Statistical Learning, Anomaly Detection, and Optimization in Self-Organizing Networks

vorgelegt von
Master of Science
Qi Liao
geb. in Nanchang

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

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzende:	Prof. Giuseppe Caire, Ph.D.
Gutachter:	Prof. Dr.-Ing. Sławomir Stańczak
Gutachter:	Prof. Wei Yu, Ph.D. (University of Toronto, Canada)
Gutachter:	Prof. Dr.-Ing. Thomas Kürner (TU Braunschweig, Germany)
Gutachter:	Dr.-Ing. Anastasios Giovanidis (CNRS, France)

Tag der wissenschaftlichen Aussprache: 21. November 2016

Berlin 2016

This thesis is dedicated to all people who have supported me all the way

My parents

Winfried & Qijing

Haotian

Acknowledgements

This thesis was written during my time as a research associate in Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute and as a doctoral candidate at Technical University of Berlin.

First and foremost, I would like to thank my supervisor, Prof. Dr.-Ing. Sławomir Stańczak, for giving me the opportunity to pursue my Doctoral studies and working with him. A Chinese proverb says, "One day's teacher, a whole life's father". I would like to thank Prof. Stańczak for being my teacher for eight years, ever since I took his course of "Resource Allocation in Wireless Networks" in graduate school in 2008, and for being an excellent example of a passionate scientist and a serious scholar.

A special thankyou to Dr. Renato L. G. Cavalcante for his valuable guidance, constructive remarks, and for taking the effort to referee this thesis. He has provided generous help, support and motivation to young researchers, ever since he joined our team in Heinrich Hertz Institute.

I would like also to thank Dr. Martin Schubert, Dr. Anastasios Giovanidis and Dr. Marcin Wiczanowski for providing interesting ideas and discussions. I have greatly enjoyed the opportunity to work with them on the topics of Self-Organizing Networks.

I would like to express my deepest gratitude to all my former colleagues in Heinrich Hertz Institute and at Technical University of Berlin for providing a comfortable and inspiring working environment. A special thankyou to Dr. Setareh Maghsudi, I miss the days when we were office-mates at Fraunhofer Mobile Communications Lab. Martin Kasparick, Jafar Mohammadi, Emmanuel Pollakis and Miguel Angel Gutierrez, thank you for a good time, I will miss your company.

The internship opportunity I had at Bell Laboratories, Alcatel-Lucent was a great opportunity for learning and professional development. I express my deepest thanks to Dr. Tim Kam Ho, Dr. Chun-Nam Yu, Dr. Carl Nuzman and Dr. Iraj Saniee for giving precious advises and guidance, and for arranging all facilities at Bell Laboratories, Murray Hill. I would also like to thank Dr. Stefan Valentin for his careful guidance for my internship at Bell Laboratories, Stuttgart.

Finally, I am grateful to my parents, my dear husband, and all my families and friends, who have never stopped believing in me, and always supported me with love and caring.

Berlin, September 2016 Qi Liao

Abstract

Self-organizing network, considered as a starting point toward self-aware cognitive network, is an automation technology designed for automated configuring, monitoring, troubleshooting and optimizing for the next generation mobile networks. Its main functionalities include: self-configuration, self-optimization and self-healing. With the emergence of new wireless devices and applications, the increasing demand for mixed types of services motivates extremely dense and heterogeneous deployments. As a result it is expected that a large amount of measurements and signaling overhead will be generated in future networks. Partial and inaccurate network knowledge, together with the increasing complexity of envisioned wireless networks, pose one of the biggest challenges for self-organizing network (SON) – maintaining perfect global network information at the level of autonomous network elements is simply illusive in large-scale, highly dynamic wireless networks. Another big challenge is the network-wide optimization of interacting or conflicting SON functionalities, with the goal of improving the efficiency of total algorithmic machinery on the network level.

This thesis studies SON in the context of erroneous and incomplete local information on network state, as well as possibly conflicting and abstractly defined objectives of different SON functions. We design novel mathematical models and statistical methods for enhancing network awareness at the locality of network elements through statistical learning, intelligent monitoring, and dynamic network feedback collection amidst network uncertainties. The extracted knowledge is used to optimize the network performance by adjusting to internal and exogenous network variations, critical network conditions, and different network anomalies.

Context-aware frameworks are proposed for automatic configuration and tuning of network elements with minimal operator intervention to achieve timely detection of network abnormal states such as coverage holes, and to carry out a network-wide optimization of different SON functions. The results prove the benefits of the developed self-healing and self-optimization functions, including cell outage detection, network state classification and anomaly detection, random access channel (RACH) optimization, mobility robustness optimization, mobility load balancing, interference reduction, and coverage and capacity optimization. We achieve timely detection and identification of network abnormal states based on the analysis of data extracted from the network. The anomaly detection algorithm automatically activates the corresponding self-healing and self-optimization algorithms for single or multiple SON use cases, which frees up operational resource and improves user-centric quality of service.

Zusammenfassung

In der nächsten Generation von Mobilfunknetzen werden selbstorganisierende Netzwerke zum Einsatz kommen, in denen die Netzwerkaufgaben: Konfiguration, Überwachung, Fehlerbehebung und Optimierung automatisiert durchgeführt werden. Mit den Eigenschaften zur Selbst-Konfiguration, Selbstoptimierung und Selbstheilung wird ein selbstorganisierendes Netzwerk auch als Vorstufe zu einem kognitiven Netzwerk betrachtet. Um die steigende Nachfrage nach mobilen Services zu erfüllen werden neue Netzinfrastrukturen ausgerollt, die zusammen mit bestehenden Netzwerken heterogene Strukturen bilden. Infolge von der Komplexität des Netzwerks werden große Mengen an zusätzlichen Protokoll-Overhead und Netzwerkkontrolldaten erhoben. Unvollständige sowie ungenaue Netzwerkkenntnisse sowie die zunehmende Komplexität stellen eine der größten Herausforderung eines selbstorganisierenden Netzwerks dar. Das Pflegen einer globalen Information über den Netzwerkzustand auf der Ebene der Netzwerkelemente ist illusorisch in großen, hochdynamischen Mobilfunknetzen. Eine weitere Herausforderung ist die netzwerkweite Optimierung der untereinander verflochtenen Eigenschaften eines selbstorganisierenden Netzwerks.

Die vorliegende Arbeit untersucht ein selbstorganisierendes Netzwerk im Zusammenhang mit fehlerhafter und unvollständiger Informationen über den Netzwerkzustand sowie unter bestimmten Bedingungen widersprüchliche und abstrakt definierte Optimierungsziele. Wir entwickeln neuartige mathematische Modelle und statistische Methoden zur Verbesserung der Netzwerk-Bewusstsein bei der Netzelementen durch statistisches Lernen, intelligente Überwachung und dynamische Netzwerk-Feedback-Sammlung inmitten Netzwerk Unsicherheiten. Die extrahierte Wissen wird verwendet durch Einstellen der internen und exogene Netzwerk Variationen, kritische Netzwerkbedingungen und verschiedenen Netzanomalien, um die Netzwerkleistung zu optimieren.

Ein Lösungsansatz wird zur Lösung der automatischen Konfiguration und Optimierung von Netzwerkeelementen mit minimalem Benutzereingriff vorgeschlagen, welches ebenfalls eine rechtzeitige Erkennung von abnormen Netzwerkzuständen beinhaltet. Die erzielten Ergebnisse belegen, dass die Netzwerkleistung profitiert von der neuen entwickelten Funktionalität der Selbstheilung und der Selbstoptimierung, einschließlich Zellausfall Erkennung, Netzwerkstatus Klassifizierung und Erkennung von Anomalien, Optimierung von Kanal mit wahlfreiem Zugriff (RACH), Mobilität Robustheit Optimierung, Mobilität Lastausgleich, Interferenzunterdrücken, und Abdeckung und Kapazitätsoptimierung. Wir erreichen rechtzeitige Erkennung und Identifizierung von Netzwerk anormale Zustände basierend auf der Analyse von Daten, die aus dem Netzwerk extrahiert werden. Die Anomalie-Detektionsalgorithmus aktiviert automatisch die entsprechenden Selbstheilung und Selbstoptimierungsalgorithmen für einzelne oder mehrere SON Szenarien, und dadurch die operativen Ressourcen entlastet und die benutzerorientierte Servicequalität verbessert.

Contents

List of Figures

List of Tables

List of Symbols

$\boldsymbol{X}$	Matrix
(x_{ij})	Matrix
$\lvert\boldsymbol{X}\rvert$	Matrix determinant
$\operatorname{diag} X$	Diagonal of matrix
$(\boldsymbol{X})_{ij}$	Matrix entry
$\mathcal{G}(\boldsymbol{X})$	Direct graph of $\boldsymbol{X} \in \mathbb{R}^{n\times n}$
$\boldsymbol{X} \circ \boldsymbol{Y}$	Hadamard product of two matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$
$\boldsymbol{X}^{-1}$	Matrix inverse
$\boldsymbol{X} \otimes \boldsymbol{Y}$	Kronecker product
$\lVert\boldsymbol{X}\rVert$	Matrix norm
$\rho(\boldsymbol{X})$	Spectrum radius
$\sigma(\boldsymbol{X})$	Matrix spectrum
$\mathrm{Tr}(\boldsymbol{X})$	Trace of matrix
$\boldsymbol{X}^T$	Transpose matrix
x	Scalar over $\mathbb{R}$
$\bar{x}$	Conjugate complex of scalar x
$\mathcal{X}$	Set
$\mathcal{A} \times \mathcal{B}$	Cartesian product of two sets $\mathcal{A}$ and $\mathcal{B}$
$\boldsymbol{x}$	Vectors
$\mathrm{diag}(\boldsymbol{x})$	Diagonal matrix with diagonal $\boldsymbol{x}$
$\langle\boldsymbol{x}, \boldsymbol{y}\rangle$	Inner product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$
$\lVert\boldsymbol{x}\rVert_p$	l_p Norm on a vector space
$\lVert\boldsymbol{x}\rVert$	Norm on a vector space
$\boldsymbol{f}^n := \boldsymbol{f} \circ \boldsymbol{f}^n$	n-fold composition of function $\boldsymbol{f} : \mathbb{R}^k_+ \to \mathbb{R}^k_+$
$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\boldsymbol{x}$ follows multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
lg	Common logarithm with base 10
log	Binary logorithm
$\mathbb{R}$	Real numbers
$\mathbb{R}_+$	Nonnegative real numbers
$\mathbb{R}_{++}$	Positive real numbers
$\mathbb{R}^n$	an n-dimensional vector space over $\mathbb{R}$

$\boldsymbol{I}$	Identity matrix
$\boldsymbol{x} \gneq \boldsymbol{y}$	$\boldsymbol{x} \geq \boldsymbol{y}$ with $\boldsymbol{x} \neq \boldsymbol{y}$
$\boldsymbol{x} + c$	entry wise addition $\boldsymbol{x} + (c, \ldots, c)$

Acronyms

3GPP	3rd generation partnership project
5G	fifth generation
ACPCI	automated configuration of physical cell identity
AIMD	Additive Increase Multiplicative Decrease
ANRF	automatic neighbor relation function
AWGN	additive white Gaussian noise
BS	base station
CBR	call blocking rate
CCO	coverage and capacity optimization
CDR	call drop rate
CHT	composite hypothesis testing
CIO	cell individual offset
CoUD	coupled uplink and downlink
CP	collision probability
CQI	channel quality indicator
CS_SR	call setup success rate
CSMA/CA	Carrier sense multiple access with collision avoidance
DeUD	decoupled uplink and downlink
DL	downlink
DMP	detection miss probability
DP	dropping probability
DUDe	downlink/uplink decoupling
E-RAB	E-UTRAN radio access bearer
eNB	evolved Node B
ERAB_SR	E-UTRAN radio access bearer setup success rate
ES	energy savings
FCM	fuzzy c-means
FDD	frequency-division duplex

GLS	generalized least square
GP	Gaussian process
HetNet	heterogeneous networks
HF	handover failure
HFR	handover failure rate
HO	handover
HO_PPR	handover ping-pong rate
HOI_SR	handover (incoming) success rate
HOM	handover margin
HOO_SR	handover (outgoing) success rate
HRQ	handover request
ICI	inter-cell interference
ICIC	inter-cell interference coordination
ID	identification
IR	interference reductions
KKT	Karush-Kuhn-Tucker
KL	Kullback-Leibler
KPI	key performance indicator
LB	load balancing
LRT	likelihood ratio test
LTE	long-term evolution
LTE-A	long-term evolution advanced
MBB	mobile broadband
MCC	mission critical communications
MIAD	Multiplicative Increase Additive Decrease
MLBO	mobility load balancing optimization
MLE	maximum-likelihood estimator
MMC	massive machine communications
MRO	mobility robustness optimization
MSE	mean square error
MTC	machine type communications
NRMSE	normalized root mean square error
OFDM	orthogonal frequency-division multiplexing
OFDMA	orthogonal frequency-division multiplexing access
OL	overloaded
PC	principal component
PCA	principal component analysis
PHY	physical layer

PPHO	ping-pong handover
PRB	physical resource block
PSD	power spectral density
QCI	QoS class identifier
QoS	quality of service
RACH	random access channel
RAT	radio access technology
RB	resource block
RLF	radio link failure
RLFR	radio link failure rate
RRC	radio resource control
RRCS_SR	RRC setup success rate
RRQ	registration request
RSRP	reference signal received power
RSRQ	reference signal received quality
RSSI	received signal strength indication
SAT	service average throughput
SC	subcarrier
SINR	signal-to-interference-plus-noise ratio
SMT	service maximum throughput
SNR	signal-to-noise ratio
SON	self-organizing network
SP	success probability
SVD	singular value decomposition
TBS	transport block size
TDD	time-division duplex
TR	target
TTI	transmission time interval
TTT	time-to-trigger
Tx	transmission
UE	user equipment
UL	uplink
VoIP	voice over IP

Part I

Introduction and Background

Chapter 1

Introduction

1.1 Motivation and Objectives

With the emergence of new wireless devices and applications, there has been a dramatic increase in demand for radio spectrum and network capacity over the past few years. This exponential trend, which is expected to continue in the coming years, together with the high costs of deploying additional base stations (BSs), motivate the development and commercialization of new types of wireless networks with a large number of network elements. These developments are expected to increase network management complexity by orders of magnitude, particularly so because these technologies release the network elements from tight network control. Efficient network management becomes a crucial priority for smooth network operation, while it accounts for a fairly significant fraction of network operating costs. The principal objective of SON is to significantly reduce the human interventions, and with it the capital and operation expenditures: less manual effort for planning, configuring, optimizing and maintaining provides clear competitive advantages in the mobile business.

Existing approaches to network management and self-organization are inadequate to cope with the growth of autonomous network elements and a paradigm shift is necessary in order to prevent a slowdown in network development due to that inadequacy. How to extract knowledge about the network states and build predictive models from large amount of collected data poses one of the biggest challenges for self-organizing wireless networks because maintaining perfect global network information at the level of autonomous network elements is simply illusive in large-scale, highly dynamic wireless networks. Another big challenge is a network-wide optimization of isolated SON functionalities to identify and avoid conflicts of different SON functionalities as well as to improve the efficiency of the total algorithmic machinery on the network level.

Many works have been carried out on the optimization of SON use cases in the EU FP7 SOCRATES project [SOC08b, SOC08a, SOC09, ALS$^+$08]. However, self-organization has not been sufficiently studied in the context of erroneous and incomplete local information, and possibly conflicting and abstractly defined objectives of different SON functionalities. Such a network perspective is necessary to uncover potential objective conflicts of different use cases, identify procedural synergies on the network level and provide insights in infrastructural and dimensioning requirements of multiple simultaneously enabled SON functionalities.

The ongoing developments show a clear trend to rethink SON and essentially redesign wireless network management by incorporating statistical learning, sensing, control and optimization theory principles; these fields are mature now and have well-defined techniques and metrics. This thesis exploits these methods to deliver novel approaches to the challenge of extracting knowledge from the network at a node level, developing node awareness about network surroundings, and leveraging it to drive the system to a desired operational point in a self-coordinated fashion, with the goal of reducing human involvement in network operational tasks for 3rd generation partnership project (3GPP) long-term evolution advanced (LTE-A) and beyond. We also develop multi-objective algorithms to jointly optimize different SON functionalities by considering network-wide interactions between them. The following network functionalities lie in the focus of this thesis:

- *Outage detection*: The objective is to automatically detect and localize unpredictable failures from collected performance measurements feedback without a priori knowledge at network elements.

- *Supervised network state inference and anomaly detection*: We target efficient network state monitoring and proactive cell anomaly detection by incorporating a priori knowledge based on historically collected information.

- *RACH optimization*: The aim is to provide a sufficient number of random access opportunities for any user equipments (UEs) or mobile devices operating within the cell, by reducing the preamble detection miss probability and contention probability of the new arrivals.

- *Mobility robustness optimization*: The objectives are to detect handover (HO)-related radio link failures and to recognize an inefficient use of network resources, and to reduce HO-related failures and the inefficient use of network resources due to unnecessary or missed handovers.

- *Load balancing*: The objectives are to identify the congested areas, to cope with the unequal traffic load, and to achieve load balancing with minimum number of handovers. The basic idea is to divert traffic in one (possibly congested) area to other (non-congested) areas by adjusting the mobility parameters.

- *Coverage optimization*: The objectives are to detect coverage holes based on the analysis of coverage related parameters like call drops and failures on random access channels, and to compensate the detected coverage holes by adjusting the network control parameters such as transmission power and antenna downtilt.

- *Capacity*: The aim is to enhance capacity of the existing network through the reallocation of wireless resources and power control for the affected BSs. To accommodate the asymmetric uplink and downlink traffic with mixed service types, our interest lies in the improvement of joint uplink and downlink performance.

1.2 Approach

In this thesis, we exploit the statistical learning, detection and optimization theory principles to design the following two types of SON functionalities:

- *Cognition, learning and detection*: Functionality of a network element with which it gradually becomes aware of its surroundings, and makes accurate and robust decisions under abnormal network states.

- *Multi-objective optimization in high dimensional space*: Functionality of a network element with which it jointly optimizes interrelated or conflicting performance metrics over interacting variables in a certain SON use case, or between multiple interacting SON use cases.

The general framework composed of the Network State Classifier/Estimator/Predictor and the Network Optimizer is shown in Fig. 1.1. The former function module collects the measurements, feedback and the extracted key performance indicators (KPIs) from the network, and achieves network inference, awareness and fast detection of the network anomalies. If one or more network anomalies are detected, the Classifier/Estimator/Predictor sends a message to the optimizer to trigger the corresponding self-healing and self-optimization functionalities. The module also learns from the collected data the mathematical and statistical model of the complex network system, and further provides the model to the optimizer for the task of network performance optimization. The latter function module, i.e., the optimizer, performs individual or multiple SON functions, that are triggered by the learning

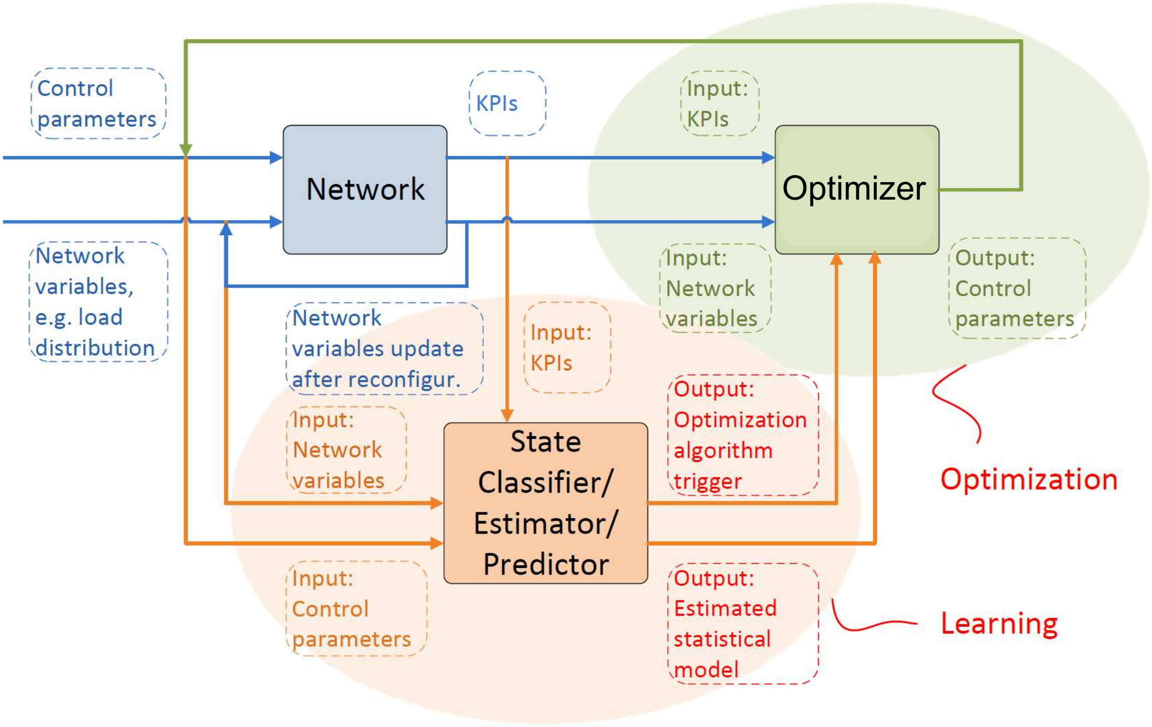

Control parameters
KPIs
Input: KPIs
Network
Optimizer
Network variables, e.g. load distribution
Input: Network variables
Output: Control parameters
Network variables update after reconfigur.
Input: KPIs
Input: Network variables
State Classifier/ Estimator/ Predictor
Output: Optimization algorithm trigger
Optimization
Input: Control parameters
Output: Estimated statistical model
Learning

Figure 1.1: Framework of learning and optimization in SON

module, by leveraging the extracted information and inferred mathematical model from the learning module.

The above mentioned self-organizing functionalities call for the development of stochastic protocols/algorithms that operate on a relatively large time scale, and therefore are based on the statistics rather than the actual information of all or some of the underlying random processes.

The work presented in this thesis provides novel ideas, mathematical models, optimization tools and related building blocks for network state inference, classification, anomaly detection, and self-optimization of multiple SON use cases. A comprehensive study covers five most challenging SON use cases: cell outage detection, RACH optimization, mobility robustness optimization, mobility load balancing, and coverage and capacity optimization. A variety of mathematical tools for modeling and optimization are developed, covering a wide range of techniques in statistics, data analysis, matrix algebra, and functional analysis.

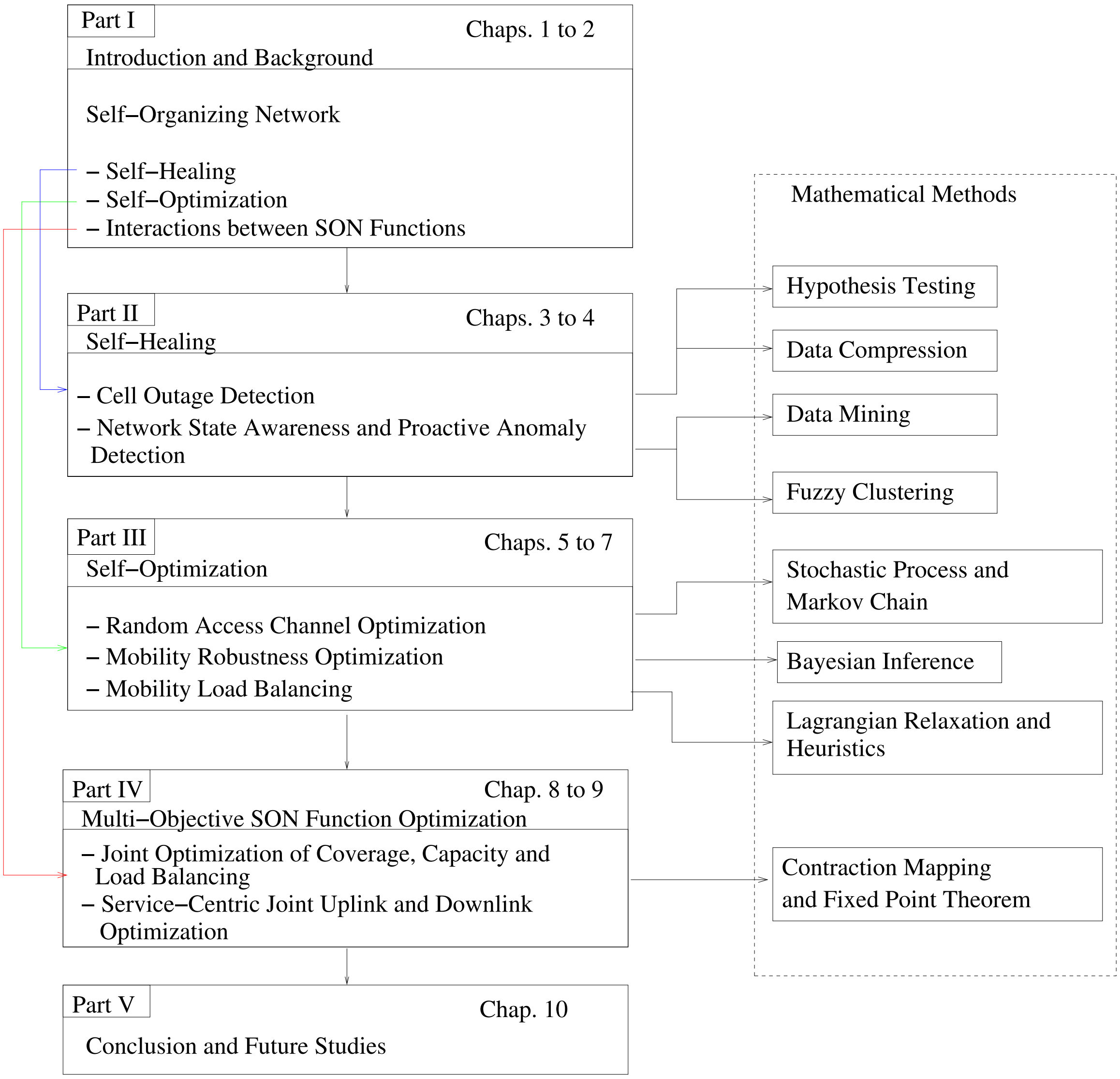

Part I
Chaps. 1 to 2
Introduction and Background
Self−Organizing Network
– Self−Healing
– Self−Optimization
– Interactions between SON Functions
Part II
Chaps. 3 to 4
Self−Healing
– Cell Outage Detection
– Network State Awareness and Proactive Anomaly Detection
Part III
Chaps. 5 to 7
Self−Optimization
– Random Access Channel Optimization
– Mobility Robustness Optimization
– Mobility Load Balancing
Part IV
Chap. 8 to 9
Multi−Objective SON Function Optimization
– Joint Optimization of Coverage, Capacity and Load Balancing
– Service−Centric Joint Uplink and Downlink Optimization
Part V
Chap. 10
Conclusion and Future Studies
Mathematical Methods
Hypothesis Testing
Data Compression
Data Mining
Fuzzy Clustering
Stochastic Process and Markov Chain
Bayesian Inference
Lagrangian Relaxation and Heuristics
Contraction Mapping and Fixed Point Theorem

Figure 1.2: Content and methodology of material

1.3 Outline and Contributions of the Thesis

Fig.1.2 shows the roadmap for this thesis, which consists of five parts, dealing with prerequisites and individual aspects of SON in particular with respect to self-healing and self-optimization. Part I provides an introduction and background knowledge on SON including the self-healing and self-optimization functions. Note that the related works and the state-of-the-art are investigated for distinct SON functionalities in each chapter respectively. In Part II we present the self healing algorithms for cell outage detection and network anomaly detection. Self optimization algorithms for use cases RACH optimization, mobility robustness optimization and mobility load balancing are presented in Part III. In Part IV we present the multi-objective optimization algorithm for joint optimizing of coverage, capacity and load balancing, and the approach for joint uplink and downlink optimization for flexible duplex-enabled fifth generation (5G) networks. The final conclusions and the outlook are presented in Part V.

This thesis begins with the introduction and background of SON. In Chapter 2, we introduce the definitions of the commonly used KPIs and network measurements according to the 3GPP standardization, as well as the objectives of the self-healing and the self-optimization functionalities. We also address the interaction and conflicts between the SON functionalities, and the challenges for joint optimization of multiple SON use cases.

Part II presents the self-healing algorithms for detecting two types of network anomalies. The first type of anomaly is usually caused by an unexpected operation fault as a rare event. Such an event is difficult to detect due to the lack of a priori knowledge. Chapter 3 presents a novel cell outage detection algorithm with *composite hypothesis testing* based on statistics and performance metrics, which enables the evolved Node B (eNB) to detect an outage of a neighbor cell, and is applicable in the lack of exact knowledge of the fault event. The second type of anomaly is caused by performance degradation, where a priori knowledge of various classes of anomalies can be found in the dataset. In Chapter 4 we propose a framework of proactive cell anomaly detection based on *dimension reduction* and *fuzzy classification techniques.* By associating the new network state to the SON use case-related clusters, we can timely detect the network anomaly and further provide guideline for self-optimizing functionalities to deal with the interaction and conflicts.

Parts of the material in this chapter were previously published in [2, 10].

Part III focuses on the optimization of individual SON use cases. Chapter 5 aims at improving RACH procedure by maximizing throughput or alternatively minimizing the user dropping rate. Protocols based on *minimization of the state-dependent stochastic drift for Markov chains* are proposed to exploit the information from measurements and user reports in order to estimate current values of the system unknowns and broadcast global action-related values to all users. Chapter 6 exploits the framework of *multivariate stochastic processes* to develop a novel method of successively choosing a sequence of multivariate training points for mobility robustness optimization (MRO). Chapter 7 suggests a novel decentralized algorithm for load balancing in the downlink based on the solution of a *mixed integer optimization problem* solved using Lagrangian - but not Linear Programming - relaxation, which allows the solution to be binary for the user assignment variables.

Parts of the material in this chapter were previously published in [3, 4, 14].

Part IV solves challenges in the joint optimization of multiple SON use cases or objectives by coordinately handling multiple control parameters. Chapter 8 aims to ensure efficient network operation by a joint optimization of coverage, capacity and load balancing based on the axiomatic framework of *standard interference functions.* To provide a service-centric network optimization, Chapter 9 proposes an optimization algorithm to jointly optimize the uplink and downlink bandwidth allocation and power control in a flexible duplex-enabled next generation wireless networks, using the *fixed point approach* for nonlinear (contraction) operators with or without monotonicity.

Parts of the material in this chapter were previously published in [15, 16].

In **Part V**, we summarize the main findings and conclusions, and discuss open research questions for future research.

Further results not included in this thesis

During my time at Fraunhofer Heinrich Hertz Institute and Bell Laboratories, we work on a broad range of problems which leverage context information to forecast the evolution of network conditions and, in turn, to improve network performance in the next generation wireless network enabled by disruptive architectures and new technologies. The following publications should be highlighted and represent a good overview of the different aspects, although they are not included in this thesis.

- *Predictive modeling for proactive optimization.* Anticipatory networking extends the idea to communication technologies by studying patterns and periodicity in human behavior and network dynamics to optimize network performance. In [17], we identify the main prediction and optimization tools adopted in this body of work and link them with objectives and constraints of the typical applications and scenarios. Understanding human mobility is an emergent research field, especially in the last few years, that has significantly benefited from the rapid proliferation of wireless devices that frequently report status and location updates. In [7, 23], we propose frameworks for predicting base station identification (ID) and staying time by using the *variable order Markov models* which includes a variety of universal lossless compression algorithms. The predicted mobility and trajectory-related context is used in [1, 5, 6, 18] to derive closed-form expressions of outage probabilities related to the events of too-early and too-late HO. By minimizing the weighted sum of the two outage probabilities, we can achieve a good trade-off between minimization of HO-related radio link failures and reduction of unnecessary HOs.

In [8,22], we develop predictive models of the physical wireless channel, i.e., the channel quality and its specific parameters, by exploiting spatial and temporal correlation in a *Bayesian framework*, so that it is possible either to take advantage of future link improvements or to counter bad conditions before they impact the system.

Despite the aforementioned works obtaining promising results for predicting lower-layer physical radio propagation-related metrics, in [9,19–21] we investigate *functional time series* prediction methods for various higher-layer performance metrics, including the transport block size, number of required physical resource blocks, and modulation and coding schemes.

- *Towards 5G technologies.* In the 5G era, besides the support for mobile broadband (MBB), the network systems should also manage machine type communications (MTC), which are mostly characterized by small packet transmissions, and have very different requirements from MBB traffic. For example, two representative use cases of MTC are massive machine communications (MMC) and mission critical communications (MCC). Handling new types of traffic has become a challenging task.

  In [12], we aim at developing a true user-centric approach that provides a flexible tradeoff between mixed types of services (where UEs generate either MBB or MCC traffic in both uplink and downlink) to meet their specific requirements in both uplink and downlink for dynamic time-division duplex (TDD) systems. The formulation of a *convex optimization problem* takes into consideration the individual requirements of each single user in terms of sustainable latency and desired throughput, thus implementing a real user-centric scheduling approach to jointly optimize: *a)* the duplexing mode, i.e., either downlink or uplink, *b)* the transmission time interval (TTI) length, and *c)* the UEs to be served and the resources allocated in each TTI.

  In [11, 13], we deal with the always-on applications and MTC which generate new types of background traffic, being more sporadic in nature. In [13], we analyze the tradeoff between the connected and idle states with respect to energy consumption and signaling cost, and develop a closed-form mathematical model of state transition process, based on the framework of *alternating renewal process.* The novel concept of user-centric mobility tracking area is proposed in [11], to minimize the core network signaling related to connection transitions, paging and handover.

A complete list of all publications can be found in the appendix.

Copyright Information

Parts of this thesis have already been published as journal articles and in conference and workshop proceedings as listed in the publication list in the appendix. These parts, which are, up to minor modifications, identical with the corresponding scientific publication, are ©2011-2016 IEEE.

Chapter 2

Background

SON is essential for today's complicated cellular networks to configure, organize, optimize performance, and to provide self healing capabilities when faults occur. The main functionality of SON includes: self-configuration, self-optimization and self-healing [3GPa]. Self-configuration is defined as the process of automatic installation and configuration of the newly deployed nodes. Self-optimization collects measurements and KPIs to auto-tune the control parameters to optimize the network performance. The features of self-healing include automatic detection and removal of failures and automatic adjustment of configuration parameters. In the following we introduce the concepts of KPIs and SON functionalities, and the interactions and conflicts between SON functionalities.

2.1 Key Performance Indicators and Network Measurements

The inference of the network states, anomaly detection and self-optimization are based on the extracted knowledge from the KPIs and reported measurements. Various KPIs are defined to describe the *accessibility*, *retainability*, *integrity*, *availability*, and *mobility* of the network [3GPb, 3GPc]. Network measurements, on the other hand, indicate the network environment, including the radio propagation environment and network traffic. The KPIs are collected at different planes or interfaces and the measurements are measured and reported at UE or eNB.

2.1.1 Control Plane KPIs

Accessibility KPIs measures the probability whether services requested by a user can be accessed within specified tolerances in the given operating conditions. One of the main procedures for accessibility KPIs is the radio resource control (RRC) connection. RRC setup success rate (RRCS_SR) can be calculated for service or signaling respectively, using the formula

$$\text{RRCS\_SR} := \frac{\#\text{RRC Connection Success}}{\#\text{RRC Connection Attempt}} \times 100\%, \tag{2.1}$$

where the symbol # denotes “the number of”hereafter. Other important accessibility KPIs are E-UTRAN radio access bearer setup success rate (ERAB_SR) and call setup success rate (CS_SR). Note that here the E-UTRAN radio access bearer (E-RAB) includes both the E-RAB radio bearer and S1 bearer.

Retainability KPIs are used to evaluate the network capability of retaining services requested by a user for a desired duration once the user is connected to the services. One example is the call drop rate (CDR) for voice over IP (VoIP). Any abnormal release on E-RAB causes call drop and is counted into the CDR, given by

$$\text{VoIP\_CDR} := \frac{\#\text{VoIP ERAB Abnormal Release}}{\#\text{VoIP ERAB Release}} \times 100\%. \tag{2.2}$$

Excluding CDR of VoIP, the retainability KPIs also include CDR for other data service.

Mobility KPIs are crucial for the user's experience. The metrics indicating the frequency of HOs are defined based on the HO types: intra-frequency, inter-frequency, and inter-radio access technology (RAT). The handover (outgoing) success rate (HOO_SR) and handover (incoming) success rate (HOI_SR) are defined as

$$\text{HOO\_SR} := \frac{\#\text{Outgoing HO Success}}{\#\text{Outgoing HO Attempt}} \times 100\% \tag{2.3}$$

$$\text{HOI\_SR} := \frac{\#\text{Incoming HO Success}}{\#\text{Incoming HO Attempt}} \times 100\%. \tag{2.4}$$

Note that HOO_SR can be defined for different types of inter-RAT HO.

The metric of handover ping-pong rate (HO_PPR) indicates the level of redundancy of the handover event based on the counting of ping-pong handovers. Ping-pong handover is a potentially undesirable phenomenon, in which the terminal performs frequent handovers between the same pair of cells back and forth. We define the HO_PPR as

$$\text{HO\_PPR} := \frac{\#\text{Ping-Pong HO}}{\#\text{Total HO Success}} \times 100\%. \tag{2.5}$$

Availability KPIs indicate the radio network availability rate. One possible KPI is the call blocking rate (CBR), provided by

$$\text{CBR} := \frac{\#\text{Call Requests} - \#\text{Admitted Request}}{\#\text{Call Requests}} \times 100\%. \tag{2.6}$$

2.1.2 User Plane KPIs

Integrity KPIs indicate the service quality provided to the end user. For example, service average throughput (SAT) and service maximum throughput (SMT) (in kbit/s) are defined for uplink (UL) and downlink (DL), and, for each QoS class identifier (QCI), respectively.

2.1.3 X2 Interface KPIs

The utilization of the resource is evaluated by load per cell. We define load as the resource block (RB) utilization rate, defined by

$$\text{Load} := \frac{\#\text{Occupied RB}}{\#\text{Available RB}} \times 100\%. \tag{2.7}$$

2.1.4 UE Measurements

In long-term evolution (LTE) or beyond radio networks, UE reports the measurements based on the reference signal for various scheme of decision making, for example, cell selection, power control and handover decisions. The most common measurements are given below.

UE sends reports of RRC measurement including reference signal received power (RSRP) in a binned format ranging from −140 to −44 dBm with 1 dBm resolution.

Unlike RSRP, which is the absolute received strength of the reference radio signals, reference signal received quality (RSRQ) is the signal-to-noise ratio. Both of them can be used as the criterion for initial cell selection or handover. RSRQ is defined from −19.5 to −3 dB with 0.5 dB resolution.

The calculation of RSRQ follows:

$$\text{RSRQ} = 10 \lg \frac{\#\text{RB} \times \text{RSRP}}{\text{RSSI}}, \tag{2.8}$$

where lg denotes the common logarithm of base 10, and received signal strength indication (RSSI) is the DL noise level measured at the UE's radio receiver antenna.

CQI is an indicator carrying the information on how good/bad the quality of communication channel is. In LTE, 15 values of CQI are defined, ranging from 1 to 15. The mapping between CQI and modulation scheme (including QPSK, 16QAM and 64QAM), code rate, and transport block size (TBS) is defined in [3GPe].

2.1.5 ENB Measurements

The traffic KPIs measured at eNB indicate the density of the users, including DL/UL traffic volume, average number of users, and maximum number of users.

2.2 SON Functionalities

2.2.1 Self-Healing Functionalities

The self-healing aims at solving or mitigating the faults which could be solved automatically by triggering appropriate recovery actions. The major functionality of self-healing is to monitor the network states and to detect the anomalies, especially the cell outage [3GPd].

- *Cell outage detection and compensation.* In the cell outage scenario, where there is a loss of total radio services in the outage cell, all the UEs cannot establish or maintain any of the radio bearers via that particular cell, i.e., all the UEs cannot establish the RRC connection in the outage cell. The objective is to timely detect the problem of cell outage and to detect the best set of cells that can compensate for the cell outage. The possible parameters to be optimized are the antenna tilt and downlink transmission power of the neighboring cells.

2.2.2 Self-Optimizing Functionalities

There are nine self-optimizing use cases defined in [3GPa]: coverage and capacity optimization (CCO), energy savings (ES), interference reductions (IR), automated configuration of physical cell identity (ACPCI), MRO, mobility load balancing optimization (MLBO), RACH, automatic neighbor relation function (ANRF) and inter-cell interference coordination (ICIC). In this thesis we focus on the functionalities related to the following topics: RACH optimization, mobility load balancing, interference reduction, mobility robustness optimization and coverage and capacity optimization.

- *RACH optimization.* RACH is an uplink unsynchronized channel, used for initial access or uplink synchronization. Random Access performance influences the call setup delay, handover delay, data resuming delay, call setup success rate and handover success rate. The objectives are reducing the delay and increasing the success rate.

- *Load balancing.* This use case aims at identifying the congested areas and achieving load balancing with fair interference distribution and minimum number of handovers. Algorithms need to be designed to adjust the distribution of the load by tuning the handover and/or cell reselection parameters such as time-to-trigger (TTT), cell individual offset (CIO) and hysteresis.

- *Interference reduction.* Capacity can be enhanced through interference reduction by switching off those cells which are not needed at some point of time, in particular home eNBs when the user is not at home. Possible solutions are automatic activation and deactivation of cells.

- *Mobility robustness optimization.* Updating the mobility related parameters after the initial deployment is too costly. The objectives are listed as following: 1) to detect handover-related radio link failures (too late or too early) and to recognize an inefficient use of network resources, and 2) to minimize the unnecessary handovers which cause a waste of resource. Possible parameters to be optimized are the handover-related parameters TTT, CIO and hysteresis.

- *Coverage and capacity optimization.* Two main objectives are: *1)* compensating the detected weak coverage region and providing optimal coverage, and *2)* enhancing the capacity of the network. While coverage optimization has higher priority than capacity optimization, the trade-off between the two is also a challenge in the optimization. The outputs of the optimization function may include the antenna tilt and downlink transmission (Tx) power.

The detection of the network anomalies related to different SON functionalities are based on a set of the KPIs and measurements, while the automatic optimization of the network is performed by tuning a set of control parameters. Table 2.1 illustrates the SON functionalities and their corresponding crucial KPIs and possible control parameters.

2.3 Interactions between SON Functionalities

From Table 2.1, we can observe strong interactions and dependencies between the SON functionalities. These interactions or dependencies can be categorized in the following three types:

- *Trigger.* The first functionality triggers other functionalities that do not need to be coordinated. In Fig. 2.1, Algorithm A adjusts Control Parameter 2, which influences KPI 1, 2 and 3, and then triggers Algorithm B as a "side effect". An example is that triggering algorithm for CCO requires optimization of the control parameters DL Tx power or/and antenna tilt, which may lead to unbalanced load, or too-early (or too-late) handover problem, and may further trigger algorithms for MLBO or/and MRO.

- *Co-operate.* The degradation of the same set of KPIs triggers multiple functionalities, that need to coordinate with each other. In Fig. 2.1, degradation of KPI 2 may trigger both Algorithm A and B. The challenge is how to coordinate both functionalities to maximize the desired performance metrics without decreasing the others. For example, increase of CDR may trigger both CCO and MRO, because the radio link failure could be caused by either poor coverage at the cell edge, or the inappropriate configuration of handover parameters. The objective is to enhance the coverage, while still satisfying the requirements of the mobility-related KPIs.

- *Co-act.* Different functionalities require to optimize the same set of control parameters, which may lead to continuously conflicting actions. For instance, both Algorithm A and B in Fig. 2.1 optimize Parameter 2. Coordination between the two functionalities is needed to avoid conflicting outputs. In practical system, Table 2.1 shows that

DL Tx power and antenna tilt are possible solutions for multiple SON functionalities, including CCO and MLBO. Thus, coordination between these functionalities is required.

So far, the SON functionalities have not been sufficiently studied in the context of the possibly conflicting objectives and solutions, which may lead to network instability. A challenging task is to resolve the conflicts over the control parameters and the KPIs.

- *Conflicts over control parameters.* Multiple SON functionalities have conflicting outputs of optimization parameters, as described in the case of "co-act".
- *Conflicts over KPIs.* Degradation of the same set of KPIs triggers several functionalities as mentioned in the case of "trigger". Another scenario is that multiple functionalities tend to optimize the same performance metric, and cause inconsistent change of the metric, as shown in the case of "co-operate".

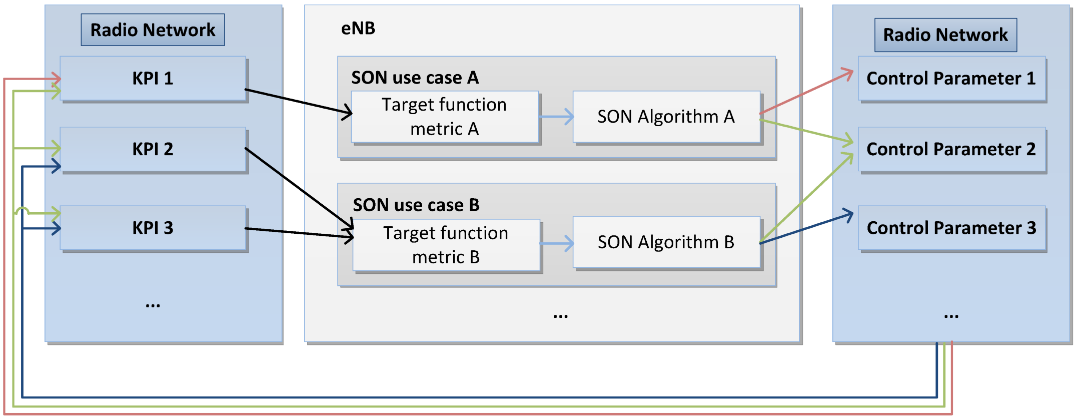


Figure 2.1: Interactions and dependencies between SON functionalities

Table 2.1: SON FUNCTIONALITIES AND CORRESPONDING PARAMETERS

Functionality	KPI and Measurement	Control Parameter
RACH optimization	· Success rate · Drop rate · Detection miss rate · HOI_SR	· Tx power · Backoff probability · Preamble allocation
Cell outage detection and compensation	· HOI_SR · RRCS_SR · CS_SR · ERAB_SR · DL/UL traffic volume · Average/maximum number of users · RSRP/RSRQ distribution	· Antenna tilt · Tx power
Coverage and capacity optimization	· VoIP_CDR · Data service CDR · UL/DL SAT · UL/DL SMT · RSRP/RSRQ distribution	· Tx power · Antenna tilt · Beamforming parameters
Mobility load balancing	· Load · CBR · SAT · RRCS_SR · DL/UL traffic volume · Average/maximum number of users	· TTT · CIO · Hysteresis · Tx power · Antenna tilt
Interference reduction	· SAT · SMT · Load · RSRP/RSRQ distribution	· BS on/off · Tx power
Mobility robustness optimization	· HOI_SR · HOO_SR · HO_PPR · VoIP_CDR · Data service CDR	· TTT · CIO · Hysteresis

Part II

Self-Healing

Chapter 3

Cell Outage Detection with Composite Hypothesis Testing

In this chapter we present a novel cell outage detection algorithm based on statistics and performance metrics, which enables an eNB to detect an outage of a neighbor cell. The algorithm is a weighted combination of three hypothesis tests based on: 1) the distribution of the CQI, 2) the time correlation of the CQI differential, and 3) the registration request (RRQ) frequency. The weights of the combined test are functions of the predicted traffic load in neighboring cells, which is motivated by the fact that the reliability of an individual test depends on the load state. To detect the change-point in the CQI distribution, we use an efficient discriminant function related to the "universal code" proposed by [Ziv88], which can be shown to be asymptotically optimal in the sense of the modified *Neyman-Pearson criterion*. The simulation results indicate that the proposed algorithm can detect the outage problem in a real-time and reliable manner.

Parts of this chapter have already been published in [2].

3.1 Motivation

Reliability and disposability are ones of the most important requirements in SONs. In this work we focus on the challenge of detecting a cell outage, which covers for instance the detection of sleeping cells or poor service in a cell caused by hardware and software failures, or external failures such as electrical power outage. Although a great deal of effort has been spent, the problem remains to design fast and robust cell outage detection algorithms. When developing such algorithms, a system designer faces several inherent challenges including:

- (*Universality*) A cell outage is usually caused by an unexpected operation fault that is a rare event.

- (*Detectability*) It may take too long for a base station to realize that there is a service outage in its cell.
- (*Separability*) It is in general difficult to separate a cell outage from other faulty events.

As far as universality is concerned, we need an algorithm that efficiently solves the hypothesis testing problem when at least one of the probability measures is unknown. Such problems are classified as composite hypothesis testing (CHT) problems, to which Bayesian or conventional hypothesis testing methods are not directly applicable because of the lack of a priori probability for faulty states. In this paper, we apply a promising CHT method, under which the abnormal state is reliably detected even if a priori knowledge of the fault state is not known. The CHT method involves an application of the universal code length function into the discriminant function. (Fast) detectability can be achieved by means of detection algorithms performed in a distributed fashion by neighbor cells. The detectability process involves the identification of a cell in outage. Finally, in order to separate a cell outage event from other events, we combine three hypothesis tests to delineate the outage in time and space.

Notice that there are three main observations at a base station if a neighbor cell is in outage: 1) The CQI distribution (especially that of cell edge users) changes due to a change in the interference structure, 2) the time correlation of the CQI differential increases, and 3) the frequency of RRQ connection reestablishment requests from users of an outage cell increases. Our algorithm is a weighted sum of hypothesis tests based on the three observations, where the weights depend on the predicted load of a neighbor cell to take into account the fact that the reliability of each test depends on the cell load. Each eNB learns its load profile and exchanges it with its neighbor cells, which in turn allows the cell to estimate the load of its neighbor cells.

3.2 Problem Statement

Consider a sequence $X_n = (x_i)_{i=1}^n \in \mathcal{A}^n$ with each x_i in a finite set $\mathcal{A}$. The sequence obeys one of the two statistical hypotheses

$$H_0 : x_i \sim P_0, i = 1, 2, \dots, n,$$
$$H_1 : x_i \sim P_1, i = 1, 2, \dots, n,$$

where P_0 and P_1 are two distinct probability distributions. We assume that $P_i, i = 0, 1$ belongs to a family of ergodic probability measures $\mathcal{P}$, which includes all finite stationary ergodic Markov processes of a finite order. Given an observation X_n, the problem is that of deciding whether its underlying source is P_0 or P_1. Let $P_j(X_n), j = 0, 1$, denote the probability of a sequence X_n under P_j. A decision rule Λ_n is a set of sequence X_n such that if $X_n \in \Lambda_n$, then X_n is classified under the distribution P_1 (faulty state), otherwise under the distribution P_0 (healthy state). There are two types of errors:

- Type I (false alarm): Let $P_0(\Lambda_n)$ denote the probability of deciding $P = P_1$ while P_0 is true.
- Type II (misdetection): Let $P_1(\overline{\Lambda}_n)$ denote the probability of deciding $P = P_0$ while P_1 is true, where $\overline{\Lambda}_n$ denotes the complement of Λ_n.

Due to the trade-off between the two error probabilities, the objective is to minimize one of them while constraining the other. If both P_0 and P_1 were known, then the *Neyman-Pearson* lemma would provide an optimality criterion on decision rule Λ_n that minimizes $P_1(\overline{\Lambda}_n)$ under the condition $P_0(\Lambda_n) \leq 2^{-\lambda n}$ for a given $\lambda > 0$ [CT91, pp. 305-306]. The optimum test is called likelihood ratio test (LRT) and the optimal decision rule is given by [CT91, pp. 304-309]

$$\Lambda_n^* = \left\{ X_n : \frac{1}{n} \log P_1(X_n) - \frac{1}{n} \log P_0(X_n) \geq T_n(\lambda) \right\}, \tag{3.1}$$

where log denotes the binary logarithm of base 2, $T_n(\lambda)$ is a threshold function, depending on n, λ, P_0 and P_1. However, in contrast to P_0 can be learned, P_1 usually remains unknown because the cell outage events are rare and in-expectable. In this work, therefore, we assume that P_1 belongs to the family $\mathcal{P}$, the hypothesis test is then $P = P_0$ against a composite alternative $P \in \mathcal{P}$. Since the LRT is not applicable in this case, Hoeffding [Hoe65] first formulated the problem by giving a generalized *Neyman-Pearson* criterion (for details see Appendix C.1.1), which follows

Problem 3.1. *Among all decision rules $\{\Lambda_n\}_{n\geq 1}$ independent of the unknown P_1, the problem is how to select a rule such that the type II error exponent $-\limsup_{n\to\infty} \frac{1}{n} \log P_1(\overline{\Lambda}_n)$ is maximized under the condition*

$$-\limsup_{n\to\infty} \frac{1}{n} \log P_0(\Lambda_n) > \lambda. \tag{3.2}$$

Note that Condition (3.2) means that the type I error exponent must be above some predefined threshold $\lambda > 0$.

3.3 Optimal Tests

Hoeffding [Hoe65] provided an optimal decision test that satisfies the criterion (3.2), by proving that a set of sequences whose Kullback-Leibler (KL) divergence from the healthy state hypothesis distribution P_0 is larger than λ defines an optimal set of hypothesis tests. In this section, we briefly describe the approach of [Ziv88], which simplified the practical implementation of Hoeffding's test by using the *Lempel-Ziv* algorithm.

Let the decision rule Λ_n be determined by a function $h : \mathcal{A}^n \to \mathbb{R}$ such that $\Lambda_n = \{x : h(x) > 0\}$, where $x \triangleq X_n$ for ease of notation. This function is called the *discriminant function* [Han81]. As P_0 can be estimated by the training samples but P_1 is unknown, the discriminant function depends only on $P_0(\cdot)$ and x, and is of the form

$$h(x, \lambda) = \frac{1}{n}\left(-\log P_0(x) - u(x)\right) - \lambda\,. \tag{3.3}$$

Here and hereafter, λ is a predefined threshold, and $u(x)$ is the length function of a universal code. Note that a code $c(x)$ of x is a mapping from $\mathcal{A}^n$ to a set of the binary sequences and the length function $u(x)$ has to satisfy the *Kraft's inequality*: $\sum_{x \in \mathcal{A}^n} 2^{-u(x)} \leq 1$ [CT91, p. 82]. Roughly speaking, a code is said to be *universal* for the family $\mathcal{P}$ if, for any source with probability measure $P \in \mathcal{P}$, the average code length converges to the entropy of P as n tends to infinity [Ziv88] (for details see Appendix C.1.2).

The following theorem for optimal discriminant function $h(x) = h(x, \lambda)$ is proved in [Ziv88] by exploiting *Kraft inequality* and the properties of universal codes with respect to the length function $u(x)$.

Theorem 3.1 ([Ziv88]). *Let $D(P_1||P_0)$ denote the KL divergence between two probability distributions P_1 and P_0 [CT91], and let $u(x)$ be the length function of a universal code for class $\mathcal{P}$. We define*

$$h(x, \lambda) \triangleq \frac{1}{n}\left(-\log P_0(x) - u(x)\right) - \lambda. \tag{3.4}$$

For every $P_0(\cdot)$, $P_1(\cdot) \in \mathcal{P}$, the type I error is then constrained by

$$P_0\left(h(x, \lambda) > 0\right) \leq 2^{-\lambda n}, \tag{3.5}$$

and the successful detection probability satisfies

$$\lim_{n \to \infty} P_1\left(h(x, \lambda) > 0\right) \geq 1 - \epsilon \tag{3.6}$$

for $0 \leq \epsilon < 1$ whenever

$$D(P_1 \| P_0) > \lambda. \tag{3.7}$$

3.4 System Model

In what follows, $\mathcal{U}_m$ is a set of UEs in active mode served by eNB m, $\mathcal{S}_m$ denotes a set of neighbor cells $s \in \mathcal{S}_m, s \neq m$ of cell m, $\mathcal{E}_m$ is the class of cell edge UEs served by eNB m, and $\mathcal{V}_m$ is used to denote a set of UEs which provides statistics for detection algorithm at eNB m. In a special case, we have $\mathcal{V}_m = \mathcal{U}_m$ or $\mathcal{V}_m = \mathcal{E}_m$. In this study, we consider a cellular wireless network, in which each eNB, say eNB m, collects CQI reports from UEs $i \in \mathcal{V}_m$ and the number of RRQs periodically. The report intervals are labeled by $n, l, r \in \mathbb{N}_+$ and are assumed to be larger than the channel coherence time. We use $t, \tau \in \mathbb{R}$ to denote the continuous time, while t_n is the time point at which the n-th interval ends. Therefore, the nth report interval corresponds to the measurements at time t with $t_{n-1} < t < t_n$. Furthermore, we assume that for every new RRQ, the ID of the preceding cell is known. ENBs cooperate in the sense that they learn their traffic load profiles per weekday and exchange them with the neighboring eNBs. The cell outage detection algorithm has a decision latitude of M report intervals and is based on the measurements and statistics of CQI reports, RRQs and traffic loads, which is discussed in the following.

3.4.1 Statistics Relevant to CQI Reports

CQI is a mapping from the signal-to-interference-plus-noise ratio (SINR) observed by a user to an N-bit integer (e.g., $N = 4$ for LTE system). In our setting, in time interval n, each user $i \in \mathcal{V}_m(n)$ reports its current CQI $Q_i(n)$ to the serving eNB. These reports are collected for a sufficiently long window of W time intervals $Q_i^W(n) = (Q_i(l))_{l=n-W+1}^{n}$ to generate a histogram Q_i at the n-th time interval, which serves as the baseline (healthy state) distribution. We drop the time index for brevity and use $H_i^q \equiv H_i^q(n) = (H_{i,1}^q(n), \ldots, H_{i,2^N}^q(n))$ to denote the histogram. Throughout the work, it is assumed that if $H_{i,j} = 0$, then $H_{i,j} = H_{i,j} + \epsilon$ for some sufficiently small $\epsilon \ll 1$. Finally, the histogram is normalized to yield $\sum_{j=1}^{2^N} H_{i,j}^q = 1$.

Instead of computing an individual histogram for each user, we can alternatively consider a weighted sum of the CQIs reported by all users

$$Q_\Sigma(n) = \sum_{i \in \mathcal{V}_m(n)} \alpha_i Q_i(n) \tag{3.8}$$

where $\alpha_i \geq 0$ is a weight of user i with $\sum_{i \in \mathcal{V}_m} \alpha_i = 1$. It reflects the relevance of a user in the sense that larger weights are assigned to cell edge users since the inter-cell interference is expected to have the strongest impact on the CQIs of such users. The histogram of Q_Σ is denoted as $H^q = (H_1^q, \ldots, H_{2^N}^q)$.

Finally, we consider the CQI differential of user i defined to be $dQ_i(n) = Q_i(n) - Q_i(n-1)$. We capture the time correlation of the CQI differential by

$$Cor(n) = \sum_{i \in \mathcal{V}_m(n)} \sum_{j \in \mathcal{V}_m(n), j \neq i} dQ_i(n) dQ_j(n), \tag{3.9}$$

and let $H^c = (H^c_1, \ldots, H^c_{2^N})$ be the histogram of Cor. An alternative to (3.9) is to consider the histogram of

$$Cor(n) = \sum_{i \in \mathcal{V}_m(n)} dQ_i(n). \tag{3.10}$$

An example of the histograms of CQI and CQI differential are shown in 3.1.

3.4.2 Statistics Relevant to RRQs

A user i that has been served by s sends a RRQ to cell m if the connection to s is lost and the user requires a handover to m. We defined the RRQ frequency to be

$$df_s(n) = \frac{1}{n - X + \delta} \tag{3.11}$$

where X is the time index of the last RRQ, $\delta \ll 1$ is a parameter used to avoid zero in the denominator. The corresponding histogram is then $H^f_s = (H^f_{1,s}, \ldots, H^f_{2^N,s})$.

Alternatively, we can use the average number of RRQs per time interval, which is calculated by averaging the number of RRQs over a short window of w intervals

$$A_s(n) = \sum_{l=0}^{w-1} a_s(n-l) \tag{3.12}$$

where $a_s(n)$ is the number of RRQs from neighboring cell s at time n. The histogram of A_s is denoted by $H^A_s = (H^A_{1,s}, \ldots, H^A_{2^N,s})$.

3.4.3 Statistics Relevant to Traffic Load

Each eNB learns its daily traffic load profile by averaging the load measurements from a number of week samples, and exchanges the profile with its neighbor cells, so that a cell can predict the load of any neighbor cell with exchanged profiles. Define the load of the j-th week sample of the d-th weekday in cell s, where $1 < d < 7$, as follows.

$$G^j_{s,d}(t) = \frac{1}{T} \int_{t-T}^{t} L^j_{s,d}(\tau) d\tau \tag{3.13}$$

where T is either some time window or decision latitude (M report intervals). $L(\tau)$ is the actual load at time τ. The load profile for d-th weekday in cell s is given by

$$\overline{G}_{s,d}(t) = \frac{1}{J} \sum_{j=1}^{J} G^j_{s,d}(t). \tag{3.14}$$

An example of a load profile is shown in Fig. 3.2.

3.5 Algorithm

We propose a cell outage detection algorithm as a weighted combination of three hypotheses based on: 1) the distribution of CQI, 2) the time correlation of CQI differential, and 3) the RRQ frequency. The weight of each discriminant function is calculated by a function of load, considering that the performance of each individual test depends on the load. In the following we present each individual test separately to present the final combined test in the last subsection.

3.5.1 Hypothesis Test on Distribution of CQI

This test is designed for early warning of changes in the distribution of CQI caused by neighbor cell outage. The approach introduced in Section 3.3 is applicable because: 1) The CQI values are taken from a finite set $\{1, 2, \ldots, 2^N\}$, 2) Although the CQI distribution under faulty state P_1 is not known, we can still assume that it belongs to a family of distributions $\mathcal{P}$, where $P_0, P_1 \in \mathcal{P}$. The decision latitude is M report intervals, $M \ll W$, where W is a long window to learn the histogram.

Denote $Q_i^M(n) = (Q_i(l))_{l=n-M+1}^n$ as the CQI reports of user i in the last M intervals. The discriminant function (3.3) for user i takes then the following form

$$h\left(Q_i^M(n), \lambda_i\right) = \frac{1}{M}\left(-\log P\left(Q_i^M(n)\right) - u_i\left(Q_i^M(n)\right)\right) - \lambda_i \tag{3.15}$$

where $P\left(Q_i^M(n)\right)$ is given by

$$P\left(Q_i^M(n)\right) = \prod_{l=n-M+1}^{n} H_{i,Q_i(l)}^q. \tag{3.16}$$

The second term $u_i\left(Q_i^M(n)\right)$ on the right-hand side of (3.15) is the length of a universal code of the sequence $Q_i^M(n)$. In this work we use the code introduced by Davisson [Dav73], inspired by *Lempel-Ziv* coding scheme [ZL78]. The calculation of the length function, which is based on finding the recurrence relations among the blocks, is provided in Appendix C.1.2. Assuming that M is divisible by $B \in \mathbb{N}_+$, the code length function $u_i\left(Q_i^M(n)\right)$ can be written as follows

$$u_i\left(Q_i^M(n)\right) = -\sum_{r=n-M+1}^{n-B} v_{i,r}\left(Q_i^M(n)\right)\log\left(v_{i,r}\left(Q_i^M(n)\right)\right) + \gamma^B \log(M/B+1) \tag{3.17}$$

where the parameter γ satisfies $\gamma \leq 2^{\lambda M}$ to keep the discriminant function optimal [Ziv88], and

$$v_{i,r}\left(Q_i^M(n)\right) = \sum_{m=1}^{M/B} \mathbf{1}\left\{(Q_i(l))_{l=r}^{r+B-1} = (Q_i(l))_{l=r+mB \bmod M}^{r+(m+1)B-1 \bmod M}\right\}. \tag{3.18}$$

The last term λ_i in (3.15) is chosen to fulfill (3.7), but it is emphasized that the divergence cannot be derived since P_1 is not known. Therefore we use instead the negative entropy of the histogram H_i^q, which is a tighter upper bound of λ_i

$$\lambda_i \leq \sum_{j=1}^{2^N} H_{i,j}^q \log H_{i,j}^q. \tag{3.19}$$

Now using the discriminant function $h(Q_i^M(n), \lambda_i)$ and the weights α_i of users defined by (3.8), the hypothesis test on CQI distribution becomes

$$\mathcal{H}_1 = 1 \text{ if } \sum_{i \in \mathcal{V}_m(n)} \alpha_i h\left(Q_i^M(n), \lambda_i\right) > 0. \tag{3.20}$$

An alternative is to use (3.8) instead of the individual $Q_i(n)$ to simplify the algorithm. In this case we formulate the discriminant function $h(Q_\Sigma^M(n), \lambda_\Sigma)$ in an analog way, and the hypothesis test is given by

$$\mathcal{H}_1 = 1 \quad \text{if} \quad h\left(Q_\Sigma^M(n), \lambda_\Sigma\right) > 0. \tag{3.21}$$

3.5.2 Hypothesis Test on Time Correlation of CQI Differential

Another symptom of neighbor cell outage is a high correlation among CQI differentials of different users, because a global influence on the CQI change is with high probability caused by a neighbor cell outage. Let the arithmetic mean of $Cor^M(n) = (Cor(l))_{l=n-M+1}^n$ be denoted by $\overline{Cor^M}(n)$, which is the average correlation among CQI differentials of different users over the last M time interval. The discriminant function $h\left(\overline{Cor^M}(n), X_c\right)$ to detect a high correlation, constraining to a small type I error probability X_c, is chosen to be

$$h\left(\overline{Cor^M}(n), X_c\right) = |\overline{Cor^M}(n) - E^c| - \sqrt{\frac{Var^c}{X_c}} \tag{3.22}$$

where $E^c = \sum_{j=1}^{2^N} jH_j^c$ is the expectation of Cor and $Vac^c = \sum_{j=1}^{2^N}(j - E^c)^2 H_j^c$ is its variance.

With this discriminant function in hand, the hypothesis test on time correlation of CQI differential takes the form

$$\mathcal{H}_2 = 1 \quad \text{if} \quad h\left(\overline{Cor^M}(n), X_c\right) > 0. \tag{3.23}$$

It is easy to show by using the *Chebyshev* bound that this test satisfies the constraint on the type I error probability

$$P_0\left(h\left(\overline{Cor^M}(n), X_c\right) > 0\right) \leq X_c. \tag{3.24}$$

3.5.3 Hypothesis Test on RRQ Frequency

An obvious indicator of a neighbor cell outage is an increase of the frequency of RRQs received by an affected cell. Denote the RRQ frequency from a neighbor cell $s \in \mathcal{S}_m$ in the last M intervals by $df_s^M(n) = (df_s(l))_{l=n-M+1}^n$, with the arithmetic mean $\overline{df_s^M}(n)$, and let the discriminant function be defined as

$$h\left(\overline{df_s^M}(n), X_f\right) = |\overline{df_s^M}(n) - E_s^f| - \sqrt{\frac{Var_s^f}{X_f}} \tag{3.25}$$

where X_f is the threshold for type I error probability, $E_s^f = \sum_{j=1}^{2^N} jH_{j,s}^f$ is the expectation of df_s, and $Vac_s^f = \sum_{j=1}^{2^N}(j - E_s^f)^2 H_{j,s}^f$ is the variance. An application of the *Chebyshev* bound shows that the type I error probability is constrained by

$$P_0\left(h\left(\overline{df_s^M}(n), X_f\right) > 0\right) \leq X_f. \tag{3.26}$$

The hypothesis on RRQ frequency is therefore given by

$$\begin{aligned} \mathcal{H}_3 = 1 \text{ if } \max_{s \in \mathcal{S}_m} h\left(\overline{df_s^M}(n), X_f\right) > 0, \\ \text{and } s^* = \arg\max_{s \in \mathcal{S}_m} h\left(\overline{df_s^M}(n), X_f\right), \end{aligned} \tag{3.27}$$

where s^* is the detected outage cell.

3.5.4 Combination of Hypothesis Tests

The decision on cell outage is made based on a hypothesis test that is a combination of the hypothesis tests introduced in Sections 3.5.1,3.5.2, and 3.5.3. We formulate the test running in eNB m on the d-th workday at time t_n as follows.

$$\mathcal{H}(d, t_n) = 1 \text{ if } \max_{s \in \mathcal{S}_m} \mathcal{H}_s(d, t_n) > 0, \tag{3.28}$$

$$\text{and } s^* = \arg\max_{s \in \mathcal{S}_m} \mathcal{H}_s(d, t_n), \tag{3.29}$$

$$\mathcal{H}_s(d, t_n) = \max_{s \in \mathcal{S}_m} \left(\frac{\beta\left(\overline{G}_{s,d}(t_n)\right)}{2} \mathcal{H}_1 + \frac{\beta\left(\overline{G}_{s,d}(t_n)\right)}{2} \mathcal{H}_2 + \left(1 - \beta\left(\overline{G}_{s,d}(t_n)\right)\right) \mathcal{H}_3^s \right). \tag{3.30}$$

In (3.30) the weight β is a monotone decreasing function of predicted traffic load $\overline{G}_{s,d}(t_n)$ to take into account the fact that the reliability of each test depends on the load state. Accordingly, the tests on CQI statistics prevail if the cell s is predicted to be lightly loaded. Since then the changes in RRQs is not significant and the test result on RRQs frequency may not be reliable. In contrast, CQI statistics still provide enough information to make reliable decisions (especially on the cell edge), because a neighbor cell outage definitely affects the interference structure in the observing cell. On the other hand, in case of a heavily loaded cell, the RRQs test is more reliable than the CQI statistic tests due to a large number of RRQs. A reasonable choice of the weight function is the normalized erfc function. We define the average load of cell s at d-th workday to be

$$G_{s,d} = \frac{1}{24h} \int_{\tau=0}^{24h} \overline{G}_{s,d}(\tau) d\tau \tag{3.31}$$

while the weight function takes the form

$$\beta(\overline{G}_{s,d}(t)) = \max\left(0.33, \frac{1}{2}\,\mathrm{erfc}\left(\frac{\overline{G}_{s,d}(t) - G_{s,d}}{\sigma^2}\right)\right) \tag{3.32}$$

where σ is a tunable parameter to choose the sensitivity of the influence of load. As shown in Fig.3.3, a small value of σ allows the algorithm to take a radical choice of weight (either 1 or 0.33) easily by deviating a little from the mean $G_{s,d}$. And a large value of σ allows a smooth evolution of the weight function. If the load is zero, the RRQ frequency test does not play a role because then the weight of $\mathcal{H}_3$ is zero.

3.6 Numerical Results

Simulations are done by implementing the algorithm into a LTE simulation environment consisted of 19 regular hexagonal sites. The CQI and RRQ reports are updated per second and the decision latitude M is a tunable parameter. The reports from the cell edge users are collected, i.e., $\mathcal{V}_m = \mathcal{E}_m$. We use the simplified version (3.8) and (3.10) to process the statistics of CQI reports. The cell outage is generated by setting the transmission power of a cell to zero at some time point.

Fig.3.4 shows that with a proper observation latitude, the hypothesis test on CQI can detect the neighbor outage cell on time. The parameter γ in (3.17) is set to be 1. A short latitude $M = 50$ leads to unreliable detection with all test results positive, while the detection based on long observing window is more promising. However, there is always a trade-off between the fast detection and reliability.

Table 3.1 records the test results of time correlation of CQI differential of the first detection latitude after the cell outage happens. We notice that the traffic load, which is indicated by user arrival rate, does not affect this test much, but a rigid type I error probability threshold X_c makes the test more conservative to give a positive detection.

Table 3.2 shows the dependency of the RRQ frequency test on load and the threshold X_f. The test is unreliable under light load state (low arrival rate) by giving the negative results (misdetection). The threshold X_f works similarly as X_c.

These results verify that when cell s is lightly loaded, the CQI statistic tests are more reliable than the RRQ frequency test. Thus, our proposal of the weight function in (3.32) is a reasonable choice, and the combination test is more robust than a single test.

