. Here, we see large potential to save energy by temporarily disengaging redundant hardware components of base stations or even entire base stations. Indeed, as pointed out by [OKLN11], reducing the number of active base stations in periods of low traffic load offers a huge potential for energy savings. It is envisioned that switching off base stations of multiple operators, that cooperate by sharing their network equipment, can lead to a reduction in energy consumption of up to 29% [OKLN11] in real network deployments. In the future, where we expect the network density to grow significantly in order to meet the growing demand for wireless access [TN12], such savings will become more pronounced.

We approach this idea and propose a novel optimization framework that considers both the load-dependent energy radiated by the antennas and the remaining forms of energy needed for operating the base stations. To this end, we divide the traffic demand into multiple time slots, and we show optimization tools for a given time slot. The objective of our optimization problem is to select network components consuming the least amount of energy while ensuring that the data rate requirements of the users are met throughout the coverage area. The starting point are techniques used for sparse optimization and we develop a majorization-minimization (MM) algorithm that gives good solutions to our energy minimization problem. The iterative algorithm is load-aware, has low computational

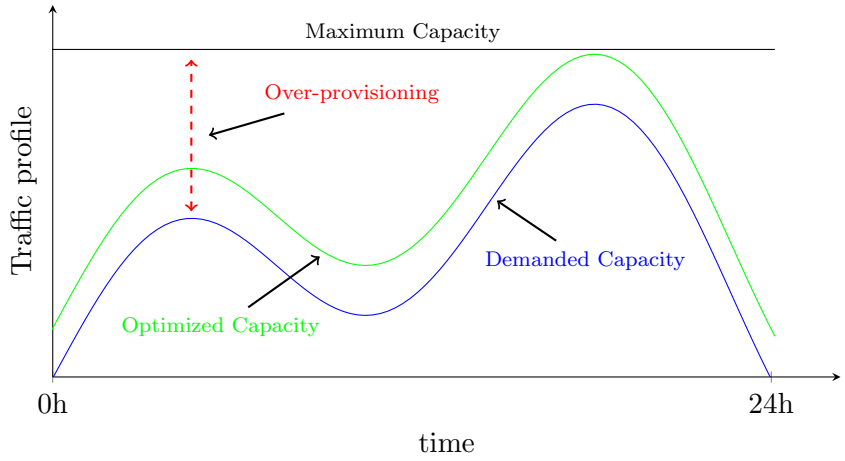


Figure 3.1: Mismatch of demanded (blue) and over-provisioned capacity (black) leads to waste of energy. Optimize provided capacity (green) to more closely resemble actual demand and save energy.

complexity, and can be implemented in an online fashion to exploit load fluctuations on a short time scale.

Remark 3. *In contrast to many studies in literature, our metric of interest is not the energy efficiency of cellular networks which is usually defined as* $\frac{C}{E}$*, where* C *is the network capacity and* E *denotes the total energy consumption of the network. Since we are interested in the actual demand rather than the capacity, the energy efficiency of the network is not a suitable metric. In fractional programming, which is usually applied to maximize the energy efficiency the capacity is to be maximized at the same time as the energy consumption is minimized. However, our interest is in a capacity that matches the actual demand and can be achieved with the least energy consumption.*

3.1 Main contribution

The main objective of our work is the minimization of the overall energy consumption in the downlink channel of mobile cellular networks. In contrast, to most existing approaches, we take into account the energy consumed by hardware and auxiliary equipment. By mitigating this shortcoming, we put ourselves in a position to boost the energy savings in cellular networks. The underlying problem is of combinatorial nature because it essentially amounts to selecting a subset of network components that consume the least amount of energy, while providing the desired coverage and capacity to the users. More precisely, motivated by [ZGY+09], we formulate a combinatorial optimization problem to find a network configuration that saves the most amount of energy while satisfying traffic demands expressed in terms of minimum data rate requirements. Our optimization framework takes into account the static energy consumed by hardware as well as the load-dependent energy spent on transmission in an optimal manner and thus, is able to balance between the different sources of energy consumption. The constraints of our optimization problem model

the technology specific capabilities and limitations to provide the desired QoS to the users. This approach enables us to apply our framework to multi-RAT systems by incorporating different RAT specific constraints and account for a broad variety of hardware energy consumption models.

The core of our optimization problem is of combinatorial nature. In fact, we show that the problem we pose is a special case of the standard bin-packing problem, which is known to be NP-hard. An immediate consequence is that even problems of moderate size are hard to solve. Thus, we follow a widely-known relaxation approach (see for instance [Lib04]) to arrive as an intermediate step at a closely related problem that can be solved efficiently. More precisely, we approximate the objective function by a concave function and relax all non-convex constraints to arrive at a problem that minimizes a concave function subject to a convex set. For this task we make use of a majorization-minimization (MM) technique [HL04]. Thus, we first point out that in contrast to most existing work in this field our approach has a theoretical justification for its good performance. But the combination of these techniques has many more major advantages over existing approaches to the problem. In particular, the proposed algorithm is able to find good solutions in a relatively short time so that it can even be used to utilize fluctuations in traffic demand on a relatively short time scale. Another major advantage is its ability to cope with a variety of network elements. It is easy to account for different energy consumption models of different hardware, e.g., base stations of different generations, sectors or even antennas as long as they can be modeled as any concave or convex function of the load. This ability puts us in the position to optimize for the network configuration that gives the smallest total network energy consumption. Furthermore, our approach is easy to modify in order to balance the minimization of the static energy consumption against the minimization of the load-dependent energy consumption. Thereby, more weight can be given to the deactivation of as many network elements as possible or an appropriate load balancing to avoid highly loaded base stations.

Starting from the core optimization framework, we show how it can be extended in several directions. First, we present how to incorporate techniques motivated by coordinated multi-point (CoMP) transmission [HV04], which actually make some relaxations superfluous and lead to further energy savings. We then delve into the application of the framework to a single-RAT LTE network followed by a brief extension to multi-RAT systems consisting of UMTS and LTE cells. Thereby, we underline the broad applicability of our optimization framework to different types of network settings. Further examples of network settings that can be addressed by our framework are sketched in Appendix B.3. Apart from that, in the main part of this chapter we present a modification of the optimization framework that makes it amenable for a distributed implementation. Such a distributed implementation significantly reduces the signaling overhead and helps to implement our algorithms in larger regions without the necessity of extensive communication with a central computation unit.

At last, we extend our optimization framework into the direction of anticipatory scheduling,

where the decisions on resource allocation and user-cell assignments are based not only on present channel state information but also on information about future propagation conditions. With such techniques, also referred to as proactive resource allocation and user assignment (PRAUA), the network performance and energy savings can be significantly enhanced. PRAUA has broad applicability in problems of smooth media streaming and traffic offloading and we focus on the latter one. PRAUA helps 'smear' the traffic requirements in time and space allowing for maintaining energy-efficient network configurations over a longer period of time. Thereby, PRAUA exploits the knowledge about users' mobility and path loss to proactively build user-cell assignment and resource allocation schedules that greatly support energy savings in cellular communication systems. In particular, we develop algorithms that schedule data transmissions for new service applications when it is favorable for energy savings. The developed mechanisms target new service types enabled by the storage capabilities at user devices such as buffered delay-sensitive applications. For such applications we investigate an optimization problem that jointly assigns users to serving cells and schedules resources over a finite time horizon with the goal of minimizing the total network energy consumption. We transfer the non-convex optimization problem into a format that makes it possible to apply reformulation and relaxation techniques introduced in the coarse of our main optimization framework of this section. In contrast to most existing work in literature, we explicitly incorporate the user-cell assignment in the optimization framework and target energy savings with a guaranteed QoS level instead of maximizing the QoS.

The chapter is organized as follows. Section 3.2 positions our work in the context of current results in literature. In Section 3.3 we provide our main optimization framework including the general system model and the energy consumption model used in our optimization problem. The general problem formulation and the derivation of the algorithm to find good solutions to the problem is presented. Section 3.3.5 outlines on modifications for including CoMP like transmission modes. In Section 3.4 and Section 3.5 we show how to apply the optimization framework to a single-RAT LTE and a mutli-RAT UMTS/LTE network respectively. A semi-decentralized implementation of the framework is sketched in Section 3.6. Section 3.7 presents numerical evaluations for the developed algorithms. The framework for PRAUA is introduced in Section 3.8, as well as numerical evaluations are presented. Finally, we conclude this chapter in Section 3.9.

3.2 Related work

The exploitation of temporal and spatial redundancies in wireless systems for energy savings has been studied by the research community over the past years. Mainly for the application in network planning references [GBGF+11, NWGY10, CCK+12] address the problem of finding an optimal number of cell sites and base station placements so as to minimize the overall energy consumption subject to some QoS requirements of users. The study in [GBGF+11] optimizes the number of base stations and their locations targeting the

minimization of the overall expected energy consumption in a wireless network employing time division multiple access (TDMA). The approach taken by the authors involves a mixed integer programming problem and proposes a scheme to solve the problem involving a simplex method together with the branch and bound algorithm. The choice of branch and bound has a sever impact on the applicability to real-time scenarios, because it may be slow, even if the underlying problem is of moderate size [JB09]. In addition, the assumption of TDMA is very limiting, since the analysis does not carry over to systems where inter-cell interference is experienced. Inter-cell interference has been identified as one of the major challenges faced by designers of modern wireless communication systems [MK10, LPGdlR+11].

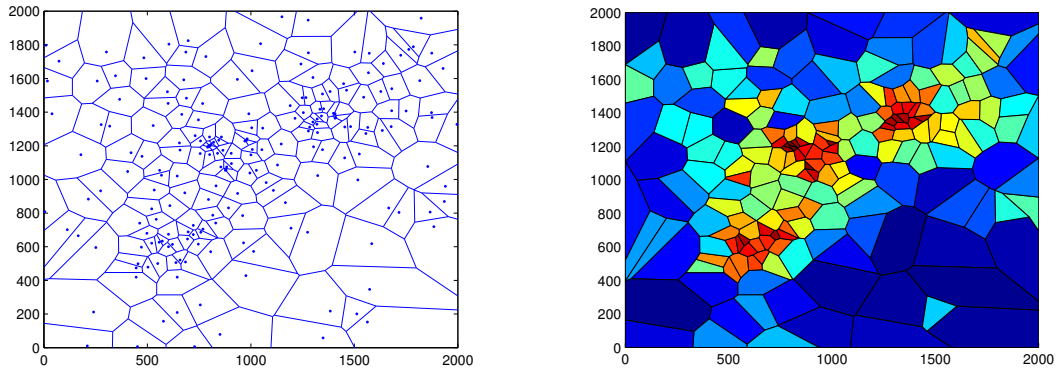
The problem of base station selection in the presence of traffic load fluctuations is addressed by references [NWGY10, ZGY+09]. They propose algorithms of centralized and decentralized nature that show good performance giving good results in reasonable time. However, the underlying system model does not allow to incorporate different sources of energy consumption, which is of utmost importance in modern wireless communication networks consisting of hierarchical structures. Furthermore, the performance is evaluated only on basis of numerical evaluation and no analytical justification for the performance is provided.

The authors of [CCK+12] argue in favor of sleep mode techniques coupled with various network planning schemes. The authors focus on a genetic algorithm to find energy-efficient network deployments by putting selected base stations into sleep mode. The major drawback of this purely heuristic approach is the lack of any mathematical justification and that the proposed approach cannot incorporate other radio technologies other than UMTS. In contrast, as mentioned before, our optimization framework is general enough to be applied to multi-RAT scenarios, including the second, third, fourth and future generations of cellular networks.

In the field of multi-RAT network optimization most work has the objective of enhancing coverage and capacity [VPRSA09, GAPRS05, YTW+11]. Mainly, the complementarity between different radio technologies is used for enhanced radio resource management (RRM) without consideration of the energy consumption. So does the work in [VPRSA09] that proposes a dynamic joint radio resource management mechanism for LTE-UMTS coexistence scenarios. For the decision on which RAT is most suitable for each user a Reinforcement Learning based algorithm is implemented at each base station. However, the work only considers networks where UMTS and LTE are collocated on the same base station. In addition, the user assignment is assumed to be given and is therefore not part of the optimization. Reference [GAPRS05] uses Fuzzy-Neural systems and neural networks for the task to coordinate the radio resource management of different RANs. The main focus lies on guaranteeing a certain QoS level to all users and at the same time keeping an acceptable value of the dropping and blocking probabilities.

Anticipatory networking which is sometimes referred to as proactive resource allocation is applied in wireless networks to optimize resource allocation and user assignments with a

large number of different objectives. Among others, it is used for anticipatory buffering, which targets smooth media streaming by filling the play-out buffer of mobile users in advance of predicted poor coverage situations. Another application is anticipatory traffic offloading, where the load at base stations is reduced by preloading or postponing the data transmission of mobile users to neighboring cells. In this context [SY12] proposes techniques based on load estimation to implement data offloading and thereby leveraging the throughput-coverage trade-off. The work highlights the importance of accurate load prediction. When the load at base stations is known in advance the resource allocation can be adjusted over the time horizon in order to optimize the load according to the desired optimization goal. In [GPMM14] potential energy savings of techniques exploiting the users' delay-tolerance are evaluated. The work switches between predetermined network configurations when the users' QoS requirements allow for delay tolerance. A key finding is the improved energy savings when users can be scheduled for a later point in time and thereby energy efficient network configurations can be operated for a longer time period. Proactive scheduling has been considered in [AzHV13] to improve the QoS of users traveling through the service area of several cells. The presented framework plans the resource allocation over a certain time horizon for fixed user-cell assignments to maximize the throughput of users. The authors of [AzHV14] propose a predictive framework for video streaming applications to increase the energy-efficiency in wireless networks. The framework is based on rate predictions for known user-cell assignments and considers the energy consumption when base stations can be switched off in deep sleep mode. The problem is formulated as a mixed integer linear program (MILP) where decisions on multiuser rate allocation, video segment quality, and base station transmit power are jointly optimized. Thus, the proposed approach trades off video quality for an improved energy-efficiency. A heuristic multi stage algorithm is used to derive solutions for the MILP problem by first allocating rates to users and then determining the segment quality and active base station set. The reasoning is based on the observation that efficient rate-allocation schemes provide power savings. Other analytical justifications for the performance are not given. By proactive resource allocation and video quality decisions the authors in [GIAT15] reduce the energy consumption of the whole network by solving a mixed integer non-linear program (MINLP) problem. An algorithm is proposed that decomposes the association and resource allocation problem in a master problem and several sub-problems to make the problem tractable. Thereby, the authors leverage energy costs and video quality taking into account backhaul costs. The resulting integer programs are solved directly by mathematical solvers and the authors argue to achieve decent scalability. However, this is achieved by assuming the allocation of an equal number of resource blocks to all users in the master problem.



(a) Exemplary test point layout.

(b) Exemplary traffic demand map.

Figure 3.2: Illustration of the test point concept.

3.3 General problem definition and solution

This section presents the general system model used throughout this chapter. We further confine ourselves to a general problem formulation that highlights the challenges in devising algorithms for energy savings.

3.3.1 System model

We consider a cellular communication network with an established network topology. The focus is on the downlink channel and we assume that there is a central network controller that is responsible for collecting measurements, executing the proposed algorithm, and propagating updated network configuration parameters throughout the network. In the most general case we assume that there are L base stations deployed in the network with one or more sectors (called cells in the following). We denote the set of all base stations by $\mathcal{L}$, and the cells belonging to base station l by $\mathcal{S}_l$. The full set of all M cells in the network is thus $\mathcal{M} := \cup_{l \in \mathcal{L}} \mathcal{S}_l$.

Assumption 5. *The cell deployment is dense enough so that coverage areas of different cells overlap creating redundancies in large geographical areas.*

By Assumption 5 users can potentially be connected to a set of neighboring cells in most cases. The particular section of the serving cell is subject to optimization in this study.

Ensuring coverage via test points

In order to ensure the desired coverage anytime and everywhere in the considered area, we impose coverage constraints by adopting the concept of test points, which is widely used in network planning and optimization [Tut98, ACMS03].

Definition 9 (Test point)**.** *A test point (TP) is a centroid of a predefined subarea that represents an aggregated QoS requirement resulting from individual QoS demands of all potential users in this subarea.*[1] *Without loss of generality, we assume N TPs with the set of all TPs denoted by $\mathcal{N} := \{1, 2, ..., N\}$.*

An interpretation of this definition is depicted in Figure 3.2 for a generic service area. A consequence of Definition 9 is that small-scale fluctuations in QoS demand at the user level are averaged out at the TP level. We assume that lower layers of the protocol stack (e.g. through adaptive modulation or coding) can compensate for those small scale fluctuations. On a large scale, the traffic demand is assumed to be static for a sufficiently large period of time for which we derive a feasible network configuration that supports this traffic demand. The duration of this period depends on the accuracy of the demand estimates and other factors such as security margins included in the optimization framework.

Assumption 6. *The QoS requirement for a TP corresponds to the aggregated expected traffic over the respective area per unit time. This traffic requirement is expressed in terms of the minimum required data rate per TP.*

Assumption 7. *If the minimum rate requirement of TP j is met, so are the requirements of the users in the associated subarea.*[2]

For services with no explicit data rate requirements (e.g., voice calls), we assume that they can be supported if a minimum data rate per service request is ensured which is a reasonable assumption in systems employing VoLTE.[3]

By Assumption 6, each TP $j \in \mathcal{N}$ introduces a rate requirement r_j to the system and we collect the rate requirements of all TPs in the vector $\boldsymbol{r} = [r_1, r_2, \ldots, r_N] \in \mathbb{R}^N_{++}$. These rates $\boldsymbol{r}$ need to be provided by the network. In general, a TP can be assigned to any cell and an assignment should be understood as follows. If TP $j \in \mathcal{N}$ is assigned to a cell $i \in \mathcal{M}$ that provides the requested data rate r_j, then all users in the respective subarea associated with TP j are served by cell i. The assignment of the TPs to the cells is subject to optimization in our work. We use $\mathbf{X} = [x_{i,j}] \in \{0, 1\}^{M \times N}$ to denote the assignment matrix, where $x_{i,j} = 1$ if TP j is assigned to cell i and $x_{i,j} = 0$ otherwise.

Assumption 8. *While each TP is assigned to exactly one cell, each cell can serve multiple TPs, and the set of TPs served by cell i under assignment $\mathbf{X}$ is denoted by $\mathcal{N}_i(\mathbf{X}) \subset \mathcal{N}$.*

We point out that this assumption has been widely used in previous studies [ZGY+09, NWGY10, ACMS03], and it is valid throughout our studies except for Section 3.3.5,

[1] A test point becomes a user if it represents a QoS requirement of one particular user, in which case the subarea is a point corresponding to the position of this user.

[2] The smaller the area represented by each TP, the better is this approximation. However, smaller areas imply an increased number of TPs, and the computational complexity of the proposed algorithm grows.

[3] In Section 3.5 where UMTS networks are considered, we present methods how to explicitly account for voice calls in the multi-RAT scenario.

where it is shown how to include scenarios in which each TP can be served by multiple cells. Assumption 8 translates to the first set of constraints $\mathcal{C}_1$ for the assignment problem

$$\sum_{i \in \mathcal{M}} x_{i,j} = 1, \quad j \in \mathcal{N}. \tag{3.1}$$

The optimization of the assignment of TPs to cells is the main task of the optimization of this work. The assignment will point out cells that can be deactivated for energy reasons. More precise, if $\mathcal{N}_i(\mathbf{X}) = \emptyset$ for some $i \in \mathcal{M}$, then cell i can be deactivated because no TP is assigned to cell i. In contrast, if $\mathcal{N}_i(\mathbf{X}) \neq \emptyset$, then cell i is active and each TP connected to it induces some amount of *cell load* $\rho_i : [0,1]^{MN} \to [0,1]$.

Definition 10 (Cell Load)**.** *Given the assignment $\tilde{\boldsymbol{x}} := \mathrm{vec}(\mathbf{X})$, the load of cell i, denoted by $\rho_i(\tilde{\boldsymbol{x}}) \in [0,1]$ or simply ρ_i for notational simplicity, is defined to be the ratio of the number of resource blocks requested by TPs served by cell $i \in \mathcal{M}$ to the total number of resource units B_i available at this cell.*[4]

We use $\boldsymbol{\rho} := [\rho_1, \ldots, \rho_M]^T \in [0,1]^M$ to denote the vector of all cell loads. From the definition of cell load, we have the following:

Fact 3. *The load at cell i satisfies $\rho_i > 0$ if and only if (iff) cell i serves at least one TP.*

By Definition 10 the cell load is the relative number of resource units requested from all TPs assigned to it. Generally, the type of resource unit is technology dependent and we detail on the specifics for selected technologies in Section 3.4 and Section 3.5. Without loss of generality we denote by $b_{i,j} > 0$ the number of resource units at cell i necessary to serve TP j with data rate r_j and provide further details in the respective Sections. For now we assume that $b_{i,j} > 0$, $\forall \mathcal{M}, \forall \mathcal{N}$ is a constant that does not depend on the assignment $\mathbf{X}$.[5] In addition, following Definition 10, the load at cells can be computed by

$$\rho_i = \sum_{j \in \mathcal{N}_i(\mathbf{X})} \frac{b_{i,j}}{B_i} \leq 1, \; i \in \mathcal{M}, \tag{3.2}$$

which gives the second set of basic requirement $\mathcal{C}_2$ for the general assignment problem.

Remark 4. *In practice, cells need to reserve some fraction of their resource units for signaling. If cell i has B_i^* resource units in total, and it needs to reserve $a_i > 0$ of its resource units for signaling, then the resource units at cell i available for allocation to TPs are $B_i = B_i^* - a_i$.*

[4] Note that for LTE networks B_i can also be interpreted as the total bandwidth available at cell i, in which case ρ_i is expressed in terms of the fraction of required and available bandwidth.

[5] We will see that this assumption is generally not fulfilled but we show practical ways how to achieve it.

Energy consumption model

Our model for energy consumption of base stations and its cells consists of two parts. On the one hand, the cell load-dependent transmit energy radiated by antennas and on the other hand remaining sources of energy consumption that are independent of the cell load as long as the cell/base station is *active*. This stands in stark contrast to most work in literature that confines itself to the transmit energy consumption.

Definition 11 (Active base station/cell)**.** *Consider a particular base station $l \in \mathcal{L}$ and its cells $i \in \mathcal{S}_l$. Let $\rho_i \in [0,1]$ be the load of cell i. We say that a cell i is* active *iff $\rho_i > 0$ and that base station l is* active *iff one of its cells is* active*, i.e. $\sum_{i\in\mathcal{S}_l} \rho_i > 0$. If a cell or base station is not active, it is said to be* inactive*.*

With Definition 11 we are in the position to define the energy consumption $E_l : [0,1]^M \to \mathbb{R}_+$ of a base station l.

Definition 12 (Energy consumption)**.** *Given a TP assignment $\mathbf{X}$ inducing a cell load $\boldsymbol{\rho}$, the energy consumption $E_l(\boldsymbol{\rho}) \geq 0$ of base station l is defined to be the power that the respective base station consumes per unit of time, where $E_l(\boldsymbol{\rho}) = 0$ iff base station l is inactive.*

The function $E_l(\boldsymbol{\rho})$ depends on several different factors and generally is different between base stations. However, we can classify three different classes of energy consumption parts at the base station:

(i) The static energy consumption of the base station $c_l > 0$ (due to shared hardware between sectors e.g. cooling, power supply, etc.),

(ii) The static energy consumption $e_i > 0$ ($i \in \mathcal{S}_l$) of its *active* cells (e.g. due to power amplifiers, signal processing units, etc.), and

(iii) the load-dependent dynamic energy consumption of its *active* cells $f_i(\rho_i)$ ($i \in \mathcal{S}_l$), where $f_i : [0,1] \to \mathbb{R}_+$ is a given continuous function relating the energy consumption to the corresponding cell load.

By these definitions and Fact 3, the energy consumption $E_l(\boldsymbol{\rho})$ of base station l is a discontinuous function of the load, and we have

$$E_l(\boldsymbol{\rho}) = \begin{cases} 0 & \text{cells } i \in \mathcal{S}_l \text{ serve no TP,} \\ c_l + \sum_{i\in\mathcal{S}_{l,\text{active}}} e_i + f_i(\rho_i) & \text{otherwise,} \end{cases}$$

where $\mathcal{S}_{l,\text{active}} \subset \mathcal{S}_l$ is the set of active cells of base station l. Therefore, the total energy consumption in a network, which is the accumulated energy consumption of all active base stations, yields

$$E(\boldsymbol{\rho}) = \sum_{l\in\mathcal{L}} E_l(\boldsymbol{\rho}) = \sum_{l\in\mathcal{L}} \left(c_l \left| \sum_{i\in\mathcal{S}_l} \rho_i \right|_0 + \sum_{i\in\mathcal{S}_l} (e_i |\rho_i|_0 + f_i(\rho_i)) \right). \tag{3.3}$$

For concreteness, we make the following assumption regarding the cell load-dependent energy consumption (see also Remark 5).

Assumption 9 (Concave dynamic energy consumption)**.** *$f_i : [0,1] \to \mathbb{R}_+ \, (i \in \mathcal{M})$, is concave and continuously differentiable.*

In particular, this assumption is satisfied by a linear dependency of the base station energy consumption and the cell load reported in current studies such as [ABG+10, CZB+10]. In fact, the optimization framework presented in the following can handle more general functions for the load-dependent dynamic energy consumption.

Remark 5. *The load dependent dynamic energy consumption can be assumed to be a convex function of the load. Moreover, we could even assume that it is a sum of convex and concave functions. The optimization framework can be straightforwardly extended to cover these cases.*

From (3.2) and Definition 10 we can observe that each load ρ_i is, in fact, a function of $\mathbf{X}$. The cell load at cell i is non-zero and $|\rho_i|_0 = 1$, if at least one TP is served by cell i (i.e., $\sum_{j\in\mathcal{N}} x_{i,j} \geq 1$). We can therefore modify the network energy consumption model (3.3) to have only $\mathbf{X}$ as a variable. We can equivalently write

$$\begin{aligned} E(\boldsymbol{\rho}) &= \sum_{l\in\mathcal{L}} \left(c_l \left| \sum_{i\in\mathcal{S}_l} \rho_i \right|_0 + \sum_{i\in\mathcal{S}_l} \left(e_i \left|\rho_i\right|_0 + f_i(\rho_i) \right) \right) \\ &= \sum_{l\in\mathcal{L}} \left(c_l \left| \sum_{i\in\mathcal{S}_l} \sum_{j\in\mathcal{N}} x_{i,j} \right|_0 + \sum_{i\in\mathcal{S}_l} \left(e_i \left| \sum_{j\in\mathcal{N}} x_{i,j} \right|_0 + f_i(\rho_i) \right) \right) \\ &= \sum_{l\in\mathcal{L}} \left(c_l \left| \boldsymbol{t}_l^T \tilde{\boldsymbol{x}} \right|_0 + \sum_{i\in\mathcal{S}_l} \left(e_i \left| \boldsymbol{s}_i^T \tilde{\boldsymbol{x}} \right|_0 + f_i(\rho_i) \right) \right), \end{aligned} \tag{3.4}$$

where $\boldsymbol{s}_i := \mathrm{vec}(\mathbf{S}_i)$ with $\mathbf{S}_i \in \{0,1\}^{M\times N}$ being a matrix of zeros, except for its ith row, which is a row of ones; $\boldsymbol{t}_l := \mathrm{vec}(\mathbf{T}_l)$ with $\mathbf{T}_l \in \{0,1\}^{M\times N}$ is a matrix of zeros, except for its rows $i \in \mathcal{S}_l$, which are rows of ones. Remember that we use $\tilde{\boldsymbol{x}}$ to denote the vector obtained by stacking up the columns of $\mathbf{X}$. The first equality in (3.4) follows from Fact 3 and the definition of the l_0-norm, which does not account for magnitudes. More precisely, if at least one TP is served by cell i (i.e., $\sum_{j\in\mathcal{N}} x_{i,j} \geq 1$), then the cell load at cell i is non-zero $\rho_i > 0$ and we have $|\rho_i|_0 = \left|\sum_{j\in\mathcal{N}} x_{i,j}\right|_0 = 1$. The second equality in (3.4) uses vector multiplication to represent the sums in a more compact way.

Definition 13. *Given the assignment $\tilde{\boldsymbol{x}}$ and the load dependent energy consumption $f_i(\rho_i(\tilde{\boldsymbol{x}}))$ of cell i, we define the function $\tilde{f}_i : [0,1]^{NM} \to \mathbb{R}_+ : \tilde{\boldsymbol{x}} \mapsto f_i(\rho_i(\tilde{\boldsymbol{x}}))$.*

Thus, the energy consumption in (3.3) is equivalently obtained by

$$\tilde{E}(\mathbf{X}) = \sum_{l\in\mathcal{L}} \left(c_l \left| \boldsymbol{t}_l^T \tilde{\boldsymbol{x}} \right|_0 + \sum_{i\in\mathcal{S}_l} \left(e_i \left| \boldsymbol{s}_i^T \tilde{\boldsymbol{x}} \right|_0 + \tilde{f}_i(\tilde{\boldsymbol{x}}) \right) \right). \tag{3.5}$$

3.3.2 Problem statement

Daytime fluctuations in traffic demand result in large spatio-temporal redundancies in coverage and capacity opening great opportunities for energy savings. In times when the traffic demand decreases and thus some entries in the rate requirement vector $\boldsymbol{r} \in \mathbb{R}^N_{++}$ become relatively small or zero, redundant cells can be deactivated reducing the provided capacity to match $\boldsymbol{r}$. Finding the optimal set of redundant cells for deactivation and thus, minimizing the total network energy consumption can be achieved by minimizing the cost function in (3.3) subject to different constraints that follow from the system model and technology specifics. By $\mathcal{X}_1$ we denote the set of assignments $\mathbf{X} \in \{0,1\}^{M \times N}$ satisfying some required constraints (e.g., capacity constraints) including $\mathcal{C}_1$ from (3.1) and $\mathcal{C}_2$ from (3.2). The specific constraints are technology dependent and are presented in Section 3.4 and Section 3.5 for single-RAT LTE and multi-RAT LTE/UMTS, respectively. For now we assume that the constraints belong to a convex set or can be relaxed to a convex set.

Assumption 10. *The set of constraints $\mathcal{X}_1$ is convex.*

Now, the problem under consideration can be formally stated as follows:

$$\text{minimize } \tilde{E}(\mathbf{X}) = \sum_{l \in \mathcal{L}} \left(c_l \left| \boldsymbol{t}_l^T \tilde{\boldsymbol{x}} \right|_0 + \sum_{i \in \mathcal{S}_l} \left(e_i \left| \boldsymbol{s}_i^T \tilde{\boldsymbol{x}} \right|_0 + \tilde{f}_i(\tilde{\boldsymbol{x}}) \right) \right) \tag{3.6a}$$

$$\text{subject to } \mathbf{X} \in \mathcal{X}_1 \tag{3.6b}$$

$$\mathbf{X} \in \{0,1\}^{M \times N} \,, \tag{3.6c}$$

where the optimization variables are $\mathbf{X}$.

We consider scenarios where the rate requirements of TPs are sufficiently low resulting in a reasonable amount of redundancies that allow for deactivation of cells. In the problem, we assume that there exists a feasible solution. If no feasible solution exists, we can add slack variables and an additional l_0 or l_1 penalty norm to the merit function in order to create feasible solutions indicating which cells are overloaded. Such extensions are presented in Appendix B.3. Moreover, if the traffic requirements in the system are sufficiently low or the number of cells is sufficiently large, $\mathbf{X}$ is expected to be *row sparse* with all zero rows specifying cells that can be deactivated.

3.3.3 Problem solution

The difficulty of Problem 3.6 lies in its combinatorial nature. In fact, Problem 3.6 can be shown to be closely related to the classical bin-packing problem which is known to be NP-hard (non-deterministic polynomial-time hard), the connection of which is shown in Appendix B.1. Consequently, the complexity is expected to grow exponentially with the

number of cells. For such problems, we can make use of branch and bound algorithms to find optimal solutions. However, such algorithms are typically slow and it takes a very long time to solve even fairly small problems [JB09]. On the positive side, Problem 3.6 has a special structure that can be exploited by majorization-minimization techniques [HL04], which have been widely used in recent years to tackle various problems in compressed sensing [CWB08] and machine learning [STL11].

Instead of finding a global solution to Problem 3.6, we will pursue a less ambitious goal. We apply majorization-minimization techniques mentioned above to develop a low-complexity anytime algorithm that has a strong analytical justification. This algorithm is expected to provide good results (in terms of low energy consumption) with low complexity. To this end, we reformulate and relax Problem 3.6 to pose it in a more tractable form.

Problem reformulation and relaxation

To obtain an optimization problem that is computationally tractable, we first replace the binary constraint in (3.6c) by[6]

$$\mathbf{X} \in [0,1]^{M \times N}. \tag{3.7}$$

The above in combination with Assumption 10 makes the set of all constraints convex and the solution to the resulting optimization problem serves as a guideline for other heuristics to make the hard discrete decisions on which cell serves a particular test point.

So now the only problem is the objective function, which is not continuous due to the l_0-norm. We also note that by Assumption 9 and Definition 13, the load-dependent term $\tilde{f}_i(\tilde{\boldsymbol{x}})$ in the objective function (3.6a) is concave and continuously differentiable for $\tilde{\boldsymbol{x}} \in [0,1]^{MN}$ since these properties are preserved under a composition with a linear function [BC11, BV06]. To address the non-continuity of the l_0-norm, we consider the following relation [CWB08, STL11]:

$$\forall_{\boldsymbol{x} \in \mathbb{R}^M} \ |\boldsymbol{x}|_0 = \lim_{\epsilon \to 0} \sum_{i=1}^{M} \frac{\log(1 + |x_i| \ \epsilon^{-1})}{\log(1+\epsilon^{-1})}, \tag{3.8}$$

which, if each component x_i of the vector $\boldsymbol{x} = [x_1, \ldots, x_M]^T$ is constrained to be non-negative as in (3.7), suggests the use of a function $f_\epsilon : \mathbb{R}_+^M \to \mathbb{R}$ to equivalently reformulate the l_0-norm by

$$\forall_{\boldsymbol{x} \in \mathbb{R}_+^M} \ |\boldsymbol{x}|_0 = \lim_{\epsilon \to 0} \sum_{i=1}^{M} \frac{\log(\epsilon + x_i) - \log(\epsilon)}{\log(1+\epsilon^{-1})}. \tag{3.9}$$

[6] This relaxation together with (3.1) leads to a communication scenario where multiple cells are allowed to serve the traffic requested at a TP by providing only a fraction each. A more detailed discussion on the implications is presented in Section 3.3.5

The choice of (3.9) is not the only choice but we will see that it has some neat properties that we can exploit. For a discussion on the choices for the l_0-norm approximations, we refer to Appendix A.1.

We can now write the cost function (3.6a) equivalently

$$\begin{aligned}&\sum_{l\in\mathcal{L}}\left(c_l\left|\boldsymbol{t}_l^T\tilde{\boldsymbol{x}}\right|_0+\sum_{i\in\mathcal{S}_l}e_i\left|\boldsymbol{s}_i^T\tilde{\boldsymbol{x}}\right|_0+\tilde{f}_i(\tilde{\boldsymbol{x}})\right)\\&=\lim_{\epsilon\to0}\sum_{l\in\mathcal{L}}\left(c_l\frac{\log(\epsilon+\boldsymbol{t}_l^T\tilde{\boldsymbol{x}})-\log(\epsilon)}{\log(1+\epsilon^{-1})}+\sum_{i\in\mathcal{S}_l}\left(e_i\frac{\log(\epsilon+\boldsymbol{s}_i^T\tilde{\boldsymbol{x}})-\log(\epsilon)}{\log(1+\epsilon^{-1})}+\tilde{f}_i(\tilde{\boldsymbol{x}})\right)\right),\end{aligned}\tag{3.10}$$

where we replaced the l_0-norm in Problem 3.6 by (3.8) and used the non-negativity of $\boldsymbol{s}_i$, $\boldsymbol{t}_i$, $\tilde{\boldsymbol{x}}$. We can therefore obtain an approximation to Problem 3.6 by replacing the objective function by the right-hand side of (3.10) for a sufficiently small but fixed $\epsilon>0$ and ignoring unnecessary constants. More precisely, for some $\epsilon>0$, the objective is to find a matrix $\mathbf{X}\in[0,1]^{M\times N}$ or, equivalently, a vector $\tilde{\boldsymbol{x}}=\mathrm{vec}(\mathbf{X})\in[0,1]^{NM}$ that solves the following problem

$$\text{minimize}\quad\sum_{l\in\mathcal{L}}\left(c_l\frac{\log(\epsilon+\boldsymbol{t}_l^T\tilde{\boldsymbol{x}})}{\log(1+\epsilon^{-1})}+\sum_{i\in\mathcal{S}_l}\left(e_i\frac{\log(\epsilon+\boldsymbol{s}_i^T\tilde{\boldsymbol{x}})}{\log(1+\epsilon^{-1})}+\tilde{f}_i(\tilde{\boldsymbol{x}})\right)\right)\tag{3.11a}$$

$$\text{subject to}\quad\mathbf{X}\in\mathcal{X}_1\tag{3.11b}$$

$$\mathbf{X}\in[0,1]^{M\times N}.\tag{3.11c}$$

Analyzing the structure of Problem 3.11 we see that solving it is not a straightforward task as we need to *minimize* a non-convex function over a convex set. However, problems of this type are amenable to majorization-minimization techniques [HL04] widely used to tackle various problems in compressed sensing [CWB08] and machine learning [STL11]. This framework provides us with a computationally efficient tool to decrease the value of the objective function. We provide the details of the majorization-minimization algorithm in Appendix C and apply it to our problem in the next section.