TABLES

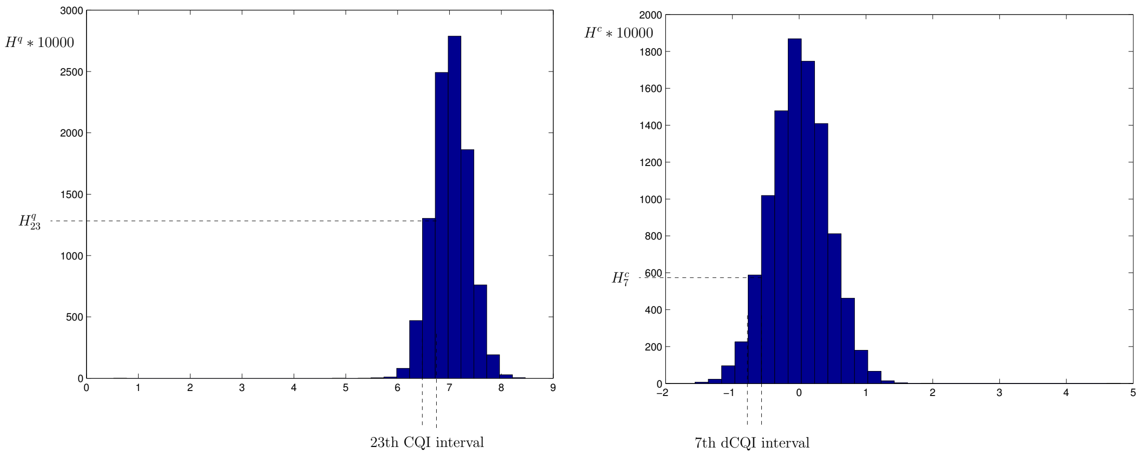
Table 3.1: HYPOTHESIS ON TIME CORRELATION OF CQI DIFFERENTIAL

	X_c			
Arrival rate	0.05	0.1	0.15	0.2
0.1 users/s	0.0593	0.1216	0.1470	0.1781
1 users/s	0.0623	0.0951	0.1217	0.4671
2 users/s	0.1121	0.1412	0.4627	0.6851

Table 3.2: HYPOTHESIS ON RRQ FREQUENCY

	X_f			
Arrival rate	0.05	0.1	0.15	0.2
0.1 users/s	-0.0392	-0.0122	0.0441	0.0624
1 users/s	0.0923	0.1946	0.2400	0.2670
2 users/s	0.1441	0.2842	0.3462	0.3832

FIGURES



(a) Histogram of CQI.

(b) Histogram of CQI differential.

Figure 3.1: Statistics of CQI

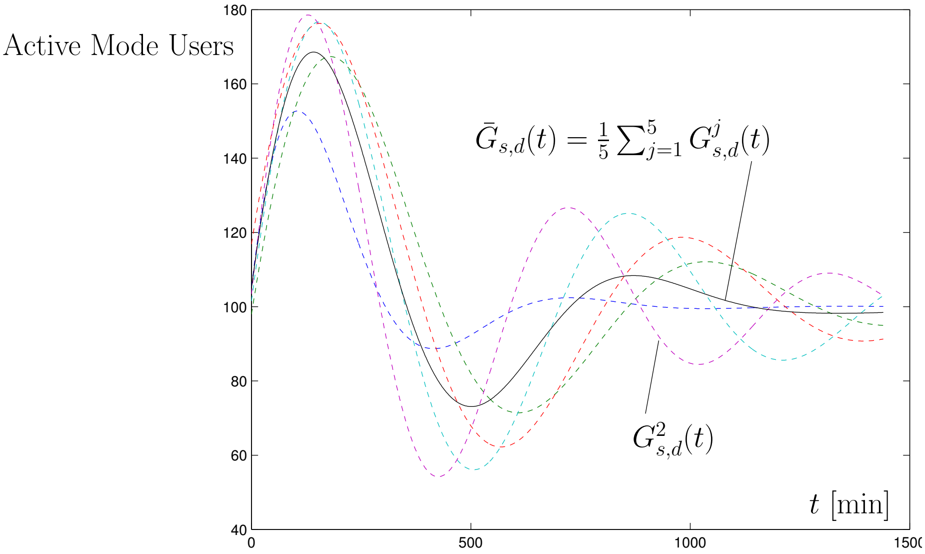


Figure 3.2: Example: load profile for cell s on d-th weekday

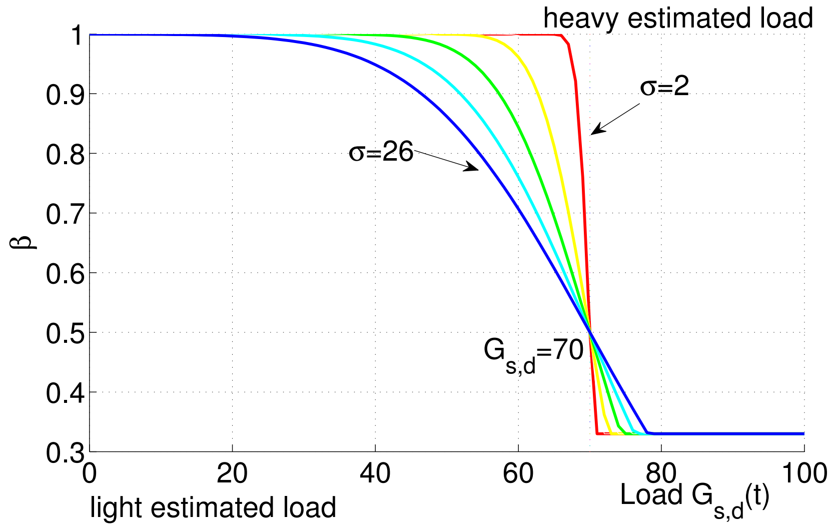


Figure 3.3: Example: weight β as erfc function of load

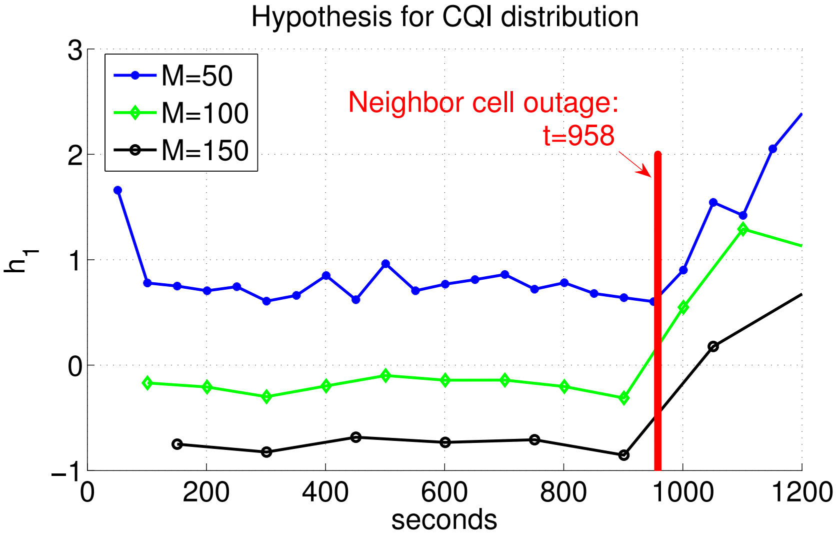


Figure 3.4: Hypothesis on CQI distribution (M is the decision latitude)

Chapter 4

Network State Awareness and Proactive Anomaly Detection

In Chapter 3 we propose a scheme for cell outage detection using the composite hypothesis testing method, which does not require the experts to have a priori knowledge. This is because, cell outage is a rare event, and the historical records may not be available before such an event occurs. However, other types of network anomalies, which occur more frequently, can be detected by exploiting the a priori knowledge. Thus, inference of network state and detection of anomaly network behavior using a priori knowledge based on historically collected information play important roles in the self-healing mechanisms for SON. In this chapter, we propose a novel framework of efficient network monitoring and proactive cell anomaly detection based on dimension reduction and fuzzy classification techniques. The enhanced semi-supervised classification algorithm allows adaptation of new behavior patterns, while incorporating a priori knowledge. The experimental results suggest that (i) our proposed method proactively detects the network anomalies associated with various fault classes, and (ii) the trajectory of the network states moving toward or away from a safe or fault class can be visualized, using the projected data onto a low-dimensional subspace.

Parts of this chapter have already been published in [10].

4.1 Introduction

We focus on automatic anomaly detection and root cause identification based on the collected KPIs , network measurements and control parameters using a priori knowledge in the network. Most prior research in this area has focused on determining the cell performance status by identifying the KPI degradation level [CLNS13, KG10, TLJ10], and providing the outputs that indicate only the severity of the degradation. We are interested in obtaining more information on classified network states associated with SON use cases, that can be further used as guidelines on self-optimization functionalities.

The major challenges to SON use case-related classification and anomaly detection are, firstly, the high-dimension of the dataset of KPIs and, secondly, the strong interactions and vague boundaries between the use cases. We propose a novel framework based on dimension reduction and semi-supervised fuzzy classification techniques to overcome the challenges. Our contributions are summarized as follows.

1) We select a set of metrics to characterize the network states, and show that the data can be mapped to a much lower dimensional space by applying the principal component analysis (PCA).

2) We enhance the kernel-based semi-supervised FCM algorithm introduced in [BP06] by optimizing the kernel parameters. The enhanced algorithm is ideally suited to deal with the vague knowledge about the classes, as it learns the hidden clustering pattern related to the SON use cases, while incorporating a priori knowledge provided by the experts.

3) We propose a proactive anomaly detection scheme based on the fuzzy classification associated with fault classes.

4) The proposed algorithms are implemented in a LTE system-level simulator. Simulation results show that the projection onto the first 3 principal components (PCs) captures the majority of the variance, and that the pattern of use case-related clusters can be observed. Thus, it is possible to visualize and to track the real-time network states in the 3-dimensional space. By analyzing the cluster memberships of the newly collected metrics, we can proactively detect the network anomalies.

4.2 Definitions and System Model

The data collected in the LTE and beyond cellular networks falls into three major groups: *control parameters*, *KPIs* and *network measurements* [HSS12]. The control parameters, such as transmission power and antenna tilt, are optimized by the self-organization solutions. Various KPIs are defined to describe the performance of accessibility, retainability, integrity, availability and mobility. The most interested KPIs are call drop rate, call blocking rate, throughput, traffic load, and mobility-related KPIs such as HO rate. Network measurements are collected at both the eNB and the UE. The statistics extracted from the network measurements indirectly reflect the traffic distribution and network environment. For example, cell-specific measurement such as estimate of UE arrival rate provides the information of the UE density. We jointly consider the KPIs and the extracted statistics from the network measurements as *network metrics*. We then use a set of network metrics to characterize the *network states*, to indicate the network performance under given network environment.

The task is to design flexible statistical methods for enhancing network awareness and for detecting network anomalies at the locality of network elements, by using the available data. We select D network metrics, and collect sample $\boldsymbol{m}_k \in \mathbb{R}^D$ at the kth observing period. Assume that we have collected a dataset of K historical samples $\mathcal{D} := \{\boldsymbol{m}_k\}_{k=1}^K$ at an eNB. Let $\boldsymbol{M} := [\boldsymbol{m}_1 \ \boldsymbol{m}_2 \ \ldots \ \boldsymbol{m}_K] \in \mathbb{R}^{D \times K}$ be the matrix formed by stacking the samples as its column vectors, with each $\boldsymbol{m}_k$ characterizing some network state.

To identify the network states, we classify them into clusters associated with different labels. In practical system, some labels can be identified based on a priori knowledge (e.g., provided by human experts) collected through historical operations. For example, the label "safe" is given to the samples if all KPIs satisfy the requirements for quality of service (QoS), and the label "coverage hole" is given if a cell outage is detected. Assume that H classes of labels are defined, and that a subset of the historical samples $\mathcal{S} \subset \mathcal{D}$ is associated with labels. For the rest of the samples, the associated labels are unknown. We define an $H \times K$ binary matrix $\boldsymbol{L} := [l_{hk}]$, where $l_{hk} = 1$ if sample k is labeled with class h, and $l_{hk} = 0$ otherwise. Note that a sample is labeled with not more than one class, we have $\sum_{h=1}^H l_{hk} \leq 1$ for each k.

4.3 Algorithmic Framework

We propose the following two steps to group the high-dimensional network states into clusters, taking into account the partially labeled samples.

1) Dimension reduction: The data of network metrics $\boldsymbol{M}$ is transformed into a new dataset $\boldsymbol{X} \in \mathbb{R}^{d \times K}$ with much lower dimensionality $d \ll D$, while retaining the geometry of the data, for the visualization purpose and for the efficiency of the classifier.

2) Semi-supervised FCM: The projected samples in dataset $\boldsymbol{X}$ are classified into C clusters, by exploring the hidden structure in data with a certain limited fraction of labeled pattern. Each cluster is associated with at most one class.

The above-mentioned two steps are described in Section 4.3.1 and 4.3.2 respectively. The proactive anomaly detection based on the classification is introduced in Section 4.3.3.

4.3.1 Dimension Reduction

We explore PCA for dimension reduction, which can be interpreted in the way of minimizing the reconstruction error between the original data and its estimates projected to the d-dimensional affine subspace [Jol02]. The details of PCA are given in Appendix C.2.

A classical solution to PCA via singular value decomposition (SVD) is as follows:

1) replacing each row of matrix $\boldsymbol{M}$ with z-scores for the row, to standardize the metrics for feature scaling,

2) performing SVD of $\boldsymbol{M}$, i.e., $\boldsymbol{M} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{W}^T$,

3) computing the solution $\boldsymbol{X} := (\boldsymbol{x}_1\ \boldsymbol{x}_2\ \dots\ \boldsymbol{x}_N) \in \mathbb{R}^{d\times N}$, where $\boldsymbol{x}_k$ is the kth column of the top $d \times K$ submatrix $\boldsymbol{\Sigma}_d\boldsymbol{W}_d^T$ of the matrix $\boldsymbol{\Sigma}\boldsymbol{W}^T$. Note that $\boldsymbol{x}_k$ as the transformation of the original data $\boldsymbol{m}_k$ can also be computed as the kth column of the top $d \times K$ submatrix $\boldsymbol{G}_d^T\boldsymbol{M}$, where $\boldsymbol{G}_d$ is exactly the first d columns of $\boldsymbol{G}$.

Matrix $\boldsymbol{X}$ is used for efficient classification in Section 4.3.2. Note that for $d \leq 3$ the network states can be visualized, which is a great advantage for monitoring the network performance.

4.3.2 Kernel-Based Semi-Supervised Fuzzy Clustering

The objective is to classify the K samples into C clusters, taking into account the limited fraction of labeled samples associated with H classes of labels. The labeled pattern is given in the binary matrix $\boldsymbol{L}$ as defined in Section 4.2. It is worth mentioning that each class h may contain a set of clusters $\mathcal{C}_h \neq \emptyset$ with cardinality $|\mathcal{C}_h| = C_h$, such that $\sum_{h=1}^{H} C_h = C$. This is because, although the experts may provide a priori knowledge, the information is incomplete and the classes are coarsely constructed. Introducing $C \geq H$ clusters achieves fine classification and further improves the anomaly detection. Although each class has at least one subordinate cluster, a cluster is associated with at most one class. If all samples assigned to a cluster are unsupervised, the cluster is associated with none of the classes and a new class is created. In this way we learn new classes to compensate for the incomplete a priori knowledge.

We enhance the kernel-based semi-supervised FCM algorithm in [BLM05] by adapting the kernel parameter, to optimize the *cluster centroids* $\boldsymbol{V} := (\boldsymbol{v}_1 \dots \boldsymbol{v}_C) \in \mathbb{R}^{d\times C}$ and *partition matrix* $\boldsymbol{U} := (u_{i,k}) \in \mathbb{R}^{C\times K}$, where each entry $u_{i,k}$ denotes the *membership degree*, which indicates the probability that sample k belongs to cluster i. The kernel-based clustering method is applied here, because it performs a nonlinear mapping that transforms nonlinearly separable data (patterns) in the input space into their linearly separable counterpart arising in the high-dimensional space. In our scenario, this corresponds to the strong nonlinear interactions between the network states related to various SON use cases.

The augmented objective function, aiming to bring together labeled and unlabeled patterns while subjected to the probabilistic constraints on membership degrees, is written as

$$
\begin{aligned}
J(\boldsymbol{U},\boldsymbol{V},\boldsymbol{\lambda}) = \alpha \sum_{i=1}^{C}\sum_{k=1}^{K} u_{i,k}^2 \|\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{\phi}(\boldsymbol{v}_i)\|^2 \\
+ (1-\alpha)\sum_{i=1}^{C}\sum_{k=1}^{K}(u_{i,k} - \tilde{u}_{i,k})^2 \|\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{\phi}(\boldsymbol{v}_i)\|^2 - \sum_{k=1}^{K}\lambda_k\left(\sum_{i=1}^{C} u_{i,k} - 1\right) \qquad (4.1)
\end{aligned}
$$

where $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_K)^T$ denotes the Lagrangian multipliers, and the *reference membership* $\tilde{u}_{i,k}$ helps to optimize the membership using the labeling information in contrast to $u_{i,k}$ as explained in (4.4). The mapping $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^F$ is a (nonlinear) mapping from a d-dimensnional space to F-dimensional space such that $d \ll F$. Note that an explicit representation for $\boldsymbol{\phi}$ is not required. Using the kernel trick [SS98, p. 38] in the inner product space $k(\boldsymbol{x}, \boldsymbol{v}) = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{v})$, and defining the Gaussian radial basis function kernel

$$
k(\boldsymbol{x}, \boldsymbol{v}) := \exp(-\|\boldsymbol{x} - \boldsymbol{v}\|^2/\sigma) \qquad (4.2)
$$

where $\sigma > 0$ is the kernel parameter, the distance between sample $\boldsymbol{x}_k$ and centroid $\boldsymbol{v}_i$ in the projected feature space is given by

$$
\begin{aligned}
\|\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{\phi}(\boldsymbol{v}_i)\|^2 &= k(\boldsymbol{x}_k, \boldsymbol{x}_k) + k(\boldsymbol{v}_i, \boldsymbol{v}_i) - 2k(\boldsymbol{x}_k, \boldsymbol{v}_i) \\
&= 2(1 - k(\boldsymbol{x}_k, \boldsymbol{v}_i)) \qquad (4.3)
\end{aligned}
$$

Thus, substituting (4.2) and (4.3) into (4.1), the objective function $J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \sigma)$ depends on variables $\{\boldsymbol{U}, \boldsymbol{V}\}$, Lagrangian multipliers $\boldsymbol{\lambda}$, and the kernel parameter σ.

To represent the labeled pattern, the reference memberships $\tilde{\boldsymbol{U}} := (\tilde{u}_{i,k})$ are iteratively updated by optimizing the objective

$$
Q(\tilde{\boldsymbol{U}}) = \sum_{h=1}^{H}\sum_{k=1}^{K}\delta_k\left(l_{h,k} - \sum_{i\in\mathcal{C}_h}\tilde{u}_{i,k}\right)^2, \ \tilde{u}_{i,k} \in [0,1] \qquad (4.4)
$$

where $\delta_k := \sum_{h=1}^{H} l_{h,k}$ takes value one if sample k is labeled and zero otherwise. The binary matrix $\boldsymbol{L} := (l_{h,k})$ indicating the labeling information is predefined according to the a priori knowledge. The set of clusters associated with class h denoted by $\mathcal{C}_h$ is iteratively updated depending on the partition matrix $\boldsymbol{U}$ as described later in this section. Ideally, when optimizing $Q(\tilde{\boldsymbol{U}})$, the sum of the reference memberships of sample k to the clusters associated with class h is one if sample i is labeled with class h, otherwise the sum is zero.

The algorithm consists of two iterative optimization phases:

- Optimize $Q(\tilde{\boldsymbol{U}})$ to update $\tilde{\boldsymbol{U}}$, and
- Optimize $J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \sigma)$ to update $\{\boldsymbol{U}, \boldsymbol{V}, \sigma\}$.

The solution based on the gradient descent and the coordinate descent methods is provided as follows.

1) Optimization of $Q(\tilde{\boldsymbol{U}})$. The matrix $\tilde{\boldsymbol{U}}$ is updated by

$$\begin{aligned}\tilde{u}_{i,k}^{(n+1)} &= \tilde{u}_{i,k}^{(n)} - \beta \frac{\partial Q(\tilde{\boldsymbol{U}})}{\tilde{u}_{i,k}} \\ &= \tilde{u}_{i,k}^{(n)} + 2\beta\delta_k \sum_{h=1}^{H} 1_{\{i \in \mathcal{C}_h^{(n)}\}} \left(l_{h,k} - \sum_{j \in \mathcal{C}_h^{(n)}} \tilde{u}_{j,k}^{(n)} \right) \end{aligned} \tag{4.5}$$

where n refers to the index of iterations, $1_{\{A\}}$ denotes the indicator function that takes value one if event A holds true, and zero otherwise, and $\beta > 0$ is the step size that controls the process of step-wise optimization over $\tilde{\boldsymbol{U}}$, which is optimized via backtracking line search.

In (4.5), set $\mathcal{C}_h$ is updated according to the partition matrix $\boldsymbol{U}$. To derive $\mathcal{C}_h$, we first define a $C \times K$ binary matrix $\boldsymbol{B} := (b_{i,k})$, such that $b_{\hat{i},k} = 1$ if $\hat{i} = \arg\max_i u_{i,k}$, and zero otherwise. Matrix $\boldsymbol{B}$ indicates whether a sample belongs to a cluster or not. We construct matrix $\boldsymbol{P} := \boldsymbol{L}\boldsymbol{B}^T \in \mathbb{R}^{H \times C}$, where $p_{h,i}$ is the number of samples in cluster i labeled with class h. Let $i \in \mathcal{C}_{\hat{h}}$ for each cluster i if $\hat{h} = \arg\max_h p_{h,i}$. Note that $\mathcal{C}_h \neq \emptyset$, if none of the clusters is assigned to class h, then h is allowed to take a cluster $\hat{i}_h = \arg\max_i p_{h,i} / \sum_{i=1}^{C} p_{h,i}$ from the other class.

2) Optimization of $J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \sigma)$. The objective function is optimized by computing the partial derivatives of (4.1) with respect to the parameters $u_{i,k}$, $\boldsymbol{v}_i$, λ_k, and σ respectively and performing the coordinate descent.

By setting $\partial J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \sigma)/\partial u_{i,k} = 0$, we have

$$u_{i,k} = \frac{\lambda_k}{4(1 - k(\boldsymbol{x}_k, \boldsymbol{v}_i))} + (1-\alpha)\tilde{u}_{i,k} \tag{4.6}$$

Setting $\partial J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \sigma)/\partial \lambda_k = 0$, we obtain the probabilistic constraint

$$\sum_{i=1}^{C} u_{i,k} = 1 \tag{4.7}$$

Substituting (4.6) into (4.7), we derive

$$\lambda_k = \frac{4\left(1 - (1-\alpha) \cdot \sum_{i=1}^{C} \tilde{u}_{i,k}\right)}{\sum_{i=1}^{C} \left(1 - k(\boldsymbol{x}_k, \boldsymbol{v}_i)\right)^{-1}} \tag{4.8}$$

We update $u_{i,k}$ by substituting (4.8) into (4.6), written as

$$u_{i,k} = \begin{cases} (1-\alpha)\tilde{u}_{i,k} + \frac{1-(1-\alpha)\sum_{j=1}^{C} \tilde{u}_{j,k}}{\sum_{j=1}^{C} \frac{1-k(\boldsymbol{x}_k, \boldsymbol{v}_i)}{1-k(\boldsymbol{x}_k, \boldsymbol{v}_j)}} & \text{if } \boldsymbol{x}_k \neq \boldsymbol{v}_i \\ 1 & \text{if } \boldsymbol{x}_k = \boldsymbol{v}_i \end{cases} \tag{4.9}$$

To update $\boldsymbol{v}_i$, we set $\partial J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \sigma)/\partial \boldsymbol{v}_i = 0$, which gives

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{K} \left(\alpha u_{i,k}^2 + (1-\alpha)(u_{i,k} - \tilde{u}_{i,k})^2\right) k(\boldsymbol{x}_k, \boldsymbol{v}_i)\boldsymbol{x}_k}{\sum_{k=1}^{K} \left(\alpha u_{i,k}^2 + (1-\alpha)(u_{i,k} - \tilde{u}_{i,k})^2\right) k(\boldsymbol{x}_k, \boldsymbol{v}_i)} \tag{4.10}$$

To update $u_{i,k}^{(n+1)}$ and $\boldsymbol{v}_i^{(n+1)}$ at the $(n+1)$th iteration, we use $\tilde{u}_{i,k}^{(n)}$, $\boldsymbol{v}_i^{(n)}$, $\boldsymbol{u}_{i,k}^{(n)}$ and $\sigma^{(n)}$ from the last iteration on the right side of the equations (4.9) and (4.10), respectively. Moreover, note that in (4.10) variable $\boldsymbol{v}_i$ also appears on the right side of the equation, a sequence of updated $\boldsymbol{v}_i$ is computed by the fixed point iteration.

Using gradient descent, the kernel parameter σ is iteratively updated as follows

$$\begin{aligned}\sigma^{(n+1)} &= \sigma^{(n)} - \rho \frac{\partial J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \sigma)}{\partial \sigma} \\ &= \sigma^{(n)} + 2\rho\alpha \sum_{i=1}^{C}\sum_{k=1}^{K} u_{i,k}^2 \frac{k(\boldsymbol{x}_k, \boldsymbol{v}_i)\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2}{{\sigma^{(n)}}^2} \\ &+ 2\rho(1-\alpha)\sum_{i=1}^{C}\sum_{k=1}^{K}(u_{i,k} - \tilde{u}_{i,k})^2 \frac{k(\boldsymbol{x}_k, \boldsymbol{v}_i)\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2}{{\sigma^{(n)}}^2}\end{aligned} \tag{4.11}$$

where $\rho > 0$, similar to β in (4.5), is the step size.

The kernel-based semi-supervised FCM algorithm with adaptive kernel parameter is provided in Algorithm 1. To determine the number of clusters C, we start with a sufficiently large value of $C^{(0)}$, and fuse the clusters iteratively, if the distance between any pair of cluster centroids is small enough.

4.3.3 Proactive Anomaly Detection

To associate the newly collected sample $\boldsymbol{m}'$ to a class, the following steps are proposed:

1) computing the normalized value $\tilde{\boldsymbol{m}}'$, with the mean and variance obtained from the z-score in Section 4.3.1,

2) computing the projection onto PCs $\boldsymbol{x}' = \boldsymbol{G}_d^T \boldsymbol{m}'$,

3) computing the membership degree to clusters $u(\boldsymbol{x}', \boldsymbol{v}_i)$ for $i = 1, \dots, C$ with (4.9).

The *class membership* is defined as $\omega_h(\boldsymbol{x}') := \sum_{i \in \mathcal{C}_h} u(\boldsymbol{x}', \boldsymbol{v}_i)$, which indicates the probability that sample $\boldsymbol{m}'$ is associated to a class h. For real-time anomaly detection, we associated the sample with class $\hat{h}$ if $\hat{h} = \arg\max_h \omega_h(\boldsymbol{x}')$.

Furthermore, by analyzing the trajectory of a sequence of recent collected samples $\{\boldsymbol{x}_{n-l}, \dots, \boldsymbol{x}_n\}$, we can predict the network anomalies. Define a metric of *percentage change* for the class memberships $\nu_{h,k} := (\omega_h(\boldsymbol{x}_k) - \omega_h(\boldsymbol{x}_{k-1}))\,/\omega_h(\boldsymbol{x}_{k-1})$. Assume that $\boldsymbol{x}_n$ is associated with the safe class h^*, i.e., $h^* = \arg\max_h \omega_h(\boldsymbol{x}_n)$. However, if the successive

Algorithm 1: Kernel-based semi-supervised FCM with adaptive kernel parameter.

```
Data: Dataset $\{\boldsymbol{x}_k\}_{k=1}^K$, labeling matrix $\boldsymbol{L}$
Result: Partition $\boldsymbol{U}$, centroids $\boldsymbol{V}$, kernel parameter $\sigma$
Initialization: number of classes $H$, number of clusters $C^{(0)}$, thresholds $\tau_1, \tau_2, \tau_3, d_0$,
maximum number of iterations $N_{\max}$, $C \leftarrow C^{(0)}$;
while $\left(C = C^{(0)}\right)$ or $(\exists i \neq j$ such that $d_{ij} < d_0)$ do
    Iteration step $n = 0$;
    Standard FCM to entire dataset to compute initial $\boldsymbol{U}^{(0)}, \boldsymbol{V}^{(0)}$;
    Determine $\mathcal{C}_h^{(0)}$ for all $h$ using $\boldsymbol{U}^{(0)}$, and $\mathcal{C}_h^{(-1)} = \emptyset$;
    Initialize $\tilde{\boldsymbol{U}}^{(0)} = \boldsymbol{U}^{(0)}$, $\sigma^{(0)} > 0$;
    while $\mathcal{C}_h^{(n)} \neq \mathcal{C}_h^{(n-1)}$ for all $h$ do
        while $\|\tilde{\boldsymbol{U}}^{(n+1)} - \tilde{\boldsymbol{U}}^{(n)}\| \geq \tau_1$ do
            Compute $\tilde{\boldsymbol{U}}^{(n+1)}$ with (4.5)
        while $\|\sigma^{(n+1)} - \sigma^{(n)}\| \geq \tau_2$ do
            Compute $\sigma^{(n+1)}$ with (4.11)
        while $\|\boldsymbol{U}^{(n+1)} - \boldsymbol{U}^{(n)}\| \geq \tau_3$ do
            a) Compute $\boldsymbol{V}^{(n+1)}$ with (4.10);
            b) Compute $\boldsymbol{U}^{(n+1)}$ with (4.9)
        Update $\mathcal{C}_h^{(n+1)}$ for all $h$;
        $n \leftarrow n + 1$;
        if $n \geq N_{max}$ then
            break
    Compute $[d_{ij}]$ where $d_{ij} := \|\boldsymbol{v}_i^{(n)} - \boldsymbol{v}_j^{(n)}\|$;
    $C \leftarrow C - 1$
```

$\{\nu_{\hat{h},k}\}_{k=n-l}^{n}$ are positive for some fault class $\hat{h}$, while $\{\nu_{h^*,k}\}_{k=n+1}^{n}$ are negative for the safe class h^*, an alarm is triggered for the potential fault class $\hat{h}$.

4.4 Experimental Results

We apply the proposed algorithms to the data collected from an OFDMA-based LTE system-level simulator aided by the IKR-Tools Library [SS10]. The IKR-Tool Library is an object-oriented class library for event-driven simulation available in both C++ and JAVA. The simulation is a wrap-around configuration of 7 hexagonal 3-sectored eNBs, with the LTE carrier bandwidth of 10 MHz. The physical layer is abstracted by simplified models that capture its characteristic with high accuracy and low complexity. The link measurements such as pathloss, shadow fading and antenna gain are modeled according to 3GPP specifications [3GPj, Table A.2.1.1-2], while the fast fading is neglected. Proportional fair scheduling algorithm with QoS constraints is implemented.

Two types of traffic are generated spatially uniformly on the playground: VoIP and data streaming traffic. The VoIP traffic has a QoS requirement of 30 kBit/s, while the data streaming user has no such requirement. With probability 0.8 the generated traffic belongs to the mobility group "pedestrian" with the speed of 3 km/h, and with probability 0.2 the traffic is generated as "urban vehicular" with the speed of 30 km/h. The traffic generator follows Poisson distribution, with configurable arrival rate for VoIP and streaming traffic. Fig. 4.5 illustrates the pixel-based number of UEs and average SINR during 500 seconds.

4.4.1 Selected Parameters and Metrics

The network system is configurable by tuning a set of control parameters (e.g., antenna tilt and transmit power) or a set of network variables (e.g., traffic arrival rate). The statistics of network metrics are collected every 500 seconds. The selected parameters and metrics are listed in Table 4.1.

1) *Control parameters.* Adaptation of antenna tilt and transmit power is the possible solution to SON functionalities CCO, ES and IR. Optimization of HO-related parameters TTT and hysteresis is among the possible solutions to MRO and MLBO.

2) *Key performance indicators.* The selected KPIs are among the most important indicators for coverage, capacity and mobility-related performance. Note that here the load indicator is defined as the fraction of the number of occupied physical resource blocks (PRBs) to the total number of the PRBs.

3) *Statistical network measurements.* The selected statistical network measurements indirectly reflect the network environment. We also include the statistics collected from the neighboring cells, to consider the interference distribution and the coupling between the sites. It is required that the neighboring eNBs exchange the following information with each other: 1) estimates of UE arrival rate, and 2) the mean and variance of RSRQ distribution. We abuse notation and compute the mean and variance of RSRQ distribution in cell b by $\bar{r}_b := (1/K_b) \cdot \sum_{k\in\mathcal{K}_b} r_k$ and $v_b := (1/K_b) \cdot \sum_{k\in\mathcal{K}_b} (r_k - \bar{r}_b)^2$ respectively, where r_k denotes the average RSRQ value of user k over an observation period, and $\mathcal{K}_b$ denotes the set of users served by cell b, with $|\mathcal{K}_b| = K_b$. The mean and variance of RSRQ distribution in all neighboring cells of cell b are calculated as $\bar{r}_{\mathcal{N}_b} := (1/K) \cdot \sum_{n\in\mathcal{N}_b} K_n \bar{r}_n$ and $v_{\mathcal{N}_b} := (1/K) \cdot \sum_{n\in\mathcal{N}_b} K_n v_n$ respectively, where $\mathcal{N}_b$ denotes the set of neighboring cells of cell b, and $K = \sum_{n\in\mathcal{N}_b} K_n$. We consider the statistical distribution of RSRQ because it indirectly indicates the signal and interference distribution.

4.4.2 Generation of Experimental Samples

The default parameters for the configuration settings are provided as follows: antenna tilt of 10 degrees, transmission power of 42 dBm, hysteresis of 0 dB and TTT of 256ms. To introduce randomness into the samples, we generate 400 random configurations, with the majority of the control parameters near from the default values. The probability mass functions of the control parameters are shown in Fig. 4.2. Among the 400 random configurations, we provide 150 labeled samples and define 6 labels, including “safe state”, “low capacity”, “low coverage”, “overload”, “too late HO”, and “too early HO”, simplified as “SAFE”, “L_COV”, “L_COV”, “L_HO” and “E_HO” respectively. Each labeled sample is associated with one of the labels according to the expert’s knowledge based on the operator-defined quality of requirement (QoS). The design principles of the labeling are shown in Table 4.2.

4.4.3 Evaluation of Algorithm

Fig. 4.5 illustrates the performance of PCA on the total number of 400 samples of 16-dimensional network metrics (including KPIs and statistical network measurements defined in Table 4.1), and shows that we can visualize the network states by using the projections onto the first 3 principal components (PCs). Fig. 4.3(a) illustrates that the first 3 eigenvalues capture over 70% of the variance. Thus, it may be adequate to use the projected data points in the 3-dimensional space for clustering. Fig. 4.3(b) shows the mean square error (MSE) for the low-rank approximation. Fig. 4.3(c) illustrates the normalized root mean square error (NRMSE) of the approximation of each network metric. We observe that some network metrics have a good approximation in 3-dimensional linear subspace, such as the average throughput of the VoIP user and the streaming user, and the mean and variance of RSRQ distribution (red circles with indices $9, 10, 13$ and 14 on x-axis in Fig.4.3(c)). Fig. 4.3(d) illustrates the contribution of the 16 network metrics to the top 3 PCs: (i) the load-related metrics (load, number of UEs) contribute most to PC1, (ii) the QoS-related metrics (RSRQ, throughput) contribute most to PC2, and (iii) the neighboring cell-related metrics (HR_in, RSRQ distribution in neighboring cells) contribute most to PC3.

The quality of the semi-supervised clustering is quantified in terms of *accuracy* and *entropy of the clusters.* The accuracy is defined as the ratio of the number of correctly classified labeled samples to the total number of the labeled samples. The entropy of cluster $i, i = 1, \dots, C$ is defined as

$$E_i = -\frac{1}{\ln H} \sum_{h=1}^{H} \frac{\tilde{K}_{i,h}}{\tilde{K}_i} \ln \frac{\tilde{K}_{i,h}}{\tilde{K}_i} \tag{4.12}$$

where $\tilde{K}_i$ denotes the number of the labeled samples in cluster i, and $\tilde{K}_{i,h}$ denotes the number of labeled samples that are associated with class h. The entropy $E_i \in [0, 1]$ measures the distribution of classes in cluster i. A low entropy is desired, which provides a good purity within the cluster. The entropy value close to one indicates a uniform distribution of classes in a cluster leading to a bad split.

By adjusting the tuning parameter α in objective function (4.1), we can minimize the number of misclassified samples. Fig. 4.4 illustrates the dependence of accuracy and entropy of cluster on α.

Fig. 4.5 shows the semi-supervised clustering with $\alpha = 0.6$. We choose $\alpha = 0.6$ to achieve a good accuracy for the labeled samples, while exploring the hidden clustering pattern in the unlabeled samples. We start with a large number of clusters $C^{(0)} = 25$ for initialization, and end up with a number of 17 clusters as shown in Fig. 4.5, by iteratively fusing the clusters if the distance between two cluster centroids is small enough.

To examine the performance of tracking and anomaly detection, we simulate a scenario of real-time detection of coverage and capacity problem, caused by the high interference received from the neighboring cells. We set the control parameters to be the default values, while step-wise increasing the average arrival rate in the neighboring cells from 0.35 to 0.75 call/sec. Fig. 4.6(a) shows the trajectory of network states, starting from a cluster associated with a SAFE class, moving toward the cluster associated with the L_COV class. The black left-pointing triangle indicates the real-time network state. The class memberships of the trajectory is shown in Fig. 4.6(b), which illustrates a significant increase in membership to class L_COV, slight increase in membership to class L_CAP, and almost constant decrease in membership to class SAFE.

4.5 Summary

we propose a novel framework of proactive anomaly detection based on dimension reduction and fuzzy classification techniques. The dimension reduction is applied for visualization purpose and for the quality and efficiency of the classification of high-dimensional data. The enhanced kernel-based semi-supervised FCM explores the complex pattern hidden in the unlabeled samples, while taking into account the a priori knowledge provided by the labeled samples. The experimental results show that the proposed framework proactively detects network anomalies associated with various fault classes.

TABLES

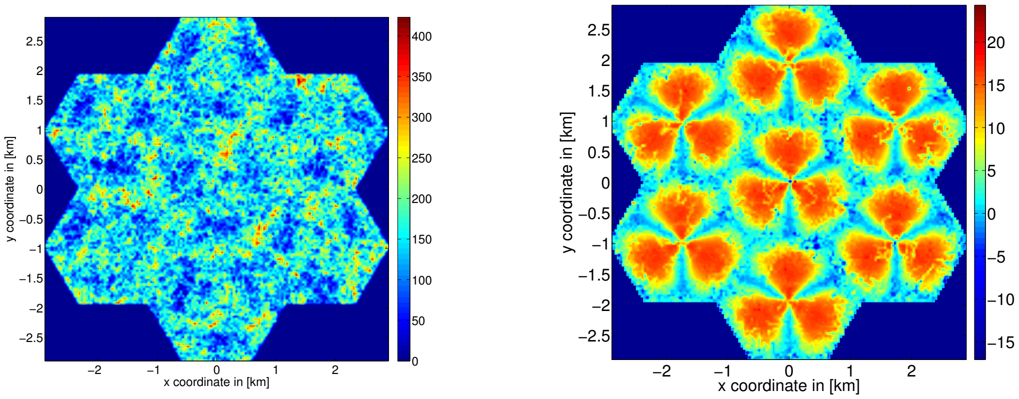
Table 4.1: SELECTED PARAMETER AND METRICS

Control Parameter	KPI	Statistical Network Measurements
1. antenna tilt	1. CDR	11. number of UEs
2. transmit power	2. CBR	12. average UEs arrival rate in neighboring cells
3. TTT	3. HOI_SR	13. mean of RSRQ distribution
4. hysteresis	4. HOO_SR	14. variance of RSRQ distribution
	5. HO_PPR 6. CS_SR	15. mean of RSRQ distribution in neighboring cells
	7. VoIP load 8. streaming load 9. VoIP SAT 10. streaming SAT	16. variance of RSRQ distribution in neighboring cells

Table 4.2: SUPERVISED CLASSES BASED ON A PRIORI KNOWLEDGE

Class	A priori knowledge
1. SAFE	all KPIs satisfy the requirements of QoS
2. L_COV	high CDR, low SAT low mean of RSRQ, high variance of RSRQ
3. L_CAP	low SAT, normal CDR
4. OL	high CBR, high load, low SAT
5. E_HO	high HO_PPR, high HOI_SR and HOO_SR
6. L_HO	low CS_SR, low HO_PPR

FIGURES



(a) Number of UEs

(b) Average SINR

Figure 4.1: Pixel-based statistics in 500 seconds.

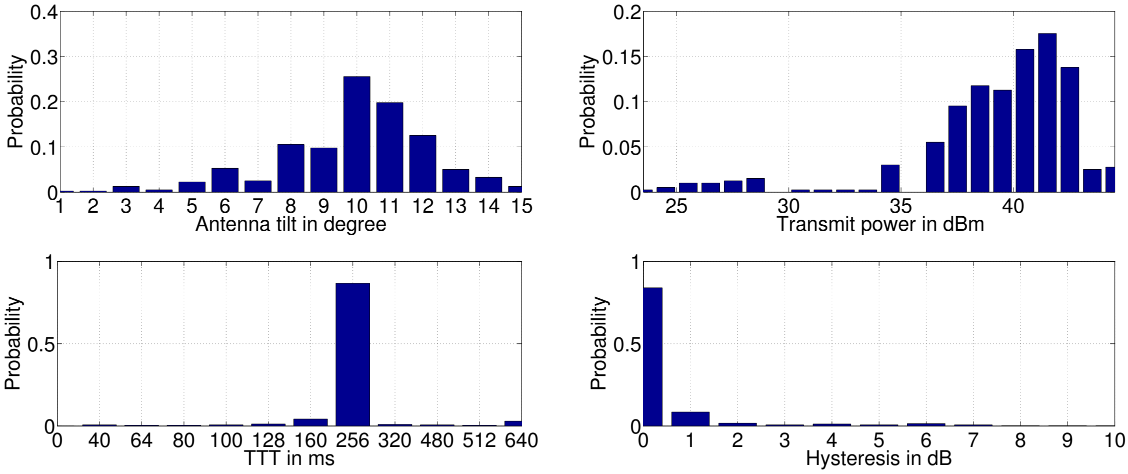
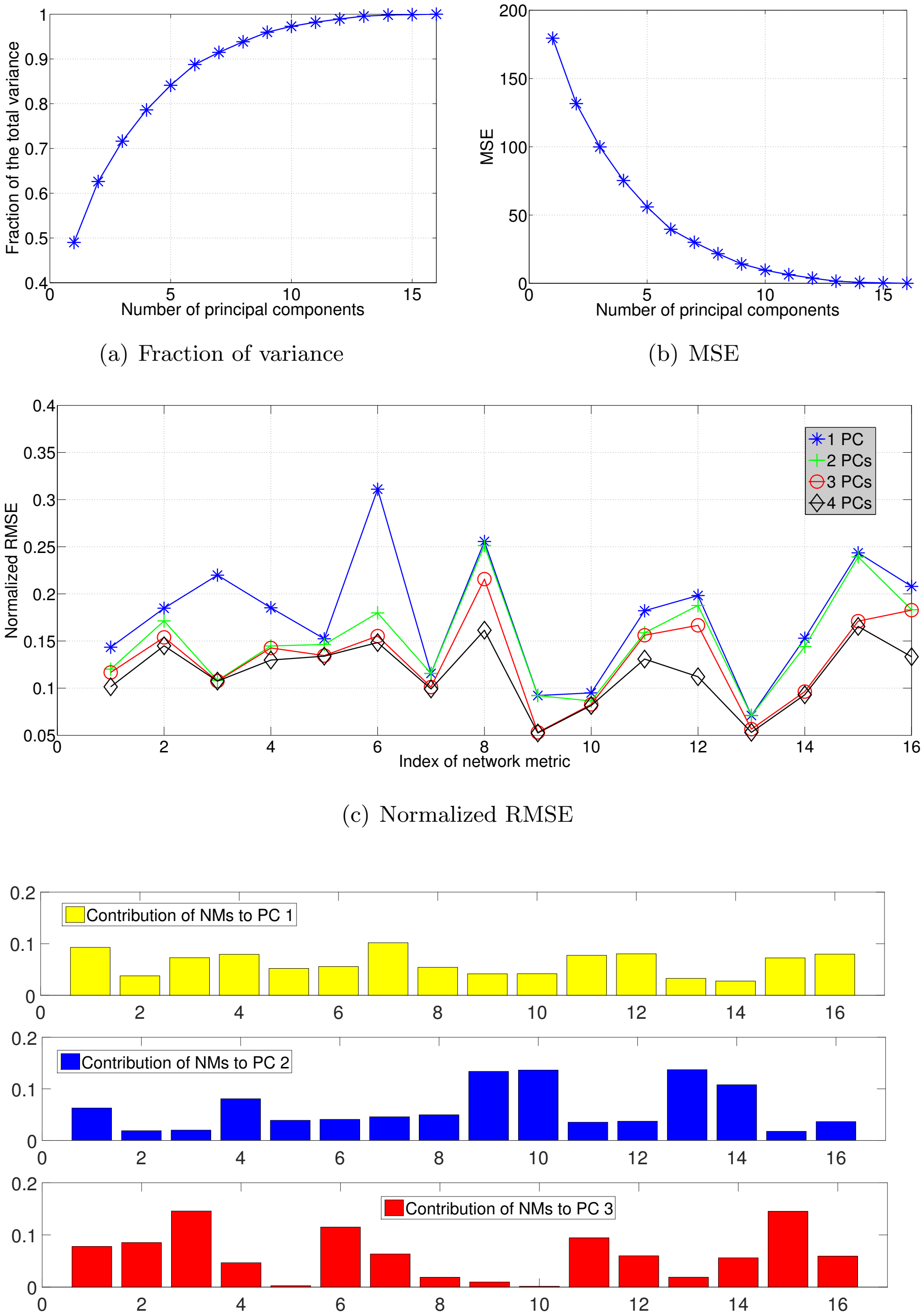


Figure 4.2: Probability mass function of control parameters



(a) Fraction of variance

(b) MSE

(c) Normalized RMSE

(d) Contribution of 16 network metrics to the top 3 PCs

Figure 4.3: Performance of PCA

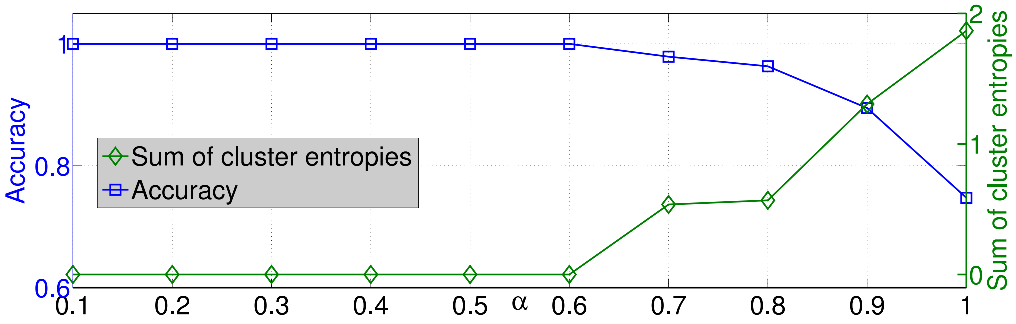


Figure 4.4: Quality of semi-supervised clustering depending on α.

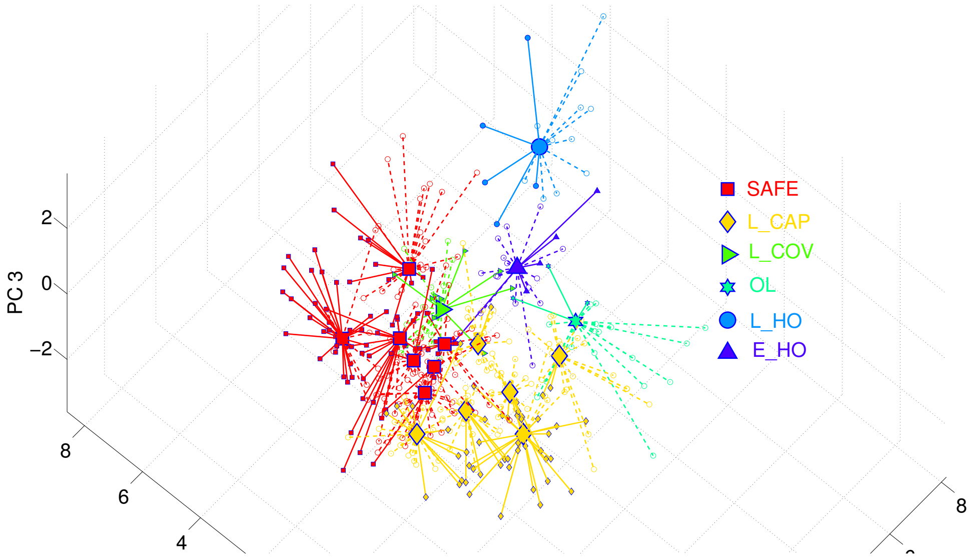
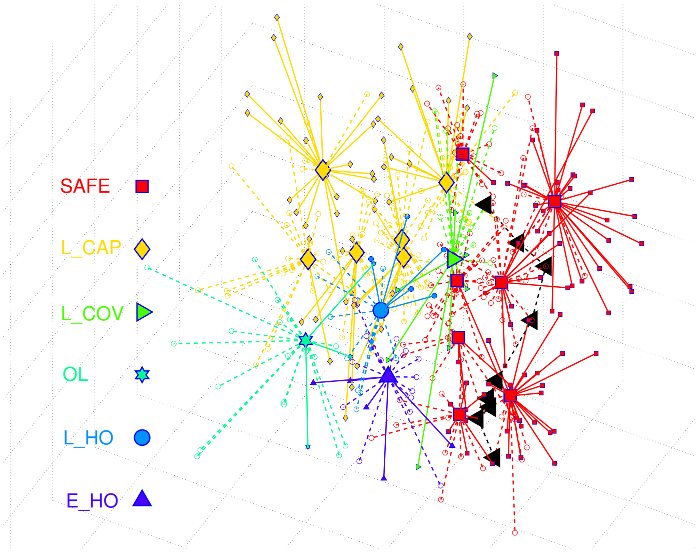
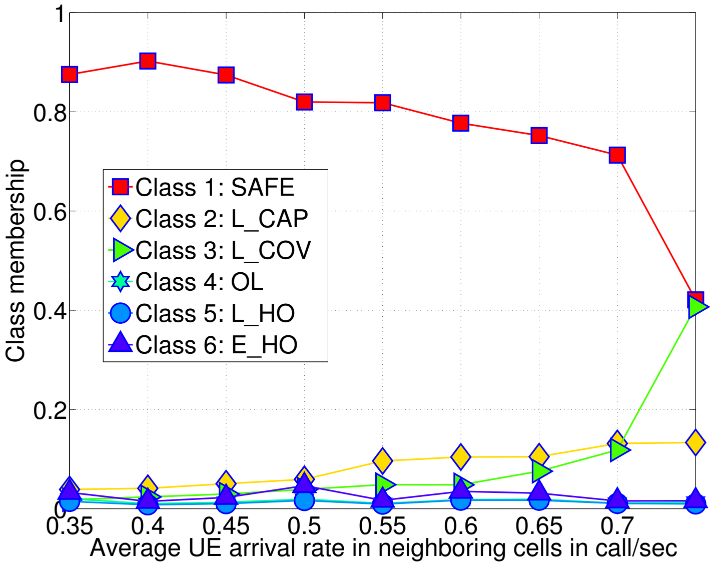


Figure 4.5: Kernel-based semi-supervised FCM with $\alpha = 0.6$. The filled markers with solid lines are the labeled samples, while unfilled circles with slashed lines stand for the unlabeled samples. Labeled samples associated to classes SAFE, L_CAP, L_COV, OL, L_HO and E_HO are represented by red square, yellow diamond, green right-pointing triangle, sea green six-pointed star, process blue circle, blue violet upward-pointing triangle respectively.



(a) Trajectory of network state



(b) Class memberships

Figure 4.6: Evolution of network state when increasing the average arrival rate in neighboring cells

Part III

Self-Optimization

Chapter 5

Measurement-Adaptive Random Access Channel Self-Optimization

In this chapter, we consider single-cell RACH in cellular wireless networks. Communications over RACH take place when users try to connect to a base station during a handover or when establishing a new connection. Within the framework of SONs, the system should self-adapt to dynamically changing environments (channel fading, mobility, etc.) without human intervention. For the performance improvement of the RACH procedure, we aim here at maximizing throughput or alternatively minimizing the user dropping rate. In the context of SON, we propose protocols which exploit information from measurements and user reports in order to estimate current values of the system unknowns and broadcast global action-related values to all users. The protocols suggest an optimal pair of user actions (transmission power and back-off probability) found by minimizing the drift of a certain function. Numerical results illustrate considerable benefits of the dropping rate, at a very low or even zero cost in power expenditure and delay, as well as the fast adaptability of the protocols to environment changes. Although the proposed protocol is designed to minimize the amount of discarded users per cell, our framework allows for other variations (power or delay minimization) as well.

Parts of this chapter have already been published in the coauthored work [14].

5.1 Introduction

Random multiple access schemes have traditionally played an important role in wireless communication systems. Their use has been established especially in cases of bursty source traffic, where a multiplicity of users requires access from a central receiver. Starting with the ALOHA protocol [Abr70], several modifications have been suggested in the years to come aiming at performance improvement [EH98]. A very common application is in wireless LANs, such as the IEEE 802.11 protocol (see [Bia00], [GSS], [SGK06] and references therein). The random access channel (RACH) is also included in the 3rd Generation Partnership Project (3GPP) as an important element within the LTE of cellular systems [3GPf], [3GPa], [3GPh].

In the case of wireless cellular networks, a very limited frequency resource is reserved for the cases when a user requests for access from a base station (BS) or in order to be synchronized for uplink/downlink data transmission. RACH communications further occur during the hand-over phase [1], because of user mobility, or when a user is (re-)initiating some new service. RACH channel can be used as well during the load balancing procedure [3], when cell-edge users are pushed to migrate to a neighboring BS after modification of the cell individual offset. Hence, *as many users as possible should be served by this limited resource, for an important number of connectivity-related actions.*

Due to limited resources, connection failure can occur in cases when the system is not well adapted to the incoming traffic. Consider for example large spaces in cities where occasionally a vast amount of requests for service can be demanded, although normaly the system is not heavily loaded (e.g. metro stations, market streets, stadiums, city squares, areas close to concert and conference halls etc.). In such places, it is very common that the system fails to support the service for all users and one of the reasons can be high collision rate in the RACH channel. It is thus necessary, within the context of SON [3GPa], [OG12] that the system can adapt to abrupt environmental changes that influence its functionality. Thus the RACH self-optmization problem is identified as an important case in the LTE standardization process [3GPa, paragraph 4.7].

Unfortunately, in all such cases, the cellular system has almost zero user-specific information. Each BS can however broadcast certain information with cell-specific access details [AFG$^+$SA], which allow the users to adapt their operation. Furthermore, carrier sensing as understood in the 802.11 is here not possible, which provides limitations to the design of high performance protocols. This is because, the possibility for a user to sense whether the channel is idle or not, is not provided and collision events cannot be avoided.

The procedure is called random access, due to the fact that the users access the channel in a random fashion. In the ALOHA case, when more than one user transmit simultaneously and their signals are detected we say that a collision occurs and all efforts are considered unsuccessful. LTE standardization, instead, provides the possibility for each user to randomly choose over a common pool of orthogonal frequencies [3GPf] and a collision takes place when at least two users make the same choice during the same transmission interval. After a failure, each source enters a back-off mode. The period of user silence is usually chosen having an exponential distribution but other possibilities can be used when such choice is adapted dynamically. This back-off time can generally be modeled in the slotted case by a per slot probability of transmission, less than 1. Using this technique, an increase in throughput is achieved at the cost of additional delay. Furthermore, since the detection or not of a user signal is also critical for the success, an important parameter is the transmission power of each user as well.

In short, the access (back-off) probability and the signal power are the two user actions, with the aim to optimally exploit the random access resource, in the sense of maximizing the rate of served users and minimizing the dropping user rate. An interesting idea to improve the decision making is to make certain global information of the system state available by broadcasting it from the base station. This is compatible with LTE standards where other type of information is already considered as globally known [3GPf]. The information should represent the current system situation, so that users may adapt their actions dynamically. In this way the delay-throughput tradeoff can be enhanced. The cost is certain signaling and computations for the updates at the BS side. Furthermore, the BS should have a way to gather relevant empirical information from its environment, related to the RACH functionality.

Based on the above idea, the current work suggests a dynamically adaptive RACH protocol for the cellular systems focused on LTE design, which maximizes a sense of throughput and minimizes dropping. Empirical information is gathered through measurements and user reports. After certain processing at the BS side global system parameters are broadcast to users who require access. The protocol suggested, which is based on adaptation of the system to changes in the environment, guarantees near-optimal performance related to a certain throughput-related metric.

5.1.1 Related Literature

Bianchi [Bia00] has been the first to provide a precise performance analysis for a random access protocol, which uses exponential back-off times. His approach considers a saturated system model, where the number of users is kept fixed to N and all have a packet to send at each time slot. The results are based on the key approximation that the collision probability of a packet transmitted is constant and independent, which decouples the evolution of the system to N 1-dimensional Markov Chains.

A different approach has been suggested by Sharma et al. [SGK06], where more general back-off strategies (generalized geometric) are considered for the IEEE 802.11 protocol in order to take service differentiation into account. One of the major differences is that the system state is described by the current number of users per effort, while the collision probability is not independent per user.

First suggestions for dynamically controlling multiple access protocols can be found in Hajek and van Loon [HvL82] as well as Lam and Kleinrock [LK75]. More recently Markov Decision Processes (MDPs) have been used in [dAF04] to derive optimal power and back-off policies for a set of backlogged users in slotted ALOHA random access systems. Cases of unknown user number have also been taken into account.

Gupta et al [GSS] have recently suggested a dynamic back-off adaptation mechanism, where contention is regulated by broadcasting a so called contention level to the users. This is similar to the idea used in our approach. Works of particular interest are also those of Liu et al [LYP$^+$09] and Cheung et al [CMRWS10] which use the framework of utility-optimization for the optimal choice of transmission probabilities.

Channel-aware scheduling approaches in conjunction with random access mechanisms (which do not find application here due to the lack of such information in the system) include [DSZ04], [TZM01], and more recently [AHBW11].

How random access works in the 3GPP-LTE systems is thoroughly described in [AFG$^+$SA], where certain suggestions are presented, related to a self-organizing mechanism with information exchange between users and the Base Station. Investigations on the RACHpower control include [LKC$^+$12] and references therein, whereas an analytical framework for RACH modeling and optimization is given in [YHH11].

Finally, rather interesting for the Carrier sense multiple access with collision avoidance (CSMA/CA) case is the dynamic adaptation mechanism suggested in [HRGD05] where users adapt their time window based on measurements and estimation of the average number of idle time slots of the random access channel. It involves an Additive Increase Multiplicative Decrease (AIMD) rule for the updates. Unfortunately, such a technique cannot be directly applied to the cellular system due to the unavailability of the sensing mechanism, it can however give ideas for application of a similar mechanism for the power updates.

5.1.2 Contributions and Outline

We investigate a saturated system model, where a number of N users are always present within a wireless cell and try to gain access to the Base Station. An effort is successful when the user transmits a certain sequence, which is detected at the Base Station and at the same time no collision occurs. The event of collision will happen when the transmitted sequence of another user is also detected. Furthermore, LTE standards allow for orthogonal sequences randomly chosen by the users, so that even when two user signals are detected, access to both may be granted.

In our analysis the miss-detection probability and collision probability are left as unknown variables. However, higher power increases the chances for detection and reduces collision probability, whereas use of access (otherwise back-off) probabilities reduces the collision events. Transmission power and access probability are the user action pair.

After description of the action space and state space, the transition probabilities are given and the evolution of the system is described by a Markov Chain. The *event of dropping*, when the users exhaust the maximum number of efforts allowed, plays a crucial role. Unfortunately, due to the unknown expression for the success probability no steady-state analysis is possible. The above are analytically presented in Section 5.2.

What we can do however, is to choose the actions myopically optimal, in the sense that they optimize the expected change in one time-slot for some function of the state space. For this we introduce in our analysis the drift of a delay-related function. To motivate further our formulation, it is shown in the Appendix B how the solution of the drift minimization problem is related to the solution of an ideal Markov Decision Problem for optimal performance in the steady-state. Our problem formulation is found in Section 5.3.

The function chosen in this work is related to a sense of throughput, and *is chosen such that the ratio of dropped users can be minimized.* Other performance measures, by choice of an appropriate function, can also be incorporated within our analysis with slight variations.

To solve the problem online a protocol is introduced. Its steps are presented in Section 5.4. The BS collects measurements as well as user reports to estimate the unknown probabilities (miss-detection, contention, success) at the Base Station side, as well as the current number of users, which is actually unknown in a real system. After solution of an optimization problem and a close-loop control problem, the BS broadcasts two values, the current *contention level* and the current *power transmission level*, so that the users can update their action pair.

Numerical simulations for the performance of the protocol in a wireless cell are presented in Section 5.5. Advantages and trade-offs in dropping rate, delay and power expenditure are discussed and explicitly illustrated in plots. Finally, Section 5.6 concludes our work.

5.2 System Model

5.2.1 General Description

We consider an arbitrary but fixed total number of N users labeled by $n = 1, \ldots, N$ trying to randomly obtain access to a cell BS over the wireless channel. The time is slotted, with each slot interval normalized to 1 and indexed by t. At each time slot all users belonging to the user set have the possibility to access the channel by transmitting a preamble sequence (as specified in the LTE standards). There are two criteria that determine the success of an attempt.

- *The signal-to-noise ratio (SNR) at the BS exceeds a predefined detection threshold* γ_d. If the SNR is below the threshold, we assume that a miss-detection occurs and the user has to retry. The **detection miss probability (DMP)** can be written as the probability of an outage event

  $$Q_n^o\left(p_n, t\right) \quad = \quad \mathbb{P}\left[\mathrm{SNR}_n\left(p_n\left(t\right), h_n\left(t\right)\right) \leq \gamma_d\right] \tag{5.1}$$

  where p_n is the chosen transmission power and the probability is taken over the random channel quantity denoted by h_n and is i.i.d. over time t. In general we will consider that the BS does not approximate somehow the expression for outage. This is reasonable since the information over the user positions and the exact fading statistics is not known a priori.

- *No collision of transmitted signals occurs.* Typically in the slotted ALOHA protocol [Abr70], when more than one user attempts to access the channel during the same time slot a collision occurs and all affected users have to repeat the effort. In more recent wireless protocols, such as those suggested in LTE standards [3GPh], a pool of orthogonal sequences (e.g. Zadoff-Chu) is made available to all users. Each user chooses one sequence from this set randomly (uniform distribution) and the probability of collision can be made less than 1 when two users transmit simultaneously.

  In our model, the probability of collision is conditional on the transmission and the detection of signals at the BS side. That is, a user may collide only if he transmits at time slot t and his signal is detected. Assuming that N users transmit at time slot t with transmission probability vector $\mathbf{1}_N := [1, \ldots, 1]^T$ and k-out-of-N (we write $k \setminus N$) are detected, the overall **collision probability (CP)** - the probability that at least one collision occurs - is an increasing function of both N and k

  $$Q^c\left(N, \mathbf{1}_N, k, t\right) \tag{5.2}$$

  As in the case of the DMP we consider that the base station does not have an exact closed form expression to calculate the CP and the above quantity is in general unknown.

5.2.2 Action Space

There are *two actions* that user n can take for transmission at time slot t.

- The choice of the **transmission power level** $p_n(t)$, which influences the detection of the transmitted signal at the BS, as shown in (5.1) and eventually the collision probability (through the number of detections k). In general Q_n^o exhibits a monotone decreasing behavior with respect to power.
- The choice of the **access (or transmission) probability** $b_n(t)$ per user, at a given slot t. This influences the number of simultaneously transmitting users in the cell and therefore directly affects the collision probability in (5.2). The *back-off* probability simply equals $1 - b_n(t)$.

The set of actions for the entire system of N users at t is denoted by the $2N$-dimensional vector $\mathbf{A}(t) := \left[\boldsymbol{b}_N(t)^T, \boldsymbol{p}_N(t)^T\right]^T$. The action space per time-slot is denoted by $\mathbb{A}$ and is the Cartesian product $[0,1]^N \times [0, P_1] \times \ldots \times [0, P_N]$, where P_n is a given individual user power constraint per slot. Furthermore, $\tilde{\mathbf{A}} = \{\mathbf{A}(1), \ldots, \mathbf{A}(t), \ldots\}$.

Until the end of the subsection, we provide a discussion on the influence of choice for the back-off probability. In the definition (5.2) no back-off action is taken, $b_n(t) = 1$, $\forall n$ and all users transmit simultaneously. On the other hand, assigning $b_n(t) \leq 1$ to some users, displaces the transmissions in time and the effect of collision is mitigated. Since less than N users simultaneously compete for the access of the medium in some slot t, the collision probability is reduced. This can also be shown analytically.

The overall collision probability of N users present within the cell, with access probability N-length vector $\mathbf{b}_N$, $b_n \leq 1$ and exactly k users detected, equals

$$Q^c(N, \mathbf{b}_N, k, t) \quad = \quad \sum_{J=0}^{N} Q^c(J, \mathbf{1}_J, k, t) \cdot Q^t(\mathbf{b}_N, J \setminus N) \tag{5.3}$$

where $Q^t(\mathbf{b}_N, J \setminus N)$ is the probability that - given a probability vector $\mathbf{b}_N$ - exactly J-out-of-N users in the cell transmit. The equality follows from the total probability theorem, since the union of events $J = 0, \ldots, N$ transmissions exhaust the sample space. The transmission probability of $J \setminus N$ users equals

$$Q^t(\mathbf{b}_N, J \setminus N) = \sum_{l=1}^{L(N,J)} \prod_{i=1}^{J} b_{q_l^{J.i}} \prod_{j=1}^{N-J} (1 - b_{\hat{q}_l^{J.j}})$$

where the summation over l is taken over all possible $L(N, J) = \left(\begin{array}{c} N \\ J \end{array}\right)$ combinations (sampling without replacement) of J users transmitting and $N - J$ users remaining silent,

$q_l^{J.i}$ is the index of user i belonging to combination l that transmits and $\hat{q}_l^{J.j}$ is the index for the user j that does not transmit.

Proposition 5.1. *Given* $\mathbf{b}_N < \mathbf{1}_N$ *(the inequality means that* $b_n < 1$ *for at least one* n*) and exactly* $1 \leq k \leq N$ *detections, we have that*

$$Q^c(N, \mathbf{b}_N, k, t) \quad < \quad Q^c(N, \mathbf{1}_N, k, t) \tag{5.4}$$

Proof. : The events $J = 0, \ldots, N$ exhaust the sample space and we have that their probability sum equals $\sum_{J=0}^{N} Q^t(\mathbf{b}_N, J \setminus N) = 1$. Furthermore, for $J < k$ it holds $Q^c(J, \mathbf{1}_J, k, t) = 0$ since there cannot be more detections than transmissions. The higher the number of transmissions, the higher the collision probability, which means $Q^c(J, \mathbf{1}_J, k, t) \leq Q^c(N, \mathbf{1}_N, k, t)$, $\forall J$ and the inequality is strict for $J < k$. From (5.3) we have

$$\begin{aligned} Q^c(N, \mathbf{b}_N, k, t) \quad &< \quad Q^c(N, \mathbf{1}_N, k, t) \cdot \sum_{J=0}^{N} Q^t(\mathbf{b}_N, J \setminus N) \\ &= \quad Q^c(N, \mathbf{1}_N, k, t) \end{aligned}$$

which concludes the proof. ■

5.2.3 Success Probability, Failure Event and Dropping

From the above, success of a transmission is an event which occurs when (i) a user transmits, (ii) the user signal is detected and (iii) no collision occurs. In the use of orthogonal sequences/preambles, it suffices that no two users sharing the same sequence collide. In general, conditioned that a user transmits, the **success probability (SP)** equals

$$Q_n^s(N, k, \mathbf{b}_N, p_n, t) \quad = \quad (1 - Q_n^o(p_n, t)) \cdot (1 - Q^c(N, \mathbf{b}_N, k, t)) \tag{5.5}$$

Observe, that the success probability of a single user does not depend only on his own action set (b_n, p_n), but also on the choices of access probabilities of the other users, as well as the number of detected users k. The latter is further dependent on the transmission power chosen for $j \neq n$, so we can instead write

$$Q_n^s(N, \mathbf{b}_N, \mathbf{p}_N, t) \tag{5.6}$$

In the case of an unsuccessful effort the user may retry. Each user is constrained to at most M *access efforts* and the efforts are indexed by m. After M unsuccessful efforts the user is considered discarded and replaced by a new-coming one, so that the total user number in the system always remains equal to N. The same holds when a user leaves the system after success. Therefore, we say that the system is *saturated.* The number of users at effort m in time slot t is denoted by $X_m(t)$ and from the above it follows that

$$\sum_{m=1}^{M} X_m(t) = N, \quad \forall t. \tag{5.7}$$

We occasionally write in the following that a user at effort $m \in \{1, \ldots, M\}$ belongs to *user class m.*

5.2.4 System States and Transition Probabilities

We define the state of user n at slot t as the current transmission effort $S_n(t) \in \{1, \ldots, M\}$, whereas the system state as the N-dimensional vector

$$\mathbf{S}(t) = [S_1(t), \ldots, S_N(t)]^T. \tag{5.8}$$

Altogether, there are M different user states and M^N different system states (e.g for a cell with 10 users and maximum 5 efforts, the number is approximately 10 million). The entire state space is denoted by $\mathcal{S}$. It is easy to verify that the system state forms an N-dimensional Markov chain.

We group the transitions for each user into (a) returning to state 1 in case of transmission and success, (b) moving to the next effort in case of transmission and failure and (c) backing-off and remaining in the same state. The expressions for the transition probabilities are given below. (Dependence of the functions on other parameters except the time index is omitted for brevity of presentation.)

- For $1 \leq m < M$:

$$\mathbb{P}[S_n(t+1) = 1 | S_n(t)] = b_n(t) \cdot Q_n^s(t) \tag{5.9}$$

$$\mathbb{P}[S_n(t+1) = S_n(t) + 1 | S_n(t)] = b_n(t) \cdot (1 - Q_n^s(t)) \tag{5.10}$$

$$\mathbb{P}[S_n(t+1) = S_n(t) | S_n(t)] = 1 - b_n(t) \tag{5.11}$$

- For the user boundary state $m = M$:

$$\mathbb{P}[S_n(t+1) = 1 | S_n(t) = M] = b_n(t) \tag{5.12}$$

$$\mathbb{P}[S_n(t+1) = M | S_n(t) = M] = 1 - b_n(t) \tag{5.13}$$

A user in state M will either back-off, in which case he remains in the same state, or transmit. When a user transmits, he will either succeed or fail. In both cases the next state is set to 1, the user is removed from the system and is replaced by a new one so that the total number is always equal to N. The transition probabilities in

(5.12)-(5.13) for $m = M$ coincide with those for $m < M$, given by (5.9)-(5.11) when $Q_n^s(t) = 1$. In other words, to keep the system saturated, the Markov Chain evolves as if transmission at state M always results in success.

This is why, it is further important for the analysis to specify the user **dropping probability (DP)**

$$Q_n^d(N, \mathbf{b}_N, \mathbf{p}_N, M, t) \quad = \quad b_n(t) \cdot (1 - Q_n^s(t)) \cdot \mathbb{P}[S_n(t) = M] \tag{5.14}$$

If the exact expressions for the DMP and CP were available, it would be possible to calculate the steady state probabilities of the system, by forming the $M^N \times M^N$ transition probability matrix and using the Perron-Frobenius theory [BP94, Ch. 2 and 8] (for details see Appendix A.3). Since the number of states is finite, and for each user the probabilities (5.9)-(5.11) and (5.12)-(5.13) sum up to $\sum_{m=1}^{M} \mathbb{P}[S_n(t+1) = m | S_n(t)] = 1$ (stochastic matrix), a steady state with probability sum equal to 1 always exists, although certain states may be transient and have zero probability.