Majorization-minimization (MM) algorithm

For notational convenience we define $\hat{c}_l:=\frac{c_l}{\log(1+\epsilon^{-1})}$ and $\hat{e}_i:=\frac{e_i}{\log(1+\epsilon^{-1})}$, and we use these definitions in (3.11a) to simplify the objective function:

$$h:\mathcal{X}\to\mathbb{R},\quad h(\tilde{\boldsymbol{x}})=\sum_{l\in\mathcal{L}}\left(\hat{c}_l\log(\epsilon+\boldsymbol{t}_l^T\tilde{\boldsymbol{x}})\right)+\sum_{i\in\mathcal{M}}\left(\hat{e}_i\log(\epsilon+\boldsymbol{s}_i^T\tilde{\boldsymbol{x}})+\tilde{f}_i(\tilde{\boldsymbol{x}})\right),\tag{3.12}$$

where $\mathcal{X}\subset\mathbb{R}^{MN}$ is the closed convex set of points satisfying the constraints (3.11b) and (3.11c) and we have used the fact that $\mathcal{M}=\cup_{l\in\mathcal{L}}\mathcal{S}_l$. It is now easy to see that for any $\epsilon>0$, (3.12) is a concave and continuously differentiable function. Remember that the load-dependent term $\tilde{f}_i(\tilde{\boldsymbol{x}})$ is also concave and continuously differentiable by

Assumption 9 and Definition 13. Following the explanations in the Appendix C, the function

$$g : \mathcal{X} \times \mathcal{X} \to \mathbb{R} : (\boldsymbol{x}, \boldsymbol{y}) \mapsto h(\boldsymbol{y}) + \nabla h(\boldsymbol{y})^T (\boldsymbol{x} - \boldsymbol{y}) \tag{3.13}$$

is used as a majorizing function of (3.12). Its gradient can be easily calculated as

$$\nabla h(\tilde{\boldsymbol{x}}) = \sum_{l \in \mathcal{L}} \hat{c}_l \frac{1}{\epsilon + \boldsymbol{t}_l^T \tilde{\boldsymbol{x}}} + \sum_{i \in \mathcal{M}} \Big(\hat{e}_i \frac{1}{\epsilon + \boldsymbol{s}_i^T \tilde{\boldsymbol{x}}} + \nabla \tilde{f}_i(\tilde{\boldsymbol{x}}) \Big). \tag{3.14}$$

which yields the core problem solved in each iteration of the MM algorithm:

$$\begin{aligned} \tilde{\boldsymbol{x}}^{(n+1)} &\in \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{X}} g(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}^{(n)}) \\ &= \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{X}} \sum_{l \in \mathcal{L}} \hat{c}_l \frac{\boldsymbol{t}_l^T \tilde{\boldsymbol{x}}}{\epsilon + \boldsymbol{t}_l^T \tilde{\boldsymbol{x}}^{(n)}} + \sum_{i \in \mathcal{M}} \left(\hat{e}_i \frac{\boldsymbol{s}_i^T \tilde{\boldsymbol{x}}}{\epsilon + \boldsymbol{s}_i^T \tilde{\boldsymbol{x}}^{(n)}} + \nabla \tilde{f}_i(\tilde{\boldsymbol{x}}^{(n)})^T \tilde{\boldsymbol{x}} \right) \end{aligned} \tag{3.15}$$

for some feasible starting point[7] $\tilde{\boldsymbol{x}}^{(0)} \in \mathcal{X}$. This concludes the derivation of the MM algorithm. We have chosen the majorizing function (3.13) such that the obtained MM algorithm solves iteratively a sequence of convex optimization problems. In fact, the optimization problem solved in every iteration is a linear programming problem (LP), which can be typically solved efficiently with standard optimization tools.

As discussed in the appendix, the sequence $\{\tilde{\boldsymbol{x}}^{(n)}\}_{n \in \mathbb{N}} \subset \mathcal{X}$ for some $\tilde{\boldsymbol{x}}^{(0)} \in \mathcal{X}$ generated by (3.15) produces a non-increasing sequence $\{h(\tilde{\boldsymbol{x}}^{(n)})\}_{n \in \mathbb{N}}$ of objective values. Therefore, as $n \to \infty$, we expect the corresponding sequence of assignment matrices $\{\boldsymbol{X}^{(n)}\}_{n \in \mathbb{N}}$ (note that $\tilde{\boldsymbol{x}}^{(n)} =: \mathrm{vec}(\mathbf{X}^{(n)})$) to evolve towards network configurations with low energy consumption. We stop the MM algorithm if the improvements in the objective value are small enough, in the sense that for some sufficiently small $\epsilon^\star > 0$, the following condition is met

$$h\big(\tilde{\boldsymbol{x}}^{(n)}\big) - h\big(\tilde{\boldsymbol{x}}^{(n+1)}\big) \leq \epsilon^\star. \tag{3.16}$$

Upon termination, the MM algorithm gives a feasible solution $\mathbf{X}^{(n)} \in [0,1]^{M \times N}$ to Problem 3.11. In order to obtain a feasible point to Problem 3.6, we need to map $\mathbf{X}^{(n)}$ to a matrix $\mathbf{X}^\star \in \{0,1\}^{M \times N}$. For this purpose, we propose a heuristic described in Algorithm 1. The main idea can be summarized as follows: we consecutively connect each test point to the cell corresponding to the respective largest entry in $\mathbf{X}^{(n)}$. If the obtained intermediate assignment matrix does not belong to the feasible set $\mathcal{X}_1 \backslash \mathcal{C}_1$ of Problem 3.6 we try to connect the test point to the cell corresponding to the second largest entry in $\mathbf{X}^{(n)}$, and so forth. If this approach fails, we activate additional cells and connect TPs to them.

Remark 6. *The solutions $\mathbf{X}^{(n)} \in [0,1]^{M \times N}$ in our simulations obtained with the help of the standard LP solver of CPLEX, are typically either zero or one. Thus, the assignment*

[7] In our experience a good starting point is derived from a feasible assignment matrix obtained by connecting each TP to the cell providing the strongest received signal strength.

Algorithm 1 Heuristic to map $[0,1]^{M\times N} \to \{0,1\}^{M\times N}$

Input: $\mathbf{X}^{(n)}$, $\mathcal{N}$, $\mathcal{M}$, set of constraints $\mathcal{X}_2$ representing (3.2) and (3.6c)
Output: final assignment matrix $\mathbf{X}^\star$

initialize: set of assigned TPs $\mathcal{A} = \emptyset$ and final assignment matrix $\mathbf{X}^\star = \mathbf{0}$.
for all $i \in \mathcal{M}, j \in \mathcal{N}$ **do**
  if $x_{i,j}^{(n)} \in \{1\}$ **then**
    $x_{i,j}^\star = x_{i,j}^{(n)}$ and $\mathcal{A} = \mathcal{A} \cup \{j\}$.
  end if
end for
Define set $\mathcal{B} = \left\{ x_{i,j}^{(n)} \in (0,1) \,|\forall i \in \mathcal{M}, \forall j \in \mathcal{N}\backslash\mathcal{A} \right\}$.
while $\mathcal{B} \neq \emptyset$ **do**
  $(i,j) = \mathrm{a}rgmax_{i,j}\{\mathcal{B}\}$
  if $x_{i,j}^\star := 1 \to \mathbf{X}^\star \in \mathcal{X}_2$ **then**
    $x_{i,j}^\star = 1$ and $\mathcal{A} = \mathcal{A} \cup \{j\}$.
    $\mathcal{B} = \mathcal{B}\backslash\{x_{i,j}^{(n)} \,|\forall i \in \mathcal{M}\}$
  else
    $\mathcal{B} = \mathcal{B}\backslash\{x_{i,j}^{(n)}\}$
  end if
end while
for all $j \notin \mathcal{A}$ **do**
  activate closest non-active cell i which yields $x_{i,j}^\star := 1 \to \mathbf{X}^\star \in \mathcal{X}_2$ and assign $x_{i,j}^\star = 1$.
  $\mathcal{A} = \mathcal{A} \cup \{j\}$.
end for

Algorithm 2 Network reconfiguration for improved energy efficient operation

Input: set of TPs, set of cells, constraints
Output: optimized network configuration according to $\mathbf{X}^\star$.

initialize $\mathbf{X}^{(0)}$ with a feasible point.
repeat
  compute $\tilde{\boldsymbol{x}}^{(n)}$ by solving (3.15)
  increment n
until (3.16) is valid
use Algorithm 1 to map $\mathbf{X}^{(n)}$ to $\mathbf{X}^\star \in \{0,1\}^{M\times N}$
connect the TPs to cells according to $\mathbf{X}^\star$.
deactivate all cells no TP is connected to.

of test points to cells with the largest entries in $\mathbf{X}^{(n)}$ *rarely results in a violation of a constraint (but we emphasize that this is not guaranteed to be true in general).*

There are many alternatives to the heuristic presented in Algorithm 1 and we present one alternative which is based on a modified objective function with penalty functions in Appendix B.2.

For convenience, we summarize the complete approach derived in this section in Algorithm 2.

3.3.4 Notes on the convergence and the complexity

Although we have no guarantee that Algorithm 2 presented in Section 3.3.3 converges to the global optimum, based on the theory of EM/MM algorithms, we can show the following characteristics. Note that

$$h(\tilde{\boldsymbol{x}}) := \sum_{l\in\mathcal{L}} \left(\hat{c}_l \log(\epsilon + \boldsymbol{t}_l^T \tilde{\boldsymbol{x}})\right) + \sum_{i\in\mathcal{M}} \left(\hat{e}_i \log(\epsilon + \boldsymbol{s}_i^T \tilde{\boldsymbol{x}}) + \tilde{f}_i(\tilde{\boldsymbol{x}})\right) \tag{3.17}$$

is monotonically decreasing by design (c.f. Appendix) and bounded from below, i.e., $h(\tilde{\boldsymbol{x}}) \geq \sum_{l\in\mathcal{L}} \hat{c}_l \log\epsilon + \sum_{i\in\mathcal{M}} \hat{e}_i \log\epsilon, \forall \tilde{\boldsymbol{x}} \in \mathcal{X}$, so the sequence $\{h(\tilde{\boldsymbol{x}}^{(n)})\}_{n\in\mathbb{N}}$ converges by the *Monotone Convergence Theorem*. We emphasize that this does not imply a convergence of $\{\tilde{\boldsymbol{x}}^{(n)}\}_{n\in\mathbb{N}}$. For further properties of the sequence $\{\tilde{\boldsymbol{x}}^{(n)}\}_{n\in\mathbb{N}}$, we refer the reader to [Wu83].

The complexity of the algorithm is of the same order of solving iteratively linear programming problems, which is a class of problems that can be solved efficiently with many standard optimization tools [BV06]. In our simulations for this task in Section 3.7, we use CPLEX, which implements the dual simplex algorithm to solve LPs [IBM15]. Typically, our proposed algorithm terminates after a few iterations ($\ll 100$) as seen in Section 3.7. The complexity of the proposed algorithm is linear in the complexity of the simplex method, which has a polynomial time complexity on average and an exponential time worst-case complexity. In contrast, integer programming problems are typically solved by branch and cut algorithms (also in CPLEX [IBM15]), which have an upper bound on the number of nodes 2^n, where n is the number of variables of the problem, and solve one LP per node resulting in an exponential complexity.

3.3.5 Serving a test point with multiple cells

By Assumption 8 the used system model outlined in Section 3.3.1 dictates that each TP is served by exactly one cell. However, this assumption may be restrictive and limits the energy savings capabilities in the network. Consider the following simple example: Two cells serve TPs in a common coverage area and both cells are heavily loaded. Now the requirements of a TP increase to a level that cannot be fulfilled by either of the cells individually. In the standard scenario an additional cell would have to be switched on in order to provide the requested service to all TPs or if no cell can be switched on the requirements cannot be met. However, if it were possible to split the traffic between the two highly loaded cells, they might have enough resources to serve all users and no additional cell would have to be switched on. This example highlights the benefit of using a joint service provision for users with multiple cells in the same TP area.

The shared service of a TP, as described above, can be easily implemented by lifting Assumption 8. The benefit of it is in fact twofold. Assumption 8 introduces the non-convex constraint (3.6c) to the optimization Problem 3.6, which motivates the relaxation (3.7)

and the heuristic mapping introduced in Algorithm 1. To avoid these heuristic approaches for which we are not guaranteed to find solutions, we assume in this section that each TP can be served by multiple cells. This assumption is implemented by using (3.7) directly instead of (3.6c) in Problem 3.6. As a result, there is no need for a relaxation of the constraints or the use of heuristic mappings such as that in Algorithm 1. We only need to approximate the cost function as done in (3.11a) and apply the MM algorithm to the resulting optimization problem. We note that these operations have a strong analytical justification.

The assumption of multiple cells serving one TP has a practical interpretation when considering Definition 9. It means that cells are allowed to serve only a fraction of the traffic generated in the area corresponding to some TP. In other words, we do not use an all-or-nothing approach, where cells should serve either all users or no users in the area corresponding to a TP.

3.4 Single-RAT network topology control

In this section we show how to apply the general load-aware energy savings framework introduced in Section 3.3 to a single-RAT LTE network. The technology specific constraints are captured by $\mathcal{C}_2$ where we now define $b_{i,j}$ in (3.2) for LTE. LTE networks implement OFDMA for the downlink communication and thus we adopt an OFDMA-based model for the spectral efficiency that is widely used in literature [MK10, MNK$^+$07, FF13]. The spectral efficiency also depends on radio propagation properties. Therefore, we associate a path-loss vector to each TP and write the path-loss vectors of all TPs as columns of the path-loss matrix $\mathbf{G} = [g_{i,j}] \in \mathbb{R}_{++}^{M \times N}$, where $g_{i,j}$ captures the long-term path loss and shadowing effects for a radio link from cell i to TP j.

Assumption 11 (Reliable path-loss estimates)**.** *A reliable estimate of* $\mathbf{G}$ *is available at the central network controller.*

Remark 7. *The problem of reliable estimation and tracking of the path-loss matrix is out of the scope of this work. However, the matrix captures only long-term fading effects, so reliable estimates of* $\mathbf{G}$ *can be obtained and tracked in practice. Promising algorithmic solutions to this estimation problem are for instance presented in [KCV$^+$16]. Moreover, in network planning problems, knowledge of* $\boldsymbol{G}$ *is a very common assumption in the literature [SY12, MK10, ACMS03].*

Now, we are in a position to define the SINR $\gamma_{i,j} : \mathbb{R}_+^M \to \mathbb{R}_+$ between cell $i \in \mathcal{M}$ and TP $j \in \mathcal{N}$ which is coupled by the load levels at other cells and is captured by [HYLS15, CSS$^+$14, MK10, SY12, FKVF13, MTHB07]:

$$\gamma_{i,j}(\boldsymbol{\rho}) = \frac{P_i \; g_{i,j}}{\sum_{k \in \mathcal{M} \setminus \{i\}} P_k \; g_{k,j} \; \rho_k + \sigma^2}, \tag{3.18}$$

where $P_i > 0$ is the transmit power per resource unit of cell i and $\sigma^2 > 0$ is the noise power per resource block. Accordingly, the link spectral efficiency $\omega_{i,j} : \mathbb{R}_+^M \to \mathbb{R}_+$ (in bits per resource block[8]) for the link from cell i to TP j is given by [MNK+07]

$$\omega_{i,j}(\boldsymbol{\rho}) = \eta_{i,j}^{\mathrm{BW}} \log_2\Big(1 + \frac{\gamma_{i,j}(\boldsymbol{\rho})}{\eta_{i,j}^{\mathrm{SINR}}}\Big), \tag{3.19}$$

where $\eta_{i,j}^{\mathrm{BW}} \in \mathbb{R}_{++}$ and $\eta_{i,j}^{\mathrm{SINR}} \in \mathbb{R}_{++}$ are suitably chosen constants, referred to as bandwidth and SINR efficiency, respectively. These constants represent system design choices, such as the chosen multi-antenna techniques, scheduling protocols, modulation and coding schemes. For realistic values of these constants, we refer the interested reader to [MK10, MNK+07] and assume in our study that they are arbitrary and fixed since the particular choice has no impact on our results.

We are now prepared to use (3.19) to compose the technology-specific constraints $\mathcal{C}_2$ for LTE by defining $b_{i,j} := \frac{r_j}{\omega_{i,j}(\boldsymbol{\rho})}$ which is used in (3.2) yielding the following system of non-linear equations

$$\rho_i = \sum_{j \in \mathcal{N}_i(\mathbf{X})} \frac{b_{i,j}}{B_i} = \sum_{j \in \mathcal{N}_i(\mathbf{X})} \frac{r_j}{B_i \omega_{i,j}(\boldsymbol{\rho})} = \sum_{j \in \mathcal{N}} \frac{r_j}{B_i \omega_{i,j}(\boldsymbol{\rho})} x_{i,j},\ i \in \mathcal{M}. \tag{3.20}$$

One of the main complications to the computation of the cell load is the interference coupling appearing in (3.20). For a fixed assignment $\mathbf{X}$, the cell load $\boldsymbol{\rho}$ in (3.20) can be efficiently computed by means of fixed-point algorithms (c.f. Chapter 4). However, the assignment of TPs to cells is the main subject of our optimization problem and thus we cannot evaluate (3.20) easily. In order to keep the complexity of the optimization problem tractable, we fix the spectral efficiency of the link connecting cell i and test point j to some constant $\tilde{\omega}_{i,j}$. We take a conservative approach and consider the worst-case interference which constitutes a lower bound on the spectral efficiency $\omega_{i,j}(\boldsymbol{\rho}) \geq \tilde{\omega}_{i,j} := \omega_{i,j}(\mathbf{1})$ for every $\boldsymbol{\rho} \in [0,1]^M$.

Assumption 12 (Worst-Case Interference)**.** *We have the worst-case interference scenario if all cells are fully loaded, i.e.* $\boldsymbol{\rho} = \mathbf{1}$.

In general, this worst-case interference assumption and the resulting lower bound on the true link spectral efficiency diminishes gains in energy savings when taking into account the energy consumption of hardware, and we show in Section 4 how to incorporate the actual link spectral efficiency to improve the energy savings. Nevertheless, having fully loaded cells as in Assumption 12 is in some cases desirable because it has been proven in [HYLS15] that full load (i.e. $\boldsymbol{\rho} = \mathbf{1}$) is optimal with respect to the transmit energy consumption (see also [CPS14]) without taking into account the energy consumed by hardware.

[8] A resource block is defined as a portion of the available time-frequency plane spanning a number of consecutive OFDM symbols in the time domain over a number of subcarriers in the frequency domain.

Remark 8. *The worst-case interference assumption cannot exploit the full potential for energy savings, but the assumption is of high practical relevance because it is an effective way to avoid coverage holes as a result of deactivating cells based, for instance, on imperfect information.*

With the above constraints used in Problem 3.6 we can state the load-aware energy savings problem for a single-RAT LTE network as follows:

$$\text{minimize} \quad \sum_{l\in\mathcal{L}} \left(c_l \left| \boldsymbol{t}_l^T \tilde{\boldsymbol{x}} \right|_0 + \sum_{i\in\mathcal{S}_l} \left(e_i \left| \boldsymbol{s}_i^T \tilde{\boldsymbol{x}} \right|_0 + \tilde{f}_i(\tilde{\boldsymbol{x}}) \right) \right) \tag{3.21a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_j}{B_i \ \tilde{\omega}_{i,j}} \ x_{i,j} \leq 1 \qquad i \in \mathcal{M} \tag{3.21b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j} = 1 \qquad j \in \mathcal{N} \tag{3.21c}$$

$$\mathbf{X} \in \{0,1\}^{M\times N}, \tag{3.21d}$$

where the optimization variables are $x_{i,j}$. In particular, (3.21b) is derived from (3.20), Assumption 12 and Definition 10 which ensures that cells are not overloaded, whereas (3.21c) together with (3.21d) come from Assumption 8.

Remark 9. *Without the worst-case interference Assumption 12, problem 3.21 has highly non convex constraints. It is however possible to relax the constraints to obtain quadratical constraints and state the problem as a quadratically constraint problem in standard form. Unfortunately, we cannot guarantee that the constraint matrix is positive-semidefinite which makes the solution of such a problem cumbersome. Therefore, we opt for another relaxation technique presented in this work that has nicer properties in terms of computational complexity.*

After applying the reformulations to the objective function and the relaxation of (3.21d) presented in Section 3.3.3 we arrive at the following problem:

$$\text{minimize} \quad \sum_{l\in\mathcal{L}} \left(c_l \frac{\log(\epsilon + \boldsymbol{t}_l^T \tilde{\boldsymbol{x}})}{\log(1+\epsilon^{-1})} + \sum_{i\in\mathcal{S}_l} \left(e_i \frac{\log(\epsilon + \boldsymbol{s}_i^T \tilde{\boldsymbol{x}})}{\log(1+\epsilon^{-1})} + \tilde{f}_i(\tilde{\boldsymbol{x}}) \right) \right) \tag{3.22a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_j}{B_i \ \tilde{\omega}_{i,j}} \ x_{i,j} \leq 1 \qquad i \in \mathcal{M} \tag{3.22b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j} = 1, \qquad j \in \mathcal{N} \tag{3.22c}$$

$$\mathbf{X} \in [0,1]^{M\times N}. \tag{3.22d}$$

The above problem is amenable to the optimization framework presented in Section 3.3.3. More precisely, we can find good assignments by generating a sequence $\{\tilde{\boldsymbol{x}}^{(n)}\}_{n\in\mathbb{N}}$ that is obtained by iteratively solving

$$\text{minimize} \quad \sum_{l\in\mathcal{L}} \frac{\hat{c}_l \boldsymbol{t}_l^T \tilde{\boldsymbol{x}}^{(n+1)}}{\epsilon + \boldsymbol{t}_l^T \tilde{\boldsymbol{x}}^{(n)}} + \sum_{i\in\mathcal{M}} \left(\frac{\hat{e}_i \boldsymbol{s}_i^T \tilde{\boldsymbol{x}}^{(n+1)}}{\epsilon + \boldsymbol{s}_i^T \tilde{\boldsymbol{x}}^{(n)}} + \nabla \tilde{f}_i(\tilde{\boldsymbol{x}}^{(n)})^T \tilde{\boldsymbol{x}}^{(n+1)} \right) \tag{3.23a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \tilde{\omega}_{i,j}}\ x_{i,j}^{(n+1)} \leq 1 \qquad i \in \mathcal{M} \tag{3.23b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j}^{(n+1)} = 1, \qquad j \in \mathcal{N} \tag{3.23c}$$

$$\mathbf{X}^{(n+1)} \in [0,1]^{M\times N}\ , \tag{3.23d}$$

with a feasible starting point $\mathbf{X}^{(0)} \in (3.23b) \cup (3.23c) \cup (3.23d)$ and using a termination criterion as in (3.16).

3.5 Multi-RAT network topology control

In this section we turn to a multi-RAT system where two separate networks of different radio access technologies are overlaid on the same geographical region. In particular, we consider the multi-RAT scenario where an UMTS system and a LTE system provide overlapping coverage to the users. To be consistent with the notation of previous sections, we use the subscript "UMTS" and "LTE" for respective base stations and their cells. More precisely, we have $L_{\text{UMTS}} = |\mathcal{L}_{\text{UMTS}}|$ and $L_{\text{LTE}} = |\mathcal{L}_{\text{LTE}}|$ denoting the number of all UMTS and LTE base stations respectively. To avoid notational clutter, we assume that each base station has exactly one cell, such that $|\mathcal{L}_{\text{UMTS/LTE}}| = |\mathcal{M}_{\text{UMTS/LTE}}| = M_{\text{UMTS/LTE}}$. In the text that follows we will only refer to base stations and note that the extension to multiple cells per base station is straightforwardly derived from findings in previous sections. We define $\mathcal{L} = \mathcal{L}_{\text{UMTS}} \cup \mathcal{L}_{\text{LTE}}$ and use $L = |\mathcal{L}| = L_{\text{UMTS}} + L_{\text{LTE}}$ to denote the total number of base stations. The test point concept is used to capture the QoS requirements of all N users in the network as outlined in Section 3.3.1. The task of our multi-RAT network optimization is to find an assignment matrix $\mathbf{X} \in \{0,1\}^{L\times N}$ that supports all user requirements and leads a low energy consumption. Thereby, the energy consumption in (3.5) of the multi-RAT network is obtained by

$$E_{\text{multi-RAT}}(\mathbf{X}) = \sum_{l\in\mathcal{L}} (c_l + e_l) \left| \boldsymbol{t}_l^T \tilde{\boldsymbol{x}} \right|_0 + \tilde{f}_l(\tilde{\boldsymbol{x}}), \tag{3.24}$$

where $\tilde{\boldsymbol{x}} = \text{vec}(\mathbf{X})$ and $\boldsymbol{t}_l := \text{vec}(\mathbf{T}_l)$ with $\mathbf{T}_l \in \{0,1\}^{L\times N}$ is a matrix of zeros, except for its rows $i \in \mathcal{S}_l$, which are rows of ones.

The LTE part of the multi-RAT system is modeled as described in Section 3.4 where the assumptions and definitions are extended to multi-RAT where needed. We now describe the system model for the UMTS part of the multi-RAT system. In contrast to the LTE system, UMTS is a WCDMA based system. Thus, we express the QoS in terms of SINR per code of the WCDMA system that is needed to provide a certain data rate over that

link. Thus, the data rate requirement of test points can be captured by the number of codes required. In more detail, the system is able to provide a data rate per code of r_{base} with a downlink SINR of $\beta_{\min}$ according to the implemented modulation and coding scheme (MCS), which we assume to be fixed. Therefore, the required data rate of test point j can be provided with $u_j = \frac{r_j}{r_{\text{base}}}$ codes. The downlink SINR $\beta_{i,j} : [0,1]^{L\times N} \to \mathbb{R}_{++}$ of a link from base station i to a test point j is computed as (c.f. [ACM03])

$$\begin{aligned}\beta_{i,j}(\mathbf{X}) &= \frac{P_{\text{RX}}}{\alpha\left(\sum_{k\in\mathcal{N}} u_k \frac{P_{\text{RX}} g_{i,j}}{g_{i,k}} x_{i,k} - P_{\text{RX}} u_j\right) + \sum_{l\neq i}\sum_{k\in\mathcal{N}} u_k \frac{P_{\text{RX}} g_{l,j}}{g_{l,k}} x_{l,k} + \sigma_{i,j}^2} \\ &= \left(\alpha\left(\sum_{k\in\mathcal{N}} u_k \frac{g_{i,j}}{g_{i,k}} x_{i,k} - u_j\right) + \sum_{l\neq i}\sum_{k\in\mathcal{N}} u_k \frac{g_{l,j}}{g_{l,k}} x_{l,k} + \tilde{\sigma}_{i,j}^2\right)^{-1},\end{aligned} \tag{3.25}$$

with P_{RX} the target receive power and $\tilde{\sigma}_{i,j}^2 = \frac{\sigma_{i,j}^2}{P_{\text{RX}}}$. For notational convenience we define $I_{i,j}^{\text{intra}}(\mathbf{X}) = \sum_{k\in\mathcal{N}} u_k \frac{g_{i,j}}{g_{i,k}} x_{i,k} - u_j$ and $I_{i,j}^{\text{inter}}(\mathbf{X}) = \sum_{l\neq i}\sum_{k\in\mathcal{N}} u_k \frac{g_{l,j}}{g_{l,k}} x_{l,k}$ for the normalized (w.r.t P_{RX}) intra- and inter-cell interference power at the test point location, respectively. The parameter α accounts for the orthogonality loss of codes within one base station. With the downlink SINR defined, we can formulate the set of constraints for the UMTS part of the system similar to [ACM03], where minimum SINR limits and power constraints are used to guarantee the target receive power P_{RX} at each test point. Each base station is constraint to the maximum transmit power $P_{\max}$ per link and thus, we obtain the first constraint for the UMTS system as

$$P_{\text{RX}} \leq g_{i,j} P_{\max}. \tag{3.26}$$

Furthermore, the SINR is lower bounded $\beta_{i,j} \geq \beta_{\min}$ to ensure the required basic data rate per link and if we denote the total available transmit power per base station by P_i^{tot}, we obtain for a power based power control mechanism

$$\sum_{j\in\mathcal{N}} u_j \frac{P_{\text{RX}}}{g_{i,j}} x_{i,j} \leq P_i^{\text{tot}}. \tag{3.27}$$

Now we have formalized all constraints to state the multi-RAT optimization problem that minimizes the energy consumption in an UMTS-LTE network while meeting the QoS constraints of all users. Using the above constraints in Problem 3.6 and combining it with

the single-RAT LTE Problem 3.21 yields

$$\begin{aligned}
\text{minimize} \quad & \sum_{l\in\mathcal{L}}(c_l+e_l)\left|\boldsymbol{t}_l^T\tilde{\boldsymbol{x}}\right|_0+\tilde{f}_l(\tilde{\boldsymbol{x}}) && && \text{(3.28a)}\\
\text{subject to} \quad & \sum_{j\in\mathcal{N}} u_j\frac{P_{\mathrm{RX}}}{g_{i,j}}x_{i,j}\le P_i^{\mathrm{tot}} && i\in\mathcal{L}_{\mathrm{UMTS}} && \text{(3.28b)}\\
& x_{i,j}P_{\mathrm{RX}}\le g_{i,j}P_{\max} && i\in\mathcal{L}_{\mathrm{UMTS}}, j\in\mathcal{N} && \text{(3.28c)}\\
& \frac{x_{i,j}}{\alpha I_{i,j}^{\mathrm{intra}}(\mathbf{X})+I_{i,j}^{\mathrm{inter}}(\mathbf{X})+\tilde{\sigma}_{i,j}^2}\ge\beta_{\min}x_{i,j} && i\in\mathcal{L}_{\mathrm{UMTS}}, j\in\mathcal{N} && \text{(3.28d)}\\
& \sum_{j\in\mathcal{N}}\frac{r_j}{B_i\ \tilde{\omega}_{i,j}}\ x_{i,j}\le 1 && i\in\mathcal{L}_{\mathrm{LTE}} && \text{(3.28e)}\\
& \sum_{i\in\mathcal{L}} x_{i,j}=1 && j\in\mathcal{N} && \text{(3.28f)}\\
& \mathbf{X}\in\{0,1\}^{L\times N}, && && \text{(3.28g)}
\end{aligned}$$

with the optimization variable being $\mathbf{X}$. Constraints (3.28b)-(3.28d) capture the constraints coming from the UMTS system with (3.28b) accounting for the total power limitation at UMTS base stations, whereas inequality (3.28c) ensures the minimum received power and (3.28d) the minimum SINR. The total bandwidth constraint for LTE base stations is captured by (3.28e). In accordance with Assumption 8, we include constraints (3.28f) and (3.28g).

In order to apply the reformulation and relaxation techniques from Section 3.3.3 to find solutions[9] to Problem 3.28, we have to express the non-convex constraint (3.28d) equivalently as

$$\beta_{\min}\Big[\alpha\big(\sum_{k\in\mathcal{N}}u_k\frac{g_{i,j}}{g_{i,k}}x_{i,k}-u_j\big)+\sum_{l\neq i}\sum_{k\in\mathcal{N}}u_k\frac{g_{l,j}}{g_{l,k}}x_{l,k}+\tilde{\sigma}_{i,j}^2\Big]-(1-x_{i,j})\,K_{i,j}\le 1 \qquad (3.29)$$

[9] We assume that the problem has a solution, i.e. is feasible.

for some sufficiently large $K_{i,j}$ (c.f. [BHM77]), which yields

$$\text{minimize} \quad \sum_{l\in\mathcal{L}}(c_l + e_l)\left|\boldsymbol{t}_l^T\tilde{\boldsymbol{x}}\right|_0 + \tilde{f}_l(\tilde{\boldsymbol{x}}) \tag{3.30a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} u_j \frac{P_{\mathrm{RX}}}{g_{i,j}} x_{i,j} \le P_i^{\mathrm{tot}} \qquad i \in \mathcal{L}_{\mathrm{UMTS}} \tag{3.30b}$$

$$x_{i,j} P_{\mathrm{RX}} \le g_{i,j} P_{\max} \qquad i \in \mathcal{L}_{\mathrm{UMTS}}, j \in \mathcal{N} \tag{3.30c}$$

$$\beta_{\min}\left[\alpha I_{i,j}^{\mathrm{intra}}(\mathbf{X}) + I_{i,j}^{\mathrm{inter}}(\mathbf{X}) + \tilde{\sigma}_{i,j}^2\right] - (1 - x_{i,j})\, K_{i,j} \le 1 \qquad i \in \mathcal{L}_{\mathrm{UMTS}}, j \in \mathcal{N} \tag{3.30d}$$

$$\sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \tilde{\omega}_{i,j}}\ x_{i,j} \le 1 \qquad i \in \mathcal{L}_{\mathrm{LTE}} \tag{3.30e}$$

$$\sum_{i\in\mathcal{L}} x_{i,j} = 1 \qquad j \in \mathcal{N} \tag{3.30f}$$

$$\mathbf{X} \in \{0,1\}^{L\times N}. \tag{3.30g}$$

Problem 3.30 has the structure to apply the relaxations from Section 3.3.3 to it and we arrive at the following problem which we solve in order to derive network configurations that save energy in multi-RAT UMTS/LTE networks.

$$\text{minimize} \quad \sum_{l\in\mathcal{L}}(c_l + e_l)\frac{\log(1 + \epsilon^{-1}\ \boldsymbol{t}_l^T\tilde{\boldsymbol{x}})}{\log(1+\epsilon^{-1})} + \tilde{f}_i(\tilde{\boldsymbol{x}}) \tag{3.31a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} u_j \frac{P_{\mathrm{RX}}}{g_{i,j}} x_{i,j} \le P_i^{\mathrm{tot}} \qquad i \in \mathcal{L}_{\mathrm{UMTS}} \tag{3.31b}$$

$$x_{i,j} P_{\mathrm{RX}} \le g_{i,j} P_{\max} \qquad i \in \mathcal{L}_{\mathrm{UMTS}}, j \in \mathcal{N} \tag{3.31c}$$

$$\beta_{\min}\left[\alpha I_{i,j}^{\mathrm{intra}}(\mathbf{X}) + I_{i,j}^{\mathrm{inter}}(\mathbf{X}) + \tilde{\sigma}_{i,j}^2\right] - (1 - x_{i,j})\, K_{i,j} \le 1 \qquad i \in \mathcal{L}_{\mathrm{UMTS}}, j \in \mathcal{N} \tag{3.31d}$$

$$\sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \tilde{\omega}_{i,j}}\ x_{i,j} \le 1 \qquad i \in \mathcal{L}_{\mathrm{LTE}} \tag{3.31e}$$

$$\sum_{i\in\mathcal{L}} x_{i,j} = 1 \qquad j \in \mathcal{N} \tag{3.31f}$$

$$\mathbf{X} \in [0,1]^{L\times N}. \tag{3.31g}$$

The solution is obtained by applying the MM algorithm from Section 3.3.3 to the objective function (omitting constant)

$$h(\tilde{\boldsymbol{x}}) := \sum_{l\in\mathcal{L}} (c_l + e_l)\log(\epsilon + \boldsymbol{t}_l^T\tilde{\boldsymbol{x}}) + \tilde{f}_l(\tilde{\boldsymbol{x}}), \tag{3.32}$$

with the majorizing function

$$g(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}^{(n)}) := \sum_{l\in\mathcal{L}}(c_l + e_l)\frac{\boldsymbol{t}_l^T\tilde{\boldsymbol{x}}}{\epsilon + \boldsymbol{t}_l^T\tilde{\boldsymbol{x}}^{(n)}} + \nabla\tilde{f}_l(\tilde{\boldsymbol{x}}^{(n)})^T\tilde{\boldsymbol{x}}. \tag{3.33}$$

The resulting optimization problem solved in each step of the adapted iterative Algorithm 2 is

$$\begin{aligned}\tilde{\boldsymbol{x}}^{(n+1)} &\in \arg \min_{\tilde{\boldsymbol{x}} \in \mathcal{X}_{\text{multi-RAT}}} g(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}^{(n)}) \\ &= \arg \min_{\tilde{\boldsymbol{x}} \in \mathcal{X}_{\text{multi-RAT}}} \sum_{l \in \mathcal{L}} (c_l + e_l) \frac{\boldsymbol{t}_l^T \tilde{\boldsymbol{x}}}{\epsilon + \boldsymbol{t}_l^T \tilde{\boldsymbol{x}}^{(n)}} + \nabla \tilde{f}_l(\tilde{\boldsymbol{x}}^{(n)})^T \tilde{\boldsymbol{x}},\end{aligned} \tag{3.34}$$

with $\mathcal{X}_{\text{multi-RAT}}$ being the set of constraints (3.31b)-(3.31g) and $\tilde{\boldsymbol{x}}^{(0)} \in \mathcal{X}_{\text{multi-RAT}}$ being a feasible starting point. The algorithm is terminated when the termination criterion in (3.16) is true.

3.6 Decentralized network topology control

The algorithm proposed in Section 3.3.3 aiming at energy savings in cellular communication networks strongly relies on Assumption 11 requiring a central network controller that has global knowledge about network parameters and propagation characteristics. However, in real networks such information might not be available in the desired granularity or the collection of which is infeasible due to the network size. Above all, such centralized algorithms require a lot of signaling which increases the overhead burden on the network. Therefore, we propose a heuristic operating in a distributed fashion where only information locally available is used and that finds feasible solutions to Problem 3.6.

To avoid notational clutter we assume that all base stations have exactly one cell $|S_l| = 1, l \in \mathcal{L}$, the load dependent part is negligible $\tilde{f}_i(\tilde{\boldsymbol{x}}) = 0$ and $e_1 = e_2 = \cdots = e_L = 0$ in (3.5). Furthermore, we define $y_i := \boldsymbol{t}_i^T \tilde{\boldsymbol{x}}$ and obtain the following reformulation of Problem 3.6

$$\text{minimize} \quad \sum_{i \in \mathcal{L}} c_i \left| y_i \right|_0 \tag{3.35a}$$

$$\text{subject to} \quad \mathbf{X} \in \mathcal{X}_D \tag{3.35b}$$

$$\sum_{i \in \mathcal{M}} x_{i,j} = 1 \qquad j \in \mathcal{N} \tag{3.35c}$$

$$y_i = \boldsymbol{t}_i^T \tilde{\boldsymbol{x}} \qquad i \in \mathcal{L} \tag{3.35d}$$

$$\mathbf{X} \in \{0,1\}^{L \times N}\,, \tag{3.35e}$$

where the optimization variables are $\mathbf{X}$ and y_i. $\mathcal{X}_D$ is the set of points satisfying the technology specific constraints accounting for the limited amount of resources at cells. Note, that $\mathcal{X}_D$ is convex or can be relaxed to convex constraints due to Assumption 10.

The main idea of the proposed scheme is based on the observation that each cell can locally decide on which TPs it allows connection (admission control) and that a TP can select from multiple cells it connects to based on some metric broadcasted by the cells in its

proximity. The proposed algorithm is a two-step iterative scheme with an initialization phase and we detail on each step in the following.

In the initialization phase of the algorithm each TP is been connected to the cell that provides the best communication link in terms of its channel gain. More precise, each TP j is connected to cell i with

$$i = \arg\min_{i \in \mathcal{L}} g_{i,j} \tag{3.36}$$

yielding the initial assignment matrix $\mathbf{X}^{(0)}$. Based on this assignment each cell i solves locally the following optimization problem:

$$\begin{aligned}
&\text{maximize } y_i^{(0)} && && \text{(3.37a)}\\
&\text{subject to } \mathbf{X} \in \mathcal{X}_D && && \text{(3.37b)}\\
&\qquad x_{i,j} \geq x_{i,j}^{(0)} && j \in \mathcal{N} && \text{(3.37c)}\\
&\qquad y_i^{(0)} = \boldsymbol{t}_i^T \tilde{\boldsymbol{x}} && && \text{(3.37d)}\\
&\qquad \mathbf{X} \in \{0,1\}^{L \times N}, && && \text{(3.37e)}
\end{aligned}$$

where the optimization variables are $\mathbf{X}$ and $y_i^{(0)}$ which counts the TPs cell i allows admission[10]. The initial set of those TPs cell i allows admission is denoted by $\tilde{\mathcal{N}}_i$. The initialization phase is concluded by all cells broadcasting $y_i^{(0)}$ and its energy consumption c_i to TPs $j \in \tilde{\mathcal{N}}_i$ where TPs compose the set of admissible cells $\mathcal{L}_j$ from all $y_i^{(0)}$ they received. Now each TP connects to a cell according to

$$i = \arg\max_{i \in \mathcal{L}_j} y_i^{(0)}, \tag{3.38}$$

giving the intermediate assignment matrix $\mathbf{X}^{(1)}$.

After the initialization phase the following two steps are iteratively performed at cells and TPs starting with $n = 1$ (Step 2 is performed locally at one TP and y_i is updated in the whole network):

1. (*Step 1, cell metric calculation*) Each cell i computes the number of TPs connected to it by $y_i^{(n)} = \sum_{i \in \tilde{\mathcal{N}}_i} x_{i,j}^{(n)}$ and broadcast $y_i^{(n)}$ and c_i to TPs $j \in \tilde{\mathcal{N}}_i$.

2. (*Step 2, TP assignment*) TP j selects cell i to connect to with

$$i = \arg\max_{i \in \mathcal{L}_j} \frac{c_{k(n-1)}}{c_i} y_i^{(n)},$$

   where c_i is the energy consumption of cell i and $c_{k(n-1)}$ is the energy consumption of the cell it was connected to in the previous iteration.

The algorithm terminates when no TP switches to another cell any more or a maximum

[10] We assume that Problem 3.37 has a solution.

Algorithm 3 Distributed network reconfiguration for improved energy efficient operation

Input: set of TPs, set of cells, constraints

Output: optimized network configuration according to $\mathbf{X}^{\star}$.

1: initialize $\mathbf{X}^{(0)}$ by assigning each TP to its closest cell.
2: Each cell i solves locally Problem 3.37 obtaining objective value $y_i^{(0)}$ and the set of TPs $\mathcal{N}_i$ allowed for admission.
3: All cells broadcast $y_i^{(0)}$ to TPs $j \in \mathcal{N}_i$.
4: TPs create the set of admissible cells $\mathcal{L}_j$ from all $y_i^{(0)}$ they received and connect to cell $i \in \mathcal{L}_j$ with largest $y_i^{(0)}$.
5: **repeat**
6: Cells count the number of TPs connected to them and broadcast the number $y_i^{(n)}$ to TPs $j \in \mathcal{N}_i$.
7: TPs j connects to cell $i = \arg\max \frac{c_{k(n-1)}}{c_i} y_i^{(n)}$.
8: Deactivate cells with no user connected to it.
9: **until** TPs are not switching cells any more or $n > n_{\max}$

number of iterations $n > n_{\max}$ is reached. The complete algorithm is summarized in Algorithm 3.

3.7 Performance evaluation

This section presents various simulation results for the algorithms presented in this chapter so far. Thereby, we exemplify the theoretical results in exemplary simulation scenarios. The simulation scenarios are chosen in such a way to illustrate specific characteristics of our developed algorithms and their applicability to a broad range of networks.

3.7.1 Numerical evaluation for LTE networks

In the following, we present a numerical evaluation of the performance of the algorithm proposed in Section 3.4 for different setups of a single-RAT LTE network. We start by outlining the basic simulation scenario followed by a comparison with two reference schemes with respect to the energy savings and computational time. Next, we present the ability of the proposed algorithm to incorporate a variety of different base station energy consumption models.

3.7.1.1 Basic simulation scenario

The simulated network is located in a square-shaped area of size 2km×2km, where L base stations are placed at locations chosen uniformly at random. Unless stated otherwise, each base station has three cells directed at 0°, 120° and 240° respectively. Traffic generated by users is represented by N TPs on an irregular grid. Hence, each TP represents the traffic requirements of an area of different size. To obtain spatially varying traffic requirements, we use the following traffic model in each run of the simulations. We define three circular hot spot areas with centers chosen uniformly at random within the area. There are two types of TPs: "hot spot TPs (HTP)" and "standard TPs (STP)". Each TP in the simulation has probability 0.3 of being a HTP and probability 0.7 of being a STP. While the position of STP is chosen uniformly at random within the whole area, a HTP can be assigned uniformly at random to one of three hot spot areas. Its final position is determined in polar coordinates by sampling the distance from the hotspot center from a normal distribution and the angle from a uniform distribution. We use a wrap around model to avoid boundary effects and determine the location of TPs to be placed outside the square-shaped area. The data rate requirements of TPs are derived from a normal distribution with $\mu_{\mathrm{d}} = 128$ kbps and variance $\sigma_{\mathrm{d}}^2 = 32$ kbps^2 with a lower bound of 1 kbps. The signal attenuation for links between cells and TPs follows the ITU propagation model for urban macro cell environments with a horizontal antenna pattern for 3-sector cell sites with fixed antenna patterns [3GP10a].

Unless otherwise stated, we use the following simulation parameters: $\epsilon^\star = 10^{-3}$, $\epsilon = 10^{-3}$, $B_i = 20$MHz, $P_i = 40dB$, $\eta_{\mathrm{SINR}} = 1$, $\eta_{\mathrm{BW}} = 0.83$, $c_i = 500$W and $e_i = 280$W. The values of the last six parameters have been chosen to mimic the behavior of commercial LTE systems. Furthermore, we use $f_i(\rho_i) = 564\ \rho_i$ to model the load-dependent energy consumption,

which is a value similar to the dynamic energy consumption of current macro cells with 6 transmit antennas [ABG+10].

The proposed algorithms are compared with a solution of the original Problem 3.6 and, where possible, with the centralized cell zooming approach from [NWGY10]. The solution to Problem 3.6 is obtained by using Matlab 2014a in combination with IBM's CPLEX v12.5 on a Intel Core i7 PC with four cores and 8GB RAM. As shown later in this section, the computational time to solve Problem 3.6 grows fast with the problem size. Therefore, to solve Problem 3.6 in a reasonable time for comparison purposes, we confine our attention to small networks with $M = 102$ cells ($L = 34$ base stations) and $N = 100$ TPs, unless otherwise stated. We obtained the 95% confidence intervals depicted in the figures by applying the bias corrected and accelerated bootstrap method [Efr87] to the outcome of 100 independent runs of the simulations. Results related to the overall network energy consumption will be normalized to the energy consumption of the network when all cells are active and fully loaded.

Definition 14 (Normalized network energy consumption)**.** *Given a TP assignment* $\mathbf{X}$ *inducing cell load* $\boldsymbol{\rho}$ *and given the resulting network energy consumption* $E(\boldsymbol{\rho})$*, the normalized network energy consumption is defined to be*

$$E_{norm}(\boldsymbol{\rho}) := \frac{E(\boldsymbol{\rho})}{E(\mathbf{1})} = \frac{E(\boldsymbol{\rho})}{\sum_{l\in\mathcal{L}} c_l + \sum_{i\in\mathcal{M}}\big(e_i + f_i(\mathbf{1})\big)}, \tag{3.39}$$

where the term in the denominator is the energy consumption for a fully loaded system ($\boldsymbol{\rho} = \mathbf{1}$)*.*

We refer to the sparsity supporting majorization-minimization algorithm developed in Section 3.4 as *"sMM"* and to any algorithm that solves Problem3.6 directly as *"MIP"* algorithm (MIP: mixed-integer programming). We refer to solutions obtained by the centralized cell zooming algorithm in [NWGY10] as *"cCZ"*.

3.7.1.2 Computational performance comparison between sMM, cCZ and MIP

The *cCZ* has limited capability to incorporate different energy consumption models and base stations with several sectors, so we confine ourselves to a simple base station model. We assume a homogeneous network model under which all base stations have only one omni-directional cell, and all base stations have the same energy consumption model. More precise, we use $|\mathcal{L}| = M = 100$, $|\mathcal{S}_l| = 1$ and (3.3) with $c_l = 500$, $e_i = 280$, $f_i(\rho_i) = 0$ ($l \in \mathcal{L}$, $i \in \mathcal{M}$).

To show trends, we start with the standard setup described above and we gradually increase the number of TPs in the system. Figure 3.3 shows the *normalized network energy consumption*.

As expected, the normalized network energy consumption for all three algorithms increase as the number of TPs increases. This is intuitive because additional TPs add extra rate

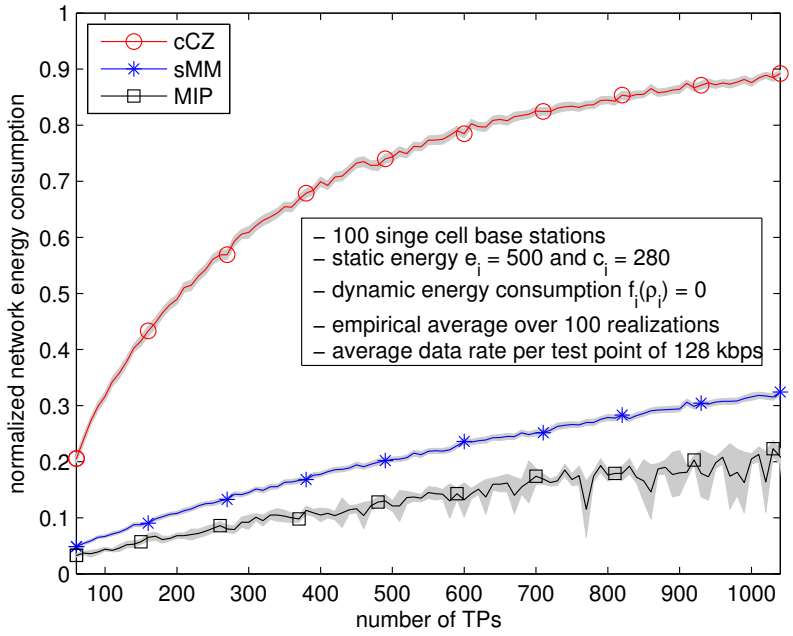


Figure 3.3: Comparison of normalized network energy consumption obtained with the *sMM* algorithm, the *cCZ* algorithm and the solution of the *MIP* problem for increasing number of TPs. Normalization with respect to the network energy consumption for a fully loaded system ($\boldsymbol{\rho} = \mathbf{1}$) when all cells are active. Results are averaged over 100 different realizations of the network and the 95% confidence intervals are provided in gray.

requirements that increase the total system load, which in turn reduces the redundancy in the network to be exploited for energy savings. The proposed *sMM* algorithm as well as the *MIP* algorithm provide network configurations that exhibit much smaller normalized network energy consumption when compared with the network configurations obtained with the *cCZ* algorithm. The smallest energy consumptions are achieved with the *MIP* algorithm, which outperforms the proposed *sMM* algorithm. For the scenario with 200 TPs, the *sMM* algorithm results in normalized network energy consumption of 12% on average. For the same number of TPs, the average normalized energy consumption under the *cCZ* and *MIP* algorithm are 49% and 7%, respectively. Similarly, for 1000 TPs, the resulting average normalized network energy consumption of 31% for the *sMM* algorithm is still larger than the 21% normalized energy consumption corresponding to the *MIP* solutions. However, it is still much smaller than *cCZ* with 88% normalized energy consumption. These results emphasize that the *sMM* algorithm is a suboptimal heuristic, which is able to find network configurations consuming low energy. Even though the resulting network energy consumption is not globally optimal, it shows much larger energy savings than the comparison scheme *cCZ*.

The main advantage of the proposed *sMM* algorithm is its fairly low computational complexity, which is directly affecting the time required to obtain an optimization result. Figure 3.4 depicts the normalized time needed to obtain the results of Figure 3.3. This time

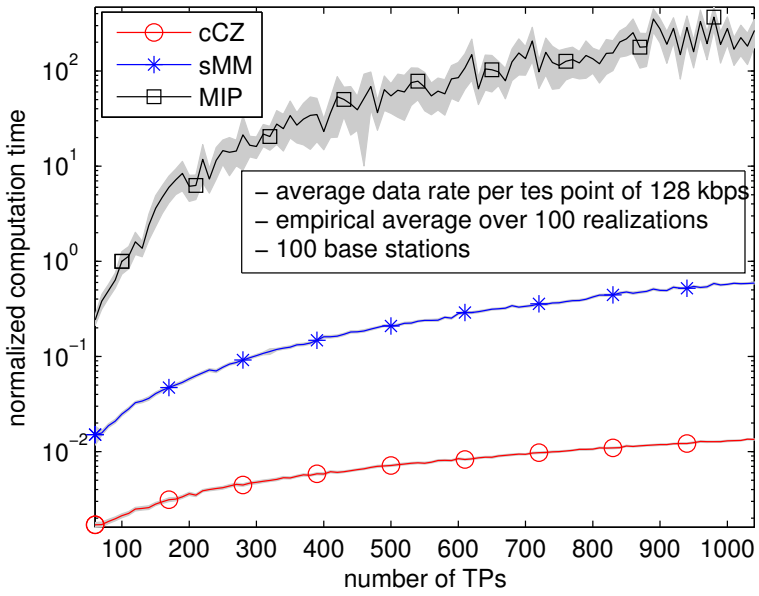


Figure 3.4: Comparison of normalized computation time to obtain results with the *sMM* algorithm, the *cCz* algorithm and the direct solution of the *MIP* problem. Normalization with respect to the empirical average of the *MIP*'s computation time for 100 cells and 100 TPs over 100 realizations. The 95% confidence intervals are provided in gray.

is normalized with respect to the computation time of the *MIP* algorithm with 100 cells and 100 TPs. The *sMM* algorithm always provides results in a substantially shorter time than the *MIP* algorithm. Even for a relatively small scenario of 100 cells and 300 TPs, the computation time is already about 200 times larger for the *MIP* algorithm compared to the proposed *sMM* algorithm. For larger setups with 1000 TPs the normalized time to solve the *MIP* was ≈ 237 compared to ≈ 0.49 for the *sMM* algorithm, which is an approximate 488 fold reduction in the computation time. We emphasize that the simulated scenarios are small and the computation of the MIP solution becomes infeasible in practical scenarios. Already for a network with 200 cells and 10,000 TPs, the sMM algorithm provided a solution in about 13s, whereas the MIP algorithm could not find a solution within one hour. Compared to the *cCZ* algorithm the proposed *sMM* algorithm takes longer time due to the lower complexity heuristic used in the *cCZ* algorithm. For a scenario of 300 TPs the average computation time is about 22 times larger for the *sMM* algorithm and with 1000 TPs it is about 43 times larger. However, with typical values of less than 1s, the computation time is still reasonably small to allow for an online implementation. Considering the advantages in energy savings, as seen from Figure 3.3, the proposed *sMM* algorithm presents a good trade off between computation time and energy savings.

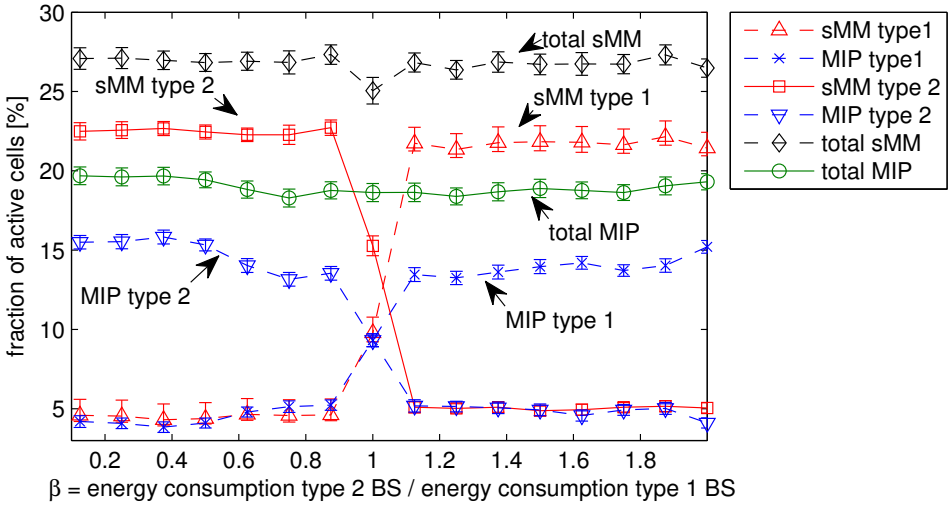


Figure 3.5: Cell selection for heterogeneous energy consumption models - random deployment. Fraction of active *type 1* and *type 2* cells in the final solution obtained with sMM and MIP. Deployment uniformly at random for *type 1* and *type 2* cells. Results are averaged over 100 different realizations of the network and the 95% confidence intervals are provided.

3.7.1.3 Cells with different sources of energy consumption

In contrast to other approaches to the problem of energy-efficient network topology control, our optimization framework can easily deal with heterogeneous networks in which cells have different static and load-dependent energy consumptions accounted for in (3.3). In other words, the proposed *sMM* algorithm can cope with different energy consumption models of cells. It can select network configurations that exhibit as low overall energy consumption as possible. To illustrate the impact of different energy consumption models on the optimization result, we start by varying the static energy consumption of all cells, while keeping the load-dependent energy consumption fixed. Later in this section, we show the impact of the load-dependent energy consumption by changing the weight of the load dependent part relative to the static part.

To study the impact of the static energy consumption of cells e_i, in the following simulations we use single cell omnidirectional base stations, and we set the load-dependent part for all cells and the common static part at base stations to zero $f(\rho_i) = 0$ and $c_l = 0$. The static energy consumption of half of the cells is varied, while the static energy consumption of the other half remains unchanged. We refer to the cells with standard fixed energy consumption as *type 1*, while *type 2* is used to refer to cells with a varying energy consumption. The energy consumption of *type 2* cells is specified relative to that of *type 1* cells. More precisely, an energy consumption relation of $\beta = 0.5$ means that if $c_i = 780$W for *type 1* cells, then $c_i = 390$W for *type 2* cells. The results for a scenario consisting of 100 cells and 100 TPs are shown in Figure 3.5.

The simulation confirms the ability of our optimization framework to incorporate different

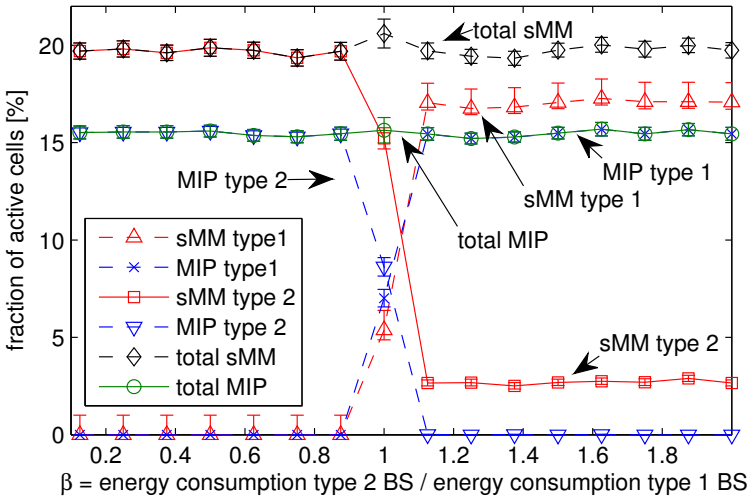


Figure 3.6: Cell selection for a heterogeneous energy consumption model - co-located deployment. Fraction of *type 1* and *type 2* cells in the final solution obtained with sMM and MIP. Deployment of *type 1* cells uniformly at random and *type 2* cells are co-located with *type 1*. Results are averaged over 100 different realizations of the network and the 95% confidence intervals are provided.

static energy consumptions. When all cells consume the same amount of energy ($\beta = 1$), the algorithm makes no difference between *type 1* and *type 2* cells. The energy consumption of *type 1* and *type 2* cells is roughly the same indicating that an equal number of *type 1* and *type 2* cells are active in the obtained solution. In contrast, if *type 2* cells consume less energy than *type 1* ($\beta < 1$), then the algorithm prefers to deactivate *type 1* cells, while attempting to keep *type 2* cells active. Obviously, if $\beta > 1$, the situation is reversed in the sense that, if possible, *type 2* cells are preferably selected for deactivation.

The differentiation becomes even more evident for cell deployments, where *type 1* and *type 2* cells are co-located. In such a case, two cells of different types are located at the same site and are "exchangeable" with respect to the service provided to the TPs (recall that we use omnidirectional cells in these simulations). In other words, if a TP is assigned to a location with two co-located cells, then it does not matter which cell is used to provide the service to the TP. This implies that the decision whether to deactivate a cell or not should depend only on the energy consumption of this cell in relation to its co-located cell[11].

The simulations with such a deployment are shown in Figure 3.6, where we see that, for $\beta < 1$, there is no active cell of *type 1*, while, for $\beta > 1$, *type 2* cells consume more energy and the simulations confirm that the algorithms clearly prefer to activate *type 1* cell.

To obtain insight into the impact of the load-dependent energy consumption, we fix the static energy consumption of a single-cell omnidirectional base station to be $e_i = 780$W

[11] Even though such setups are unlikely in practice, we use it for reasons of illustration.

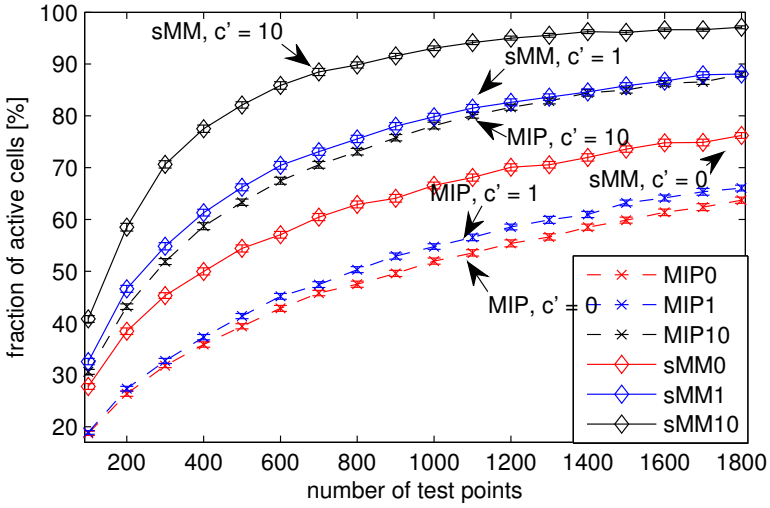


Figure 3.7: Fraction of active cells for different dynamic energy consumption c' with increasing number of TPs. Results are averaged over 100 different realizations of the network and the 95% confidence intervals are provided.

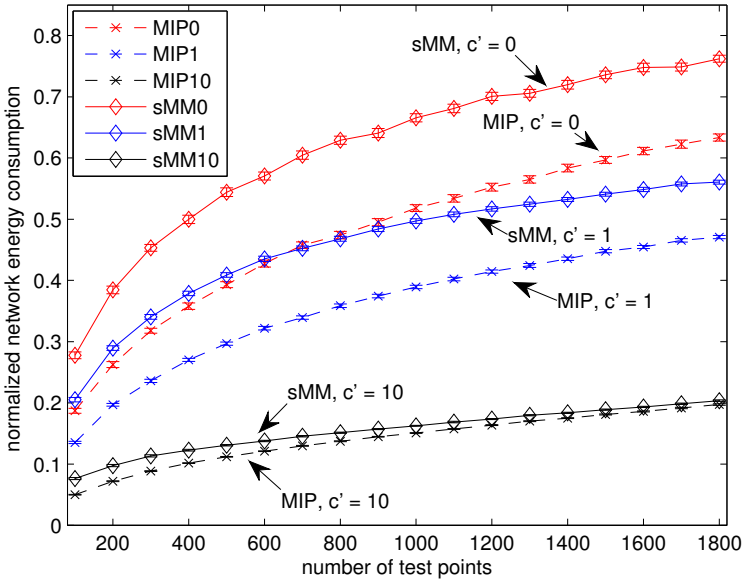


Figure 3.8: Normalized network energy consumption for different dynamic energy consumption weights c' with increasing number of TPs. Normalization with respect to the energy consumption when all cells are active. Results are averaged over 100 different realizations of the network and the 95% confidence intervals are provided.

and $c_l = 0$W, and we vary the load-dependent energy consumption $f_i(\boldsymbol{\rho}) = 564\, c'\, \rho_i$ by letting c' take values on $c' \in \{0, 1, 10\}$. For an increasing number of TPs, Figure 3.7 shows the fraction of active cells, while the normalized network energy consumption is shown in Figure 3.8.

First we observe that the network energy consumption always increases with an increasing

number of TPs, which is in fact no surprise. Moreover, the fraction of active cell increases with c' for both the *sMM* algorithm and the *MIP* algorithm. An examination of the objective function in (3.21a) shows that this is what we expect. If the ratio of the load-dependent energy consumption becomes larger relative to the static one, then the algorithm tends to increase the fraction of active cells for an improved load balancing in order to keep the load of each active cell at a relatively low level. In other words, instead of deactivating as many cells as possible to minimize the static energy consumption, the algorithm deactivates the cells to find the best possible balance between the static and load-dependent energy consumption. This can be observed in Figure 3.7 We can see that the higher the load-dependent energy consumption (which is reflected by $c' \geq 0$), the more cells are activated under both the *sMM* algorithm and *MIP* algorithm. In particular, if $c' = 10$, the fraction of active cells is significantly increased compared to the situation in which the load-dependent energy consumption is negligible ($c' = 0$).

3.7.2 Numerical evaluation for UMTS/LTE networks

We now present simulation results that show the performance of the algorithm for multi-RAT UMTS/LTE networks proposed in Section 3.5. Since the underlaying algorithm that finds good solutions to the stated optimization problem is the same as for the single-RAT problem, we confine ourselves with a reduced set of simulations that highlight the algorithm's applicability and its good performance for multi-RAT problems. In contrast to the considered scenarios in the previous section we have two networks with different technologies (UMTS and LTE) serving the same coverage area. In more detail, the simulation environment consists of 50 UMTS and 50 LTE base stations deployed uniformly at random in a 2 km by 2 km area. For simplicity, we consider that each UMTS and each LTE base station has only one omnidirectional cell and thus $|\mathcal{L}_{\text{UMTS}}| = M_{\text{UMTS}} = 50$, $|\mathcal{L}_{\text{LTE}}| = M_{\text{LTE}} = 50$, $|\mathcal{S}_l| = 1$. The traffic in the network is generated by users and their data rate requirements are represented by TPs. These requirements are modeled in a similar way as done in Section 3.7.1 for the single-RAT simulations. We define three circular hotspot areas with radius 500m and place TPs in one of these areas with a probability of 5% each. Its final position is determined in polar coordinates by sampling the distance from the hotspot center of a normal distribution and the angle from a uniform distribution. All remaining TPs are placed uniformly in the whole area. All TPs support both technologies which enables them to connect to a UMTS base station or a LTE base station. The A wrap around model is used to avoid boundary effects and to place TPs that have a location outside the 2 km by 2 km area. The signal power gain of each link from all base stations to all test points follows the ITU propagation model for urban macro cell environments with fixed antenna patterns [3GP10a].

Unless otherwise stated we use the following simulation parameters: $r_{\text{base}} = 64\text{kb/s}$, $B_i = 20\ \text{MHz}\ \forall i \in \mathcal{M}_{\text{LTE}}$, u_i uniformly distributed in $\{1, 2\}$, $P_{\min} = -100\text{dBm}$, $P_{\text{tot}} = 40\text{W}$, $\text{SIR}_{\min} = 12\text{dB}$ $P_{\max} = 30\text{dBm}$, $K_{i,j} = 10^6$, $\alpha = 0.5$ and $c_i = 400\text{W}$.

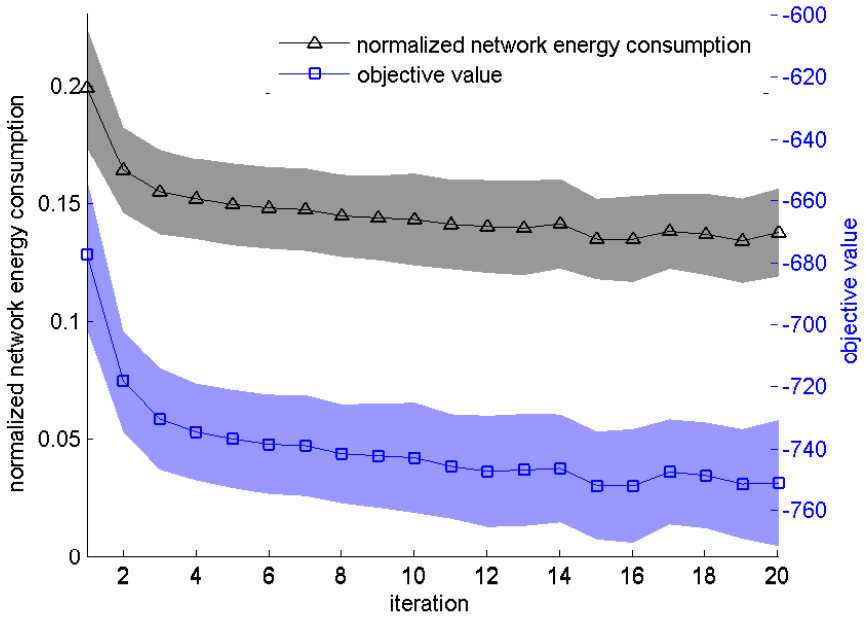


Figure 3.9: Objective value of (3.33) and number of active base stations as a function of the number of iterations before applying Algorithm 1. The results are averaged over 100 realizations, and the confidence intervals indicate the estimated error of the mean (95% confidence interval). The blue curve with square markers shows the objective in (3.33) with corresponding y-axis on the right. The black curve with triangle markers shows the corresponding normalized network energy consumption for the result of each iteration (legend on the left side).

The algorithm uses the stopping criterion (3.16) with $\epsilon^\star = 10^{-3}$ and in LP (3.34) $\epsilon = 10^{-3}$ is used.

To obtain a better insight in the behavior of the proposed algorithm we evaluate the results for an exemplary realization with $N = 300$ test points. Our algorithm identified 10 UMTS base stations and 7 LTE base stations to be necessary to serve the QoS requirements of the 300 test points. Thereby, 163 test points have been connected to UMTS base stations and 137 test points have been connected to LTE base stations. Figure 3.9 shows the evolution of the objective value and the corresponding normalized network energy consumption for each iteration while executing the proposed algorithm. The results are averaged over 100 realizations, and the confidence intervals indicate the estimated error of the mean (95% confidence interval) The algorithm shows the property to have a fast decaying objective value; for this scenario, it needs only a small number of iterations to terminate. We further observe that the objective value (3.33) decreases within the first five iterations and improves only marginally after. Additionally, we plot the normalized network energy consumption as defined in Definition 14 for the solution vector of each iteration (i.e., the results of Algorithm 1 after the corresponding number of iterations of Algorithm 2). We observe that it follows the same trend, and only a large reduction of the objective value leads to a significant decrease in the normalized network energy consumption. Simulations for different

energy consumption [W] UMTS / LTE	active BS UMTS / LTE
200 / 400	10.5 / 1.6
400 / 400	7.2 / 7.7
600 / 400	6.1 / 9.6
800 / 400	3.9 / 14.5

Table 3.1: Simulation results for different energy consumption of UMTS and LTE base stations. Results are averaged over 30 simulations for $M_{\text{UMTS}} = 50$, $M_{\text{LTE}} = 50$ and $N = 300$.

base station numbers and test points have shown similar patterns. A smaller $\epsilon^\star$ (which leads to more iterations) does not lead to an improved solution.

To evaluate the practical performance of the algorithm, which iteratively solves LP (3.34), we compare it to the original Problem 3.28, solved by using Matlab 2014a in combination with IBM's CPLEX on a Intel Core i7 PC with four cores and 4GB RAM. For small problem instances of 150 test points in a service area of 20 UMTS and 20 LTE base stations, the computation time was around 15 minutes to solve Problem 3.28. The obtained solution identified a set of six active base stations to serve the test points. In contrast, our algorithm found a solution with the same number of active base stations in only 0.8 seconds. For larger problems with 50 UMTS, 50 LTE base stations and 300 test points, our algorithm obtained a solution within a few seconds, whereas Problem 3.28 could not be solved in reasonable time (less than four hours). These results are in line with performance evaluations for single-RAT networks (c.f. Section 3.7.1). Note however, that they are not directly comparable due to implementation specifics and different hardware setup used for simulations.

Finally we demonstrate the ability of our algorithm to confine the traffic to base stations that consume only little energy when both UMTS and LTE base stations can serve the test points equally. We alternate the basic energy consumption c_i of UMTS and LTE base stations and record the averaged results in Table 3.1. Hereby, the results are averaged over 30 simulation runs for the setup with $M_{\text{UMTS}} = 50$, $M_{\text{LTE}} = 50$, and $N = 300$. It can be observed that the algorithm tends to keep active base stations with low energy consumption, and it tends to switch off base stations with high energy consumption, as expected and already observed in the single-RAT case (see Section 3.7.1).

3.7.3 Numerical evaluation of the decentralized approach

The performance of the decentralized scheme proposed in Section 3.6 is empirically evaluated by means of simulations. We compare the resulting network energy consumption to the ones obtained by the MIP, sMM and cell-zooming algorithm. For the performed simulations we reuse the simulation scenario and parameters from the single-RAT centralized simulations in Section 3.7.1. For convenience we summarize the key simulation parameters

Parameter	Value
Simulated network area	2km × 2km
Number of base stations l	$\lvert\mathcal{L}\rvert = 34$
Number of cells M	$\lvert\mathcal{M}\rvert = 102$, $\lvert\mathcal{S}_l\rvert = 3$
Energy consumption	$c_i = 500$W, $e_i = 280$W, $f_i(\rho_i) = 0$
Number of resource units per cell	$B_i = 20$MHz
Data rate requirement per TP r_j	$\mathcal{N}(128 \text{ kbps}, 32 \text{ kbps}^2)$
Stopping criterion	$\epsilon^\star = 10^{-3}$
Relaxation parameter in (3.10)	$\epsilon = 10^{-3}$

Table 3.2: Key simulation parameters used in empirical evaluation of decentralized algorithm for single-RAT LTE network energy optimization.

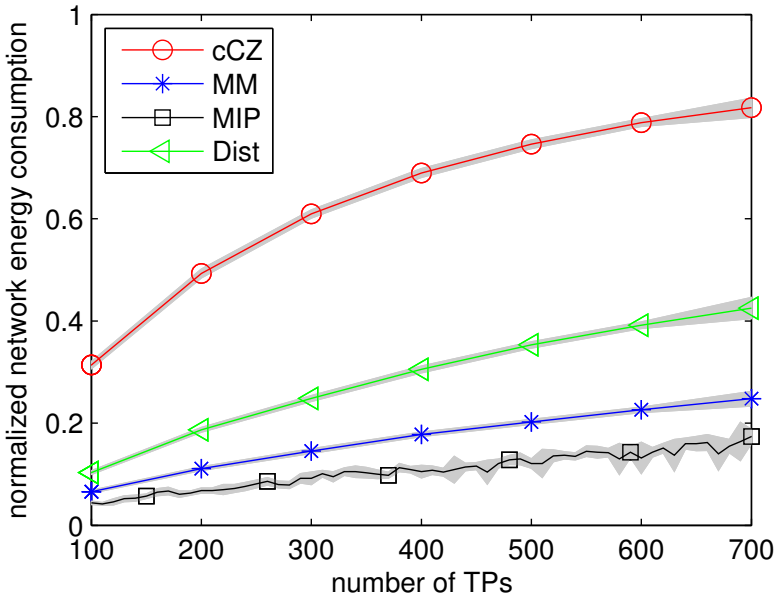


Figure 3.10: Comparison of sMM, MIP, cell-zooming and our decentralized approach. 95% confidence interval calculated for 100 random network realizations according to the setup explained in Section 3.7.1

in Table 3.2. The comparison of the simulation results for different number of test points in the system is provided in Figure 3.10 showing the normalized network energy consumption defined in Definition 14. The results obtained with our proposed distributed algorithm are following the same trend as the centralized reference schemes. For an increase in traffic demand, represented by a higher number of test points with a constant data rate requirement, the network energy consumption increases since there are less redundancies in the network that can be exploited for energy savings. Compared to the optimal solution and the centralized *sMM* solutions the decentralized algorithm gives network setups that consume more energy when providing the same service to the test points. However, our decentralized scheme has still a much better performance than the cell zooming scheme from [NWGY10].