5.3 Problem Statement as Drift Minimization

Since the exact expressions for the detection miss probability Q_n^o as well as contention probability Q^c are unknown (hence the success probability Q_n^s, which appears in (5.9) and (5.10)), it is not possible to use the standard steady-state analysis as followed in [TK85], [BKMS87], [PYC08], [PVP$^+$07], [KL75] and [LYP$^+$09] (among others) to derive long-term performance measures and optimize the system. Even if this would be possible however, the solution of a system of such an immense number of variables would bring difficulties (remember the number of 10 million variables for $N = 10$ and $M = 5$). The same problems are met in a Markov Decision Problem (MDP) formulation, as followed e.g. in [LK75] and [dAF04].

Furthermore, in a realistic setting, we would like to propose a protocol, which takes into consideration the fact that within the wireless cell, users appear and leave the system after a while, whereas the fading situation changes unpredictably. These two factors greatly influence the miss-detection and collision probabilities, which do not remain fixed until infinity, but exhibit large fluctuations over time. This falls within the concept of SON's which should self-adapte and self-optimize the wireless system parameters as a reaction to such unpredictable changes from outside without human intervention.

For the above reasons we make use of the notion of *drift for the Markov Chain* under study, in order to achieve an improvement in the system performance by appropriate choice of actions. The idea of drift is commonly used in the literature of stability of systems with infinite states [TE93], [TE92], [NMR03], [NMR05]. In such cases, if we can find, for a given positive Lyapunov function, an action policy which keeps the drift negative for the entire state space - except possibly for some finite subspace - the system is guaranteed to remain stable. This comes from direct application of Foster's theorem (see [Asm00, Prop. 5.3(ii)]). Intuitively the negative drift gives the function of states a tendency to decrease in expectation at each step, as long as it is outside the aforementioned subspace, so that in the long run the value a state can take will not be unbounded (and the stability is guaranteed). In our case the state space is finite due to the finiteness of M. However, since the amount of users that exceed M efforts are eventually dropped, stability of the system refers to keeping the number of dropped users finite. (Alternative application of the drift minimization to a problem with $M \to \infty$ and no dropping does not change much the policy and results).

The drift equals per definition, the expected change in the Lyapunov function from t to $t+1$. By choosing an appropriate non-negative function of the system state $V(\mathbf{S}(t))$ related to some performance criterion, we can choose actions that optimize performance at each time-slot. Since it is impossible to know how the system will evolve in future slots, and since expressions for DMP and CP are not available, the best thing we can do is to provide an one-step look-ahead (**myopic**) policy for the system, given its current state and measurements performed on time t, which estimate unknown parameters. Specifically, given that the system state at t is $\mathbf{S}(t)$, the drift is defined as

$$D(V(\mathbf{S}(t)), \mathbf{A}(t)) \;:=\; \mathbb{E}\left[V(\mathbf{S}(t+1)) - V(\mathbf{S}(t)) \,|\mathbf{S}(t)\right] \tag{5.15}$$

and is also a function of the action set $\mathbf{A}(t)$, since the actions control the system state transition probabilities $p_{s_t \to s_{t+1}}$.

The function V to be used is the sum of user states and is linear. It can be rewritten as the sum of cardinalities of users at a state, weighted by their effort index.

$$V(\mathbf{S}(t)) = \sum_{n=1}^{N} S_n(t) = \sum_{m=1}^{M} m \cdot X_m(t) \tag{5.16}$$

A user who is currently at a higher effort, contributes more to the function, than users at lower ones. By **minimizing** the drift of such function we wish to choose appropriate actions in order to have success with as few efforts as possible. This has following objectives:

- keep a good trade-off between power consumption and delay until success per user
- diminish the proportion of users who are dropped
- maximize a notion of total system throughput

To understand the last point, observe that each user n contributes a ratio $\frac{1}{m_n^*}$ to the total system throughput if $m_n^* \leq M$ efforts are required for success and contributes nothing if the user is dropped. Consider now as a single virtual user, the set N of users in the network. By use of the Renewal-Reward theorem [GWB08], the long-term throughput of such a virtual user (considering only number of efforts and not the total number of time-slots required including user silence slots) will be the ratio $\frac{N}{\mathbb{E}[V(S)]}$. Alternative Lyapunov function could change the objective of the minimization, giving emphasis to total delay or power consumption and can be understood as alternative formulations of the same general problem and solution methodology.

Let us consider state-dependent, rather than user-dependent actions, in the sense that all users who are at class m in slot t should make the same choice for transmission power and back-off. The specific drift expression can now be derived to yield

$$\begin{aligned}
D\left(V\left(\mathbf{S}\left(t\right)\right), \mathbf{A}\left(t\right)\right) \quad &= \quad \sum_{n=1}^{N} \{1 \cdot \mathbb{P}\left[S_n\left(t+1\right) = 1 | S_n\left(t\right)\right] + \\
&\quad \left(S_n\left(t\right)+1\right) \cdot \mathbb{P}\left[S_n\left(t+1\right) = S_n\left(t\right)+1 | S_n\left(t\right)\right] + \\
&\quad S_n\left(t\right) \cdot \mathbb{P}\left[S_n\left(t+1\right) = S_n\left(t\right) | S_n\left(t\right)\right] - S_n\left(t\right)\} \\
&\overset{(5.9)-(5.13)}{=} \sum_{n=1}^{N} b_n\left(t\right) \cdot \left[1 - S_n\left(t\right) \cdot Q_n^s\left(N, \mathbf{b}_N, \mathbf{p}_N, t\right)\right] \\
&\overset{state\_dep.}{=} \sum_{m=1}^{M} X_m\left(t\right) b_m\left(t\right) \cdot \left[1 - m Q_m^s\left(N, \mathbf{b}_N, \mathbf{p}_N, t\right)\right] \qquad (5.17)
\end{aligned}$$

The drift minimization problem at each time slot t is

$$\begin{array}{ll} \mathbf{min} & D\left(V\left(\mathbf{S}\left(t\right)\right), \mathbf{A}\left(t\right)\right) \\ \mathbf{s.t.} & \mathbf{A}\left(t\right) \in \mathbb{A} \end{array} \qquad (5.18)$$

A further motivation to pose the problem as a drift minimization is provided in the Appendix B. It is shown that (5.18) is a myopic solution of an MDP with objective the minimization of the expected Lyaponov function at the steady-state (for $t \to \infty$). For the formulation and solution of the MDP, the expression for Q_n^s, $\forall n$ should be available and the channel/user statistics should remain unchanged over the entire time horizon.

What is needed to solve the above problem per slot? It follows from (5.17) that the following information should be available at the BS side:

1. The cardinality $X_m\left(t\right)$ of users at each effort m.

2. The current value of $Q_m^o\left(t\right)$ at each m.

3. The current value of $Q^c\left(t\right)$.

Using 2. and 3. and the product in (5.5) the actual value of $Q_m^s(t)$ can be obtained. Although the BS does not know these values it may estimate the variables and with it approximate the objective function, using *measurements* related to channel and service quality, as well as *information* reported directly by the user set. The goal is to use these estimates for optimization, in order to achieve significant performance gains, while keeping an additional overhead of exchanged information as small as possible.

In this way, a sequence of problems with different numbers of users, contention and miss-detection probabilities can be solved over time, which help the cell to follow and adapt to dynamic unpredictable changes. The steps of the proposed adaptive protocol are summarized in Table 5.1.

5.4 Five Steps of the Protocol

Before proceeding to the algorithm, we first discuss over the action pair of access probabilities and transmission powers. Considering the access probabilities, we adopt the approach in [GSS] (similar functions are also found in [LYP+09] and references therein), with per effort probability given by

$$b_m(t) = \min\left\{\frac{f(m)}{L(t)}, 1\right\}, \ \forall m. \tag{5.19}$$

Here and hereafter, L is called *contention level* and $f(m)$ is some fixed function of the transmission effort. In this way, a simple variable L can simultaneously define the entire set of transmission probabilities. By choosing f to be monotone increasing in m, priority is given to users with higher efforts, while such users obtain lower priorities when f is strictly monotone decreasing. Typical back-off protocols follow the exponential rule, which reduces by half the probability of accessing the channel after each failure, so in this case $f(m) = 2^{-m+1}$ and $b_1 = 1/L$. Other possible choice could be $f(m) = m^{-a}$, $a \in \mathbb{R}_+$ (in this work and the simulations to follow the case $a = 1$ is mostly used). Exponents $a > 1$ will lead to an overly conservative system with large delays for users in higher states, whereas $a << 1$ tends to treat users of all classes with the same priority. In the following, the expression in (5.19) will sometimes be replaced by $b_m(t) = f(m)/L(t)$ and the constraint $b_m(t) \leq 1$ is taken into account in the constraint set of the minimization problem.

We consider, furthermore, the transmission power to vary per effort as a ramping function. This approach is often considered in practice (for related approaches, the reader is referred to [AFG+SA] and references therein). The power level for the first effort is given by p and for all efforts by the expression

$$p_m(t) = p(t) + (m-1) \cdot \Delta p, \ \forall m \tag{5.20}$$

where Δp is the ramping step with a fixed (tunable) value. Thus, analogously to the case of the backoff probabilities, the vector of power actions can be defined by appropriate choice of the *power level* $p\left(t\right)$ per time slot.

5.4.1 Step 1: Measurements and User Reports

When users attempt to randomly access the channel, we assume that the BS counts the overall number of detected user efforts, as well as the overall number of successful efforts. Given an observation window of length W, both the quantities depend on the time interval $[t-W+1,t]$ and are denoted by $N_d\left(t\right)$ and $N_s\left(t\right)$ respectively. Furthermore, after every successful effort, the users are assumed to *report* to the BS, the total number of trials required to get access. In this way, the BS can keep track of the number of successes at effort m, within the observation window, denoted by $n_{s,m}\left(t\right), \ \forall m$. The reports over the success state also provide information over the overall number of transmissions of users being at some state m. As an example, if within the observation period two users report success at effort 3 and 2 respectively, the BS can estimate the number of transmissions at state $m=1$ by 2, at $m=2$ by 2 and at state $m=3$ by 1, without considering users that have yet not declared success, or are dropped. We denote these estimates by $n_{t,m}\left(t\right), \ \forall m$ and their sum, which equals approximately the number of access efforts within the observation window, by $N_t\left(t\right) = \sum_{m=1}^{M} n_{t,m}$. Altogether, the set of gathered empirical information, updated per time slot, is represented by

$$\mathcal{I}\left(t\right) := \{N_d(t),\ N_s(t),\ N_t(t),\ n_{s,m}(t), \forall m,\ n_{t,m}(t), \forall m\} \tag{5.21}$$

5.4.2 Step 2: Estimation of Unknowns in the Objective function

Using the above counters, we can now approximate the unknowns in the expression (5.17) that are briefly discussed in points 1. - 3. in the previous Section.

As far as the unknowns in 2. and 3. are concerned, the actual overall contention probability $Q^c\left(t\right)$ and per effort success probability $Q^s_m\left(t\right)$ in (5.5), can be estimated by contention and success **rates**, an idea which has already appeared in [AFG$^+$SA]. Observe that the additional information about the per effort miss-detection probability $Q^o_m\left(t\right)$ cannot be deduced from the above measurements. What can be calculated, instead, is an overall rate of miss-detection (DMR), without differentiating between efforts, which we denote by

$R^o(t)$.

$$R^c(t) = 1 - \frac{N_s(t)}{N_d(t)} \qquad (contention\ rate) \tag{5.22}$$

$$R^s_m(t) = \frac{n_{s,m}(t)}{n_{t,m}(t)}, \quad \forall m \quad (success\ rate\ per\ effort) \tag{5.23}$$

$$R^o(t) = 1 - \frac{N_d(t)}{N_t(t)} \qquad (miss - detection\ rate). \tag{5.24}$$

Regarding the number of users currently within the cell (discussed in 1.) and their estimation, we proceed as follows. Instead of attempting to find integer values, we consider arrival rates. As the total arrival rate of users we consider the ratio $\frac{N_s(t)}{W}$, which is the time dependent ratio of accepted users, divided by the observation window. The above is used under the assumption that only a very small fraction of the users are dropped throughout the process, so that almost all users appearing within the cell, will eventually have at some point a success. Taking dropped users into account requires an additive correcting term that may be deduced from empirical observations.

The window is considered long enough, so that the resulting success rates per state, $R^s_m(t)$ in (5.23), approach the actual success probability per effort. These can replace the entries in the one-step transition probability matrix in equations (5.9)-(5.11) and (5.12)-(5.13). The steady state probability distribution is found by solving the system $\pi = \pi \cdot \hat{\mathbf{P}}_{\mathbf{M}}$, where π is the row vector of the unknown probabilities for the M states with $||\pi||_1 = 1$ and $\hat{P}_M$ is the transition probability matrix. The solution equals

$$\pi_1(t) = \left(1 + \sum_{i=2}^{M} \frac{b_1}{b_i}(1 - R^s_1(t)) \cdot \ldots \cdot (1 - R^s_{i-1}(t))\right)^{-1} \tag{5.25}$$

$$\pi_m(t) = \pi_1(t) \cdot \left(\frac{b_1}{b_m}(1 - R^s_1(t)) \cdot \ldots \cdot (1 - R^s_{m-1}(t))\right), \ 2 \leq m \leq M. \tag{5.26}$$

The ratios of the unknown backoff probabilities b_1/b_m are involved in the expression above. From the previous discussion $b_1/b_m = f(1)/f(m)$, which is known since the function f is chosen a priori. With these observations and definitions at hand, we can estimate the user arrivals per effort according to

$$\frac{X_m(t)}{W} \approx \pi_m(t) \cdot \frac{N_s(t)}{W} \tag{5.27}$$

where the π_m's are the probabilities given by (5.25) and (5.26).

5.4.3 Step 3: Solving the Problem

Once step 2 is performed, we can formulate the objective function to approximately solve problem (5.18) and with it find the optimal actions per time slot. To this end, we break down the problem into two subproblems and propose two sub-algorithms based on the measurements and estimated quantities described above.

Backoff Probability Problem: The objective function at the base station is estimated by

$$\hat{D}\left(V\left(\mathbf{S}\left(t\right)\right),L\left(t\right)\right):=\frac{1}{L\left(t\right)}\cdot\left[\sum_{m=1}^{M}\pi_{m}\frac{N_{s}\left(t\right)}{W}f\left(m\right)\cdot\left(1-m\cdot R_{m}^{s}\left(t\right)\right)\right], \tag{5.28}$$

where the success probability Q_m^s is substituted by the success rate R_m^s in (5.23) and the average user number $\frac{X_m}{W}$ by the expression in (5.27). As long as such estimates are close to the actual values and are considered reliable, the BS can solve a problem with parameters adapted to the changing environment.

When the expression in brackets above $[\ldots]$ is positive, the objective function is convex and decreasing in the contention level variable L (behaves as $+\frac{1}{L}$). When $[\ldots]$ is negative, the objective is concave and increasing in L (behaves as $-\frac{1}{L}$). Due to the monotonicity and concavity/convexity, the optimization will have as a result either maximum or minimum value of L depending on the sign of the term inside the square brackets.

In the following we provide the boundary values $L_{\min}$ and $L_{\max}$ of the domain of L. The lower bound on L follows from the fact that all access probabilities are less than or equal to 1:

$$\frac{f\left(m\right)}{L\left(t\right)}\leq 1,\ \forall m \quad\Rightarrow\quad L\left(t\right)\geq L_{\min}:=\max\left\{f(m)\right\}. \tag{5.29}$$

To obtain an upper bound, we further provide a constraint on the probability of a time slot being idle (no user transmits). This probability is less than or equal to $\mathcal{A}$, which is a design factor for the system.

$$\begin{aligned}\mathbb{P}\left[IDN\right]=\prod_{m=1}^{M}\left(1-\frac{f(m)}{L\left(t\right)}\right)^{\frac{X_{m}(t)}{W}} &\leq\mathcal{A}\qquad\Rightarrow\\ \sum_{m=1}^{M}\pi_{m}\frac{N_{s}\left(t\right)}{W}\cdot\log\left(1-\frac{f(m)}{L\left(t\right)}\right) &\leq\log(\mathcal{A})\quad .\end{aligned} \tag{5.30}$$

The left handside is increasing with L, thus the inequality provides an upper bound on L. If we solve (5.30) for equality, we then derive the value of $L_{\max}$. Notice furthermore that, all values of L within the interval $[L_{\min}, L_{\max}]$ are feasible solutions of the contention level.

Proposition 5.2. *Considering the problem of minimizing $\hat{D}$ in (5.28) subject to the upper and lower bound constraints on L, the following necessary and sufficient optimality conditions hold:*

- *if* $\left[\sum_{m=1}^{M}\pi_m\frac{N_s(t)}{W}f\left(m\right)\cdot\left(1-m\cdot R_m^s\left(t\right)\right)\right]\geq 0$ *then the optimal contention level equals* $L_{\max}$ *and is found by solving*

$$\sum_{m=1}^{M}\pi_m\frac{N_s\left(t\right)}{W}\cdot\log\left(1-\frac{f(m)}{L^*\left(t\right)}\right) \quad = \quad \log(\mathcal{A}) \tag{5.31}$$

- *if* $\left[\sum_{m=1}^{M}\pi_m\frac{N_s(t)}{W}f\left(m\right)\cdot\left(1-m\cdot R_m^s\left(t\right)\right)\right]< 0$ *then the optimal contention level equals* $L_{\min}$

$$L^*\left(t\right)=\max\left\{f(m)\right\}. \tag{5.32}$$

Power Control Problem: In order to identify optimal transmission levels, one could proceed along similar lines as above, to formulate an optimization problem, given the back-off probabilities $f(m)/L^*(t)$ and the contention rates $R^c(t)$ from (5.22). In order to determine the objective function based on (5.17), which is denoted by $\tilde{D}\left(V\left(\mathbf{S}\left(t\right)\right),p\left(t\right)\right)$, the closed form expression for the detection-miss probability $Q_m^o\left(t\right)$ as a function of power may be necessary. It is however unlikely that the channel's fading behavior in practical systems can be accurately represented by a closed-form expression, especially since in the random access cellular system the user position is not known to the BS.

A different approach - which is adopted here - is to use a Multiplicative Increase Additive Decrease (MIAD) control rule, as in the case of congestion control protocols in TCP [CJ89]. In this way, the BS reacts to the change of the estimated DMR stepwise, by increasing or decreasing the power level $p(t)$ per time slot, depending on the current value $R^o\left(t\right)$. We set two levels of action, a high detection-miss level DMR^H and a low one DMR^L. The control loop then works as follows: When DMR^H is exceeded, the power level is increased by multiplication with a tunable factor $1+\delta_1$. This action increases considerably the transmission power since miss-detection is highly non-desirable. When the ratio falls under the low level DMR^L, which is considered satisfactory for the system performance, the power is reduced in a conservative way, to reduce the energy consumption on the mobile devices, by subtracting a constant tunable amount of δ_2. For instance δ_2 can be set equal to the ramping step Δp in (5.20). The control loop is then described by the power updates

$$p^*\left(t\right) \quad = \quad \begin{cases} p^*\left(t-1\right)\cdot\left(1+\delta_1\right), & \text{if } R^o\left(t\right)>\mathrm{DMR}^H \\ p^*\left(t-1\right)-\delta_2, & \text{if } R^o\left(t\right)<\mathrm{DMR}^L \end{cases}. \tag{5.33}$$

Obviously, updates on the per-effort ramping steps or user-specific power control could be much more beneficial instead of the update in the global power level $p\left(t\right)$. Furthermore, it is obvious that by varying $p\left(t\right)$ globally, power consumption will increase not only for users in higher efforts but also for those in their first effort, which may not be necessary. However, there are certain difficulties in providing a different type of feedback. Most importantly, there is no user channel state information available at the BS and channel adaptation is impossible. Furthermore, based on the possible approximations that - given the measurements and the reports - are suggested, only a global miss-detection rate R^o can be estimated in (5.24) and no state-specific or user-specific rates (say R^o_m). We cannot approximate, in other words, the rate of miss-detection for a user at different states and as a result we cannot suggest different state-dependent power levels. Finally, state-dependent power control would increase considerably the feedback information broadcast to all users. For all the above reasons, the suggestion of the MIAD rule was considered more appropriate.

5.4.4 Step 4 and 5: Broadcast of Information to the Users and Action Calculation

The last two steps of the proposed algorithm involve the broadcasting of the action-related information to the users and the choice of appropriate actions by them. The broadcast information includes the pair consisting of the contention level and the power level

$$\mathcal{J}(t) \quad := \quad \{L^*(t), p^*(t)\}. \tag{5.34}$$

Let us assume that the expressions in (5.19) and (5.20) for the success probability and the power level per effort are known a priori to the mobile stations. Since each user is aware of its current individual state $S_n(t)$, calculation of its own action pair is possible, according to

$$A_n(S_n(t), \mathcal{J}(t)) = (b_n(t), p_n(t)) = \left(\frac{f(S_n(t))}{L^*(t)},\ p^*(t) + S_n(t)\Delta p\right). \tag{5.35}$$

Note that if the required power and access functions ($f(\bullet)$ and the ramping step Δp) is not available at the mobiles, the BS could broadcast the entire vector of computed transmission powers and access probabilities to the users so that they choose the actions according to their current effort.

A remark considering implementation issues of such protocols is that the updates of these two levels are not expected to take place very frequently, but rather only at the rate of estimated change of user traffic and fading conditions. Furthermore, user reports and broadcast feedback from the BS is already suggested in standardization reports, so that the proposed protocol complies fully with the existing standardization literature [3GPf], [3GPa], [3GPh], without introducing additional protocol information.

5.5 Numerical results

5.5.1 Description of the Simulations Setting

The proposed algorithm has been implemented in a single cell scenario. The users are randomly positioned, with a 2D uniform distribution and the algorithm is initially evaluated for the cases of $N = 1, 2, \ldots, 14$ [users/time slot] present in the cell. Considering the transmission scenario, each user randomly chooses at each attempt one sequence, out of a pool of 10 orthogonal sequences, and transmits with a chosen backoff probability and transmission power. The number 10 is used for simulation purposes, whereas the actual number suggested in the LTE literature equals 64; however not all users have access to the entire pool of sequences (see [3GPf]) since the sequence allocation procedure is more complicated than the simple uniform choice we use here.

The signal experiences path loss due to the user-BS distance. Fast fading is initially not modeled (this will be considered in the second part of the Section for the power consumption evaluation) but the channel is considered additive white Gaussian noise (AWGN) with noise mean equal to -133.2 dBm. We have to note that in case fast-fading were also implemented, a further randomness in the channel would affect the signal detection and the protocol performance. To keep things simple, we consider first only the randomness of user positioning which affects the slow-fading coefficients - also unknowns during the procedure. The evaluation of the protocol's performance will not change much by adding more randomness factors.

An effort is successful when among the detected sequences there exists no pair that collides, in the sense that no two detected users choose the same sequence for transmission. A user is dropped when the effort fails at the maximum access effort $M = 5$. After a success or an event of dropping, users are removed from the waiting-for-transmission list, and the same number of newly arriving users are added, each given a random position on the plane.

Power and access probability for the users are computed per slot equal to the action pair in (5.35), for $f(m) = m^{-1}$. The choice of exponent -1 is not conservative (whereas a higher exponent would be) while at the same time it takes class differentiation into account. Important is to notice that the expression of the function f greatly affects the delay. On the other hand, the delay can be controlled by the parameter $\mathcal{A}$ which is system-operator-dependent and tunes the expected idle period. The set of values for the parameters of the system simulation are summarized in Table 5.2.

Several factors for the protocol design have been left open for choice. One of them, as mentioned already, has been the desired idle probability $\mathcal{A}$. The higher factor $\mathcal{A}$ is, the more the delay suffered by the system but the higher the benefits in dropping rate and power consumption are. Other important parameters are the steps δ_1, δ_2 and bounds DMR^H, DMR^L of the MIAD rule, the access function f and the adaptive window length W, which defines how fast should the protocol adjust to environmental changes. A summary of these tunable factors and how they are chosen within the simulation setting under consideration is provided in Table 5.3.

5.5.2 Comparison to a Fixed "Open Loop" Power Fixed Backoff Protocol

The suggested algorithm is compared to a scenario, where access probabilities and target power are held fixed, while the ramping step for the transmission power is predifined and same for all efforts. The fixed scenario is in other words an "open-loop" control scheme, with predefined constant $(p, \Delta p)$. The choice for the fixed backoff probability in the comparison scenario, equals $[b_1, b_2, b_3, b_4, b_5] = [0.5, 0.4, 0.3, 0.2, 0.1]$ and is such that the average occurance of an idle slot is less than $\mathcal{A} = 0.05$, hence the channel is kept busy with user efforts for access during most of the time . In this sense, the comparison between the adaptive-protocol suggested and a fixed protocol is more fair for a tunable factor of $\mathcal{A} = 0.05$ or less. How the average idle probability changes between $\mathcal{A} = \{0.05, 0.25, 0.5\}$ and the fixed case can be seen in Fig. 5.1. We refer the reader to the Parameter Table 5.2 for the actual values used throughout these simulations. The above fixed scenario is denoted by (FPFB) for Fixed Power Fixed Backoff. Two types of protocols are used for performance comparison:

- **Fixed Power Dynamic Backoff (FPDB) protocols**. In this case the "open loop" power control of the protocol is the same as in the fixed scenario FPFB case. The backoff mechanism adapts to measurements as suggested in the protocol description of this work (Paragraph 4.3, Backoff Probability Problem).

- **Dynamic Power Dynamic Backoff (DPDB) protocols**. In this case both backoff and power are adapted as the protocol suggests in Paragraph 4.3. The backoff comes from the solution of the drift minimization problem, while the target power p is adapted according to the MIAD rule.

5.5.3 Performance Evaluation: Lyapunov Function and Number of Efforts

The performance of the scheme and its comparison to the fixed scenario FPFB is initially illustrated in the plots of the performance metric in Fig.5.2 and the plots of the average number of access efforts until success in Fig.5.3. The two figures show a close relation to each other, due to the choice of the specific Lyapunov function V. Since V was chosen as the sum of user efforts, lower values translate into better performance for the protocol. In all six curves, our protocol outperforms the FPFB scenario in the metric chosen as well as in the average number of user efforts. Furthermore, all DPDB cases show improved performance compared to FPDB, given a certain value of the parameter $\mathcal{A}$. The higher the value of tunable factor $\mathcal{A}$, the better the performance and the less the average efforts required up to packet reception.

5.5.4 Performance Evaluation: Delay, Power Consumption and Dropping Rate

The three most important performance measures in random access that can illustrate the improvements of the suggested protocol are the total delay suffered by a packet until success (including backoff slots), the total transmission power used until success as well as the percentage of users dropped because the maximum number M of efforts is exceeded. These are shown in Fig.5.4(a), 5.4(b), 5.5(a), 5.5(b) and 5.6(a), 5.6(b) respectively, for (a) the FPDB case and (b) the DPDB case.

From the plots, it is illustrated how an increase of the parameter $\mathcal{A}$ influences positively power consumption and dropping rate at the cost of delay. Furthermore, the DPDB schemes perform better than the FPDB schemes in terms of delay and dropping, but have a cost in power consumption. Altogether, the performance of the protocol is tunable, to the requirements of the service provider. If the delay is not an issue, power can be considerably saved and the number of users dropped is reduced. As long as delay becomes an issue, transmission power can still be saved by using only the FPDB protocols. The dropping rate is also improved in such a case.

The most important observation is the fact that the suggested protocol in all cases considerably reduces the dropping rate of the incoming users. Hence, the random access resource is better exploited than in the FPFB case. This is due to the specific choice of performance function that we chose to incorporate in the drift minimization (sum of states). Other functions could potentially minimize different system performance measures (e.g. power or delay). Dynamic backoff, in our protocol, generally allows the system to remain stable - in the sense that the rate of dropped users does not tend to ”explode” - for a higher value of N. The behavior of this measure also improves for higher $\mathcal{A}$, which is reasonable since allowing a higher idle probability, distributes the transmissions of users among a larger number of time-slots.

A more detailed comparison of the schemes is given in the following figures. Specifically, Fig.5.7(a) and Fig.5.7(b) illustrate the beneficial use of the MIAD power control for the detection miss ratio, which leads to a drastic reduction of the average number of miss-detected signals in the system for DPDB protocols. Obviously the miss-detection curves for FPDB are similar to the FPFB case, since no power control is applied. Furthermore, considering the contention ratio CR, both Fig.5.8(a) and Fig.5.8(b) show benefits compared to the fixed FPFB case. Interestingly, the DPDB cases are slightly worse than the FPDB. This is because a higher number $N_d(t)$ is detected for the same window size W, so that the CR calculated as in (5.22) appears higher.

5.5.5 Protocol Temporal Adaptation to Channel Fluctuations and Deep Fades

In the current subsection, we further illustrate the performance of our protocol - which operates with parameters given in Table 5.3 - for a scenario with fluctuations and abrupt changes of the fading conditions. Such investigation shows how fast and with which cost in power expenditure can the protocol adapt to environmental changes. Specifically, we use a factor β to multiply the long-term fading of each user. Initially the factor has an expectation 1 and its value fluctuates uniformly within the interval $[0.7, 1.3]$. After a certain time-interval we initiate a sudden deterioration of the channel to an average of 0.8, which returns to 1 after some time. The realization of such fading scenario for a given user is presented in Fig. 5.9(a).

Very important here is to show how the protocol performs over time and adapts to the changes. Compared to the fixed power scenario, our suggested protocol can react very fast to the changes by an increase in power consumption during the period of the deep fade, which keeps the DMR always within the defined interval $\left[\mathrm{DMR}^L, \mathrm{DMR}^H\right]$. This can be observed in Fig.5.9(b) and Fig.5.9(c).

5.5.6 Protocol Temporal Adaptation to Traffic Load Fluctuations

To complete the evaluation of our protocol, we illustrate the temporal behavior of the DPDB protocol compared to the fixed case FPFB, when the arrival traffic load varies with time. The chosen idle parameter is $\mathcal{A} = 0.25$. All other parameters follow Table 5.3, noticing that the window size is $W = 200$ slots. Specifically, we consider a scenario where from 0 to 1000 time slots the users arrive in the cell with an average value of 5 [users/sec], the average arrival rate increases to 10 [users/sec] from 1000 to 2000 slots and reduces again to 10 [users/sec] from 2000 to 3000 slots. The traffic scenario over time can be found in Fig. 5.10(a) and the temporal evaluation of FPFB and DPDB in Fig. 5.10(b), 5.10(c), 5.10(d), 5.10(e).

Specifically, the improvement of DPDB compared to the FPFB over the performance measure is evident in Fig. 5.10(b). As a consequence of the chosen performance function, a considerable improvement in the dropping rate is shown in Fig. 5.10(e), where the dropping rate, even with the abrupt change of the average traffic load from 5 to 10 [users/slot] does not exceed the 0.1% for DPDB. This is achieved with almost zero cost in power consumption as shown in Fig. 5.10(d) and usually even better delay as shown in Fig. 5.10(c) compared to the FPFB case. As the plots show, our protocol functions as promised with reference to the dropping rate and hence the optimal exploitation of the available resources, in order to serve the maximum possible rate of incoming users.

One may observe an overshoot and a delayed response in Fig. 5.10(c) and 5.10(d) starting at the beginnings of the abrupt changes from 5 to 10 [users/sec] and from 10 to 5 [users/sec]. The reason is the choice of a long window $W = 200$ slots, and the power control factors δ_1 and δ_2 which we left as in the previous evaluation plots - for coherence reasons - and shown in Table 5.3. If we optimally select these values and choose the parameter $\mathcal{A}$ appropriately, we can adapt our protocol to different scenarios of traffic load variations. Furthermore, we may choose whether we wish to save in power or delay, while aiming for maximum user service, but this depends on the system needs.

5.6 Conclusions

We have suggested a dynamically adaptive protocol which updates the user access probabilities and transmission powers in cellular random access communications for LTE systems, with the aim to maximize the served load of the cell. The protocol is based on measurements and user reports at the base station side, which allow for an estimation of the number of users present within the cell, as well as the quantities of detection-miss and contention probability. The protocol updates take place per time slot in a myopic fashion. By solving a drift minimization problem for the contention level and using closed loop updates for the transmission power level by a MIAD rule, the BS coordinates the actions chosen by the users, by broadcasting the pair $(L^* (t), p^* (t))$.

The protocol was constructed based on a specific choice of performance function - *the sum of system states*. This function aimed at maximizing the usage of the restricted random access resource in the cellular system and consequently at minimizing the ratio of dropped users. Simulations results have shown the considerable performance increase of the protocol with minimum cost and occasionally even benefit in delay and power consumption. The performance of our protocol is tunable with paramaters that can be controlled by a system designer, such as the idle parameter $\mathcal{A}$ and the power steps δ_1, δ_2 and Δp to achieve the desired performance depending on the actual scenario.

The algorithmic steps, together with the methodology of the drift minimization for a certain measure of interest, provide a general suggestion to treat problems of self-organization in wireless networks. Considering the specific scheme, a large variation of algorithms can be extracted, by choosing e.g. some different state function for the performance measure, or by introducing other kinds of user reports, which may provide more information to the central receiver, at the cost of increase in signaling. Furthermore, a larger action set can definitely provide a higher performance, compared to the proposed one - which introduces two possible values for the contention level (high/low) and two actions for the power level (increase/decrease). Even in this scheme however, which is characterized by an “economy” of signaling and information exchange, the results - as illustrated by numerical examples - are very beneficial, especially as the user number in the cell increases.

TABLES

Table 5.1: GENERAL SELF-OPTIMIZATION ALGORITHM

STEP 1	Gather empirical information $\mathcal{I}$ at the BS.
STEP 2	Estimate unknown factors (see 1. - 3. above).
STEP 3	Solve the resulting optimization problem in (5.18).
STEP 4	Broadcast action-related information $\mathcal{J}$.
STEP 5	Calculate at the user side the required actions, based on $\mathcal{J}$.

Table 5.2: PARAMETER TABLE

Parameters	Value
Wireless Network	Single cell
User distribution	Uniform within cell
Number of users in cell	$\{1, 2, \ldots, 14\}$
Sequence pool size	10
Fixed Tx Power	250 mW
Power ramping step Δp	20 mW
Maximum Tx Power	500 mW
Path loss PL	$128.1 + 37.6 \log(D\ km)$ dB
Noise	-133.2 dBm
SNR threshold	8 dB
Maximum effort M	5
Fixed backoff probability	$[0.5, 0.4, 0.3, 0.2, 0.1]$
Number of slots	15000 slots

Table 5.3: TUNABLE FACTORS TABLE

Tunable Factors	Value
Window length W	200 slots
Backoff factor A	$\{0.05, 0.25, 0.5\}$
Access Function $f(m)$	m^{-1}
Power control factor δ_1	2×10^{-4}
Power control factor δ_2	8 mW
DMR^H	3.5%
DMR^L	2.5%

FIGURES

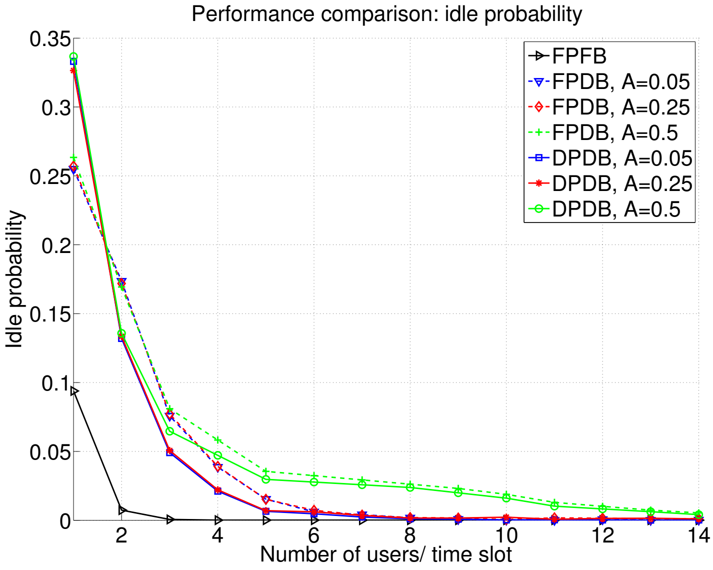


Figure 5.1: Comparison of the average occurence of idle slot per scheme. The dynamic scenario with $\mathcal{A} = 0.05$ is the closest to follow the chosen fixed one.

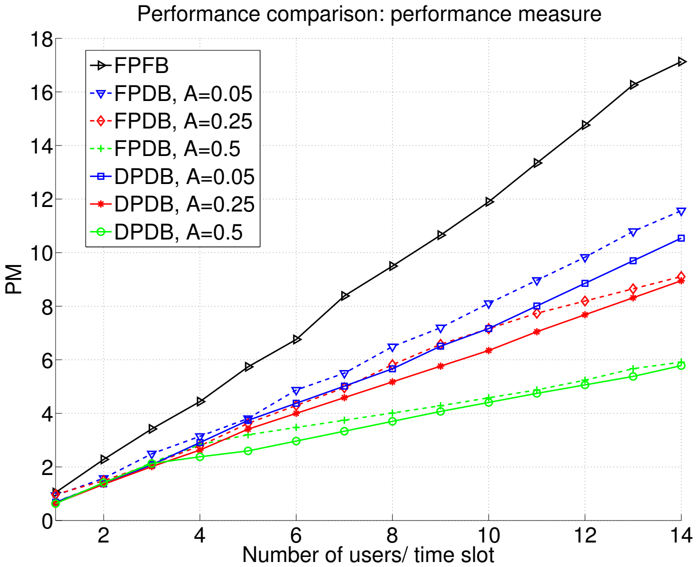


Figure 5.2: Comparison of performance measure, equal to the chosen function V as $t \to \infty$. The measure improves with increasing idle probability bound $\mathcal{A}$. Furthermore, all DPDB schemes outperform the FPDB ones.

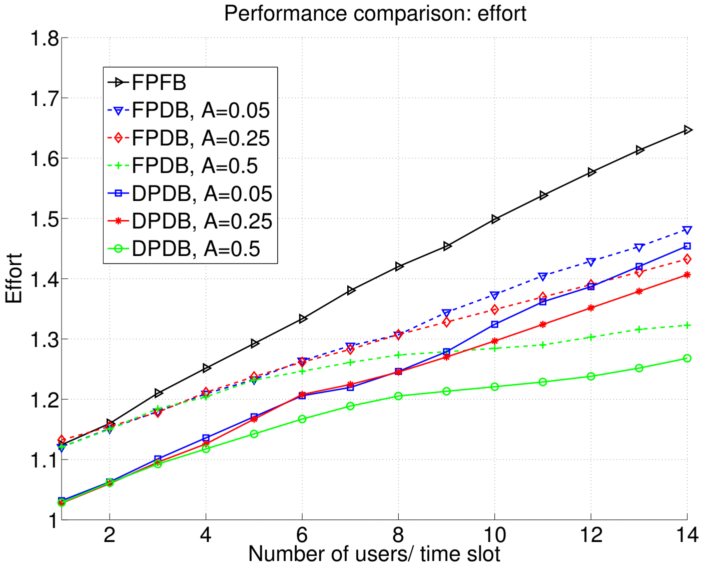
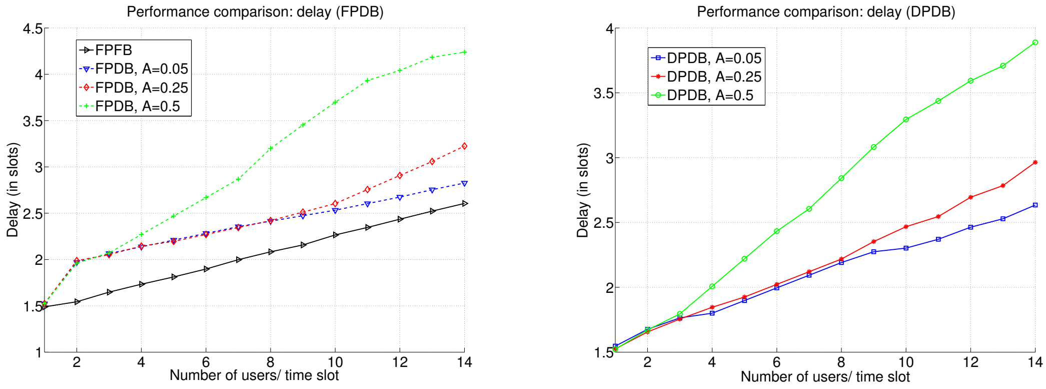


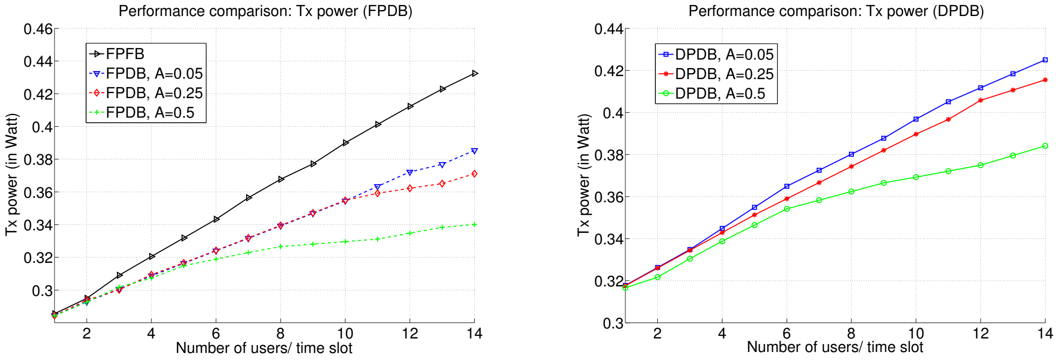
Figure 5.3: Comparison of the average number of efforts until success. The behaviour of these curves follows closely the performance metric curves, due to the specific choice of the Lyapunov function V as sum of user states.



(a) Total delay in FPDB protocols.

(b) Total delay in DPDB protocols.

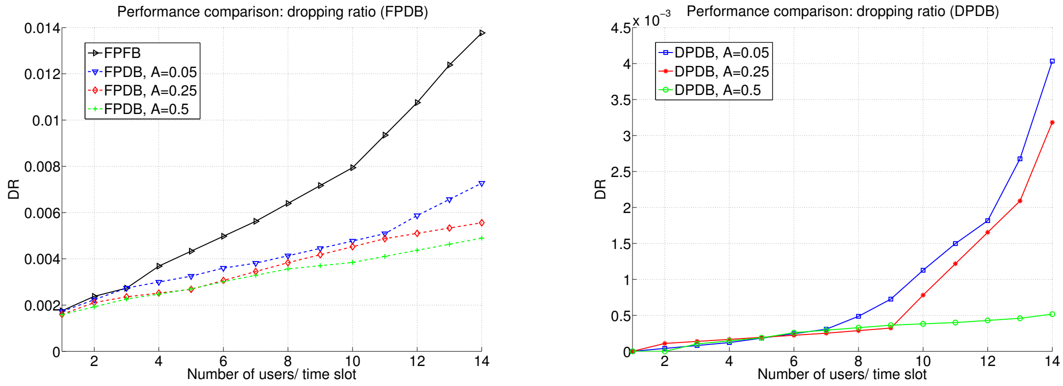
Figure 5.4: Evaluation of total average delay up to success (including backoff slots) in the case of (a) FPDB protocols and (b) DPDB protocols. The higher the parameter $\mathcal{A}$, the higher the allowed delay. For $\mathcal{A} = 0.05$, the protocol delay approaches the one of the FPFB protocol. In general power control improves the delay.



(a) Tx power in FPDB protocols.

(b) Tx power in DPDB protocols.

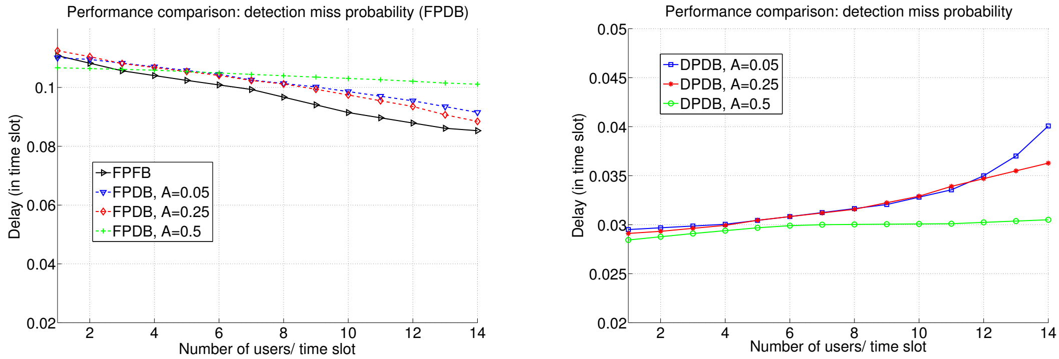
Figure 5.5: Evaluation of average Tx Power consumption up to success in the case of (a) FPDB protocols and (b) DPDB protocols. In the case of FPDB, the consumed power is always lower than the FPFB case. Both cases exhibit benefits in Tx power.



(a) Dropping rate in FPDB protocols.

(b) Dropping Rate in DPDB protocols.

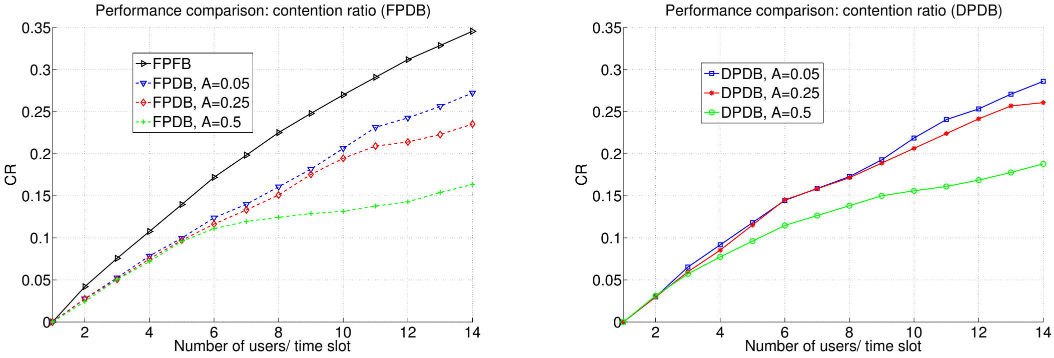
Figure 5.6: Comparison of the average dropping rate (DR) in the case of (a) FPDB protocols and (b) DPDB protocols.. The abrupt increase of the rate after a certain user number is an indicator that the system is not anymore stable for a further increase in the cell user number. Higher values of $\mathcal{A}$ can increase the point when the instability appears, at the cost of delay. (For a single user, the dropping rate may be non-zero if the event of miss-detection occurs M consecutive times due to bad channel conditions and poor transmission power.)



(a) Miss-detection rate in FPDB.

(b) Miss-detection rate in DPDB

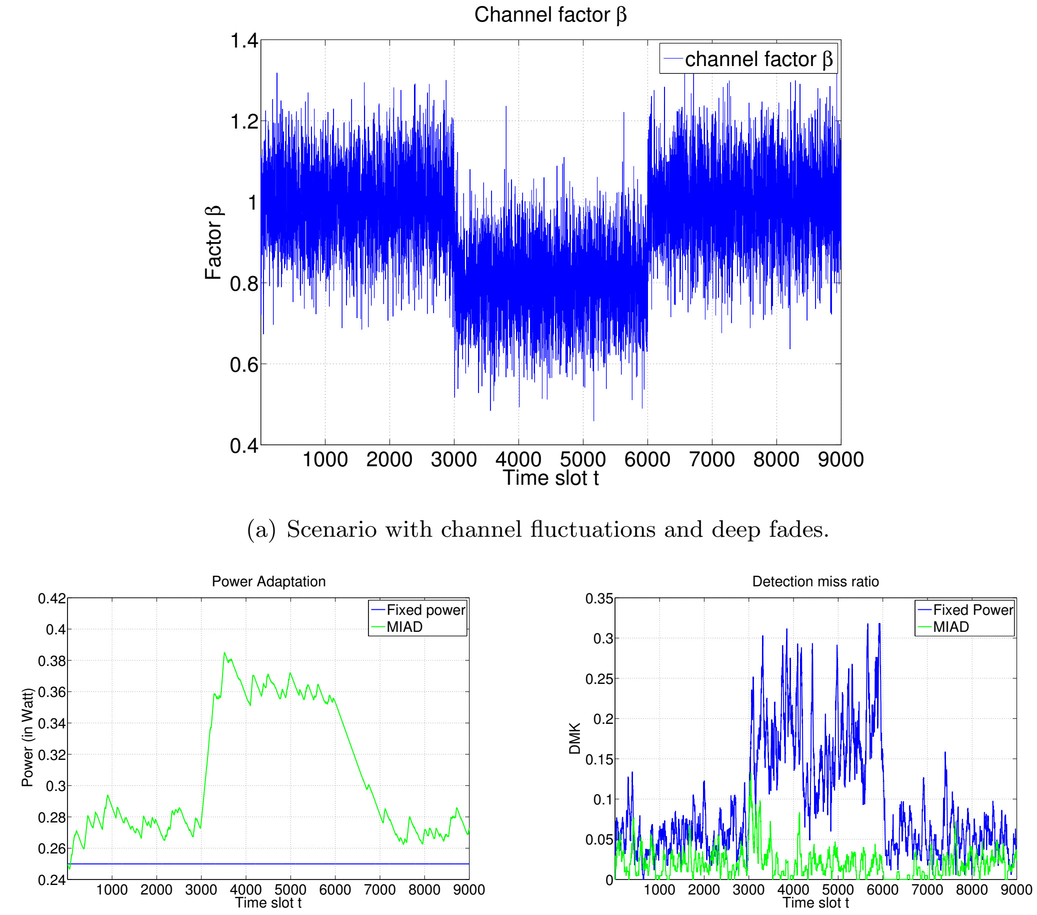
Figure 5.7: Comparison of miss-detection rate DMR for the two protocols (a) FPDB and (b) DPDB. Benefits are evident only in the case (b) where the MIAD rule is applied.



(a) Contention rate rate in FPDB.

(b) Contention rate in DPDB

Figure 5.8: Comparison of contention rate CR for the two protocols (a) FPDB and (b) DPDB. Both schemes exhibit improvements compared to the FPFB case, due to the backoff optimal choices. The case DPDB is slightly worse than the FPDB due to the fact that a larger number of packets are detected, so that the CR appears lower.

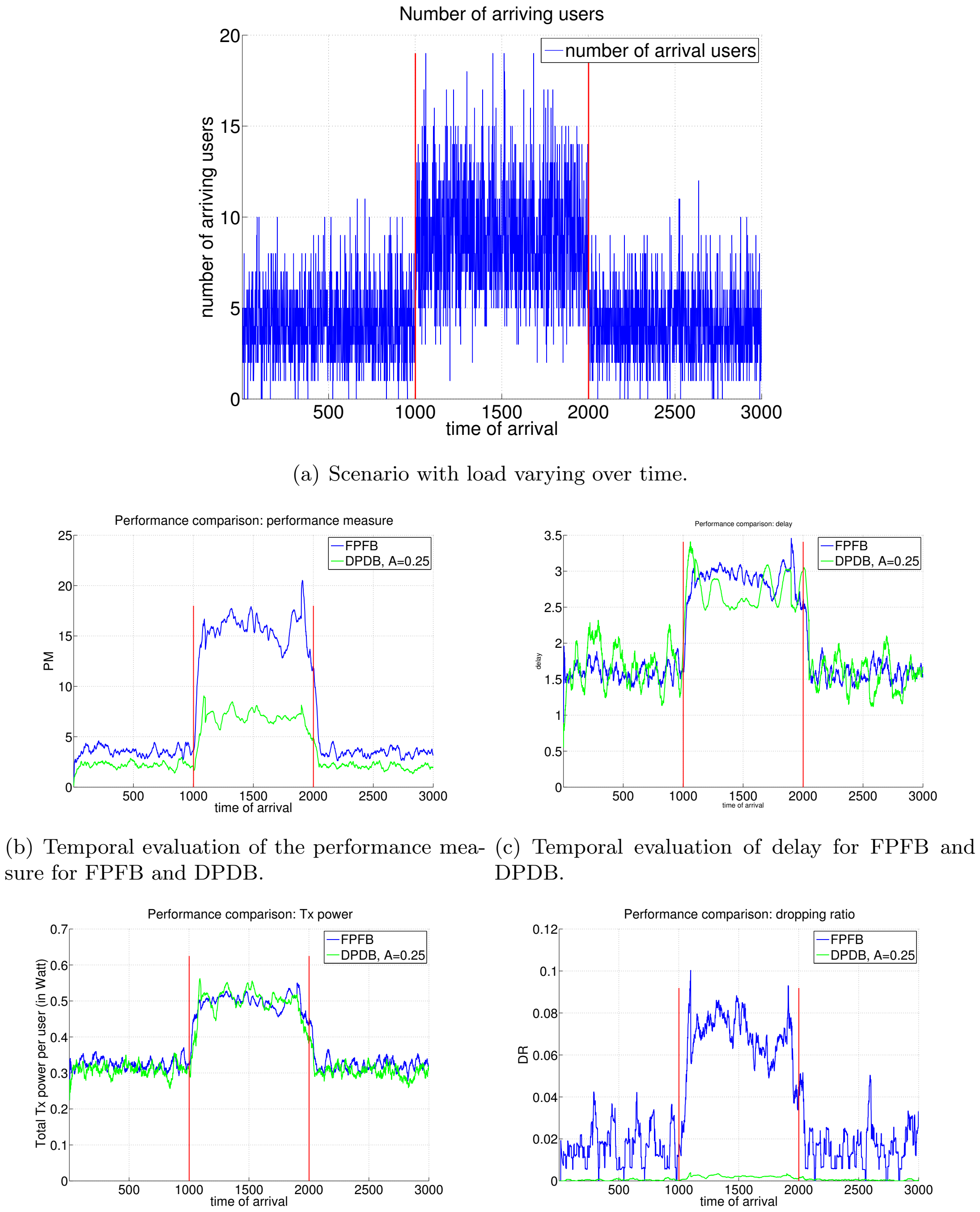


(a) Scenario with channel fluctuations and deep fades.

(b) Temporal adaptation of transmission power to a deep fade.

(c) Temporal variation of the DMR.

Figure 5.9: Protocol adaptation with respect to power and DMR



(a) Scenario with load varying over time.

(b) Temporal evaluation of the performance measure for FPFB and DPDB.

(c) Temporal evaluation of delay for FPFB and DPDB.

(d) Temporal evaluation of power consumption for FPFB and DPDB.

(e) Temporal evaluation of dropping rate for FPFB and DPDB.

Figure 5.10: Protocol adaptation over time when the traffic load varies from an average of 5 [users/sec] to an average of 10 [users/sec] and back. Value of idle parameter $\mathcal{A} = 0.25$ and chosen window size $W = 200$ slots. The benefits of the protocol over the fixed case are apparent for the delay and dropping rate, with almost the same power consumption. The DPDB case is definitely superior compared to the FPFB case regarding the performance measure in (b). A certain overshoot and delayed response in both (c) and (d) is due to the choice of large window size W and the power step Δp, which can be further optimally tuned to adapt to each scenario of expected traffic change.

Chapter 6

Mobility Robustness Optimization

The MRO problem in LTE SON is a multi-objective optimization problem, which involves a set of non-convex contradicting objective functions that depend on multiple variables such as handover (HO) parameters and user mobility classes. In this chapter we exploit the framework of stochastic processes to develop a novel method of successively choosing a sequence of multi-variate training points for multi-objective optimization. Combined with the collected statistics and a priori knowledge, the proposed method is used in the design of an efficient MRO algorithm. The performance of the algorithm is evaluated by simulations to illustrate significant improvements with respect to both HO-related radio link failure (RLF) and unnecessary HOs.

Parts of this chapter have already been published in [4]

6.1 Motivation and Related Work

A key objective of MRO is to improve the HO performance by reducing the number of HO-related RLFs and the number of unnecessary or missed handovers caused by incorrect HO decisions. The main desired functionalities include detection of “too early HO”and “too late HO”, and improving the overall handover performance by tuning the HO-related parameters.

Although some approaches to the problem have already been proposed, most of them are not based on systematic methods but rather on engineering intuition and simulations. Second, most of the existing algorithms such as those in [Jea10, Jea11, Bea11] adjust only the two global HO parameters hysteresis and TTT so that they impact the HO performance in the whole cell. Such approaches are therefore inadequate to cope with HO problems that pertain only to a specific cell pair, in which case it is more appropriate to adjust the local HO parameter such as CIO. Last but not least, the HO performance of a user strongly depends on the mobility class to which the user belongs. The authors of [SWZZ10, Lea11, Kea11] take the mobility classes into account, but they do not differentiate between local and global HO problems, and consider only the global HO parameters.

We are motivated to formulate the MRO problem as a multi-objective optimization problem, in which the objective functions are in general unknown, non-convex, and depend on multiple variables. The unknown functions can be explored at selected training points by taking measurements (called trials). The training points can possibly be corrupted by some Gaussian noise due to the missing or delayed measurements. The maximum allowable number of trials is strongly restricted, because each trail results in a relative high cost, for instance, in terms of wireless resources. We therefore consider an extension of the so-called P-algorithm which was introduced by Kushner [Kus64] and Žilinskas [Ž85] for single-objective global optimization; this algorithm, which models an unknown function as a stochastic process defined by the noisy training set, has been shown to be an efficient method for minimizing unknown functions. Recently, using Gaussian processes for statistical modeling, the P-algorithm has been generalized to multi-objective optimization [Ž12]. In this work, however, all components of the multi-objective functions are assumed to be independent processes, which is not satisfied in our MRO scenario since different HO performance measures are highly dependent on each other. For this reason, using the framework of multivariate Gaussian process (GP), we extend the method of [Ž12] to incorporate the inter-dependencies between different HO performance measures. The algorithm provides optimized local and global HO parameters per user mobility class. The collected local statistics and a priori knowledge are utilized to improve the efficiency of the algorithm. Simulation results show significant performance gains.

6.2 System Model and Problem Statement

We consider a multi-cell scenario consisting of one central (serving) cell surrounded by m neighbor cells $j \in \mathcal{S}, |\mathcal{S}| = m$. Let the set of users served by the central cell be denoted by $\mathcal{K}$. In the remainder of this section, we briefly describe the HO process, introduce HO metrics and parameters, and state the optimization problem.

6.2.1 HO Process and Parameters

A HO process of user $k \in \mathcal{K}$ from the serving cell to cell j is illustrated in Fig.6.1. UE reports the raw measurement of RSRP from each detected cell j at physical layer (PHY) layer $q_j(n)$ at the n-th time unit, and provides results to RRC layer for averaging once every N_0 ms. A nominal measurement period from L3 point of view is $N_0 = 200$ ms [LPGC12].

The filtered RSRP $P_j(n)$ is computed with

$$P_j(n) := (1-\beta)P_j(n-1) + \beta q_j(n), \tag{6.1}$$

where $P_j(0) := q_j(0)$ and parameter $\beta := 2^{-k/4}$ depending on the filter coefficient k is optionally signaled to UE in RRC measurement configuration message.

While moving towards cell j, UE waits for a time t_1 to trigger a counter for handover request (HRQ) until the HO condition $P_j(n) \geq P_0(n) + M_j$ is satisfied, where P_j is the filtered RSRP of user k from neighbor cell j, P_0 is the filtered RSRP from serving cell, and M_j is the handover margin (HOM) given by

$$M_j = H - O_j, \tag{6.2}$$

Here and hereafter H is the hysteresis in serving cell to ensure strong signals from the candidate cells, and O_j is the pairwise CIO to give a higher preference to a candidate cell to take over the user.

If the condition holds for a time $t_2 = T$ called TTT, then a HRQ is sent to cell j. A HRQ is considered *successful* if after requesting it, the user moves into a coverage area (a region where $\Pr\{\mathrm{SINR} \geq \gamma_0\} \geq \lambda$ is satisfied for some predefined thresholds γ_0 and λ) of cell j; otherwise we have a HO failure. In contrast, a HO-related RLF occurs when a user leaves the coverage area of the serving cell before a successful HO is completed [1]. This is the case when t_1 or t_2 is too long for the velocity v_k. Hereafter for brevity we use RLF to represent the HO-related RLF in the serving cell. Finally, a ping-pong handover (PPHO) is defined to be a handover to a neighbor cell that returns to the original cell after a short time T_{crit}. Fig. 6.2 illustrates the examples of a normal HO process, a RLF caused by too-late HO, a HF caused by too-early HO, and a PPHO (unnecessary HO) caused by too-early HO.

6.2.2 Handover Metrics

The HO performance is generally evaluated by three HO metrics: radio link failure rate (RLFR) denoted by R_1, handover failure rate (HFR) denoted by R_2 and HO_PPR denoted by R_3. According to [3GPa], these are defined as

$$R_1 = \frac{N_{\mathrm{RLF}}}{|\mathcal{K}|},\ R_2 = \frac{N_{\mathrm{HF}}}{N_{\mathrm{HRQ}}},\ R_3 = \frac{N_{\mathrm{PPH}}}{N_{\mathrm{HRQ}}}\,. \tag{6.3}$$

Here and hereafter, $|\mathcal{K}|$ is the cardinality of $\mathcal{K}$, while $N_{(\cdot)}$ is used to denote the number of occurrences of event $(\cdot)$.[2] The HO metrics in (6.3) are global metrics for the entire

[1] In [3GPa] a handover failure (HF) is also defined as a RLF which occurs in the target cell after the HO process. To distinguish the *too-late* and *too-early* indicators, in this chapter we name the RLF in the serving cell before sending a HRQ as RLF, whereas the RLF in the target cell after sending a HRQ as HF.

[2] For instance, N_{HRQ} is the number of handover requests, while N_{HRQ_j} used in (6.4) is the number of handover requests to neighbor cell j.

serving cell. In contrast, the HO performance between the serving cell and neighbor cell j is expressed in terms of local HO metrics defined to be

$$r_{j,1} = \frac{N_{\mathrm{RLF}_j}}{|\mathcal{K}_j|}, r_{j,2} = \frac{N_{\mathrm{HF}_j}}{N_{\mathrm{HRQ}_j}}, r_{j,3} = \frac{N_{\mathrm{PPH}_j}}{N_{\mathrm{HRQ}_j}}. \tag{6.4}$$

Since $|\mathcal{K}| = \sum_{j=1}^{m} |\mathcal{K}_j|$ and $N_{\mathrm{HRQ}} = \sum_{j=1}^{m} N_{\mathrm{HRQ}_j}$, the global metrics can be seen as the weighted average of the local metrics:

$$R_i = \sum_{j=1}^{m} a_{j,i} r_{j,i}, \text{ where } a_{j,i} = \begin{cases} \frac{|\mathcal{K}_j|}{|\mathcal{K}|}, & i = 1 \\ \frac{N_{\mathrm{HRQ}_j}}{N_{\mathrm{HRQ}}}, & i = 2, 3\,. \end{cases} \tag{6.5}$$

While the estimates of $r_{j,2}$ and $r_{j,3}$ can be obtained from HRQs between the cells as proposed in [3GPa], the estimate of $r_{j,1}$ cannot be directly obtained from the measurements. Therefore, we propose that each user k reports the cell ID of the best neighbor $j^* = \arg\max_j \bar{P}_j$ periodically, where $\bar{P}_j$ is the averaged value of P_j over the last predefined τ time frames (e.g., in simulations, $\tau = 10$). During an observation time period, we estimate $|\mathcal{K}_j|$ and N_{RLF_j} as follows:

- If a call is dropped and the last report before the call drop is j, increment both N_{RLF_j} and $|\mathcal{K}_j|$ by 1.
- Increment $|\mathcal{K}_j|$ by 1 either if a call is handed over to j-th neighbor cell, or a call is ended and the last report is j, or if a call remains in serving cell and the latest report is j.

6.2.3 Problem Statement and Our Approach

Our objective is to minimize the HO metrics while satisfying some given requirements on them. Once a violation of the requirements is detected and the HO problem is identified/classified, a MRO algorithm is initiated with appropriate parameters to resolve the problem by adapting the HO control parameters, including the global parameters $\{H, T\}$ (hysteresis and TTT) and the local parameter O_j (CIO).[3] To this end, we model the unknown relationship between the HO performance metrics and the HO control parameters as a multivariate Gaussian process and apply different multi-objective P-algorithms of [Ž85] (see Section 6.4 for more detail). The choice of the algorithm and its initial parameters depend on the type of a detected HO problem. As described in Section 6.3.1, we differentiate between global and local problems on the one hand, and too-late and too-early problems on the other hand.

[3] Note that the global parameters affect the HO performance at all cell edges, while O_j has impact only on the jth cell edge.

Finally, we point out that we differentiate between $C = 3$ user mobility classes classified based on users' reported mobility states as suggested in [3GPg]: *normal, medium* and *high.* The HO metrics are collected per mobility class so that the optimization problem decomposes in C independent sub-optimization problems with HO parameters defined per user mobility class. For convenience, however, we confine our attention in Section 6.3 to one arbitrary mobility class and point out that the following problem formulation and optimization strategies based on the collected local statistics can be applied individually to each mobility class. Thus, the output of the algorithm is a set of optimized HO parameters per user mobility class per cell.

6.3 MRO Algorithm

6.3.1 Handover Problem Detection

As aforementioned, HO problems are classified in two groups, either of which contains two sub-groups:

1. *Too-late and too-early HO problems*: Larger values of $t_1 + t_2$ (see Fig.6.1) lead to too-late decisions and higher RLFs, while smaller values of $t_1 + t_2$ result in too-early HO decisions in strongly overlapped serving area, thereby increasing HFR and HO_PPR.

2. *Global and local HO problems*: Roughly speaking, there is a global HO problem if there are sufficiently many local HO problems of the same type, while other boundaries do not suffer from a conflicting type of local HO problems; otherwise a local HO problem is declared to be dealt with local HO control parameters.

Note that for some predefined requirements $\delta_i > 0, i = 1, 2, 3$, there is a local HO problem associated with the jth neighbor cell if either $r_{j,1} > \delta_1$ (too many RLFs caused by too-late decisions) or if $\sum_{i=2,3} r_{j,i} > \sum_{i=2,3} \delta_i$ (too many RLFs and HO_PPRs due to too-early decisions). Based on this, given $\{r_{j,i}\}$, the proposed detection algorithm summarized in Algorithm 2 classifies detected HO problems in four classes.

Based on the output of this algorithm, we tune either global or local parameters at each step. The distinction between too-late and too-early HO problems allows us to confine the search domain to certain regions.

6.3.2 Handover Optimization

We introduce the following assumption, which is justified at the network level where optimization periods are relatively long.

Algorithm 2: HO problem detection and classification

```
loop
  Collect $\{r_{j,i} : j = 1, \ldots, m, i = 1, 2, 3\}$
  Find the sets $\mathcal{B}^{(l)} = \{j : r_{j,1} > \delta_1\}$ and $\mathcal{B}^{(e)} = \{j : \sum_{i=2,3} r_{j,i} > \sum_{i=2,3} \delta_i\}$
  if $|\mathcal{B}^{(l)}| \geq m/2$ and for $\hat{j} \notin \mathcal{B}^{(l)}$, $\sum_{i=2,3} r_{\hat{j},i} \leq \sum_{i=2,3} \delta_i - \epsilon_2$ then
    Global, too-late
  else if $|\mathcal{B}^{(e)}| \geq m/2$ and for $\hat{j} \notin \mathcal{B}^{(e)}, r_{\hat{j},1} \leq \delta_1 - \epsilon_1$ then
    Global, too-early
  else if $\mathcal{B}^{(l)} \neq \emptyset$, for each $j$ in $\mathcal{B}^{(l)}$ then
    Local, too-late, boundary $j$
  else if $\mathcal{B}^{(e)} \neq \emptyset$, for each $h$ in $\mathcal{B}^{(e)}$ then
    Local, too-early, boundary $h$
  else if $\mathcal{B}^{(l)} \cup \mathcal{B}^{(e)} = \emptyset$ then
    Normal
  end if
end loop
```

Assumption 6.1. *The moving direction and speed of each mobility class are random stationary processes over every optimization period.*

Under Assumption 6.1, for each boundary j, the local metrics defined in (6.4) depend only on $\boldsymbol{v}_j = (M_j, T)^T$. Let us denote the global HO control vector by $\boldsymbol{x} = (H, T)^T \in \mathcal{X}_0 = [H_{min}, H_{max}] \times [T_{min}, T_{max}]$, while $\boldsymbol{z}_j = (O_j, 0)^T \in \mathcal{O}_0 = [O_{min}, O_{max}] \times \{0\}$ contains only the local HO control parameter.[4] The functions

$$f_{j,i}(\boldsymbol{v}_j) = f_{j,i}(\boldsymbol{x} - \boldsymbol{z}_j), i \in \{1, 2, 3\}, 1 \leq j \leq m\,.$$

determine the relationship between $r_{j,i}$ and the HO control parameters $\boldsymbol{x}$ and $\boldsymbol{z}_j$.

6.3.3 Global MRO Algorithm

We define $F(\boldsymbol{x}) = (\boldsymbol{f}_1(\boldsymbol{x} - \hat{\boldsymbol{z}}_1), \ldots, \boldsymbol{f}_m(\boldsymbol{x} - \hat{\boldsymbol{z}}_m))^T$ for any fixed $\{\hat{\boldsymbol{z}}_j\}_{j=1}^m$ where

$$\boldsymbol{f}_j(\boldsymbol{x} - \hat{\boldsymbol{z}}_j) = (f_{j,i}(\boldsymbol{x} - \hat{\boldsymbol{z}}_j) : i = 1, 2, 3)^T \tag{6.6}$$

contains the local HO metrics for boundary j. Then the global MRO problem is given by

$$\min_{\boldsymbol{x} \in \mathcal{X}_0} F(\boldsymbol{x}) \tag{6.7}$$

To apply the multi-objective version of P-algorithm introduced in Section 6.2.3, the following assumption is made.

Assumption 6.2. *During each optimization period, the observations of $F(\boldsymbol{x})$ are assumed to be a Gaussian random field $\Psi(\boldsymbol{x})$. The components $\{\boldsymbol{\psi}_j(\boldsymbol{x})\}_{j=1}^m$ are independent and each $\boldsymbol{\psi}_j(\boldsymbol{x})$ is considered an tri-variate GP.*

Algorithm 3: Searching strategy for global MRO problem.

```
Input: The predefined system performance requirements for RLFR, HFR and HO_PPR:
    $\delta_i > 0, i = 1, 2, 3$
Collect $n$ initial sample points of local HO metrics, including the input set
    $\mathcal{V}_{j,n} = \{\boldsymbol{v}_{j,l} = (\boldsymbol{x}_l - \boldsymbol{z}_{j,l})\}_{l=1}^n$ and the output set $\mathcal{Y}_{j,n} = \{\boldsymbol{y}_{j,l}\}_{l=1}^n$, where the $l$-th
    observation is $\boldsymbol{y}_{j,l} = (r_{j,i}^{(l)} : i = 1, 2, 3)^T \in \mathbb{R}^3$.
loop
  if global too-late HO problem is detected then
    Confine the search domain $\mathcal{X} = \mathcal{X}_0 \setminus [H_n, H_{max}] \times [T_n, T_{max}]$, where $(H_n, T_n)$
      denotes the HO global parameters at the $n$th observation.
  else if global too-early HO problem is detected then
    Search domain $\mathcal{X} = \mathcal{X}_0 \setminus [H_{min}, H_n] \times [T_{min}, T_n]$
  end if
  Choose the next observation point
      $$\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \prod_{j=1}^{m} \Pr\{\boldsymbol{\psi}_j(\boldsymbol{x}) \le \boldsymbol{y}_j^{on} | \mathcal{V}_{j,n}, \mathcal{Y}_{j,n}\}\,. \quad (6.8)$$
    $\boldsymbol{y}_j^{on} = (y_{j,1}^{on}, y_{j,2}^{on}, y_{j,3}^{on})^T$, $y_{j,i}^{on} = \max\{y_{j,i}^{min}, \frac{\delta_i}{a_{j,i}m}\}$, and $y_{j,i}^{min} = \min_{1 \le l \le n} r_{j,i}^{(l)}$.
  $n \leftarrow n + 1$, collect new sample and update $\mathcal{V}_{j,n}, \mathcal{Y}_{j,n}$.
  Stops if $R_i \le \delta_i, \forall i$.
end loop
```

The assumption implies that each HO performance metric is a smooth function corrupted by Gaussian noise. Moreover, it captures the fact that HO metrics for different boundaries are jointly independent, whereas the observation processes for RLFR, HFR and HO_PPR are dependent for any boundary j – indeed, $f_{j,1}$ and $(f_{j,2}, f_{j,3})^T$ are contradicting objective functions of the same variables.