3.8 Anticipatory scheduling for improved energy savings

In previous sections we have presented schemes that enable network reconfigurations to reduce energy consumptions. In particular, we devised algorithms that deactivate network elements in a coordinated manner to provide service to delay-sensitive applications at any given point in time. Even though those mechanisms exhibit good energy saving gains they are limited due to the fundamental trade-off between energy efficiency and delay constraints which is still an open problem [CZXL11].

The recent technological development towards higher throughput in wireless communication systems enables mobile users to use services with high data rate requirements like HD video streaming on the go. For this example a seamless video playback is supported by buffering data in the local memory of the user terminal. Thereby, a previously loaded fraction of the movie can be replayed even in areas with bad or no connection. This buffering capability builds the basis for our work in this section. We exploit the possibility to preload and save data on user devices which will serve as an enabling concept to save energy in the communication system by disengaging certain network components[12].

We now turn towards further improvements of these energy savings by exploiting users' tolerance for delay. The techniques of previous sections are based on the load variations in the network. It turns out that such spatial and temporal load fluctuations can be "shaped" to some extent. We drop the delay-sensitivity assumption of applications which is true for a class of applications e.g. machine-to-machine communications, file transfer, buffered media/video streaming etc. The service provision of moving wireless devices (e.g. cars, pedestrians, trains) can be diverted to areas where available capacity is high enough or to times that are favorable to save energy. We present a framework that helps prolong the periods when the network can be operated in a state of low energy consumption. This concept falls into the class of proactive resource allocation schemes and we extend it to include also the user assignment to cells. The resulting class of algorithms is termed Proactive Resource Allocation and User Assignment (PRAUA).

Consider the following example for better illustration. We have a dense cellular communication system consisting of cells with largely overlapping coverage area. We want to operate the cellular communication network in a low energy consumption state and thus need to identify cells that are not needed to provide the QoS demand of users. In our example we have a single user moving along a known route. Along the way the user wants to stream a video without any stalling and on its route the user passes through the service area of three cells. The example setup is depicted in Figure 3.11. Without any delay tolerant service scheduling, all three cells would have to be active in order to stream the video to the user. However, it is beneficial to exploit the user's mobility and buffering capabilities by allocating more resources to the user in the first cell such that the video could be preloaded

[12]Notice, that we are looking only at capacity aiding base stations to turn off. The coverage has to be secured at all times. We consider therefore a basic coverage by some legacy network and propose a possible shutting down of capacity providing base stations.

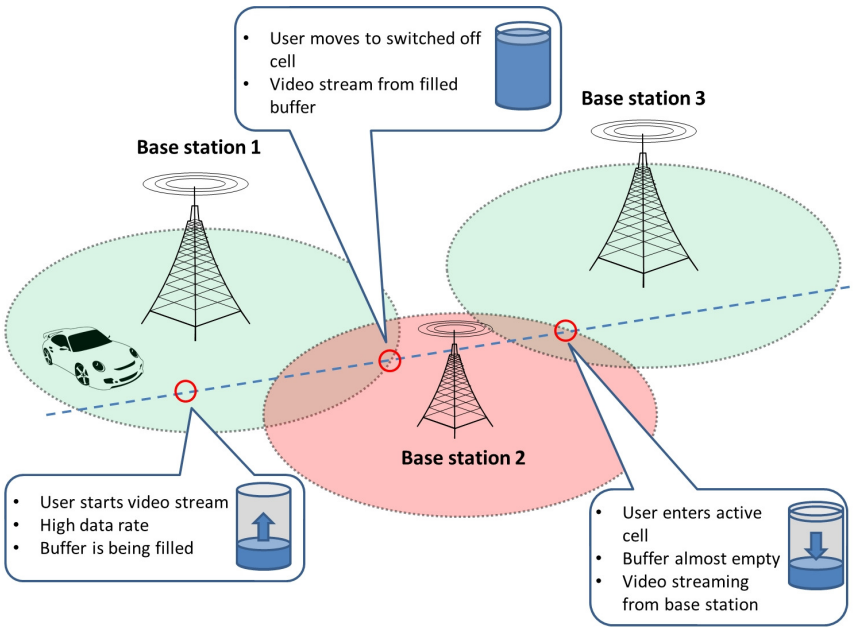


Figure 3.11: Exemplary illustration of a delay tolerant scheduling

in the user's buffer. When moving through the service area of the second cell the user streams the video content from its buffer instead of streaming it from the cell. Once entering the service area of the third cell, the buffer streaming can continue from cell three. In this simple example cell two can stay inactive resulting in energy savings.

The underlying idea is as follows. We apply the network reconfiguration techniques and incorporate the delay-tolerance to allow for intermediate degradation in service provisioning in order to improve the service before and after so that the energy saving state has more degrees of freedom. By delaying the service provision of some users we may avoid to activate network elements that are only needed when the traffic demand is of bursty nature. The result is a data transmission schedule sending data to users for buffering in high capacity cells whereas it avoids access to cells that are overloaded or switched of for reasons of energy savings. The result is a resource usage that lets users buffer data in high capacity cells whereas it avoids access to cells that are overloaded or switched off for reasons of energy savings. We use the predicted routes of users and the learned path loss coverage maps to find such a user-cell association and rate allocation policy under the exploitation of the users' buffers, i.e., when a user is predicted to pass an area without coverage it will be allocated more resources right before, so that the data can be loaded in the buffer (bridging the coverage hole).

3.8.1 Scenario and system model

We consider the downlink of a cellular heterogeneous communication system employing an OFDM-based resource allocation. We are interested in switching off capacity units of

the network, e.g. sectors, cells or the entire base station[13]; the corresponding decisions are performed at a central network controller.

Assumption 13. *Coverage is ensured by a legacy communication system.*

As a consequence of Assumption 13 LTE cells can be switched off without loosing coverage in that area. The set of all base stations is denoted by $\mathcal{L} = \{1, 2, ..., L\}$, and the cells belonging to base station l by $\mathcal{S}_l$. The full set of all M cells in the network is thus $\mathcal{M} := \cup_{l \in \mathcal{L}} \mathcal{S}_l$. Each cell i has total number of resource blocks B_i to allocate to its users. There are N users in the system to be served and we denote the set of all users as $\mathcal{N} = \{1, 2, ..., N\}$. The time is divided into K time slots of equal duration Δ_k and we denote the set of all time slots by $\mathcal{K} = \{1, \ldots, K\}$. For each time slot, the objective is to find a resource allocation and user-cell association. Each user is equipped with a first-in-first-out (FIFO) buffer and we denote the buffer level (in bits) of user j in slot k by $d_j^{(k)} \geq 0$ with $d_j^{(0)} = 0$ (empty buffer at start). We assume a sufficiently large buffer and refer to the technology specific spectral efficiency per resource block of the link from cell i to user j in slot k as $\tilde{\omega}_{i,j}^{(k)}$.

Assumption 14. *A reliable estimate of the users' routes and the supported spectral efficiency per resource block along those routes is available at the central controller.*

Assumption 14 can be implemented by using side information or estimated mobility trajectories based on the high regularity in human mobility [SQBB10] together with techniques to learn path loss maps, e.g., [KCV+16]. If the supported spectral efficiency per resource block is not available with sufficient accuracy, Assumption 12 can be used to lower bound the spectral efficiency of a link between cell i and user j at time point t by $\tilde{\omega}_{i,j}^{(k)} \geq \omega_{i,j}^{(k)} := \omega_{i,j}(\mathbf{1})$.

Remark 10. *The task of learning various radio maps like coverage, path-loss or interference maps is considered as a major challenge in radio network design [PSG13] and thus is a hot topic in research as well as in industry. The retrieval of such maps is out of the scope of this thesis but we refer to adaptive online learning mechanisms like Adaptive Projected Sub-gradient Method (APSM) which are investigated in e.g. [KCV+16] and references therein.*

The task of our optimization framework is to provide a schedule of resource allocations satisfying the rate requirements of users while trying to reduce the energy consumption. Similar to the optimization task in Section 3.4 we want to find a user cell assignment that allows for deactivation of network elements giving the largest possible energy savings. In this chapter we additionally enable users to buffer data locally to be used at a later point in time. Therefore, we do not impose constraints on a fixed rate per user per time slot received from a particular cell. If a user j is served by cell i in time slot k we denote the

[13] In the text that follows we will use cells as a placeholder for any type of network element that can be switched on/off independently.

effective transmit data rate as $r_{i,j}^{(k)} := b_{i,j}^{(k)} \omega_{i,j}^{(k)}$ where $b_{i,j}^{(k)}$ is the number of resource units allocated to user j by cell i in slot k. We collect the rates allocated by cell i to all users at time k in vector $\boldsymbol{r}_i^{(k)} = \left[r_{i,1}^{(k)}, r_{i,2}^{(k)}, ..., r_{i,N}^{(k)}\right]^T$. We further use $\mathbf{R}^{(k)} = \left[\boldsymbol{r}_1^{(k)}, \boldsymbol{r}_2^{(k)}, ..., \boldsymbol{r}_M^{(k)}\right]$ to refer to all rates allocated over all cells to all users in slot k. We extend Definition 10 to the time domain and obtain the following.

Definition 15 (Instantaneous Cell Load)**.** *Given the rate assignment matrix $\mathbf{R}^{(k)}$, the instantaneous load of cell i at time slot k, denoted by $\rho_i^{(k)}\left(\mathbf{R}^{(k)}\right) \in [0,1]$ or simply $\rho_i^{(k)}$ for notational simplicity, is defined to be the ratio of the number of resource blocks allocated to users served by cell $i \in \mathcal{M}$ in slot k to the total number of resource blocks B_i available at this cell, i.e., $\rho_i^{(k)} = \frac{\sum_{j \in \mathcal{N}} b_{i,j}^{(k)}}{B_i}$.*

We use $\boldsymbol{\rho}_i := [\rho_i^{(1)}, \ldots, \rho_i^{(K)}]^T \in [0,1]^K$ to denote the vector of cell loads at cell i for all time slots and denote the collection of all cell loads over time by $\mathbf{P} := [\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_M]^T \in [0,1]^{M \times K}$. A consequence of Definition 15 is the following fact:

Fact 4. *The load at cell i satisfies $\rho_i^{(k)} > 0$ if and only if (iff) cell i serves at least one user in slot k.*

In other words, $|\boldsymbol{\rho}_i\ \mathbf{1}|_0 = 0$ iff cell i serves no user in all time slots K, where $\mathbf{1} \in \mathbb{R}^K$ is a vector of ones and $|\cdot|_0$ is the l_0-norm. If $|\boldsymbol{\rho}_i\ \mathbf{1}|_0 = 0$ cell i can be deactivated for energy saving reasons.

Remark 11. *For simplicity we say that the service demand of all users starts at the same time $k = 0$ and endures the same time K. This simplification is no limitation to our scheme but more to make the formulas more readable.*

We consider buffered delay-sensitive applications which are characterized by a strict per time slot data rate requirement of each user. In more detail, the data transmitted to the user is stored in its buffer from which the delay-sensitive application reads with constant data rate $r_j^{\min}$. If the scheduling algorithm allocates a higher data rate to a user in time slot k, i.e., $r_{i,j}^{(k)} > r_j^{\min}$, then the additional transferred data remains in the users buffer $d_j^{(k)} = d_j^{(k-1)} + \Delta_k(r_{i,j}^{(k)} - r_j^{\min})$ and is read in the next time slot. If the user is not allocated a sufficiently high rate in slot k, i.e., $r_{i,j}^{(k)} < r_j^{\min}$, the buffer level decreases as $d_j^{(k)} = d_j^{(k-1)} - \Delta_k(r_j^{\min} - r_{i,j}^{(k)})$. In every time slot k the aggregate data rate from the buffer and streamed from a cell has to be large enough which yields the constraint[14]

$$\sum_{i \in \mathcal{M}} r_{i,j}^{(k)} + \frac{d_j^{(k-1)}}{\Delta_k} \geq r_j^{\min}. \tag{3.40}$$

The buffer level of user j at the end of time slot k is therefore described by

$$0 \leq d_j^{(k)} = d_j^{(k-1)} + \sum_{i \in \mathcal{M}} \Delta_k r_{i,j}^{(k)} - \Delta_k r_j^{\min}. \tag{3.41}$$

[14] Note, that this definition allows users to be served by multiple cells as well as the buffer in a time slot. In such cases fountain coding can be used to implement mutual information combing.

Since each base station has only B_i resource units to allocate to users we have the constraint

$$\sum_{j \in \mathcal{N}} \frac{b_{i,j}^{(k)}}{B_i} = \sum_{j \in \mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i \omega_{i,j}^{(k)}} \leq \rho_i^{(k)}. \tag{3.42}$$

3.8.2 Problem statement

We are now in the position to state the optimization problem that aims at finding the optimal set of active cells, user-cell assignments and rate allocations while consuming the least amount of energy. The objective function $E : [0,1]^{M \times K} \to \mathbb{R}_+$ is a combination of static and dynamic sources of energy consumption. In more detail, each active cell has static energy consumption of e_i per time slot and a load dependent part which is captured by a concave or convex function $f_i(\boldsymbol{\rho}_i) \in \mathbb{R}_+$ or simply f_i. The total network energy consumption in a so called *on/off scheme* is

$$E_{\text{on/off}}(P) = \sum_{i \in \mathcal{M}} K \; e_i \, |\boldsymbol{\rho}_i \; \mathbf{1}|_0 + f_i. \tag{3.43}$$

The above model assumes that cells are deactivated before the first time slot and stay inactive for all K time slots. The model can easily be adapted to modes of operation where so called micro-sleeps of cells are allowed. In such cases, a cell can be deactivated for a single time slot k in order to save energy and be activated in the next time slot $k+1$. The energy consumption of such a mode of operation is captured by

$$E_{\text{sleep}}(P) = \sum_{i \in \mathcal{M}} \sum_{k=1}^{K} e_i \left| \rho_i^{(k)} \right|_0 + f_i. \tag{3.44}$$

We refer to the scheme where micro-sleep of cells is allowed as *micro sleep scheme*.

The complete optimization problem for *buffered delay-sensitive applications* with the *on/off scheme* can be composed as

$$\begin{aligned}
\text{minimize} \quad & \sum_{i \in \mathcal{M}} K \; e_i \, |\boldsymbol{\rho}_i \; \mathbf{1}|_0 + f_i && && \text{(3.45a)} \\
\text{subject to} \quad & \sum_{j \in \mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i \omega_{i,j}^{(k)}} \leq \rho_i^{(k)} && i \in \mathcal{M}, k \in \mathcal{K} && \text{(3.45b)} \\
& \sum_{i \in \mathcal{M}} r_{i,j}^{(k)} + \frac{d_j^{(k-1)}}{\Delta_k} \geq r_j^{\min} && j \in \mathcal{N}, k \in \mathcal{K} && \text{(3.45c)} \\
& d_j^{(k-1)} + \sum_{i \in \mathcal{M}} \Delta_k r_{i,j}^{(k)} - \Delta_k r_j^{\min} = d_j^{(k)} && j \in \mathcal{N}, k \in \mathcal{K} && \text{(3.45d)} \\
& 0 \leq d_j^{(k)} && j \in \mathcal{N}, k \in \mathcal{K}, && \text{(3.45e)}
\end{aligned}$$

where the optimization variables are $r_{i,j}^{(k)} \in \mathbb{R}_+$ and $\rho_i^{(k)} \in [0,1]$. Thereby, (3.45b) assures

that cells are not overloaded and (3.45c) guarantees that users receive the required instantaneous data rate. Constraint (3.45d) represents the buffer level increase or decrease at each time slot k.

Problem (3.45) can be written in a more compact form as

$$\text{minimize} \quad \sum_{i\in\mathcal{M}} K\ e_i\,|\boldsymbol{\rho}_i\ \mathbf{1}|_0 + f_i \tag{3.46a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \leq \rho_i^{(k)} \qquad i\in\mathcal{M}, k\in\mathcal{K} \tag{3.46b}$$

$$\sum_{l=1}^{k}\left(\sum_{i\in\mathcal{M}} r_{i,j}^{(l)} - r_j^{\min}\right) \geq 0 \qquad j\in\mathcal{N}, k\in\mathcal{K}, \tag{3.46c}$$

since the buffer level at the end of time slot k can be stated as the data surplus of the aggregated data transmitted up to time slot k.

In a similar way we can state the problem for the *micro sleep scheme* which uses (3.44) and we obtain

$$\text{minimize} \quad \sum_{i\in\mathcal{M}}\sum_{k=1}^{K} e_i\left|\rho_i^{(k)}\right|_0 + f_i \tag{3.47a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \leq \rho_i^{(k)} \qquad i\in\mathcal{M}, k\in\mathcal{K} \tag{3.47b}$$

$$\sum_{l=1}^{k}\left(\sum_{i\in\mathcal{M}} r_{i,j}^{(l)} - r_j^{\min}\right) \geq 0 \qquad j\in\mathcal{N}, k\in\mathcal{K}, \tag{3.47c}$$

where the optimization variables are $r_{i,j}^{(k)} \in \mathbb{R}_+$ and $\rho_i^{(k)} \in [0,1]$.

Remark 12. *By adjusting the buffer size it is possible to introduce some sort of delay-robustness. The larger the buffer, the more data can be preloaded and the longer can a user be without any service from any cell. In an extreme case all data is sent in the first time slot and the rest of the time it is not served by any base station. In practice such scenarios are unrealistic. User behavior often skips videos forward or changes to request other data sets and in such cases the preloaded data would have wasted resources that could have been used otherwise.*

3.8.3 Network topology control for buffered delay-sensitive applications

Due to their non-convex objective function Problem 3.46 and Problem 3.47 are in general hard to solve. Fortunately, both problems exhibit a structure that can be exploited by low complexity algorithms to find good user-cell association and rate schedules consuming low energy. In more detail, we apply a relaxation of the l_0-norm in combination with the

Majorization-Minimization method as proposed in Section 3.3.3. In the following we sketch the solution approach.

The objective functions of Problem 3.46 and Problem 3.47 are not continuous due to the involved l_0-norm. To obtain an optimization problem that is mathematically tractable, we address the non-continuity problem of the l_0-norm by considering the following relation [CWB08]:

$$\forall_{\boldsymbol{z}\in\mathbb{R}^K} \; |\boldsymbol{z}|_0 = \lim_{\epsilon\to 0} \sum_{k=1}^{K} \frac{\log(1+|z_k|\;\epsilon^{-1})}{\log(1+\epsilon^{-1})}, \tag{3.48}$$

Thus, (3.46a) can be equivalently written as

$$\sum_{i\in\mathcal{M}} K\; e_i \,|\boldsymbol{\rho}_i\; \mathbf{1}|_0 + f_i = \lim_{\epsilon\to 0} \sum_{i\in\mathcal{M}} K\; e_i\; \frac{\log(1+\epsilon^{-1}\;\boldsymbol{\rho}_i\;\mathbf{1})}{\log(1+\epsilon^{-1})} + f_i. \tag{3.49}$$

Similarly, the objective function of the *micro sleep scheme* (3.47a) can be equivalently written as

$$\sum_{i\in\mathcal{M}}\sum_{k=1}^{K} e_i \left|\rho_i^{(k)}\right|_0 + f_i = \lim_{\epsilon\to 0} \sum_{i\in\mathcal{M}}\sum_{k=1}^{K} e_i\; \frac{\log(1+\epsilon^{-1}\;\rho_i^{(k)})}{\log(1+\epsilon^{-1})} + f_i. \tag{3.50}$$

We obtain a relaxed version of Problem 3.46 and Problem 3.47 by replacing the objective function by the right-hand side of (3.49) and (3.50), respectively and fixing $\epsilon > 0$ to a sufficiently small value which yields

$$\text{minimize} \sum_{i\in\mathcal{M}} K\; e_i\; \frac{\log(1+\epsilon^{-1}\;\boldsymbol{\rho}_i\;\mathbf{1})}{\log(1+\epsilon^{-1})} + f_i \tag{3.51a}$$

$$\text{subject to} \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \le \rho_i^{(k)} \qquad i\in\mathcal{M}, k\in\mathcal{K} \tag{3.51b}$$

$$\sum_{l=1}^{k}\sum_{i\in\mathcal{M}} (r_{i,j}^{(l)} - r_j^{\min}) \ge 0 \qquad j\in\mathcal{N} \tag{3.51c}$$

for the *on/off scheme*. Accordingly, the we state the relaxed form of the *micro sleep scheme* as

$$\text{minimize} \sum_{i\in\mathcal{M}}\sum_{k=1}^{K} e_i\; \frac{\log(1+\epsilon^{-1}\;\rho_i^{(k)})}{\log(1+\epsilon^{-1})} + f_i \tag{3.52a}$$

$$\text{subject to} \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \le \rho_i^{(k)} \qquad i\in\mathcal{M}, k\in\mathcal{K} \tag{3.52b}$$

$$\sum_{l=1}^{k}\sum_{i\in\mathcal{M}} (r_{i,j}^{(l)} - r_j^{\min}) \ge 0 \qquad j\in\mathcal{N}. \tag{3.52c}$$

Problem 3.51 and Problem 3.52 are still not easy to solve because we need to *minimize* a non-convex function over a convex set. Fortunately, [CWB08] presents an optimization framework based on the MM technique [HL04] to handle problems of this type. The

Table 3.3: Network parameters of the simulation [MET13, Sect. 4.2].

Parameter	Value
Antenna height of macro cells	53 m
Antenna height of micro cells	10 m
Carrier frequency of macro cells	800 MHz
Carrier frequency of micro cells	2500 MHz
Max. transmit power of macro cells	43 dBm
Max. transmit power of micro cells	30 dBm
e_i of macro cells	400 W
e_i of micro cells	100 W
B_i	100
Noise power spectral density	-145.1 dBm/Hz
ϵ	10^{-2}

framework can be used to decrease the value of the objective function in a computationally efficient way. For notational convenience we define the sets $\mathcal{X}_1$ and $\mathcal{X}_2$ to be the set of rate allocations $\mathbf{R} \in \mathbb{R}_+^{N\times M\times K}$ and cell loads $\mathbf{P} \in \mathbb{R}_+^{M\times K}$ satisfying constraints (3.51b)-(3.51c) and (3.52b)-(3.52c), respectively. In addition, we define the constant $\hat{e}_i := \frac{e_i}{\log(1+\epsilon^{-1})}$. Applying the MM technique to Problem 3.51 and Problem 3.52 yields a fast algorithm that iteratively solves

$$[(\mathbf{R},\mathbf{P})]^{[n+1]} \in \arg\min_{(\mathbf{R},\mathbf{P})\in\mathcal{X}_1} \sum_{i\in\mathcal{M}} \left(K\ \hat{e}_i \frac{\boldsymbol{\rho}_i \mathbf{1}}{\epsilon + [\boldsymbol{\rho}_i]^{[n]}\mathbf{1}} + \nabla f_i \left([\boldsymbol{\rho}_i]^{[n]} \right)^T \boldsymbol{\rho}_i \right) \tag{3.53}$$

for some feasible starting point and where we used the notation $[\cdot]^{[n]}$ to refer to the respective variable in the n-th iteration of the MM algorithm.

Analogous to (3.53) we arrive at the following iterative algorithm for Problem 3.52

$$[(\mathbf{R},\mathbf{P})]^{[n+1]} \in \arg\min_{(\mathbf{R},\mathbf{P})\in\mathcal{X}_2} \sum_{i\in\mathcal{M}} \left(\sum_{k=1}^{K} \hat{e}_i \frac{\rho_i^{(k)}}{\epsilon + [\rho_i^{(k)}]^{[n]}} + \nabla f_i \left([\boldsymbol{\rho}_i]^{[n]} \right)^T \boldsymbol{\rho}_i \right) \tag{3.54}$$

for some feasible starting point.

3.8.4 Empirical evaluation

In this section we present simulation results for the presented buffered delay-sensitive optimization framework for the *on/off scheme* and the *micro sleep scheme*. The numerical evaluation is based on the macro and micro layers of the "dense urban information society" scenario proposed in the METIS project [MET13, Sect. 4.2]. The basic layout is depicted in Figure 3.12 where three macro and 12 pico cells are deployed on rooftops and in the streets, respectively. The main parameters of the simulation scenario are listed in Table 3.3. To obtain the information used in Assumption 14 we use the channel gain data provided by the METIS consortium [MET15] for the scenario depicted in Figure 3.12. Additionally,

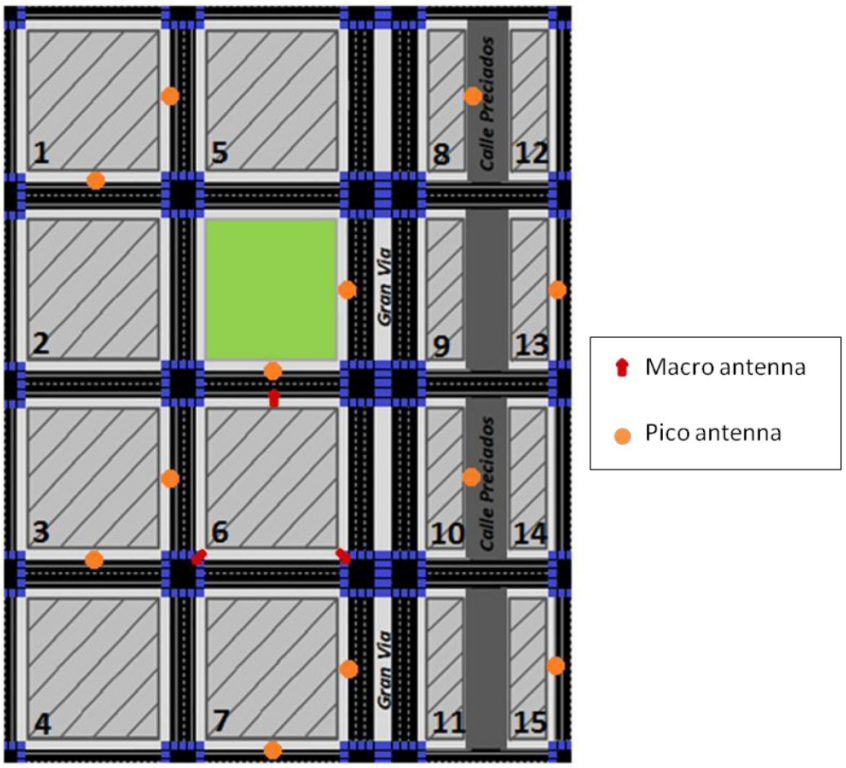


Figure 3.12: Basic deployment model of METIS TC2 [MET13] used in simulations for the buffered delay-sensitive optimization framework.

the provided mobility traces for cars are used to generate the traffic in our simulations. We focus on car users only due to their high mobility which will serve well to illustrate the gains achievable with anticipatory resource allocation. The data set for car users provides mobility traces of $N = 420$ different car users in the central deployment and a wrap around model is used to avoid boundary effects. The data set provides mobility traces for a time duration of 3600 seconds for each car user. Further details on the scenario and mobility model can be found in [MET13].

We normalize the energy consumption achievable by each scheme to the energy consumption of the network topology where all cells are active the whole time. In more detail, we normalize the network energy consumption by $E_{\text{on/off}}(\mathbf{1}) = E_{\text{sleep}}(\mathbf{1}) = K\left(\sum_{i\in\mathcal{M}} e_i + f_i(\mathbf{1})\right)$ and evaluate for our schemes

$$E'_{\text{on/off}}(\mathbf{R}) = \frac{E_{\text{on/off}}(\mathbf{R})}{K\left(\sum_{i\in\mathcal{M}} e_i + f_i(\mathbf{1})\right)} \tag{3.55}$$

and

$$E'_{\text{sleep}}(\mathbf{R}) = \frac{E_{\text{micro sleep}}(\mathbf{R})}{K\left(\sum_{i\in\mathcal{M}} e_i + f_i(\mathbf{1})\right)}. \tag{3.56}$$

In order to show the isolated effective gains from the anticipatory scheduling framework we neglect the dynamic energy consumption of hardware and use $f_i(\boldsymbol{\rho}) = 0,\ \forall i$. The evaluation of the effect of the dynamic energy consumption on the energy savings has been studied in Section 3.7.1.3 for energy saving algorithms without proactive resource allocation. We

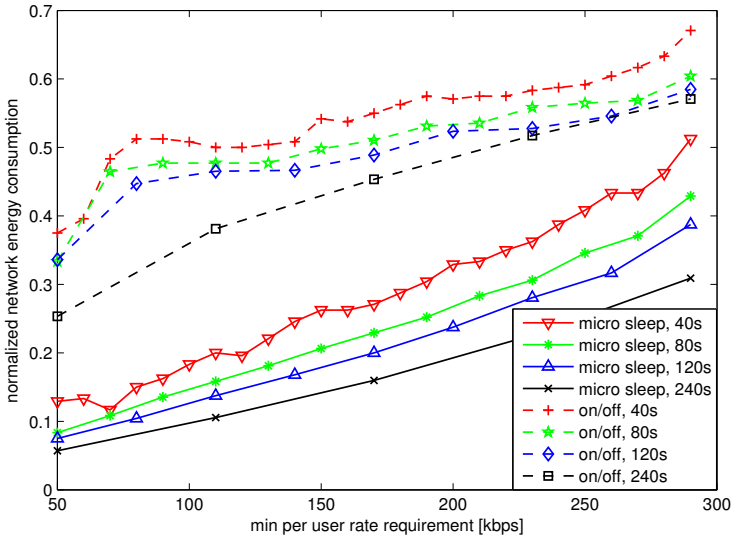


Figure 3.13: Normalized network energy consumption of the *micro sleep scheme* and the *on/off scheme* with increasing minimum per user rate requirement and for different optimization windows. Normalization with respect to all cells active the whole time.

expect the effects to carry over to proactive resource allocation schemes and the evaluation of which will be part of future studies.

We evaluate the proposed algorithm for the *on/off scheme* and the *micro sleep scheme* in terms of energy savings capabilities for the described scenario focusing on the influence of the number of time slots the optimization is done for. More precise, we find tuples of $(\mathbf{R}, \mathbf{P})$ for a different number of time slots $K \in \{40, 80, 120, 240\}$ with time slot duration $\Delta_k = 1s$. The starting point k_0 in the data set where we apply our optimization framework is selected uniformly at random from $\{1, 3600 - K + 1\}$ and the optimization window is selected as $\{k_0, \ldots, k_0 + K - 1\}$.

Figure 3.13 depicts the achievable normalized network energy consumption with increasing per user data rate requirement for the *on/off scheme* and the *micro sleep scheme*. We observe the intuitive result that with an increasing minimum per user data rate requirement the normalized network energy consumption increases. The more user demand there is in the network the less redundancies are in the system that can be exploited for energy savings. When comparing the two energy savings schemes it can be seen that the energy savings potentials with the *micro sleep scheme* are larger than the ones with the *on/off scheme*. The reason lies in the nature of the schemes that the *micro sleep scheme* allows the deactivation and activation of cells on a smaller time scale which in turn enables the scheme to save energy even for shorter time periods whereas the *on/off scheme* finds a set of active cells only for the full optimization window. The effect of the size of the optimization window is also evident from Figure 3.13. We can see that for both schemes the normalized network energy consumption is higher for a smaller optimization window.

A reason for this observation can be found in the better chances of larger optimization windows to preallocate resources for more data transmission in advance in order to free some cells from service provision.

3.9 Conclusion and final remarks

We have introduced an optimization framework for enhancing the energy efficiency of cellular networks. In cellular communication systems, problems of this type are hard to solve because of their combinatorial nature and the nontrivial interference coupling among cells. Indeed, even with a simplifying assumption of the worst-case interference, the energy saving problem is a mixed integer programming problem that is strongly related to the bin-packing problem, which in turn is known to be NP-hard. As a result, we cannot expect to find optimal solutions quickly. Instead we focused on fast sub-optimal heuristics. Unlike many existing approaches in the literature, the proposed methods can naturally consider both the dynamic and static energy consumption of base stations with multiple cells in heterogeneous networks. In the proposed heuristic, we relaxed the mixed integer programming problem to a form suitable for the application of majorization-minimization techniques. The resulting algorithm requires the solution of a series of linear programming problems that can be efficiently solved with standard mathematical solvers. Therefore, it can be applied to large-scale problems, and it is also suitable for online operation. By means of simulations we showed that good sparse solutions are obtained with only few iterations of the algorithm. We have also shown that our technique can outperform recent methods from literature in practical scenarios. In several simulation scenarios of different kind we highlighted the ability of the proposed algorithms to consider heterogeneous networks where base stations have different static and dynamic energy consumption. Finally, we showed the application of our algorithms to a network consisting of LTE and UMTS base stations.

Motivated by the good energy saving results we presented a second optimization framework that exploits the knowledge of user-cell trajectories and learned path-loss maps to find network topologies with low energy consumption. It finds user-cell association and rate schedules that provide the requested data rate of users and at the same time reduces the energy consumption of cellular communication networks. The used energy consumption model is general enough to capture static energy consumption for cooling, basic power conversion etc. as well as dynamic load dependent energy consumption. The key ingredient to these algorithms is the exploitation of end user devices' storage capabilities to implement buffered delay-sensitive applications with proactive resource allocation and user assignment. Thereby, we stretch the applicability of cell sleep and switching on/off techniques in the time horizon leading to significant energy savings. We have formalized the problem as a non-convex optimization problem and have presented relaxation techniques that are able to give good solutions to this problem in reasonable time making it amenable for online implementation. We evaluated the proposed *on/off scheme* and the *micro sleep scheme* in a realistic network deployment with realistic traffic demands and user mobility models. Results for the METIS TC2 deployment show good energy saving potentials for both schemes and we have empirically evaluated the influence of the optimization windows with respect to energy savings.

4 The Application of Interference Calculus to Network Topology Control for Load Aware Energy Savings

In Chapter 3 we have developed algorithms for energy savings in current communication networks that rely on the well-predicted behavior of traffic over time. To make those algorithms mathematically tractable we used a conservative approach which is not able to exploit the full potential of energy savings. The algorithms are designed to provide feasible solutions supporting the QoS requested by users even in the worst case scenario where the system is fully loaded and there is a lot of inter-cell interference. However, this assumption is very conservative since in practice, especially in low load situations, the interference is much lower and the QoS requirements could be provided with less resources of the cells. So by considering the true load values in the network topology optimization problem opens further opportunities to improve system performance in terms of either more energy savings or increased total system throughput. This becomes even more pronounced in todays and future networks where network designers need to address the ever-increasing data rate requirements in wireless networks. So by using such techniques, radio engineers are put into the position to evaluate if a given network configuration (base station tilts, power, etc.) is able to support the expected traffic demand. First results in this respect are available in literature [MK10, SY12] providing iterative techniques to compute the true cell load for particular interference coupling models in LTE systems. Alongside with these results, there has been considerable efforts to advance the theory in the field of interference calculus [SB12, Yat95]. This general framework is typically used to study interdependencies among users in interference-coupled systems. First results [FKVF13, CSS$^+$14, HYLS15] connect these two fields, namely the load/interference coupling models and interference calculus. Recent results show that the total transmit power in coupled cellular communication networks can be reduced without degrading the users' QoS when allowing the load to increase.

Based on these studies we present a unified framework to compute the actual load of cells which will put us in the position to answer questions of feasibility of given network configurations and the effectiveness of changes in the configuration (e.g., a change in antenna tilts). We combine the load based model of [SY12] for interference-coupled cellular networks with the framework of standard interference functions [Yat95]. Thereby, we are able to obtain a better approximation of the mutual interference and thus more accurate

cell load levels allowing for larger energy savings at the cost of slightly increased complexity and overhead. We are also able to derive novel algorithms for power computation that do not require nested iterative techniques such as that in [HYLS15]. In particular, if we are given the task to recompute power assignments to increase load (e.g., to decrease the transmit energy as discussed above), we can derive simple algorithms that can give information about the precision of the power estimates at each iteration. We also combine these techniques with the topology optimization techniques described in Chapter 3 in order to exploit the energy savings gains that were previously omitted due to the worst-case interference levels. Finally, we show how we can include the optimization of different system parameters such as antenna tilts. Even though, the resulting algorithms are based on heuristics, we demonstrate the energy savings gains compared to the pure topology optimization from Chapter 3.

4.1 Contribution

We build upon the energy savings techniques introduced in Chapter 3 and extend them towards realistic load models based on the general framework of interference calculus. Starting from available results for standard interference functions we use the result that the computation of load in cellular communication systems considered in our study can be posed as the problem of finding fixed points of standard interference mappings. This provides us with readily available, computational efficient techniques (standard iterative algorithms) to find the load at cells for a given network setup and puts us in the position to identify feasible network settings. These results are used to improve and extend the optimization algorithms for energy savings developed in Chapter 3. Finally, we present a framework which allows to incorporate finding feasible cell configurations (e.g., antenna tilts) supporting the energy savings algorithm that deactivate redundant network elements.

In the following we highlight the main contribution of the chapter:

- We show how the framework of standard interference functions can be applied to our problem of identifying feasible network setups. Thereby, the involved computations are of low complexity due to the involved fixed point algorithms that are used.
- We use interference mappings that are able to capture many practical limitations, such as an upper bound for the spectral efficiency or a constant signaling overhead.
- We present a non-heuristic stopping criterion to compute fixed points of our interference mappings that lead to an arbitrary but fixed precision.
- We show how the framework of standard interference functions is used to compute the actual load for a fixed network setup and how this information is used to improve the performance of our topology control algorithms from Chapter 3.
- With the help of heuristic arguments we combine the improved algorithm for topology control with the optimization of cell configurations. In particular, we show how the optimization of down tilts can help to save energy.

4.2 Related work

The problem of identifying the interference coupling in wireless communication systems has been a long standing problem. A significant body of work [MNK+07, MTHB07, MK10, SY12, FF12, FKVF13, HYS14, CSS+14, CSS15] relates the interference in the system to the cell load which is defined as the fraction of resource blocks used for transmission. Early research results [Gee08, MK10, SY12] introduce a model that computes the load via the solution of a system of non-linear equations for fixed network settings such as down-link powers and antenna tilts. The work in [MK10] is about wireless network design and optimization. The authors establish a model for cell load computation in large radio networks where the downlink bitrate follows a behavior approximating the Shannon-Hartley formula for achievable bitrates in a noisy transmit channel. This is used to deduce the network capacity and the suggested iterative algorithm is reasonably fast. The result is a local search algorithm to make decisions on site selections for network planning. The work in [SY12] presents a similar model characterizing the coupling relation between the cell load factors. Thereby, the intercell interference is taken into account for arbitrary network topologies enabling a performance analysis of the whole network in terms of resource consumption. The problem at hand is posed as a system of non-linear equations. A first attempt to solve these equations based on linearization and bounding is presented but ultimately, only sufficient and necessary conditions are derived for the existence of a solution.

In parallel, there has been advances in the theory of standard interference functions [Yat95, SWB09, SB12]. This framework provides computationally efficient fixed point algorithms to find solutions to fixed point problems of functions with specific properties. The applications of this framework are very broad. Its original application is rooted in power control in code division multiple access systems [Yat95] but it is applicable to many modeling and analyzing interference scenarios in multi-user systems.

First results that highlight the connection of load computation and interference calculus can be found in [FF12, FKVF13, CPS13a, CSS+14, CSS15]. It was shown that the problem of load computation can be posed as a fixed point problem involving standard interference mappings. The work in [FF12] identified the load coupling model as a standard interference function and proposed its solution by standard fixed point algorithms. In a parallel stream of work [CPS13a, CSS+14], concave mappings for network planning using load computation and interference have been proposed. Specific mappings have been proposed and the wider applicability has been highlighted. The solutions are obtained by iteratively solving fixed point problems involving standard interference mappings. The study [CSS15] improves the convergence speed of the standard fixed point iteration by constructing a matrix that has spectral radius strictly smaller than one if the corresponding mapping has a fixed point. The application is of particular interest in large network planning tasks where the computation of the fixed point may require time-consuming iterative methods and fast solutions are desired.

In [FKVF13], it was shown how to apply these results to a user association and antenna tilt optimization problem targeting at load balancing and maximizing the average resource available at cells. Motivated by this work we seek applications for energy savings as outlined later in this chapter.

[HYLS15] has shown that, the revers estimation problem, where the power is to be found inducing a particular load is of high practical relevance. A key result in this respect is that the total transmit power reaches its minimum when the load is at its maximum. It was shown that the reversed problem can also be posed as a fixed point problem but the involved interference mappings could not be obtained in closed form, which made nested iterative algorithms necessary. Such algorithms are likely to converge only after a large number of iterations.

4.3 Preliminaries and basic concepts

We include some basic definitions and concepts from convex analysis ([BV06]) and the theory of interference functions ([Yat95, SB12]) that are used in this chapter.

Definition 16 (Convex set)**.** *A set C is said to be* convex *if $\theta\boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2 \in C$ for every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ and $0 \leq \theta \leq 1$. If, in addition to being convex, C contains all its boundary points, then C is a* closed convex set*.*

Definition 17 (Convex and concave functions)**.** *A function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be* convex *if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ and $0 \leq \theta \leq 1$,*

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y}).$$

A function f is said to be concave *if $-f$ is convex.*

Fact 5 (*Selected properties of convex/concave functions*)**.** *For convex functions, we have the following properties [BC11, Ch. 8.2][BV06]:*

1. *The set of convex functions is closed under addition and multiplication by non-negative constants.*
2. *If $f_1 : \mathbb{R}^M \to \mathbb{R}$ is a convex function, then $f_2(\boldsymbol{x}) := f_1(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is a convex function, where $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, $\boldsymbol{x} \in \mathbb{R}^N$, and $\boldsymbol{b} \in \mathbb{R}^M$*
3. *If $f_1, f_2 : \mathbb{R}^M \to \mathbb{R}$ are two convex functions, then their point-wise maximum $f : \mathbb{R}^M \to \mathbb{R}$ defined by*
$$f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$$
is also convex.
4. *(Perspective) Let $f_1 : \mathbb{R}^M \to \mathbb{R}$ be a convex function, then the function $f_2 : \mathbb{R}^M \times \mathbb{R}_{++} \to \mathbb{R}$ defined by:*
$$f_2(\boldsymbol{x}, y) = y f_1\left(\frac{\boldsymbol{x}}{y}\right)$$

is convex.

5. *(Dimension reduction) If $f_1 : \mathbb{R}^M \to \mathbb{R}$ is a convex function, then $f_2 : \mathbb{R}^{M-1} \to \mathbb{R} : \boldsymbol{x} \mapsto f_1([\boldsymbol{x}^T\ 1]^T)$, which is obtained by fixing the last element argument of the function f_1 to one, is a convex function.*

The following concepts and definitions play an important role in explanations of this chapter. The presentation is based on [Yat95, SB12] and we refer to these references for a more detailed discussion about interference functions.

Definition 18 (Standard interference functions [Yat95])**.** *A function $I : \mathbb{R}_+^M \to \mathbb{R}_{++}$ is said to be a standard interference function if the following holds:*

1. *(*Positivity*) $I(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}_+^M$.*
2. *(*Scalability*) $\alpha I(\boldsymbol{x}) > I(\alpha \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}_+^M$ and all $\alpha > 1$.*
3. *(*Monotonicity*) $I(\boldsymbol{x}_1) \geq I(\boldsymbol{x}_2)$ if $\boldsymbol{x}_1 \geq \boldsymbol{x}_2$.*

Fact 6. *Concave functions $I : \mathbb{R}_+^M \to \mathbb{R}_{++}$ are standard interference functions [CSS+14].*

Definition 19 (Standard interference mappings)**.** *A mapping $\mathcal{J} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ is a standard interference mapping if it can be written as $\mathcal{J}(\boldsymbol{x}) = [I_1(\boldsymbol{x}) \dots I_M(\boldsymbol{x})]^T$, where each function $I_i : \mathbb{R}_+^M \to \mathbb{R}_{++}$ $(i = 1, \dots, M)$ is a standard interference function.*

Definition 20 (Fixed points)**.** *If a point $\boldsymbol{x} \in \mathbb{R}^M$ satisfies $\mathcal{J}(\boldsymbol{x}) = \boldsymbol{x}$ for a given mapping $\mathcal{J} : \mathbb{R}^M \to \mathbb{R}^M$, then $\boldsymbol{x}$ is said to be a fixed point of $\mathcal{J}$. The set of all fixed points of $\mathcal{J}$ is denoted by* $\mathrm{Fix}(\mathcal{J}) := \{\boldsymbol{x} \in \mathbb{R}^M \mid \boldsymbol{x} = \mathcal{J}(\boldsymbol{x})\}$.

Fact 7 (Existence and uniqueness of fixed points [Yat95])**.** *Let $\mathcal{J} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ be a standard interference mapping. If the mapping $\mathcal{J}$ has a fixed point, then the fixed point is unique. Furthermore, the fixed point exists if and only if $\mathcal{J}(\boldsymbol{x}') \leq \boldsymbol{x}'$ for some $\boldsymbol{x}' \in \mathbb{R}_{++}^M$.*

Fact 8 (Computation of fixed points [Yat95])**.** *Let $\mathcal{J} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ be a standard interference mapping having a fixed point and denote this fixed point by $\boldsymbol{x}^\star \in \mathbb{R}_{++}^M$. By Fact 7, the fixed point is unique. Then $\boldsymbol{x}^\star$ is the limit of the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ generated by*

$$\boldsymbol{x}_{n+1} = \mathcal{J}(\boldsymbol{x}_n), \tag{4.1}$$

where $\boldsymbol{x}_{-1} \in \mathbb{R}_+^M$ is arbitrary. In particular, if $\boldsymbol{x}_{-1} = \boldsymbol{0}$, then the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ is monotone non-decreasing ($\boldsymbol{x}_n \leq \boldsymbol{x}_{n+1}$ for every $n \in \mathbb{N}$). In contrast, for $\boldsymbol{x}_{-1} \geq \boldsymbol{x}'$, where $\boldsymbol{x}' \in \mathbb{R}_{++}^M$ is any point satisfying $\mathcal{J}(\boldsymbol{x}') \leq \boldsymbol{x}'$, then the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ is monotone non-increasing ($\boldsymbol{x}_n \geq \boldsymbol{x}_{n+1}$ for every $n \in \mathbb{N}$).

Fact 9. *Let $\mathcal{J} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ be a mapping with $\mathcal{J}(\boldsymbol{x}) =: [I_1(\boldsymbol{x}) \dots I_M(\boldsymbol{x})]^T$. If, for each $i \in \{1, \dots, M\}$, the function I_i is concave (and, by definition, strictly positive on the domain $\mathbb{R}_+^M$), then the following holds:*

1. *$\mathcal{J}$ is a standard interference mapping.*

2. *The fixed point, if it exists, is the solution to the following convex optimization problem [SB12]:*

$$\begin{aligned} &\text{maximize} && \mathbf{1}^T \boldsymbol{x} \\ &\text{subject to} && \mathbf{0} \leq \boldsymbol{x} \leq \mathcal{J}(\boldsymbol{x}) \end{aligned}$$

Fact 10 (Operations that preserve standard interference functions)**.** *For standard interference functions we have the following properties [Yat95, SB12]:*

1. *The set of interference functions is closed under addition and multiplication by strictly positive constants.*
2. *If* $\mathcal{I}_i : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ *(*$i \in \{1, \ldots, M\}$*) are standard interference functions, then* $\mathcal{I}'(\boldsymbol{x}) := \min_{i \in \{1,\ldots,M\}} \mathcal{I}_i(\boldsymbol{x})$ *and* $\mathcal{I}''(\boldsymbol{x}) := \max_{i \in \{1,\ldots,M\}} \mathcal{I}_i(\boldsymbol{x})$ *are standard interference functions.*

4.4 Standard system model

We use a similar system model as presented in Chapter 3. For convenience we summarize the main points here. Our system consists of M cells and N test points forming a cellular radio communication network. We denote the set of cells by $\mathcal{M} = \{1, 2, \ldots, M\}$ and the set of test points by $\mathcal{N} = \{1, 2, \ldots, N\}$. The QoS requirements are stated in terms of minimum data rate $r_j \in \mathbb{R}_{++}$. The assignment matrix $\boldsymbol{X} =: [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \{0, 1\}^{M \times N}$ assigns test points to cells, i.e., the component $x_{i,j}$ of $\boldsymbol{X}$ takes the value $x_{i,j} = 1$ if test point j is assigned to cell i or the value $x_{i,j} = 0$ otherwise.

In this chapter we incorporate a parameterized model to allow for different antenna settings at cells. In general, we use a number of Q parameters that can be adjusted (e.g., antenna tilt, antenna height, etc.). The set of possible configurations is denoted by $\mathcal{X}_i \subset \mathbb{R}^Q$. For a particular cell i let $\boldsymbol{\theta}_i \in \mathcal{X}_i$ be a vector representing its used antenna configuration. All antenna configuration of the whole network are collected in $\boldsymbol{\Theta} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) \in \mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

We denote the long term channel state information of the link between cell i and test point j by $g_{i,j} : \mathbb{R}^D \to \mathbb{R}_{++}$. Besides antenna configuration settings, each cell chooses a fixed $P_i \in \mathbb{R}_+$, which denotes the transmit power spectral density per minimum resource unit and we define $\boldsymbol{p} := [P_1, \ldots, P_M]^T$. The number of available resource blocks at cell i is limited to $B_i \leq B$ with $B \in \mathbb{R}_{++}$ the total number of resource blocks in the system. We reuse Definition 10 and denote the vector of cell loads by $\boldsymbol{\rho} := [\rho_1, \ldots, \rho_M]^T$ with ρ_i being the cell load of cell $i \in \mathcal{M}$. We have shown in Chapter 3 that the load vector can be characterized with the help of the link spectral efficiency per resource block $\omega'_{i,j} : [0, 1]^M \times \mathcal{X} \to \mathbb{R}_+$ obtained by the modified Shannon capacity equation [MNK+07]

which we define by

$$\begin{aligned}\omega'_{i,j}(\boldsymbol{\rho},\boldsymbol{\Theta}) &:= \eta^{\mathrm{BW}}_{i,j}\log_2\left(1+\frac{\gamma_{i,j}(\boldsymbol{\rho},\boldsymbol{\Theta})}{\eta^{\mathrm{SINR}}_{i,j}}\right)\\ &= \eta^{\mathrm{BW}}_{i,j}\log_2\left(1+\frac{P_i\ g_{i,j}(\boldsymbol{\theta}_i)}{\eta^{\mathrm{SINR}}_{i,j}\sum_{l\in\mathcal{M}\setminus\{i\}}P_l\ g_{l,j}(\boldsymbol{\theta}_l)\ \rho_l+\sigma^2}\right),\end{aligned} \tag{4.2}$$

where σ^2 is the noise density per resource block and $\eta^{\mathrm{BW}}_{i,j} \in \mathbb{R}_{++}$ ($\eta^{\mathrm{SINR}}_{i,j} \in \mathbb{R}_{++}$) is a constant accounting for the bandwidth (SINR) efficiency of the link between cell i and test point j. The last equation in (4.2) uses the definition of the SINR $\gamma_{i,j} : \mathbb{R}^M_+ \times \mathcal{X} \to \mathbb{R}_+$ between cell i and test point j [SY12, FF12]:

$$\gamma_{i,j}(\boldsymbol{\rho},\boldsymbol{\Theta}) := \frac{P_i\ g_{i,j}(\boldsymbol{\theta}_i)}{\sum_{l\in\mathcal{M}\setminus\{i\}}P_l\ g_{l,j}(\boldsymbol{\theta}_l)\ \rho_l+\sigma^2}, \tag{4.3}$$

The cell load ρ_i of cell i is computed by the fraction of resource blocks allocated to its users plus some fixed fraction of resource blocks $s_i \in [0,1]$ for signaling when the cell is active. More precise, with the link spectral efficiency at hand we can now compute the cell load ρ_i of cell i for a given assignment matrix $\mathbf{X}$ as the solution to the following system of nonlinear equations:

$$\begin{aligned}\rho_i &= \sum_{j\in\mathcal{N}}\frac{r_j}{B_i\omega'_{i,j}(\boldsymbol{\rho},\boldsymbol{\Theta})}x_{i,j} + s_i\left|\sum_{j\in\mathcal{N}}x_{i,j}\right|_0\\ &= \sum_{j\in\mathcal{N}}\frac{\lambda_{i,j}\ x_{i,j}}{\log_2\left(1+\frac{P_i\ g_{i,j}(\boldsymbol{\theta}_i)}{\eta^{\mathrm{SINR}}_{i,j}\left(\sum_{l\in\mathcal{M}\setminus\{i\}}P_l\ g_{l,j}(\boldsymbol{\theta}_l)\ \rho_l+\sigma^2\right)}\right)} + s_i\left|\sum_{j\in\mathcal{N}}x_{i,j}\right|_0\\ &\quad i\in\mathcal{M},\end{aligned} \tag{4.4}$$

with constants $\lambda_{i,j} := \frac{r_j}{B_i\eta^{\mathrm{BW}}_{i,j}}$. Unfortunately, solving the system of nonlinear equations in (4.4) is not a trivial task. In fact, even when we fix the cell configurations $\boldsymbol{\Theta}$ and the test point assignment matrix $\mathbf{X}$, the solution of (4.4), if it exists, can only be computed by means of iterative methods. If not explicitly stated, we assume that the cell configuration is not object to optimization and we use a known and constant configuration $\boldsymbol{\Theta}$ and drop the variable for notational convenience. Section 4.8 will address the task of how to choose $\boldsymbol{\Theta}$.

4.5 Network feasibility analysis

In the process of planning a network it becomes of utmost importance to identify if the designed network is feasible. Also when operating a network one needs to know if any changes to the configuration will make the network unstable. In this section we follow up on the question of how to identify if a network is feasible, i.e. if the coverage and capacity

requirements defined can be satisfied. The main results of this section were also presented by research work [SY12], using different arguments. For convenience, we formally define our notion of a feasible network.

Definition 21 (Feasible Network)**.** *For a given network configuration $\boldsymbol{\Theta} \in \mathcal{X}$ and assignment matrix $\boldsymbol{X} \in \{0,1\}^{M\times N}$, we say that the network is* feasible *if all test points can be served while guaranteeing that the load in each base station i is strictly smaller than its maximum allowed value; i.e., $\rho_i < \frac{B_i}{B}$ for every $i \in \mathcal{M}$.*

We use the following assumption in this section to be able to identify feasible networks in the sense of Definition 21.

Assumption 15. *The network configuration variable $\boldsymbol{\Theta} \in \mathcal{X}$, and assignment matrix $\boldsymbol{X} \in \{0,1\}^{M\times N}$ are fixed and arbitrary (test points can be assigned to only one cell; i.e., $\sum_{i\in\mathcal{M}} x_{i,j} = 1$ for every $j \in \mathcal{N}$).*

Since by Assumption 15 the network configuration variable is fixed we omit the variable $\boldsymbol{\Theta} \in \mathcal{X}$ and write $g_{i,j} := g_{i,j}(\boldsymbol{\theta}_i)$.

Assumption 16. *All cells considered in the feasibility analysis are active, i.e., $\sum_{j\in\mathcal{N}} x_{i,j} > 0, i \in \mathcal{M}$.*

Assumption 16 is no limitation to our framework. It can be easily assured by removing all cells from set $\mathcal{M}$ that do not serve any test points.

With these assumptions we define the mapping:

$$\mathcal{J} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M : \boldsymbol{\rho} \mapsto [I_1(\boldsymbol{\rho}) \ \dots \ I_M(\boldsymbol{\rho})]^T , \tag{4.5}$$

where $I_i : \mathbb{R}_+^M \to \mathbb{R}_+$ $(i \in \mathcal{M})$ is given by

$$I_i(\boldsymbol{\rho}) := \sum_{j\in\mathcal{N}} \frac{\lambda_{i,j}\ x_{i,j}}{\log_2\left(1 + \frac{P_i\ g_{i,j}}{\eta_{i,j}^{\mathrm{SINR}}\ \sum_{l\in\mathcal{M}\setminus\{i\}} P_l\ g_{l,j}\ \rho_l + \sigma^2}\right)} + s_i \left|\sum_{j\in\mathcal{N}} x_{i,j}\right|_0 . \tag{4.6}$$

By means of the above mapping the cell load $\boldsymbol{\rho}$ in (4.4) can be equivalently written as

$$\begin{aligned} \rho_1 &= I_1(\boldsymbol{\rho}) \\ &\vdots \\ \rho_M &= I_M(\boldsymbol{\rho}). \end{aligned} \tag{4.7}$$

The system in (4.7) allows us to compute the load as the fixed-point $\boldsymbol{\rho} \in \mathrm{Fix}(\mathcal{J})$, if it exists, i.e., $\mathrm{Fix}(\mathcal{J}) \neq \emptyset$. Thus, the study of network feasibility amounts to studying the properties of the mapping $\mathcal{J}$. In particular, the question of existence and uniqueness of $\mathrm{Fix}(\mathcal{J})$ has to be answered.

Proposition 2. *Let the network configuration and test point assignment be fixed, then I_i, $i \in \mathcal{M}$ in (4.6) is a standard interference function and $\mathcal{J}$ in (4.5) is a standard interference mapping ([FKVF13, CSS+14, HYLS15]).*

Proof. The proof follows from the observation that the function $f : \mathbb{R} \to \mathbb{R} : u \mapsto \frac{1}{\log_2(1+\frac{1}{u})}$ is a concave function for $u > 0$ since it has has properties $f'(u) > 0$ and $f''(u) < 0$. Applying certain convexity preserving operations from Fact 5 we obtain the concavity of I_i, $i \in \mathcal{M}$:

$$
\begin{aligned}
& f(u) \ \text{ concave} \\
\underset{\text{(Fact 5.2)}}{\Rightarrow} & f\left(\frac{\eta_{i,j}^{\text{SINR}}}{P_i\ g_{i,j}}\left(\sum_{l\in\mathcal{M}\setminus\{i\}} P_l\ g_{l,j}\ \rho_l + \sigma^2\right)\right) \ \text{ concave } \ (\text{w.r.t. } \boldsymbol{\rho}) \\
\underset{\text{(Fact 5.1)}}{\Rightarrow} & \sum_{j\in\mathcal{N}} \lambda_{i,j} x_{i,j} f\left(\frac{\eta_{i,j}^{\text{SINR}}}{P_i\ g_{i,j}}\left(\sum_{l\in\mathcal{M}\setminus\{i\}} P_l\ g_{l,j}\ \rho_l + \sigma^2\right)\right) + s_i \left|\sum_{j\in\mathcal{N}} x_{i,j}\right|_0 \ \text{ concave.}
\end{aligned}
$$

Incorporating Fact 6 shows that I_i is a standard interference function. Furthermore, $\mathcal{J}$ is strictly positive; i.e., $\mathcal{J}(\boldsymbol{\rho}) > \mathbf{0}$ for every $\boldsymbol{\rho} \in \mathbb{R}_+^M$ (due to Assumption 16) and it follows from Fact 9.1. that $\mathcal{J}$ is a standard interference mapping. □

From the result that we identified $\mathcal{J}$ as a standard interference mapping many conclusions immediately follow. We know from Fact 7 that the fixed point of $\mathcal{J}$, if it exists, is unique and thus, also the load vector of the network $\boldsymbol{\rho} \in \text{Fix}(\mathcal{J})$. Additionally, we have low complexity schemes at hand to compute the fixed point by Fact 8 and Fact 9.2. In detail, there exist a plethora of efficient numerical methods to solve the convex optimization problem shown in Fact 9.2 [BV06, BC11, YO04, SB12, YYY11]. The test for network feasibility then reduces to checking Definition 21 for the obtained fixed point, i.e., $\rho_i < B_i/B$ for every $i \in \mathcal{M}$.

Unfortunately, because of Fact 7 the standard interference mapping in (4.5) only has a fixed point if there exists a $\boldsymbol{\rho}' \in \mathbb{R}_{++}^M$ satisfying $\mathcal{J}(\boldsymbol{\rho}') \leq \boldsymbol{\rho}'$. In order to find such a $\boldsymbol{\rho}'$ we define a new mapping $\underline{\mathcal{J}} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$

$$
\underline{\mathcal{J}}(\boldsymbol{\rho}) := \min\left(\mathcal{J}(\boldsymbol{\rho}), [\Gamma_1, \ldots, \Gamma_M]^T\right), \tag{4.8}
$$

that is guaranteed to have a fixed point. In (4.8) $\boldsymbol{\Gamma} := [\Gamma_1, \ldots, \Gamma_M]^T$ are large constants and the min-operator should be understood as a component-wise operation. In words, we impose limits on the maximum load of each cell to derive the mapping $\underline{\mathcal{J}}(\boldsymbol{\rho})$ from our original mapping $\mathcal{J}(\boldsymbol{\rho})$

We know $\underline{\mathcal{J}}(\boldsymbol{\rho})$ is also a standard interference mapping (Fact. 10) and that it is also bounded from above. Now all conditions of Fact 7 are fulfilled and we conclude that the fixed-point always exists and is unique.

Algorithm 4 Non-heuristic stopping criterion to compute fixed points of interference mappings

Input: Standard interference mapping $\mathcal{J} : \mathbb{R}_{+}^{M} \to \mathbb{R}_{++}^{M}$
Point $\boldsymbol{\rho}'$ satisfying $J(\boldsymbol{\rho}') \leq \boldsymbol{\rho}'$
Desired precision $\epsilon > 0$.

Output: Point $\boldsymbol{\rho} \in \mathbb{R}_{++}^{M}$ satisfying $\|\boldsymbol{\rho} - \boldsymbol{\rho}^{\star}\|_{\infty} \leq \epsilon$, where $\boldsymbol{\rho}^{\star} \in \mathrm{Fix}(\mathcal{J})$.

$\underline{\boldsymbol{\rho}}^{(-1)} = \mathbf{0}$
$\overline{\boldsymbol{\rho}}^{(-1)} = \boldsymbol{\rho}'$
$n = -1$
while $\|\overline{\boldsymbol{\rho}}^{(n)} - \underline{\boldsymbol{\rho}}^{(n)}\|_{\infty} > \epsilon$ **do**
  $\underline{\boldsymbol{\rho}}^{(n+1)} = \mathcal{J}(\underline{\boldsymbol{\rho}}^{(n)})$
  $\overline{\boldsymbol{\rho}}^{(n+1)} = \mathcal{J}(\overline{\boldsymbol{\rho}}^{(n)})$
  $n \leftarrow n + 1$
end while
return either $\underline{\boldsymbol{\rho}}^{(n)}$ (if a lower bound of the fixed point with the desired precision is desired) or $\overline{\boldsymbol{\rho}}^{(n)}$ (if an upper bound of the fixed point with the desired precision is desired).

In the remainder of this section we present a low-complexity scheme to compute the fixed point $\boldsymbol{\rho}^{\star} \in \mathrm{Fix}(\mathcal{J})$. In particular, we use the iterative method following from Fact 8 and introduce a non-heuristic stopping criterion to achieve an arbitrary precision $\epsilon > 0$. Starting from $\underline{\boldsymbol{\rho}}^{(-1)} = \mathbf{0}$ and $\overline{\boldsymbol{\rho}}^{(-1)} = \boldsymbol{\Gamma}$ we construct two sequences $\{\underline{\boldsymbol{\rho}}^{(n)}\}_{n\in\mathbb{N}}$ and $\{\overline{\boldsymbol{\rho}}^{(n)}\}_{n\in\mathbb{N}}$ derived from (4.1). The sequences $\{\underline{\boldsymbol{\rho}}^{(n)}\}_{n\in\mathbb{N}}$ and $\{\overline{\boldsymbol{\rho}}^{(n)}\}_{n\in\mathbb{N}}$ are monotonically increasing and decreasing respectively and by Fact 8 their limits are the unique fixed point $\boldsymbol{\rho}^{\star}$ of $\mathcal{J}$. The non-heuristic stopping criterion is based on the observation that we have $\underline{\rho}_i^{(n)} \leq \rho_i^{\star} \leq \overline{\rho}_i^{(n)}$ for every $n \in \mathbb{N}$ and $i \in \mathcal{M}$ by construction (Fact 8). We stop the algorithm at iteration $N \in \mathbb{N}$ when

$$\max_{i\in\mathcal{M}}(\overline{\rho}_i^{(N)} - \underline{\rho}_i^{(N)}) \leq \epsilon \tag{4.9}$$

is valid. We can use $\underline{\boldsymbol{\rho}}^{(N)}$ or $\overline{\boldsymbol{\rho}}^{(N)}$ as an approximation of $\boldsymbol{\rho}^{\star}$ because the relations

$$\|\underline{\boldsymbol{\rho}}^{(N)} - \boldsymbol{\rho}^{\star}\|_{\infty} = \max_{i\in\mathcal{M}} \left|\underline{\rho}_i^{(N)} - \rho_i^{\star}\right| \leq \epsilon \tag{4.10}$$

and

$$\|\overline{\boldsymbol{\rho}}^{(N)} - \boldsymbol{\rho}^{\star}\|_{\infty} = \max_{i\in\mathcal{M}} \left|\overline{\rho}_i^{(N)} - \rho_i^{\star}\right| \leq \epsilon \tag{4.11}$$

are valid for all $n \geq N$ and all components of the vectors $\underline{\boldsymbol{\rho}}^{(N)}$ and $\underline{\boldsymbol{\rho}}^{(N)}$ are within ϵ from their corresponding components of the fixed point $\boldsymbol{\rho}^{\star}$. The application of the non-heuristic stopping criterion is summarized in Algorithm 4.

Numerical evaluation of network feasibility

In this section we present the application of the network feasibility test and perform a numerical evaluation to show its performance in a practical scenario. We deploy a total

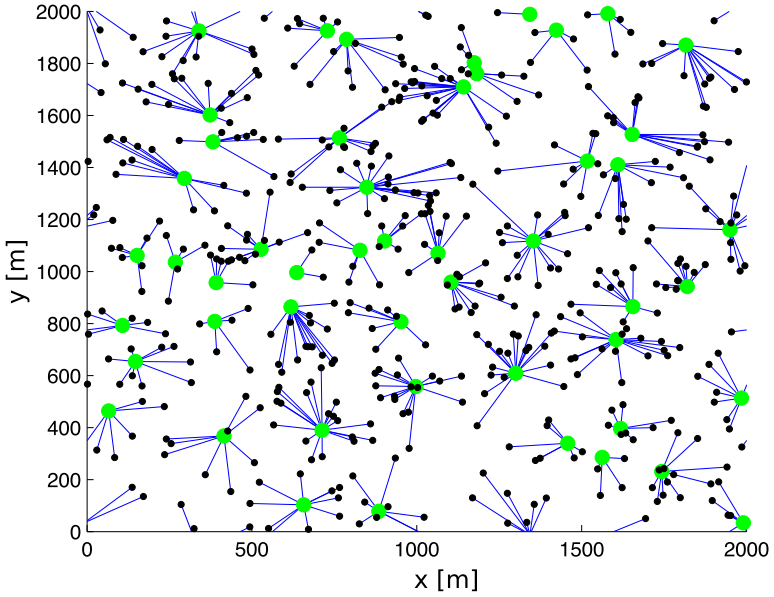


Figure 4.1: Exemplary network configuration used in the analysis. Solid green circles correspond to the cells that are deployed to provide service to the test points. The solid blue lines show the resulting test point - cell assignments. Note that we use a wrap-around model.

number of $M = 50$ cells in a square shaped area of size 2x2km according to a uniform distribution. Users are represented by $N = 500$ test points dropped uniformly at random in the area. A wrap around model is used to avoid boundary effects. We use the ITU propagation model for urban macro cell environments with fixed antenna patterns [3GP10a] and the test point assignment is based on a strongest serving cell metric. For illustration we depict an exemplary deployment in Figure 4.1. We apply Algorithm 4 to the described simulation scenario and evaluate the load evolution as a function of the number of iterations. We choose the starting points $\underline{\boldsymbol{\rho}}^{(-1)} = \mathbf{0}$ and $\overline{\boldsymbol{\rho}}^{(-1)} = \boldsymbol{\Gamma} = \mathbf{1}$ where $\mathbf{1}$ is a vector of all ones of appropriate size. Further simulation parameters are presented in Table 4.1 if not stated otherwise. As a reference we also track the load evolution obtained by (4.1) with a random initialization $\widetilde{\boldsymbol{\rho}}^{(-1)} \in [0,1]^M$ with each element drawn from a uniform distribution $\mathcal{U}(0,1)$.

For an exemplary cell the simulated load evolution with increasing number of iterations is depicted in Figure 4.2. The load evolution for $\underline{\boldsymbol{\rho}}^{(-1)}$, $\overline{\boldsymbol{\rho}}^{(-1)}$ and $\widetilde{\boldsymbol{\rho}}^{(-1)}$ is shown in green, blue and red, respectively. By observation we can confirm the monotonicity of the sequences $\{\underline{\boldsymbol{\rho}}^{(n)}\}_{n\in\mathbb{N}}$ and $\{\overline{\boldsymbol{\rho}}^{(n)}\}_{n\in\mathbb{N}}$ and that they are lower and upper bounds, respectively. Furthermore, we observe that in this particular scenario, the termination criterion (4.9) is satisfied after only three iterations resulting in an accuracy of $\epsilon = 10^{-5}$. The load evolution at other cells exhibits a similar behavior.

We now turn to the actual task of identifying feasible network configurations. To do so, we modify the same scenario as before to artificially simulate overload situations. In more

	Parameter	Value
Communication area	square shaped	2km × 2km
Number of test points	N	500
Number of cells	M	50
Data rate	r_j	16 kbit/s
Transmit power	P_i	46 dBm, $\forall i \in \mathcal{M}$
Bandwidth efficiency	$\eta_{i,j}^{\mathrm{BW}}$	1.5
SINR efficiency	$\eta_{i,j}^{\mathrm{SINR}}$	0.6
Signaling overhead	s_i	0.1
System bandwidth	B	5MHz
Upper bound of cell load	Γ_i	1, $\forall i \in \mathcal{M}$
Accuracy level	ϵ	10^{-5}
Lower starting point	$\underline{\boldsymbol{\rho}}^{(-1)}$	$\mathbf{0}$
Upper starting point	$\overline{\boldsymbol{\rho}}^{(-1)}$	$\boldsymbol{\Gamma}$

Table 4.1: Standard simulation parameters for network feasibility evaluation.

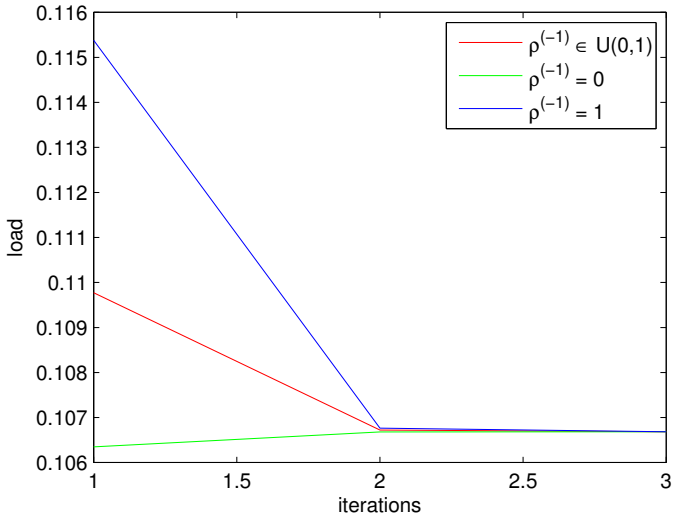


Figure 4.2: Load values generated by the fixed point algorithm as a function of the number of iterations for three initial starting points for an arbitrarily chosen cell (cell 36).

detail, we gradually increase the data rate requirements r_j of each test point to induce more demand in the network. In addition, we use $\Gamma_i > 1,\ i \in \mathcal{M}$ to raise the upper bound on the cell load in our algorithm. In particular, we use $\Gamma_1 = \ldots = \Gamma_M = 10$ for these simulations. The graph in Figure 4.3 shows the load after n iterations with accuracy of $\epsilon = 10^{-5}$ for an increasing system sum rate requirement which is the aggregated data rate demand from all test points. We first observe that the cell load of cells is increasing with increasing rate requirement at test points. Three out of the five randomly chosen cells (cell 1, cell 11 and cell 38) exhibit feasible load values ($\rho_i \leq 1$) for all selected rate requirements, whereas cell 50 exceeds load level $\rho_{50} > 1$ at a system sum rate requirement of $\approx$ 370Mbit/s and cell 21 has $\rho_{21} > 1$ at around $\approx$ 320Mbit/s sum rate requirement.

To identify at which system sum data rate requirement the first cell is overloaded we

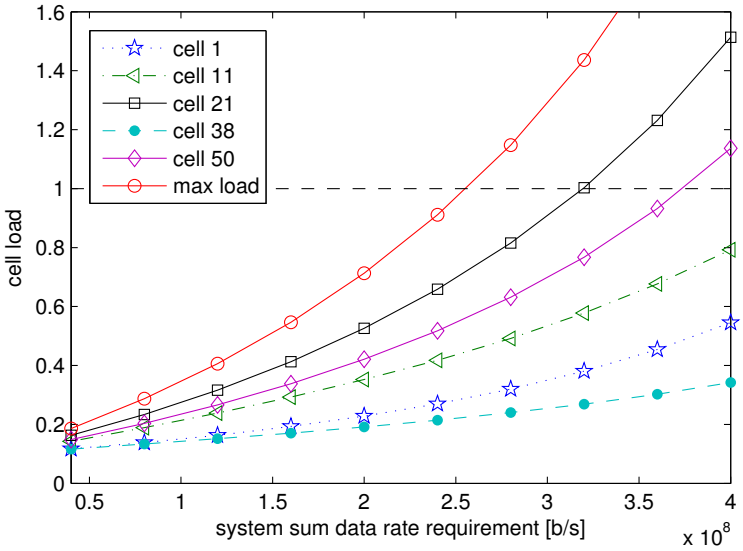


Figure 4.3: Cell loads as a function of the system sum rate for five randomly chosen cells. In addition the cell load of the cell with the highest load is shown (red curve with circle markers).

evaluate $\max_{i\in\mathcal{M}} \rho_i > 1$ which is also plot in Figure 4.3. The feasibility of this particular system setup and data rate demand is given up to a system sum rate requirement of $\approx$ 260Mbit/s.

4.6 Transmit power planning in LTE systems

A considerable body of work has investigated the load characterization in networks based on standard load coupling models for LTE-like systems [MK10, SY12]. A recent study [HYLS15] has highlighted the importance of a problem strongly related to that, namely, that of computing the downlink transmit power inducing a given load profile. The authors of that study prove that the users' rate requirements can be satisfied with lower transmit power if the load at each base station is allowed to increase. In addition, by using the framework of standard interference functions the power assignment of base stations inducing a given load profile can be computed as the fixed point of an interference mapping. Motivated by these results we derive, in this section, an interference mapping having as its fixed point the power allocation inducing a given load profile and propose an algorithm for power computation. In particular, if we are given the task to recompute power assignments to increase load (e.g., to decrease the transmit energy as discussed above), we can derive simple algorithms that can give information about the precision of the power estimates at each iteration.