With Assumption 6.2, the algorithm described in Section 6.4 is applied to the global MRO problem in (6.7). In more detail, a search strategy is formulated in Algorithm 3.

With independence assumption on $\boldsymbol{\psi}_j(\boldsymbol{x})$, we can easily compute (6.8) based on the independence model in Section 6.4.3. Since the HO parameters are chosen from a set of finite size [3GPi], the conditional probability in (6.8) can be computed numerically according to the multivariate GP modeling in Section 6.4.4. The differentiation between too-late and too-early HO problems provides additional constraints on the search domain. Since $f_{j,1}$ on the one hand and $(f_{j,2}, f_{j,3})^T$ on the other one are contradicting objectives, and therefore difficult to minimize at the same time, we use $\boldsymbol{y}_j^{on}$ instead of $\boldsymbol{y}_j^{min}$ in (6.11) to enforce $r_{j,i} \le \frac{\delta_i}{a_{j,i}m}, \forall j, i$, from which we have $\forall i, R_i = \sum_{j=1}^{m} a_{j,i} r_{j,i} \le \delta_i$. The algorithm is stopped when all global metrics defined in (6.3) fall below the threshold $\delta_i, i = 1, 2, 3$.

[4] The second entry of $\boldsymbol{z}_j$ is 0 so that we can write $\boldsymbol{v}_j = \boldsymbol{x} - \boldsymbol{z}_j$ (recall that $M_j = H - O_j$).

6.3.4 Local MRO Algorithm

If a HO problem is detected at boundary $\hat{j}$, with fixed global parameter $\hat{\boldsymbol{x}}$, the local MRO algorithm is triggered of the form

$$\begin{aligned} &\min_{\boldsymbol{z}_{\hat{j}} \in \mathcal{O}} \boldsymbol{f}_{\hat{j}}(\boldsymbol{z}_{\hat{j}}) \\ &\boldsymbol{f}_{\hat{j}}(\boldsymbol{z}_{\hat{j}}) = (f_{\hat{j},i}(\hat{\boldsymbol{x}} - \boldsymbol{z}_{\hat{j}}) : i = 1, 2, 3)^T. \end{aligned} \tag{6.9}$$

The problem is equivalent to that stated in (6.10), and can be approached by the algorithm in (6.11). Similar to Algorithm 3, the search domain $\mathcal{O}$ is constrained based on the a priori knowledge about the type of detected HO problems. Accordingly, if too-late problem is detected, then $\boldsymbol{z}_{\hat{j}} \in \mathcal{O}$, $\mathcal{O} = \mathcal{O}_0 \setminus [O_{min}, O_{\hat{j},n}] \times \{0\}$, where $O_{\hat{j},n}$ is the current CIO assigned to boundary $\hat{j}$. Also the cumulative distribution function is calculated up to $\boldsymbol{y}^{on}_{\hat{j}}$, where $y^{on}_{\hat{j},i} = \max\{y^{min}_{\hat{j},i}, \delta_i\}$. The algorithm is stopped if the system requirements $\boldsymbol{\delta}$ on local metrics in (6.4) are satisfied. The system requirements $\boldsymbol{\delta}$ are the same for global and local metrics, since the global metrics are the weighted average of the local metrics, as shown in (6.5).

6.3.5 Interaction between Global and Local MRO Algorithms

The global MRO algorithm improves the general HO performance but it may lead to some side effects on a few boundaries. For example, if a "global, too late" problem is detected, the global MRO algorithm is triggered, and the HO performance on most boundaries is improved. However, a few boundaries may suffer from this global optimization and have "too early" problem, in which case a local MRO algorithm is then triggered to compensate the detrimental impact of the global changes. This does not affect the HO performance on other boundaries due to the independence according to Assumption 6.2. Thus, the overall HO performance benefits from a global optimization followed by some local compensation actions.

6.4 Extended Multi-Objective P-Algorithm

6.4.1 Multi-Objective P-Algorithm

Consider the following optimization

$$\min_{\boldsymbol{x} \in \mathcal{A}} \boldsymbol{f}(\boldsymbol{x}), \ \boldsymbol{f}(\boldsymbol{x}) = \big(f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x})\big)^T \tag{6.10}$$

where $\mathcal{A} \subset \mathbb{R}^d$ denotes a feasible set of $d \geq 1$ control parameters, and $\boldsymbol{f} : \mathcal{A} \to \mathbb{R}^m, m \geq 1$, is the unknown vector-valued objective function. Since $\boldsymbol{f}$ is unknown, we model this function using a random field $\boldsymbol{\psi} : \mathcal{A} \to \mathbb{R}^m$ so that $\boldsymbol{\psi}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{A}$, is a random vector. Now we can define the multi-objective P-algorithm.

Definition 6.1. *(Multi-objective P-algorithm [Ž12]) Suppose* $\mathcal{X}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathcal{A}$ *are available training points up to step* n*, and* $\mathcal{Y}_n = \{\boldsymbol{y}_1 = \boldsymbol{\psi}(\boldsymbol{x}_1), \ldots, \boldsymbol{y}_n = \boldsymbol{\psi}(\boldsymbol{x}_n)\}$ *are realizations of* $\boldsymbol{\psi}(\boldsymbol{x})$ *on these points where* $\forall i\, \boldsymbol{\psi}(\boldsymbol{x}_i) = (y_{i,1} = \psi_1(\boldsymbol{x}_i), \ldots, y_{i,m} = \psi_m(\boldsymbol{x}_i))^T$*. The P-algorithm is then defined by the following iteration*

$$\boldsymbol{x}_{n+1} = \underset{\boldsymbol{x} \in \mathcal{A}}{\arg\max} \Pr\big\{\boldsymbol{\psi}(\boldsymbol{x}) \leq \boldsymbol{y}^{min} | \mathcal{X}_n, \mathcal{Y}_n\big\}, n \in \mathbb{N} \tag{6.11}$$

where $\boldsymbol{y}^{min} = (y_1^{min}, \ldots, y_m^{min})$*, and* $y_j^{min} = \min_{1 \leq i \leq n} y_{i,j}$*.*

Note that at step n, the P-algorithm chooses the next test point $\boldsymbol{x}_{n+1}$ so as to maximize the conditional probability for $\boldsymbol{y}_{n+1} = \boldsymbol{\psi}(\boldsymbol{x}_{n+1}) \leq \boldsymbol{y}^{min}$, where $\boldsymbol{y}^{min}$ is a vector containing the minimum values among all observed values up to step n.

6.4.2 Modeling with Gaussian Processes

In [Ž12], the iteration in (6.11) is performed under the assumption that the components of $\boldsymbol{\psi}(\boldsymbol{x})$ are independent Gaussian random variables for every $\boldsymbol{x} \in \mathcal{A}$. Since this assumption is not necessarily satisfied in the MRO context due to strong dependencies between different components, we extend the model of [Ž12] to include the interdependencies.

To this end, assume that $\boldsymbol{\psi}(\boldsymbol{x})$ is a multivariate GP

$$\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{\phi}(\boldsymbol{x}) + \boldsymbol{b} \tag{6.12}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ is a symmetric positive definite matrix which determines the variance-covariance matrix of $\boldsymbol{\psi}(\boldsymbol{x})$, $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x}))^T$ is used to denote a vector of mutually independent stationary GP with zero mean and unit variance, and $\boldsymbol{b} \in \mathbb{R}^m$ is the mean of the process $\boldsymbol{\psi}(\boldsymbol{x})$. It is assumed that the correlation function of $\psi_j(\boldsymbol{x})$ yields

$$c_j(\boldsymbol{x}_l, \boldsymbol{x}_k) = \exp\left(-\frac{1}{2}(\boldsymbol{x}_l - \boldsymbol{x}_k)^T \boldsymbol{M}_j (\boldsymbol{x}_l - \boldsymbol{x}_k)\right) \tag{6.13}$$

where $\boldsymbol{M}_j = \mathrm{diag}(\boldsymbol{\theta}_j), \boldsymbol{\theta}_j \in \mathbb{R}^d$. The parameters $\{A, \boldsymbol{b}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ are called *hyperparameters* and can be freely chosen. Reference [RW06] provides various methods to determine the hyperparameters and one possible method is to optimize the marginal likelihood. By (6.13), we have

$$\boldsymbol{K}_{lk} := \mathrm{Cov}(\boldsymbol{y}_l, \boldsymbol{y}_k) = \mathrm{Cov}\left(\boldsymbol{\psi}(\boldsymbol{x}_l), \boldsymbol{\psi}(\boldsymbol{x}_k)\right) = \boldsymbol{A}\boldsymbol{C}_m\boldsymbol{A}^T \tag{6.14}$$

where $\boldsymbol{C}_m = \mathrm{diag}(c_1(\boldsymbol{x}_l, \boldsymbol{x}_k), \ldots, c_m(\boldsymbol{x}_l, \boldsymbol{x}_k)) \in \mathbb{R}^{m \times m}$. Note that if $\boldsymbol{x}_l = \boldsymbol{x}_k$, then

$$\boldsymbol{\Sigma}_0 := \mathrm{Cov}(\boldsymbol{y}_l, \boldsymbol{y}_l) = \boldsymbol{A}\boldsymbol{A}^T. \tag{6.15}$$

Notice that the assumption of the correlation function in (6.13) leads to $\boldsymbol{K}_{lk} = \boldsymbol{K}_{kl}$, the covariance matrix $\boldsymbol{\Sigma}_{mn} \in \mathbb{R}^{mn \times mn}$ of the output vector $\boldsymbol{y}_{mn} := (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T \in \mathbb{R}^{mn}$ is given by

$$\boldsymbol{\Sigma}_{mn} = \boldsymbol{\Sigma}(\boldsymbol{y}_{mn}, \boldsymbol{y}_{mn}) = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{K}_{12} & \ldots & \boldsymbol{K}_{1n} \\ \boldsymbol{K}_{12} & \boldsymbol{\Sigma}_0 & \ldots & \boldsymbol{K}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{K}_{1n} & \boldsymbol{K}_{2n} & \ldots & \boldsymbol{\Sigma}_0 \end{pmatrix} . \tag{6.16}$$

Given a test input $\boldsymbol{x}_*$, let the corresponding output be denoted by $\boldsymbol{y}_*$. The $m \times mn$ covariance of the test point output $\boldsymbol{y}_*$ and the training output vector $\boldsymbol{y}_{mn}$ is given by

$$\boldsymbol{\Sigma}_{*,mn} = \boldsymbol{\Sigma}(\boldsymbol{y}_*, \boldsymbol{y}_{mn}) = (\mathrm{Cov}(\boldsymbol{y}_*, \boldsymbol{y}_1), \ldots, \mathrm{Cov}(\boldsymbol{y}_*, \boldsymbol{y}_n)) \ .$$

In Section 6.3 we model the objective function using both dependent and independent models. The following section introduces the *independence* model based on [Ž12], whereas the *non-separable dependence model* is considered thereafter.

6.4.3 Independence Model

If the components of $\boldsymbol{\psi}(\boldsymbol{x})$ are independent, $\boldsymbol{A} = \mathrm{diag}(\sigma_1^2, \ldots, a_m^2)$, in which case the regression of $\boldsymbol{f}(\boldsymbol{x})$ decomposes in m separate GP regressions: $\psi_j(\boldsymbol{x})$ for each $f_j(\boldsymbol{x})$. In this special case, the covariances matrix for each process $\psi_j(\boldsymbol{x})$ gives $\boldsymbol{\Sigma}_n \in \mathbb{R}^{n \times n}$, with the (l,k)-th entry equal to $c_j(\boldsymbol{x}_l, \boldsymbol{x}_k)$, and $\boldsymbol{\Sigma}_0 = c_j(\boldsymbol{x}_l, \boldsymbol{x}_l) = 1$. Given training points $(\mathcal{X}_n, \mathcal{Y}_n)$ for a test point $\boldsymbol{x}_*$, it follows that $\boldsymbol{\Sigma}_{*,n} = (c_j(\boldsymbol{x}_*, \boldsymbol{x}_1), \ldots, c_j(\boldsymbol{x}_*, \boldsymbol{x}_n))$. So the joint conditional distribution of $y_{*,j}$ and $\boldsymbol{y}_{n,j} = (y_{1,j}, \ldots, y_{n,j})^T$ given $\boldsymbol{b}$ is

$$\left[\begin{pmatrix} y_{*,j} \\ \boldsymbol{y}_{n,j} \end{pmatrix} \middle| b_j, \mathcal{X}_n, \boldsymbol{x}_* \right] \sim \mathcal{N}\left(b_j \mathbf{1}, \begin{bmatrix} 1 & \boldsymbol{\Sigma}_{*,n} \\ \boldsymbol{\Sigma}_{*,n}^T & \boldsymbol{\Sigma}_n \end{bmatrix} \right) \tag{6.17}$$

where $\boldsymbol{b}$ can be estimated by the generalized least square (GLS) estimators:

$$\hat{b}_j = \frac{\mathbf{1}^T \boldsymbol{\Sigma}_n^{-1} \boldsymbol{y}_{n,j}}{\mathbf{1}^T \boldsymbol{\Sigma}_n^{-1} \mathbf{1}}. \tag{6.18}$$

With estimated hyperparameters $\{\hat{b}_j, \hat{\sigma}_j, \hat{\boldsymbol{\theta}}_j\}$, given the test point input, the conditional mean of the test point output $m_j(\boldsymbol{x}_*|\cdot) := m(y_{*,j}|\boldsymbol{y}_{n,j}, \mathcal{X}_n, \boldsymbol{x}_*)$ and the conditional variance of the test point output $s_j(\boldsymbol{x}_*|\cdot) := s(y_{*,j}|\boldsymbol{y}_{n,j}, \mathcal{X}_n, \boldsymbol{x}_*)$ (the predictive equations for single variate GP regression) are

$$\begin{aligned} y_{*,j}|\boldsymbol{y}_{n,j}, \mathcal{X}_n, \boldsymbol{x}_* \sim & \mathcal{N}\left(m_j(\boldsymbol{x}_*|\cdot), s_j^2(\boldsymbol{x}_*|\cdot)\right) \\ \text{where } m_j(\boldsymbol{x}_*|\cdot) =& \hat{b}_j + \boldsymbol{\Sigma}_{*,n}^T \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{y}_{n,j} - \mathbf{1}\hat{b}_j) & (6.19) \\ s_j^2(\boldsymbol{x}_*|\cdot) =& \sigma_j^2 \left[1 - \boldsymbol{\Sigma}_{*,n}^T \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\Sigma}_{*,n} + \frac{(1 - \mathbf{1}^T \boldsymbol{\Sigma}_n^{-1} \mathbf{1})^2}{\mathbf{1}^T \boldsymbol{\Sigma}_n^{-1} \mathbf{1}} \right] . & (6.20) \end{aligned}$$

The above the conditional distributions are derived with the properties of the multivariate Gaussian distribution described in Appendix C.3. Thus, the probability in (6.11) can be computed as

$$\begin{aligned}\Pr\{\boldsymbol{\psi}(\boldsymbol{x}) \leq \boldsymbol{y}^{min}|\mathcal{X}_n, \mathcal{Y}_n\} &= \prod_{j=1}^{m} \Pr\{\psi_j(\boldsymbol{x}) \leq y_j^{min}|\cdot\} \\ &= \prod_{j=1}^{m} G\left(\frac{y_j^{min} - m_j(\boldsymbol{x}|\cdot)}{s_j(\boldsymbol{x}|\cdot)}\right) \end{aligned} \tag{6.21}$$

where $G(\cdot)$ denotes the CDF of the standard normal distribution, while $m_j(\boldsymbol{x}|\cdot)$ and $s_j(\boldsymbol{x}|\cdot)$ are given by (6.19) and (6.20), respectively.

6.4.4 Non-Separable Dependence Model

For the non-separable dependence model, let $\boldsymbol{A}$ be the unique square root of $\boldsymbol{\Sigma}_0$, which is by definition positive semidefinite. We therefore use the Cholesky decomposition $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ and ensure that all the elements on the main diagonal of $\boldsymbol{L}$ are non-negative, i.e., the constraints $l_{i,i} \geq 0, i = 1, \ldots, m$ are given in the maximum-likelihood estimator (MLE). The joint distribution of $\boldsymbol{y}_*$ and $\boldsymbol{y}_{mn}$ is

$$\left[\begin{pmatrix}\boldsymbol{y}_* \\ \boldsymbol{y}_{mn}\end{pmatrix}\middle| \boldsymbol{b}, \mathcal{X}_n, \boldsymbol{x}_*\right] \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{I}_m \\ \boldsymbol{I}_{mn}\end{bmatrix}\boldsymbol{b}, \begin{bmatrix}\boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_{*,mn} \\ \boldsymbol{\Sigma}_{*,mn}^T & \boldsymbol{\Sigma}_{mn}\end{bmatrix}\right) \tag{6.22}$$

where $\boldsymbol{I}_{mn} := \boldsymbol{1}_n \otimes \boldsymbol{I}_m$. Assume that $\boldsymbol{b}$ follows a non-informative uniform distribution; then the conditional mean and variance of $\boldsymbol{\psi}(\boldsymbol{x}_*) = \boldsymbol{y}_*$, given a set of training points $(\mathcal{X}_n, \mathcal{Y}_n)$, denoted by $\boldsymbol{m}(\boldsymbol{x}_*|\cdot) := \boldsymbol{m}(\boldsymbol{y}_*|\mathcal{X}_n, \mathcal{Y}_n, \boldsymbol{x}_*)$ and $\boldsymbol{S}(\boldsymbol{x}_*|\cdot) := \boldsymbol{S}(\boldsymbol{y}_*|\mathcal{X}_n, \mathcal{Y}_n, \boldsymbol{x}_*)$ respectively, yields

$$\boldsymbol{\psi}(\boldsymbol{x}_*)|\mathcal{X}_n,\mathcal{Y}_n \sim \mathcal{N}\left(\boldsymbol{m}(\boldsymbol{x}_*|\cdot), \boldsymbol{S}(\boldsymbol{x}_*|\cdot)\right) \tag{6.23}$$

$$\boldsymbol{m}(\boldsymbol{x}_*|\cdot) = \hat{\boldsymbol{b}} + \boldsymbol{\Sigma}_{*,mn}\boldsymbol{\Sigma}_{mn}(\boldsymbol{y}_{mn} - \boldsymbol{I}_{mn}\hat{\boldsymbol{b}}) \tag{6.24}$$

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{x}_*|\cdot) =& \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{*,mn}\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{\Sigma}_{*,mn}^T + (\boldsymbol{I}_m - \boldsymbol{\Sigma}_{*,mn}\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{I}_{mn}) \\ & \times (\mathcal{I}^T\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{I}_{mn})^{-1} \times (\boldsymbol{I}_m - \boldsymbol{\Sigma}_{*,mn}\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{I}_{mn})^T\end{aligned} \tag{6.25}$$

where $\hat{\boldsymbol{b}} = (\boldsymbol{I}_{mn}^T\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{I}_{mn})^{-1}\boldsymbol{I}_{mn}^T\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{y}_{mn}$. The mathematical properties in Appendix C.3 is used to derived the above conditional probabilities. The hyperparameters $\{\boldsymbol{A}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ are estimated by the MLE. With the conditional mean in (6.24) and the conditional variance in (6.25), we can estimate the conditional probability in Algorithm 6.1 (via a cumulative distribution function) numerically (e.g., multivariate normal cumulative distribution function implemented in MATLAB is based on [Dre94, GB02]). For details of the Gaussian identities please refer to Appendix C.3.

In summary, the steps of the multi-objective version of P-algorithm are

1. Collect initial training points $\{\mathcal{X}_n, \mathcal{Y}_n\}$.

2. Estimate hyperparameters $\{\boldsymbol{b}, \boldsymbol{A}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ with the training points by maximizing the marginal likelihood.

3. Find $\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x} \in \mathcal{A}} \Pr\{\boldsymbol{\psi}(\boldsymbol{x}) \leq \boldsymbol{y}^{min} | \mathcal{X}_n, \mathcal{Y}_n\}$.

4. Evaluate $\boldsymbol{y}_{n+1}$, increment n, repeat step 2) with updated training sample set, until a stopping criterion is met.

6.5 Experimental Results

We consider a highway scenario as shown in Fig.6.3, where 50 users are randomly distributed on a highway moving both ways at a speed of 150 km/h. The trajectory of these users follows a wrap-around property, i.e., once a user moves out of the area, it appears on the other side of the highway in the next time slot. There are 150 users uniformly distributed on the playground, moving with random direction. The velocity distribution of the playground users with three mobility classes: low (3 km/h), medium (50 km/h), and high (150 km/h) is $(0.4, 0.4, 0.2)$. The number of the users in each time slot are fixed, i.e., if a user is dropped, or moves out of the playground, a new user is generated within the playground. HO parameters H, T, O_j are chosen from the predefined pool $T \in \{4, 64, 80, 100, 128, 160, 256, 320, 480, 512, 640, 1024, 2560, 5120\}$ in [ms], $0 \leq H \leq 15, H \in \mathbb{Z}$ in dB, $-24 \leq O_j \leq 24, O_j \in \mathbb{Z}, \forall j$ in [dB] [3GPi], and ping-pong criteria time is set as $T_{crit} = 5$s.

The system is started with 25 uniformly distributed grid as initial training points, and with initial state $T_0 = 64$ms, $H_0 = 0$dB, and $O_j, \forall j$ randomly chosen in $[0, 6]dB$. The thresholds are set as $\delta_1 = 0.02$, $\delta_2 = \delta_3 = 0.04$. The mobility dependent MRO algorithm proposed in Section 6.3 is implemented, and its performance is compared against a conventional scheme, which stepwise decreases or increases the same global parameter for all mobility classes if a “too late” or “too early” HO problem is detected. The optimization interval is 120s. The simulation results in Fig.6.5 shows that “global too early” problem is detected first, and the global MRO algorithm is activated to minimize the HFR and HO_PPR. The trade off between the RLFR and HFR, HO_PPR leads to the local problem on the highway-boundaries to neighbor cell 3 and 6, and the local MRO algorithm is triggered to further optimize the local HO performance. Fig.6.5(a) shows that our algorithm outperforms the conventional stepwise method.

6.6 Summary

We consider the MRO problem as a multi-objective optimization problem, where the objective functions are unknown except for a limited number of training samples. To solve the problem, we modify the multi-objective version of P-algorithm by exploiting the framework of multi-variate Gaussian processes, so that the algorithm is suitable for dependence model in theMRO scenario. We present respectively the detection and optimization strategies for global and local MRO problems based on the proposed local statistics. The algorithm is implemented per user mobility class. Simulation results show significant improvements on reduction of the RLFRs and unnecessary HOs.

FIGURES

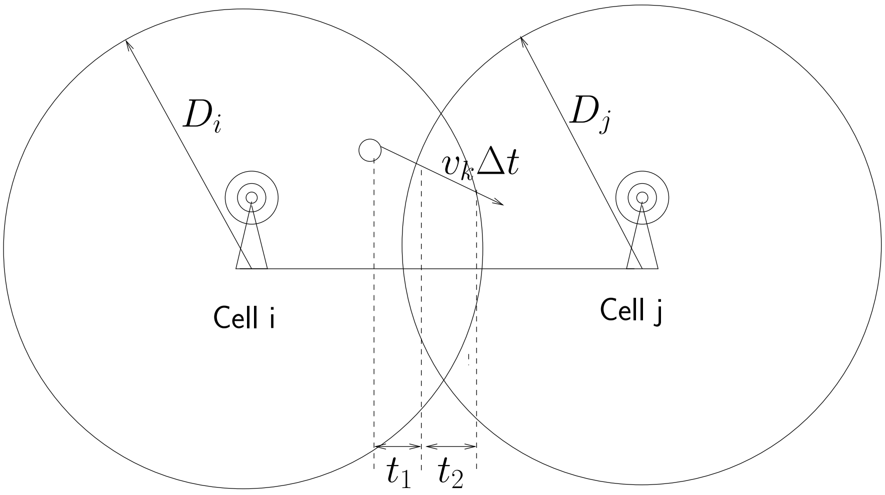
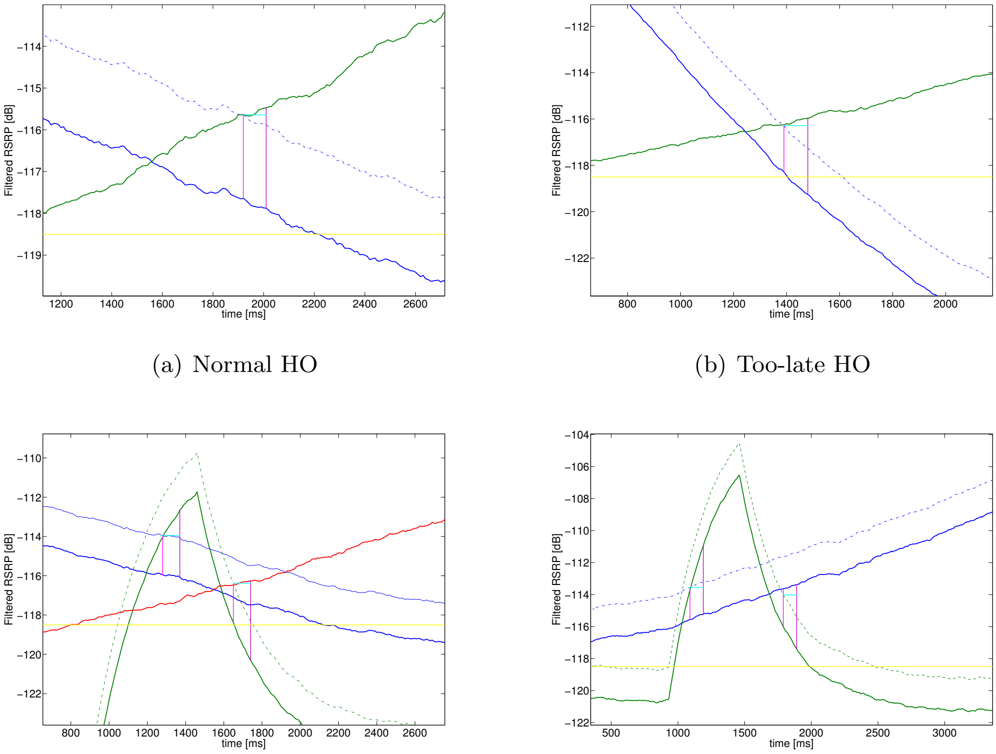


Figure 6.1: Illustration of a handover process

Figure 6.4 shows the empirical curves derived by Monte Carlo experiments. Two observations are made: 1) HO performance depends on the user mobility. With the same HO parameters, RLFR increases with the increase of the user mobility, while the HFR and HO_PPR decrease. The values of optimal TTT and HOM or a higher mobility class are generally lower than those for a lower mobility class. 2) As expected, HO metrics turn out to be inter-dependent. The RLFR and HFR,HO_PPR are contradicting objectives.



(a) Normal HO

(b) Too-late HO

(c) Too-early HO

(d) Ping-pong HO

Figure 6.2: HO process: blue solid curve - source pilot; green solid curve - first candidate pilot; red solid curve - second candidate pilot; blue dashed curve - source pilot + HOM; magenta vertical lines - TTT counting started; purple vertical lines - TTT counting terminated; cyan horizontal line - TTT

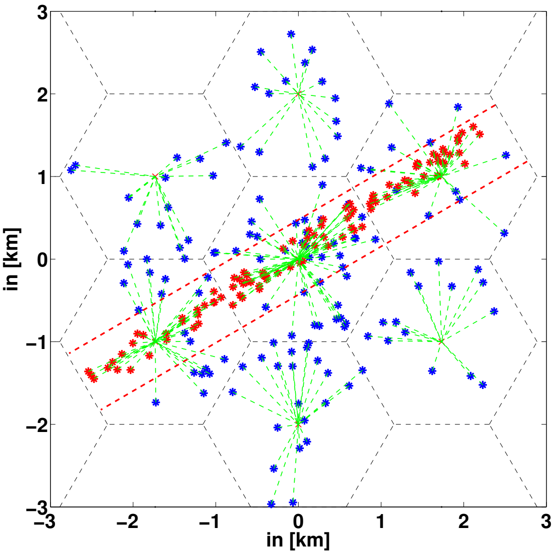
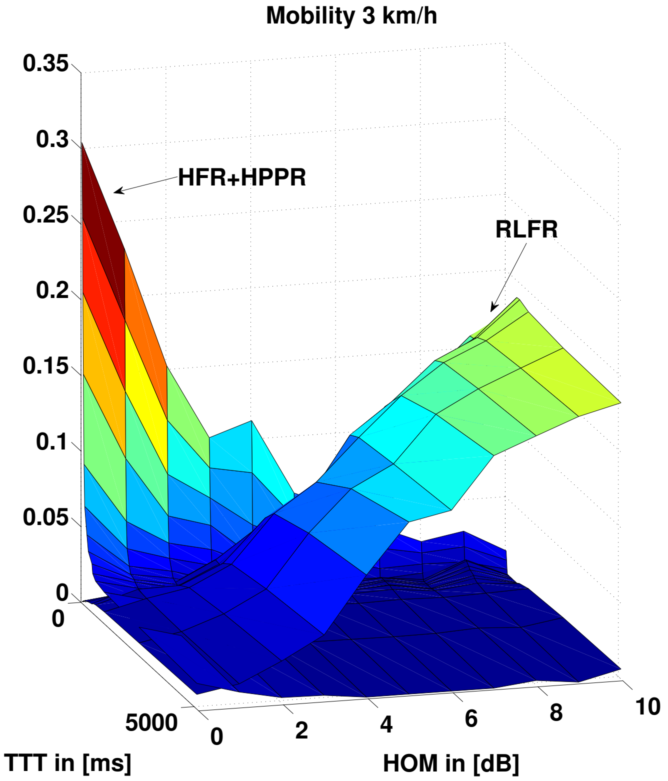
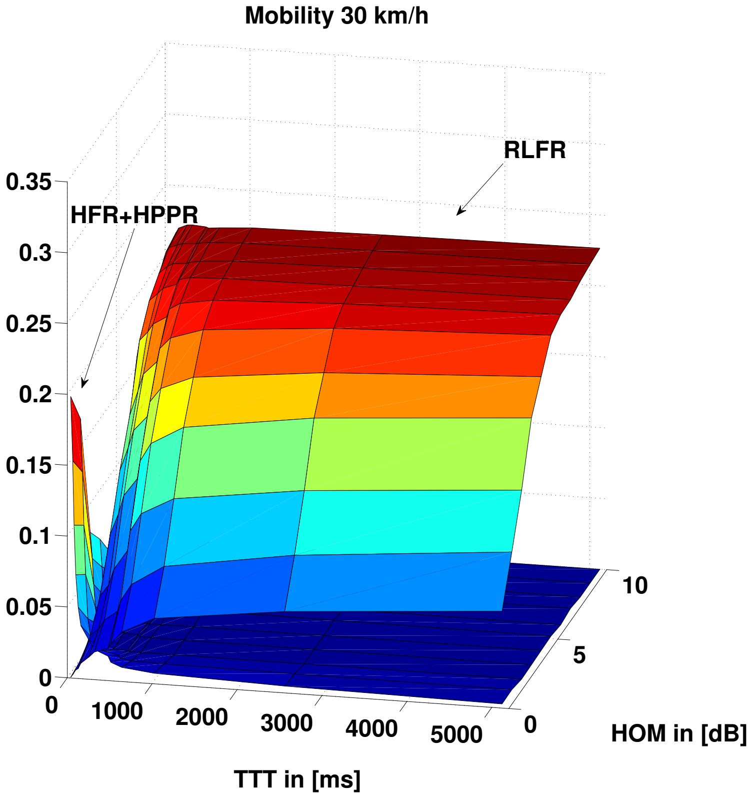


Figure 6.3: Simulation scenario

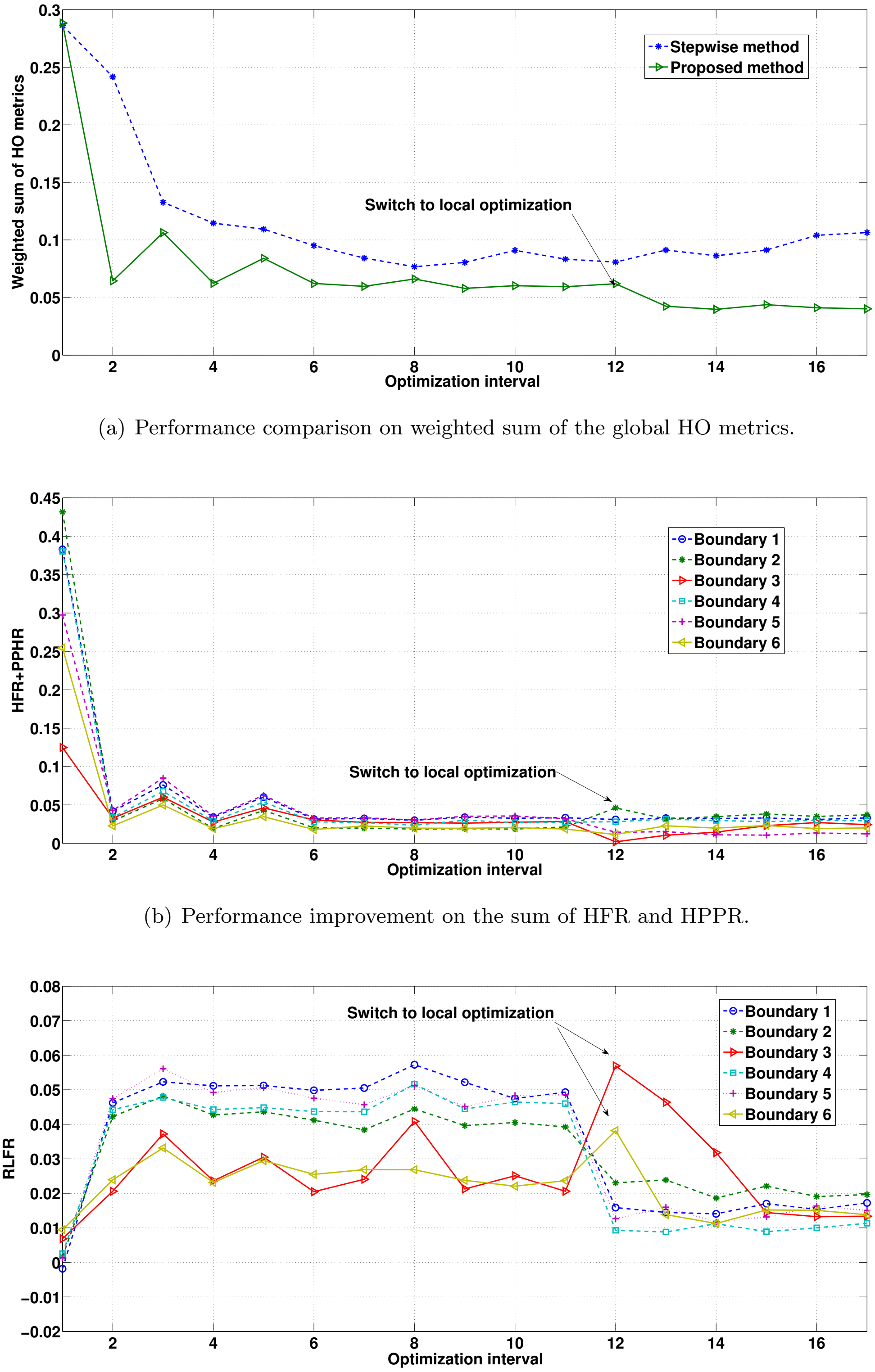


(a) HO metrics with mobility 3km/h.



(b) HO metrics with mobility 30km/h.

Figure 6.4: HO metrics depending on mobility classes.



(a) Performance comparison on weighted sum of the global HO metrics.

(b) Performance improvement on the sum of HFR and HPPR.

(c) Performance of RLFR.

Figure 6.5: Performance comparison.

Chapter 7

Distributed Interference-Aware Mobility Load balancing Algorithm

Within a cellular wireless network the unbalanced user load among cells, together with inter-cell interference (ICI), constitute major factors responsible for poor overall performance. In this chapter, we suggest a novel decentralized algorithm for Load Balancing in the downlink.

There are two major novelties in the analysis. (i) The algorithm is based on the solution of a mixed integer optimization problem solved using Lagrangian - but not Linear Programming - relaxation, which allows the solution to be binary for the user assignment variables. (ii) Its implementation is based on exchange of certain prices among base stations and allows each of them to make choices individually without the aid of a central controller. The cell handover parameters are further adequately adjusted to enforce cell-edge users to migrate to their optimal base station.

The algorithm aims at optimally balancing the load, while at the same time guaranteeing low levels of ICI. Its performance is evaluated through simulations, which illustrate the improvements provided on aggregate system utility.

Parts of this chapter have already been published in the coauthored work [3].

7.1 Introduction

In LTE networks, orthogonal frequency-division multiplexing access (OFDMA) eliminates intra-cell interference by assigning users to orthogonal subcarriers. However, the high frequency reuse factor among cells leads to ICI, which is a major cause of performance degradation, especially for cell-edge users. Imbalanced load among cells further intensifies the ICI problem, since a heavily loaded BS can cause strong interference to neighboring cells, while at the same time not being able to provide full service to its own users.

In LTE SON [3GPa], load balancing (LB) aims at balancing the load among different cells by adapting the cell reselection/HO parameters. In a conventional LB scheme, a pair of overloaded (OL) and target (TR) cell is initially specified, and afterwards some cell-edge users of the OL cell are handed over to the TR cell by modifying the HO parameters. In [Lea10], a cell is defined to be overloaded if the sum of the required bandwidth from users is larger than the available total bandwidth. TR cell is chosen to be the one with best RSRP for the cell edge users. In [SZC07], the OL cell selection is based on certain congestion metrics from admission control, and the TR cell is chosen to be the lightest loaded neighboring cell. Further relevant references include [SWMG08, ZRC+08, SAR+10].

However, none of them have considered that distributing the traffic load equally - while assigning users to the BS with best channel quality - is not enough to improve the overall spectral efficiency. This is because load balancing may cause severe ICI at the cell edge and eventually deteriorate the overall performance.

Our objective in this work is to better balance the load among cells, while taking ICI into consideration and introducing a utility per BS to model its satisfaction. For this purpose, after presenting in Section 7.2 the system model under study, we pose in Section 7.3 a mixed integer optimization problem together with an equivalent transformation. The Lagrangian relaxation of certain constraints and its decomposition into simpler subproblems is presented in Section 7.4. Several properties of the optimal Lagrangian solution are derived in 7.5, which depend on the value of a *load price* and *interference cost* per BS. After relating the mathematical model and solution more precisely to the actual network parameters, we propose in Section 7.7 a novel LB scheme which maximizes the aggregate utility function of the modified *spectral efficiency.* The algorithm allows the BSs to communicate in pairs and make individual decisions. It results in an improved total BS satisfaction by ICI mitigation and appropriate user re-assignment which is illustrated in Section VII by means of simulations.

7.2 System Model

We consider the downlink of an LTE multi-cell network with a set of BSs (or cells) $\mathcal{M} = \{1, \ldots, M\}$ and a set of users $\mathcal{N} = \{1, \ldots, N\}$. Let $\mathcal{N}_m$ denote the set of users assigned to BS $m \in \mathcal{M}$. The binary assignment indicator $a_{n,m} \in \{0, 1\}$ takes the value 1 if user n is assigned to BS m, otherwise it is equal to 0. A user can be assigned to exactly one BS, and therefore

$$\forall m \in \mathcal{M}, \sum_{m=1}^{M} a_{n,m} = 1. \tag{7.1}$$

The system under study implements an OFDMA scheme where all BSs share the same spectrum W to support the users. If $a_{n,m} = 1$, that is user n is assigned to BS m, then the BS should allocate a part of the bandwidth, denoted by $w_{n,m}$ for transmission. The spectrum allocation is considered here to be random and uniformly distributed over the entire W, as a statistical effect of timely varying frequency selective channel and the frequency hopping spread spectrum method. We have

$$\forall m \in \mathcal{M}, \sum_{n=1}^{N} w_{n,m} \leq W. \tag{7.2}$$

The transmission power from the BS to the user has a fixed value per unit of frequency equal to p and measured in [Joule/sec/Hz]. The total transmission power in [Joule/sec] destined for a specific user n equals $p \cdot w_{n,m}$. The total power budget per BS thus equals $p \cdot W$.

ICI arises when two or more neighboring cells operate on the same sub-carrier. The closed form of ICI depends on the underlying sub-carrier allocation scheduling scheme. Under the underlying model, the amount of interference created by BS $s \neq m$ to user n is a strictly increasing function of the allocated bandwidth in base station $s \neq m$. This is reasonable since the more frequency resources are utilized by a BS the more probable it is that the same subcarriers (SCs) are occupied by another BS, in which case inter-cell interference appears.

The power density of the interference caused to user n in cell m from a neighboring BS s is considered here an affine function $I_{n,s}$ of the *SC utilization ratio* $\frac{\sum_{j=1}^{N} w_{j,s}}{W}$ at s

$$I_{n,s} = p \cdot h_{n,s} \cdot \Big(\frac{\sum_{j=1}^{N} w_{j,s}}{W}\Big) \tag{7.3}$$

In the above, $h_{n,s} > 0$ is the long-term channel gain from BS s to user n, possibly estimated by RSRP measurements.

Considering a unit of bandwidth, the SINR of user n when served by BS m and with σ_n the thermal noise power spectral density equals

$$\mathrm{SINR}_{n,m} := \frac{p \cdot h_{n,m}}{\sum_{s \in \mathcal{M} \setminus \{m\}} I_{n,s} + \sigma_n}, \ \forall n, m. \tag{7.4}$$

We set a minimum rate requirement for each user n, denoted by γ_n. Using the Shannon capacity formula, the rate requirement of n should satisfy the constraint $w_{n,m} \cdot \log(1 + \mathrm{SINR}_{n,m}) \geq a_{n,m} \cdot \gamma_n$. The constraint is always fulfilled when $a_{n,m} = 0$, in which case n is not connected to m. Taking into account that the RSRQ is distributed within $[-19.5, -3]$ dB for typical services such as voice [KG10], the approximation $\log(1 + x) \approx x$ is valid, in which case the constraint can be written as

$$w_{n,m} \cdot \mathrm{SINR}_{n,m} \geq a_{n,m} \cdot \gamma_n, \ \forall n, m. \tag{7.5}$$

The modified *spectral efficiency* for cell m is defined as $x_m = \sum_{n=1}^{N} a_{n,m} \cdot \frac{\gamma_n}{w_{n,m}}$.We use the minimum rate requirement γ_n instead of the actual rate. In this way the expression takes the targeted user load implicitly into account. The problem with such a definition is however that the nonlinear relationship between variables $a_{n,m}$ and $w_{n,m}$ for each $n \in \mathcal{N}$ and $m \in \mathcal{M}$ makes the optimization problem difficult to handle. Moreover, x_m becomes unbounded when $w_{n,m} \to 0$. An alternative definition is therefore proposed, which is well defined due to linearity

$$x_m = \sum_{n=1}^{N} \big(a_{n,m} \cdot \gamma_n - \delta \cdot w_{n,m}\big), \ \forall m, \tag{7.6}$$

where $\delta \geq 0$ is a tuning parameter. The higher the δ, the higher the cost of the bandwidth resource. In the following we will use the terms "spectral efficiency" and "BS load" interchangeably, when referring to x_m. Since the load should be positive we get a further limitation on $w_{n,m}$ according to

$$0 \leq w_{n,m} \leq a_{n,m} \cdot \min\left\{\frac{\gamma_n}{\delta}, W\right\}, \forall n, m. \tag{7.7}$$

7.3 Problem Formulation

Each BS is assigned to a utility function of the spectral efficiency $U_m(x_m)$ which reflects the level of satisfaction. The function is strictly increasing in x_m and also strictly concave to discourage the assignment of additional resources to BSs which already have relatively high load. Different choice in the utility functions for the same set of constraints leads naturally to a different operational point. Potential choices are [SWB09]

$$U(x) = \begin{cases} \frac{x^{1-\beta}}{1-\beta} & \beta > 1 \\ \log x & \beta = 1 \end{cases}. \tag{7.8}$$

The general optimization problem is to maximize the aggregate utility function, subject to certain operational constraints:

$$\begin{array}{ll} \max_{\mathbf{x},\mathbf{a},\mathbf{w}} & \sum\limits_{m=1}^{M} U_m(x_m) \\ \textbf{s.t.} & (7.1),\ (7.2),\ (7.5),\ (7.6),\ (7.7) \\ & a_{n,m} \in \{0,1\},\ w_{n,m}, x_m \in \mathbf{R}_+. \end{array} \tag{7.9}$$

Problem (7.9) is a *mixed integer program* which can be solved by a centralized optimizer. However, we aim in this work at a distributed operation of the BSs which can approximate the maximum of the objective function.

7.3.1 Linearization of the Constraint Set

We first observe that the inequalities in (7.5) are non-linear for the variables $a_{n,m}$ and $w_{n,s}$, for some $s \neq m$. We have to transform the constraint set into a set of linear inequalities so that the problem obtains a form more easy to handle.

We have that for binary assignment variables $a_{n,m}$, the inequalities in (7.5) are equivalent to linear constraints using the so-called *big-M factor*

$$\begin{aligned} p \cdot h_{n,m} w_{n,m} &\geq \gamma_n a_{n,m} \left[\textstyle\sum_{s \in \mathcal{M} \setminus \{m\}} \mathcal{I}_{n,s} + \sigma_n \right] \Leftrightarrow \\ (1 - a_{n,m}) \cdot \mathbf{M}_{n,m} + p \cdot h_{n,m} w_{n,m} &\geq \gamma_n \left[\textstyle\sum_{s \in \mathcal{M} \setminus \{m\}} \mathcal{I}_{n,s} + \sigma_n \right], \forall n, m \end{aligned} \tag{7.10}$$

where

$$\mathbf{M}_{n,m} \; := \; \gamma_n \left[\sum_{s \in \mathcal{M} \setminus \{m\}} \mathcal{I}^{\max}_{n,s} + \sigma_n \right] \tag{7.11}$$

and $\mathcal{I}^{\max}_{n,s}$ is the maximum value that the interference function from s to user n can take - considering assignment of user n to BS m. This equals

$$\mathcal{I}^{\max}_{n,s} \overset{(7.3)}{=} p \cdot h_{n,s} \cdot 1. \tag{7.12}$$

Then (7.10) can be understood as follows. When $a_{n,m} = 1$ the QoS requirements for user n should be satisfied by BS m. When $a_{n,m} = 0$ the constraint is automatically satisfied due to the positive term activated at the left-hand side of the inequality. This is definitely greater or equal to the right-hand side irrespective of $w_{n,m}$. Then for $a_{n,m} \in \{0, 1\}$ the two inequalities are equivalent and the mixed-integer problem (7.9) can be rewritten as

$$\begin{array}{ll} \max_{\mathbf{x},\mathbf{a},\mathbf{w}} & \sum_{m=1}^{M} U_m(x_m) \\ \textbf{s.t.} & (7.1), (7.2), (7.10), (7.6), (7.7) \\ & a_{n,m} \in \{0,1\},\ w_{n,m}, x_m \in \mathbf{R}_+. \end{array} \tag{7.13}$$

The above constraint set is denoted by $\mathcal{F}$. Let us further denote the optimal value of the objective by Z^*. Observe that by relaxing the binary constraint for the assignment variables (which we **do not** do however in our approach) so that $a_{n,m} \in [0, 1]$ the above optimization problem is a *convex program* with a concave objective function and linear constraint set which can be solved by known techniques [BV04]. The optimal solution however does not exhibit integrality. The optimal value of the objective for the linear relaxation is denoted by Z^*_L.

7.4 Lagrangian Relaxation

We proceed by relaxing the equality in (7.6) which defines the BS load and the inequality in (7.10) for the transformed QoS constraints. To do this we relate with each equality a real Lagrange multiplier $\lambda_m \in \mathbf{R}$ and with each inequality a real non-negative Lagrange multiplier $\mu_{n,m} \in \mathbf{R}_+$ and add them to the objective. For given $\lambda_m, \mu_{n,m}, \forall n, m$, we get the Lagrangian of our problem

$$\begin{aligned} q(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \max_{\mathbf{x}} \sum_m \left[U_m(x_m) - \lambda_m x_m\right] \\ &+ \max_{\mathbf{a}} \sum_{m,n} \left[\lambda_m \gamma_n a_{n,m} + \mu_{n,m}(1 - a_{n,m}) \mathbf{M}_{n,m}\right] \\ &+ \max_{\mathbf{w}} \sum_{m,n} \Bigg[- \delta \lambda_m w_{n,m} + \mu_{n,m} p h_{n,m} w_{n,m} - \\ & \quad -\mu_{n,m}\gamma_n \left(\sum_{s \in \mathcal{M} \setminus \{m\}} I_{n,s} + \sigma_n \right) \Bigg] \end{aligned} \tag{7.14}$$

and the maximization above is taken over the constraint set

$$\mathcal{F}_{LR} := \{a_{n,m} \in \{0,1\}, w_{n,m}, x_m \in \mathbf{R}_+, \forall n, m | (7.1), (7.2), (7.7)\}$$

An important property is that $\forall (\boldsymbol{\lambda}, \boldsymbol{\mu})$ it holds that $q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \geq Z^*$ and hence the *weak duality* property [BT97] holds

$$Z^*_{LR} := \min q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \geq Z^*. \tag{7.15}$$

7.4.1 Decomposition

In $\mathcal{F}_{LR}$ the variables $a_{n,m}$ and $w_{n,m}$ are related through (7.7). The constraint actually states that when $a_{n,m} = 0$ then necessarily the bandwidth variable is also $w_{n,m} = 0$ otherwise

$$0 \leq w_{n,m} \leq \min\left\{\frac{\gamma_n}{\delta}, W\right\} \tag{7.16}$$

In other words the solution is not allowed to give positive bandwidth when there is no assignment. We can consider however an enlarged constraint set

$$\mathcal{F}'_{LR} := \{\mathbf{a}, \mathbf{w}, \mathbf{x} | (7.1), (7.2), (7.16)\} \supseteq \mathcal{F}_{LR}$$

where we replace the constraint in (7.7) by (7.16). By solving (7.14) over this, we can see that the solution for the assignment variables will not be influenced. The possibility of allocating positive bandwidth to (n, m) pairs where there is no assignment is now allowed. We will see however in the next section that the optimality conditions do not allow such a case and the solution of the two problems is the same. A direct gain by this change in constraints is that we achieve a decomposition of the problem into subproblems which are easier to handle.

Proposition 7.1. *Consider the mixed integer problem in (7.13) and replace inequality (7.7) by (7.16). Then the Lagrangian of the problem which results by relaxing the constraints (7.6) and (7.10) decomposes into three subproblems:*

- ***Load Distribution**: The optimal load per BS is given by solving over x_m, $\forall m$*

$$\max_{x_m} \quad U_m\left(x_m\right) - \lambda_m x_m \tag{7.17}$$

- ***BS Assignment**: The optimal assignment of each user n to a single BS is derived by solving $\forall n$ over $\mathbf{a}_n := \left[a_{n,1}, \ldots, a_{n,M}\right]$*

$$\begin{array}{ll} \max_{\boldsymbol{a}_n} & \sum\limits_m \left[\lambda_m \gamma_n a_{n,m} + \mu_{n,m}\left(1 - a_{n,m}\right)\mathbf{M}_{n,m}\right] \\ \boldsymbol{s.t.} & \sum\limits_m a_{n,m} = 1 \end{array} \tag{7.18}$$

- ***Bandwidth Allocation**: The optimal bandwidth allocation is derived by solving over* **w**

$$\begin{array}{ll} \max_{\mathbf{w}} & \sum\limits_m \sum\limits_n \left[-\delta\lambda_m w_{n,m} + \mu_{n,m} p h_{n,m} w_{n,m} - \mu_{n,m}\gamma_n\left(\sum\limits_{s \in \mathcal{M}\setminus\{m\}} I_{n,s} + \sigma_n\right)\right] \\ \boldsymbol{s.t.} & \sum\limits_n w_{n,m} \leq W, \ \forall m \\ & 0 \leq w_{n,m} \leq \min\left\{\frac{\gamma_n}{\delta}, W\right\}, \ \forall n, m \end{array} \tag{7.19}$$

7.5 A Lagrangian Relaxation Approach

7.5.1 Solution for Given Prices

Given a set of Lagrange multipliers $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ named from now on also *prices*, we can find the optimal values on load, BS assignment and bandwidth allocation by solving each one of the above subproblems respectively.

• For the load distribution of the problem the optimal solution is given by solving (7.17), which for a fixed value λ_m of *load price* per BS satisfies the expression

$$\frac{dU_m\left(x_m\right)}{dx_m} \quad = \quad \lambda_m \tag{7.20}$$

Using as an example $U_m\left(x_m\right) = \log\left(x_m\right)$, $\forall m$, the above results in the solution $x_m = \lambda_m^{-1}$.

• The BS assignment problem is solved for each user. Problem (7.18) is a discrete optimization problem which can be rephrased into finding for each user n the BS m_n which maximizes the expression

$$\begin{aligned} m_n &= \arg\max_m \left\{ \lambda_m + \sum_{k \in \mathcal{M}\setminus\{m\}} \mu_{n,k} \frac{\mathbf{M}_{n,k}}{\gamma_n} \right\} \\ &\overset{(a)}{=} \arg\max_m \left\{ c + \lambda_m - \mu_{n,m} \frac{\mathbf{M}_{n,m}}{\gamma_n} \right\} \\ &\overset{(7.11),(b)}{=} \arg\max_m \left\{ \lambda_m - \mu_{n,m} \left(\sum_{s \in \mathcal{M}\setminus\{m\}} I_{n,s}^{\max} + \sigma_n \right) \right\} \end{aligned} \tag{7.21}$$

where (a) comes by adding and subtracting the term $\mu_{n,m}\mathbf{M}_{n,m}/\gamma_n$ and c is a term constant and equal to $c := \sum_{k=1}^{M} \mu_{n,k}\mathbf{M}_{n,k}/\gamma_n$ and hence can be removed from the objective. (b) results by further substituting the expression for the big-M factor.

It is obvious from (7.21.b) that user n is assigned to the BS with a maximum linear combination of (i) positive load price and (ii) negative sum of maximum interference from the other BSs, weighted by the price $\mu_{n,m} \geq 0$. This is reasonable because the user should be given to a BS which still has enough "room" to accept users (this is better understood by considering the log-utility expression, where $\lambda_m = x_m^{-1}$) and at the same time suffers by as low interference as possible from the rest of the system.

• Considering the bandwidth allocation problem in (7.19), we simplify the constraints by assuming that the total bandwidth available is large enough $W >> 1$ so that the constraint (7.2) is always satisfied with strict inequality. Then our subproblem can be solved for each $w_{n,m}$, $n \in \mathcal{N}$ and $m \in \mathcal{M}$. More specifically, by differentiating the objective in (7.19) over $w_{n,m}$ we get the expression

$$\epsilon_{n,m} := -\delta\lambda_m + \mu_{n,m} p h_{n,m} - \mathcal{J}_m \tag{7.22}$$

where $\mathcal{J}_m$ is a characteristic value for each BS m, given a vector $\boldsymbol{\mu}$, and will be called from now on the *interference cost*

$$\begin{aligned} \mathcal{J}_m &:= \sum_{s \neq m} \sum_j \mu_{j,s}\gamma_j \frac{\partial I_{j,m}}{\partial w_{n,m}} \\ &\overset{(7.3)}{=} \sum_{s \neq m} \sum_j \mu_{j,s}\gamma_j \frac{p \cdot h_{j,m}}{W} \geq 0 \end{aligned} \tag{7.23}$$

Then the power allocation follows the rule:

$$w_{n,m} = \begin{cases} \min\left\{\frac{\gamma_n}{\delta}, W\right\} & \text{if } \epsilon_{n,m} > 0 \\ 0 & \text{if } \epsilon_{n,m} < 0 \\ \omega \in \left(0, \min\left\{\frac{\gamma_n}{\delta}, W\right\}\right) & \text{if } \epsilon_{n,m} = 0 \end{cases} . \tag{7.24}$$

which is easy to be understood since the sign of $\epsilon_{n,m}$ defines the monotonicity of the objective function depending on the Lagrange multipliers. To further get an intuition for the result in (7.24), we see that the case $\epsilon_{n,m} \geq 0$ gives the condition

$$ph_{n,m} \quad \geq \quad \frac{\delta\lambda_m + \mathcal{J}_m}{\mu_{n,m}}$$

which is a threshold rule for bandwidth assignment. If the power of the received signal is above a price-dependent threshold and $\mu_{n,m} \neq 0$, then the user is allocated the maximum possible bandwidth from BS m, otherwise 0. If $\mu_{n,m} = 0$ and either λ_m or one of the $\mu_{n,s},\ s \neq m$ multipliers is not zero then the assignment is always 0 bandwidth.

To summarize the results we provide the following proposition.

Proposition 7.2. *Given a price vector* $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ *for the relaxed problem (7.14) under the constraint set* $\mathcal{F}'_{LR}$ *and assuming* $W >> 1$*, the optimal load per BS is the solution to*

$$U'_m(x_m) = \lambda_m. \tag{7.25}$$

Furthermore, each user n *is assigned to BS* m_n *s.t.*

$$m_n = \arg\max_m \left\{ \lambda_m - \mu_{n,m} \left(\sum_{s \in \mathcal{M} \setminus \{m\}} ph_{n,s} + \sigma_n \right) \right\} \tag{7.26}$$

and is allocated bandwidth $w_{m,n} = \frac{\gamma_n}{\delta}$ *(or* ω *- see (7.24)) for each BS with channel quality above the threshold*

$$\mu_{n,m} ph_{n,m} \geq \delta\lambda_m + \mathcal{J}_m \quad \& \quad \mu_{n,m} \neq 0. \tag{7.27}$$

If $\mu_{n,m} = 0$ *then necessarily* $w_{n,m} = 0$.

We observe here that there may be a certain inconsistency between the assignment of a user n to a single BS satisfying (7.26) and the allocation of positive bandwidth to possibly more than one BS satisfying the thresholding rule in (7.27). The reason for this is the change of the constraint set from $\mathcal{F}_{LR}$ to $\mathcal{F}'_{LR}$, which replaced (7.7) by (7.16). In the following subsection we will see how this is resolved.

7.5.2 Optimal Solution

Denote by $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ and by $(\boldsymbol{x}^*, \boldsymbol{a}^*, \boldsymbol{w}^*)$ the optimal primal and dual solution of the Lagrangian problem in (7.15). Then the following *complementary slackness conditions*, related to the relaxed QoS constraints $\forall n, m$, should be satisfied, $\mu^*_{n,m} \geq 0$

$$\mu^*_{n,m} \cdot \left[\left(1 - a^*_{n,m}\right) \cdot \mathbf{M}_{n,m} + p \cdot h_{n,m} w^*_{n,m} - \gamma_n \left(\sum_{s \in \mathcal{M} \setminus \{m\}} \mathcal{I}^*_{n,s} + \sigma_n \right) \right] = 0. \tag{7.28}$$

The equality is fulfilled when

$$\begin{cases} \text{Case I:} & \mu^*_{n,m} = 0 \\ \text{Case II:} & \mu^*_{n,m} > 0 \quad \& \ \left(a^*_{n,m} = 1 \ \& \ \mathrm{SINR}^*_{n,m} = \gamma_n\right) \end{cases}$$

The above implies that for the optimal solution there exists no difference between using $\mathcal{F}_{LR}'$ instead of $\mathcal{F}_{LR}$. To see this, let for a user n be optimal not to be assigned to some BS m, then $a^*_{n,m} = 0$. The quantity in brackets (7.28) is non-zero and $\mu^*_{n,m} = 0$ necessarily. But by Proposition 7.2 this further suggests that $w^*_{n,m} = 0$. Another interesting property for the optimal bandwidth allocation is given below.

Proposition 7.3. *If $a^*_{n,m} = 1$ then either $w^*_{n,m} = 0$ or $w^*_{n,m} = \omega \in \left(0, \min\left\{\frac{\gamma_n}{\delta}, W\right\}\right]$, such that $\mathrm{SINR}^*_{n,m} = \gamma_n$.*
If $a^*_{n,m} = 0$ then $w^*_{n,m} = 0$.

Proof. For $a^*_{n,m} = 1$ the complementary slackness condition in (7.15) is satisfied either when $\mu^*_{n,m} = 0$ or when $\mu^*_{n,m} > 0$ and $\mathrm{SINR}_{n,m} = \gamma_n$. If $\mu^*_{n,m} = 0$ then by Prop. 7.2 the bandwidth $w^*_{n,m} = 0$. In the other case the bandwidth is chosen such that the QoS constraint is fulfilled with equality. For the case of $a^*_{n,m} = 0$ the arguments are given above the Proposition. ■

Proposition 7.4. *For a user n the optimal BS to be assigned to is the one for which*

$$m^*_n = \arg \min_{m \in \mathcal{M}_\lambda} \mathcal{J}^*_m \tag{7.29}$$

where

$$\mathcal{M}_\lambda = \left\{m : \lambda^*_m = \max_m \lambda^*_m\right\} \tag{7.30}$$

Proof. Consider a user n and let m^*_n be the optimal BS assignment $a^*_{n,m^*_n} = 1$. Furthermore let $\hat{m} \neq m^*_n$ be one BS for which $a^*_{n,\hat{m}} = 0$. From the note above Prop.7.3 we have that $\mu^*_{n,\hat{m}} = 0$. Then (7.26) implies that

$$\lambda^*_{m^*_n} - \mu^*_{n,m^*_n}\left(\sum_{s \neq m^*_n} p h_{n,s} + \sigma_n\right) \geq \lambda^*_{\hat{m}} \quad \Rightarrow$$

$$\lambda^*_{m^*_n} - \lambda^*_{\hat{m}} \geq \mu^*_{n,m^*_n}\left(\sum_{s \neq m^*_n} p h_{n,s} + \sigma_n\right) \quad \geq 0$$

The above inequality implies that $\lambda^*_{m^*_n} \geq \lambda^*_{\hat{m}}$ and the user is assigned to the base station with maximum λ^*_m.

In the case that more than one BSs satisfy the above inequality for the case when $\lambda^*_{m^*_n} = \lambda^*_{\hat{m}} = \lambda^*$, we turn to the condition for the bandwidth allocation. Observe from (7.24) that non-negative bandwidth is assigned to the base station with non-negative derivative $\epsilon^*_{n,m}$. Since we would like to choose only one, this is the BS for which

$$m^*_n \quad = \quad \arg\max_m \left\{-\delta\lambda^* + \mu^*_{n,m} p h_{n,m} - \mathcal{J}^*_m\right\}$$

For all other $\hat{m}$ we have $\mu^*_{n,\hat{m}} = 0$ and following inequality holds

$$\begin{aligned} -\delta\lambda^* + \mu^*_{n,m^*_n} p h_{n,m^*_n} - \mathcal{J}^*_{m^*_n} \geq -\delta\lambda^* - \mathcal{J}^*_{\hat{m}} \qquad &\Rightarrow \\ \mathcal{J}^*_{m^*_n} - \mathcal{J}^*_{\hat{m}} \leq \mu^*_{n,m^*_n} p h_{n,m^*_n}, \quad &\forall \hat{m} \neq m^*_n \end{aligned}$$

Since the right-hand side is non-negative, the above set of inequalities will definitely be satisfied if we choose as m^*_n the BS with minimum $\mathcal{J}^*_m$. ■

7.5.3 Ascent Method

Consider again the initial problem in (7.13) with the concave objective and linear constraints and discrete assignment variables, which we rewrite here as

$$\max f(\boldsymbol{y}) \quad \textbf{s.t.} \quad \boldsymbol{y} \in \mathcal{F} \tag{7.31}$$

In the above $f(\boldsymbol{y}) := \sum_m U_m(x_m)$ and $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{w})$.

The solution of the Lagrangian relaxation which was investigated in the previous sections, provides only an upper bound for the optimal value. Furthermore, the decomposition is valid by assuming $W >> 1$, so that the constraint for total bandwidth per BS was considered always satisfied with strict inequality. Hence, a feasible solution is not guaranteed when the W takes some realistic restricted values.

The Lagrangian solution however provides guidelines over the structure of the optimal solution. To derive an algorithm which solves the problem, we will use in the following a variation of the so-called *ascent methods* proposed in [BV04] and adapted here to the mixed integer setting we have to deal with. Given any feasible vector $\boldsymbol{y} \in \mathcal{F}$, which is not the optimal solution, we will call a *feasible ascent direction* $\boldsymbol{d} = \Delta\boldsymbol{y} = \hat{\boldsymbol{y}} - \boldsymbol{y}$ any $\boldsymbol{d}$ which fulfills

$$\begin{aligned} \boldsymbol{y} + \boldsymbol{d} \in \mathcal{F} \quad &\& \quad f(\boldsymbol{y} + \boldsymbol{d}) \geq f(\boldsymbol{y}) \Rightarrow \\ \hat{\boldsymbol{y}} \in \mathcal{F} \quad &\& \quad f(\hat{\boldsymbol{y}}) \geq f(\boldsymbol{y}) \end{aligned} \tag{7.32}$$

What we aim for is to generate a sequence of feasible vectors $\left\{\boldsymbol{y}^k\right\}$, $k = 0, 1, \ldots$ which step-wise increases the value of the objective for the problem. The vector $\boldsymbol{y}^k$ describes a state of the system with assignment variables $\boldsymbol{a}^k$ and bandwidth allocation $\boldsymbol{w}^k$. To choose a feasible direction we will work as follows:

- Choose an appropriate pair of *overloaded* BS (OL) and *target* BS (TR), using the guidelines from the Lagrangian solution.

- Define all possible subsets of users $\mathcal{C}_q$, which at iteration k are assigned to OL and is possible to be shifted to TR, as long as the new vector is feasible $\hat{\boldsymbol{y}}_q \in \mathcal{F}$.

- Find the subset $\mathcal{C}_{q^*}$, which if reallocated provides the maximum improvement, in other words

$$\boldsymbol{y}^{k+1} = \hat{\boldsymbol{y}}_{q^*} = \arg\max_{q} \left\{ f\left(\hat{\boldsymbol{y}}_q\right) - f\left(\boldsymbol{y}^k\right) \right\} \tag{7.33}$$

- Continue the iteration as long as no more improvement in the objective is possible.

Since we aim at providing an algorithm possible to be implemented in LTE advanced cellular networks, the users should be encouraged to change cell by proper adaptation of the HO parameters per cell. In the following sections we will explain how the HO parameters work in the network and which adaptation is necessary to fulfill (7.33). An appropriate algorithm will be finally derived.

7.6 Cellular Network Aspects

We consider that each step k of the algorithm depends on the following variables, which will be explained in more detail in the following paragraphs

$$\mathcal{S}^k \quad := \quad \left\{ m_+^k, m_-^k, \boldsymbol{\lambda}^k, \boldsymbol{J}^k, W_{m_+^k}^k, \mathcal{C}_{q^*}^k \right\}. \tag{7.34}$$

7.6.1 Choice of OL-TR Pair

A first issue for the implementation of the algorithm suggested above is the choice of an appropriate OL-TR pair of BSs. Then users from the OL cell could be removed towards the TR cell for a better balance of the load. We will use the guidelines of Prop.7.4 which gives the Lagrangian optimal BS assignment.

Based on that, during iteration step k a cell m_+^k is activated as a TR cell if

$$m_+^k \quad = \quad \arg\min_{m \in \mathcal{M}_{\lambda^k}} \mathcal{J}_m^k \tag{7.35}$$

$$\mathcal{M}_{\lambda^k} \quad = \quad \left\{ m : \lambda_m^k = \max_m \lambda_m^k \right\} \tag{7.36}$$

which means that we choose the cell with maximum utility derivative equal to the load price and minimum interference cost towards the neighboring BSs. An alternative way used in the algorithm an simulations section on this work is by choosing the BS which maximizes the linear combination

$$m_{+}^{k\,\prime} \;=\; \arg\max_{m\in\mathcal{M}} \left\{\lambda_m^k - \alpha\cdot\mathcal{J}_m^k\right\} \tag{7.37}$$

The OL cell is chosen anti-symmetrically as

$$m_{-}^{k} \;=\; \arg\min_{m\in\mathcal{M}} \left\{\lambda_m^k - \beta\cdot\mathcal{J}_m^k\right\} \tag{7.38}$$

where $\alpha,\ \beta \geq 0$ are tuning factors giving higher or lower weight on the interference cost.

For such choices to be made, knowledge of the vectors $\boldsymbol{\lambda}^k, \boldsymbol{J}^k$ at the BS side is necessary. We know that each BS can calculate its load price λ_m^k using (7.25). For this it needs to calculate the current load value x_m^k based on the subset of users it supports $\mathcal{N}_m^k$.

Considering the interference costs, we see from (7.23) that for each BS m these depend on the Lagrangian dual variables $\mu_{n,s}$ for $\forall s \neq m, \forall n$. Furthermore, we know that for the Lagrangian solution $\mu_{n,m} = 0$ if $a_{n,m} = 0$. Setting all activated $\mu_{n,m} = 1$ we get that the value of $\mathcal{J}_m^k$ can be written as

$$\mathcal{J}_m^k \;=\; \sum_{n\in\mathcal{N}\setminus\mathcal{N}_m^k} \gamma_n \frac{p\cdot h_{n,m}}{W} \tag{7.39}$$

which can be calculated by BS m if knowledge over the channel $h_{n,m}$ through RSRP measurements is available.

7.6.2 Handover Criterion

The assignment of users to cells is controlled by the handover parameters of the cells. Using the notation conventional in the 3GPP literature, $\mathrm{RSRP}_{n,m}$ denotes the filtered received signal strength (for more details see Section 6.2.1) of user n from BS m and is an indicator of the SINR, Hys_m is a cell-related hysteresis factor and $\mathrm{CIO}_{s\to m}$ is a control parameter for the ordered BS pair (s,m) called Cell Individual Offset. Furthermore, let us define the difference

$$\Delta\mathrm{RSRP}_n(s,m) \;:=\; \mathrm{RSRP}_{n,m} - \mathrm{RSRP}_{n,s} \tag{7.40}$$

A user belonging to BS m_-^k (and we write $n \in \mathcal{N}_{m_-^k}^k$), can be handed over to BS m_+^k if the following criterion is satisfied

$$\mathrm{CIO}_{m_-^k\to m_+^k} \;\geq\; -\Delta\mathrm{RSRP}_n(m_-^k, m_+^k) + \mathrm{Hys}_{m_-^k} \tag{7.41}$$

The above inequality says that a user n will be handed over to the TR cell if the value of the control parameter denoted by $\mathrm{CIO}_{m_-^k\to m_+^k}$, is set greater or equal to the negative difference of channel qualities for user n, increased by the hysteresis factor at the OL cell.

To avoid the so called ping-pong effect, which would allow the user n already handed-over, to return to its OL cell, the following condition for the mirror-parameter $\mathrm{CIO}_{m_+^k \to m_-^k}$ should be satisfied

$$\mathrm{CIO}_{m_+^k \to m_-^k} \leq \Delta\mathrm{RSRP}_n(m_-^k, m_+^k) + \mathrm{Hys}_{m_+^k}. \tag{7.42}$$

7.6.3 Candidate User Subsets

A user $n \in \mathcal{N}_{m_-^k}$ is included to the candidate set $\mathcal{C}^k$ if the required bandwidth for reallocation from the OL to the TR cell - while the QoS criterion is fulfilled (see also Prop.7.3) - satisfies the inequality

$$w_{n,m_+^k} \leq \min\left\{\frac{\gamma_n}{\delta}, W_{m_+^k}^k\right\}. \tag{7.43}$$

where $W_{m_+^k}^k$ is the available free bandwidth in BS m_+^k. We denote the cardinality of this set by $|\mathcal{C}^k|$.

We construct $|\mathcal{C}^k|$ candidate subsets, each denoted as $\mathcal{C}_q^k$, $q \in \{1, \ldots, |\mathcal{C}^k|\}$ by the following procedure. We order the elements (users) of the set $\mathcal{C}^k$ by decreasing channel differences. The order $n_1, n_2, \ldots, n_{|\mathcal{C}^k|}$ refers to the order $\Delta\mathrm{RSRP}_{n_1}(m_-^k, m_+^k) \geq \Delta\mathrm{RSRP}_{n_1}(m_-^k, m_+^k) \geq \ldots \geq \Delta\mathrm{RSRP}_{n_{|\mathcal{C}^k|}}(m_-^k, m_+^k)$. From this, following sets can be constructed

$$\begin{aligned} \mathcal{C}_1^k &= \{n_1\} \\ \mathcal{C}_2^k &= \{n_1, n_2\} \\ \cdots & \quad \cdots \\ \mathcal{C}_{|\mathcal{C}^k|}^k &= \left\{n_1, n_2, \ldots, n_{|\mathcal{C}^k|}\right\} \end{aligned}$$

The HO parameters are then mapped to the above sets, so that (7.41) and (7.42) are satisfied after the handover for all users belonging to some subset $\mathcal{C}_k^q$. Which will be the optimal subset chosen will be defined in the following paragraph. The appropriate CIO parameters become

$$\mathrm{CIO}_{m_-^k \to m_+^k}^q = -\Delta\mathrm{RSRP}_{n_q}(m_-^k, m_+^k) + \mathrm{Hys}_{m_-^k} \tag{7.44}$$

$$\mathrm{CIO}_{m_+^k \to m_-^k}^q = \Delta\mathrm{RSRP}_{n_q}(m_-^k, m_+^k) + \mathrm{Hys}_{m_+^k} \tag{7.45}$$

7.6.4 Optimal User Subset

For all candidate user subsets, the vectors $\hat{\boldsymbol{y}}_q^k = (\boldsymbol{x}_q^k, \boldsymbol{a}_q^k, \boldsymbol{w}_q^k)$ can be easily calculated for each q given the vector $\hat{\boldsymbol{y}}^k = (\boldsymbol{x}^k, \boldsymbol{a}^k, \boldsymbol{w}^k)$, by changing the assignment and bandwidth variables for the possible handed-over users and re-calculating the load. The optimal user subset $\mathcal{C}_{q^*}^k$ is chosen such that (7.33) is satisfied, in other words as the one with maximum increase of the objective.

7.6.5 Distributed Algorithm

Based on the above we present in what follows an algorithm for the optimal load balancing among BSs of a cellular wireless network, taking ICI and adaptation of the HO parameters into consideration. The steps are given below

```
Algorithm 4: Distributed load balancing algorithm
  Input: A possibly unbalanced but feasible BS-User association and BW allocation
         $\boldsymbol{y}^0$
  Output: Enhanced sum of utilities and adequate reconfiguration of the system HO
          parameters
  Initialization: Initial user assignment $\boldsymbol{a}^0$ and bandwidth allocation $\boldsymbol{w}^0$. All users $\mathcal{N}$
  gain knowledge over the channel through RSRP measurements. Afterwards they
  communicate their channel quality vector $\boldsymbol{h}_n := [h_{n,1}, \ldots, h_{n,M}]$ and QoS demand $\gamma_n$
  to all BSs $\mathcal{M}$. The channel is considered constant throughout the iterations.
  Repeat at each step $k$
    1. Each BS has knowledge of its set of assigned users $\mathcal{N}_m^k$. Then it calculates:
         • The current load $x_m^k$ using (7.6).
         • The current load price $\lambda_m^k$ using (7.25).
         • The current interference cost $\mathcal{J}_m^k$ using (7.39).
    2. The BSs exchange the current values of $\lambda_m^k$ and $\mathcal{J}_m^k$ with their direct neighbors.
    3. Using (7.35), (7.36) (or (7.37) alternatively) and (7.38) and the knowledge over the
       other prices, each BS can decide whether it is a TR or OL cell for its neighborhood.
    4. The OL cell initiates a communication process with the TR cell.
    5. All possible candidate user subsets $\mathcal{C}_q^k$ are defined using also (7.43) and the TR and
       OL BSs calculate the possible change in load $\hat{x}_{\mathrm{OL},q}^k, \hat{x}_{\mathrm{TR},q}^k$ and utility
       $\Delta U\left(\hat{\boldsymbol{x}}_q^k, \boldsymbol{x}^k\right) = U_{\mathrm{TR}}\left(\hat{x}_{\mathrm{TR},q}^k\right) + U_{\mathrm{OL}}\left(\hat{x}_{\mathrm{OL},q}^k\right) - \left(U_{\mathrm{TR}}\left(x_{\mathrm{TR}}^k\right) + U_{\mathrm{OL}}\left(x_{\mathrm{OL},q}^k\right)\right)$
    6. The user set $\mathcal{C}_{q^*}^k$ which maximizes $\Delta U\left(\hat{\boldsymbol{x}}_q^k, \boldsymbol{x}^k\right)$ is chosen.
    7. The CIOs are reconfigured based on (7.44) and (7.45) to force users to migrate form
       OL to TR.
    8. Update variables $\boldsymbol{y}^{k+1} = \boldsymbol{y}_{q^*}^k$
  Until $\boldsymbol{\lambda}^k = \boldsymbol{\lambda}^{k-1}$ and $\boldsymbol{J}^k = \boldsymbol{J}^{k-1}$ for some $k \geq 1$.
```

7.7 Simulation Results

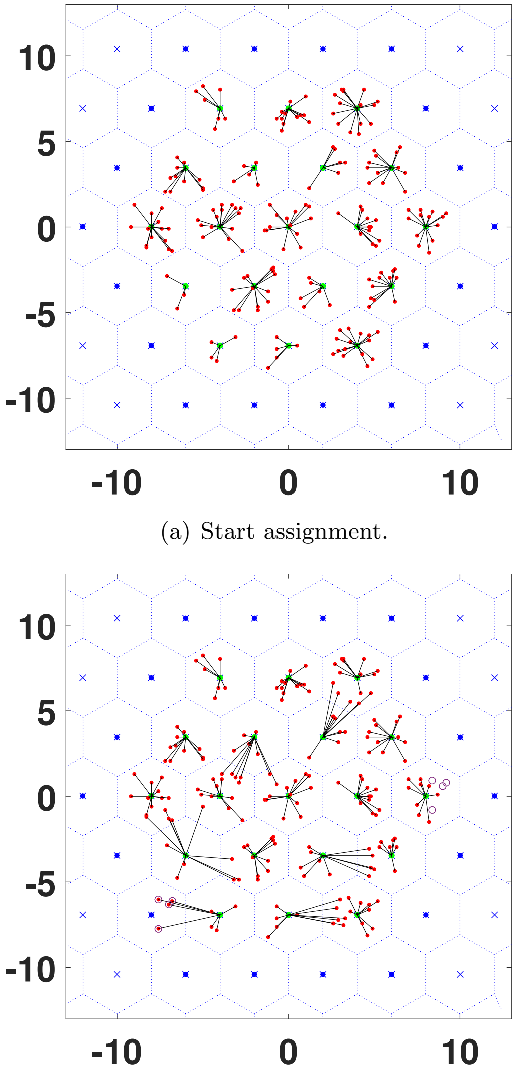
The algorithm is implemented on an LTE cellular network model with 19 cells artificially wrapped around at the border, so that the edge-cells include the cells on the opposite side in their neighborhood. Users with QoS requirement $\gamma_n = 14.4$ kbit/s are assumed to be static but randomly distributed on the plane with average number per cell $|\mathcal{N}_{cl}| = 10$. The channel quality per user-BS pair is a random realization with Rayleigh distribution. The transmission power density p is fixed and normalized to 1 Joule/Hz/s. The total bandwidth W per cell is equal to 0.5 MHz and shared among all BSs. The utility function is chosen for the implementations equal to $U(x) = \log(x)$.

The UE assignments before and after applying the proposed LB algorithm are presented in Fig.7.1. The colored small circles represent the handed-over users, i.e., the initial assignment and the optimized assignment. Fig.7.2(a) and Fig.7.2(b) illustrates how the prices λ_m and the load x_m for all BSs converge after just a few iterations. Thus, the algorithm can be very practical and robust in real system implementations. Furthermore, in Fig.7.2(c) the impact on the performance of the algorithm by modification of the tuning factors δ in (7.6) and α in (7.37) is demonstrated. A higher δ makes the algorithm more conservative considering bandwidth allocation, hence less re-assignments are performed while the total utility exhibits a reduced value. By choosing δ small, the BSs are more flexible to offer the free resource (to accept the handover users), as shown in Fig. 7.2(d). Higher α chooses BSs as TR cells with the priority focused on low interference cost. We see that the benefits are better for lower α since the reallocation of users becomes more dynamic by choosing TR cells with emphasis on the load price λ. However, although not illustrated here, there is the danger of exploiting very large amount of frequency resources for providing he desired QoS when α is low, which could lead to infeasibility very fast as the number of users increases.

7.8 Summary

The chapter starts with a thorough investigation on the state of art of the LB scheme for the self-organizing LTE networks. Notations and definitions are introduced with the system model. The general problem and the relaxed convex optimization problem are formulated, and the optimal solution is provided by solving the decomposited sub-problems with Karush-Kuhn-Tucker (KKT) conditions and the steepest decent method, which helps to choose the cell-pair distributedly and to select the UE groups to handover. The criterion for HO parameter adaptation is presented. The algorithm is proposed with a flowchart, followed by the simulation results and a complete analysis on the effects of the tuning factors δ and α. The paper ends with conclusions of the work and the future studies.

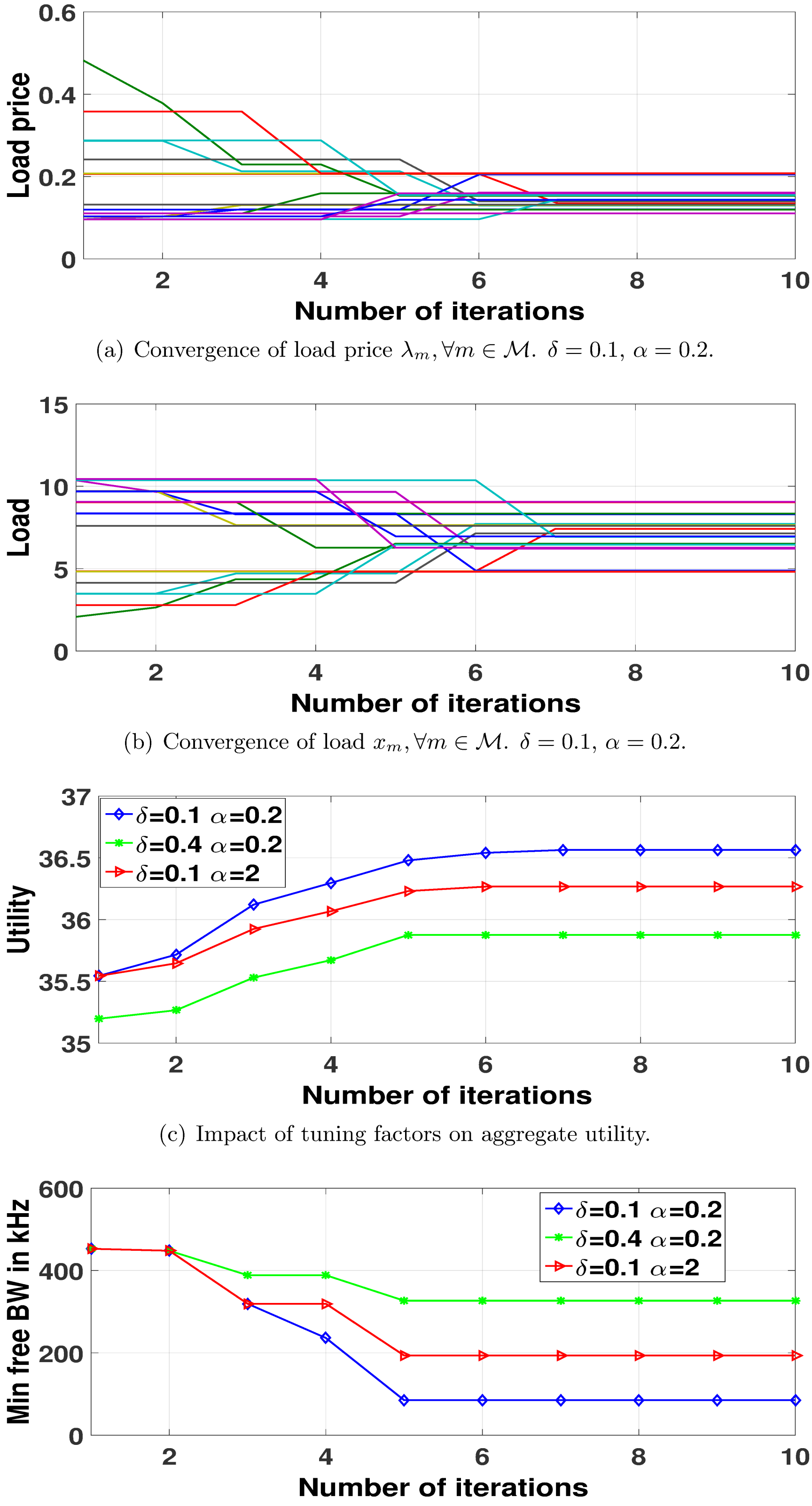
FIGURES



(a) Start assignment.

(b) Balanced assignment for small δ. $\delta = 0.1$, $\alpha = 0.2$.

Figure 7.1: Assignments.



(a) Convergence of load price $\lambda_m, \forall m \in \mathcal{M}$. $\delta = 0.1$, $\alpha = 0.2$.

(b) Convergence of load $x_m, \forall m \in \mathcal{M}$. $\delta = 0.1$, $\alpha = 0.2$.

(c) Impact of tuning factors on aggregate utility.

(d) Convergence of minimum free bandwidth.

Figure 7.2: Convergence of algorithm and aggregate utility improvement.

Part IV

Multi-Objective SON Function Optimization

Chapter 8

Joint Optimization of Coverage, Capacity and Load Balancing

This chapter develops an optimization framework for multi-objective optimization in SON. The objective is to ensure efficient network operation by a joint optimization of coverage, capacity and load balancing. Based on the axiomatic framework of standard interference functions, we formulate an optimization problem for the uplink and propose a two-step optimization scheme: i) per base station antenna tilt optimization and power allocation, and ii) cluster-based base station assignment of users and power allocation. We then consider the downlink, which is more difficult to handle due to the coupled variables, and show downlink-uplink duality relationship. As a result, a solution for the downlink is obtained by solving the uplink problem. Simulations show that our approach achieves a good trade-off between coverage and capacity.

Parts of this chapter have already been published in [15].

8.1 Introduction

A major challenge towards SON is the joint optimization of multiple SON use cases by coordinately handling multiple configuration parameters. Widely studied SON use cases include CCO, MLBO and MRO [3GPa]. However, most of these works study an isolated single use case and ignore contradictions among performance metrics [RKC10, 3].

In contrast, in this chapter we consider a joint optimization of multiple SON functionalities. The objective of this paper is to achieve a good trade-off between coverage and capacity performance, while achieving load-balanced network. The SON functionalities are usually implemented at the network management layer and are designed to deal with "long-term" network performance. Short-term optimization of individual users is left to lower layers of the protocol stack. To capture long-term global changes in a network, we consider a cluster-based network scenario, where users served by the same BS with similar SINR distribution are adaptively grouped into clusters. Our objective is to jointly optimize the following variables:

1) Per-cluster BS assignment and power allocation.

2) Per-BS antenna tilt optimization and power allocation.

The joint optimization of antenna tilt, transmit power and BS assignment in multi-cell scenario is an inherently challenging problem. The interference and the resulting performance measures depend on these variables in a complex and intertwined manner. A few studies have investigated joint optimization of multiple antenna configurations. For example, in [Kea12] a problem of jointly optimizing antenna tilt and cell selection to improve the spectral and energy efficiency is stated. In [FKVF13] the authors propose the algorithms that jointly adapt user association policies and antenna tilts based on an interference model. In [SVY06] the authors address automated optimization of service coverage and antenna configuration with three configuration parameters: transmit power, antenna tilt and antenna azimuth. However, in this paper we try to take one more step in multi-objective optimization based on the modeling of interference coupling. We aim to achieve a good tradeoff between coverage and capacity and to achieve load balancing by jointly optimizing antenna tilt, transmit power and BS assignment.

We propose a robust algorithmic framework built on a utility model, which enables fast and optimal uplink solutions and sub-optimal downlink solutions by exploiting three properties: i) the monotonic property of standard interference functions, ii) decoupled property of the antenna tilt and BS assignment optimization in the uplink network, and iii) uplink-downlink duality. The first property admits global optimal solution with fixed-point iteration for utility-based max-min fairness problems, while the second and third properties enable decomposition of the high-dimensional optimization problem. Our main contributions in this work can be summarized as follows:

1) We tackle a multi-objective optimization problem over a high dimensional action space. More specifically, We propose a max-min utility balancing algorithm for capacity-coverage trade-off optimization over antenna tilts, BS assignments and transmit powers. By distributing the interference fairly among the cells, load-balanced network is also achieved.

2) We provide an efficient algorithm to provide the optimal solution in the uplink by exploiting the interference patterns of standard interference function. Then, we decompose the high-dimensional optimization problem in downlink by utilizing uplink-downlink duality, and propose an efficient sub-optimal solution in downlink. Unlike

other studies which analyze the uplink-downlink duality for power control and beamforming in a max-min SINR fairness problem [BS06, SB05, HTR13, HHY$^+$12], we formulate the utility function as a convex combination of the coverage and the capacity metrics to jointly optimize transmit powers, antenna tilts and BS assignments.

8.2 System Model

We consider a multi-cell wireless network composed of a set of BSs $\mathcal{N} := \{1, \ldots, N\}$ and a set of users $\mathcal{K} := \{1, \ldots, K\}$. Using fuzzy C-means clustering algorithm [BEF84], we group users with similar SINR distributions[1] and served by the same BS into clusters. The clustering algorithm is beyond the scope of this paper. Let the set of user clusters be denoted by $\mathcal{C} := \{1, \ldots, C\}$, and let $\boldsymbol{A}$ denote a $C \times K$ binary user/cluster assignment matrix whose columns sum to one. The BS/cluster assignment is defined by a $N \times C$ binary matrix $\boldsymbol{B}$ whose columns also sum to one.

Throughout the paper, we assume a frequency flat channel. The average/long-term downlink path attenuation between N BSs and K users are collected in a channel gain matrix $\boldsymbol{H} \in \mathbb{R}^{N \times K}$. We introduce the cross-link gain matrix $\boldsymbol{V} \in \mathbb{R}^{K \times K}$, where the entry $v_{lk}(\theta_j)$ is the cross-link gain between user l served by BS j, and user k served by BS i, i.e., between the transmitter of the link (j, l) and the receiver of the link (i, k). Note that $v_{lk}(\theta_j)$ depends on the antenna downtilt θ_j. Let the BS/user assignment matrix be denoted by $\boldsymbol{J}$ so that we have $\boldsymbol{J} := \boldsymbol{B}\boldsymbol{A} \in \{0, 1\}^{N \times K}$, and $\boldsymbol{V} := \boldsymbol{J}^T\boldsymbol{H}$. We denote by $\boldsymbol{r} := [r_1, \ldots, r_N]^T$, $\boldsymbol{q} := [q_1, \ldots, q_C]^T$ and $\boldsymbol{p} := [p_1, \ldots, p_K]^T$the BS transmission power budget, the cluster power allocation and the user power allocation, respectively.

8.2.1 Inter-Cluster and Intra-Cluster Power Sharing Factors

We introduce the inter-cluster and intra-cluster power sharing factors to enable the transformation between two power vectors with different dimensions. Let $\boldsymbol{b} := [b_1, \ldots, b_C]^T$ denote the serving BSs of clusters $\{1, \ldots, C\}$. We define the vector of the inter-cluster power sharing factors to be $\boldsymbol{\beta} := [\beta_1, \ldots, \beta_C]^T$, where $\beta_c := q_c/r_{b_c}$. With the BS/cluster assignment matrix $\boldsymbol{B}$, we have $\boldsymbol{q} := \boldsymbol{B}_{\boldsymbol{\beta}}^T\boldsymbol{r}$, where $\boldsymbol{B}_{\boldsymbol{\beta}} := \boldsymbol{B}\,\mathrm{diag}\{\boldsymbol{\beta}\}$. Since users belonging to the same cluster have similar SINR distribution, we allocate the cluster power uniformly to the users in the cluster. The intra-cluster sharing factors are represented by $\boldsymbol{\alpha} := [\alpha_1, \ldots, \alpha_K]^T$ with $\alpha_k = 1/|\mathcal{K}_{c_k}|$ for $k \in \mathcal{K}$, where $\mathcal{K}_{c_k}$ denotes the set of users belonging to cluster c_k, while c_k denotes the cluster with user k. We have $\boldsymbol{p} := \boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q}$, where $\boldsymbol{A}_{\boldsymbol{\alpha}} := \boldsymbol{A}\,\mathrm{diag}\{\boldsymbol{\alpha}\}$. The transformation between BS power $\boldsymbol{r}$ and user power $\boldsymbol{p}$ is then $\boldsymbol{p} := \boldsymbol{T}\boldsymbol{r}$ where the transformation matrix $\boldsymbol{T} := \boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{B}_{\boldsymbol{\beta}}^T$.

[1] We assume the KL divergence as the distance metric

8.2.2 Signal-to-Interference-Plus-Noise Ratio

Given $\boldsymbol{V}$, the downlink SINR of the kth user depends on all transmission powers and is given by

$$\mathrm{SINR}_k^{\mathrm{DL}} := \frac{p_k \cdot v_{kk}(\theta_{n_k})}{\sum_{l \in \mathcal{K} \backslash k} p_l \cdot v_{lk}(\theta_{n_l}) + \sigma_k^2}, k \in \mathcal{K} \tag{8.1}$$

where n_k denotes the serving BS of user k, σ_k^2 denotes the noise power received in user k. Likewise, the uplink SINR is

$$\mathrm{SINR}_k^{\mathrm{UL}} := \frac{p_k \cdot v_{kk}(\theta_{n_k})}{\sum_{l \in \mathcal{K} \backslash k} p_l \cdot v_{kl}(\theta_{n_k}) + \sigma_k^2}, k \in \mathcal{K} \tag{8.2}$$

Assuming that there is no self-interference, the cross-talk terms can be collected in a matrix

$$[\tilde{\boldsymbol{V}}]_{lk} := \begin{cases} v_{lk}(\theta_{n_l}), & l \neq k \\ 0, & l = k \end{cases}. \tag{8.3}$$

Thus the downlink interference received by user k can be written as $I_k^{\mathrm{DL}} := [\tilde{\boldsymbol{V}}^T \boldsymbol{p}]_k$, while the uplink interference is given by $I_k^{\mathrm{UL}} := [\tilde{\boldsymbol{V}} \boldsymbol{p}]_k$.

A crucial property is that the uplink SINR of user k depends on the BS assignment n_k and the single antenna tilt θ_{n_k} alone, while the downlink SINR depends on the BS assignment vector $\boldsymbol{n} := [n_1, \ldots, n_K]^T$, and the antenna tilt vector $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_N]^T$. The decoupled property of uplink transmission has been widely exploited in the context of uplink and downlink multi-user beamforming [BS06] and provides a basis for the optimization algorithm in this paper.

The notation used in this paper is summarized in Table 8.1.

8.3 Utility Definition and Problem Formulation

As mentioned, the objective is to jointly optimize the performance of coverage, capacity and load balancing. We capture coverage by the worst-case SINR, while the average SINR is used to represent capacity. The load balancing can be achieved by distributing the inter-cell interference fairly among the cells. Given the cluster/user assignment, the network performance depends on: i) BS power allocation $\boldsymbol{r}$ and antenna downtilt $\boldsymbol{\theta}$, and ii) cluster power allocation $\boldsymbol{q}$ and BS/cluster assignment $\boldsymbol{b}$.[2]

In the following, we formulate a two-stage power allocation problem and then develop an iterative algorithm for optimizing BS variables $(\boldsymbol{r}, \boldsymbol{\theta})$ and cluster variables $(\boldsymbol{q}, \boldsymbol{b})$. We start with the problem statement and algorithmic approaches for the uplink. We then discuss the downlink in Section 8.5.

[2] The reader should note that user-specific variables $(\boldsymbol{p}, \boldsymbol{n})$ can be derived directly from cluster-specific variables $\boldsymbol{q}$ and $\boldsymbol{b}$, provided that cluster/user assignment $\boldsymbol{A}$ and intra-cluster power sharing factor $\boldsymbol{\alpha}$ are given.

Table 8.1: NOTATION SUMMARY

$\mathcal{N}$	set of BSs
$\mathcal{K}$	set of users
$\mathcal{C}$	set of user clusters
$\boldsymbol{A}$	cluster/user assignment matrix
$\boldsymbol{B}$	BS/cluster assignment matrix
$\boldsymbol{J}$	BS/user assignment matrix
c_k	cluster that user k is subordinated to
$\mathcal{K}_c$	set of users subordinated to cluster c
$\boldsymbol{H}$	channel gain matrix
$\boldsymbol{V}$	interference coupling matrix
$\tilde{\boldsymbol{V}}$	interference coupling matrix without intra-cell interference
$\tilde{\boldsymbol{V}}_{\boldsymbol{b}}$	interference coupling matrix depending on BS assignments $\boldsymbol{b}$
$\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}$	interference coupling matrix depending on antenna tilts $\boldsymbol{\theta}$
$\boldsymbol{r}$	BS power budget vector
$\boldsymbol{q}$	cluster power vector
$\boldsymbol{p}$	user power vector
$\boldsymbol{\alpha}$	intra-cluster power sharing factors
$\boldsymbol{\beta}$	inter-cluster power sharing factors
$\boldsymbol{A}_{\boldsymbol{\alpha}}$	transformation from $\boldsymbol{q}$ to $\boldsymbol{p}$, $\boldsymbol{p} := \boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q}$
$\boldsymbol{B}_{\boldsymbol{\beta}}$	transformation from $\boldsymbol{r}$ to $\boldsymbol{q}$, $\boldsymbol{q} := \boldsymbol{B}_{\boldsymbol{\beta}}^T\boldsymbol{r}$
$\boldsymbol{T}$	transformation from $\boldsymbol{r}$ to $\boldsymbol{p}$, $\boldsymbol{p} := \boldsymbol{T}\boldsymbol{r}$
$\boldsymbol{\theta}$	BS antenna tilt vector
$\boldsymbol{b}$	serving BSs of clusters
b_c	serving BS of cluster c
$\boldsymbol{n}$	serving BSs of the users
n_k	serving BS of user k
$\boldsymbol{\sigma}$	noise power vector
$P^{\max}$	sum power constraint

8.3.1 Cluster-Based BS Assignment and Power Allocation

Assume the per-BS variables $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{\theta}})$ are fixed, let the interference coupling matrix depend on BS assignment $\boldsymbol{b}$ in (8.3) be denoted by $\tilde{\boldsymbol{V}}_{\boldsymbol{b}}$. We define two utility functions indicating capacity and coverage per cluster respectively.

Average SINR Utility (Capacity)

With the intra-cluster power sharing factor introduced in Section 8.2.1, we have $\boldsymbol{p} := \boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q}$. Define the noise vector $\boldsymbol{\sigma} := [\sigma_1^2, \ldots, \sigma_K^2]^T$, the average SINR of all users in cluster c is written as

$$\begin{aligned}\bar{U}_c^{\mathrm{UL},1}(\boldsymbol{q},\boldsymbol{b}) &:= \frac{1}{|\mathcal{K}_c|}\sum_{k\in\mathcal{K}_c}\mathrm{SINR}_k^{\mathrm{UL}} \\ &= \frac{1}{|\mathcal{K}_c|}\sum_{k\in\mathcal{K}_c}\frac{q_c\alpha_k v_{kk}}{\left[\tilde{\boldsymbol{V}}_{\boldsymbol{b}}\boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q}+\boldsymbol{\sigma}\right]_k} \\ &\geq \frac{1}{|\mathcal{K}_c|}\frac{q_c\sum_{k\in\mathcal{K}_c}\alpha_k v_{kk}}{\sum_{k\in\mathcal{K}_c}\left[\tilde{\boldsymbol{V}}_{\boldsymbol{b}}\boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q}+\boldsymbol{\sigma}\right]_k} = U_c^{\mathrm{UL},1}(\boldsymbol{q},\boldsymbol{b}) \qquad (8.4)\end{aligned}$$

The uplink capacity utility of cluster c denoted by $U_c^{\mathrm{UL},1}$ is measured by the ratio between the total useful power and the total interference power received in the uplink in the cluster. Utility $U_c^{\mathrm{UL},1}$ is used instead of $\bar{U}_c^{\mathrm{UL},1}$ because of two reasons: First, it is a lower bound for the average SINR. Second, it has certain monotonicity properties (introduced in Definition D.8 in Appendix D.3.2) which are useful for optimization.

Introducing the cluster coupling term $\overline{\boldsymbol{G}}_{\boldsymbol{b}}^{\mathrm{UL}} := \boldsymbol{\Psi}\boldsymbol{A}\tilde{\boldsymbol{V}}_{\boldsymbol{b}}\boldsymbol{A}_{\boldsymbol{\alpha}}^T$, where $\boldsymbol{\Psi} := \mathrm{diag}\{|\mathcal{K}_1|/g_1,\ldots,|\mathcal{K}_c|/g_C\}$ and $g_c := \sum_{k\in\mathcal{K}_c}\alpha_k v_{kk}$ for $c\in\mathcal{C}$; and the noise term $\overline{\boldsymbol{z}} := \boldsymbol{\Psi}\boldsymbol{A}\boldsymbol{\sigma}$, the capacity utility is simplified as

$$U_c^{\mathrm{UL},1}(\boldsymbol{q},\boldsymbol{b}) := \frac{q_c}{\mathcal{J}_c^{(\mathrm{UL},1)}(\boldsymbol{q},\boldsymbol{b})} \qquad (8.5)$$

$$\text{where } \mathcal{J}_c^{(\mathrm{UL},1)}(\boldsymbol{q},\boldsymbol{b}) := \left[\overline{\boldsymbol{G}}_{\boldsymbol{b}}^{\mathrm{UL}}\boldsymbol{q}+\overline{\boldsymbol{z}}\right]_c. \qquad (8.6)$$

Worst-Case SINR Utility (Coverage)

Roughly speaking, the coverage problem arises when a certain number of the SINRs are lower than the predefined SINR threshold. Thus, to improve the coverage performance is equivalent to maximize the worst-case SINR such that the worst-case SINR achieves the desired SINR target. We then define the uplink coverage utility for each cluster as

$$\begin{aligned}U_c^{\mathrm{UL},2}(\boldsymbol{q},\boldsymbol{b}) &:= \min_{k\in\mathcal{K}_c}\mathrm{SINR}_k^{\mathrm{UL}} = \min_{k\in\mathcal{K}_c}\frac{q_c\alpha_k v_{kk}}{\left[\tilde{\boldsymbol{V}}_{\boldsymbol{b}}\boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q}+\boldsymbol{\sigma}\right]_k} \\ &= \frac{q_c}{\max_{k\in\mathcal{K}_c}\left[\boldsymbol{\Phi}\tilde{\boldsymbol{V}}_{\boldsymbol{b}}\boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q}+\boldsymbol{\Phi}\boldsymbol{\sigma}\right]_k} \qquad (8.7)\end{aligned}$$

where $\boldsymbol{\Phi} := \mathrm{diag}\{1/\alpha_1 v_{11},\ldots,1/\alpha_K v_{KK}\}$. We define a $C\times K$ matrix $\boldsymbol{X} := [\boldsymbol{x}_1|\ldots|\boldsymbol{x}_C]^T$, where $\boldsymbol{x}_c := \boldsymbol{e}_K^j$ and $\boldsymbol{e}_i^j$ denotes an i-dimensional binary vector which has exact one entry (the j-th entry) equal to 1. Introducing the term $\underline{\boldsymbol{G}}_{\boldsymbol{b}}^{\mathrm{UL}} := \boldsymbol{\Phi}\tilde{\boldsymbol{V}}_{\boldsymbol{b}}\boldsymbol{A}_{\boldsymbol{\alpha}}^T$, and the noise term $\underline{\boldsymbol{z}} := \boldsymbol{\Phi}\boldsymbol{\sigma}$, the coverage utility is given by

$$U_c^{\mathrm{UL},2}(\boldsymbol{q},\boldsymbol{b}) := \frac{q_c}{\mathcal{J}_c^{(\mathrm{UL},2)}(\boldsymbol{q},\boldsymbol{b})} \qquad (8.8)$$

$$\text{where } \mathcal{J}_c^{(\mathrm{UL},2)}(\boldsymbol{q},\boldsymbol{b}) := \max_{\boldsymbol{x}_c:=\boldsymbol{e}_K^j,\, j\in\mathcal{K}_c}\left[\boldsymbol{X}\underline{\boldsymbol{G}}_{\boldsymbol{b}}^{\mathrm{UL}}\boldsymbol{q}+\boldsymbol{X}\underline{\boldsymbol{z}}\right]_c. \qquad (8.9)$$

Cluster-Based Max-Min Utility Balancing

Let $\boldsymbol{\gamma} := [\gamma_1, \ldots, \gamma_C]^T$ denote the cluster utility targets. To achieve optimal load balancing, we propose a power-constrained max-min utility balancing problem in the uplink in below.

Problem 8.1 (Cluster-Based Utility Balancing)**.**

$$C^{UL}(P^{max}) = \max_{\boldsymbol{q}\geq 0, \boldsymbol{b}\in\mathcal{N}^C} \min_{c\in\mathcal{C}} \frac{U_c^{UL}(\boldsymbol{q}, \boldsymbol{b})}{\gamma_c}, s.t.\ \|\boldsymbol{q}\| \leq P^{max} \tag{8.10}$$

where $C^{UL}(P^{max})$ denotes the achievable balanced margin given fixed sum power contraint P^{max}. $\|\cdot\|$ is an arbitrary monotone norm, i.e., $\boldsymbol{q} \leq \boldsymbol{q}'$ implies $\|\boldsymbol{q}\| \leq \|\boldsymbol{q}'\|$, P^{max} denotes the power constraint, and the joint utility $U_c^{UL}(\boldsymbol{q}, \boldsymbol{b})$ is defined as

$$U_c^{UL}(\boldsymbol{q}, \boldsymbol{b}) := \frac{q_c}{\mathcal{J}_c^{UL}(\boldsymbol{q}, \boldsymbol{b})} \tag{8.11}$$

$$\text{where } \mathcal{J}_c^{UL}(\boldsymbol{q}, \boldsymbol{b}) := \mu\mathcal{J}_c^{(UL,1)}(\boldsymbol{q}, \boldsymbol{b}) + (1-\mu)\mathcal{J}_c^{(UL,2)}(\boldsymbol{q}, \boldsymbol{b}). \tag{8.12}$$

In other words, the joint interference $\mathcal{I}_c^{UL}$ is a convex combination of $\mathcal{I}_c^{UL,1}$ in (8.6) *and $\mathcal{I}_c^{UL,2}$ in* (8.9)*. The algorithm optimizes the performance of capacity when we set the tuning parameter $\mu = 1$ (utility is equivalent to the capacity utility in* (8.5)*), while with $\mu = 0$ it optimizes the performance of coverage (utility equals to the coverage utility in* (8.8)*). By tuning μ properly, we can achieve a good trade-off between the performance of coverage and capacity.*

8.3.2 BS-Based Antenna Tilt Optimization and Power Allocation

The user transmission power $\boldsymbol{p}$ and the BS assignment $\boldsymbol{n}$ can be directly deduced from $(\boldsymbol{q}, \boldsymbol{b})$ optimized on a per-cluster basis. However, the antenna tilt and BS power budget need to be optimized per base station. Given the fixed $(\hat{\boldsymbol{b}}, \hat{\boldsymbol{q}})$, we compute the intra-cluster power sharing factor $\boldsymbol{\beta}$, given by $\beta_c := \hat{q}_c / \sum_{c\in\mathcal{C}_{b_c}} \hat{q}_c$ for $c \in \mathcal{C}$. We denote the interference coupling matrix depending on $\boldsymbol{\theta}$ by $\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}$. In the following we formulate the BS-based max-min utility balancing problem such that it has the same physical meaning as the problem stated in (8.10). We then introduce the BS-based capacity and the coverage utilities interpreted by $(\boldsymbol{r}, \boldsymbol{\theta})$.

BS-Based Max-Min Utility Balancing

To be consistent with our objective function $C^{\mathrm{UL}}(P^{\max})$ in (8.10), we transform the cluster-based optimization problem to the BS-based optimization problem:

Problem 8.2 (BS-Based Utility Balancing)**.**

$$\begin{aligned} C^{(u)}(P^{max}) &= \max_{\boldsymbol{r}\geq 0, \boldsymbol{\theta}\in\Theta^N} \min_{c\in\mathcal{C}} \frac{U_c^{UL}(\boldsymbol{r},\boldsymbol{\theta})}{\gamma_c} \\ &= \max_{\boldsymbol{r}\geq 0, \boldsymbol{\theta}\in\Theta^N} \min_{n\in\mathcal{N}} \left(\min_{c\in\mathcal{C}_n} \frac{U_c^{UL}(\boldsymbol{r},\boldsymbol{\theta})}{\gamma_c} \right) \\ &= \max_{\boldsymbol{r}\geq 0, \boldsymbol{\theta}\in\Theta^N} \min_{n\in\mathcal{N}} \widehat{U}_n^{UL}(\boldsymbol{r},\boldsymbol{\theta}) \\ s.t.\ & \|\boldsymbol{r}\| \leq P^{max} \end{aligned} \tag{8.13}$$

where Θ denotes the predefined space for antenna tilt configuration. It is shown in (8.13) that by defining

$$\widehat{U}_n^{\mathrm{UL}}(\boldsymbol{r},\boldsymbol{\theta}) := \min_{c\in\mathcal{C}_n} \frac{U_c^{\mathrm{UL}}(\boldsymbol{r},\boldsymbol{\theta})}{\gamma_c} = \frac{r_n}{\widehat{\mathcal{J}}_n^{\mathrm{UL}}(\boldsymbol{r},\boldsymbol{\theta})} \tag{8.14}$$

$$\widehat{\mathcal{J}}_n^{\mathrm{UL}}(\boldsymbol{r},\boldsymbol{\theta}) := \max_{c\in\mathcal{C}_n} \frac{\gamma_c}{\beta_c} \mathcal{J}_c^{\mathrm{UL}}(\boldsymbol{r},\boldsymbol{\theta}), \tag{8.15}$$

the cluster-based problem is transferred to the BS-based problem, where $\mathcal{J}_c^{\mathrm{UL}}(\boldsymbol{r},\boldsymbol{\theta})$ is obtained from $\mathcal{J}_c^{\mathrm{UL}}(\boldsymbol{q},\boldsymbol{b})$ in (8.12) by substituting $\boldsymbol{q}$ with $\boldsymbol{q} := \boldsymbol{B}_{\boldsymbol{\beta}}^T\boldsymbol{r}$, and $\tilde{\boldsymbol{V}}_{\boldsymbol{b}}$ with $\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}$.

The utility functions corresponding to (8.4) and (8.7) are provided below.

Average SINR Utility (Capacity)

According to (8.14), the capacity utility of BS n is defined as the minimum of the ratios of cluster-based capacity utilities to the utility targets of the clusters assigned to BS n. With (8.4), (8.5) and (8.6), and the power transformation $\boldsymbol{p} := \boldsymbol{T}\boldsymbol{r}$, we have

$$\begin{aligned} \widehat{U}_n^{\mathrm{UL},1}(\boldsymbol{r},\boldsymbol{\theta}) &:= \min_{c\in\mathcal{C}_{b_c}} \frac{U_c^{\mathrm{UL},1}(\boldsymbol{r},\boldsymbol{\theta})}{\gamma_c} \\ &= \frac{r_n}{\max_{c\in\mathcal{C}_{b_c}} \frac{\gamma_c}{\beta_c} \left[\boldsymbol{\Psi}\boldsymbol{A}\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}\boldsymbol{T}\boldsymbol{r} + \overline{\boldsymbol{z}} \right]_c} \end{aligned} \tag{8.16}$$

Define a $N \times C$ matrix $\boldsymbol{S} := [\boldsymbol{s}_1|\dots|\boldsymbol{s}_N]^T$, where $\boldsymbol{s}_n := \boldsymbol{e}_C^j$. Introducing the term $\overline{\boldsymbol{\Lambda}}_{\boldsymbol{\theta}}^{\mathrm{UL}} := \boldsymbol{D}\boldsymbol{\Psi}\boldsymbol{A}\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}\boldsymbol{T}$ and the noise term $\overline{\boldsymbol{\eta}} := \boldsymbol{D}\overline{\boldsymbol{z}}$, where $\boldsymbol{D} := \mathrm{diag}\{\gamma_1/\beta_1, \dots, \gamma_C/\beta_C\}$, utility in (8.16) can be simplified as

$$\widehat{U}_n^{\mathrm{UL},1}(\boldsymbol{r},\boldsymbol{\theta}) := \frac{r_n}{\max_{\boldsymbol{s}_n := \boldsymbol{e}_C^j, j\in\mathcal{C}_n} \left[\boldsymbol{S}\overline{\boldsymbol{\Lambda}}_{\boldsymbol{\theta}}^{\mathrm{UL}}\boldsymbol{r} + \boldsymbol{S}\overline{\boldsymbol{\eta}} \right]_n} \tag{8.17}$$

Worst-Case SINR Utility (Coverage)

The coverage utility of BS n is defined by

$$\begin{aligned}\widehat{U}_n^{\mathrm{UL},2}(\boldsymbol{r},\boldsymbol{\theta}) &:= \min_{c\in\mathcal{C}_n} \frac{U_c^{\mathrm{UL},2}(\boldsymbol{r},\boldsymbol{\theta})}{\gamma_c} \\ &= \frac{r_n}{\max_{c\in\mathcal{C}_n}\left\{\frac{\gamma_c}{\beta_c}\max_{k\in\mathcal{K}_c}\left[\boldsymbol{\Phi}\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}^{\mathrm{UL}}\boldsymbol{T}\boldsymbol{r}+\underline{\boldsymbol{z}}\right]_k\right\}} \\ &= \frac{r_n}{\max_{k\in\mathcal{K}^n}\left[\widehat{\boldsymbol{D}}\boldsymbol{\Phi}\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}^{\mathrm{UL}}\boldsymbol{T}\boldsymbol{r}+\widehat{\boldsymbol{D}}\underline{\boldsymbol{z}}\right]_k}\end{aligned} \tag{8.18}$$

where $\widehat{\boldsymbol{D}} := \mathrm{diag}\{\boldsymbol{A}^T\boldsymbol{\Gamma}\boldsymbol{\beta}\}$, and $\boldsymbol{\Gamma} := \mathrm{diag}\{\boldsymbol{\gamma}\}$. Define a $N \times K$ matrix $\widehat{\boldsymbol{X}} := [\widehat{\boldsymbol{x}}_1|\ldots|\widehat{\boldsymbol{x}}_N]^T$, where $\widehat{\boldsymbol{x}}_n := \boldsymbol{e}_K^j$. Introducing the coupling term $\underline{\boldsymbol{\Lambda}}_{\boldsymbol{\theta}}^{\mathrm{UL}} := \widehat{\boldsymbol{D}}\boldsymbol{\Phi}\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}^{\mathrm{UL}}\boldsymbol{T}$ and the noise term $\underline{\boldsymbol{\eta}} := \widehat{\boldsymbol{D}}\underline{\boldsymbol{z}}$, we can write the coverage utility in (8.18) as

$$\widehat{U}_n^{\mathrm{UL},2}(\boldsymbol{r},\boldsymbol{\theta}) := \frac{r_n}{\max_{\widehat{\boldsymbol{x}}_n:=\boldsymbol{e}_K^j, j\in\mathcal{K}^n}\left[\widehat{\boldsymbol{X}}\underline{\boldsymbol{\Lambda}}_{\boldsymbol{\theta}}^{\mathrm{UL}}\boldsymbol{r}+\widehat{\boldsymbol{X}}\underline{\boldsymbol{\eta}}\right]_k} \tag{8.19}$$

8.4 Optimization Algorithm

We developed our optimization algorithm based on the fixed-point iteration algorithm proposed by Yates [YH95], by exploiting the properties of the *standard interference function* (see Definition D.8 in Appendix D.3.2).

Theorem 8.1. *[Yat95] If $\boldsymbol{\mathcal{I}}(\boldsymbol{p})$ is a standard interference function, and the utility target $\boldsymbol{\gamma} := [\gamma_1,\ldots,\gamma_K]^T$ is feasible, under a sum-power constraint, then for an arbitrary initialization $\boldsymbol{p}^{(0)} \geq 0$, the iteration*

$$p_k^{(t+1)} = \gamma_k \cdot \mathcal{I}_k(\boldsymbol{p}^{(t)}), \forall k \tag{8.20}$$

converges to the optimum of the power minimization problem

$$\inf_{\boldsymbol{p}>0} \|\boldsymbol{p}\|, \textit{s.t.}\ \frac{p_k}{\mathcal{I}_k(\boldsymbol{p})} \geq \gamma_k, \forall k. \tag{8.21}$$

Define the utility $U_k(\boldsymbol{p}) := p_k/\mathcal{I}_k(\boldsymbol{p})$, the solution of (8.21) indirectly solves the following max-min fairness problem

$$\max_{\boldsymbol{p}>0}\min_{1\leq k\leq K}\frac{U_k(\boldsymbol{p})}{\gamma_k}, \text{s.t. } \|\boldsymbol{p}\| \leq P^{\max} \tag{8.22}$$

by scaling the utility target γ_k iteratively (for example, the one-dimensional bisection search method) until the max-min utility boundary is achieved.

8.4.1 Joint Optimization Algorithm

We aim on jointly optimizing both problems, by optimizing $(\boldsymbol{q}, \boldsymbol{b})$ in Problem 8.1 and $(\boldsymbol{r}, \boldsymbol{\theta})$ in Problem 8.2 iteratively with the fixed-point iteration. In the following we present some properties that are required to solve the problem efficiently and to guarantee the convergence of the algorithm.

Decoupled Variables in Uplink

In uplink the variables $\boldsymbol{b}$ and $\boldsymbol{\theta}$ are decoupled in the interference functions (8.12) and (8.15), i.e., $\mathcal{J}_c^{\mathrm{UL}}(\boldsymbol{q}, \boldsymbol{b}) := \mathcal{J}_c^{\mathrm{UL}}(\boldsymbol{q}, b_c)$ and $\widehat{\mathcal{J}}_n^{\mathrm{UL}}(\boldsymbol{r}, \boldsymbol{\theta}) := \widehat{\mathcal{J}}_n^{\mathrm{UL}}(\boldsymbol{r}, \theta_n)$. Thus, we can decompose the BS assignment (or tilt optimization) problem into sub-problems that can be independently solved in each cluster (or BS), and the interference functions can be modified as functions of the power allocation only:

$$\mathcal{J}_c^{\mathrm{UL}}(\boldsymbol{q}) := \min_{b_c \in \mathcal{N}} \mathcal{J}_c^{\mathrm{UL}}(\boldsymbol{q}, b_c) \tag{8.23}$$

$$\widehat{\mathcal{J}}_n^{\mathrm{UL}}(\boldsymbol{r}) := \min_{\theta_n \in \Theta} \widehat{\mathcal{J}}_n^{\mathrm{UL}}(\boldsymbol{r}, \theta_n) \tag{8.24}$$

Standard Interference Function

The modified interference function (8.23) and (8.24) are *standard*. Using the following three properties: 1) an affine function $\boldsymbol{\mathcal{I}}(\boldsymbol{p}) := \boldsymbol{V}\boldsymbol{p} + \boldsymbol{\sigma}$ is standard, 2) if $\boldsymbol{\mathcal{I}}(\boldsymbol{p})$ and $\boldsymbol{\mathcal{I}}'(\boldsymbol{p})$ are standard, then $\beta\boldsymbol{\mathcal{I}}(\boldsymbol{p}) + (1-\beta)\boldsymbol{\mathcal{I}}'(\boldsymbol{p})$ are standard, and 3) If $\boldsymbol{\mathcal{I}}(\boldsymbol{p})$ and $\boldsymbol{\mathcal{I}}'(\boldsymbol{p})$ are standard, then $\boldsymbol{\mathcal{I}}^{\min}(\boldsymbol{p})$ and $\boldsymbol{\mathcal{I}}^{\max}(\boldsymbol{p})$ are standard, where $\boldsymbol{\mathcal{I}}^{\min}(\boldsymbol{p})$ and $\boldsymbol{\mathcal{I}}^{\max}(\boldsymbol{p})$ are defined as $\mathcal{I}_j^{\min}(\boldsymbol{p}) := \min\{\mathcal{I}_j(\boldsymbol{p}), \mathcal{I}'_j(\boldsymbol{p})\}$ and $\mathcal{I}_j^{\max}(\boldsymbol{p}) := \max\{\mathcal{I}_j(\boldsymbol{p}), \mathcal{I}'_j(\boldsymbol{p})\}$ respectively [Yat95], we can easily prove that (8.23) and (8.24) are standard interference functions.

Substituting (8.23) and (8.24) in Problem 8.1 and Problem 8.2, define $U_c^{\mathrm{UL}}(\boldsymbol{q}) := q_c/\mathcal{I}_c^{\mathrm{UL}}(\boldsymbol{q})$ and $U_n^{\mathrm{UL}}(\boldsymbol{r}) := r_n/\mathcal{I}_n^{\mathrm{UL}}(\boldsymbol{r})$, we can write both problems in the general framework of the max-min fairness problem (8.22):

Problem 1. $\max_{\boldsymbol{q}\geq 0} \min_{c\in\mathcal{C}} U_c^{\mathrm{UL}}(\boldsymbol{q})/\gamma_c, \|\boldsymbol{q}\| \leq P^{\max}$.

Problem 2. $\max_{\boldsymbol{r}\geq 0} \min_{n\in\mathcal{N}} U_n^{\mathrm{UL}}(\boldsymbol{r}), \|\boldsymbol{r}\| \leq P^{\max}$

The above two properties enables us to solve each problem efficiently with two iterative steps: 1) find optimum variable b_c (or θ_n) for each cluster c (or each BS n) independently, 2) solve the max-min balancing power allocation problem with fixed-point iteration.

Connections between Two Problems

Problem 8.1 and Problem 8.2 have the same objective achievable balanced margin $C^{\mathrm{UL}}(P^{\max})$ as stated in (8.10) and (8.13), i.e., given the same variables $(\hat{\boldsymbol{q}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{\theta}})$, using (8.14), we have $\min_{c\in\mathcal{C}} U_c^{\mathrm{UL}}/\gamma_c = \min_{n\in\mathcal{N}} \widehat{U}_n^{\mathrm{UL}}$. Both problems are under the same sum power constraint. However, the convergence of the two-step iteration requires two more properties: 1) the BS power budget $\boldsymbol{r}$ derived by solving Problem 8.2 at the previous step should not be violated by the cluster power allocation $\boldsymbol{q}$ found by optimizing Problem 8.1, and 2) when optimizing Problem 8.2, the inter-cluster power sharing factor $\boldsymbol{\beta}$ should be consistent with the derived cluster power allocation $\boldsymbol{q}$ in Problem 8.1.

To fulfill the first requirement, we introduce an individual cluster power constraint $P_c^{\max}$ depending on the BS power budget r_{b_c} in Problem 8.1. Moreover, we propose a scaled version of fixed point iteration similar to the one proposed in [VS11], to iteratively scale the cluster power vector and achieve the power-constrained max-min utility boundary, as stated below.

$$q_c^{(t+1)} = \lambda^{(t)} \min\{P_c^{\max(t)}, \gamma_c \mathcal{I}_c^{\mathrm{UL}}(\boldsymbol{q}^{(t)})\} \tag{8.25}$$

where the scaling factor is given by $\lambda^{(t)} = \max_{c\in\mathcal{C}} \mathcal{I}_c^{\mathrm{UL}}(\boldsymbol{q}^{(t)})/P_c^{\max(t)}$. To fulfill the second requirement, once $\boldsymbol{q}^{(n+1)}$ is derived, the power sharing factors $\boldsymbol{\beta}$ need to be updated for solving Problem 8.2 at the next step, provided as

$$\boldsymbol{\beta}^{(n+1)} := \boldsymbol{Q}^{-1}\boldsymbol{B}^T\boldsymbol{r}^{(n)}, \text{where } \boldsymbol{Q} = \mathrm{diag}\{\boldsymbol{q}^{(n+1)}\} \tag{8.26}$$

The individual power constraint $P_c^{\max}$ is updated at the previous step of optimizing Problem 8.2. The scaled fixed-point iteration to optimize Problem 8.2 is provided by

$$r_n^{(t+1)} = \frac{\widehat{\mathcal{I}}_n^{\mathrm{UL}}(\boldsymbol{r}^{(t)})}{\|\widehat{\boldsymbol{\mathcal{I}}}^{\mathrm{UL}}(\boldsymbol{r}^{(t)})\|}. \tag{8.27}$$

Alternatively, if per-BS power constraint $\widehat{P}_n^{\max}$ for each BS $n \in \mathcal{N}$ is required by the system instead of the sum power constraint $P^{\max}$, we can apply

$$r_n^{(t+1)} = \widehat{\lambda}^{(t)} \min\{\widehat{P}_n^{\max}, \widehat{\mathcal{I}}_n^{\mathrm{UL}}(\boldsymbol{r}^{(t)})\} \tag{8.28}$$

where the scaling factor follows $\widehat{\lambda}^{(t)} = \max_{n\in\mathcal{N}} \widehat{\mathcal{I}}_n^{\mathrm{UL}}(\boldsymbol{r}^{(t)})/\widehat{P}_n^{\max}$, and $\boldsymbol{P}^{\max} = [P_1^{\max}, \ldots, P_C^{\max}]^T$ should be calculated with

$$\boldsymbol{P}^{\max(n+1)} = \mathrm{diag}\{\boldsymbol{\beta}^{(n)}\}\boldsymbol{B}^T\boldsymbol{r}^{(n+1)}. \tag{8.29}$$

The joint optimization algorithm is given in Algorithm 5.

Algorithm 5: Joint Optimization of Problem 8.1 and 8.2

```
broadcast the information required for computing $\boldsymbol{V}$, predefined constraint $P^{\max}$ and
  thresholds $\epsilon_1, \epsilon_2, \epsilon_3$
arbitrary initial power vector $\boldsymbol{q}^{(t)} > 0$ and iteration step $t := 0$
**repeat** {joint optimization of Problem 8.1 and 8.2}
  **repeat** {fixed-point iteration for every cluster $c \in \mathcal{C}$}
    broadcast $\boldsymbol{q}^{(t)}$ to all base stations
    **for** all assignment options $b_c \in \mathcal{N}$ **do**
      compute $\mathcal{I}_c^{\mathrm{UL}}(\boldsymbol{q}^{(t)}, b_c)$ with (8.12)
    **end for**
    compute $\mathcal{I}_c^{\mathrm{UL}}(\boldsymbol{q}^{(t)})$ with (8.23) and update $b_c^{(t+1)}$
    update $q_c^{(t+1)}$ with (8.25)
    $t := t + 1$
  **until** convergence: $\left|q_c^{(t+1)} - q_c^{(t)}\right| / q_c^{(t)} \leq \epsilon_1$
  update $\boldsymbol{\beta}^{(t)}$ with (8.26)
  **repeat** {fixed-point iteration for every BS $n \in \mathcal{N}$}
    broadcast $\boldsymbol{r}^{(t)}$ to all base stations
    **for** all antenna tilt options $\theta_n \in \Theta$ **do**
      compute $\widehat{\mathcal{I}}_n^{\mathrm{UL}}(\boldsymbol{r}^{(t)}, \theta_n)$ with (8.15)
    **end for**
    compute $\widehat{\mathcal{I}}_n^{\mathrm{UL}}(\boldsymbol{r}^{(t)})$ with (8.24) and update $\theta_n^{(t+1)}$
    update $r_c^{(n+1)}$ with (8.27) or (8.28)
    $t := t + 1$
  **until** convergence: $\left|r_n^{(t+1)} - r_n^{(t)}\right| / r_n^{(t)} \leq \epsilon_2$
  update $\boldsymbol{P}^{\max(t)}$ with (8.29)
  compute $l^{(t+1)} := \min_{n \in \mathcal{N}} \widehat{U}_n^{\mathrm{UL}}(\boldsymbol{r}^{(n+1)})$
**until** convergence: $|l^{(t+1)} - l^{(t)}| / l^{(t)} \leq \epsilon_3$
```

8.5 Uplink-Downlink Duality

We state the joint optimization problem in uplink in Section 8.3 and propose an efficient solution in Section 8.4 by exploiting the decoupled property of $\boldsymbol{V}$ over the variables $\boldsymbol{\theta}$ and $\boldsymbol{b}$. The downlink problem, due to the coupled structure of $\boldsymbol{V}^T$, is more difficult to solve. As extended discussion we want to address the relationship between the uplink and the downlink problem, and to propose a sub-optimal solution for downlink which can be possibly found through the uplink solution.

Let us consider cluster-based max-min capacity utility balancing problem in Section

8.3.1 as an example. In the downlink the optimization problem is written as

$$\begin{aligned}
&\max_{\boldsymbol{q},\boldsymbol{b}} \min_{c} \frac{U_c^{(\mathrm{d},1)}(\boldsymbol{q},\boldsymbol{b})}{\gamma_c} \\
&U_c^{(\mathrm{d},1)} := \frac{q_c}{[\boldsymbol{\Psi}\boldsymbol{A}\tilde{\boldsymbol{V}}_{\boldsymbol{b}}^T\boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q} + \boldsymbol{\Psi}\boldsymbol{z}^{\mathrm{DL}}]} \\
&\text{s.t. } \|\boldsymbol{q}\|_1 \leq P^{\max}
\end{aligned} \tag{8.30}$$

The cluster-based received noise is written as $\boldsymbol{z}^{\mathrm{DL}} := \boldsymbol{A}\boldsymbol{\sigma}^{\mathrm{DL}}$.

In the following we present a virtual dual uplink network in terms of the feasible utility region for the downlink network in (8.30) via Perron-Frobenius theory, such that the solution of problem (8.30) can be derived by solving the uplink problem (8.31) with the algorithm introduced in Section 8.4.

Proposition 8.1. *Define a virtual uplink network where the link gain matrix is modified as* $\boldsymbol{W}_{\boldsymbol{b}} := \mathrm{diag}\{\boldsymbol{\alpha}\}\tilde{\boldsymbol{V}}_{\boldsymbol{b}}\,\mathrm{diag}^{-1}\{\boldsymbol{\alpha}\}$, *i.e.,* $w_{lk} := v_{lk}\frac{\alpha_l}{\alpha_k}$, *and the received uplink noise is denoted by* $\boldsymbol{\sigma}^{UL} := [\sigma_1^{2\,UL}, \ldots, \sigma_K^{2\,UL}]^T$, *where* $\sigma_k^{2\,UL} := \frac{\Sigma_{tot}}{|\mathcal{K}_{c_k}|\cdot C}$ *for* $k \in \mathcal{K}$, *and assume* $\Sigma_{tot} := \|\boldsymbol{\sigma}^{UL}\|_1 = \|\boldsymbol{\sigma}^{DL}\|_1$ *(which means, the sum noise is equally distributed in clusters, while in each cluster the noise is equally distributed in the subordinate users). The dual uplink problem of problem* (8.30) *is given by*

$$\begin{aligned}
&\max_{\boldsymbol{q},\boldsymbol{b}} \min_{c} \frac{U_c^{(u,1)}(\boldsymbol{q},\boldsymbol{b})}{\gamma_c} \\
&U_c^{(u,1)} := \frac{q_c}{[\boldsymbol{\Psi}\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{b}}\boldsymbol{A}_{\boldsymbol{\alpha}}^T\boldsymbol{q} + \boldsymbol{\Psi}\boldsymbol{z}^{UL}]} \\
&s.t.\ \|\boldsymbol{q}\|_1 \leq P^{max}
\end{aligned} \tag{8.31}$$

where $\boldsymbol{z}^{UL} := \boldsymbol{A}\boldsymbol{\sigma}^{UL}$.

The proof of Proposition 8.1 is given in Appendix A.3.1.

Note that the optimizer $\boldsymbol{b}^*$ for BS assignment in downlink can be equivalently found by minimizing the spectral radius $\boldsymbol{\Lambda}^{(\boldsymbol{u})}(\boldsymbol{b})$ in the uplink. Once $\boldsymbol{b}^*$ is found, the associate optimizer for uplink power $\boldsymbol{q}^{\mathrm{UL}^*}$ is given as the dominant right-hand eigenvector of matrix $\boldsymbol{\Lambda}^{\mathrm{UL}}(\boldsymbol{b}^*)$, while the associate optimizer for downlink power $\boldsymbol{q}^{\mathrm{DL}^*}$ is given as the dominant right-hand eigenvector of matrix $\boldsymbol{\Lambda}^{\mathrm{DL}}(\boldsymbol{b}^*)$. Proposition 8.1 provides an efficient approach to solve the downlink problem with two iterative steps (as the one proposed in [BS06]): 1) for a fixed power allocation $\hat{\boldsymbol{q}}$, solve the uplink problem and derive the assignment $\boldsymbol{b}^*$ that associated with the spectral radius of extend coupling matrix $\boldsymbol{\Lambda}^{\mathrm{UL}}$, and 2) for a fixed assignment $\hat{\boldsymbol{b}}$, update the power $\boldsymbol{q}^*$ as the solution of (A.10).

Note that although we are able to find a dual uplink problem for the downlink problem in (8.30) with our proposed utility functions *under sum power constraints*, we are not able to construct a dual network with decoupled properties for the modified problem *under individual power constraints* (8.25). However, numerical experiments show that our approach to the downlink based on the proposed uplink solution does improve the network performance, although the duality does not hold between the downlink problem and our proposed uplink problem under the individual power constraints.

8.6 Numerical Results

We consider a hexagonal network composed of 7 tri-sectored BSs with site-to-site distance of 1 km. The pathloss is modeled with Okumura Hata model for urban areas. The SINR threshold is defined as -6.5 dB. The power constraint per BS is 46dBm.

Fig. 8.1 illustrates the convergence of the algorithm. Our algorithm achieves the max-min utility balancing, and improves the feasibility level $C^{(u)}(P^{\max})$ by each iteration step.

In Fig.8.2 we show that the trade-off between coverage and capacity can be adjusted by tuning parameter μ. By increasing μ we give higher priority to capacity utility (which is proportional to the ratio between total useful power and total interference power), while for better coverage utility (defined as minimum of SINRs) we can use a small value of μ instead.

Fig. 8.3, 8.4 and 8.5 illustrate the improvement of coverage and capacity performance and decreasing of the energy consumption in both uplink and downlink systems when the numbers of the users per BS are $\{15, 20, 25, 30, 35\}$, by applying the proposed algorithm. In Fig. 8.4 we further show that by optimizing the capacity utility, the actual average SINR indicating the performance of capacity can be improved as well. Fig. 8.5 shows that by applying the proposed algorithm, the BS power budgets can be adaptively adjusted. Thus, compared to the fixed BS power budget scenario, our algorithm is more energy efficient. Compared to the near-optimal uplink solutions, less improvements are observed for the downlink solutions as shown in Fig. 8.3, 8.4 and 8.5. This is because we derive the downlink solution by exploiting an uplink problem which is not exactly its dual due to the individual power constraints (as described in Section 8.5). However, the sub-optimal solutions still provide significant performance improvements.

8.7 Conclusions and Further Research

We present an efficient and robust algorithmic optimization framework build on the utility model for joint optimization of the SON use cases coverage and capacity optimization and load balancing. The max-min utility balancing formulation is employed to enforce the fairness across clusters. We propose a two-step optimization algorithm in the uplink based on fixed-point iteration to iteratively optimize the per-base station antenna tilt and power allocation as well as the per-cluster BS assignment and power allocation. We then analyze the network duality via Perron-Frobenius theory, and propose a sub-optimal solution in the downlink by exploiting the solution in the uplink. Simulation results show significant improvements in performance of coverage, capacity and load balancing in a power-efficient way, in both uplink and downlink. In our follow-up papers we will further propose a more complex interference coupling model and the optimization framework where frequency band assignment is taken into account. We will also examine the suboptimality under more general form of power constraints.

FIGURES

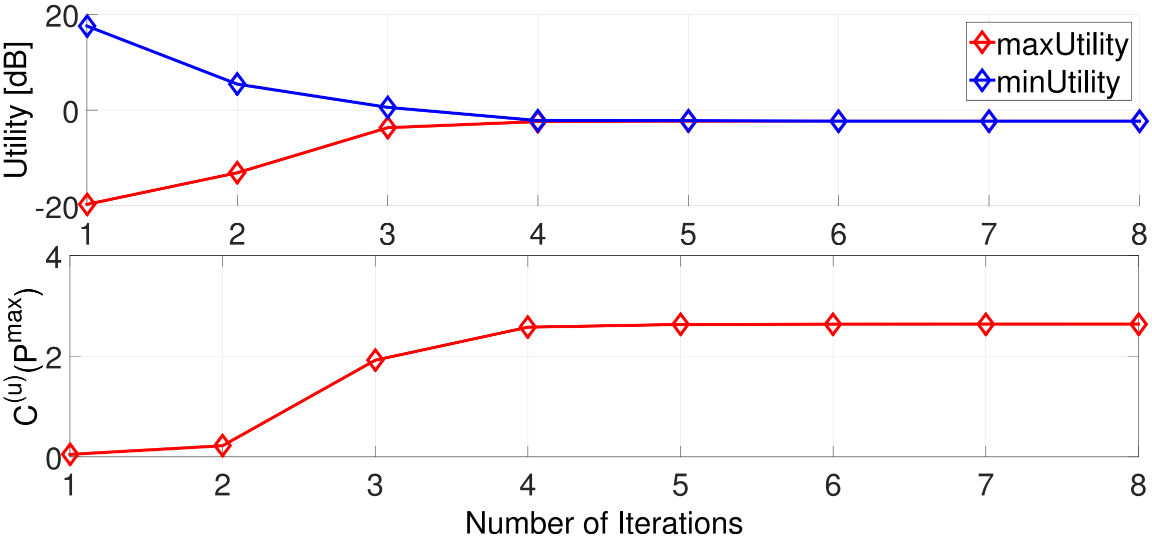


Figure 8.1: Algorithm convergence.

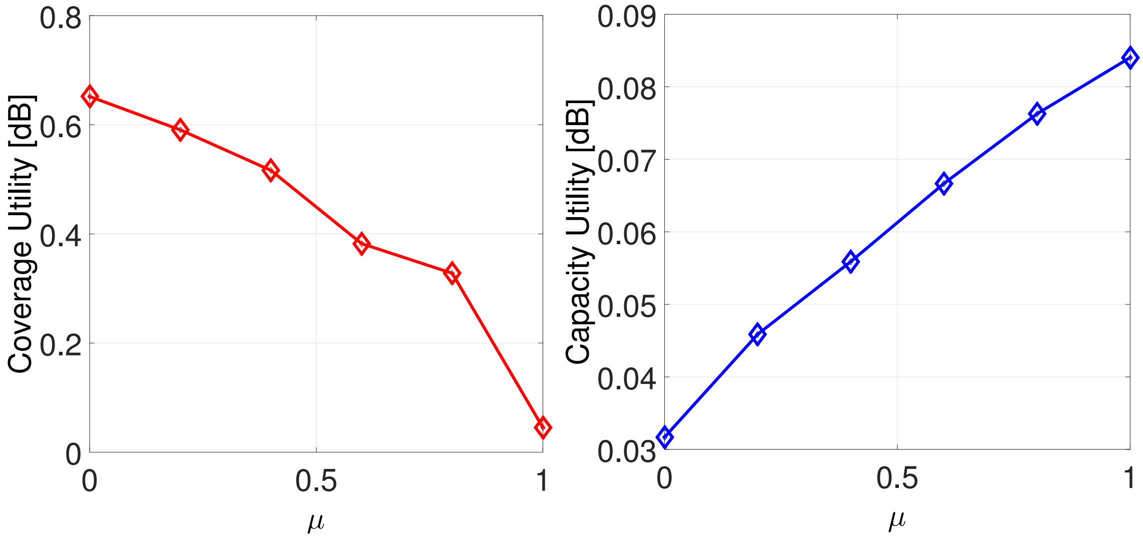


Figure 8.2: Trade-off between utilities depending on μ.

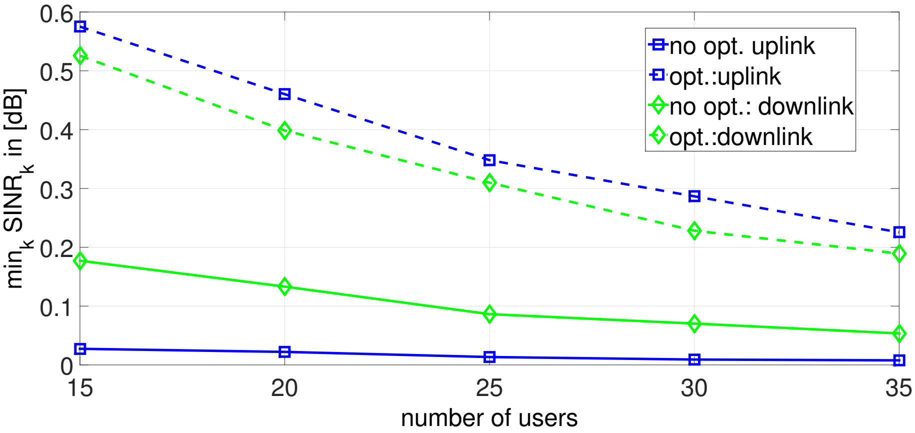


Figure 8.3: Performance of proposed algorithm: coverage.

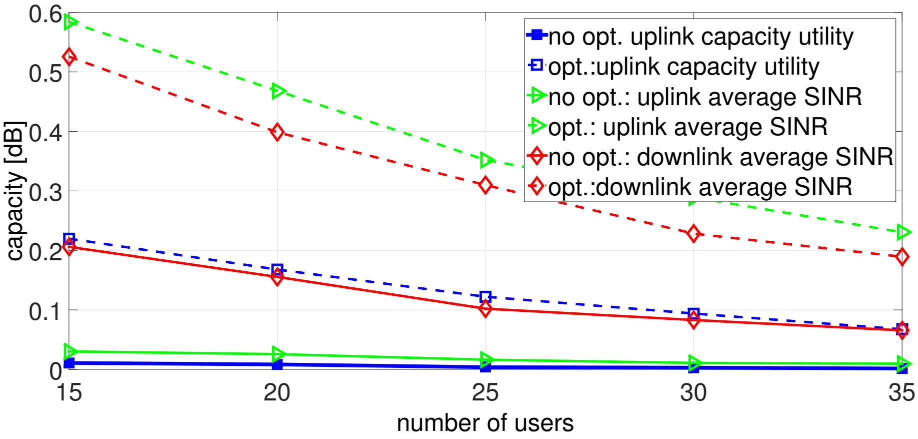


Figure 8.4: Performance of proposed algorithm: capacity.

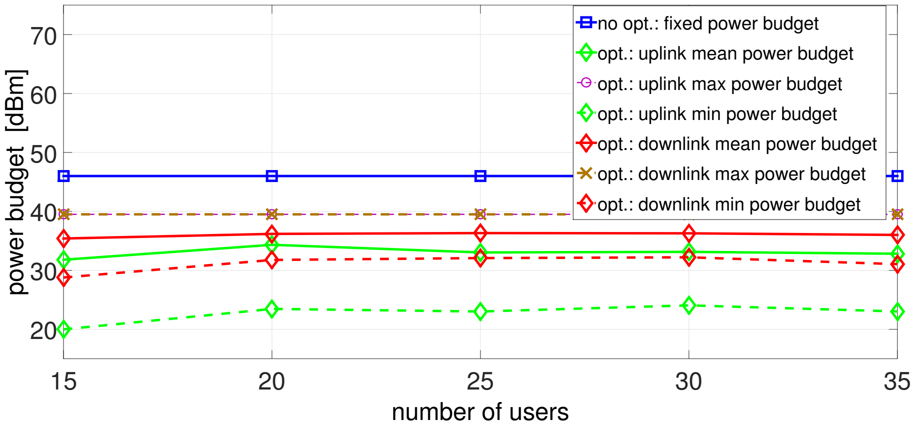


Figure 8.5: Performance of proposed algorithm: per-BS power budget.