Recall that the load ρ_i at cell $i \in \mathcal{M}$ is the fraction of resources used to serve its users. Furthermore, we have shown in Section 4.5 that for a fixed power assignment $\boldsymbol{p} \in \mathbb{R}^M_{++}$, the load

vector can be obtained by computing the fixed point of the mapping

$$\mathcal{J} : \mathbb{R}_{+}^{M} \to \mathbb{R}_{++}^{M} : \boldsymbol{\rho} \mapsto [I_1(\boldsymbol{\rho}), \ldots, I_M(\boldsymbol{\rho})]^T, \tag{4.12}$$

with I_i defined in (4.6). Since the mapping $\mathcal{J}$ was identified to be a standard interference mapping [CSS+14, FKVF13, HYLS15] we can apply standard techniques to compute its fixed point (if it exists); i.e., we can use the standard iterative algorithm $\boldsymbol{\rho}_{n+1} = \mathcal{J}(\boldsymbol{\rho}_n)$ with an arbitrary starting point $\boldsymbol{\rho}_1 \in \mathbb{R}_{+}^{M}$ as outlined in Section 4.5. Once we have obtained the load $\boldsymbol{\rho}^\star \in \mathrm{Fix}(\mathcal{J})$ the total transmit power radiated by all cells in the network is given by $P_{\Sigma}^{\mathrm{TX}} := \sum_{i \in \mathcal{M}} B_i \rho_i^\star P_i$.

In [HYLS15] the authors have highlighted that a fully loaded system, i.e., $\boldsymbol{\rho} = \mathbf{1}$ is optimal in terms of transmit energy consumption. In other words, to have the smallest transmit energy consumption, all cells should use all their resources. So in what follows we investigate how to set the downlink transmit power vector $\boldsymbol{p}$ to have a feasible system that consumes the least amount of transmit energy P_{Σ}^{TX}. To answer the above question we need to solve the system of nonlinear equations (c.f. (4.4) ignoring unnecessary constants)

$$\rho_i = \sum_{j \in \mathcal{N}_i} \frac{r_j}{B_i \omega'_{i,j}(\boldsymbol{\rho}, \boldsymbol{p})}, \; i \in \mathcal{M}, \tag{4.13}$$

for the power vector $\boldsymbol{p} \in \mathbb{R}_{++}^{M}$ and a fixed but given cell load $\boldsymbol{\rho} \in \mathbb{R}_{++}^{M}$. We define a function $\tilde{\mathcal{P}}_{\boldsymbol{\rho},i} : \mathbb{R}_{++}^{M} \to \mathbb{R}_{++}$, that is obtained by multiplying both sides of (4.13) by $\frac{P_i}{\rho_i} > 0$:

$$P_i = \frac{P_i}{\rho_i} \sum_{j \in \mathcal{N}_i} \frac{r_j}{B_i \omega'_{i,j}(\boldsymbol{\rho}, \boldsymbol{p})} =: \tilde{\mathcal{P}}_{\boldsymbol{\rho},i}(\boldsymbol{p}), \quad i \in \mathcal{M}. \tag{4.14}$$

We continue by showing that the solution to (4.13) and (4.14), if it exists, can be computed as a fixed point of a standard interference mapping. The particular standard interference mapping relies on the following proposition.

Proposition 3. *The function $\tilde{\mathcal{P}}_{\boldsymbol{\rho},i} : \mathbb{R}_{++}^{M} \to \mathbb{R}_{++}$ defined in (4.14) is concave for every $i \in \mathcal{M}$.*

Proof. First notice that the function $f^{(1)} : \mathbb{R} \to \mathbb{R} : u \mapsto \frac{1}{\log_2(1+\frac{1}{u})}$ is concave. Let $\boldsymbol{p}_{-i} \in \mathbb{R}_{++}^{M-1}$ be a power vector obtained by excluding the ith component of the power vector $\boldsymbol{p}$, where i is arbitrary. Now we can verify that

$$f_{i,j}^{(2)} : \mathbb{R}_{++}^{M} \to \mathbb{R}_{++} : \begin{bmatrix} \boldsymbol{p}_{-i} \\ \sigma^2 \end{bmatrix} \mapsto \frac{r_j}{B_i} f^{(1)} \left(\sum_{k \in \mathcal{M} \setminus \{i\}} \rho_k P_k g_{k,j} + \sigma^2 \right),$$

is concave for $i \in \mathcal{M}$ and $j \in \mathcal{N}$ by Fact 5.2. Furthermore, by Fact 5.4, the function

$$f_{i,j}^{(3)} : \mathbb{R}_{++}^{M+1} \to \mathbb{R}_{++} : \begin{bmatrix} \boldsymbol{p} \\ \sigma^2 \end{bmatrix} \mapsto P_i f_{i,j}^{(2)} \left(\frac{1}{P_i} \begin{bmatrix} \boldsymbol{p}_{-i} \\ \sigma^2 \end{bmatrix} \right)$$

is concave. Applying Fact 5.5 to $f_{i,j}^{(3)}$ we arrive at a new concave function $f_{i,j}^{(4)} : \mathbb{R}^M \to \mathbb{R} : \boldsymbol{p} \mapsto \frac{P_i r_j}{(B_i \omega'_{i,j}(\boldsymbol{\rho},\boldsymbol{p}))}$. Concavity of $\tilde{\mathcal{P}}_{\boldsymbol{\rho},i}$ now follows from this last result and Fact 5.1. □

For $\tilde{\mathcal{P}}_{\boldsymbol{\rho},i}$ to be standard interference function we have to extend it to the closure of its domain. For every $i \in \mathcal{M}$, the concave function $\tilde{\mathcal{P}}_{\boldsymbol{\rho},i} : \mathbb{R}_{++}^M \to \mathbb{R}_{++}$ can be continuously extended to the domain $\mathbb{R}_+^M$. This extension, denoted by $\mathcal{P}_{\boldsymbol{\rho},i} : \mathbb{R}_+^M \to \mathbb{R}_{++}$, which is also a concave function, is given by [CSZZ16]

$$\mathcal{P}_{\boldsymbol{\rho},i}(\boldsymbol{p}) = \begin{cases} \dfrac{P_i}{\rho_i} \sum_{j \in \mathcal{N}_i} \dfrac{r_j}{B_i \omega'_{i,j}(\boldsymbol{\rho},\boldsymbol{p})}, & \text{if} \quad P_i \neq 0 \\ \sum_{j \in \mathcal{N}_i} \dfrac{r_j \ln 2}{B_i g_{i,j} \rho_i} \left(\sum_{k \in \mathcal{M} \setminus \{i\}} \rho_k P_k g_{k,j} + \sigma^2 \right), \text{otherwise.} \end{cases} \tag{4.15}$$

With the above extension, we have that $\mathcal{P}_{\boldsymbol{\rho},i}$ is also a positive concave function for every $i \in \mathcal{M}$ and as a result, we can apply Fact 6 to conclude that the mapping $\mathcal{P}_{\boldsymbol{\rho}} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ by $\mathcal{P}_{\boldsymbol{\rho}}(\boldsymbol{p}) := [\mathcal{P}_{\boldsymbol{\rho},1}(\boldsymbol{p}), \ldots, \mathcal{P}_{\boldsymbol{\rho},M}(\boldsymbol{p})]^T$ is a standard interference mapping. By Fact. 7, the fixed point $\boldsymbol{p}^\star \in \mathrm{Fix}(\mathcal{P}_{\boldsymbol{\rho}})$, if it exists, is unique and strictly positive. These facts imply the equivalence between the solution of the nonlinear system in (4.14) and the fixed point $\boldsymbol{p}^\star$ of $\mathcal{P}_{\boldsymbol{\rho}}$.

A practical consequence of the above is that the power assignment $\boldsymbol{p}$ inducing a given load $\boldsymbol{\rho}$ (if it exists) is the limit of the sequence $\{\boldsymbol{p}_n\}$ generated by $\boldsymbol{p}_{n+1} = \mathcal{P}_{\boldsymbol{\rho}}(\boldsymbol{p}_n)$, where $\boldsymbol{p}_1 \in \mathbb{R}_+^M$ is arbitrary. Note that this simple iterative scheme eliminates the need for the bisection technique required by the scheme in [HYLS15]. In the remainder of this section we develop an iterative algorithm that provides the power vector $\boldsymbol{p}$ in order to increase the load of the current network configuration. In words, for a given power assignment $\boldsymbol{p}'$ that induces a load $\boldsymbol{\rho}'$ we iteratively find a power vector $\boldsymbol{p}''$ that increase the load from $\boldsymbol{\rho}'$ to $\boldsymbol{\rho}'' \geq \boldsymbol{\rho}'$. All other parameters of the model are fixed. We will see that the algorithm has the property $\boldsymbol{p}'' < \boldsymbol{p}'$ and that $\rho_i'' P_i'' < \rho_i' P_i'$ for every $i \in \mathcal{M}$ which shows that the users' data rate requirements can be satisfied with lower transmit power if we allow the load to increase[1]. The proposed algorithm is based on Proposition 4, and the proof of which requires the following lemma.

Lemma 1. *The function $f : \mathbb{R}_{++} \to \mathbb{R}_{++} : x \mapsto x \ln(1 + 1/x)$ is strictly increasing; i.e., $y, x \in \mathbb{R}_{++}$ with $y > x$ implies $f(y) > f(x)$.*

Proof. First recall that $\dfrac{y}{y+1} < \ln(1+y)$ for every $y > 0$ [Lov80]. Now, for $y = 1/x \in \mathbb{R}_{++}$, we deduce: $0 < \ln\left(1 + \dfrac{1}{x}\right) - \dfrac{1}{1+x} = f'(x)$ for every $x \in \mathbb{R}_{++}$, which implies the desired result. □

Proposition 4. *Let $\boldsymbol{\rho}' \in \mathbb{R}_{++}^M$ be the load corresponding to the power assignment $\boldsymbol{p}' \in \mathbb{R}_{++}^M$; i.e., $\boldsymbol{\rho}' \in \mathrm{Fix}(\mathcal{J}_{\boldsymbol{p}'})$, or, equivalently, $\boldsymbol{p}' \in \mathrm{Fix}(\mathcal{P}_{\boldsymbol{\rho}'})$. Choose an arbitrary vector satisfying*

[1] Note that this conclusion has been originally obtained in [HYLS15].

$\boldsymbol{\rho}'' \geq \boldsymbol{\rho}'$ and $\boldsymbol{\rho}' \neq \boldsymbol{\rho}''$, and define $\alpha_i = \rho_i''/\rho_i' \geq 1$ for $i \in \mathcal{M}$. Then the interference mapping $\mathcal{P}_{\boldsymbol{\rho}''} : \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ has a uniquely existing fixed point $\boldsymbol{p}'' \in \mathbb{R}_{++}^M$. Furthermore, we have $\mathbf{0} < \boldsymbol{p}'' < \boldsymbol{p} < \boldsymbol{p}'$ and $\rho_i'' P_i'' < \rho_i' P_i'$ for every $i \in \mathcal{M}$, where the ith element of the vector $\boldsymbol{p} = [P_1, \ldots, P_M]^T$ is given by $P_i := P_i'/\alpha_i$. Moreover, the sequence $\{\mathcal{P}_{\boldsymbol{\rho}''}^n(\boldsymbol{p})\}_{n\in\mathbb{N}}$, which converges to $\boldsymbol{p}''$, is monotonously decreasing.

Proof. By definition, $P_i \rho_i'' = P_i' \rho_i'$ for every $i \in \mathcal{M}$. As a result, by Lemma 1 and $\boldsymbol{p}' \in \mathrm{Fix}(\mathcal{P}_{\boldsymbol{\rho}'})$, we deduce

$$\mathcal{P}_{\boldsymbol{\rho}'',i}(\boldsymbol{p}) = \frac{P_i'}{\alpha_i \rho_i'} \sum_{j\in\mathcal{N}_i} \frac{r_j}{\alpha_i B_i \log_2\left(1 + \dfrac{P_i' g_{i,j}}{\alpha_i \left(\sum_{k\in\mathcal{M}\setminus\{i\}} \rho_k' P_k' g_{k,j} + \sigma^2\right)}\right)}$$
$$\leq \frac{P_i'}{\alpha_i \rho_i'} \sum_{j\in\mathcal{N}_i} \frac{r_j}{B_i \omega_{i,j}(\boldsymbol{\rho}', \boldsymbol{p}')} = \frac{1}{\alpha_i} \mathcal{P}_{\boldsymbol{\rho}',i}(\boldsymbol{p}') = \frac{P_i'}{\alpha_i} = P_i, \quad (4.16)$$

and the inequality is strict if and only if $i \in \mathcal{I} := \{k \in \mathcal{M} \mid \alpha_k > 1\} \neq \emptyset$. Therefore, $\mathcal{P}_{\boldsymbol{\rho}''}(\boldsymbol{p}) \leq \boldsymbol{p}$, which is already enough to show by Fact 7 that the fixed point $\boldsymbol{p}''$ of the mapping $\mathcal{P}_{\boldsymbol{\rho}''}$ exists, it is unique, and it satisfies $\boldsymbol{p}'' \leq \mathcal{P}_{\boldsymbol{\rho}''}^n(\boldsymbol{p}) \leq \boldsymbol{p}$ for every $n \in \mathbb{N}$. This last inequality and Fact 8 also show that the sequence $\{\mathcal{P}_{\boldsymbol{\rho}''}^n(\boldsymbol{p})\}_{n\in\mathbb{N}}$ is monotonously decreasing (and converges to $\boldsymbol{p}'' \in \mathrm{Fix}(\mathcal{P}_{\boldsymbol{\rho}''})$). From (4.16) and the assumption that $g_{i,j} > 0$ for every $i \in \mathcal{M}$ and $j \in \mathcal{M}$,[2] we observe that $\mathcal{P}_{\boldsymbol{\rho}'',i}(\boldsymbol{p})\rho_i'' < P_i \rho_i'' = P_i' \rho_i'$ for every $i \in \mathcal{I}$ (for $i \notin \mathcal{I}$, we have $\mathcal{P}_{\boldsymbol{\rho}'',i}(\boldsymbol{p}) = P_i$). We can now verify that $\mathcal{P}_{\boldsymbol{\rho}''}(\boldsymbol{p}) < \boldsymbol{p}$, which, by Fact 8, shows that $\boldsymbol{p}'' < \boldsymbol{p}$, and we conclude that $P_i'' \rho_i'' < P_i \rho_i'' = P_i' \rho_i'$ for every $i \in \mathcal{M}$. □

With the findings from Proposition 4 and based on [CSS+14, Remark 1] we derive a simple algorithm to compute new power assignments to increase the load of a given network configuration (as proved above, and also in [HYLS15], by doing so we decrease the transmit power).

We start with $\boldsymbol{p}' \in \mathrm{Fix}(\mathcal{P}_{\boldsymbol{\rho}'})$ and $\boldsymbol{\rho}' \in \mathrm{Fix}(\mathcal{J}_{\boldsymbol{p}'})$ as the power and load for the current network configuration, respectively and fix all other parameters to a constant value (e.g., the users' data rates). To compute a new power assignment $\boldsymbol{p}''$ inducing a load $\boldsymbol{\rho}'' \geq \boldsymbol{\rho}'$, we construct in parallel two sequences $\{\overline{\boldsymbol{p}}_n\}$ and $\{\underline{\boldsymbol{p}}_n\}$ where $\underline{\boldsymbol{p}}_0 := \mathbf{0}$, $\overline{\boldsymbol{p}}_0 := \boldsymbol{p}$, and $\boldsymbol{p}$ is the vector defined in Proposition 4. Fact 8 and Proposition 4 show that the sequences $\{\overline{\boldsymbol{p}}_n\}$ and $\{\underline{\boldsymbol{p}}_n\}$ are monotonously decreasing and increasing, respectively, and both sequences converge to $\boldsymbol{p}'' \in \mathrm{Fix}(\mathcal{P}_{\boldsymbol{\rho}''}) \neq \emptyset$. As a result, $\underline{\boldsymbol{p}}_n \leq \boldsymbol{p}'' \leq \overline{\boldsymbol{p}}_n$ for every $n \in \mathbb{N}$, and the monotonously decreasing sequence $\{\epsilon_n := \|\underline{\boldsymbol{p}}_n - \overline{\boldsymbol{p}}_n\|_\infty\}$ provides us with information about the numerical precision obtained at each iteration n because we have both $\|\underline{\boldsymbol{p}}_n - \boldsymbol{p}''\|_\infty \leq \epsilon_n$ and $\|\overline{\boldsymbol{p}}_n - \boldsymbol{p}''\|_\infty \leq \epsilon_n$. Based on these facts we propose Algorithm 5.

[2] If we replace this assumption by the weaker assumption that only the pathloss between users and their serving base stations are not zero, then the next strict inequalities should be replaced by their corresponding non-strict inequalities.

Algorithm 5 Transmit power assignment for desired load

Input: Current load $\boldsymbol{\rho}'$, current power assignment $\boldsymbol{p}'$, desired load $\boldsymbol{\rho}'' \geq \boldsymbol{\rho}'$, maximum number of iterations m, vector $\boldsymbol{p}$ defined in Proposition 4, and desired numerical precision $\epsilon > 0$ of the power assignment $\boldsymbol{p}''$ inducing the load $\boldsymbol{\rho}''$.

Output: Power assignment $\widetilde{\boldsymbol{p}}$ and numerical precision $\widetilde{\epsilon}$ satisfying $\|\widetilde{\boldsymbol{p}} - \boldsymbol{p}''\|_\infty \leq \widetilde{\epsilon}$.

1: Initialization: $\underline{\boldsymbol{p}}_0 \leftarrow \mathbf{0}$, $\overline{\boldsymbol{p}}_0 \leftarrow \boldsymbol{p}$, $n \leftarrow 0$, $\widetilde{\epsilon} = \|\boldsymbol{p}\|_\infty$.
2: **while** While $\widetilde{\epsilon} > \epsilon$ and $n \leq m$ **do**
3: Compute value of lower sequence $\underline{\boldsymbol{p}}_n \leftarrow \overline{\mathcal{P}}(\underline{\boldsymbol{p}}_{n-1})$
4: Compute value of upper sequence $\overline{\boldsymbol{p}}_n \leftarrow \overline{\mathcal{P}}(\overline{\boldsymbol{p}}_{n-1})$
5: Compute numerical precision $\widetilde{\epsilon} \leftarrow \|\underline{\boldsymbol{p}}_n - \overline{\boldsymbol{p}}_n\|_\infty$
6: Increment $n \leftarrow n + 1$
7: **end while**
8: Return the power assignment $\widetilde{\boldsymbol{p}} \leftarrow \overline{\boldsymbol{p}}_{n-1}$ and the numerical precision $\widetilde{\epsilon}$.

We note that, by Fact 8, the above algorithm terminates after a finite number of iteration even if we set $m = \infty$, in which case $\widetilde{\epsilon} \leq \epsilon$ upon termination.

4.7 Improved network topology control via load computation

In Chapter 3 we have introduced an optimization problem which addresses the task of selecting the set of network elements consuming the least amount of energy while satisfying the service demand from users. In this section we combine the algorithms developed to find solutions to this problem with our findings from the framework of standard interference functions. For convenience we recall the optimization problem for an LTE based communication system (c.f. Problem 3.21).

$$\text{minimize} \quad \sum_{l \in \mathcal{L}} \left(c_l \left| \boldsymbol{t}_l^T \tilde{\boldsymbol{x}} \right|_0 + \sum_{i \in \mathcal{S}_l} \left(e_i \left| \boldsymbol{s}_i^T \tilde{\boldsymbol{x}} \right|_0 + \tilde{f}_i(\tilde{\boldsymbol{x}}) \right) \right) \tag{4.17a}$$

$$\text{subject to} \quad \sum_{j \in \mathcal{N}} \frac{r_j}{B_i \, \omega_{i,j}(\boldsymbol{\rho})} \, x_{i,j} \leq 1 \qquad i \in \mathcal{M} \tag{4.17b}$$

$$\sum_{i \in \mathcal{M}} x_{i,j} = 1 \qquad j \in \mathcal{N} \tag{4.17c}$$

$$\mathbf{X} \in \{0,1\}^{M \times N} . \tag{4.17d}$$

The above problem has been shown to be NP-hard because of (4.17b) and (4.17c). It can not be solved in a straightforward and computationally tractable manner. We presented several simplifications and relaxations to be able to approach the problem with techniques based on majorization-minimization theory. A key step to do so is Assumption 12 which installs a conservative lower bound on the spectral efficiency in order to make (4.17b) tangible. However, this practice to lower bound the spectral efficiency can be very pessimistic leading to sub-optimal performance in terms of energy savings. With a too low estimate of the spectral efficiency users get allocated more resource blocks than required. An immediate

consequence of this is that the cell load is estimated higher than it actually is resulting in more active network elements.

Therefore, we seek to derive an algorithm that is able to produce better results based on more accurate values for the spectral efficiency. Fortunately, we have now a method at hand to compute the cell load and thereby the link spectral efficiency more accurate by the scheme outlined in Section 4.5. We now have the tools at hand to concurrently address two non-trivial challenges:

1. The assignment of test points to cells leading to a reduced energy consumption
2. Computation of the actual cell load for a given network setup.

In the following we present how to combine the results from Section 4.5 and the algorithm developed in Chapter 3. This forms our basis for the following two-step alternating iterative scheme and to find better solutions for Problem 4.17. Starting with the worst-case interference assumption $\boldsymbol{\rho}^{(-1)} = \mathbf{1}$:

Step 1 Compute the link spectral efficiency $\omega_{i,j}^{(n)}(\boldsymbol{\rho}^{(n-1)})$ at iteration n by (3.19) with the help of the cell load vector $\boldsymbol{\rho}^{(n-1)}$ obtained in the previous iteration. Use $\omega_{i,j}^{(n)}(\boldsymbol{\rho}^{(n-1)})$ in (4.17b) and derive test point cell assignment $\mathbf{X}^{(n)}$ for Problem 4.17 with the framework described in Section 3.3.3.

Step 2 Use the obtained test point cell assignment $\mathbf{X}^{(n)}$ and find the fixed point $\boldsymbol{\rho}^{(n)} \in \mathrm{Fix}(\underline{\mathcal{J}})$ for the standard interference mapping $\underline{\mathcal{J}}$ in (4.8).

To compute the fixed point in Step 2 we can use Algorithm 4. In words, the first step assigns test points to cells for a fixed link spectral efficiency with the goal of reducing the overall network energy consumption. The second step then computes the actual link spectral efficiency for the resulting assignment from step one. These two steps will be executed in an alternating manner to arrive at feasible network configurations consuming low total energy.

The complete alternating iterative scheme is summarized in Algorithm 6.

Algorithm 6 Energy minimization with the actual spectral efficiency of links

Input: Worst-case spectral efficiency $\boldsymbol{\omega}^{(-1)} = \boldsymbol{\omega}(\mathbf{1})$. Maximum number of iterations Z.
Output: Network configuration $\mathbf{X}^{(Z)}$ with low energy consumption.

1: **for** $n = 0 : Z$ **do**
2: Use $\boldsymbol{\omega}^{(n-1)}$ to construct Problem (3.6).
3: Use Algorithm 2 to obtain $\mathbf{X}^{(n)}$ and remove deactivated cells from the set of cells to be considered in subsequent iterations.
4: Compute the new link spectral efficiency $\boldsymbol{\omega}^{(n)}$ for the assignment $\mathbf{X}^{(n)}$ by computing the fixed point of the standard interference mapping $\mathcal{J}$.
5: **end for**
6: Return the network configuration resulting from $\mathbf{X}^{(Z)}$.

	Parameter	Value
Communication area	square shaped	2km × 2km
Number of test points	N	500
Number of cells	M	100
Data rate	r_j	16 kbit/s
Transmit power	P_i	46 dBm, $\forall i \in \mathcal{M}$
Bandwidth efficiency	$\eta_{i,j}^{\mathrm{BW}}$	1.5
SINR efficiency	$\eta_{i,j}^{\mathrm{SINR}}$	0.6
Signaling overhead	s_i	0.1
Carrier frequency	f_c	1800MHz
System bandwidth	B	5MHz
Upper bound of cell load	Γ_i	1, $\forall i \in \mathcal{M}$
Accuracy level	ϵ	10^{-5}
Lower starting point	$\underline{\boldsymbol{\rho}}^{(-1)}$	$\mathbf{0}$
Upper starting point	$\overline{\boldsymbol{\rho}}^{(-1)}$	$\boldsymbol{\Gamma}$ with $\Gamma_i = 10$
Number of alternating iterations	Z	15

Table 4.2: Standard simulation parameters for improving energy savings with the framework of interference calculus.

Remark 13. *The convergence of this Algorithm 6 follows from the convergence of Step 1 for a fixed load as outlined in Section 3.3.3 and the property of the algorithm that we only allow for deactivation of cells (we do not allow reactivation of cells deactivated in a previous iteration step). The latter property guarantees that the assignment from the previous iteration is always feasible in the current iteration. The complexity of the algorithm follow analogous to the explanations in Section 3.3.4.*

Numerical evaluation of network topology control with load computation (Algorithm 6)

Before we detail on the gains in energy consumption with the proposed alternating algorithm we present a numerical study about the inaccuracy of the cell load based on the worst case lower bound on the spectral efficiency and the more accurate value obtained by the framework of standard interference functions presented in Section 4.5. We reuse the simulation scenario from Section 4.5 but deploy a total number of $M = 100$ cells. This will give sufficient redundancies to identify cells for deactivation even with the worst-case interference assumption. The complete set of simulation parameters is presented in Table 4.2.

The results are produced by the following procedure:

1. Deploy M cells and N test points in the area.
2. Execute Algorithm 2 from Section 2 to generate a feasible test point assignment $\mathbf{X}^\star$ for Problem 4.17 and deactivate cells with no test point connected to it.
3. For $\mathbf{X}^\star$ compute the load vector $\boldsymbol{\rho}^\star$ by (4.4) using the worst case interference $\tilde{\omega}_{i,j} = \omega_{i,j}(\mathbf{1})$.

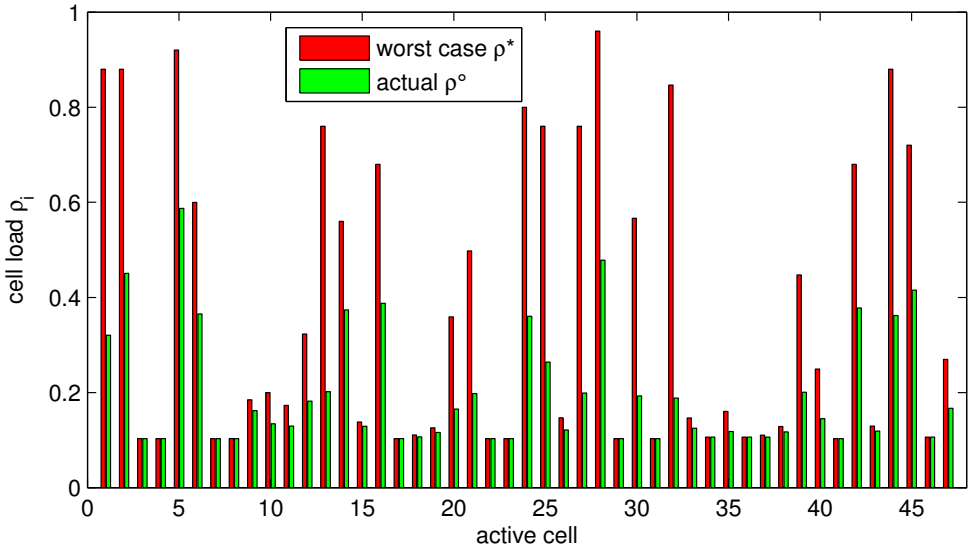


Figure 4.4: Load values $\boldsymbol{\rho}^{\star}$ at cells by considering the worst case spectral efficiency and by computing the actual load $\boldsymbol{\rho}^{\circ}$ with the fixed point iterations in Algorithm 4 (with precision $\epsilon = 10^{-5}$).

4. Apply Algorithm 4 to calculate the more accurate cell load $\boldsymbol{\rho}^{\circ}$.

For an exemplary simulation run the first execution of Step 2 of Algorithm 2 reduced the initial set of 100 cells to a total of 47 necessary to accommodate all test point demands. For this reduced set of active cells Figure 4.4 shows the load values $\boldsymbol{\rho}^{\star}$ obtained with the worst case interference assumption and the load values $\boldsymbol{\rho}^{\circ}$ computed by Algorithm 4. Note that we do only display the load value for active cells since the load at deactivated cells is zero. Although, the result is depicted only for one particular simulation setup, other simulations show a similar behavior. The cell load for the worst case interference scenario (indicated by the red bars) is always larger or equal to the more accurate values (indicated by the green bars). We can see that the gap in many cases is significant. Furthermore, in the case of worst case interference assumption there are several cells with large load values $\boldsymbol{\rho}^{\star}$ indicating that almost all resources are used. Therefore, these cells are not able to provide service to additional test points which in turn could lead to the deactivation of additional cells. This observation also supports the correct operation of Algorithm 2. However, when the actual load is considered it becomes evident that most cells are actually lightly loaded with many cells not even half loaded. This fact shows that there exists a great potential for further energy savings by reassigning test points and thereby freeing other cells. Another way to exploit this observation is that test points have significant room to increase their data rate requirements and at the same time the system is feasible (c.f. Section 4.5).

We now proceed with the empirical evaluation of the full Algorithm 6. The standard simulation setting described earlier in this section is used and we limit the total number of alternating iterations to $Z = 10$. To obtain results for different system loads we simulate

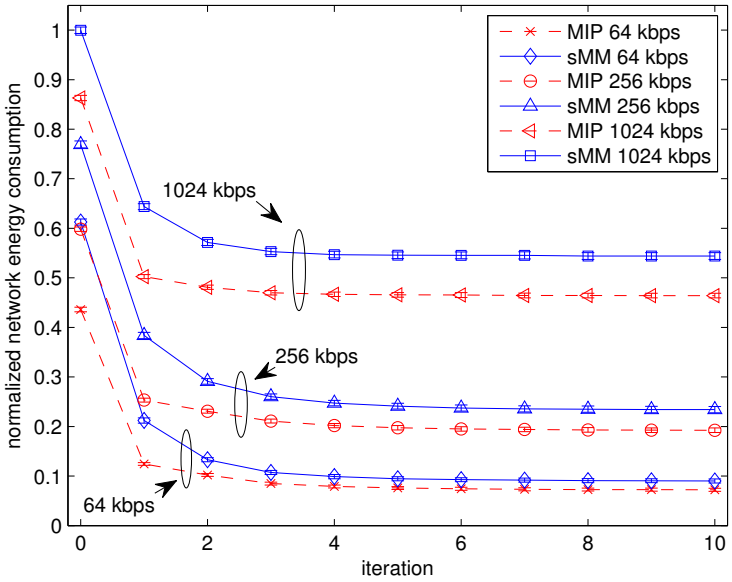


Figure 4.5: Alternating *sMM* algorithm. Normalized network energy consumption when applying Algorithm 6 with the sMM algorithm compared to the use of the MIP solution in each iteration. Normalization with respect to the energy consumption when all cells are active. Results are averaged over 100 different realizations of the network and the 95% confidence intervals are provided.

the system for different mean data rates μ_d at TPs and, for each mean data rate, we use 100 different realizations of the simulation scenario. Recall, that the starting point of Algorithm 6 uses the worst-case spectral efficiency $\boldsymbol{\omega}^{(-1)} = \boldsymbol{\omega}(\boldsymbol{\rho}^\star)$ with $\boldsymbol{\rho}^\star = \mathbf{1}$. Figure 4.5 depicts the resulting normalized network energy consumption together with the 95% confidence levels for the application of Algorithm 6 with the *sMM* and alternatively with the *MIP* algorithms from Chapter 3. The results have been normalized with respect to the energy consumption when all cells are active and fully loaded, i.e., $E_{\text{norm}} = \frac{E(\boldsymbol{\rho})}{E(\mathbf{1})}$. A significant increase in energy savings is observed when applying Algorithm 6 to the system for our *sMM* algorithm from Chapter 3 as well as for the optimum solution MIP. In all simulations the largest reduction in energy consumption is observed in the first iteration which corresponds to the first time use of the more accurate spectral efficiency after using the worst-case interference assumption. We can conclude that there are huge potentials for additional energy savings when taking into account a better estimate of the link spectral efficiency compared to the worst-case spectral efficiency. The worst-case assumption is very conservative and diminishes performance gains.

4.8 Joint base station configuration and network topology control

In this section we extend our analysis of identifying a network topology that saves the most amount of energy from the cell selection problem towards the consideration of different cell settings of active elements. Thereby, we try to answer the question of how to choose the settings of identified active cells in such an topology control optimization. The ultimate goal is to jointly optimize the topology of a network and the cell configuration such as antenna settings to attain configurations supporting the energy savings by topology control. Based on our standard system model we can write our standard optimization problem (4.17) as

$$\text{minimize} \quad \sum_{l\in\mathcal{L}} \left(c_l \left| \sum_{i\in\mathcal{S}_l} \rho_i \right|_0 + \sum_{i\in\mathcal{S}_l} e_i \left|\rho_i\right|_0 \right) \tag{4.18a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \omega'_{i,j}(\boldsymbol{\rho},\boldsymbol{\Theta})}\ x_{i,j} = \rho_i \qquad i\in\mathcal{M} \tag{4.18b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j} = 1 \qquad j\in\mathcal{N} \tag{4.18c}$$

$$\mathbf{X} \in \{0,1\}^{M\times N} \tag{4.18d}$$

$$\boldsymbol{\rho} \in [0,1]^M \tag{4.18e}$$

$$\boldsymbol{\Theta} \in \mathcal{X}, \tag{4.18f}$$

where we used relation (3.4) in the objective function (4.18a) and omitted the load dependent energy consumption $f_i(\rho_i)$ to avoid notational clutter[3]. $\mathcal{X}$ in (4.18f) is the antenna configuration of the whole network defined in Section 4.4. We define the set of feasible solutions as $D := \{(\mathbf{X},\boldsymbol{\rho},\boldsymbol{\Theta}) \in \mathbb{R}^{M\times N}\times\mathbb{R}^M\times\mathbb{R}^{K\times M} | (4.18b)-(4.18f)\}$.

Now recall, that standard interference functions are monotone functions (Definition 18) and thus we can assume the following, based on the definition of the spectral efficiency (4.2):

Assumption 17. *If* $\boldsymbol{\rho}_1,\boldsymbol{\rho}_2 \in [0,1]^M$ *satisfies* $\boldsymbol{\rho}_1 \le \boldsymbol{\rho}_2$ *then* $\omega'_{i,j}(\boldsymbol{\rho}_1,\boldsymbol{\Theta}) \ge \omega'_{i,j}(\boldsymbol{\rho}_2,\boldsymbol{\Theta})$.

The practical interpretation of Assumption 17 is that a decrease in the load of neighboring cells cannot reduce the spectral efficiency of a link. This assumption has been implicitly used before to calculate the worst-case spectral efficiency $\tilde{\omega}'_{i,j}(\boldsymbol{\Theta}) := \omega'_{i,j}(\mathbf{1},\boldsymbol{\Theta})$ and with fixed $\boldsymbol{\Theta}$, $\tilde{\omega}_{i,j} := \omega_{i,j}(\mathbf{1})$.

Problem 4.18 is difficult to solve since we have highly non-convex constraints, i.e. (4.18b). Therefore, to obtain good solutions to the problem we devise a greedy two-step approach. The first step uses fixed antenna configuration and load values to find an assignment matrix $\mathbf{X}$ resulting in a low network energy consumption. The second step fixes the assignment to

[3]The load dependent energy consumption can be straightforwardly included.

$\mathbf{X}$ found in the previous step and chooses an antenna configuration that is likely to help decrease the objective function of the optimization problem with fixed antenna configuration in the next iteration. The procedure is repeated until no more sites can be eliminated or the maximum number of iterations is reached.

In detail, we define a cost function $z_D(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\Theta}) : \mathbb{R}^{M\times N} \times \mathbb{R}^M \times \mathbb{R}^{K\times M} \to \mathbb{R}_+$ as

$$z_D(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\Theta}) = \sum_{l\in\mathcal{L}} \left(c_l \left| \sum_{i\in\mathcal{S}_l} \rho_i \right|_0 + \sum_{i\in\mathcal{S}_l} e_i \left|\rho_i\right|_0 \right) + \delta_D(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\Theta}), \tag{4.19}$$

where we use the indicator function

$$\delta_D(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\Theta}) := \begin{cases} 0, & \text{if } (\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\Theta}) \in D \\ \infty, & \text{otherwise.} \end{cases} \tag{4.20}$$

With (4.19) and (4.20), problem (4.18) amounts to finding a point in the set

$$\underset{(\mathbf{X},\boldsymbol{\rho},\boldsymbol{\Theta})\in\mathbb{R}^{M\times N}\times\mathbb{R}^M\times\mathbb{R}^{K\times M}}{\arg\inf} z_D(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\Theta}). \tag{4.21}$$

To find such a point we propose a scheme that produces a sequence $\{(\mathbf{X}^{(n)}, \boldsymbol{\rho}^{(n)}, \boldsymbol{\Theta}^{(n)})\}_{n\in\mathbb{N}} \subset \mathbb{R}^{M\times N} \times \mathbb{R}^M \times \mathbb{R}^{K\times M}$ monotonically decreasing (4.19), i.e.,

$$z_D(\mathbf{X}^{(n+1)}, \boldsymbol{\rho}^{(n+1)}, \boldsymbol{\Theta}^{(n+1)}) \leq z_D(\mathbf{X}^{(n)}, \boldsymbol{\rho}^{(n)}, \boldsymbol{\Theta}^{(n)}). \tag{4.22}$$

The main idea of our scheme is summarized in Algorithm 7 and we proceed by detailing on each step of the algorithm. In order to find $\mathbf{X}^{(n)}$ and $\boldsymbol{\rho}'^{(n)}$ in step 1 we can use any

Algorithm 7 Joint site selection and configuration for energy saving

Input: Tuple $(\mathbf{X}^{(-1)}, \boldsymbol{\rho}^{(-1)}, \boldsymbol{\Theta}^{(-1)})$ satisfying $z_D(\mathbf{X}^{(-1)}, \boldsymbol{\rho}^{(-1)}, \boldsymbol{\Theta}^{(-1)}) < \infty$. Number of iterations L_o

1: **for** $n = 0 : L_\mathrm{o}$ **do**

2: **(Step 1)** Set $\mathbf{X}^{(n)}$ and $\boldsymbol{\rho}'^{(n)}$ to any value satisfying $z_D(\mathbf{X}^{(n)}, \boldsymbol{\rho}'^{(n)}, \boldsymbol{\Theta}^{(n-1)}) \leq z_D(\mathbf{X}^{(n-1)}, \boldsymbol{\rho}^{(-1)}, \boldsymbol{\Theta}^{(n-1)})$.