Chapter 9

Service-Centric Joint Uplink and Downlink Optimization for Uplink and Downlink Decoupling-Enabled HetNets

The concept of user-centric and personalized service in the 5G mobile networks encourages technical solutions such as dynamic asymmetric uplink/downlink resource allocation and elastic association of cells to users with decoupled uplink and downlink (DeUD) access. In this chapter we develop a joint uplink and downlink optimization algorithm for DeUD-enabled wireless networks for adaptive joint uplink and downlink bandwidth allocation and power control, under different link association policies. Based on a general model of inter-cell interference, we propose a three-step optimization algorithm to jointly optimize the uplink and downlink bandwidth allocation and power control, using the fixed point approach for nonlinear operators with or without monotonicity, to maximize the minimum level of quality of service satisfaction per link, subjected to a general class of resource (power and bandwidth) constraints. We present numerical results illustrating the theoretical findings for network simulator in a real-world setting, and show the advantage of our solution compared to the conventional proportional fairness resource allocation schemes in both the coupled uplink and downlink (CoUD) access and the novel link association schemes in DeUD.

Parts of this chapter have already been published in [16].

9.1 Introduction

The high rate of growth in global mobile data traffic drives the operators to set foot on the path of delivering the 5G of mobile networks, for user-centric and personalized service supporting diverse and often conflicting KPIs, such as high-speed, low-latency, high reliability, high mobility, and low cost/energy consumption.

In the 5G era, the evolution of heterogeneous networkss (HetNets) results in cell densification with cells of different sizes. Due to the time- and spatial-dependent service requirements and traffic patterns, it is expected to have time-varying asymmetric traffic load in both UL and DL in different cells (as shown in Fig. 9.1). Many optimization strategies are designed to provide seamless coverage and QoS in DL, while little interest has been shown in UL. However, the importance of UL grows along with the evolution of social networking and information/resource sharing system. Therefore, it is of great interest to develop a general framework for joint UL/DL optimization of resource allocation and power control, to adapt to the traffic asymmetry between UL and DL.

Apart from dynamic UL/DL resource splitting, flexible UL/DL traffic distribution among the cells with different transmission ranges is also crucial for improvement of joint UL/DL performance. As proposed in [And13, BHL+14], one way to enable the flexible UL/DL traffic distribution is to allow the user terminal to be associated to two different radio access nodes in UL and DL, respectively. Such a DeUD access has the potential benefits including improvement of performance in UL (without degradation of performance in DL), reduction of energy consumption in mobile terminal, and network load balancing.

The joint UL/DL optimization framework can benefit from the user-centric context-aware communication environment in 5G networks. More specifically, this includes dynamic splitting resources and distributing network traffic between UL and DL, based on the awareness of the heterogeneity of UL and DL channel conditions and traffic demands.

The focus of this paper is to develop a general model of joint UL/DL interference, and to design a joint UL/DL optimization algorithm for adaptive UL/DL bandwidth allocation and power control under different association policies for DeUD-enabled wireless networks. The objective is to optimize the minimum level of QoS satisfaction across all service links, using the fixed point approach for nonlinear operators with or without monotonicity.

9.1.1 Related Work

Joint Uplink and Downlink Optimization

Although much work has been done on the joint UL/DL resource allocation in conventional network with coupled uplink and downlink (CoUD) association [SHWL07, SB05, EHDS12, AKAKDT11, CLL+09, KRC10], to the best of the author's knowledge, none of the authors has worked on the problem for the next-generation networks with disruptive architectural design such as DeUD. For example, both of authors in [CH12] and [LCCZ15] propose user association schemes in CoUD. The goal of the former is to jointly maximize the system capacity in DL and to minimize transmitting power consumption in UL, while the aim of the latter is to minimize the sum of UL and DL average traffic delay and to reduce the overall UL and DL power consumption.

Another restriction of the existing works is that they concern with the intra-cell communication either in the standard OFDMA-based networks or in the static or dynamic TDD-based networks. For example, the authors in [EHDS12] proposed a subcarrier allocation algorithm to maximize a utility function that captures the joint UL/DL QoS requirements, by formulating the problem as a two-sided stable matching game. In [KL09], a network utility maximization framework is proposed to solve the joint UL/DL resource allocation problem considering systems with frequency-division duplex (FDD) or static TDD through the user-level satisfaction.

Decoupled Uplink and Downlink Access

The concept of downlink/uplink decoupling (DUDe)[1] is introduced in [And13, ADF$^+$13, BHL$^+$14, BAE$^+$15]. The recent contributions can be classified in three groups.

The first group of articles focuses on the architectural design and realization. The pioneering contributions [BHL$^+$14, BAE$^+$15] identify and explain some key arguments in favor of DUDe based on a blend of theoretical, experimental, and logical arguments.

The second group proposes varies link association policies and show the performance gain with simulations based on LTE field trial network. In [EBDI14a], the notion of DUDe is studied, where the downlink cell association is based on the downlink received power while the uplink is based on the pathloss. The follow-up work [EBDI14b] considers the cell-load as well as the available backhaul capacity during the association process. One other idea for range extension of small cells in UL is to add a cell selection offset to the reference signals, to increase the priority of the small cells to be selected [Qua08].

Last but not least, the third group of articles studies on the analytical evaluation of a predefined association policy. The work in [SEP$^+$14, SPG15] focuses on the analytical characterization of the decoupled access by using the framework of stochastic geometry, applying the same association criteria as in [EBDI14a]. In [SZA14], the authors propose a model to characterize the uplink SINR and rate distribution as a function of the association rules (assuming weighted pathloss for both UL and DL association) and power control parameters (assuming fractional pathloss-inversion based power control).

[1] In this paper, we use a different term DeUD for "decoupled uplink/downlink", in consistency with the term CoUD for "coupled uplink/downlink".

Fixed-Point Based Framework for Max-Min Utility Maximization

Yates [Yat95, YH95] proposed a framework of power control that is based on the notions of positivity, monotonicity, and scalability of standard interference functions (for details see Appendix D.3.2), to solve the SIR balancing problem. Since then, the framework of interference calculus is widely studied for the utility maximization involving only power and rate control. In [UY98, LUE03, LUE05], the authors extend Yates' framework to stochastic power control algorithms.

The authors in [CB04, BSSW05, SBS05, BS08, SWB09] studied the max-min utility fairness problem with deterministic interference function involving power or rate control, and characterized the feasibility using the Perron-Frobenius theorem [FFFF12]. Recent work [ZT14, HTZ^{+}14] leverages the nonlinear Perron-Frobenius theory [LN12] and overcome the non-convexity or non-monotonicity in special cases of wireless utility maximization. In [ZT14], examples of SINR- or reliability-related non-convex utility optimization were introduced involving power control only. In [HTZ^{+}14], the author proposes a general framework that enables rigorous treatment of nonlinear monotonic constraints in the utility fairness resource allocation problems.

In [Nuz07], the properties of standard interference function are re-examined from a contraction mapping point of view, where the convergence to a unique fixed point follows by a version of the Banach fixed point theorem [Sma80]. The theory provided in [Nuz07] can be extended to certain non-monotonic functions.

Interference Model Based on Power and Load Coupling

The above-mentioned work typically addresses the inter-cell interference model with power coupling. In [SY12, Reaar, HYS14], the authors consider the inter-cell interference characterized by the load coupling model, where cell load measures the average level of resource usage in the cell and implies the probability of generating interference from a transmitter to a receiver in orthogonal frequency-division multiplexing (OFDM) sytsems. The interaction between power and load coupling are analyzed in [CPS14,HYLSon]. The authors in [CPS14] derive an interference mapping having as its fixed point the power allocation including a given load profile. The authors in [HYLSon] address an energy minimization problem, and prove that operating at fill load is optimal in minimizing the sum energy.

9.1.2 Contribution

The main contributions of this paper are listed as follows.

We consider the next-generation wireless HetNets with disruptive architectural design with respect to dynamic splitting of UL/DL resource and link association. A common set of resource blocks are considered joint resource for both UL and DL services, and adaptive resource partitioning between UL and DL is enabled to adapt to the link-specific traffic demand. The decoupled UL and DL access is further introduced to adapt to the link-specific channel condition (as shown in Fig. 9.5).

We introduce a general model of inter-cell interference for joint UL/DL system. It includes the inter-link interference between UL and DL and is power and load coupling-aware. A general class of resource constraint is then formulated, applicable for various types of power or load constraints. For example, the sum per-cell power budget constraint in the downlink depends on both the power per resource block and the number of assigned RB in the downlink. The per-cell load constraint depends on the number of RBs assigned both in the uplink and downlink. We then develop a framework involving a fixed-point class with nonlinear contraction operators (mainly motivated by the work in [Nuz07]), and an optimizer for the utility of QoS satisfaction level, subjected to a general class of resource constraints. A three-step optimization algorithm is proposed, to find the local optimum of the joint variables bandwidth allocation and power spectral density on a per-link basis, corresponding to the different link association policies.

To adapt the framework to the practical interest, we extend the work to cover the following aspects: 1) per-transmitter power control instead of per-link power control, and 2) energy efficient power control.

The rest of the chapter is organized as follows. In Section 9.2 we introduce some basic notations and system model. In Section 9.3, we present the utility fairness problem and its decomposition into two subproblems. The solution to the subproblem of adaptive joint UL/DL bandwidth allocation is provided in Section 9.4, while of joint UL/DL power control (including the extension to the per-transmitter power control and energy efficient power control) in Section 9.5. The joint algorithm to solve the main optimization problem is summarized in Section 9.6. The performance of the proposed algorithms are evaluated numerically in Section 9.7. We conclude the study in Section 9.8.

9.2 System Model

In this paper, we use the following standard definitions. The nonnegative and positive orthant in k dimensions are denoted by $\mathbb{R}^k_+$ and $\mathbb{R}^k_{++}$, respectively. Let $\boldsymbol{x} \leq \boldsymbol{y}$ denote the component-wise inequality between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. And let $\mathrm{diag}(\boldsymbol{x})$ denote a diagonal matrix with the elements of $\boldsymbol{x}$ on the main diagonal. For a function $\boldsymbol{f}: \mathbb{R}^k \to \mathbb{R}^k$, $\boldsymbol{f}^n$ denotes the n-fold composition so that $\boldsymbol{f}^n := \boldsymbol{f} \circ \boldsymbol{f}^n$. The $k \times k$ identity matrix is denoted by $\boldsymbol{I}_k$ and the $n \times k$ zero matrix is denoted by $\boldsymbol{0}_{n\times k}$. The k-dimensional all-ones (all-zeros) vector is denoted by $\boldsymbol{1}_k$ ($\boldsymbol{0}_k$). The horizontal concatenation of two matrices $\boldsymbol{A} \in \mathbb{R}^{n\times k}$, $\boldsymbol{B} \in \mathbb{R}^{n\times l}$ is written as $[\boldsymbol{A} \mid \boldsymbol{B}]$, while the vertical concatenation of two matrices $\boldsymbol{A} \in \mathbb{R}^{n\times k}$, $\boldsymbol{B} \in \mathbb{R}^{m\times k}$ is written as $[\boldsymbol{A};\boldsymbol{B}]$. The cardinality of set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. The notation that will be used in this paper is summarized in Table 9.1.

We consider an OFDM-based wireless system consisting of a set of BSs $\mathcal{N}$ with $|\mathcal{N}| = N$ and a set of UEs $\mathcal{K}$ with $|\mathcal{K}| = K$. We drop the usual assumption in wireless system design that UL and DL transmissions are associated with the same BS, and assume that they can be split. Let the UL(DL) cell-UE association matrix be denoted by $\boldsymbol{A}^{\mathrm{UL}} \in \{0,1\}^{N\times K}(\boldsymbol{A}^{\mathrm{DL}} \in \{0,1\}^{N\times K})$.

We assume the reciprocal UL and DL channels. The set of all links (including ULs and DLs) is denoted by $\overline{\mathcal{K}} := \mathcal{K}^{\mathrm{UL}} \cup \mathcal{K}^{\mathrm{DL}}$, where $\mathcal{K}^{\mathrm{UL}}$ and $\mathcal{K}^{\mathrm{DL}}$ are the sets of ULs and DLs, respectively. Because ULs and DLs have different transmitters and receivers, we have that $\mathcal{K}^{\mathrm{UL}} \cap \mathcal{K}^{\mathrm{DL}} = \emptyset$. Without loss of generality, we assume that $|\mathcal{K}^{\mathrm{UL}}| = |\mathcal{K}^{\mathrm{DL}}| = K$ and $|\mathcal{K}| = 2K$. We define the power spectral density (PSD) to be the transmit power assigned per RB, and we use $\boldsymbol{p}^{\mathrm{UL}} \in \mathbb{R}_{+}^{K}$ and $\boldsymbol{p}^{\mathrm{DL}} \in \mathbb{R}_{+}^{K}$ to denote the vectors of uplink and downlink PSDs, respectively. Accordingly, $\boldsymbol{w}^{\mathrm{UL}} \in [0,1]^{K}$ is used to denote fraction of the allocated RBs (normalized by dividing the number of allocated RBs by the total number of the available RBs), while $\boldsymbol{w}^{\mathrm{DL}} \in [0,1]^{K}$ is the vector for such fractions in the downlink. We collect $\boldsymbol{p}^{\mathrm{UL}}$ and $\boldsymbol{p}^{\mathrm{DL}}$ in one power vector $\boldsymbol{p} := [\boldsymbol{p}^{\mathrm{UL}};\boldsymbol{p}^{\mathrm{DL}}] \in \mathbb{R}_{+}^{2K}$, and collect $\boldsymbol{w}^{\mathrm{UL}}$ and $\boldsymbol{w}^{\mathrm{DL}}$ in $\boldsymbol{w} := [\boldsymbol{w}^{\mathrm{UL}};\boldsymbol{w}^{\mathrm{DL}}] \in [0,1]^{2K}$. Let the total number of the RBs be denoted by W_0.

We consider the flexible duplex mode that allows UL and DL transmissions to share a joint set of RBs and to dynamically split between the RBs allocated to UL and DL. The split ratio is time-variant and cell-specific. Flexible duplex mode is proposed as the next step of FDD/TDD convergence in 5G networks [All15, DMP+14]. The rapid evolution of subband-based splitting and filtering [ZM15] and full duplex technology [BJK14] makes dynamic splitting of spectrum allocated to UL and DL realizable in the near future. The main drawback results from the need for coping with more intricate inter-cell interference structures: the interference is not only restricted to UL-to-UL and DL-to-DL interference, but also includes the inter-link interference between UL and DL, as shown in Fig. 9.3.

Remark 9.1 (Adaptation to Dynamic TDD)**.** *Although in this paper the system model and optimization algorithm are developed based on forward-looking assumption of flexible duplex, they can be well adapted to more practical system with dynamic TDD configuration, by interpreting $\boldsymbol{w}^{UL}$ and $\boldsymbol{w}^{DL}$ as fraction of time frames dedicated to UL and DL, respectively. In this incident, we can see the resource on the horizontal axis in Fig.9.3 as time frames instead of frequency subbands, and the inter-cell inter-link interference appears in the central frames that are used for UL transmission in BS j, while for DL transmission in another BS i.*

Table 9.1: NOTATION SUMMARY

$\mathcal{N}$	set of (macro and pico) BSs
$\mathcal{K}$	set of UEs
$\mathcal{K}^{\mathrm{UL}}$ ($\mathcal{K}^{\mathrm{DL}}$)	set of ULs (DLs)
$\overline{\mathcal{K}}$	set of all service links
$\boldsymbol{A}^{\mathrm{UL}}$ ($\boldsymbol{A}^{\mathrm{DL}}$)	BS assignment matrix for ULs (DLs)
$\boldsymbol{A}$	BS assignment matrix for all service links
Π	set of link association policies
b_k^{UL} (b_k^{DL})	BS associated to the kth UL (DL)
$\boldsymbol{p}^{\mathrm{UL}}$ ($\boldsymbol{p}^{\mathrm{DL}}$)	PSD for ULs (DLs)
$\boldsymbol{p}$	PSD for all service links
$\boldsymbol{q}^{\mathrm{DL}}$	cell-specific PSD in DL
$\overline{\boldsymbol{p}}$	per-transmitter PSD
$\boldsymbol{w}^{\mathrm{UL}}$ ($\boldsymbol{w}^{\mathrm{DL}}$)	fraction of allocated RBs for ULs (DLs)
$\boldsymbol{w}$	fraction of allocated RBs for all service links
d_l	traffic demand (bit rate) of the lth link, $l \in \overline{\mathcal{K}}$
r_l	spectral efficiency of the lth link, $l \in \overline{\mathcal{K}}$
W_0	total number of RBs
$\boldsymbol{V}$	link gain coupling matrix
$\tilde{\boldsymbol{V}}$	link gain coupling matrix without intra-cell interference
$g_1(\boldsymbol{w})$	constraint function implying the constraint on load
$g_2(\boldsymbol{w}, \boldsymbol{p})$	contraint function implying the contraint on transmit power
λ	objective utility

9.2.1 Constrained Per-Cell Load and Per-Transmitter Power

Since the UL and DL transmissions share a common set of resource blocks, we define the *cell load* to be the fraction of the total occupied frequency resource (in UL and DL) per cell. We collect the per-cell loads in a vector $\boldsymbol{\nu} := \boldsymbol{A}\boldsymbol{w} \in [0,1]^N$, where $\boldsymbol{A} := \left[\boldsymbol{A}^{\mathrm{UL}} \mid \boldsymbol{A}^{\mathrm{DL}}\right] \in \{0,1\}^{N \times 2K}$ is the binary association matrix. Since the per-cell load is bounded above by 1, we have

$$\mathbb{R}_+^{2K} \to [0,1]: \; g_1(\boldsymbol{w}) := \|\boldsymbol{A}\boldsymbol{w}\|_\infty \leq 1. \tag{9.1}$$

This implies that for each cell, the sum of the fractions of allocated RBs for both UL and DL is constrained, i.e., $\forall n \in \mathcal{N}$ we have $\sum_{k \in \mathcal{K}} \left(a_{n,k}^{\mathrm{UL}} w_k^{\mathrm{UL}} + a_{n,k}^{\mathrm{DL}} w_k^{\mathrm{DL}}\right) \leq 1$.

Let $\boldsymbol{p}_{\max}^{\mathrm{UL}} \in \mathbb{R}_{++}^K$ and $\boldsymbol{q}_{\max}^{\mathrm{DL}} \in \mathbb{R}_{++}^N$ denote the maximum UL transmit power per UE and the maximum DL transmit power per BS for the whole frequency band, respectively. Note that the maximum transmit power of a macro BS and a pico BS can vastly differ from each other in HetNets. We define the extended maximum power vector by $\boldsymbol{p}_{\max}^{\mathrm{ext}} := [\boldsymbol{p}_{\max}^{\mathrm{UL}}; \boldsymbol{q}_{\max}^{\mathrm{DL}}] \in \mathbb{R}_{++}^{K+N}$ and the extended assignment matrix for transmitter-to-link association by $\boldsymbol{A}^{\mathrm{ext}} := [\boldsymbol{I}_K \mid \boldsymbol{0}_{K \times K}; \boldsymbol{0}_{N \times K} \mid \boldsymbol{A}^{\mathrm{DL}}] \in \{0,1\}^{(K+N) \times 2K}$. The per-transmitter (including both UEs and

BSs) power constraints imply that

$$\begin{aligned}&\mathbb{R}_+^{2K} \times \mathbb{R}_+^{2K} \to \mathbb{R}_+ :\\ &g_2(\boldsymbol{w}, \boldsymbol{p}) := W_0 \| \operatorname{diag}(\boldsymbol{p}_{\max}^{\text{ext}})^{-1} \boldsymbol{A}^{\text{ext}} \operatorname{diag}(\boldsymbol{w}) \boldsymbol{p} \|_\infty \leq 1, \end{aligned} \tag{9.2}$$

which is equivalent to $\sum_{k \in \mathcal{K}} a_{n,k}^{\text{DL}} (W_0 w_k^{\text{DL}}) p_k^{\text{DL}} \leq q_{\max,n}^{\text{DL}}$, $\forall n \in \mathcal{N}$, and $(W_0 w_k^{\text{UL}}) p_k^{\text{UL}} \leq p_{\max,k}^{\text{UL}}$, $\forall k \in \mathcal{K}$. This means that the total transmit power per transmitter, computed as PSD[2] multiplied by the total number of occupied RBs, is constrained by the predefined maximum power budget. Note that $\operatorname{diag}(\boldsymbol{w})\boldsymbol{p}$ and $\operatorname{diag}(\boldsymbol{p})\boldsymbol{w}$ are interchangeable. Moreover, for any fixed $\hat{\boldsymbol{p}}$ or $\hat{\boldsymbol{w}}$, the function g_2 over the joint variable $(\boldsymbol{w}, \boldsymbol{p})$ can be written as $g_{2,\hat{\boldsymbol{w}}}(\boldsymbol{p}) : \mathbb{R}_+^{2K} \to \mathbb{R}_+$ or $g_{2,\hat{\boldsymbol{p}}}(\boldsymbol{w}) : \mathbb{R}_+^{2K} \to \mathbb{R}_+$.

9.2.2 Link Gain Coupling Matrix

The interference coupling between users (as shown in Fig. 9.5) is characterized by a link gain coupling matrix. To define this matrix, we define three channel gain matrices $\boldsymbol{H}_0 \in \mathbb{R}_{++}^{N \times K}$, $\boldsymbol{H}_1 \in \mathbb{R}_{++}^{N \times N}$ and $\boldsymbol{H}_2 \in \mathbb{R}_{++}^{K \times K}$ to indicate BS-to-UE, BS-to-BS, and UE-to-UE channel gain, respectively. The link gain coupling matrix between the $2K$ transmission links (UL and DL) is then defined to be

$$\begin{aligned}\boldsymbol{V} &:= \begin{bmatrix} \boldsymbol{V}^{\text{UL}\leftarrow\text{UL}} & \boldsymbol{V}^{\text{UL}\leftarrow\text{DL}} \\ \boldsymbol{V}^{\text{DL}\leftarrow\text{UL}} & \boldsymbol{V}^{\text{DL}\leftarrow\text{DL}} \end{bmatrix} \quad (9.3)\\ &= \begin{bmatrix} \boldsymbol{A}^{\text{UL}^T} \boldsymbol{H}_0 & \boldsymbol{A}^{\text{UL}^T} \boldsymbol{H}_1 \boldsymbol{A}^{\text{DL}} \\ \boldsymbol{H}_2 & \boldsymbol{H}_0^T \boldsymbol{A}^{\text{DL}} \end{bmatrix}. \quad (9.4)\end{aligned}$$

The matrices $\boldsymbol{V}^{\text{X}\leftarrow\text{Y}} := \left(v_{k,j}^{\text{X}\leftarrow\text{Y}}\right) \in \mathbb{R}_{++}^{K \times K}$, $\text{X}, \text{Y} \in \{\text{UL}, \text{DL}\}$, determine the cross-link couplings. For example, $v_{k,j}^{\text{UL}\leftarrow\text{DL}}$ denotes the channel gain coupling between the transmitter of the downlink to UE j and the receiver of the uplink from UE k as shown in Fig. 9.5. Note that $\boldsymbol{V}^{\text{UL}\leftarrow\text{UL}}$, $\boldsymbol{V}^{\text{UL}\leftarrow\text{DL}}$ and $\boldsymbol{V}^{\text{DL}\leftarrow\text{DL}}$ are in general not symmetric, while $\boldsymbol{V}^{\text{DL}\leftarrow\text{UL}}$ is symmetric.

We assume that each base station employs an OFDM-based scheme for resource allocation to schedule its users on orthogonal resources. As a result, there is no intra-cell interference and the interference coupling is completely described by the modified link gain matrix $\tilde{\boldsymbol{V}}$, which is defined by (9.3) with $\boldsymbol{V}^{\text{X}\leftarrow\text{Y}}$ replaced by $\tilde{\boldsymbol{V}}^{\text{X}\leftarrow\text{Y}} := \left(\tilde{v}_{k,j}^{\text{X}\leftarrow\text{Y}}\right)$ where

$$\tilde{v}_{k,l}^{\text{X}\leftarrow\text{Y}} := \begin{cases} v_{k,l}^{\text{X}\leftarrow\text{Y}} & \text{if } b_l^{\text{Y}} \neq b_k^{\text{X}} \\ 0 & \text{o.w.} \end{cases}. \tag{9.5}$$

Here and hereafter, b_k^{X}, $\text{X} \in \{\text{UL}, \text{DL}\}$ denotes the serving BS of UE k in UL or DL.

[2] Note that in this chapter the unit of PSD is Watt per RB.

9.2.3 Models of SINR and Rate

To capture the dynamic inter-cell interference in OFDM systems, it is reasonable to assume that the inter-cell interference increases as the fraction of the allocated RBs at the interfering BSs increases as well. We interpret $\boldsymbol{w}$ as the probability of generating interference from the transmitter of a link to the receiver of the other link (on any RB) [MNK$^+$07]. More precisely, we assume that the DL and UL SINR per RB of UE k is given by (respectively)

$$\mathrm{SINR}_k^{\mathrm{DL}} := \frac{p_k^{\mathrm{DL}} v_{k,k}^{\mathrm{DL}\leftarrow\mathrm{DL}}}{\sum\limits_{i\in\mathcal{K}} \tilde{v}_{k,i}^{\mathrm{DL}\leftarrow\mathrm{DL}} w_i^{\mathrm{DL}} p_i^{\mathrm{DL}} + \sum\limits_{j\in\mathcal{K}} \tilde{v}_{k,j}^{\mathrm{DL}\leftarrow\mathrm{UL}} w_j^{\mathrm{UL}} p_j^{\mathrm{UL}} + \sigma^2}$$

$$\mathrm{SINR}_k^{\mathrm{UL}} := \frac{p_k^{\mathrm{UL}} v_{k,k}^{\mathrm{UL}\leftarrow\mathrm{UL}}}{\sum\limits_{i\in\mathcal{K}} \tilde{v}_{k,i}^{\mathrm{UL}\leftarrow\mathrm{DL}} w_i^{\mathrm{DL}} p_i^{\mathrm{DL}} + \sum\limits_{j\in\mathcal{K}} \tilde{v}_{k,j}^{\mathrm{UL}\leftarrow\mathrm{UL}} w_j^{\mathrm{UL}} p_j^{\mathrm{UL}} + \sigma^2}$$

where $\sigma^2 > 0$ denotes the background-noise power spectral density, which is assumed to be the same for all receivers. Note that in this expression of SINRs $\boldsymbol{w}$ as the probability has no unit, and both the numerator and denominator have the same units Watt per RB. Let us define $\boldsymbol{\sigma} := \sigma^2 \mathbf{1}_{2K}$, and collect the uplink and downlink SINR in a vector $\mathbf{SINR} := [\mathrm{SINR}_1^{\mathrm{UL}}; \ldots; \mathrm{SINR}_K^{\mathrm{UL}}; \mathrm{SINR}_1^{\mathrm{DL}}; \ldots; \mathrm{SINR}_K^{\mathrm{DL}}] \in \mathbb{R}_{++}^{2K}$. Using (9.3), (9.4), and (9.5), the above expressions of SINR can be written in a general form

$$\mathrm{SINR}_l(\boldsymbol{p}, \boldsymbol{w}) := \frac{p_l}{\left[\boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{V}} \operatorname{diag}\{\boldsymbol{p}\}\boldsymbol{w} + \boldsymbol{\sigma}\right)\right]_l}, \ l \in \overline{\mathcal{K}}, \tag{9.6}$$

where $\boldsymbol{D} := \operatorname{diag}\{v_{1,1}^{\mathrm{UL}\leftarrow\mathrm{UL}}, \ldots, v_{K,K}^{\mathrm{UL}\leftarrow\mathrm{UL}}, v_{1,1}^{\mathrm{DL}\leftarrow\mathrm{DL}}, v_{K,K}^{\mathrm{DL}\leftarrow\mathrm{DL}}\} \in \mathbb{R}_+^{2K}$ is a diagonal matrix. For $l = 1, \ldots, K$, (9.6) is equal to the UL SINR, while the DL SINR is given by (9.6) for $l = K+1, \ldots, 2K$.

We further assume that the spectral efficiency (bit rate per RB) of the virtual UEs (includes both UL and DL transmission) is a strictly increasing function of the SINR given by

$$r_l(\boldsymbol{p}, \boldsymbol{w}) := B \log_2(1 + \mathrm{SINR}_l(\boldsymbol{p}, \boldsymbol{w})), \ l \in \overline{\mathcal{K}}, \tag{9.7}$$

where B denotes the effective bandwidth per RB.

Given the per-UE uplink and downlink traffic demands (bit rate)

$$\boldsymbol{d} := [d_1^{\mathrm{UL}}, \ldots, d_K^{\mathrm{UL}}, d_1^{\mathrm{DL}}, \ldots, d_K^{\mathrm{DL}}]^T \in \mathbb{R}_{++}^{2K},$$

it follows from (9.7) that the traffic demands are satisfied if and only if (note that $w_l \cdot W_0$ is equal to the number of RBs used by link l)

$$w_l \geq \frac{d_l}{W_0 r_l(\boldsymbol{p}, \boldsymbol{w})}, \ l \in \overline{\mathcal{K}}. \tag{9.8}$$

Remark 9.2 (Full Overlap or Partial Overlap)**.** *The SINR modeled in* (9.6) *is based on the strategy that each UL or DL transmission is allocated a number of RBs in a joint frequency band for both UL and DL, regardless of the location of the band. However, this may result in a full overlap of frequency bands used by UL and DL transmissions leading to high probability of inter-link interference. A more reasonable strategy is to allow only partial overlap, as shown in Fig. 9.3, where the DL is preferred to allocated at the head of the band while the UL at the tail of the band, or vice versa. In this case, the inter-link interference only exists on the overlapping band, and the above-presented model overestimates the probability of receiving inter-link interference. A more accurate readjustment is to multiply the term of inter-link interference with an additional overlap factor. Some possible methods to define the overlap factor are given in Appendix D.3.1. In the remainder of this paper, the analysis and algorithms are still presented with the interference model in* (9.6) *for the simplicity of the form. However, without loss of generality, we can easily adjust the model by introducing the overlap factor into the coupling matrix* $\tilde{\boldsymbol{V}}$*.*

9.2.4 Link Association Policies

Assume that there are a finite set of link association policies $\Pi := \{\pi_m : m = 1, \ldots, M\}$ implemented in the network, which can be dynamically selected by an operator. Each policy defines the BS-UE assignment matrices $\boldsymbol{A}^{\mathrm{UL}}(\pi_m)$ and $\boldsymbol{A}^{\mathrm{DL}}(\pi_m)$, and further defines the link gain coupling matrix $\tilde{\boldsymbol{V}}(\pi_m)$ and link gain matrix $\boldsymbol{D}(\pi_m)$ in (9.6).

As examples, in the following we list one conventional UL/DL coupled user association policy and two types of decoupled UL/DL link association policies, respectively.

(1) **CoUD**: Conventional coupled UL/DL user association based on reference signal received power (RSRP) in DL is given by

$$b_k^{\mathrm{UL}} = b_k^{\mathrm{DL}} = \arg\max_{n\in\mathcal{N}} \mathrm{RSRP}_{n,k}, \ \forall k \in \mathcal{K}. \tag{9.9}$$

(2) **DeUD_O**: Decoupled UL/DL link association assisted with cell selection offset [Qua08]. A cell selection offset is added to the reference signals of the small cells to increase their coverage in UL in order to offload some traffic from the macro cell. This can be formalized as follows

$$b_k^{\mathrm{X}} = \arg\max_{n\in\mathcal{N}} \mathrm{RSRP}_{n,k} + \mathrm{offset}_n^{\mathrm{X}}, \ \forall k \in \mathcal{K}, \mathrm{X} \in \{\mathrm{UL}, \mathrm{DL}\} \tag{9.10}$$

where $\mathrm{offset}_n^{\mathrm{X}} > 0$ (in dB) if $\mathrm{X} = \mathrm{UL}$ and n is a small cell BS with low transmit power; otherwise the offset is set to zero if $\mathrm{X} = \mathrm{DL}$ or n is a macro cell BS.

(3) **DeUD_P**: Decoupled UL/DL link association based on DL received power and UL

pathloss respectively [EBDI14a], where the association criteria in DL and UL are given by (respectively)

$$b_k^{\mathrm{DL}} = \arg\max_{n\in\mathcal{N}} \mathrm{RSRP}_{n,k}, \tag{9.11}$$

$$b_k^{\mathrm{UL}} = \arg\max_{n\in\mathcal{N}} \mathrm{PL}_{n,k},\ \forall k \in \mathcal{K}, \tag{9.12}$$

where $\mathrm{PL}_{n,k}$ denotes the pathloss estimate between BS n and UE k.

Note that in (9.10), by setting $\mathrm{offset}_n^{\mathrm{X}} = 0$ for all $n \in \mathcal{N}$ and X = UL, the association policy is equivalent to CoUD. And, by setting the offset (in dB) of the small cell BS in UL as the difference between the transmit power (in dBm) of the macro cell BS and the small cell BS, DeUD_O is equivalent to DeUD_P.

9.3 Problem Formulation

To achieve the service-centric network fairness, we define the objective utility λ to be the *minimum level of QoS satisfaction* among all links, where the level of QoS satisfaction is equal to the ratio of the per-link feasible transmission rate to the required traffic demand. So we have

$$\lambda := \min_{l\in\overline{\mathcal{K}}} \frac{W_0 w_l r_l(\boldsymbol{p},\boldsymbol{w})}{d_l}, \tag{9.13}$$

where $r_l(\boldsymbol{p},\boldsymbol{w})$ is given by (9.7).

Given a certain link association policy π' and its corresponding UL(DL) assignment matrix $\boldsymbol{A}^{\mathrm{UL}}(\pi')(\ \boldsymbol{A}^{\mathrm{DL}}(\pi'))$, coupling matrix $\tilde{\boldsymbol{V}}(\pi')$, and link gain matrix $\boldsymbol{D}(\pi')$, the objective is to maximize the utility λ over the joint space of loads and powers subject to the constraints on the maximum per-cell load (9.1) and the maximum per-transmitter power (9.2). Moreover, if the optimized utility satisfies $\lambda \geq 1$, then the vector of link-specific traffic demands $\boldsymbol{d}$ is feasible; otherwise, the traffic demand cannot be satisfied for every service link. Formally, the problem of interest in this paper can be stated as follows.

Problem 9.1.

$$\max_{\boldsymbol{w}\in\mathbb{R}_+^{2K},\boldsymbol{p}\in\mathbb{R}_+^{2K}} \quad \lambda \tag{9.14a}$$

$$\textit{subject to } \boldsymbol{w} \geq \lambda \boldsymbol{f}(\boldsymbol{p},\boldsymbol{w}) \tag{9.14b}$$

$$f_l(\boldsymbol{p},\boldsymbol{w}) := \frac{d_l}{W_0 r_l(\boldsymbol{p},\boldsymbol{w})}, \forall l \in \overline{\mathcal{K}} \tag{9.14c}$$

$$(9.1), (9.2), \tag{9.14d}$$

where the vector function $\boldsymbol{f} : \mathbb{R}_+^{2K} \to \mathbb{R}_{++}^{2K}$ *in* (9.14b) *is a collection of* f_l *defined in* (9.14c)*, i.e.,* $\boldsymbol{f} := [f_1, \ldots, f_{2K}]^T$*.* The utility λ depends on the joint variable $(\boldsymbol{w},\boldsymbol{p}) \in \mathbb{R}_+^{2K} \times \mathbb{R}_+^{2K}$

in an inextricably intertwined manner, which is due to the nonlinear power and resource coupling relationship between links. We decompose Problem 9.1 into two subproblems in Problem 9.2b by alternately optimizing over $\boldsymbol{w}$ or $\boldsymbol{p}$, and provide computationally efficient locally optimal solution to Problem 9.1, based on the optimal solution to each of the subproblems.

Problem 9.2.

9.2a *Given fixed $\boldsymbol{p}' \in \mathbb{R}_+^{2K}$, find $\boldsymbol{w}' := \boldsymbol{w}'(\boldsymbol{p}')$ such that*

$$\boldsymbol{w}' = \arg\max_{\boldsymbol{w}\in\mathbb{R}_+^{2K}} \lambda \tag{9.15a}$$

$$\text{subject to } \boldsymbol{w} \geq \lambda \boldsymbol{f}_{\boldsymbol{p}'}(\boldsymbol{w}) \tag{9.15b}$$

$$g_1(\boldsymbol{w}) \leq 1, g_{2,\boldsymbol{p}'}(\boldsymbol{w}) \leq 1, \tag{9.15c}$$

where $\boldsymbol{f}_{\boldsymbol{p}'}$, g_1, and $g_{2,\boldsymbol{p}'}$ are obtained by replacing $\boldsymbol{p}$ with $\boldsymbol{p}'$ in (9.14c), (9.1) *and* (9.2), *respectively.*

9.2b *Given fixed $\boldsymbol{w}' \in \mathbb{R}_+^{2K}$ satisfying $g_1(\boldsymbol{w}') \leq 1$, find $\boldsymbol{p}' := \boldsymbol{p}'(\boldsymbol{w}')$ such that*

$$\boldsymbol{p}' = \arg\max_{\boldsymbol{p}\in\mathbb{R}_+^{2K}} \lambda \tag{9.16a}$$

$$\text{subject to } \boldsymbol{w}' \geq \lambda \boldsymbol{f}_{\boldsymbol{w}'}(\boldsymbol{p}) \tag{9.16b}$$

$$g_{2,\boldsymbol{w}'}(\boldsymbol{p}) \leq 1, \tag{9.16c}$$

where $\boldsymbol{f}_{\boldsymbol{w}'}$ and $g_{2,\boldsymbol{w}'}$ are obtained by replacing $\boldsymbol{w}$ with $\boldsymbol{w}'$ in (9.14c) *and* (9.2), *respectively.*

Prob.9.2a and Prob.9.2b are formulated in such a way that a common desired utility λ is maximized subject to the common load and power constraints. Thus, for a given link association policy π', by sequentially solving Prob.9.2a and Prob.9.2b, we improve λ in each step and achieve a local optimum of λ with respect to π'.

In Section 9.4 and 9.5 we provide the optimal solution to Prob.9.2a and Prob.9.2b, respectively. The joint algorithm is summarized in Section 9.6.

9.4 Joint Uplink and Downlink Resource Allocation

In this section we develop the algorithms for joint UL/DL bandwidth allocation. In Section 9.4.1 we develop an algorithm for Prob.9.2a in Prop. 9.1. Since a solution $\boldsymbol{w}$ to Prob.9.2a must fulfill $\max\{g_1(\boldsymbol{w}), g_{2,\boldsymbol{p}'}(\boldsymbol{w})\} \leq 1$, some free resources may still be available, i.e., $g_1(\boldsymbol{w}) < 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}) = 1$, even under optimal power allocation (in the sense of Prob.

9.2a). Therefore, an additional step involving power scaling and bandwidth updating is introduced in Prop. 9.2 in Section 9.4.2, to further improve the desired utility λ. Another case of $g_1(\boldsymbol{w}) = 1$ and $g_{2,\mathbf{p}'}(\boldsymbol{w}) \leq 1$ is discussed in Prop. 9.3 in Section 9.5.

9.4.1 Algorithm for Bandwidth Allocation

The following lemma proves a key property of the vector function $\boldsymbol{f}_{\mathbf{p}'}$, which is necessary to solve Prob. 9.2a.

Lemma 9.1. *Given a fixed power vector $\mathbf{p}'$, the function $\boldsymbol{f}_{\mathbf{p}'} : \mathbb{R}_{+}^{2K} \to \mathbb{R}_{++}^{2K}$ defined in Prob. 9.1 is a standard interference function.*

The definition and some selected properties of standard interference function (SIF) are provided in Appendix D.3.2. The proof of Lemma 9.1 following the proof of [Reaar, Ex. 2] is provided in Appendix D.3.3.

We further prove the following theorem.

Theorem 9.1. *Suppose*

- $g(\boldsymbol{x}) : \mathbb{R}_{++}^{k} \to \mathbb{R}_{++}$ *is monotonic, and homogeneous of degree* 1 *(i.e.,* $g(\alpha \boldsymbol{x}) = \alpha g(\boldsymbol{x})$ *for all* $\alpha > 0$*),*
- $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}_{+}^{k} \to \mathbb{R}_{++}^{k}$ *is a SIF.*

Then, for each $\theta > 0$ there is exactly one eigenvector $\boldsymbol{x}' \in \mathbb{R}_{++}^{k}$ and associate eigenvalue ρ' of $\boldsymbol{f}$ such that $\rho' \boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}')$ and $g(\boldsymbol{x}') = \theta$. The repeated iteration

$$\boldsymbol{x}^{(t+1)} = \frac{\theta \boldsymbol{f}(\boldsymbol{x}^{(t)})}{g \circ \boldsymbol{f}(\boldsymbol{x}^{(t)})}, \; t \in \mathbb{N}, \tag{9.17}$$

converges to the unique vector $\boldsymbol{x}'$, which is called the fixed point of $\boldsymbol{f}$. The associate eigenvalue is $\rho' = g \circ \boldsymbol{f}(\boldsymbol{x}')/\theta$.

The proof of Theorem 9.1 is referred to Appendix D.3.4. It is an extension of the proof of [Nuz07, Th. 3.2], where g was defined as any monotonic norm $\|\cdot\|$, while we define two properties monotonicity, homogeneity and positivity on $\mathbb{R}_{++}^{k}$. Note that the function in (9.17) $\boldsymbol{\psi} := \theta \boldsymbol{f}/g \circ \boldsymbol{f} : \mathbb{R}_{+}^{k} \to \mathbb{R}_{++}^{k}$ is non-monotonic, while it preserves the property of scalability of the mapping $\boldsymbol{f}$.

Using Lemma 9.1 and Theorem 9.1, we prove the following proposition, which gives rise to an algorithmic solution to Prob.9.2a.

Proposition 9.1. *Given a fixed* $\boldsymbol{p}' \in \mathbb{R}_+^{2K}$*, let the set of solutions to Prob.9.2a be denoted by* $\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{p}')$*. There exists one* $\boldsymbol{w}' \in \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{p}')$ *such that* $\boldsymbol{w}' \leq \boldsymbol{w}$ *for all* $\boldsymbol{w} \in \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{p}')$*. Moreover,* $\boldsymbol{w}'$ *is an eigenvector of* $\boldsymbol{f}_{\boldsymbol{p}'}$ *satisfying* $\max\{g_1(\boldsymbol{w}'), g_{2,\boldsymbol{p}'}(\boldsymbol{w}')\} = 1$ *and can be obtained by performing the following fixed point iteration:*

$$\boldsymbol{w}^{(t+1)} = \frac{\boldsymbol{f}_{\boldsymbol{p}'}(\boldsymbol{w}^{(t)})}{g_{\boldsymbol{p}'} \circ \boldsymbol{f}_{\boldsymbol{p}'}(\boldsymbol{w}^{(t)})}, \; t \in \mathbb{N}, \tag{9.18a}$$

$$\textit{where } g_{\boldsymbol{p}'}(\boldsymbol{w}) := \max\{g_1(\boldsymbol{w}), g_{2,\boldsymbol{p}'}(\boldsymbol{w})\}. \tag{9.18b}$$

The iteration in (9.18) *converges to* $\boldsymbol{w}'$*, and* $\lambda_{\boldsymbol{p}'} = 1/g_{\boldsymbol{p}'} \circ \boldsymbol{f}_{\boldsymbol{p}'}(\boldsymbol{w}')$*.*

The proof of Prop. 9.1 is provided in Appendix D.3.5.

9.4.2 Optimization to Achieve Maximum Load

As aforementioned, Prop.9.1 provides an algorithm that converges to the optimal solution to Prob.9.2a. Let $\boldsymbol{w}'$ be this solution. Since $\max\{g_1(\boldsymbol{w}'), g_{2,\boldsymbol{p}'}(\boldsymbol{w}')\} = 1$, it is possible that $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$, while $g_1(\boldsymbol{w}') < 1$, i.e., the maximum power per transmitter is satisfied with equality, while free resources are still available. In this case, we propose an additional step to further optimize λ by iteratively scaling down the fixed power vector $\boldsymbol{p}'$, until $g_1(\boldsymbol{w}') = 1$ is achieved.

Proposition 9.2. *Let* $\boldsymbol{w}' \in \mathbb{R}_+^{2K}$ *be the solution to Prob.9.2a and suppose that* $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$ *and* $g_1(\boldsymbol{w}') < 1$*. Starting from* $\boldsymbol{p}^{(0)} = \boldsymbol{p}'$ *and* $\boldsymbol{w}^{(0)} = \boldsymbol{w}'$*, by iteratively performing the following two steps:*

(1) scaling down $\boldsymbol{p}$ *by*

$$\boldsymbol{p}^{(t+1)} = g_1(\boldsymbol{w}^{(t)}) \cdot \boldsymbol{p}^{(t)}, \tag{9.19}$$

(2) updating $\boldsymbol{w}^{(t+1)}$*, as the unique fixed point of iteration* (9.18)*, with updated* $\boldsymbol{p}' = \boldsymbol{p}^{(t+1)}$*,*

the sequence of utility λ *is monotone increasing, until the maximum load constraint* $g_1(\boldsymbol{w}) = 1$ *is satisfied.*

The proof of Prop. 9.2 is provided in Appendix D.3.6.

The optimization step provided in Prop. 9.2 further improves our desired utility λ if the solution to Prob.9.2a $\boldsymbol{w}'$ satisfies $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$ and $g_1(\boldsymbol{w}') < 1$. Now assume the algorithm defined in Prop. 9.2 converges to $(\boldsymbol{p}^\star, \boldsymbol{w}^\star)$. Then, in addition to the full utilization of resources in the sense that $g_1(\boldsymbol{w}^\star) = 1$, we have $g_2(\boldsymbol{p}^\star, \boldsymbol{w}^\star) \leq 1 = g_{2,\boldsymbol{p}'}(\boldsymbol{w}')$, which means that the allocation obtained under Prop. 9.1 is more power efficient than that of Prop. 9.1.

Remark 9.3. *It is worth mentioning that Ho [HYLSon] formulates a power minimization problem, based on the cell-specific load and power coupling in the DL, and concludes that if the minimum required rate is feasible, then the optimal solution to the power minimization problem satisfies that the system is fully loaded [HYLSon, Th. 1]. In this paper, we formulate a utility maximization problem, based on the link-specific bandwidth and power coupling framework in joint UL/DL, with per-cell load and per-transmitter power constraints, and conclude that if some minimum utility is feasible with cell load lower than one, we can scale down the power vector using the algorithm presented in Prop. 9.2, to further increase the desired utility, until the per-cell load constraint holds with equality.*

9.5 Joint Uplink and Downlink Power Control

Now let us consider the problem of power control. In this section, we first present the optimal solution to Prob.9.2b introduced in Section 9.5.1. Then, in Section 9.5.2 and 9.5.3, we further examine two alternative algorithms for cell-specific power control and energy efficient power control, respectively.

9.5.1 Algorithm for Link-Specific Power Control

Let us first consider Prob.9.2b. Given some fixed $\boldsymbol{w}' \in [0,1]^{2K}$, we first rewrite the rate constraints in (9.16b). For $\boldsymbol{p} \in \mathbb{R}^{2K}_{++}$, we have

$$\boldsymbol{w}' \geq \lambda \boldsymbol{f}_{\boldsymbol{w}'}(\boldsymbol{p}) \Leftrightarrow p_l \geq \lambda \frac{p_l f_{\boldsymbol{w}',l}(\boldsymbol{p})}{w'_l} \text{ for } l \in \overline{\mathcal{K}}. \tag{9.20}$$

We further define the following vector function using (9.20)

$$\begin{aligned} \tilde{\boldsymbol{f}}_{\boldsymbol{w}'} :& \mathbb{R}^{2K}_{++} \to \mathbb{R}^{2K}_{++} : \boldsymbol{p} \mapsto \left[\tilde{f}_{\boldsymbol{w}',1}(\boldsymbol{p}), \ldots, \tilde{f}_{\boldsymbol{w}',2K}(\boldsymbol{p})\right]^T \\ \text{where } \tilde{f}_{\boldsymbol{w}',l}(\boldsymbol{p}) :=& \frac{p_l}{w'_l} f_{\boldsymbol{w}',l}(\boldsymbol{p}),\ l \in \overline{\mathcal{K}}. \end{aligned} \tag{9.21}$$

Note that the domain of $\tilde{\boldsymbol{f}}_{\boldsymbol{w}'}$ defined in (9.21) is the positive orthant $\mathbb{R}^{2K}_{++}$. To extend it to the non-negative orthant $\mathbb{R}^{2K}_{+}$, we define the following extension for each $l \in \overline{\mathcal{K}}$:

$$\boldsymbol{f}'_{\boldsymbol{w}',l}(\boldsymbol{p}) := \begin{cases} \tilde{\boldsymbol{f}}_{\boldsymbol{w}',l}, & \text{if } p_l \neq 0 \\ \dfrac{d_l \ln 2}{W_0 B w'_l} \mathcal{I}_{\boldsymbol{w}',l}(\boldsymbol{p}) & \text{o.w.} \end{cases}, \tag{9.22}$$

$$\text{where } \mathcal{I}_{\boldsymbol{w}',l}(\boldsymbol{p}) := \left[\boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{V}} \operatorname{diag}\{\boldsymbol{w}'\}\boldsymbol{p} + \boldsymbol{\sigma}\right)\right]_l. \tag{9.23}$$

The domain extension is derived by leveraging the linear approximation $\log_2(1+x) \approx x/\ln 2$ for $x \to 0$. As shown in (9.22), this approximation is only used for $p_l = 0$ (which further leads to $\text{SINR}_l = 0$), otherwise if $p_l \neq 0$, the nonlinear closed-form of $\tilde{\boldsymbol{f}}_{\boldsymbol{w}',l}$ (9.21) is used.

With (9.20), (9.22), and (9.23), Prob.9.2b is rewritten as

$$\max_{\boldsymbol{p}\in\mathbb{R}_+^{2K}} \lambda, \text{ s.t. } \boldsymbol{p} \geq \lambda \boldsymbol{f}'_{\boldsymbol{w}'}(\boldsymbol{p}),\ g_{2,\boldsymbol{w}'}(\boldsymbol{p}) \leq 1 \tag{9.24}$$

The following lemma shows that $\boldsymbol{f}'_{\boldsymbol{w}'}$ has the same key property as $\boldsymbol{f}_{\boldsymbol{p}'}$, which is shown for $\boldsymbol{f}_{\boldsymbol{p}'}$ in Lemma 9.1.

Lemma 9.2. *The vector function* $\boldsymbol{f}'_{\boldsymbol{w}'} : \mathbb{R}_+^{2K} \to \mathbb{R}_{++}^{2K}$ *defined in* (9.22) *is SIF.*

Proof. The proof follows directly from the previous results in [CPS14, Prop. 1], where a cell-specific utility function over the cell-specific power vector in DL is shown to be positive concave, and thus a SIF [Reaar, Prop. 1]. It is easy to see that our defined link-specific function $\boldsymbol{f}'_{\pi',\boldsymbol{w}'}$ shares the same form with the cell-specific function introduced in [CPS14, Prop. 1]. Thus, we omit the details here and conclude that it is also a SIF. ■

Note that in the expression of per-transmitter power constraint (9.2), the term diag$(\boldsymbol{w})\boldsymbol{p}$ and diag$(\boldsymbol{p})\boldsymbol{w}$ are interchangeable. With some fixed $\boldsymbol{w}'$, the function $g_{2,\boldsymbol{w}'}$ defined in (9.24) is monotonic, positive and homogeneous of degree 1 on $\mathbb{R}_{++}^{2K}$. Thus, by leveraging Lemma 9.2 and Theorem 9.1, we can argue along similar lines as in Prop. 9.1 to conclude the following: starting from an arbitrary $\boldsymbol{p}^{(1)} \in \mathbb{R}_+^{2K}$, the following fixed point iteration

$$\boldsymbol{p}^{(t+1)} = \frac{\boldsymbol{f}'_{\boldsymbol{w}'}(\boldsymbol{p}^{(t)})}{g_{2,\boldsymbol{w}'} \circ \boldsymbol{f}'_{\boldsymbol{w}'}(\boldsymbol{p}^{(t)})},\ t \in \mathbb{N} \tag{9.25}$$

converges to the solution of Prob.9.2b, denoted by $\boldsymbol{p}''$. And the utility $\lambda_{\boldsymbol{p}''}$ corresponding to $\boldsymbol{p}''$ is given by $\lambda_{\boldsymbol{p}''} = 1/g_{2,\boldsymbol{w}'} \circ \boldsymbol{f}'_{\boldsymbol{w}'}(\boldsymbol{p}'')$.

Using (9.25), we can iteratively approach arbitrarily close to solution to Prob.9.2b given fixed $\boldsymbol{w}'$ as the solution to Prob.9.2a. However, for joint optimization over $(\boldsymbol{w}, \boldsymbol{p})$, we are interested in whether or not this solution further improves the desired utility derived from the solution to Prob.9.2a. We present the relationship between $\lambda'' := \lambda_{\boldsymbol{p}''}$ and $\lambda' := \lambda_{\boldsymbol{p}'}$ in Prop. 9.3.

Proposition 9.3. *For some fixed* $\boldsymbol{p}'$*, let* $\boldsymbol{w}' \in \mathbb{R}_{++}^{2K}$ *be the solution to Prob.9.2a and* λ' *the corresponding utility. Moreover, given* $\boldsymbol{w}'$*, let* $\boldsymbol{p}'' \in \mathbb{R}_{++}^{2K}$ *be the solution to Prob.9.2b and* λ'' *the corresponding utility. Then,* λ' *and* λ'' *are related as follows.*

- *If* $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$*, then, we have* $\lambda'' = \lambda'$ *and* $\boldsymbol{p}'' = \boldsymbol{p}'$
- *If* $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$*, then, we have* $\lambda'' > \lambda'$

The proof of Prop. 9.3 can be found in Appendix D.3.7.

Prop. 9.3 implies that given the solution $(\boldsymbol{w}', \boldsymbol{p}')$ derived from the bandwidth updating step (Prop. 9.1) or the power scaling step (Prop. 9.2), with fixed $\boldsymbol{w}'$ at hand, solving Prob.9.2b (by performing (9.25)) can further improve the desired utility only if $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$; otherwise if $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$ the solution to Prob.9.2b with respect to $\boldsymbol{w}'$ is equivalent to $\boldsymbol{p}'$.

Remark 9.4. *In this section, we rewrite the rate constraints $\boldsymbol{w}' \geq \lambda \boldsymbol{f_p}(\boldsymbol{w}')$ in Prob. 9.2b into a system of nonlinear inequalities $\boldsymbol{p} \geq \lambda \boldsymbol{f}'_{\boldsymbol{w}'}(\boldsymbol{p})$ as shown in* (9.20)*-*(9.23)*. Hence both the fixed point iterations in* (9.18) *and* (9.25) *(to solve Prob. 9.2a and Prob. 9.2b, respectively) converge to the solutions that maximize the same λ defined in Prob. 9.3. Note that if we treat the power control problems separately, as stated for instance in [BS05], the rate constraint $r_l(\boldsymbol{p}, \boldsymbol{w}') \geq \lambda d_l/(w'_l W_0)$ for all $l \in \overline{\mathcal{K}}$ can be directly translate into a SINR constraint by taking the exponential function of both sides. We write* (9.20) *into a system of linear inequalities in powers:*

$$p_l \geq \eta(\lambda) \boldsymbol{f}''_{\boldsymbol{w}'}(\boldsymbol{p})$$

where $\eta(\lambda) := 2^{\frac{\lambda d_l}{W_0 B w'_l}} - 1$ is monotone increasing for any $\lambda \in \mathbb{R}^{2K}_{+}$, and $f''_{\boldsymbol{w}'} : \mathbb{R}^{2K}_{+} \to \mathbb{R}^{2K}_{++}$ is of form of an affine transformation $\boldsymbol{p} \mapsto \boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{V}} \operatorname{diag}(\boldsymbol{w}')\boldsymbol{p} + \boldsymbol{\sigma}\right)$. We can agree along similar lines as in Prop. 9.1 to maximize η by performing the fixed point iteration $\boldsymbol{p} = f''_{\boldsymbol{w}'}(\boldsymbol{p})/(g_{2,\boldsymbol{w}'} \circ f''_{\boldsymbol{w}'}(\boldsymbol{p}))$ and thus indirectly maximize λ.

9.5.2 Algorithm for Cell-Specific Power Control

So far we have considered the case that the PSD $\boldsymbol{p}$ can be specified per service link. In the practical system, however, in DL a transmitter determines constant cell-specific energy per resource element across all DL bandwidth and subframes until it needs to be updated [3GPe], while in UL a distinct transmission power can be assigned to each UE. Without loss of generality, the developed power control algorithm can be easily modified to meet this practical requirement. The objective is to optimize the per-transmitter PSD as a collection of the per-UE UL and per-BS DL power vectors

$$\overline{\boldsymbol{p}} := [\boldsymbol{p}^{\mathrm{UL}}; \boldsymbol{q}^{\mathrm{DL}}]^T \in \mathbb{R}^{K+N}_{+}, \tag{9.26}$$

where $\boldsymbol{q}^{\mathrm{DL}} \in \mathbb{R}^{N}_{+}$ is the cell-specific PSD in DL, and the nth entry of q^{DL}_l denotes the PSD of all the DLs associated to cell n. Since all DLs served by the same cell share the same PSD, we have

$$\boldsymbol{p}^{\mathrm{DL}} = \boldsymbol{A}^{\mathrm{DL}^T} \boldsymbol{q}^{\mathrm{DL}}. \tag{9.27}$$

The transformation between $\boldsymbol{p}$ and $\overline{\boldsymbol{p}}$ is then given by

$$\boldsymbol{p} = \boldsymbol{\Lambda}\overline{\boldsymbol{p}}, \text{ with } \boldsymbol{\Lambda} := \begin{bmatrix} \boldsymbol{I}_K & \boldsymbol{0}_{K\times N} \\ \boldsymbol{0}_{K\times K} & {\boldsymbol{A}^{\mathrm{DL}}}^T \end{bmatrix}. \tag{9.28}$$

In the following, we collect the per-UE rate constraint in UL and per-cell sum rate constraint in DL depending on $\overline{\boldsymbol{p}}$ in a set of $K+N$ nonlinear inequalities, where for $j \in \mathcal{K}$ the jth inequality implies the UL rate constraint for UE j, while for $j \in \overline{\mathcal{N}} := \{K+1,\ldots,K+N\}$, the jth inequality implies the DL sum rate constraint for cell $n = j-K$.

Per-UE Rate Constraint in Uplink

Substituting (9.28) into (9.6), SINR of UE j in UL is simply given by

$$\mathrm{SINR}_j(\overline{\boldsymbol{p}}, \boldsymbol{w}') := \frac{\overline{p}_j}{\overline{\mathcal{I}}_{\boldsymbol{w}',j}(\overline{\boldsymbol{p}})}, \text{ for } j \in \mathcal{K}, \tag{9.29}$$

$$\text{where } \overline{\mathcal{I}}_{\boldsymbol{w}',j}(\overline{\boldsymbol{p}}) := \left[\boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{V}}\operatorname{diag}\{\boldsymbol{w}'\}\boldsymbol{\Lambda}\overline{\boldsymbol{p}} + \boldsymbol{\sigma}\right)\right]_j. \tag{9.30}$$

Substituting (9.29) into (9.7) and (9.8), the per-UE rate constraint in UL depending on $\overline{\boldsymbol{p}}$ is given by

$$\overline{p}_j \geq \frac{\overline{p}_j}{w_j} \cdot \frac{d_j}{W_0 r_j(\overline{\boldsymbol{p}}, \boldsymbol{w}')} =: \tilde{f}_{\boldsymbol{w}',j}(\overline{\boldsymbol{p}}), \text{ for } j \in \mathcal{K}. \tag{9.31}$$

Per-Cell Sum Rate Constraint in Downlink

Substituting (9.28) into (9.6), the DL SINR of UE k associated with cell n (depending on $\overline{\boldsymbol{p}}$) can be rewritten as:

$$\mathrm{SINR}^{\mathrm{DL}}_{n,l}(\overline{\boldsymbol{p}}, \boldsymbol{w}') := \frac{\overline{p}_{K+n}}{\overline{\mathcal{I}}_{\boldsymbol{w}',l}(\overline{\boldsymbol{p}})}, \ \forall l \in \overline{\mathcal{K}}^{\mathrm{DL}}_n, \tag{9.32}$$

where $\overline{\mathcal{I}}_{\boldsymbol{w}',l}(\overline{\boldsymbol{p}})$ is defined in (9.30), $\overline{\mathcal{K}}^{\mathrm{DL}}_n$ denotes the set of DL transmissions associated with cell n, and $\overline{p}_{K+n}$ as the $(K+n)$th entry of $\overline{\boldsymbol{p}}$ denotes the PSD in DL in cell n.

The spectral efficiency of UE k associated with cell n in DL and denoted by $r^{\mathrm{DL}}_{n,l}(\overline{\boldsymbol{p}}, \boldsymbol{w}')$ is computed by substituting (9.32) into (9.7). Then, using (9.8), the sum rate constraint per cell in DL (depending on $\overline{\boldsymbol{p}}$) yields

$$\nu'_n = \sum_{l \in \overline{\mathcal{K}}^{\mathrm{DL}}_n} w'_l \geq \sum_{l \in \overline{\mathcal{K}}^{\mathrm{DL}}_n} \frac{d_l}{W_0 r^{\mathrm{DL}}_{n,l}(\overline{\boldsymbol{p}}, \boldsymbol{w}')}, \ \forall n \in \mathcal{N} \tag{9.33}$$

$$\begin{aligned} \Rightarrow \overline{p}_j &\geq \frac{\overline{p}_j}{\nu'_{j-K}} \sum_{l \in \overline{\mathcal{K}}^{\mathrm{DL}}_{j-K}} \frac{d_l}{W_0 r^{\mathrm{DL}}_{j-K,l}(\overline{\boldsymbol{p}}, \boldsymbol{w}')} \\ &=: \tilde{f}_{\boldsymbol{w}',j}(\overline{\boldsymbol{p}}), \text{ for } j \in \overline{\mathcal{N}} \end{aligned} \tag{9.34}$$

where ν'_n denotes fraction of the total allocated RBs of cell n in DL, note that for $j \in \overline{\mathcal{N}}$, the jth entry of $\overline{\boldsymbol{p}}$ is equal to the PSD of cell $n = j - K$ in DL.

Note that (9.34) defines the jth entry of function $\tilde{f}_{\boldsymbol{w}',j}$ for $j = K+1, \ldots, K+N$, while for $j = 1, \ldots, K$, the expression of $\tilde{f}_{\boldsymbol{w}',j}$ is given in (9.31).

Joint Downlink Cell-Specific and Uplink UE-specific Power Control

With (9.31) and (9.34) in hand, using the same techniques as shown in (9.20)-(9.23), the optimization problem is written as

$$\max_{\overline{\boldsymbol{p}} \in \mathbb{R}_{+}^{K+N}} \lambda, \text{ s.t. } \overline{\boldsymbol{p}} \geq \lambda \overline{\boldsymbol{f}}_{\boldsymbol{w}'}(\overline{\boldsymbol{p}}), \ \overline{g}_{2,\boldsymbol{w}'}(\overline{\boldsymbol{p}}) \leq 1 \tag{9.35}$$

where $\overline{g}_{2,\boldsymbol{w}'}(\overline{\boldsymbol{p}})$ is obtained by substituting (9.28) into (9.2), and $\overline{\boldsymbol{f}}_{\boldsymbol{w}'}(\overline{\boldsymbol{p}})$ is given by

$$\begin{aligned} &\overline{f}_{\boldsymbol{w}',j}(\overline{\boldsymbol{p}}) := \\ &\begin{cases} \tilde{f}_{\boldsymbol{w}',j}(\overline{\boldsymbol{p}}) & \text{if } \overline{p}_j \neq 0 \\ \dfrac{d_l \ln 2}{W_0 B w'_j} \overline{\mathcal{I}}_{\boldsymbol{w}',j}(\overline{\boldsymbol{p}}) & \text{if } \overline{p}_j = 0, j \in \mathcal{K} \\ \displaystyle\sum_{l \in \mathcal{K}_{j-K}^{\text{DL}}} \dfrac{d_l \ln 2}{W_0 B \nu'_{j-K}} \overline{\mathcal{I}}_{\boldsymbol{w}',l}(\overline{\boldsymbol{p}}) & \text{if } \overline{p}_j = 0, j \in \overline{\mathcal{N}} \end{cases} \end{aligned} \tag{9.36}$$

Proceeding long similar lines as in Lemma 9.2, it is easy to show that $\overline{\boldsymbol{f}}_{\boldsymbol{w}'} : \mathbb{R}_{+}^{K+N} \rightarrow \mathbb{R}_{++}^{K+N}$ is SIF, while $\overline{g}_{2,\boldsymbol{w}'} : \mathbb{R}_{++}^{K+N} \rightarrow \mathbb{R}_{++}$ is monotonic and homogeneous with degree 1. Therefore, we can compute the solution to (9.35) by means of the fixed point iteration in (9.25), and with $\boldsymbol{f}'_{\boldsymbol{w}'}(\boldsymbol{p})$ replaced by $\overline{\boldsymbol{f}}_{\boldsymbol{w}'}(\overline{\boldsymbol{p}})$.

9.5.3 Algorithm for Energy Efficient Power Control

If the following assumption holds, the rate requirements are strictly feasible for all UL and DL transmissions.

Assumption 9.1. *The solution to Prob. 9.2* $(\boldsymbol{w}^\star, \boldsymbol{p}^\star)$ *satisfies* $\lambda^\star > 1$.

Under Assumption 9.1, the problem of interest in the context of energy efficient networks is that, instead of consuming high energy to achieve $\lambda > 1$, how to minimize the sum transmit power, such that the per-link rate constraint is just satisfied, i.e., $\lambda = 1$. The power minimization problem subjected to the rate and power constraints are defined in Problem 9.3

Problem 9.3.

$$\min_{\boldsymbol{p} \in \mathbb{R}_{+}^{2K}} \psi(\boldsymbol{p}), \ \textit{s.t. } \boldsymbol{p} \geq \boldsymbol{f}'_{\boldsymbol{w}^\star}(\boldsymbol{p}), \ g_{2,\boldsymbol{w}^\star}(\boldsymbol{p}) \leq 1 \tag{9.37}$$

where $\psi : \mathbb{R}_+^{2K} \to \mathbb{R}_+$ can be any monotonic function (in each coordinate, i.e., $\psi(\boldsymbol{x}) \geq \psi(\boldsymbol{y})$ iff $x_i \geq y_i$ for each i) that is non-decreasing. For example, by setting $\psi(\boldsymbol{p}) = \| \operatorname{diag}\{\boldsymbol{w}^\star\}\boldsymbol{p}\|_1$, we aim at minimizing the sum transmit power over all occupied RBs and all transmitters.

Since $\boldsymbol{f}'_{\boldsymbol{w}^\star}$ is SIF, Prob. 9.3 is a classical power minimization problem introduced in [YH95], and we provide the solution in Prop. (9.4). We omit the proof because it follows directly from [YH95, Thm. 2].

Proposition 9.4. *Under Assumption 9.1, the fixed point iteration*

$$\boldsymbol{p}^{(t+1)} = \boldsymbol{f}'_{\boldsymbol{w}^\star}\left(\boldsymbol{p}^{(t)}\right), t \in \mathbb{N} \tag{9.38}$$

converges to the optimum solution $\boldsymbol{p}^{\star\star}$ to Prob. 9.3.

Note that without loss of generality, (9.37) can be easily translated to the power minimization problem over $\overline{\boldsymbol{p}}$ by substituting (9.28) into (9.37) and replacing $\boldsymbol{f}'_{\boldsymbol{w}^\star}$ with $\overline{\boldsymbol{f}}_{\boldsymbol{w}'}$.

9.6 Algorithm for Joint Optimization

Now we provide an algorithm for joint optimization of bandwidth allocation $\boldsymbol{w}$ and power control $\boldsymbol{p}$ per link, with respect to any fixed link association policy $\pi' \in \Pi$. Based on Prop. 9.1, 9.2, and 9.3, we can compute the locally optimum of $(\boldsymbol{w}(\pi'), \boldsymbol{p}(\pi'))$. In the following we explain in more detail the three main steps (S1, S2 and S3) of the algorithm.

S1. Updating Bandwidth

The algorithm starts with optimizing the bandwidth allocation $\boldsymbol{w}$, given an initial PSD $\boldsymbol{p}'$. Prop. 9.1 provides the optimal solution $\boldsymbol{w}'$ in the sense of maximizing λ for any fixed $\boldsymbol{p}'$. The algorithm converges to a solution $\boldsymbol{w}'$, satisfying $\max\{g_1(\boldsymbol{w}'), g_{2,\boldsymbol{p}'}(\boldsymbol{w}')\} = 1$, i.e., either $g_1(\boldsymbol{w}') = 1$, or $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$, or both. Therefore, it remains to consider the following three cases

(1) $g_1(\boldsymbol{w}') < 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$

(2) $g_1(\boldsymbol{w}') = 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$

(3) $g_1(\boldsymbol{w}') = 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$

Note that once the third condition is achieved, $(\boldsymbol{w}', \boldsymbol{p}')$ is a local optimum. In contrast, in the first case and the second case the algorithm is designed to further improve the utility by proceeding with S2 and S3 (see Algorithm 6), respectively.

S2. Power Scaling to Achieve The Full Load Condition

The first condition leads to the power scaling step as described in Prop. 9.2. At this step, power scaling (9.19) and bandwidth updating (9.18) are performed iteratively, until the solution $(\boldsymbol{p}', \boldsymbol{w}')$ converges and satisfies $g_1(\boldsymbol{w}') = 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') \leq 1$.

(1) If $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$, then $(\boldsymbol{p}', \boldsymbol{w}')$ is considered the local optimum.

(2) If $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$, then the algorithm moves to the power updating step S3.

S3. Updating Power Budget

As shown in Prop. 9.3, the power updating step improves the utility if $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$, where $(\boldsymbol{w}', \boldsymbol{p}')$ are derived from the bandwidth updating step S1. Therefore, the algorithm moves to S3 if either of the following conditions holds.

(1) S1 returns $g_1(\boldsymbol{w}') = 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$, and the algorithm moves directly to S3.

(2) S1 returns $g_1(\boldsymbol{w}') < 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$, and the algorithm moves to S2. If S2 returns $g_1(\boldsymbol{w}') = 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$, then, algorithm further moves to S3.

Remark 9.5 (Selection of The Initial Point)**.** *The initial point has in general a significant impact on the outcome of the algorithm. We use the transmit power budget defined in the 3GPP specification [3GPe] as the reference to compute the initial PSD* $\mathbf{p}'$*, such that the optimized solution of* $(\boldsymbol{w}, \boldsymbol{p})$ *is guaranteed to provide a better performance than the standard configuration. The power spectral density in dBm (per RB) of link* $l \in \overline{\mathcal{K}}$ *is defined by* $PSD_l = \min\{PSD_{\max}, \mathrm{SNR}_l^{tar} + P_{noise} + \alpha PL_l\}$*, where* $PSD_{\max}$ *denotes the maximum PSD,* SNR_l^{tar} *is the open loop SNR target for the* l*th link,* P_{noise} *is the noise PSD in the receiver,* α *is the pathloss compensate factor, and* $PL_l := PL_{b_l,l}$ *is the pathloss estimate of the link* l *served by BS* b_l*.*

9.7 Numerical Results

In this section, we verify the propositions presented in Section 9.4 and 9.5, show the convergence of Algorithm 6, and compare the performance with the proposed algorithm to the conventional resource allocation schemes under different association policies presented in Section 9.2.4 through simulations.

9.7.1 Simulation Parameters

To obtain practically relevant results, we study the real-world scenario as shown in Fig. 9.6. This map shows the center of Berlin, Germany in the WGS 84 coordinate system. There are 81 BSs, among which 45 of them are macro cell BSs (1 BS per sector) with directional antenna and maximum transmit power of 43 dBm, while 36 of them are pico cell BSs with

Algorithm 6: Joint Allocation of Bandwidth and Power

```
input  : $\boldsymbol{p}' \leftarrow \hat{\boldsymbol{p}} \in \mathbb{R}^{2K}_{++}$, $\boldsymbol{w}' \leftarrow \hat{\boldsymbol{w}} \in \mathbb{R}^{2K}_{++}$, $\boldsymbol{w} \leftarrow \boldsymbol{0}$, $\lambda \leftarrow 0$, $\pi' \in \Pi$, $\epsilon_1$, $\epsilon_2$, $\epsilon_3$
output: $\boldsymbol{w}^\star$, $\boldsymbol{p}^\star$
Compute $\boldsymbol{A}^{\mathrm{UL}}(\pi')$, $\boldsymbol{A}^{\mathrm{DL}}(\pi')$, $\tilde{\boldsymbol{V}}(\pi')$ and $\boldsymbol{D}(\pi')$;
% S1: Update $\boldsymbol{w}$ based on Prop.9.1;
while $\|\boldsymbol{w}' - \boldsymbol{w}\|_\infty \geq \epsilon_2$ do
    $\boldsymbol{w} \leftarrow \boldsymbol{w}'$;
    % Fixed point iteration (9.18);
    $\boldsymbol{w}' \leftarrow$ UpdateBandwidth($\boldsymbol{p}', \boldsymbol{w}$);
% S2: Update $\boldsymbol{w}$ to achieve full load based on Prop.9.2;
if $g_1(\boldsymbol{w}') < 1 \& g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$ then
    while $g_1(\boldsymbol{w}') < 1$ do
        $\boldsymbol{p} \leftarrow \boldsymbol{p}'$;
        % Power scaling in (9.19);
        $\boldsymbol{p}' \leftarrow$ ScalePower($\boldsymbol{w}', \boldsymbol{p}$);
        while $\|\boldsymbol{w}' - \boldsymbol{w}\|_\infty \geq \epsilon_2$ do
            $\boldsymbol{w} \leftarrow \boldsymbol{w}'$;
            % Fixed point iteration (9.18);
            $\boldsymbol{w}' \leftarrow$ UpdateBandwidth($\boldsymbol{p}', \boldsymbol{w}$);
% S3: Update $\boldsymbol{p}$;
if $g_1(\boldsymbol{w}') = 1 \& g_{2,\boldsymbol{p}'}(\boldsymbol{w}') < 1$ then
    $\boldsymbol{p} \leftarrow \boldsymbol{0}$;
    while $\|\boldsymbol{p}' - \boldsymbol{p}\|_\infty \geq \epsilon_3$ do
        $\boldsymbol{p} \leftarrow \boldsymbol{p}'$;
        % Fixed point iteration (9.25);
        $\boldsymbol{p}' \leftarrow$ UpdatePower($\boldsymbol{w}', \boldsymbol{p}$);
$\boldsymbol{w}(\pi') \leftarrow \boldsymbol{w}'$; $\boldsymbol{p}(\pi') \leftarrow \boldsymbol{p}'$; $\lambda(\pi') \leftarrow \lambda'$;
```

omni-directional antenna and maximum transmit power of 30 dBm. We assume that a total bandwidth of 5 MHz is subdivided into 25 RBs of 12 subcarriers each, and that the frequency reuse factor is 1. The color map refers to the pathloss in dB. For each pixel of 50×50m size, the channel gain over all received downlink signals from the macro cell BSs is given according to the measured data of pathloss from [MOM04]. The pico cell BSs are randomly placed on the cell edge of the macro cells. Based on the 3GPP LTE model provided in [3GPj], we obtain the pathloss between the pico BSs and the UEs to compute $\boldsymbol{H}_0$ (joint with the macro-to-UE pathloss), the pathloss between the BSs to compute $\boldsymbol{H}_1$, and the pathloss between the mobile terminals to compute $\boldsymbol{H}_2$. On top of this realistic pathloss, we implement uncorrelated fast fading characterized by Rayleigh distribution. We assume reciprocal uplink and downlink channels.

The users are uniformly randomly distributed in the playground. The maximum transmit power of the user terminal is 22 dBm. We define 5 service classes, with the downlink rate requirements of $[300, 25, 50, 10, 0.01]$ Mbit/s, and the corresponding uplink rate requirements of $[50, 50, 25, 10, 0.01]$ Mbit/s. These classes imply the following 5 services: 1) cloud service video and other digital service, 2) HD video/photo sharing, 3) high-resolution video and other digital services, 4) broadband data allowing video email and web surfing, and 5) text, voice or video messages.

9.7.2 Convergence of the Algorithm

Let us first examine the convergence behavior of the algorithms presented in Prop. 9.1, 9.2 and 9.3 (corresponding to S1, S2, and S3) in Algorithm 6, respectively. In Fig.9.7 we verify the propositions and show the convergence of the algorithm 6 with the fixed association policy DeUD_P, at a single simulation snapshot (i.e., the users are assumed to be static within one time interval). The number of users is $K = 500$. The desired numerical precisions are set to $\epsilon_i = 1e - 7$, for $i = 1, 2, 3$.

Fig. 9.7(a) illustrates the convergence behavior of three successive steps S1, S2, and S3. The algorithm starts at step S1, where $g_1(\boldsymbol{w}^{(0)}) < 1$ and $g_2(\boldsymbol{p}^{(0)}, \boldsymbol{w}^{(0)}) < 1$. The initial power $\boldsymbol{p}^{(0)}$ is chosen as described in Rem. 9.5, where $\text{PSD}_{\text{max}} = 12$ dBm, $\text{SNR}^{\text{tar}} = 12.2$ dB, $\alpha = 1$, and $P_{\text{noise}} = -121.45$ dBm. The initial bandwidth allocation is defined as $\boldsymbol{w}^{(0)} = \boldsymbol{0}$. After performing the fixed point iteration (9.18) at S1, it converges to the fixed point $\boldsymbol{w}'$ such that $g_2(\boldsymbol{p}^{(0)}, \boldsymbol{w}') = 1$ while $g_1(\boldsymbol{w}')$ is extremely small (approximately 0.01). The algorithm moves therefore to S2 of power scaling. The algorithm at S2 converges to the point $(\boldsymbol{w}'', \boldsymbol{p}')$, where $g_1(\boldsymbol{w}') = 1$ and $g_2(\boldsymbol{w}'', \boldsymbol{p}') < 1$, which causes the algorithm to move to S3. By the end of S3, the fixed point iteration (9.25) converges to $\boldsymbol{p}''$ such that $g_1(\boldsymbol{w}'') = g_2(\boldsymbol{w}'', \boldsymbol{p}'') = 1$, and the algorithm terminates. At each step, the iteration improves the desired utility λ monotonically.

An interesting observation we have made concerning the relationship between per-cell power constraint and the feasible utility is illustrated in Fig. 9.7(b). The motivation is to find out the tradeoff between the power consumption and the improvement of the utility. Fig. 9.7(b) shows the increase of the utility as we increase the power constraint factor θ (θ increases from 0.01 to 1.01 with step size of 0.01), under different self-noise power σ. As shown in Thm. 9.1, θ is the scaling factor of the monotonic constraint $g(\boldsymbol{x})$. As for S3, in particular, θ is scaling factor of the maximum power constraint such that $g_{2,\boldsymbol{w}'}(\boldsymbol{p}) \leq \theta$. For small value of σ (i.e., in an interference-dominant system), small value of θ is sufficient for the feasible utility, and increase of θ only leads to minor increase of utility (blue and red curves for the noise power of -121 dBm and -100 dBm, respectively). Conversely, for the large value of σ (i.e., in a noise-dominant system), increase of θ has a stronger effect on improving utility (green and black curves for the noise power of -80 dBm and -70 dBm, respectively). The above observation can help us to choose a proper operation point, to provide a good tradeoff between the total power consumption and the desired utility.

Fig. 9.7(c) and 9.7(d) are provided to illustrate the performance of algorithms presented in Section 9.5.2 and 9.5.3. Fig. 9.7(c) shows a case that restricting cell-specific DL power results in approximately 16% degradation of utility achieved by UE-specific DL power. Fig. 9.7(d) shows a specific example that for a certain snap shot of the network, over 90% of power consumption can be saved if we only target at required utility $\lambda = 1$ instead of the maximum feasible λ, by performing the step of energy efficient power control presented in Section 9.5.3.

9.7.3 Network Performance Evaluation

Selection of Association Policy

Now let us examine the performance of Algorithm 6 under different link association policies. The set of association policies Π, including CoUD, DeUD_O (with variety of offsets) and DeUD_P as introduced in Section 9.2.4, is defined as follows. Note that all macro cell BSs have maximum transmit power of 43 dBm, while all small cell BSs of 30 dBm. Thus, by setting $\text{offset}_n^{\text{UL}} = 13$ dB for n as small cell BS, the policy DeUD_O is equivalent to DeUD_P, while by setting $\text{offset}_n^{\text{UL}} = 0$ for all $n \in \mathcal{N}$, the policy DeUD_O is equivalent to CoUD. The set of policy Π is then defined as a set of DeUD_O policies with offsets $\{0, 1, 3, 5, \ldots, 51\}$ of the small cell BSs in UL, where 0 corresponding to CoUD and 13 corresponding to DeUD_P.

Fig. 9.8 shows the average performance of the algorithm under each policy $\pi \in \Pi$ using the Monte Carlo techniques. We run 500 independent tests, with uniform user distribution of 100 static users in each test. Fig. 9.8(a) shows the percentage of the counts that a fixed policy provides the utility among the top three maximum utilities achieved by all policies. Fig. 9.8(b) shows the average utility of a fixed policy over the 500 tests (the high value of utility is due to the lower number of the users compared to Fig. 9.7). The following two observations are made. 1) Proper selection of DeUD policy can achieve approximately 2× improvements on desired utility, compared against CoUD. 2) Although DeUD_P is not always the best policy that provides maximum utility, it has a high chance to provide relatively good performance (approximately 73% of counts among the top three maximum utility). Thus, in case the operator wants to save the computational cost of exhaustive searching for optimal association policies, always selecting DeUD_P provides a suboptimal compromises. However, we shall remind that in many cases, DeUD_P is not the best association policy with respect to maximizing the desired utility, as shown in the two examples of the single trial in Fig. 9.8(c) and Fig. 9.8(d) respectively.

Effects of Overlapping Uplink/Downlink Frequency Bands

Note that in Section 9.7.3, the frequency band allocation follows the rule that only *partial overlap* between UL/DL frequency band is allowed to mitigate the inter-link interference, as shown in Rem. 9.2. Computation of the overlap factor is provided by Appendix D.3.1. Since the overlap factor is estimated based on the historical measurements, the actual utility λ derived using optimized $(\boldsymbol{p}, \boldsymbol{w})$ may not be as high as the computed λ in Algorithm 6. On the other hand, if *full overlap* is allowed (i.e., each transmission can be allocated to any of the RBs, regardless of whether it is in UL or DL), then, the overlap factor is one, and the utility achieved by Algorithm 6 can be much lower due to the strong inter-link interference.

In Fig. 9.9(a) we show the utility achieved by our proposed joint UL/DL optimization algorithm (represented by “Jo”), with the strategy of partial or full overlap. The three subplots from left to right illustrate the utility when the association policies “Best”, “DeUD_P”and “CoUD”are applied, respectively. Policy “Best” denotes the policy where the offset provides the maximum value of λ, i.e., $\pi^\star = \arg\max_{\pi\in\Pi} \lambda(\pi)$. For scenario of partial overlap, the blue dashed line expresses the optimized λ computed with our algorithm, while the green and red solid lines express the actual λ in UL and DL, respectively. Although the algorithm aims at achieving fair user-specific UL and DL utility, a small gap between the UL and DL utility can be observed due to the biased estimation of the overlap factor. For scenario of full overlap, the magenta solid line expresses the achieved λ for both UL and DL. Because the interference coupling model in (9.6) is accurate under the assumption of full overlap, there is no gap between the computed λ and the actual achievable λ.

Furthermore, we make the following observations. 1) Using optimized $(\boldsymbol{w}, \boldsymbol{p})$ based on estimated overlap factor, we can achieve the actual utility in DL only about $2\% - 3\%$ lower than the computed maximum feasible λ from the proposed algorithm, and in UL about $10\% - 30\%$ lower. 2) By regulating the frequency band allocated to UL and DL transmission with partial overlap, we achieve a $50\% - 100\%$ increase in utility than allowing the full overlap. 3) By enabling UL and DL decoupling, we can achieve a two-fold increase in the utility, compared to CoUD. Although DeUD_P may not be the best association policies, it still provides $60\% - 75\%$ increase. The same conclusion is reached by the analysis on association policies in Section 9.7.3.

Comparison against QoS-Based Proportional Fairness

We use the proportional fairness (PF) algorithm as a baseline for evaluating the utility benefits provided by our algorithm. To provide a fair comparison between the PF algorithm and our proposed algorithm, instead of the rate-based PF algorithm [NH06], we replace the rate with the metric of level of QoS satisfaction, i.e., $W_0 w_l r_l / d_l$ for link $l \in \overline{\mathcal{K}}$ presented in

(9.13). We run PF algorithm under default UL/DL bandwidth ratio under both association policies CoUD and DeUD_P, to compare with the proposed joint UL/DL optimization algorithm. The default UL/DL bandwidth ratio is set to be 9 : 16, i.e., out of 25 RBs, 9 of them are assigned for UL transmission while 16 for DL transmission.

Fig. 9.9(b) shows the performance comparison between our proposed algorithm and the PF algorithm under DeUD_P and CoUD. Conventional PF algorithm achieves fairness in UL and DL independently, and the fixed ratio of UL/DL bandwidth ratio causes a large gap between the achievable utility in UL and DL. Our proposed Algorithm 6 outperforms the PF algorithm, in the sense that it jointly optimizes the level of QoS satisfaction in UL and DL to the best closing levels. The utility in UL achieves three-fold increase than the PF algorithm in both DeUD_P and CoUD. We still observe a $20\% - 50\%$ increase in DL utility in DeUD_P, while in CoUD we sacrifice some DL utility to achieve a higher gain in UL. However, as more UEs are served in the system, even in CoUD we achieve better utility in both UL and DL than the QoS-based PF algorithm.

Another observation in reference to Fig. 9.9(b) is that, for both algorithms, by splitting the UL/DL access, the performance can be further improved by about $60\% - 70\%$. It is worth mentioning that the gain of UL/DL decoupling is not as high as expected in [BAE$^+$15, EBDI14a] (more than two-fold increase). Our explanation is that although the strength of the useful signal is increased by offloading more uplinks in small cells, the received signal strength of the interference may also be increased because the small cells are normally located on the cell edge. Therefore, it increases the need for the joint UL/DL optimization algorithm allowing flexible UL/DL bandwidth ratio, as we proposed in Algorithm 6.

9.8 Conclusion

We studied the utility maximization problem for the uplink and downlink decoupling-enabled HetNet, to jointly optimize the uplink and downlink bandwidth allocation and power control, under different association policies. The utility is modeled as the minimum level of the QoS satisfaction, to achieve fair service-centric performance. We develop a general model of inter-cell interference, that includes inter-link interference between uplink and downlink, with properties of power coupling and load coupling. Based on the interference model, we develop a three-step optimization algorithm using the fixed point approach for nonlinear operators with or without monotonicity. The algorithm benefits from the user-centric context-aware communication environment in 5G networks, adapts the bandwidth allocation and power spectral density according to the channel condition and traffic demand in both UL and DL, and achieves jointly optimized utility in both UL and DL. Numerical results show that the performance of our algorithm outperforms the QoS-based proportional fairness algorithm, and it is robust against heavily loaded system with high traffic demand.

FIGURES

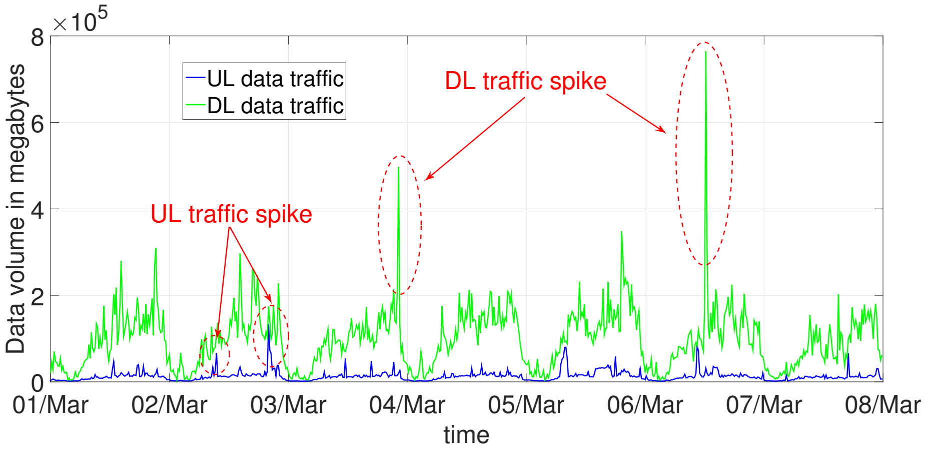


Figure 9.1: Time-varying UL and DL data traffic volume (aggregated every 15 minutes) for a week from Mar. 01 to Mar. 08, 2015 in a spatial grid in Rome, Italy. Data source from Telecom Italia's Big Data Challenge [Tel15].

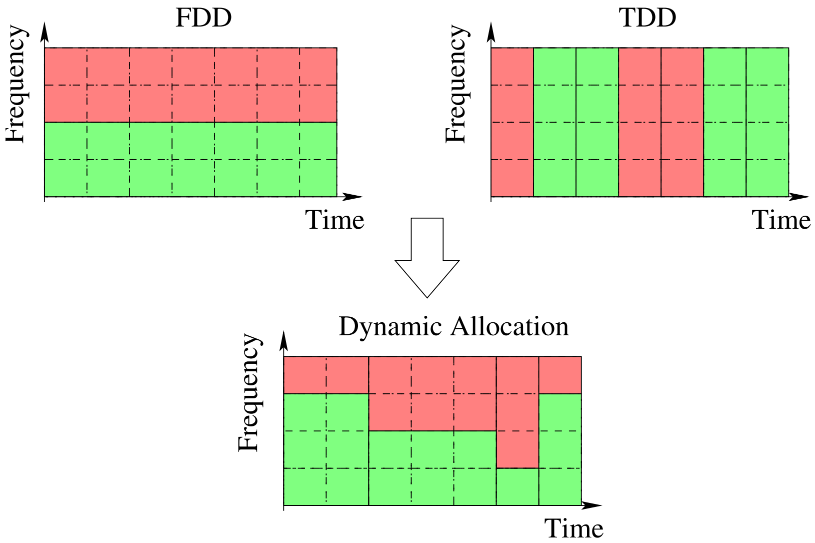


Figure 9.2: Difference between the traditional FDD (or TDD) technology and proposed dynamic UL/DL resource partitioning. The RBs assigned to UL is colored in red while to DL in green. The guard band and guard interval are not plotted.

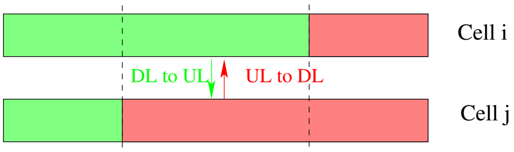


Figure 9.3: Inter-cell inter-link interference between UL (red) and DL (green). The guard band is not displayed.

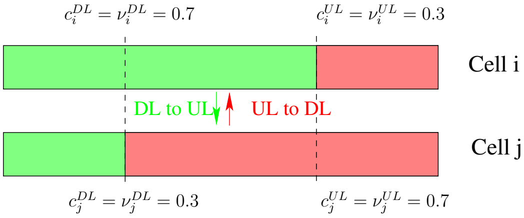

$c_i^{DL} = \nu_i^{DL} = 0.7$
$c_i^{UL} = \nu_i^{UL} = 0.3$
Cell i
DL to UL
UL to DL
Cell j
$c_j^{DL} = \nu_j^{DL} = 0.3$
$c_j^{UL} = \nu_j^{UL} = 0.7$

Figure 9.4: One possible approach to estimate the overlap factor based on the historical load measurements. The overlap factor between downlinks served by cell i and the uplinks served by cell j is computed by $c_i^{\text{DL}} c_j^{\text{UL}} = 0.49$, while the overlap factor between the uplinks served by cell i and the downlinks served by cell j is computed by $c_i^{\text{UL}} c_j^{\text{DL}} = 0.09$.

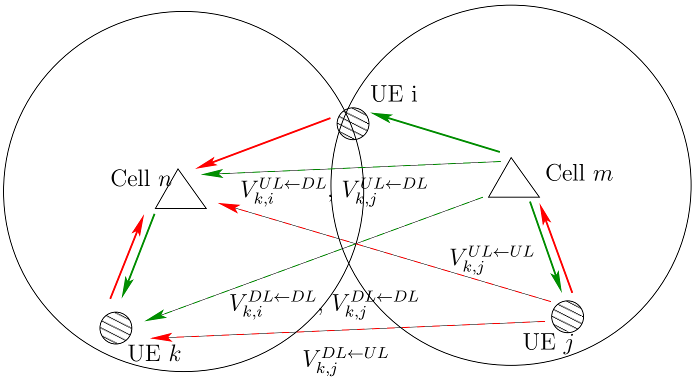

UE i
Cell n
Cell m
$V_{k,i}^{UL\leftarrow DL}, V_{k,j}^{UL\leftarrow DL}$
$V_{k,j}^{UL\leftarrow UL}$
$V_{k,i}^{DL\leftarrow DL}, V_{k,j}^{DL\leftarrow DL}$
UE k
UE j
$V_{k,j}^{DL\leftarrow UL}$

Figure 9.5: Inter-cell interference coupling on the per-user basis. UE i is associated to n in UL and to cell m in DL.

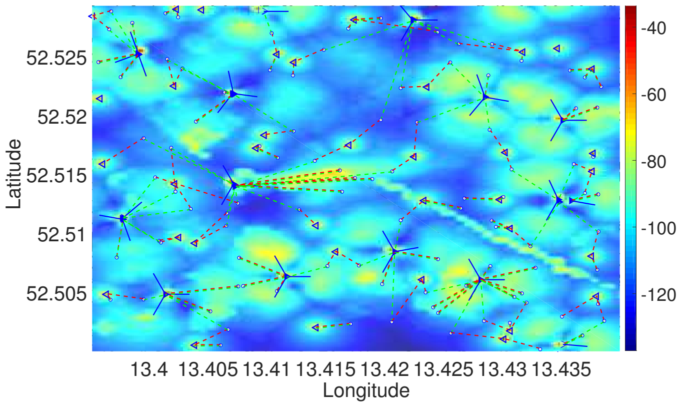
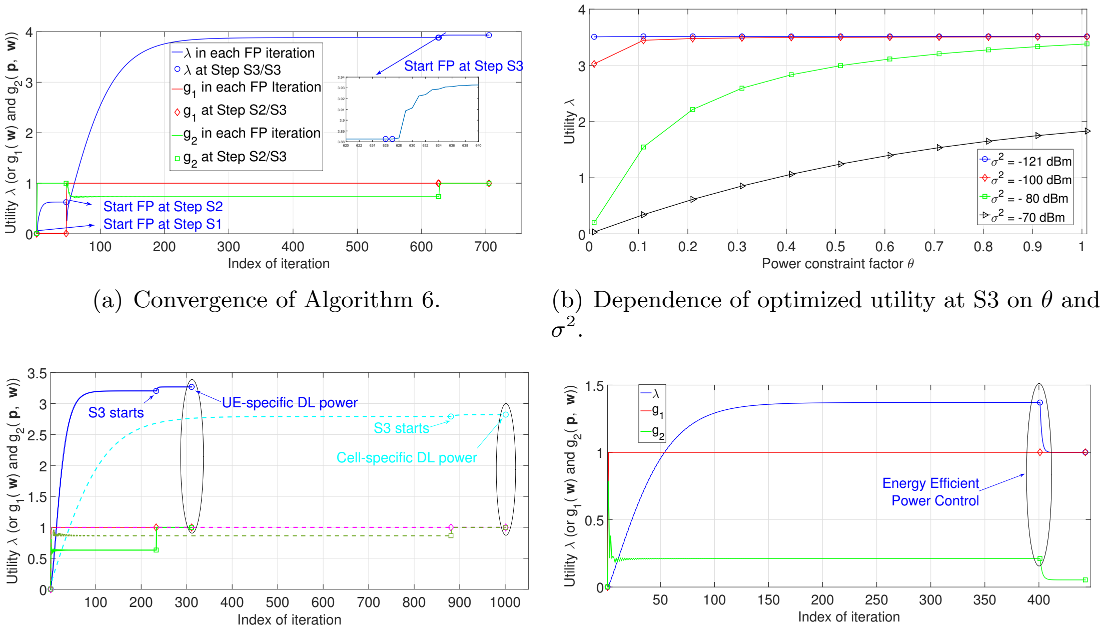


Figure 9.6: DeUD-enabled wireless network. Macro BSs - blue solid triangles; pico cells - blue hollow triangles; UEs - white circle with blue edge; downlink association - green dashed line; uplink association - red dashed line.



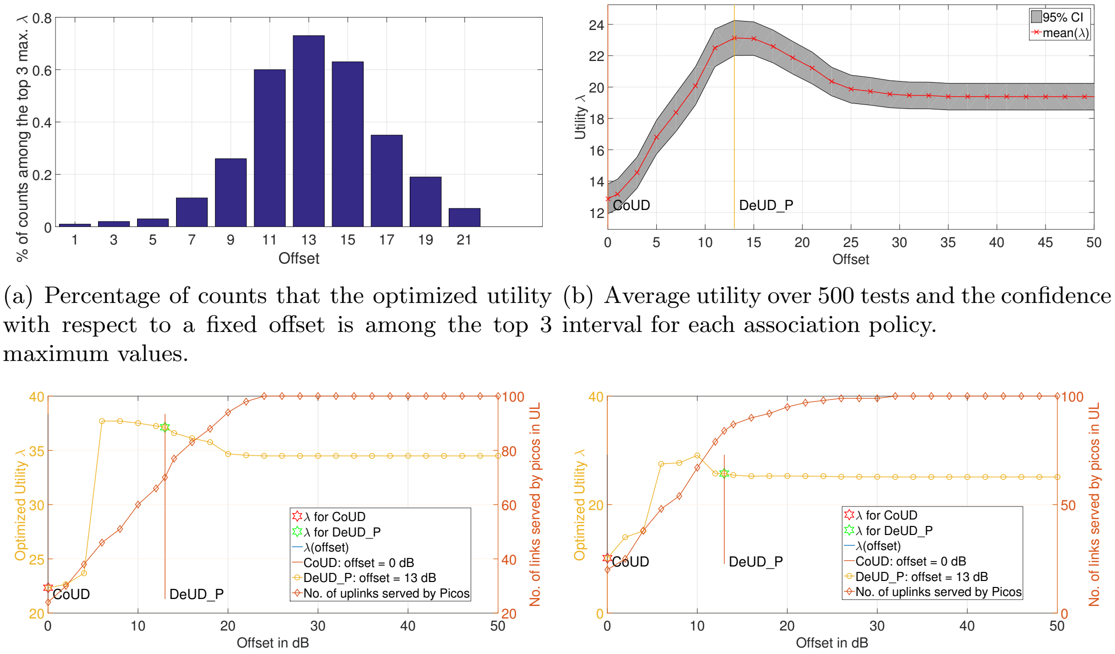
(a) Convergence of Algorithm 6.

(b) Dependence of optimized utility at S3 on θ and σ^2.

(c) Comparison between UE-specific power control and cell-specific power control in DL.

(d) Energy efficient power control.

Figure 9.7: Algorithm convergence ($K = 500$, DeUD_P).



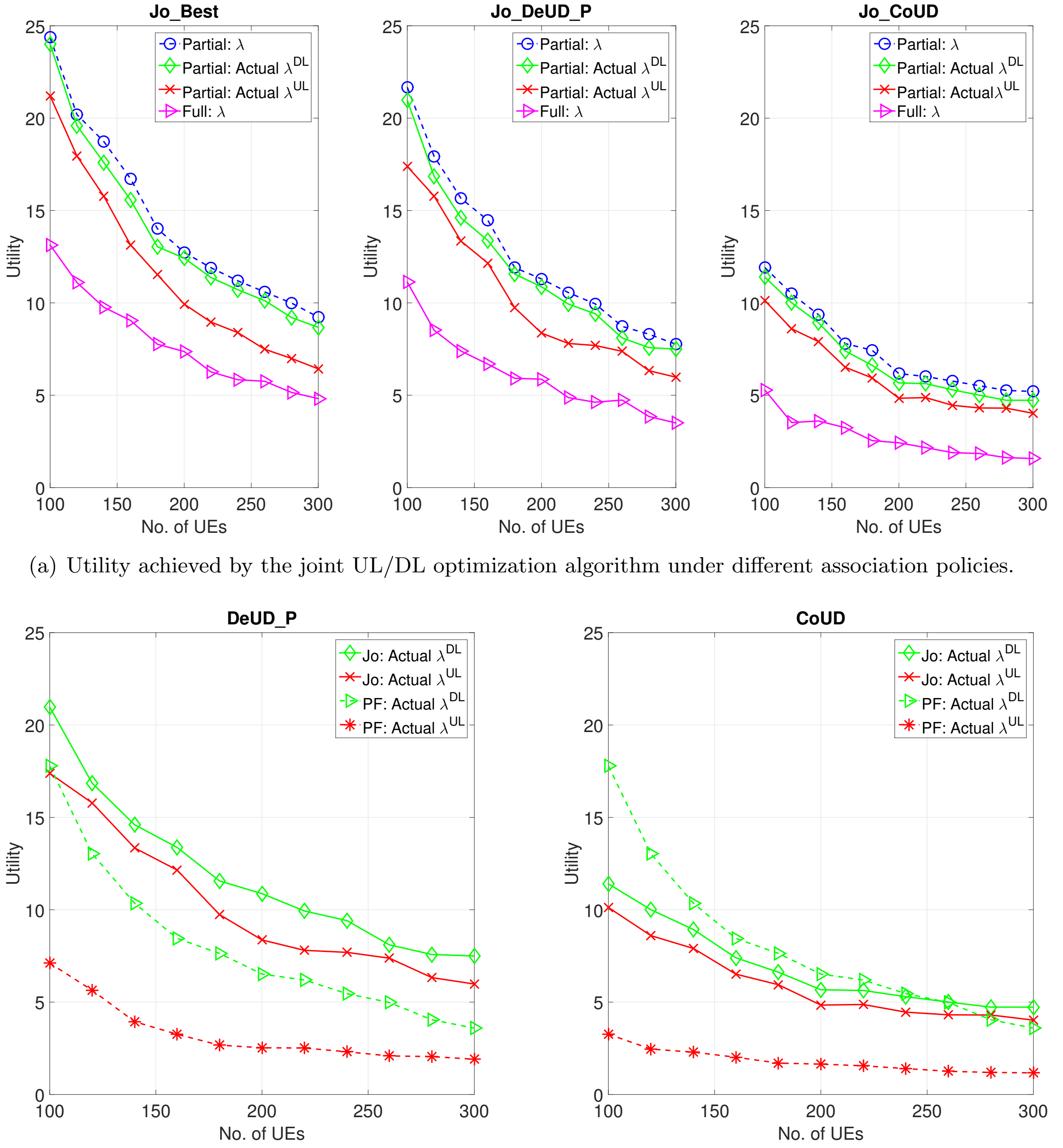
(a) Percentage of counts that the optimized utility with respect to a fixed offset is among the top 3 maximum values.

(b) Average utility over 500 tests and the confidence interval for each association policy.

(c) Example trial #1.

(d) Example trial #2.

Figure 9.8: Optimized utility depending on association policy ($K = 100$).



(a) Utility achieved by the joint UL/DL optimization algorithm under different association policies.

(b) Performance comparison between the joint UL/DL optimization algorithm and the QoS-based PF algorithm under different policies.

Figure 9.9: Performance evaluation of Algorithm 6.

Part V

Conclusion

Chapter 10

Conclusion and Future Studies

10.1 Summary

The main functionalities of SON include: *self-configuration*, *self-optimization* and *self-healing*. This thesis investigates multiple stages of self-organizing networks with respect to *self-healing* and *self-optimization* by introducing novel inference, anomaly detection and optimization techniques for the following functionalities:

- cognition, learning and detection for self-healing functions;
- context-aware statistical modeling and optimization for isolated SON functionalities;
- multi-objective optimization in high dimensional space for joint optimization of multiple SON functionalities.

The key to transform the SON paradigm from reactive to proactive is to exploit the knowledge of the network states extracted from the available data. In the first part of the thesis, we treated the problem of information extraction and model inference. Based on the collected network measurements, self-healing algorithms are developed for detecting two types of network anomalies. The first type of anomaly is usually caused by an unexpected operation fault that is a rare event such as cell outage. To detect the anomaly without a priori knowledge, we propose an information theory based anomaly detection algorithm, using the composite hypothesis testing technique. We develop an efficient discriminant function related to the *universal code* based on the modified Neyman-Pearson criterion, which can be shown to be asymptotically optimal. The second type of anomaly is usually caused by performance degradation, where a priori knowledge of the various classes of anomalies can be found by analyzing a large set of data collected from the network. A framework of proactive cell anomaly detection is proposed based on dimension reduction and fuzzy classification techniques. The dimension reduction is applied for visualization purpose and for the efficiency of the classification of high-dimensional data. The enhanced kernel-based semi-supervised FCM explores the complex pattern hidden in the unlabeled samples, while taking into account a priori knowledge contained in the labeled samples. The experimental results show that the proposed framework proactively detects network anomalies associated with various fault classes.

Based on the extracted knowledge, the system should self-adapt to dynamically changing environments (channel fading, mobility, load distribution, etc.). The second part of the thesis presents statistical modeling and optimization techniques that are used to develop robust algorithms against time-varying network environments and noisy feedback for isolated SON functionalities RACH optimization, MLBO, and MRO respectively. For RACH optimization, we suggest an algorithm for decentralized control of user back-off probabilities and transmission powers in random access communications. The algorithm is based on measurements and user reports at the base station side, which allows for an estimation of the number of users present within the cell, as well as the quantities of detection-miss and contention probability. By solving a drift minimization problem for the contention level and using closed loop updates for the transmission power level by an MIAD rule, the base station coordinates the actions chosen by the users, by broadcasting the information pair of contention level and power level. The algorithmic steps, as well as the methodology of the drift minimization for a certain measure of interest referring to the steady state, provide a general suggestion to treat problems of self-organization in wireless networks. For the use case of mobility robustness optimization, we exploit the framework of stochastic processes to develop a novel method of successively choosing a sequence of multi-variate training points for multi-objective optimization that involves a set of non-convex contradicting objective functions depending on multiple variables such as HO parameters and user mobility classes. The unknown functions can be explored at selected training points by taking measurements (called trials). The training points can be corrupted by some Gaussian noise due to the missing or delayed measurements. The maximum allowable number of trials is strongly restricted, because each trail results in a relative high cost, for instance, in terms of wireless resources. We therefore consider an extension of the so-called P-algorithm by Kushner and Žilinskas for single-objective global optimization. Using the framework of multi-variate GP, we extend the method of P-algorithm with single objective to incorporate the inter-dependencies between multiple objectives of HO performance measures. The algorithm provides optimized local and global HO parameters per user mobility class, and achieves reduced number of HO-related radio link failures and number of unnecessary or missed handovers caused by incorrect HO decisions. The collected local statistics and a priori knowledge are utilized to improve the efficiency of the algorithm. To achieve the mobility load balancing, together with inter-cell interference mitigation, we propose a mixed integer optimization problem solved using Lagrangian – but not Linear Programming – relaxation, which allows the solution to be binary for the user assignment variables. Several properties of the optimal Lagrangian solution are derived, which depend on the value of a load price and interference cost per BS. The implementation of the algorithm is based on exchange of certain prices among base stations and allows each of them to make choices individually without the aid of a central controller. The cell HO parameters are further adequately adjusted to enforce cell-edge users to migrate to their optimal BS.

After solving problems for individual SON use cases, the next challenge is to ensure the efficient and robust network operation by a joint optimization of multiple interacting or conflicting SON use cases. Last but not least, the problems of multi-objective optimization over a high dimensional action space are tackled in the final part of the thesis. In this part, we mainly focus on the fixed point theory-based approach, as it is a powerful tool to prove the existence and to determine uniqueness of solutions to dynamical multi-agent systems. We first study on the problem of joint optimization of coverage, capacity and load balancing. A robust algorithmic framework is built on a utility model, which enables fast and optimal uplink solutions and sub-optimal downlink solutions by exploiting three properties: *a)* the monotonic property of standard interference functions, *b)* decoupled property of the antenna tilt and BS assignment optimization in the uplink network, and *c)* uplink-downlink duality. The first property allows obtaining the global optimal solution with fixed-point iteration for two specific problems: utility-constrained power minimization and power-constrained max-min utility balancing. The second and third properties enable decomposition of the high-dimensional optimization problem, such as the joint beamforming and power control. Based on the three properties, we propose a max-min utility balancing algorithm for capacity-coverage trade-off over a joint space of antenna tilts, BS assignments and power in uplink. Then, to include the downlink, we analyze the uplink-downlink duality by using the Perron-Frobenius theory. Utilizing optimized variables in the dual uplink allows us to decompose the high-dimensional optimization problem and to obtain an efficient sub-optimal solution for downlink. A further step is to jointly optimize uplink and downlink performance with joint uplink and downlink resource allocation and power control. Due to the time- and spatial-dependent service requirements and traffic patterns, it is expected to have time-varying asymmetric traffic load in both uplink and downlink in different cells. Apart from dynamic uplink/downlink resource splitting, flexible uplink/downlink traffic distribution among the cells with different transmission ranges is also crucial for improvement of joint uplink/downlink performance. One way to enable the flexible uplink/downlink traffic distribution is to allow the user terminal to be associated to two different radio access nodes in uplink and donwlink, respectively – so called DUDe. Such a DUDe access has the potential benefits including improvement of performance in uplink (without degradation of performance in downlink), reduction of energy consumption in mobile terminal, and network load balancing. We introduce a general model of inter-cell interference for joint uplink/downlink system, which includes the inter-link interference between uplink and downlink and is both power and load coupling-aware. We then develop a framework involving a fixed-point class with nonlinear contraction operators, *with or without monotonicity*, and an optimizer for the utility of QoS satisfaction level, subjected to a general class of resource (in both frequency and power domain) constraints. A three-step optimization algorithm is proposed, to find the local optimum of the joint variables bandwidth allocation and power spectral density on a per-link basis, corresponding to the different link association policies. The algorithm benefits from the user-specific context-aware communication environment in 5G networks, adapts the bandwidth allocation and power spectral density according to the channel condition and traffic demand in both uplink and downlink, and achieves jointly optimized utility in both uplink and downlink.

10.2 Future Research

The results presented in this thesis have demonstrated the effectiveness of our proposed learning, detection and optimization algorithms. However, we would like to point out open problems and research directions that are related to or result from the presented research.

The actual network will provide a critical role in providing the almost-real-time access to data from a multitude of sensors and a augmented intelligence tools running on a massive distributed set of muliti-dimensional resources. As the cost the data sets tends to decrease, the hyperbole of the big data phenomenon will transition into new, small data applications that provide real knowledge. As stated in [Wel16], *big data will become "small"*. How to extract "just enough" data to make an informed and proper decision remains an open question.

How to deal with error in modeling is another challenge. The limitation of deriving accurate model is based on mathematical and statistical fact: the introduction of noise increases the number of required observation samples for a reliable model. Further more, what is more important is the decision making about the future based on the predictive model. How to further utilize the predictive models obtained by self-healing to improve the proactive anticipatory self-organizing networks attracts our attention. In the presented framework, the inferred predictive models are used for proactively detecting the abnormal network states to trigger the self-optimization functions. Introducing the predicted network conditions and the KPIs into the optimization framework may enhance the performance of self optimization.

Last but not least, the concept of 5G networks enables new potential technologies and a set of new configuration control parameters such as adaptive waveforms, scalable TTI and numerologies, and flexible duplex. The service-centric requirements of the network define the new KPIs such as reliability, security and extreme low latency. Formulating the new objective functions under more dynamic and flexible network conditions brings numerous challenges into the future self-organizing networks.

Appendices

Appendix A

Some Concepts and Results from Matrix Analysis

A.1 Scalars, Vectors and Matrices

Throughput the dissertation, vectors and matrices are defined over the field of real numbers $\mathbb{R}$, unless something otherwise stated. Elements of $\mathbb{R}$ are called scalars. We use $\mathbb{R}_+$ and $\mathbb{R}_{++}$ to denote the set of nonnegative and positive reals, respectively. We denote the scalars with italic lower case letter, vectors with boldface lowercase letter, and matrix with boldface uppercase letters. For example, x, $\boldsymbol{x}$ and $\boldsymbol{X}$ denote a scalar, a vector and a matrix, respectively. For any $\boldsymbol{x} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, the notation $\boldsymbol{x} + c$ is used throughout the thesis to denote $\boldsymbol{x} + (c, \ldots, c)$, where $(c, \ldots, c) \in \mathbb{R}^n$. Similar convention is also used for matrices.

The Euclidean n-space denoted by $\mathbb{R}^n$ is a n-dimensional vector space over the field $\mathbb{R}$. For two (column) vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the partial ordering on $\mathbb{R}^n$ is defined as follows:

$$\begin{aligned}
&\boldsymbol{x} \geq \boldsymbol{y} \Leftrightarrow \forall_{1 \geq i \geq n}\ x_i \geq y_i, \quad \boldsymbol{x} > \boldsymbol{y} \Leftrightarrow \forall_{1 \geq i \geq n}\ x_i > y_i, \\
&\boldsymbol{x} = \boldsymbol{y} \Leftrightarrow \forall_{1 \geq i \geq n}\ x_i = y_i, \quad \boldsymbol{x} \gneq \boldsymbol{y} \Leftrightarrow \forall_{1 \geq i \geq n}\ x_i \geq y_i \text{ and } \boldsymbol{x} \neq \boldsymbol{y}.
\end{aligned}$$

All the norms used in this dissertation are l^p-norms and the maximum norm. For any $p \leq 1$, the l^p-norm and the maximum norm of $\boldsymbol{x} \in \mathbb{R}^n$, denoted by $\|\boldsymbol{x}\|_p$ and $\|\boldsymbol{x}\|_\infty$ respectively, are defined to be

$$\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^{n} |\boldsymbol{x}_i|^p \right)^{\frac{1}{p}} \text{ and } \|\boldsymbol{x}\|_\infty := \max(|x_1|, \ldots, |x_n|). \tag{A.1}$$

respectively.

A $n \times m$ matrix is denoted by $\boldsymbol{X} := (x_{i,j})_{1 \leq i \leq n, q \leq j \leq m}$ or simply $\boldsymbol{X}$:=(x_{ij}). The entries of $\boldsymbol{X}$ are denoted as $(\boldsymbol{X})_{ij}$. The $n \times n$ diagonal matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{X}$:=diag($\boldsymbol{x}$):= diag$(x_1, \ldots, x_n)$. The diagonal of a matrix $\boldsymbol{X}$ is denoted by diag X. In particular, $\boldsymbol{I}$:=

$\operatorname{diag}(\mathbf{1}) = \operatorname{diag}(1, \ldots, 1)$ denotes the identity matrix. A block diagonal matrix has the form

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{X}_n \end{bmatrix}.$$

We denote the transpose of matrix $\boldsymbol{X}$ by $\boldsymbol{X}^T$. Consider a $n \times n$ square matrix $\boldsymbol{X}$, we denote the trace of matrix $\boldsymbol{X}$ by $\operatorname{Tr}(\boldsymbol{X}) := \sum_{i=1}^n x_{i,i}$, the inverse of the matrix by $\boldsymbol{X}^{-1}$ if it exists, the determinant of $\boldsymbol{X}$ by $|\boldsymbol{X}|$. For any two matrix $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times m}$, the Hadamard product $\boldsymbol{X} \circ \boldsymbol{Y}$ is the entry-wise product of matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$. For ant two matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{Y} \in \mathbb{R}^{i\times j}$, the Kronecker product of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is denoted by $\boldsymbol{X} \otimes \boldsymbol{Y}$.

Given a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, a matrix norm of $\boldsymbol{X}$ is denoted by $\|\boldsymbol{X}\|$. General matrix norm satisfies (A.1), with the vector $\boldsymbol{x}$ replaced by some matrix. Additionally, if $\boldsymbol{XY}$ exists, we have

$$\|\boldsymbol{XY}\| \leq \|\boldsymbol{X}\|\|\boldsymbol{Y}\|.$$

The Frobenius norm of matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ is given by

$$\|\boldsymbol{X}\|_F^2 := \sum_{i,j} |x_{i,j}|^2 = \operatorname{Tr}(\boldsymbol{X}^T\boldsymbol{X}). \tag{A.2}$$

Lemma A.1 (Matrix Inversion Lemma)**.** *The matrix inversion lemma, also known as the Woodbury formula [PTVF96, p. 75], is given by*

$$(\boldsymbol{Z} + \boldsymbol{UWV})^{-1} = \boldsymbol{Z}^{-1} - \boldsymbol{Z}^{-1}\boldsymbol{U}(\boldsymbol{W}^{-1} + \boldsymbol{V}^T\boldsymbol{Z}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{Z}^{-1} \tag{A.3}$$

assuming the relevant inverse all exist. Here $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$, $\boldsymbol{W} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times m}$.

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{R} & \boldsymbol{S} \end{bmatrix}, \quad \boldsymbol{A}^{-1} = \begin{bmatrix} \tilde{\boldsymbol{P}} & \tilde{\boldsymbol{Q}} \\ \tilde{\boldsymbol{R}} & \tilde{\boldsymbol{S}} \end{bmatrix}, \tag{A.4}$$

where $\boldsymbol{P}, \tilde{\boldsymbol{P}} \in \mathbb{R}^{n_1\times n_1}$ and $\boldsymbol{S}, \tilde{\boldsymbol{S}} \in \mathbb{R}^{n_2\times n_2}$, with $n = n_1 + n_2$. The submatrices of $\boldsymbol{A}^{-1}$ are found by either the formulas [PTVF96, p. 77]

$$\left.\begin{aligned} \tilde{\boldsymbol{P}} &= \boldsymbol{P}^{-1} + \boldsymbol{P}^{-1}\boldsymbol{QMRP}^{-1} \\ \tilde{\boldsymbol{Q}} &= -\boldsymbol{P}^{-1}\boldsymbol{QM} \\ \tilde{\boldsymbol{R}} &= -\boldsymbol{MRP}^{-1} \\ \tilde{\boldsymbol{S}} &= \boldsymbol{M} \end{aligned}\right\} \text{where } \boldsymbol{M} = (\boldsymbol{S} - \boldsymbol{RP}^{-1}\boldsymbol{Q})^{-1}$$

or equivalently

$$\left.\begin{aligned} \tilde{\boldsymbol{P}} &= \boldsymbol{N} \\ \tilde{\boldsymbol{Q}} &= -\boldsymbol{NQS}^{-1} \\ \tilde{\boldsymbol{R}} &= -\boldsymbol{S}^{-1}\boldsymbol{RN} \\ \tilde{mS} &= \boldsymbol{S}^{-1} + \boldsymbol{S}^{-1}\boldsymbol{RNQS}^{-1} \end{aligned}\right\} \text{where } \boldsymbol{N} = (\boldsymbol{P} - \boldsymbol{QS}^{-1}\boldsymbol{R})^{-1}$$

A.2 Matrix Spectrum and Spectral Radius

Definition A.1 (Matrix Spectrum)**.** *The set of distinct eigenvalues of $\boldsymbol{X}$ is referred to as the spectrum of $\boldsymbol{X}$ and is denoted by $\sigma(\boldsymbol{X})$.*

Since the root s of a polynomial with real coefficients occur in conjugate pairs, $\lambda \in \sigma(\boldsymbol{X})$ implies that $\bar{\lambda} \in \sigma(\boldsymbol{X})$ where $\bar{x}$ denotes the conjugate complex. Furthermore, we have [Mey00, p. 498]

$$\sigma(\boldsymbol{X}) = \sigma(\boldsymbol{X}^T) \tag{A.5}$$

Definition A.2 (Spectral Radius)**.** *For any square matrix $\boldsymbol{X} \in \mathbb{R}^n \times n$, we define $\rho : \mathbb{R}^{n\times n} \to \mathbb{R}$ as*

$$\rho(\boldsymbol{X}) := \max\{\|\lambda\| : \lambda \in \sigma(\boldsymbol{X})\}. \tag{A.6}$$

The real number $\rho(\boldsymbol{X})$ is called the spectral radius of $\boldsymbol{X}$.

If $\|\cdot\|$ is any matrix norm, then $\rho(\boldsymbol{X}) = \lim_{k\to\infty} \|\boldsymbol{X}^k\|^{1/k}$. A rather crude (but cheap) upper bound on $\rho(\boldsymbol{X})$ is obtained by observing that $\rho(\boldsymbol{X}) \leq \|\boldsymbol{X}\|$ for every matrix norm [Mey00, p. 497].

Theorem A.1 ([SWB09, p. 355])**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be arbitrary. Then, the following statements are equivalent.*

(i) $\sum_{k=0}^{\infty} \boldsymbol{X}^k$ *converges.*

(ii) $\rho(\boldsymbol{X}) < 1$.

(iii) $\lim_{k\to\infty} \boldsymbol{X}^k = 0$.

In these cases, $(\boldsymbol{I} - \boldsymbol{X})^{-1}$ exists, and $(\boldsymbol{I} - \boldsymbol{X})^{-1} = \sum_{k=0}^{\infty} \boldsymbol{X}^k$.

A.3 Perron-Frobenius Theory of Nonnegative Matrices

Definition A.3 (Nonnegative matrix)**.** *Any square matrix $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{n\times n}$ with $x_{ij} \in \mathbb{R}_+$ for $1 \leq i, j \leq n$ (or denoted by $\boldsymbol{X} \geq 0$) is called a nonnegative matrix. If $x_{ij} \in \mathbb{R}_{++}$ for $1 \leq i, j \leq n$ holds, then $\boldsymbol{X}$ is called a positive matrix.*

Definition A.4 (Irreducible matrix)**.** *The graph of $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, denoted by $\mathcal{G}(\boldsymbol{X})$, is the* **direct graph** *of the nodes $\{N_1, \ldots, N_n\}$ in which there is a directed edge leading from N_i to N_j if and only if $x_{ij} \neq 0$. Graph $\mathcal{G}(\boldsymbol{X})$ is* **strongly connected** *if for each pair of nodes (N_i, N_k), there is a sequence of directed edges leading from N_i to N_k. The matrix $\boldsymbol{X}$ is said to be* **reducible** *if there exists a permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{P} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{C} \end{pmatrix}$,* *where* $\boldsymbol{A}$ *and* $\boldsymbol{C}$ *are both square matrices, and* $\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{P}$ *is the* **symmetric permutation** *of* $\boldsymbol{X}$*. Otherwise,* $\boldsymbol{X}$ *is said to be* **irreducible**. $\mathcal{G}(\boldsymbol{X})$ *is strongly connected if and only if* $\boldsymbol{X}$ *is irreducible.*

Theorem A.2 (Perron's Theorem of Positive Matrices [Mey00, p. 667])**.** *If* $\boldsymbol{X}_{n\times n} > 0$ *with* $r = \rho(\boldsymbol{X})$*, then the following statements are true.*

(i) $r > 0$.

(ii) $r \in \sigma(\boldsymbol{X})$ *(r is called the* **Perron root***).*

(iii) $\operatorname{alg\,mult}_{\boldsymbol{X}}(r) = 1$*, where* $\operatorname{alg\,mult}_{\boldsymbol{X}}(r)$*, denoting the* **algebraic multiplicities** *of* r*, is the number of times* r *is repeated as a root of the characteristic polynomial.*

(iv) There exists an eigenvector $\boldsymbol{p} > 0$ *such that* $\boldsymbol{X}\boldsymbol{p} = r\boldsymbol{p}$.

(v) The **Perron vector** *is the unique vector defined by*

$$\boldsymbol{X}\hat{\boldsymbol{p}} = r\hat{\boldsymbol{p}}, \boldsymbol{p} > 0, \text{ and } \|\hat{\boldsymbol{p}}\|_1 = 1, \tag{A.7}$$

and, except for positive multiples of $\hat{\boldsymbol{p}}$*, there are no other nonnegative eigenvectors for* $\boldsymbol{X}$*, regardless of the eigenvalue.*

(vi) r *is the only eigenvalue on the spectral circle of* $\boldsymbol{X}$.

(vii) $r = \max_{\boldsymbol{p}\in\mathcal{N}} f(\boldsymbol{p})$ *(***Collatz–Wielandt formula***), where*

$$f(\boldsymbol{p}) := \min_{\substack{1\le i\le n \\ p_i \neq 0}} \frac{(\boldsymbol{X}\boldsymbol{p})_i}{p_i} \text{ and } \mathcal{N} := \{\boldsymbol{p} | \boldsymbol{p} \geq \boldsymbol{0} \text{ with } \boldsymbol{p} \neq \boldsymbol{0}\}. \tag{A.8}$$

Theorem A.3 (Perron-Frobenius Theorem of Nonnegative Matrices [Mey00, p. 673])**.** *If* $\boldsymbol{X}_{n\times n} \geq 0$ *is irreducible with* $r = \rho(\boldsymbol{X})$*, then the following statements are true.*

(i) $r \in \sigma(\boldsymbol{X})$ *and* $r > 0$.

(ii) $\operatorname{alg\,mult}_{\boldsymbol{X}}(r) = 1$

(iii) There exists an eigenvector $\boldsymbol{p} > 0$ *such that* $\boldsymbol{X}\boldsymbol{p} = r\boldsymbol{p}$.

(iv) The **Perron vector** *is the unique vector defined by*

$$\boldsymbol{X}\hat{\boldsymbol{p}} = r\hat{\boldsymbol{p}}, \boldsymbol{p} > 0, \text{ and } \|\hat{\boldsymbol{p}}\|_1 = 1,$$

and, except for positive multiples of $\hat{\boldsymbol{p}}$*, there are no other nonnegative eigenvectors for* $\boldsymbol{X}$*, regardless of the eigenvalue.*

(v) The **Collatz–Wielandt formula** $r = \max_{\boldsymbol{p}\in\mathcal{N}} f(\boldsymbol{p})$*, where*

$$f(\boldsymbol{p}) := \min_{\substack{1\le i\le n\\ p_i\neq 0}} \frac{(\boldsymbol{X}\boldsymbol{p})_i}{p_i} \text{ and } \mathcal{N} := \{\boldsymbol{p}|\boldsymbol{p}\geq \boldsymbol{0} \text{ with } \boldsymbol{p}\neq \boldsymbol{0}\}\,.$$

Theorem A.3 shows how adding irreducibility to nonnegativity recovers most of the Perron properties in Theorem A.2. The only property in Theorem A.2 that irreducibility is not able to salvage is (vi), which states that there is only one eigenvalue on the spectral circle. The property of having (or not having)only one eigenvalue on the spectral circle divides the set of nonnegative irreducible matrices into two important classes: primitive matrices and imprimitive matrices, as defined as follows.

Theorem A.4 ([SWB09, p. 371])**.** *. Let* $\boldsymbol{X}_{n\times n} \geq 0$ *be arbitrary, and let* $\alpha > 0$ *be any scalar. A necessary and sufficient condition for a solution* $\boldsymbol{p} \gneq 0$*, to*

$$(\alpha\boldsymbol{I} - \boldsymbol{X})\boldsymbol{p} = \boldsymbol{b} \tag{A.9}$$

to exist for any $\boldsymbol{b} > 0$ *is that* $\alpha > r = \rho(\boldsymbol{X})$*. In this case, there is only one solution* $\boldsymbol{p}$*, which is strictly positive and given by* $\boldsymbol{p} = (\alpha\boldsymbol{I} - \boldsymbol{X})^{-1}\boldsymbol{b}$*.*

A.3.1 Proof of Proposition 8.1

For any fixed BS assignment $\hat{\boldsymbol{b}}$, denote $\hat{\boldsymbol{W}} := \boldsymbol{W}_{\hat{\boldsymbol{b}}}$ and $\hat{\boldsymbol{V}} := \tilde{\boldsymbol{V}}_{\hat{\boldsymbol{b}}}$ for convenience, the optimal downlink power solution $\hat{\boldsymbol{q}}^{\mathrm{DL}}$ for problem (8.30) satisfies [SWB09]

$$\boldsymbol{\Lambda}^{\mathrm{DL}}\hat{\boldsymbol{q}}^{\mathrm{DL}} = \frac{1}{C^{\mathrm{DL}}(\hat{\boldsymbol{b}}, P^{\max})}\hat{\boldsymbol{q}}^{\mathrm{DL}}, \hat{\boldsymbol{q}}^{\mathrm{DL}} \in \mathbb{R}_{+}^{C} \tag{A.10}$$

where $\boldsymbol{\Lambda}^{\mathrm{DL}} \in \mathbb{R}_{+}^{C\times C}$ is defined as

$$\boldsymbol{\Lambda}^{\mathrm{DL}} := \boldsymbol{\Gamma\Psi}\left[\boldsymbol{A}\hat{\boldsymbol{V}}^{T}\boldsymbol{A}_{\boldsymbol{\alpha}}^{T} + \frac{1}{P^{\max}}\boldsymbol{z}^{\mathrm{DL}}\boldsymbol{1}_{C}^{T}\right]. \tag{A.11}$$

we denote $\boldsymbol{\Gamma} := \mathrm{diag}\{\gamma_1, \ldots, \gamma_C\}$, $C^{\mathrm{DL}}(\hat{\boldsymbol{b}}, P^{\max}) = \max_{\boldsymbol{q}\geq 0}\min_c U_c^{(\mathrm{d},1)}/\gamma_c$ subject to $\|\boldsymbol{q}\|_1 \leq P^{\max}$, and $\boldsymbol{1}_C$ is a C-dimensional all-one vector. (A.10) and (A.11) are derived by writing the utility fairness $U_c^{(\mathrm{d},1)}/\gamma_c = C^{\mathrm{DL}}(\hat{\boldsymbol{b}}, P^{\max})$ for all $c \in \mathcal{C}$ and the power constraint $\|\boldsymbol{q}^{\mathrm{DL}}\|_1 = P^{\max}$ with matrix notation. Targets $\boldsymbol{\gamma}$ is feasible if and only if $C^{\mathrm{DL}}(\hat{\boldsymbol{b}}, P^{\max}) > 1$.

Similarly, the optimal uplink power solution $\hat{\boldsymbol{q}}^{\mathrm{UL}}$ for uplink problem (8.31) needs to satisfy

$$\boldsymbol{\Lambda}^{\mathrm{UL}}\hat{\boldsymbol{q}}^{\mathrm{UL}} = \frac{1}{C^{\mathrm{UL}}(\hat{\boldsymbol{b}}, P^{\max})}\hat{\boldsymbol{q}}^{\mathrm{UL}}, \hat{\boldsymbol{q}}^{\mathrm{UL}} \in \mathbb{R}_{+}^{C} \tag{A.12}$$

where $\boldsymbol{\Lambda}^{\mathrm{UL}} \in \mathbb{R}_{+}^{C\times C}$ is defined as

$$\boldsymbol{\Lambda}^{\mathrm{UL}} := \boldsymbol{\Gamma\Psi}\left[\boldsymbol{A}\hat{\boldsymbol{W}}\boldsymbol{A}_{\boldsymbol{\alpha}}^{T} + \frac{1}{P^{\max}}\boldsymbol{z}^{\mathrm{UL}}\boldsymbol{1}_{C}^{T}\right]. \tag{A.13}$$

where $\boldsymbol{z}^{\mathrm{UL}} := \boldsymbol{A}\boldsymbol{\sigma}^{\mathrm{UL}}$, i.e., $z_c^{\mathrm{UL}} = \Sigma_{\mathrm{tot}}/C$ for all $c \in \mathcal{C}$.

The balanced level $C^{\mathrm{DL}}(\hat{\boldsymbol{b}}, P^{\max})$ and $C^{\mathrm{UL}}(\hat{\boldsymbol{b}}, P^{\max})$ are the reciprocal spectral radius of the nonnegative extended coupling matrix $\boldsymbol{\Lambda}^{\mathrm{DL}}$ and $\boldsymbol{\Lambda}^{\mathrm{UL}}$. Moreover, according to Perron-Frobenius theorem, if both $\boldsymbol{\Lambda}^{\mathrm{DL}}$ and $\boldsymbol{\Lambda}^{\mathrm{UL}}$ are irreducible, they have unique real spectral radius and their corresponding eigenvectors (power allocation) have strictly positive components. By comparing the interference terms in (A.11) and (A.13), we have $(\boldsymbol{A}\hat{\boldsymbol{V}}^T\boldsymbol{A}_{\boldsymbol{\alpha}}^T)^T = \boldsymbol{A}_{\boldsymbol{\alpha}}\hat{\boldsymbol{V}}\boldsymbol{A}^T = \boldsymbol{A}\,\mathrm{diag}\{\boldsymbol{\alpha}\}\hat{\boldsymbol{V}}\boldsymbol{I}\boldsymbol{A}^T = \boldsymbol{A}\,\mathrm{diag}\{\boldsymbol{\alpha}\}\hat{\boldsymbol{V}}\,\mathrm{diag}^{-1}\{\boldsymbol{\alpha}\}\,\mathrm{diag}\{\boldsymbol{\alpha}\}\boldsymbol{A}^T = \boldsymbol{A}\hat{\boldsymbol{W}}^T\boldsymbol{A}_{\boldsymbol{\alpha}}^T$. By comparing the noise terms we have $\boldsymbol{z}^{\mathrm{UL}} = \frac{1}{C}\mathbf{1}_C{\boldsymbol{z}^{\mathrm{DL}}}^T\mathbf{1}_C$ (by using $z_c^{\mathrm{UL}} = \Sigma_{\mathrm{tot}}/C$ for all $c \in \mathcal{C}$), thus $\boldsymbol{z}^{\mathrm{UL}}\mathbf{1}_C^T = \frac{1}{C}\mathbf{1}_C{\boldsymbol{z}^{\mathrm{DL}}}^T\mathbf{1}_C\mathbf{1}_C^T = \mathbf{1}_C{\boldsymbol{z}^{\mathrm{DL}}}^T = (\boldsymbol{z}^{\mathrm{DL}}\mathbf{1}_C^T)^T$. By using the properties of spectral radius $\rho(\boldsymbol{X}) = \rho(\boldsymbol{X}^T)$ and $\rho(\boldsymbol{X}\boldsymbol{Y}) = \rho(\boldsymbol{Y}\boldsymbol{X})$ we have that $\rho(\boldsymbol{\Lambda}^{\mathrm{DL}}) = \rho(\boldsymbol{\Lambda}^{\mathrm{UL}})$ and thus $C^{\mathrm{DL}}(\hat{\boldsymbol{b}}, P^{\max}) = C^{\mathrm{UL}}(\hat{\boldsymbol{b}}, P^{\max})$. Notice that the network duality holds for any given BS assignment $\hat{\boldsymbol{b}}$, the achievable utility regions are the same for both the downlink problem (8.30) and uplink problem (8.31).

Appendix B

Some Concepts and Results from Markov Problem Solution

In this chapter, we show how the solution of the drift minimization problem is related to the solution of an ideal Markov Decision Problem for optimal performance in the steady-state in Section 5.3.

We begin by considering an ideal setting, meaning that all expressions are known and the system is fully controllable by the choice of actions. Let $V\left(\mathbf{S}\left(t\right)\right)$ be a non-negative function of the system state and let $\mathcal{M}\left(V,\tilde{\mathbf{A}}\right)$ be a performance metric related to the steady state reached when $t\rightarrow\infty$, if the initial state is $\mathbf{S}\left(0\right)$. The metric is a function of the entire set of actions $\tilde{\mathbf{A}}$

$$\mathcal{M}\left(V,\tilde{\mathbf{A}}\right) \;:=\; \lim_{t\rightarrow\infty}\mathbb{E}\left[V\left(\mathbf{S}\left(t\right)\right)|\mathbf{S}\left(0\right)\right]. \tag{B.1}$$

If the actions are chosen per time-slot t from the set $\mathbf{A}\left(t\right)$, the following general MDP can be posed:

$$\begin{array}{ll}\mathbf{min} & \mathcal{M}\left(V,\tilde{\mathbf{A}}\right)\\ \mathbf{s.t.} & \mathbf{A}\left(t\right)\in\mathbb{A},\, t=0,1,\ldots\end{array} \tag{B.2}$$

B.1 Relationship between Solution of Markov Decision Problem and Solution of Drift Minimization Problem

Proposition B.1. *The MDP in (B.2) can be solved using the dynamic programming tools. The optimal solution satisfies Bellman's equation [Put05]*

$$J\left(\mathbf{S}\right) \;=\; \min_{\mathbf{A}\in\mathbb{A}}\left\{D\left(V\left(\mathbf{S}\right),\mathbf{A}\right)+\sum_{\mathbf{S}'\in\mathcal{S}}p_{s\rightarrow s'}J\left(\mathbf{S}'\right)\right\},\;\forall\mathbf{S}\in\mathcal{S} \tag{B.3}$$

for the cost-to-go function $J\left(\mathbf{S}\right)$, where $\mathbf{S}'$ is the possible state at the next time slot, while the transition probabilities $p_{s\rightarrow s'}$ are functions of the actions chosen. The solution is state-dependent, meaning that the optimal actions depend on the system state and not on time.

Corollary B.1. *The solution of the drift minimization problem (5.18) at each time slot t, is a suboptimal solution to the MDP in (B.2). It is called one-stage look-ahead (myopic), in the sense that the actions are chosen per slot, considering only the transition to the next state and not the entire cost-to-go.*

B.1.1 Proof of Proposition B.1

We first need the following lemma

Lemma B.1. *The performance measure can be written as an infinite sum of expected drifts over the discrete time axis, given the initial state* $\mathbf{S}(0)$

$$\mathcal{M}\left(V, \tilde{\mathbf{A}}\right) = V\left(\mathbf{S}(0)\right) + \sum_{t=0}^{\infty} \mathbb{E}\left[D\left(V\left(\mathbf{S}(t)\right), \mathbf{A}(t)\right) | \mathbf{S}(0)\right]. \tag{B.4}$$

Proof. : Let $\mathcal{F}^{(t)} := \{\mathbf{S}(0), \ldots, \mathbf{S}(t)\}$ be the information over the system realizations up to slot t. Obviously $\mathcal{F}^{(0)} \subseteq \mathcal{F}^{(t)}$ (formally we call $\left\{\mathcal{F}^{(t)},\ t \geq 0\right\}$ a filtration and $\mathcal{F}^{(0)}$ is a sub-σ-algebra of $\mathcal{F}^{(t)}$) and the tower property for expectations [Wil91, p.88] holds. Hence,

$$\begin{aligned}\mathbb{E}\left[V\left(\mathbf{S}(t+1)\right) | \mathbf{S}(0)\right] &\overset{Tower}{=} \mathbb{E}\left[\mathbb{E}\left[V\left(\mathbf{S}(t+1)\right) | \mathcal{F}^{(t)}\right] | \mathcal{F}^{(0)}\right] \\ &\overset{Markov}{=} \mathbb{E}\left[\mathbb{E}\left[V\left(\mathbf{S}(t+1)\right) | \mathbf{S}(t)\right] | \mathbf{S}(0)\right] \\ &\overset{(5.15)}{=} \mathbb{E}\left[D\left(V\left(\mathbf{S}(t)\right), \mathbf{A}(t)\right) | \mathbf{S}(0)\right] + \mathbb{E}\left[V\left(\mathbf{S}(t)\right) | \mathbf{S}(0)\right]\end{aligned}$$

and by repeating the process for $t, \ldots, 0$ and taking the limits for $t \to \infty$ we reach the result. ■

Now we can continue with the proof of the Proposition. Consider the series in (B.4) up to a finite horizon $T+1$ and denote the related sum by $\mathcal{M}_T\left(V, \tilde{\mathbf{A}}\right)$. Then the expected drift term for some $\tau \leq T$ equals

$$\begin{aligned}\mathbb{E}\left[D\left(V\left(\mathbf{S}(\tau)\right), \mathbf{A}(\tau)\right) | \mathbf{S}(0)\right] &= \\ \sum_{\mathbf{S}(1)} \cdots \sum_{\mathbf{S}(\tau)} p_{s_o \to s_1} \ldots p_{s_{\tau-1} \to s_\tau} D\left(V\left(\mathbf{S}(\tau)\right), \mathbf{A}(\tau)\right)\end{aligned}$$

It can be observed that $p_{s_{\tau-1} \to s_\tau}$, which can be controlled by the actions $\mathbf{A}(\tau-1)$ appear in all summands of $\mathcal{M}_T\left(V, \tilde{\mathbf{A}}\right)$, for $\tau \leq \hat{t} \leq T$ and not for $0 \leq t \leq \tau - 1$. Following this observation, the optimal choice of actions $p^*_{s_T \to s_{T+1}}$ are found by solving $\min_{\mathbf{A}(T) \in \mathbb{A}}$ $\mathcal{M}_T\left(V, \tilde{\mathbf{A}}\right)$, the cost-to-go at T.

The cost-to-go can be verified to satisfy the recursion, $\forall \mathbf{S}(\tau-1) \in \mathcal{S}$:

$$J\left(\mathbf{S}(\tau-1)\right) = \min_{\mathbf{A}(\tau-1) \in \mathbb{A}} \sum_{\mathbf{S}(\tau)} p_{s_{\tau-1} \to s_\tau}\left(V\left(\mathbf{S}(\tau)\right) - V\left(\mathbf{S}(\tau-1)\right) + J\left(\mathbf{S}(\tau)\right)\right).$$

The expression holds as well, when we let the horizon $T \to \infty$. Thus taking $\tau \to \infty$ results in (B.3).

Appendix C

Some Concepts and Results from Statistical Learning

C.1 Composite Hypothesis Testing

C.1.1 Generalization of Stein's Lemma

Theorem C.1 (Generalization of Stein's Lemma [Hoe65]). *For any $P_0, P_1 \in \mathcal{P}$, let the discriminant function $h(x)$ be such that*

$$P_0(h(x) > 0) \leq 2^{-\lambda n}. \tag{C.1}$$

Then,

$$\lim_{n\to\infty} P_1(h(x) > 0) \geq 1 - \epsilon, \tag{C.2}$$

for some $\epsilon < 1$ if and only if

$$D(P_1||P_0) > \lambda, \tag{C.3}$$

and condition (C.3) *is sufficient for achieving* (C.2) *for all $\epsilon > 0$ (i.e. achieving $P_1(h > 0) \to 1$) if $h(x)$ is the optimal discriminant function, provided as*

$$h(x) \triangleq h(x, \lambda) \triangleq \frac{1}{n} \log \frac{P_1(x)}{P_0(x)} - \lambda. \tag{C.4}$$

The *divergence* $D(P_1||P_0)$ in Theorem C.1 is defined by

$$D(P_1||P_0) \triangleq \lim_{n\to\infty} \frac{1}{n} \sum_{\mathcal{A}^n} P_1(x) \log \frac{P_1(x)}{P_0(x)}. \tag{C.5}$$

C.1.2 Universal Code

Definition C.1 (Universal Code)**.** *A "universal code"for the class $\mathcal{P}$ is a sequence of codes $c(n), n = 1, 2, \ldots,$ such that for every $P(\cdot) \in \mathcal{P}$,*

$$\lim_{n\to\infty} P\left[x : \frac{1}{n}u(x) \leq -\frac{1}{n}\log P(x) + \epsilon\right] = 1 \tag{C.6}$$

for any $\epsilon > 0$.

The expectation of $\frac{1}{n}u(x)$ approaches the minimal possible value as $n \to \infty$, this value being the entropy for $P(\cdot)$, given by

$$H \triangleq -\lim_{n\to\infty} \frac{1}{n} \sum_{\mathcal{A}^n} P(x) \log P(x). \tag{C.7}$$

For this reason, we say that every universal code is asymptotically optimal.

We introduce in below an example of universal code. Let $x \triangleq x^M$, $x_l \in \mathcal{A}, l = 1, 2, \ldots, M$. Assume that B divides M to m blocks, and denote $x_r^B = (x_l)_{l=r}^{r+B-1}$, $t_{r,m}^B = (x_l)_{l=r+mB \mod M}^{r+(m+1)B-1 \mod M}$. There exists a universal code for the class $\mathcal{P}$ with length function $u(x)$ given by [Dav73]:

$$u(x) = \frac{M}{B} H\left(v^B(x)\right) + \gamma^B \log\left(\frac{M}{B} + 1\right), \tag{C.8}$$

where γ is a constant,

$$H\left(v^B(x)\right) = -\sum_{r=1}^{M-B} v_r^B(x_r^B) \log\left(v_r^B(x_r^B)\right), \tag{C.9}$$

and $v_r^B(x_r^B)$ is defined as:

$$v_r^B(x_r^B) = \frac{B}{M} \sum_{m=1}^{M/B} \mathbf{1}\left\{t_{r,m}^B = x_r^B\right\}, \tag{C.10}$$

where the indicator function $\mathbf{1}\{\cdot\}$ is equal to 1 if $\{\cdot\}$ is true, and 0 otherwise.