3: **(Step 2)** Set $\boldsymbol{\Theta}^{(n)}$ and $\boldsymbol{\rho}^{(n)}$ to any value satisfying $z_D(\mathbf{X}^{(n)}, \boldsymbol{\rho}^{(n)}, \boldsymbol{\Theta}^{(n)}) \leq z_D(\mathbf{X}^{(n)}, \boldsymbol{\rho}^{(n)}, \boldsymbol{\Theta}^{(n-1)})$ and that are likely to help Step 1 to make progress in the next iteration.

4: (If progress cannot be made for too many iterations, exit by returning the current feasible configuration)

5: **end for**

6: **return** $(\mathbf{X}^{(L_\mathrm{o})}, \boldsymbol{\rho}^{(L_\mathrm{o})}, \boldsymbol{\Theta}^{(L_\mathrm{o})})$

suitable heuristic. One possibility is to use the heuristic described in Section 4.7 that gives good solutions in reasonable time. Step 2 consists of finding an antenna configuration $\boldsymbol{\Theta}^{(n)}$ supporting (4.22) which is a challenging task because only changing antenna configurations at cells (without reassigning test points) is not sufficient enough to free cells for deactivation.

However, the antenna configuration can be optimized in order to help Algorithm 7 make progress in Step 1.

Examining the objective function (4.18a) of Problem 4.18 we can identify that good solutions attaining low energy consumption are characterized by a sparse cell load vector $\boldsymbol{\rho} \in [0,1]^M$. In (4.18a) the sparsity is perused by the sparsity promoting l_0 operator $|\cdot|_0$. Unfortunately, such highly non-convex functions are difficult to handle but we have presented a way in Chapter 3 how to approach this problem. The basic approach is to replace the non-convex function by its convex envelope, i.e., to replace $|x|_0$ on $\{x \in \mathbb{R} \mid |x|_\infty \leq C\}$ by the function $f(x) = |x|_1/C$. Now, recall that $\boldsymbol{\rho} \in [0,1]$ and we can formulate the convex envelope of the objective function (4.18a) as

$$\sum_{l \in \mathcal{L}} \left(\frac{c_l}{|\mathcal{S}_l|} \sum_{i \in \mathcal{S}_l} \rho_i + \sum_{i \in \mathcal{S}_l} e_i \ \rho_i \right) = \sum_{l \in \mathcal{L}} \sum_{i \in \mathcal{S}_l} \left(\frac{c_l}{|\mathcal{S}_l|} + e_i \right) \rho_i. \tag{4.23}$$

Note that the cell load in (4.23) is a function of the antenna configuration $\boldsymbol{\Theta}$ (c.f. (4.18b)). Thus, we intuitively obtain a method to be applied in Step 2 of Algorithm 7 by choosing an antenna configuration $\boldsymbol{\Theta}$ that decreases (4.23). By this approach we try to enforce sparsity in the cell load vector $\boldsymbol{\rho}$.

The techniques to optimize (4.23) depend on the adopted site antenna configuration model and we present in Algorithm 8 a simple method to be applied for antenna configuration models with a discrete set of possible configurations $\mathcal{X}_i$, $i \in \mathcal{M}$. We present the algorithm for site configuration to be applied in Step 2 of Algorithm 7 in Algorithm 8. The difficulty of

Algorithm 8 Site configuration for Step 2 of Algorithm 7

Input: Configuration $(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\theta}_1^{(-1)}, \ldots, \boldsymbol{\theta}_M^{(-1)})$ satisfying $\delta_D(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\theta}_1^{(-1)}, \ldots, \boldsymbol{\theta}_M^{(-1)}) < \infty$. Number of iterations L_{in}. (For notational convenience, the configuration of a site $i \in \mathcal{M}$ at iteration n is denoted by $\boldsymbol{\theta}_i^{(n)}$.)

1: **for** $n = 0 : L_{\text{in}}$ **do**
2: **for** $i = 1 : M$ **do**
3: Update $\boldsymbol{\theta}_i^{(n)}$ according to

$$(\boldsymbol{\theta}_i^{(n)}, \boldsymbol{\rho}) \in \underset{\boldsymbol{\theta} \in \mathcal{X}_i,\ \boldsymbol{\rho} \in \mathbb{R}_+^M}{argmin} z_D(\mathbf{X}, \boldsymbol{\rho}, \boldsymbol{\theta}_1^{(n)}, \ldots, \boldsymbol{\theta}_{i-1}^{(n)}, \boldsymbol{\theta}, \boldsymbol{\theta}_{i+1}^{(n-1)}, \ldots, \boldsymbol{\theta}_M^{(n-1)}) \tag{4.24}$$

4: **end for**
5: **end for**
6: **return** $\boldsymbol{\theta}_1^{(L_{\text{in}})}, \ldots, \boldsymbol{\theta}_M^{(L_{\text{in}})}$

Algorithm 8 lies in solving the optimization Problem 4.24. Since it is of combinatorial nature, to find the solution might involve exhaustive search. However, if the cardinality of the sets $\mathcal{X}_i$ ($i \in \mathcal{M}$) is small, the application of a grid search is practically. Furthermore, the value of the indicator function δ_D can be computed by evaluating the current antenna configuration for its feasibility by means of feasibility analysis of fixed network configurations as shown in Section 4.5.

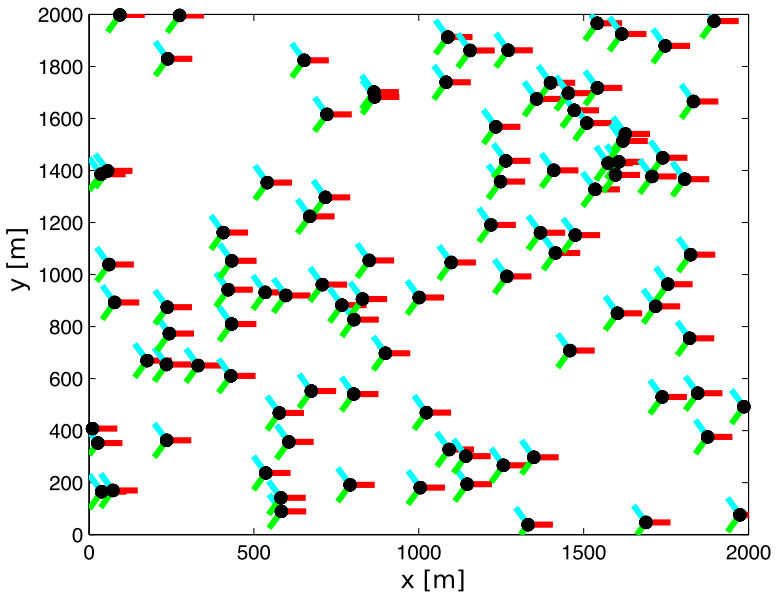


Figure 4.6: Network configuration. The network consists of a single layer LTE network at 1800MHz with 300 cells at 100 different sites. Sites are depicted as red dots and the azimuth direction of each cell is shown by a colored line departing at the corresponding site.

Numerical evaluation of joint base station configuration and network topology control (Algorithm 7)

We now evaluate the energy savings gains achieved by applying an antenna tilt optimization phase to the load aware topology control mechanism presented in Section 4.7. This corresponds to finding the optimal downtilt at each cell, i.e., $\theta_i \in \mathcal{X}_i \subset \mathbb{R}$. For this purpose we use our standard simulation setup with a total number of 100 three-sectored LTE base stations distributed in a uniformly at random manner. Each base station site has three co-located cells with different azimuth direction $(0°, 120°, 240°)$. For simplicity, we assume that all cells consume the same static energy $e_i = 400W$ while active and the additional static base station energy consumption is zero, i.e. $c_l = 0$. An exemplary layout is depicted in Figure 4.6, where the main azimuth direction of a sector is indicated by a line originating from the cell. The user distribution is according to our three hotspot model. In order to incorporate the effect of different antenna settings we use the standard 3rd Generation Partnership Project (3GPP) spatial channel model for a typical urban scenario with 3D antenna patterns [3GP10b]. The search space for downtilts at cells is $\theta_i \in [0, 20]$ and we limit Algorithm 7 to a total number of $L_{\text{o}} = 10$ alternating iterations and $L_{\text{in}} = 1$ inner iterations for Algorithm 8. The remaining simulation parameters are according to our standard simulation settings for energy savings with the framework of interference calculus summarized in Table 4.3.

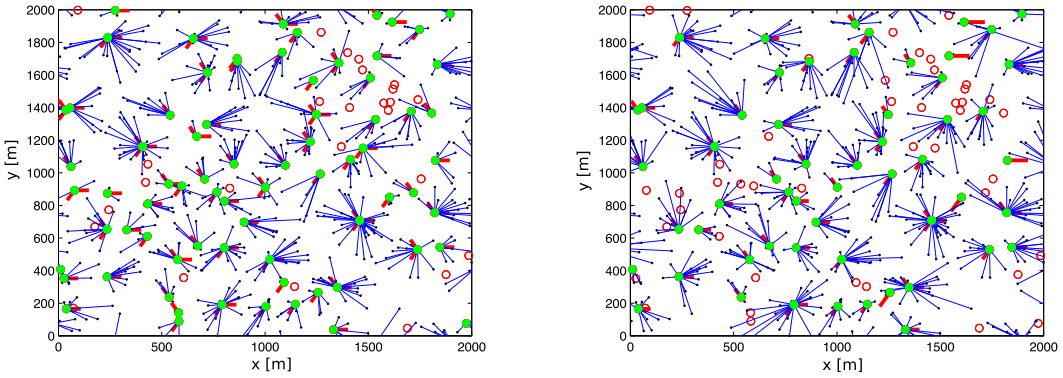
The results for the exemplary deployment are depicted in Figure 4.6 and the results obtained by applying Algorithm 7 to it are provided in Figure 4.7. Figure 4.7(a) shows the

	Parameter	Value
Communication area	square shaped	2km × 2km
Number of test points	N	500
Number of cells	M	100
Data rate	r_j	16 kbit/s
Transmit power	P_i	46 dBm, $\forall i \in \mathcal{M}$
Bandwidth efficiency	$\eta_{i,j}^{\mathrm{BW}}$	1.5
SINR efficiency	$\eta_{i,j}^{\mathrm{SINR}}$	0.6
Signaling overhead	s_i	0.1
Carrier frequency	f_c	1800MHz
System bandwidth	B	5MHz
Upper bound of cell load	Γ_i	1, $\forall i \in \mathcal{M}$
Accuracy level	ϵ	10^{-5}
Lower starting point	$\underline{\boldsymbol{\rho}}^{(-1)}$	$\mathbf{0}$
Upper starting point	$\overline{\boldsymbol{\rho}}^{(-1)}$	$\boldsymbol{\Gamma}$ with $\Gamma_i = 10$
Number of alternating iterations	Z	15

Table 4.3: Standard simulation parameters for the evaluation of Algorithm 7.

final test point assignment after the first time Step 1 of Algorithm 7 was executed. At this point the tilt has not been optimized yet, which is illustrated by the red bars having all the same length. The downtilt of all cells is still set to its initial value of $\boldsymbol{\theta}_i = 15, i \in \mathcal{M}$. In Figure 4.7(b) the final test point assignment and tilt settings are depicted after $L_{\mathrm{o}} = 10$ alternating steps. After executing Step 1 of Algorithm 7 the first time ($L_{\mathrm{o}} = 0$) the final network configuration results in 100 active cells at 74 different locations. Step 1 was terminated because solely by reassigning test points it is not possible to deactivate more cells. Thus Step 2 aims at helping the algorithm in Step 1 to further reduce this number of active cells. For this purpose Algorithm 8 is used in Step 2 of Algorithm 7 in order to find a good set of tilts resulting in a feasible setup that opens opportunities to deactivate more cells. Remember that for each antenna reconfiguration θ_i we recompute the channel gain $g_{i,j}(\theta_i)$ to all test points $j \in \mathcal{N}$ with the help of the 3GPP spatial channel model.

Comparing Figure 4.7(a) and Figure 4.7(b) it can be observed that Step 2 helps Step 1 to further reduce the number of active cells. After $L_{\mathrm{o}=10}$ alternating executions of Step 1 and Step 2 of Algorithm 7 the number of active cells is reduced to 65 at 56 different locations and the tilts have been adjusted as indicated by the length of the red lines in Figure 4.7(b). The effect of the tilt optimization opened opportunities in Step 1 to reassign test points to other cells and free resources leading to cell deactivation. In consequence, the algorithm is able to steer the downtilt in the direction of the test points that are served by the cell. For cells with many connected test points far away the downtilt is low and where test points are close the downtilt is high. Note, that in some cases a cell serves a test point very close to its own location and the resulting tilt is very low. This is due to the used 3GPP spatial channel model which caps the maximum loss for the antenna tilt mismatch resulting in all



(a) After Step 1 of the first alternating iteration L_o. The tilts have not been optimized ($\boldsymbol{\theta}_i = 15, i \in \mathcal{M}$). The configuration consists of 100 active cells at 74 different locations.

(b) Final tilt settings and test point assignment ($L = 10$). The final configuration consists of 65 active cells at 56 different locations.

Figure 4.7: Network configuration when applying Algorithm 7. One inactive site is depicted by a red circle. The final antenna tilt of a cell is indicated by the length of its red line. Long lines stand for tilts close to 0 degree and short lines indicate a high downtilt (in the order of 15-20 degrees).

tilts $\theta_i \in [0, 20]$ used in our simulations being equally bad. The choice of a low downtilt results in the least interference to other users in the system.

To characterize the additional energy savings gains achievable with Algorithm 7 a number of simulations with different random network deployments have been performed. The simulation scenario outlined above is used and 50 different cell and test point deployments were generated. Figure 4.8 shows the average normalized network energy consumption and the corresponding 95% confidence intervals for each alternating iteration $L_o = 0 \ldots 10$.

The results for three different per test point data rate requirements are provided. It can be observed that significant energy savings in all three cases are achievable. The savings are largest in the first few alternating iterations and gains diminish with increasing number of iterations L_o. In the case of data rate requirements 1024Kb/s the average normalized network energy consumption can be reduced from 0.396 to 0.300 and for data rate requirements of 512Kb/s (256Kb/s) from 0.273 (0.217) down to 0.150 (0.100). Note that the confidence interval for 1024Kb/s case is larger since it can be observed that in such cases the general cell load is higher and the antenna configuration of Step 2 is not able to succeed in helping Step 1 of the algorithm in all instances. In other words the optimization of downtilts did not improve the interference scenario sufficiently much such that additional cells can be deactivated.

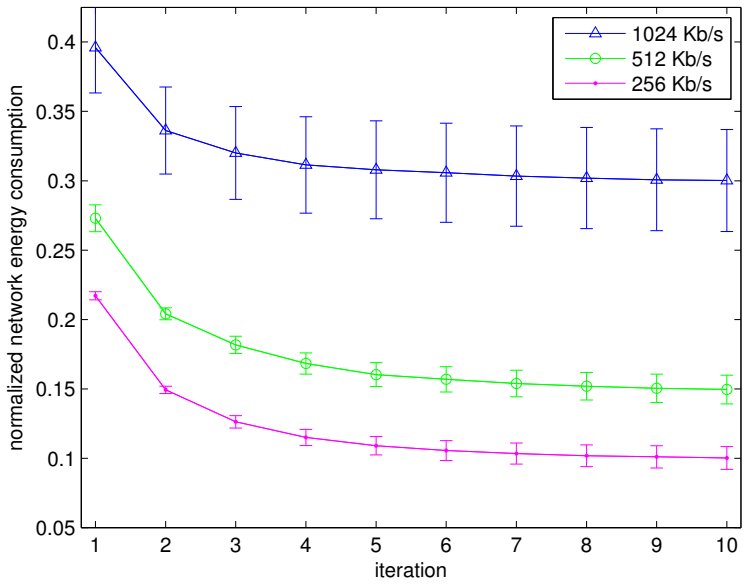


Figure 4.8: Normalized network energy consumption when applying Algorithm 8 to a network with 100 three-sectored LTE cells and 500 test points. Normalization with respect to the energy consumption when all cells are active. Results are averaged over 30 different realizations of the network and the 95% confidence intervals are provided.

4.9 Conclusion

Starting from available results in the field of standard interference functions we exploited that the computation of load in cellular communication systems can be posed as the problem of finding fixed points of standard interference mappings. Thereby, we leveraged the existing body of work to compute fixed points of standard interference mappings with the general framework of interference calculus which gave us computationally inexpensive techniques to compute more accurate network loads. We enhanced our models used in the topology control for energy savings framework presented in Chapter 3 to more realistically model the interference coupling between cells. These models use interference mappings that are able to capture many practical limitations, such as an upper bound for the spectral efficiency or a constant signaling overhead. With the help of these interference mappings and the derived iterative algorithms we showed how the framework can be used to easily crosscheck if an arbitrary network configuration is feasible in the sense that it is able to provide all users with the desired QoS. Being based on the general framework of interference calculus, the involved computations are of low complexity due to the involved fixed point algorithms that are used. We presented a non-heuristic stopping criterion to compute fixed points of our interference mappings that lead to an arbitrary but fixed precision.

Building upon the results obtained from applying interference calculus to the problem of finding the interference coupling and cell load in a cellular communication system,

we derived a standard interference mapping in closed form that has as its fixed point the power allocation inducing a given load in LTE-like systems. We showed that well-known techniques to compute fixed points become readily available, and these iterative techniques are remarkably simpler than previous methods that, for example, require nested iterative approaches. In particular, the proposed iterative techniques is able to give accurate information about the precision of the power assignment vector obtained at each iteration.

Having at hand the tools for power allocation and cell load computation, we derived a low-complexity algorithm that makes use of the general framework of standard interference functions for the task of identifying network configurations consuming low energy. In particular, we extended the algorithms for energy savings from Chapter 3 in the direction of load computation which opens more opportunities to reduce the energy in cellular networks. Unlike the approach in Chapter 3 where the worst case interference assumption was used in combination with relaxations and majorization-minimization techniques, we incorporated the actual cell load for the calculation of the interference coupling. The actual cell load is computed by fixed point algorithms that guarantee convergence asymptotically. This new enhanced topology control framework for energy savings has the ability to exploit the full energy savings potential of a cellular communication system in low load scenarios and incorporates computationally inexpensive algorithms that retain the frameworks amenability for online implementation.

Lastly, we turned towards finding configurations of active network elements that support the developed topology control framework. With the help of heuristic arguments we combined the improved framework for topology control with the optimization of configurations of active cells. An algorithm was developed that is able to identify different cell configurations that help to save even more energy. In particular, we showed how the optimization of down tilts can help to save energy. As a result, these algorithms are substantially better and highly relevant for practical systems.

5 Conclusions

In this thesis, we proposed solutions targeting at energy savings in wireless communication systems. In particular we focused on the problem of network topology control that is concerned with the energy consumption of all active network elements. Thereby, we tried to give answers to the scaling of energy consumption in cellular communication systems and proposed solutions how to minimize the energy consumption of the network by disengaging certain network elements. In the first chapter of this thesis, we gave an insight in the growth behavior of wireless communication systems when the number of users tends to infinity. We started from a scaling law analysis of throughput as a function of the number of users having some QoS requirement and extended the results into the direction of embodied network energy consumption. In our analysis we specifically included the energy consumed by hardware, for, e.g., cooling or processing, which is typically neglected in theoretical studies from information theory. We derived three scaling regimes for *hybrid wireless networks* where ad hoc communication is supported by infrastructure nodes. The identified scaling regimes are further investigated by simulations that allow to verify the scaling results predicted by theory and to assess the speed of convergence of the asymptotic results. Our analysis and simulation results point out the fact that the results of vanishing throughput in random hybrid wireless communication systems for the asymptotic regime are already observable for a finite number of users. Thus the problem of diminishing per node throughput is a problem to be addressed even for smaller number of nodes. The second main result from the first chapter is that the energy-per-bit of random hybrid networks with n wireless nodes and $m(n)$ infrastructure nodes operating under a practical communication scheme as outlined in [ZdV05] satisfies an upper bound. The energy consumption and energy per bit results are not of general applicability, but we conjecture that they provide scaling laws for specific communication schemes currently used in cellular networks such as GSM, UMTS, LTE and future 5G. From the results of the first chapter we concluded that: 1) for a practical communication scheme the infrastructure nodes need to scale accordingly to have the best energy per bit scaling; 2) Under our considered communication scheme, a too large infrastructure scaling is not beneficial in terms of energy per bit; 3) More advanced radio access technologies are needed to fully exploit the potential of cellular networks with very dense infrastructure and to achieve a better scaling tradeoff between throughput and energy efficiency. These three conclusions serve as motivation for the work presented in the proceeding chapters of this thesis.

The main objective of the remaining chapters was the minimization of the overall energy consumption in the downlink channel of mobile cellular networks including the energy

consumed by hardware and auxiliary equipment. We followed an approach that tries to select the set of network elements consuming the least amount of energy while satisfying the QoS requirements of all users in the system. The underlying problem is of combinatorial nature and we formulated an optimization problem to find feasible network configurations that show these properties. Our optimization framework takes into account the static energy consumed by hardware as well as the load-dependent energy spent on transmission in an optimal manner and thus, is able to balance between the different sources of energy consumption. Furthermore, we modeled the technology specific capabilities and limitations to provide the desired QoS to the users as constraints to our optimization problem. Thereby, we were able to tailor the optimization problem to address networks of multiple radio access technologies. As an example we included the problem definition of an LTE single-RAT network and an UMTS/LTE multi-RAT network. Since the developed optimization framework includes a combinatorial problem it is in general hard to solve. In fact, we showed that the problem is related to the standard bin-packing problem, which is known to be NP-hard. In order to obtain solutions in a computationally efficient manner we applied relaxation techniques to approximate the objective function by a concave function and relaxed all non-convex constraints which led to a problem that is amenable for the majorization-minimization (MM) technique. With these steps we were able to propose an optimization framework that can find good solutions to the combinatorial problem in relatively short time. A major advantage is its ability to cope with a variety of network elements. It is easy to account for different energy consumption models of different hardware, e.g., base stations of different generations, sectors or even antennas as long as they can be modeled as any concave or convex function of the load. We provided extensive simulation results to show the framework's performance. Finally, we included several extensions to account for novel aspects of networks such as CoMP transmission, the desire for a distributed implementation or to take into account buffering capabilities at user side. The latter led to an extension into the direction of anticipatory scheduling, where the decisions on resource allocation and user-cell assignments are based not only on present channel state information but also on information about future propagation conditions. Thereby, we were able to exploit the knowledge about users' mobility and path loss to proactively build user-cell assignment and resource allocation schedules that greatly support energy savings in cellular communication systems. We showed by simulations that in such a scenario the energy savings can be increased if the requested data of users can be preloaded in its buffer.

In the last chapter we improved a drawback in the modeling of the technology specific constraints from previous chapters, namely the worst-case interference assumption which was necessary to arrive at convex set of constraints. This resulted in conservative network configurations that were feasible solutions satisfying all user constraints but missed out to exploit the full energy saving potential. For the purpose of having a more accurate interference estimation in the network that could potentially help our optimization framework to save more energy, we turned to the general framework of interference calculus.

Starting from available results for standard interference functions we used the result that the computation of load in cellular communication systems considered in our study can be posed as the problem of finding fixed points of standard interference mappings. This provided us with readily available, computational efficient techniques (standard iterative algorithms) to find the load at cells for a given network setup which in turn led to a more accurate estimation of the interference in the network. Based on this result, in the first step we developed a tool to classify the feasibility of network settings for a given user-cell assignment. The results obtained from applying interference calculus to the problem of finding the interference coupling and cell load in a cellular communication system, allowed us to derive a standard interference mapping in closed form that has as its fixed point the power allocation inducing a given load in LTE-like systems. These tool were then combined with the network topology control optimization framework in order to increase the energy savings. Finally, we presented a framework which allows to incorporate finding feasible cell configurations (e.g., antenna tilts) supporting the energy savings algorithm that deactivate redundant network elements. In other words, with the help of heuristic arguments we combine the improved algorithm for topology control with the optimization of cell configurations. In particular, we show how the optimization of down tilts can help to save energy.

A Comments on Sparsity

The notion of sparsity has been used differently in literature and thus we comment on one interpretation that is used in this study. We denote a vector or matrix as sparse when most of its elements are equal to zero and only very few are non-zero. A common mean of describing the sparsity of a vector $\boldsymbol{x} \in \mathbb{R}$ is to say that $\boldsymbol{x}$ is k-sparse if only k elements of $\boldsymbol{x}$ are non-zero. To count the non-zero elements of a vector we have introduced the $|\cdot|_0$-operator and have used it throughout our study in several sparsity promoting optimization problems.

In a wider sense, the question of sparsity naturally arises in the context of find structured solutions to under-determined systems of linear equations. Out of this research question the fields of matrix completion and compressive sensing have evolved. These fields provide a mathematically rich theory for the described task using convex optimization techniques. Applications are very broad and can be found for example in medical imaging [CRT06, LDP07], digital photography [DDT+08], image processing [CW08], sensor networks [BHSN06], and many others.

In Chapter 3 we have introduced a problem that needs to find sparse solutions satisfying a set of constraints. In its most general from this problem can be stated as

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \;\; |\boldsymbol{x}|_0 \tag{A.1a}$$

$$\text{subject to } \mathbf{A}\boldsymbol{x} = \boldsymbol{y}\,, \tag{A.1b}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a constraint matrix and $\boldsymbol{y} \in \mathbb{R}^m$ is the limit vector for the constraints. However, this problem is generally difficult to solve due to the non-convex $|\cdot|_0$-operator and its discontinuity. For small problem instances an exhaustive combinatorial search will find a solution but for intermediate and large problems it becomes intractable.

A typical way to address this problem is to relax the problem to arrive at a convex optimization problem which can be solved efficiently with standard numerical solvers. A representative of this class is used with Basis pursuit [CDS98], where the cost function is replaced by the l_1-norm. Mathematically speaking, the problem is recast to

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \;\; |\boldsymbol{x}|_1 \tag{A.2a}$$

$$\text{subject to } \mathbf{A}\boldsymbol{x} = \boldsymbol{y}\,. \tag{A.2b}$$

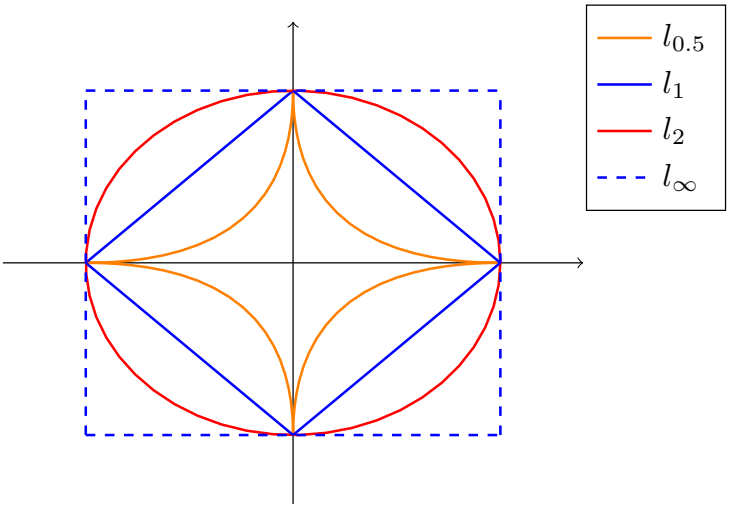


Figure A.1: The unit ball in different normed vector spaces.

It has been shown [CT06] that under certain conditions Problem A.1 and Problem A.2 are equivalent. The relaxation used to arrive at Problem A.2 uses the convex envelope of the $|\boldsymbol{x}|_0$ for $x \in [0, 1]$. A natural questions arises, whether other relaxations giving convex optimization problems can be used to achieve the same results with improved performance. We comment on this question in Section A.1 and present selected sparsity measure concluding this Chapter with comments on the measure of sparsity used in our study in Section A.2.

A.1 Measures of sparsity and $|\cdot|_0$-operator approximations

Even though literature provides a huge variety of relaxation functions and approximations for the $|\cdot|_0$-operator, the l_1-norm has become the most used relaxation for the $|\cdot|_0$-operator. It has good sparsity promoting properties while being a convex function which makes problems involving the l_1-norm amenable for recast to linear problems. However, it also has certain drawbacks; one being the lack of being able to produce a sparse solution when the magnitude of each element of the solution vector is very small. The reason is also the key difference between the $|\cdot|_0$-operator and the l_1-norm, namely the dependence of the magnitude of the latter norm. The $|\cdot|_0$-operator does not penalize large non-zero entries which the l_1-norm does. Figure A.1 illustrates the unit ball of different normed vector spaces from which the influence of the argument's magnitude of a l_p-norm can be observed.

In the following we comment on some continuous functions that are used to approximate the $|\cdot|_0$-operator. The $|\cdot|_0$-operator for a vector $\boldsymbol{x} \in \mathbb{R}^n$ can be written as

$$|\boldsymbol{x}|_0 = \sum_{i=1}^{n} \delta(x_i), \tag{A.3}$$

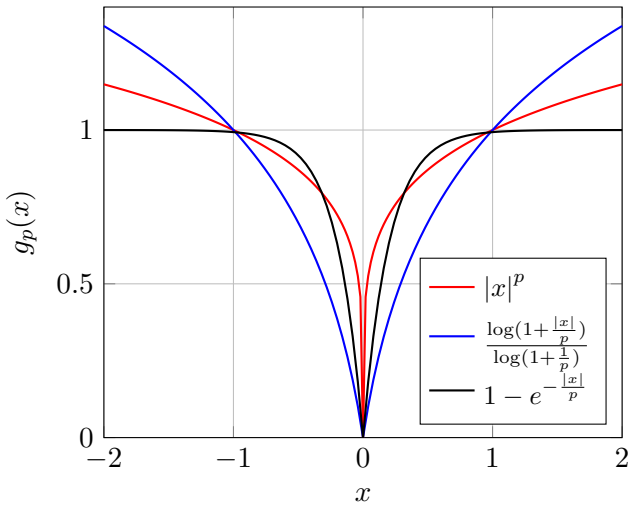


Figure A.2: Three surrogate functions $g_{1,p}(x)$, $g_{2,p}(x)$ and $g_{3,p}(x) =$ that are used to approximate $|x|_0$, $p = 0.2$ [SBP15].

where $\delta : \mathbb{R} \to \{0, 1\}$ is the indicator function

$$\delta(x) := \begin{cases} 1 & \text{if } x \in \mathbb{R}\backslash\{0\} \\ 0 & \text{if } x \equiv 0. \end{cases} \tag{A.4}$$

A way to incorporate other approximations for the $|\cdot|_0$-operator is to replace the indicator function $\delta(x)$ by some continuous function $g_p : \mathbb{R} \to \mathbb{R}$ with desired sparsity promoting property. The parameter p is used to control the quality of the approximation. Reference [SBP15] analyzes the following three surrogate functions:

1. $g_{1,p}(x) = |x|^p, 0 < p \leq 1,$
2. $g_{2,p}(x) = \frac{\log(1+\frac{|x|}{p})}{\log(1+\frac{1}{p})}, p > 0,$
3. $g_{3,p}(x) = 1 - e^{-\frac{|x|}{p}}, p > 0.$

The first one is commonly used in compressed sensing problems [GR97, CY08] and we refer to it as the p-norm for $p < 1$. The second one was proposed in [STL11] for the sparse generalized eigenvalue problem and the third one appears in [Man96] in the context of feature selection problems. The different surrogate functions are plotted in Figure A.2 for fixed $p = 0.2$. Each surrogate function has its own strengths and downsides. The first surrogate function $g_{1,p}(x) = |x|^p, 0 < p \leq 1$ has favorable properties for $x << 1$ since it shows a fast decay around $x = 0$. However, its mathematical properties make it less favorable for cases where linear relaxations are pursued. The function $g_{3,p}(x) = 1 - e^{-\frac{|x|}{p}}, p > 0$ has good properties for $x > 1$ since it does not give high weights to large arguments. Function $g_{2,p}(x) = \frac{\log(1+\frac{|x|}{p})}{\log(1+\frac{1}{p})}, p > 0$ is a good compromise due to its acceptable weights for larger arguments and its decaying property for $x \to 0$.

A.2 Notes on the $|\cdot|_0$-operator reformulation

Throughout our study we have used a particular reformulation and relaxation of the $|\cdot|_0$-operator. In particular, in Section 3.3.3 we have used the following relation:

$$|\mathbf{h}|_0 = \sum_{i=1}^{M} |h_i|_0 = \sum_{i=1}^{M} \lim_{\epsilon \to 0} \frac{\log(1 + |h_i|\epsilon^{-1})}{\log(1 + \epsilon^{-1})}. \tag{A.5}$$

In the following we show the equivalence of the left-hand side and the right-hand side of (A.5). To show this equivalence, we make use of the rule of l'Hopital which says the following. Assume the two functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are differentiable on an open interval with endpoint x_0 and $g'(x) \neq 0$ on the interval. Then, if $\lim_{x \to x_0} \frac{f'(x)}{g'(x)} = L$ exists, then

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{f'(x)}{g'(x)} = L. \tag{A.6}$$

This becomes useful when $\frac{f(x)}{g(x)}$ leads to an undefined expression such as $\frac{0}{0}$, i.e. $\lim_{x \to x_0} f(x) = 0$ and $\lim_{x \to x_0} g(x) = 0$ or $\lim_{x \to x_0} |f(x)| = \infty$ and $\lim_{x \to x_0} |g(x)| = \infty$. Applied to the summands in the right-hand side of (A.5) for $h_i \neq 0$ we have

$$\lim_{\epsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} = \lim_{\epsilon \to 0} \frac{\log(1 + |h_i|\epsilon^{-1})}{\log(1 + \epsilon^{-1})}, \tag{A.7}$$

with

$$f(\epsilon) = \log(1 + |h_i|\epsilon^{-1}) \tag{A.8}$$

$$g(\epsilon) = \log(1 + \epsilon^{-1}), \tag{A.9}$$

where both functions are ∞ for $\epsilon = 0$; i.e., $f(0) = \infty$ and $g(0) = \infty$. The respective derivatives of (A.8) and (A.9) are

$$f'(\epsilon) = -\frac{|h_i|}{(|h_i|\epsilon^{-1} + 1)\,\epsilon^2} \tag{A.10}$$

$$g'(\epsilon) = -\frac{1}{(\epsilon^{-1} + 1)\,\epsilon^2} \tag{A.11}$$

and we can apply the rule of l'Hopital for $h_i \neq 0$ which yields

$$\begin{aligned} \lim_{\epsilon \to 0} \frac{\log(1 + |h_i|\epsilon^{-1})}{\log(1 + \epsilon^{-1})} &= \lim_{\epsilon \to 0} \frac{-\frac{|h_i|}{(|h_i|\epsilon^{-1}+1)\epsilon^2}}{-\frac{1}{(\epsilon^{-1}+1)\epsilon^2}} \\ &= \lim_{\epsilon \to 0} \frac{|h_i| + |h_i|\epsilon}{|h_i| + \epsilon} \\ &= \frac{|h_i|}{|h_i|} \\ &= 1. \end{aligned} \tag{A.12}$$

For $h_i = 0$ we have the trivial result

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\log(1+0\epsilon^{-1})}{\log(1+\epsilon^{-1})} &= \lim_{\epsilon\to 0}\frac{\log(1)}{\log(1+\epsilon^{-1})}\\ &= \lim_{\epsilon\to 0}\frac{0}{\log(1+\epsilon^{-1})}\\ &= \frac{0}{\infty}\\ &= 0.\end{aligned} \tag{A.13}$$

Combining (A.12) and (A.13) we obtain

$$\lim_{\epsilon\to 0}\frac{\log(1+|h_i|\epsilon^{-1})}{\log(1+\epsilon^{-1})} = \begin{cases}1 & \text{if } h_i \neq 0\\ 0 & \text{if } h_i \equiv 0\end{cases}, \tag{A.14}$$

which is exactly $|h_i|_0$.