C.2 Principal Component Analysis

Given a matrix $\boldsymbol{X} := [\boldsymbol{x}_1 \ldots \boldsymbol{x}_k] \in \mathbb{R}^{D\times k}$, denoting a collection of k D-dimensional data samples, we interpreted PCA in the way of minimizing the reconstruction error between the original data $\boldsymbol{X}$ and its estimates projected to the d-dimensional affine subspace $\boldsymbol{Y} \in \mathbb{R}^{d\times k}$, with $d \ll D$ [1].

[1] There are two other ways to formulate the problem: 1) maximizing the variance of projection, and 2) Maximum likelihood estimates of a parameter in a probabilistic model.

Let each point $\boldsymbol{x}_k \in \mathbb{R}^D$ be approximated by the affine projection of $\boldsymbol{y}_k$ in a d-dimensional subspace, represented as

$$\boldsymbol{x}_k = (\boldsymbol{x}_0 + \boldsymbol{U}_d\boldsymbol{y}_0) + \boldsymbol{U}_d(\boldsymbol{y}_k - \boldsymbol{y}_0) = \boldsymbol{x}_0 + \boldsymbol{U}_d\boldsymbol{y}_k \tag{C.11}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^D$ is a fixed point, $\boldsymbol{U}_d \in \mathbb{R}^{D\times d}$ is composed of d orthonormal column vectors, and $\boldsymbol{y}_k \in \mathbb{R}^d$ is the vector of new coordinates of $\boldsymbol{x}_k$ in the subspace. In order to obtain a unique solution, we impose the constraint $\bar{\boldsymbol{y}} := (1/K)\sum_{k=1}^K \boldsymbol{y}_k = 0$, and the optimization problem is to minimize the sum of squared error between $\boldsymbol{x}_k$ and its projection on the subsapce, given by

$$\min_{\boldsymbol{x}_0,\boldsymbol{U}_d,\{\boldsymbol{y}_k\}} \sum_{k=1}^{N} \|\boldsymbol{x}_k - (\boldsymbol{x}_0 + \boldsymbol{U}_d\boldsymbol{y}_k)\|^2 \tag{C.12}$$
$$\text{s.t. } \boldsymbol{U}_d^T\boldsymbol{U}_d = \boldsymbol{I} \text{ and } \bar{\boldsymbol{y}} = \boldsymbol{0}$$

Assuming $\boldsymbol{U}_d$ is fixed, differentiating the objective function with respect to $\boldsymbol{x}_0$ and $\boldsymbol{y}_k$ and setting the derivatives to be zero, we have $\hat{\boldsymbol{x}}_0 = \bar{\boldsymbol{x}} = (1/K)\sum_{k=1}^K \boldsymbol{x}_k$ and $\hat{\boldsymbol{y}}_k = \boldsymbol{U}_d^T(\boldsymbol{x}_k - \bar{\boldsymbol{x}})$. Substituting $\hat{\boldsymbol{x}}_0$ and $\hat{\boldsymbol{y}}_k$ into (C.12), and defining $\tilde{\boldsymbol{x}}_k := \boldsymbol{x}_k - \bar{\boldsymbol{x}}$, the original problem becomes one of finding an orthogonal matrix $\boldsymbol{U}_d$ that solves the problem

$$\min_{\boldsymbol{U}_d} \sum_{k=1}^{K} \|\tilde{\boldsymbol{x}}_k - \boldsymbol{U}_d\boldsymbol{U}_d^T\tilde{\boldsymbol{x}}_k\|^2, \text{ s.t. } \boldsymbol{U}_d^T\boldsymbol{U}_d = \boldsymbol{I} \tag{C.13}$$

A classical solution to PCA via SVD is provided in Theorem C.2.

Theorem C.2 (PCA via SVD [Jol02])**.** *Let $\tilde{\boldsymbol{X}} := [\tilde{\boldsymbol{x}}_1 \dots \tilde{\boldsymbol{x}}_K] \in \mathbb{R}^{D\times K}$ be the matrix formed by stacking the (zero-mean) data samples as its column vectors. Let $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the SVD of the matrix $\tilde{\boldsymbol{X}}$. Then for any $d < D$, a solution to* (C.13)*, $\hat{\boldsymbol{U}}_d$ is exactly the first d columns of $\boldsymbol{U}$; and $\hat{\boldsymbol{y}}$ is the kth column of the top $d \times K$ submatrix $\boldsymbol{\Sigma}_d\boldsymbol{V}_d^T$ of the matrix $\boldsymbol{\Sigma}\boldsymbol{V}^T$.*

C.3 Gaussian Identities

The *multivariate Gaussian (normal) distribution* is “non-degenerate” when the symmetric *covariance matrix* $\boldsymbol{\Sigma}$ is positive definite. In this case the joint probability density is given by

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{C.14}$$

where $|\boldsymbol{X}|$ denotes the matrix determinant, and $\boldsymbol{\mu} \in \mathbb{R}^D$ denotes the mean vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ is the symmetric, positive definite covariance matrix. As a shorthand we write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be jointly Gaussian random vectors

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^T & \boldsymbol{B} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{C}} \\ \tilde{\boldsymbol{C}}^T & \tilde{\boldsymbol{B}} \end{bmatrix}^{-1}\right), \tag{C.15}$$

then the *marginal distribution* of $\boldsymbol{x}$ and the *conditional distribution* of $\boldsymbol{x}$ given $\boldsymbol{y}$ are (see [VM14, sec. 9.3] and Equation (A.4) in Appendix A.1)

$$\begin{aligned} \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{A}), \text{ and } \boldsymbol{x}|\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^T) \\ \text{or } \boldsymbol{x}|\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{\mu}_x - \tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{C}}(\boldsymbol{y} - \boldsymbol{\mu}_y), \tilde{\boldsymbol{A}}^{-1}). \end{aligned} \tag{C.16}$$

Appendix D

Some Concepts and Results from Contraction Mapping

D.1 Mathematical Spaces

Definition D.1 (Metric Space)**.** *A metric space is a pair $(\mathcal{X}, d)$, where $\mathcal{X}$ is a set and d is a metric on $\mathcal{X}$ (or distance function on $\mathcal{X}$), that is, a function defined*[1] *on $\mathcal{X} \times \mathcal{X}$ such that for all $x, y, z \in \mathcal{X}$ we have:*

$$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \textit{ (Non-negative, real)}, \quad d(x, y) = 0 \Leftrightarrow x = y \textit{ (Identity of indiscernibles)},$$
$$d(x, y) = d(y, x) \textit{ (Symmetry)}, \quad d(x, y) \leq d(x, z) + d(z, y) \textit{ (Triangle inequality).}$$

Definition D.2 (Vector Space)**.** *A vector space over a field $\mathcal{K}$ is a nonempty set $\mathcal{X}$ of elements $\boldsymbol{x}, \boldsymbol{y}, \ldots$ (called vectors) together with two algebraic operations: vector addition and multiplication of vectors by scalars.*

Definition D.3 (Normed Space, Banach Space)**.** *A* normed space *$\mathcal{X}$ is a vector space with a norm defined on it. A Banach space is a complete normed space. Here a norm on Euclidean n-space $\mathbb{R}^n$ is a real-valued function on $\mathbb{R}^n$ whole value at an $\boldsymbol{x} \in \mathbb{R}^n$ is denoted by $\|\boldsymbol{x}\|$, and which has the properties*

$$\forall_{\boldsymbol{x}\in\mathbb{R}^n} \ \|\boldsymbol{x}\| \geq 0, \quad \forall_{\alpha\in\mathbb{R},\boldsymbol{x}\in\mathbb{R}^n} \|\alpha\boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\|, \quad \|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0},$$
$$\forall_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n} \|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \textit{ (Triangle inequality).}$$

Definition D.4 (Inner Product Space, Hilbert Space)**.** *An inner product space (or pre-Hilber space) is a vector space $\mathcal{X}$ with an inner product defined on $\mathcal{X}$. A Hilbert space is a*

[1] The symbol × denotes the *Cartesian product* of sets $\mathcal{A} \times \mathcal{B}$.

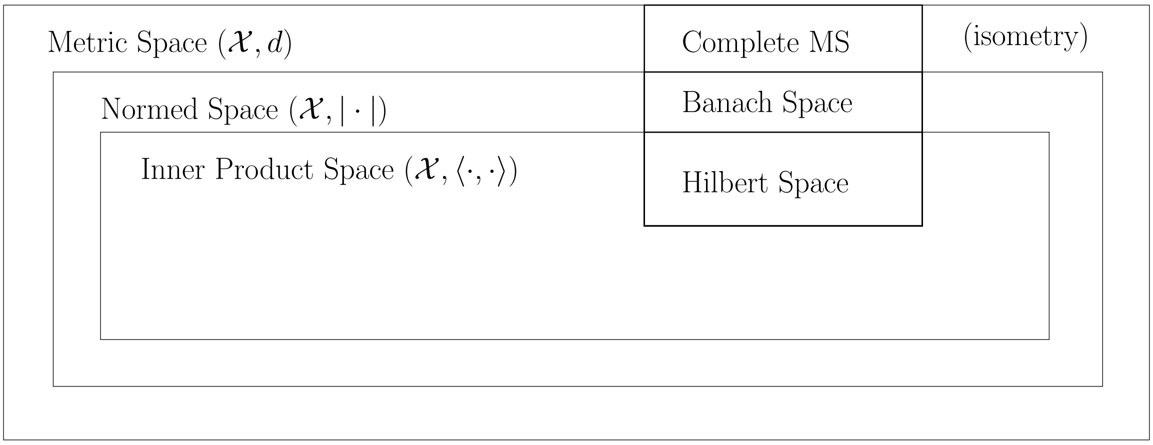

Metric Space $(\mathcal{X}, d)$
Complete MS
(isometry)
Normed Space $(\mathcal{X}, |\cdot|)$
Banach Space
Inner Product Space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$
Hilbert Space

Figure D.1: Representation of mathematical spaces

complete inner product space (complete in the metric defined by the inner product). Here, an inner product on $\mathcal{X}$ is a mapping of $\mathcal{X} \times \mathcal{X}$ into the scalar field $\mathcal{K}$ of $\mathcal{X}$; that is, with every pair of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ there is associated a scalar which is written as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle$$

and is called the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$, such that for all vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ and scalars α we have

$$\begin{aligned}
&\langle \boldsymbol{x}+\boldsymbol{y}, \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle + \langle \boldsymbol{y}, \boldsymbol{z} \rangle, \quad \langle \alpha\boldsymbol{x}, \boldsymbol{y} \rangle = \alpha\langle \boldsymbol{x}, \boldsymbol{y} \rangle && \textit{(Linearity)},\\
&\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \overline{\langle \boldsymbol{y}, \boldsymbol{x} \rangle} && \textit{(Conjugate symmetry)},\\
&\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0, \quad \langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \Leftrightarrow \boldsymbol{x} = 0 && \textit{(Positive-definiteness)}.
\end{aligned}$$

An inner product on $\mathcal{X}$ defines a norm on $\mathcal{X}$ given by

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \tag{D.1}$$

and a metric on $\mathcal{X}$ given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}. \tag{D.2}$$

Hence, inner product space are normed space, and Hilbert spaces are Banach spaces.

A visual representation of the above-mentioned spaces is illustrated in Fig. D.1.

D.2 Fixed Point Theorems

Definition D.5 (Nonexpansive, shrinking, contraction [Kre89]). *A mapping $\boldsymbol{f} : \mathcal{X} \to \mathcal{X}$ from a metric space $(\mathcal{X}, d)$ to itself is said to be*

- *nonexpansive if $d(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{y})) \leq d(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$;*
- *shrinking (or contractive) if $d(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{y})) < d(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{x} \neq \boldsymbol{y} \in \mathcal{X}$;*
- *a contraction if there is $c < 1$ such that $d(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{y})) \leq cd(\boldsymbol{x}, \boldsymbol{y})$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}$.*

Theorem D.1 (Banach Contraction Mapping [Kre89]). *Let $(\mathcal{X}, d)$ be a complete metric space and $\boldsymbol{f} : \mathcal{X} \to \mathcal{X}$ be a contraction. Then $\boldsymbol{f}$ has a unique fixed point $\boldsymbol{x}^* \in \mathcal{X}$, and for any $\boldsymbol{x} \in \mathcal{X}$ the sequence of iterations $\boldsymbol{f}^n(\boldsymbol{x})$ converges to $\boldsymbol{x}^*$.*

Theorem D.2 (Edelstein Contractive Mapping [Ede62]). *Let $(\mathcal{X}, d)$ be a compact metric space and $\boldsymbol{f} : \mathcal{X} \to \mathcal{X}$ be a contractive. Then $\boldsymbol{f}$ has a unique fixed point $\boldsymbol{x}^* \in \mathcal{X}$, and for any $\boldsymbol{x} \in \mathcal{X}$ the sequence of iterations $\boldsymbol{f}^n(\boldsymbol{x})$ converges to $\boldsymbol{x}^*$.*

Definition D.6 (Hilbert's Projective Metric). *Let C be a convex cone in a real vector space $\mathcal{X}$, and we have $C = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} \geq 0\}$. We define Hilbert's (projective) metric [Bir57,KP82], $d_H : C \times C \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ on C, as follows: $d_H(\mathbf{0}, \mathbf{0}) = 0$; when $\boldsymbol{x}, \boldsymbol{y} \geq \mathbf{0}, d_H(\boldsymbol{x}, \mathbf{0}) = d_H(\mathbf{0}, \boldsymbol{y}) = \infty$ and*

$$d_H(\boldsymbol{x}, \boldsymbol{y}) \equiv \log \frac{M(\boldsymbol{x}, \boldsymbol{y})}{m(\boldsymbol{x}, \boldsymbol{y})} \tag{D.3}$$

where

$$M(\boldsymbol{x}, \boldsymbol{y}) \equiv \inf\{\lambda \geq 0 : \boldsymbol{x} \leq \lambda \boldsymbol{y}\} = \max_i x_i/y_i \tag{D.4}$$

$$m(\boldsymbol{x}, \boldsymbol{y}) \equiv \sup\{\lambda \geq 0 : \boldsymbol{x} \geq \lambda \boldsymbol{y}\} = \min_i x_i/y_i \tag{D.5}$$

clearly we have $m(\boldsymbol{x}, \boldsymbol{y}) = 1/M(\boldsymbol{x}, \boldsymbol{y})$ and d_H can be written as

$$d_H(\boldsymbol{x}, \boldsymbol{y}) \equiv \max_i \log x_i/y_i + \max_i \log y_i/x_i \tag{D.6}$$

The metric d_H is called projective on C because d_H is constant on rays, that is, $d_H(\lambda \boldsymbol{x}, \mu \boldsymbol{y}) = d_H(\boldsymbol{x}, \boldsymbol{y})$ for $\lambda, \mu > 0$, and $d_H(\boldsymbol{x}, \boldsymbol{y}) = 0$ iff $\boldsymbol{x} = \lambda \boldsymbol{y}$ for some $\lambda > 0$. Using the metric d_H, Birkhoff [Bir57, Theorem 3] observe that every linear transformation with a positive matrix may be viewed as a contraction mapping on the nonnegative orthant, and this observation turns the Perron-Frobenius theorem into a special case of the Banach contraction mapping theorem.

In the following we introduce the other two metrics motivated by the projective metric d_H, which are important in the generalizations of Perron-Frobenius theory to monotonic and subhomogeneous functions.

Definition D.7. *We define the metrics d_S and d_M for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^K$ as follows:*

- $d_S(\boldsymbol{x}, \boldsymbol{y}) \equiv \max_i |\log x_i/y_i|$
- $d_M(\boldsymbol{x}, \boldsymbol{y}) \equiv \max_i(\log x_i/y_i)^+ + \max_i(\log y_i/x_i)^+$

where $(x)^+$ *denotes* $\max\{x, 0\}$.

Note that d_S is derived by taking the component-wise logarithm of the supremum norm $\rho_S = \|\boldsymbol{x} - \boldsymbol{y}\|_\infty = \max_i |x_i - y_i|$, and d_M is obtained by taking the component-wise logarithm of $\rho_M = \max_i(x_i - y_i)^+ + \max_i(y_i - x_i)^+$. The component-wise logarithm defines an isomorphism between $(\mathbb{R}_+^K, d)$ to $(\mathbb{R}^K, \rho)$.

D.3 Contractive Mappings with or without Monotonicity

This section includes some concept and proofs from the max-min fairness problem using contractive operators with or without monotonicity introduced in Chapter 9.

D.3.1 Approximation of Overlap Factor

One possible method is to compute the overlap factor proportional to the fraction of the overlapping band. For example, the cell-pairwise directional overlap factor $o_{i,j}^{\mathrm{X}\leftarrow\mathrm{Y}}$ for $\mathrm{X}, \mathrm{Y} \in \{\mathrm{UL}, \mathrm{DL}\}$ and $i, j \in \mathcal{N}, i \neq j$ can be define by $o_{i,j}^{\mathrm{X}\leftarrow\mathrm{Y}} := \max\{0, (\nu_j^{\mathrm{Y}} + \nu_i^{\mathrm{X}} - 1)/\nu_i^{\mathrm{X}}\}$ if $\mathrm{X} \neq \mathrm{Y}$, to express the probability that a RB in cell i receives interference in UL (DL) from any DL (UL) transmission signal in cell j (inter-cell inter-link interference); and $o_{i,j}^{\mathrm{X}\leftarrow\mathrm{Y}} := \max\{1, \nu_j^{\mathrm{Y}}/\nu_i^{\mathrm{X}}\}$ if $\mathrm{X} = \mathrm{Y}$, to express the probability that a RB in cell i receives interference in UL (DL) from any UL (DL) transmission signal in cell j (inter-cell intra-link interference). For example, assuming $\nu_i^{\mathrm{DL}} = 0.7, \nu_i^{\mathrm{UL}} = 0.3$ for cell i and $\nu_j^{\mathrm{DL}} = 0.3, \nu_j^{\mathrm{UL}} = 0.7$ (as shown in Fig. 9.4), we have $o_{i,j}^{\mathrm{DL}\leftarrow\mathrm{UL}} = \max\{0, (\nu_j^{\mathrm{UL}} + \nu_i^{\mathrm{DL}} - 1)/\nu_i^{\mathrm{DL}}\} = \max\{(0.7 + 0.7 - 1)/0.7, 0\} \approx 0.57$, while $o_{ij}^{\mathrm{UL}\leftarrow\mathrm{DL}} = \max\{0, (\nu_j^{\mathrm{DL}} + \nu_i^{\mathrm{UL}} - 1)/\nu_i^{\mathrm{UL}}\} = 0$. Let us define the overlap matrix $\boldsymbol{O}^{\mathrm{X}\leftarrow\mathrm{Y}} := (o_{i,j})^{\mathrm{X}\leftarrow\mathrm{Y}} \in [0, 1]^{N\times N}$, for $\mathrm{X}, \mathrm{Y} \in \{\mathrm{UL}, \mathrm{DL}\}$. To transform $\boldsymbol{O}^{\mathrm{X}\leftarrow\mathrm{Y}}$ to the per-link basis matrix (between the UL and DL), we define $\tilde{\boldsymbol{O}}^{\mathrm{X}\leftarrow\mathrm{Y}} := (\boldsymbol{A}^{\mathrm{X}})^T \boldsymbol{O}^{\mathrm{X}\leftarrow\mathrm{Y}} \boldsymbol{A}^{\mathrm{Y}}$. The cross-link coupling matrix is then modified by computing the Hadamard product (element-wise product) of $\tilde{\boldsymbol{V}}^{\mathrm{X}\leftarrow\mathrm{Y}}$ and $\tilde{\boldsymbol{O}}^{\mathrm{X}\leftarrow\mathrm{Y}}$, for $\mathrm{X}, \mathrm{Y} \in \{\mathrm{UL}, \mathrm{DL}\}$.

Unfortunately, the fraction of the overlapping bands depends on the cell-specific loads $\boldsymbol{\nu}^{\mathrm{UL}}$ and $\boldsymbol{\nu}^{\mathrm{DL}}$, which further depend on the dynamic UL and DL resource allocation $\boldsymbol{w}$

(as the variable to be optimized in Prob. 9.1). Thus, introducing such a modification dramatically complicates optimization problem.

A compromise approach is to use the historical measurements of load $\boldsymbol{\nu}^{\mathrm{UL}}$ and $\boldsymbol{\nu}^{\mathrm{DL}}$ as estimates to compute the *cell-pairwise overlap factor* $o_{ij}^{\mathrm{X}\leftarrow\mathrm{Y}}$ for $\mathrm{X},\mathrm{Y} \in \{\mathrm{UL},\mathrm{DL}\}$, $i,j \in \mathcal{N}$ as described above.

An alternative to the *cell-pairwise overlap factor* $o_{ij}^{\mathrm{X}\leftarrow\mathrm{Y}}$ is to define a *cell-specific overlap factor* c_i^{X}, for $\mathrm{X} \in \{\mathrm{UL},\mathrm{DL}\}$, $i \in \mathcal{N}$ to express how likely a transmission in cell i causes inter-link interference to the transmission in another cell, while the computation of intra-link overlap factor remains the same as the approach above. This approach is more error-tolerant in the sense that it does not return zero probability for inter-cell inter-link interference. We define two vectors with constant values $\boldsymbol{c}^{\mathrm{UL}} \in [0,1]^N$ and $\boldsymbol{c}^{\mathrm{DL}} \in [0,1]^N$, which can be chosen proportional to the historical measurements of $\boldsymbol{\nu}^{\mathrm{UL}}$ and $\boldsymbol{\nu}^{\mathrm{DL}}$, respectively. Further we can modify the cross-link coupling matrix by defining $\boldsymbol{V}^{\mathrm{UL}\leftarrow\mathrm{DL}} := (\boldsymbol{A}^{\mathrm{UL}})^T \operatorname{diag}(\boldsymbol{c}^{\mathrm{UL}})\boldsymbol{H}_1 \operatorname{diag}(\boldsymbol{c}^{\mathrm{DL}})\boldsymbol{A}^{\mathrm{DL}}$, and $\boldsymbol{V}^{\mathrm{DL}\leftarrow\mathrm{UL}} := \operatorname{diag}\left((\boldsymbol{A}^{\mathrm{DL}})^T\boldsymbol{c}^{\mathrm{DL}}\right)\boldsymbol{H}_2 \operatorname{diag}\left((\boldsymbol{A}^{\mathrm{UL}})^T\boldsymbol{c}^{\mathrm{UL}}\right)$, such that the coupling between UL and DL is proportional to the multiplication of the cell UL and DL overlap factors. For example, the overlap factor between the downlinks in cell i and the uplinks in cell j is proportional to $c_i^{\mathrm{DL}}c_j^{\mathrm{UL}}$ as shown in Fig. 9.4.

D.3.2 Standard Interference Function

Definition D.8. *A vector function $\boldsymbol{f} : \mathbb{R}_+^k \to \mathbb{R}_{++}^k$ is said to be a standard interference function (SIF) if the following axioms hold:*

1. *(Monotonicity) $\boldsymbol{x} \leq \boldsymbol{y}$ implies $\boldsymbol{f}(\boldsymbol{x}) > 0 \leq \boldsymbol{f}(\boldsymbol{y})$*

2. *(Scalability) for each $\alpha > 1$, $\alpha\boldsymbol{f}(\boldsymbol{x}) > \boldsymbol{f}(\alpha\boldsymbol{x})$*

The original definition of standard interference function is stated in [Yat95], which also requires positivity. In Definition D.8 we drop the positivity $\boldsymbol{f}(\boldsymbol{x}) > 0$ for $\boldsymbol{x} \in \mathbb{R}_+^k$ because it is a consequence of the other two properties [LSWL04].

Lemma D.1 (Selected Properties of SIF [Yat95])**.** *Let $\boldsymbol{f} : \mathbb{R}_+^k \to \mathbb{R}_{++}^k$ be a SIF. Then*

1. *There is at most one fixed point $\boldsymbol{x} \in Fix(\boldsymbol{f}) := \{\boldsymbol{x} \in \mathbb{R}_{++}^k | \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})\}$.*

2. *The fixed point exists if and only if there exists $\boldsymbol{x}' \in \mathbb{R}_{++}^k$ satisfying $\boldsymbol{f}(\boldsymbol{x}') \leq \boldsymbol{x}'$.*

3. *If a fixed point exists, then it is the limit of the sequence $\{\boldsymbol{x}^{(n)}\}$ generated by $\boldsymbol{x}^{(n+1)} = \boldsymbol{f}(\boldsymbol{x}^{(n)})$, $n \in \mathbb{N}$, where $\boldsymbol{x}^{(1)} \in \mathbb{R}_+^k$ is arbitrary. If $\boldsymbol{x}^{(1)} = \mathbf{0}$, then the sequence is monotonically increasing (in each component). In contrast, if $\boldsymbol{x}^{(1)}$ satisfies $\boldsymbol{f}(\boldsymbol{x}^{(1)}) \leq \boldsymbol{x}^{(1)}$, then the sequence is monotonically decreasing (in each component).*

D.3.3 Proof of Lemma 9.1

The essential steps of the proof follow those in the proof of [Reaar, Ex. 2]. First we show that $f_{\boldsymbol{p}',l}(\boldsymbol{w}) := d_l / \left(W_0 B \log(1 + \mathrm{SINR}_l(\boldsymbol{w}))\right)$ is positive and concave. Function $f_{\boldsymbol{p}',l}(\boldsymbol{w})$ is positive concave, because of the following facts: i) $h(x) := 1/\log_2(1 + 1/x)$ is a concave function on $\mathbb{R}_{++}$, ii) composition of concave functions with affine transformations (see the interference term in (9.6)) preserves concavity, and iii) a set of concave functions is closed under multiplication and addition. Then, because a positive concave function is proved to be a SIF in [Reaar, Prop. 1], $f_{\boldsymbol{p}',l}$ is SIF. As a collection of $\{f_{\boldsymbol{p}',l}\}$, the vector function $\boldsymbol{f}_{\boldsymbol{p}'}$ is SIF.

D.3.4 Proof of Theorem 9.1

Since the essential steps follow those in the proof of [Nuz07, Th. 3.2], we describe only proof outlines and mention crucial lemmas in this paper, for lack of space. Using [Nuz07, Lem. 3.3], we know that $\boldsymbol{h} := \boldsymbol{x}/g(\boldsymbol{x})$ is non-expansive (see details in Definition D.5) on $(\mathbb{R}^k_{++}, d_M)$, where the metric d_M is defined in Definition D.7. Because $\boldsymbol{f}$ is SIF, by virtue of [Nuz07, Lem. 2.2], $\boldsymbol{\psi} = \theta \boldsymbol{h} \circ \boldsymbol{f} = \theta \boldsymbol{f}/(g \circ \boldsymbol{f})$ in (9.18) is shrinking (or contractive, see details in Definition D.5) with respect to d_M.

If $\boldsymbol{\psi}$ is a contractive mapping on a compact metric space on $(\mathbb{R}^k_{++}, \mu_s)$, there exists a unique fixed point $\boldsymbol{x} \in \mathbb{R}^k_{++}$ with $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{x}$ [Sma80, Th.5.2.3]. In the following we show that $\boldsymbol{\psi}$ is a mapping of a compact space to itself. For any input, since g is homogeneous on $\mathbb{R}^k_{++}$, we have $g \circ \boldsymbol{\psi} = (\theta / g \circ \boldsymbol{f}) \cdot (g \circ \boldsymbol{f}) = \theta$. Because a monotonic vector function has bounded level sets, we have that $\boldsymbol{\psi}(\boldsymbol{x}) \leq \boldsymbol{b}$ for some finite $\boldsymbol{b} > \boldsymbol{0}$. With $\boldsymbol{\psi}(\boldsymbol{x}) \leq \boldsymbol{b}$ and $\boldsymbol{f}(\boldsymbol{x}) \geq \boldsymbol{f}(\boldsymbol{0})$ for all $\boldsymbol{x} \in \mathbb{R}^k_{+}$, we have $\boldsymbol{\psi}^2(\boldsymbol{x}) \geq \theta \boldsymbol{f}(0)/(g \circ \boldsymbol{f}(\boldsymbol{b})) = \boldsymbol{a} > \boldsymbol{0}$, and we see that the range of $\boldsymbol{\psi}^n$ falls inside the finite positive rectangle $\mathrm{R}(\boldsymbol{a}, \boldsymbol{b})$ for $n \geq 2$. Hence, there is exactly one eigenvector $\boldsymbol{x} \in \mathbb{R}^k_{++}$ to satisfy $\boldsymbol{x}' = \rho' \boldsymbol{f}(\boldsymbol{x}')$ where the associate eigenvalue is given by $\rho' = \theta/(g \circ \boldsymbol{f}(\boldsymbol{x}'))$, such that $g(\boldsymbol{x}') = g(\boldsymbol{\psi}(\boldsymbol{x}')) = \theta$.

D.3.5 Proof of Prop. 9.1

In the following part of this proof, for simplicity of notation, we omit the dependency on $\boldsymbol{p}'$, and denote $\boldsymbol{f} := \boldsymbol{f}_{\boldsymbol{p}'}$, $g := g_{\boldsymbol{p}'}$ and $\lambda := \lambda_{\boldsymbol{p}'}$.

It is obvious that g defined in (9.18b) is positive and homogeneous of degree 1 on $\mathbb{R}^{2K}_{++}$. By virtue of Theorem 9.1 and Lemma 9.1, we have that for $\theta = 1$, there exist a unique fixed point $\boldsymbol{w}' = \lambda' \boldsymbol{f}(\boldsymbol{w}')$ such that $g(\boldsymbol{w}') = 1$, where λ' can be computed with iteration (9.18a).

Then we show that there exists no $\lambda'' > \lambda'$ to satisfy $\boldsymbol{w}'' \geq \lambda'' \boldsymbol{f}(\boldsymbol{w}'')$ and $g(\boldsymbol{w}'') \leq 1$. We proceed by contradiction. Suppose that there exists a $\lambda'' > \lambda'$ to satisfy $\boldsymbol{w}'' \geq \lambda'' \boldsymbol{f}(\boldsymbol{w}'')$ such that $g(\boldsymbol{w}'') \leq 1$. Let us define a function $\boldsymbol{f}' := \lambda' \boldsymbol{f}$. Because $\boldsymbol{f}$ is a SIF, $\boldsymbol{f}'$ is also a SIF. We then have $\boldsymbol{f}'(\boldsymbol{w}'') = \lambda'\boldsymbol{f}(\boldsymbol{w}'') < \lambda''\boldsymbol{f}(\boldsymbol{w}'') \leq \boldsymbol{w}''$, i.e., $\boldsymbol{w}''$ is a feasible point with respect to the SIF $\boldsymbol{f}'$. Thus, the sequence starting from $\boldsymbol{w}''$ decreases monotonically to $\boldsymbol{w}'$ (by using the third property of SIF stated in Lemma D.1). Then we have $\boldsymbol{w}' \leq \boldsymbol{f}'(\boldsymbol{w}'') < \boldsymbol{w}''$. Since $g(\boldsymbol{w})$ is monotone increasing on $\mathbb{R}_+^{2K}$, we have $g(\boldsymbol{w}'') > g(\boldsymbol{w}') = 1$, which contradicts the earlier statement $g(\boldsymbol{w}'') \leq 1$.

Knowing that λ' is the maximum feasible utility, now we show that for all $\boldsymbol{w} \in \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{p}')$ satisfying $\boldsymbol{w} \geq \lambda'\boldsymbol{f}(\boldsymbol{w}) = \boldsymbol{f}'(\boldsymbol{w})$, we have $\boldsymbol{w}' \leq \boldsymbol{w}$. Because $\boldsymbol{f}'$ is also a SIF, $\boldsymbol{w} \geq \boldsymbol{f}'(\boldsymbol{w})$ implies that the sequence $\boldsymbol{w}$ decreases monotonically to $\boldsymbol{w}'$ satisfying $\boldsymbol{w}' = \boldsymbol{f}'(\boldsymbol{w}') = \lambda'\boldsymbol{f}(\boldsymbol{w}')$. Thus., $\boldsymbol{w}' \leq \boldsymbol{w}$.

D.3.6 Proof of Prop. 9.2

We will prove by induction that by using algorithm in Prop. 9.2, the sequence λ is monotonically increasing until $g_1(\boldsymbol{w}) = 1$ is satisfied.

At the base step, suppose the solution to Prob.9.2a yields $\boldsymbol{w}' = \lambda'\boldsymbol{f}_{\boldsymbol{p}'}(\boldsymbol{w}')$ where $\lambda' := 1/g_{\boldsymbol{p}'}(\boldsymbol{w}')$ and $g_{\boldsymbol{p}'}(\boldsymbol{w}') = \max\{g_1(\boldsymbol{w}'), g_{2,\boldsymbol{p}'}(\boldsymbol{w}')\}$, with $g_1(\boldsymbol{w}') < 1$ and $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = 1$. Let us define $g_1(\boldsymbol{w}') = a < 1$ and $\boldsymbol{p}'' = a\boldsymbol{p}'$. With fixed $\boldsymbol{p}''$, using Theorem 9.1, iteration (9.18) converges to a unique fixed point $\boldsymbol{w}''$, satisfying

$$\boldsymbol{w}'' = \lambda''\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{w}'') \tag{D.7}$$

$$\text{such that } \max\{g_1(\boldsymbol{w}''), g_2(\boldsymbol{p}'', \boldsymbol{w}'')\} = 1 \tag{D.8}$$

It is clear that $\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{w}') < \boldsymbol{f}_{\boldsymbol{p}'}(\boldsymbol{w}') = \boldsymbol{w}'/\lambda'$, by dividing both the numerator and denominator by a in (9.6), and substituting (9.6) in (9.7) and (9.14c). Now let us define $\boldsymbol{v}' = \boldsymbol{w}'/a > \boldsymbol{w}'$. Moreover, knowing that $\boldsymbol{f}_{\boldsymbol{p}''}$ is also a SIF, we have $\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{v}') = \boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{w}'/a) < \boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{w}')/a$ due to the scalability, that further leads to $\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{v}') < \boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{w}')/a < \boldsymbol{f}_{\boldsymbol{p}'}(\boldsymbol{w}')/a = \boldsymbol{w}'/(a\lambda') = \boldsymbol{v}'/\lambda'$. In other words, there exists $\boldsymbol{v}'$ such that $\lambda'\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{v}') < \boldsymbol{v}'$, and $\boldsymbol{v}'$ is a feasible point with respect to the SIF $\boldsymbol{f}'_{\boldsymbol{p}''} := \lambda'\boldsymbol{f}_{\boldsymbol{p}''}$. Thus, starting from $\boldsymbol{v}'$, the sequence of $\boldsymbol{v}$ decrease monotonically to a unique fixed point (by using the third property of SIF stated in Lemma D.1)

$$\boldsymbol{v}'' = \boldsymbol{f}'_{\boldsymbol{p}''}(\boldsymbol{v}'') < \boldsymbol{f}'_{\boldsymbol{p}''}(\boldsymbol{v}') < \boldsymbol{v}' \tag{D.9}$$

Due to the monotonicity and homogeneity of g_1 with respect to $\boldsymbol{w}$, and the same properties of g_2 with respect to both $\boldsymbol{p}$ and $\boldsymbol{w}$, we have

$$g_1(\boldsymbol{v}'') < g_1(\boldsymbol{v}') = g_1(\boldsymbol{w}'/a) = g_1(\boldsymbol{w}')a = 1 \tag{D.10}$$

$$g_2(\boldsymbol{p}'', \boldsymbol{v}'') < g_2(a\boldsymbol{p}', \boldsymbol{v}') = g_2(a\boldsymbol{p}', \boldsymbol{w}'/a) = 1 \tag{D.11}$$

We prove $\lambda'' > \lambda'$ by contradiction. Suppose $\lambda'' \leq \lambda'$, then we have $\lambda''\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{v}'') \leq \lambda'\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{v}'') = \boldsymbol{v}''$, using (D.9). By defining $\boldsymbol{f}''_{\boldsymbol{p}''} := \lambda''\boldsymbol{f}_{\boldsymbol{p}''}$ which is also a SIF, since $\boldsymbol{f}''_{\boldsymbol{p}''}(\boldsymbol{v}'') \leq$

$\boldsymbol{v}''$, starting from $\boldsymbol{v}''$, the sequence of $\boldsymbol{w}$ is monotonically decreasing to the unique fixed point $\boldsymbol{v}^\star$ satisfying $\boldsymbol{v}^\star = \boldsymbol{f}''_{\boldsymbol{p}''}(\boldsymbol{v}^\star) = \lambda''\boldsymbol{f}_{\boldsymbol{p}''}(\boldsymbol{v}^\star)$. Because $\boldsymbol{v}^\star$ is unique (by using the first and second properties of SIF stated in Lemma D.1), using (D.7), we have $\boldsymbol{w}'' = \boldsymbol{v}^\star \leq \boldsymbol{v}''$, which further leads to $\max\{g_1(\boldsymbol{v}''), g_2(\boldsymbol{p}'', \boldsymbol{v}'')\} \geq \max\{g_1(\boldsymbol{w}''), g_2(\boldsymbol{p}'', \boldsymbol{w}'')\} = 1$. This contradicts the inequalities (D.10) and (D.11). Thus, we have that $\lambda'' > \lambda'$ if $g_1(\boldsymbol{w}') < 1$.

For the further iteration step, using (D.8), it remains to consider cases in which $g_1(\boldsymbol{w}'') = 1$, or $g_1(\boldsymbol{w}'') < 1, g_2(\boldsymbol{p}'', \boldsymbol{w}'') = 1$. The former case directly leads to $g_1(\boldsymbol{w}'') = 1$, and the algorithm stops at $\lambda'' > \lambda'$. The latter case yields $g_1(\boldsymbol{w}'') < 1$, The proof above shows that the iteration step further increases λ, with scaled $\boldsymbol{p}''' = g_1(\boldsymbol{w}'')\boldsymbol{p}''$.

D.3.7 Proof of Prop. 9.3

The solution to P.2a satisfies $\boldsymbol{p}' = \lambda' \boldsymbol{f}_{\boldsymbol{w}'}(\boldsymbol{p}')$ using the reformulation in (9.20). Since the variables $\boldsymbol{p}$ and $\boldsymbol{w}$ are interchangeable in g_2, we have $g_{2,\boldsymbol{p}'}(\boldsymbol{w}') = g_{2,\boldsymbol{w}'}(\boldsymbol{p}')$.

Therefore, if $g_{2,\boldsymbol{w}'}(\boldsymbol{p}') = 1$, Theorem 9.1 implies that there is exactly one eigenvector λ and associate eigenvector $\boldsymbol{p}$ of $\boldsymbol{f}_{\boldsymbol{w}'}$ such that $g_{2,\boldsymbol{w}'}(\boldsymbol{p}') = 1$, and we have $\lambda'' = \lambda'$ and $\boldsymbol{p}'' = \boldsymbol{p}'$.

Then we consider the case when $g_{2,\boldsymbol{w}'}(\boldsymbol{p}') < 1$. Because $\boldsymbol{p}''$ is the optimal solution to P.2b, if we can find a $\hat{\boldsymbol{p}} \in \mathbb{R}^{2K}_{++}$ such that $\hat{\lambda} := \min_{l \in \overline{\mathcal{K}}} \hat{p}_l / f_{\boldsymbol{w}',l}(\hat{\boldsymbol{p}})$, $g_{2,\boldsymbol{w}'}(\hat{\boldsymbol{p}}) \leq 1$ and $\hat{\lambda} > \lambda'$, then we have $\lambda'' \geq \hat{\lambda} > \lambda'$. Thus, the remaining task is to find an arbitrary $\hat{\boldsymbol{p}}$ that fulfills the above mentioned conditions. Let us define $\alpha = 1/g_{2,\boldsymbol{w}'}(\boldsymbol{p}') > 1$ and $\hat{\boldsymbol{p}} := a\boldsymbol{p}'$. Then, we have

$$\hat{\lambda} = \min_{l \in \overline{\mathcal{K}}} \frac{\alpha p'_l}{f_{\boldsymbol{w}',l}(\alpha \boldsymbol{p}')} > \min_{l \in \overline{\mathcal{K}}} \frac{\alpha p'_l}{\alpha f_{\boldsymbol{w}',l}(\boldsymbol{p}')} = \lambda'$$

The above inequality is due to the scalability of the SIF $\boldsymbol{f}_{\boldsymbol{w}'}$.

List of Publications

[1] Q. Liao, M. Kaliszan, and S. Stańczak, "A virtual soft handover method based on base station cooperation with fountain codes," in *11th European Wireless Conference 2011-Sustainable Wireless Technologies (European Wireless).* VDE, 2011, pp. 1–6.

[2] Q. Liao, M. Wiczanowski, and S. Stańczak, "Toward cell outage detection with composite hypothesis testing," in *International Conference on Communications (ICC).* IEEE, 2012, pp. 4883–4887.

[3] A. Giovanidis, Q. Liao, and S. Stańczak, "A distributed interference-aware load balancing algorithm for LTE multi-cell networks," in *Smart Antennas (WSA), 2012 International ITG Workshop on.* IEEE, 2012, pp. 28–35.

[4] Q. Liao, S. Stańczak, and F. Penna, "A statistical algorithm for multi-objective handover optimization under uncertainties," in *Wireless Communications and Networking Conference (WCNC), 2013 IEEE.* IEEE, 2013, pp. 1552–1557.

[5] Z. Ren, P. Fertl, Q. Liao, F. Penna, and S. Stańczak, "Street-specific handover optimization for vehicular terminals in future cellular networks," in *Vehicular Technology Conference (VTC Spring).* IEEE, 2013, pp. 1–5.

[6] Q. Liao, F. Penna, S. Stańczak, Z. Ren, and P. Fertl, "Context-aware handover optimization for relay-aided vehicular terminals," in *14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC).* IEEE, 2013, pp. 555–559.

[7] Q. Liao, T. K. Ho, C. Yu, and S. Stańczak, "Future locations and staying time prediction of mobile subscribers over wireless networks," in *The 1st KuVS Workshop on Anticipatory Networks*, 2014.

[8] Q. Liao, S. Valentin, and S. Stańczak, "Channel gain prediction in wireless networks based on spatial-temporal correlation," in *16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC).* IEEE, 2015, pp. 400–404.

[9] Z. Sayeed, Q. Liao, D. Faucher, E. Grinshpun, and S. Sharma, "Cloud analytics for wireless metric prediction-framework and performance," in *8th International Conference on Cloud Computing.* IEEE, 2015, pp. 995–998.

[10] Q. Liao and S. Stańczak, "Network state awareness and proactive anomaly detection in self-organizing networks," in *GLOBECOM International Workshop on Emerging Technologies for 5G Wireless Cellular Networks.* IEEE, 2015.

[11] D. Aziz, H. Bakker, A. Ambrosy, and Q. Liao, "Signaling minimization framework for short data packet transmission in 5G," in *VTC Fall, accepted.* IEEE, 2016.

[12] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD systems," in *GLOBECOM International Workshop on Emerging Technologies for 5G Wireless Cellular Networks, accepted.* IEEE, 2016.

[13] Q. Liao and D. Aziz, "Modeling of mobility-aware RRC state transition for energy-constrained signaling reduction," in *GLOBECOM, accepted.* IEEE, 2016.

[14] A. Giovanidis, Q. Liao, and S. Stańczak, "Measurement-adaptive cellular random access protocols," *Wireless networks, Springer*, vol. 20, no. 6, pp. 1495–1514, 2014.

[15] Q. Liao, D. A. Awan, and S. Stańczak, "Joint optimization of coverage, capacity and load balancing in self-organizing networks," 2016. [Online]. Available: http://arxiv.org/abs/1607.04754

[16] Q. Liao, D. Aziz, and S. Stańczak, "Dynamic joint uplink and downlink optimization for uplink and downlink decoupling-enabled 5G heterogeneous networks," *IEEE Trans. Wireless Communications, submitted*, 2016. [Online]. Available: http://arxiv.org/abs/1607.05459

[17] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer, "Anticipatory networking in future generation mobile networks: A survey," *IEEE Commun. Surveys and Tutorials, submitted*, 2016. [Online]. Available: http://arxiv.org/abs/1606.00191

List of Patents

[18] Q. Liao, F. Penna, S. Stańczak, Z. Ren, and P. Fertl, "Verfahren zur berechnung von übergabeparametern für ein kommunikationsgerät, verfahren zur kommunikation und kommunikationsgerät hierfür," Patent DE102 013 211 130 A1, Jan., 2013.

[19] Q. Liao, E. Grinshpun, and S. Zulfiquar, "System and method for mitigating network congestion using fast congetsion detection in a wireless radio access network (RAN)," Patent US20 160 227 434, July, 2016.

[20] ——, "System and method for controlling an application for classifying an application type using data bearer characteristics," Patent US20 160 226 703, July, 2016.

[21] S. Zulfiquar, Q. Liao, and E. Grinshpun, "System and method for controlling an operation of an application by forecating a smoothed transport block size," Patent US20 160 219 563, July, 2016.

[22] S. Valentin and Q. Liao, "Predicting the state of wireless links based on radio maps," Patent Filing Number DE 15 305 429.1, Mar., 2015.

[23] ——, "Predicting the trajectory of vehicular users based on road maps and mobility history," Patent Filing Number DE 15 305 428.3, Mar., 2015.

Bibliography

[3GPa] 3GPP TR 36.902(V 9.3.0) Self-Configuring and Self-Optimizing Network (SON) use cases and solutions.

[3GPb] 3GPP TS 32.450 (V 12.0.0) Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions.

[3GPc] 3GPP TS 32.451 (V 12.0.0) Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Requirements.

[3GPd] 3GPP TS 32.541 (V 12.0.0) Telecommunication Management; Self-organizing Networks (SON); Self-healing concepts and requirements.

[3GPe] 3GPP TS 36.213 (V 12.5.0) Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures): Requirements.

[3GPf] 3GPP TS 36.300 (V 12.5.0) Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2.

[3GPg] 3GPP TS 36.304 (V 12.4.0) Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode.

[3GPh] 3GPP TS 36.321 (V 12.5.0) Evolved universal terrestrial radio access (E-UTRA); Medium Access Control (MAC) protocol specification.

[3GPi] 3GPP TS 36.331 (V 12.5.0) Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification.

[3GPj] 3GPP TS 36.814 (V 9.0.0) Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects.

[Abr70] N. Abramson. The ALOHA system - Another Alternative for Computer Communications. *Proc. AFIPS Fall Joint Comput. Conf.*, 27, 1970.

[ADF+13] David Astely, Erik Dahlman, Gabor Fodor, Stefan Parkvall, and Joachim Sachs. LTE release 12 and beyond. *Communications Magazine, IEEE*, 51(7):154–160, 2013.

[AFG+SA] M. Amirijoo, P. Frenger, F. Gunnarsson, J. Moe, and K. Zetterberg. On Self-Optimization of the Random Access Procedure in 3G Long Term Evolution. *Proc. IEEE Integrated Network Management-Workshops, 2009.*, pages 177–184, Jun. 2009, New York, NY, USA.

[AHBW11] Y. Al-Harthi, S. Borst, and P. Whiting. Distributed adaptive algorithms for optimal opportunistic medium access. *Mobile Netw Appl (Springer)*, 16, Issue 2:217–230, April 2011.

[AKAKDT11] Amin Abdel Khalek, Lina Al-Kanj, Zaher Dawy, and George Turkiyyah. Optimization models and algorithms for joint uplink/downlink UMTS radio network planning with SIR-based power control. *Vehicular Technology, IEEE Transactions on*, 60(4):1612–1625, 2011.

[All15] NGMN Alliance. 5G white paper. *Next Generation Mobile Networks, White paper*, 2015.

[ALS+08] Mehdi Amirijoo, Remco Litjens, Kathleen Spaey, Martin Döttling, Thomas Jansen, Neil Scully, and Ulrich Türke. Use cases, requirements and assessment criteria for future self-organising radio access networks. In *International Workshop on Self-Organizing Systems*, pages 275–280. Springer, 2008.

[And13] Jeffrey G Andrews. Seven ways that HetNets are a cellular paradigm shift. *Communications Magazine, IEEE*, 51(3):136–144, 2013.

[Asm00] S. Asmussen. *Applied Probability and Queues.* Springer, NY, 2000.

[BAE+15] Federico Boccardi, Jeffrey Andrews, Hisham Elshaer, Mischa Dohler, Stefan Parkvall, Petar Popovski, and Sarabjot Singh. Why to decouple the uplink and downlink in cellular networks and how to do it. *arXiv preprint arXiv:1503.06746*, 2015.

[Bea11] I. Balan and et al. Enhanced weighted performance based handover optimization in LTE. In *Future Network & Mobile Summit*, 2011.

[BEF84] James C Bezdek, Robert Ehrlich, and William Full. FCM: The fuzzy C-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203, 1984.

[BHL$^+$14] Federico Boccardi, Robert W Heath, Aurelie Lozano, Thomas L Marzetta, and Petar Popovski. Five disruptive technology directions for 5G. *Communications Magazine, IEEE*, 52(2):74–80, 2014.

[Bia00] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE JSAC*, 18, issue:3:535–547, Mar. 2000.

[Bir57] Garrett Birkhoff. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, pages 219–227, 1957.

[BJK14] Dinesh Bharadia, Kiran Joshi, and Sachin Katti. Robust full duplex radio link. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 147–148. ACM, 2014.

[BKMS87] Robert R. Boorstyn, Aaron Kershenbaum, Basil Maglaris, and Veli Sahin. Throughput analysis in multihop CSMA packet radio networks. *IEEE Trans. on Communications*, COM-35, no.3:267–274, March 1987.

[BLM05] H.A. Boubacar, S. Lecoeuche, and S. Maouche. Self-adaptive kernel machine: online clustering in RKHS. In *Neural Networks, IJCNN '05*, volume 3, pages 1977 – 1982 vol. 3, July 2005.

[BP94] Abraham Berman and Robert J. Plemmons. *Nonnegative matrices in the mathematical sciences.* SIAM Classics in Applied Mathematics, 1994.

[BP06] Abdelhamid Bouchachia and Witold Pedrycz. Enhancement of fuzzy clustering by mechanisms of partial supervision. *Fuzzy Sets and Systems*, 157(13):1733–1759, 2006.

[BS05] H. Boche and M. Schubert. Duality theory for uplink and downlink multiuser beamforming. In *Smart Antennas–State-of-the-Art*, EURASIP Book Series on Signal Processing and Communications, pages 545–575. Hindawi Publishing Corporation, 2005.

[BS06] H. Boche and M. Schubert. *Smart Antennas: State of the Art*, chapter Duality theory for uplink downlink multiuser beamforming. Hindawi Publishing Corporation, 2006.

[BS08] Holger Boche and Martin Schubert. The structure of general interference functions and applications. *Information Theory, IEEE Transactions on*, 54(11):4980–4990, 2008.

[BSSW05] Holger Boche, Martin Schubert, Slawomir Stanczak, and Marcin Wiczanowski. An axiomatic approach to resource allocation and interference balancing. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, March 18-23 2005.

[BT97] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CB04] Mung Chiang and Jason Bell. Balancing supply and demand of bandwidth in wireless cellular networks: utility maximization over powers and rates. In *INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 4, pages 2800–2811. IEEE, 2004.

[CH12] Xue Chen and Rose Hu. Joint uplink and downlink optimal mobile association in a wireless heterogeneous network. In *Global Communications Conference (GLOBECOM), 2012 IEEE*, pages 4131–4137. IEEE, 2012.

[CJ89] Dah-Ming Chiu and Raj Jain. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. *Computer Networks and ISDN Systems*, 17, North Holland:1–14, 1989.

[CLL$^+$09] Chih-He Chiang, Wanjiun Liao, Tehuang Liu, Iam Kin Chan, and Hsi-Lu Chao. Adaptive downlink and uplink channel split ratio determination for TCP-based best effort traffic in TDD-based WiMax networks. *Selected Areas in Communications, IEEE Journal on*, 27(2):182–190, 2009.

[CLNS13] Gabriela F Ciocarlie, Ulf Lindqvist, Szabolcs Nováczki, and Henning Sanneck. Detecting anomalies in cellular networks using an ensemble method. In *9th CNSM*, pages 171–174. IEEE, 2013.

[CMRWS10] Man Hon Cheung, Amir-Hamed Mohsenian-Rad, Vincent W.S. Wong, and Robert Schober. Random access for elastic and inelastic traffic in WLANs. *IEEE Trans. on Wireless Communications*, 9, no. 6:1861–1866, June 2010.

[CPS14] Renato L. G. Cavalcante, Emmanuel Pollakis, and Slawomir Stanczak. Power estimation in LTE systems with the general framework of standard interference mappings. In *GlobalSIP'14*, pages 818–822. IEEE, 2014.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience New York, NY, USA, 1991.

[dAF04] Guillermo del Angel and Terrence L. Fine. Optimal power and retransmission control policies for random access systems. *IEEE/ACM Trans. on Networking*, 12, no. 6:1156 – 1166, Dec. 2004.

[Dav73] Lee D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783–795, 1973.

[DMP$^+$14] Erik Dahlman, Gunnar Mildh, Stefan Parkvall, Janne Peisa, Joachim Sachs, and Yngve Selén. 5g radio access. *Ericsson Review*, 6:2–7, 2014.

[Dre94] Z. Drezner. Computation of the trivariate normal integral. *Mathematics of Computation*, 63:289–294, 1994.

[DSZ04] G. Dimic, N. D. Sidiropoulos, and R. Zhang. Medium Access Control - Physical Cross-Layer Design. *IEEE Signal Processing Magazine*, 4, Sep. 2004.

[EBDI14a] Hisham Elshaer, Federico Boccardi, Mischa Dohler, and Ralf Irmer. Downlink and uplink decoupling: a disruptive architectural design for 5G networks. In *GLOBECOM'14*, pages 1798–1803. IEEE, 2014.

[EBDI14b] Hisham Elshaer, Federico Boccardi, Mischa Dohler, and Ralf Irmer. Load & backhaul aware decoupled downlink/uplink access in 5G systems. *arXiv preprint arXiv:1410.6680*, 2014.

[Ede62] Michael Edelstein. On fixed and periodic points under contractive mappings. *Journal of the London Mathematical Society*, 1(1):74–79, 1962.

[EH98] A. Ephremides and B. Hajek. Information theory and communication networks: an unconsummated union. *IEEE Trans. on Inf. Theory*, 44, no. 6:2416–2434, Oct. 1998.

[EHDS12] Ahmad M El-Hajj, Zaher Dawy, and Walid Saad. A stable matching game for joint uplink/downlink resource allocation in OFDMA wireless networks. In *Communications (ICC), 2012 IEEE International Conference on*, pages 5354–5359. IEEE, 2012.

[FFFF12] Georg Ferdinand Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, and Ferdinand Georg Frobenius. *Über Matrizen aus nicht negativen Elementen.* Königliche Akademie der Wissenschaften, 1912.

[FKVF13] Albrecht J Fehske, Henrik Klessig, Jens Voigt, and Gerhard 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, 2013.

[GB02] A. Genz and F. Bretz. Comparison of methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics*, 11(4):950–971, 2002.

[GSS] P. Gupta, Y. Sankarasubramaniam, and A. Stolyar. Random-Access Scheduling with Service Differentiation in Wireless Networks. *INFOCOM 2005*, 3:1815 – 1825.

[GWB08] A. Giovanidis, G. Wunder, and H. Boche. A short-term throughput measure for communications using ARQ protocols. *Proc. 7th ITG Conf. on SCC*, 2008.

[Han81] D.J. Hand. *Discrimination and Classification.* Wieley New York, 1981.

[HHY$^+$12] Shiwen He, Yongming Huang, Luxi Yang, Arumugam Nallanathan, and Pingxiang Liu. A multi-cell beamforming design by uplink-downlink max-min sinr duality. *Wireless Communications, IEEE Transactions on*, 11(8):2858–2867, 2012.

[Hoe65] W. Hoeffding. Asymptotically optimal test for multinomial distributions. *The Annals of Mathematical Statistics*, 36:369–401, 1965.

[HRGD05] M. Heusse, F. Rousseau, R. Guillier, and A. Duda. Idle sense: An optimal access method for high throughput and fairness in rate diverse Wireless LANs. *Proc. ACM SIGCOMM'05*, Philadelphia, Pennsylvania, USA, Aug. 21-26 2005.

[HSS12] Seppo Hämäläinen, Henning Sanneck, and Cinzia Sartori. *LTE self-organising networks (SON): network management automation for operational efficiency.* John Wiley & Sons, 2012.

[HTR13] Yichao Huang, Chee Wei Tan, and Bhaskar D Rao. Joint beamforming and power control in coordinated multicell: Max-min duality, effective network and large system transition. *Wireless Communications, IEEE Transactions on*, 12(6):2730–2742, 2013.

[HTZ$^+$14] Y-W Peter Hong, Chee Wei Tan, Liang Zheng, Cheng-Lin Hsieh, and Chia-Han Lee. A unified framework for wireless max-min utility optimization with general monotonic constraints. In *INFOCOM, 2014 Proceedings IEEE*, pages 2076–2084. IEEE, 2014.

[HvL82] B. Hajek and T. van Loon. Decentralized dynamic control of a multiaccess broadcast channel. *IEEE Trans. on Automatic Control*, AC-27, no. 3:559–569, June 1982.

[HYLSon] C Ho, Di Yuan, Lei Lei, and Sumei Sun. On power and load coupling in cellular networks for energy optimization. *IEEE Trans. Wireless Commun.*, 2014, accepted for publication.

[HYS14] Chin Keong Ho, Di Yuan, and Sumei Sun. Data offloading in load coupled networks: A utility maximization framework. *Wireless Communications, IEEE Transactions on*, 13(4):1921–1931, 2014.

[Jea10] T. Jansen and et al. Handover parameter optimization in LTE self-organizing networks. In *Proceedings of the IEEE 72nd VTC 2010-Fall*, 2010.

[Jea11] T. Jansen and et al. Weighted performance based handover parameter optimization in LTE. In *Proceedings of the IEEE 73rd VTC 2011-Spring*, 2011.

[Jol02] Ian Jolliffe. *Principal component analysis.* Wiley Online Library, 2002.

[Kea11] K. Kitagawa and et al. A handover optimization algorithm with mobility robustness for LTE systems. In *Proceedings of the IEEE 22nd International Symposium on PIMRC*, pages 1647 – 1651, 2011.

[Kea12] Henrik Klessig and et al. Improving coverage and load conditions through joint adaptation of antenna tilts and cell selection rules in mobile networks. In *ISWCS*, pages 21–25. IEEE, 2012.

[KG10] Ralf Kreher and Karsten Gaenger. *LTE signaling: troubleshooting and optimization.* John Wiley & Sons, 2010.

[KL75] Leonard Kleinrock and Simon S. Lam. Packet switching in a multiaccess broadcast channel: Performance evaluation. *IEEE Trans. on Communications*, COM-23, no.4:410–423, April 1975.

[KL09] Sungyeon Kim and Jang-Won Lee. Joint resource allocation for uplink and downlink in wireless networks: A case study with user-level utility functions. In *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, pages 1–5. IEEE, 2009.

[KP82] Elon Kohlberg and John W Pratt. The contraction mapping approach to the Perron-Frobenius theory: why Hilbert's metric? *Mathematics of Operations Research*, 7(2):198–210, 1982.

[KRC10] Samian Kaur, Alexander Reznik, and Douglas R Castor. Method and apparatus for a multi-radio access technology layer for splitting downlink-uplink over different radio access technologies, August 20, 2010. US Patent App. 12/859,863.

[Kre89] Erwin Kreyszig. *Introductory functional analysis with applications*, volume 81. wiley New York, 1989.

[Kus64] H. Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Eng.*, 86:97–106, 1964.

[LCCZ15] Dantong Liu, Yue Chen, Kok Keong Chai, and Tiankui Zhang. Backhaul aware joint uplink and downlink user association for delay-power trade-offs in HetNets with hybrid energy sources. *Transactions on Emerging Telecommunications Technologies*, 2015.

[Lea10] Andreas Lobinger and et al. Load balancing in downlink LTE self-optimizing networks. In *Proc. 71st IEEE VTC'10-Spring*, Taipei, Taiwan, May 2010.

[Lea11] L. Luan and et al. Handover parameter optimization of LTE system in variational velocity environment. In *Proceedings of the IET International Conference on ICCTA*, pages 395 – 399, 2011.

[LK75] Simon S. Lam and Leonard Kleinrock. Packet switching in a multiaccess broadcast channel: Dynamic control procedures. *IEEE Trans. on Communications*, COM-23, no. 9:891–904, Sept. 1975.

[LKC$^+$12] Wonbo Lee, Dongmyoung Kim, Seunghyun Choi, Kyung-Joon Park, Sunghyun Choi, and Ki-Young Han. Self-optimization of RACH power considering multi-cell outage in 3GPP LTE systems. *Proc. of the 75th VTC Spring*, 2012.

[LN12] Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory.* Number 189. Cambridge University Press, 2012.

[LPGC12] David Lopez-Perez, Ismail Guvenc, and Xiaoli Chu. Mobility management challenges in 3GPP heterogeneous networks. *Communications Magazine, IEEE*, 50(12):70–78, 2012.

[LSWL04] Kin Kwong Leung, Chi Wan Sung, Wing Shing Wong, and Tat-Ming Lok. Convergence theorem for a general class of power-control algorithms. *Communications, IEEE Transactions on*, 52(9):1566–1574, 2004.

[LUE03] J Luo, S Ulukus, and A Ephremides. Probability one convergence in joint stochastic power control and blind mmse interference suppression. *Positivity*, 1:0, 2003.

[LUE05] Jie Luo, Sennur Ulukus, and Anthony Ephremides. Standard and quasi-standard stochastic power control algorithms. *Information Theory, IEEE Transactions on*, 51(7):2612–2624, 2005.

[LYP$^+$09] J. Liu, Y. Yi, A. Proutiere, M. Chiang, and H.V. Poor. Towards utility-optimal random access without message passing. *Wirel. Commun. Mob. Comput. (published online)*, 10(1):1–12, 2009.

[Mey00] Carl D Meyer. *Matrix analysis and applied linear algebra.* Siam, 2000.

[MNK$^+$07] Preben Mogensen, Wei Na, István Z Kovács, Frank Frederiksen, Akhilesh Pokhariyal, Klaus I Pedersen, Troels Kolding, Klaus Hugl, and Markku Kuusela. Lte capacity compared to the shannon bound. In *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, pages 1234–1238. IEEE, 2007.

[MOM04] MOMENTUM. Models and simulations for network planning and control of UMTS. `http://momentum.zib.de`, 2004.

[NH06] Tien-Dzung Nguyen and Youngnam Han. A proportional fairness algorithm with QoS provision in downlink OFDMA systems. *IEEE Communications Letters*, 10(11):760–762, 2006.

[NMR03] M. J. Neely, E. Modiano, and C. E. Rohrs. Power Allocation and Routing in Multibeam Satellites with Time-Varying Channels. *IEEE/ACM Trans. on Networking*, 11(1), Feb. 2003.

[NMR05] M. J. Neely, E. Modiano, and C. E. Rohrs. Dynamic Power Allocation and Routing for Time-Varying Wireless Networks. *IEEE JSAC*, 23(1), Jan 2005.

[Nuz07] Carl J Nuzman. Contraction approach to power control, with non-monotonic applications. In *GLOBECOM'07*, pages 5283–5287. IEEE, 2007.

[OG12] Olav Osterbo and Ole Grondalen. Benefits of Self-Organizing Networks (SON) for mobile operators. *Hindawi Publishing Corporation, Journal of Computer Networks and Communications,*, 2012.

[PTVF96] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes in C*, volume 2. Citeseer, 1996.

[Put05] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley & Sons, 2005.

[PVP$^+$07] I. Papapanagiotou, J.S. Vardakas, G.S. Paschos, M.D. Logothetis, and S.A. Kotsopoulos. Performance evaluation of IEEE 802.11e based on ON-OFF traffic model. *Proc. of the 3rd international conference on Mobile multimedia communications (MobiMedia)*, 2007.

[PYC08] Alexandre Proutiere, Yung Yi, and Mung Chiang. Throughput of random access without message passing. *Proc. 42nd Annual Conference on Information Sciences and Systems, (CISS).*, 2008.

[Qua08] Qualcomm. Range expansion for efficient support of heterogeneous networks. *3GPP TSG-RAN WG1 R1-083813*, 2008.

[Reaar] Cavalcante. R. and et al. Toward energy-efficient 5G wireless communication technologies. *Signal Processing Mag.*, to appear.

[RKC10] Rouzbeh Razavi, Siegfried Klein, and Holger Claussen. Self-optimization of capacity and coverage in LTE networks using a fuzzy reinforcement learning approach. In *PIMRC, 2010 IEEE*, pages 1865–1870, 2010.

[RW06] C. E. Rasmussen and C. K. I. Williams. *Gaussian Process for Machine Learning.* MIT press, Cambridge, MA, 2006.

[SAR$^+$10] Mohamed Salem, Abdulkareem Adinoyi, Mahmudur Rahman, Halim Yanikomeroglu, David Falconer, and Young-Doo Kim. Fairness-aware radio resource management in downlink OFDMA cellular relay networks. *Wireless Communications, IEEE Transactions on*, 9(5):1628–1639, 2010.

[SB05] Martin Schubert and Holger Boche. Iterative multiuser uplink and downlink beamforming under SINR constraints. *Signal Processing, IEEE Transactions on*, 53(7):2324–2334, 2005.

[SBS05] Martin Schubert, Holger Boche, and Slawomir Stanczak. Joint power control and multiuser receiver design–fairness issues and cross-layer optimization. In *Proc. IST Summit 2005*, Dresden, Germany, June 19-23 2005.

[SEP$^+$14] Katerina Smiljkovikj, Hisham Elshaer, Petar Popovski, Federico Boccardi, Mischa Dohler, Liljana Gavrilovska, and Ralf Irmer. Capacity analysis of decoupled downlink and uplink access in 5G heterogeneous systems. *arXiv preprint arXiv:1410.7270*, 2014.

[SGK06] Gaurav Sharma, Ayalvadi Ganesh, and Peter Key. Performance analysis of contention based medium access control protocols. *INFOCOM 2006*, pages 1–12, Apr. 2006.

[SHWL07] Guan-Ming Su, Zhu Han, Min Wu, and KJ Liu. Joint uplink and downlink optimization for real-time multiuser video streaming over WLANs. *Selected Topics in Signal Processing, IEEE Journal of*, 1(2):280–294, 2007.

[Sma80] David R Smart. *Fixed point theorems*. Number 66. CUP Archive, 1980.

[SOC08a] SOCRATES, European Research Project. Review of use cases and framework. *Deliverable 2.5, EU-Project SOCRATES (INFSO-ICT-216284)*, 2008.

[SOC08b] SOCRATES, European Research Project. Self-optimisation and self-configuration in wireless networks. `http://www.fp7-socrates.eu`, 2008.

[SOC09] SOCRATES, European Research Project. Review of use cases and framework ii. *Deliverable 2.6, EU-Project SOCRATES (INFSO-ICT-216284)*, 2009.

[SPG15] Katerina Smiljkovikj, Petar Popovski, and Liljana Gavrilovska. Analysis of the decoupled access for downlink and uplink in wireless heterogeneous networks. *Wireless Communications Letters, IEEE*, 4(2):173–176, 2015.

[SS98] Alex J Smola and Bernhard Schölkopf. *Learning with kernels.* Citeseer, 1998.

[SS10] Jörg Sommer and Joachim Scharf. IKR simulation library. In *Modeling and Tools for Network Simulation*, pages 61–68. Springer, 2010.

[SVY06] Iana Siomina, Peter Varbrand, and Di Yuan. Automated optimization of service coverage and base station antenna configuration in UMTS networks. *Wireless Communications, IEEE*, 13(6):16–25, 2006.

[SWB09] Slawomir Stanczak, Marcin Wiczanowski, and Holger Boche. *Fundamentals of resource allocation in wireless networks: theory and algorithms*, volume 3. Springer, 2009.

[SWMG08] Aimin Sang, Xiaodong Wang, Mohammad Madihian, and Richard D Gitlin. Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems. *Wireless Networks*, 14(1):103–120, 2008.

[SWZZ10] D. Su, X. Wen, H. Zhang, and W. Zheng. A self-optimizing mobility management scheme based on cell ID information in high velocity environment. In *Proceedings of the 2nd International Conference on ICCNT*, pages 285 – 288, 2010.

[SY12] Iana Siomina and Di Yuan. Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization. In *Communications (ICC), 2012 IEEE International Conference on*, pages 1357–1361. IEEE, 2012.

[SZA14] Sarabjot Singh, Xinchen Zhang, and Jeffrey G. Andrews. Joint rate and SINR coverage analysis for decoupled uplink-downlink biased cell associations in HetNets. *CoRR*, abs/1412.1898, 2014.

[SZC07] W. Song, W. Zhuang, and Yu Cheng. Load balancing for cellular/WLAN integrated networks. *IEEE Network*, 21:27–33, January 2007.

[TE92] L. Tassiulas and A. Ephremides. Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks. *IEEE trans. on Automatic Control*, 37(12), Dec 1992.

[TE93] Leandros Tassiulas and A. Ephremides. Dynamic Server Allocation to Parallel Queues with Randomly Varying Connectivity. *IEEE Trans. on Inf. Theory*, 39(2), March 1993.

[Tel15] Telecom Italia. Big data challenge 2015. 2015.

[TK85] Hideaki Takagi and Leonard Kleinrock. Throughput analysis for persistent CDMA systems. *IEEE Trans. on Communications*, COM-33, no. 7:627–638, July 1985.

[TLJ10] Marina Thottan, Guanglei Liu, and Chuanyi Ji. Anomaly detection approaches for communication networks. In *Algorithms for Next Generation Networks*, pages 239–261. Springer, 2010.

[TZM01] L. Tong, Q. Zhao, and G. Mergen. Multipacket reception in random access wireless networks: From signal processing to optimal medium access control. *IEEE Comm. Magazine*, pages 108–112, Nov. 2001.

[UY98] S. Ulukus and R. Yates. Stochastic power control for cellular radio systems. *IEEE Trans. Commun.*, 46(6):784–798, 1998.

[VM14] Richard Von Mises. *Mathematical theory of probability and statistics.* Academic Press, 2014.

[VS11] Nikola Vucic and Martin Schubert. Fixed point iteration for max-min sir balancing with general interference functions. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 3456–3459. IEEE, 2011.

[Ž85] A. Žilinskas. Axiomatic characterization of a global optimization algorithm and investigation of its search strategy. *Operations Research Letters*, 4(1):35–39, 1985.

[Ž12] A. Žilinskas. A statistical model-based algorithm for black box multi-objective optimisation. *International Journal of Systems Science, accepted*, 2012.

[Wel16] Marcus K Weldon. *The Future X Network: A Bell Labs Perspective.* Crc Press, 2016.

[Wil91] D. Williams. *Probability with Martingales.* Cambridge, 1991.

[Yat95] R. D. Yates. A framework for uplink power control in cellular radio systems. *IEEE J. Select. Areas Commun.*, 13(7):1341–1347, September 1995.

[YH95] R. D. Yates and C. Y. Huang. Integrated power control and base station assignment. *IEEE Trans. Veh. Technol.*, 44(3):638–644, August 1995.

[YHH11] Osman N. C. Yilmaz, Jyri Hamalainen, and Seppo Hamalainen. Self-optimization of Random Access Channel in 3GPP LTE. *Proc. 7th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2011.

[Ziv88] Jacob Ziv. On classification with empirically observed statistics and universal data compression. *IEEE Transactions on Information Theory*, 34(2):278–286, 1988.

[ZL78] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

[ZM15] Xi Zhang and Jia Ming. Filtered-OFDM enabler for flexible waveform in the 5th generation cellular networks. In *Global Communications Conference (GLOBECOM), 2015 IEEE*, pages 1–6. IEEE, 2015.

[ZRC^{+}08] Yu Zhou, Yanxia Rong, H-A Choi, Jae-Hoon Kim, Jung-Kyo Sohn, and Hyeong In Choi. Utility-based load balancing in WLAN/UMTS internetworking systems. In *Radio and Wireless Symposium*, pages 587–590. IEEE, 2008.

[ZT14] Liang Zheng and Chee Wei Tan. Optimal algorithms in wireless utility maximization: Proportional fairness decomposition and nonlinear Perron-Frobenius theory framework. *Wireless Communications, IEEE Transactions on*, 13(4):2086–2095, 2014.