B Characteristics and Variations of the Network Topology Control Problem

B.1 Bin-packing problem

In this section we show the relation of our standard problem from Chapter 3 to the well known bin-packing problem [MT90]. The results have been partially published in [CPS+13b]. For convenience we restate Problem 3.6 with the specific constraints for an LTE network:

$$\text{minimize} \quad \sum_{l\in\mathcal{L}} \left(c_l \left| \boldsymbol{t}_l^T \tilde{\boldsymbol{x}} \right|_0 + \sum_{i\in\mathcal{S}_l} \left(e_i \left| \boldsymbol{s}_i^T \tilde{\boldsymbol{x}} \right|_0 + \tilde{f}_i(\tilde{\boldsymbol{x}}) \right) \right) \tag{B.1a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \tilde{\omega}_{i,j}}\ x_{i,j} \leq 1 \qquad i \in \mathcal{M} \tag{B.1b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j} = 1 \qquad j \in \mathcal{N} \tag{B.1c}$$

$$\mathbf{X} \in \{0,1\}^{M\times N}. \tag{B.1d}$$

Problem B.1, which is akin to those in [ZGY+09, NWGY10, GBGF+11], is a generalized bin-packing problem. In other words, the standard bin-packing problem, which is known to be NP-hard, is a special case of Problem B.1. In particular, note that we can recover the standard bin-packing problem from Problem B.1 by simplifying it to single cell base stations ($|S_l| = 1$) with equal static energy consumption same $c_1 = \ldots = c_M = e_1 = \ldots = e_M$ and assuming that $\tilde{\omega}_{i,j}$ depends only on j (i.e., $\tilde{\omega}_{i,j} =: \omega_j$ for some $\omega_j \in \mathbb{R}_{++}$, $j \in \mathcal{N}$). We simplify our objective function by assuming $f_i(\rho_i) = 0$ and $|\mathcal{S}_l| = 1, \quad \forall l \in \mathcal{L}$ and introducing the auxiliary variable $y_i := \sum_{j\in\mathcal{N}} x_{i,j}$. The resulting simplified problem is stated as

$$\text{minimize} \quad \sum_{i\in\mathcal{M}} y_i \tag{B.2a}$$

$$\text{subject to} \sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \omega_j}\ x_{i,j} \leq y_i \qquad i \in \mathcal{M} \tag{B.2b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j} = 1, \qquad j \in \mathcal{N} \tag{B.2c}$$

$$\mathbf{X} \in \{0,1\}^{M\times N} \tag{B.2d}$$

$$y_i \in \{0,1\} \qquad i \in \mathcal{M}. \tag{B.2e}$$

Problem B.2 is the generalized bin-packing problem and as such is also NP-hard, so we cannot expect to obtain computationally efficient optimization schemes that are always guaranteed to obtain optimal solutions.

B.2 Alternatives for the heuristic mapping $[0,1] \to \{0,1\}$

The developed framework of network topology control for energy savings in mobile communication networks from the main part of this study (e.g. Section 3.3.3) incorporates a step to map the solution matrix from $\mathbf{X}^\star \in [0,1]^{M\times N}$ to $\mathbf{X} \in \{0,1\}^{M\times N}$. For this step we have used a heuristic that had reasonably low complexity and good performance in practice. To overcome the necessity of a heuristic algorithm, this section gives some ideas how to find better algorithms that have analytical justification to map $\mathbf{X}^\star \in [0,1]^{M\times N}$ to $\mathbf{X} \in \{0,1\}^{M\times N}$.

We start by replacing the discrete constraint in Problem B.2 by a convex closed set. The idea presented in this section can be straightforwardly used in other discrete linear programming problems. For convenience we recall a simple version of our original discrete optimization problem:

$$\text{minimize} \quad \sum_{i\in\mathcal{M}} c_i \left|\boldsymbol{t}_i^T \tilde{\boldsymbol{x}}\right|_0 \tag{B.3a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \tilde{\omega}_{i,j}}\ x_{i,j} \leq 1 \qquad i \in \mathcal{M} \tag{B.3b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j} = 1, \qquad j \in \mathcal{N} \tag{B.3c}$$

$$\mathbf{X} \in \{0,1\}^{M\times N}. \tag{B.3d}$$

With the above formulation, base station i can be switched off if $\left|\boldsymbol{t}_i^T \tilde{\boldsymbol{x}}\right|_0 = 0$. Now, define $\mu := \sum_{i\in\mathcal{M}} c_i$ (i.e., μ is the maximum static energy consumption, which corresponds to the case where all base stations are switched on), and consider the following optimization problem with convex constraints:

$$\underset{\mathbf{X}\in\mathbb{R}^{M\times N}}{\text{minimize}} \sum_{i\in\mathcal{M}} c_i \left|\boldsymbol{x}_i^T \mathbf{1}\right|_0 + \mu \left|\mathbf{X}\right|_0 \tag{B.4a}$$

$$\text{subject to} \sum_{j\in\mathcal{N}} \frac{r_j}{B_i\ \tilde{\omega}_{i,j}}\ x_{i,j} \leq 1 \qquad i \in \mathcal{M} \tag{B.4b}$$

$$\sum_{i\in\mathcal{M}} x_{i,j} = 1, \qquad j \in \mathcal{N} \tag{B.4c}$$

$$\mathbf{X} \in [0,1]^{M\times N}. \tag{B.4d}$$

In the optimization problem above we have included a l_0-penalty $\mu\left|\mathbf{X}\right|_0$ inducing maximum sparsity to promote sparse solutions also in the optimization variable. The inclusion of such a l_0-penalty is widely used in literature. It has been considered in, for example, sparse linear

regression [SS12, BYD07, BD08, Nik13, ZDL13, DZ13, MUH15], sparse signal recovery [Nik13], PCA and low rank matrix completion [US11, USM15].

Proposition 5. *If Problem B.3 has a solution, then the optimal solution of Problem B.4 is also the optimum solution of Problem B.3.*

We now sketch a proof of Proposition 5. For notational convenience, denote the sets of matrices in $\mathbb{R}^{M\times N}$ satisfying (B.3b), (B.3c), (B.3d), and (B.4d) by, respectively, $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, and $\mathcal{C}_4$.Furthermore, denote the functions $f_1, f_2 : \mathbb{R}^{M\times N} \to \mathbb{R}$ in (B.3a) and (B.4a) by $f_1(\mathbf{X}) := \sum_{i\in\mathcal{M}} c_i \left|\boldsymbol{x}_i^T\mathbf{1}\right|_0$ and $f_2(\mathbf{X}) := \sum_{i\in\mathcal{M}} c_i \left|\boldsymbol{x}_i^T\mathbf{1}\right|_0 + \mu\left|\mathbf{X}\right|_0$. In doing so, Problem (B.3) and Problem (B.4) can be compactly written as, respectively, $\min\limits_{\text{s.t. } \mathbf{X}\in\mathcal{C}_1\cap\mathcal{C}_2\cap\mathcal{C}_3} f_1(\mathbf{X})$ and $\min\limits_{\text{s.t. } \mathbf{X}\in\mathcal{C}_1\cap\mathcal{C}_2\cap\mathcal{C}_4} f_2(\mathbf{X})$. To show the equivalence of these two problems, we use extensively the simple observation below:

Remark 14. *Assume that* $\mathbf{X}\in\mathcal{C}_2$*. Then* $\left|\mathbf{X}\right|_0 \geq N$*, and the equality* $\left|\mathbf{X}\right|_0 = N$ *is achieved if and only if* $\mathbf{X}\in\mathcal{C}_2$ *also belongs to* $\mathcal{C}_3$ *(i.e.,* $\mathbf{X}\in\mathcal{C}_2\cap\mathcal{C}_3$*).*

Proof. The inequality $\left|\mathbf{X}\right|_0 \geq N$ follows directly from the definition of the set $\mathcal{C}_2$, which guarantees that there exists at least one nonzero component in each column of an arbitrary matrix $\mathbf{X}\in\mathcal{C}_2\subset\mathbb{R}^{M\times N}$. Furthermore, the equality $\left|\mathbf{X}\right|_0 = N$ (which characterizes the set of sparsest matrices in $\mathcal{C}_2$) is achieved if and only if a single component of each column of $\mathbf{X}$ is one and the remaining components are zeros. Such matrices in the set $\mathcal{C}_2$ also belong to $\mathcal{C}_3$, by definition of $\mathcal{C}_3$. □

We are now ready to show the equivalence of the problems:

Proposition 6. *Assume that Problem (B.3) has at least one solution. Then an assignment matrix* $\mathbf{X}^\star$ *solves Problem (B.3) if and only if* $\mathbf{X}^\star$ *also solves Problem (B.4).*

Proof. Define $\mathcal{D} := (\mathcal{C}_1\cap\mathcal{C}_2\cap\mathcal{C}_4)\backslash\mathcal{C}_3$, then we deduce:[1]

$$\begin{aligned}
\inf_{\mathbf{X}\in\mathcal{D}} f_2(\mathbf{X}) &= \inf_{\mathbf{X}\in\mathcal{D}} \left(f_1(\mathbf{X}) + \mu\left|\mathbf{X}\right|_0\right) && \text{(B.5)}\\
&\geq \inf_{\mathbf{X}\in\mathcal{D}} f_1(\mathbf{X}) + \mu N + \mu && \text{(B.6)}\\
&\geq \inf_{\mathbf{X}\in\mathcal{D}} f_1(\mathbf{X}) + \min_{\mathbf{X}\in\mathcal{C}_1\cap\mathcal{C}_2\cap\mathcal{C}_3} \left(f_1(\mathbf{X}) + \mu N\right) && \text{(B.7)}\\
&= \inf_{\mathbf{X}\in\mathcal{D}} f_1(\mathbf{X}) + \min_{\mathbf{X}\in\mathcal{C}_1\cap\mathcal{C}_2\cap\mathcal{C}_3} \left(f_1(\mathbf{X}) + \mu\left|\mathbf{X}\right|_0\right) && \text{(B.8)}\\
&> \min_{\mathbf{X}\in\mathcal{C}_1\cap\mathcal{C}_2\cap\mathcal{C}_3} \left(f_1(\mathbf{X}) + \mu\left|\mathbf{X}\right|_0\right) && \text{(B.9)}\\
&= \min_{\mathbf{X}\in\mathcal{C}_1\cap\mathcal{C}_2\cap\mathcal{C}_3} f_2(\mathbf{X}). && \text{(B.10)}
\end{aligned}$$

□

[1] For an arbitrary function g, we define $\inf_{\boldsymbol{x}\in\mathcal{D}} g(\boldsymbol{x}) = \infty$ if $\mathcal{D} = \emptyset$

In the first inequality, we use the fact that $\mathbf{X} \in \mathcal{C}_2$ and $\mathbf{X} \notin \mathcal{C}_3$ implies $|\mathbf{X}|_0 \geq N+1$ (Remark 14); in the second inequality, we use $f_1(\mathbf{X}) \leq \mu$ for every $\mathbf{X} \in \mathbb{R}^{M \times N}$ and the assumption of the existence of a solution to Problem (B.3); in the second equality, we use $|\mathbf{X}|_0 = N$ for every $\mathbf{X} \in \mathcal{C}_2 \cap \mathcal{C}_3$ (Remark 14); in the last inequality we use $f_1(\mathbf{X}) > 0$ for every $\mathbf{X} \in \mathcal{C}_2$. (Note: all above relations are trivially satisfied if $\mathcal{D} = \emptyset$.)

In words, by $\mathcal{C}_3 \subset \mathcal{C}_4$, the above derivation shows that the set of solutions to Problem B.4 is a superset of $\mathcal{C}_3$ if a solution to Problem B.3 exists:

$$\begin{aligned} \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_4}{\arg\min} f_2(\mathbf{X}) &= \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3 \cap \mathcal{C}_4}{\arg\min} f_2(\mathbf{X}) && \text{(B.11)} \\ &= \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3}{\arg\min} f_2(\mathbf{X}) && \text{(B.12)} \end{aligned}$$

with (B.11) conditioned on the existence of a solution to Problem B.3. We arrive at the desired result by using $|\mathbf{X}|_0 = N$ for $\mathbf{X} \in \mathcal{C}_2 \cap \mathcal{C}_3$ (Remark 14) and by removing unnecessary constants:

$$\begin{aligned} \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_4}{\arg\min} f_2(\mathbf{X}) &= \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3}{\arg\min} f_2(\mathbf{X}) && \text{(B.13)} \\ &= \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3}{\arg\min} \left(f_1(\mathbf{X}) + \mu \, |\mathbf{X}|_0\right) && \text{(B.14)} \\ &= \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3}{\arg\min} \left(f_1(\mathbf{X}) + \mu N\right) && \text{(B.15)} \\ &= \underset{\mathbf{X} \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3}{\arg\min} f_1(\mathbf{X}). && \text{(B.16)} \end{aligned}$$

again conditioned on the existence of a solution to Problem B.3.

Further investigations and the detailed proof is subject to future work.

B.3 Additional problem variation for handling infeasible constraints

The underlaying assumption when trying to solve problems akin to Problem B.2 is that it always has a feasible solution. However, depending on the data rate requirements $\boldsymbol{r}$ of users, which typically have to be estimated, we cannot guarantee the existence of a feasible solution. To handle infeasible cases, we can allow the optimization problem to add bandwidth $z_i \geq 0$ to base station $i \in \mathcal{M}$ if required. To do so we replace Problem B.2 with:

$$\text{minimize} \sum_{i \in \mathcal{M}} y_i + \delta \left| \boldsymbol{z} \right|_1 \tag{B.17a}$$

$$\text{subject to} \sum_{j \in \mathcal{N}} \frac{b_{i,j}}{B_i} x_{i,j} \leq y_i + \frac{z_i}{B_i} \qquad i \in \mathcal{M} \tag{B.17b}$$

$$\sum_{i \in \mathcal{M}} x_{i,j} = 1, \qquad j \in \mathcal{N} \tag{B.17c}$$

$$x_{i,j} \in \{0, 1\} \qquad i \in \mathcal{M}, j \in \mathcal{N} \tag{B.17d}$$

$$y_i \in \{0, 1\} \qquad i \in \mathcal{M}, \tag{B.17e}$$

where $\boldsymbol{z} := [z_i \cdots z_M]$, $x_{i,j}$, y_i are the optimization variables; and δ is a parameter penalizing the addition of bandwidth to base stations. Note that, for $\boldsymbol{z} \in \mathbb{R}^N$, the largest convex function being an underestimator of $|\boldsymbol{z}|_0$ over the l_∞ unit ball is the l_1 norm, which shows that the l_1 norm is a sparsity-promoting norm (a property that has been widely exploited in, among other fields, machine learning, geophysics, statistics, and signal processing [CWB08]). As a result, if δ is sufficiently large, then we can expect that a solution to the above optimization problem has a vector $\boldsymbol{z}$ with a large number of zero components (or values close to zero). Note that solutions with large z_i values typically indicate that base station i is overloaded.

C The Majorization-Minimization algorithm

In this section we briefly summarize the majorization-minimization (MM) algorithm [HL04], which can be seen as a generalization of the well- known expectation-maximization (EM) algorithm. The presentation that follows is heavily based on that in the study in [STL11] (see also [PCS12][CSS$^+$14]).

Suppose that the objective is to minimize a function $h : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^N$. Assume that there exists a solution to this optimization problem, and let $\mathbf{x}^\star \in \mathcal{X}$ be a global minimizer of h; i.e., $h(\mathbf{x}^\star) \leq h(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$. Unless h has a special structure that can be exploited (e.g. convexity), finding $\mathbf{x}^\star$ is computationally intractable in general [Roc70]. Hence, we typically have to content ourselves with generating a sequence of vectors with non-increasing objective value. To this end, we can use the majorization-minimization (MM) technique, which drives h downhill with the help of a majorizing function $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In more detail, we say that g is majorizing function for h if it satisfies the following properties:

1. g majorizes h at every point in $\mathcal{X}$, i.e.

$$h(\mathbf{x}) \leq g(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \tag{C.1}$$

2. g and h coincide at $(\mathbf{x}, \mathbf{x})$ so that

$$h(\mathbf{x}) = g(\mathbf{x}, \mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{C.2}$$

By starting from a feasible point $\mathbf{x}^{(0)} \in \mathcal{X}$, the MM algorithm generates a sequence $\left\{\mathbf{x}^{(n)}\right\}_{n\in\mathbb{N}} \subset \mathcal{X}$ with monotone decreasing function values $h(\mathbf{x}^{(n)})$ according to (we assume that the optimization problems have a solution)

$$\mathbf{x}^{(n+1)} \in \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{x}^{(n)}) \,. \tag{C.3}$$

Irrespective of the choice of g, we can easily verify monotonicity of the objective value with the help of (C.1), (C.2) and (C.3): $h(\mathbf{x}^{(n)}) = g(\mathbf{x}^{(n)}, \mathbf{x}^{(n)}) \geq g(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)}) \geq g(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}) = h(\mathbf{x}^{(n+1)})$. Therefore, since the function h is bounded below when restricted to $\mathcal{X}$ by assumption, we can conclude that $h(\mathbf{x}^{(n)}) \to c \in \mathbb{R}$ for some $c \geq h(\mathbf{x}^\star)$ as $n \to \infty$. However, we emphasize that this in general does not imply the convergence of the sequence $\left\{\mathbf{x}^{(n)}\right\}$.

The choice of the function g is problem dependent, but it should be sufficiently structured in order to make the optimization problem in C.3 tractable. In particular, in our study we deal with concave and continuously differentiable functions h. In such cases, a natural choice for g satisfying (C.1) and (C.2) is

$$g(\mathbf{x}, \mathbf{y}) = h(\mathbf{y}) + \nabla h(\mathbf{y})^T(\mathbf{x} - \mathbf{y}). \tag{C.4}$$

This particular choice is common in, for example, sparse signal recovery [CWB08].

Remark 15. *We note that, instead of solving the optimization problem in* C.3 *exactly, it is sufficient for the monotonicity of the sequence* $\{h(\boldsymbol{x}^{(n)})\}$ *that* $g(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)}) \leq g(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})$ *for every* $n \in \mathbb{N}$*. This observation is relevant if the right-hand side of* C.3 *can only be solved asymptotically, in which case the iteration can be truncated whenever the above inequality is satisfied.*

Publication List

[BPS13] J. Bühler, E. Pollakis, and S. Stanczak. Information-theoretic framework for estimating the energy consumption. *GreenNets Project, FP7.SME.2011.1*, 2013. GreenNets Deliverable 4.2, Tech Report.

[CPS13a] R.L.G. Cavalcante, E. Pollakis, and S. Stanczak. Energy-efficient network topology configuration. *GreenNets Project, FP7.SME.2011.1*, 2013. GreenNets Deliverable 5.2, Tech Report.

[CPS+13b] R.L.G. Cavalcante, E. Pollakis, S. Stanczak, S. Stefanski, R. Nowak, T. Kürner, A. Eisenblätter, and D. Montvila. Energy savings in cellular networks. *COST IC1004*, 2013.

[CPS14] R.L.G. Cavalcante, E. Pollakis, and S. Stanczak. Power estimation in LTE systems with the general framework of standard interference mappings. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pages 818–822, December 2014.

[KPS11] M. Kaliszan, E. Pollakis, and S. Stanczak. Efficient beamforming algorithms for MIMO multicast with application-layer coding. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 928 –932, August 2011.

[KPS12] M. Kaliszan, E. Pollakis, and S. Stanczak. Multigroup multicast with application-layer coding: Beamforming for maximum weighted sum rate. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pages 2270 –2275, April 2012.

[PCS12] E. Pollakis, R.L.G. Cavalcante, and S. Stanczak. Base station selection for energy efficient network operation with the majorization-minimization algorithm. In *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*, pages 219–223, June 2012.

[PCS13] E. Pollakis, R.L.G. Cavalcante, and S. Stanczak. Enhancing energy efficient network operation in multi-rat cellular environments through sparse optimization. In *Signal Processing Advances in Wireless Communications (SPAWC), 2013 IEEE 14th Workshop on*, pages 260–264, June 2013.

[PCS16] E. Pollakis, R.L.G. Cavalcante, and S. Stanczak. Traffic demand-aware topology control for enhanced energy-efficiency of cellular networks. *EURASIP Journal on Wireless Communications and Networking*, 2016(1):1–17, 2016.

[PCSP12] E. Pollakis, R.L.G. Cavalcante, S. Stanczak, and F. Penna. Robust interference identification for multi-RAT optimization in wireless cellular networks. In *New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2012 IEEE Symposium on*, October 2012.

[PS16] E. Pollakis and S. Stanczak. Anticipatory networking for energy savings in 5G systems. In *WSA 2016; 20th International ITG Workshop on Smart Antennas; Proceedings of*, March 2016.

References

[3GP10a] 3GPP. Further advancements for EUTRA: Physical layer aspects (release 9), TR 36.814 v2.0.1, March 2010.

[3GP10b] 3GPP. Further advancements for EUTRA: Physical layer aspects (release 9), TR 36.814 v2.0.1, March 2010.

[ABG+10] G. Auer, O. Blume, V. Giannini, I. Godor, M.A. Imran, Y. Jading, E Katranaras, M. Olsson, D. Sabella, P. Skillermark, and W. Wajda. D2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. Technical report, INFSO-ICT-247733 EARTH (Energy Aware Radio and NeTwork TecHnologies), December 2010.

[ACM03] E. Amaldi, A. Capone, and F. Malucelli. Planning UMTS base station location: Optimization models with power control and algorithms. *Wireless Communications, IEEE Transactions on*, 2(5):939 – 952, sept. 2003.

[ACMS03] E. Amaldi, A. Capone, F. Malucelli, and F. Signori. Optimization models and algorithms for downlink umts radio planning. In *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*, volume 2, pages 827–831 vol.2, 2003.

[AK04] A. Agarwal and P.R. Kumar. Capacity bounds for ad hoc and hybrid wireless networks. *ACM SIGCOMM Computer Communication Review*, 34(3):71–81, 2004.

[AL08] Alcatel-Lucent. 9900 wireless network guardian, 2008. Tech. White Paper.

[AL15] Alcatel-Lucent. Strategic white paper: 5G is coming - are you prepared? Technical report, Alcatel-Lucent, March 2015.

[AzHV13] H. Abou-zeid, H.S. Hassanein, and S. Valentin. Optimal predictive resource allocation: Exploiting mobility patterns and radio maps. In *Global Communications Conference (GLOBECOM), 2013 IEEE*, pages 4877–4882, Dec 2013.

[AzHV14] H. Abou-zeid, H.S. Hassanein, and S. Valentin. Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks. In *IEEE Transactions on Vehicular Technology*, volume 63/, pages 2013–2026, Jun 2014.

[BC11] H.H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[BD08] T. Blumensath and M.E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

[BHM77] S.P. Bradley, A.C. Hax, and T.L. Magnanti. *Applied mathematical programming*. Addison-Wesley Pub. Co., 1977.

[BHSN06] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak. Compressive wireless sensing. In *Proceedings of the 5th International Conference on Information Processing in Sensor Networks*, IPSN '06, pages 134–142, New York, NY, USA, 2006. ACM.

[BV06] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, Cambridge, U.K., 2006.

[BYD07] T. Blumensath, M. Yaghoobi, and M.E. Davies. Iterative hard thresholding and l_0 regularisation. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 3, pages III–877–III–880, April 2007.

[CCK+12] L. Chiaraviglio, D. Ciullo, G. Koutitas, M. Meo, and L. Tassiulas. Energy-efficient planning and management of cellular networks. In *Wireless On-demand Network Systems and Services (WONS), 2012 9th Annual Conference on*, pages 159 –166, Jan. 2012.

[CDS98] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[CH08] A. Corliano and M. Hufschmid. Energieverbrauch der mobilen Kommunikation - Schlussbericht. Technical report, Bundesamt für Energie, Schweizerische Eidgenossenschaft, Bern, Swiss, February 2008. (in German).

[Cis15] Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019. White Paper, February 2015.

[CJ08] V.R. Cadambe and S.A. Jafar. Interference alignment and degrees of freedom of the k-user interference channel. *IEEE Trans. Inform. Theory*, 54(8):3425 –3441, Aug. 2008.

[CRT06] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, Feb 2006.

[CSS+14] R.L.G. Cavalcante, S. Stanczak, M. Schubert, A. Eisenblätter, and U. Türke. Toward energy-efficient 5G wireless communications technologies: Tools for decoupling the scaling of networks from the growth of operating power. *IEEE Signal Processing Mag.*, 31(6):24–34, Nov. 2014.

[CSS15] R.L.G. Cavalcante, Y. Shen, and S. Stanczak. Elementary properties of positive concave mappings with applications to network planning and optimization. *IEEE Trans. Signal Processing*, 64(7):1774–1783, April 2015.

[CSZZ16] R. L. G. Cavalcante, S. Stanczak, J. Zhang, and H. Zhuang. Low complexity iterative algorithms for power estimation in ultra-dense load coupled networks. *IEEE Transactions on Signal Processing*, 64(22):6058–6070, Nov 2016.

[CT06] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, Dec 2006.

[CW08] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

[CWB08] E.J. Candes, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.*, 14(5):877–905, Dec. 2008.

[CY08] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3869–3872, March 2008.

[CYZ+11] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman. Network energy saving technologies for green wireless access networks. *Wireless Communications, IEEE*, 18(5):30–38, October 2011.

[CZB+10] L.M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Godor, G. Auer, and L. Van der Perre. Challenges and enabling technologies for energy aware mobile radio networks. *Communications Magazine, IEEE*, 48(11):66–72, November 2010.

[CZXL11] Y. Chen, S. Zhang, S. Xu, and G.Y. Li. Fundamental trade-offs on green wireless networks. *IEEE Communications Magazine*, 49(6):30–37, June 2011.

[DDT+08] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, Ting Sun, K.F. Kelly, and R.G. Baraniuk. Single-pixel imaging via compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):83–91, March 2008.

[DZ13] B. Dong and Y. Zhang. An efficient algorithm for l_0 minimization in wavelet frame based image restoration. *Journal of Scientific Computing*, 54(2-3):350–368, 2013.

[Efr87] B. Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

[FF12] A.J. Fehske and G.P. Fettweis. Aggregation of variables in load models for interference-coupled cellular data networks. In *Communications (ICC), 2012 IEEE International Conference on*, pages 5102–5107, June 2012.

[FF13] A.J. Fehske and G.P. Fettweis. On flow level modeling of multi-cell wireless networks. In *Modeling Optimization in Mobile, Ad Hoc Wireless Networks (WiOpt), 2013 11th International Symposium on*, pages 572–579, May 2013.

[FKVF13] A.J. Fehske, H. Klessig, J. Voigt, and G.P. Fettweis. Concurrent load-aware adjustment of user association and antenna tilts in self-organizing radio networks. *Vehicular Technology, IEEE Transactions on*, 62(5):1974–1988, Jun 2013.

[Gal87] R.G. Gallager. Energy limited channels: Coding, multiaccess, and spread spectrum. *Proc. Conf. Information Sciences and Systems (CISS)*, page 372, March 1987.

[GAPRS05] L. Giupponi, R. Agusti, J. Perez-Romero, and O. Sallent. Joint radio resource management algorithm for multi-RAT networks. In *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, volume 6, pages 3851–3855, Dec. 2005.

[GBGF+11] P. Gonzalez-Brevis, J. Gondzio, Yijia Fan, H.V. Poor, J. Thompson, I. Krikidis, and Pei-Jung Chung. Base station location optimization for minimal energy consumption in wireless networks. In *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pages 1 –5, May 2011.

[Gee08] H.-F. Geerdes. *UMTS radio network planning: Mastering cell coupling for capacity optimization.* PhD thesis, Technical University of Berlin, February 2008. Ph.D. dissertation.

[GIAT15] A. Galanopoulos, G. Iosifidis, A. Argyriou, and L. Tassiulas. Green video delivery in LTE-based heterogeneous cellular networks. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2015 IEEE 16th International Symposium on a*, pages 1–9, June 2015.

[GK00] P. Gupta and P.R. Kumar. The capacity of wireless networks. *IEEE Trans. Inform. Theory*, 46(2):388–404, March 2000.

[GK03] P. Gupta and P.R. Kumar. Towards an information theory of large networks: An achievable rate region. *IEEE Trans. Inform. Theory*, 49(8):1877–1894, August 2003.

[GPMM14] S. Gamboa, A. Pelov, P. Maille, and N. Montavont. Exploiting user delay-tolerance to save energy in cellular network: An analytical approach. In *Personal, Indoor, and Mobile Radio Communication (PIMRC), 2014 IEEE 25th Annual International Symposium on*, pages 1426–1431, Sept 2014.

[GR97] I.F. Gorodnitsky and B.D. Rao. Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3):600–616, Mar 1997.

[GT02] M. Grossglauser and D. Tse. Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Transactions on Networking*, 10(4):477–486, Aug 2002.

[GW02] A.J. Goldsmith and S.B. Wicker. Design challenges for energy-constrained ad hoc wireless networks. *Wireless Communications, IEEE*, 9(4):8–27, Aug 2002.

[HHA+11] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P.M. Grant, H. Haas, J.S. Thompson, I. Ku, Cheng-Xiang Wang, Tuan Anh Le, M.R. Nakhai, Jiayi Zhang, and L. Hanzo. Green radio: radio techniques to enable energy-efficient wireless networks. *IEEE Commun. Mag.*, 49(6):46 –54, June 2011.

[HKD11] J. Hoydis, M. Kobayashi, and M. Debbah. Green small-cell networks. *IEEE Veh. Technol. Magazine*, 6(1):37–43, March 2011.

[HL04] D.R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, Feb. 2004.

[HV04] H. Huang and S. Venkatesan. Asymptotic downlink capacity of coordinated cellular networks. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on*, volume 1, pages 850–855 Vol.1, Nov 2004.

[HYLS15] C.K. Ho, D. Yuan, L.Lei, and S. Sun. Power and load coupling in cellular networks for energy optimization. *Wireless Communications, IEEE Transactions on*, 14(1):509–519, Jan 2015.

[HYS14] C. K. Ho, D. Yuan, and S. Sun. Data offloading in load coupled networks: A utility maximization framework. *Wireless Communications, IEEE Transactions on*, 13(4):1921–1931, April 2014.

[IBM15] IBM Corp. IBM ILOG CPLEX optimization studio - CPLEX user's manual. *Tech. manual*, 2015.

[Int15] GSMA Intelligence. The Mobile Economy 2015. Technology Report, 2015.

[JB09] S. Joshi and S. Boyd. Sensor selection via convex optimization. *IEEE Trans. Signal Processing*, 57(2):451–462, Feb. 2009.

[JKV11] A. Jain, S.R. Kulkarni, and S. Verdu. Multicasting in large wireless networks: Bounds on the minimum energy per bit. *Information Theory, IEEE Transactions on*, 57(1):14–32, Jan 2011.

[KCV+16] M. Kasparick, R. L. G. Cavalcante, S. Valentin, S. Stanczak, and M. Yukawa. Kernel-based adaptive online reconstruction of coverage maps with side information. *IEEE Transactions on Vehicular Technology*, 65(7):5461–5473, July 2016.

[KNG11] S. Kamath, U. Niesen, and P. Gupta. The capacity per unit energy of large wireless networks. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 1618–1622, Jul./Aug. 2011.

[KT03] U.C. Kozat and L. Tassiulas. Throughput capacity of random ad hoc networks with infrastructure support. In *ACM/MOBICOM-2003*, San Diego, Sep 2003.

[LDP07] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[Lib04] L.S. Liberti. *Reformulation and Convex Relaxation Techniques for Global Optimization*. PhD thesis, Imperial College London, Department of Chemical Engineering and Chemical Technology, South Kensington Campus, London SW7 2AZ, 2004.

[LLT03] B. Liu, Z. Liu, and D. Towsley. On the capacity of hybrid wireless networks. In *IEEE INFOCOM 2003*, volume 2, pages 1543–1552, Mar 2003.

[Lov80] E.R. Love. Some logarithm inequalities. *The Mathematical Gazette*, 67(439):55–57, Mar. 1980.

[LPGdlR+11] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T.Q.S. Quek, and J. Zhang. Enhanced intercell interference coordination challenges in heterogeneous networks. *Wireless Communications, IEEE*, 18(3):22–30, June 2011.

[Man96] O.L. Mangasarian. Machine learning via polyhedral concave minimization. In H. Fischer, B. Riedmüller, and S. Schäffler, editors, *Applied Mathematics and Parallel Computing*, pages 175–188. Physica-Verlag HD, 1996.

[MET13] METIS Project. Deliverable D6.1: Simulation guidelines. *METIS Project (Mobile and wireless communications Enablers for the Twenty-twenty Information Society)*, ICT-317669-METIS/D6.1, 2013.

[MET15] METIS. METIS ray tracing files. *https://www.metis2020.com/documents/simulations/*, 2015. Accessed on 20.03.2015.

[MK10] K. Majewski and M. Koonert. Conservative cell load approximation for radio networks with shannon channels and its application to LTE network planning. In *Telecommunications (AICT), 2010 Sixth Advanced International Conference on*, pages 219 –225, May 2010.

[MNK+07] P. Mogensen, Wei Na, I.Z. Kovacs, F. Frederiksen, A. Pokhariyal, K.I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela. LTE capacity compared to the shannon bound. In *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, pages 1234–1238, 2007.

[MT90] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Inc., New York, NY, USA, 1990.

[MTHB07] K. Majewski, U. Türke, X. Huang, and B. Bonk. Analytical cell load assessment in OFDM radio networks. In *Proc. IEEE PIMRC'07*, pages 1–5, 2007.

[MUH15] G. Marjanovic, M.O. Ulfarsson, and A.O. Hero. Mist: l_0 sparse linear regression with momentum. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 3551–3555, April 2015.

[Nik13] M. Nikolova. Description of the minimizers of least squares regularized with l_0-norm. Uniqueness of the global minimizer. *SIAM Journal on Imaging Sciences*, 6(2):904–937, 2013.

[NM05] M.J. Neely and E. Modiano. Capacity and delay tradeoffs for ad-hoc mobile networks. *IEEE Trans. Inform. Theory*, 51(6):1917–1937, Jun 2005.

[NWGY10] Z. Niu, Y. Wu, J. Gong, and Z. Yang. Cell zooming for cost-efficient green cellular networks. *IEEE Commun. Mag.*, 48(11):74 –79, November 2010.

[OBB+14] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M.A. Uusitalo, B. Timus, and M. Fallgren. Scenarios for 5G mobile and wireless communications: the vision of the METIS project. *Communications Magazine, IEEE*, 52(5):26–35, May 2014.

[OKLN11] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu. Toward dynamic energy-efficient operation of cellular network infrastructure. *Communications Magazine, IEEE*, 49(6):56 –61, june 2011.

[OLT07] A. Ozgur, O. Leveque, and D.N.C. Tse. Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *Information Theory, IEEE Transactions on*, 53(10):3549–3572, Oct 2007.

[PSG13] C. Phillips, D. Sicker, and D. Grunwald. A survey of wireless path loss prediction and coverage mapping methods. *Communications Surveys Tutorials, IEEE*, 15(1):255–270, First 2013.

[Qur15] R. Qureshi. Ericsson Mobility Report. Technology Report, June 2015.

[Roc70] R.T. Rockafellar. *Convex Analysis*. Princeton Mathematical Series. University Press, 1970.

[SB12] M. Schubert and H. Boche. *Interference Calculus - A General Framework for Interference Management and Network Utility Optimization.* Springer, Berlin, 2012.

[SBP15] J. Song, P. Babu, and D.P. Palomar. Sparse generalized eigenvalue problem via smooth optimization. *Signal Processing, IEEE Transactions on*, 63(7):1627–1642, April 2015.

[SQBB10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[SS12] A.J. Seneviratne and V. Solo. On vector l_0 penalized multivariate regression. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3613–3616, March 2012.

[STL11] B.K. Sriperumbudur, D.A. Torres, and G.R.G. Lackriet. A majorization-minimization approach to the sparse generalized eigenvalue problem. *Machine Learning*, 85(1-2):3–39, Oct. 2011.

[SWB09] S. Stanczak, M. Wiczanowski, and H. Boche. *Fundamentals of resource allocation in wireless networks: theory and algorithms.* Springer, 2 edition, 2009.

[SY12] I. Siomina and D. Yuan. Analysis of cell load coupling for LTE network planning and optimization. *IEEE Trans. Wireless Commun.*, 11(6):2287–2297, June 2012.

[TN12] R. Wood T. Norman. LTE infrastructure: Worldwide demand drivers and base station forecast 2012-2017. Technical report, Analysys Mason Group, May 2012.

[Tut98] K. Tutschku. Demand-based radio network planning of cellular mobile communication systems. In *INFOCOM '98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 3, pages 1054–1061 vol.3, Mar 1998.

[US11] M.O. Ulfarsson and V. Solo. Vector l_0 sparse variable PCA. *Signal Processing, IEEE Transactions on*, 59(5):1949–1958, May 2011.

[USM15] M.O. Ulfarsson, V. Solo, and G. Marjanovic. Sparse and low rank decomposition using l_0 penalty. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 3312–3316, April 2015.

[Ver02] S. Verdu. Recent results on the capacity of wideband channels in the low-power regime. *IEEE Wireless Commun. Mag.*, 9:40–45, August 2002.

[VPRSA09] N. Vucevic, J. Perez-Romero, O. Sallent, and R. Agusti. Joint radio resource management for LTE-UMTS coexistence scenarios. In *Personal,*

Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on, pages 12 –16, Sept. 2009.

[WMBW09] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz. Primary user behavior in cellular networks and implications for dynamic spectrum access. *IEEE Commun. Mag.*, 47(3):88 –95, March 2009.

[Wu83] C.F.J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[XK04] L.-L. Xie and P. R. Kumar. A network information theory for wireless communication: scaling laws and optimal operation. *IEEE Trans. Inform. Theory*, 50(5):748–767, May 2004.

[Yat95] R.D. Yates. A framework for uplink power control in cellular radio systems. *IEEE J. Select. Areas Commun.*, 13(7):pp. 1341–1348, Sept. 1995.

[YO04] I. Yamada and N. Ogura. Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings. *Numer. Funct. Anal. Optim.*, 27(7/8):619–655, 2004.

[YTW+11] S. Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson. Capacity and coverage enhancement in heterogeneous networks. *Wireless Communications, IEEE*, 18(3):32–38, June 2011.

[YYY11] I. Yamada, M. Yukawa, and M. Yamagishi. *Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings*. Springer-Verlag, 2011. IN: Fixed-Point Algorithms for Inverse Problems in Science and Engineering and H.H. Bauschke and R. Burachick and P.L. Combettes and V. Elser and D. R. Luke and H. Wolkowicz.

[ZDL13] Y. Zhang, B. Dong, and Z. Lu. l_0 minimization for wavelet frame based image restoration. *Mathematics of Computation*, 82(282):995–1015, 2013.

[ZdV05] A. Zemlianov and G. de Veciana. Capacity of ad hoc wireless networks with infrastructure support. *IEEE Journal on Selected Areas in Communications*, 23(3):657–667, Mar. 2005.

[ZGY+09] S. Zhou, J. Gong, Z. Yang, Z. Niu, and P. Yang. Green mobile access network with dynamic base station energy saving. In *ACM MobiCom*, September 2009.